Internet & Cyberspace Computing

(Development, Elements, Uses & Applications)

Antonia Race
Makaila Romeo

First Edition, 2012

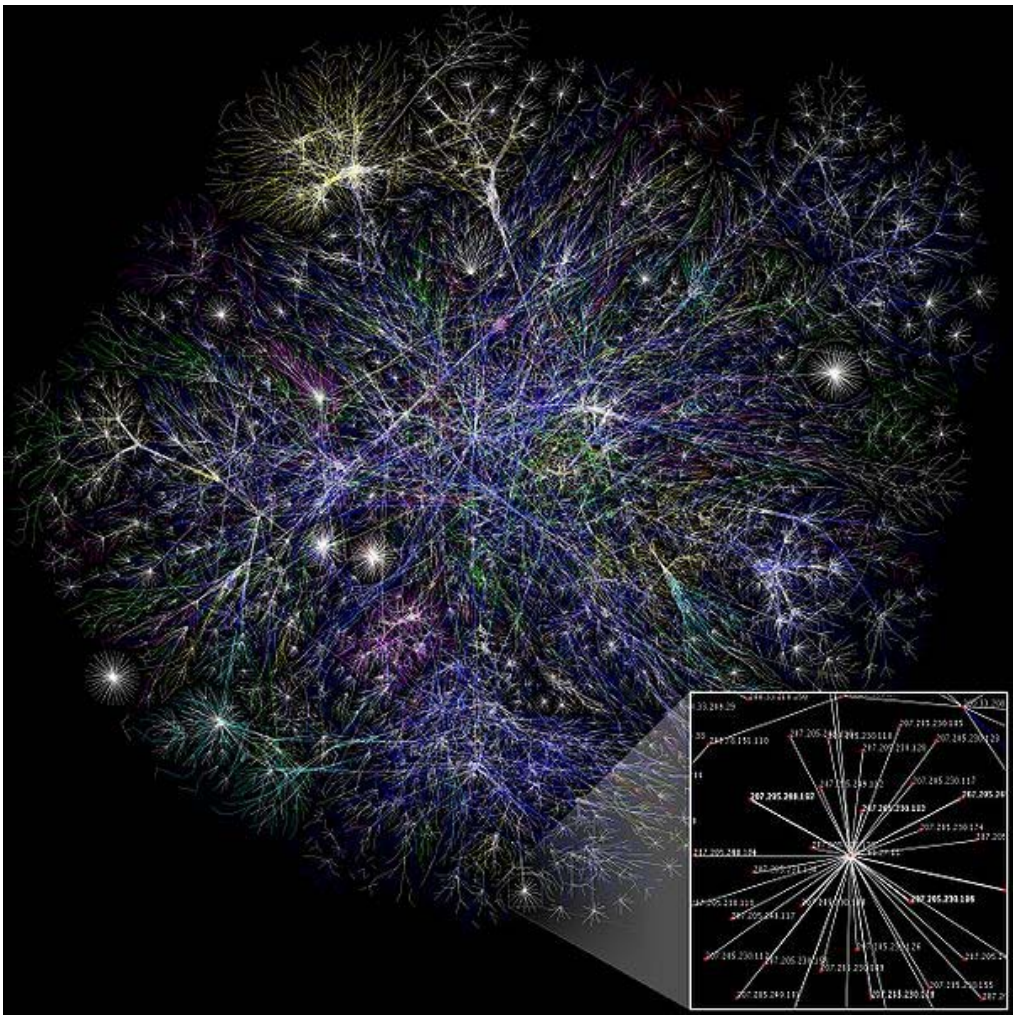
ISBN 978-81-323-1281-9

Published by:
College Publishing House
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter 1

Introduction to Internet



Visualization of the various routes through a portion of the Internet. From 'The Opte Project'

The **Internet** is a global system of interconnected computer networks that use the standard Internet Protocol Suite (TCP/IP) to serve billions of users worldwide. It is a *network of networks* that consists of millions of private, public, academic, business, and

government networks, of local to global scope, that are linked by a broad array of electronic and optical networking technologies. The Internet carries a vast range of information resources and services, such as the inter-linked hypertext documents of the World Wide Web (WWW) and the infrastructure to support electronic mail.

Most traditional communications media including telephone, music, film, and television are being reshaped or redefined by the Internet. Newspaper, book and other print publishing are having to adapt to Web sites and blogging. The Internet has enabled or accelerated new forms of human interactions through instant messaging, Internet forums, and social networking. Online shopping has boomed both for major retail outlets and small artisans and traders. Business-to-business and financial services on the Internet affect supply chains across entire industries.

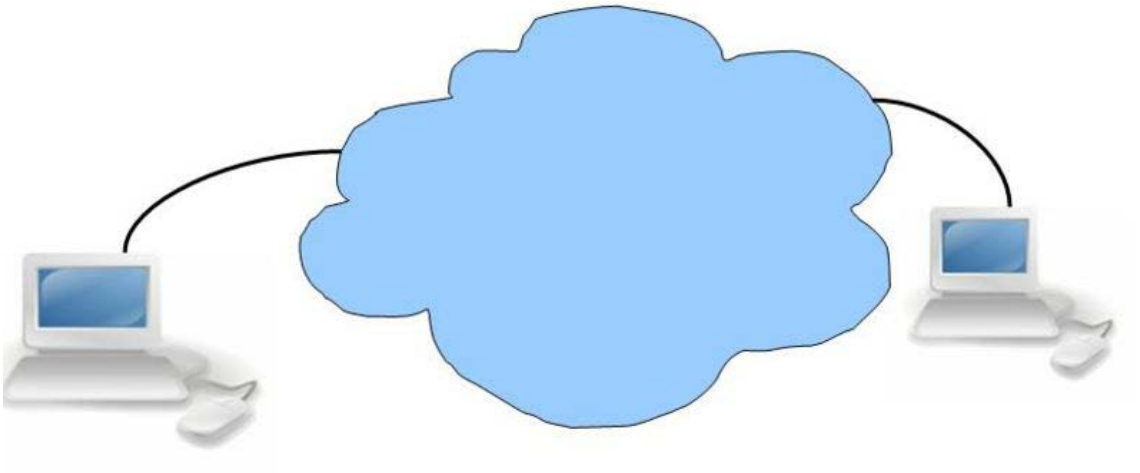
The origins of the Internet reach back to the 1960s with both private and United States military research into robust, fault-tolerant, and distributed computer networks. The funding of a new U.S. backbone by the National Science Foundation, as well as private funding for other commercial backbones, led to worldwide participation in the development of new networking technologies, and the merger of many networks. The commercialization of what was by then an international network in the mid 1990s resulted in its popularization and incorporation into virtually every aspect of modern human life. As of 2009, an estimated quarter of Earth's population used the services of the Internet.

The Internet has no centralized governance in either technological implementation or policies for access and usage; each constituent network sets its own standards. Only the overreaching definitions of the two principal name spaces in the Internet, the Internet Protocol address space and the Domain Name System, are directed by a maintainer organization, the Internet Corporation for Assigned Names and Numbers (ICANN). The technical underpinning and standardization of the core protocols (IPv4 and IPv6) is an activity of the Internet Engineering Task Force (IETF), a non-profit organization of loosely affiliated international participants that anyone may associate with by contributing technical expertise.

Terminology

Internet is a short form of the technical term internetwork, the result of interconnecting computer networks with special gateways (routers). The Internet is also often referred to as *the net*.

The term *the Internet*, when referring to the entire global system of IP networks, has traditionally been treated as a proper noun and written with an initial capital letter. In the media and popular culture a trend has developed to regard it as a generic term or common noun and thus write it as "the internet", without capitalization.



Depiction of the Internet as a *cloud* in network diagrams

The terms *Internet* and *World Wide Web* are often used in everyday speech without much distinction. However, the Internet and the World Wide Web are not one and the same. The Internet is a global data communications system. It is a hardware and software infrastructure that provides connectivity between computers. In contrast, the Web is one of the services communicated via the Internet. It is a collection of interconnected documents and other resources, linked by hyperlinks and URLs.

In many technical illustrations when the precise location or interrelation of Internet resources is not important, extended networks such as the Internet are often depicted as a cloud. The verbal image has been formalized in the newer concept of cloud computing.

Technology

Structure

The Internet structure and its usage characteristics have been studied extensively. It has been determined that both the Internet IP routing structure and hypertext links of the World Wide Web are examples of scale-free networks. Similar to the way the commercial Internet providers connect via Internet exchange points, research networks tend to interconnect into large subnetworks such as GEANT, GLORIAD, Internet2 (successor of the Abilene Network), and the UK's national research and education network JANET. These in turn are built around smaller networks.

Many computer scientists describe the Internet as a "prime example of a large-scale, highly engineered, yet highly complex system". The Internet is extremely heterogeneous; for instance, data transfer rates and physical characteristics of connections vary widely.

The Internet exhibits "emergent phenomena" that depend on its large-scale organization. For example, data transfer rates exhibit temporal self-similarity. The principles of the routing and addressing methods for traffic in the Internet reach back to their origins the 1960s when the eventual scale and popularity of the network could not be anticipated. Thus, the possibility of developing alternative structures is investigated.

Modern uses

The Internet is allowing greater flexibility in working hours and location, especially with the spread of unmetered high-speed connections and web applications.

The Internet can now be accessed almost anywhere by numerous means, especially through mobile Internet devices. Mobile phones, datacards, handheld game consoles and cellular routers allow users to connect to the Internet from anywhere there is a wireless network supporting that device's technology. Within the limitations imposed by small screens and other limited facilities of such pocket-sized devices, services of the Internet, including email and the web, may be available. Service providers may restrict the services offered and wireless data transmission charges may be significantly higher than other access methods.

Educational material at all levels from pre-school to post-doctoral is available from websites. Examples range from CBeebies, through school and high-school revision guides, virtual universities, to access to top-end scholarly literature through the likes of Google Scholar. In distance education, help with homework and other assignments, self-guided learning, whiling away spare time, or just looking up more detail on an interesting fact, it has never been easier for people to access educational information at any level from anywhere. The Internet in general and the World Wide Web in particular are important enablers of both formal and informal education.

The low cost and nearly instantaneous sharing of ideas, knowledge, and skills has made collaborative work dramatically easier, with the help of collaborative software. Not only can a group cheaply communicate and share ideas, but the wide reach of the Internet allows such groups to easily form in the first place. An example of this is the free software movement, which has produced, among other programs, Linux, Mozilla Firefox, and OpenOffice.org. Internet "chat", whether in the form of IRC chat rooms or channels, or via instant messaging systems, allow colleagues to stay in touch in a very convenient way when working at their computers during the day. Messages can be exchanged even more quickly and conveniently than via e-mail. Extensions to these systems may allow files to be exchanged, "whiteboard" drawings to be shared or voice and video contact between team members.

Version control systems allow collaborating teams to work on shared sets of documents without either accidentally overwriting each other's work or having members wait until they get "sent" documents to be able to make their contributions. Business and project teams can share calendars as well as documents and other information. Such collaboration occurs in a wide variety of areas including scientific research, software

development, conference planning, political activism and creative writing. Social and political collaboration is also becoming more widespread as both Internet access and computer literacy grow. From the flash mob 'events' of the early 2000s to the use of social networking in the 2009 Iranian election protests, the Internet allows people to work together more effectively and in many more ways than was possible without it.

The Internet allows computer users to remotely access other computers and information stores easily, wherever they may be across the world. They may do this with or without the use of security, authentication and encryption technologies, depending on the requirements. This is encouraging new ways of working from home, collaboration and information sharing in many industries. An accountant sitting at home can audit the books of a company based in another country, on a server situated in a third country that is remotely maintained by IT specialists in a fourth. These accounts could have been created by home-working bookkeepers, in other remote locations, based on information e-mailed to them from offices all over the world. Some of these things were possible before the widespread use of the Internet, but the cost of private leased lines would have made many of them infeasible in practice. An office worker away from their desk, perhaps on the other side of the world on a business trip or a holiday, can open a remote desktop session into his normal office PC using a secure Virtual Private Network (VPN) connection via the Internet. This gives the worker complete access to all of his or her normal files and data, including e-mail and other applications, while away from the office. This concept has been referred to among system administrators as the Virtual Private Nightmare, because it extends the secure perimeter of a corporate network into its employees' homes.

Services

Information

Many people use the terms *Internet* and *World Wide Web*, or just the *Web*, interchangeably, but the two terms are not synonymous. The World Wide Web is a global set of documents, images and other resources, logically interrelated by hyperlinks and referenced with Uniform Resource Identifiers (URIs). URIs allow providers to symbolically identify services and clients to locate and address web servers, file servers, and other databases that store documents and provide resources and access them using the Hypertext Transfer Protocol (HTTP), the primary carrier protocol of the Web. HTTP is only one of the hundreds of communication protocols used on the Internet. Web services may also use HTTP to allow software systems to communicate in order to share and exchange business logic and data.

World Wide Web browser software, such as Microsoft's Internet Explorer, Mozilla Firefox, Opera, Apple's Safari, and Google Chrome, let users navigate from one web page to another via hyperlinks embedded in the documents. These documents may also contain any combination of computer data, including graphics, sounds, text, video, multimedia and interactive content including games, office applications and scientific demonstrations. Through keyword-driven Internet research using search engines like

Yahoo! and Google, users worldwide have easy, instant access to a vast and diverse amount of online information. Compared to printed encyclopedias and traditional libraries, the World Wide Web has enabled the decentralization of information.

The Web has also enabled individuals and organizations to publish ideas and information to a potentially large audience online at greatly reduced expense and time delay. Publishing a web page, a blog, or building a website involves little initial cost and many cost-free services are available. Publishing and maintaining large, professional web sites with attractive, diverse and up-to-date information is still a difficult and expensive proposition, however. Many individuals and some companies and groups use *web logs* or blogs, which are largely used as easily updatable online diaries. Some commercial organizations encourage staff to communicate advice in their areas of specialization in the hope that visitors will be impressed by the expert knowledge and free information, and be attracted to the corporation as a result. One example of this practice is Microsoft, whose product developers publish their personal blogs in order to pique the public's interest in their work. Collections of personal web pages published by large service providers remain popular, and have become increasingly sophisticated. Whereas operations such as Angelfire and GeoCities have existed since the early days of the Web, newer offerings from, for example, Facebook and MySpace currently have large followings. These operations often brand themselves as social network services rather than simply as web page hosts.

Advertising on popular web pages can be lucrative, and e-commerce or the sale of products and services directly via the Web continues to grow.

When the Web began in the 1990s, a typical web page was stored in completed form on a web server, formatted with HTML, ready to be sent to a user's browser in response to a request. Over time, the process of creating and serving web pages has become more automated and more dynamic. Contributors to these systems, who may be paid staff, members of a club or other organization or members of the public, fill underlying databases with content using editing pages designed for that purpose, while casual visitors view and read this content in its final HTML form. There may or may not be editorial, approval and security systems built into the process of taking newly entered content and making it available to the target visitors.

Communication

E-mail is an important communications service available on the Internet. The concept of sending electronic text messages between parties in a way analogous to mailing letters or memos predates the creation of the Internet. Today it can be important to distinguish between internet and internal e-mail systems. Internet e-mail may travel and be stored unencrypted on many other networks and machines out of both the sender's and the recipient's control. During this time it is quite possible for the content to be read and even tampered with by third parties, if anyone considers it important enough. Purely internal or intranet mail systems, where the information never leaves the corporate or organization's network, are much more secure, although in any organization there will be IT and other

personnel whose job may involve monitoring, and occasionally accessing, the e-mail of other employees not addressed to them. Pictures, documents and other files can be sent as e-mail attachments. E-mails can be cc-ed to multiple e-mail addresses.

Internet telephony is another common communications service made possible by the creation of the Internet. VoIP stands for Voice-over-Internet Protocol, referring to the protocol that underlies all Internet communication. The idea began in the early 1990s with walkie-talkie-like voice applications for personal computers. In recent years many VoIP systems have become as easy to use and as convenient as a normal telephone. The benefit is that, as the Internet carries the voice traffic, VoIP can be free or cost much less than a traditional telephone call, especially over long distances and especially for those with always-on Internet connections such as cable or ADSL. VoIP is maturing into a competitive alternative to traditional telephone service. Interoperability between different providers has improved and the ability to call or receive a call from a traditional telephone is available. Simple, inexpensive VoIP network adapters are available that eliminate the need for a personal computer.

Voice quality can still vary from call to call but is often equal to and can even exceed that of traditional calls. Remaining problems for VoIP include emergency telephone number dialing and reliability. Currently, a few VoIP providers provide an emergency service, but it is not universally available. Traditional phones are line-powered and operate during a power failure; VoIP does not do so without a backup power source for the phone equipment and the Internet access devices. VoIP has also become increasingly popular for gaming applications, as a form of communication between players. Popular VoIP clients for gaming include Ventrilo and Teamspeak. Wii, PlayStation 3, and Xbox 360 also offer VoIP chat features.

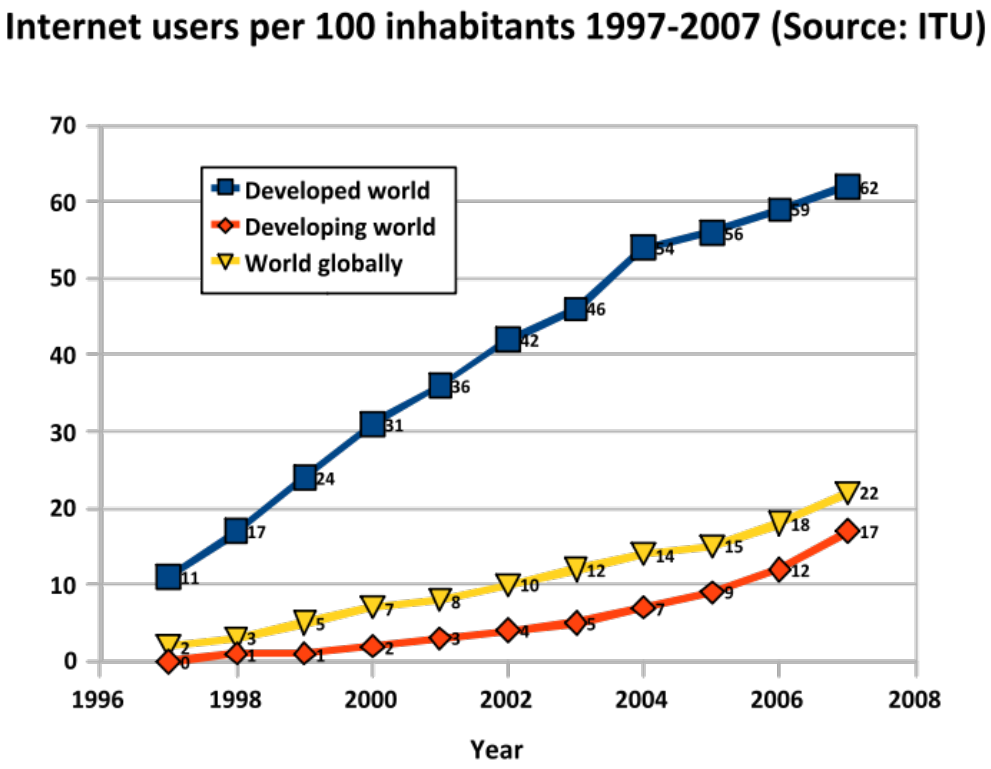
Data transfer

File sharing is an example of transferring large amounts of data across the Internet. A computer file can be e-mailed to customers, colleagues and friends as an attachment. It can be uploaded to a website or FTP server for easy download by others. It can be put into a "shared location" or onto a file server for instant use by colleagues. The load of bulk downloads to many users can be eased by the use of "mirror" servers or peer-to-peer networks. In any of these cases, access to the file may be controlled by user authentication, the transit of the file over the Internet may be obscured by encryption, and money may change hands for access to the file. The price can be paid by the remote charging of funds from, for example, a credit card whose details are also passed—usually fully encrypted—across the Internet. The origin and authenticity of the file received may be checked by digital signatures or by MD5 or other message digests. These simple features of the Internet, over a worldwide basis, are changing the production, sale, and distribution of anything that can be reduced to a computer file for transmission. This includes all manner of print publications, software products, news, music, film, video, photography, graphics and the other arts. This in turn has caused seismic shifts in each of the existing industries that previously controlled the production and distribution of these products.

