ITIF
INFORMATION TECHNOLOGY
& INNOVATION FOUNDATION

Paid Prioritization: Why We Should Stop Worrying and Enjoy the “Fast Lane”

BY DOUG BRAKE | JULY 2018

Prioritization is the only economical way to radically improve the performance of broadband given the wide diversity of different applications networks support.

Data traffic prioritization ranks up there with GMOs and nuclear power as one of the most unfairly maligned technologies. Caricaturing commonplace network management techniques as “fast lanes,” net neutrality activists warn that introducing the option of paying for specific performance levels of Internet traffic will destroy the characteristic “openness” of the web. This is false. Prioritization and other mechanisms that differentiate data traffic are the only economical ways to significantly improve the performance of broadband, given the wide diversity of different applications that broadband networks must support, and would encourage further innovation throughout the Internet.

This is not to say prioritization is always benign, either. The nightmare scenarios dreamed up by activists of prioritized services squeezing out noncommercial speech, for example, are, if very unlikely, at least imaginable. But with simple governance and oversight, there is no reason differential treatment of Internet traffic cannot safely improve the overall performance of the Internet to everyone’s benefit.

When it comes to discussions of potential net neutrality legislation, there is fairly widespread agreement on what the substance of rules should look like.[1] There is rough consensus around blocking and throttling—namely, that baseline rules should flatly ban Internet Service Providers (ISPs) from blocking legal Internet traffic or slowing or throttling data to extract payment. The main political controversy is around the scope of authority Congress should give the Federal Communications Commission (FCC).

But the rough consensus around substantive rules admittedly breaks down at paid prioritization. For example, consider the opening statements of Rep. Marsha Blackburn (R-TN) and Rep. Frank Pallone (D-NJ) at a recent Subcommittee on Communications and Technology hearing on prioritization.[2] Blackburn noted, “Despite what some of my colleagues sometimes seem to think, 'prioritization' is not a dirty word. The Internet is based on it.… Prioritization of data on the network is not unique, or uniquely harmful."[3] Pallone, on the other hand, said, "Despite these latest attempts to muddy the water and create confusion, banning paid prioritization is not a new issue. The FCC solved this problem when it passed net neutrality in 2015. At that time, the FCC correctly banned these fast lanes."[4] The breakdown in consensus is not uniformly partisan, but there is clearly a divide when it comes to how paid prioritization should be treated.

Prioritization is actually a relatively narrow, specific type of traffic differentiation. Prioritization in a network engineering sense refers to specific traffic scheduling techniques as data traffic is routed through the Internet. But within the broader policy discussion, the term "prioritization" often acts as a stand-in for any type of differential treatment of Internet traffic.

It should not be controversial that prioritization or other forms of traffic differentiation can optimize Internet Protocol-based (IP-based) networks for different purposes. Indeed, from the Internet's very beginning, architects have designed protocols for traffic differentiation precisely because the Internet is a best-effort network, in which not all applications work optimally without differentiation. Unlike circuit-switched residential telephony, where there was one wire and one application, the Internet supports a dizzying array of applications, many of which work just fine over a traditional "neutral" network—while others would be greatly improved by prioritization.

Moreover, prioritization is not a zero-sum proposition—that is, it is possible to see a benefit without an offsetting loss. In some cases, this is because slight delays in packet delivery are far less noticeable to consumers for some applications than others. For example, nobody would notice email packets being delayed by 50 milliseconds, yet such a delay would make a big difference in video conferencing applications. Prioritization works by optimizing the timing or scheduling of packets for different applications, allowing tradeoffs to not detrimentally affect the quality perceived by a user. It is possible to send email, stream videos, and build new tools for communication, continuing the characteristic openness of the Internet that has generated such tremendous innovation in recent decades, all while allowing for traffic differentiation that enables real-time applications with very strict performance requirements, especially related to the delay or variability in delay of the data flow.

Most all agree there should be room for "specialized services" that run over the same infrastructure as the Internet, such as for telehealth applications. The question is how strict or permissive restrictions should be on traffic differentiation. A relatively permissive regulatory regime would allow for all companies—large and small—to contract for prioritization services, and not have to navigate a bureaucratic process at the FCC or arrange for specialized billing with broadband providers.

With simple governance rules and ongoing oversight, a "non-neutral" network can unlock new, real-time services without harming general "best-effort" traffic and preventing for anticompetitive consequences.

TRAFFIC DIFFERENTIATION CAN DRAMATICALLY IMPROVE PERFORMANCE OF IP-BASED NETWORKS

Various forms of traffic differentiation are necessary to move the utility of IP-based networks beyond traditional web-based services to enable support for low-latency systems that cannot be reliably provided by the Transmission Control Protocol (TCP). TCP, one of the main underlying protocols of the Internet, is responsible for resolving network congestion and ensuring reliable, error-free communication of data. It emphasizes the reliability of data transferred via the network over the speed of the transmission. Other protocols attempt to minimize latency, but come at the cost of reliability. Broadband networks are the future of all communications, and should be allowed to be intelligent enough to compensate for architectural differences to support higher order systems. Prioritization is the only obvious, cost-effective way to achieve this goal.

