Explorations in Cyber International Relations
Massachusetts Institute of Technology Harvard University

Characterizing Cyberspace: Past, Present and Future

David D. Clark

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology

March 12, 2010

This material is based on work supported by the U.S. Office of Naval Research, Grant No. N00014-09-1-0597. Any opinions, findings, conclusions or recommendations therein are those of the author(s) and do not necessarily reflect the views of the Office of Naval Research.

Characterizing cyberspace: past, present and future

David Clark
MIT CSAIL

Version 1.2 of March 12, 2010

This work is funded by the Office of Naval Research under award number N00014091059 and N000140810898. Any opinions, findings, and conclusions or recommendations expressed in this email are those of the author (s) and do not necessarily reflect the views of the Office of Naval Research.

Introduction

In general terms, most practitioners share a working concept of cyberspace—it is the collection of computing devices connected by networks in which electronic information is stored and utilized, and communication takes place[1]. Another way to understand the nature of cyberspace is to articulate its purpose, which I will describe as the processing, manipulation and exploitation of information, the facilitation and augmentation of communication among people, and the interaction of people and information. Both information and people are central to the power of cyberspace. If we seek a better understanding of what cyberspace might be, one approach is to identify its salient characteristics: a catalog of its characteristics may be more useful than a list of competing definitions.

A four layer model

In this note, I will attempt to capture the character of cyberspace using a model with four layers. From the top down, the important layers are:

- The people who participate in the cyber-experience—who communicate, work with information, make decisions and carry out plans, and who themselves transform the nature of cyberspace by working with its component services and capabilities.
- The information that is stored, transmitted, and transformed in cyberspace.
- The logical building blocks that make up the services and support the platform nature of cyberspace.
- The physical foundations that support the logical elements.

It is not the computer that creates the phenomenon we call cyberspace. It is the interconnection that makes cyberspace—an interconnection that affects all the layers in our model. Today, we associate the phenomenon with the Internet, with its particular approach to interconnection, but there could be many alternative cyberspaces, defined (and created) by different approaches to interconnection. Indeed, in his book *The Victorian Internet*[2], Tom Standage argues that the mode of interconnection created by the

[1] The term was coined by a science fiction writer, William Gibson, and popularized in his book Neuromancer (1984).

[2] Standage, Tom. *The Victorian Internet.* Berkley Trade (October 15, 1999)

telegraph was as transformative in its time as the Internet is today. But the structure (and the structural implications) of the telegraph and the Internet could not be more different, as I will argue below.

Looking at the layers

The physical layer

The physical layer of cyberspace is the foundation of cyberspace—the physical devices out of which it is built. Cyberspace is a space of interconnected computing devices, so its foundations are PCs and servers, supercomputers and grids, sensors and transducers, and the Internet and other sorts of networks and communications channels. Communications may occur over wires or fibers, via radio transmission, or by the physical transport of the computing and storage devices from place to place. The physical layer is perhaps the easiest to grasp; since it is tangible, its physicality gives it a grounded sense of location. Physical devices such as routers or data centers exist in a place and thus sit within a jurisdiction. Some physical components, such as residential access networks, are capital-intensive, and the industries that produce them are as much construction companies as telecommunications companies. These firms are the traditional targets of telecommunications regulation, since their physical assets make them "easy to find".

The logical layer

The physical foundations of cyberspace are important—cyberspace is a real artifact build out of real elements, not a fantastical conception with no grounding. But the nature of cyberspace—its strengths and its limitations, derive more from the decisions made at the logical level than the physical level. The Internet, for example, provides a set of capabilities that are intentionally divorced to a great extent from the details of the technology that underpins it. If one wants to understand why some of the Internet vulnerabilities exist—why it allows phishing or denial of service attacks, for example, it is correct but not very useful to point out that computers and communications are subject to the laws of physics. It would have been possible to build a very different Internet within the constraints of the same physics. The decisions that shape the Internet arise at the higher layer—the logical layer where the platform nature of the Internet is defined and created. So that layer is going to be central to many of the considerations that arise when we analyze cyberspace, as will the layers that deal with information and with people.

The design of the Internet leads to a cyberspace that is build out of components that provide services, and these services are designed so that they can be composed and combined to form more complex services. Low level services include program execution environments, mechanisms for data transport, and standards for data formats. Out of this are build applications, such as a word processor, a database or the Web. By combining these, more complex services emerge. For example, by combining a database with the Web, we get dynamic content generation and active Web objects. On top of the Web, we now see service such as Facebook that are themselves platforms for further application development. The nature of cyberspace is the continuous and rapid evolution of new capabilities and services, based on the creation and combination of new logical constructs, all running on top of the physical foundations. Cyberspace, at the logical level, is thus a series of *platforms,* on each of which new capabilities are constructed, which in turn become a platform for the next innovation. Cyberspace is very *plastic*, and it can be described as *recursive*; platforms upon platforms upon platforms. The platforms may differ in detail, but they share the common feature that they are the foundation for the next platform above them.

