The Impact of Resource Reservation for Real-Time Internet Services

Henning Schulzrinne
Dept. of Computer Science
Columbia University
schulzrinne@cs.columbia.edu

October 22, 1996

Abstract

Real-time services and changing consumer Internet usage patterns may force the introduction of resource reservation as well as usage-based charging. The motivations and implications are discussed.

1 Introduction

Real-time services are expected to contribute a significant fraction of the future workload of the Internet, in particular, if the Internet is to replace some fraction of the current telephone and radio/television infrastructure. Due to changes particularly in the way consumers are likely to access the Internet, some form of usage-based charges are likely, at least for services with any type of QOS commitment. We discuss both technical implications of using resource reservations where charging is involved, as well as some of the economic issues.

2 Characteristics of Real-Time Services

We can roughly distinguish four traffic classes in the Internet today, listed here in increasing order of the demands they make on the network:

Bulk data: Bulk data is comprised mainly of network news, email, unattended bulk data transfer such as software downloads and remote backups and, to a lesser degree, off-line web preloading (Freeloader and similar applications). Delays and packet loss are not important, as long as a minimum throughput, averaged over possibly hours, is maintained. This is an example of an almost perfectly *elastic* application.

Interactive data: Web browsing, IRC and telnet fall into the category of interactive data, with minimal throughput demands and some sensitivity to delay. Since many web pages are designed to be displayed in reasonable time for those using 14.4 or 28.8 kb/s modem, throughput below around 10 kb/s will cause annoying download delays and aborted web page loads. Interactive data applications currently dominate Internet traffic.

Media retrieval: Delivery of media streams, mainly audio at this point, but also low-frame-rate video such as CuSeeMe, requires throughput that remains reasonably constant and above about 14 kb/s for extended periods of time. Communication-quality speech can be supported with audio codecs running at 5.6 and 6.3 kb/s (G.723.1), but packet header overhead increases that by about 30%. FM-quality audio pushes the bandwidth requirement to at least 56 kb/s. Packet losses are less significant since the playout delay can be long enough to permit retransmission.

Interactive multimedia: Interactive communications such as Internet telephony requires about the same amount of bandwidth as media retrieval, but above a delay of 150 ms and losses of about 5%, sound quality is no longer commercially viable.

	sensitivity	cost/month ($)		
item		total	current	continuous 28.8
T1 line with major carrier	traffic	1,500	0.75	29
T3 line with major carrier	traffic	65,000	0.75	29
business line access charge	connect time	18	1.80	12
modem depreciation	connect time	11	1.10	7
terminal server depreciation	connect time	3	0.30	2
router depreciation	traffic	277	0.03	0
sum			3.98	50

Table 1: Monthly costs for consumer internet access; equipment costs are based on a depreciation of 36 months; with per-customer costs for both current multiplexing and continuous usage of 16 h/day, for 28.8 kb/s modems

We will refer to retrieval and interactive multimedia applications as *continuous media* services. We note that the peak bandwidth that is useable for web retrievals and for continuous multimedia applications is fairly similar, however, clearly the average bandwidth is much higher for the latter. The lower burstiness of packet audio and video, however, allows for higher network utilization for the same amount of buffering.

While congestion control for point-to-point bulk and interactive data are well established (if only recently formally written down [2]), the control of continuous-media applications is an open problem, with some initial experiments [3, 4, 5], but no established algorithms. Other traffic, such as DNS, SNMP or routing, contributes a significant fraction of traffic, but cannot really be flow-controlled and requires sufficiently low packet loss, as it is mainly UDP-based.

3 Pricing

3.1 Pricing for Commercial Internet Access

Commercial or institutional access is characterized by a LAN feeding one or more access connections to the global Internet. This mode applies to universities and corporations, whether they resell connectivity to others or not. (Universities and corporations have become large ISPs with a captive customer base, with their permanently connected users foreshadowing the future of the consumer market.) These access lines are usually charged on the basis of average utilization, at least for T3 and higher access speeds, i.e., those speeds close to the backbone speed. Thus, reservations and real-time services can be easily reflected back into charging, without any new infrastructure.

The operation of the Internet depends on the cooperation of transport protocol implementations all sharing bandwidth using the same algorithm. Bandwidth sharing using TCP or any loss-based mechanism is not likely to be fair by any criterion, as one's share of bandwidth depends on the number of hops, distance and any number of other factors. However, since these factors affect everybody more or less equally, with some penalty for long-distance traffic, there is the perception of long-term fairness. Fortunately, two factors help to ensure this: first, protocol stack vendors have little incentive to make their implementations more aggressive, at least if they are likely to be used by a large number of customers. Secondly, administrative controls at the institutional level tend to discourage abuse.

