What QoS Research Hasn't Understood About Risk

Benjamin Teitelbaum
ben@internet2.edu

Stanislav Shalunov
shalunov@internet2.edu

ABSTRACT

In an effort to meet application needs, network researchers have designed internet quality of service (QoS) architectures intended to deliver end-to-end performance guarantees to applications. Because such services offer *guarantees* and must elevate and isolate select traffic from competing traffic, we refer to them as "elevated services" (this is in contrast to forms of QoS that differentiate two or more best-effort service classes none of which is strictly elevated over another [3]). Despite two decades of vigorous research and standards activity, elevated services have failed to deploy. We argue that a fundamental reason for this failure is a confusion between QoS as an underlying technology and QoS as a service offering. Neither customers nor Internet service providers need or want hard performance guarantees. Instead, each wants tools to understand and manage risk.

Categories and Subject Descriptors

C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design—*Packet-switching networks*; C.2.5 [**Computer-Communication Networks**]: Local and Wide-Area Networks—*Internet*

1. INTRODUCTION

Successful technologies are designed or adapted to solve particular problems. What then are the design goals for internet quality of service (QoS)? What business need is the technology intended to solve?

S. Keshav has succinctly articulated the fundamental design goal for elevated QoS. He wrote: *"The Holy Grail of computer networking is to design a network that has the flexibility and low cost of the Internet, yet offers the end-to-end quality-of-service guarantees of the telephone network"* [6]. This ideal is certainly attractive, but it has remained elusive.

We argue that engineering true end-to-end QoS guarantees necessarily destroys the cost advantages of the Internet. We further argue that what both customers and providers need are not true guarantees, but rather tools to manage risk.

At the level of packet forwarding, risk management tools might include adequate capacity provisioning or, perhaps, statistically provisioned QoS techniques aimed at providing a level of assurance without the full cost of a *guarantee*. However, engineering is merely one technique to manage risk. Economic means, including warranties, insurance policies, and risk markets, must also be considered.

2. NETWORK ECONOMICS

Understanding why Keshav's Holy Grail is unattainable involves understanding the economics of the default best-effort internet service and those of the circuit switched network.

2.1 Best-Effort Network Economics

The Internet's default best-effort service is remarkably lax. Packets are forwarded toward their destination with no assurances about delay, packet loss probability, reordering, corruption, or duplication. The absence of performance requirements has made it possible to run IP over any link layer, build IP routers with virtually any internal switching design, and run an IP network largely unattended and with very simple peering and settlement agreements. IP is dumb and cheap and, consequently, scalable to very high speeds and global reach.

Few providers in countries with fiber deployment have had trouble provisioning their networks to meet demand; what exceptions there have been, have been in environments with significant negative externalities such as last-mile access monopolies or extreme geographic isolation. Admittedly, the task of provisioning has been made artificially easy. Most senders abide by the social contract of TCP Reno congestion control [5], are damped in aggregate by the prevalence of performance faults in hosts and LAN wiring [7] [4], and are ultimately constrained by the bottleneck of the last-mile access circuit. Even if these constraints were to vanish and best-effort network providers found themselves dramatically more vulnerable to unpredictable traffic loads (e.g., from distributed denial of service (DoS) attacks), providers could restore a high-quality best-effort internet service to users who need it without resorting to engineered service guarantees (e.g., by employing usage-based pricing feedback to senders, possibly with differentiated service levels [1] [8]).

The low cost of the Internet stems not just from the simplicity of the best-effort service model, but from the end-to-end design principle [9] as well. Internet applications are de-

ACM SIGCOMM 2003 Workshops August 25 & 27, 2003, Karlsruhe, Germany.

signed to exploit the abundance of cheap CPU and memory resources at customer endpoints to achieve application-level performance objectives. Examples of this include: TCP, media playback buffers, and loss-tolerant media codecs (e.g., packet loss concealment algorithms, error-correcting codes).