One could build a very different system by taking a different approach to the logic of interconnection. Using the same sorts of physical elements, one could design a closed, essentially fixed function system such as an air traffic control system. Earlier examples of interconnected systems tended to have this character—fixed function and closed; the telegraph and the early telephone system had this character. Both these systems predate the computer, and indeed predate essentially all of what we call electronics—not just the transistor but the vacuum tube and the relay. It is the interconnection that makes cyberspace, but it is the programmability and generality of the computer that makes possible the flexible logical structure I am associating with cyberspace.

The drive for increased productivity and comparative competitive advantage shapes many aspects of cyberspace as we see it, but most particularly it drives toward a characteristic that allows for rapid innovation, new patterns of exploitation and so on. The logical layer—the "platform/plasticity" component of cyberspace—enhances this capability, and this fact in turn fuels the emphasis and attention given to this layer by the market and by developers of cyberspace. But we must not restrict our attention there, because another aspect of this drive is the expectations this rapid change implies for the "people" layer—rapid change will bring forward and advance people who can recognize new opportunities to exploit ICT, who value rather than fear change and new ways of doing things, and so on.

The information layer

As noted above, there are many aspects to cyberspace, including the technology-mediated interconnection of people. But clearly the creation, capture, storage and processing of information is central to the experience. Information in cyberspace takes many forms—it is the music and videos we share, the stored records of businesses, and all of the pages in the world wide web. It is online books and photographs. It is information about information (meta-data). It is information created and retrieved as we search for other information (as is returned by Google).

The character of information in cyberspace (or "on the net") has changed greatly since computers first started working with data sets. Data has been processed by isolated computers well before we had capabilities for interconnection. Data lived in card decks, on tapes, and later on disks. Initially, data was normally thought of as static, stored and retrieved as needed. Books are static products of authors, images are static, and so on. Massive archives of static information still exist, such as corporate transaction records that are now stored in "data warehouses" and "mined" for further information. But more and more, information is created dynamically on demand, blurring the boundaries between storage and computation. Web pages are now often made on demand, tailored to each user, based on component information stored in data bases. Information is now becoming more a personal experience, not a communal one. Issues of ownership, authenticity, and dependability are all critical as more and more information moves online.

The top layer—people

People are not just the passive users of cyberspace, they define and shape its character by the ways they choose to use it. The people and their character, which may vary from region to region, is an important part of the character of cyberspace. If people contribute to Wikipedia, then Wikipedia exists. If people tweet, then Twitter exists.

Alfred Thayer Mahan, in *The Influence of Sea Power Upon History*, wrote:
"The history of the seaboard nations has been less determined by the shrewdness and foresight of governments than by conditions of position, extent, configuration, number and character of their people,--by what are called, in a word, natural conditions." As we contemplate the nature of cyberspace, and the position of different countries with respect to their place and power in cyberspace, this same observation will certainly apply. So we must recognize people as an important component of cyberspace, just as we must recognize wires and protocols.

One of the reasons why the U.S. has led in the cyber-revolution is our enthusiasm for innovation and experiment, our willingness to risk failure to find gain, and our respect for risk-takers. This national trait should serve us well if advantage can be gained from "out-innovating" the enemy. The military has recognized in general the need for innovative thinking, rapid reaction, and delegation of responsibility along the chain of command. But this mode of thinking may not be as deeply associated in the military with cyberspace and IT, where the evolution of cyberspace may be seen as a procurement problem.

Changes overseas may shift this balance in significant ways. For example, the $100 laptop project (OLPC), if it could be successful in meeting a need for children in the developing world, would create, within 10 years, millions of military-age young adults that are fully conversant with the power of cyberspace tools. The current argument for the $100 laptop is centered in peacetime, and on social as well as economic motivations, but it can have implications as well for state confrontation. The $100 laptop, because it reveals more of the "platform/plasticity" dimension of cyberspace than (say) current cell-phones, may have more of an impact than if the cell-phone is the successful technology for the developing world.

All of these layers are important. As a specific example, if one wants to understand the security of cyberspace, one cannot focus on just one of these layers. Attacks can come at all layers, from destruction of physical components to compromise of logical elements to corruption of information to corruption of the people. So defense must similarly be based on an understanding of all these layers.

Coming and going

The term cyberspace implies that it is indeed a space of some sort, where people might *go*. It implies a volume, or locus, or destination. But this is definition by analogy, and may not always be the right image. In many cases, the “go to” image of cyberspace is quite accurate. Participants in multiplayer games “go to” their virtual world, and have a rich experience there—they establish a persona, make money, make friends, argue about governance, and have a rich sense of the “reality” of that virtual space.

But if cyberspace is a place to “go to”, it is also an experience that comes to us. To define by analogy again, the term “air power” implies that “air space” is a distinct place to which one goes, but at the same time, the surface of the land and sea is universally touched by the air. The air comes to us, and there is a powerful relationship between what happens in air space and what happens on the surface of the earth. One could argue about which is more important for air space—the unique nature of air space or its interface to the land and sea. The same is true of cyberspace. With the increasing deployment of sensors, RFID tags, embedded computers, and ubiquitous wireless networking, cyberspace may be as close to us as the air. The right image for cyberspace may be a thin veneer that is drawn over “real” space, rather than a separate space that one “goes to”.

