How Can Routers Help Internet Economics?

John M. Schnizlein
Cisco Systems
170 West Tasman Drive
San Jose, CA95134-1706 USA
+1 301 567 7126
john.schnizlein@cisco.com

1. ABSTRACT

Statistical sharing enables remarkable network efficiency in internets, compared to circuit-switched networks, but complicates economic efficiency associating traffic priority with users' valuations. How can the network (routers) differentiate service so that it justifies differential pricing? One approach is integrating internets with reservations which can be billed, like calls, based on duration and capacity. A more recent approach is differential treatment of packets marked for different types of service. Peak traffic rates over a negotiated time period can be either measured or controlled. These peak rates aggregate, with some degree of asynchrony, to the capacity limit of the network. At this limit, routers protect the network from congestive collapse by dropping additional traffic based on the type of service marks. Although modern TCP end stations respond by reducing their network loads to a fair share, non-responsive applications threaten the integrity of the Internet. Can billing for congestion effectively control this threat?

1.1 Keywords

Network, economics, differentiated, billing, congestion

2. INTRODUCTION

Statistical sharing in packet-based networks, especially internets, has produced unprecedented network efficiency. This efficiency has not been automatic, but depends on the behavior of end systems. The Internet effectively collapsed in 1986 before TCP (transmission control protocol) was redesigned to avoid congestion [13]. This efficiency, and the rapidly growing sharing of resources on connected computers, made the Internet successful beyond the research community.

In contrast, the Internet has not accomplished economic efficiency. The combination of remarkable network efficiency with research funding rapid deployment made economic efficiency so unimportant that it still offers essentially a single grade of best-effort service. Although the service is often quite good, it is worst when most in demand. Efforts to build internets with better service than the global Internet have focussed on private (e.g. Frame Relay) networks that serve smaller communities.

Economists have warned that the current problem of competition suppressing prices, and revenue required for expansion, could result from economic in-efficiency. An economically efficient system generates revenue for expansion by matching prices to users' valuation of the service [16, 24].

However, economic optimality may not be as important as finding pricing structures that can be deployed [22]. One practical realization is that pricing mechanisms will be concentrated at the edges of routing domains. Researchers now advocate experimental implementation of various pricing policies in place of economic optimality research. Uniform pricing policies are unlikely when multiple competing routing domains are responsible for carrying traffic between two endpoints. "In the context of this edge pricing paradigm, usage-based pricing and flat pricing are not radically different but instead both reside along the single continuum of usage-constraining policies." The kind of usage constraint we should seek is, as [10] applied to queuing, "Not only does this allow the current generation of flow control algorithms to function more effectively, but it creates an environment where users are rewarded for devising more sophisticated and responsive algorithms."

This paper explores how the routers that compose the Internet can help, and asks what mechanisms should be provided. This depends as much on what is practical to implement in routers throughout the network as on economic theory. We review developments up to this point and seek direction on how ideas discussed in Internet economics research could be applied.

3. RESERVATIONS

Networks based on circuit leasing or circuit switching have demonstrated economic efficiency. Because the capacity of the circuit is dedicated to the user, either for a contract period or during the call for switched service, the quality of service can be specified precisely. Both competition and variable pricing exist in telephone networks. Although complex regulatory issues are involved in the transition of the telephone market from a regulated monopoly to a competitive market, circuit-switched communication does not have the problems of a single (best-effort) service at various access rates and competition mostly on price.

The quality commitment of a dedicated circuit can be provided without the waste of unused capacity by supporting reserved capacity in the Internet. Billing for reservations in the same terms as calls supports users' need for higher quality and the revenue to provide it. The Integrated Service [4] extension of the Internet was designed to integrate guaranteed and predictive service quality reservations with the best-effort service of the Internet. The protocol supporting reservations in the Integrated Services Internet is RSVP [26].