The Internet Engineering Task Force explains that with only best-effort traffic, there are fundamental limitations to the performance of real-time applications.

Best-Effort Traffic Is a Powerful Engine of Innovation, but Is Not Optimal for All Uses

A fundamental question rarely addressed by those advocating for "neutral" networks is, "'Neutral' with respect to what?" Many conceptualize neutrality simply as a lack of ISP intervention—as if there were neutral networks discovered in the wild. Of course, decisions regarding basic network architecture, routing, interconnection, and design have long affected the performance of networks, and have to be made if the Internet is to work, but ultimately result in different functionalities being elevated or suppressed.

Perhaps the clearest notion of a neutral net is one characterized by best-effort data traffic, which can be understood as data traffic that does not rely on differential treatment of traffic flows.[5] This is usually what net neutrality advocates have in mind when they call for a neutral network—one free from traffic differentiation amongst various data flows. In other words, if email packets are flowing along the network with packets from a video conference call, the bits at the front of the line simply go first—first in, first out.

The term "best efforts" can be somewhat misleading, sometimes giving the mistaken impression that the best possible attempt is made to deliver traffic. Rather, it is better conceptualized as each step along a particular data flow's path, as it is routed across the Internet, simply sends packets forward toward their ultimate destination, and then hopes for the best.

This is unlike the case of a circuit-switched network like telephony, wherein an entire dedicated circuit is formed for each call and optimized only for telephone voice traffic—the only type of traffic on the network. The advantage of this circuit-switched network is lack of contention: Each call gets what it needs from the network. The disadvantage is its extreme inefficiency. Circuit-switched networks are like driving across the country on the Interstate system with no other cars allowed anywhere on a particular route. The Internet, a packet-switched network, on the other hand, is like millions of people transporting: some of them on bikes, some in trucks, some in ambulances, some on trains, some on a deadline, some out for a leisurely drive. This might all work well at 2 a.m., but on a Memorial Day weekend, problems would arise.

Early use cases of the Internet, such as applications like file-sharing between academics, posting of text, and email, all work just fine over undifferentiated best-effort routing. Indeed, best-effort traffic has a lot going for it. It makes minimal technical demands on network infrastructure, and it keeps the economic relations throughout the Internet system relatively simple.[6] With some exceptions, it is how the Internet has developed to date.

But neutral, best-effort traffic has its limits. As explained in a memo by the Internet Engineering Task Force (IETF), the organization responsible for developing Internet standards, "Simple best-effort traffic is not optimal."[7] The memo adds:

> With only simple best-effort traffic, there would be fundamental limitations to the performance that real-time applications could deliver to users. In addition to the obvious needs for high bandwidth, low delay or jitter, or low packet drop rates, some applications would like [other specialized treatment]. There are severe limitations to how effectively these requirements can be accommodated by simple best-effort service in a congested environment.[8]

This means overly strong net neutrality regulations that ban prioritization risk limiting the growth of real-time applications in order to lock in an architecture that favors plain-text webpages.

The most value from prioritization will be gained by exactly these sorts of dynamic, latency-sensitive applications the Internet's current best-effort architecture discriminates against. One good example of the type of application that would benefit from prioritization is high-definition video conferencing. Video conferencing requires feedback from two users in real time—notably different from video that is buffered and not in real time, like popular streaming services. In other words, video streaming does not need prioritization on the lion's share of broadband networks because the packets load fast enough to keep the buffer full. But other applications that are more two-way in nature, such as remote controls, interactive gaming, two-way video, and future data-intensive, real-time cloud services, would potentially benefit from—if not require—traffic differentiation techniques.

The Internet Is Not Inherently Neutral

There is a misconception the Internet has always been neutral. In fact, early design decisions explicitly allowed for differentiation of services.

Some advocating for neutrality elevate what is known as the "end-to-end" heuristic, in an attempt to secure best-effort traffic as somehow inherent to early Internet designs.[9] This end-to-end principle was originally formulated as a design decision for how to "guide placement of functions," such as error correction, in distributed computer systems.[10] The idea was one of efficiency. Instead of building redundant error correction protocols throughout the network, it would make more sense to abstract that function, move it further up the stack, and have it only performed at the endpoints. Generally, once one end received a full data transmission, it would only then check for errors, rather than implementing redundant error correction algorithms at each individual node of the network.

