ENGINEERING BACKGROUND

A. The Development of the Internet

The "Internet" is not a single network, but is instead a loose confederation of thousands upon thousands of networks, most of them built and operated with private risk capital, with no guaranteed returns. Without government compulsion or intervention, each of these constituent networks has voluntarily adopted a common protocol and addressing scheme—the Internet Protocol—that enables its customers to communicate with customers connected to other networks for purposes of exchanging higher-layer applications and content.[1] "The Internet," as that term is commonly used, is a conceptual aggregation of these mostly private IP-based networks spread across the world.

The Internet Protocol and its predecessors were first formulated several decades ago by academics and consultants funded by the Advanced Research Projects Agency ("ARPA"), a subagency within the U.S. Department of Defense. The development of the Internet Protocol was (and continues to be) overseen by the Internet Engineering Task Force ("IETF"), a private entity.[2] For many years after its inception, the Internet was restricted to academic and governmental institutions and their consultants, and commercial transactions were strictly prohibited. In the early 1990s, the U.S. government fully "privatized" the Internet by selling key infrastructure assets, including an integral backbone network known as NSFNET, to private network operators. Since then, the Internet has developed to its current advanced state, largely unrestricted by government regulation.

B. Overview of the Internet's Constituent IP Networks and the Blurring Distinction Among Backbone, Access, and Edge Functionalities

The intertwined private networks of the Internet are all part of an evolving global ecosystem. A given network's role in that ecosystem is complex and dynamic, and the network may play several roles at once. Nonetheless, popular discussions of the Internet tend to classify its constituent networks into three basic categories: backbone networks; access/aggregation networks; and edge networks. Despite their name, the "edge" networks play as central a role as conventional access and backbone networks in ensuring that application and content providers can reach end users quickly and reliably.

1. Backbone Networks

In this context, the term "backbone network" denotes the very high-capacity portion of a network operator's facilities, typically consisting of very high-speed routers and fiber-optic links stretching across large geographic areas. That backbone network serves two main functions.

[1] *See* Resolution of the Federal Networking Council, Oct. 24, 1995 (quoted in Barry M. Leiner *et al.*, *A Brief History of the Internet*, ISOC, http://www.isoc.org/internet/history/brief.shtml).

[2] *See* IETF, *The Tao of IETF: A Novice's Guide to the Internet Engineering Task Force* (Nov. 30, 2009), http://www.ietf.org/tao.html.

First, it connects the various access/aggregation networks that the provider has deployed to reach its end user customers, which may range from residential households to large enterprise businesses, including Internet content and application providers. Second, each provider's backbone network interconnects with other providers' backbone networks. The conceptual accumulation of all network operators' individual backbones is sometimes referred to collectively (and somewhat misleadingly) in the singular as "the Internet backbone."

The bilateral agreements that enable traffic to travel between two different backbone networks commonly follow one of two different business models: *peering* and *transit*. The choice between these two models turns in part on the relative value that each of the two networks brings to the interconnection arrangement.

Under *peering* agreements, each network interconnects for the purpose of terminating packets sent from the other peer to end points served by the terminating peer's network. Such arrangements typically anticipate, among other things, that the traffic exchanged between the two networks will be roughly equal in volume, such that each backbone network will incur roughly the same costs in handling the traffic originated by the other network. To avoid administrative overhead, parties to these bilateral peering agreements typically forgo the mutual exchange of compensation and peer on a settlement-free basis. But in some cases, where the traffic volumes exchanged are unequal, or where one network otherwise falls short of the other's peering criteria, the parties may enter into a paid peering arrangement. Under paid peering, the networks still exchange traffic through high-capacity peering links, but the "non-compliant" network makes payments to the other network.

Under *transit* arrangements, Network X pays Network Y to arrange delivery of Network X's packets to any destination on the Internet and to accept delivery of packets destined for Network X's customers from any location on the Internet.[3] Rather than exchanging traffic through peering links with Network Y, Network X typically buys a robust, enterprise-class Internet access service from Network Y, which supplies the interconnection facilities.

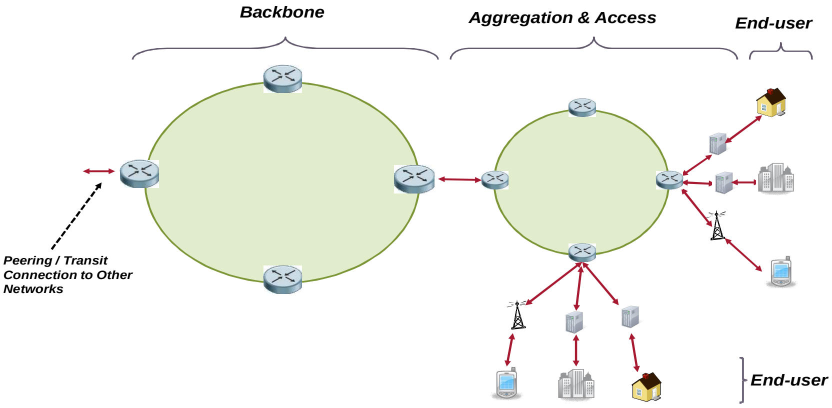
From their inception, these peering and transit relationships have been unregulated, and the market for peering and transit has functioned with great efficiency. A key reason is that the larger backbones "compete for the transit business of smaller backbones in order to increase their revenues," and this competition has driven transit prices down significantly over the last decade, from approximately $1200/Mbps in 1998 to less than $12/Mbps in 2008.[4] At the same time, the growing volume of traffic on the Internet, which we discuss below, will require content and backbone providers alike to explore new technologies and business models for the cost-effective delivery of high-bandwidth and performance-sensitive content.

[3] *See* Michael Kende, *The Digital Handshake: Connecting Internet Backbones*, FCC, Office of Plans and Policy, OPP Working Paper No. 32, at 7 (Sept. 2000), http://www.fcc.gov/Bureaus/OPP/working_papers/oppwp32.pdf.

[4] *See id.* at 20; DrPeering, Why care about Transit Pricing?, http://drpeering.net/a/Peering_vs_Transit___The_Business_Case_for_Peering.html.

2. Access/Aggregation Networks

End users—from residential subscribers to large enterprise customers—connect to the Internet through the "access" portion of an ISP's network.[5] Broadband access networks perform two key functions within the Internet ecosystem. First, they provide the last mile (or last several miles) to end-user locations through a variety of technologies, ranging from DSL or coaxial cable links to wireless spectrum to OCn-level fiber-optic cables. Second, at one or more points along the way to the ISP's backbone network, they *aggregate* the traffic of progressively larger sets of different users and transmit this aggregated traffic over increasingly higher-capacity facilities. This portion of an access network—the bridge between the "last mile" and a backbone network—is sometimes known as an "aggregation" network. While the boundaries between access facilities, aggregation facilities and backbone facilities vary from network to network and are not always easy to identify with precision, the following diagram provides a general approximation of the three network segments:



Different broadband networks require different degrees of network management to function properly for consumers. Wireless broadband, in particular, poses formidable and ever-changing network-management challenges. These arise from, among other things, the unique nature of radio of spectrum, such as hard limits on available spectrum and the physics of radio propagation, and the revolutionary transformation of wireless broadband technology itself as network engineers complete their conversions from 2G to 3G—and then begin converting today's 3G networks into tomorrow's 4G LTE networks.

For many years, broadband providers have offered quality-of-service ("QoS") enhancements to enterprise customers, including application and content providers. For example, broadband providers have long allowed content providers and other enterprise customers to designate certain packets for priority handling during periods of congestion, depending on (among other variables) whether those packets are associated with real-time or other unusually delay-sensitive applications. A broadband provider will then ensure special

[5] These comments use the terms "broadband Internet access provider" and "ISP" interchangeably.

handling for those packets throughout the QoS-enabled portions of its network. Those network facilities also typically carry non-QoS-enhanced ("best effort") Internet traffic from both enterprise and residential customers. These networks are engineered to meet the performance requirements of each class of traffic while allowing the network operator and its customers to reap the tremendous cost efficiencies of shared packet-switched facilities.