Because a reservation commits resources, admission control for reservations is the logical place to handle commitment to pay for those resources. The policy for admission control is in a policy server separated from the routing functions. Routers would request a policy decision from the policy server prior to allocating a requested reservation. The current working draft for the interaction of routers and policy servers is Common Open Policy Service (COPS) [3]. Strong security is necessary both between routers and policy servers and between policy servers and the billing system that connects policies to economics because their interaction implies financial transactions.

Keeping the state for each reservation in all the intervening routers is expensive. The potential number of reservations is larger than all pairs of communicating computers because reservations can be specified for individual flows from any application on any computer to another. In the core of the Internet, routers take the place occupied by simple but fast multiplexers in the circuit-based network. The feasibility of supporting the potentially huge number of reservations aggregated near the center of the Internet is questioned [18]. However, enabling users to make and pay for committed resources fits the demands of economic efficiency, and RSVP is being included in user (client) software.

Billing for reservations would be essentially like billing for phone calls, based on duration and capacity, with some interesting variations. Receivers establish reservations for information flowing (one way) from sources they specify, in contrast to a caller establishing a two-way channel in traditional telephony. Internet reservations can specify more detail than just a standard increment of bandwidth, in contrast with just multiples of the basic (DS-0) telephone channel. What is similar is that charges are based on usage as with telephone calls. This usage-based billing is not the dominant tradition in Internet access.

4. ACCESS RATE

The traditional economic model for Internet pricing has been charges based on access rate. Since the maximum rate at which a user can load an internet is limited by the access circuit, the access rate determines the worst-case provisioning requirements for circuits carrying traffic aggregated among subscribers. The cost of these circuits for aggregated traffic is one of the largest costs borne by service providers. Provisioning for the worst case is neither economically feasible nor necessary because of the statistical sharing of trunk capacity. But at least some congestion will occur where traffic aggregated from subscriber circuits exceeds trunk capacity.

As seen in Figure 1, recent Internet access prices [23] show significant economies of scale. The approximately 800-times bandwidth ratio from the smallest standard (DS-0) circuit to the largest (DS-3) only costs 50 times as much. Because the standard units of capacity, fixed by the existing telephone multiplexing hierarchy, increase by factors of 24 and 30, large increases in capacity and cost are required to obtain the scale economies. Greater variation in price at higher capacity levels reflects pricing alternatives to simple access rate already available from service providers. The most common alternative is usage-based billing in addition to a lower monthly charge.

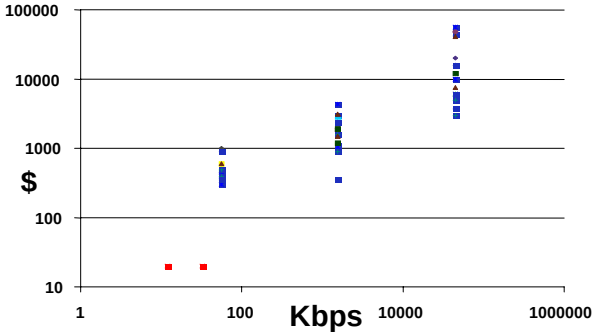

Figure 1. Price (US dollars) per month for access rates at 56 Kbps, 1.54 Mbps, 45 Mbps, and dial-in at 12 – 33 Kbps.

It is worth noting how the economy of scale for dedicated access circuits extrapolates to the common $20 per month dial access price. Extrapolating the price/capacity pattern above to 20 – 28 Kbps yields a monthly price around $200 – 500, which is 20 – 25 times the typical dial subscription price. This suggests an over-subscription factor of at least 20 for dial access. This implies 20 subscribers contending for each available access modem, unless the cost per capacity is significantly less for dial ports than dedicated ports. Since dial access is circuit-switched, billing for the duration of access calls could fairly allocate the over-subscribed resources.

There are essentially two ways to reduce prices to compete for subscribers whose needs fit in between the standard circuit capacities: measure use or control it. Control can be implemented in the access circuit through fractional (multiplexer) rates or at the routers to which the access circuits connect. The advantage of fractional rate circuits is simplicity at the router, which is balanced by the complexity of involving the circuit provider in any capacity changes. New technology for access circuits, such as Digital Subscriber Line (DSL) and cable-modems are expected to offer additional access circuit capacities to subscribers who use dial access within a few years. Since these subscribers are not likely to pay monthly fees around $1000 in Figure 1 for access rates around 1 Mbps, different pricing options will be desired.