Finding implications

Power and control

Layering is a traditional approach used by technologists to design (or to understand) a system by breaking it up into parts, and by controlling the dependencies among those parts. In a system that is strictly defined by layers (in a technical sense), the implication is that the upper layers build upon (or depend on) the lower layers, as a house sits on its foundation, but that the lower layers are not influenced by the upper layers. The functional dependency is from high to low, and not the other way around. In the layering proposed here, this relationship is much less precise. For example, people use information, but they create it.

When looking at a system like this from a social, rather than a technical point of view, a perhaps more interesting question is to ask about the locus of power and control implied by the design—is power centralized (and if so, to what actor) or diffuse? Does the design create points of control or avoid them? I will use an informal method I call *control point analysis* to try to capture and understand the power relationships created by specific design decisions in cyberspace—today’s or tomorrow’s.

Open platforms

The word “open” is often used but ill-defined. In general, it is a word with a strong positive implication (“open conversation”, “open expression”), and it implies the free ability to access, pass through or use. With respect to a platform, as in the basic service of the Internet, the term open is used to imply that anyone can exploit or build upon that platform (that service interface) without restriction, license or uncertain negotiation. The

Internet, both to it users and to the developers of applications and services on top of it, should be available to all comers. This issue (and the possibility that Internet Service Providers or ISPS might restrict usage and erode the open nature of the Internet) is at the root of the current debates about *network neutrality*.

The protocols and standards that define the Web are similarly "open", in that anyone can exploit those standards freely. Anyone can design a web page, anyone can start a web hosting service, and the like. The open nature of the Web protocols is complemented by the availability of "open source software" to build a web server, where the word "open" again means free to use by all without license or fee.

However, as one looks upward through the logical layer at the series of platforms that exist, not all are equally open. For example, Facebook is not just a web application, it is a platform for development of further applications by third parties. Application developers using the Facebook development platform are expected to conform to a rather long set of policies[3], and Facebook reserves the right, at its total discretion, to disable applications that violate these policies. At the present time, the policies seem to have the benign objectives of protecting the rights of the user (e.g. privacy), preserving the stability of the user experience, and preventing the mis-appropriation of Facebook branding. The policies do not seem to reflect a desire to exercise discrimination among applications based on business objectives. None the less, any developer building an application on top of Facebook must take into account that their use of this platform is at the sole discretion of Facebook. Policies could change.

In many parts of cyberspace today, there can be found tussles about "open" vs. "closed", about who controls a platform and whether that control is benign, whether it is predictable, and whether it is in the best interest of cyberspace generally, or in the interest of a controlling actor, e.g. a corporate entity. Platforms have a "tippy" quality, and successful platforms can achieve enough of a market share to give their owner some market power, as has been argued about the application development interface on the various Microsoft operating systems.

Control point analysis of the current Internet

Control point analysis, as I use the term here, is not a well-developed and rigorous methodology. It is a set of tools to help think in a rigorous way about the design of a system from a particular perspective—that of determining which actors obtain power by virtue of control over key components of the system. One can proceed in several ways, which complement each other.

- One can assemble a catalog of all the parts of the system, and make note of the pattern of control that applies to them.
- One can trace the steps of common actions (e.g., in the case of the Internet, retrieving a Web page), and ask at each step if one has encountered a significant point of control. This method can help identify points in the system that might have been overlooked in the initial catalog.

[3] See http://developers.facebook.com/policy/, visited December 30, 2009

- One can look at each of the actors in the ecosystem and ask what forms of control they exercise. This provides a third cut at the same set of issues.

In general, systems that are "open" seem to be associated with weak or decentralized control, and limited allocation of power to one actor. Systems that are more closed tend to allocate power to the owner of the system. (One must ask, of a system that is deemed to be open, whether this nature derives fundamentally from the manner of its creation, or the granting of "openness" by a creator might some day be revoked. For example, the original Ethernet technology was patented by Xerox. Xerox built a consortium with Digital Equipment and Intel, and made substantial declarations that they would allow the use of the patent free of licensing encumbrances, in order to have Ethernet made a standard by the IEEE.)

If we look at the major actors that make up the Internet, we see that different actors hold different points of control with different powers—if there is engagement or tussle among these actors using their powers, it is asymmetric engagement.

- Application designers, by their decisions, define the available patterns of communication, and what is revealed or concealed in the messages being communicated. With respect to encryption, which can occur at different levels in the system (link, end-to-end or application), only the application level can discriminate among different parts of the communicated information, encrypting some but revealing other parts. For example, the design of email protocols reveals the headers of the email, even if the body is encrypted. This decision actually reveals a great deal, but at the same time permits staged delivery, which in turn allows for the "out-sourcing" of virus checking and spam filtering.
- Users and their end-node computers control the initiation of activity. To the extent they have choice in the selection of service providers, they can use that choice as a discipline to select for trustworthy providers.
- Internet Service Providers (ISPs) build and operate regions of the network. Within their regions, they control topology and completion of connections. (e.g. who talks to whom under what circumstances.) There is no monolithic network, but different parts controlled by different ISPs that may be trusted (or not) to different extents. Examples of control include physical topology and routing (making sure traffic actually passes through a firewall or point of censorship). Networks (less abstractly, the devices in them) can see what is not encrypted and change what is not signed.
- The operating system designer provides the platform for execution of end-node software. While some platforms today are more open (Linux) and some are more controlled by their owners (e.g. Windows), most operating systems today are viewed as raising few fears of exercise of control.

