2

The Challenge of New Patterns in Internet Usage

By Christopher S. Yoo,[1] University of Pennsylvania

Introduction

The Internet unquestionably represents one of the most important technological developments in recent history. It has revolutionized the way people communicate with one another and obtain information and has created an unimaginable variety of commercial and leisure activities. Many policy advocates believe that the Internet's past success depended in no small part on its architecture and have argued that its continued success depends on preserving that architecture in the future.[2]

Interestingly, many members of the engineering community see the Internet in starkly different terms. For them, the Internet began as a military network, which caused it to reflect tradeoffs that would have been made quite differently had the Internet been designed as a commercial network from the outset.[3] Moreover, engineers often observe that the current network is ill-suited to handle the demands that end users are placing on it.[4] Indeed, engineering researchers often describe the network as ossified and impervious to significant architectural change.[5] As a result, the U.S. government has launched a series of initiatives to support research into alternative network architectures.[6] The European Commission has followed a similar course,[7] and university-based researchers in both the U.S. and Europe are pursuing a variety of "clean slate" projects studying how the Internet might be different if it were designed from scratch today.[8]

This essay explores some emerging trends that are transforming the way end users are using the Internet and examines their implications both for network architecture and public policy. Identifying future trends is inherently speculative and in retrospect will doubtlessly turn out to be mistaken in a number of important respects. Still, I hope that these ruminations and projections will yield some insights into range of possible evolutionary paths that the future Internet may take.

Internet Protocol Video

The development that has generated the most attention from policymakers and the technical community is the use of Internet-based technologies to distribute video programming. "Over-the-top" services (such as YouTube and Hulu) rely on the public Internet to distribute video. Other services, such as AT&T's U-verse, use the protocols developed for the Internet to distribute video through proprietary networks. Verizon's fiber-based FiOS service and many cable television providers already rely on these protocols to provide video on demand and are making preparations to begin using Internet-based technologies to distribute their regular video channels as well. Because these services are often carried in whole or in part by private networks instead of the public Internet, they are called "Internet Protocol (IP) Video" or IPTV. Industry observers have long predicted that video will represent an increasing proportion of total network traffic.

The growing use of IP-based protocols to distribute video has raised a number of technical and policy challenges. Not only will the growth of IPTV require more bandwidth, it may also require more basic changes in the architecture and regulatory regimes governing the network.

Bandwidth and Quality of Service

Industry observers have long disputed how large the video-induced spike in network demand will actually be. A recent industry report estimates that Internet video now represents over one-third of all consumer Internet traffic and will grow to more than 90 percent of all consumer traffic by 2014.[9] Experts disagree about what the future holds. Some industry observers have long predicted the coming of a video-induced "exaflood" that would require a sharp increase in capital spending.[10] The Minnesota Internet Traffic Studies (also known as MINTS and headed by Andrew Odlyzko) disagrees, pointing to the lack of any sign of such an upsurge in traffic.[11] Other observers challenge MINTS's conclusions, arguing that in focusing solely on traffic patterns at public peering points, MINTS fails to take into account the sizable proportion of the overall traffic that now bypasses the public backbone.[12] Moreover, even if the shift to IP-based video distribution has not yet by itself created a spike in the demand for bandwidth, the wide-scale deployment of high definition (and the looming emergence of ultra-high definition), 3D, and multiscreen technologies may cause the rate of traffic growth to increase in the future.

Aside from increased bandwidth, video requires network services that are qualitatively different in many ways from those required by the applications that formed the bulk of first-generation Internet usage. On the one hand, video is more tolerant of packet loss than web browsing and e-mail. On the other hand, unlike the performance of e-mail and web browsing, which depends solely on when the last packet is delivered, video quality depends on the timing with which every intermediate packet is delivered.

Specifically, video is more sensitive to *jitter*, which is variations in spacing between intermediate packets in the same stream and which typically arises

Christopher S. Yoo is Professor of Law, Communication, and Computer and Information Science, and Director of the Center for Technology, Innovation, and Competition at the University of Pennsylvania.

His research focuses on exploring the insights that the principles of network engineering and the economics of imperfect competition can provide into the regulation of the Internet and other forms of electronic communications. He has been a leading voice in the "network neutrality" debate that has dominated Internet policy over the past several years. He is also pursuing research on copyright theory as well as the history of presidential power.

He is the author (with Daniel F. Spulber) of *Networks in Telecommunications: Economics and Law* (Cambridge, 2009) and (with Steven G. Calabresi) of *The Unitary Executive: Presidential Power from Washington to Bush* (Yale, 2008). Yoo testifies frequently before Congress, the Federal Communications Commission, and the Federal Trade Commission.

He received his J.D. from Northwestern University, his M.B.A. from UCLA, and his A.B. from Harvard University.

when a stream of packets traverses routers that are congested. Jitter can cause video playback to freeze temporarily, which degrades the quality of the viewers' experience.

The usual solution to jitter is to delay playback of the video until the receiver can buffer enough packets to ensure that playback proceeds smoothly. This solution has the drawback of exacerbating another dimension of quality of service that is relevant for video, which is *delay* or *latency*, defined as the amount of time that it takes for playback to commence after it has been requested. Interestingly, viewers' tolerance of latency varies with the type of content being transmitted. While viewers of static video typically do not mind waiting five to 10 seconds for playback to begin, such delays are not acceptable for interactive video applications, such as video conferencing.[13] Some content providers reduce latency by using data centers or content delivery networks to position their content in multiple locations, thereby shortening the distance between the content and end users. Storing content in multiple locations only works for static content that does not change. It cannot work for interactive content, such as videoconferencing or online gaming, which changes dynamically.[14]