It is possible to bill for any usage measure. Although prices need not be tied to costs, it is dangerous to price services out of line with the underlying costs of the service. Any resulting subsidies may be exploited by subscribers who resell services that are offered below cost. Successful providers will constrain usage patterns toward more efficiency through their pricing policy.

An easy usage price is total packets or bytes through the subscriber's interface. But since the cost to provide comparable service is higher when their internet is near peaks (greater capacity needed to avoid congestion and packet drops), competitive pressure would encourage discounts for off-peak usage. For example, per-byte charges could be discounted based on time of day, as long as the complexity of the price schedule does not drive customers toward simpler subscriptions. Discounts on time of day might require adjustment as subscribers time-shift their traffic, and possibly the traffic peak.

There is a different reason to focus on usage-constraining pricing rather than total volume pricing. Billing for total traffic may have the perverse effect of encouraging use when networks are congested. [8 section 4.4] TCP will increase its rate until congestion signals the limit of capacity. Total traffic moves more quickly when the network is unloaded - at no effective cost - which increases charges per unit of time. Users may prefer to use the network during peak periods when traffic is slower and it is easier to limit the volume of their traffic. Total volume billing thus creates a disincentive for users to prefer off-peak use although shifting to off-peak use improves both network efficiency and user response time.

One way to avoid the overhead cost of managing and discounting accounting data is to shift the goal of minimizing peak usage to the customer by pricing on the peak usage rate. Since the access circuit is 100% used during packet transmission, peak-rate measurement requires a measurement interval. Adjusting this interval, over which the peak is averaged, discounts traffic bursts but charges for sustained rates. Discounting bursts is no problem because internet routers are designed to accommodate just such bursts. UUNET and Digex are among the service providers already offering this kind of pricing. A competitive advantage of this approach is that it is particularly easy for customers to increase their subscription levels under this schedule.

Unfortunately, the implementation of burstable peak-rate pricing now requires sampling the rate frequently to determine its peak. Computing peak values at the router's interface where the statistics are collected could reduce the amount of data being moved through the network for accounting purposes. A design question for efficient distributed systems is whether to move data or the software needed to process it. For peak rate measurement, moving a single parameter for the averaging window to the subscriber's interface in the router could replace hundreds (288 = 12 5-minute samples * 24 hours per day) of samples from the router. The router must also be able to handle the additional per-interface calculation.

Measuring peak rates would be part of a (variable) usage-based billing policy. An alternative is to control the peak rate by limiting transmissions from routers by either shaping or dropping excess traffic. Routers already have features to limit output streams to the rate of a frame-relay virtual circuit, which can be much lower than the rate of the interface. Packets can be queued for transmission at the specified rate, shaping the output to the rate limit. Or, at a reduced cost, because it does not require memory for the queue, packets can be dropped if they exceed the configured rate limit, simulating the effect of congestion due to the rate limit of the interface. Where bursts can be managed downstream, un-shaped traffic avoids queuing delays.

5. DIFFERENTIATED SERVICE

Differentiated service can justify the differentiated pricing that is consistent with economic efficiency without re-introducing the call/reservation model of traditional telephony.

5.1 Assured Service

In the Expected Capacity model [7] of differentiated service, the concept of rate-limiting a subscriber's interface is combined with the Internet's tradition of making as much capacity available to any user in order to provide higher value rather than lower cost access. This model more directly responds to the economic goal summarized as "In the public Internet, where commercial providers offer service for payment, the feedback will most often be different prices charged to customers with different requirements. This allows the providers to charge differential prices to users that attach greater value to their Internet access, and thus fund the deployment of additional resources to better serve them." [16]

