5-12-2012

Network Neutrality and the Need for a Technological Turn in Internet Scholarship

Christopher S. Yoo
University of Pennsylvania Carey Law School

Network Neutrality and the Need for a Technological Turn in Internet Scholarship

Christopher S. Yoo*

ABSTRACT

To most social scientists, the technical details of how the Internet actually works remain arcane and inaccessible. At the same time, convergence is forcing scholars to grapple with how to apply regulatory regimes developed for traditional media to a world in which all services are provided via an Internet-based platform. This chapter explores the problems caused by the lack of familiarity with the underlying technology, using as its focus the network neutrality debate that has dominated Internet policy for the past several years. The analysis underscores a surprising lack of sophistication in the current debate. Unfamiliarity with the Internet's architecture has allowed some advocates to characterize prioritization of network traffic as an aberration, when in fact it is a central feature designed into the network since its inception. The lack of knowledge has allowed advocates to recast pragmatic engineering concepts as supposedly inviolable architectural principles, effectively imbuing certain types of political advocacy with a false sense of scientific legitimacy. As the technologies comprising the network continue to change and the demands of end users create pressure on the network to further evolve, the absence of technical grounding risks making the status quo seem like a natural construct that cannot or should not be changed.

* John H. Chestnut Professor of Law, Communication, and Computer & Information Science and Founding Director of the Center for Technology, Innovation and Competition, University of Pennsylvania. This Chapter is adapted from comments the author submitted on November 4, 2010, in the FCC proceeding on "Preserving the Open Internet."

INTRODUCTION

Academic scholarship about the Internet now includes a broad array of interdisciplinary perspectives, encompassing such fields as communications, economics, sociology, political science, history, anthropology, and law (Nissenbaum and Price 2004). Yet to most social scientists, the technical details of how the Internet actually works remain arcane and inaccessible (Sandvig 2009). At the same time, convergence is forcing scholars to grapple with how to apply regulatory regimes developed for traditional media such as broadcasting, telephony, and cable television to a world in which all voice, video, and text services are provided via an Internet-based platform. Grappling with how to reconcile existing law with these changes requires some degree of familiarity with the intricacies of the technology. The required level of technical expertise is likely to continue to increase even further as the Internet matures both as an industry and as a field of study.

This chapter explores the problems caused by the lack of familiarity with the underlying technology by way of illustration. It focuses on the network neutrality debate that has dominated Internet policy for the past several years, beginning with four historical architectural commitments to permitting prioritization and then examining ten modern examples of non-neutral, prioritized architectures.

The analysis underscores just how surprising the relative lack of sophistication reflected in the current debate actually is. Unfamiliarity with the Internet's architecture has allowed some advocates to characterize prioritization of network traffic as an aberration, when in fact it is a

central feature designed into the network since its inception. Moreover, despite the universal recognition of the need to accommodate technical concepts such as network security and congestion management, many people involved in the debate have only the barest notion of how the Internet manages security and congestion. At the same time, the lack of knowledge has allowed advocates to recast pragmatic engineering concepts as supposedly inviolable architectural principles, effectively imbuing certain types of political advocacy with a false sense of scientific legitimacy (Blumenthal 2002; Gillespie 2006).

Lastly, those without an understanding of the network's design will find it difficult to appreciate the significance of changes in the way people are using the network. Video is now a significant component of network traffic, with other innovations, such as cloud computing, sensor networks, and the advent of fourth-generation (4G) wireless broadband waiting in the wings. The radical changes in the technologies comprising the network and the demands that end users are placing on it is creating pressure on the network to evolve in response (Yoo 2012). The absence of some technical grounding risks making the status quo seem like a natural construct that cannot or should not be changed.

HISTORICAL EXAMPLES OF PRIORITIZATION

The current policy debate often tries to depict network owners' recent efforts to prioritize certain traffic as new and aberrant deviations from the status quo. A brief review of the history of the Internet reveals that prioritization is a feature that has been built in from the beginning. Moreover, the years that followed witnessed sustained and persistent efforts to extend and enhance network operators' ability to engage in sophisticated network management.

A. The Type of Service Flag in the Original Internet Protocol

The heart of the Internet is the Internet Protocol (IP), which a leading textbook on computer networking aptly describes as "[t]he glue that holds the entire Internet together" (Tanenbaum 2003: 432). IP is designed to provide a single common language that enables a diverse range of different network technologies to interconnect with one another seamlessly (Cerf & Kahn 1974). As a matter of principle, IP was kept as simple as possible, being "specifically limited in scope to provide the functions necessarily to deliver a package of bits . . . from a source to a destination over an interconnected systems of networks" (Information Science Institute 1981: 1). IP was thus designed to include only the bare minimum needed for the network to function properly (Leiner *et al.* 1985).

The need to keep IP as simple and robust as possible made the protocol architects' decision to include an eight-bit *type of service field* in the IP header particularly telling (Zhu 2007). The type of service field was designed to allow networks to attach different levels of priority to particular packets. The first three bits permitted the assignment of three varying levels of precedence to the packet. The next three bits allowed the specification of the three different dimensions of precedence: delay, throughput, and reliability (Information Science Institute 1981; see also Information Science Institute 1979). A separate standard documented how to map the flags in the type of service field onto the actual service provided by the networks comprising the Internet (Postel 1981).

This field was included explicitly to "support . . . a variety of types of service . . . distinguished by differing requirements for such things as speed, latency and reliability" (Clark 1988: 108). Indeed, the document establishing IP specifically notes that they included the type of service field to "capitalize on the services of its supporting networks to provide various types and

qualities of service" (Info. Sci. Inst. 1981: 1). The specification later noted, "Several networks offer service precedence, which somehow treats high precedence traffic as more important than other traffic." The decision to include the type of service flag in the IP header reflects a belief in the importance in supporting this type of functionality. The protocol designers explicitly recognized that prioritization inevitably gave rise to tradeoffs: "In many networks better performance for one of these parameters is coupled with worse performance on another." The existence of such costs counseled in favor of using prioritization judicially rather than prohibiting it altogether (*id.* at 12).