Looking at a typical Web transaction yields another point of view about control points. The normal sequence of steps would be as follows.

- The user acquires a URL by some means, perhaps by typing it in or by clicking on a link in another web page. This action is carried out using "browser" software.

Today, there are open source browsers (Firefox), and more closed browser like Internet Explorer. There have been claims in the past that certain browsers would not allow the user to use certain URLs, but this does not seem to be a major concern today. The browser presents other options for control; see below.

- The DNS name in the URL must be translated into the IP address of a server, which requires the use of the Domain Name System, or DNS. The DNS is specified in open, IETF standards, and for a long time was not seen as an important point of control. However, with the emergence of DNS names as object of commercial value, and the contentious role of ICANN as an overseer of the set of acceptable top level domain (TLD) names, control of the DNS system has become a very important issue in the stability of the Internet. The DNS system itself is highly decentralized, with most ISPs operating a server for their clients. For this reason, each ISP (or other service provider, such as a hotel or hot-spot) has a very high degree of control over what address is returned in response to a lookup. Many mis-directions occur in practice today using this point of control, and these DNS servers have as well been the target of attackers, who install mis-directions of their own.
- Assuming that the DNS has returned an IP address, the browser opens a connection (in tech-speak, a TCP connection) to that address.
- The routers at each hop along the path to the server look up the preferred path toward that address and forward the packet. They thus have absolute control over whether the client can reach the server. If the router has no route computed to that server, no connection can occur. (This outcome does not normally arise, but can during transient outages.) More significantly, if the router deliberately mis-directs the packet, or changes the destination address in the packet, the packet will arrive at the wrong server[4].)
- If the web site uses secure protocols (signaled by the prefix HTTPS rather than HTTP at the beginning of the URL), the server will return to the browser a *certificate* attesting to the identity of the server. The certificate is "signed" or validated by one of a number of *certificate authorities* or CAs. There are a number of important points of control surrounding these CAs. Different CAs may exercise different degrees of care before signing a certificate. But more interesting, all browsers today have built into them a list of CAs that are deemed trustworthy, which gives the browser designers a significant measure of control[5].

[4] This sort of mis-direction may seem unlikely, but it (or a related mis-direction involving the DNS) occurs all the time. It is a common experience to open a browser window in a hotel or "hot-spot", and attempt a connection to some page, only to get a page instead inviting the user to pay a fee for access. This happens only because of some intentional, if mostly benign, mis-direction occurring within the hotel/hot-spot system.

[5] If a user encounters a certificate signed by a CA that is not included in the list included in the browser, as for example with certificates issued by MIT, strange error messages arise, and the user is instructed to take inexplicable actions, such as "downloading a new authority certificate". This outcome degrades usability.

The use of certificates can detect some of the forms of mis-direction that occur at the DNS or router level. That is, if the DNS or router have mis-directed the browser, this can be detected. The consequence is that the browser will raise some sort of alert or alarm to the user. However, most users have no idea what to make of these alarms, and often proceed to connect to the wrong server, to their peril.

- The web site then finds the stored content associated with the URL (or prepares the content dynamically if this is what is required) and returns this content to the browser over the network. The content may have "embedded content": URLs of further pages that are to be retrieved by the browser and displayed as part of the page. This process requires the browser to repeat all of the above steps for each of those embedded content links. Each of them may be subjected to mis-direction by the same points of control.
- It might seem that the site hosting the content has (as one would expect) ultimate control over the format and content of the page. However, for a number of reasons, this is not correct. First, some of the embedded content may be hosted on other servers. The most common example of this is advertizing. Advertizing raises specific risks aside from annoyance—since malware can be hidden in innocent-looking web pages, a web server that includes third-party ads on its web site must trust that the site generating the ads is trustworthy and has not been infiltrated. There is no way for the original web server to check. Second, since the ISPs along the path from the server to the browser control the routing, any of them can redirect the returning web page to an intermediate node that performs arbitrary modifications to it before sending it on to the browser. Unless secure connections are used (as described above), the power of the ISP to control routing gives it the ultimate control. Examples of modifications that have occurred today include reformatting the page to fit onto the small display of a mobile device (which seems somewhat benign) and finding embedded content URLs and replacing them—for example replacing the ads selected by the web server with ads selected by the ISP. This behavior is not normally considered benign.

This sort of control point analysis reveals that the Internet, although sometimes described by its creators as "simple", contains a rich mix of points of control, and a range of design principles that different actors use to "blunt the instruments of control" by other actors. Encrypting content is the most obvious example—an ISP cannot change what is signed, and cannot see what is encrypted. Other approaches used to blunt the controls of the ISPs include the use of Virtual Private Networks (VPNs), a common tool for travelers using ISPs, hotels and hot-spots they do not trust. A VPN does more than encrypt, it then sends all traffic from the client host back to a trustworthy relay point (e.g. at the travelers corporate headquarters), where the traffic is then injected into the Internet as if it originated there. The IP addresses that the client machine is trying to reach anr encrypted until they reach the trustworthy relay site, so intermediate untrustworthy routers cannot see them.