Streaming media refers to the act that many existing radio and television broadcasters promote Internet "feeds" of their live audio and video streams (for example, the BBC). They may also allow time-shift viewing or listening such as Preview, Classic Clips and Listen Again features. These providers have been joined by a range of pure Internet "broadcasters" who never had on-air licenses. This means that an Internet-connected device, such as a computer or something more specific, can be used to access on-line media in much the same way as was previously possible only with a television or radio receiver. The range of available types of content is much wider, from specialized technical webcasts to on-demand popular multimedia services. Podcasting is a variation on this theme, where—usually audio—material is downloaded and played back on a computer or shifted to a portable media player to be listened to on the move. These techniques using simple equipment allow anybody, with little censorship or licensing control, to broadcast audio-visual material worldwide.

Webcams can be seen as an even lower-budget extension of this phenomenon. While some webcams can give full-frame-rate video, the picture is usually either small or updates slowly. Internet users can watch animals around an African waterhole, ships in the Panama Canal, traffic at a local roundabout or monitor their own premises, live and in real time. Video chat rooms and video conferencing are also popular with many uses being found for personal webcams, with and without two-way sound. YouTube was founded on 15 February 2005 and is now the leading website for free streaming video with a vast number of users. It uses a flash-based web player to stream and show video files. Registered users may upload an unlimited amount of video and build their own personal profile. YouTube claims that its users watch hundreds of millions, and upload hundreds of thousands of videos daily.

Access



Graph of Internet users per 100 inhabitants between 1997 and 2007 by International Telecommunication Union

The prevalent language for communication on the Internet is English. This may be a result of the origin of the Internet, as well as English's role as a lingua franca. It may also be related to the poor capability of early computers, largely originating in the United States, to handle characters other than those in the English variant of the Latin alphabet. After English (28% of Web visitors) the most requested languages on the World Wide Web are Chinese (23%), Spanish (8%), Japanese (5%), Portuguese and German (4% each), Arabic, French and Russian (3% each), and Korean (2%). By region, 42% of the world's Internet users are based in Asia, 24% in Europe, 14% in North America, 10% in Latin America and the Caribbean taken together, 5% in Africa, 3% in the Middle East and 1% in Australia/Oceania. In Asia, South Korea has the biggest internet penetration with 81.1% users (as comparison Japan with 78.2% and USA with 77.3%). The Internet's technologies have developed enough in recent years, especially in the use of Unicode, that good facilities are available for development and communication in the world's widely used languages. However, some glitches such as *mojibake* (incorrect display of some languages' characters) still remain.

Common methods of Internet access in homes include dial-up, landline broadband (over coaxial cable, fiber optic or copper wires), Wi-Fi, satellite and 3G technology cell

phones. Public places to use the Internet include libraries and Internet cafes, where computers with Internet connections are available. There are also Internet access points in many public places such as airport halls and coffee shops, in some cases just for brief use while standing. Various terms are used, such as "public Internet kiosk", "public access terminal", and "Web payphone". Many hotels now also have public terminals, though these are usually fee-based. These terminals are widely accessed for various usage like ticket booking, bank deposit, online payment etc. Wi-Fi provides wireless access to computer networks, and therefore can do so to the Internet itself. Hotspots providing such access include Wi-Fi cafes, where would-be users need to bring their own wireless-enabled devices such as a laptop or PDA. These services may be free to all, free to customers only, or fee-based. A hotspot need not be limited to a confined location. A whole campus or park, or even an entire city can be enabled. Grassroots efforts have led to wireless community networks. Commercial Wi-Fi services covering large city areas are in place in London, Vienna, Toronto, San Francisco, Philadelphia, Chicago and Pittsburgh. The Internet can then be accessed from such places as a park bench. Apart from Wi-Fi, there have been experiments with proprietary mobile wireless networks like Ricochet, various high-speed data services over cellular phone networks, and fixed wireless services. High-end mobile phones such as smartphones generally come with Internet access through the phone network. Web browsers such as Opera are available on these advanced handsets, which can also run a wide variety of other Internet software. More mobile phones have Internet access than PCs, though this is not as widely used. An Internet access provider and protocol matrix differentiates the methods used to get online.

Social impact

The Internet has enabled entirely new forms of social interaction, activities, and organizing, thanks to its basic features such as widespread usability and access. Social networking websites such as Facebook, Twitter and MySpace have created new ways to socialize and interact. Users of these sites are able to add a wide variety of information to pages, to pursue common interests, and to connect with others. It is also possible to find existing acquaintances, to allow communication among existing groups of people. Sites like LinkedIn foster commercial and business connections. YouTube and Flickr specialize in users' videos and photographs.

In the first decade of the 21st century the first generation is raised with widespread availability of Internet connectivity, bringing consequences and concerns in areas such as personal privacy and identity, and distribution of copyrighted materials. These "digital natives" face a variety of challenges that were not present for prior generations.

The Internet has achieved new relevance as a political tool, leading to Internet censorship by some states. The presidential campaign of Howard Dean in 2004 in the United States was notable for its success in soliciting donation via the Internet. Many political groups use the Internet to achieve a new method of organizing in order to carry out their mission, having given rise to Internet activism. Some governments, such as those of Iran, North Korea, Myanmar, the People's Republic of China, and Saudi Arabia, restrict what people in their countries can access on the Internet, especially political and religious content.

This is accomplished through software that filters domains and content so that they may not be easily accessed or obtained without elaborate circumvention.

In Norway, Denmark, Finland and Sweden, major Internet service providers have voluntarily, possibly to avoid such an arrangement being turned into law, agreed to restrict access to sites listed by authorities. While this list of forbidden URLs is only supposed to contain addresses of known child pornography sites, the content of the list is secret. Many countries, including the United States, have enacted laws against the possession or distribution of certain material, such as child pornography, via the Internet, but do not mandate filtering software. There are many free and commercially available software programs, called content-control software, with which a user can choose to block offensive websites on individual computers or networks, in order to limit a child's access to pornographic materials or depiction of violence.

The Internet has been a major outlet for leisure activity since its inception, with entertaining social experiments such as MUDs and MOOs being conducted on university servers, and humor-related Usenet groups receiving much traffic. Today, many Internet forums have sections devoted to games and funny videos; short cartoons in the form of Flash movies are also popular. Over 6 million people use blogs or message boards as a means of communication and for the sharing of ideas. The pornography and gambling industries have taken advantage of the World Wide Web, and often provide a significant source of advertising revenue for other websites. Although many governments have attempted to restrict both industries' use of the Internet, this has generally failed to stop their widespread popularity.

One main area of leisure activity on the Internet is multiplayer gaming. This form of recreation creates communities, where people of all ages and origins enjoy the fast-paced world of multiplayer games. These range from MMORPG to first-person shooters, from role-playing games to online gambling. This has revolutionized the way many people interact while spending their free time on the Internet. While online gaming has been around since the 1970s, modern modes of online gaming began with subscription services such as GameSpy and MPlayer. Non-subscribers were limited to certain types of game play or certain games. Many people use the Internet to access and download music, movies and other works for their enjoyment and relaxation. Free and fee-based services exist for all of these activities, using centralized servers and distributed peer-to-peer technologies. Some of these sources exercise more care with respect to the original artists' copyrights than others.

Many people use the World Wide Web to access news, weather and sports reports, to plan and book vacations and to find out more about their interests. People use chat, messaging and e-mail to make and stay in touch with friends worldwide, sometimes in the same way as some previously had pen pals. The Internet has seen a growing number of Web desktops, where users can access their files and settings via the Internet.

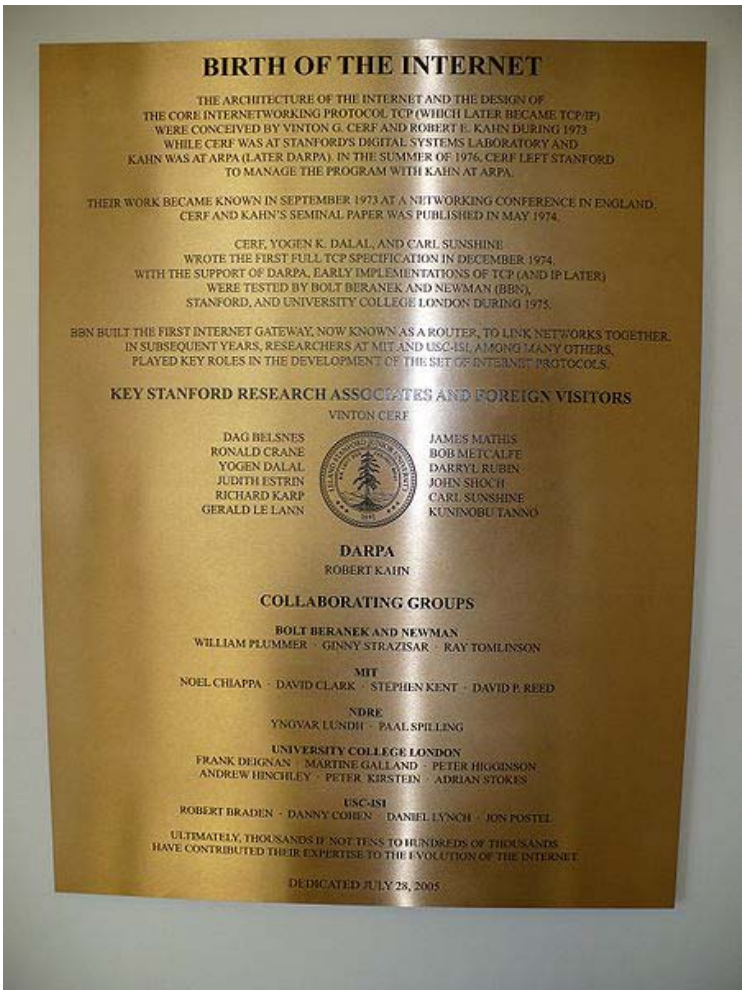
Cyberslacking can become a drain on corporate resources; the average UK employee spent 57 minutes a day surfing the Web while at work, according to a 2003 study by

Peninsula Business Services. Internet addiction disorder is excessive computer use that interferes with daily life. Some psychologists believe that Internet use has other effects on individuals for instance interfering with the deep thinking that leads to true creativity.

Internet usage has also shown a strong connection to loneliness. Lonely people tend to use the internet for an outlet for their feelings and to share their stories with other lonely people, such as in the "I am lonely will anyone speak to me" thread.

Chapter 2

History of the Internet



Commemorative plaque listing some of the early Internet pioneers

The concept of data communication - transmitting data between two different places, connected via some kind of electromagnetic medium, such as radio or an electrical wire - actually predates the introduction of the first computers. Such communication systems were typically limited to point to point communication between two end devices. Telegraph systems and telex machines can be considered early precursors of this kind of

communication. The earlier computers used the technology available at the time to allow communication between the central processing unit and remote terminals. As the technology evolved new systems were devised to allow communication over longer distances (for terminals) or with higher speed (for interconnection of local devices) that were necessary for the mainframe computer model. Using these technologies it was possible to exchange data (such as files) between remote computers. However, the point to point communication model was limited, as it did not allow for direct communication between any two arbitrary systems; a physical link was necessary. The technology was also deemed as inherently unsafe for strategic and military use, because there were no alternative paths for the communication in case of an enemy attack.

As a response, several research programs started to explore and articulate principles of communications between physically separate systems, leading to the development of the packet switching model of digital networking. These research efforts included those of the laboratories of Vinton G. Cerf at Stanford University, Donald Davies (NPL), Paul Baran (RAND Corporation), and Leonard Kleinrock at MIT and at UCLA. The research led to the development of several packet-switched networking solutions in the late 1960s and 1970s, including ARPANET, Telenet, and the X.25 protocols. Additionally, public access and hobbyist networking systems grew in popularity, including *unix-to-unix copy* (UUCP) and FidoNet. They were however still disjointed separate networks, served only by limited gateways between networks. This led to the application of packet switching to develop a protocol for internetworking, where multiple different networks could be joined together into a super-framework of networks. By defining a simple common network system, the Internet Protocol Suite, the concept of the network could be separated from its physical implementation. This spread of internetworking began to form into the idea of a global network that would be called the Internet, based on standardized protocols officially implemented in 1982. Adoption and interconnection occurred quickly across the advanced telecommunication networks of the western world, and then began to penetrate into the rest of the world as it became the de-facto international standard for the global network. However, the disparity of growth between advanced nations and the third-world countries led to a digital divide that is still a concern today.

Following commercialization and introduction of privately run Internet service providers in the 1980s, and the Internet's expansion for popular use in the 1990s, the Internet has had a drastic impact on culture and commerce. This includes the rise of near instant communication by electronic mail (e-mail), text based discussion forums, and the World Wide Web. Investor speculation in new markets provided by these innovations would also lead to the inflation and subsequent collapse of the Dot-com bubble. But despite this, the Internet continues to grow, driven by commerce, greater amounts of online information and knowledge and social networking known as Web 2.0.

Three terminals and an ARPA

In the 1950s and early 1960s, before the widespread inter-networking that led to the Internet, most communication networks were limited in that they only allowed communications between the stations on the network. Some networks had gateways or

bridges between them, but these bridges were often limited or built specifically for a single use. One prevalent computer networking method was based on the central mainframe method, simply allowing its terminals to be connected via long leased lines. This method was used in the 1950s by Project RAND to support researchers such as Herbert Simon, at Carnegie Mellon University in Pittsburgh, Pennsylvania, when collaborating across the continent with researchers in Sullivan, Illinois, on automated theorem proving and artificial intelligence.

A fundamental pioneer in the call for a global network, J.C.R. Licklider, articulated the ideas in his January 1960 paper, Man-Computer Symbiosis.

"A network of such [computers], connected to one another by wide-band communication lines [which provided] the functions of present-day libraries together with anticipated advances in information storage and retrieval and [other] symbiotic functions."
—J.C.R. Licklider,

In August, 1962, Licklider and Welden Clark published the paper "On-Line Man Computer Communication", one of the first descriptions of a networked future.

In October, 1962, Licklider was hired by Jack Ruina as Director of the newly established IPTO within DARPA, with a mandate to interconnect the United States Department of Defense's main computers at Cheyenne Mountain, the Pentagon, and SAC HQ. There he formed an informal group within DARPA to further computer research. He began by writing memos describing a distributed network to the IPTO staff, whom he called "Members and Affiliates of the Intergalactic Computer Network". As part of the information processing office's role, three network terminals had been installed: one for System Development Corporation in Santa Monica, one for Project Genie at the University of California, Berkeley and one for the Compatible Time-Sharing System project at the Massachusetts Institute of Technology (MIT). Licklider's identified need for inter-networking would be made obvious by the apparent waste of resources this caused.

"For each of these three terminals, I had three different sets of user commands. So if I was talking online with someone at S.D.C. and I wanted to talk to someone I knew at Berkeley or M.I.T. about this, I had to get up from the S.D.C. terminal, go over and log into the other terminal and get in touch with them. [...]

I said, it's obvious what to do (But I don't want to do it): If you have these three terminals, there ought to be one terminal that goes anywhere you want to go where you have interactive computing. That idea is the ARPAnet."

—Robert W. Taylor, co-writer with Licklider of "The Computer as a Communications Device", in an interview with the New York Times,

Although he left the IPTO in 1964, five years before the ARPANET went live, it was his vision of universal networking that provided the impetus that led his successors such as

Lawrence Roberts and Robert Taylor to further the ARPANET development. Licklider later returned to lead the IPTO in 1973 for two years.

Packet switching

Packet switching is a digital networking communications method that groups all transmitted data – regardless of content, type, or structure – into suitably-sized blocks, called *packets*. Packet switching features delivery of variable-bit-rate data streams (sequences of packets) over a shared network. When traversing network adapters, switches, routers and other network nodes, packets are buffered and queued, resulting in variable delay and throughput depending on the traffic load in the network.

Packet switching contrasts with another principal networking paradigm, circuit switching, a method which sets up a limited number of dedicated connections of constant bit rate and constant delay between nodes for exclusive use during the communication session. In case of traffic fees, for example in cellular communication, circuit switching is characterized by a fee per time unit of connection time, even when no data is transferred, while packet switching is characterized by a fee per unit of information.

Two major packet switching modes exist; connectionless packet switching, also known as datagram switching, and connection-oriented packet switching, also known as virtual circuit switching. In the first case each packet includes complete addressing or routing information. The packets are routed individually, sometimes resulting in different paths and out-of-order delivery. In the second case a connection is defined and preallocated in each involved node before any packet is transferred. The packets include a connection identifier rather than address information, and are delivered in order. See below.

Packet mode communication may be utilized with or without intermediate forwarding nodes (packet switches). In all packet mode communication, network resources are managed by statistical multiplexing or dynamic bandwidth allocation in which a communication channel is effectively divided into an arbitrary number of logical variable-bit-rate channels or data streams. Each logical stream consists of a sequence of packets, which normally are forwarded by the multiplexers and intermediate network nodes asynchronously using first-in, first-out buffering. Alternatively, the packets may be forwarded according to some scheduling discipline for fair queuing or for differentiated or guaranteed quality of service, such as pipeline forwarding or time-driven priority (TDP). Any buffering introduces varying latency and throughput in transmission. In case of a shared physical medium, the packets may be delivered according to some packet-mode multiple access scheme.

History

The concept of switching small blocks of data was first explored by Paul Baran in the early 1960s. Independently, Donald Davies at the National Physical Laboratory in the UK had developed the same ideas (Abbate, 2000).

Leonard Kleinrock conducted early research in queueing theory which would be important in packet switching, and published a book in the related field of digital message switching (without the packets) in 1961; he also later played a leading role in building and management of the world's first packet switched network, the ARPANET.

Baran developed the concept of message block switching during his research at the RAND Corporation for the US Air Force into survivable communications networks, first presented to the Air Force in the summer of 1961 as briefing B-265 then published as RAND Paper P-2626 in 1962 , and then including and expanding somewhat within a series of eleven papers titled On Distributed Communications in 1964. Baran's P-2626 paper described a general architecture for a large-scale, distributed, survivable communications network. The paper focuses on three key ideas: first, use of a decentralized network with multiple paths between any two points; and second, dividing complete user messages into what he called *message blocks* (later called packets); then third, delivery of these messages by store and forward switching.

Baran's study made its way to Robert Taylor and J.C.R. Licklider at the Information Processing Technology Office, both wide-area network evangelists, and it helped influence Lawrence Roberts to adopt the technology when Taylor put him in charge of development of the ARPANET.

Baran's work was similar to the research performed independently by Donald Davies at the National Physical Laboratory, UK. In 1965, Davies developed the concept of packet-switched networks and proposed development of a UK wide network. He gave a talk on the proposal in 1966, after which a person from the Ministry of Defense told him about Baran's work. A member of Davies' team met Lawrence Roberts at the 1967 ACM Symposium on Operating System Principles, bringing the two groups together.