B. Prioritization of Terminal Sessions over File Transfer Sessions on the NSFNET

One of the earliest examples of prioritized service occurred in 1987 on the Internet predecessor known as the National Science Foundation Network (NSFNET) when end users first began to connect to the network through personal computers (PCs) instead of dumb terminals. Terminal sessions are an extremely interactive application, in which every key stroke is immediately transmitted and which requires constant, real-time interaction with the network. Any delay causes the terminal to lock up temporarily. File transfers are considerably less interactive. Particularly given the 56 kbps backbone speeds of the time, end users would typically expect file transfers to last several minutes.

The advent of PCs made it much easier for end users to transfer files, which in turn increased the intensity of the demands that end users were placing on the network to the point where the network slowed to a crawl. The resulting congestion caused terminal sessions to run agonizingly slowly, and the fact that fixed cost investments could not be made instantaneously created an unavoidable delay in adding network capacity.

NSFNET's interim solution was to reprogram its routers to give traffic running the application protocol associated with terminal sessions (telnet) higher priority than traffic running the application associated with file transfer sessions (File Transfer Protocol or FTP) until additional bandwidth could be added. In short, intelligence in the core of the network looked inside packets and gave a higher priority to interactive, real-time traffic and deprioritized traffic that was less sensitive to delay. The network also made wider use of prioritization in the type of service field in the IP header (MacKie-Mason and Varian 1994).

This episode demonstrates why forecasting the amount of network capacity is so difficult. The spike in traffic was driven not by any change within the network itself, but rather a major innovation in a complementary technology (the PC) that changed the ways people used the network. In this sense, it bears a striking resemblance to the state of affairs in 1995 and 1996, when the simultaneous development of HTML and Mosaic, the first graphically-oriented browser, caused Internet traffic to grow at an annual rate of 800 per cent to 900 per cent and to turn the network into what many dubbed "the World Wide Wait" (Yoo 2012). As difficult as it is to correctly anticipate developments within the network, it is even harder to foresee game-changing improvements in complementary technologies.

This episode also demonstrates the beneficial role that network management can play in providing a better end user experience. Indeed, prioritization actually might have been able to offer better service to users of terminal sessions without degrading the experience of file transfer users. This is because the performance of file transfer sessions depends entirely on when the last packet arrives. Interactive applications (particularly streaming applications), in contrast, are very sensitive to the speed and spacing with which intermediate packets arrive. So long as the delivery time of the last packet is not affected, the network can rearrange the delivery schedule for

intermediate packets associated with terminal sessions without adversely affecting overall performance file transfer sessions.

At the same time, this episode demonstrates how core-based solutions that explicitly route traffic based on the application layer protocol with which it is associated can benefit consumers. Although this example represented a short-run solution, in theory such solutions need not be temporary. Indeed, in a technologically dynamic world, one would expect at times that employing network management techniques would be cheaper than adding bandwidth, and vice versa. Moreover, one would also expect the relative cost of these alternative solutions (and the balance that they imply) to change over time.

C. The Shift to BGP to Enable Policy Based Routing

The emergence of Border Gateway Protocol (BGP) also reflects the historic importance of allowing greater control over the way certain packets travel over the Internet. Before BGP emerged, the primary routing protocol was known as the Exterior Gateway Protocol (EGP). EGP suffered from a number of shortcomings. For example, it could not accommodate more complex topologies in which a particular network (also called an autonomous system) was available via more than one route (Rekhter 1989).

In addition, a network running EGP only informed neighboring networks about the length of the path through which it could reach to particular addresses without providing any specific information about the path that particular packets would traverse. A network that was interested only in delivering packets as quickly as possible could simply examine the length of the routes advertised by its neighbors and opt for the shortest option. The problem is that networks are often interested in more than just the length of the path. For example, until 1991, the standard acceptable use policy prohibited using the NSFNET for conveying commercial traffic. As a

result, networks sending commercial traffic needed some way to know whether particular advertised routes traversed the NSFNET and sometimes to forego a shorter route in order comply with the NSFNET's commercialization restrictions (Huitema 1995). Others may prefer certain routes because the existence of peering agreements with particular networks or the need to keep certain traffic within certain ratios may make it more cost efficient to route traffic along a particular path. Still others might prefer to avoid certain paths because of security concerns. A leading textbook gives the following examples of such routing policies (Tanenbaum 2003: 460):

1. No transit traffic through certain networks.
2. Never put Iraq on a route starting at the Pentagon.
3. Do not use the United States to get from British Columbia to Ontario.
4. Only transit Albania if there is no alternative to the destination.
5. Traffic starting or ending at IBM should not transit Microsoft.

Unfortunately, because EGP only provided information about path length without identifying the particular networks traversed, it did not provide sufficient information to support such polices.

Instead of following EGP's approach of having routers exchange information only about the length of the path by which they could reach a particular address, routers running BGP notify their neighbors about the precise path used. Every router running BGP examines the advertised routes and uses a proprietary scoring system to calculate to each location through every particular neighbor and transmits packets bound for that location via the shortest path.

One advantage of providing complete path information about particular routes is that it provides much stronger support for routing policies. A router conveying commercial traffic during the early days of the NSFNET could easily examine the precise paths comprising particular routes and decline to use any that traversed the NSFNET. Indeed, it is a simple matter

to assign any route that violates a policy a score of infinity, thereby guaranteeing that that route will not be used (Tanenbaum 2003). “In nontechnical terms, this means AT&T routers can make discriminatory routing decisions such as treating traffic from Sprint more favorably than traffic from Verizon, or even rejecting Verizon Traffic altogether” (Zhu 2007: 635).

The desire to provide better support for routing policies is widely recognized as one of the primary motivations driving the shift from EGP to BGP. Indeed, as the initial standard describing BGP noted, creating a routing system “from which policy decisions at an [autonomous system] level may be enforced” was one of the central design goals underlying BGP (Lougheed and Rekhter 1989: 1). All traffic subject to a routing policy would necessarily have to travel along a longer route (and thus take a longer time) than traffic between the same two points that was not subject to the policy.