Looking to the future

The nature of cyberspace can be expected to evolve and mutate, driven by a complex interplay of forces. The forces of innovation would normally tend to drive rapid mutation.

However, successful platforms, by their nature, accumulate a constituency of users that depend on them, and this constituency becomes a limiter on the rate of innovation—if the platform changes in ways that generate costs for users without compensating benefits, users will complain or move away from the platform. None the less, over a period of years, we can expect even the Internet itself may mutate and evolve in ways that may change its platform quality, and thus change the distribution of power and control.

The U.S. National Science Foundation is funding a project called Future Internet Design, or FIND, which invites the research community to envision what the Internet of 15 years from now might be. Here are a number of ideas drawn from that community, and the future-looking network research community more generally.

Virtualization—an alternative to the present Internet

Today's Internet's service can be implemented over a wide range of network technologies, since it only requires the technology to carry strings of bytes as packets. If these technologies have other capabilities, the applications sitting on top of the IP layer cannot exploit them. The simplicity of the IP layer is both powerful (in terms of generality) and limiting (in terms of exploiting technology features). For example, radio is intrinsically a broadcast medium, but the broadcast feature cannot be exploited in the Internet. Fiber is capable of carrying a very high bandwidth analog signal with very low distortion, but this capability cannot be exploited using the Internet design.

A different approach, which might lead to a very different global network in 10 to 15 years, attempts to make the advanced features of any network technology directly available to the application. This approach is based on the concept of *virtualized resources.* Virtualization is a very familiar concept in computer science and electrical engineering. It is a set of techniques that start with one physical set of technology resources, and divide them up among multiple uses in a way that gives each use (or user) the illusion of having a separate and distinct copy of these resources, essentially identical except that the copies run slower than the actual technology base that is being virtualized. We know how to take a computer and add a layer of software that creates a number of *virtual machines.* We can divide up and virtualize circuits, memory, and so on. Now research seeks to take all the resources found in a global network, virtualize them all, and then give *virtual global networks* to different classes of users or uses.[6] There is no commitment to a common packet format, to packets at all, or to any of the other current conventions of the Internet. Different applications can make use of the resources in ways best suited to their needs: one set of users might run something like the Internet in one virtual global network, while another set of users might use a virtualized fiber-optic path to carry analog signals, and a third set of users might use a virtual global network for global dissemination of bulk data (broadcasting, in effect), with a totally

[6] Two proposals currently funded by the National Science Foundation (NSF) that explore this concept are Jon Turner, et al., "An Architecture for a Diversified Internet," <http://www.nets-find.net/DiversifiedInternet.php>; and [need FIRST NAMES please] Feamster, Rexford, and Gao, "CABO, Concurrent Architectures are Better than One," <http://www.nets-find.net/Cabo.php>.

different approach to packetization and routing. This approach, if successful, could lead to a network in which applications are able to exploit the native capabilities of current and future network technologies.

Control point analysis: Compared to today's Internet, this approach moves the point of agreement—the definition of the service platform—down, to the technology base. It demands of each technology, not that it be able to carry packets, but that it be able to "virtualize itself." In such a scheme, the owner/operator of the physical devices has much less control than an ISP does today. In fact, it may take away so much control that the facilities owner/operator is not actually motivated to enter into the business. The response might be that the virtualization layer would not end up being "open" in the way the Internet platform layer is, but much more closed, with the facility owner/operator entering into to selective/exclusive business relationships with higher-level providers. This business approach, however, is hindered by the fact that there would be hundreds or thousands of facilities owners around the globe, who would all have to enter into these business arrangements in common if a global network were to be built. This sort of bargaining would seem to call for a very different industry structure.

Thus, if the virtualization approach brings to the next higher-level platform designers a much greater range of generality, it also raises fundamental architectural questions about the division of function between the virtualization layer and the layer above it—questions about security, management, the economics of the industry, and so on. And it invites the question of whether the associated generality is actually more valuable than the generality of the Internet of today. But this concept, if it proves itself and moves out of the research community, could offer a very different future global network.

Future Architectural Concepts at a Higher Level

Current research may lead to higher-level services that might define an Internet of tomorrow. These services would involve not just simple point-to-point data carriage, but more complex functions that could better match the needs of sophisticated applications. I describe three examples of such research: support for an information architecture; support for the creation and management of services that depend on functions placed on servers distributed across the network; and support for communication among nodes and regions of the network that are connected only intermittently.

These examples of current research are important for two reasons. First, of course, they suggest the range of functions we may see in a future network system. They also raise questions about what lower-level functions a future Internet should support, if the purpose of those lower-level functions is not to support applications directly, but to support intermediate services such as these. It may be that some basic features of the present Internet, such as global end-node interconnectivity, will not be as important in a future where most machines only communicate via application-level servers. Such a shift would have major implications for security, resilience, locus of control, and generality.
A future network may be characterized by an architecture that is defined at this higher level of abstraction: closer to what the user is trying to do, and further away from moving bytes. We discuss three examples of current lines of research that illustrate this trend.

