High Tech Forum
869 South Cole Drive
Lakewood, Colorado

Hearing Before United States House of Representatives

Committee on Energy and Commerce

Subcommittee on Communications and Technology

April 17, 2018

Testimony of Richard Bennett

Founder, High Tech Forum

Summary[1]

The Internet has undergone substantial change since it was initially designed in the 1970s. A network built to allow academics to use remote computers is now open to the public. The Internet has disrupted a plethora of major industries, and it has in turn been disrupted by advances in networking technology and applications. The backbone-oriented, end-to-end architecture of the research Internet has given way to today's Content Delivery Network (CDN) model. Large firms such as Google, Amazon, and Netflix have gained control over their network traffic by playing the roles of both network and application.

The traditional regulatory model that separated content from communications no longer fits; large content interests own worldwide networking facilities, only connecting to

[1] This is a lightly edited version of comments filed on April 13, 2018; most changes are for clarity and style.

Internet Service Providers (ISPs) to perform the relatively simple task of delivering streams of Internet packets over the last mile.

Regulators have struggled to keep up with the transformation of electronic communications from the telephone network to the Internet. While some infrastructure providers are highly regulated, others – such as Cloudflare and the pure CDNs – are almost completely free to behave as they please. While net neutrality once promised regulators a short-cut around the complexity of traditional competition law and economics, it has become all but impossible to reach consensus on its most troubling part, the presumptive ban on "paid prioritization" introduced in the FCC's 2015 Open Internet Order.

The naïve view of the Internet as a magically self-organizing system, enabling all applications to utilize the resources they desire without active engagement by network operators, has not proved out in the real world. While the web is still a dominant application, the Internet continues to spawn novel applications and better ways of enabling traditional ones with less expense and greater reliability.

A diverse pool of users and applications competes for access to critical network resources such as bandwidth, latency, packet loss, and jitter. This competition is central to the Internet's packet switching design and is therefore unavoidable. ISPs – both fixed-line and mobile – distinguish themselves largely in their expertise at managing network resources in optimal ways that meet consumer expectations.

Contrary to popular misconception, optimizations that improve the experience of users of real-time applications need not perceptibly degrade traditional applications such as video streaming, the web, or email. This is because traditional applications have extremely high

tolerance for variations in packet stream delivery speed because they buffer. Internet optimization is therefore *not* a zero-sum game.

New networking product lines such as Wide Area Network (WAN) Edge Infrastructure, Software Defined Networking (SDN), and Managed Services overcome shortcomings in the Internet's design by enabling dynamic routing or path selection. These capabilities enable the Internet to replace costly private lines for many enterprise applications.

ISPs can do the best job of traffic optimization when they can identify the nature and requirements of individual packet streams. The most effective way to do this – while preserving privacy – is to allow application developers to register applications requiring special treatment, and even to pay for such treatment in some circumstances. If this is not allowed, ISPs, CDNs, and dominant firms will not face robust competition.

A tremendous amount of misinformation is afoot about the factors that determine the performance of web pages, much of it by well-meaning advocates. The reality is that the speed at which web pages load on US broadband networks is largely controlled by choices made by web page owners. While the average speed of US broadband networks has increased 35% per year for a decade, web performance has remained stagnant, even decreasing in 2016.

The interests of innovators are best served when they are able to purchase the network services they need without undertaking the breathtaking expense of building the networks of data centers owned by the five largest US firms. A generally permissive approach to the design and sale of innovative network services – with proper oversight by well-informed regulators – is the best way forward for the Internet.

While net neutrality – and especially the ban on Internet optimization for a fee – has been held out to policy makers as a silver bullet that solves the problem of keeping the Internet on track, it is a false hope. In reality, nearly all forms of network optimization are good in come contexts and bad in others. Our regulators need to develop the wisdom to tell the difference.

Contents

Introduction

Chairman Walden, Chairman Blackburn, Ranking Member Pallone, Ranking Member Doyle, and Members of the Committee,

Internet traffic management – short of blocking and throttling – has proven to be the most difficult element of the network neutrality construct to commit to regulation. Net neutrality emerged in the early part of the century as a potentially simple way to craft Internet regulations that balanced efficiency and fairness, provided a short cut to enforcing provisions against anti-competitive conduct, and encouraged infrastructure owners to invest in resource upgrades. In practice, net neutrality regulations are anything but simple because broad bans on behavior that can be either constructive or destructive depending on context are necessarily riddled with exceptions, loopholes, and special circumstances. A brief summary of the FCC's actions is instructive.

The Original Idea of Net Neutrality

In the first articulation of net neutrality *per se*, Professor Tim Wu attempted to split the baby.[2] On the one hand, he allowed broadband Internet service providers (ISPs) to manage traffic as they saw fit for the applications they provided (such as voice and video), but on the other he insisted they manage their "Internet gateway" in a blindly nondiscriminatory manner:

> *…absent evidence of harm to the local network or the interests of other users, broadband carriers should not discriminate in how they treat traffic on their broadband network on the basis of inter-network criteria.*[3]

By way of illustration, Wu explained that users of resource-intensive applications such as online gaming should pay for sufficient resources to run these applications successfully; but he did not approve of ISPs levying tolls for the use of gaming apps or other classes of applications.

Wu regarded his approach as superior to "open access" regimes such as Title II broadband unbundling that did nothing to remedy the "Internet's greatest deviation from network neutrality…[the] favoritism of data applications, as a class, over latency-sensitive applications involving voice or video." Wu proposed to permit ISPs to actively manage Internet traffic as long as they didn't do so arbitrarily.

---

[2] Tim Wu, "Network Neutrality, Broadband Discrimination," *Journal of Telecommunications and High Technology Law* 2 (2003): 141, https://doi.org/10.2139/ssrn.388863.

[3] Wu.

FCC Policy Statements Preceding Regulation

FCC Chairman Michael Powell proposed a generally similar approach a year after Wu presented his paper at a Silicon Flatirons conference in Boulder, Colorado. Also speaking in Boulder, Powell proposed a bill of rights for Internet users that came to be known as the "Four Freedoms of the Internet."[4] Like Wu, Powell insisted that ISPs have the power to actively manage traffic, but went further in declaring that this right should only be limited by disclosure and the ability of users to run the applications of their choice:

> *I recognize that network operators have a legitimate need to manage their networks and ensure a quality experience, thus reasonable limits sometimes must be placed in service contracts. Such restraints, however, should be clearly spelled out and should be as minimal as necessary.*[5]

In 2005, the Kevin Martin FCC endorsed a slightly modified version of the *Four Freedoms* as the Internet Policy Statement, but refrained from issuing regulations.[6]

FCC Traffic Management Regulations

In 2010, the Genachowski FCC issued the Open Internet Order, America's first set of direct regulations over ISP traffic management practices. This order banned "unreasonable

---

[4] Michael K. Powell, "PRESERVING INTERNET FREEDOM: GUIDING PRINCIPLES FOR THE INDUSTRY" (Federal Communications Commission, February 8, 2004), https://apps.fcc.gov/edocs_public/attachmatch/DOC-243556A1.pdf.

[5] Powell.

[6] "FCC Adopts Policy Statement," Federal Communications Commission, August 5, 2005, https://apps.fcc.gov/edocs_public/attachmatch/FCC-05-151A1.pdf.

discrimination" and permitted the sale of "specialized services" as long as such services were not used to access the Internet:[7]

> *We recognize that broadband providers may offer other services over the same last-mile connections used to provide broadband service. These "specialized services" can benefit end users and spur investment, but they may also present risks to the open Internet. We will closely monitor specialized services and their effects on broadband service to ensure, through all available mechanisms, that they supplement but do not supplant the open Internet.*

These services were considered to include enterprise VoIP and similar applications that did not touch the web.