BGP is not without its shortcomings. For example, although it allows the advertisement of multiple paths to the same network, it permits only one of those paths to be used at any particular time. When multiple routers connect two networks, BGP does not support balancing the load across all of those routers. Moreover, because route information is exchanged between adjacent networks, information about changes in routes and topology can take time to propagate through the system. During the time when routing information has not yet reached equilibrium, different routers may be referencing routing information that is incorrect or inconsistent (Comer 2006). None of these considerations, however, alters the fundamental fact that BGP was specifically designed to allow individual networks to give preference to traffic associated with particular sources and destinations and to avoid certain networks altogether.

D. IETF Standards for Integrated Services, Differentiated Services, and MultiProtocol Label Switching

The development of the IP header and the deployment of BGP did not represent the only way in which the engineering community attempted to support prioritization on the Internet. Over the past two decades, the engineering community has developed a series of potential solutions to provide applications with different levels of quality of service.

The first initiative, developed in 1991, is known as Integrated Services (IntServ). IntServ allows end users to send a reservation message inquiring whether sufficient resources exist at that time to provide a particular level of service. If the capacity is available, each router notes the reservation and reserves the resources needed to transmit the communication (Braden *et al.* 1994). The principal downside was that implementing IntServ would require substantial changes to the router infrastructure. Specifically, each router would have to be reconfigured to be able to signal the end user whether the requested resources are available and to have some means for reserving those resources. In addition, the need to set up each flow in advance can be quite complicated, requires routers to maintain per-flow state, and violates the store-and-forward principles requiring that each router make route each individual packet independently.

The second initiative, standardized in 1998, is known as Differentiated Services (DiffServ). DiffServ divides traffic into particular routing classes, with each class denoted by a Differentiated Services Code Point (DSCP) stored in a reconfigured type-of-service field in the IP header (Blake *et al.* 1998). Disassociating the type-of-service field from the three dimensional, three-level semantics in the original design of the type-of-service field allows DiffServ to support a broader range of quality of service. Many companies have begun to use DiffServ in their internal networks to ensure that delay-sensitive traffic is delivered in a timely manner. For example, Comcast is using DiffServ to prevent delays in its voice service, and

AT&T is using DiffServ to ensure that there are no delays in its video service. Unlike the resource reservation and admission control approach employed by IntServ, the prioritization-oriented approach employed by DiffServ cannot guarantee quality of service. It can only increase the probability that a particular packet will arrive within a particular time.

Another solution known as MultiProtocol Label Switching (MPLS) incorporates the features of technologies such as Asynchronous Transfer Mode (ATM) that were designed to increase routers' forwarding speed. Instead of routing based on IP addresses, MPLS adds a label to the front of each packet and routes on the basis of that label. In addition, each flow (known as a *Forwarding Equivalence Class*) is assigned specific path through the network. Information about the label and the associated route are prorogated to other MPLS-enabled routers (Rosen *et al*. 2001). Because labels are shorter than IP addresses, routers can direct traffic more rapidly. The fact that the route that a particular flow is defined in advance gives end users greater control over security. In addition, MPLS can support load balancing simply by dividing traffic between the same two endpoints into two separate Forwarding Equivalence Classes and assigning them different paths. Most importantly for the purposes of this chapter, in determining the particular path that a particular flow will travel, the MPLS router can match the quality of service demanded by the flow with the resources available along possible paths (Stallings 2001). By establishing what are tantamount to virtual circuits, MPLS exists in considerable tension with many architectural principles that many regard as central to the Internet. That said, MPLS is now being widely deployed by network providers and represents still another way in which the standards comprising the existing architecture are designed explicitly to support prioritizing certain traffic over other traffic.

CONTEMPORARY EXAMPLES OF PRIORITIZATION

The historical examples enumerated above have been followed by more contemporary examples in which network management is being used to increase the network's functionality or reliability. These examples can serve useful reference points when determining what constitutes unreasonable discrimination, reasonable network management, specialized services, and other key concepts with respect to network neutrality.

A. The 700 MHz Auction

The auction of the spectrum recovered from television broadcasters following the migration to digital transmission (commonly known as the 700 MHz auction) provides another prime example of the benefits of network management. The FCC divided the spectrum in this auction into five blocks, labeled "A" through "E." Blocks A and E were the least commercially attractive, as they were subject to interference issues; the E block contained 6 MHz of spectrum, roughly half the amount needed to provide high-quality service.

Of the remaining blocks, which represented the most commercially attractive opportunities, the B block contained 12 MHz of spectrum and was subject to the fewest restrictions. The C block contained 22 MHz of spectrum, but was subject to the requirement that the licensee provide open access to all applications and devices that do not cause harm to the network (FCC 2007).

The D block contained 11 MHz of spectrum subject to the requirement that the licensee build a single network shared both by public safety users and commercial users. Public safety users would be given the unconditional right to preempt commercial traffic during emergency situations. Commercial users would operate on a secondary basis that must accept interference

from primary users at all times and must not interfere with the primary users. In essence, the rules established two classes of service operating on the same network, with the primary, higher value use being given unconditional priority over all secondary uses (FCC 2007).

The Commission's justification for this decision offers one of the clearest statements of the benefits associated with allowing multiple tiers of service. First, the higher utilization from combining two different types of uses can reduce costs by allowing the network to realize economies of scale. Second, sharing spectrum with multiple users promotes more efficient use of spectrum. Third, the addition of secondary uses will help defray the cost of building out the network without adversely affecting the needs of the higher value, primary uses (FCC 2006; FCC 2007). The FCC's decisions with respect to the D block in the 700 MHz auction provide a powerful exposition of the benefits of allowing networks to give priority to higher value traffic traveling in the same pipe as lower value traffic. Doing so both yields consumer benefits and promotes competition in areas where broadband is already available. Making it easier for network providers to cover the costs of constructing new networks also promotes entry into currently unserved areas.

The 700 MHz auction was held on March 18, 2008, with wide variations in the amounts generated by each block. To normalize for the different amounts of spectrum included in each block, the standard practice is to analyze the total price paid on a per-MHz basis. In addition, license value is also determined by the size of the population encompassed within its boundaries, since licenses in densely populated areas are more valuable than licenses providing the same amount of spectrum in more sparsely populated areas. Therefore, license value is also determined on a per population basis, with the cost of particular spectrum being measured in terms of megahertz/population ("MHz/pop").