For interactive applications, the engineering community has focused on two other means for providing higher levels of quality of service. One solution is for network owners to overprovision bandwidth and switching capacity. When combined with distributed architectures for content delivery (such as caching and content delivery networks), this surplus capacity can give networks the headroom they need to handle any transient bursts in traffic without any congestion-related delays.[15] Overprovisioning is subject to a number of limitations, however. Wireless networks cannot simply add capacity to meet demand. Moreover, even networks that can increase bandwidth cannot do so instantaneously. Forecasting errors are inevitable, and in those instances where a network provider has failed to anticipate a key demographic shift or the emergence of a key application, device, or other complementary technology, it may sometimes find itself unable to expand capacity quickly enough to meet this increase in demand.[16] Overprovisioning also only increases the probability that particular traffic will pass through the network without delay. It does not guarantee the quality of service that any particular traffic will receive.[17] Finally, overprovisioning inherently requires networks to guarantee quality of service through capital expenditures (CapEx) rather than through operating expenditures (OpEx). As the difficulty in raising capital in the current economic downturn eloquently demonstrates, the relative costs of CapEx and OpEx solutions typically vary across time. Simple economics thus militate against locking network providers into one or the other option.[18]

The other alternative to provide higher-quality video service is to engage in increasingly sophisticated forms of network management that either reduce congestion or provide some means for providing higher levels of quality of service. Over the past two decades, the engineering community has developed a wide range of potential solutions, including Integrated Services (IntServ),[19] Differentiated Services (DiffServ),[20] MultiProtocol Label Switching (MPLS),[21] and Explicit Congestion Notification (ECN).[22] Other new initiatives, such as Low Extra Delay Background Transport (LEDBAT), also show promise.[23] Other engineers disagree with this approach, complaining that adding quality of service to the network would require devoting processing power in routers that would make the network too expensive and too slow.[24]

Leading engineering textbooks recognize that the engineering community is split over which solution—overprovisioning or network management—would be better.[25] The fact that the engineering community has yet to reach consensus counsels against regulatory intervention mandating either approach.

Congestion Management

The advent of IPTV may also require fundamental changes to the way the network deals with

congestion. The current approach to congestion management was developed in the late 1980s, shortly after the Internet underwent a series of congestion collapses. Because congestion is a network-level problem, in many ways the logical solution would have been to address it through a network-level solution, as was done in the original ARPANET, networks running asynchronous transfer mode (ATM), and many other early corporate networks. However, the router hardware of the early 1980s made implementing solutions at the network level prohibitively expensive. On the other hand, although edge-based congestion management is feasible, the hosts operating at the edge of the network typically lack the information to know when the network is congested.

Van Jacobson devised an ingenious mechanism that allows hosts operating at the edge of the network to infer when the core of the network has become congested.[26] This solution takes advantage of a particular feature of the Transmission Control Protocol (TCP). TCP ensures reliability by requiring the receiving host to send an acknowledgment every time it receives a packet. If the sending host does not receive an acknowledgement within the expected timeframe, it presumes that the packet was lost and resends it. Jacobson noted that packet loss typically occurs for one of two reasons: (1) transmission errors, or (2) discard by a router where congestion caused its buffer to become full. Because wireline networks rarely dropped packets due to transmission errors, hosts operating at the edge of the network could take the failure to receive an acknowledgement within the expected time as a sign of congestion and a signal to slow down their sending rates exponentially.[27]

This edge-based approach is now required of every computer attached to the Internet and continues to serve as the primary mechanism for managing congestion today. The problem is that TCP does not represent the only transport protocol commonly used on the network. In particular, by resending every packet that fails to receive an acknowledgement within the expected timeframe, TCP implicitly prioritizes reliability over delay. The DARPA protocol architects recognized from the Internet's earliest years that applications such as packet voice cannot tolerate such delays and would prefer to avoid them even if it meant sacrificing reliability altogether. To support these applications, the DARPA protocol architects created the User Datagram Protocol (UDP), which forgoes the use of acknowledgements altogether. UDP has now become the primary transport protocol for transmitting the data traffic associated with Voice over Internet Protocol (VoIP). Because IPTV makes the same tradeoff, UDP also has become the primary protocol for IPTV as well.

Because the mechanism for managing congestion described above depends on acknowledgements to signal when the network is congested, it does not work for protocols like UDP that do not use acknowledgements. While this was not a problem when UDP represented only a small proportion of bandwidth demand, the growing importance of VoIP and IPTV has caused UDP to become an increasingly significant component of network traffic. Consequently, engineers have sought to ensure that UDP acts in a way that is "TCP-friendly," measured in terms of whether a UDP-based application consumes more network resources than would a similar TCP-based application.[28] Some of these solutions require the receiving hosts to send acknowledgements in a manner somewhat reminiscent of TCP, which threatens to force UDP to run unacceptably slowly.[29] Others would require reconfiguring routers to send information about congestion to sending hosts, which had historically been rejected because of cost.[30] More recently, other engineers have organized a more fundamental attack on TCP-friendliness as the benchmark for evaluating bandwidth allocation, arguing that it allocates more bandwidth to users running applications that steadily generate small amounts of traffic than to users running applications that generate traffic in short bursts, even when the total amount of bandwidth consumed by both types of applications is exactly the same. It also tolerates allowing end users to seize more of the bandwidth simply by initiating multiple TCP sessions.[31]

Simply put, because video relies on UDP, the growth of video is putting pressure on the way the network manages congestion. Considerable disagreement remains over the best means for addressing this problem and also over the basis for evaluating the fairness or optimality of any particular solution. As a result, it is likely that different actors will pursue different solutions. Under these circumstances, policymakers must be careful to avoid the temptation to intervene to establish a uniform solution and should instead allow this debate to run its course.

Multicasting

TCP and UDP are *unicast* protocols, in that they transmit data between a single sender and a single receiver, with each destination receiving a separate stream of packets. While such an approach makes sense for person-to-person communications like e-mail or file transfer, it makes less sense for mass communications. Consider, for example, what occurs if an IPTV provider uses UDP to transmit video to one million viewers. Unicast technologies require that the provider transmit one million duplicate packets to its first hop router. The first hop router must in turn pass those packets on to downstream routers that serve multiple customers, even though many of those packets are duplicates as well.