3.2 Pricing for Consumer Internet Access

Table 1 provides a rough overview of some of the costs of providing consumer Internet access, excluding billing and customer service.

The current hourly or monthly pricing model is based on access capacity, where the ISP assumes that a particular user generates or receives, on average, only a very small fraction of the modem bandwidth and is only on-line a small fraction of the day[1].

Connect time and bytes/month are both increasing. Flat-rate pricing and second phone lines allow almost continuous connectivity, which is further encouraged by the chance of a busy modem line and applications like Internet telephony. The average number of bytes per month is increasing due to a number of factors, including continuous-media applications that do not require interaction with the computer.

[1]The current guidelines for ISPs [1] suggest a ratio of 200 28.8 kb/s lines for a T1 connection, with a ratio of 10:1 subscribers to modem lines for a larger ISP. This corresponds roughly to the average of 20 hours that a web user is on-line per week.

If home computers are more or less permanently connected to the ISP, at least during the evening hours, but receiving packets only intermittently ("standby"), only the costs labelled "connect time" in Table 1 would be affected. However, phone companies are already concerned about call holding times way above phone switch design points and are likely to seek compensation [**?**]. Costs labeled "traffic" would increase if the personal computer became a replacement for radio stations ("Internet radio"). Billing and customer service are major cost components, but not likely to be affected by these changes.

4 Resource Reservation to the Rescue?

Based on the discussion in Section 3.2, one could argue that resource reservation and usage-based charging are primarily needed at the edges of the network, i.e., for customers of an ISP or users in a corporate or institutional setting sharing the access bandwidth, since that is where the sum of possible resource demands greatly exceeds the bandwidth of the Internet access. This would allow to offer lower pricing to the standby users, while accomodating the "Internet radio" subscribers.

For the corporate environment, this greatly simplifies matters, since there are usually either administrative ("no reservations for undergraduates") or charge-back mechanisms ("please include your grant number in your request") in place.

Whether one subscribes to this view, resource reservation poses a number of open problems:

Security: RSVP relies on the IPSEC security model. To simplify authentication, it would probably be easiest to have the ingress network provider act as the CA, so that there is no need to check an on-line customer database for each reservation request.

Scaling: Large-scale experiments with RSVP are only just beginning. Due to its soft-state nature, RSVP is more resource intensive than hard-state methods. By default, a PATH and RESV message is generated every 30 seconds. For all but extremely low bit-rate services, the overhead in traffic, computation and state should be minimal, with a packet arrival rate of 48 packets/second for a T3 link booked solid with voice flows (720 conversations). The refresh rate can be reduced if routers detect routing changes and implement local repair.

Advance reservation: Currently, RSVP has no notion of advance reservation. If reservations have a significant chance of failing, advance reservations will be necessary – one would not want to set up a teleconference days in advance, just to have it fail at its start time due to lack of network resources.

Provider selection: In the current PSTN, a subscriber can select the long-distance carrier on a subscription or call-by-call basis. In the near future, selection from several local carriers may also be available, although it is not clear at what timescale. In the Internet, currently the choice of a local ISP determines the choice of a (set of) long-distance carriers. True provider selection would require that the interdomain routing tables, with cost as metric, for different traffic and QOS classes are made available to the customer. However, it seems unlikely that Internet services will evolve to a scenario of "let's take MCI from New York to Chicago, AT&T from Chicago to Reno and LDI from Reno to LA."

Charging and billing: The current version of RSVP has no provision for carrying charging, billing and price sensitivity information. However, like all IP packets, RSVP RESV messages could be source-routed for provider-selection.

Hot-potato routing: If charges are not distance-sensitive, there is a strong incentive to pass reservations and traffic to the nearest other provider that advertises reachability for the destination, even though network loading and distance may be unfavorable.

Charging has to have both a reservation and traffic component, i.e., even if a flow does not use a reservation, some charge has to accumulate. Otherwise, reservations could be easily used for denial-of-service attacks. The level of this reservation charge depends on the fraction of reserved traffic using the network. If it is high, the carrier may suffer true opportunity costs by having to reject a reservation request by some other subscriber.

For unicast, RSVP could implement both receiver- and sender-based charges, simply by putting charging authorizations into either the RESV or PATH messages, respectively. To limit complexity, it seems sensible to only have the ingress point of a reservation request create billing records, rather than each router along the path.