In 2015, the Wheeler FCC strengthened the presumption against "paid prioritization" in order to ban it:[8]

> *A person engaged in the provision of broadband Internet access service, insofar as such person is so engaged, shall not engage in paid prioritization.*
>
> *"Paid prioritization" refers to the management of a broadband provider's network to directly or indirectly favor some traffic over other traffic,*

[7] Federal Communications Commission, "Report and Order: Preserving the Open Internet" (2010), http://www.fcc.gov/Daily_Releases/Daily_Business/2010/db1223/FCC-10-201A1.pdf.

[8] Federal Communications Commission, "Report and Order on Remand, Declaratory Ruling, and Order in the Matter of Protecting and Promoting the Open Internet" (FCC, February 26, 2015), https://apps.fcc.gov/edocs_public/attachmatch/FCC-15-24A1.pdf.

> *including through use of techniques such as traffic shaping, prioritization, resource reservation, or other forms of preferential traffic management, either (a) in exchange for consideration (monetary or otherwise) from a third party, or (b) to benefit an affiliated entity.*[9]

The Wheeler order substantially departed from the FCC's (by then) long-standing light-touch approach by banning practices that had been regarded as constructive (with some caveats) by previous commissions of both parties. On its face, the paid prioritization ban could make services that compete with carrier-provided Voice over LTE (VoLTE) unlawful because such services would need resource reservation (using the IETF's Integrated Services standard[10]) to be competitive, especially at higher definition.

Restoring Internet Freedom

The Pai FCC's 2017 *Restoring Internet Freedom Order* erased Wheeler's ban on paid prioritization:[11]

> *We also decline to adopt a ban on paid prioritization. The transparency rule we adopt, along with enforcement of the antitrust and consumer protection laws, addresses many of the concerns regarding paid prioritization raised in this record. Thus, the incremental benefit of a ban on*

[9] Federal Communications Commission.

[10] R. Braden, D. Clark, and S. Shenker, "RFC 1633 - Integrated Services in the Internet Architecture: An Overview" June 1994, http://tools.ietf.org/rfc/rfc1633.txt.

[11] Federal Communications Commission, "Restoring Internet Freedom: Declaratory Ruling, Report and Order, and Order" (Federal Communications Commission, December 14, 2017), https://apps.fcc.gov/edocs_public/attachmatch/FCC-17-166A1.pdf.

> *paid prioritization is likely to be small or zero. On the other hand, we expect that eliminating the ban on paid prioritization will help spur innovation and experimentation, encourage network investment, and better allocate the costs of infrastructure, likely benefiting consumers and competition. Thus, the costs (forgone benefits) of the ban are likely significant and outweigh any incremental benefits of a ban on paid prioritization.*[12]

As noted, the *Restoring Internet Freedom Order* required disclosure of paid prioritization while permitting the practice. Thus, it is consistent with the *Four Freedoms*.

The history of FCC regulation of traffic management shows a general acceptance of Internet optimization – even for a fee – with proper disclosure, apart from the 2015 order. But even in the orders and statements that support the practice, we see significant variation in presumptions and general reasoning. In part, this variation reflects differences in prevailing technologies and practices; it also reflects varying degrees of technology awareness on the part of Commission staff and leadership.

While all FCC Internet regulations have been forward-looking to some extent, it's fair to say that the Wheeler order pays greater attention to historical policies and practices. It was especially attentive to agency actions such as the Computer Inquiries dating back to the 1960s.[13] The Pai order is the only one to mention 5G; this technology is crucially important

---

[12] Federal Communications Commission.

[13] "FCC Computer Inquiries," *Wikipedia*, September 26, 2017, https://en.wikipedia.org/w/index.php?title=FCC_Computer_Inquiries&oldid=802553161. See also Tom Wheeler, "Remarks of Tom Wheeler at Aspen Institute 2016 Communication Policy Conference" (Federal Communications Commission, August 14, 2016).

because it's the most likely path to more robust, facilities-based competition for residential Internet services. Most observers agree that competition is a more effective means than regulation for ensuring constructive organizational behavior.

Why Optimize the Internet?

The Internet has become the world's primary communication medium. As such, it is called-upon to carry information for a variety of applications, such as:

- The well-known World Wide Web;
- Internet Protocol-based telephone calls[14] intersecting with the traditional telephone network;
- Large private networks intersecting with the public Internet at several points (the five largest US corporations maintain such networks);
- Private communications between offices of organizations that use the Internet for Wide-Area Network (WAN) connectivity in lieu of purchasing Business Data Services;
- Public safety communications between dispatchers and first responders;
- Real-time communications among Internet of Things (IoT) devices, gamers, or specialized applications such as air traffic control.

This wide range of usage patterns requires a great deal of agility. In fact, there is no "one-size-fits-all" traffic management technique that efficiently meets the needs of all applications. The only approach that has ever worked is to treat each application with a high

[14] Whether over the Internet or on a private intra-domain (intra-network-providers) or inter-domain (cross-provider) basis.

degree of sensitivity to its requirements. This is why Service Level Agreements (SLA) specifying precise Quality of Service (QoS) parameters are the norm for commercial Internet service agreements.[15]

The Internet is a Statistical System

Unlike the traditional telephone network, the Internet is a statistical system that shares resources among a broad pool of applications. The telephone network ensures that each portion of each phone call is delivered with the same fidelity, latency, and quality within the limits of the communications medium (wire or radio), but this quality assurance extracts a high price in terms of network efficiency. The telephone network divides resources into fixed buckets or channels of capacity and then allocates one per call. Because it is sized for peak load, under normal load most of the network's capacity goes to waste.

The cable television network uses similar design logic. It consists of a number of 6 MHz radio frequency channels, each of which is statically assigned to a television channel, a voice channel, or an Internet channel. A typical cable TV network once assigned a pair of channels for Internet access, but now it will tend to assign 24 channels for downstream Internet and 2 for upstream[16]. Channels carrying unwatched TV channels effectively go to waste.

The Internet Uses Packet Switching

The Internet uses a technology known as "packet switching" that allows multiple users and applications to share a single very large communication channel. While a traditional

[15] Network Quality of Service is measured in terms of data volume, latency (delay), jitter (variations in latency) and packet loss.

[16] Or even more channels, as this changes routinely as bandwidth requirements, and hence speeds to end users, increase.

network with 24 channels would permit 24 users to communicate in parallel, each would only be allowed to use 1/24th of the network's design capacity.

Packet switching would allow the 24 to use the channel in series, one after another. Each user's packets would transit the network 24 times faster, but they would have to wait behind packets already in flight or waiting for transmission. When fewer than 24 users are active, they will be able to access more bandwidth on the packet switched network than they would on the traditional circuit-switched network. When all 24 are active, their ability to access bandwidth would be the same, but some would experience more delay (latency) than they would on the traditional network.

Packet switched networks are typically provisioned with sufficient capacity that most users experience high speeds most of the time[17]. The packet switching design is efficient in terms of bandwidth allocation – it permits applications that need high bandwidth to obtain it – but they don't provide the same consistency of delay as circuit switching arrangements.

Packet switching was determined to be the preferred technology for computer applications as early as the 1960s because of its ability to provide flexible service. Computers run applications, and applications have a variety of different communications requirements. Hence, a flexible network tends to serve their interests better than network designed to support a single application such as telephone calls or TV viewing.

Hence, the Internet does two things that previous networks did not do:

1) It allows multiple users to dynamically share common network facilities; and:

[17] This is largely proven out by results recorded for ISPs since 2010 by the FCC's "Measuring Broadband America" program.