The debate over the design principles of the Internet, and what they mean for net neutrality policy, is old, but largely settled. Barbara van Schewick, associate professor of law at Stanford Law School and director of Stanford Law School's Center for Internet and Society, has long argued for a relatively strict version of network neutrality.[11] But she recognizes that neutral or best-effort Internet traffic is not required by the end-to-end principle, and in her book calls out a series of arguments that overextend various design principles in the name of net neutrality.[12]

We have been able to muddle through with this best-effort system, resigned to freezing frames or choppy voices, but many of the exciting innovations around the corner will increasingly require reliable low-latency connections.

An examination of the history of Internet architecture reveals the earliest design decisions built flexibility and adaptability into the network. Neutrality was not a design feature of the technology itself. As Christopher S. Yoo, law professor and founder of the Center for Technology, Innovation and Competition at the University of Pennsylvania has explained:

> Unfamiliarity with the Internet's architecture has allowed some advocates to characterize prioritization of network traffic as an aberration, when in fact it is a central feature designed into the network since its inception.… Lack of knowledge has allowed advocates to recast pragmatic engineering concepts as supposedly inviolable architectural principles, effectively imbuing certain types of political advocacy with a false sense of scientific legitimacy.[13]

Professor Yoo examined a series of historical and contemporary examples of different types of prioritization, dating back to the most fundamental "type of service" (TOS) flag within the Internet Protocol. He convincingly has argued against the Internet as historically or designed to be neutral in building his case that "the radical changes in the technologies comprising the network and the demands … place[d] on it is creating pressure on the network to evolve in response."[14]

Similarly, Richard Bennett, formerly a senior research fellow at ITIF, has stated:

> Network neutrality advocates err when they suppose that the power of the Internet to extend political participation depends on the relatively narrow elements of system design bracketed within the scope of the end-to-end arguments. Network neutrality advocates have apparently reached deep inside the Internet's technical structure less to understand it than to find the authority to regulate it.[15]

The end-to-end arguments have limited import when it comes to issues like throughput or delay, wherein features within the network itself make all the difference.[16] Prioritization to provide a reliable low-latency connection must be implemented within network nodes and cannot be achieved at the application layer. Recognizing this, some neutrality advocates, such as Professor van Schewick, call for "user-directed" quality of service (QoS).[17] We believe user-directed prioritization is not sufficient to see real-time innovation flourish, but the point here is simply that the more-sober net neutrality advocates generally accept that

QoS has real benefits and is not generally at odds with the fundamental design of the Internet.

Packet-Switched Networks Gain in Efficient Use of Bandwidth but Lose in Predictability

Users' demand for broadband capacity is "bursty" in that it rapidly changes as they perform different tasks on the Internet.[18] Moving from a single user's link deeper into the network, different IP-based communications are joined together in a process called multiplexing.[19] But each piece of equipment and link within a network have a limitation on the capacity they can handle. In order to economize performance, operators make sophisticated predictions about how much capacity will be needed at any given point in the network, setting it to a level such that total instantaneous peak demand will only occasionally exceed capacity. That is much more efficient than building the equivalent of a 20-lane freeway to handle the occasional Sunday traffic coming from the football game. But even given these predictions, random combinations of traffic spikes come together to create congestion.[20]

In most circumstances, for most applications, unpredictable congestion is no cause for alarm. In fact, it is a fundamental feature of the congestion control algorithms built into the most basic protocols of the Internet. Whenever TCP detects a packet has been dropped, likely as a result of congestion, the sender automatically retransmits the information. However, these protocols are not a cure-all. In fact, sometimes the relatively large control loop built into these protocols can exacerbate the problem.[21] The result is a considerable increase in delay as a result of a few lost packets. This is of course not a problem for most uses of the Internet, as users do not notice or care if it takes a few extra milliseconds to, say, receive an email.

The key problem is some applications simply cannot tolerate too much delay (latency) or variance in delay (jitter). Applications that provide real-time services—such as Voice over Internet Protocol (VoIP) or teleconferencing—operate under relatively tight performance criteria. We are of course able to make real-time services work to some degree without specialized prioritization—Skype and FaceTime are obvious examples. But we are also all probably familiar with a Skype call gone awry. If you are on a poor connection, or calling someone far away, it is difficult to use real-time applications with much confidence. For example, a high-resolution teleconference with attendees in different countries is unlikely to go well without some form of prioritization. But even if both parties enjoy a robust broadband connection, such real-time services can often experience annoying hiccups due to normal, intermittent congestion.

To date, we have been able to muddle through with this best-effort system, resigned to frames freezing or voices that are sometimes choppy, but many of the exciting innovations around the corner will increasingly require reliable low-latency connections. And while some applications affirmatively need prioritization or some kind of differentiation, other applications can easily tolerate delay or jitter. Bulk file transfers such as software and operating system updates do not care about delay. Whether your operating system is updated now or two minutes from now—or even two hours from—now makes no difference to most users, and hence the vast majority of applications, including the "next Google" born in a garage will be more than satisfied with the best-effort Internet, especially given broadband speeds have been consistently increasing on the order of 25 percent annually.[22] But for the "next Google" attempting to design and market a service that performs suboptimally with a best-effort Internet, not being allowed to have prioritization can be the difference between success and failure.