The B block, which was the least encumbered by regulatory requirements, proved to be the most valuable, selling for $2.68 per MHz/pop. The C block, which was encumbered by open access requirements, sold for $0.76 per MHz/pop. The D block, which was required to share spectrum and give priority to public safety services, failed to meet its reserve price of $1.33 billion (the equivalent of $0.44 per MHz/pop) and thus did no sell at all. In fact, the largest bid for the D block was only $472 million, or roughly $0.17 per MHz/pop. For comparison, the interference-encumbered A and E blocks sold for $1.16 and $0.74 per MHz/pop respectively (Kirby 2008).

The actual results of the 700 MHz auction provide specific empirical evidence of the economic impact of open access requirements. Auction results for specific blocks may reflect variations in the amount of contiguous spectrum in each block, differences in the propagation characteristics of particular portions of the spectrum, disparities in the size of the geographic area being served, and the case-specific dynamics of which firms happen to pursue particular licenses. That said, the fact that the block encumbered by open access requirements (the C block) sold for 72 per cent less than the least unencumbered block (the B block) suggests that the reduction in value associated with open access is substantial. Put another way, these prices imply that providers may have to commit up to 3.5 times more spectrum if they are to provide service on an open-access basis. The failure of the D block to meet its reserve price reveals that not all forms of prioritized service are necessarily beneficial. By giving the public safety community the right to interrupt service effectively guarantees that other users will lose connectivity during times of crisis. In the words of one financial analyst, the advertising slogan for such a service would essentially be, “Guaranteed not to work when you need it most” (Bazelon 2008).

B. Load Balancing

When the Internet first emerged, the entities that comprised it interconnected through a relatively simple and uniform set of business relationships. End users purchased a service from a single last-mile provider. Last-mile Internet service providers (ISPs) exchanged traffic with a single regional ISP. Regional ISPs handed off all their traffic to a single backbone provider. The one-to-one nature of these relationships caused the network to assume a topology known as a *spanning tree*, in which any two endpoints were only connected by a single path. Routing decisions in a spanning tree are relatively simple, because the uniqueness of the path connecting any two endpoints means that there is only a single route that traffic between those points can take.

Over time, the networks that form the Internet began to interconnect with one another in ways that deviated from the spanning tree topology. End users and ISPs began to *multihome* by purchasing connections to more than one upstream provider. In addition, regional ISPs entered into *secondary peering* arrangements in which they exchanged traffic directly with one another without traversing any backbones (Yoo 2012).

One result of these topological innovations is that many endpoints began to be connected by multiple paths. The presence of multiple paths between endpoints naturally means that someone must decide along which path to route the traffic. Although most networks choose routes that minimize the number of hops, networks may sometimes find it beneficial to route traffic along longer routes in order to satisfy other requirements of their interconnection relationships. For example, a network may seek to reduce congestion and minimize transit costs by balancing the loads between the two available paths. Alternatively, a network may intentionally route traffic over a longer path if doing so will help it maintain its traffic within the

ratios mandated by its peering contract. Some load balancing systems may also increase the networks effective performance by testing the throughput rates provided by each path and sending the traffic that is the most sensitive to delay along the faster connection (Yoo 2010).

Again, the effect is to introduce significant variance in the speed with which similarly situated packets will arrive at their destination and the cost that similarly situated packets will have to bear. This variance results not from anticompetitive motives, but rather from networks' attempts to minimize costs and to ensure quality of service in the face of a network topology that is increasingly heterogeneous.

C. AT&T's U-verse

As the Commission noted when first seeking comment on its Open Internet proposal in October 2009, AT&T's U-verse represents an important example of a specialized service (FCC 2009). The benefits provided by prioritization and reserving bandwidth are vividly illustrated by comparing U-Verse with Verizon's FiOS network. Verizon is investing $23 billion to create a last-mile network, the fiber-based FiOS network, which offers up to 100 Mbps and holds the promise of providing up to 10 Gbps of service. In contrast, AT&T's strategy leverages the existing telephone network by deploying a DSL-based technology known as VDSL. U-verse provides smaller amounts of bandwidth, ranging from 20 to 32 Mbps depending on a particular customer's location, but at the much lower cost of $6 to $7 billion (Yoo 2012).

The problem is that U-verse does not have enough bandwidth to provide video in the same manner as cable companies and FiOS. Thus, in order to avoid the delays that can render video programming unwatchable, U-verse reserves bandwidth for its own proprietary video offerings and gives its video traffic priority over other traffic.

In many ways, AT&T's practices represent precisely the type of conduct that gives network neutrality proponents pause. It prioritizes a single application (video) from a single source (AT&T) and runs the risk of allowing AT&T to gain a competitive advantage by favoring its own content over others. And yet, these practices have allowed AT&T to avoid having to spend an additional $17 billion needed to deploy fiber-based solutions like FiOS.

Given the ever-growing demand for bandwidth and the tightening of the capital markets associated with the ongoing recession, policymakers should avoid regulations that make higher capital investments the only solution to the problem of video-induced traffic growth and should instead permit networks to use prioritization to employ more efficiently the capacity that already exists. Placing regulatory restrictions on network management would not only degrade the service of existing customers. If network providers are not permitted to use network management to ensure adequate quality of service, their only option is to build larger networks to ensure that capacity never reaches saturation. Increasing the amount of capacity needed to support a particular number of customers would increase the per capita expense of building new networks. This de facto increase in cost would limit broadband deployment in rural and other low-density populations, as demonstrated by the large number of filings by public officials and business leaders from rural areas and small towns opposing the Open Internet initiative.

D. The Amtrak Acela

Another interesting form of network management occurs on the wireless broadband service provided on the Amtrak Acela. Amtrak makes clear in its terms that file downloads are limited to 10 MB. Not only does it block high-volume uses; it targets specific applications by explicitly "block[ing] access to streaming media." It does so because "[t]he explosion of the Internet and the use of Wi-Fi have created incredible demands for connectivity." Only by

managing its network in this manner can Amtrak "maximize the amount of onboard bandwidth available to all passengers" (Amtrak n.d.).