The use of adaptive and loss-tolerant applications is merely one way that customers manage their exposure to variable Internet performance. Other current practices include multi-homing with several providers, installing QoS appliance middle boxes in front of access circuits, enforcing per-host usage quotas, and employing rigorous host security practices to reduce the likelihood of disruptive DoS attacks.

The best-effort service model and the end-to-end design principle have supported the fantastic growth of the Internet. Although unreliable, the best-effort internet is fast and cheap and allows most applications to meet their needs most of the time. To retrofit this design to provide guarantees without destroying these qualities is a tall order.

2.2 Circuit Switched Network Economics

The public switched telephony network (PSTN) provides global reach and hard performance assurances. PSTN carriers have evolved mostly as regulated monopolies and significant attention has been made to billing customers based on usage and to accurate inter-provider settlements. Telecommunications carriers have extended their service offerings from 64 kb/s voice channels and bundled voice trunks to ISDN and ATM services with similar global reach and hard performance assurances.

The PSTN demonstrates that hard QoS assurances are possible in a global network. It fails, however, when it comes to price and efficiency. Circuit switched services are more expensive per bit than internet services, particularly when engineered to be comparably resilient. In part, these services are expensive because of lack of statistical multiplexing, leaving network resources idle much of the time. An even bigger factor, however, are the administrative costs of maintaining call detail records, customer billing, and a careful accounting of inter-provider settlements.

3. FROM ASSURANCE TO INSURANCE

An IP QoS guarantee would do little to change packet forwarding under normal conditions. Most of the Internet is well-provisioned and provides a cheap and global communications service that performs very well in the typical case. But elevated services are not about the typical case. They are about the worst case.

Although typical performance may be very good day in and day out, customers may want to bound their exposure to risk from failed networked business transactions. What major call center would want the possibility of a major service disruption due to internet congestion? Who would tolerate tele-robotic surgery that could be disrupted by IP packet loss events?

The goal of guaranteed QoS designs has been to eliminate or bound the risk that preferred traffic will experience congestion. But congestion is merely one component of risk. Customers are also concerned with other sources of network failure that could affect their businesses (e.g., equipment failures, fiber cuts). Elevated QoS must be seen, therefore, as a technology used to manage a component of risk, not as the end service sold to customers. Customers want an insurance policy against unacceptable network performance.

3.1 Managing Risk

Bounding the risk of one's exposure to the worst case is not the same as designing a system that can provably avoid the worst case. In architecture, aircraft design, freight delivery, and other businesses, customers don't insist on infallible designs. Instead, they purchase insurance to hedge against the possibility of failure. Systems reach a point where it is cheaper to reduce risk through economic means than it is to "add more nines" to the reliability of the underlying technology. Actually bounding worst-case performance may only be cost-effective for the most critical business applications. For the vast majority of applications, customers will be satisfied to absorb some uncertainty if they know they'll be financially compensated for network performance failures.

3.2 Warranted QoS

In practice, the service that is advertised and sold to the customer ("Premium service with zero loss and jitter!") can not be identical to the service that is actually engineered by the provider. Businesses built around service assurances (e.g., overnight package delivery, business class air travel, frame relay) do not strive for 100% service reliability. By separating the advertised service from the engineered service, these businesses have the flexibility to trade off statistical over-booking and operational corner-cutting against the probability that customer assurances will not be met and that money will have to be refunded to the customer.

To maximize profit, elevated services must be explained to the customer in simple terms, but engineered carefully by the provider with a strong understanding of the statistical nature of the offered traffic, the performance of the network elements, and the exposure of the provider to DoS attacks. Of course, the statistical nature of traffic is always changing as new applications emerge and older ones fade away; so, this effort would have to be ongoing. There is insufficient theoretical understanding of how to do this kind of traffic modeling well for IP networks and to reason about statistically provisioned services. Even perfect models would not suffice. As we have already noted, IP networks are vulnerable to denial of service (DoS) attacks with which an adversary can undermine any statistical model of load.