Instead of dropping traffic that is outside the rate for which the subscriber is paying, the Expected Capacity model marks traffic that is within the rate profile for relative protection from congestion elsewhere in the internet. Traffic above the profile rate is transmitted as before. When congestion occurs anywhere in the internet, routers would drop ordinary traffic in preference to traffic marked for assured service. The packet dropping mechanism needed throughout the internet is Random Early Detection (RED) with In/Out enhancements (RIO). RED [11] is a queue management mechanism that manages congestion due to the bursty nature of TCP in core routers better than tail-drop, the previous mechanism. The improvement from RED is so important its deployment is recommended throughout the Internet [5]. RIO extends RED so that packets marked out-of-profile are dropped before those marked in-profile. Packets marked for assured service are protected from the effects of congestion by ordinary packets, which are dropped preferentially at congestion. A single bit in the Type-of-Service (TOS) byte, designed into the Internet Protocol but little used until now, provides value and price differentiation while maintaining both network efficiency and the tradition of fixed pricing based on access rate. Two parameters, the rate and allowable burst for assured service, are needed in the subscriber's interface to police the service level agreement.

5.2 Premium Service

Just one more bit in the TOS byte is required for the 2-bit model for differentiated service [20] which provides a different kind of differentiation. Although assured services protect higher-valued traffic from the risk of congestion-induced drops, this traffic could still wait in queues in several routers that are absorbing traffic bursts or avoiding congestion with moderate average queue lengths. Applications such as interactive voice might be willing to pay for a premium service that minimizes delay. Not only would this service provide better than best effort, it would minimize delay through two necessary components: guaranteed provisioning of sufficient bandwidth and a separate queue which will always be very short. Since congestion and queues only form if there is not enough capacity for all traffic, over-provisioning all circuits so that there is more capacity than the aggregate of premium traffic avoids a persistent queue for this traffic. Under that guarantee, a separate queue for traffic marked premium can be given absolute priority over the queue for other (ordinary and assured) traffic. The rate limits for premium traffic must enforce the over-provisioning requirement for this to work, and traffic over the rate limit MUST be dropped. Dropping this traffic is consistent with its requirement for low latency because later delivery of the traffic would be worthless.

5.3 Settlement

Since the service provider would presumably get more revenue for assured service, the next provider in the traffic path might want a share of that revenue. The second provider could treat the first as a subscriber with a rate limit, or settlements between providers could be based on the volume of traffic that is preferentially protected from congestion. Agreements between providers to support premium traffic would require strict provisioning guarantees, as is necessary for circuit-based networking. Since the capacity for premium service will be automatically used by assured and ordinary traffic, the network efficiency of packet switching is maintained. This efficiency, along with the scale economies observed earlier, suggest large economic advantages from a packet-switched infrastructure for a wide variety of applications.

The Assured and Premium service models are promising enough that an IETF working group has been working to standardize the use of the TOS byte, which will be renamed the Differentiated Service (DS) byte. "Differentiated services are intended to provide scalable service discrimination in the Internet without the need for per-flow state and signaling at every hop. The differentiated services approach to providing quality of service in networks employs a small, well-defined set of building blocks from which a variety of services may be built. ... A differentiated-services-capable network node includes a classifier that selects packets based on the TOS octet and is capable of delivering the treatment corresponding to that marking of the TOS octet. Setting of the TOS octet and other conditioning of the dynamic behavior of marked packets need only be performed at network boundaries and may vary in complexity." [19]

5.4 Smart Market

Several levels of precedence are included in drafts being considered by the Differentiated Services working group. The precedence value could specify the priority for dropping packets during congestion. The lowest precedence packets would be dropped at lower levels of congestion, with packets at higher precedence dropped only

if congestion becomes worse, as with RIO. This preferential dropping of traffic by precedence in response to congestion is a key feature of the Smart Market [17] model for Internet economics. A brief summary of the Smart Market is that routers queue packets based on the user's price bid, with lower-bid packets dropped in response to congestion. Attractive economic results were shown when the price for all packets is set at the highest bid dropped due to congestion.