For multicast, sender-based charges are more difficult to implement if they are to depend on the number of receivers, as no single point in the network knows the number of receivers. Simply counting the number of receivers is also not a good reflection of the cost to the network, as the number of edges spanned by the tree may vary widely. If the network charges for each link traversed by the multicast tree, random nodes of the network have to be prevented from sending RESV messages up the tree, for example, by having the PATH messages set up a password, encrypted with the provider's public key, that all RESV messages have to "know".

There are several possible models how the use of resource reservation, based on RSVP or successor protocols, might evolve:

Internet virtual private networks (IVPN): Corporations and research universities might use the Internet much like they use private lines or frame relay networks today, reserving a minimum amount of capacity between two or more sites, probably using the *controlled load* service within the Internet integrated services framework. The controlled load service makes no guarantees about delay, but offers the perception of an unloaded network of a given capacity to the user. These IVPNs offer the advantage of higher peak capacity, while only paying close to the average utilization. As long as most traffic is not of this IVPN type or the other reserved classes below, utilization can remain high and costs low. The IVPNs may be used for PBX interconnection, but also to ensure that data applications have predictable performance. This mode, probably the first to be deployed, does not require elaborate authentication and billing systems, since providers rather than customers probably provision these IVPNs.

Use for high-bandwidth applications: Here, only high-bandwidth applications, i.e., those using a significant fraction of shared network resources, would apply resource reservation, with everybody else being expected to operate in back-off (adaptive) mode. It is likely that Internet telephony would not use reservations in most parts of the Internet, for example, since if it is to work at all, it has to be designed to be a mass service. Just like the current phone system, the likelihood of a resource request being turned down has to be close to zero.

Per-application signaling: In this scenario, each user-level application sets up reservations. Since reserved bandwidth will have to cost more than best-effort bandwidth, there is an incentive for applications to use best effort services unless the network is congested. There is also the opportunity to combine best effort and reserved services in parallel: an application would reserve the minimal amount of bandwidth necessary for communication; any improved quality of service would be transmitted "at risk", with bandwidth reduced when network congestion is detected. To ensure some notion of long-term fairness, these back-off algorithms will have to be generally agreed upon.[2]

5 Market Issues

Despite predictions to the contrary, it is unlikely that flat-rate charging will disappear from the Internet. If anything, other competitive communications services are moving into that direction: witness the disappearance of distance as a factor of long-distance telephony [6] or the popularity of flat-rate ("10c/minute anywhere") long-distance services, with very little dependence on time-of-day.

Even cellular telephony billing, currently probably the consumer service with the most arcane pricing structure, is moving into that direction, with the intent that almost all calls in a month are covered by the built-in allowance of the calling plan.

Any transition from a pure flat-rate or hourly model to volume- and QOS-sensitive charging also greatly complicates the billing infrastructure and raises the likelihood of billing disputes requiring human intervention. Also, flat-rate billing greatly simplifies internal chargeback accounting.

As pointed out earlier, 56k, ISDN and cable modems in particular will make the current consumer charging model difficult to sustain. A possible model might be a per-byte charge for guaranteed traffic, with adaptive services (with a low long-term average) covered by the monthly charge. In addition, "radio/TV" services retrieved from a server local to the provider would not be subject to additional charges – similar to the cable or satellite TV model.

For QOS charging, it is difficult to see how anything other than bandwidth-based charges can be sustained, except possibly for a high-delay "radio/TV" charge and a, somewhat higher, low-delay "telephone" charge.

[2] One such rule might be the "no impact" rule, i.e., that losses with and without an individual application should be the same.

References

[1] D. Dennis, “Internet provider resources,” web page, Amazing, 1996.

[2] W. Stevens, “Tcp slow start, congestion avoidance, fast retransmit, and fast recovery algorithms,” Internet Draft, Internet Engineering Task Force, Mar. 1996. Work in progress.

[3] I. Busse, B. Deffner, and H. Schulzrinne, “Dynamic QoS control of multimedia applications based on RTP,” *Computer Communications*, Jan. 1996.

[4] J.-C. Bolot and A. V. Garcia, “Control mechanisms for packet audio in the internet,” in *Proceedings of the Conference on Computer Communications (IEEE Infocom)*, (San Fransisco, California), Mar. 1996.

[5] J.-C. Bolot, T. Turletti, and I. Wakeman, “Scalable feedback control for multicast video distribution in the internet,” in *SIGCOMM Symposium on Communications Architectures and Protocols*, (London, England), pp. 58–67, ACM, Aug. 1994.

[6] Anonymous, “The death of distance – a survey of telecommunications,” *The Economist*, Sept. 1995.