Interestingly, Davies had chosen some of the same parameters for his original network design as Baran, such as a packet size of 1024 bits. In 1966 Davies proposed that a network should be built at the laboratory to serve the needs of NPL and prove the feasibility of packet switching. The NPL Data Communications Network entered service in 1970. Roberts and the ARPANET team took the name "packet switching" itself from Davies's work.

The first computer network and packet switching network deployed for computer resource sharing was the Octopus Network at the Lawrence Livermore National Laboratory that began connecting four Control Data 6600 computers to several shared storage devices (including an IBM 2321 Data Cell in 1968 and an IBM Photostore in

1970) and to several hundred ASR-33 Teletype terminals for time sharing use starting in 1968.

Connectionless and connection-oriented packet switching

The service actually provided to the user by networks using packet switching nodes can be either connectionless (based on datagram messages), or virtual circuit switching (also known as connection oriented). Some connectionless protocols are Ethernet, IP, and UDP; connection oriented packet-switching protocols include X.25, Frame relay, Asynchronous Transfer Mode (ATM), Multiprotocol Label Switching (MPLS), and TCP.

In connection oriented networks, each packet is labeled with a connection ID rather than an address. Address information is only transferred to each node during a connection set-up phase, when the route to the destination is discovered and an entry is added to the switching table in each network node through which the connection passes. The signalling protocols used allow the application to specify its requirements and the network to specify what capacity etc. is available, and acceptable values for service parameters to be negotiated. Routing a packet is very simple, as it just requires the node to look up the ID in the table. The packet header can be small, as it only needs to contain the ID and any information (such as length, timestamp, or sequence number) which is different for different packets.

In connectionless networks, each packet is labeled with a destination address, source address, and port numbers; it may also be labeled with the sequence number of the packet. This precludes the need for a dedicated path to help the packet find its way to its destination, but means that much more information is needed in the packet header, which is therefore larger, and this information needs to be looked up in power-hungry content-addressable memory. Each packet is dispatched and may go via different routes; potentially, the system has to do as much work for every packet as the connection-oriented system has to do in connection set-up, but with less information as to the application's requirements. At the destination, the original message/data is reassembled in the correct order, based on the packet sequence number. Thus a virtual connection, also known as a virtual circuit or byte stream is provided to the end-user by a transport layer protocol, although intermediate network nodes only provides a connectionless network layer service.

Packet switching in networks

Packet switching is used to optimize the use of the channel capacity available in digital telecommunication networks such as computer networks, to minimize the transmission latency (i.e. the time it takes for data to pass across the network), and to increase robustness of communication.

The most well-known use of packet switching is the Internet and local area networks. The Internet is implemented by the Internet Protocol Suite using a variety of Link Layer technologies. For example, Ethernet and Frame Relay are common. Newer mobile phone technologies (e.g., GPRS, I-mode) also use packet switching.

X.25 is a notable use of packet switching in that, despite being based on packet switching methods, it provided virtual circuits to the user. These virtual circuits carry variable-length packets. In 1978, X.25 was used to provide the first international and commercial packet switching network, the International Packet Switched Service (IPSS). Asynchronous Transfer Mode (ATM) also is a virtual circuit technology, which uses fixed-length cell relay connection oriented packet switching.

Datagram packet switching is also called connectionless networking because no connections are established. Technologies such as Multiprotocol Label Switching (MPLS) and the resource reservation protocol (RSVP) create virtual circuits on top of datagram networks. Virtual circuits are especially useful in building robust failover mechanisms and allocating bandwidth for delay-sensitive applications.

MPLS and its predecessors, as well as ATM, have been called "fast packet" technologies. MPLS, indeed, has been called "ATM without cells". Modern routers, however, do not require these technologies to be able to forward variable-length packets at multigigabit speeds across the network.

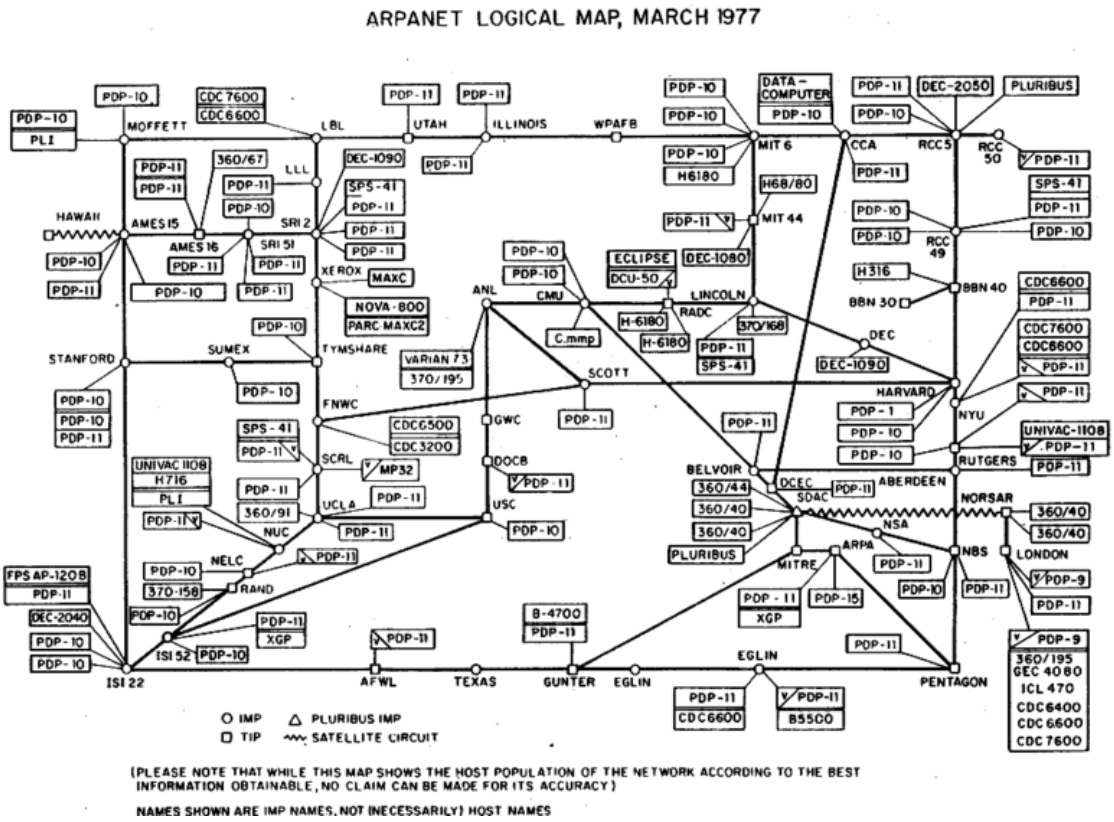
X.25 vs. Frame Relay packet switching

Both X.25 and Frame Relay provide connection-oriented packet switching, also known as virtual circuit switching. A major difference between X.25 and frame relay packet switching are that X.25 is a reliable protocol, based on node-to-node automatic repeat request, while Frame Relay is a non-reliable protocol, maximum packet length is 1000 bytes. Any retransmissions must be carried out by higher layer protocols. The X.25 protocol is a network layer protocol, and is part of the X.25 protocol suite, also known as the OSI protocol suite. It was widely used in relatively slow switching networks during the 1980s, for example as an alternative to circuit mode terminal switching, and for automated teller machines. Frame relay is a further development of X.25. The simplicity of Frame relay made it considerably faster and more cost effective than X.25 packet switching. Frame relay is a data link layer protocol, and does not provide logical addresses and routing. It is only used for semi-permanent connections, while X.25 connections also can be established for each communication session. Frame relay was used to interconnect LANs or LAN segments, mainly in the 1990s by large companies that had a requirement to handle heavy telecommunications traffic across wide area networks. (O'Brien & Marakas, 2009, p. 250) Despite the benefits of frame relay packet switching, many international companies are staying with the X.25 standard. In the United States, X.25 packet switching was used heavily in government and financial networks that use mainframe applications. Many companies did not intend to cross over to frame relay packet switching because it is more cost effective to use X.25 on slower

networks. In certain parts of the world, particularly in Asia-Pacific and South America regions, X.25 was the only technology available. (Girard, 1997)

Networks that led to the Internet

ARPANET



ARPANET logical map, March 1977

ARPANET (**Advanced Research Projects Agency Network**), created by a small research team at the head of the Massachusetts Institute of Technology and the Defense Advanced Research Projects Agency (DARPA) of the United States Department of Defense, was the world's first operational packet switching network, and one of the networks that came to compose the global Internet. The packet switching of the ARPANET was based on designs by Lawrence Roberts, of the Lincoln Laboratory.

Packet switching, now the dominant basis for data communications worldwide, was a new concept at the time. Data communications had been based on the idea of circuit switching, as in the old, typical telephone circuit, wherein a dedicated circuit is occupied

for the duration of the telephone call, and communication is possible only with the single party at the far end of the circuit.

With packet switching, a data system could use one communications link to communicate with more than one machine by disassembling data into datagrams, then gather these as packets. Thus, not only could the link be shared (much as a single post box can be used to post letters to different destinations), but each packet could be routed independently of other packets.

History

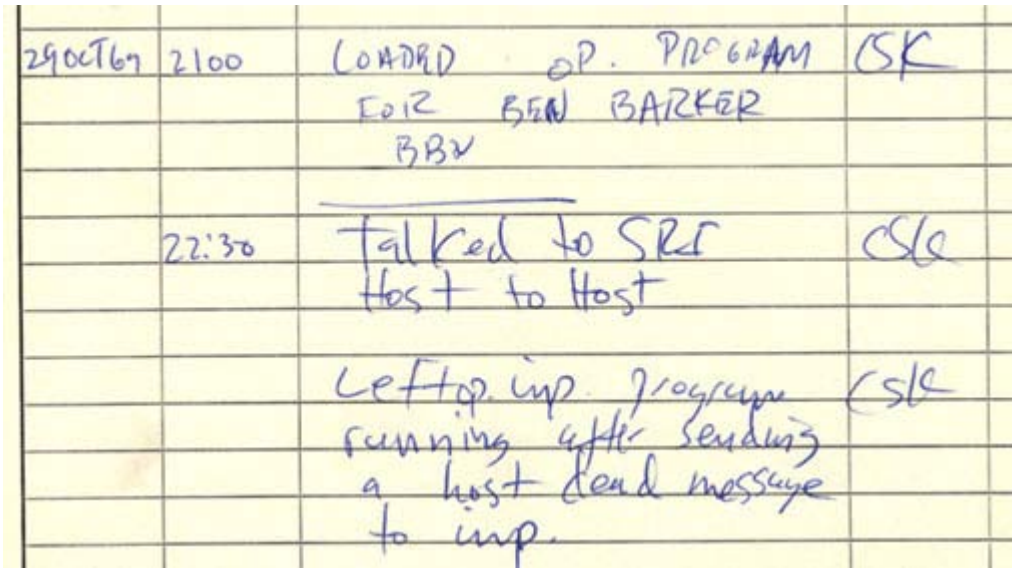
The earliest ideas for a computer network intended to allow general communications among computer users were formulated by the computer scientist J. C. R. Licklider, of the Bolt, Beranek and Newman (BBN) company, in August 1962, in memoranda discussing his concept for an “Intergalactic Computer Network”. Those ideas contained almost everything that composes the contemporary Internet. In October 1963, at the United States Department of Defense, Licklider was appointed head of the Behavioral Sciences and Command and Control programs at the Advanced Research Projects Agency — ARPA (the initial ARPANET acronym). He then convinced Ivan Sutherland and Bob Taylor that this computer network concept was very important, meriting development, although he left ARPANET before anyone worked on his concept. ARPA and Bob Taylor continued their interest in creating such a computer communications network, in part, to allow ARPA-sponsored researchers at various corporate and academic locales to put to use the computers ARPA was providing them, and, in part, to make new software and other computer science results quickly and widely available. In his office, Taylor had three computer terminals, each connected to separate computers, which ARPA was funding: the first, for the System Development Corporation (SDC) Q-32, in Santa Monica; the second, for Project Genie, at the University of California, Berkeley; and the third, for Multics, at MIT. Taylor recalls the circumstance: "For each of these three terminals, I had three different sets of user commands. So, if I was talking online with someone at S.D.C., and I wanted to talk to someone I knew at Berkeley, or M.I.T., about this, I had to get up from the S.D.C. terminal, go over and log into the other terminal and get in touch with them. I said, “Oh Man!”, it’s obvious what to do: If you have these three terminals, there ought to be one terminal that goes anywhere you want to go. That idea is the ARPANET". Somewhat contemporaneously, several other people had (mostly independently) worked out the aspects of “packet switching”, with the first public demonstration presented by the National Physical Laboratory (NPL), on 5 August 1968, in the United Kingdom.

Creation

By mid-1968, Taylor had prepared a complete plan for a computer network, and, after ARPA’s approval, a Request For Quotation (RFQ) was sent to 140 potential bidders. Most computer science companies regarded the ARPA–Taylor proposal as outlandish, and only twelve submitted bids to build the network; of the twelve, ARPA regarded only

four as top-rank contractors. At year's end, ARPA considered only two contractors, and awarded the contract to build the network to BBN Technologies on 7 April 1969. The initial, seven-man BBN team were much aided by the technical specificity of their response to the ARPA RFQ — and thus quickly produced the first working computers. The BBN-proposed network closely followed Taylor's ARPA plan: a network composed of small IMP computers, Interface Message Processors (contemporary routers). At each site, the IMPs performed store-and-forward packet switching functions, and were interconnected with modems that were connected to leased lines (initially running at 50 kbit/second). The host computers were connected to the IMPs via custom serial interfaces connecting to the ARPANET. The system, including the hardware and the packet switching software, was designed and installed in nine months. To build the first-generation IMPs, BBN Technologies initially used a rugged computer version of the Honeywell DDP-516 computer (originally) configured with 24 kB of (expandable) core memory, and a 16-channel Direct Multiplex Control (DMC) direct memory access control unit. The DMC established custom interfaces with each of the host computers and modems. In addition to the front-panel lamps, the DDP-516 computer also features a special set of 24 indicator-lamps showing the status of the IMP communication channels. Each IMP could support up to four local hosts, and could communicate with up to six remote IMPs via leased lines.

ARPA deployed



Historical document: First ARPANET IMP log: the first message ever sent via the ARPANET, 10:30 PM, 29 October 1969. This IMP Log excerpt, kept at UCLA, describes setting up a message transmission from the UCLA SDS Sigma 7 Host computer to the SRI SDS 940 Host computer

The initial ARPANET consisted of four IMPs installed at:

1. University of California, Los Angeles (UCLA), where Leonard Kleinrock had established a Network Measurement Center, with an SDS Sigma 7 being the first computer attached to it;
2. The Stanford Research Institute's Augmentation Research Center, where Douglas Engelbart had created the ground-breaking NLS system, a very important early hypertext system (with the SDS 940 that ran NLS, named "Genie", being the first host attached);
3. University of California, Santa Barbara (UCSB), with the Culler-Fried Interactive Mathematics Centre's IBM 360/75, running OS/MVT being the machine attached;
4. The University of Utah's Computer Science Department, where Ivan Sutherland had moved, running a DEC PDP-10 running TENEX.

The first message transmitted over the ARPANET was sent by UCLA student programmer Charley Kline, at 10:30 p.m, on October 29, 1969. Supervised by Prof. Leonard Kleinrock, Kline transmitted from the university's SDS Sigma 7 Host computer to the Stanford Research Institute's SDS 940 Host computer. The message text was the word "login"; the "l" and the "o" letters were transmitted, but the system then crashed. Hence, the literal first message over the ARPANET was "lo". About an hour later, having recovered from the crash, the SDS Sigma 7 computer effected a full "login". The first permanent ARPANET link was established on November 21, 1969, between the IMP at UCLA and the IMP at the Stanford Research Institute. By December 5, 1969, the entire four-node network was connected.

The contents of the first e-mail transmission in 1971 have been forgotten; in the Frequently Asked Questions section of his Web site, the sender, Ray Tomlinson, who sent the message between two computers sitting side-by-side, claims that the contents were "entirely forgettable, and I have, therefore, forgotten them", and speculates that the message likely was "QWERTYUIOP" or some such.

Software & protocols

The starting point for host-to-host communication on the ARPANET was the 1822 protocol, which defined how a host computer transmitted messages to an ARPANET IMP. The message format was designed to work unambiguously with a broad range of computer architectures. An 1822 message essentially consisted of (i) a message type, (ii) a numeric host address, and (iii) a data field. To send a data message to another host, the transmitting host would format a data message containing the destination host's address and the data message being sent, and then transmit the message through the 1822 hardware interface. The IMP then delivered the message to its destination address, either by delivering it to a locally connected host, or by delivering it to another IMP. When the message was ultimately delivered to the destination host, the receiving IMP would transmit a *Ready for Next Message* (RFNM) acknowledgement to the sending, host IMP.

Unlike modern Internet datagrams, the ARPANET was designed to reliably transmit 1822 messages, and to inform the host computer when it loses a message; the contemporary IP is unreliable, whereas the TCP is reliable. Nonetheless, the 1822

protocol proved inadequate for handling multiple connections among different applications residing in a host computer. This problem was addressed with the Network Control Program (NCP), which provided a standard method to establish reliable, flow-controlled, bidirectional communications links among different processes in different host computers. The NCP interface allowed application software to connect across the ARPANET by implementing higher-level communication protocols, an early example of the *protocol layering* concept incorporated to the OSI model. In 1983, TCP/IP protocols replaced NCP as the ARPANET's principal protocol, and the ARPANET then became one component of the early Internet.

Network Applications

NCP provided a standard set of network services that could be shared by several applications running on a single host computer. This led to the evolution of *application protocols* that operated, more or less, independently of the underlying network service. When the ARPANET migrated to the Internet protocols in 1983, the major application protocols migrated with it.

- **E-mail:** In 1971, Ray Tomlinson, of the BBN company sent the first network e-mail. By 1973, e-mail constituted 75 per cent of ARPANET traffic

- **File transfer:** By 1973, the File Transfer Protocol (FTP) specification had been defined and implemented, enabling file transfers over the ARPANET

- **Voice traffic:** The Network Voice Protocol (NVP) specifications were defined in (RFC 741), then implemented, but, because of technical shortcomings, conference calls over the ARPANET never worked well; the contemporary Voice over Internet Protocol (packet voice) was decades away

Growth

In March, 1970, the ARPANET reached the east coast of the United States, when a BBN company IMP was connected to the network. Thereafter, the ARPANET grew: 9 IMPs by June 1970 and 13 IMPs by December 1970, then 18 by September 1971 (when the network included 23 university and government hosts); 29 IMPs by August 1972, and 40 by September, 1973. By June 1974, there were 46 IMPs, and in July 1975, the network numbered 57 IMPs. By 1981, the number was 213 host computers, with another host connecting approximately every twenty days.

In 1968, two satellite links, traversing the Pacific and Atlantic oceans, to Hawaii and Norway, one, the Norwegian Seismic Array (NORSAR), were connected to the ARPANET. Moreover, from Norway, a terrestrial circuit added a London IMP to the network in 1973.

Given that its primary function was funding research and development, the ARPA, in 1975, transferred ARPANET control to the Defense Communications Agency, a

component of the U.S. Department of Defense. In 1983, the U.S. military sub-networks of the ARPANET became the discrete Military Network (MILNET) for unclassified defense department communications; separating the civil and military networks reduced the 113-node ARPANET by 68 nodes.

