11-6-2013

Wireless Networks: Technological Challenges and Policy Implications

Christopher S. Yoo
University of Pennsylvania Carey Law School

Wireless Networks: Technological Challenges and Policy Implications

Christopher S. Yoo*

ABSTRACT

Since June 2012, mobile wireless has emerged as the largest and fast growing medium for broadband service. At the same time, mobile wireless networks have proven considerably more difficult to manage than wireline networks. The primary causes are the rapid growth in demand for wireless bandwidth and the greater susceptibility of wireless networks to poor quality of service because of the omnidirectional propagation of wireless signals, bad handoffs, local congestion, and the susceptibility to complex interference patterns caused by multipath propagation. Moreover, the central inference underlying the primary form of congestion management is not valid for wireless networks. As a result, wireless networks adopt different approaches to error correction and congestion management than do wireline networks, which results in significantly heavier network management in ways that violate the Internet's commitment to the absence of per-flow state and its supposed adherence to the absence of prioritization.

In addition, mobile networks put significant pressure on the routing architecture by requiring the use of Internet gateways for 3G networks, accelerating the pace with which the routing architecture changes, fragmenting the compactness of the address space, and relying on a mobile IP solution that depends on a home agent to serve as a proxy in the core of the network. Proposed solutions, such as the identity/locator split, represent significant deviations from the universal address architecture around which the current architecture is designed. These considerations support the Federal Communications Commission's decision to subject wireless broadband to a less restrictive version of its rule against unreasonable discrimination in its Open Internet Order.

* 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 paper benefited from presentations at ETH Zurich and the Free State Foundation. Responsibility for any errors remains with the author.

INTRODUCTION

Network neutrality has dominated the debate over U.S. broadband policy for the past several years.[1] The initial stages of the debate focused almost exclusively on wireline communications, which was natural given that until recently the overwhelming majority of broadband connections occurred over a wireline technology, such as digital subscriber lines (DSL), cable modem systems, or fiber to the home (FTTH). In recent years, however, the telecommunications industry has become increasingly dominated by wireless technologies. With respect to conventional telephony, the number of U.S. wireless subscribers surpassed the number of wireline subscribers in 2004.[2] By 2008, the number of wireless subscribers more than

[1] For an overview of the early history of the debate, see Christopher S. Yoo, *Network Neutrality and the Economics of Congestion*, 94 GEO. L.J. 1847 (2006). For citations to more recent developments, see Daniel F. Spulber & Christopher S. Yoo, *Rethinking Broadband Internet Access*, 22 HARV. J.L. & TECH. 1, 4, 16–19 (2008).

[2] FCC Industry Analysis Technology Division, Wireline Competition Bureau, Local Telephone Competition: Status as of December 31, 2004, at 1, 3, 5 tbl.1, 17 tbl.13 (July 2005), *available at* http://www.fcc.gov/Bureaus/Common_Carrier/Reports/FCC-State_Link/IAD/lcom0705.pdf.

doubled the number of wireline subscribers, with more than twenty percent of all subscribers relying exclusively on their wireless phone for voice service as of mid-2009.[3]

Wireless broadband has followed a similar pattern, enjoying meteoric grown once they began to be widely deployed in 2005. According to the Federal Communications Commission (FCC), when measured at the lowest tier of service as of June 2012, U.S. mobile wireless broadband had captured 153 million subscribers, more than three times the number of those subscribing to the next largest technology, cable modem service, and was growing four times faster.[4] Even at the FCC's benchmark of 3 Mbps downstream and 768 kbps upstream, mobile broadband surpassed cable modem for the first time in June 2012.[5] The impending deployment of fourth generating wireless technologies such as Long Term Evolution (LTE) and the emergence of wireless as the leading broadband platform abroad both suggest that wireless broadband will become increasingly important in the years to come.[6]

The growing importance of wireless broadband naturally led regulators to show greater interest in how it should be regulated. The FCC initially took a hands-off approach, reflected in 2007 its decision characterizing wireless broadband as an "information service," a category long associated with deregulation,[7] and its refusal to rule on Skype's petition asking the FCC to mandate that all wireless broadband providers permit end users to run applications and attach nonharmful devices as they see fit.[8] Later that year, when preparing to auction off licenses to

---

[3] Implementation of Section 6002(b) of the Omnibus Budget Reconciliation Act of 1993, Fourteenth Report, 25 FCC Rcd. 11407, 11504 ¶ 155, 11603 ¶¶ 339–40 (2010).

[4] Fed. Commc'ns Comm'n, Internet Access Services: Status as of June 30, 2012, at 1 (May 2013), http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-321076A1.pdf.

[5] *Id.* at 25 tbl.7.

[6] *See Technological Determinism and Its Discontents*, 127 HARV. L. REV. (forthcoming 2014).

[7] Appropriate Regulatory Treatment for Broadband Access the Internet over Wireless Networks, Declaratory Ruling, 22 FCC Rcd. 5901, 5909–11 ¶¶ 22–27 (2007).

[8] Skype Communications S.A.R.L. Petition to Confirm a Consumer's Right to Use Internet Communications Software and Attach Devices to Wireless Networks, RM 11631 (filed Feb. 20, 2007), *available at* http://fjallfoss.fcc.gov/ecfs/document/view?id=6518909730.

the spectrum formerly used for analog television recovered following the transition to digital television, the FCC required that the winner of the spectrum band known as the C Block refrain from blocking, degrading, or interfering with end users' ability to run applications.[9] In so ruling, the FCC specifically noted that it had not yet decided whether to subject all wireless broadband networks to a similar requirement.[10]

The FCC began the process of addressing whether to impose network neutrality on wireless broadband networks in its October 2009 Notice of Proposed Rulemaking initiating the Open Internet proceeding by seeking comment as to whether the network neutrality rules should apply to wireless broadband.[11] After seeking additional comment specifically relating to wireless broadband,[12] the FCC issued its order in this proceeding in late December 2010 in which took the intermediate position of requiring that wireless broadband providers comply with the FCC's transparency and no-blocking rules, but exempting them from the nondiscrimination rules imposed on wireline broadband providers.[13]

The FCC's decision to exempt wireless broadband networks from the nondiscrimination mandate has proven quite controversial, drawing criticism from network neutrality advocates,[14] members of Congress,[15] and Democratic Commissioners.[16] As soon as the order was published

---

[9] Service Rules for the 698–746, 747–762 and 777–792 MHz Bands, Second Report and Order, 22 FCC Rcd. 15289, 15361–65 ¶¶ 195–206 (2007).

[10] *Id.* at 15363 n.463.

[11] Preserving the Open Internet, Notice of Proposed Rulemaking, 24 FCC Rcd. 13604, 13117–24 ¶¶ 154–174 (2009).

[12] Further Inquiry into Two Under-Developed Issues in the Open Internet Proceeding, 25 FCC Rcd. 12637, 12640–42 (2010). The FCC also solicited additional comments on specialized services. *Id.* at 12638–40.

[13] Preserving the Open Internet, Report and Order, 25 FCC Rcd. 17905, 17956–63 ¶¶ 93–107 (2010) [hereinafter Open Internet Order].

[14] *See, e.g.*, Tim Karr, *FCC Caves on Net Neutrality*, SAVE THE INTERNET, Dec. 21, 2010, http://www.savetheinternet.com/blog/10/12/21/fcc-caves-net-neutrality.

[15] *See* Alex Kingsbury, *FCC Sets Internet Rules*, U.S. NEWS, Dec. 24, 2010, at 7 (quoting Senator Al Franken as lamenting that exempting wireless broadband rendered the rules "worse than nothing").

[16] Open Internet Order, *supra* note 13, at 18047 (Copps, Comm'r, concurring); *id.* at 18082 (Clyburn, Comm'r, approving in part, concurring in part).

in the *Federal Register*, advocacy groups asked the U.S. Courts of Appeals to overturn the FCC's decision to treat wireless and wireline technologies differently.[17]

Thus far, the academic commentary has focused almost exclusively on the economics of wireless network neutrality, debating whether wireless broadband providers have the ability and incentive to use prioritization to harm competition.[18] While one can debate the economic merits of prohibiting discrimination and prioritization, to date none of the literature has grappled with whether applying the same rules to wireline and wireless broadband providers is even technically feasible.

An examination of the way wireless broadband networks actually works reveals that extending to wireless the prohibition on discrimination that the FCC developed for wireline technologies would raise serious problems. For example, wireless broadband networks manage congestion and reliability in a manner that is fundamentally different from the mechanisms used on the wireline Internet. The engineering literature is replete with observations listing the support for mobility as one of the key network functions that the current architecture does not perform well.[19] Indeed, the National Science Foundation's Future Internet Architecture program

---

[17] *See* Petition for Review, Access Humboldt v. FCC, No. 11-72849 (9th Cir. Sept. 26, 2011), *available at* http://accesshumboldt.net/site/files/AHvFCCPetitionForReview26September2011.pdf; Petition for Review, Free Press v. FCC, No. 11-2123 (1st Cir. Sept. 28, 2011), *available at* http://www.freepress.net/files/Petition_for_review.pdf.

[18] The debate was initiated by Tim Wu, *Wireless* Carterfone, 1 INT'L J. COMM. 389 (2007), *available at* http://ijoc.org/ojs/index.php/ijoc/article/view/152/96. For later discussions, see Babette E.L. Boliek, *Wireless Net Neutrality Regulation and the Problem with Pricing: An Empirical, Cautionary Tale*, 16 MICH. TELECOMM. TECH. L. REV. 1 (2009), http://www.mttlr.org/volsixteen/boliek.pdf; Rob Frieden, *Hold the Phone: Assessing the Rights if Wireless Handset Owners and Carriers*, 69 U. PITT. L. REV. 675 (2008); Robert W. Hahn, Robert E. Litan & Hal J. Singer, *The Economics of Wireless Net Neutrality*, 3 J. COMPETITION L. & ECON. 399 (2007); George S. Ford, et al., *Wireless Net Neutrality: From* Carterfone *to Cable Boxes* (Phoenix Ctr. for the Advanced Legal & Econ. Pub. Pol'y. Studies Pol'y Paper No. 17, Apr. 2007) *available at* http://www.phoenix-center.org/PolicyBulletin/PCPB17Final.pdf; Greg Rosston, An Antitrust Analysis of the Case for Wireless Network Neutrality (Stanford Institute for Economic Policy Research Discussion Paper No. 08-040, Aug. 2009), *available at* http://www.stanford.edu/group/siepr/cgi-bin/siepr/?q=system/files/shared/pubs/papers/pdf/08-040.pdf; Marius Schwartz & Federico Mini, *Hanging Up on* Carterfone*: The Economic Case Against Access Regulation in Mobile Wireless*, (May 2, 2007) (unpublished manuscript available at http://ssrn.com/abstract=984240).