The use restrictions cannot plausibly be attributed to anticompetitive motives. Indeed, Amtrak provides WiFi service for free in its stations and on many of its most popular routes. Instead, its primary motivation is to promote minimum levels of quality of service by preventing a small group of users from consuming all of the available bandwidth.

E. PlusNet

The innovative network management techniques employed by British DSL provider PlusNet provide another example of the potential benefits of specialized services. PlusNet uses deep packet inspection (DPI) to divide the data stream into multiple different levels of priority. In so doing, PlusNet has served as a model of public disclosure, explaining how what it is doing to prioritize traffic, why connection speeds vary in particular cases, and offering meaningful guidance as to expected speeds during different times of day. Prioritizing traffic in this manner has enabled PlusNet to win numerous industry awards for the quality of their network connections and for customer satisfaction (PlusNet 2011a, 2011b, 2011c, n.d.).

In many ways, DPI has generated undeserved criticism. Sometimes denigrated as a deviation from network norms, DPI is widely used by most (if not all) major ISPs to examine samples of traffic to search for security threats. PlusNet provides a particularly telling example of why a reflexive hostility toward DPI is unwarranted.

F. Internet2's Interoperable On-demand Network (ION)

One of the central tenets underlying the Internet is that routers should operate on a pure store-and-forward basis without having to keep track of what happens to packets after they have

been passed on. This commitment is reflected in the Internet's general hostility toward virtual circuits and the belief that routers should not maintain per-flow state. Opponents of network management often point to the Senate testimony offered by officials of Internet2—a nonprofit partnership of universities, corporations, and other organizations devoted to advancing the state of the Internet—noting that, although their network designers initially assumed that ensuring quality of service required building intelligence into the network, "all of [their] research and practical experience supported the conclusion that it was far more cost effective to simply provide more bandwidth" (Bachula 2006: 66).

To a certain extent, this longstanding hostility toward virtual circuits is an artifact of the Internet's military origins that has less relevance for the Internet of today. DARPA protocol architect David Clark has pointed out that the belief that routers operating in the core of the network should not maintain per-flow state derived largely from the high priority that military planners placed on survivability (Clark 1988). Clark notes, however, that survivability does not represent a significant concern for the modern Internet. Moreover, technologies such as IntServ and MPLS, both of which are governed by accepted standards promulgated by the Internet Engineering Task Force (IETF), employ what amount to virtual circuits to enhance quality of service and to increase network efficiency to allow greater control over routing, functions that the original design prioritized below survivability. Although IntServ has not achieved widespread acceptance, interest in MPLS appears to be growing.

These developments can be seen as part of a broader move away from viewing routers as static devices that always operate in a particular way and toward looking at the network as a programmable switching fabric that can be reconfigured from store-and-forward routers into virtual circuits as needed. For example, Internet2 (which, as noted earlier, is often held out as

proof of the engineering community's conviction that network management is unnecessary) now offers a service that it calls its Interoperable On-demand Network (ION) that allows researchers to establish dedicated point-to-point optical circuits to support large data transfers and other bandwidth-intensive applications. Internet2 notes that the "advanced science and engineering communities . . . are already straining against the limits of today's network capabilities—and capacities" and that advanced media and telepresence applications often need the type of dedicated circuits previously regarded as anathema (Internet2 2009).

Given the greater flexibility and functionality of today's routers and the increasingly intense demands being placed on them, there seems little reason to require that they always operate in a single, predetermined manner. That said, effective utilization of these new capabilities will doubtlessly require the development of new technical and institutional arrangements. Such innovations and changes may be inevitable if end users are to enjoy the full range of the network's technical capabilities.

G. Peha's Real-Time Secondary Markets for Spectrum

Another interesting example of a specialized service was proposed by Jon Peha in a paper co-authored with one of his graduate students before he became the FCC's Chief Technologist. The paper takes as its starting point the classic tradeoff between licensed and unlicensed uses of spectrum. Because exclusive licensing typically restricts access to a limited number of users, it tends to use spectrum inefficiently when those users connect to the network sporadically and the resource lays fallow whenever those particular parties are not using the network. At the same time, exclusive licensing does enable the network to offer guaranteed levels of quality of service. Unlicensed spectrum reverses this tradeoff. The fact that any number of users can share the same spectrum allows the resource to be used more efficiently. At the same time, unlicensed spectrum

is unable to provide guaranteed levels of quality of service, in part because the openness of the resource provides little incentive for users to conserve the amount of bandwidth used and in part because there is no way to limit the number of devices connected to the network (Peha and Panichpapiboon 2004).

Peha proposed a hybrid system that "offer[s] both the efficiencies of sharing with the possibility of quality of service guarantees." Under this approach, spectrum is exclusively licensed. Secondary users can request permission to share the spectrum for a fee, but the license holder would be allowed to deny access if the network is already saturated with prior calls. The ability to generate additional revenue without degrading existing sources of revenue makes it easier for networks to break even, which should promote buildout and help alleviate the digital divide. Although secondary users would receive lower priority, they benefit from paying prices estimated to be only one third of those paid by primary users (Peha and Panichpapiboon 2004).

In essence, this proposal is simply a form of prioritized service, in which the provider divides a single pipe into two tiers, each offering different levels of quality of service and different prices. The addition of a lower quality, lower price tier allows for more efficient use of the spectrum and makes service more affordable by allowing it to be offered at a lower price point. At the same time, it provides consumers who wish to run more demanding applications with the choice of a service able to offer better quality of service guarantees.

H. Low Extra Delay Background Transport (LEDBAT)

Low Extra Delay Background Transport (LEDBAT) is a new IETF congestion management initiative that shows tremendous promise. It is designed to address problems caused by applications that transmit large amounts of data over long periods of time. When this traffic passes through routers that forward on a first in, first out basis without engaging in any active

queue management, it imposes heavy delays on all other applications. LEDBAT is designed to address these problems by allowing this high volume, low priority traffic to avoid competing with other best-efforts traffic for its share of the available bandwidth. Instead, LEDBAT permits low priority traffic to step out of the way whenever it encounters any other traffic (Shalunov *et al.* 2011).