3. "Edge"/Overlay Networks, CDNs, and the Rise of the Content "Hyper Giants"

In the Internet's early years, the stereotypical "edge" network used by an application or content provider consisted of a server or two operated by a small entrepreneur working in a garage or in low-rent office space. Today's leading edge networks have evolved into something radically different: transnational facilities-based networks with an unprecedented combination of transmission capacity, processing power, and data storage.[6] Among the largest are the massive "server farms" and caching networks developed by companies as diverse as service providers Akamai and Level 3, on-line retailers Amazon.com and eBay, Internet portals Yahoo! and MSN, and—largest of them all—Google. These "overlay" or "content-delivery networks" (CDNs) use much the same technology and perform many of the same routing and long-haul transmission functions as Internet backbones and allow application and content providers to direct customer requests to the closest cache server that has both the requested content and the capacity to serve the request at the instant it is received.

Google, for example, maintains a sprawling network consisting of hundreds of thousands of servers, many of them clumped in massive data centers or server farms, connected by high-capacity fiber-optic cable.[7] Combined with Google's multi-billion-dollar investment in data storage and processing power, this "overlay" CDN enables Google to outperform its rivals in the delivery of (for example) split-second search results to end users throughout the world. Google's success exemplifies the growing power of CDNs on the Internet. Traditionally known as "caching" networks, CDNs distribute and store copies of content on servers at multiple locations across the Internet (typically located near ISP backbone networks) and thus enable end users to gain access to that content more quickly and reliably than in a conventional "unicast" arrangement, where each end user must communicate directly with a single centralized server. For example, when a typical end user types "www.apple.com," "www.facebook.com," or "news.google.com" into an Internet browser,[8] the data request is directed to a nearby CDN cache server, where the content of those websites has been stored, thus enabling the end user and the

[6] *See* George Ou, *Two Hypocrites in a Garage*, Digital Society, Nov. 23, 2009, http://www.digitalsociety.org/2009/11/the-hypocrisy-of-google-and-skype/.

[7] *See* George Gilder, *The Information Factories*, Wired, Oct. 2006, http://www.wired.com/wired/archive/14.10/cloudware_pr.html. In addition to Google, other major Internet companies, including Microsoft and Yahoo!, are likewise constructing enormous networks of their own and, like Google, are revolutionizing the role of these ostensible "edge" networks within the Internet.

[8] *See* Akamai, *Customer List*, http://www.akamai.com/html/customers/customer_list.html.

cache server to exchange data far more quickly and efficiently than if the data were stored on a single, centrally located server far from the end user.[9]

The bottom line is that, all else held equal, end users have better experiences in their interactions with CDN-equipped content providers than with content providers that do not use CDN functionality. This in turn means that well-funded content and application providers that can afford to purchase (or self-provision) CDN services have a substantial advantage over less-well-funded rivals in the battle to bring end users top-quality Internet experiences.

The success of Google, Limelight, and other overlay networks also marks an unprecedented shift of power within the Internet ecosystem. Even four years ago, analysts recognized that Google had begun "building a network so massive that several service provider specialists believe it could end up with one of the world's largest core transport networks, effectively building its own private Internet" and "controlling distribution of much of the world's Internet traffic."[10] Today, that process is nearing completion. A recent study conducted by the University of Michigan and Arbor Networks cites the rise of Google and other content "hyper giants" as evidence of a fundamental shift in power relationships within the Internet ecosystem:

> Five years ago, Internet traffic was proportionally distributed across tens of thousands of enterprise managed web sites and servers around the world. Today, most content has increasingly migrated to a small number of very large hosting, cloud and content providers. Out of the 40,000 routed end sites in the Internet, 30 large companies – "hyper giants" like Limelight, Facebook, Google, Microsoft and YouTube – now generate and consume a disproportionate 30% of all Internet traffic.[11]

This development has upended Internet business models. Rather than relying upon conventional Internet backbone networks to deliver their content to "eyeball" networks, these hyper giants have grown so large and powerful that they can "cut out the middle man" and obtain settlement-free (zero-priced) peering directly with some end-user broadband networks.

A related harbinger of change within the Internet ecosystem is the emergence of so-called *reverse-blocking:* the practice by certain content providers of withholding their must-have Web content from end users unless the broadband providers for those end users agree to pay extra for it. For example, Disney currently blocks access to its premium sports programming site,

[9] Although Google and a number of other large Internet companies self-provision their own CDNs, many application and content providers outsource this functionality by hiring third-party CDN providers such as Akamai, Limelight, Level 3, and AT&T.

[10] R. Scott Raynovich, *Google's Own Private Internet*, Light Reading, Sept. 20, 2005, http://www.lightreading.com/document.asp?doc_id=80968.

[11] Arbor Networks, *Two-Year Study of Global Internet Traffic Will be Presented at NANOG47*, Oct. 13, 2009, http://www.arbornetworks.com/en/arbor-networks-the-university-ofmichigan- and-merit-network-to-present-two-year-study-of-global-int-2.html.

ESPN360, from consumers whose broadband providers do not pay fees to Disney, and it explicitly steers those disappointed consumers to rival providers that have paid up:[12]

Google similarly blocks access to YouTube from Internet-enabled set top boxes sold by vendors that do not enter into high-priced advertising arrangements with Google.[13] There is no reason to suppose that these will be isolated incidents. As illustrated by recent cable retransmission-consent deals, content providers often have more market clout than distribution networks and can now successfully charge those networks substantial fees for the privilege of carrying their content.[14] Over time, the reverse-blocking phenomenon may force ISPs to pass through charges to the specific subscribers who use the content in question—rather than to all subscribers indiscriminately—by establishing different content-based tiers of Internet access service: those for end users who order various combinations of premium applications and content, and those who do not.

There is no clear reason why such overt "balkanization" of the Internet should concern policymakers less than the much more benign prospect that money will sometimes flow in the opposite direction as well, when a content provider voluntarily pays a broadband provider for QoS enhancements for unusually performance-sensitive content. Certainly considerations of market power cannot support this disparity in regulatory treatment, because broadband providers are often the less powerful parties in the relevant (national) market than the application/content providers they must deal with.

C. The Internet Is Not Now a "Neutral" Place, and Proposals to Convert Broadband Networks into a Collection of "Dumb Pipes" Would Make It *Less* Neutral in Its Treatment of Competing Applications and Content

The rise of CDNs and the content hyper-giants is one of several phenomena that explodes a popular myth underlying much net neutrality advocacy: the notion that as a platform for commerce, the Internet does not distinguish between a budding entrepreneur in a student dormitory room and a Fortune 500 company. In fact, application and content providers with the

[12] As Disney explains on the ESPN360 website: "ESPN360.com is available nationwide, but you must subscribe to a participating high speed internet service provider. . . . Click here to find out more on how you can request access to ESPN360.com or switch your service to a participating high speed internet service provider." *ESPN360.com FAQ*, http://espn.go.com/broadband/espn360/faq#4.

[13] *See* Eliot Van Buskirk, *YouTube Blocks Non-Partner Device Syabas as Allegations Fly*, Wired, Nov. 20, 2009, http://www.wired.com/epicenter/2009/11/youtube-blocks-non-partner-device-syabas-as-allegations-fly/.

[14] For example, "News Corp sought as much as $1 a month per Time Warner Cable subscriber for rights to Fox, home of 'The Simpsons' and 'American Idol". . . . If other networks seek similar terms, cable operators may have to fork out as much as $5 billion a year—and would probably pass the cost on to subscribers, said Craig Moffett, an analyst at Sanford C. Bernstein in New York." Kelly Riddell, *Fox-Time Warner Cable deal could mean billions for broadcasters*, Washington Post, Jan. 4, 2010.

capital resources needed to buy CDN services—or to build out their own global networks, as Google has done—will provide consumers with far better performance than can any "mom-and-pop site" or "budding entrepreneur in a dorm room" that lacks such resources and cannot obtain capital financing. As Akamai explains in a 2002 white paper, the competitive advantage that well-funded providers gain from CDN services has long justified the price of purchasing them.[15]

In contrast, application and content providers that cannot afford to buy CDN services must rely on traditional access/aggregation and backbone services to send their traffic through many potentially congested routers and links en route to other Internet users, with accompanying increases in the potential for latency, jitter, and packet loss. No one claims that the government should intervene to neutralize this disparity, because no one who understands the Internet ecosystem believes the populist "equality" rhetoric underlying much of the advocacy for net neutrality regulation.