[19] *See, e.g.*, Jon Crowcroft, *Net Neutrality: The Technical Side of the Debate*, COMPUTER COMM. REV., Jan. 2007, at 49, 51; Mark Handley, *Why the Internet Only Just Works*, 24 BT TECH. J. 119, 120 (2006); Raj Jain,

is sponsoring a project to explore how the Internet might need to be redesigned to accommodate wireless.[20]

Many of these solutions to these problems in the wireless world violate many central tenets of the Internet's architecture, either by changing the semantics or by changing the basic principles around which the Internet is currently designed. If adopted, such changes would reduce the interoperability of the network and create a much tighter integration between end users and the network. Even less transformative proposals are likely to affect different applications and end users differently and inevitably cause traffic on wireless and wireline networks to behave in a strikingly different manner. Understanding the technical space is thus essential to understanding whether differential regulatory treatment between wireline and wireless networks may be justified, precisely how broad the wireless exception might be, and what would be lost if no distinction were drawn between wireless and wireline networks.

I. BASIC INTERNET PRINCIPLES

A full appreciation of the ways in which wireless broadband networks deviate from the traditional architecture requires a basic understanding of the architectural commitments that were incorporated into the Internet's design. Only then is it possible to comprehend when deviations actually occur, what is motivating them, and how those changes might detrimentally affect the Internet. Some of these commitments fall outside the scope of this paper.[21] For our purposes, it suffices to focus on three in particular:

Internet 3.0: Ten Problems with Current Internet Architecture and Solutions for the Next Generation, PROC. MIL. COMM. CONF. (MILCOM 2006) (2007), http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4086425; Thrasyvoulos Spyropoulos et al., *Future Internet: Fundamentals and Measurement*, 37 COMPUTER COMM. REV., Apr. 2007, at 101; Sixto Ortiz, Jr., *Internet Researchers Look to Wipe the Slate Clean*, COMPUTER, Jan. 2008, at 12.

[20] MobilityFirst Future Internet Architecture Project, http://mobilityfirst.winlab.rutgers.edu/.

[21] One prime example is the idea of protocol layering. For a more complete discussion, see Christopher S. Yoo, Modularity Theory and Internet Policy (forthcoming 2014).

- the absence of per-flow state and
- the assignment of unique, universal address to every machine (known as an Internet Protocol address or IP address) visible to all other machines

Some commentators also argue that the Internet also reflects a fundamental commitment not to prioritize traffic based on its source, destination, content, or the application with which it is associated.

A. The Absence of Per-Flow State

One of the central commitments around which the Internet is designed is that the routers operating in the core of the network simply store packets as they arrive and forward them toward their final destination. Two corollaries of this principle are that each router makes its own decision about the direction to route any particularly packet and that each packet travels through the network independent of the packets preceding or following it in the data stream. This represented a sharp change from the architecture around which the telephone network was designed, which established dedicated circuits between end users and channeled all of the data associated with that communication along that circuit. The nodes in the core of such a circuit-switched network must necessarily retain a lot of information about each flow passing through the network. This information about where packets came from or where they are routed to is called *per-flow state*.[22]

The Internet's origins as a military network meant that the architects placed the highest priority on *survivability*, measured by the network's continuing ability to operate despite the loss of nodes within the network.[23] Networks that rely on a large amount of per-flow state tend not to be particularly robust in this manner. Consider what occurs when a switch in the middle of a

[22] Brian E. Carpenter, Architectural Principles of the Internet (IETF Network Working Group Request for Comments 1958, June 1996), *available at* http://tools.ietf.org/pdf/rfc1958 [hereinafter RFC 1958].

[23] David D. Clark, *The Design Philosophy of the DARPA Internet Protocols*, COMPUTER COMM. REV., Aug. 1988, at 106, 107.

telephone network fails. When the switch is lost, so is all of the information maintained by the switch with respect to each flow. The loss of this per-flow state means that neither the network nor the end user can recover from this event. As a result, the communication fails, and the only way to reestablish it is by placing a new call. Designing the network to avoid per-flow state in the core of the network increased the network's survivability.[24]

That said, some entity in connection with the communication has to keep monitoring it to see if it is delivered, and the failure of that entity necessarily causes the communication to fail. The Internet architects assigned responsibility for these function to the computers operated by end users at the edge of the network, called *hosts*. Their justification for having the hosts maintain per-flow state became known as *fate sharing*, which presumes that it is okay for the success of the communication to depend on the continuing survival of the sending and receiving hosts, since if those hosts collapse, there will likely be no remaining in completing the communication anyway.[25]

Although survivability represented the original justification for avoiding having routers operating in the core of the network maintain per-flow state, this rationale has little applicability to the modern Internet. While the loss of nodes may be a common occurrence in the hostile environments in which the military operates, the destruction of nodes is not a major concern in commercial networks.[26] Instead, the modern rationale for avoiding the maintenance of per-flow state in the core of the network is to facilitate the interconnection of networks that operate on very different principles.

The manner in which the absence of per-flow state facilitates interconnection is well illustrated by the history of the ARPANET, which represents of the predecessors of the Internet.

---

[24] *Id.* at 108.
[25] *Id.*; RFC 1958, *supra* note 22.
[26] Clark, *supra* note 23, at 107.

In the ARPANET, all of the routers operating in the core of the network (called Interface Message Processors or IMPs) were manufactured by a single company based on the same computer and ran the same software and were interconnected by the same technology (telephone lines).[27] The IMPs were responsible for a wide variety of tasks. For example, consistent with the standard approach of day,[28] IMPs were responsible for making sure that the packets were successfully delivered to the next IMP and, if not, for correcting any errors by resending the packets.[29] In addition, IMPs were responsible for congestion control.[30]

The result was that IMPs had to maintain a large amount of information about the current status of the packets passing through its network. Although these tasks were often quite complex, the fact that all IMPs were constructed of the same technology and operated on the same principles made them very easy to interconnect. The architects encountered greater problems when they attempted to interconnect the ARPANET with the two other packet network sponsored by the Defense Department: the San Francisco Bay Area Packet Radio Network (PRNET) and the Atlantic Packet Satellite Network (SATNET). Differences in transmission technologies, throughput rates, packet sizes, and error rates made these networks remarkably difficult to interconnect. In addition, every network would have to maintain the same state information as the other network with which it wanted to interconnect and would have to understand its expected response when receiving a communication from another router.[31]

---

[27] F.E. Heart et al., *The Interface Message Processor for the ARPA Computer Network*, 36 AFIPS CONF. PROC. 551, 551 (1970).

[28] *See* Geoff Huston, *The End of End to End?*, THE ISP COLUMN 1 (May 2008), http://www.potaroo.net/ispcol/2008-05/eoe2e.pdf (noting that the predominant approach to digital networking during the 1970s and 1980s required that each switch in a path store a local copy of the data until it received confirmation that the downstream switch has received the data).

[29] John M. McQuillan & David C. Walden, *The ARPANET Design Decisions*, 1 COMPUTER NETWORKS 243, 282 (1977).

[30] Christopher S. Yoo, *Protocol Layering and Internet Policy*, 161 U. PA. L. REV. 1707, 1758 (2013).

[31] JANET ABBATE, INVENTING THE INTERNET 121–22 (1999).

The International Network Working Group (INWG) considered a variety of solutions to these problems.[32] It rejected as too cumbersome and error-prone approaches that would require every host to run every protocol used by other types of networks simultaneously[33] or require each system to translate the communication into another format whenever it crossed a boundary between autonomous systems.[34] Instead, Cerf and Kahn established a single common language that all networks could understand.[35] To facilitate its use by multiple networks, this common language was kept as simple as possible and included only the minimum information needed to transmit the communication.[36] All of this information was placed in an internetwork header that every gateway could read without modifying.[37] The fact that all of the information needed to route a packet was contained in the packet itself eliminated the need for any router to know anything about the design of the upstream network delivering the packet to it or anything about the design of the downstream network to which it was delivering the packet.

This in turn meant that functions that used to be handled by routers (such as reliability) were now assigned to the hosts operating at the edge of the network. Even friendly observers have conceded that at the time this approach was regarded as "heresy,"[38] "unconventional,"[39] and "odd."[40] Over time, it has become an accepted feature of the network.

---

[32] *Id.* at 131–32.

[33] Vinton G. Cerf & Robert E. Kahn, *A Protocol for Packet Network Interconnection*, 22 IEEE TRANSACTIONS ON COMM. 637, 638 (1974) ("The unacceptable alternative is for every HOST or process to implement every protocol . . . that may be needed to communicate with other networks.").

[34] *See* ABBATE, *supra* note 31, at 128; Vinton G. Cerf & Peter T. Kirstein, *Issues in Packet-Network Interconnection*, 66 PROC. IEEE 1386, 1399 (1978).

[35] Cerf & Kahn, *supra* note 33, at 638.

[36] *See* Barry M. Leiner et al., *The DARPA Internet Protocol Suite*, IEEE COMM., Mar. 1985, at 29, 31 ("The decision on what to put into IP and what to leave out was made on the basis of the question 'Do gateways need to know it?'".).

[37] Cerf & Kahn, *supra* note 33, at 638–39.

[38] Huston, *supra* note 28, at _.

[39] ABBATE, *supra* note 31, at 125.

[40] Ed Krol & Ellen Hoffman. FYI on "What Is the Internet?," IETF Network Working Group Request for Comments 1462, May 1993), *available at* http://tools.ietf.org/pdf/rfc1462.

B. The Assignment of a Unique, Universal Address to Every Machine Visible to All Other Machines

The interconnection of different networks was also complicated by the fact that each network tended to employ its own idiosyncratic scheme for assigning addresses to individual hosts and routers.[41] The Internet's architects solves this problem by requiring that that all networks employ a single, uniform addressing scheme common to all networks.[42] They included the address information in the header of every IP packet so that every router could access the address information directly instead of having to maintain per-flow state. Moreover, hosts operating at the edge of the network must make their IP addresses visible to the rest of the network.[43]

C. The (Supposed) Absence of Prioritization/Quality of Service

Network neutrality advocates often assert that requiring routers not to prioritize traffic represents another fundamental commitment incorporated into the Internet's architecture.[44] As a matter of history, this claim is problematic.[45] Since its inception, the IP header has contained a six-bit *type of service field* designed to allow the attachment of different levels of priority to

---

41 Cerf & Kahn, *supra* note 33, at 637.

42 *See* Cerf & Kirstein, *supra* note 34, at 1393, 1399 (discussing the common internal address structure required for packet-level interconnectivity); Cerf & Kahn, *supra* note 33, at 641 ("A uniform internetwork TCP address space, understood by each GATEWAY and TCP, is essential to routing and delivery of internetwork packets.").