Providers can avoid the inefficiency of distributing mass communications in this manner by using a *multicast* protocol. Instead of sending multiple copies of duplicate packets to the first hop router, multicasting sends a single stream of packets and depends on each downstream router to create duplicates as necessary.[32]

Although more efficient in terms of bandwidth usage, multicasting presents a number of challenges.[33] Multicasting requires the deployment of special routers in the core of the network that are capable of processing group information and duplicating packets as necessary. It also requires group management processes to inform routers when individual hosts tune in and out of the multicast stream. Effective group management also requires the security to ensure that multicasting is not used by unauthorized senders or receivers. Multicast flows are also typically not TCP friendly, so widespread use of multicasting may degrade unicast traffic and may even contribute to congestion collapse. Multicasting also presents routing challenges that are quite distinct from and more complicated than routing for unicast protocols. The lack of management tools has inhibited the deployment of multicasting that spans multiple domains. That said, many companies employ multicasting within their proprietary networks. Most notably for our purposes, AT&T relies on multicasting to distribute video through its U-verse network.

Multicasting is likely to play an increasingly important role as other video providers migrate their distribution systems to IP-based technologies. The need to deploy different hardware in the core of the network, group management tools, and routing algorithms create challenges that represent a significant change to the network's architecture.

Regulatory Classifications

More widespread use of IPTV is also likely to raise questions about its proper regulatory classification. Traditional multichannel video program distribution systems, such as cable television, are regulated as "cable services."[34] As such, they are required to pay franchising fees, provide leased access and PEG access, and to provide open access to their set-top boxes, among other requirements. Internet-based services have traditionally been classified as "information services" that have largely been exempt from such regulation.

What is the proper regulatory classification for IP-based video distribution systems? New services provided by telephone companies that use Internet technologies to distribute video over their own

proprietary networks, such as AT&T's U-verse and Verizon's FiOS services, are classified as cable services. Other video distribution platforms, such as YouTube and Hulu, do not own any access networks of their own. Instead, they distribute content over the public backbone and whatever last-mile connectivity individual end users have obtained. To date, these so-called "over the top" services have been exempt from regulation as cable services.

The increasing variety of IP video distribution platforms is starting to raise difficult definitional questions. For example, Internet-enabled gaming systems now support multiplayer gaming as well as direct interaction through video chat features. In addition, gaming systems are now important sources of over-the-top video services, such as Netflix. The convergence of gaming into the Internet ecosystem has raised the question of whether carrying over-the-top video turns these gaming platforms into cable services.

Wireless Broadband

Another emerging trend that is transforming U.S. Internet policy is the emergence of wireless as a technological platform for broadband service. The most recent data released by the FCC indicates that wireless has already captured nearly 25 percent of the market for high-speed lines as of the end of 2008, as compared with just over 40 percent for cable modem and just under 30 percent for ADSL.[35] The expansion of the U.S. wireless broadband market since 2008 and the emergence of wireless as the leading broadband platform in other countries both suggest that wireless broadband will become increasingly important in the years to come.

Policymakers sometimes suggest that the same principles applying to other broadband technologies should simply be extended to wireless. These suggestions overlook key technological differences between wireless and wireline technologies that policymakers must take into account.

Bandwidth Limits and Local Congestion

Wireless technologies face limitations that are quite different from those faced by wireline technologies. As noted earlier, wireless broadband is subject to bandwidth constraints that are much stricter than those confronted by wireline technologies. While wireless providers can increase capacity by relying on smaller cell sites operating at lower power, they cannot add capacity in the same manner as wireline providers.

In addition, because wireless technologies share bandwidth locally, they are more susceptible to local congestion than many fixed-line services, such as ADSL.[36] These problems are exacerbated by the fact that, in wireless networks, data and voice traffic typically share bandwidth, in contrast with telephone and cable companies, which typically place data traffic in a separate channel. Thus, excess data traffic can degrade wireless providers' core business to an extent not possible for other broadband technologies.

The Physics of Wave Propagation

Those who took physics in high school will recall that waves have some unique characteristics. They can reinforce each other in unexpected ways, as demonstrated by unusual echoes audible in some locations in a room, but not others, and by whispering corners, where the particular shape of the room allows sound to travel from one corner to the other even though a person speaks no louder than a whisper. As noise-reducing headphones marketed by Bose and other companies demonstrate, waves can also cancel each other out. Waves also vary in the extent to which they can bend around objects and pass through small openings, depending on their wavelength.

The unique features of waves can cause wireless technologies to face interference problems with which wireline technologies do not have to contend. For example, wireless signals attenuate much more rapidly with distance than do wireline signals. Moreover, in contrast to wireline technologies, there is an absolute limit to the density of wireless users that can operate in any particular area. Shannon's Law dictates that the maximum rate with which information can be transmitted given limited bandwidth is a function of the signal-to-noise ratio.[37] Unlike wireline transmissions, which travel in a narrow physical channel, wireless signals propagate in all directions and are perceived as noise by other receivers. At some point, the noise becomes so significant that the addition of any additional wireless radios becomes infeasible.

Wireless transmissions also suffer from what are known as multipath problems resulting from the fact that terrain and other physical features can create reflections that can cause the same signal to arrive at the same location multiple times. Unless the receiver is able to detect that it is receiving the same signal multiple times, it will perceive multipathing as an increase in the noise floor, which in turn reduces the available bandwidth. If the signal arrives 180° out of phase, it can even cancel out the original signal completely. Although smart receivers can avoid these problems if they know the exact location of each source, they cannot do so if the receiver or the other sources are mobile devices whose locations are constantly changing.

For these reasons, many wireless providers implement protocols that give priority to time-sensitive applications during times when subscribers are in areas of low bandwidth (such as by holding back e-mail while continuing to provide voice service). Other wireless providers rate limit or ban video or peer-to-peer downloads in order to prevent a small number of users from rendering the service completely unusable.[38]

Congestion Management

Wireless technologies also require a significantly different approach to congestion management. As noted earlier, the Internet's primary mechanism for managing congestion is based on the inference that, because wireline networks rarely drop packets due to transmission errors, any observed packet loss is likely to be due to congestion. The problem is that this inference is invalid for wireless networks, which drop packets due to transmission error quite frequently, either because of a bad handoff as a mobile user changes cells or because of the interference problems discussed above. When a packet is dropped due to transmission error, reducing the sending rate exponentially is precisely the wrong response. Instead, the sending host should resend the dropped packet as quickly as possible without slowing down. In other words, the optimal response for wireless networks may well be the exact opposite of the optimal response for wireline networks.