3.3 Third Party QoS Insurance

Savvy customers will recognize that service providers have incentives to over-book capacity and cheat on their assurances. They will also recognize that their providers are vulnerable to DoS attack. These customers might demand accurate outage reports or strong recourse for service failure (penalties to the provider more severe than honoring a "your money back" warranty). However, this is generally not something that service providers have accepted in other businesses, unless required by law.

A logical option is for neutral third-party insurers to fill this void. Insurance is a device that allows customers to trade off money for lowered risk and allows insurers to manage pools of risk, making a profit on the "float" of the premiums they collect. In an efficient insurance market, this trade-off occurs with economic efficiency. Such insurers might require audits of a provider's "performance record," as well as of its engineering and operational practices, including the use of common-sense DoS prevention mechanisms.

Internet service providers might be required to carry a certain amount of network performance insurance by law or insurance providers could offer a continuum of products directly to customers allowing them to select their level of risk and steering them towards "preferred provider" carriers.

This paper is not the first to propose markets for trading risk and service level assurances. Indeed, the proposal for "risk brokers" in the Market Managed Multiservice Internet (M3I) project [2] is quite similar. B. Schneier advocates a broader view of security threats as risks to be managed through such instruments as, in particular, insurance [10] [11].

One might ask, however, why a market for network performance insurance has not developed. A potential explanation is that there has been insufficient research into new internet pricing and business models, while at the same time, far too much research on engineered QoS designs that do not address the real need of providers and customers to manage risk flexibly.

4. CONCLUSION

In conclusion, QoS researchers have far too often ignored network economics and the real needs of providers and customers. Researchers, having invented novel scheduling and policing disciplines, have been far too quick to assume that they may be used to implement services that are deployable, marketable, and useful. We have argued that what is really needed are tools to understand and manage risk. Engineered QoS mechanisms may be employed to help manage a component of risk—namely the risk of unacceptable network performance due to congestion—but they must be complemented by economic instruments and market mechanisms. Warranties and insurance policies, in particular, are two such mechanisms that have been discussed.

5. REFERENCES

[1] E. Anderson, F. Kelly, and R. Steinberg. A Contract and Balancing Mechanism for Sharing Capacity in a Communication Network. *Working Paper LSEOR 02.50, London School of Economics, Department of Operational Research* (current draft: http://www.statslab.cam.ac.uk/ frank/aks.pdf), July 2002.

[2] B. Briscoe. Market Managed Multi-service Internet (M3I) overview and architecture. *Proceedings of M3I QofIS'2000 Workshop on Internet Charging* (http://www.m3i.org/workshop/workshop/pdf/briscoe.pdf), Berlin, September 27, 2000.

[3] P. Hurley, J.-Y. Le Boudec, P. Thiran, and M. Kara. ABE: Providing a Low-Delay Service within Best Effort. *IEEE Network Magazine*, v 15, no 3, May/June 2001.

[4] Internet2 NetFlow Weekly Reports (http://netflow.internet2.edu/weekly/).

[5] V. Jacobson and M. J. Karels. Congestion Avoidance and Control. *ACM Computer Communication Review; Proceedings of the Sigcomm '88 Symposium in Stanford, CA*, August, 1988.

[6] S. Keshav. *An Engineering Approach to Computer Networking*, Addison-Wesley, 1997.

[7] C. de Luna. Auto-sensing, Auto-negotiation, and Duplexing at JPL. http://performance.jpl.nasa.gov/Auto010115.PDF, September 12, 2001.

[8] A. Odlyzko. Paris metro pricing for the internet. *Proceedings of the first ACM conference on Electronic commerce*, pp 140–147, 1999.

[9] J. Saltzer, D. Reed, and D. Clark. End-to-End Arguments In System Design. *ACM Transactions in Computer Systems*, November, 1984.

[10] B. Schneier. *Secrets and Lies: Digital Security in a Networked World*, John Wiley *&* Sons, 2000.

[11] B. Schneier. The Insurance Takeover. *Information Security*, February 2001.