QoS Mechanisms and New Routing Techniques Can Improve the End-User Experience

A wide variety of innovative "edge" applications will greatly benefit—if not outright require—specialized treatment. Take, for example, virtual reality (VR) and augmented reality (AR). Leading experts in VR and AR, such as John Carmack at OculusVR and Michael Abrash at Valve, have explained that extremely low latency is fundamental to effective VR and AR experiences.[23] A steady total-system latency of no more than 20 milliseconds is generally considered necessary for acceptable experience without nausea.[24] Reliably provided cloud-based, shared VR experiences with significant bandwidth requirements with such low latency will likely require specialized treatment of traffic. Engineers are hard at work to provide networks that can support a variety of reliable low-latency services, such as factory automation, intelligent transportation services, robotics and telepresence, competitive gaming, virtual and augmented reality, and dynamic control over smart-grid technologies.[25] Other, similar examples abound.[26]

Engineers are working to support a variety of low-latency services, such as factory automation, intelligent transportation services, robotics, telepresence, competitive gaming, virtual and augmented reality, and dynamic control over smart-grid technologies.

Prominent academics have raised concerns about potential unintended consequences of overly broad net neutrality regimes that limit the ability of networks to offer differentiated levels of service through what are known as QoS mechanisms. For example, David Clark of MIT and kc claffy of UC San Diego have argued that coordination problems combined with "current regulatory resistance to enhanced services" will force a "natural response" by ISPs to shift capital to private networks enhanced with QoS mechanisms to meet "the nation's need for a stable network infrastructure."[27] Such a shift, were it to happen, would represent a significant waste of valuable societal resources—the equivalent of building all those extra lanes for the Sunday football game.

Johannes Bauer and Günter Knieps, of Michigan State University and the University of Freiburg respectively, made a similar point in their recent paper "Complementary Innovation and Network Neutrality."[28] Discussing the popular "permission-free" innovation at the application layer, they wrote:

> Yet, not all types of Internet-based innovation fit into this framework. The growing relevance of video, cloud computing, and Internet of Things (IoT) applications requires that innovations in traffic service networks meet quality of service (QoS) needs, which cannot be met in the historical, best-effort network. Because such innovations will become more important in the future, the provision of differentiated QoS in traffic networks becomes a precondition for expanding the innovation opportunities for applications and services in higher layers.[29]

The innovation we can imagine from increased bandwidth is limited. Most of the challenges requiring prioritization have nothing to do with bandwidth, but rather with the delay between endpoints (latency) and the variability in the delay (jitter). Indeed, many of the exciting innovations around the corner require very strict performance criteria around scheduling and latency to enable real-time responsiveness.

THE "FAST LANE/SLOW LANE" ANALOGY IS FUNDAMENTALLY INACCURATE

Many of the sound-bites around net neutrality portray a fundamentally inaccurate, but politically appealing, scenario. The functions of Internet routers and switches occur at a timescale that resists easy analogy to everyday circumstances.

Traffic Differentiation Is Not Zero-Sum

Many net neutrality advocates claim that a fast lane by definition means everyone else is in a slow lane, and then paint a socially dystopic picture with the most vulnerable Americans in this proverbial slow lane. This is not the case. With cooperation among applications, networks can intelligently trade off the performance criteria of different applications such that the end-user experience is improved for at least some—and no one is made worse off. In other words, there are no slow lanes—only preexisting "regular lanes" and "better lanes" for some purposes.

In its report, "Differentiated Treatment of Internet Traffic," the Broadband Internet Technical Advisory Group observed, as a technical fact, differentiated treatment of traffic can produce a net improvement in the quality of users' experiences.[30]

> 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.[31]

But even for some net neutrality advocates this is not acceptable. For these egalitarians, every Internet consumer and producer should have exactly the same experience. So even if there are no slow lanes, there certainly should not be fast lanes. Of course, this ignores the reality of virtually every broadband provider in the world selling different speed services to businesses and consumers, with faster speeds costing more. What is more, it is common for enterprise networks—broadband sold to businesses—to prioritize services such as video conferencing, and sometimes even deprioritize buffered services such as employees streaming music—which, again, does not affect the employees' ability to enjoy their Spotify.

There is one other problem with the "fast lane/slow lane" analogy: Traffic differentiation does not "speed up" a users' broadband connections. Many seem to think prioritization would make a very crude distinction, with a fast lane necessary for all businesses or successful speech on the Internet, lest it fall into the dreaded slow lane. Prioritization is for services with low delay requirements, so remaining best-effort bandwidth would be more than sufficient for continued successful use, especially now that the average household broadband speed in the United States is 94 Mbps—far more than enough to watch a high-definition video stream on a television (Netflix encodes 4K at 15 Mbps).[32] The download speed of a network is based on the tier of service a user buys from their broadband provider. Traffic differentiation addresses a set of different challenges related principally to latency and jitter.