These differences have caused wireless networks to manage congestion and packet loss in different ways. Some solutions place a "snoop module" at the base station that serves as the gateway used by wireless hosts to connect to the Internet and that keeps copies of all packets that are transmitted and monitors acknowledgments passing in the other direction. When the base station detects that a packet has failed to reach a wireless host, it resends the packet locally instead of having the sending host do so.[39] Other solutions call for the sending host to be aware of when its transmission is carried in part by a wireless link and to distinguish between losses due to congestion and losses due to transmission errors. Still other solutions call for a split connection, in which the sending host establishes one TCP connection with an IP gateway in the middle of the network where the transmission shifts to wireless and a separate TCP connection between the IP gateway and the receiving host.[40] Some of these solutions violate the semantics of IP. All of them require introducing traffic management functions into the core of the network to a greater extent than originally envisioned by the Internet's designers.

The Heterogeneity of Devices

Wireless technologies do not vary only in terms of transmission technologies. They also vary in terms of end-user devices. Instead of relying on a personal computer, wireless broadband subscribers connect to the network through a wide variety of smartphones. These devices are much more sensitive to power consumption than are PCs, which sometimes leads wireless network providers to disable certain functions that shorten battery life to unacceptable levels. In addition, wireless devices have much less processing capacity and employ less robust operating systems than do the laptop and personal computers typically connected to wireline services. As a result, they are more sensitive to conflicts generated by multiple applications, which can cause providers to be much more careful about which applications to permit to run on them.

Wireless devices also tend to be much more heterogeneous in terms of operating systems and input interfaces (including keyboards and touch screens). As a result, the dimensions and levels of functionality offered by particular wireless devices vary widely. It seems too early to predict with any confidence which platform or platforms will prevail. Furthermore, as noted earlier, many wireless networks address bandwidth scarcity by giving a higher priority to time-sensitive applications, which typically requires close integration between network and device. These features underscore the extent to which variations in particular devices are often an inextricable part of the functionality of the network.[41]

Even more fundamentally, wireless devices interconnect with the network in a manner that is quite different from devices connected to wireline networks. Devices connected to wireline networks have IP addresses that are visible to all other Internet-connected hosts. Wireless devices, in contrast, do not have IP addresses. Instead, Internet connectivity is provided by an IP gateway located in the middle of the network that connects to individual wireless devices using a telephone-based technology rather than IP. Stated in technical terms, wireless broadband devices operate at Layer 2 rather than Layer 3 of the Internet protocol stack. Wireless devices will eventually connect through the Internet protocol once fourth-generation wireless technologies such as LTE are deployed. Until that time, wireless devices necessarily will connect to the Internet on different and less open terms than devices connected through wireline networks.

Routing and Addressing

Another problem confronting wireless broadband results from the fact that the Internet developed at a time when computers did not move. As a result, the architecture could use a single address to specify both the identity of a particular machine as well as where that machine was connected to the network. The advent of mobility has caused the unity of identity and location to break down. A single mobile device may now connect to the network through any number of locations. Although the network could constantly update the routing table to reflect the host's current location, doing so would require propagating the updated information to every router in the network as well as to an unacceptably large number of programs and databases.

Instead, mobile devices typically designate a router on its home network that has a fixed, permanent IP address as a "home agent" that serves as the initial contact point for all IP-based communications. Anyone seeking to contact a mobile device would first send the packets to the home agent, which would then encapsulate the packets in another packet and forward them to wherever the mobile host is currently located. Managing mobile communications in this manner is surprisingly complex and requires protocols for home agents to notify others of its location, to encapsulate traffic bound for the mobile host, and to allow mobile hosts to register and deregister their current location with their home agents, notify the foreign network that they are currently attached to it, and decapusulate the packets as they arrive. Sending communications via the home

An Example of Route Architecture

Wireless technologies are also causing pressure on the amount of resources that the network must spend on keeping track of Internet addresses. Tier 1 ISPs necessarily must maintain complete routing tables that contain routes to the IP address for every host connected to the Internet. The current system relies on route aggregation to keep routing tables from growing too large. This mechanism can be illustrated by an analogy to the telephone system. Consider a party in Los Angeles who is attempting to call the main telephone number for the University of Pennsylvania, which is (215) 898-5000. So long as all calls to the 215 area code pass through the same outbound link, a phone switch in Los Angeles could represent all telephone numbers in that area code with a single entry in its routing table. Similarly, so long as all telephone numbers in the 898 directory are connected to the same central office, switches within Philadelphia need not maintain separate entries for each phone number in that directory. Instead, they can represent all telephone numbers located in (215) 898-xxxx with a single entry.

The Internet employs a similar system known as Classless InterDomain Routing (CIDR) to aggregate routes. CIDR is even more flexible. It can aggregate routes at any number of digits rather than being limited in the manner of area codes and directories, with the digits of course being represented in binary.

This strategy depends on the address space remaining compact. In other words, this approach will fail if the 215 area code includes phone numbers that are not located in Philadelphia. If that is the case, the routing table will have to use separate entries to keep track of every single address. The problem is that true mobile addressing fragments the geographic compactness of the address space.

Another problem is somewhat more subtle. The current architecture is built on the implicit assumption that Internet addresses change on a slower timescale than do communication sessions. So long as the address architecture changes at a slower timescale, any particular Internet-based communication may take the address architecture as given. Mobility, however, increases the rate at which the address architecture changes. In addition, because addressing is handled on a decentralized basis, information about changes in the address architecture takes time to spread across the Internet. Increases in the rate with which the address space changes can cause communications sessions to fail and create the need for a new way to manage addresses.