The Smart Market might be approximated using drop precedences of Differentiated Services. Some compromises between practical router mechanisms and the original Smart Market model are expected. A single queue, with RED, weighted at several levels rather than just the two needed for RIO, could provide ordered packet drops without the latency advantages implied by actually ordering the queue by bid. Since there is not room for monetary bids in the small precedence field, a billing server would have to map bids into precedence values. The bid manager's could compress values of bids into a small range of values of precedence by focussing on the values at which RED-drops signal congestion. Because replacing existing TCP/IP software is infeasible, and because some system is needed to authorize payment, separating the bidding process from packet forwarding is a reasonable approximation of the Smart Market design. A bid manager would authorize a user's edge router to mark priority for traffic as negotiated with the user. Precedence bid prices would be another component of a variable monthly bill, which would still include access-rate, circuit/call, and other charges. Open questions include how service providers would settle between their price bidding systems, and if users would actually want free-market pricing for network service.

5.5 Integrated and Differentiated Services

Differentiated Service need not be an alternative to Integrated Services. Integrated Services features such as application-specific reservations and admission control policy could be used where its scale is feasible, with Differentiated Services operating at larger scales where it is not. An Internet draft [2] has identified the characteristics of Differentiated Services necessary to connect regions of Integrated Services.

6. BILLING-RATIONING SERVICE

Metered components as well as fixed-price components are valid in billing for network services. Where reservations are made across internets, or circuits (calls) are dedicated to a subscriber for a period of time, billing based on time and capacity is well accepted, based on traditions in telephony. Other measured components of network billing are identified in this paper. However, subscribers seem to value fixed monthly bills [25] to the extent that as many as 40% of flat-rate subscribers for local telephone service would pay less with metered billing. The fixed budgets of some subscribers, such as government agencies and university departments, discourage metered-price services. Occasionally, these subscribers have extra funding for special purposes they might want to spend on network service. A network service might reconcile the values of controlled costs and economic efficiency with a little help from the routers.

A token bucket, which is a common technique in network traffic control, could control a subscriber's network access account. This control mechanism supports the intrinsically bursty nature of network use within specified bounds. Usually the bucket is filled at a constant rate, with tokens consumed by variable demands. Network billing could be modeled with payments filling a token bucket and metered charges, as well as fixed-rate charges, consuming the tokens. The contents of the bucket, and projected usage rates would provide the feedback in the subscriber's economic control loop. Billing in arrears could be modeled as credit for a payment cycle. Payments in addition to the contractually fixed charges simply add tokens to the bucket. Because disconnecting access is not the desired response to token depletion, more gradual mechanisms would ration network services. For example, admission control for reservations would be blocked when the bucket has too few tokens. If Assured Service marking were part of the service, the rate and/or burst limits could be reduced to conserve tokens until replenished by the scheduled payment. If reductions in better-than-best effort service were insufficient to avoid token depletion, the effective rate limit of the subscriber's interface could be reduced. In effect, the quality of service would diminish as payment tokens are depleted.

For rationing to work, the accounting system would need an efficient mechanism to change the rate-limit configuration of the subscriber's network interface. To operate as an effective economic control loop, delays in enforcing limits, as well as in accounting for measured usage, must be shorter than the resource consumption decisions of the user. This accounting system would serve the needs of subscribers who can manage variable network payments as well; they just never trigger rationing enforcement. All subscribers would value notification when rationing was about to be applied. This notice would have the status of a bill for variable-payment subscribers.

This kind of interactive interface to the economic status of a network subscriber's account has been discussed in general terms as an Expenditure Controller Interface [9]. User preferences among a wide variety of network quality and price combinations, and more detailed control mechanisms than suggested above, are being studied empirically in the Internet Demand Experiment (INDEX) [25], for which user interfaces and quality control have been developed. If electronic payments are added to the interactive accounting and quality control, this process enables electronic

commerce for the network infrastructure that facilitates electronic commerce more generally.

7. CONGESTION

Economists have focussed on congestion costs because congestion is the key limitation of internets. "Most of the costs of providing the Internet are more-or-less independent of the level of usage of the network; i.e., most of the costs are fixed costs. If the network is not saturated the incremental cost of sending additional packets is essentially zero." [17]. Since traffic at congestion drives the need for expansion, and increased cost, why not charge specifically for this traffic to fund expansion?