2) It allows applications to compete with each other for access to a common pool of network resources such as bandwidth (capacity), delay (latency), jitter (variations in latency) and packet loss.

This mode of operation raises issues with access to network resources that previous technologies did not face, at least not to the same degree.

Resource Contention is Unavoidable on the Internet

The Internet is a system in which multiple users run a variety of applications over shared infrastructure. Even when the last mile cable is unshared – as is the case for DSL and Ethernet – the rest of the system, after that brief first hop link, is shared. Overall, the Internet consists of several levels of traffic aggregation and disaggregation. Sharing is inherent in the Internet's design.

Access to shared resources of any kind implies the development and implementation of a sharing policy. For the Internet, this policy can take various forms, each of which has varying degrees of Quality of Service impact on different applications:[18]

1. ISPs generally parcel bandwidth into service tiers that simply pre-allocate fractional portions of total network bandwidth to each account. While a cable system may provide one gigabit per second of bandwidth to a neighborhood, each subscriber is limited to using their **subscribed quota** of 50, 100, or 250 Mbps, for example. While the total of the subscriptions in a neighborhood will exceed actual capacity, network engineering

[18] See a detailed examination of sharing policies in my paper: Richard Bennett, "Arrested Development: How Policy Failure Impairs Internet Progress" (Washington, D.C.: American Enterprise Institute, December 2015), http://www.aei.org/publication/arrested-development-how-policy-failure-impairs-internet-progress/.

sensitive to usage patterns ensures that the expected speed is generally achieved by all active users. Some degree of over-subscription is essential to Internet economics, which depend on statistical multiplexing.

2. When networks are lightly used, operators may simply forward each packet of information on a **first-come, first-served** basis with no regard for its specific needs such as urgency or necessity. Because the Internet carries both "elastic" and "non-elastic" data[19], this policy is sub-optimal under most load conditions.
3. Operators commonly employ policies that seek to identify application types in order to apply **smart queueing policies** under moderately high load conditions. Voice packets have strict time requirements – less than 150 ms from end-to-end. But packets containing software code, such as patch updates, have very loose time requirements. Hence, it is sensible for operators to prioritize individual voice packets over software patches by **moving voice packets to the head of the transmission queue** feeding a moderately loaded data link. This practice does not impair the patch application because it doesn't alter the time at which its final packet arrives or it is not essential for the download to finish in a particular second of time, whereas the user of the voice application will soon audibly noticed any delays in that application. File transfer applications (of which the web is one) are not impacted by the delivery time of intermediate packets, only the final one. But each voice packet is impaired by delay.

[19] Elastic data is buffered at the destination before use and is often stored, while non-elastic data is used immediately and then discarded. When packets of elastic data are lost, they are re-transmitted. When packets of non-elastic data are lost, the application simply moves on to the next packet in the stream.

4. When network load increases from moderate to high, it becomes necessary for operators to discard data packets in order to maintain the efficiency of the Internet's Transmission Control Protocol (TCP). Because non-elastic applications don't use TCP, there is no value in dropping their packets. Operators have a number of choices for dropping TCP packets, such as dropping the newest, the oldest, or a random selection. They may also apply drop quotas to particular streams, especially very heavy ones. A common method is **Random Early Detection** (RED), an algorithm that discards packets at random but at a rate that corresponds to the degree of load on the network segment.[20] Classical RED is insensitive to Quality of Service, hence more sophisticated versions such as Weighted RED and Adaptive RED have been developed.[21]

   In more recent years, various forms of Active Queue Management (AQM) have been developed as well, such as CoDel (Controlled Delay) and Proportional Integral controller Enhanced (PIE) queueing. To overcome related problems, Google has also developed and deployed the new Quick UDP Internet Connections (QUIC) protocol, as an alternative to TCP flows for web-based and other Internet traffic.

5. Operators also commonly impose **quotas on the amounts of data** users are permitted to send or receive over a given period of time, especially when resource contention is high. Quota-based discard policies drop packets from heavy users before dropping those from light users. Alternatively, high load application streams – such as code downloads

---

[20] S. Floyd and V. Jacobson, "Random Early Detection Gateways for Congestion Avoidance," *IEEE/ACM Transactions on Networking* 1, no. 4 (August 1993): 397–413, https://doi.org/10.1109/90.251892.

[21] Kevin Wallace, "Weighted Random Early Detection (WRED)," in *Cisco IP Telephony Flash Cards* (Cisco, 2005), http://www.ciscopress.com/articles/article.asp?p=352991&seqNum=8.

that seek to saturate network capacity – would be subject to dropping before light load streams, such as narrowband voice, that moderate their consumption of network resources. Most Internet data today is video streaming, an application that regulates its resource consumption[22].

We sometimes hear claims that network congestion can be eliminated by increasing the capacity of data pipes to some arbitrary bandwidth. If this were the case, the research literature on managing packet networks under load would not be so rich. In reality, building our way out of network congestion is impossible because increases in the capacity of one part of the Internet create more congestion in some other part.

It is also impossible to fully coordinate upgrades because the Internet is not simply a single system managed by the single agent; it's a loose federation of networks that undergo increases in load and capacity at their own rates.

Each time we add capacity to alleviate downstream congestion, we enable more upstream congestion. So the battle for congestion-free networks is never-ending. This is especially true for wireless networks.

Brief Overview of Internet Traffic Management

Some facially fair and uniform traffic management policies have been proved to be very harmful to the Internet. The Internet consists of a mesh of Internet Exchanges in major cities,

[22] Video servers do this using techniques such as Dynamic Adaptive Streaming over HTTP (DASH), where the server can dynamically – every few seconds - change between various quality levels to adapt to changing throughput conditions. Servers can adapt from SD quality when there is very little capacity to HD or 4K when there is an abundance. Servers also adapt to changes in their own CPU and storage resources due to load.

Content Delivery Networks inside Internet Exchanges and ISP networks, middle mile facilities, and last mile networks.