Development: hardware

Support for inter-IMP circuits of up to 230.4 kbit/s was added in 1970, although considerations of cost and IMP processing power meant this capability was not actively used.

1971 saw the start of the use of the non-ruggedized (and therefore significantly lighter) Honeywell 316 as an IMP. It could also be configured as a Terminal IMP (TIP), which added support for up to 63 ASCII serial terminals through a multi-line controller in place of one of the hosts. The 316 featured a greater degree of integration than the 516, which made it less expensive and easier to maintain. The 316 was configured with 40 kB of core memory for a TIP. The size of core memory was later increased, to 32 kB for the IMPs, and 56 kB for TIPs, in 1973.

In 1975, BBC introduced IMP software running on the Pluribus multi-processor. These appeared in a small number of sites. In 1981, BBC introduced IMP software running on its own C/30 processor product.

The original IMPs and TIPs were phased out as the ARPANET was shut down after the introduction of the NSFNet, but some IMPs remained in service as late as 1989.

Senator Albert Gore, Jr. began to craft the High Performance Computing and Communication Act of 1991 (commonly referred to as "The Gore Bill") after hearing the 1988 report toward a National Research Network submitted to Congress by a group chaired by Leonard Kleinrock, professor of computer science at UCLA. The bill was passed on December 9, 1991 and led to the National Information Infrastructure (NII) which Al Gore called the "information superhighway".

The ARPANET under nuclear attack

Common ARPANET lore posits that the computer network was designed to survive a nuclear attack. In *A Brief History of the Internet*, the Internet Society describe the coalescing of the technical ideas that produced the ARPANET:

It was from the RAND study that the false rumor started, claiming that the ARPANET was somehow related to building a network resistant to nuclear war. This was never true of the ARPANET, only the unrelated RAND study on secure voice considered nuclear war. However, the later work on Internetting did emphasize robustness and survivability, including the capability to withstand losses of large portions of the underlying networks.

Although the ARPANET was designed to survive subordinate-network losses, the principal reason was that the switching nodes and network links were unreliable, even without any nuclear attacks. About the resources scarcity that spurred the creation of the ARPANET, Charles Herzfeld, ARPA Director (1965–1967), said:

The ARPANET was not started to create a Command and Control System that would survive a nuclear attack, as many now claim. To build such a system was, clearly, a major military need, but it was not ARPA's mission to do this; in fact, we would have been severely criticized had we tried. Rather, the ARPANET came out of our frustration that there were only a limited number of large, powerful research computers in the country, and that many research investigators, who should have access to them, were geographically separated from them.

Retrospective

The support and management of ARPA contributed to the successful creation of the ARPANET. To wit, the *ARPANET Completion Report*, jointly published by the BBN company and ARPA, concludes that:

... it is somewhat fitting to end on the note that the ARPANET program has had a strong and direct feedback into the support and strength of computer science, from which the network, itself, sprang. [4]

In the wake of ARPANET being formally decommissioned on the 28th of February, 1990, Vinton Cerf wrote the following lamentation, entitled, "Requiem of the ARPANET":

It was the first, and being first, was best,
but now we lay it down to ever rest.
Now pause with me a moment, shed some tears.
For auld lang syne, for love, for years and years
of faithful service, duty done, I weep.
Lay down thy packet, now, O friend, and sleep.

-Vinton Cerf

The ARPANET in film and other media

- A 1969 Walt Disney movie, *The Computer Wore Tennis Shoes*
- A 1983 movie, *WarGames*, is a story of possible nuclear warfare averted by an intelligent but anxious teen who cracks into a government command and control system, the WOPR (a reference to the actual military WMCCS).
- A 1985 episode of the U.S. television sitcom *Benson* includes a scene in which ARPANET is accessed. This is believed to be the first incidence of a popular TV show referencing the Internet or its progenitors.

- In *Let the Great World Spin: A Novel*, published in 2009 but set in 1974 and written by Colum McCann, a character named The Kid and others use ARPANET from a Palo Alto computer to dial phone booths in New York City in order to hear descriptions of Philippe Petit's tight rope walk between the World Trade Center Towers.
- In *Metal Gear Solid 3: Snake Eater*, a character named Sigint takes part in the development of ARPANET after the events depicted in the game.
- The *Doctor Who* Past Doctor Adventures novel *Blue Box*, written in 2003 but set in 1981, includes a character predicting that by the year 2000 there will be four hundred machines connected to ARPANET.
- There is an electronic music artist known as *Arpanet*, Gerald Donald, one of the members of Drexciya. The name is formatted as a word instead of an acronym, but is still a clear nod to ARPANET. The artist's 2002 album *Wireless Internet* features commentary on the expansion of the internet via wireless communication, with songs such as *NTT DoCoMo*, dedicated to the mobile communications giant based in Japan.
- In numerous *The X-Files* episodes ARPANET is referenced and usually hacked into by The Lone Gunmen. This is most noticeable in the episode "Unusual Suspects".
- Thomas Pynchon's 2009 novel Inherent Vice, set in southern California circa 1970, contains a character who accesses the ARPANET throughout the course of the book. ARPANET is spelled therein as 'ARPAnet.'

X.25 and public access

Based on ARPA's research, packet switching network standards were developed by the International Telecommunication Union (ITU) in the form of X.25 and related standards. While using packet switching, X.25 is built on the concept of virtual circuits emulating traditional telephone connections. In 1974, X.25 formed the basis for the SERCnet network between British academic and research sites, which later became JANET. The initial ITU Standard on X.25 was approved in March 1976.

The British Post Office, Western Union International and Tymnet collaborated to create the first international packet switched network, referred to as the International Packet Switched Service (IPSS), in 1978. This network grew from Europe and the US to cover Canada, Hong Kong and Australia by 1981. By the 1990s it provided a worldwide networking infrastructure.

Unlike ARPANET, X.25 was commonly available for business use. Telenet offered its Telemail electronic mail service, which was also targeted to enterprise use rather than the general email system of the ARPANET.

The first public dial-in networks used asynchronous TTY terminal protocols to reach a concentrator operated in the public network. Some networks, such as CompuServe, used X.25 to multiplex the terminal sessions into their packet-switched backbones, while others, such as Tymnet, used proprietary protocols. In 1979, CompuServe became the

first service to offer electronic mail capabilities and technical support to personal computer users. The company broke new ground again in 1980 as the first to offer real-time chat with its CB Simulator. Other major dial-in networks were America Online (AOL) and Prodigy that also provided communications, content, and entertainment features. Many bulletin board system (BBS) networks also provided on-line access, such as FidoNet which was popular amongst hobbyist computer users, many of them hackers and amateur radio operators.

NPL

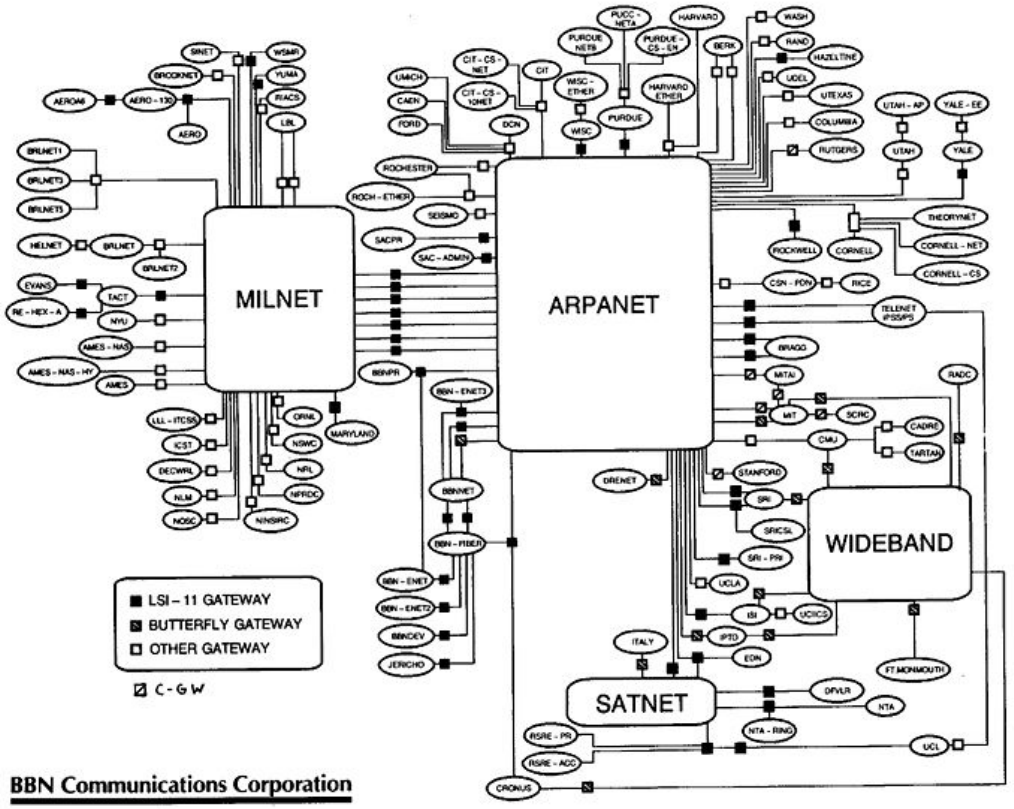
In 1965, Donald Davies of the National Physical Laboratory (United Kingdom) proposed a national data network based on packet-switching. The proposal was not taken up nationally but by 1970 he had designed and built a packet-switched network to meet the needs of the multidisciplinary laboratory and prove the technology under operational conditions. By 1976 12 computers and 75 terminal devices were attached and more were added until the network was replaced in 1986.

Merging the networks and creating the Internet

ARPANET to several federal wide area networks: MILNET, NSI, and NSFNet

After the ARPANET had been up and running for several years, ARPA looked for another agency to hand off the network to; ARPA's primary mission was funding cutting edge research and development, not running a communications utility. Eventually, in July 1975, the network had been turned over to the Defense Communications Agency, also part of the Department of Defense. In 1983, the U.S. military portion of the ARPANET was broken off as a separate network, the MILNET. MILNET subsequently became the unclassified but military-only NIPRNET, in parallel with the SECRET-level SIPRNET and JWICS for TOP SECRET and above. NIPRNET does have controlled security gateways to the public Internet.

The networks based on the ARPANET were government funded and therefore restricted to noncommercial uses such as research; unrelated commercial use was strictly forbidden. This initially restricted connections to military sites and universities. During the 1980s, the connections expanded to more educational institutions, and even to a growing number of companies such as Digital Equipment Corporation and Hewlett-Packard, which were participating in research projects or providing services to those who were.



BBN Technologies TCP/IP internet map early 1986

Several other branches of the U.S. government, the National Aeronautics and Space Agency (NASA), the National Science Foundation (NSF), and the Department of Energy (DOE) became heavily involved in Internet research and started development of a successor to ARPANET. In the mid 1980s, all three of these branches developed the first Wide Area Networks based on TCP/IP. NASA developed the NASA Science Network, NSF developed CSNET and DOE evolved the Energy Sciences Network or ESNet.

In 1984 NSF developed CSNET exclusively based on TCP/IP. CSNET connected with ARPANET using TCP/IP, and ran TCP/IP over X.25, but it also supported departments without sophisticated network connections, using automated dial-up mail exchange. This grew into the NSFNet backbone, established in 1986, and intended to connect and provide access to a number of supercomputing centers established by the NSF.

Transition towards the Internet

The term "internet" was adopted in the first RFC published on the TCP protocol (RFC 675: Internet Transmission Control Program, December 1974) as an abbreviation of the term *internetworking* and the two terms were used interchangeably. In general, an *internet* was any network using TCP/IP. It was around the time when ARPANET was

interlinked with NSFNet in the late 1980s, that the term was used as the name of the network, Internet, being a large and global TCP/IP network.

As interest in wide spread networking grew and new applications for it were developed, the Internet's technologies spread throughout the rest of the world. The network-agnostic approach in TCP/IP meant that it was easy to use any existing network infrastructure, such as the IPSS X.25 network, to carry Internet traffic. In 1984, University College London replaced its transatlantic satellite links with TCP/IP over IPSS.

Many sites unable to link directly to the Internet started to create simple gateways to allow transfer of e-mail, at that time the most important application. Sites which only had intermittent connections used UUCP or FidoNet and relied on the gateways between these networks and the Internet. Some gateway services went beyond simple e-mail peering, such as allowing access to FTP sites via UUCP or e-mail.

Finally, the Internet's remaining centralized routing aspects were removed. The EGP routing protocol was replaced by a new protocol, the Border Gateway Protocol (BGP), in order to allow the removal of the NSFNet Internet backbone network. In 1994, Classless Inter-Domain Routing was introduced to support better conservation of address space which allowed use of route aggregation to decrease the size of routing tables. The picture on the right hand side shows a system made with the help of the high-tech company BBN.

TCP/IP becomes worldwide

CERN, the European Internet, the link to the Pacific and beyond

Between 1984 and 1988 CERN began installation and operation of TCP/IP to interconnect its major internal computer systems, workstations, PCs and an accelerator control system. CERN continued to operate a limited self-developed system CERNET internally and several incompatible (typically proprietary) network protocols externally. There was considerable resistance in Europe towards more widespread use of TCP/IP and the CERN TCP/IP intranets remained isolated from the Internet until 1989.

In 1988 Daniel Karrenberg, from Centrum Wiskunde & Informatica (CWI) in Amsterdam, visited Ben Segal, CERN's TCP/IP Coordinator, looking for advice about the transition of the European side of the UUCP Usenet network (much of which ran over X.25 links) over to TCP/IP. In 1987, Ben Segal had met with Len Bosack from the then still small company Cisco about purchasing some TCP/IP routers for CERN, and was able to give Karrenberg advice and forward him on to Cisco for the appropriate hardware. This expanded the European portion of the Internet across the existing UUCP networks, and in 1989 CERN opened its first external TCP/IP connections. This coincided with the creation of Réseaux IP Européens (RIPE), initially a group of IP network administrators who met regularly to carry out co-ordination work together. Later, in 1992, RIPE was formally registered as a cooperative in Amsterdam.

At the same time as the rise of internetworking in Europe, ad hoc networking to ARPA and in-between Australian universities formed, based on various technologies such as X.25 and UUCPNet. These were limited in their connection to the global networks, due to the cost of making individual international UUCP dial-up or X.25 connections. In 1989, Australian universities joined the push towards using IP protocols to unify their networking infrastructures. AARNet was formed in 1989 by the Australian Vice-Chancellors' Committee and provided a dedicated IP based network for Australia.

The Internet began to penetrate Asia in the late 1980s. Japan, which had built the UUCP-based network JUNET in 1984, connected to NSFNet in 1989. It hosted the annual meeting of the Internet Society, INET'92, in Kobe. Singapore developed TECHNET in 1990, and Thailand gained a global Internet connection between Chulalongkorn University and UUNET in 1992.

Digital divide

While developed countries with technological infrastructures were joining the Internet, developing countries began to experience a digital divide separating them from the Internet. On an essentially continental basis, they are building organizations for Internet resource administration and sharing operational experience, as more and more transmission facilities go into place.

Africa

At the beginning of the 1990s, African countries relied upon X.25 IPSS and 2400 baud modem UUCP links for international and internetwork computer communications.

In August, 1995, InfoMail Uganda, Ltd., a privately held firm in Kampala now known as InfoCom and NSN Network Services of Avon, Colorado, sold in 1997 and now known as Clear Channel Satellite, established Africa's first native TCP/IP high-speed satellite Internet services. The data connection was originally carried by a C-Band RSCC Russian satellite which connected InfoMail's Kampala offices directly to NSN's MAE-West point of presence using a private network from NSN's leased ground station in New Jersey. InfoCom's first satellite connection was just 64kbps, serving a Sun host computer and twelve US Robotics dial-up modems.

In 1996 a USAID funded project, the Leland initiative, started work on developing full Internet connectivity for the continent. Guinea, Mozambique, Madagascar and Rwanda gained satellite earth stations in 1997, followed by Côte d'Ivoire and Benin in 1998.

Africa is building an Internet infrastructure. AfriNIC, headquartered in Mauritius, manages IP address allocation for the continent. As do the other Internet regions, there is an operational forum, the Internet Community of Operational Networking Specialists.

There are a wide range of programs both to provide high-performance transmission plant, and the western and southern coasts have undersea optical cable. High-speed cables join

North Africa and the Horn of Africa to intercontinental cable systems. Undersea cable development is slower for East Africa; the original joint effort between New Partnership for Africa's Development (NEPAD) and the East Africa Submarine System (Eassy) has broken off and may become two efforts.

Asia and Oceania

The Asia Pacific Network Information Centre (APNIC), headquartered in Australia, manages IP address allocation for the continent. APNIC sponsors an operational forum, the Asia-Pacific Regional Internet Conference on Operational Technologies (APRICOT).

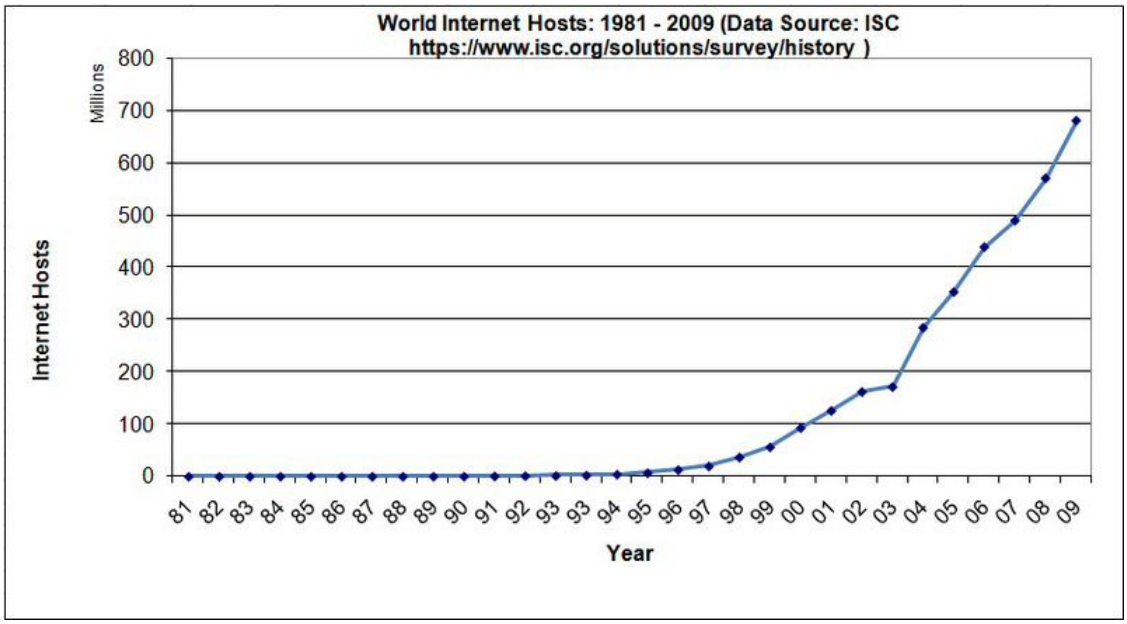
In 1991, the People's Republic of China saw its first TCP/IP college network, Tsinghua University's TUNET. The PRC went on to make its first global Internet connection in 1994, between the Beijing Electro-Spectrometer Collaboration and Stanford University's Linear Accelerator Center. However, China went on to implement its own digital divide by implementing a country-wide content filter.

Latin America

As with the other regions, the Latin American and Caribbean Internet Addresses Registry (LACNIC) manages the IP address space and other resources for its area. LACNIC, headquartered in Uruguay, operates DNS root, reverse DNS, and other key services.