One example of high bandwidth, low priority application would be a service that allows end users to use the Internet to back up their hard disks to remote locations. The end user would likely not care if the service took several hours or even several days. Technologies like LEDBAT permit these end users to run these applications without taking up a disproportionate share of the available capacity or causing network congestion. Peer-to-peer applications similarly generate large amounts of traffic over sustained periods of time.

LEDBAT underscores the analytical emptiness of attempting to distinguish between prioritization and degradation. In essence, LEDBAT provides for a level of priority that is worse than best efforts routing. Whether or not it is regarded as degradation depends on what level of service is taken as the relevant baseline. While there is a temptation to regard the current level of service reflected in the current status quo as the natural baseline for comparison, the history of the Internet as well as ongoing debates in the engineering community reveal that there is nothing natural about this level. It is instead simply one of many choices made.

Approaches like LEDBAT reduce the cost of networking by allowing providers to offer higher levels of quality of service without having to expand network capacity. It permits a fairer allocation of bandwidth without requiring all providers to reconfigure their routers to actively manage their queues. It does represent a form of tiered service that will almost certainly involve different levels of pricing. Again, it underscores how permitting such differentials can benefit

consumers while simultaneously promoting the goals of increasing the amount of network capacity available to all citizens.

I. Internet Protocol Version 6 (IPv6)

At the time the Internet exploded onto the scene, the unifying protocol was the Internet Protocol version 4 (IPv4). The IPv4 header allocated 32 bits for addresses, which made it possible to assign approximately 4.3 billion unique addresses. At the time, the Internet was viewed as an academic rather than a mass-market phenomenon, and this amount of addresses was considered more than enough to satisfy all future needs. The Internet's commercial success has exhausted the available IPv4 addresses. For this reason, the network is transitioning to a next-generation protocol known as IP version 6 (IPv6), which by allocating 128 bits to the address field permits the allocation of 79 nonillion times more unique addresses. This is sufficient to provide a septillion addresses per square meter of the earth and is widely regarded as enough to satisfy the needs for the foreseeable future.

Dealing with IP address exhaustion was only one of the goals of the transition to IPv6. Another explicit goal was to provide greater support for real-time applications. Among the topics listed in the initial solicitation for white papers was the use of flows and resource reservation to provide better support for time-critical processes as well as support for policy-based routing (Bradner and Mankin 1993).

Consistent with this emphasis, the document creating IPv6 included within it an 8-bit field for *traffic class* to allow "originating nodes and forwarding routers to identify and distinguish between different classes or priorities of IPv6 packets." IPv6 was thus designed to support differentiated services in the same way as the type of service field in IPv4. In addition, the specification included a new field for *flow labels* (a concept developed for the network

technology known as Asynchronous Transfer Mode and now attached to MPLS). The document describes flow labels as enabling "send requests [for] special handling, such as non-default quality of service or 'real-time' service" (Deering and Hinden 1998: 2; see also Hinden 1996). Although experimental at that time, flow labels were fully specified in a later document (Rajahalme *et al*. 2004).[1]

The inclusion of these fields and the supporting documentation makes clear that quality of service is not a relic. Indeed, providing better support for quality of service (particularly for real-time data) was identified as one of the major goals of the transition to IPv6 (Bradner and Mankin 1993).

J. MetroPCS

Another example of non-neutrality is the new 4G offering by MetroPCS. This example is particularly important, as network neutrality proponents have already indicated their desire to challenge MetroPCS's practices.

MetroPCS is a regional wireless provider in the United States. In January 2011, MetroPCS revised the service plans for its new fourth generation (4G) wireless service. Its $40-per-month plan offered unlimited talk, text, and 4G Web browsing including unlimited YouTube access. Its $50-per-month plan added additional features[2] as well as 1 GB of "data access," defined to include multimedia streaming and video on demand services. Its $60-per-month plan offered unlimited data access (MetroPCS 2011a).

One week later, a consortium of advocacy groups submitted a letter calling for the FCC to investigate whether MetroPCS's proposed service plans violated the FCC's Open Internet

[1] An updated version was proposed in November 2011 and is still pending (Amante *et al.* 2011).

[2] These features included international and premium text messaging, GPS, mobile instant messaging, corporate e-mail, caller identity screening, and WiFi access to its service for full-track music downloads and premium video content (known as MetroSTUDIO).

order. Their primary complaint was that MetroPCS's $40 and $50 per month plans permitted unlimited access to YouTube, while categorizing other video services, such as Netflix, as data access subject to bandwidth limits (Free Press 2011). Consumers Union (2011) filed a similar letter 11 days later.[3]

Anyone evaluating these claims must take into account several realities. Specifically, MetroPCS controls far less spectrum than its rivals, typically deploying 4G on as little as 1.4 MHz of spectrum, while its rivals typically use 20 MHz of spectrum to offer 4G service. MetroPCS's limited spectrum resources mean that it has to be more innovative in offering a competitive service. For example, consumer requests led MetroPCS to search for ways to make YouTube available despite its bandwidth limitations. Because video delivered to mobile devices do not require the same resolution as full-sized television screens, MetroPCS was able to compress YouTube video so that it would work effectively despite MetroPCS's bandwidth limitations. Moreover, because MetroPCS was already offering unlimited YouTube access on its 1G data plans, it felt it had to include YouTube service in all of its 4G offerings. If new customers selected 1G over 4G, the increase in 1G traffic would overwhelm its 1G network and force the company to invest in an infrastructure it was planning to retire. MetroPCS emphasized that it facilitated access to YouTube in response to customer demand and that it lacked any financial arrangements that provided it with any incentive to favor YouTube. It also claimed that no other YouTube competitors had ever sought access to the MetroPCS network (MetroPCS 2011b).

[3] Some of these organizations also complained that MetroPCS's initial LTE deployments did not support VoIP because no VoIP clients were available for BREW. The arrival of an Android-based handset in early February 2011 allowed all MetroPCS 4G LTE customers to access VoIP so long as their handset was technically capable of supporting a VoIP client.