In other respects as well, it is wrong to suggest that the Internet would be "neutral" in its treatment of different applications and content if only broadband networks were turned into a collection of dumb pipes. Indeed, many of the outcomes that the pro-regulation advocates would impose on the Internet would make it, if anything, *less* neutral under any meaningful definition of that term.

One reason relates to a content or application provider's choice of a transport protocol for its outgoing traffic.[16] The Internet is often described as using the "TCP-IP protocol suite," with IP at Layer 3 (the "network" layer) and the transport control protocol ("TCP") at Layer 4 (the "transport" layer). But some Internet traffic does not in fact use TCP; instead, application providers sometimes choose the alternative "user datagram protocol" ("UDP") at Layer 4. When used appropriately, UDP's attributes can be beneficial for a range of purposes, including Domain Name System (DNS) queries. At the same time, the choice between these transport-layer protocols has significant implications for how finite bandwidth is allocated among competing uses during periods of congestion. TCP is considered a "polite" transport protocol because it can sense congestion and "throttles back" transmission rates until after the congestion lifts. In contrast, UDP omits the error-correction functions of TCP and, unlike TCP, does not throttle back in the face of network congestion. And precisely because UDP applications "send out data as fast as [they] can," even when they encounter congestion, "while [conventional] TCP-friendly

[15] Akamai White Paper, *Why Performance Matters*, (2002), http://www.akamai. com/dl/whitepapers/Akamai_Why_Performance_Matters_ Whitepaper.pdf (emphasis added).

[16] *See generally* James Kurose & Keith Ross, *Computer Networking: A Top-Down Approach* 50-54 (5th ed. 2010) ("*Kurose & Ross*") (discussing Internet protocol layering). As discussed below, providers also use, in addition to the Layer 3 and Layer 4 mechanisms discussed in the text, a variety of prioritization techniques on other, non-IP layers of data communications, including Layer 2 (*e.g.*, Ethernet, ATM), Layer 2.5 (MPLS), and even Layer 7 (SPDY), all of which affect how end users experience the Internet.

applications deliberately send fewer and fewer packets," the latter applications may end up "starved of network resources."[17]

Moreover, even if application designers choose TCP for their packets, they can structure their applications to elbow other applications aside in a quest for a greater share of the limited bandwidth across congested links. Indeed, "BitTorrent" sessions are so named precisely because they aggressively consume disproportionate amounts of upstream subscriber bandwidth by opening up multiple connection streams to seize capacity for themselves.[18] As one academic study has shown, "as few as 15 BitTorrent users" on a cable modem network "can significantly reduce the service quality experienced by other subscribers."[19] As the inventor of BitTorrent has explained, this was intentional: "My whole idea was, 'Let's use up a lot of bandwith.' . . . I had a friend who said, 'Well, ISPs won't like that.' And I said, 'Why should I care?'"[20] BitTorrent Inc. recently acknowledged the need to be more network-friendly and, to that end, launched a new implementation of the BitTorrent protocol: uTorrent 2.0. According to recent tests, however, the efficacy of this solution is still in doubt.[21]

In short, *passive* management of the IP platform would produce *non*-neutral outcomes among the packets associated with different applications, because it would allow applications with "selfish" protocols to trump those with "polite" protocols in the contest for finite bandwidth.

Second, even if all transport-layer protocols were equally polite, passive management of the IP platform *still* would not produce "neutral" results in any meaningful sense, because it can hardly be "neutral" for network engineers to ignore the vast disparities in the QoS needs of emerging Internet applications. Although the Internet Protocol was designed from the beginning to be capable of providing enhanced service quality (see below), many Internet access networks designed for residential users were initially optimized to process the traditionally most prevalent type of communication: non-latency-sensitive data applications, such as the delivery of email or the downloading of ordinary webpages. One of the most important and pro-consumer developments of the past five years has been the rapid convergence of *all* electronic

[17] Jon M. Peha, *The Benefits and Risks of Mandating Network Neutrality, and the Quest for a Balanced Policy*, at 7 (2006), http://web.si.umich.edu/tprc/papers/2006/574/Peha_balanced_net_neutrality_policy.pdf ("*Benefits and Risks*").

[18] *See, e.g.*, Bob Briscoe, *Flow rate fairness: Dismantling a Religion*, 37 Computer Commc'n Rev. 63 (2007), http://portal.acm.org/citation.cfm?doid=1232919.1232926.

[19] James J. Martin and James M. Westall, *Assessing the Impact of BitTorrent on DOCSIS Networks*, at 1 (2007) (http://www.cs.clemson.edu/~jmarty/papers/bittorrentBroadnets.pdf).

[20] David Downs, *BitTorrent, Comcast EFF Antipathetic to FCC Regulation of P2P Traffic*, San Francisco Weekly, Jan. 23, 2008 (http://www.sfweekly.com/2008-01-23/news/bittorrent-comcast-eff-antipathetic-to-fcc-regulation-of-p2p-traffic).

[21] George Ou, *Analysis of BitTorrent uTP congestion avoidance*, Nov. 22, 2009 (http://www.digitalsociety.org/2009/11/analysis-of-bittorrent-utp-congestion-avoidance).

communications around the IP platform, including applications—such as real-time high-definition video—that will not function properly during periods of congestion unless network providers accompany them with quality-of-service enhancements that non-performance-sensitive applications do not need in order to function well.

Any requirement that networks treat all packets exactly the same, irrespective of the QoS needs of their associated applications would flatly discriminate against QoS-sensitive applications like real-time video and VoIP. If required to treat all packets identically, a broadband network "might at times transmit 100 P2P packets before it transmits a single VoIP packet," causing "many of the VoIP packets . . . to wait so long that they expire and cause dropped audio," an outcome that "is blatantly unfair and destructive to the VoIP application."[22] Even some proponents of net neutrality regulation therefore agree that any sensible view of "neutrality" must account for these application-specific disparities in QoS needs.[23] If anything, therefore, network-management techniques designed to allocate finite network resources to the latency-sensitive applications that actually need them are *pro*-neutrality – and unquestionably are pro-consumer.

D. The Rapid Convergence of All Electronic Communications Around the IP Platform Poses Critical Engineering Challenges

The rapid convergence of multiple services onto a single IP platform carries many advantages beyond the obvious economies of scale and scope derived from building one network rather than several. It also allows for the integration of voice, video, and text into feature-rich multimedia applications and it facilitates greater competition among service providers. For example, cable and telephone companies, once siloed from mutual competition because of their single-purpose networks, may now compete fiercely to offer the "triple play" of voice, video, and Internet access services. It also creates opportunities for independent application and content providers to offer a variety of innovative services to a wide range of customers, including residential, small/medium business, and enterprise customers. Such services would be economically infeasible if individual services required separate networks.

But the many advantages of IP convergence come with a critical engineering challenge: how to make all of these applications, with their quite different QoS needs, function as well as possible over a shared and sometimes congested network infrastructure.

[22] George Ou, *Debunking the Myth that Prioritized Networks Are Harmful*, Digital Society, Nov. 12, 2009, http://www.digitalsociety.org/2009/11/debunking-the-myth-that-prioritizednetworks-are-harmful/

[23] See Wu, *Network Neutrality, Broadband Discrimination*, J. Telecomm. & High Tech. L. 141, __ (2003) ("the Internet's greatest deviation from network neutrality" has consisted of its traditional "favoritism of data applications, as a class, over latency-sensitive applications involving voice or video").

1. Managing the Phenomenon of Convergence Requires Not Just Higher-Capacity Pipes, but Smarter Networks

Virtually all commercial IP networks are "shared" among different users and also different *uses*. This sharing is one of the greatest advantages of IP networks as compared to traditional circuit-switched networks. It lowers costs for users as compared to dedicated networks, and it permits maximum utilization of broadband infrastructure. But sharing presents trade-offs.

The benefits of sharing are best understood by contrasting IP networks with traditional circuit-switched networks. In a conventional telephone network, a fixed amount of bandwidth must be dedicated to a continuous path (the "circuit") between the two end points to the call, and that circuit must be kept open for the entire call. While this approach ensures highly predictable performance, it "wastes" capacity. For example, even during pauses in a voice conversation or data transmission, the reserved capacity on the circuit is unavailable for any other use.