More Accurate Analogies Make Clear the Case for Traffic Differentiation

The "fast lane/slow lane" or "tollbooths" analogies, where imagined Internet users would have to cough up money to avoid a traffic jam, are fundamentally inaccurate. Consider the differences in data-flow characteristics in, for example, streaming video. Streaming high-definition video is like a freight train—a constant torrent of data. In 2016, streaming video accounted for 73 percent of all Internet traffic.[33] And Cisco estimates 82 percent of IP traffic will be video by 2021.[34] Streaming video, such as the services offered by Netflix or YouTube, is buffered, meaning it is not real time. Although users may care about the initial packets arriving as quickly as possible so their video starts right away, after that, as long as the buffer stays full, the video will continue to play smoothly.

Networks can intelligently trade off the performance criteria of different applications such that the end-user experience is improved for at least some uses—and no one is made worse off.

As this streaming video traffic—more than 70 percent of all Internet traffic—navigates across the physical infrastructure of the Internet, at each hop from router to router, it spends milliseconds in line waiting to be sent to the next hop. At each of these points, the data flow is quite literally waiting in line to wait in line. Once it gets to its final destination—the streaming device—it sits in the buffer waiting for its turn to flash on the screen. A latency-sensitive application could go first—dramatically improving that application's performance—with no downside to streaming video flow. This is probably the largest opportunity to trade off around different applications' requirements.

Imagine you are in traffic, stopped at a red light. There is another red light at the next intersection. An ambulance cuts through the intersection. No one even has to pull over. Is this such a travesty? Now consider the scale of this intersection compared with computer networks. The cars would be freight trains the size of skyscrapers, and the ambulance would be a tiny bicycle. The wait imposed on the drivers would be a blink of an eye, and that next red light at which they would have to wait anyway would be several minutes long.

PRIORITIZATION SHOULD BE PAID FOR

Some prominent voices in the pro-net neutrality camp understand the technical merits of prioritization and how it can dramatically improve performance. However, they simply do not want it to be paid for by a third party. Rather, they call for "user-directed" prioritization. It is not clear how user-directed prioritization would work for real-time, bidirectional services (where end-to-end control is preferred), but is nevertheless innocuous and should be allowed. Regardless, there are a number of reasons prioritization should be paid for.

First of all, neutrality advocates lament the "tollbooth" of differentiated services—this image gets the issue all wrong. Differentiated service or prioritization would be an optional service that does something different, and would by no means be required for web traffic or data flows without unique needs from the network.

Rules should allow for a variety of business models to develop. For example, customers should be able to select a premium video conferencing service and pay by the minute, or the company providing the conferencing could approach the broadband provider directly to ensure their product is prioritized without the customers' involvement. It is not immediately clear how traffic differentiation payments would flow through the value chain, and rules would have to accommodate a variety of schemes.

Pricing of Limited Resources Leads to Efficient Allocation

The fundamental congestion-control algorithms of the Internet depend on dropping packets when a particular link or router is congested. Congestion is built into the system: Bandwidth is filled until it is congested, at which point traffic flows back off. As Daniel Lyons, professor at Boston College Law School asked, "How can we address that congestion? One can drop packets randomly, which seems to align with net neutrality's ethos…. Alternatively, we can use the price mechanism, which is the way we generally allocate scarce resources (like bandwidth) in a capitalist society."[35]

In other words, guaranteed low-latency connections—wherein packets are not dropped but shuffled to the front of the queue—are of a limited number. We need a way to determine who would derive the greatest value from having their service prioritized, and a price mechanism is the obvious choice. It is conceivable that users would want multiple services prioritized over a given link, in which case the network would need to have some way of sorting the priority, or the degree of urgency of competing data flows. Willingness to pay is the most obvious way to do this.

Another relatively simple reason prioritization should be paid for is there are real costs in enabling the network to provide these services that need to be recouped. Providing these services is not cost-free, and give users real value, so expecting payment under these circumstances should not be controversial.

The Market for Prioritized Services Is Not Likely to Swamp Best-Effort Services

It is fair to say most large edge providers are skeptical of paid prioritization, having built services that work quite well over current best-effort routing, especially with the help of extensive Content Delivery Networks (CDNs). A CDN is a distributed network of servers or data centers that host copies of popular web or streaming content geographically closer to the end users. This reduces the physical distance that, for example, the images on a webpage or a YouTube video have to travel, reducing latency and page-load times, and reducing overall bandwidth costs for a given service. In many ways, especially for static or buffered content, CDNs have similar effects as prioritization would.