Moreover, according to the most recent FCC data (2011), MetroPCS had 8.2 million subscribers at the end of 2010, which represented less than 3 per cent of the market and lagged far behind Verizon (96 million), AT&T (94 million), Sprint (50 million), and T-Mobile (34 million), which are the four national providers. It is thus hard to see how any policy implemented by a firm of MetroPCS's size could hurt consumers or competition.

Lastly, MetroPCS deployed 4G through a device known as the Samsung Craft, a feature phone that is significantly cheaper and more limited than the typical smartphone. Unlike smartphones, which run open operating systems with open application programming interfaces that anyone can use to write applications, feature phones typically run proprietary operating systems that support a much narrower range of third-party software. Basing its service around feature phones inevitably means that MetroPCS's phones support a more limited range of applications than its competitors. As Tom Keys, MetroPCS's chief operating officer, said, "We didn't build this network or this device to be all things to all people" (Fitchard 2010). In an era in which spectrum is limited, increasing the competitiveness of the market depends on allowing wireless providers like MetroPCS to experiment with innovative forms of network management, especially for providers like MetroPCS that specialize in offering low-cost plans. Any evaluation of MetroPCS's compliance with the requirements of the Open Internet Order must take these technical realities into account.

* * *

This brief overview only touches on a few of the innovative ways that providers are deploying specialized services to use bandwidth more efficiently, reduce cost, and provide better service. Providers will need to become even more innovative as the universe of end users continues to become more heterogeneous and as market saturation causes providers to focus on

delivering greater value to each customer (Yoo 2012). Most importantly, reducing network costs can help promote the next generation of capacity expansion and reduce the digital divide by reducing the number of subscribers needed for an upgrade to the available capacity to break even.

Interestingly, these changes may be just the tip of the iceberg. The U.S. government, the EU, and university-based researchers are pursuing "clean slate" initiatives exploring how the architecture might differ radically if it were designed from scratch today. Within the scope of these projects are even more extensive usage of specialized services to deal with such emerging functionalities as security, mobility, and cloud computing (Pan *et al.* 2011).

CONCLUSION

The specific examples recounted above demonstrate how the details of the underlying technology can affect the assessment of a wide variety of types of network management. Far from being an aberration, such practices have been essential features that have been baked into the Internet's design since the very beginning. In addition, they demonstrate how network providers are experimenting with these techniques to provide affordable service in the face of rapid growth in network traffic, widescale deployment of applications that demand increasingly higher levels of quality of service, and severe limitations in the spectrum available to support wireless broadband services. Finally, the variety of possible technical solutions reveals the potential benefits of embracing a network diversity principle that permits network providers to experiment with a variety of different forms of network management unless and until the evidence indicates that those practices are harming consumers or competition (Yoo 2005). Internet scholars and regulatory authorities must deepen their appreciation for the technological

context surrounding these practices if they are to make an intelligent assessment of their likely impact and whether they represent good policy.

REFERENCES

Amante, S. *et al*. (2011) 'IPv6 flow label specification', Network Working Group Request for Comments 6437. Online. Available HTTP: http://tools.ietf.org/pdf/rfc6437 (accessed 15 February 2012).

Amtrak (n.d.) 'Journey with Wi-Fi'. Online. Available HTTP: www.amtrak.com/wi-fi (accessed 15 February 2012).

Bachula, G. (2006) *Net Neutrality: Hearing Before the Senate Committee on Commerce, Science, and Transportation*, 109th Congress, 2d Session 63–68.

Bazelon, C. (2008) *Oversight of the Federal Communications Commission—The 700 MHz Auction: Hearing Before the Subcommittee on Telecommunications and the Internet of the House Committee on Energy and Commerce*, 110th Congress, 2d Session. 185–210.

Blake, S. *et al*. (1998) 'An architecture for differentiated services', Network Working Group Request for Comments 2475. Online. Available HTTP: http://tools.ietf.org/pdf/rfc2475 (accessed 15 February 2012).

Blumenthal, M.S. (2002) 'End-to-end and subsequent paradigms', *Law Review of Michigan State University Detroit College of Law*, 709–717.

Braden, R. *et al*. (1994) 'Integrated services in the Internet architecture: An overview', Network Working Group Request for Comments 1633. Online. Available HTTP: http://tools.ietf.org/pdf/rfc1633 (accessed 15 February 2012).

Bradner, S. and Mankin, A. (1993) 'IP: Next generation (IPng) white paper solicitation', Network Working Group Request for Comments 1550. Online. Available HTTP: http://tools.ietf.org/pdf/rfc1550 (accessed 15 February 2012).

Cerf, V.G. and Kahn, R.E. (1974) 'A protocol for packet network intercommunication', *IEEE Transactions on Communications*, 22: 637–48.

Clark, D.D. (1988) 'The design philosophy of the DARPA Internet protocols', *ACM SIGCOMM Computer Communication Review*, 18(4): 106–14.

Comer, D.E. (2006) *Internetworking with TCP/IP, Volume 1: Principles, Protocols, and Architecture*, 5th edn, Upper Saddle River, N.J.: Pearson Prentice Hall.

Consumers Union (21 January 2011) Letter from Parul P. Desai and Mark Cooper to Julius Genachowski. Online. Available HTTP: http://fjallfoss.fcc.gov/ecfs/document/view.action?id=7021026388 (accessed 15 February 2012).

Deering, S. and Hinden, R. (1998) 'Internet Protocol, version 6 (IPv6) specification', Network Working Group Request for Comments 2460. Online. Available HTTP: http://tools.ietf.org/pdf/rfc2460 (accessed 15 February 2012).

Federal Communications Commission (2006) 'Implementing a nationwide, broadband, interoperable public safety network in the 700 MHz band', ninth notice of proposed rulemaking, *Federal Communications Commission Record*, 21: 14837–62.

——— (2007) 'Service rules for the 698–746, 747–762 and 777–792 MHz bands', second report and order, *Federal Communications Commission Record*, 22: 15289–575.

——— (2009) 'Preserving the open Internet', notice of proposed rulemaking, *Federal Communications Commission Record*, 24: 13064–170.