In contrast, the Internet's constituent IP networks use packet-switched rather than circuit-switched technology, do not typically establish fixed end-to-end paths between two points, and do not reserve capacity for a particular communication stream. Rather, IP networks break the stream into data packets, each of which contains a "header" (an initial series of bits) that identifies, among other things, the packet's ultimate destination. Each router examines the address in the packet's header and directs it to the next router, selected on the basis of predictions about the most efficient route to the packet's ultimate destination. A conventional "best-effort" IP network makes such routing decisions on a packet-by-packet basis without "knowing" what higher-layer application any packet is associated with or whether that application is performance-sensitive.

Modern Internet access networks are typically engineered to high standards that accommodate sharing among a wide range of applications even on such a "best-effort" basis. This has enabled companies like Vonage, Skype, and Vuze to use such networks to offer highly competitive voice and video services that hundreds of millions of consumers have embraced. Indeed, Skype alone has more than 520 million registered users worldwide.[24] But all packet-switched, shared networks are inherently susceptible to several forms of service degradation during peak periods of congestion, which affect some applications far more than others.

First, the packets associated with any given application are subject to *latency:* the delays that result from, among other things, "the accumulation of transmission, processing, and queuing delays in [the multiple] routers" between two end users in an Internet data session.[25] Second, Internet applications can suffer from *jitter:* variations in delays among associated packets, such that different packets arrive unpredictably and sometimes out of order. Third, applications can suffer from outright *packet loss*, which—as its name implies—occurs when the buffers in

[24] eBay Inc., Form 10-Q, at 25 (filed Oct. 27, 2009), http://files.shareholder.com/downloads/ebay/797758946x0xS1193125-09-214947/1065088/filing.pdf (noting that there were 520.8 million Skype users as of September 30, 2009).

[25] *Kurose & Ross*, at 618.

congested routers fill to capacity and the network "loses" the additional incoming packets. For example, "[i]f one of the links is congested because other packets need to be transmitted at the same time, then [a given] packet will have to wait in a buffer at the sending side of the transmission link, and suffer a delay. If the wait time is too long, the buffer overflows and the packet is 'lost.' The Internet makes its *best effort* to deliver packets in a timely manner, but it does not make any guarantees."[26]

Applications differ enormously in their relative sensitivity to latency, jitter, and packet loss and their ability to compensate for them. For example, "in many multimedia applications" such as real-time video streaming, "packets that incur a sender-to-receiver delay of more than a few hundred milliseconds are essentially useless to the receiver. On the other hand, networked multimedia applications are for the most part loss-tolerant—occasional loss only causes occasional glitches in the audio-video playback, and these losses can often be partially or fully concealed. These delay-sensitive but loss-tolerant characteristics are clearly different from those of elastic applications such as the Web, e-mail, FTP, and Telnet," for which delays are tolerable but substantial packet loss is not.[27]

While the best-effort Internet has sufficed to support VoIP and some other performance-sensitive services so far, the growing popularity of such services, together with escalating consumer demand for real-time high-definition video and other premium services, poses a fundamental engineering challenge. How can engineers structure a unified IP platform to maintain the cost-reducing *efficiency* of packet-switched IP networks while also assuring the *quality of service* that consumers demand for real-time services, such as voice and video, now that the signals for those services no longer travel on service-specific transmission networks? The answer cannot be that IP networks must blindly treat all packets alike by subjecting them equally to the best-effort delivery principles used today for downloading ordinary web pages or delivering e-mails. That approach would produce unacceptably poor quality for real-time applications like voice and video and would thwart the promise of convergence.

The answer likewise cannot be that network providers, on top of the tens of billions of dollars they have already invested in next-generation networks, must so radically enlarge the capacity of their IP networks as to give *all* packets—including those associated with *non*-real-time applications that are reasonably tolerant of latency and jitter—the same guarantees of nearly instantaneous delivery needed for high-quality video services. Network engineers keep usage affordable by scaling the network's routers and transmission links to meet desired performance levels for different classes of traffic under foreseeable conditions. Raw bandwidth, in the form of extremely-high-capacity routers and other data-processing and transport infrastructure,

[26] *Kurose & Ross, supra*, at 27. Wireless broadband networks (and applications designed for them) must accommodate the unusually high levels of packet loss encountered in wireless transmissions, and are also severely constrained in the bandwidth they may deploy for end users in particular transmission cells. This is one of many respects in which network engineers in the wireless context face network-management challenges more severe than their wireline counterparts.

[27] *Id.* at 598. In these comments, we use the term "latency-sensitive" as a shorthand to denote sensitivity to latency, jitter, or both.

remains very costly. Network engineers therefore do not—and could not economically—oversupply capacity to ensure instantaneous delivery of all packets at all times, particularly since random events can trigger unpredictable spikes in usage. Indeed, forcing them to take that approach would rob IP networks of the efficiency characteristics that make Internet usage affordable in the first place. Economic studies have thus shown that, as IP video services escalate in popularity, any single-minded reliance on "fat, dumb pipes" as a solution to QoS requirements in this environment of rapidly escalating Internet usage would dramatically raise network costs and cause end-user rates to skyrocket.[28]

Moreover, this overcapacity approach might well be futile even if money were no object for broadband networks and their customers. Experience has shown that as networks increase the capacity of given links on the Internet, usage on that link—particularly in the form of peer-to-peer file-transfer applications—rapidly expands to fill the new capacity. For example, Japan, "with widely marketed 100 Mbps connections, still has concerns with congestion and has adopted multiple strategies to cope with problems related to network neutrality. This indicates that, contrary to the views of some proponents of national broadband policies, greater investment in broadband infrastructure alone is unlikely to eliminate the role of traffic management by network operators."[29]

In short, the solution to this engineering challenge lies not only in more networks and higher-capacity pipes, but in greater network intelligence as well, including an ability to identify and provide the appropriate level of performance required by different applications traversing the network so that users can receive the service quality they desire. Fortunately, the designers of the Internet Protocol perceived a need for precisely such differentiation of traffic into latency-sensitive and non-latency-sensitive applications, and they built the capacity for such differentiation into IP. The following sections describe the history and technology of "DiffServ," its common use in the provision of IP services to enterprise customers, and its increasing use within the consumer marketplace as well.

2. The Internet Protocol, and Broadband Networks in General, Have Always Been Designed to Support Differential Treatment of Traffic to Satisfy Quality-of-Service Needs

Much of the advocacy for net neutrality regulation rests on a creative misreading of a 25-year-old white paper by three highly regarded network engineering experts—Jerome Saltzer,

[28] *See, e.g.*, George Ford et al., *The Efficiency Risk of Network Neutrality Rules*, Phoenix Center Policy Bulletin No. 16 (2006), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=925347; Richard N. Clarke, *Costs of Neutral/Unmanaged IP Networks* 21 (2006), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=903433); Steven Pociask, *Net Neutrality and the Effects on Consumers*, American Consumer Institute 14 (2007), http://www.theamericanconsumer.org/ACI%20NN%20Final.pdf.

[29] Scott J. Wallsten & Stephanie Hausladen, *Net Neutrality, Unbundling, and Their Effects on International Investment in Next-Generation Networks*, 8 Rev. Network Econ. 90, 101-02 (March 2009), http://www.techpolicyinstitute.org/files/wallsten_unbundling_march_2009.pdf; *see also id.* at 110-11 (citing Japanese government reports detailing congestion problems).

David Clark, and David Reed—concerning the so-called "end-to-end" (or "e2e") principle.[30] Many pro-regulation advocates cite this paper as a policy manifesto for reducing every IP network to a collection of "dumb pipes" that should be forever consigned to treating every IP packet exactly the same, oblivious to whether the packet is associated with a performance-sensitive application or not. The paper is nothing of the kind. It is instead an early description of how key error-correction and related functions in communications across different networks can usually, for most data applications, be conducted more efficiently and effectively by end-user devices on each end of a data session than by the routers in between. For example, errors occurring toward the edge of the network (*e.g.*, between the end point and the first network router) can be noticed and corrected by the end point, but may escape notice and correction if those functions are performed by the router instead.