Opening the network to commerce

The interest in commercial use of the Internet became a hotly debated topic. Although commercial use was forbidden, the exact definition of commercial use could be unclear and subjective. UUCPNet and the X.25 IPSS had no such restrictions, which would eventually see the official barring of UUCPNet use of ARPANET and NSFNet connections. Some UUCP links still remained connecting to these networks however, as administrators cast a blind eye to their operation.



During the late 1980s, the first Internet service provider (ISP) companies were formed. Companies like PSINet, UUNET, Netcom, and Portal Software were formed to provide service to the regional research networks and provide alternate network access, UUCP-based email and Usenet News to the public. The first commercial dialup ISP in the United States was The World, opened in 1989.

In 1992, Congress allowed commercial activity on NSFNet with the Scientific and Advanced-Technology Act, 42 U.S.C. § 1862(g), permitting NSFNet to interconnect with commercial networks. This caused controversy amongst university users, who were outraged at the idea of noneducational use of their networks. Eventually, it was the commercial Internet service providers who brought prices low enough that junior colleges and other schools could afford to participate in the new arenas of education and research.

By 1990, ARPANET had been overtaken and replaced by newer networking technologies and the project came to a close. In 1994, the NSFNet, now renamed ANSNET (Advanced Networks and Services) and allowing non-profit corporations access, lost its standing as the backbone of the Internet. Both government institutions and competing commercial providers created their own backbones and interconnections. Regional network access points (NAPs) became the primary interconnections between the many networks. The final commercial restrictions ended in May 1995 when the National Science Foundation ended its sponsorship of the Internet backbone.

Internet Engineering Task Force

Requests for Comments (RFCs) started as memoranda addressing the various protocols that facilitate the functioning of the Internet and were previously edited by the late Dr. Postel as part of his IANA functions.

The IETF started in January 1985 as a quarterly meeting of U.S. government funded researchers. Representatives from non-government vendors were invited starting with the fourth IETF meeting in October of that year. In 1992, the Internet Society, a professional membership society, was formed and the IETF was transferred to operation under it as an independent international standards body.

NIC, InterNIC, IANA and ICANN

The first central authority to coordinate the operation of the network was the Network Information Centre (NIC) at Stanford Research Institute (SRI) in Menlo Park, California. In 1972, management of these issues was given to the newly created Internet Assigned Numbers Authority (IANA). In addition to his role as the RFC Editor, Jon Postel worked as the manager of IANA until his death in 1998.

As the early ARPANET grew, hosts were referred to by names, and a HOSTS.TXT file would be distributed from SRI International to each host on the network. As the network grew, this became cumbersome. A technical solution came in the form of the Domain Name System, created by Paul Mockapetris. The Defense Data Network—Network Information Center (DDN-NIC) at SRI handled all registration services, including the top-level domains (TLDs) of .mil, .gov, .edu, .org, .net, .com and .us, root nameserver administration and Internet number assignments under a United States Department of Defense contract. In 1991, the Defense Information Systems Agency (DISA) awarded the administration and maintenance of DDN-NIC (managed by SRI up until this point) to Government Systems, Inc., who subcontracted it to the small private-sector Network Solutions, Inc.

Since at this point in history most of the growth on the Internet was coming from non-military sources, it was decided that the Department of Defense would no longer fund registration services outside of the .mil TLD. In 1993 the U.S. National Science Foundation, after a competitive bidding process in 1992, created the InterNIC to manage the allocations of addresses and management of the address databases, and awarded the contract to three organizations. Registration Services would be provided by Network Solutions; Directory and Database Services would be provided by AT&T; and Information Services would be provided by General Atomics.

In 1998 both IANA and InterNIC were reorganized under the control of ICANN, a California non-profit corporation contracted by the United States Department of Commerce to manage a number of Internet-related tasks. The role of operating the DNS system was privatized and opened up to competition, while the central management of name allocations would be awarded on a contract tender basis.

Internet governance

Policies and mechanisms for **Internet governance** have been topics of debate between many different Internet stakeholders, some of whom have very different opinions for how and indeed whether the Internet should facilitate free communication of ideas and information.

Definition

The definition of Internet governance has been contested by differing groups across political and ideological lines. One of the main debates concerns the authority and participation of certain actors, such as national governments, corporate entities and civil society, to play a role in the Internet's governance.

A Working group established after a United Nations-initiated World Summit on the Information Society (WSIS) proposed the following definition of Internet governance as part of its June 2005 report:

Internet governance is the development and application by Governments, the private sector and civil society, in their respective roles, of shared principles, norms, rules, decision-making procedures, and programmes that shape the evolution and use of the Internet.

Law professor Yochai Benkler developed a conceptualization of Internet governance by the idea of three "layers" of governance: the "physical infrastructure" layer through which information travels; the "code" or "logical" layer that controls the infrastructure; and the "content" layer, which contains the information that signals through the network.

History

To understand how the Internet is managed today, it is necessary to know some of the main events of Internet governance.

The original ARPANET, one of the components which evolved eventually into the Internet, connected four Universities: University of California Los Angeles, University of California Santa Barbara, Stanford Research Institute and Utah University. The IMPs, interface minicomputers, were built during 1969 by Bolt, Beranek and Newman in accord with a proposal by the US Department of Defense Advanced Research Projects Agency, which funded the system as an experiment. By 1973 it connected many more systems and included satellite links to Hawaii and Scandinavia, and a further link from Norway to London. ARPANET continued to grow in size, becoming more a utility than a research project. For this reason during 1975 it was transferred to the US Defense Communications Agency.

During the development of ARPANET, a numbered series of Request for Comments (RFCs) memos documented technical decisions and methods of working as they evolved. The standards of today's Internet are still documented by RFCs, produced through the very process which evolved on ARPANET.

Outside of the USA the dominant technology was X.25. The International Packet Switched Service, created during 1978, used X.25 and extended to Europe, Australia, Hong Kong, Canada, and the USA. It allowed individual users and companies to connect to a variety of mainframe systems, including Compuserve. Between 1979 and 1984, an system known as Unix to Unix Copy Program grew to connect 940 hosts, using methods like X.25 links, ARPANET connections, and leased lines. Usenet News, a distributed discussion system, was a major use of UUCP.

The Internet protocol suite, developed between 1973 and 1977 with funding from ARPA, was intended to hide the differences between different underlying networks and allow many different applications to be used over the same network.

RFC 801 describes how the US Department of Defense organized the replacement of ARPANET's Network Control Program by the new Internet Protocol during January 1983. During the same year, the military systems were removed to a distinct MILNET, and the Domain Name System was invented to manage the names and addresses of computers on the "ARPA Internet". The familiar top-level domains .gov, .mil, .edu, .org, .net, .com, and .int, and the two-letter country code top-level domains were deployed during 1984.

Between 1984 and 1986 the US National Science Foundation created the NSFNET backbone, using TCP/IP, to connect their supercomputing facilities. The combined network became generally known as the Internet.

By the end of 1989 Australia, Germany, Israel, Italy, Japan, Mexico, the Netherlands, New Zealand, and the United Kingdom had connected to the Internet, which now contained over 160,000 hosts.

During 1990, ARPANET formally terminated, and during 1991 the NSF ended its restrictions on commercial use of its part of the Internet. Commercial network providers began to interconnect, extending the Internet.

Today almost all Internet infrastructure is provided and owned by the private sector. Traffic is exchanged between these networks, at major interconnect points, in accordance with established Internet standards and commercial agreements.

Actors

During 1979 the Internet Configuration Control Board was founded by DARPA to oversee the network's development. During 1984 it was renamed the Internet Advisory Board (IAB), and during 1986 it became the Internet Activities Board.

The Internet Engineering Task Force (IETF) was formed during 1986 by the US Government to develop and promote Internet standards. It consisted initially of researchers, but by the end of the year participation was available to anyone, and its business was performed largely by email.

From the early days of the network until his death during 1998, Jon Postel oversaw address allocation and other Internet protocol numbering and assignments in his capacity as Director of the Computer Networks Division at the Information Sciences Institute of the University of Southern California, under a contract from the Dept. of Defense. This function eventually became known as the Internet Assigned Numbers Authority (IANA), and as it expanded to include management of the global Domain Name System (DNS) root servers, a small organization grew. Postel also served as RFC Editor.

Allocation of IP addresses was delegated to four Regional Internet Registries (RIRs):

- American Registry for Internet Numbers (ARIN) for North America
- Réseaux IP Européens - Network Coordination Centre (RIPE NCC) for Europe, the Middle East, and Central Asia
- Asia-Pacific Network Information Centre (APNIC) for Asia and the Pacific region
- Latin American and Caribbean Internet Addresses Registry (LACNIC) for Latin America and the Caribbean region

In 2004 a new RIR, AfriNIC, was created to manage allocations for Africa.

After Jon Postel's death during 1998, the IANA became part of the Internet Corporation for Assigned Names and Numbers (ICANN), a newly created Californian non-profit corporation, initiated during September 1998 by the US Government and awarded a contract by the US Department of Commerce. Initially two board members were elected by the Internet community at large, though this was changed by the rest of the board during 2002 in a little- attended public meeting in Accra, in Ghana.

During 1992 the Internet Society (ISOC) was founded, with a mission to *"assure the open development, evolution and use of the Internet for the benefit of all people throughout the world"*. Its members include individuals (anyone may join) as well as corporations, organizations, governments, and universities. The IAB was renamed the Internet *Architecture* Board, and became part of ISOC. The Internet Engineering Task Force also became part of the ISOC. The IETF is overseen currently by the Internet Engineering Steering Group (IESG), and longer term research is carried on by the Internet Research Task Force and overseen by the Internet Research Steering Group.

During 2002, a restructuring of the Internet Society gave more control to its corporate members.

At the first World Summit on the Information Society (WSIS) in Geneva 2003 the topic of Internet governance was discussed. ICANN's status as a private corporation under

contract to the U.S. government created controversy among other governments, especially Brazil, China, South Africa and some Arab states. Since no general agreement existed even on the definition of what comprised Internet governance, United Nations Secretary General Kofi Annan initiated a Working Group on Internet Governance (WGIG) to clarify the issues and report before the second part of the World Summit on the Information Society in Tunis 2005. After much controversial debate, during which the US delegation refused to consider surrendering the US control of the Root Zone file, participants agreed on a compromise to allow for wider international debate on the policy principles. They agreed to establish an Internet Governance Forum, to be convened by United Nations Secretary General before the end of the second quarter of the year 2006. The Greek government volunteered to host the first such meeting.

Controversy

The position of the US Department of Commerce as the controller of the Internet gradually attracted criticism from those who felt that control should be more international. A hands-off philosophy by the US Dept. of Commerce helped limit this criticism, but this was undermined in 2005 when the Bush administration intervened to help kill the .xxx top level domain proposal.

When the IANA functions were given to a new US non-profit Corporation called ICANN, controversy increased. ICANN's decision-making process was criticised by some observers as being secretive and unaccountable. When the directors' posts which had previously been elected by the "at-large" community of Internet users were abolished, some feared the worst. ICANN stated that they were merely streamlining decision-making processes, and developing a structure suitable for the modern Internet.

Other topics of controversy included the creation and control of generic top-level domains the control of country-code domains, recent proposals for a large increase in ICANN's budget and responsibilities, and a proposed "domain tax" to pay for the increase.

There were also suggestions that individual governments should have more control, or that the International Telecommunication Union or the United Nations should have a function in Internet governance.

Use and culture

E-mail and Usenet

E-mail is often called the killer application of the Internet. However, it actually predates the Internet and was a crucial tool in creating it. E-mail started in 1965 as a way for multiple users of a time-sharing mainframe computer to communicate. Although the history is unclear, among the first systems to have such a facility were SDC's Q32 and MIT's CTSS.

The ARPANET computer network made a large contribution to the evolution of e-mail. There is one report indicating experimental inter-system e-mail transfers on it shortly after ARPANET's creation. In 1971 Ray Tomlinson created what was to become the standard Internet e-mail address format, using the @ sign to separate user names from host names.

A number of protocols were developed to deliver e-mail among groups of time-sharing computers over alternative transmission systems, such as UUCP and IBM's VNET e-mail system. E-mail could be passed this way between a number of networks, including ARPANET, BITNET and NSFNet, as well as to hosts connected directly to other sites via UUCP.

In addition, UUCP allowed the publication of text files that could be read by many others. The News software developed by Steve Daniel and Tom Truscott in 1979 was used to distribute news and bulletin board-like messages. This quickly grew into discussion groups, known as newsgroups, on a wide range of topics. On ARPANET and NSFNet similar discussion groups would form via mailing lists, discussing both technical issues and more culturally focused topics (such as science fiction, discussed on the sflovers mailing list).

During the early years of the Internet, e-mail and similar mechanisms were also fundamental to allow people to access resources that were not available due to the absence of online connectivity. UUCP was often used to distribute files using the 'alt.binary' groups. Also, FTP e-mail gateways allowed people that lived outside the US and Europe to download files using ftp commands written inside e-email messages. The file was encoded, broken in pieces and sent by e-mail; the receiver had to reassemble and decode it later, and it was the only way for people living overseas to download items such as the earlier Linux versions using the slow dial-up connections available at the time. After the popularization of the Web and the HTTP protocol such tools were slowly abandoned.

From gopher to the WWW

As the Internet grew through the 1980s and early 1990s, many people realized the increasing need to be able to find and organize files and information. Projects such as Gopher, WAIS, and the FTP Archive list attempted to create ways to organize distributed data. Unfortunately, these projects fell short in being able to accommodate all the existing data types and in being able to grow without bottlenecks.

One of the most promising user interface paradigms during this period was hypertext. The technology had been inspired by Vannevar Bush's "Memex" and developed through Ted Nelson's research on Project Xanadu and Douglas Engelbart's research on NLS. Many small self-contained hypertext systems had been created before, such as Apple Computer's HyperCard. Gopher became the first commonly-used hypertext interface to the Internet. While Gopher menu items were examples of hypertext, they were not commonly perceived in that way.

This NeXT Computer was used by Sir Tim Berners-Lee at CERN and became the world's first Web server.

In 1989, while working at CERN, Tim Berners-Lee invented a network-based implementation of the hypertext concept. By releasing his invention to public use, he ensured the technology would become widespread. For his work in developing the World Wide Web, Berners-Lee received the Millennium technology prize in 2004. One early popular web browser, modeled after HyperCard, was ViolaWWW.

A potential turning point for the World Wide Web began with the introduction of the Mosaic web browser in 1993, a graphical browser developed by a team at the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign (NCSA-UIUC), led by Marc Andreessen. Funding for Mosaic came from the *High-Performance Computing and Communications Initiative*, a funding program initiated by the *High Performance Computing and Communication Act of 1991* also known as the *Gore Bill*. Indeed, Mosaic's graphical interface soon became more popular than Gopher, which at the time was primarily text-based, and the WWW became the preferred interface for accessing the Internet. (Gore's reference to his role in "creating the Internet", however, was ridiculed in his presidential election campaign.)

Mosaic was eventually superseded in 1994 by Andreessen's Netscape Navigator, which replaced Mosaic as the world's most popular browser. While it held this title for some

time, eventually competition from Internet Explorer and a variety of other browsers almost completely displaced it. Another important event held on January 11, 1994, was *The Superhighway Summit* at UCLA's Royce Hall. This was the "first public conference bringing together all of the major industry, government and academic leaders in the field [and] also began the national dialogue about the *Information Superhighway* and its implications."

24 Hours in Cyberspace, "the largest one-day online event" (February 8, 1996) up to that date, took place on the then-active website, *cyber24.com.* It was headed by photographer Rick Smolan. A photographic exhibition was unveiled at the Smithsonian Institution's National Museum of American History on January 23, 1997, featuring 70 photos from the project.

Search engines

Even before the World Wide Web, there were search engines that attempted to organize the Internet. The first of these was the Archie search engine from McGill University in 1990, followed in 1991 by WAIS and Gopher. All three of those systems predated the invention of the World Wide Web but all continued to index the Web and the rest of the Internet for several years after the Web appeared. There are still Gopher servers as of 2006, although there are a great many more web servers.

As the Web grew, search engines and Web directories were created to track pages on the Web and allow people to find things. The first full-text Web search engine was WebCrawler in 1994. Before WebCrawler, only Web page titles were searched. Another early search engine, Lycos, was created in 1993 as a university project, and was the first to achieve commercial success. During the late 1990s, both Web directories and Web search engines were popular—Yahoo! (founded 1994) and Altavista (founded 1995) were the respective industry leaders. By August 2001, the directory model had begun to give way to search engines, tracking the rise of Google (founded 1998), which had developed new approaches to relevancy ranking. Directory features, while still commonly available, became after-thoughts to search engines.

Database size, which had been a significant marketing feature through the early 2000s, was similarly displaced by emphasis on relevancy ranking, the methods by which search engines attempt to sort the best results first. Relevancy ranking first became a major issue circa 1996, when it became apparent that it was impractical to review full lists of results. Consequently, algorithms for relevancy ranking have continuously improved. Google's PageRank method for ordering the results has received the most press, but all major search engines continually refine their ranking methodologies with a view toward improving the ordering of results. As of 2006, search engine rankings are more important than ever, so much so that an industry has developed ("search engine optimizers", or "SEO") to help web-developers improve their search ranking, and an entire body of case law has developed around matters that affect search engine rankings, such as use of trademarks in metatags.

The sale of search rankings by some search engines has also created controversy among librarians and consumer advocates. As of June 3, 2009, Microsoft launched its own search engine. Bing became immediately popular with the masses searching the internet. It has multiple sites belonging to separate countries e.g. the United States version is different from the Australian version. In the US, Bing ranked 17th among all websites out of over 450,000 websites, up from 5120 the week before the official launch when the website was merely a placeholder. Within the Search Engines category, Bing ranked 4th out of the search engines tracked by Hitwise and Bing Image Search ranked 15th for the week ending June 6, 2009.

Dot-com bubble

Suddenly the low price of reaching millions worldwide, and the possibility of selling to or hearing from those people at the same moment when they were reached, promised to overturn established business dogma in advertising, mail-order sales, customer relationship management, and many more areas. The web was a new killer app—it could bring together unrelated buyers and sellers in seamless and low-cost ways. Visionaries around the world developed new business models, and ran to their nearest venture capitalist. While some of the new entrepreneurs had experience in business in economics, the majority were simply people with ideas, and didn't manage the capital influx prudently. Additionally, many dot-com business plans were predicated on the assumption that by using the Internet, they would bypass the distribution channels of existing businesses and therefore not have to compete with them; when the established businesses with strong existing brands developed their own Internet presence, these hopes were shattered, and the newcomers were left attempting to break into markets dominated by larger, more established businesses. Many did not have the ability to do so.

The dot-com bubble burst on March 10, 2000, when the technology heavy NASDAQ Composite index peaked at 5,048.62 (intra-day peak 5,132.52), more than double its value just a year before. By 2001, the bubble's deflation was running full speed. A majority of the dot-coms had ceased trading, after having burnt through their venture capital and IPO capital, often without ever making a profit.

Online population forecast

A study conducted by JupiterResearch anticipates that a 38 percent increase in the number of people with online access will mean that, by 2011, 22 percent of the Earth's population will surf the Internet regularly. The report says 1.1 billion people have regular Web access. For the study, JupiterResearch defined online users as people who regularly access the Internet from dedicated Internet-access devices, which exclude cellular telephones.