43 Tony Hain, Architectural Implications of NAT 7–8, 18 (IETF Network Working Group Request for Comments 2993), *available at* http://tools.ietf.org/pdf/rfc2993.

44 *See, e.g.*, Open Internet Order, *supra* note 13, at 17947 ¶ 76 ("pay for priority would represent a significant departure from historical and current practice"); LAWRENCE LESSIG, THE FUTURE OF IDEAS 37 (2002) (arguing that "the design effects a neutral platform—neutrality the sense that the network owner can't discriminate against some packets while favoring others").

45 *See* Clark, *supra* note 23, at 108 ("The second goal [of the DARPA architecture after survivability] is that it should support . . . a variety of types of service. Different types of service are distinguished by differing requirements for such things as speed, latency and reliability."); *see also* Kai Zhu, Note, *Bringing Neutrality to Net Neutrality*, 22 BERKELEY TECH. L.J. 615, 619–21, 634–38 (2007) (tracing the history of the engineering community's efforts to support quality of service).

particular packets.[46] The original design accommodated three levels of precedence as well as additional flags for particular needs regarding delay, throughput, and reliability, although subsequent changes allow this field to be used even more flexibly.[47]

Moreover, claims that the Internet is hostile toward prioritization ignore certain realities about the routing architecture. Indeed, enabling networks to engage in policy-based routing that alters the path that traffic takes based on its source or destination represented one of the principal motivations behind deploying Border Gateway Protocol (BGP), which remains the mechanism for routers to share routing information with one another on the Internet.[48]

Nor did efforts to support prioritization end there. Throughout the Internet's history, the IETF has issued standards designed to allow networks to provide differential levels of quality of service, including Integrated Services (IntServ),[49] Differentiated Services (DiffServ),[50] MultiProtocol Label Switching (MPLS),[51] and such modern initiatives as Low Extra-Delay Batch Transport (LEDBAT).[52] Providing better support for quality of service (particularly for

---

[46] Info. Sci. Inst., *Internet Protocol: DARPA Internet Program Protocol Specification* 8, 18, 35–36 (IETF Network Working Group Request for Comments 791, Sept. 1981), *available at* http://tools.ietf.org/pdf/rfc791; *see also* Info. Sci. Inst., *DoD Standard Internet Protocol* 12, 26–27 (Internet Engineering Note (IEN) 123, Dec. 1979), *available at* http://128.9.160.29/ien/txt/ien123.txt.

[47] ANDREW S. TANENBAUM, COMPUTER NETWORKS 434 (4th ed. 2003).

[48] Kirk Lougheed, A Border Gateway Protocol (BGP) 1 (IETF Network Working Group Request for Comments 1105, 1981), *available at* http://tools.ietf.org/pdf/rfc1105; CHRISTIAN HUITEMA, ROUTING IN THE INTERNET 195 (1995). A leading textbook gives the following examples of policy-based routing: "1. No transit traffic through certain [Autonomous Systems]. 2. Never put Iraq on a route starting at the Pentagon. 3. Do not use the United States to get form 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." TANENBAUM, *supra* note 47, at 460.

[49] *See* Robert Braden et al., *Integrated Services in the Internet Architecture: An Overview* (IETF Network Working Group Request for Comments 1633, July 1994), *available at* http://www.rfc-editor.org/rfc/rfc1633.pdf.

[50] *See* Steven Blake et al., *An Architecture for Differentiated Services* (IETF Network Working Group Request for Comments 2475, Dec. 1998), *available at* http://www.rfc-editor.org/rfc/pdfrfc/rfc2475.txt.pdf.

[51] *See* Eric C. Rosen et al., *Multiprotocol Label Switching Architecture* (IETF Network Working Group Request for Comments 3031, Jan. 2001), *available at* http://www.rfc-editor.org/rfc/pdfrfc/rfc3031.txt.pdf.

[52] *See Low Extra Delay Background Transport (LEDBAT) Working Group Charter*, INTERNET ENGINEERING TASK FORCE, http://www.ietf.org/html.charters/ledbat-charter.html.

real-time data) was identified as one of the major goals of the transition to IPv6.[53] Indeed, the IPv6 includes a *traffic class* filed that is analogous to the type of service filed in IPv4.[54]

To say that the desire for quality of service has long historical pedigree is not to say it has won the day. To be sure, just as quality of service has its advocates within the engineering community, it also has its detractors. My point is not to take sides in the debate. Indeed, if the presentations in the leading textbooks are network engineering are any guide, the controversy over quality of service shows no signs of abating, with many holding strong views on both sides of the argument.[55] My point is more limited. Those who support prioritization as the better solution will be untroubled by the fact the current regulatory regime permits wireless networks to prioritize traffic associated with certain applications over traffic associated with other applications. Those who are concerned about prioritization must bear in mind how traffic growth is adding new urgency to the arguments in favor of quality of service and how limiting wireless broadband providers' ability to prioritize creation applications over others risks reducing the functionality of the network.

II. DIFFERENCES IN TRAFFIC GROWTH AND BANDWIDTH CONSTRAINTS

One of the biggest challenges confronting wireless networks is the sharp increase in bandwidth consumption. Not only does the number of wireless broadband subscribers exceed the number of subscribers of all other broadband technologies combined.[56] Industry observers

---

[53] Scott Bradner & Allison Mankin, *IP: Next Generation (IPng) White Paper Solicitation* 4 (IETF Network Working Group Request for Comments 1550, Dec. 1993), *available at* http://tools.ietf.org/pdf/rfc1550; *accord* 1 DOUGLAS E. COMER, INTERNETWORKING WITH TCP/IP 563 (5th ed. 2006); LARRY L. PETERSON & BRUCE S. DAVIE, COMPUTER NETWORKS: A SYSTEMS APPROACH 319 (4th ed. 2007); TANENBAUM, *supra* note 47, at 465.

[54] Stephen E. Deering & Robert M. Hinden, Internet Protocol, Version 6 (IPv6) Specification 25 (IETF Network Working Group 2660), *available at* http://tools.ietf.org/pdf/rfc2460.

[55] *See* COMER, *supra* note 53, at 510, 515; JAMES F. KUROSE & KEITH W. ROSS, COMPUTER NETWORKING: A TOP-DOWN APPROACH 602–04 (5th ed. 2010).

[56] *See supra* note _ and accompanying text.

estimate that wireless traffic will grow at an annual rate of 66% from 2012 to 2017, as compared with a growth rate of 20% to 21% forecasted for other networks.[57] When traffic saturates the available capacity, packets are forced wait in queues. These queues become sources of jitter and delay, which degrades the quality of service provided by the network.

There are two classic approaches to managing explosive traffic growth. One solution is simply to increase network capacity.[58] The presence of additional headroom makes it less likely that spikes in traffic will saturate the network, which in turn allows the packets to pass through the network without any delay. The other solution employs network management to give a higher priority to traffic associated with those applications that are most sensitive to delay.[59]

For example, traditional Internet applications, such as email and web browsing, are essentially file transfer applications. Because file transfer applications typically display their results only after the last packet is delivered, delays in the delivery of intermediate packets typically do not adversely affect their performance. This stands in stark contrast with real-time, interactive applications, such as voice over Internet Protocol (VoIP), video conferencing, and virtual worlds, which are becoming increasingly important on the Internet. The performance of these applications depends on the arrival time and spacing of every intermediate packet, with delays of as little as one third of a second being enough to render the service unusable.[60] As such, these applications are considerably more vulnerable to network congestion.[61]

---

[57] *See* CISCO SYS., INC., CISCO VISUAL NETWORKING INDEX: GLOBAL MOBILE DATA TRAFFIC FORECAST UPDATE, 2012–2017, at 6 tbl.1 (2013), *available at* http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360.pdf.

[58] For a representative statement appearing in the engineering literature, see Yaqing Huang & Roch Guerin, *Does Over-Provisioning Become More or Less Efficient as Networks Grow Larger?*, *in* PROC. 13TH IEEE INT'L CONF. ON NETWORK PROTOCOLS (ICNP) 225 (2005). For similar statements appearing in the legal literature, see LESSIG, *supra* note 44, at 47.

[59] Christopher S. Yoo, *Beyond Network Neutrality*, 19 HARV. J.L. & TECH. 1, 21–23 (2005).

[60] International Telecommunication Union, ITU-T Recommendation G.114 (2003).

[61] The problem is most acute for interactive video, such as video conferencing. For linear video (whether prerecorded or live), media players can ameliorate the jitter caused by congestion by delaying playback to buffer a

Networks can help protect the operation of time-sensitive applications either by expanding capacity or by giving their packets a higher priority. In the latter case, it is conceivable that the network need only rearrange the order of the intermediate packets without affecting when the last packet will arrive. If so, network management can improve the performance of the time-sensitive application without having any adverse impact on the application that is less time sensitive. Even if small delays occur, with file-transfer applications, delays of a fraction of a second are virtually undetectable.

A review of leading textbooks reveals that the choice between these two approaches has long been a source of controversy in the engineering community with respect to wireline networks.[62] In the wireline context, engineering studies indicate that the amount of headroom needed to preserve quality of service without prioritization can be substantial.[63] Expanding bandwidth thus maintains simplicity, but requires the incurrence of significant capital costs. The additional cost associated with nonprioritized solutions increases the number of subscribers that a bandwidth expansion needs to breakeven, which in turn limits broadband deployment in ways that are likely to exacerbate the digital divide.[64] Network management substitutes operating costs for capital costs, which allows them to be recovered as they are incurred. It does have the side effect of adding complexity to the network.

However one strikes the balance between these two approaches in the wireline context, the tradeoff between these two approaches plays out much differently in the context of wireless networking. As an initial matter, wireless networks face limits on the number of end users that

---

quantity of packets so they may be released in a steady stream. Christopher S. Yoo, *The Changing Patterns of Internet Usage*, 63 FED. COMM. L.J. 67, 71 (2010).

[62] *See* COMER, *supra* note 53, at 510, 515; KUROSE & ROSS, *supra* note 55, at 602–04.