——— (2011) ‘Implementation of section 6002(b) of the Omnibus Budget Reconciliation Act of 1993’, fifteenth report, *Federal Communications Commission Record*, 26: 9664–971.

Fitchard, K. (21 September 2010) ‘LTE launches in the U.S.—MetroPCS style’, *Connected Planet*. Online. Available HTTP: http://connectedplanetonline.com/3g4g/news/metropcs-launches-lte-092110/ (accessed 15 February 2012).

Free Press (10 January 2011) Letter from M. Chris Riley to Julius Genachowski. Available HTTP: http://fjallfoss.fcc.gov/ecfs/document/view?id=7021025490 (accessed 15 February 2012).

Gillespie, T. (2006) ‘Engineering a principle: “End-to-end” in the design of the Internet’, *Social Studies of Science*, 36: 427–457.

Hinden, R.M. (1996) ‘IP next generation overview’, *Communications of the ACM*, 39(6): 61–71.

Huitema, C. (1995). *Routing in the Internet*, Englewood Cliffs, N.J.: Prentice Hall PTR.

Information Sciences Institute (1979) ‘DoD standard internet protocol’, Internet Engineering Note 123. Online. Available HTTP: http://128.9.160.29/ien/txt/ien123.txt (accessed 15 February 2012).

——— (1981) ‘Internet protocol: DARPA Internet program protocol specification’, Network Working Group Request for Comments 791. Online. Available HTTP: http://tools.ietf.org/pdf/rfc791 (accessed 15 February 2012).

Internet2 (2009), *Internet2 ION*. Available HTTP: http://www.internet2.edu/pubs/200909-IS-ION.pdf (accessed 15 February 2012).

Kirby, P. (21 April 2008) ‘Verizon Wireless, AT&T big winners in record auction of 700 MHz frequencies’, *Telecommunications Reports*, 3–_.

Leiner, B.M. et al. (1985) 'The DARPA Internet protocol suite', *IEEE Communications Magazine* 23(3): 29–34.

Lougheed, K. and Rekhter, J. (1989) 'A border gateway protocol (BGP)', Network Working Group Request for Comments 1105. Online. Available HTTP: http://tools.ietf.org/pdf/rfc1105 (accessed 15 February 2012).

MacKie-Mason, J.K. and Varian, H. (1994) 'Economic FAQs about the Internet', *Journal of Economic Perspectives*, 8(3): 75–96.

MetroPCS (3 January 2011a) 'MetroPCS' new 4G LTE plans offer unprecedented value and choice with prices starting at just $40'. Available HTTP: http://www.metropcs.com/presscenter/newsreleasedetails.aspx?id=1 (accessed 15 February 2012).

——— (14 February 2011b) Letter from Carl W. Northrop to Julius Genachowski. Online. Available HTTP: http://fjallfoss.fcc.gov/ecfs/document/view.action?id=7021029361 (accessed 15 February 2012).

Nissenbaum, H. and Price, M. (eds) (2004) *Academy and the Internet*, New York: Peter Lang.

Pan, J., Paul, S., and Jain, R. (2011) 'A survey of the research on future Internet architectures', *IEEE Communications Magazine*, 49(7): 26–36

Peha, J.M. and Panichpapiboon, S. (2004) 'Real-time secondary markets for spectrum', *Telecommunications Policy*, 28: 603–18.

PlusNet (19 April 2011a) 'Broadband: Your broadband speed – The basics'. Available HTTP: http://www.plus.net/support/broadband/speed_guide/speed_basics.shtml (accessed 15 February 2012).

——— (19 May 2011b) ‘Broadband: Broadband download speeds’. Available HTTP: http://www.plus.net/support/broadband/speed_guide/download_speeds.shtml (accessed 15 February 2012).

——— (8 July 2011c) ‘Broadband: All about traffic management’. Available HTTP: http://www.plus.net/support/broadband/speed_guide/traffic_management.shtml (accessed 15 February 2012).

——— (n.d.) *What we’ve won*. Available HTTP: http://www.plus.net/press/awards.shtml (accessed 15 February 2012).

Postel, J. (1981) ‘Service mapping’, Network Working Group Request for Comments 795. Online. Available HTTP: http://tools.ietf.org/pdf/rfc795 (accessed 15 February 2012).

Rajahalme, J. et al. (2004) ‘IPv6 flow label specification’, Network Working Group Request for Comments 397. Online. Available HTTP: http://tools.ietf.org/pdf/rfc3697 (accessed 15 February 2012).

Rekhter, J. (1989) ‘EGP and policy based routing in the new NSFNET backbone’, Network Working Group Request for Comments 1092. Online. Available HTTP: http://tools.ietf.org/pdf/rfc1092 (accessed 15 February 2012).

Rosen, E.C. *et al*. (2001) ‘Multiprotocol label switching architecture’, Network Working Group Request for Comments 3031. Online. Available HTTP: http://tools.ietf.org/pdf/rfc3031 (accessed 15 February 2012).

Sandvig, C. (2009) ‘How technical is technology research? Acquiring and deploying technical knowledge in social research projects’, in Hargittai, E. (Ed.) *Research Confidential*, Ann Arbor, MI: University of Michigan Press.

Shalnuov S. *et al*. (2011) Low extra delay background transport (LEDBAT), LEDBAT Working Troup Internet Draft. Online. Available http://datatracker.ietf.org/doc/draft-ietf-ledbat-congestion/ (accessed 15 February 2012).

Stallings W. (2001) 'MPLS', *Internet Protocol Journal*, 4(3): 2–14.

Tanenbaum, A.S. (2003) *Computer Networks*, 4th edn, Upper Saddle River, N.J.: Prentice Hall PTR.

Yoo, C.S. (2005) 'Beyond network neutrality', *Harvard Journal of Law and Technology*, 17: 1–77.

——— (2010) 'Innovations in the Internet's architecture that challenge the status quo', *Journal on Telecommunications and High Technology Law*, 8: 79–99.

——— (2012) *The Dynamic Internet: How Technology, Users, and Business Are Transforming the Network*, Washington, D.C.: AEI Press.

Zhu, K. (2007) 'Bringing neutrality to net neutrality', *Berkeley Technology Law Journal*, 22: 615–45.