A future architecture for information

Some current research aims at an architecture for information as opposed to an architecture for byte carriage. Such research is based on the observation that what people care about is not devices, but content—information. People exchange content, and create it for others to use. The device (e.g. the computer) is a low-level aid to this task. Such a system would focus on such issues as efficient information naming, dissemination, retrieval and authenticity.

On the Internet today, the authenticity and provenance of an information object is validated by where it comes from. For example, in today's Internet, the only way a user can be sure that a Web page that appears to be from CNN is legitimate is to try to retrieve it from CNN. Once it is retrieved, there is no trustworthy information associated with the page itself that captures and conveys the provenance of the information. If one user downloads a page and then sends it on to a second person, there is no way for that second person to be sure it is not a forgery. If the validity of the information were associated with the information itself, perhaps using some sort of encryption scheme to sign the information, then we could use a much richer set of mechanisms for information dissemination. These might include peer-to-peer systems, third-party caching, archiving and forwarding, casual user-to-user transmission via e-mail, and so on. This richness would provide a network that was more diverse and efficient.

An architecture for information would presumably include some means to name information. Current Internet names (domain names) and addresses do not name information objects; rather they name physical end-points, that is, computers attached to the edge of the network. The Web provides one scheme for naming objects, the URL. There has been much research on other schemes for naming objects, ranging from IETF research on Universal Resource Names (URNs)[7] and centrally managed schemes such as the Handle scheme[8], to highly decentralized schemes that allow users to pick local names and share them.[9] All of these schemes imply some solution to the location problem: given a name for an object, can I find the object itself? Some solutions to the location problem would involve a centrally managed catalog of objects; some solutions would use a more decentralized approach, with location services run by different operators in different regions of the network; some schemes would be completely decentralized, based on a requestor broadcasting a request for an object until a copy is found. These different schemes have very different implications for control, resilience, performance, and utility.

The problem of *search* is distinct from the problem of *location.* The location problem is to find a copy of an object, once the particular object has been identified. The analog in

[7] See K. Sollins, ed., "*Architectural Principles of Uniform Resource Name Resolution*," RFC 2276, Internet Engineering Task Force (IETF), 1998. Also see <http://www.w3.org/TR/uri-clarification/> for a discussion of confusion over terms.

[8] See <http://www.handle.net/>.

[9] See, for example, <http://publications.csail.mit.edu/abstracts/abstracts06/baford/baford.html>.

the real world is to ask what library or bookstore has the closest copy of a specific book. The *search* problem is to look for objects that may match certain selection criteria, analogous to seeking a book on the history of baseball that costs less than $20. Google's approach for finding a good match to given search criteria, for example, is quite complex and sophisticated, which gives Google a high degree of control over what search results are returned, and in what order. Another problem that has a critical influence on search is the emergence of dynamic information, mentioned above, that is generated on demand for each user. It is not clear that current centralized search systems reliably or consistently index information that only exists when a particular user asks to retrieve it. There is ongoing research in the design of advanced search engines that could search or otherwise deal with dynamic content.

A specific example of an information architecture is Content-centric Networking[10]. This proposal makes the addresses in packets correspond to information elements, rather than machines. One sends a packet to a piece of information, asking that it be returned, and gets back a packet from that address containing the requested information.

This reformulation of what addresses means leads directly to new modes of dissemination and in-network storage, such as any-to-any dissemination, demand-driven buffering and directed diffusion distribution. Information is signed, so validity and provenance are explicit properties of the scheme. And since one does not receive information except when asking for it, one only receives information that is relevant. So certain security attributes are intrinsic to this scheme. The traditional service of the Internet today, conversations between explicit endpoints, can be derived as a tiny part of the total space.

Control point analysis: Depending on exactly how information architectures are designed, they can create (or not) new centralized points of control. Since naming and authentication are central to these schemes, the design of these mechanism (analogous to DNS and certificate authorities today) will have a critical influence over balance of power. Content-centric networking, in particular, by embedding the names of information into the packets, seems to give a considerable measure of control over information access to the owner/operator of the packet switching devices—the routers, which might seem to empower the ISPs at the expense of the other actors. Of course, one could implement this scheme on top of a virtualization scheme, which might lead to competing schemes for information dissemination. Alternative schemes for information dissemination are designed to sit on top of the current packet-level service similar to today's Internet, which would allow competing schemes to exist, and weaken the ISP's options for control .

The importance of this research to the future of cyberspace is that, depending on how such efforts succeed, we may be able in the future to place either much more confidence, or much less, in the validity, authenticity, completeness, and persistence of information in cyberspace. Fraud and forgery could destroy all important uses of on-line information or, at the other possible extreme, cyberspace could become the only important and relevant source of information.

[10] See http://www.ccnx.org/, visited 12/31/2009

An architecture for services and service construction

The architecture of the current Internet really recognizes only two sorts of devices: the devices that make up the Internet itself—the routers that forward packets — and everything else, including all devices attached to the edges of the network, between and among which the Internet packets are transported. The dominant examples of edge devices now are the PCs that belong to the users, and the servers that provide content to users. In the Internet architecture, there is no consideration of what these servers do, what patterns of communication connect them, nor of how applications actually exploit servers to build up their services.