[63] *See* M. Yuksel et al., *Quantifying Overprovisioning vs. Class-of-Service: Informing the Net Neutrality Debate*, *in* 2010 PROC. 9TH INT'L CONF. ON COMPUTER COMM. & NETWORKS (2010), *available at* http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5560131&tag=1.

[64] Christopher S. Yoo, *Network Neutrality, Consumers, and Innovation*, 2008 UNIVERSITY OF CHICAGO LEGAL FORUM 179, 188, 229–32.

can be served in a particular area that wireline networks do not. A person connected to the Internet via a wireline technology (whether fiber, coaxial cable, or twisted pairs of copper) employs a signal that is narrowly channeled through space. This geographic limitation allows multiple end users to avoid interfering with one another even if they are sitting side by side.[65]

Wireless signals propagate quite differently. Unlike wireline signals, which travel in a confined path between the end user and the network node through which they are accessing the Internet, wireless signals propagate in an unchanneled manner in all directions.[66] The signals of one user are thus perceived as noise by other end users. As Claude Shannon recognized in 1948, the increase in noise reduces the amount of usable bandwidth available to those other users.[67] The greater the density of users becomes, the more constricted the bandwidth becomes. The implication is that there is an absolute limit to the density of end users who can use wireless broadband in any particular geographic area.[68]

Even more importantly, the options for wireless providers are much more limited than they are for wireline networking. Wireless providers can increase bandwidth by deploying a larger number of microwave base stations operating at lower power or by deploying increasingly sophisticated receiving equipment. Such solutions are typically quite costly. Moreover, the gains from such strategies are finite. Once they are exhausted, the restrictions on the amount of

---

[65] The fact that any electrical current creates some degree of radio frequency interference does mean that adjacent usage does create some interference. Any such interface occurs at very low power and can be minimized by proper shielding of the cables and the equipment.

[66] Piyush Gupta & P.R. Kumar, *The Capacity of Wireless Networks* , 46 IEEE TRANSACTIONS ON INFO. THEORY 388 (2000).

[67] Claude F. Shannon, *Communication in the Presence of Noise*, 37 PROC. INST. RADIO ENGINEERS 10 (1949).

[68] Gupta & Kumar, *supra* note 66, at _. Wireless operators can reduce this interference by using directional transmitters and receivers. Such solutions work only if you know the location of every sender and receiver. As such, they are poorly suited to wireless networking of mobile devices.

spectrum allocated to any particular service sharply limits network providers' ability to expand capacity any further.[69]

These bandwidth limitations dictate that wireless networks typically engage in extensive network management.[70] Specifically, if a subscriber in a low-bandwidth location is speaking on the telephone, the wireless network will prioritize the voice traffic and hold all email and other data traffic until the subscriber moves to a higher-bandwidth location or ends the call. A prohibition on prioritization based on applications would obstruct these types of network enhancements from being deployed. This approach requires tight integration of the network and the device. And as the FCC noted when repealing the regulation barring network providers from bundling telecommunications services with the devices used by end-user (also known as customer premises equipment or CPE), the equipment increasingly serve as enhancements to the network that requires sophisticated interactions between the network and the device that was being impeded by the unbundling requirement.[71] In other words, the device was part of the functionality of the network itself, and prohibitions on bundling devices and network services.

III. QUALITY OF SERVICE AND RELIABILITY

Another key difference between wireline and wireless broadband networks is their reliability. As anyone who has suffered through dropped calls on their mobile telephone recognizes, wireless technologies suffer much higher levels of packet loss than do wireline technologies. Part of the problem is the result of the difficulty of seamlessly handing off a communication when a mobile wireless user transfers from one base station to another. Other

---

[69] Charles Jackson et al., *Spread Spectrum Is Good—But It Does Not Obsolete* NBC v. U.S.!, 58 FED. COMM. L.J. 245, 253–59 (2006).

[70] *See* Charles L. Jackson, *Wireless Efficiency Versus Net Neutrality*, 63 FED. COMM. L.J. 445, 477 (2011).

[71] Policy and Rules Concerning the Interstate, Interexchange Marketplace, Report and Order, 16 FCC Rcd. 7418, 7427 ¶ 16 (2001).

problems are due to the physics of wave propagation, which cause interference in wireless networks to arise in much more transient and unpredictable ways than in wireline networks.

These differences in reliability in turn have implications for many basic architectural decisions in the Internet. For example, although the current network relies on hosts to correct errors by resending packets that are dropped, in a wireless world it is often more efficient to assign responsibility for those functions to routers operating in the core of the network. In addition, wireline networks rely on hosts to manage congestion on the Internet. For reasons discussed below, wireless networks' lack of reliability means that that the traditional approach to congestion management will not work well on wireless. The result is that such basic functions as recovery from errors and managing congestion—two of the most fundamental functions performed by the network—will operate far differently on wireless networks than on wireline networks.

A. Different Dimensions of Quality of Service

Most commentators discuss quality of service in terms of guaranteed throughput rates. As a preliminary matter, it bears mentioning that the engineering community typically views quality of service as occupying more dimensions than mere bandwidth. In addition, networks vary in terms of their reliability (i.e., the accuracy with which they convey packets), delay or latency (i.e., the amount of time it takes for the application to begin functioning after the initial request is made), and jitter (i.e., variations in the regularity of the spacing between packets).[72]

Interestingly, applications vary widely in the types of quality of service they demand. For example, the transfer of health records is not particularly bandwidth intensive and can accept millisecond latencies and jitter without much trouble, but is particularly demanding in terms of

[72] TANENBAUM, *supra* note 47, at 397.

reliability. Voice over Internet Protocol (VoIP) is also not bandwidth intensive and tolerates unreliability, but is quite sensitive to latency and jitter. Financial transactions have low bandwidth requirements, but must have latency guarantees in the microseconds and perfect reliability. Interactive video applications (such as video conferencing and virtual worlds) are bandwidth intensive and intolerant of jitter and latency, but can allow a degree of unreliability.

Interestingly, network systems can improve certain dimensions of quality of service, but only at the expense of degrading other dimensions. For example, streaming video works best when packets arrive in a steady stream. As a result, it is quite sensitive to jitter. Irregularities in the spacing between packets can be largely eliminated by placing all of the arriving packets in a buffer for some length of time and beginning to release them later. The presence of an inventory of backlogged packets allows them to be released in a nice even pattern. The cost, however, is to create a delay before the application begins to run.

B. Causes of Poor Quality of Service

Quality of service on wireless broadband networks can degrade for a wide variety of reasons not applicable to wireline networks. These reasons include bad handoffs between base stations, local congestion, and the physics of wave propagation.

1. Bad Handoffs

In order to receive service, a wireless device must typically establish contact with some base station located nearby. Circumstances may require a device to transfer its connection from one base station to another. For example, the mobile host may have moved too far away from the original base station. Alternatively, the current base station may have become congested or environmental factors may have caused the signal strength between the current base station and

the mobile host to have deteriorated.[73] For reasons discussed more fully below, transferring responsibility for a mobile host from one base station to another has proven to be quite tricky. It is not unusual for wireless networks to make a bad handoff, which can cause the communication to be dropped.

2. Local Congestion

In addition, because wireless technologies share bandwidth locally, they are more susceptible to local congestion than many fixed-line services, such as DSL and FTTH. Local congestion makes end users acutely sensitive to the downloading behavior of their immediate neighbors. Other technologies, such as cable modem systems, are also subject to local congestion. The more restrictive bandwidth limitations make this problem worse for wireless networks, as does the fact that wireless networks are typically designed so that data and voice traffic share bandwidth, unlike wireline telephone and cable modem systems which place their data traffic in a different channel from their core business offerings. As a result, wireless broadband networks are particularly susceptible to spikes in demand.

These limits have led many wireless providers rate limit or ban bandwidth intensive applications (such as video and peer-to-peer downloads) in order to prevent a small number of users from rendering the service completely unusable. For example, some providers using unlicensed spectrum to offer wireless broadband in rural areas have indicated that they bar users from operating servers for this reason.[74] Amtrak similarly blocks video and restricts large

---

[73] KUROSE & ROSS, *supra* note 55, at 581–82.

[74] *See, e.g.*, *Ensuring Competition on the Internet: Net Neutrality and Antitrust: Hearing Before the Subcomm. on Intellectual Prop., Competition, and the Internet of the H. Comm. on the Judiciary*, 112th Cong. 55 (2011) (prepared testimony of Laurence Brett ("Brett") Glass, Owner & Founder, Lariat).

downloads on its Acela trains, while permitting such traffic in its stations where bandwidth is less restricted.[75]

3. The Physics of Wave Propagation

Anyone who has studied physics knows that waves have some unique characteristics. They can reinforce each other in unexpected ways, as demonstrated by unusual echoes audible in some locations in a room and by whispering corners, where the particular shape of the room allows sound to travel from one corner to the other even though a person speaks no louder than a whisper. As noise-reducing headphones and cars demonstrate, waves can also cancel each other out. Waves also vary in the extent to which they can bend around objects and pass through small openings, depending on their wavelength. The discussion that follows is necessarily simplified, but is sufficient to convey the intuitions underlying some of the considerations that make wireless networking so complex.

The unique features of waves can cause wireless technologies to face interference problems that are more complex and fast-changing than anything faced by wireline technologies. For example, wireless signals attenuate much more rapidly with distance than do wireline signals, which makes bandwidth much more sensitive small variations in how distant a particular user is from the nearest base station. This requires wireless to allocate bandwidth by dynamically requiring individual transmitters to adjust their power. The physics of wireless transmission can also create what is known as the "near-far" problem, where a transmitter can completely obscure the signal of another transmitter located directly behind it by broadcasting

[75] Yoo, *supra* note 61, at 79 n.39.

too loudly.[76] WiFi networks similarly adjust the power of individual users dynamically to help allocate bandwidth fairly.[77]

Again, the solution is to require the nearer transmitter to reduce its power (and its available bandwidth) in order for the other transmitter to be heard.

Moreover, in contrast to wireline technologies, there is an absolute limit to the density of wireless users that can operate in any particular area. Shannon's Law dictates that the maximum rate with which information can be transmitted given limited bandwidth is a function of the signal-to-noise ratio.[78] Unlike wireline transmissions, which travel in a narrow physical channel, wireless signals propagate in all directions and are perceived as noise by other receivers. At some point, the noise becomes so significant that the addition of any additional wireless radios becomes infeasible.