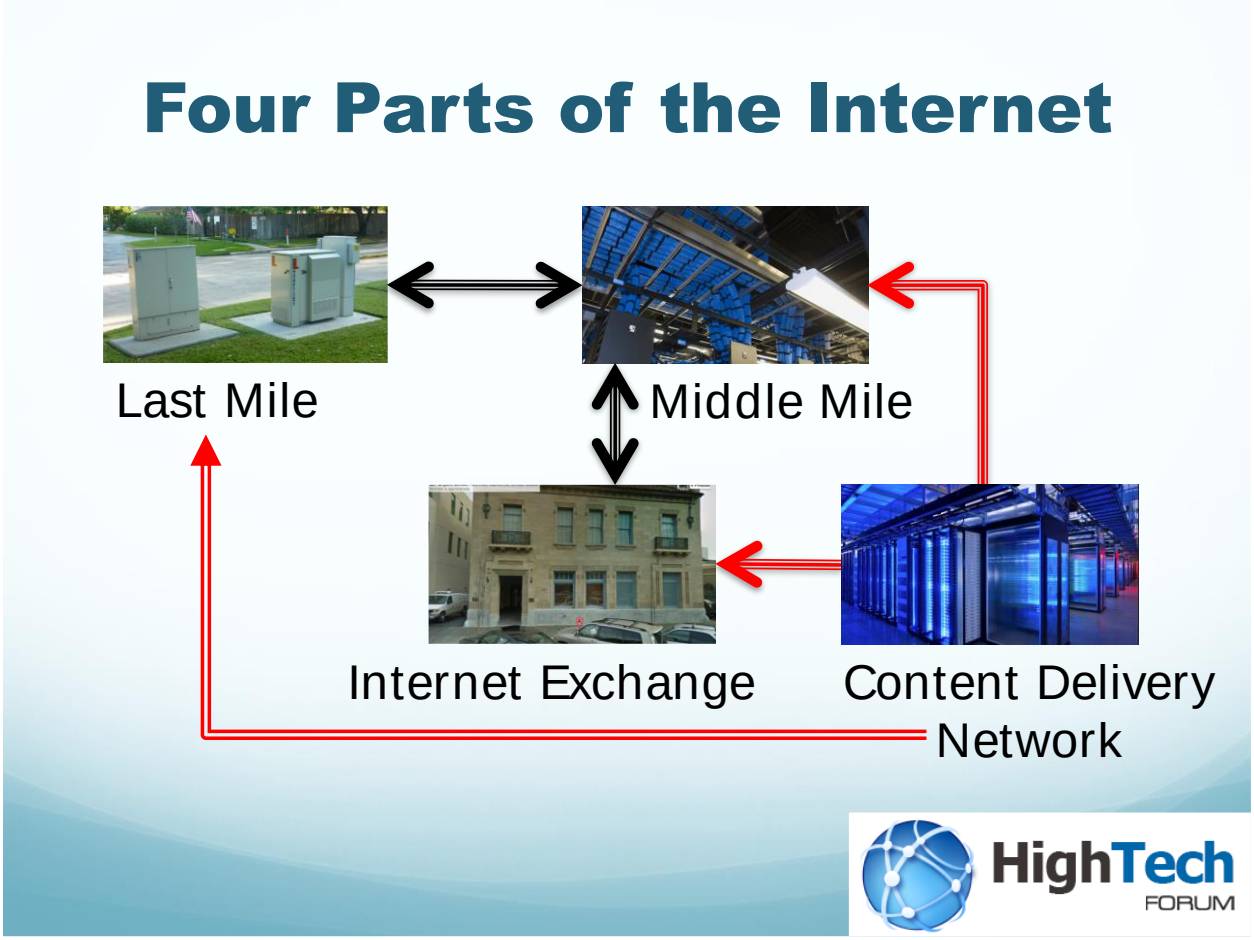


Figure 1: Four Parts of the Internet

Inside each of the four parts, we find Ethernet switches and Internet routers. The parts are interconnected by network circuits or "datalinks", which are typically wavelengths of light transmitted through fiber optic cables. Where datalinks meet routers, computer memories known as queues hold information packets awaiting transmission. The five resource management techniques described in the previous section apply to these queues.

The most simple, uniform management technique that can be applied to Internet queues simply drops new packets when the queue is full. The effect of this method – often

praised by net neutrality advocates[23] – is called "global Internet synchronization". Packet loss is a signal to TCP to slow down; when a series of application streams are told to slow down at the same time, the utilization of the datalink decreases from 100 percent to 50 percent.

Following this sudden decrease, the rate of TCP streams gradually increases until it reaches 100 percent again, and the process repeats. Because the Internet datalink is rarely at 100%, adding bandwidth is relatively costly and ineffective.

Techniques such as RED were developed to prevent synchronization. Because RED has unfortunate side effects on some applications, it has been supplemented and/or replaced with more advanced techniques such as CoDel, QUIC, and SPDY intended to ensure more efficient bandwidth utilization. This is an ongoing research area in computer science.

Optimization is Not a Zero Sum Game.

In the absence of perfect techniques to manage Internet contention, ISPs, transit networks, and both public and private CDNs differentiate traffic according to application type in order to optimize Quality of Experience. The Broadband Internet Technical Advisory Group (of which I am a member) published an excellent report on this topic in 2015, *Differentiated Treatment of Internet Traffic*.[24] The report demonstrates that traffic differentiation is not a zero-sum game due to the nature of Quality of Experience.

The BITAG report observes:

---

[23] M. Chris Riley and Robb Topolski, "The Hidden Harms of Application Bias" (Free Press, November 2009), http://conference.freepress.net/sites/default/files/resources/The_Hidden_Harms_of_Application_Bias.pdf.

[24] Broadband Internet Technical Advisory Group, Inc., "Differentiated Treatment of Internet Traffic" (Boulder: BITAG, October 2015), http://www.bitag.org/documents/BITAG_-_Differentiated_Treatment_of_Internet_Traffic.pdf.

Differentiated treatment can produce a net improvement in Quality of Experience (QoE).

When differentiated treatment is applied with an awareness of the requirements for different types of traffic, it becomes possible to create a benefit without an offsetting loss. For example, some differentiation techniques improve the performance or quality of experience (QoE) for particular applications or classes of applications without negatively impacting the QoE for other applications or classes of applications. The use and development of these techniques has value.[25]

This is especially true when large amounts of video streaming (Netflix, YouTube, Amazon Instant) traffic are present in a residential broadband network. Advocates who argue that it's impossible to prioritize one application without impairing another fail to examine prioritization in proper technical detail.

Internet applications interact with users at the level of transactions, each of which has a beginning, a middle, and an end. Downloading a web page is a transaction; watching a movie is a transaction, and making a Skype video call is a transaction.

Each transaction consists of a stream of packets, from a few in the case of an email to some number of millions in the case of some movies. When two Internet transactions take place on a residential Internet connection, their packet streams are intermingled. If we examine

[25] Broadband Internet Technical Advisory Group, Inc., page iii.

the Internet datalink, we might find clumps of a few hundred video packets interspersed with an occasional Skype packet.

When several hundred video packets are enqueued for transmission inside an ISP network when a Skype packet arrives, it is reasonable to move the Skype packet to the head of the queue. This situation can arise because of the way video streaming services operate. They are connected to ISP networks through very high-capacity datalinks, often 10 – 100 gigabits per second. These datalinks are much faster than last mile datalinks connecting to consumer premises.

They also tend to deliver traffic in an idiosyncratic matter, filling network queues as fast as they can for short periods of time and then waiting before filling it again.

As the report explains:

> ***Managing the impact of streaming video on other traffic***
>
> *A typical video stream, as sent by a server, consists of a series of large bursts of traffic, or "chunks," where each chunk consists of multiple packets transmitted as quickly as possible. Sequential chunks are separated by time periods that can span seconds. The transmission rate for each chunk is much higher than the average rate of the encoded stream, which is a function of the average chunk size and the time between chunks. The video client buffers the chunks and then plays them out at the encoded rate.*
>
> *When a chunk from a video stream arrives at a bottleneck link, it can cause significant delay and jitter for other traffic sharing the same link,*

causing severe degradation in the QoE of time-sensitive applications such as interactive voice. This problem can be mitigated via a technique known as pacing, in which the video stream is differentiated and traffic shaped to a rate equal to or greater than the stream's average rate, but still lower than the bottleneck link's rate. Pacing spaces out the video packets in time, allowing other traffic in between the chunks and in doing so may reduce the latency and jitter experienced by other traffic. Since the first packet in each chunk is not delayed, the net effect of pacing on streaming video is to deliver video packets to the receiver at a more consistent rate without creating any additional delay in video playback. In effect, network pacing performs the same "smoothing" function in the received video content that the receive buffer in the video client would have performed had the chunks been received in discrete high speed bursts, so the QoE for the streaming video may be maintained because the content in each chunk is still received before the decoder needs it.

Pacing is an example of differentiated treatment that is implemented in mobile networks and that acts on the traffic within Internet access services. It may also be implemented by the sending service or application, reducing the need for differentiation in the network. As noted above, this technique can improve the QoE for other traffic without degrading the QoE for OTT video streams.

Pacing replaces the stair-step traffic pattern with a line that more accurately represents the average rate line. By introducing gaps into the clump of video packets, Internet optimization in the form of pacing allows time-sensitive applications such as Skype to more happily coexist with Netflix. Prioritizing Skype packets by moving them to the head of the transmission queue does not impair Netflix because streaming is impacted more by the *volume* of competing data than by its *placement*.