To the extent that users value this traffic less than their congestion prices, billing for congestion would encourage them to shift that use to uncongested times and locations, also improving network efficiency. Congestion pricing can provide the economic feedback advocated [16] for an economic control loop encouraging both network and economic efficiency. Including users' valuations in the control loop is essential for economic efficiency. Including the ends in control has also been a design principle [6] of the Internet.

Congestive collapse remains an ever-present danger. Internets naturally operate on the verge of congestion because TCP will exploit as much bandwidth as it can get. Although modern TCP avoids congestion well, it is not reasonable [14] to rely on the ideal (TCP) behavior in the network; but mechanisms of

1. packet scheduling,
2. buffer management,
3. feedback, and
4. end adjustments

may be necessary and sufficient to control congestion. Some applications do not respond to congestion signals as well as TCP, and they threaten congestive collapse [5, 12] if they are not controlled. Without economic consequences, what incentive would new application designers have to solve the complex problem of responding to congestion appropriately in their applications? Non-responsive greedy applications would decrease the Internet's effective capacity for those that share properly.

There are proposals to protect internets from non-responsive traffic flows within the routers, using the first two mechanisms [14] above. One research team [12] proposes identifying and regulating high-bandwidth flows that are non-responsive to congestion signals. These dangerous flows are identified from analysis of the RED drop history. Constrained scheduling on just these flows is more efficient than per-flow scheduling, which cannot solve the non-responsive flow problem by itself and may not scale for the large number of flows in core routers. Another research team [15] proposes a modification, called Fair RED (FRED), which includes per-flow accounting in RED's queue management mechanism. Flows that attempt to queue more than their (burstable) fair share of packets are limited to the average number of packets per flow in the queue. FRED is more efficient than per-flow scheduling because it operates entirely by dropping packets from a single queue, and performs accounting only for flows that have packets in the queue.

RED appears to be a good measure of congestion. RED samples usage. Sampling is how NSFnet reduced the burden of accounting data for core routers. RED is fair to the extent that "the fraction of marked packets for each connection is roughly proportional to that connection's share of the bandwidth." [11] Improvements in RED's fairness such as FRED are compatible with its use as a measure of congestion. If RED is weighted to reflect access-rate limits, as in RIO, or higher delivery precedence, the better-than-best-effort [1] traffic is protected from congestion pricing it the same way it is protected from drops; presumably the price premium has already been applied.

There are essentially two alternatives to measure congestion, in the middle or at the edge of the network. Measuring in the middle requires accounting in very busy places; measuring at the edges requires propagating details of congestion to the edge where the scale of the accounting process is reduced. A method for propagating RED signals of congestion to the receiving edge of the internet, Explicit Congestion Notification (ECN) [21], has already been proposed. Unfortunately, ECN would not apply to all traffic, just TCP sessions between end systems that honestly mark traffic for which they take responsibility to reduce traffic when signaled, and receivers that return congestion signals in through TCP. That ECN could be a component of the economic control loop for non-responsive flows appears unlikely.

An advantage of accounting for congestion where RED occurs is that the location of the congestion, as well as the source and destination in the packet, would be captured. This information can guide the deployment of additional resources. Since it is at the location where expansion needs will be identified that the accounting is needed, and a congested router is already busy, the accounting mechanism must be efficient. More important than its efficiency is that it not slow the forwarding path through the router. Instead of simply dropping a RED packet, the router could queue its header to an accounting process, which would reduce its storage requirement to a count increment for that flow. Because the number of flows in the core is huge, the accounting process would have to aggregate flows to reduce the data further. The order in which attributes of the flow are obscured by aggregation would be configurable, depending on the provider's charge allocation policy.