Wireless transmissions also suffer from what are known as *multipath* problems resulting from the fact that terrain and other physical features can create reflections that can cause the same signal to arrive at the same location multiple times. Unless the receiver is able to detect that it is receiving the same signal multiple times, it will perceive multipathing as an increase in the noise floor that reduces the available bandwidth.

When reflections cause the same signal to arrive by different paths, the signal can arrive either in phase (with the peaks and the valleys of the wave form from the same signal arriving at exactly the same time) or out of phase (with the peaks and the valleys of the wave form from the same signal arriving at different times). When waves reflecting off a hard surface arrive in

---

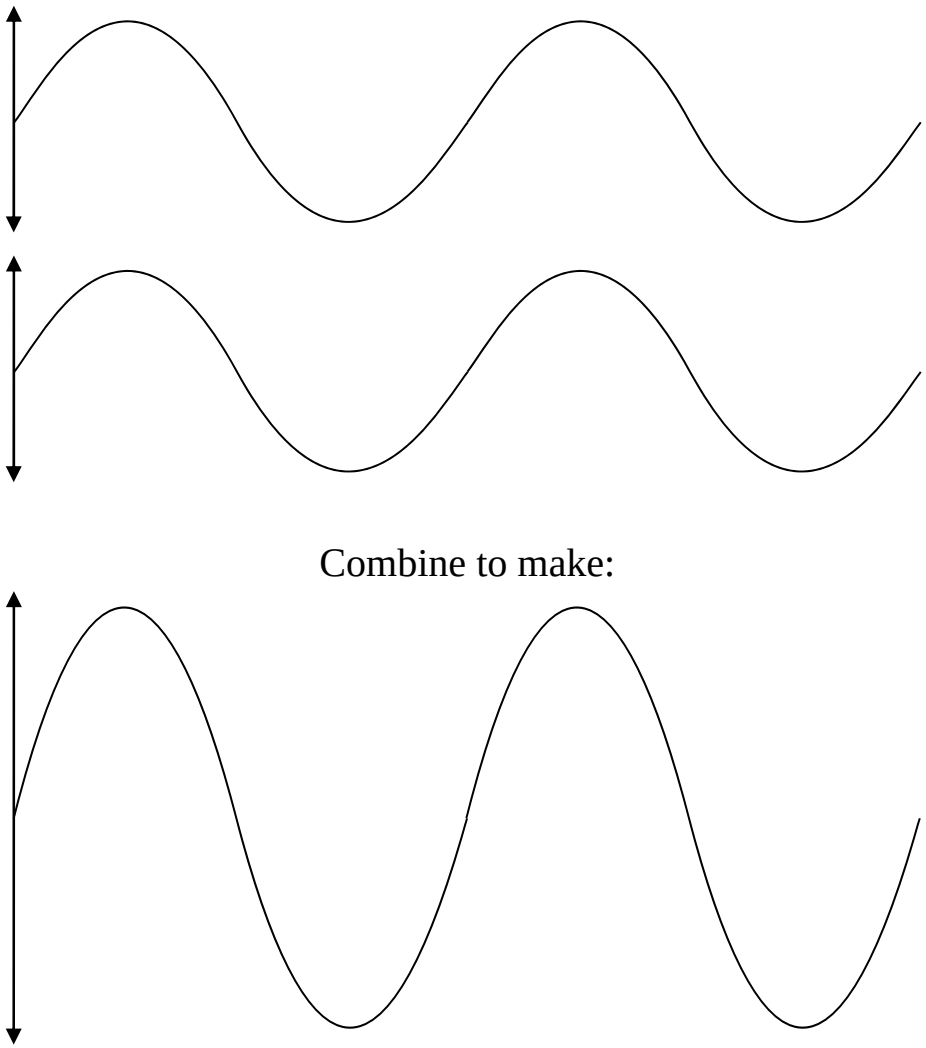
[76] *See, e.g.*, Mahesh K. Varanasi & Behnaam Aazhang, *Optimally Near-Far Multiuser Detection in Differentially Coherent Synchronous Channels*, 37 IEEE TRANSACTIONS ON INFO. THEORY 1006 (1991).

[77] *See, e.g.*, Huazhi Gong & JongWon Kim, *Dynamic Load Balancing Through Association Control of Mobile Users in WiFi Networks*, 54 IEEE TRANSACTIONS ON CONSUMER ELEC. 342 (2008).

[78] C.E. Shannon, *A Mathematical Theory of Communication* (pt. 1), 27 BELL SYS. TECH. J. 379 (1948); C.E. Shannon, *A Mathematical Theory of Communication* (pt. 2), 27 BELL SYS. TECH. J. 623 (1948).

phase, the signal reinforces itself, creating a localized hot spot in which signal is unusually strong.

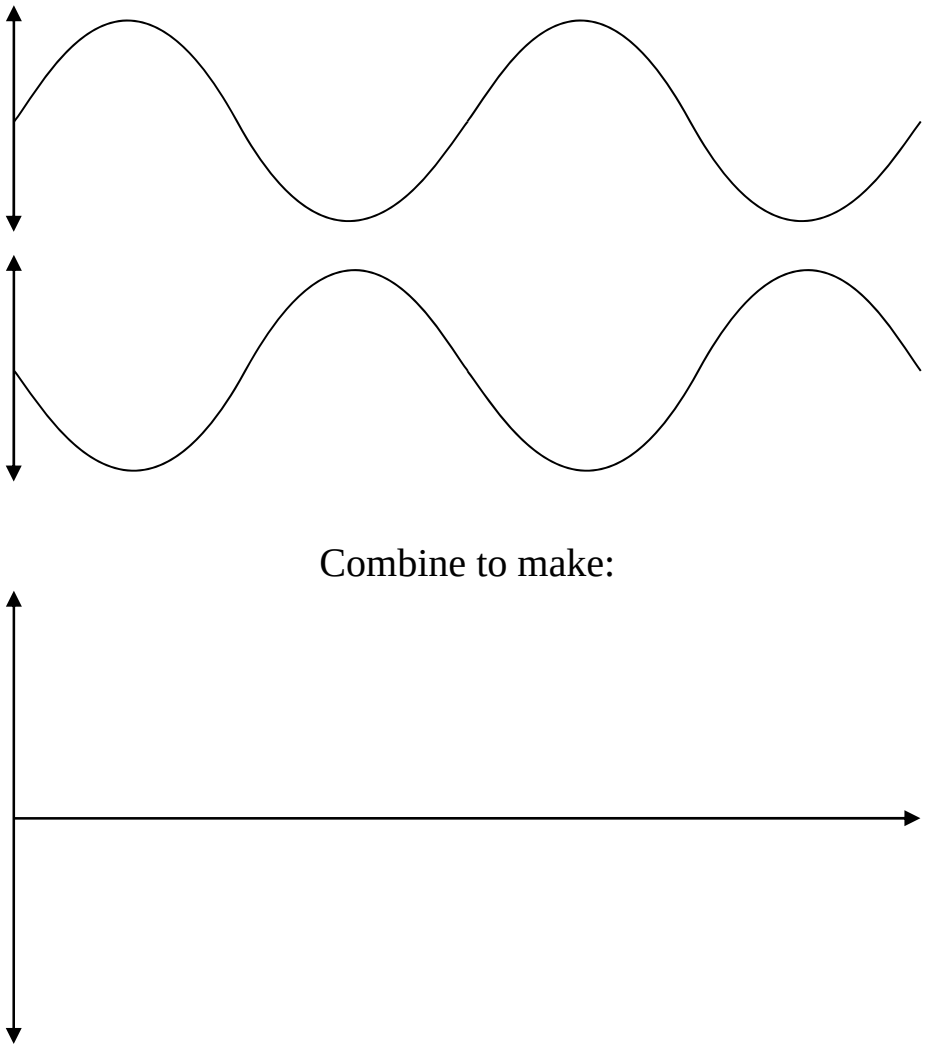
Figure 1: Reinforcement of Two Wave Forms That Are in Phase



When reflected waves arrive out of phase, they can dampen the signal. When they arrive perfectly out of phase (i.e., 180º out of phase), the reflection can create a dead spot by canceling

out the wave altogether. Although smart transmitters and receivers can avoid these problems if they know the exact location of each source and can even use the additional signal to extend the usable transmission range, they cannot do so if the receiver or the other sources are mobile devices whose locations are constantly changing.

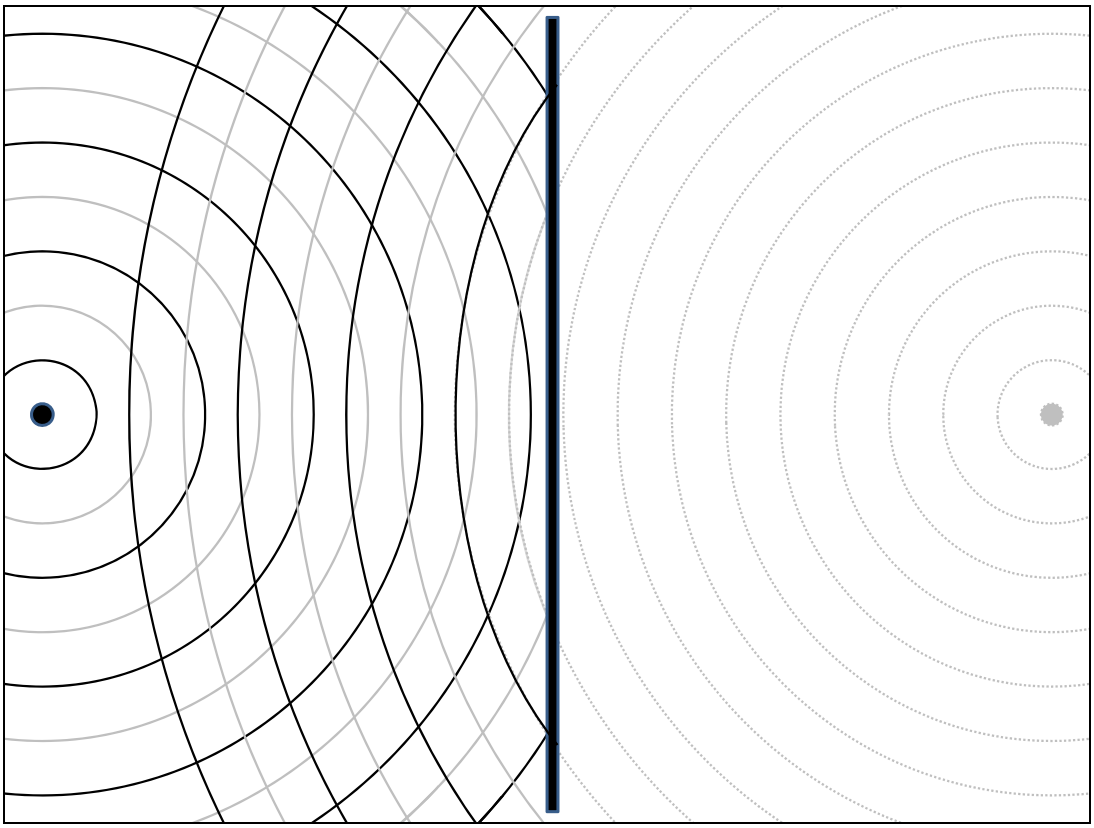
Figure 2: Cancellation by Two Wave Forms That Are 180º Out of Phase



A standard result in any physics textbook is that a reflection creates waves that are identical to a point source that is equidistantly located on the other side of the reflective surface.