The following diagram from an academic paper illustrates the effect of pacing on a typical video stream.

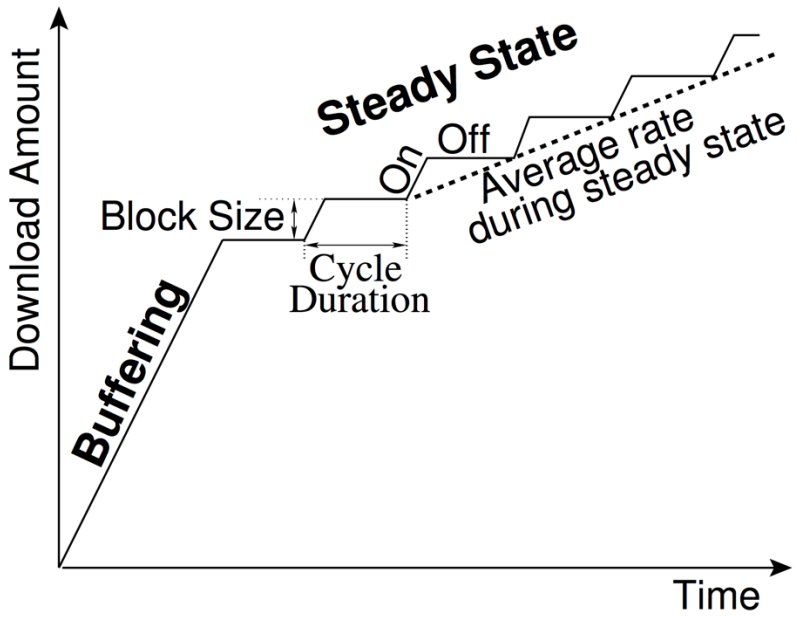


Figure 2: Rao et al., "Network Characteristics of Video Streaming Traffic".[26]

Thus, optimization in the form of prioritization is **not a zero-sum game** in which every boost in queue position in favor of one application harms another application. In fact, this sort

[26] Kenjiro Cho and Association for Computing Machinery, *Proceedings of the Seventh COnference on Emerging Networking EXperiments and Technologies* (New York, NY: ACM, 2011).

of prioritization helps applications that need protection from the high-density, high capacity packet streams associated with video servers and CDNs without producing a perceptible effect on the video stream.

Software-Defined Networks

The Internet is a highly-connected mesh composed of a number of interconnected networks. Between any given pair of Internet endpoints, a number paths exist by which the endpoints can communicate with each other.

For example, a visit to the House of Representatives' website from my office in Colorado travels directly from Denver to Washington through the Level 3 transit network, while one to the Senate website goes from Denver to the Dallas Internet Exchange over the Comcast network, where it is handed off to the AT&T transit network for delivery to an Akamai CDN server in Chicago.

The standard way for the Internet to select paths uses the Border Gateway Protocol (BGP). This is a system for exchanging routing information that was developed in the mid-90s to facilitate the conversion of the Internet from a research network to a commercial one. BGP allows networks to communicate with each other over a variety of intermediaries according to business arrangements; it performs "policy-based routing".

BGP is sub-optimal because it selects paths mainly on distance rather than dynamic criteria such as occupancy, packet loss, latency, jitter, throughput, historical reliability and maximum peer capacity. Because some applications – especially real-time applications such as voice and video conferencing – require consistent low latency, services have developed means of overcoming BGP's shortcomings by continually assessing link quality and selecting optimal

paths dynamically. While BGP provides static routing, the optimal approach provides dynamic routing.

The pioneer of dynamic routing over the Internet was ITXC, a firm founded by Tom and Mary Evslin in 1997.[27] ITXC delivered ordinary telephone calls on a wholesale basis using the Internet as an intermediary between traditional telephone networks. Its method consisted of evaluating current link quality and selecting paths that could best meet the quality requirements of telephony over the statistical Internet.[28]

The ITXC dynamic routing method is now the basis of an entire software defined wide-area network (SD-WAN) industry segment known by the names "WAN Edge Infrastructure", "Software Defined Networking (SDN)", "Virtualized customer premises equipment (vCPE)", and Managed Network Services (MNS).

A recent Gartner Group report identifies 16 of more than 40 firms offering these products.[29] Their general value proposition lies in allowing customers to save money by using the Internet as a substitute for Business Data Services or private lines. Gartner reports that SD-WANs may be deployed by organizations on a "DIY" basis as ITXC did; but they may also be offered by network service providers, system integrators, or specialized Managed Service Providers.

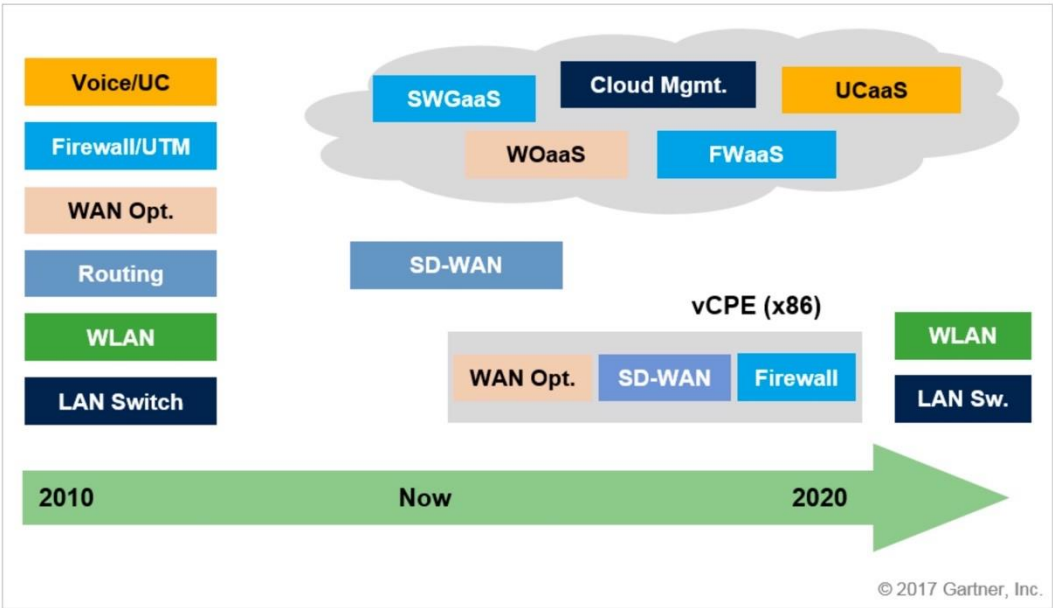
---

[27] "ITXC Corporation," *Wikipedia*, March 7, 2015, https://en.wikipedia.org/w/index.php?title=ITXC_Corporation&oldid=650311017.

[28] For an explanation of the ITXC method, see this video podcast: Richard Bennett, "Internet Pioneers Discuss Network Architecture and Regulation," *High Tech Forum* (blog), August 16, 2017, http://hightechforum.org/internet-pioneers-discuss-architecture-regulation/.

[29] Andrew Lerner and Neil Rickard, "Market Guide for WAN Edge Infrastructure," Gartner Reprint (Gartner Group, March 23, 2017), https://www.gartner.com/doc/reprints?id=1-3X6W6KF&ct=170404&st=sb.

In many cases, SD-WANs are hybrids of ordinary commercial Internet services with managed services and even private lines. All of these networking facilities use Internet standard protocols, TCP and IP, and equipment either similar to or identical with standard Internet routers.