Eventually, the aggregated usage information would be transferred to billing servers, which would manage subscribers' accounts and settlements with other providers. A smart-market bid-mapping server would also need the precedence level of the RED-drops, if that concept were deployed. In order for congestion prices to operate as an effective feedback signal for (as yet unspecified) cost-avoiding applications, the charges must be propagated to subscribers' accounts, with dynamic user interaction, within the timescale of the congestion period.

Congestion billing can be seen as billing for waste, since the particular packets counted are not delivered. This kind of billing policy would benefit not those who pay congestion charges, but the subscribers who would pay much less for best-effort service because they do not subsidize expansion for network that fail to share properly when congestion signals the limit of capacity. If the hidden hand of market economics works, best-effort use would not degrade so badly because congestion prices discourage the deployment of applications that badly fit the statistical sharing model of the Internet.

8. SUMMARY and QUESTIONS

Existing and potential features of the routers that compose the infrastructure of internets include

- measuring or controlling transmission rates,
- marking traffic for better than best effort protection from congestion or delay,
- reserving specific transmission characteristics for particular or aggregates of flows,
- and protecting network capacity from rapacious flows that respond to congestion badly.

Systems are conceivable that

- enable competitive bidding for better internet service
- and control service levels to meet the payment objectives of subscribers.

These potential capabilities raise questions economics may better answer than engineering. Which of the mechanisms underpinning the billing process will service providers actually use? As important as what service providers want, is the question of what they need from the routers they deploy to sustain their economic success. It probably requires both economists and network operators to guide the choice of features to be implemented.

How important are fixed regular bills for network subscribers? Can those needs be adequately met with rate-limited interfaces? Will subscribers choose rationing in order to combine usage-priced services with fixed bills?

Would enough subscribers choose a measured peak rate billing option for the peak computation to be worthwhile implementing in the router interface?

Would any network service provider actually try deploying a system in which users bid for priority protection from congestion? If someone is developing this, how many levels of precedence are needed, and how are packets marked?

Are there enough subscribers who value low-cost, best-effort network service to justify the development of congestion billing systems?

9. REFERENCES

[1] Baker, F. IP Quality of Service: Better Than Best Effort. *Business Communications Review*, 28(3), March 1998. http://www.bcr.com/bcrmag/03/98p28.htm

[2] Bernet, Y., Yavatkar, R., Ford, P., Baker, F. and Zhang, L. A Framework for End-to-End QoS Combining RSVP/Intserv and Differentiated Services. Internet Draft. March 1998. http://diffserv.lcs.mit.edu/Drafts/draft-bernet-intdiff-00.txt

[3] Boyle, J., Cohen, R., Durham, D., Herzog, S., Rajan, R., and Sastry, A. The COPS (Common Open Policy Service) Protocol. Internet Draft. March 1998. ftp://ds.internic.net/internet-drafts/draft-ietf-rap-cops-01.txt

[4] Braden, B., Clark, D., Shenker, S. Integrated Services in the Internet Architecture: an Overview. RFC 1633. June 1994.

[5] Braden, B., Clark, D., Crowcroft, J., Davie, B., Deering, S., Estrin, D., Floyd, S., Jacobson, V., Minshall, G., Partridge, C., Peterson, L., Ramakrishnan, K. K., Shenker, S., Wroclawski, J., and Zhang, L. Recommendations on Queue Management and Congestion Avoidance in the Internet. RFC 2309. April 1998.

[6] Carpenter, B. Architectural Principles of the Internet. RFC 1958. June 1996.

[7] Clark, D.B., and Fang, W. Explicit Allocation of Best Effort Packet Delivery Service. November 1997. http://diffserv.lcs.mit.edu/Papers/exp-alloc-ddc-wf.pdf

[8] D. Clark, D.B., and Wroclawski, J. An Approach to Service Allocation in the Internet. Internet Draft. July 1997. http://diffserv.lcs.mit.edu/Drafts/draft-clark-diff-svc-alloc-00.txt

[9] Danielsen, K., and Weiss, M. User Control and IP Allocation. In *Internet Economics*, ed. Lee McKnight and Joseph Bailey. Cambridge, MA: MIT Press, 1997. http://www.press.umich.edu/jep/works/DanieContr.html

[10] Demers, A., Keshav, S., and Shenker, S. Analysis and Simulation of a Fair Queueing Algorithm. Proceedings SIGCOMM '89 reprinted in *Computer*

Communications Review, 25(1), January 1995. page 185.