The paper makes clear that the authors never intended this now-unremarkable guideline to be an "absolute rule" even as an engineering matter, let alone any sort of normative policy judgment.[31] As network engineer Richard Bennett observes, "the end-to-end arguments of network engineering differ significantly from network neutrality advocates' idiosyncratic end-to-end principle, a demand for a low-function, 'stupid' network."[32] And because those advocates have "failed to stay up-to-date with the engineering community's ongoing discussions about Internet architecture," they "have consistently asked regulators to require network operators to employ engineering solutions within the Internet that are more appropriate to the traditional, single-function telephone network, such as over-provisioning. . . . Applied blindly, end-to-end can become a dogma that limits network efficiency, increases costs, and constrains opportunities to innovate."[33]

More fundamentally, this rigidly prescriptive misuse of the end-to-end guideline also runs headlong into thirty years of development of the Internet Protocol itself, which has always recognized the need for and utility of IP-layer network intelligence to account for differences in application type. As early as September 1981, the IETF established a mechanism for marking packets by handling class so that networks could give applications within each class at least the minimum level of performance they need. Known as the "Type of Service" (ToS) field in the packet header, the purpose of this mechanism was designed to help IP networks "offer service precedence" under which a network would "treat[] high precedence traffic as more important than other traffic (generally by accepting only traffic above a certain precedence at time of high load)."[34] Thus, "[e]ven three decades ago, the vision of providing different levels of service to different levels of traffic was clear[.]"[35]

[30] J.H. Saltzer, D.P. Reed & D.D. Clark, *End-to-End Arguments in System Design* (Nov. 1984), http://web.mit.edu/Saltzer/www/publications/endtoend/endtoend.pdf (originally published in 2 ACM Transactions in Computer Systems 277 (Nov. 1984)).

[31] *See id.* at 7 ("Thus the end-to-end argument is not an absolute rule, but rather a guideline that helps in application and protocol design analysis; one must use some care to identify the end points to which the argument should be applied.").

[32] Bennett, *Designed for Change* at 2.

[33] *Id.* at 4.

[34] *Internet Protocol – DARPA Internet Program Protocol Specification*, RFC 791, at 11

That vision started to became a significant commercial reality by the 1990s. In 1994, another RFC noted that, in addition to the "simple priority" described in the 1981 RFC, more work needed to be done to facilitate latency-sensitive Internet applications: "[R]eal-time applications often do not work well across the Internet because of variable queuing delays and congestion losses," and thus "[b]efore real-time applications such as remote video, multimedia, conferencing, visualization, and virtual reality can be broadly used, the Internet infrastructure must be modified to support real-time QoS."[36] The 1994 RFC thus endorsed a mechanism that would enable network operators "to divide traffic into a few administrative classes and assign to each a minimum percentage of the link bandwidth under conditions of overload, while allowing 'unused' bandwidth to be available at other times."[37]

In 1998, building on RFC 791 and other RFCs, RFC 2474 adopted an updated version of ToS, known as Differentiated Services or DiffServ, that uses the Differentiated Services Code Point (DSCP) to mark and prioritize packets at the IP layer.[38] Today, bits 8-15 within an IPv4 packet are devoted to DSCP functionality. DiffServ operates at the IP layer (Layer 3) and permits differentiated service handling wherever routers are equipped to recognize and act upon the DSCP field.[39]

AT&T and other providers have long used DiffServ in conjunction with analogous mechanisms at other layers, including Ethernet and ATM at Layer 2 and MPLS at Layer 2.5, to ensure differentiated service handling across diverse network facilities.[40] For example, AT&T offers an enterprise-grade Internet access service, known as Managed Internet Service ("MIS"), that combines DiffServ and MPLS-based class-of-service mechanisms to ensure enhanced performance for traffic that MIS customers designate for special handling.[41] AT&T and other

(Sept. 1981), http://www.ietf.org/rfc/rfc0791.txt?number=791.

[35] *Kurose & Ross* at 648.

[36] R. Braden *et al.*, *Integrated Services in the Internet Architecture: an Overview*, RFC 1633, at 1 (June 1994), http://www.ietf.org/rfc/rfc1633.txt?number=1633.

[37] *Id.*

[38] K. Nichols *et al.*, *Definition of the Differentiated Services Field (DS Field) In the IPv4 and IPv6 Headers*, RFC 2474 (Dec. 1998), http://www.ietf.org/rfc/rfc2474.txt?number=2474.

[39] A. Retana *et al.*, *Using 31-Bit Prefixes on IPv4 Point-to-Point Links*, RFC 3021 (Dec. 2000), http://www.ietf.org/rfc/rfc3021.txt?number=3021; *see also* RFC 2914, *supra*. Figure 5 is taken from Wikipedia, *IPv4*, http://en.wikipedia.org/wiki/IPv4 (last accessed Dec. 12, 2009).

[40] *See* Nortel, *Introduction to Quality of Service (QoS)* (2003), http://www.nortel.com/products/02/bstk/switches/bps/collateral/56058.25_022403.pdf; Ralph Santitoro, *Metro Ethernet Services – A Technical Overview*, Metro Ethernet Forum, at 9 (Apr. 2003), http://metroethernetforum.org/PDF_Documents/metro-ethernet-services.pdf ("DiffServ . . . provide[s] more robust QoS capabilities when compared to the simple forwarding-based priority of IP TOS[.]").

[41] *See* AT&T Wholesale, Managed Internet Service, http://www.business.att.com/wholesale/Family/ip-solutions-wholesale/managed-internet-service-wholesale/.

network providers sell such services to a range of enterprise customers, including content providers that wish to purchase prioritized handling for performance-sensitive content throughout core network facilities.

AT&T likewise combines Layer 3 DiffServ functionality with Layer 2 mechanisms to separate its U-verse "triple play" platform into logically discrete voice, video, and Internet access streams and guarantee each service the network performance it needs to meet customer expectations.[42] The top Internet access speed available over the shared U-verse platform—24 Mbps—is several times the top speed attainable under AT&T's legacy DSL service, even though the copper infrastructure used for that service was *not* shared with any managed video service. AT&T's Internet access customers have thus benefited from the extensive fiber deployments that permit such dramatically higher-speed services. But those multi-billion-dollar deployments have made economic sense in the first place precisely because the new infrastructure *is shared*—because it supports voice and video services in addition to Internet access. [43]

The diagram above illustrates the DSCP field in today's standard version of the Internet Protocol: IPv4. The Internet community has now adopted and is beginning to implement a successor protocol—IPv6—which, among other things, permits many times the number of unique IP addresses and thus accommodates the exploding global demand for such addresses. The designers of IPv6 not only retained IPv4's differentiated-services functionality within the updated protocol, but significantly expanded on it by making provision for differences both in "traffic class" and "flow":

> RFC 1752 and RFC 2460 state that [the flow header] allows "labeling of packets belonging to *particular flows for which the sender requests special handing*, such as a nondefault quality of service or real-time service." For example, *audio and video transmission might likely be treated as a flow*. On the other hand, the more traditional applications, such as file transfer and e-mail, might not be treated as flows. . . . The IPv6 header also has an 8-bit traffic class field. This field, like the TOS field in IPv4, can be used to *give priority to certain datagrams within a flow*, or it can be used to *give priority to datagrams from certain applications . . . over datagrams from other applications*[.][44]

Like other aspects of the Internet Protocol, each of these "service handling" mechanisms (ToS, DiffServ, MPLS, and others) was developed by network engineering experts through the time-tested, consensus-building RFC process. They represent the collective wisdom of the

[42] AT&T's U-verse service recently surpassed 2 million subscribers. AT&T, Press Release, *AT&T U-verse TV Marks 2 Million Customer Milestone*, Dec. 9, 2009, http://www.att.com/gen/press-room?pid=4800&cdvn=news&newsarticleid=30203.

[43] The success of this model has led Frost and Sullivan to choose AT&T U-verse as its "2009 North American Consumer Communications Service Product of the Year." *See* http://www.att.com/Common/merger/files/pdf/Frost_Sullivan_2009_Consumer_Product_of_the_Year.pdf.