The result is signal strength that is quite unpredictable. Consider the simple diagram in Figure 3, in which that the black circles represent the peaks of the wave form, while the grey circles represent the valleys. The points where two black circles or two grey circles cross represent hot spots where signals reinforce one another. The locations where a black circle crosses a grey circle represent dead spots where waves tend to cancel one another out.

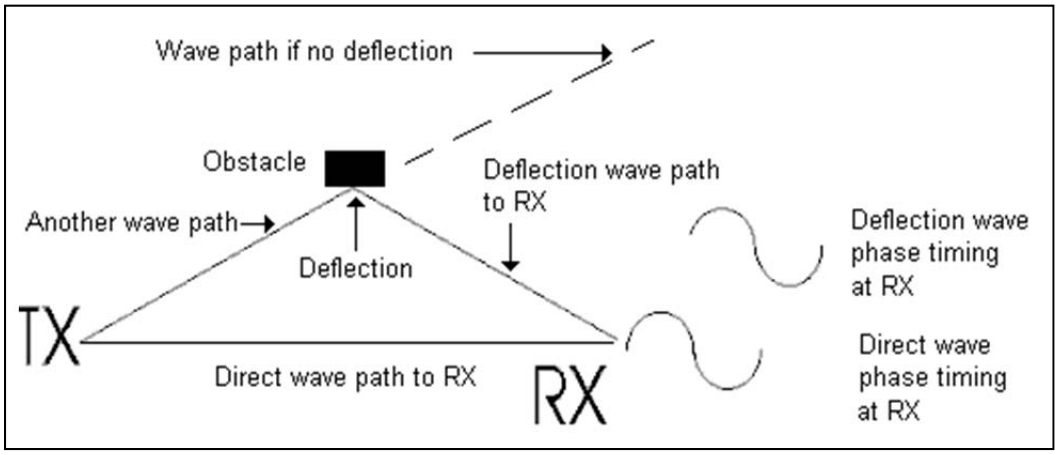
Figure 3: The Problem of Multipath Propagation



Obviously individuals traversing a room might pass through a variety of hot and cold spots. In addition, wave reflections can result not only from immobile objects, such as terrain and buildings, but also from mobile objects, such as cars and trucks. The result is that the amount of bandwidth available can change dynamically on a minute-by-minute basis. A participant at a May 2010 conference held at the University of Pennsylvania related a particularly vivid example of this phenomenon. While living in London, he had an apartment overlooking

the famous Speakers' Corner in Hyde Park. Thinking that those in the Speakers Corner might enjoy having WiFi service, he established a WiFi hotspot and pointed a directional antenna at the location only to find that his signal was intermittently blocked even though nothing ever passed directly between his apartment and the Corner. He eventually discovered that the interference arose whenever a double-decker bus was forced to stop at a nearby traffic light. Even though the bus did not directly obstruct with the waves travelling to and from the Speakers' Corner, it created a multipath reflection that periodically cancelled out the direct signal.[79]

Figure 4: The Problem of Multipath Propagation



Source: Dirk Grunwald

The result is that interference from other sources can be quite unpredictable and change rapidly from minute to minute. For these reasons, many wireless providers implement protocols that dynamically manage their networks based on the available bandwidth, giving priority to time-sensitive applications during times when subscribers are in areas of low bandwidth (such as

[79] Christian Sandvig, Associate Professor of Communication, University of Illinois, Remarks presented at the Center for Technology, Innovation and Competition's Conference on "Rough Consensus and Running Code: Integrating Engineering Principles into the Internet Policy Debates": How to See Wireless (May 7, 2010). For a description of the project, see PHILIP N. HOWARD, NEW MEDIA CAMPAIGNS AND THE MANAGED CITIZEN xi–xii (2006).

by holding back email while continuing to provide voice service). They have to do so much more aggressively and dynamically than do wireline providers.

C. Implications

1. Error Correction

Wireless networks sometimes run afoul of the standard approach to ensuring reliability on the wireline Internet. The standard approach to error correction in the Transmission Control Protocol (TCP) calls for every host to set a retransmission timer based on the expected roundtrip time between the sending host and the receiving host.[80] Receiving hosts are supposed to send acknowledgements for every packet they successfully receive. If the sending host does not receive an acknowledgment when its retransmission timer expires, it resends the packet and repeats the process until it is successfully transmitted.[81]

In many ways, relying on feedback loops and end-to-end retransmission is quite inefficient. Resending packets from the source requires the consumption of significant network resources. In addition, waiting for the retransmission timer to expire can cause significant delays. Such overhead costs become higher as the packet loss rates increase. If loss rates become sufficiently high, it may make sense for networks to employ network-based error recovery mechanisms instead of relying on end-to-end error recovery. For the reasons stated above, wireless networks tend to be considerably less reliable than wireline networks. For example, PRNET employed a network-based reliability system known as forward-error correction.[82] The higher loss rates in wireless technologies also explains why wireless broadband networks are increasingly deploying network-based reliability systems, such as

---

80 TANENBAUM, *supra* note 47, at 552.

81 *Id.*

82 Robert E. Kahn et al., *Advances in Packet Radio Technology*, 66 PROC. IEEE, 1468, 1492 (1978).

Automatic Repeat reQuest (ARQ), that detect transmission errors and retransmit the missing data from the core without waiting for the host-based retransmission timer to expire and without consuming the additional network resources needed to retrieve the packet all the way from the host.[83] Other techniques that allow routers in the core to participate in the transport layer exist as well.[84]

2. Congestion Management

The lack of reliability also requires that wireless technologies employ a significantly different approach to managing congestion. The primary mechanism for controlling congestion on the Internet was developed in the late 1980s shortly after the Internet underwent a series of congestion collapses. As noted earlier, TCP requires that receiving hosts send acknowledgments every time they successfully receive a packet. If the sending host does not receive an acknowledgement within the expected timeframe, it presumes that the packet was lost and resends it.[85] The problem is that the host now has sent twice the number of packets into a network that was already congested. Once those packets also failed to arrive, the host introduced still another duplicate packet. The result was a cascade that brought the network to a stop.

Because congestion is a network-level problem that is the function of what multiple end users are doing simultaneously rather than the actions of any one end user, some proposed addressing it address it through a network-level solution, as was done in the original ARPANET, networks running asynchronous transfer mode (ATM), and many other early corporate

---

[83] KUROSE & ROSS, *supra* note 55, at 219–27; TANENBAUM, *supra* note 47, at 208–11.

[84] *See* TANENBAUM, *supra* note 47, at 553–55 (exploring indirect TCP and the inclusion of snooping agents as possible solutions to the problem).

[85] *Id.* at 552.

networks.[86] However, the router hardware of the time made network-based solutions prohibitively expensive. On the other hand, hosts can also stop congestion collapse if they cut their sending rates in half or more whenever they encounter congestion. The problem is that congestion is the product of what multiple hosts are doing, whereas any individual host only knows what it is doing. Thus the hosts operating at the edge of the network typically lack the information to know when the network is congested.

Van Jacobson devised an ingenious mechanism by which hosts operating at the edge of the network can infer when the core of the network has become congested based on the information they were able to see.[87] Jacobson noted that packet loss typically occurs for only two reasons: (1) transmission errors or (2) discard by a router where congestion has caused its buffer to become full. Because wireline networks rarely drop packets due to transmission errors, hosts operating at the edge of the network could infer that the failure to receive an acknowledgement within the expected time was a sign of congestion and take this as a signal to reduce congestion by slowing down their sending rates exponentially.[88]

The problem is that this inference is invalid for wireless networks, which drop packets due to transmission error quite frequently, either because of a bad handoff as a mobile user changes cells or because of the interference problems discussed above. When a packet is dropped due to a transmission error, reducing the sending rate exponentially only serves to degrade network performance. Instead, the sending host should resend the dropped packet as quickly as possible without slowing down. In other words, the optimal response for wireless networks may well be the exact opposite of the optimal response for wireless networks.

---

[86] Raj Jain et al., Digital Equip. Corp., Congestion Avoidance in Computer Networks with a Connectionless Network Layer 6–7 (1997), *available at* http://www1.cse.wustl.edu/~jain/papers/ftp/cr5.pdf.

[87] Van Jacobson, *Congestion Avoidance and Control*, 18 COMPUTER & COMM. REV. 314 (1988).

[88] *Id.* at 319.

D. Solutions

In short, the deployment of wireless broadband is putting pressure on the traditional mechanisms for managing error correction and congestion, two of the most basic functions performed by the network. The higher loss rates make edge-based and feedback-based error recovery more expensive and make it impossible to regard packet loss as a sign of congestion.

As a result, the engineering community is experimenting with a variety of alternative approaches.[89] One approach allows local recovery of bit errors through some type of forward error recovery.[90] One such solution places a "snoop module" at the base station that serves as the gateway used by wireless hosts to connect to the Internet that keeps copies of all packets that are transmitted and monitors acknowledgments passing in the other direction. When the base station detects that a packet has failed to reach a wireless host, it resends the packet locally instead of having the sending host do so.[91] A second approach calls for the sending host to be aware of when its transmission traverses wireless links. Dividing the transaction into to two internally homogeneous sessions makes it easier to infer the current status of the network.[92] A third approach splits the wireless and the wireline approaches into separate TCP or UDP session.[93]

Many of these approaches violate the semantics of TCP, since the packets are not addressed to the receiving hosts. Many of them introduce intelligence into the core of the network and violate the principle of avoiding per-flow state. The split connection approach

---

89 KUROSE & ROSS, *supra* note 55, at 586; TANENBAUM, *supra* note 47, at 553–54.

90 Ender Ayanoglu et al., *AIRMAIL: A Link-Layer Protocol for Wireless Networks*, 1 WIRELESS NETWORKS 47 (1995).

91 Hari Balakrishnan et al., *Improving Reliable Transport and Handoff Performance in Cellular Wireless Networks*, 1 WIRELESS NETWORKS 469 (1995).