Mobile phones and the Internet

The first mobile phone with Internet connectivity was the Nokia 9000 Communicator, launched in Finland in 1996. The viability of Internet services access on mobile phones

was limited until prices came down from that model and network providers started to develop systems and services conveniently accessible on phones. NTT DoCoMo in Japan launched the first mobile Internet service, i-Mode, in 1999 and this is considered the birth of the mobile phone Internet services. In 2001 the mobile phone email system by Research in Motion for their Blackberry product was launched in America. To make efficient use of the small screen and tiny keypad and one-handed operation typical of mobile phones, a specific document and networking model was created for mobile devices, the Wireless Application Protocol (WAP). Most mobile device Internet services operate using WAP. The growth of mobile phone services was initially a primarily Asian phenomenon with Japan, South Korea and Taiwan all soon finding the majority of their Internet users accessing resources by phone rather than by PC. Developing countries followed, with India, South Africa, Kenya, Philippines and Pakistan all reporting that the majority of their domestic users accessed the Internet from a mobile phone rather than a PC. The European and North American use of the Internet was influenced by a large installed base of personal computers, and the growth of mobile phone Internet access was more gradual, but had reached national penetration levels of 20–30% in most Western countries. The cross-over occurred in 2008, when more Internet access devices were mobile phones than personal computers. In many parts of the developing world, the ratio is as much as 10 mobile phone users to one PC user.

Historiography

Some concerns have been raised over the historiography of the Internet's development. Specifically that it is hard to find documentation of much of the Internet's development, for several reasons, including a lack of centralized documentation for much of the early developments that led to the Internet.

"The Arpanet period is somewhat well documented because the corporation in charge - BBN - left a physical record. Moving into the NSFNET era, it became an extraordinarily decentralized process. The record exists in people's basements, in closets. [...] So much of what happened was done verbally and on the basis of individual trust."
—Doug Gale (2007),

Chapter 3

World Wide Web

World Wide Web

The Web's historic logo designed by Robert Cailliau

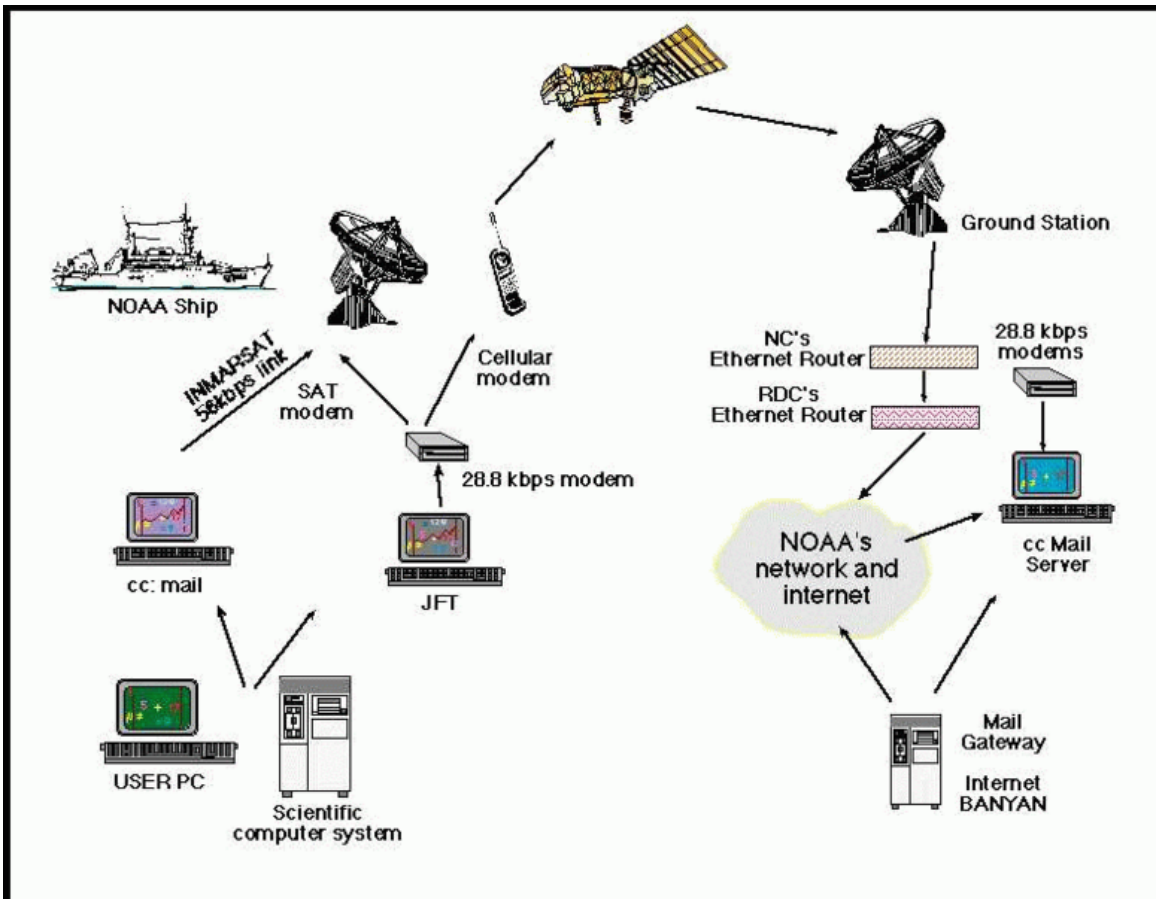
Inventor	Sir Tim Berners-Lee
Launch year	1990
Company	CERN
Availability	Worldwide

The **World Wide Web**, abbreviated as **WWW** and commonly known as **the Web**, is a system of interlinked hypertext documents accessed via the Internet. With a web browser, one can view web pages that may contain text, images, videos, and other multimedia and navigate between them by using hyperlinks. Using concepts from earlier hypertext systems, English engineer and computer scientist Sir Tim Berners-Lee, now the Director of the World Wide Web Consortium, wrote a proposal in March 1989 for what would eventually become the World Wide Web. At CERN in Geneva, Switzerland, Berners-Lee and Belgian computer scientist Robert Cailliau proposed in 1990 to use "HyperText [...]

to link and access information of various kinds as a web of nodes in which the user can browse at will", and publicly introduced the project in December.

"The World-Wide Web (W3) was developed to be a pool of human knowledge, and human culture, which would allow collaborators in remote sites to share their ideas and all aspects of a common project."

History of the World Wide Web



Today, the Web and the Internet allow connectivity from literally everywhere on earth—even ships at sea and in outer space.

The **World Wide Web** ("**WWW**" or simply the "**Web**") is a global information medium which users can read and write via computers connected to the Internet. The term is often mistakenly used as a synonym for the Internet itself, but the Web is a service that operates over the Internet, as e-mail does. The history of the Internet dates back significantly further than that of the World Wide Web.

The hypertext portion of the Web in particular has an intricate intellectual history; notable influences and precursors include Vannevar Bush's Memex, IBM's Generalized Markup Language, and Ted Nelson's Project Xanadu.

The concept of a home-based global information system goes at least as far back as "A Logic Named Joe", a 1946 short story by Murray Leinster, in which computer terminals, called "logics," were in every home. Although the computer system in the story is centralized, the story captures some of the feeling of the ubiquitous information explosion driven by the Web.

1980–1991: Development of the World Wide Web

The NeXTcube used by Tim Berners-Lee at CERN became the first Web server

In 1980, Tim Berners-Lee, an independent contractor at the European Organization for Nuclear Research (CERN), Switzerland, built ENQUIRE, as a personal database of people and software models, but also as a way to play with hypertext; each new page of information in ENQUIRE had to be linked to an existing page.

In 1984 Berners-Lee returned to CERN, and considered its problems of information presentation: physicists from around the world needed to share data, and with no common machines and no common presentation software. He wrote a proposal in March 1989 for "a large hypertext database with typed links", but it generated little interest. His boss, Mike Sendall, encouraged Berners-Lee to begin implementing his system on a newly acquired NeXT workstation. He considered several names, including *Information Mesh*,

The Information Mine (turned down as it abbreviates to TIM, the WWW's creator's name) or *Mine of Information* (turned down because it abbreviates to MOI which is "Me" in French), but settled on *World Wide Web*.

Robert Cailliau, Jean-François Abramatic and Tim Berners-Lee at the 10th anniversary of the WWW Consortium.

He found an enthusiastic collaborator in Robert Cailliau, who rewrote the proposal (published on November 12, 1990) and sought resources within CERN. Berners-Lee and Cailliau pitched their ideas to the European Conference on Hypertext Technology in September 1990, but found no vendors who could appreciate their vision of marrying hypertext with the Internet.

By Christmas 1990, Berners-Lee had built all the tools necessary for a working Web: the HyperText Transfer Protocol (HTTP) 0.9, the HyperText Markup Language (HTML), the first Web browser (named WorldWideWeb, which was also a Web editor), the first HTTP server software (later known as CERN httpd), the first web server and the first Web pages that described the project itself. The browser could access Usenet newsgroups and FTP files as well. However, it could run only on the NeXT; Nicola Pellow therefore created a simple text browser that could run on almost any computer called the Line Mode Browser. To encourage use within CERN, Bernd Pollermann put the CERN telephone directory on the web — previously users had had to log onto the mainframe in order to look up phone numbers.

Tim Berners-Lee's account of the exact locations at CERN where the Web was invented, is here.

Paul Kunz from the Stanford Linear Accelerator Center visited CERN in September 1991, and was captivated by the Web. He brought the NeXT software back to SLAC, where librarian Louise Addis adapted it for the VM/CMS operating system on the IBM mainframe as a way to display SLAC's catalog of online documents; this was the first web server outside of Europe and the first in North America.

On August 6, 1991, Berners-Lee posted a short summary of the World Wide Web project on the alt.hypertext newsgroup. This date also marked the debut of the Web as a publicly available service on the Internet.

The WorldWideWeb (WWW) project aims to allow all links to be made to any information anywhere. [...] The WWW project was started to allow high energy physicists to share data, news, and documentation. We are very interested in spreading the web to other areas, and having gateway servers for other data. Collaborators welcome!" —from Tim Berners-Lee's first message

An early CERN-related contribution to the Web was the parody band Les Horribles Cernettes, whose promotional image is believed to be among the Web's first five pictures.

1992–1995: Growth of the WWW

In keeping with its birth at CERN, early adopters of the World Wide Web were primarily university-based scientific departments or physics laboratories such as Fermilab and SLAC.

Early websites intermingled links for both the HTTP web protocol and the then-popular Gopher protocol, which provided access to content through hypertext menus presented as a file system rather than through HTML files. Some sites were also indexed by WAIS, enabling users to submit full-text searches similar to the capability later provided by search engines.

There was still no graphical browser available for computers besides the NeXT. This gap was filled in April 1992 with the release of Erwise, an application developed at Helsinki University of Technology, and in May by ViolaWWW, created by Pei-Yuan Wei, which included advanced features such as embedded graphics, scripting, and animation. ViolaWWW was originally an application for HyperCard. Both programs ran on the X Window System for Unix.

Students at the University of Kansas adapted an existing text-only hypertext browser, Lynx, to access the web. Lynx was available on Unix and DOS, and some web designers, unimpressed with glossy graphical websites, held that a website not accessible through Lynx wasn't worth visiting.

Early browsers

The turning point for the World Wide Web was the introduction of the Mosaic web browser in 1993, a graphical browser developed by a team at the National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana-Champaign (UIUC), led by Marc Andreessen. Funding for Mosaic came from the *High-Performance Computing and Communications Initiative*, a funding program initiated by then-Senator Al Gore's *High Performance Computing and Communication Act of 1991* also known as the *Gore Bill*.

The origins of Mosaic had begun in 1992. In November 1992, the NCSA at the University of Illinois (UIUC) established a website. In December 1992, Andreessen and Eric Bina, students attending UIUC and working at the NCSA, began work on Mosaic. They released an X Window browser in February 1993. It gained popularity due to its strong support of integrated multimedia, and the authors' rapid response to user bug reports and recommendations for new features.

The first Microsoft Windows browser was Cello, written by Thomas R. Bruce for the Legal Information Institute at Cornell Law School to provide legal information, since more lawyers had more access to Windows than to Unix. Cello was released in June 1993.

After graduation from UIUC, Andreessen and James H. Clark, former CEO of Silicon Graphics, met and formed Mosaic Communications Corporation to develop the Mosaic browser commercially. The company changed its name to Netscape in April 1994, and the browser was developed further as Netscape Navigator.

Web organization

In May 1994 the first International WWW Conference, organized by Robert Cailliau, was held at CERN; the conference has been held every year since. In April 1993 CERN had agreed that anyone could use the Web protocol and code royalty-free; this was in part a reaction to the perturbation caused by the University of Minnesota announcing that it would begin charging license fees for its implementation of the Gopher protocol.

In September 1994, Berners-Lee founded the World Wide Web Consortium (W3C) at the Massachusetts Institute of Technology with support from the Defense Advanced Research Projects Agency (DARPA) and the European Commission. It comprised various companies that were willing to create standards and recommendations to improve the quality of the Web. Berners-Lee made the Web available freely, with no patent and no royalties due. The W3C decided that their standards must be based on royalty-free technology, so they can be easily adopted by anyone.

By the end of 1994, while the total number of websites was still minute compared to present standards, quite a number of notable websites were already active, many of whom are the precursors or inspiring examples of today's most popular services.

1996–1998: Commercialization of the WWW

By 1996 it became obvious to most publicly traded companies that a public Web presence was no longer optional. Though at first people saw mainly the possibilities of free publishing and instant worldwide information, increasing familiarity with two-way communication over the "Web" led to the possibility of direct Web-based commerce (e-commerce) and instantaneous group communications worldwide. More dotcoms, displaying products on hypertext webpages, were added into the Web.

1999–2001: "Dot-com" boom and bust

The low interest rates in 1998–99 helped increase the start-up capital amounts. Although a number of these new entrepreneurs had realistic plans and administrative ability, most of them lacked these characteristics but were able to sell their ideas to investors because of the novelty of the dot-com concept.

Historically, the dot-com boom can be seen as similar to a number of other technology-inspired booms of the past including railroads in the 1840s, automobiles in the early 20th century, radio in the 1920s, television in the 1940s, transistor electronics in the 1950s, computer time-sharing in the 1960s, and home computers and biotechnology in the early 1980s.

In 2001 the bubble burst, and many dot-com startups went out of business after burning through their venture capital and failing to become profitable. Many others, however, did survive and thrive in the early 21st century. Many companies which began as online retailers blossomed and became highly profitable. More conventional retailers found online merchandising to be a profitable additional source of revenue. While some online entertainment and news outlets failed when their seed capital ran out, others persisted and eventually became economically self-sufficient. Traditional media outlets (newspaper publishers, broadcasters and cablecasters in particular) also found the Web to be a useful and profitable additional channel for content distribution, and an additional vehicle to generate advertising revenue. The sites that survived and eventually prospered after the bubble burst had two things in common; a sound business plan, and a niche in the marketplace that was, if not unique, particularly well-defined and well-served.

2002–present: The Web becomes ubiquitous

In the aftermath of the dot-com bubble, telecommunications companies had a great deal of overcapacity as many Internet business clients went bust. That, plus ongoing investment in local cell infrastructure kept connectivity charges low, and helping to make high-speed Internet connectivity more affordable. During this time, a handful of companies found success developing business models that helped make the World Wide Web a more compelling experience. These include airline booking sites, Google's search engine and its profitable approach to simplified, keyword-based advertising, as well as ebay's do-it-yourself auction site and Amazon.com's online department store.

This new era also begot social networking websites, such as MySpace and Facebook, which, though unpopular at first, very rapidly gained acceptance in becoming a major part of youth culture.

Web 2.0

Beginning in 2002, new ideas for sharing and exchanging content ad hoc, such as Weblogs and RSS, rapidly gained acceptance on the Web. This new model for information exchange, primarily featuring DIY user-edited and generated websites, was coined Web 2.0.

The Web 2.0 boom saw many new service-oriented startups catering to a new, democratized Web. Some believe it will be followed by the full realization of a Semantic Web.

Tim Berners-Lee originally expressed the vision of the Semantic Web as follows:

I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A 'Semantic Web', which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The 'intelligent agents' people have touted for ages will finally materialize.

– *Tim Berners-Lee, 1999*

Predictably, as the World Wide Web became easier to query, attained a higher degree of usability, and shed its esoteric reputation, it gained a sense of organization and unsophistication which opened the floodgates and ushered in a rapid period of popularization. In 2005, 3 ex-PayPal employees formed a video viewing website called YouTube. Only a year later, YouTube was proven the most quickly popularized website in history, and even started a new concept of user-submitted content in major events, as in the CNN-YouTube Presidential Debates.

The popularity of YouTube and similar services, combined with the increasing availability and affordability of high-speed connections has made video content far more common on all kinds of websites. Many video-content hosting and creation sites provide an easy means for their videos to be embedded on third party websites without payment or permission.

This combination of more user-created or edited content, and easy means of sharing content, such as via RSS widgets and video embedding, has led to many sites with a typical "Web 2.0" feel. They have articles with embedded video, user-submitted comments below the article, and RSS boxes to the side, listing some of the latest articles from other sites.

Continued extension of the World Wide Web has focused on connecting devices to the Internet, coined Intelligent Device Management. As Internet connectivity becomes ubiquitous, manufacturers have started to leverage the expanded computing power of their devices to enhance their usability and capability. Through Internet connectivity, manufacturers are now able to interact with the devices they have sold and shipped to their customers, and customers are able to interact with the manufacturer (and other providers) to access new content.

Lending credence to the idea of the ubiquity of the web, Web 2.0 has found a place in the global English lexicon. On June 10, 2009 the Global Language Monitor declared it to be the one-millionth English word.

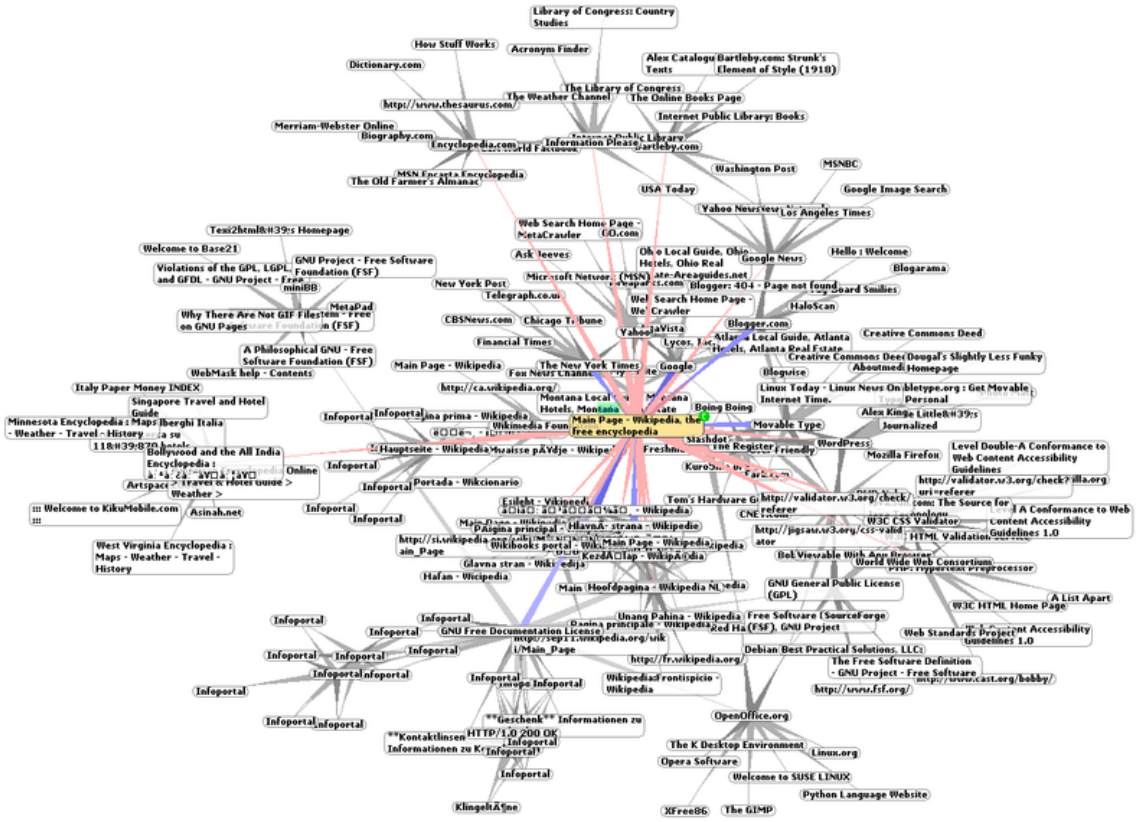
Function

The terms Internet and World Wide Web are often used in every-day speech without much distinction. However, the Internet and the World Wide Web are not one and the same. The Internet is a global system of interconnected computer networks. In contrast, the Web is one of the services that runs on the Internet. It is a collection of interconnected documents and other resources, linked by hyperlinks and URLs. In short, the Web is an application running on the Internet. Viewing a web page on the World Wide Web normally begins either by typing the URL of the page into a web browser, or by following a hyperlink to that page or resource. The web browser then initiates a series of communication messages, behind the scenes, in order to fetch and display it.