[44] *Kurose & Ross* at 367 (emphasis added).

global Internet engineering community, as embodied in IETF, and they are intended to meet the needs of the global user community. Regulators have historically, and very wisely, left the resolution of engineering debates to that community and have never proposed to take this evolving and highly nuanced set of engineering judgments about IP architecture, freeze it to suit the policy preferences of particular advocates, and stamp it with the coercive authority of law. That, however, is what the proponents of net neutrality regulation seek.

3. The Importance of QoS Enhancements in the Market Today

Any prohibition of prioritization agreements would not only foreclose many *future* pro-consumer services, but also draw a range of *existing* services into doubt and disrupt current arrangements throughout the Internet.

IP networks currently honor requests from enterprise customers (including content providers) for prioritized handling of designated content beginning on the access/aggregation links serving those customers across the network's core backbone network links—and, in some cases, all the way through that network for end-to-end QoS-enhanced data sessions between enterprise customers. At present, the network capabilities needed to provide such end-to-end QoS enhancements for Internet traffic are more prevalent in the access/aggregation networks deployed primarily to serve business customers rather than in those deployed in more residential areas.[45]

One type of end-to-end QoS arrangement in the enterprise space involves the use of network-based *virtual private networks*. Such VPNs often make use of MPLS at Layer 2.5 to "encapsulate" traffic from defined customer locations and route them transparently over prescribed paths to other such locations. "The customer experiences direct communication to their sites as though they had their own private network, even though their traffic is traversing a public network infrastructure and they are sharing that infrastructure with other businesses."[46] Network providers use various QoS techniques to establish priorities among "multiple classes of service within a VPN, as well as priorities *among* VPNs."[47]

[45] As with many technologies that are first made available to business users, it is reasonable to expect that these QoS capabilities will also become increasingly available to residential consumers. For example, while AT&T's U-verse high speed Internet access service is offered on a best-effort basis today, AT&T's network is technically capable of supporting multiple classes of service in the future. Similarly, the standards for wireless LTE-based broadband services, which will serve both business and residential users, contain a very robust set of QoS mechanisms. And it will be essential to use those mechanisms in order to efficiently provide, among other things, the voice quality that consumers demand of their mobile devices, given that voice appears as just one IP application among many in the LTE environment.

[46] Cisco, *Introduction to Cisco MPLS VPN Technology*, at 1-3 (http://www.cisco.com/en/US/docs/net_mgmt/vpn_solutions_center/1.1/user/guide/VPN_UG1.pdf).

[47] *Id.* at 1-5 (emphasis added).

Although many network-based VPNs are specific to given enterprise customers, network operators can and do configure them to encompass groups of multiple customers. The engineering community has thus deployed methods for merging "two or several VPNs . . . to a single VPN."[48]

To our knowledge, no one has suggested that such enterprise-to-enterprise arrangements might be problematic, nor could they plausibly make that argument in a marketplace where networks have long provided such QoS enhancements to willing business customers. Instead, the focus has always been on prioritization of Internet traffic in the last mile to "consumers" or, in the industry vernacular, "eyeball" customers. However, a regulator could not reasonably draw regulatory distinctions between "business" or "content-producing" customers (for whom last-mile prioritization would be permitted) and "eyeball" customers (for whom it presumably would not be). Assigning Internet users to such regulatory silos would be ill-conceived because, among other considerations, every user is potentially *both* a content provider *and* a set of eyeballs.

Moreover, these innovations are not, and should not be, confined to the business space to begin with. In the residential space as well, providers use the same DSCP-based prioritization (and related mechanisms) to provide QoS to performance-sensitive services, like IPTV and VoIP, that share a converged IP platform with best-effort Internet access. As even pro-regulation advocates have conceded, it would make no sense to prohibit such prioritization.[49] Such a ban could only give broadband providers perverse incentives to keep their voice and video networks physically separate from the IP networks used for Internet access: that is, to create redundant networks in order to ensure that their consumers retain the service quality they need for applications that must be run on a managed network. That result—if economically achievable at all—would introduce radical inefficiencies into the communications market: It would lead to higher prices for all network customers, who must ultimately pay for these unnecessary costs; it would defeat the promise of convergence by forcing different services back onto physically distinct, "siloed" platforms; and it would deter the roll-out of video competition for incumbent cable television companies.

More generally, just as it is efficient and pro-consumer to logically (rather than physically) segregate the dedicated IPTV stream from best-effort Internet traffic, so too is it efficient and pro-consumer to permit different classes of service for different types of applications and content *within* the Internet portion of the pipe—as, again, broadband providers have long done for enterprise customers.

Thus, a regulator cannot ban applications-specific differential service handling without either (i) seriously disrupting the industry by prohibiting commercial arrangements that are already common in the enterprise space, such as the sale of QoS enhancements to content providers and other enterprise customers; or (ii) creating new regulatory silos dividing "content-producing" customers from "eyeball" customers. The first option should be unthinkable. And the second would be unwise, both because there is no valid reason to deprive "residential"

[48] *Id.* at 1-12.

[49] *See* Letter from Timothy Wu and Lawrence Lessig to Marlene H. Dortch, CS Docket No. 02-52, at 14 (Aug. 22, 2003).

customers of the advanced capabilities now available to "enterprise" customers and because every network user is potentially *both* a consumer *and* a producer of Internet content.

E. The Market for Service Enhancements

As exemplified most prominently by the rise of CDNs, the Internet ecosystem features an entire market for *service enhancements:* methods that allow performance-sensitive applications and content to function well even during periods of congestion. One of the key questions in this regard is whether broadband Internet access providers should be barred from fully competing with CDNs and other vendors in that market, which is national and indeed global in scope. Understanding this point requires some background in the various technologies for managing competing demands on finite bandwidth.

The following discussion summarizes a number of key methods that network engineers at broadband and content providers alike can use to ensure higher-quality end-user experiences in an environment of increasing network congestion.[50]

1. Bandwidth Provisioning

Every broadband end user, from a suburban household to the largest global content provider, chooses the bandwidth of the broadband "pipe" or pipes that connect it to the Internet. For example, end users can purchase different tiers of AT&T U-verse broadband Internet access, with download speeds ranging from 1.5 Mbps (the "express" tier) to 24 Mbps (the "max turbo" tier). And enterprise businesses, including application and content providers, choose from a vast range of different enterprise broadband services offered by a variety of providers.

The bandwidth an end user chooses will depend, of course, on the volume of traffic it expects to exchange with other end points on the Internet, both upstream and downstream. While broadband providers continually upgrade their networks to give customers the bandwidth they desire (consistent with their terms of service), virtually all Internet traffic crosses shared facilities at some point in its end-to-end transmission path. As a result, the access "bandwidth" an end user purchases, no matter how great, cannot insulate it from the service degradation caused by congestion on shared links, ranging from aggregation facilities in the access network to peering points connecting Internet backbones. As discussed, moreover, network providers cannot economically serve their customers by radically over-provisioning bandwidth throughout their networks to guarantee the same low-latency, low-jitter, and low-loss performance at all times for all applications, whether those applications are performance-sensitive or not.[51]

[50] This discussion is meant to be illustrative rather than comprehensive. For example, content providers also can reduce data-transfer times through digital compression technologies.

[51] Indeed, even on the circuit-switched PSTN, carriers cannot economically over-provision capacity so that all calls by all callers can be completed at all times, which is why some callers receive a "fast busy" signal during certain peak calling periods.

2. Differentiated Service Handling, Buffering, and Queuing

As discussed, network engineers manage QoS for real-time applications such as streaming video—which are often highly sensitive to latency and jitter—by configuring routers to prioritize packets with DSCP-field ("DiffServ") markings that indicate how the packets should be handled in the event of congestion.[52] Routers typically implement this task through *buffering* and *queuing* techniques. The costs of a network that employ DiffServ techniques are substantially lower for all users than the costs of a network that addresses performance needs solely through increases in capacity.[53] Indeed, Cisco estimates that these techniques, when used to prioritize up to 10% of a network's traffic, will more than double the network's bandwidth in real terms.[54]

Although queuing and buffering techniques are complex, the following captures the basics. Routers transfer packets between links in a network in time intervals typically measured in a few milliseconds. It is not uncommon, however, for the packet load on a particular link (*i.e.*, the number of packets attempting to access the link) to spike briefly above the link's capacity. When this happens, more packets may arrive at the link than can be placed immediately on the link. To handle this situation, network engineers equip routers with "buffers," which very briefly store excess packets until capacity on the link becomes available. If enough packets arrive to fill up the buffer, newly arriving packets are dropped and may be resent.