Others have proposed radical changes in the addressing and routing architecture. One approach would replace the single address now employed in the network with two addresses: one to identify the particular machine and the other to identify its location. Whatever solution is adopted would represent a fundamental change in the network layer that unifies the entire Internet.

agent also suffers from the inefficiency of what is sometimes called "triangle routing," because, instead of passing directly from the sending host to the receiving host, traffic must travel first from the sending host to the home agent and then from the home agent to the receiving host. In the extreme case, a communication between two mobile hosts located next to one another in a conference room on the West Coast might have to travel back and forth across the country if one of them has a home agent located on the East Coast. The home agent can eliminate triangle routing by passing the mobile host's current location on to the sender so that the sender may forward subsequent packets to it directly. The initial communications must still bear the inefficiency of triangle routing. Moreover, such solutions become much more difficult to implement if the mobile agent is constantly on the move.[42]

Cloud Computing

Cloud computing represents one of the hottest topics in today's information technology (IT) community. Under the traditional paradigm, end users run applications on data stored locally on the host computer's hard disk. Under cloud computing, data resides in the network and is accessed on demand.

The National Institute of Standards and Technology divides the type of services offered by cloud computing providers into three categories:[43]

- **Software as a Service (SaaS)** providers offer finished applications that end users can access through a thin client (typically a web browser). Prominent examples of SaaS include Gmail, Google Docs, and Salesforce.com. The only computing power that an end user needs to access SaaS is a netbook capable of running a web browser. The end user has limited control over the design of the application, such as minor customization and configuration. It has no control over the servers, networking, and storage infrastructure.
- **Platform as a Service (PaaS)** providers offer suites of programming languages and software development tools that customers can use to develop their own applications. Prominent examples include Microsoft Windows Azure and Google App Engine. PaaS gives end users control over application design, but does not give them control over the physical infrastructure.
- **Infrastructure as a Service (IaaS)** providers offer end users direct access to processing, storage, and other computing resources and allow them to deploy their own operating systems and to configure those resources as they see fit. Examples of IaaS include Amazon Elastic Compute Cloud (EC2), Rackspace, and IBM Computing on Demand.

Cloud computing can also be understood in terms of the business needs motivating its adoption. Customers are often reluctant to abandon their entire corporate intranets and to rely exclusively on cloud computing. One business case that stops short of fully embracing cloud computing is *disaster recovery*, in which customers back up their data remotely on the network. It may also involve the functionality to access that data on a short-term basis should the customer's internal network fail. Another classic scenario is called *cloud bursting*, in which the customer relies on cloud computing to provide overflow capacity to cover spikes in demand.

Proponents of cloud computing predict that it will yield substantial benefits.[44] Assuming that data centers allow multiple customers to share the same hardware, cloud computing should allow smaller companies to take advantage of scale economies that they could not realize on their own. Even companies that are large enough to achieve minimum efficient scale by themselves may see advantages. The fact that hardware represents discrete (often significant) investments that must typically be provisioned in advance means that companies risk running out of capacity should demand grow more rapidly than anticipated. Conversely, they may face the burden of underutilized resources should demand grow unexpectedly slowly. The fact that companies must provision hardware for peak demand also means that cloud computing is particularly helpful when demand is highly variable, since aggregating demand lowers variability.[45] The greater dispersion made possible by virtualization can reduce latency and increase reliability.

Predictions about the future of cloud computing run the gamut, with some forecasting that all IT will eventually migrate into the cloud[46] and others arguing that it is nothing more than overhyped repackaging of existing technologies.[47] What is even more poorly understood is what increasing use of cloud computing would mean for the network architecture.

End User Connectivity

Cloud computing customers need different services from the network that provides the physical connectivity to end users (often called the "last mile"). Since the software and data needed to run applications no longer reside on end users' hard disks, cloud computing needs more ubiquitous connectivity and more substantial uptime guarantees than previously required. Because data processing no longer occurs locally, reliance on cloud computing also increases demand for the quantity of bandwidth as well as its ubiquity.

Moreover, because cloud computing provides services that used to be delivered by corporate intranets, cloud computing users may well demand higher levels of quality of service from their last-mile networks. These demands will likely vary from company to company. For example, financial services companies typically require perfect transactions with latency guarantees measured in microseconds. In addition, the provider must be able to verify the delivery time of each and every transaction after the fact. The fact that information that used to reside exclusively within an end user's hard disk and processor must now be transmitted over the network also means that cloud computing customers are likely to demand higher levels of security from their last-mile networks.

Data Center Connectivity

The advent of cloud computing also requires improvements in data center connectivity. As an initial matter, customers establishing new cloud computing instances must provision their data to the data center. Because datasets in the terabyte range would take weeks to upload, many cloud computing providers recommend that customers download their data onto a physical storage medium and send it via an overnight mail service, such as FedEx.[48]

The agility and virtualization demanded by cloud computing also requires the flexibility to move large amounts of data between data centers very quickly. The best-efforts architecture of the current Internet cannot offer the guaranteed levels of quality of service that these functions require. For this reason, many cloud computing providers interconnect their data centers through dedicated private lines. Others have begun outsourcing these services to other networks, partially to gain the economies of sharing resources with other firms and partially out of concern that operating these networks will lead them to be classified as common carriers.

Cloud computing is also placing new demands on the network's approach to routing. The BGP-based system responsible for routing traffic on the current Internet employs an algorithm that by default sends traffic along the path that transverses the fewest autonomous systems. Most cloud computing providers need greater control over the paths taken by key traffic. As a result, many rely on MPLS or some other protocol to manage routing. On a more radical level, some industry observers note that the identity/locator split discussed above, with mobile computing assigning separate addresses to each individual machine and to the location that the machine is currently connected to the network, should be augmented still further with a third address to mark the location where the application that the machine is accessing resides.[49]

Privacy and Security

Finally, cloud computing has fairly significant implications for privacy and security. As an initial matter, cloud computing often requires large amounts of data that previously did not leave a corporate campus to be shifted from one data center to another. In addition, virtualization necessarily envisions that this data will reside on the same servers as other companies' data. As a result, the hardware located in these data centers and the networks interconnecting them require a higher level of security than previously necessary. Industry participants are also often very protective of information about the volume and pattern of their transactions. They are thus likely to impose stringent requirements on what data can be collected about their operations and how that data is used.