First, the server-name portion of the URL is resolved into an IP address using the global, distributed Internet database known as the Domain Name System (DNS). This IP address is necessary to contact the Web server. The browser then requests the resource by sending an HTTP request to the Web server at that particular address. In the case of a typical web page, the HTML text of the page is requested first and parsed immediately by the web browser, which then makes additional requests for images and any other files that complete the page image. Statistics measuring a website's popularity are usually based either on the number of page views or associated server 'hits' (file requests) that take place.

While receiving these files from the web server, browsers may progressively render the page onto the screen as specified by its HTML, Cascading Style Sheets (CSS), or other page composition languages. Any images and other resources are incorporated to produce the on-screen web page that the user sees. Most web pages contain hyperlinks to other related pages and perhaps to downloadable files, source documents, definitions and other web resources. Such a collection of useful, related resources, interconnected via hypertext links is dubbed a *web* of information. Publication on the Internet created what Tim Berners-Lee first called the *WorldWideWeb* (in its original CamelCase, which was subsequently discarded) in November 1990.

Linking



Graphic representation of a minute fraction of the WWW, demonstrating hyperlinks

Over time, many web resources pointed to by hyperlinks disappear, relocate, or are replaced with different content. This makes hyperlinks obsolete, a phenomenon referred to in some circles as link rot and the hyperlinks affected by it are often called dead links. The ephemeral nature of the Web has prompted many efforts to archive web sites. The Internet Archive, active since 1996, is one of the best-known efforts.

Dynamic updates of web pages

JavaScript is a scripting language that was initially developed in 1995 by Brendan Eich, then of Netscape, for use within web pages. The standardized version is ECMAScript. To overcome some of the limitations of the page-by-page model described above, some web applications also use Ajax (asynchronous JavaScript and XML). JavaScript is delivered with the page that can make additional HTTP requests to the server, either in response to user actions such as mouse-clicks, or based on lapsed time. The server's responses are used to modify the current page rather than creating a new page with each response. Thus the server only needs to provide limited, incremental information. Since multiple Ajax requests can be handled at the same time, users can interact with a page even while data is being retrieved. Some web applications regularly poll the server to ask if new information is available.

WWW prefix

Many domain names used for the World Wide Web begin with *www* because of the long-standing practice of naming Internet hosts (servers) according to the services they provide. The hostname for a web server is often *www*, in the same way that it may be *ftp* for an FTP server, and *news* or *nntp* for a USENET news server. These host names appear as Domain Name System (DNS) subdomain names, as in `www.example.com`. The use of 'www' as a subdomain name is not required by any technical or policy standard; indeed, the first ever web server was called `nxoc01.cern.ch`, and many web sites exist without it. Many established websites still use 'www', or they invent other subdomain names such as 'www2', 'secure', etc. Many such web servers are set up such that both the domain root and the *www* subdomain refer to the same site; others require one form or the other, or they may map to different web sites.

The use of a subdomain name is useful for load balancing incoming web traffic by creating a CNAME record that points to a cluster of web servers. Since, currently, only a subdomain can be cname'ed the same result cannot be achieved by using the bare domain root.

When a user submits an incomplete website address to a web browser in its address bar input field, some web browsers automatically try adding the prefix "www" to the beginning of it and possibly ".com", ".org" and ".net" at the end, depending on what might be missing. This feature started appearing in early versions of Mozilla Firefox, when it still had the working title 'Firebird' in early 2003. It is reported that Microsoft was granted a US patent for the same idea in 2008, but only for mobile devices.

The scheme specifier in URIs refers to the Hypertext Transfer Protocol and to HTTP Secure respectively and so defines the communication protocol to be used for the request and response. The HTTP protocol is fundamental to the operation of the World Wide Web, and the encryption involved in HTTPS adds an essential layer if confidential information such as passwords or banking information are to be exchanged over the public Internet. Web browsers usually prepend the scheme to URLs too, if omitted.

In English, *www* is pronounced by individually pronouncing the name of characters (*double-u double-u double-u*). Although some technical users pronounce it *dub-dub-dub* this is not widespread. The English writer Douglas Adams once quipped in The Independent on Sunday (1999): "The World Wide Web is the only thing I know of whose shortened form takes three times longer to say than what it's short for," with Stephen Fry later pronouncing it in his "Podgrammes" series of podcasts as "wuh wuh wuh." In Mandarin Chinese, *World Wide Web* is commonly translated via a phono-semantic matching to *wàn wéi wăng* (万维网), which satisfies *www* and literally means "myriad dimensional net", a translation that very appropriately reflects the design concept and proliferation of the World Wide Web. Tim Berners-Lee's web-space states that *World Wide Web* is officially spelled as three separate words, each capitalized, with no intervening hyphens.

Privacy

Computer users, who save time and money, and who gain conveniences and entertainment, may or may not have surrendered the right to privacy in exchange for using a number of technologies including the Web. Worldwide, more than a half billion people have used a social network service, and of Americans who grew up with the Web, half created an online profile and are part of a generational shift that could be changing norms. Facebook progressed from U.S. college students to a 70% non-U.S. audience, and in 2009 estimated that only 20% of its members use privacy settings. In 2010 (six years after co-founding the company), Mark Zuckerberg wrote, "we will add privacy controls that are much simpler to use".

Privacy representatives from 60 countries have resolved to ask for laws to complement industry self-regulation, for education for children and other minors who use the Web, and for default protections for users of social networks. They also believe data protection for personally identifiable information benefits business more than the sale of that information. Users can opt-in to features in browsers to clear their personal histories locally and block some cookies and advertising networks but they are still tracked in websites' server logs, and particularly web beacons. Berners-Lee and colleagues see hope in accountability and appropriate use achieved by extending the Web's architecture to policy awareness, perhaps with audit logging, reasoners and appliances.

In exchange for providing free content, vendors hire advertisers who spy on Web users and base their business model on tracking them. Since 2009, they buy and sell consumer data on exchanges (lacking a few details that could make it possible to de-anonymize, or identify an individual). Hundreds of millions of times per day, Lotame Solutions captures what users are typing in real time, and sends that text to OpenAmplify who then tries to determine, to quote a writer at *The Wall Street Journal*, "what topics are being discussed, how the author feels about those topics, and what the person is going to do about them".

Microsoft backed away in 2008 from its plans for strong privacy features in Internet Explorer, leaving its users (50% of the world's Web users) open to advertisers who may make assumptions about them based on only *one click* when they visit a website. Among services paid for by advertising, Yahoo! could collect the most data about users of commercial websites, about 2,500 bits of information per month about each typical user of its site and its affiliated advertising network sites. Yahoo! was followed by MySpace with about half that potential and then by AOL–TimeWarner, Google, Facebook, Microsoft, and eBay.

Security

The Web has become criminals' preferred pathway for spreading malware. Cybercrime carried out on the Web can include identity theft, fraud, espionage and intelligence gathering. Web-based vulnerabilities now outnumber traditional computer security concerns, and as measured by Google, about one in ten web pages may contain malicious

code. Most Web-based attacks take place on legitimate websites, and most, as measured by Sophos, are hosted in the United States, China and Russia. The most common of all malware threats is SQL injection attacks against websites. Through HTML and URIs the Web was vulnerable to attacks like cross-site scripting (XSS) that came with the introduction of JavaScript and were exacerbated to some degree by Web 2.0 and Ajax web design that favors the use of scripts. Today by one estimate, 70% of all websites are open to XSS attacks on their users.

Proposed solutions vary to extremes. Large security vendors like McAfee already design governance and compliance suites to meet post-9/11 regulations, and some, like Finjan have recommended active real-time inspection of code and all content regardless of its source. Some security vendors like Commtouch monitor new threats and provide reporting tools for malware outbreaks, spam, and zombie trends along with real-time outbreak monitors.

Some have argued that for enterprise to see security as a business opportunity rather than a cost center, "ubiquitous, always-on digital rights management" enforced in the infrastructure by a handful of organizations must replace the hundreds of companies that today secure data and networks. Jonathan Zittrain has said users sharing responsibility for computing safety is far preferable to locking down the Internet.

Standards

Web standards

Web standards is a general term for the formal standards and other technical specifications that define and describe aspects of the World Wide Web. In recent years, the term has been more frequently associated with the trend of endorsing a set of standardized best practices for building web sites, and a philosophy of web design and development that includes those methods.

Many interdependent standards and specifications, some of which govern aspects of the Internet, not just the World Wide Web, directly or indirectly affect the development and administration of web sites and web services. Considerations include the interoperability, accessibility and usability of web pages and web sites. Web standards, in the broader sense, consist of the following:

- *Recommendations* published by the World Wide Web Consortium (W3C)
- *Internet standard* (STD) documents published by the Internet Engineering Task Force (IETF)
- *Request for Comments* (RFC) documents published by the Internet Engineering Task Force
- *Standards* published by the International Organization for Standardization (ISO)
- *Standards* published by Ecma International (formerly ECMA)
- *The Unicode Standard* and various *Unicode Technical Reports* (UTRs) published by the Unicode Consortium

- Name and number registries maintained by the Internet Assigned Numbers Authority (IANA)

Common usage

When a web site or web page is described as complying with web standards, it usually means that the site or page has valid HTML, CSS and JavaScript. The HTML should also meet accessibility and semantic guidelines.

When web standards are discussed, the following publications are typically seen as foundational:

- Recommendations for markup languages, such as Hypertext Markup Language (HTML), Extensible Hypertext Markup Language (XHTML), Scalable Vector Graphics (SVG), and XForms, from W3C.
- Recommendations for stylesheets, especially Cascading Style Sheets (CSS), from W3C.
- Standards for ECMAScript, more commonly JavaScript, from Ecma International.
- Recommendations for Document Object Models (DOM), from W3C.
- Properly formed names and addresses for the page and all other resources referenced from it (URIs), based upon RFC 2396, from IETF.
- Proper use of HTTP and MIME to deliver the page, return data from it and to request other resources referenced in it, based on RFC 2616, from IETF.

Web accessibility is normally based upon the Web Content Accessibility Guidelines published by the W3C's Web Accessibility Initiative.

Work in the W3C toward the Semantic Web is currently focused by publications related to the Resource Description Framework (RDF), Gleaning Resource Descriptions from Dialects of Languages (GRDDL) and Web Ontology Language (OWL).

Standards publications and bodies

A W3C Recommendation is a specification or set of guidelines that, after extensive consensus-building, has received the endorsement of W3C Members and the Director.

An IETF Internet Standard is characterized by a high degree of technical maturity and by a generally held belief that the specified protocol or service provides significant benefit to the Internet community. A specification that reaches the status of Standard is assigned a number in the IETF STD series while retaining its original IETF RFC number.

Non-standard and vendor-proprietary pressures

In the current Working Draft of the HTML 5 proposed standard document, the W3C has a section entitled "Relationship to Flash, Silverlight, XUL and similar proprietary

languages" that says, "In contrast with proprietary languages, this specification is intended to define an openly-produced, vendor-neutral language, to be implemented in a broad range of competing products, across a wide range of platforms and devices. This enables developers to write applications that are not limited to one vendor's implementation or language. Furthermore, while writing applications that target vendor-specific platforms necessarily introduces a cost that application developers and their customers or users will face if they are forced to switch (or desire to switch) to another vendor's platform, using an openly-produced and vendor neutral language means that application authors can switch vendors with little to no cost."

Web development tools

Many websites are designed using WYSIWYG HTML-generation programs such as Adobe Dreamweaver or Microsoft FrontPage. Microsoft FrontPage often generates non-standard HTML by default, hindering the work of the World Wide Web Consortium in promulgating standards, specifically with XHTML and Cascading Style Sheets (CSS), which are used for page layout. Dreamweaver and other more modern Microsoft HTML development tools such as Microsoft Expression Web and Microsoft Visual Studio conform to the W3C standards.

Accessibility

Web accessibility

Web accessibility refers to the inclusive practice of making websites usable by people of all abilities and disabilities. When sites are correctly designed, developed and edited, all users can have equal access to information and functionality. For example, when a site is coded with semantically meaningful HTML, with textual equivalents provided for images and with links named meaningfully, this helps blind users using text-to-speech software and/or text-to-Braille hardware. When text and images are large and/or enlargable, it is easier for users with poor sight to read and understand the content. When links are underlined (or otherwise differentiated) as well as coloured, this ensures that color blind users will be able to notice them. When clickable links and areas are large, this helps users who cannot control a mouse with precision. When pages are coded so that users can navigate by means of the keyboard alone, or a single switch access device alone, this helps users who cannot use a mouse or even a standard keyboard. When videos are closed captioned or a sign language version is available, deaf and hard of hearing users can understand the video. When flashing effects are avoided or made optional, users prone to seizures caused by these effects are not put at risk. And when content is written in plain language and illustrated with instructional diagrams and animations, users with dyslexia and learning difficulties are better able to understand the content. When sites are correctly built and maintained, all of these users can be accommodated while not impacting on the usability of the site for non-disabled users.

The needs that Web accessibility aims to address include:

- **Visual:** Visual impairments including blindness, various common types of low vision and poor eyesight, various types of color blindness;
- **Motor/Mobility:** e.g. difficulty or inability to use the hands, including tremors, muscle slowness, loss of fine muscle control, etc., due to conditions such as Parkinson's Disease, muscular dystrophy, cerebral palsy, stroke;
- **Auditory:** Deafness or hearing impairments, including individuals who are hard of hearing;
- **Seizures:** Photoepileptic seizures caused by visual strobe or flashing effects.
- **Cognitive/Intellectual:** Developmental disabilities, learning disabilities (dyslexia, dyscalculia, etc.), and cognitive disabilities of various origins, affecting memory, attention, developmental "maturity," problem-solving and logic skills, etc.;

Assistive technologies used for web browsing

Individuals living with a disability use assistive technologies such as the following to enable and assist web browsing:

- Screen reader software, which can read out, using synthesized speech, either selected elements of what is being displayed on the monitor (helpful for users with reading or learning difficulties), or which can read out everything that is happening on the computer (used by blind and vision impaired users).
- Braille terminals, consisting of a Refreshable Braille display which renders text as Braille characters (usually by means of raising pegs through holes in a flat surface) and either a QWERTY or Braille keyboard.
- Screen magnification software, which enlarges what is displayed on the computer monitor, making it easier to read for vision impaired users.
- Speech recognition software that can accept spoken commands to the computer, or turn dictation into grammatically correct text - useful for those who have difficulty using a mouse or a keyboard.
- Keyboard overlays, which can make typing easier and more accurate for those who have motor control difficulties.

Guidelines on accessible web design

Web Content Accessibility Guidelines

In 1999 the Web Accessibility Initiative, a project by the World Wide Web Consortium (W3C), published the Web Content Accessibility Guidelines WCAG 1.0. In recent years, these have been widely accepted as the definitive guidelines on how to create accessible websites.

On 11 December 2008, the WAI released the WCAG 2.0 as a Recommendation. WCAG 2.0 aims to be up to date and more technology neutral.

Criticism of WAI guidelines

For a general criticism of the W3C process, read Putting the user at the heart of the W3C process. There was a formal objection to WCAG's original claim that WCAG 2.0 will address requirements for people with learning disabilities and cognitive limitations headed by Lisa Seeman and signed by 40 organisations and people. In articles such as WCAG 2.0: The new W3C guidelines evaluated, To Hell with WCAG 2.0 and Testability Costs Too Much, the WAI has been criticised for allowing WCAG 1.0 to get increasingly out of step with today's technologies and techniques for creating and consuming web content, for the slow pace of development of WCAG 2.0, for making the new guidelines difficult to navigate and understand, and other argued failings.

Other guidelines

Canada

Canada has the Common Look and Feel Standards requiring federal government internet websites to meet Web Content Accessibility Guidelines (WCAG) 1.0 Checkpoints Priorities 1 and 2 (Double A conformance level). The standards have existed since 2000 and were updated in 2007.

Philippines

As part of the Web Accessibility Initiatives in the Philippines, the government through the National Council for the Welfare of Disabled Persons (NCWDP) board approved the recommendation of forming an adhoc or core group of webmasters that will help in the implementation of the Biwako Millennium Framework set by the UNESCAP.

The Philippines was also the place where the Interregional Seminar and Regional Demonstration Workshop on Accessible Information and Communications Technologies (ICT) to Persons with Disabilities was held where eleven countries from Asia - Pacific were represented. The Manila Accessible Information and Communications Technologies Design Recommendations was drafted and adopted in 2003.

Spain

In Spain, UNE 139803 is the norm entrusted to regulate web accessibility. This standard is based on Web Content Accessibility Guidelines 1.0.

Sweden

In Sweden, Verva, the Swedish Administrative Development Agency is responsible for a set of guidelines for Swedish public sector web sites. Through the guidelines, Web

accessibility is presented as an integral part of the overall development process and not as a separate issue.

The Swedish guidelines contain criteria which cover the entire lifecycle of a website; from its conception to the publication of live web content. These criteria address several areas which should be considered, including:

- accessibility
- usability
- web standards
- privacy issues
- information architecture
- developing content for the web
- Content Management Systems (CMS) / authoring tools selection.
- development of web content for mobile devices.

An English translation was released in April 2008: Swedish National Guidelines for Public Sector Websites

The translation is based on the latest version of Guidelines which was released in 2006.

United Kingdom

In the UK, the Disability Rights Commission (DRC) in collaboration with BSI have published Pas 78 which outlines good practice in commissioning accessible websites.

Japan

Web Content Accessibility Guidelines in Japan was established in 2004 as JIS (Japanese Industrial Standards) X 8341-3. JIS X 8341-3 will be revised within 2009 by adopting WCAG 2.0. New version will have the same 4 principles, 12 guidelines, and 61 success criteria as WCAG 2.0 has.

Essential Components of Web Accessibility

In order for the web to be accessible, 7 components must be included:

1. the content on Web pages must be natural information (text, images, and sound),
2. Web browsers and media players,
3. assistive technologies,
4. users' knowledge and experience using the Web,
5. developers,
6. authoring tools
7. evaluation tools

These components interact with each other to create an environment that is accessible to people with disabilities.