"Queuing" involves deciding the order in which buffers release packets from a router onto a link.[55] If a router supports DiffServ, each service class is assigned to separate buffers. Network engineers typically design the buffers for *loss-sensitive* service classes to accommodate more packets than do the buffers for *non*-loss-sensitive service classes. And buffers designed for *latency*-sensitive service classes will be "polled" more frequently to release their packets quickly onto the link. If the buffer is empty, the polling process moves to the next buffer. All buffers are polled often enough to give each service class the opportunity to consume at least its prescribed minimum amount of bandwidth.

[52] *See generally* Murat Yuksel, *et al.*, *Value of Supporting Class-of-Service in IP Backbones* (2007) (*"RPI Study"*) (http://www.cse.unr.edu/~yuksem/my-papers/iwqos07.pdf).

[53] *See RPI Study, supra; see also* RPI Press Release, "Undifferentiated Networks Would Require Significant Extra Capacity," July 2, 2007 (quoting RPI professor Shivkumar Kalyanaraman (coauthor of the RPI study): "The study makes clear that there are substantial additional costs for the extra capacity required to operate networks in which all traffic is treated alike, and carrying traffic that needs to still be assured performance as specified in service level agreements (SLAs).").

[54] Cisco, A Discussion with the FCC on the Open Internet, at 17 (Dec. 8, 2009), (http://www.openinternet.gov/workshops/docs/ws_tech_advisory_process/Cisco%20FCC%20Network%20Management%20Presentation%20120809.pdf).

[55] *See, e.g.*, Chuck Semeria, *Supporting Differentiated Service Classes: Active Queue Memory Management*, at 5, Juniper Networks (2002) (http://www.juniper.net/solutions/literature/white_papers/200021.pdf); OpenBSD, *PF: Packet Queuing and Prioritization* (2007) (http://www.openbsd.org/faq/pf/queueing.html).

Because latency and jitter impair real-time applications much more than non-real-time applications, this technique ensures the most efficient and pro-consumer allocation of scarce network resources—the link capacity between two routers or between a router and an end point. Again, this technique assures that every service class may "claim" at least the minimum bandwidth needed to support normal operations for that class, even during periods of network congestion. In addition, when the network is *not* congested, buffers for less performance-sensitive service classes may claim unused capacity that has not been claimed by the buffers for the more performance-sensitive classes. Since congestion tends to be sporadic and momentary, the division of traffic into these classes of service has no effect on any class the vast majority of the time.

Choices among queuing techniques—the algorithms that determine the manner in which buffers sequentially deliver traffic to transport links—are inherently provider-specific, and there "are no real industry standards."[56] Moreover, queuing methodologies are highly dynamic: equipment vendors and network providers are constantly improving existing methodologies and inventing new ones. Thus, each network provider must balance the costs and benefits of the various queuing methodologies to select the one that best meets the needs of its customers.

In addition to "prioritization" at the IP layer (*i.e.*, DiffServ), many other protocols at other layers also allow network operators, content providers, and others to "enhance" or "prioritize" particular data, including data consisting of Internet access traffic. As discussed, these include differential-handling techniques at Layer 2 (*e.g.*, Ethernet, ATM, and Frame Relay) and Layer 2.5 (MPLS). Likewise, some Layer 7 protocols, such as the new SPDY protocol created and promoted by Google, appear to enable content providers to prioritize some HTTP data streams over others so that some content (perhaps Google-sponsored advertisements) will appear first when a webpage downloads.[57] These and similar practices are widespread; all are "non-neutral" in that they prioritize some traffic over others; and any "nondiscrimination" rule would draw many of them into question for the first time. Thus, to the extent any regulator proposes to regulate "prioritization" that affects the Internet, it is wading into a vast ocean of technologies and commercial relationships.

3. Congestion Avoidance

Content Delivery Networks. As explained in our discussion of CDN services, one effective way a content provider can surpass its rivals in on-line performance is to minimize the number of hops its packets must make en route to end users, thereby reducing processing- and congestion-related delays. Under the most prevalent such method, a provider caches its data (such as webpages and media files) in multiple locations near the regional ISPs serving its geographically dispersed end users. When an end user requests the data, a cache server can convey the requested packets quickly and reliably from its nearby location, thereby sparing them

[56] Semeria, *Supporting Differentiated Service Classes*, *supra*, at 4.

[57] *See* SPDY: An experimental protocol for a faster web, http://dev.chromium.org/spdy/spdy-whitepaper; Mike Belshe & Roberto Peon, *SPDY Protocol*, http://dev.chromium.org/spdy/spdy-protocol.

a long, multiple-hop trip through potential bottlenecks on any of several different networks. As discussed, some companies, such as Akamai and Limelight, provide this CDN service commercially to third parties, whereas others, such as Google, build CDNs of their own.

CDN Collocation. Some content providers and broadband networks have begun exploring content distribution methods that would involve direct interconnection and caching of content not just *close to* the broadband provider's access/aggregation networks, but *within* those networks as well. Such arrangements, known as "CDN collocation," eliminate the need to deliver content through a transit or peering link when the end user requests it. Depending on the context, this approach often allows content providers to reach end users more economically and with superior performance as compared to more conventional CDN peering or transit arrangements. For example, Google is reportedly negotiating an arrangement to pay British Telecom to store Google's content within BT's (and other ISPs') access networks for efficient transmission to end users.[58] Such arrangements "enable[] ISPs to store content within *their own* networks," such that "[t]he media companies would pay them, rather than the likes of Akamai, and get a guaranteed service even at peak times."[59]

Paid Peering. Traditionally, large content providers and CDNs have entered into comprehensive transit relationships with large backbone providers to convey their traffic to many different ISPs within the Internet. Backbone providers have often implemented these arrangements by selling these customers enterprise-class Internet access service and interconnecting with them by means of robust, high-capacity facilities. If a content provider wishes to interconnect directly at the peering links of an ISP to obtain closer network proximity to its end users, but does not meet the criteria for settlement-free peering, it may enter into bilateral *paid peering* arrangements with certain ISPs. Under such arrangements, the content provider pays the network operator for such interconnection but at rates lower than it would pay under the traditional transit model if it had contracted with a backbone provider to deliver its traffic throughout the Internet.[60] Moreover, as explained by the University of Michigan study noted above, Google and other dominant content providers have assumed sufficient market clout that they have now begun interconnecting with ISPs on a settlement-free basis.

IP Multicast. When providing high-definition video streams of popular events in real time, content providers face prohibitive costs if they must arrange for the transport of many redundant streams on an end-to-end *unicast* basis: *i.e.*, as separate streams from a centralized source to each of the many end users that wishes to receive the content. As discussed, a content provider can reduce those costs by hiring or building CDNs to replicate and disperse its content-transmitting nodes closer to an ISP's end-users and thereby reduce the total network resources that each individual stream must consume en route to a given end user. CDNs, however, require

[58] Richard Wray, *BT and Google in talks over creating video delivery network for ISPs*, The Guardian, Dec. 7, 2009 (http://www.guardian.co.uk/business/2009/dec/07/bt-google-isp-digital-video).

[59] *Id.* (emphasis added).

[60] *See* George Ou, *FCC NPRM ban on Paid Peering harms new innovators*, Nov. 10, 2009 (http://www.digitalsociety.org/2009/11/fcc-nprm-ban-on-paid-peering-harms-new-innovators/).

substantial investments in cache servers to store all of this content, along with other infrastructure to transport content to all of these cache servers. And from a network resource perspective, too, CDNs can be sub optimally efficient for the distribution of any content that many users in the same area wish to obtain at the same time, such as streaming real-time video: each cache server must transmit hundreds or thousands of redundant streams to all geographically proximate users that request it.