But some technologies do not benefit from CDNs. Consider Aira Tech Corporation, represented by Paul Schroeder at a recent hearing on prioritization.[36] This start-up company aims to connect people suffering from visual impairments with a sighted assistant through mobile video in real time. A blind person navigating, say, an airport, could stream video to an assistant, who would then read back gate information or directions. Mobile bandwidth is generally limited in the upstream. Thankfully Aira is working with leading operators to prioritize this traffic so, for example, an assistant can instruct a blind traveler to turn left instead of right in real time.[37] Do we really think such a service should be on an equal footing with an average Snap sent by an impatient traveler? Do we care if that Snap is broadcast a quarter second later than it otherwise would? Does that quarter second really imperil free speech as net neutrality advocates claim?

These use cases, while valuable, are relatively narrow. The bandwidth, they would claim, is likely quite small compared with something like streaming video or general static webpages, which again, would benefit little by prioritization. Best-effort traffic will continue to be the dominant mode of the Internet, and nobody will be forced to buy a so-called fast lane.

Paid Prioritization Would Not Hurt Small Businesses or Start-ups—and Could Help Them

The example of Aira also demonstrates how paid prioritization would not hurt small business or start-ups. Net neutrality advocates claim that fast lanes would only be affordable by deep-pocketed corporations, which would prevent small businesses from being able to compete. This is untrue for several reasons.

We need a way to determine who would derive the greatest value from having their service prioritized, and a price mechanism is the obvious choice.

First, very few applications will need prioritization—certainly not a small business website or email service. There is simply no reason to think prioritization would be worth the cost for applications that do not benefit from it. Again, best-effort would still be available, and works well for most circumstances.

Second, there are other limitations small Internet start-ups face that dwarf the possibility of paid prioritization. Beyond the normal challenges small business face, like tiny advertising budgets, limited economies of scale, etc., most large tech companies deploy extensive CDNs. By caching content or processing closer to the user, traffic does not have to traverse as far a geographic distance—effectively "prioritizing" the service. Companies like Google, Facebook, and Netflix have extensive delivery networks, essentially backing into the last-mile networks. This is not a bad thing—quite the contrary, CDNs should be celebrated as making the operation of the Internet far more efficient. But these extensive internally controlled networks are very expensive, and outside the reach of a small businesses, which instead purchase content hosting as a service from companies like Akamai.

A third reason paid prioritization would not hurt small businesses is it would likely be offered on a volume basis. It would make sense to sell traffic differentiation by how much traffic received the service. If the rare small business that needed to purchase prioritization wanted to start small and then scale up their operations, the initial costs would likely be low.

Fourth, as noted above, lack of ability to purchase traffic differentiation services would harm start-ups and small businesses that want to introduce the kind of next-generation applications that cannot reliably work well without prioritization. Prioritization would help these start-ups gain scale, not inhibit them.

SIMPLE RULES CAN PREVENT ABUSE

Admittedly, embarking on paid prioritization would be an adventure into uncharted territory. Net neutrality advocates and many users do not place a great deal of trust in ISPs. What is more, both companies and users of the Internet deserve the confidence that their application, communications, or speech will continue to traverse the globe unimpeded. With simple rules, paid prioritization that will not harm the openness of the Internet can be allowed.

There are many possible ways to contain the possibility of abuse of traffic differentiation, ranging from relatively light oversight against unfair or deceptive trade practices from the Federal Trade Commission (the regime in place today) to a strict ban on paid prioritization. The right answer lies somewhere in the middle, with an expert, independent agency like the FCC overseeing prioritization arrangements on a case-by-case basis, informed and constrained by antitrust principles. An example of one such framework was proposed by a working group of respected academics and think tanks under the Digital Age Communications Act project at the Progress and Freedom Foundation.[38]

In terms of potential net neutrality legislation, Congress has the opportunity to craft very fine-tuned institutional arrangements and specific oversight processes. It is important this be done in a way that provides lasting certainty and clarity for all actors in the Internet ecosystem, and also preserves the continued evolution of networks to support increasingly important real-time services.

Differentiation Should Not Be Exclusive

Traffic differentiation should not be only available to a select few firms. These services should be offered openly, ideally through a transparent API. No one should feel forced to purchase prioritization—or be prevented from doing so. Note that as a general matter, exclusive dealing is allowed under antitrust laws and can often be pro-competitive. However, we believe the unique political sensitivities around net neutrality, and the potential harm to competition should an actor be denied the ability to compete with prioritized services justifies a ban on exclusive dealing.

Differentiation Should Be Offered on Similar Terms for Similar Users

We recommend regulators go further than simply banning exclusive dealing and import some limited concepts from common carriage—specifically, treating like customers the same way—is justified. Such a rule ensures traffic differentiation is open to anyone who would benefit from it, and is not limited to a select few firms.

With these two simple rules, enforced by the FCC, we can enable more intelligent broadband networks while still preserving the wonder that is the open Internet.

CONCLUSION

The Internet supports an unbelievable array of different applications, many of which work just fine over a traditional neutral network—while conceivably others would only be enabled by prioritization.