92 Ajay Bakre & B.R. Badrinath, *I-TCP: Indirect TCP for Mobile Hosts*, 1995 PROC. INT'L CONF. ON DISTRIB. COMPUTING SYS. (ICDCS) 136; Hari Balakrishnan et al., *A Comparison of Mechanisms for Improving TCP Performance Over Wireless Links*, 5 IEEE/ACM TRANSACTIONS ON NETWORKING 756 (1997).

93 Wei Wei et al., *Inference and Evaluation of Split-Connection Approaches in Cellular Data Networks*, PROC. ACTIVE & PASSIVE MEASUREMENT WORKSHOP (2006); Raj Yavatkar & Namrata Bhagwat, *Improving End-to-End Performance of TCP over Mobile Internetworks*, PROC. WORKSHOP ON MOBILE COMPUTING SYS. & APPLICATIONS 146 (1994).

violates the principle of end-to-end connectivity. All of them require introducing traffic management functions into the core of the network to a greater extent than originally envisioned by the Internet's designers.

IV. THE HETEROGENEITY OF DEVICES

Wireless technologies do not vary only in terms of transmission technologies. For example, Verizon's wireless broadband network is based on a protocol known as Evolution-Data Optimized (EV-DO) operating in the traditional cellular portion of the spectrum. Sprint's wireless broadband network also employs EV-DO, but operates in the band of spectrum originally allocated to the second generation wireless technology known as Personal Communications Services (PCS). AT&T's wireless broadband networks use a different format known as High Speed Packet Access (HSPA). Each of them has different technical characteristics. Indeed, the greater compatibility of HSPA with the iPhone is part of what led Apple initially to deploy the iPhone exclusively through AT&T.

Instead of relying on a personal computer, wireless broadband subscribers connect to the network through a wide variety of smart phones. These devices are much more sensitive to power consumption than are PCs, which sometimes leads wireless network providers to disable certain functions that shorten battery life to unacceptable levels, for example because they either employ analog transmission or search constantly for an available connection. In addition, wireless devices have much less processing capacity and employ less robust operating systems that do the laptop and personal computers typically connected to wireline services. As a result, they are more sensitive to conflicts generated by multiple applications, which can cause providers to be much more careful about which applications to permit to run on them. This

compels wireless broadband networks to manage devices and applications to a greater extent than wireline networks.

Wireless devices also tend to be much more heterogeneous in terms of operating systems and input interfaces (including keyboards and touch screens). As a result, the dimensions and levels of functionality offered by particular wireless devices vary widely. It seems too early to predict with any confidence which platform or platforms will prevail. Furthermore, as noted earlier, many wireless networks address bandwidth scarcity by giving a higher priority to time-sensitive applications, which typically requires close integration between network and device. These features underscore the extent to which variations in particular devices are often an inextricable part of the functionality of the network.[94]

V. ROUTING

Routing on wireless broadband is also very different from wireline networks. Because of their technical aspects, wireless does things differently. In the process, wireless violates the principles of unique universal addresses and simple store and forward routing.

A. The Use of Internet Gateways

Recall that one of the Internet's foundational principles is that each host connected to the Internet has a unique IP address that is visible and accessible to all other hosts. In addition, all of the routers within the network are supposed to route on the basis of this address.

It bears mentioning that until recently, wireless networks have not routed traffic in this manner. Unlike devices connected to wireline networks, which have IP addresses that are visible to all other Internet-connected hosts, wireless devices do not have IP addresses. Instead, Internet

[94] Charles L. Jackson, *Wireless Efficiency versus Net Neutrality*, 63 FED. COMM. L.J. (forthcoming March 2011).

connectivity is provided by an IP gateway located in the middle of the network that connects to individual wireless devices using a legacy telephone-based technology rather than IP. Stated in technical terms, wireless broadband devices operate at Layer 2 rather than Layer 3 of the Internet protocol stack. This means that all current wireless devices do not have the end-to-end visibility enjoyed by true Internet-enabled devices. They also necessarily depend on a virtual circuit between the Internet gateway and the wireless device. Wireless devices will eventually connect through the Internet protocol once fourth-generation wireless technologies such as LTE are deployed. Until that time, wireless devices necessarily will connect to the Internet on different and less open terms than devices connected through wireline networks.

This violates the principle that each device have a unique IP address that is visible to all others. It also route traffic through the last connection based on a different address system and on principles that may deviate from store and forward. Simply put, traffic bound for and received from wireless devices will not pass through the network on the same terms as traffic going to and from hosts connected to the network through wireline technologies.

All of this will change with the deployment of fourth-generation wireless technologies, such as Long Term Evolution (LTE). Unlike third-generation wireless technologies, LTE does route traffic based on IP addresses. Until that occurs, wireless and wireline traffic will travel through the network on distinctly different terms.

B. Acceleration in the Pace of Changes in Routing Architecture

Another feature of the current routing architecture is that it is updated on a decentralized basis. Every backbone router periodically informs its adjacent neighbors of the best routes by which it can reach every location on the Internet. This means that initially any changes to the network architecture will only be advertised locally. During the next update cycle, routers that

have been informed of the change will inform the routers located the next level away. Over time, the information will spread out in all directions until the entire network is aware of the change. When this occurs, the routing table is said to have reached equilibrium.

Before the routing table has reached equilibrium, however, some parts of the network may not know of certain changes that have occurred in other parts of the network. Suppose, for example, that one host in one corner of the network drops off the network. A host in a distant corner will not find out about that for quite some time. In the meantime, it could keep sending packets to a host that is no longer there, which wastes resources and unnecessarily adds to network congestion.

The efficient functioning of the network thus depends on the routing architecture being able to reach equilibrium. Whether it does so is largely a function of the speed with which locations change compared to the speed with which information about that change can propagate through the entire network. Moreover, the current architecture is built on the implicit assumption that Internet addresses change on a slower timescale than do communication sessions. So long as the address architecture changes at a slower timescale, any particular Internet-based communication may take the address architecture as given.

Mobility, however, increases the rate at which the address architecture changes. In addition, because addressing is handled on a decentralized basis, information about changes in the address architecture takes time to spread across the Internet. Increases in the rate with which the address space changes can cause communications sessions to fail and create the need for a new way to manage addresses.

C. Compactness of the Address Space

As a separate matter, wireless technologies are also causing pressure on the way the amount of resources that the network must spend on keeping track of Internet addresses. To understand why this is the case, one must keep in mind that routers typically follow one of two strategies in keeping routes. Some routers keep *global routing tables* that identify the outbound link that represents the most direct path to every single host on the Internet. Other routers avoid the burden of maintaining complete routing tables by only keeping track of a limited number of paths. All traffic bound for locations for which this router does not maintain specific information is sent along a *default route* to a *default router*, which is responsible for identifying the route for delivering all other traffic to its final destination.

The presence of default routes in a routing can give rise to a potential problem. For example, routers using default routes could point at one another (either directly or in a loop), which would cause the packets to pass back and forth indefinitely. The Internet ensures that traffic does not travel indefinitely through the network by assigning a *time to live* to each packet that limits the total number of hops that any packet may traverse before dropping off the network. Eventually, any packet caught in such a cycle will reach its maximum and drop off the network.[95]

The best way to prevent such roads to nowhere is to ensure that at least some actors maintain global routing tables, which by definition are routing tables that do not include any default routes. This role is traditionally played by the major backbone providers (also known as

[95] Paul Milgrom et al., *Competitive Effects of Internet Peering Policies*, *in* THE INTERNET UPHEAVAL 175, 179–80 (Ingo Vogelsang & Benjamin M. Compaine eds., 2000).

Tier 1 ISPs). Indeed, more than the economic relationships (such as peering), many regard the maintenance of default free routing tables as the defining characteristic of Tier 1 ISPs.[96]

Maintaining a global routing table that maintained a separate entry for the best path to every location on the Internet proved to be very difficult. The growth of the Internet meant that the size of the routing table was growing at a very fast rate. In fact, it grew faster than the routers could keep up.[97]

The solution was an innovation called Classless InterDomain Routing (CIDR).[98] The important part for our purposes is that CIDR allowed routers to use *route aggregation* to prevent routing tables from growing out of control. This mechanism can be illustrated by analogy to the telephone system. Consider a party in Los Angeles who is attempting to call the main telephone number for the University of Pennsylvania, which is (215) 898-5000. So long as all phones with phone numbers in the 215 area code are located in Philadelphia and all traffic bound for Philadelphia exist Los Angeles on the same link, a phone switch in Los Angeles could represent all telephone numbers in that area code ((215) xxx-xxxx) with a single entry in its routing table. Indeed, one can think of all ten million telephone numbers in the 215 area code as lying within the cone of telephone numbers represented by that entry.

Similarly, so long as all telephone numbers in the 898 directory within the 215 area code are connected to the same central office, switches within Philadelphia need not maintain separate entries for each phone number in that directory. Instead, they can represent the cone of all ten thousand telephone numbers located in (215) 898-xxxx with a single entry.

---

[96] Peyman Faratin et al., *The Growing Complexity of Internet Interconnection*, 72 COMM. & STRATEGIES 51, 54 (2008).

[97] Huston, *supra* note 28, at _.

[98] Yoo, *supra* note 61, at 82.

CIDR adopts a similar strategy to reduce the size of the routing tables maintained by Tier 1 ISPs. For example, the University of Pennsylvania has been assigned all of the addresses in the 128.91.xxx.xxx prefix (covering 128.91.0.0 to 128.91.255.255). Various locations have individual addresses falling within this range, with the main website for the University of Pennsylvania being covered by 128.91.34.233 and 128.91.34.234. Assuming that all of the hosts associated with these IP addresses are located in the same geographic area, a Tier 1 ISP could cover all of the one million addresses within this prefix with a single entry.

The success of this strategy depends on the address space remaining compact. In other words, this approach will fail if the 215 area code includes phone numbers that are not located in Philadelphia. If the telephones associated with those numbers sometimes lie outside the Philadelphia area, the telephone company will have to maintain separate entries in its call database for all phones located outside the area. Similarly, if some hosts with the 128.91.xxx.xxx prefix reside outside the Philadelphia area, Tier 1 ISPs will have to track those locations with additional entries in their routing tables.