Most applications today display a somewhat complex structure of servers and services to implement their functions. Content may be duplicated and pre-positioned on servers near the ultimate recipient to improve performance and resilience.[11] Communication among end-nodes is often relayed through intermediate servers to realize such functions as mutual assurance of identity, selective hiding of certain information, logging and archiving of communication, or insertion of advertising. Some servers, such as e-mail servers, hold information such as e-mail for recipients until they choose to connect to the network to retrieve it. This allows nodes to communicate even if they are not simultaneously connected and active.

The emergence (and recognition) of systems that explicitly recognize servers and services may signal a move of processing and control from the "edge" of the network—the computers of the end-users — toward the "center," locations operated by providers. Providers might include those who supply the basic packet carriage of the Internet, as well as specialty providers of specific application services. Both types of players may become important parts of the Internet experience of tomorrow.

Some proposals in this general category are motivated by the observation that a network architecture for tomorrow needs to take more explicit account of the various stake-holders that make up the system. Customers, service providers, application developers, governments, and so on have concerns and objectives. The current "narrow waist" of the Internet works fine as a way to create a useful data plane, but does not allow these various actors to establish policy—it masks what is going on inside the network, and does not provide the right set of management tools. This reality requires that we "re-factor" the basic functions of the network.

Control point analysis: At the moment, there is no evidence that the movement of function "toward the center" implies the emergence of new monopolists, but this is a factor to analyze and monitor, along with potential government exploitation of such centralization. Players at many layers of the Internet may have significant market power now or in the future, including providers of residential broadband service, providers of core software products such as Microsoft, or higher-layer providers that deal in information, such as Google. Areas that will warrant attention in the future may include advertising, marketing, and those who collect and manage information about consumers.

[11] This function is provided today by third-party providers such as Akamai.

At the same time, we can see the possibility of a future with increasing regional differences, and a possible Balkanization of the network at these higher layers. Even at the basic packet-carriage layer of the Internet, which is defined by a very simple standard that is "the same everywhere," there have been attempts to control the flow of information across regions of the Internet. These higher-layer services are more complex, with richer structure and more visible signals that may reveal the intentions of the users, and they provide many more options for points of control. The experience may differ depending on the user's location; many actors may try to shape and limit the user's experience.[12] Differentiation and fragmentation of users' experiences has many implications for balance of power.

One of the specific issues that arises in these more complex, server-based applications is that the physical location of the servers is relevant, and indeed sometimes critical, to the effective operation of the application. If the purpose of the server is to pre-position content close to the user to increase performance and reliability, then the server, because it is physically close to the user, may be more likely to be under the same legal or governmental jurisdiction within which the user resides. It may be that services that are designed for performance and for resilience in the face of network failures will have a quite different design than services that position the content at a distance from the user in order to avoid the imposition of local controls and Balkanization.[13] Research could help to enable such geographic independence. A continuing tension between technical mechanisms and legal or other nontechnical mechanisms is likely, given competing interests in disseminating or restricting specific kinds of information in specific locales.

An architecture for relayed delivery of content

A central assumption of the current Internet architecture is that communication between pairs of nodes is interactive in real time, and since any pairs of nodes may want to establish such interactive communication, the network provides this capability universally. The dominant protocol for data transfer on the Internet, the Transmission Control Protocol (TCP), is based on the immediate confirmation back to the sender of the delivery of each packet. But if applications can define restricted patterns of communication involving intermediate servers, delivery from the original source to the ultimate destination is not necessarily immediate but may instead be delayed, staged, or relayed. This pattern of communication seems to be common to many sensor applications and many information dissemination applications, so it is posited that this pattern should also be recognized as a part of the core Internet architecture. A current line of research involves the investigation of DTNs, Delay (or Disruption) Tolerant Networking, which

[12] The Open Net Initiative at http://www.opennetinitiative.org/ is one of several organizations tracking current content controls on the Internet.

[13] See examples noted in Reporters Without Borders, *Handbook for Bloggers and Cyber-Dissidents*, September 14, 2005, <http://www.rsf.org/rubrique.php3?id_rubrique=542>.

generalize this concept of "store and forward" networking (in contrast to "end-to-end interactive" networking).[14]

Control point analysis: DTNs raise many important questions about security (how much must we trust the intermediate nodes), resilience (how the service might recover from failed intermediates), routing (who has control over which intermediates are used) and assurance of delivery (whether the design provides confirmation of delivery that can be relayed back from recipient to sender?). At the same time, they allow the network designer great freedom in dealing with challenging contexts such as poor and intermittent connectivity or hosts with intermittent duty cycles, and they allow the application designer to hand off part of the design problem to the network.

A Long-term Outcome: Revolutionary Integration of New Architecture Ideas

Above we described an alternative network design based on *virtualization,* in which the basic network service is not packet carriage, but access to the network technology at a lower-level and more technology-specific way. This lower-layer idea gains importance when it is combined with a higher-level architecture for services and delay-tolerant delivery. If information is being delivered in stages, not directly from source to destination, then the details of how the data is transmitted can be different in each stage. The requirement for a uniform and universal commitment to a single packet modality, for example, is much reduced if different parts of the network talk to each other only via a higher-level server. The result, in 10 or 15 years, might blend all of these ideas together in a global network which, at a lower level, is based on virtualized network technology, and at a higher level, on application-aware servers that connect together different parts of the network in ways that are matched to the features of the technology in that region of the network.