FWaaS: firewall as a service; LAN Sw.: LAN switch; SWGaaS: secure web gateway as a service; UC: unified communications; UCaaS: unified communications as a service; UTM: unified threat management; WAN Opt.: WAN optimization; WOaaS: WAN optimization as a service

Source: Gartner (March 2017)

Figure 3: The Transformation of the WAN Edge

If a firm has Cisco or Citrix routers on premise with some ports connected to an ordinary ISP and others connected to managed services and still others connected to private wires, where does the Internet end and the private network begin? The demarcation point, if there is one, is somewhere inside the router; but the router has a single configuration that regards all three ports simply as routes of particular quality to general destinations.

First Responder Network Authority (FirstNet)

On October 18-19, 2017, I attended an R & D summit on Highly Mobile Deployable Networks sponsored by the National Institute of Standards and Technology's Public Safety Communications Research Program at the Commerce Department's Boulder (Colorado) Lab. The summit covered ongoing research projects designed to increase the capacity of FirstNet to meet the needs of first responders in the aftermath of disasters.

Of particular interest was NIST's Public Safety Communication Research (PSCR) division deployable network testbed, a means of testing new products and applications in the deployables space intended to aid first responders.[30] The deployables discussed are LTE-based temporary facilities that connect to permanent LTE facilities provided by FirstNet and other LTE carriers. FirstNet is an LTE network with extensive Quality, Pre-emption and Prioritization (QPP) mechanisms to help ensure greater reliability for first responder communications across a wide range of scenarios ranging from full backhaul to intermittent backhaul to no backhaul at all; if paid prioritization were not permitted, FirstNet would not be lawful.

These are challenging problems. We identified the following gaps in current products and research:[31]

- Lack of tools/analytics and standards to measure, model, and predict network coverage, capabilities, load, and reliability in real time to inform decision making for self-organized networks

[30] "Deployables" are networks composed of drones, balloons, vehicle-mounted towers, and cell towers on wheels that can be used to stand up a temporary communications infrastructure until a more permanent one can be recreated.

[31] Emily Nunez, "Deployable Networks R&D Summit Highlights," Text, NIST, November 1, 2017, https://www.nist.gov/news-events/news/2017/11/deployable-networks-rd-summit-highlights.

- Inability for deployables from different vendors / agencies to recognize and synchronize with each other (discovery problem)
- Need to determine when to rely on deployable resources vs. core resources
- Need to determine an architecture or process for data storage and processing at the edge to minimize backhaul reliance and to balance network load
- "ICAM on the fly" -- How to register with other non-federated Identity, Credential, and Access Management (ICAM) services (i.e., mutual aid)
- Common Services/ Standardization for Applications
- Need to optimize size, weight, power of deployable hardware for different tasks, agencies, and environments

In the limited backhaul scenarios we examined, we discovered a need to connect deployables to standard LTE networks until a genuine FirstNet backhaul could be restored. Because FirstNet relies heavily on QPP, it stands to reason that the presence of such features on standard LTE networks would also be useful.

Both previous and current research for FirstNet include a number of topics related to prioritization and Quality of Service that would be adversely affected by an over-broad ban on optimization:[32]

- Establishment and Modification of LTE Bearers with Specific QPP[33] Requirements
- Prioritization of Traffic across Backhaul with Limited Bandwidth or Congestion

[32] "Annual Report Outlines Year of Progress Leading to Launch of FirstNet Network | First Responder Network Authority," accessed April 13, 2018, https://firstnet.gov/newsroom/blog/annual-report-outlines-year-progress-leading-launch-firstnet-network.

[33] Preston Kelley, "End-to-End Quality Priority and Pre-Emption," NIST, August 12, 2016, https://www.nist.gov/programs-projects/end-end-quality-priority-and-pre-emption.

- Prioritization of Encrypted Traffic (e.g., mobile Virtual Private Network [mVPN])
- Prioritization of Traffic processed through In-Vehicle Routers
- Technologies and Methods for the Interface between Local Control and the Network QPP Systems
- Assessment of EPS Bearer[34] Capabilities to Prioritize Encrypted Traffic

FirstNet builds on the "bearer" capabilities inherent in LTE. Bearers are a concept that comes to LTE from the Internet through the IETF Integrated Services (IntServ) standards. These standards were developed in the 1990s to ease the convergence of voice and data on a single network, the Internet.[35] While IntServ was relatively dormant for many years, it became a vital part of the Internet with the advent of LTE. It would be shame not to recognize its importance.

Other Applications for Optimized Networking

As we complete the transition from the monopoly circuit-switched network of the past to the privately owned, decentralized, dynamic, packet switched networks of and future, it's important to bear in mind the fact that the Internet is becoming ubiquitous.

While the net neutrality controversy has encouraged policy makers to divide networking between an "Internet segment" frozen to the status quo of the 1990s and a "not-the-Internet segment" free to grow and evolve, this is an unwise separation. Our grandchildren will not recognize any form of networking as "not the Internet". Even if we stipulate, *arguendo*, that the traditional Internet was a "best-efforts" system that treated all packets the same, it does not

[34] "LTE in WIRELESS: Bearers in LTE," accessed April 13, 2018, http://lteinwireless.blogspot.com/2012/12/bearers-in-lte.html.

[35] Bennett, "Arrested Development."

follow that the future Internet should follow such a model. The traditional Internet was small, highly specialized research network, while the Internet of the future must be all things to all people because it will be the only game in town.

For example, a myriad of applications than once required private lines – such as high-definition video conferencing, wholesale voice transport, high volume transaction processing, video entertainment, and industrial process control applications – can now run over Internet facilities with proper support from network service providers.

Consequently, we should not divide network applications into those suitable for the standard Internet and those that require private networks for specialized treatment. Rather, we should ensure that the Internet is sufficiently robust to handle the needs of *all applications, all the time*.

Why Charge for Optimization?

Advocates of strict net neutrality with a far-reaching ban on what they call "paid prioritization" have to walk a very fine line to endorse "reasonable network management". One example of this reasoning is the blog post on today's hearing by Phillip Berenbroick, Senior Policy Counsel at Public Knowledge.[36] Berenbroick denounces prioritization as "inefficient and unnecessary from a traffic management standpoint," while also claiming "it makes sense for [some] services to work in real time, while email does not need to appear in your inbox instantaneously."

[36] Phillip Berenbroick, "House Commerce Takes on Paid Prioritization, an Essential Tenet to the Open Internet," Public Knowledge, accessed April 13, 2018, https://www.publicknowledge.org/news-blog/blogs/house-commerce-takes-on-paid-prioritization-an-essential-tenet-to-the-open-internet.

This is a curious juxtaposition.

The only way real time applications can leap-frog email is for network service providers, end user operating systems, or other network devices[37] to prioritize. Hence, prioritization is good whether provided by users, CDNs, or ISPs, and even when provided by ISPs in the form of for-fee CDN services. The issue seems to be the payment of fees rather than the act of prioritizing. So we have to examine the role that fees pay in network management.

Reasonable network management of the type permitted by the FCC's 2010 and 2015 Open Internet Orders requires two distinct operations: first, the ISP has to recognize the types of applications generating the packet streams it sees. This can sometimes be very simple: Skype packet streams consist of packets much shorter than most Internet packets, and they're very evenly spaced. When an ISP sees a series of three short packets at a constant interval, it's very likely that they represent some sort of Voice over IP application that should be prioritized. This is the case even if the packets are encrypted.[38]

Similarly, when an ISP sees three clumps of full size packets, each consisting of a few hundred packets separated by a common interval, there's a high probability that they represent a video stream, but there are caveats. At the beginning of a movie streaming transaction, senders are testing the subscriber's network capacity while also examining their own network for load conditions and the presence of the desired title at the desired resolution. This creates quite a bit of network chaos, so application identification may take as much as a minute.[39]

---

[37] Such as a home router gateway device, an Ethernet switch, or a Wi-Fi Access Point (AP).