One promising solution is IP multicast, "a bandwidth-conserving technology specifically designed to reduce traffic by simultaneously delivering a single stream of information to potentially thousands of corporate recipients or homes," while requiring only a single stream (rather than one per viewer) at the content source.[61] Suppose a content provider wants to stream video coverage of a highly popular sports event over the Internet simultaneously to thousands of subscribers in the same geographic area. Under an IP multicast approach, the content provider arranges with the ISP for the routers in an ISP's access/aggregation network to instantaneously replicate copies of the incoming packets and transmit them to multiple local users simultaneously, depending on which users have requested the relevant content. No caching is required, and redundancy is enormously reduced by moving the packet duplication as close as possible to the ultimate recipients. IP multicast thus dramatically lowers the cost of high-quality distribution by "minimiz[ing] the burden on both sending and receiving hosts and reduc[ing] overall network traffic."[62] And if multicast is used in conjunction with CDN technology (i.e., a CDN cache server transmits content to a multicast-enabled router), even greater bandwidth efficiencies may be possible, which opens up new opportunities for content and application providers to deliver higher-quality services over the Internet. Indeed, multicast already plays a vital role in the efficient delivery of *non*-Internet-based IPTV services, such as AT&T's U-verse video service.

Paid peering, CDN collocation, and multicast arrangements are unambiguously pro-consumer and should be welcomed. CDN collocation and multicast in particular will be essential to the distribution of affordable streaming high-definition video over the Internet. These and the similar technologies discussed above illustrate a broader point. By targeting QoS enhancements to QoS-sensitive applications, network operators can facilitate the development of innovative Internet applications that would not be feasible to provide otherwise. The use of such techniques thus expands both the business opportunities available to application and content providers and, in turn, the applications and content available to consumers. This virtuous cycle—smarter networks supporting QoS-sensitive applications and content, thereby increasing consumer welfare—will fuel enormous economic growth *if* policymakers encourage the deployment of shared, multi-purpose broadband platforms that are capable of delivering a range of QoS capabilities to content and application providers.

[61] Cisco White Paper, *IP Multicast Technical Overview*, at 1 (Aug. 2007) ("*Cisco Multicast White Paper*") (emphasis omitted); *see also* Metaswitch Networks, *IP Multicast Explained*, at 2 (2004).

[62] *Cisco Multicast White Paper* at 1.

Unfortunately, a broad “nondiscrimination” rule could prohibit such QoS arrangements insofar as they would involve payments by content providers for especially efficient and high-quality distribution of their content within specific access/aggregation networks.[63]

4. P2P Content Distribution

Under traditional content-distribution methods, a complete copy of a content file (such as a song or a feature-length movie) is stored on servers and distributed from there to the end users that request it. In contrast, P2P technologies disassemble content into small files and widely distribute them to different end-user computers for storage and subsequent retrieval and reassembly by other end users.[64] The result is the functional equivalent of a massively distributed server network, in which each end user’s computer acts as an individual server for a portion of the content being distributed. Although P2P technology has been used (and continues to be used) by some parties for the unlawful distribution of pirated content, it has also been adopted as a mechanism for the distribution of lawful content by a variety of companies. Vuze, for example, claims that it “has attracted over 100 content partners, including A&E, BBC, CBC, G4 TV, The History Channel, Ministry of Sound, National Geographic, PBS, Showtime, Starz Media, The Poker Channel, TV Guide Channel, and many more.”[65]

In the past, content providers (and their distribution partners) have traditionally borne the costs of maintaining enough centralized storage and server capacity to convey their content to end users. By converting end-user devices into content caches for other end users, however, P2P technology offers a way to shift those costs to end users and their network providers. But while P2P distribution may thereby offer content providers a relatively cheap storage and distribution mechanism, most current implementations of P2P applications impose enormous upstream and downstream traffic burdens on broadband networks, particularly with the rise of shared video. As network-engineering scholars have explained, this “network-oblivious peering strategy . . . may cause traffic to scatter and unnecessarily traverse multiple links within a provider’s network, leading to much higher load on some backbone links” and producing “inefficiencies for both P2P applications and network providers.”[66]

None of this is to say that P2P technologies are inherently inefficient in all instances. Quite to the contrary, the distributed, peer-based content-delivery model underlying today’s P2P technologies could bring tremendous benefits for content providers, network operators and consumers alike—faster distribution at lower cost in some circumstances—*if* the industry can

[63] *See* Ou, *FCC NPRM ban, supra.*

[64] *See, e.g.*, Detlef Schoder, Kai Fischbach, & Christian Schmitt, *Core Concepts in Peer-to-Peer Networking* (2005) (http://www.idea-group.com/downloads/excerpts/Subramanian01.pdf).

[65] Petition for Rulemaking, *Vuze Inc. Petition to Establish Rules Governing Network Management Practices by Broadband Network Operators, Broadband Industry Practices*, WC Docket No. 07-52, at 5-6 (Nov. 14, 2007).

[66] Haiyong Xie et al., *P4P: Explicit Communications for Cooperative Control Between P2P and Network Providers*, Distributed Computing Industry Ass’n, at 1 (May 2007) (“*P4P: Explicit Communications*”) (http://www.dcia.info/documents/P4P_Overview.pdf).

resolve the current inefficiencies in that model. To that end, AT&T is part of a new industry-wide working group—composed of representatives from BitTorrent, LimeWire, Cisco, Verizon, Verisign, and researchers from Yale and Washington Universities, among others—that is trying to develop an efficient, network-*aware* peer-to-peer technology. Known as "P4P," this new generation of technology is being developed to optimize network resources rather than hoard them.[67]

5. Security Screening

Finally, protection from spam, worms, viruses, distributed denial-of-service attacks, and other malicious behavior on the Internet is critically important to network management, and no net neutrality advocate seriously contends otherwise. An important but often overlooked benefit of these robust network security practices is that keeping harmful traffic out of a network in the first place can significantly reduce network congestion by conserving network resources for traffic from legitimate sources. According to Verizon Wireless, for example, a single spammer tried in 2007 to send 12 million text messages to its wireless customers.[68] As Verizon Wireless explained, wireless spam "impairs the delivery of legitimate messages, and because spam is often sent in high volume over short periods of time, it can place a strain on overall performance of the wireless network," and "[t]here's a lot of time and money that goes into blocking all of that."[69]

With multiple petabytes of data passing through its network each business day, the first crucial step to effective network security for AT&T or any other network provider is rapid identification of illegitimate packets. By closely monitoring the traffic coming into and out of its network, a network provider like AT&T can take steps to detect the early stages of attacks on network integrity and activate mechanisms to minimize the effects of those attacks. "Before a worm strikes, technicians see strange spikes of traffic going to normally obscure ports, as malware developers test and tweak their code. A sudden, sharp increase in the amount of Web traffic worldwide could mean breaking news—or a distributed denial-of-service attack being lobbed at a single company halfway around the world."[70] For example, "AT&T security analysts knew about the 2003 Slammer worm before it hit, because of strange traffic going to port 1434."[71]

Wireless broadband providers may also employ additional techniques to safeguard the security of wireless networks. AT&T, for example, uses a technique called "Code Signing" to

[67] *See id.*

[68] *See* Verizon, Press Release, *Wireless Spammer Target Of Legal Action By Verizon Wireless*, June 1, 2007, http://news.vzw.com/news/2007/06/pr2007-06-01b.html.

[69] Howard Buskirk, *Verizon Wireless Says Filters Cut Wireless Spam's Impact*, Communications Daily, June 4, 2007.

[70] Sarah D. Scalet, *Introducing AT&T, Your Internet Security Company*, CIO, May 17, 2007, http://www.cio.com/article/110250/Introducing_AT_T_Your_Internet_Security_Company.

[71] *Id.*

control access to the network at the device and application lawyer. AT&T-partnered devices are configured to allow third-party applications to access the network only once AT&T has been reassured (either through testing or through the developer's affirmative, contractual representation) that the application will not introduce malicious code or some other intrusive agent into the network. This "certification" process also helps prevent the introduction of applications that inappropriately access customer data (e.g., contact lists, location information) and violate customers' reasonable privacy expectations.

Any net neutrality regulations that would restrict the wide latitude network providers have to perform such critical functions would strike a serious blow to network security and consumer safety.