Conclusions

The control point analysis of these various future schemes is necessarily sketchy, since the schemes themselves are at an early stage of development. The specific details are less important than the summary observation that the network research community is exploring a number of approaches to networking that would shift the points of control and reallocate power among the different actors that make up the network ecosystem.

The catalog of actors in this discussion has focused on those most directly involved with the creation and operation of cyberspace—the ISPs, the designers of applications, services and content, the users and so on. Tlhis discussion has not addressed the actors that sit a bit removed, but which have substantial concerns about the shape of cyberspace, most obviously governments, and as well large industries that find themselves being

[14] Applications being explored range from outer space to the battlefield. For pointers to current research, see http://www.dtnrg.org/wiki and http://www.darpa.mil/sto/solicitations/DTN/.

influenced or reshaped by the cyberspace experience. A detailed discussion of this larger sweep of actors must be left for another document, but a few words may be in order.

As we look across nations to see sources of national variation, one can look at the different layers for examples. Even though the physical layer is (in many countries) a private-sector artifact, governments can influence national differences is at the physical foundations of cyberspace. National regulatory policy (usually the domain of the civilian rather than military parts of the government) can have a strong influence over the nature of cyberspace, especially at the physical level. In the days when ATT was the single regulated monopoly provider of telecommunications, the government could negotiate with it to establish a level of investment that achieved a desired level of resilience and survivability. The options for this sort of negotiation (and the levels of resilience of the "old" phone system) have eroded with the regulatory shift toward competition rather than regulation as the tool to shape our infrastructure. Different countries may take very different approaches in this space.

The use of open standards and COTS technology at the logical (platform/plasticity) level might seem to imply a universal commonality to cyberspace at this level. But this conclusion only applies only to the extent we ignore the nature of the people that deploy, use and shape that layer.

So any analysis of national variation and governmental control must look at all the layers of cyberspace to find points where strong differentiation suggests the opportunities for action.

A final comment about national variation is to note the tendencies of certain nations to draw national boundaries more strongly in cyberspace. The French decision to harden national boundaries with respect to information (regulation of Nazi memorabilia) is one example, but is it should not be seen as an isolated event. This is a trend to watch. In any case, it suggests an analysis of the extent to which governments can seize control over the points of control I have cataloged above. A point of control that can be removed altogether from the system cannot be appropriated by a government or any other actor. In the U.S, our traditional view is that open access to information will encourage movement (cultural, social, etc.) in "good" directions. National boundaries seem to oppose this. So we have favored a cyberspace with very weak manifestations of national boundaries. Should this be a consistent national policy, or are there countervailing considerations?

A historical note

I observed above that different decisions, especially at the logical level, can greatly affect the nature of cyberspace, and that different designs for cyberspace can have major implications for the balance of power among the various interested actors. This consequence may not be obvious to all network designers, but it has been very clear, at least to some. In the 1970's, there was a substantial debate between advocates of two sorts of network, called "datagram" and "virtual circuit". Datagram networks have a simpler core, with more functions shifted to the hosts at the edge. Virtual circuit network have more function in the core of the net, and thus more power and control shifted to the network operator. The Internet is a datagram network; the ARPAnet was more a virtual circuit network, and the data network standard developed by the telephone industry, Asynchronous Transfer Mode, or ATM, is a virtual circuit network.

One of the vocal advocates of the datagram approach was Louis Pouzin, who was building a datagram network called Cyclades in France at the same time that the Internet was being first built. In 1976, he published a paper with the following conclusion[15]:

> *The controversy DG vs. VC in public packet networks should be placed in its proper context.*
>
> *First, it is a technical issue, where each side has arguments. It is hard to tell objectively what a balanced opinion should be, since there is no unbiased expert. This paper argues in favor of DG's, but the author does not pretend being unbiased. Even if no compromise could be found, the implications would be limited to some additional cost in hardware and software at the network interface. So much resources are already wasted in computing and communications that the end result may not be affected dramatically.*
>
> *Second, the political significance of the controversy is much more fundamental, as it signals ambushes in a power struggle between carriers and computer industry. Everyone knows that in the end, it means IBM vs. Telecommunications, through mercenaries. It may be tempting for some governments to let their carrier monopolize the data processing market, as a way to control IBM. What may happen, is that they fail in checking IBM but succeed in destroying smaller industries. Another possible outcome is underdevelopment, as for the telephone. It looks as if we may need some sort of peacemaker to draw up boundary lines before we call get it trouble.*

In contrast to the Internet, Pouzin's Cyclades network was not ultimately successful. Its failure is often (if speculatively) attributed to the hostility and resistance of the French PTT.

[15] Pouzin, L. 1976. Virtual circuits vs. datagrams: technical and political problems. In *Proceedings of the June 7-10, 1976, National Computer Conference and Exposition* (New York, New York, June 07 - 10, 1976). AFIPS '76. ACM, New York, NY, 483-494. DOI= http://doi.acm.org/10.1145/1499799.1499870