[38] "Very likely" does not mean certain; they could be sensor readings as well.

[39] Because most significant network events take place in thousandths or millionths of a second, a minute is a very large chunk of network time.

Video streams are often embedded in web pages as ads these days, so it can also be tricky to distinguish a genuine movie or TV show from some advertising fluff on a web page. Online games also need special treatment, but they often begin by downloading a great deal of imagery that can appear to be video streaming at the network level.

So identifying applications and guessing their requirements can be difficult unless applications announce themselves and specify their requirements. Internet standards such as IntServ and DiffServ specify mechanisms for this sort of identification, but they're not always operational across internetwork boundaries. Wi-Fi also employs such mechanisms according to the IEEE 802.11e standard.[40]

In the case of Wi-Fi, applications announce themselves to the network access point by sending an "Admission Control" message informing it of their upper and lower bounds on latency and data volume. If the network has sufficient resources, it will reserve the required amount and allow the application to proceed. As the application runs, it is allowed to obtain high priority network access. When complete, it notifies the network so the resources can be recycled.

This negotiation requires a bit of policy – the network can only give high priority to a limited number of applications, and it must trust the application to use resources responsibly. So this procedure is generally carried out on private networks. When it is, experience shows

[40] IEEE Computer Society et al., *IEEE Std 802.11e™-2005: IEEE Standard for Information Technology Telecommunications and Information Exchange between Systems--Local and Metropolitan Area Networks--Specific Requirements. Part 11, Amendment 8, Part 11, Amendment 8,* (New York, NY: Institute of Electrical and Electronics Engineers, 2003), http://ieeexplore.ieee.org/servlet/opac?punumber=10328.

that networks are capable of successfully carrying four times as many phone calls as they could carry when every packet stream has the same priority.[41]

Payment for priority treatment (whether by queue re-ordering, resource reservation, or traffic shaping) serves the same purpose on a commercial ISP network that our Admission Control message does on the enterprise network. It is impossible for a network to give all packets high priority on any network, just as all the children in Lake Wobegon cannot really be above average. Regardless of its capacity, the network has a limited supply of low priority (or low loss, or low jitter) transmissions per second. These can either be parceled out at random – as net neutrality supporters appear to demand – or according to some sort of plan, the accuracy of which is dubious unless requestors have incentives to only seek low latency when they need it. As net neutrality advocates correctly point out, if all packet streams are prioritized, none are.

The most straightforward way to align user incentives with network capacity is to charge a consideration – perhaps a small fee – for the privilege of jumping the queue over packets of unknown character. The consideration might involve a barter, a quota, or some other non-monetary form. One idea that was suggested nearly a decade ago at a policy panel in Washington called for a quota of low-delay packets per month, similar to the minutes of use once attached to cellular telephone services.[42]

---

[41] This narrative is based on personal experience with creating the IEEE 802.11e standard while working at Sharp Labs and then implementing enterprise Wi-Fi products that carried it out at Trapeze Networks.

[42] Richard Bennett and Brett Glass, "Forum on Network Management" (ITIF Forum on Network Management, Washington, D.C, March 12, 2008), https://itif.org/events/2008/03/12/forum-network-management.

The value to the user and to the network provider is that today's guesswork would be replaced with a more certain system in which users and developers would register applications requiring treatment with ISPs so the latter could handle them with great precision.

Net neutrality advocates are worried that payment for specialized treatment could lead to abuse. This fear can be allayed through disclosure requirements addressing the amount of special treatment sold. If ISPs are selling more special treatment than they can reasonably provide, or selling much, much less than their peers, the regulator will understandably have issues. In any case, the potential problems of financial abuse are better resolved by empirical analysis and ex-post enforcement than by pre-emptive emotional reactions.

What about Free Speech?

Arguments for banning Internet optimization generally rest on a faulty understanding of Internet performance.

Many advocates have insisted that allowing ISPs to sell optimized delivery services would mute the voices of non-profit advocates and impose barriers to startups. For example, Malkia Cyril warned that "fast lanes" and "slow lanes" are limits on speech:

> *Internet Service Providers want to break the internet into fast and slow lanes that sell public voice to the highest bidder. If we lose that vote, the most democratic communications platform the world has ever seen could become*

more like cable TV, a fairly scary place that reproduces the economic gaps and racial hierarchies of the offline world.[43]

Cyril's fears are based on a misinterpretation of a reality of Internet commerce.[44] While it's true that users abandon web e-commerce sites that are slow to load, the key to a fast site does not rest on buying speed from ISPs. My research on the relative speeds of broadband networks and websites indicates that the websites generally use a small fraction of the speed ISPs make available to them today.[45]

Broadband Speed vs. Web Speed

While broadband network speed has increased by an average of 35% per year for ten years, web page load times have remained fairly stagnant.[46] In 2016, for example, web page load times were worse than they were in 2015. The FCC's *Measuring Broadband America* reports have consistently shown that ultra-fast broadband speeds do not make web pages load faster than they do over 12 – 15 Mbps networks.[47]

---

[43] Malkia Cyril, "Only Net Neutrality Can Protect the Internet from Becoming like TV: White, Middle-Class and Exclusive," the Guardian, February 26, 2015, http://www.theguardian.com/commentisfree/2015/feb/26/only-net-neutrality-can-protect-the-internet-from-becoming-like-tv-white-middle-class-and-exclusive.

[44] Rick Whittington, "Is A Slow Website Costing You Sales?," accessed April 10, 2018, https://www.rickwhittington.com/blog/is-a-slow-website-costing-you-sales/.

[45] Richard Bennett, "You Get What You Measure: Internet Performance Measurement as a Policy Tool" (American Enterprise Institute, November 2017), http://www.aei.org/publication/you-get-what-you-measure-internet-performance-measurement-as-a-policy-tool/.

[46] Web page performance measurement is not as precise as we would like it to be, so it's possible that the first screens of web pages load much faster than they used to.

[47] Early MBA reports reported this fact correctly, but those issued after the FCC redefined "broadband" to 25 Mbps have claimed a threshold value of 25 Mbps even though the underlying data have not changed; see FCC's Office of Engineering and Technology and Consumer and Governmental Affairs Bureau, "Measuring Broadband America," Measuring Broadband America (Washington, DC: Federal Communications Commission, 2016 2011), https://www.fcc.gov/general/measuring-broadband-america; FCC's Office of Engineering and Technology and Consumer and Governmental Affairs Bureau, "2012 Measuring Broadband America: July Report" (Washington, DC: Federal Communications Commission, July 2012), http://transition.fcc.gov/cgb/measuringbroadbandreport/2012/Measuring-Broadband-America.pdf; FCC's Office

Putting a website in a "slow lane" would require the ISP to reduce the speed of the service it offers to the user from its native speed – an average of 85 Mbps in the US – to a speed less than 15 Mbps.[48] This would raise serious false advertising issues.

So there is clearly more to the speed at which web pages load than the service provided by ISPs. Other factors include the size and complexity of web pages themselves, because pages must be processed by browsers before they can be seen. Mozilla Firefox has proved that simply re-arranging web page content can cut load time in half.[49] The reason for this is that ads are slower to load than native content.

Content Delivery Networks (CDNs) such as Akamai, Amazon AWS, and Fastly also make web pages load faster than they would from a single web server. In fact, speeding up websites is the fundamental value proposition for CDNs. It's worth noting that very large companies operate their own CDNs. Each of America's five largest companies by market cap – Apple, Google, Facebook, Microsoft, and Amazon – operates its own CDN as well as its own large scale IP network ("pipes").