The fact that data may be shifted from one data center to another also potentially makes that data subject to another jurisdiction's privacy laws. Because customers are ultimately responsible for any such violations, they are likely to insist on a significant degree of control over where it resides at any particular moment.

Programmable Networking

One of the primary architectural principles underlying the Internet is that routers should operate on a pure store-and-forward basis without having to keep track of what happens to packets after they have been passed on. This commitment is reflected in the Internet's general hostility toward virtual circuits and the belief that routers should not maintain per-flow state. Opponents of network management often point to the Senate testimony offered by officials of Internet2 (a nonprofit partnership of universities, corporations, and other organizations devoted to advancing the state of the Internet) noting that, although their network designers initially assumed that ensuring quality of service required building intelligence into the network, "all of [their] research and practical experience supported the conclusion that it was far more cost-effective to simply provide more bandwidth."[50]

To a certain extent, this long-standing hostility toward virtual circuits is an artifact of the Internet's military origins that has less relevance for the Internet of today. DARPA protocol architect David Clark has pointed out that the belief that routers operating in the core of the network should not maintain per-flow state derived largely from the high priority that military planners placed on survivability.[51] As noted earlier, survivability does not represent a significant concern for the modern Internet.

Moreover, technologies such as IntServ and MPLS, both of which are governed by accepted IETF standards, use virtual circuits to simplify packet forwarding and to support a fairer and more efficient allocation of traffic. Although IntServ has not achieved widespread acceptance, interest in MPLS appears to be growing.

These developments can be seen as part of a broader move away from viewing routers as static devices that always operate in a particular way and toward looking at the network as a programmable switching fabric that can be reconfigured from store-and-forward routers into virtual circuits as needed. For example, Internet2 (which, as noted earlier, is often held out as proof of the engineering community's conviction that network management is unnecessary) offers a service that it calls its Interoperable On-demand Network (ION) that allows researchers to establish dedicated point-to-point optical circuits to support large data transfers and other bandwidth-intensive applications. Internet2 notes that the "advanced science and engineering communities … are already straining the limits of today's network capabilities—and capacities" and that advanced media and telepresence applications often need the type of dedicated circuits previously regarded as anathema.[52]

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

Pervasive Computing and Sensor Networks

The last development that I will discuss that promises to effect some fundamental changes to the network architecture is the deployment of pervasive computing and sensor networks.[53] Computer chips are being incorporated into an ever-widening array of devices. In addition, the growth of what the ITU calls "the Internet of things" means that more and more objects are being outfitted with radio frequency identification (RFID) chips, which combine a wireless antenna with a tiny amount of memory.[54] Although RFID chips do not have their own power sources, the wireless signal from a sensor can provide enough power to activate them.

Most of the literature on sensor networks has focused on the privacy implications. What has been largely overlooked is the extent to which sensor networks and pervasive computing will require different services from the network. As an initial matter, pervasive computing and RFID chips may require a greater degree of last-mile connectivity than the network currently provides. Sensor networks also necessarily involve machine-to-machine communications, which are typically more intensive and follow patterns that are quite different from those that occur when human beings initiate the communications. These developments also represent a significant increase in the number of devices that will require network visibility, which will increase the pressure on the network to migrate to IPv6. In addition, the mobility of many of these endpoints may accelerate the rate of change within the address space, which may cause changes in routing and addressing systems.

Equally importantly, these developments represent a fairly significant increase in the heterogeneity of devices attached to the network. The current network model implicitly assumes that the network interconnects a series of general purpose devices. Pervasive computing and sensor networks involve more specialized devices that perform a narrower range of functions. As such, they may require a different approach to networking. For example, these devices may not be able to send acknowledgements in the manner envisioned by TCP. Unleashing the functionality of these stripped-down devices may also require a much tighter integration with the network.

Consequently, these devices may not have individual IP addresses. Instead, they may reside behind an IP gateway and communicate with one another through a lower-layer protocol. If so, they may require more wide-scale deployment of the middlebox architecture that has proven so controversial. That said, it is probably too early to offer any reliable predictions of the impact that deployment of these technologies will have on the architecture of the network.

Conclusion

One recurrent theme in the debates over Internet policy is the claim that the Internet's future success depends on preserving the architecture that has made it successful in the past. This claim has always struck me as inherently conservative and potentially Panglossian.[55] Policymakers must be open to the possibility that fundamental changes in the way people are using the network may require the network to evolve in new directions. Indeed, a significant portion of the engineering community believes that the time is ripe for a clean-slate approach aimed at creating a network that is starkly different from the one we have today. It is also possible that the network may not have a single response to these developments. Instead, as what people want from the network becomes increasingly heterogeneous, different portions of the network will respond in different ways to meet this demand.

Exactly what architectural changes will be required to meet these new challenges is difficult to foresee. Instead of creating regulations that lock in any particular vision of the network's architecture, policymakers should create regulatory structures that give industry actors the latitude they need to experiment with different solutions. In so doing, policymakers would do well to recognize that, while disruptive, change is inevitable and to keep in mind the aphorism that in a technologically dynamic environment, businesses are either moving forward or moving backward; there is no standing still.