[11] Floyd, S. and Jacobson, V. Random Early Detection gateways for Congestion Avoidance. *IEEE/ACM Transactions on Networking*, 1(4), August 1993, p. 397-413. ftp://ftp.ee.lbl.gov/papers/early.pdf

[12] Floyd, S., and Fall, K. Promoting the Use of End-to-End Congestion Control in the Internet. February 1998. ftp://ftp.ee.lbl.gov/papers/collapse.feb98.pdf

[13] Jacobson, V. Congestion Avoidance and Control. *Proceedings SIGCOMM '88*, 18(4), August 1998.

[14] Lefelhocz, C., Lyles, B., Shenker, S., and Zhang, L. Congestion Control for Best-Effort Service: Why We Need a New Paradigm. *IEEE Network*, 10(1), January 1996.

[15] Lin, D., and Morris, R. Dynamics of Random Early Detection. *Proceedings SIGCOMM '97*, October 1997. http://www.acm.org/sigcomm/sigcomm97/papers/p078.pdf

[16] MacKie-Mason, J.K., Murphy, L., and Murphy, J. The Role of Responsive Pricing in the Internet. In *Internet Economics*, ed. Lee McKnight and Joseph Bailey. Cambridge, MA: MIT Press, 1997. http://www.press.umich.edu/jep/works/MackieResp.html

[17] MacKie-Mason, J.K., and Varian, H.R. Pricing the Internet. In *Public Access to the Internet*, ed. Brian Kahin and James Keller. Englewood Cliffs, N.J.: Prentice-Hall. February 1994. ftp://alfred.sims.berkeley.edu/pub/Papers/Pricing_the_Internet.ps.Z

[18] Mankin, A., Baker, F., Braden, B., Bradner, S., O'Dell, M., Romanow, A., Weinrib, A., and Zhang, L. Resource ReSerVation Protocol (RSVP) Version 1 Applicability Statement Some Guidelines on Deployment. RFC 2208. September 1997.

[19] Nichols, K. and Blake, S. Differentiated Services Operational Model and Definitions, Internet Draft, February 1998. ftp://ftp.ietf.org/internet-drafts/draft-nichols-dsopdef-00.txt

[20] Nichols, K., Jacobson, V., and Zhang, L. A Two-bit Differentiated Services Architecture for the Internet. Internet Draft November 1997. http://diffserv.lcs.mit.edu/Drafts/draft-nichols-diff-svc-arch-00.pdf

[21] Ramakrishnan, K.K., Floyd, S. A Proposal to add Explicit Congestion Notification (ECN) to IPv6 and to TCP. Internet Draft. November 1997. ftp://ds.internic.net/internet-drafts/draft-kksjf-ecn-00.txt

[22] Shenker, S., Clark, D., Estrin, D., and Herzog, S. Pricing in Computer Networks: Reshaping the Research Agenda. *Telecommunications Policy*, 20(1), 1996 and Computer Communications Review, 26(2), April 1996. http://www.statslab.cam.ac.uk/~frank/PRICE/scott.ps

[23] Sweet, L. ISP SuperGuide. *Internet Computing*. November 1997. http://www.zdnet.com/icom/zdlabs/superguide/

[24] Varian, H. Differential Pricing and Efficiency. *First Monday*, 1(2), August 1995. http://www.firstmonday.dk/issues/issue2/different/index.html

[25] Varian, H., and Varaiya P. INDEX Workshop. Berkeley, CA. March 1998. http://www.INDEX.Berkeley.EDU/public/index.phtml

[26] Wroclawski, J. The Use of RSVP with IETF Integrated Services. RFC 2210. September 1997. Schwartz, M., and Task Force on Bias-Free Language. Guidelines for Bias-Free Writing. Indiana University Press, Bloomington IN, 1995.