The advent of mobile telephony and mobile computing means, of course, that telephones and laptops will often connect to the network outside their home locations. This in turn threatens to cause the routing tables to grow faster again. Other developments, including multihoming, the use of provider independent addresses, and the deployment of IPv6, are also placing upward pressure on the routing table. That said, wireless broadband remains a major cause.

D. Mobile IP

The most straightforward approach to addressing mobility is to assign a mobile host a new IP address whenever it changes location. This would put a lot of strain on the network by requiring that it inform the rest of the network about the change. To the extent that it disrupts the

compactness of the address space, it can cause the put pressure on the routing architecture by causing the routing table to grow. In addition, dynamically changing IP addresses in the middle of an application can cause many applications to fail.[99]

How, then, do we handle mobility without having to update the routing tables constantly and without cause the size of routing tables to grow out of control? The Internet currently solves these problems through a regime known as *mobile IP*. Under mobile IP, each mobile user has a *home network*, with all other network being called *foreign network*. The mobile host designates a router located on its home network as the contact point for all IP-based communications directed to the mobile host. This contact point is called the *home agent*. Anyone seeking to contact the mobile host (called the *correspondent*) simply sends the packets to the home agent, which then forwards the communication to the mobile host. If the mobile host moves from one foreign network to another, it simply notifies its home agent, which the routes any new packets it receives to the new location.

Although this solution sounds relatively simple, actually implementing can be quite complex. For example, the home agent has to know to where the mobile host is currently located. This is relatively easy when the mobile host initiates the transaction. It is more complicated when a third party is attempting to contact the mobile host. Stated in the example of mobile telephony, networks can easily discover where a particular cellular user is located when it is that user that is imitating the call. The simple fact of establishing contact with the local microwave tower announces the location. The situation is different when the mobile user is receiving the call. To terminate this call, the network has to know where the mobile user is even when it is just sitting around waiting.

---

[99] PETERSON & DAVIES, *supra* note 53, at 290.

This means that if a mobile host is to receive traffic, it must constantly announce to the network serving its current location so that that network knows that it is there. This can be accomplished by designating a router located on the foreign network as the *foreign agent* responsible for managing mobile IP. Every mobile host must regularly register with the foreign agent serving the local foreign network in order to receive communications. This can happen by the foreign agent sending an advertisement notifying mobile nodes located in its service area that it is prepared to facilitate mobile IP or by the mobile node sending a solicitation to see if any foreign agents are located nearby capable of supporting mobile IP. Once a foreign agent registers the presence of a mobile host, it must then notify the home agent about the mobile host's current whereabouts so that the home agent knows where to forward any packets that it receives. Mobile IP works best if mobile nodes deregister when they leave the foreign network.

So how does the home agent send the packets to the foreign agent for delivery? It could alter the IP address contained in the packet. But this is a bad idea – prone to errors and we want the communication to be transparent to the sending host. Instead, the home agent encapsulates these packets in another IP packet addressed to the foreign agent where the mobile host is currently located. That way the application receiving the datagram does not know that the datagram was forwarded by the home agent. Once the foreign agent de-encapsulates the packet, it cannot simply send it to the address contained in the IP header. That would cause the packets to be routed back to the home network. Instead, it checks to see if the packets are addressed to a mobile host that has registered locally. It then uses a Layer 2 technology to route the packets to the mobile host.

Mobile IP thus requires that the network perform three distinct functions:

- A protocol by which mobile nodes can register and deregister with foreign agents.

- A protocol by which foreign agents can notify home agents where the mobile node is currently located.
- Protocols for home agents and foreign agents to encapsulate and decapsulate datagrams they receive.

Unfortunately, this approach suffers from a number of well-known inefficiencies.

1. Security

The ability to register from remote locations raises major security concerns. For example, a malicious user could attempt to mislead the home agent into thinking it was the proper recipient. If so, it could receive all of the packets address to the IP address.[100]

2. Handoffs

Mobile IP also has must find a way to manage the network when a mobile host moves from one base station to another. One solution is to can update the home agent. Any tardiness in the update can cause packets to become lost. Another solution is to designate the first foreign agent in a particular transaction as the *anchor foreign agent* that will be the location where the home agent will send all packets. Should the mobile host shift to a different foreign network, the anchor foreign agent can forward the packets to the new location.

3. Triangle Routing

By envisioning that all traffic will travel to the home agent and then be forwarded to the foreign agent, mobile IP employs a form of indirect routing that can be very inefficient. For example, a person's whose home network is located in Philadelphia travels to Los Angeles and the person seated next to her in a conference room attempts to forward a document to her, that

[100] KUROSE & ROSS, *supra* note 55, at 575; PETERSON & DAVIES, *supra* note 53, at 294; TANENBAUM, *supra* note 47, at 464.

document will have to travel all the way across the country to the home agent located in Philadelphia and then be rerouted back to Los Angeles. This can result in the inefficiency of what is sometimes called "triangle routing."[101]

The home agent can eliminate triangle routing by passing the mobile host's current location on to the sender so that the sender may forward subsequent packets to it directly. The initial communications must still bear the inefficiency of triangle routing. Moreover, such solutions become much more difficult to implement if the mobile agent is constantly on the move.[102] The network must have some way to notify the correspondent that the mobile host has changed location. The usual solution is that much as the home network and the foreign network have agents, the correspondent attempting to contact the mobile host also has a *correspondent agent*. The correspondent agent queries the home agent to learn the location of the mobile host. It then encapsulates the datagram in a new datagram addressed to the foreign agent. The foreign agent then decapuslates the new datagram and passes the original datagram to the mobile host.

The problem arises if the mobile host moves from one foreign network to another. Under indirect routing, the mobile host simply notifies its home agent of the change of location. Under direct routing, however, the correspondent agent that is responsible for encapsulating datagrams and forwarding them to the mobile host, not the home agent. At this point, the mobile node needs a way to update the correspondent agent as to its new location. This in turn requires two more protocols.

- A protocol by which correspondent agents can query the home agent as to the mobile node's current location.
- A protocol by which the mobile host that changes foreign networks can notify the correspondent agent about its new location.

---

101 PETERSON & DAVIES, *supra* note 53, at 293.

102 COMER, *supra* note 53, at 339–46; KUROSE & ROSS, *supra* note 55, at 566–77; TANENBAUM, *supra* note 47, at 372–75, 462–64.

The additional complexity is sufficiently difficult to implement that direct routing was not included in the upgrade to IPv6.

E. The Identity/Locator Split

The most radical solution to these problems known as the identity/locator split.[103] The idea gained new impetus by the Report from the Internet Architecture Board (IAB) Workshop on Routing and Addressing, which reflected a consensus that such a split was necessary.[104] The International Telecommunication Union (ITU) has also embraced the need for the ID/locator split in Next Generation Networks (NGNs) ITU.[105] It is also the focus of a major research initiative sponsored by the National Science Foundation's Future Internet Architecture Program.[106]

The proposal is based on the insight that IP addresses currently serve two distinct functions. It simultaneously serves as an *identifier* that identifies a machine as well as a *locator* that identifies where that machine is currently attached to the network topology. When all hosts connected to the Internet via fixed telephone lines, the fact that a single address combined both functions was unproblematic. The advent of mobility has caused the unity of identity and location to break down. A single mobile device may now connect to the network through any number of locations. Although the network could constantly update the routing table to reflect the host's current location, doing so would require propagating the updated information to every router in the network as well as an unacceptably large number of programs and databases.

---

[103] For an early statement, see Jerome H. Saltzer, On the Naming and Binding of Network Destinations (IETF Network Working Group Request for Comments 1498, 1993), *available at* http://tools.ietf.org/pdf/rfc1498.

[104] David Meyer et al., Report from the IAB Workshop on Routing and Addressing 22–23 (IETF Network Working Group Request for Comments 4984, 2007), *available at* http://tools.ietf.org/pdf/rfc4984.

[105] International Telecommunication Union Telecommunication Standardization Sector, General Requirements for ID/Locator Separation in NGN (Recommendation ITU-T Y.2015, 2009), *available at* http://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-Y.2015-200901-I!!PDF-E&type=items.

[106] MobilityFirst, *supra* note 20.

Others have proposed radical changes in the addressing and routing architecture. One approach would replace the single address now employed in the network with two addresses: one to identify the particular machine and the other to identify its location. A number of proposals advance just such a solution, including the Host Identity Protocol (HIP) (RFC 4423), the Locator Identifier Separation Protocol (LISP), Level-3 Shim for IPv6 (Shim6), and Six/One.[107] Others criticize such proposals as unnecessarily complicated.[108]

If deployed, the identity/locator split would represent a radical deviation from the existing architecture. Whatever solution is adopted would represent a fundamental change in the network layer than unifies the entire Internet. It would require a change in the way we approach routing and addressing and require reconfiguring every device attached to the network. If implemented, it would eliminate some of the asymmetries in the way that routing to mobile hosts is done and wireline hosts.

As of right now, it has not yet come to pass. And even if did, there would probably an extended transition time where things ran both.

CONCLUSION

The net result is that mobile wireless broadband networks operate on principles that are quite different from those governing the rest of the Internet. Bandwidth limitations require that wireless providers manage their networks more intensively than those operating networks based on other technologies. The fact that smartphones do not have IP addresses and the higher incidence of packet loss requires that wireless networks employ virtual circuits and embed

---

[107] *See* Chakchai So-In, *Virtual ID: ID/Locator Split in a Mobile IP Environment for Mobility, Multihoming and Location Privacy for the Next Generation Wireless Networks*, 5 INT'L J. INTERNET PROTOCOL TECH. 142 (2010) (surveying alternative approaches to the ID/locator split).

[108] *See, e.g.*, Dave Thaler, Keynote Address at the 3rd ACM International Workshop on Mobility in the Evolving Internet Architecture (MobiArch 2008): Why Do We Really Want an ID/Locator Split Anyway? (Aug. 22, 2008), *available at* http://conferences.sigcomm.org/sigcomm/2008/workshops/mobiarch/slides/thaler.pdf.

intelligence in the network to provide Internet access and to handle the problems of congestion. The unpredictability of signal strength resulting from the physics of wave propagation can necessitate more extensive supervision than other technologies require, as do the realities of system conflicts and power consumption. Lastly, mobility is placing pressure on the routing and addressing space that may soon require more fundamental changes. The industry has not yet reached consensus on the best approach for addressing all of these concerns. In its consideration of regulatory interventions, the Commission must be careful to create a regime that takes these differences into account.