Endnotes

1. Professor of Law, Communication, and Computer and Information Science and Founding Director of the Center for Technology, Innovation, and Competition, University of Pennsylvania.
2. *See, e.g.*, Preserving the Open Internet, Notice of Proposed Rulemaking, 24 F.C.C.R. 13064, 13065-67 ¶¶ 3-8 (2009); *Net Neutrality: Hearing Before the S. Comm. on Commerce, Science, & Transportation*, 109th Cong. 54-56 (2006) (statement of Lawrence Lessig), *available at* http://commerce.senate.gov/public/?a=Files.Serve&File_id=c5bf9e54-b51f-4162-ab92-d8a6958a33f8.
3. For example, David Clark's seminal description of the priorities animating the Internet's initial design (and which represents one of the most frequently cited articles in the literature) notes that the Internet's origins as a Defense Department initiative led the protocol architects to place a high priority on certain concerns that would be relatively unimportant to the commercial Internet (such as survivability in a hostile environment) and to downplay other priorities that would prove critical once the Internet became a mainstream phenomenon (such as efficiency and cost allocation).
   See David D. Clark, *The Design Philosophy of the DARPA Internet Protocols*, 18 ACM Computer & Comm. Rev. 106, 107, 110 (1988).
4. These lists typically include such major concerns as security, mobility, quality of service, multicasting, and multihoming. *See, e.g.*, Mark Handley, *The Internet Only Just Works*, 24 BT Tech. J. 119, 123, 126-27 (2006); Jon Crowcroft, *Net Neutrality: The Technical Side of the Debate: A White Paper*, Computer & Comm. Rev., Jan. 2007, at 49, 50-51, 52.
5. *See, e.g.*, Paul Laskowski & John Chuang, *A Leap of Faith? From Large-Scale Testbed to the Global Internet* 2 (unpublished manuscript presented at TPRC's 37th Research Conference on Communication, Information, and Internet Policy Sept. 27 2009), *available at* http://www.tprcweb.com/images/stories/papers/Laskowski_2009.pdf (collecting sources); *see also* Olivier Martin, *State of the Internet & Challenges Ahead* 1, 29 (2007), *available at* http://www.ictconsulting.ch/reports/NEC2007-OHMartin.doc (noting that "there appears to be a wide consensus about the fact that the Internet has stalled or ossified").
6. One example is DARPA's New Arch initiative. *See* David Clark et al., *New Arch: Future Generation Internet Architecture* 4, 13 (Final Technical Report Dec. 31, 2003), http://www.isi.edu/newarch/iDOCS/final.finalreport.pdf. The National Science Foundation is pursuing similar initiatives. One is known as the Global Environment for Network Innovations (GENI). *See* National Science Foundation, Global Environment for Network Innovation (GENI), http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=501055. Another was originally known as the Future Internet Design (FIND) project. *See* Vint Cerf et al., *FIND Observer Panel Report* (Apr. 9, 2009), http://www.nets-find.net/FIND_report_final.pdf. FIND was subsequently folded into the NSF's Networking Technology and Systems (NeTS) program. *See* National Science Foundation, Networking Technology and Systems (NeTS), http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503307. The NSF's major current initiative is the Future Internet Architectures program. *See* National Science Foundation, Future Internet Architectures (FIA), http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503476.
7. *See, e.g.*, European Commission Directorate-General for the Information Society and Media Project SMART 2008/0049, Internet Development Across the Ages (Jan. 2010), http://cordis.europa.eu/fp7/ict/fire/docs/executive-summary_en.pdf; European Commission, FIRE – Future Internet and Experimentation, (July 6, 2010), http://cordis.europa.eu/fp7/ict/fire/home_en.html.
8. *See, e.g.,* Jon Crowcroft & Peter Key, *Report from the Clean Slate Network Research Post-SIGCOMM 2006 Workshop*, Computer & Comm. Rev., Jan. 2007, at 75; Anja Feldmann, *Internet Clean-Slate Design: What and Why?*, Computer & Comm. Rev., July 2007, at 59; 100x100 Clean Slate Project, http://100x100network.org/; Stanford University Clean Slate Project, Program Goals, http://cleanslate.stanford.edu/index.php.
9. Cisco Systems, Inc., Cisco Visual Networking Index: Forecast and Methodology, 2009-2014, at 2 (June 2, 2010), http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-481360.pdf.
10. *See, e.g.*, Brett Swanson & George Gilder, *Estimating the Exaflood: The Impact of Video and Rich Media on the Internet* (Jan. 2008), http://www.discovery.org/a/4428.
11. *See* University of Minnesota, Minnesota Internet Traffic Studies, http://www.dtc.umn.edu/mints/home.php (estimating that Internet traffic was continuing to grow at the previous annual rate of 40 percent to 50 percent as of the end of 2009).