Prioritization is not a zero-sum game. That is, it is possible to both see a benefit without an offsetting loss, and maintain the characteristic openness of the Internet that has generated so much innovation in recent decades, while also allowing for traffic differentiation that enables new, real-time applications with very strict performance requirements.

There should be room for "specialized services" that run over the same infrastructure as the Internet. With simple governance rules and ongoing oversight, a non-neutral network can unlock new, real-time services without harming general best-effort traffic.

ENDNOTES

1. Doug Brake, “Why We Need Net Neutrality Legislation and What It Should Look Like,” (Information Technology and Innovation Foundation, May 2018), https://itif.org/publications/2018/05/07/why-we-need-net-neutrality-legislation-and-what-it-should-look.
2. *From Core to Edge: Perspective on Internet Prioritization,* 115th Cong. (2018) https://energycommerce.house.gov/hearings/from-core-to-edge-perspective-on-internet-prioritization/.
3. Ibid. (statement of Marsha Blackburn, Chairman House Subcommittee on Communications and Technology).
4. Ibid. (statement of Frank Pallone, Ranking Member of Energy and Commerce Committee)
5. S. Floyd and M. Allman, “Comments on the Usefulness of Simple Best-Effort Traffic,” IETF Network Working Group RFC 5290 (July 2008), https://tools.ietf.org/html/rfc5290.
6. Ibid.
7. Ibid.
8. Ibid.
9. For the original paper on the end-to-end principle, Saltzer et al. (1984) “End-to-End Arguments in System Design” *ACM Transactions on Computer Systems* 2.4, pp. 277-288, http://web.mit.edu/Saltzer/www/publications/endtoend/endtoend.pdf.
10. Ibid.
11. Barbara van Schewick, *Internet Architecture and Innovation* (Cambridge, Massachusetts: The MIT Press, 2010).
12. Id. at 104-112
13. Christopher S. Yoo, “Network Neutrality and the Need for a Technological Turn in Internet Scholarship,” *Penn Law Public Law and Legal Theory Research Paper Series, Research Paper No. 12-35*.
14. Ibid.
15. Richard Bennett, “Designed For Change: End-to-End Arguments, Internet Innovation, and the Net Neutrality Debate” (Information Technology and Innovation Foundation 2009), https://itif.org/publications/2009/09/25/designed-change-end-end-arguments-internet-innovation-and-net-neutrality.
16. When it comes specifically to the end-to-end argument, the most thorough (and entertainingly digressive) account can be found in the dissertation of Matthias Bärwolff, *End-to-End, Arguments in the Internet: Principles, Practices, and Theory.* Matthias Bärwolff, *End-to-End, Arguments in the Internet: Principles, Practices, and Theory*, PhD thesis Department of Electrical Engineering and Computer Science at Technische Universität Berlin (2010).
17. Barbara van Schewick, “Network Neutrality and Quality of Service: What a Non-Discrimination Rule Should Look Like,” Center for Internet and Society (June 11, 2012).
18. For example, loading a web page will open several data streams at once, fetching several images, advertisements, and text from a web page’s server. The load on the consumer’s broadband pipe peaks as those resources are loaded, then returns to baseline as the consumer scrolls through the page.

    Streaming video adds considerably to this load, putting a constant stream of data onto the channel. As consumer’s broadband subscriptions have been increasing in speed, the demands of the average websites have grown, with more and more streaming video advertisements and resource-intensive designs, resulting in large, random spikes of capacity demand over short time scales.
19. This technique allows for a single communication channel to be shared by a number of different data streams at the same time. Multiplexing results in tremendous gains to efficiency, as the down time between any one user’s bursts of traffic are smoothed out and available capacity is more fully taken advantage of.

20. Similarly, denial-of-service attacks, network failures or unexpected routing changes, or cell overload in the mobile context can all cause congestion. For the most part, users' instantaneous capacity demands are not correlated—we are talking much shorter time scales than the general ebb and flow of traffic patterns throughout the day. Intermittent congestion is perhaps more likely when everyone comes home from work and flips on their favorite streaming service, but this type of congestion, while fleeting, is largely unpredictable, and different from the more lasting congestion at points of interconnection that has been in the news. For good discussion of congestion at points of interconnection, see David Clark et al., "Measurement and Analysis of Internet Interconnection and Congestion," (paper, CSAIL, MIT & CAIDA, University of California, San Diego, September 9, 2014), http://groups.csail.mit.edu/ana/Measurement-and-Analysis-of-Internet-Interconnection-and-Congestion-September2014.pdf.

21. In TCP, oversimplifying a bit, the sender waits for an acknowledgement from the intended receiver. If it is not received in time, the sender backs off its sending rate, assuming packets are lost due to congestion. However, the control loop is so large, and the effect of a few dropped packets so severe, that this mechanism serves more to avoid congestion collapse for the overall system and does not optimize for any one application.