---

of Engineering and Technology and Consumer and Governmental Affairs Bureau, "2013 Measuring Broadband America: February Report," Measuring Broadband America (Washington, DC: Federal Communications Commission, February 2013), http://transition.fcc.gov/cgb/measuringbroadbandreport/2013/Measuring-Broadband-America-feb-2013.pdf; FCC Office of Engineering and Technology and Consumer and Governmental Affairs Bureau, "Measuring Broadband America - 2014," Measuring Broadband America (Washington, DC: Federal Communications Commission, 2014), http://www.fcc.gov/reports/measuring-broadband-america-2014; FCC's Office of Engineering and Technology, "Measuring Broadband America Fixed Report - 2015," Federal Communications Commission, December 22, 2015, https://www.fcc.gov/reports-research/reports/measuring-broadband-america/measuring-broadband-america-2015; FCC's Office of Engineering and Technology, "Measuring Fixed Broadband Report - 2016," Federal Communications Commission, September 29, 2016, https://www.fcc.gov/reports-research/reports/measuring-broadband-america/measuring-fixed-broadband-report-2016.

[48] "United States's Mobile and Broadband Internet Speeds," Speedtest Global Index, accessed April 10, 2018, http://www.speedtest.net/global-index/united-states#fixed.

[49] Richard Bennett, "The Firefox Fast Lane," *High Tech Forum* (blog), December 21, 2017, http://hightechforum.org/the-firefox-fast-lane/.

Advertising also has a significant effect on the speed at which webpages load. Many ad networks – such as Google – rely on real-time auctions to sell ads, and these auctions sometimes take more time than they do at others. Web pages, especially commercial ones, tend to include tracking by a variety of parties (most commonly Google and Facebook) that requires the execution of code while the page is loading. Executing this code requires CPU resources in the user's computer; this is why pages load faster when we upgrade our laptops and mobile devices.[50]

Consequently, "fast lane" services worthy of the name would not consist of high-speed data pipes within ISP networks. Rather, they would be comprised of fast CDN web servers located close to ISP networks – or even inside them, as many are – and pages free of advertising, tracking, and auto-play videos.

Hence, advocates who insist that optimization must be banned in order to protect free speech have failed to analyze the issue correctly. And despite the absence of benefit to the web from optimization, the practice is extremely useful – and even necessary – to the host of non-web applications that exist on the Internet today and will exist tomorrow.

What about Innovation?

The great myth about innovation stresses the lone inventor cobbling together a better mousetrap in a dorm room. While this is sometimes the case, the majority of significant innovations and inventions are made by large firms and well-funded startups. Bell Labs, for

[50] I recommend the use of the "Ghostery" plugin for seeing the impact of trackers on web page tracking and performance; see https://www.ghostery.com/

example, created the integrated circuit, fiber-optic communications, the Unix operating system, and the C programming language; the microprocessor was devised by Texas Instruments and Intel Corporation; the Internet was designed by researchers funded by the US Department of Defense; the Google search algorithm was developed by recipients of National Science Foundation (NSF) grants working in a well-equipped Stanford University lab; and the SPDY and QUIC protocols that enable web pages to load faster were devised by Google within the last six years. While 2016 was a down year for investment in the entrepreneurial ecosystem, venture capital flow for the year is estimated at $69.1 billion in the US alone.[51]

Consequently, the first question for potential innovations in the Internet ecosystem is whether the necessary capabilities are available to entrepreneurs at any price. If an innovation cannot be done on the networks of the day, it will obviously not be done.

The Internet has proved to have something of an affinity for consolidation. It has spawned an effective duopoly of Google and Facebook in the advertising market, for example. Similarly, single firms dominate the markets for search, Internet retail, social networking, mobile apps marketplaces, desktop operating systems, and video streaming.[52] The dominant positions maintained by America's five most valuable companies – Apple, Facebook, Alphabet (Google), Microsoft, and Amazon – are protected by the private content delivery networks owned by each of these firms.

---

[51] "After Peaking in 2015, Venture Investment Activity Normalizes in 2016, According to PitchBook," NVCA, accessed April 13, 2018, https://nvca.org/pressreleases/peaking-2015-venture-investment-activity-normalizes-2016-according-pitchbook-nvca-venture-monitor/.

[52] Richard Bennett, "Internet Monopoly Platform Crisis," *High Tech Forum* (blog), October 26, 2017, http://hightechforum.org/internet-monopoly-platform-crisis/.

Facebook, for example has two billion users. A would-be competitor to this firm would need to reach similar scale because the value proposition of a social network is its ability to connect us to others. Hosting content and facilitating communication between two billion users requires enormous investment, outstripping the capacity of even our rich venture capital establishment. But would-be competitors may boot-strap their way to success by using commercial CDNs such as Akamai and Cloudflare, keeping costs in rough alignment with business growth and then gradually replacing CDN contracts with private investment, as Netflix did.

But the CDN model only works for applications that don't require real-time communication. Real-time applications such as voice and video conferencing rely on entirely separate networks – such as Cisco's Webex and the VoIP exchanges – that bypass the constraints of the public Internet. Real-time networks practice a form of admission control and strict internal management to guarantee low latency delivery across large portions of the planet.

Building networks of the size and scale of Webex is not practical for small entrepreneurs. And prices are high because there is limited competition for such services. A much more efficient technical and economic model would allow real-time startups to contract directly with ISPs for low-delay transport to ISP customers, in much the same way that firms purchase Business Data Services (BDS). But it's essential to allow entrepreneurs to reach potential customers who are not users of BDS. The public Internet is available to ordinary users at relatively low prices because the grade of service required by popular applications such as Netflix and Facebook is relatively low.

Upgrading to higher quality services increases costs to ISPs. Ideally, these would be borne by users rather than startups. But it's extremely difficult to persuade users to upgrade to a higher service grade to run an application with which they have no experience. Consequently, it's sensible to allow third parties to purchase upgrades on behalf of potential customers so that they an appreciate the value of new services for gaming, video communication, augmented reality, virtual reality, and even holographic conferencing.

These applications are a poor fit to the service model of today's Internet, but they probably do represent the future of electronic communications. That future depends on regulators not strangling the baby in the cradle, however.

Conclusion and Recommendations

The "traditional" Internet as understood by policymakers (and many senior technologists no longer involved in day to day network engineering and operations) – an end-to-end system organized around a common backbone – is a thing of the past. Today's Internet is simply the universal network for all forms of electronic communication. Application service networks are often directly connected to last mile networks, and the role of transit and backbones has diminished, making old notions of "tier 1" and "tier 2" networks moot. To meet the new demands, the Internet has developed a new architecture consisting of content delivery networks, large private networks, parallel networks, and multiply-connected networks. Traditional protocols such as HTTP 1.1 and TCP have given way to SPDY/2 and QUIC, and traditional algorithms such as RED have given way to CoDel and PIE. The Internet is now entrusted with applications ranging from cat videos to safety-of-life first responder networks.

Regulators must display wisdom by recognizing the new mandates and the new technologies. Trying to stuff the Internet back into the traditional mold is simply an exercise in futility. Internet Service Providers, innovators, regulators, and dominant (for the moment) service providers have always worked together according to the so-called "Internet model" of multi-stakeholder governance. This model has generally worked, and we should continue to rely on it.

Abandoning the multi-stakeholder model in favor of a top-down, micromanagement model assumes that regulators will be possessed not just of the wisdom born of experience but that they will indeed be endowed with god-like power to see far into the future. It's best for policy makers to limit the scope of regulatory power to the range of affairs that can be performed by mere human beings.

It's perfectly fine for regulators to permit first and sanction only in the presence of meaningful evidence of harm. The alternative stifles new technologies and applications before we've had the opportunity to test them. That is not the path to a better tomorrow.