12. Nemertes Research, Internet Interrupted: Why Architectural Limitations Will Fracture the 'Net 34-35 (Nov. 2008). For example, regional ISPs that are too small to peer with backbone providers are now peering with each other in a practice known as "secondary peering," which allows them to exchange traffic without employing the public backbone. Content delivery networks such as Akamai and Limelight now use "content delivery networks" to store information at thousands of locations around the world, often in places where they can deliver traffic without traversing the backbone. Lastly, large content and application providers are building large server farms that similarly allow them to distribute their content without touching the public backbone. *See* Christopher S. Yoo, *Innovations in the Internet's Architecture that Challenge the Status Quo*, 8 J. on Telecomm. & High Tech. L. 79, 84-90 (2010).
13. Andrew S. Tanenbaum, Computer Networks 395-98 (4th ed. 2003); James F. Kurose & Keith W. Ross, Computer Networking: A Top-Down Approach 618-19, 622 (5th ed. 2010).
14. Kurose & Ross, *supra* note 13, at 626-29.
15. *Id.* at 603.
16. Christopher S. Yoo, *Beyond Network Neutrality*, 19 Harv. J.L. & Tech. 1, 22-23, 70-71 (2005).
17. Kurose & Ross, *supra* note 13, at 664.
18. Yoo, *supra* note 16, at 23, 71.
19. *See* Bob Braden et al., *Integrated Services in the Internet Architecture: An Overview* (1994) (RFC 1633).
20. *See* Steven Blake et al., *An Architecture for Differentiated Services* (1998) (RFC 2475).
21. *See* Eric C. Rosen et al., *Multiprotocol Label Switching Architecture* (2001) (RFC 3031).
22. *See* K.K. Ramakrishnan, *The Addition of Explicit Congestion Notification (ECN) to IP* (2001) (RFC 3168).
23. *See* Internet Engineering Task Force, Low Extra Delay Background Transport (LEDBAT) Working Group Charter (2009), http://www.ietf.org/html.charters/ledbat-charter.html.
24. Douglas E. Comer, Internetworking with TCP/IP 510 (5th ed. 2006).
25. The leading engineering textbook on TCP/IP notes the continuing existence of a "major controversy" over whether quality of service is necessary and feasible. *Id.* at 510, 515. Another textbook describes the "continuing debate" between those who would use network management to provide quality of service guarantees and those who believe that increases in bandwidth and the use of content distribution networks can obviate the need for network management. Kurose & Ross, *supra* note 13, at 602-04.
26. Van Jacobson, *Congestion Avoidance and Control*, 18 Computer & Comm. Rev. 314 (1988).
27. *Id.* at 319.
28. Jamshid Madhavi & Sally Floyd, TCP-Friendly Unicast Rate-Based Flow Control (Jan. 1997) (unpublished manuscript), *available at* http://www.psc.edu/networking/papers/tcp_friendly.html.
29. *See* Randall R. Stewart, *Stream Control Transmission Protocol* (2000) (RFC 2960).
30. Sally Floyd & Kevin Fall, *Promoting the Use of End-to-End Congestion Control in the Internet*, 6 IEEE/ACM Transactions on Networking, 458 (Oct. 1998).
31. *See* Bob Briscoe, *A Fairer, Faster Protocol*, IEEE Spectrum, Dec. 2008, at 38; Jon Crowcroft, TCP Friendly Considered Unfriendly (Dec. 6, 2001), http://www.cl.cam.ac.uk/~jac22/otalks/TCP_Too_Friendly_files/frame.htm.
32. In addition, *broadcast* protocols exist that transmit packets to every host connected to the network. Broadcasting is inefficient if only a fraction of the hosts are interested in the message.
33. *See* Ian Brown et al., *Internet Multicast Tomorrow*, Internet Protocol J., Dec. 2002, at 2; Christophe Diot et al., *Deployment Issues for the IP Multicast Service and Architecture*, IEEE Network, Jan./Feb. 2000, at 78.
34. 47 U.S.C. § 522(6).
35. FCC Indus. Analysis & Tech. Div., Wireline Competition Bur., High-Speed Services for Internet Access: Status as of December 31, 2008, at 9 chart 2 (Feb. 2010), http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-296239A1.pdf.
36. Because cable modem systems also share bandwidth locally, they are similarly susceptible to local congestion. *See* Christopher S. Yoo, *Network Neutrality, Consumers, and Innovation*, 2008 U. Chi. Legal F. 179, 199-202.
37. C.E. Shannon, *A Mathematical Theory of Communication* (pt. 1), 27 Bell Sys. Tech. J. 379 (1948); C.E. Shannon, *A Mathematical Theory of Communication* (pt. 2), 27 Bell Sys. Tech. J. 623 (1948).
38. A recent, eloquent demonstration of this strategy is the placards aboard the Amtrak Acela express trains asking passengers to refrain from using the WiFi service to download video.
39. Hari Balakrishnan et al., *Improving Reliable Transport and Handoff Performance in Cellular Wireless Networks*, 1 Wireless Networks 469 (1995).

40. Kurose & Ross, *supra* note 13, at 585-86; Tanenbaum, *supra* note 13, at 553-54.
41. Charles L. Jackson, *Wireless Handsets Are Part of the Network* (Apr. 30, 2007), http://files.ctia.org/pdf/Comments_CTIA_SkypeOpposition_AppendixC_43007.pdf.
42. Comer, *supra* note 24, at 339-46; Kurose & Ross, *supra* note 13, at 566-77; Tanenbaum, *supra* note 13, at 372-75, 462-64.
43. Peter Mell & Tim Grance, The NIST Definition of Cloud Computing 2 (version 15 Oct. 7, 2009), http://csrc.nist.gov/groups/SNS/cloud-computing/cloud-def-v15.doc.
44. *See* Joe Weinman, *The 10 Laws of Cloudonomics*, Bloomberg Bus. Week, Sept. 6, 2008, http://www.businessweek.com/technology/content/sep2008/tc2008095_942690.htm.
45. In short, aggregating flows that are not perfectly correlated reduces variability. Christopher S. Yoo, *Rethinking the Commitment to Free, Local Television*, 52 Emory L.J. 1579, 1707-08 & n.471 (2003). As one industry commentator observes, "The peak of the sum is never greater than the sum of the peaks." Weinman, *supra* note 44.
46. Nick Carr, The Big Switch (2007).
47. Oracle CEO Larry Ellison Bashes "Cloud Computing" Hype, http://www.youtube.com/watch?v=UOEFXaWHppE.
48. Jon Brodkin, *Amazon Cloud Uses FedEx Instead of the Internet to Ship Data*, Network World, June 10, 2010, http://www.networkworld.com/news/2010/061010-amazon-cloud-fedex.html.
49. Nemertes Research, *supra* note 12, at 53-54.
50. *Net Neutrality: Hearing Before the S. Comm. on Commerce, Science, and Transportation*, 109th Cong. 64, 66 (2006) (statement of Gary R. Bachula, Vice President, External Affairs, Internet2), *available at* http://www.gpo.gov/fdsys/pkg/CHRG-109shrg605/pdf/CHRG-109shrg605.pdf.
51. Clark, *supra* note 3, at 107-08.
52. Internet2, Internet2 ION (Sept. 2009), http://www.internet2.edu/pubs/200909-IS-ION.pdf.
53. For the seminal work, see Mark Weiser, *The Computer for the 21st Century*, Sci. Am., Sept. 1991, at 94.
54. Int'l Telecomm. Union, ITU Internet Reports 2005: The Internet of Things (Nov. 2005).
55. Pangloss was the teacher in Voltaire's *Candide* who often remarked that "we live in the best of all possible worlds." Voltaire, Candide and Other Stories 4, 14, 88 (Roger Pearson trans., 2006) (1759).