22. "US Broadband: Speeds Have Been Rising 25% Annually," Cartesian, accessed July 26, 2018, http://www.cartesian.com/us-broadband-speeds-have-been-rising-25-annually/.

23. Michael Abrash, "Latency – the *Sine Qua Non* of AR and VR," *Ramblings in Valve Time*, December 29, 2012, http://blogs.valvesoftware.com/abrash/latency-the-sine-qua-non-of-ar-and-vr/; "John Carmack Delivers Some Home Truths on Latency," *Oculus Rift-Blog*, accessed October 23, 2015, http://oculusrift-blog.com/john-carmacks-message-of-latency/682/.

24. Ibid.

25. Imtiaz Parvez et al., "A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions," Accepted in *IEEE Communications Surveys and Tutorials* (May 2018) *available at* https://arxiv.org/abs/1708.02562.

26. An ITU report examines the coming "Tactile Internet"—applications that rely on secure, robust, and extremely low-latency connections. The report identifies, in addition to VR and AR, robotics, telepresence, telemedicine, connected cars, smart grid, and a wide variety of interactive sensors and actuators as technologies that will place extreme demands on networks. These are innovations that we should encourage, not slow with bureaucracy. "The Tactile Internet," (ITU-T Technology Watch Report, August 2014), http://www.itu.int/dms_pub/itu-t/oth/23/01/T23010000230001PDFE.pdf.

27. David Clark and kc claffy, "Adding Enhanced Services to the Internet: Lessons for History," (paper presented at the 43rd Research Conference on Communication, Information and Internet Policy, September 8, 2015), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2587262.

28. Johannes Bauer and Günter Knieps, "Complementary Innovation and Network Neutrality," Telecommunications Policy (December 2017), https://www.sciencedirect.com/science/article/pii/S0308596117304615.

29. Ibid.

30. "Differentiated Treatment of Internet Traffic," Broadband Internet Technology Advisory Group (BITAG), (uniform agreement report, October 2015), http://www.bitag.org/documents/BITAG_-_Differentiated_Treatment_of_Internet_Traffic.pdf; Thomas W. Struble, "On the Relationship Between QoS & QoE: Why Differential Traffic Management on the Internet Is Not a Zero-Sum Practice" *TPRC 44* (August 2016), http://docs.techfreedom.org/Paid_Prioritization_TPRC_2016.pdf.

31. Ibid.

32. "Speedtest Global Index" Ookla Speedtest, June 2018, http://www.speedtest.net/global-index; Karl Bode, "Netflix: 4K Video Will Need at Least 15 Mbps," DSL Reports, September 23, 2013, http://www.dslreports.com/shownews/Netflix-4K-Video-Will-Need-At-Least-15-Mbps-125924.

33. Rani Molla, “An explosion of online video could triple bandwidth consumption again in the next five years” *Recode* (June 2017), https://www.recode.net/2017/6/8/15757594/future-internet-traffic-watch-live-video-facebook-google-netflix.

34. Ibid.

35. Daniel Lyons, “Paid prioritization: Debunking the myth of fast and slow lanes,” *AEIdeas* (April 2018), http://www.aei.org/publication/paid-prioritization-debunking-the-myth-of-fast-and-slow-lanes/.

36. Paul W Schroeder, “Statement Before the Committee on Energy and Commerce Subcommittee on Communications and Technology Hearing, ‘From Core to Edge: Perspective on Prioritization’” (April 2018), https://docs.house.gov/meetings/IF/IF16/20180417/108168/HHRG-115-IF16-Wstate-SchroederP-20180417-U31.pdf.

37. Ibid.

38. Progress and Freedom Foundation, “DACA Group Opposes Net Neutrality Mandate Regulatory Working Group Prefers Competition-Based Standard,” news release, March 20, 2006, http://www.pff.org/news/news/2006/032006dacastmt.html.

ABOUT THE AUTHOR

Doug Brake is director of broadband and spectrum policy at the Information Technology and Innovation Foundation. He specializes in broadband policy, wireless enforcement, and spectrum-sharing mechanisms.

He previously served as a research assistant at the Silicon Flatirons Center at the University of Colorado, where he sought to improve policy surrounding wireless enforcement, interference limits, and gigabit network deployment. Prior to that, he served as a Hatfield scholar at the Federal Communications Commission, assisting with the implementation of the advanced communications services section of the 21st Century Communications and Video Accessibility Act.

Brake holds a law degree from the University of Colorado Law School and a bachelor's degree in English literature and philosophy from Macalester College.

ABOUT ITIF

The Information Technology and Innovation Foundation (ITIF) is a nonprofit, nonpartisan research and educational institute focusing on the intersection of technological innovation and public policy. Recognized as the world's leading science and technology think tank, ITIF's mission is to formulate and promote policy solutions that accelerate innovation and boost productivity to spur growth, opportunity, and progress.

FOR MORE INFORMATION, VISIT US AT WWW.ITIF.ORG.