Web **developers** usually use **authoring tools** and evaluation tools to create Web **content**. **People** ("**users**") use Web **browsers**, **media players**, **assistive technologies** or other "**user agents**" to get and interact with the **content**."

Guidelines for Different Components

Authoring Tool Accessibility Guidelines (ATAG)

- ATAG contains 28 checkpoints that provide guidance on:
  - producing accessible output that meets standards and guidelines
  - promoting the content author for accessibility-related information
  - providing ways of checking and correcting inaccessible content
  - integrating accessibility in the overall look and feel
  - making the authoring tool itself accessible to people with disabilities

Web Content Accessibility Guidelines

Web Content Accessibility Guidelines (**WCAG**) are part of a series of Web accessibility guidelines published by the W3C's Web Accessibility Initiative. They consist of a set of guidelines on making content accessible, primarily for disabled users, but also for all user agents, including highly limited devices, such as mobile phones. The current version is 2.0.

WCAG 1.0

The WCAG 1.0 were published and became a W3C recommendation on May 5, 1999. They have since been superseded by WCAG 2.0

WCAG 1.0 has three *priority levels:*

- Priority 1: Web developers **must** satisfy these requirements, otherwise it will be impossible for one or more groups to access the Web content. Conformance to this level is described as *A*.
- Priority 2: Web developers **should** satisfy these requirements, otherwise some groups will find it difficult to access the Web content. Conformance to this level is described as *AA* or *Double-A*.
- Priority 3: Web developers **may** satisfy these requirements, in order to make it easier for some groups to access the Web content. Conformance to this level is described as *AAA* or *Triple-A*.

WCAG Samurai

In February 2008, The WCAG Samurai, a group of developers independent of the W3C, and led by Joe Clark, published corrections for, and extensions to, the WCAG1.0.

WCAG 2.0

WCAG 2.0 was published as a W3C Recommendation on December 11, 2008. The lengthy consultation process prior to this encouraged participation in editing (and responding to the hundreds of comments) by the Working Group, with diversity assured by inclusion of accessibility experts and members of the disability community.

The Web Accessibility Initiative is also working on guidance for migrating from WCAG 1.0 to WCAG 2.0. A comparison of WCAG 1.0 checkpoints and WCAG 2.0 success criteria is already available.

WCAG 2.0 uses the same three *levels of conformance* as WCAG 1.0, but has redefined them. The WCAG working group maintains an extensive list of web accessibility techniques and common failure cases for WCAG 2.0.

User Agent Accessibility Guidelines (UAAG)

- UAAG contains a comprehensive set of checkpoints that cover:
  - access to all content
  - user control over how content is rendered
  - user control over the user interface
  - standard programming interfaces

Legally-required web accessibility

A growing number of countries around the world have introduced legislation which either directly addresses the need for websites and other forms of communication to be accessible to people with disabilities, or which addresses the more general requirement for people with disabilities not to be discriminated against.

Australia

In 2000, an Australian blind man won a court case against the Sydney Organizing Committee of the Olympic Games (SOCOG). This was the first successful case under Disability Discrimination Act 1992 because SOCOG had failed to make their official website, Sydney Olympic Games, adequately accessible to blind users. The Human Rights and Equal Opportunity Commission (HREOC) also published World Wide Web Access: Disability Discrimination Act Advisory Notes. All Governments in Australia also have policies and guidelines that require accessible public websites; Vision Australia maintain a complete list of Australian web accessibility policies.

Ireland

In Ireland, the Disability Act 2005 was supplemented with the National Disability Authority's Code of Practice on Accessible Public Services in July 2006. It is a practical guide to help all Government Departments and nearly 500 public bodies to comply with their obligations under the Disability Act 2005.

United Kingdom

In the UK, the Disability Discrimination Act 1995 (DDA) does not refer explicitly to website accessibility, but makes it illegal to discriminate against people with disabilities. The DDA applies to anyone providing a service; public, private and voluntary sectors. The Code of Practice: Rights of Access - Goods, Facilities, Services and Premises document published by the government's Disability Rights Commission to accompany the Act does refer explicitly to websites as one of the "services to the public" which should be considered covered by the Act.

Website accessibility audits

A growing number of organizations, companies and consultants offer *website accessibility audits*. These audits, a type of system testing, identify accessibility problems that exist within a website, and provide advice and guidance on the steps that need to be taken to correct these problems.

A range of methods are used to audit websites for accessibility:

- Automated tools are available which can identify some of the problems that are present.
- Expert technical reviewers, knowledgeable in web design technologies and accessibility, can review a representative selection of pages and provide detailed feedback and advice based on their findings.
- User testing, usually overseen by technical experts, involves setting tasks for ordinary users to carry out on the website, and reviewing the problems these users encounter as they try to carry out the tasks.

Each of these methods has its strengths and weaknesses:

- Automated tools can process many pages in a relatively short length of time, but can only identify some of the accessibility problems that might be present in the website.
- Technical expert review will identify many of the problems that exist, but the process is time consuming, and many websites are too large to make it possible for a person to review every page.
- User testing combines elements of usability and accessibility testing, and is valuable for identifying problems that might otherwise be overlooked, but needs

to be used knowledgeably to avoid the risk of basing design decisions on one user's preferences.

Ideally, a combination of methods should be used to assess the accessibility of a website.

Accessible Web applications and WAI-ARIA

For a Web page to be accessible all important semantics about the page's functionality must be available so that assistive technology can understand and process the content and adapt it for the user. However as content becomes more and more complex, the standard HTML tags and attributes become inadequate in providing semantic reliably. Modern Web applications often apply scripts to elements to control their functionality and to enable them to act as a control or other dynamic component. These custom components or widgets do not provide a way to convey semantic information to the user agent. WAI-ARIA (Accessible Rich Internet Applications) is a specification published by the World Wide Web Consortium that specifies how to increase the accessibility of dynamic content and user interface components developed with Ajax, HTML, JavaScript and related technologies. ARIA enables accessibility by enabling the author to provide all the semantics to fully describe its supported behaviour. It also allows each element can expose its current states and properties and its relationships between other elements. Accessibility problems with the focus and tab index are also corrected.

Internationalization

The W3C Internationalization Activity assures that web technology will work in all languages, scripts, and cultures. Beginning in 2004 or 2005, Unicode gained ground and eventually in December 2007 surpassed both ASCII and Western European as the Web's most frequently used character encoding. Originally RFC 3986 allowed resources to be identified by URI in a subset of US-ASCII. RFC 3987 allows more characters—any character in the Universal Character Set—and now a resource can be identified by IRI in any language.

Statistics

According to a 2001 study, there were massively more than 550 billion documents on the Web, mostly in the invisible Web, or deep Web. A 2002 survey of 2,024 million Web pages determined that by far the most Web content was in English: 56.4%; next were pages in German (7.7%), French (5.6%), and Japanese (4.9%). A more recent study, which used Web searches in 75 different languages to sample the Web, determined that there were over 11.5 billion Web pages in the publicly indexable Web as of the end of January 2005. As of March 2009, the indexable web contains at least 25.21 billion pages. On July 25, 2008, Google software engineers Jesse Alpert and Nissan Hajaj announced that Google Search had discovered one trillion unique URLs. As of May 2009, over 109.5 million websites operated. Of these 74% were commercial or other sites operating in the `.com` generic top-level domain.

Speed issues

Frustration over congestion issues in the Internet infrastructure and the high latency that results in slow browsing has led to a pejorative name for the World Wide Web: the *World Wide Wait*. Speeding up the Internet is an ongoing discussion over the use of peering and QoS technologies. Other solutions to reduce the congestion can be found at W3C. Standard guidelines for ideal Web response times are:

- 0.1 second (one tenth of a second). Ideal response time. The user doesn't sense any interruption.
- 1 second. Highest acceptable response time. Download times above 1 second interrupt the user experience.
- 10 seconds. Unacceptable response time. The user experience is interrupted and the user is likely to leave the site or system.

Caching

If a user revisits a Web page after only a short interval, the page data may not need to be re-obtained from the source Web server. Almost all web browsers cache recently obtained data, usually on the local hard drive. HTTP requests sent by a browser will usually only ask for data that has changed since the last download. If the locally cached data are still current, it will be reused. Caching helps reduce the amount of Web traffic on the Internet. The decision about expiration is made independently for each downloaded file, whether image, stylesheet, JavaScript, HTML, or whatever other content the site may provide. Thus even on sites with highly dynamic content, many of the basic resources only need to be refreshed occasionally. Web site designers find it worthwhile to collate resources such as CSS data and JavaScript into a few site-wide files so that they can be cached efficiently. This helps reduce page download times and lowers demands on the Web server.

There are other components of the Internet that can cache Web content. Corporate and academic firewalls often cache Web resources requested by one user for the benefit of all. Some search engines also store cached content from websites. Apart from the facilities built into Web servers that can determine when files have been updated and so need to be re-sent, designers of dynamically generated Web pages can control the HTTP headers sent back to requesting users, so that transient or sensitive pages are not cached. Internet banking and news sites frequently use this facility. Data requested with an HTTP 'GET' is likely to be cached if other conditions are met; data obtained in response to a 'POST' is assumed to depend on the data that was POSTed and so is not cached.

Chapter 4

Internet Protocol Suite

The **Internet Protocol Suite** is the set of communications protocols used for the Internet and other similar networks. It is commonly also known as **TCP/IP** named from two of the most important protocols in it: the Transmission Control Protocol (TCP) and the Internet Protocol (IP), which were the first two networking protocols defined in this standard. Modern IP networking represents a synthesis of several developments that began to evolve in the 1960s and 1970s, namely the Internet and local area networks, which emerged during the 1980s, together with the advent of the World Wide Web in the early 1990s.

The Internet Protocol Suite consists of four abstraction layers. From the lowest to the highest layer, these are the Link Layer, the Internet Layer, the Transport Layer, and the Application Layer. The layers define the operational scope or reach of the protocols in each layer, reflected loosely in the layer names. Each layer has functionality that solves a set of problems relevant in its scope.

The Link Layer contains communication technologies for the local network the host is connected to directly, the link. It provides the basic connectivity functions interacting with the networking hardware of the computer and the associated management of interface-to-interface messaging. The Internet Layer provides communication methods between multiple links of a computer and facilitates the interconnection of networks. As such, this layer establishes the Internet. It contains primarily the Internet Protocol, which defines the fundamental addressing namespaces, Internet Protocol Version 4 (IPv4) and Internet Protocol Version 6 (IPv6) used to identify and locate hosts on the network. Direct host-to-host communication tasks are handled in the Transport Layer, which provides a general framework to transmit data between hosts using protocols like the Transmission Control Protocol and the User Datagram Protocol (UDP). Finally, the highest-level Application Layer contains all protocols that are defined each specifically for the functioning of the vast array of data communications services. This layer handles application-based interaction on a process-to-process level between communicating Internet hosts.

History

The Internet Protocol Suite resulted from research and development conducted by the Defense Advanced Research Projects Agency (DARPA) in the early 1970s. After initiating the pioneering ARPANET in 1969, DARPA started work on a number of other data transmission technologies. In 1972, Robert E. Kahn joined the DARPA Information Processing Technology Office, where he worked on both satellite packet networks and ground-based radio packet networks, and recognized the value of being able to communicate across both. In the spring of 1973, Vinton Cerf, the developer of the existing ARPANET Network Control Program (NCP) protocol, joined Kahn to work on open-architecture interconnection models with the goal of designing the next protocol generation for the ARPANET.

By the summer of 1973, Kahn and Cerf had worked out a fundamental reformulation, where the differences between network protocols were hidden by using a common internetwork protocol, and, instead of the network being responsible for reliability, as in the ARPANET, the hosts became responsible. Cerf credits Hubert Zimmerman and Louis Pouzin, designer of the CYCLADES network, with important influences on this design.

The design of the network included the recognition that it should provide only the functions of efficiently transmitting and routing traffic between end nodes and that all other intelligence should be located at the edge of the network, in the end nodes. Using a simple design, it became possible to connect almost any network to the ARPANET, irrespective of their local characteristics, thereby solving Kahn's initial problem. One popular expression is that TCP/IP, the eventual product of Cerf and Kahn's work, will run over "*two tin cans and a string.*"

A computer, called a router, is provided with an interface to each network. It forwards packets back and forth between them. Originally a router was called *gateway*, but the term was changed to avoid confusion with other types of gateways.

The idea was worked out in more detailed form by Cerf's networking research group at Stanford in the 1973–74 period, resulting in the first TCP specification. The early networking work at Xerox PARC, which produced the PARC Universal Packet protocol suite, much of which existed around the same period of time, was also a significant technical influence.

DARPA then contracted with BBN Technologies, Stanford University, and the University College London to develop operational versions of the protocol on different hardware platforms. Four versions were developed: TCP v1, TCP v2, a split into TCP v3 and IP v3 in the spring of 1978, and then stability with TCP/IP v4 — the standard protocol still in use on the Internet today.

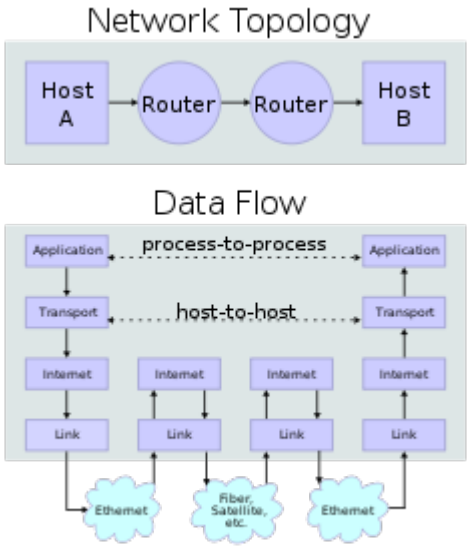
In 1975, a two-network TCP/IP communications test was performed between Stanford and University College London (UCL). In November, 1977, a three-network TCP/IP test was conducted between sites in the US, UK, and Norway. Several other TCP/IP

prototypes were developed at multiple research centres between 1978 and 1983. The migration of the ARPANET to TCP/IP was officially completed on flag day January 1, 1983, when the new protocols were permanently activated.

In March 1982, the US Department of Defense declared TCP/IP as the standard for all military computer networking. In 1985, the Internet Architecture Board held a three day workshop on TCP/IP for the computer industry, attended by 250 vendor representatives, promoting the protocol and leading to its increasing commercial use.

Layers in the Internet Protocol Suite

The concept of layers



Instantiations of the TCP/IP stack operating on two hosts each connected to its router on the Internet. Shown is the flow of user data through the layers used at each hop.

The Internet Protocol Suite uses encapsulation to provide abstraction of protocols and services. Encapsulation is usually aligned with the division of the protocol suite into layers of general functionality. In general, an application (the highest level of the model) uses a set of protocols to send its data down the layers, being further encapsulated at each level.

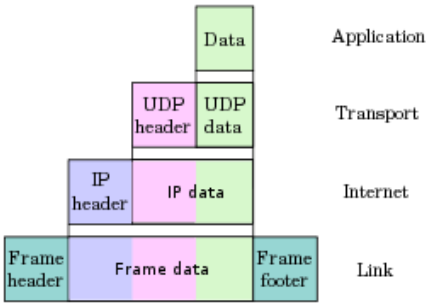
According to RFC 1122, the Internet Protocol Suite organizes the functional groups of protocols and methods into four layers, the Application Layer, the Transport Layer, the Internet Layer, and the Link Layer. This model was not intended to be a rigid reference model into which new protocols have to fit in order to be accepted as a standard.

The role of layering in TCP/IP may be illustrated by an example network scenario (right-hand diagram), in which two Internet host computers communicate across local network

boundaries constituted by their internetworking routers. The application on each host executes read and write operations as if the processes were directly connected to each other by some kind of data pipe, every other detail of the communication is hidden from each process. The underlying mechanisms that transmit data between the host computers are located in the lower protocol layers. The Transport Layer establishes host-to-host connectivity, meaning it handles the details of data transmission that are independent of the structure of user data and the logistics of exchanging information for any particular specific purpose. The layer simply establishes a basic data channel that an application uses in its task-specific data exchange. For this purpose the layer establishes the concept of the *port*, a numbered logical construct allocated specifically for each of the communication channels an application needs. For many types of services, these port numbers have been standardized so that client computers may address specific services of a server computer without the involvement of service announcements or directory services.

The Transport Layer operates on top of the Internet Layer. The Internet Layer is not only agnostic of application data structures as the Transport Layer, but it also does not distinguish between operation of the various Transport Layer protocols. It only provides an unreliable datagram transmission facility between hosts located on potentially different IP networks by forwarding the Transport Layer datagrams to an appropriate next-hop router for further relaying to its destination. With this functionality, the Internet Layer makes possible internetworking, the interworking of different IP networks, and it essentially establishes the Internet. The Internet Protocol is the principal component of the Internet Layer, and it defines two addressing systems to identify network hosts computers, and to locate them on the network. The original address system of the ARPANET and its successor, the Internet, is Internet Protocol Version 4 (IPv4). It uses a 32-bit IP address and is therefore capable of identifying approximately four billion hosts. This limitation was eliminated by the standardization of Internet Protocol Version 6 (IPv6) in 1998, and beginning production implementations in approximately 2006.

The lowest layer in the Internet Protocol Suite is the Link Layer. It comprises the tasks of specific networking requirements on the local link, the network segment that a hosts network interface is connected to. This involves interacting with the hardware-specific functions of network interfaces and specific transmission technologies.



Successive encapsulation of application data descending through the protocol stack before transmission on the local network link.

As the user data, first manipulated and structured in the Application Layer, is passed through the descending layers of the protocol stack each layer adds encapsulation information as illustrated in the diagram (right). A receiving host reverses the encapsulation at each layer by extracting the higher level data and passing it up the stack to the receiving process.

Layer names and number of layers in the literature

The following table shows the layer names and the number of layers of networking models presented in RFCs and textbooks in widespread use in today's university computer networking courses.

RFC 1122	Tanenbaum	Cisco Academy	Kurose Forouzan	Comer Kozierok	Stallings	Arpanet Reference Model 1982 (RFC 871)
Four layers	*Four layers*	*Four layers*	*Five layers*	*Four+one layers*	*Five layers*	*Three layers*
"Internet model"	"TCP/IP reference model"	"Internet model"	"Five-layer Internet model" or "TCP/IP protocol suite"	"TCP/IP 5-layer reference model"	"TCP/IP model"	"Arpanet reference model"
Application	Application	Application	Application	Application	Application	Application/Process
Transport	Transport	Transport	Transport	Transport	Host-to-host or transport	Host-to-host
Internet	Internet	Internetwork	Network	Internet	Internet	
Link	Host-to-network	Network interface	Data link	Data link (Network interface)	Network access	Network interface
			Physical	(Hardware)	Physical	

These textbooks are secondary sources that may contravene the intent of RFC 1122 and other IETF primary sources.

Different authors have interpreted the RFCs differently regarding the question whether the Link Layer (and the TCP/IP model) covers Physical Layer issues, or if a hardware layer is assumed below the Link Layer. Some authors have tried to use other names for the Link Layer, such as *network interface layer*, in view to avoid confusion with the Data Link Layer of the seven layer OSI model. Others have attempted to map the Internet Protocol model onto the OSI Model. The mapping often results in a model with five layers where the Link Layer is split into a Data Link Layer on top of a Physical Layer. In literature with a bottom-up approach to Internet communication, in which hardware issues are emphasized, those are often discussed in terms of Physical Layer and Data Link Layer.

The Internet Layer is usually directly mapped into the OSI Model's Network Layer, a more general concept of network functionality. The Transport Layer of the TCP/IP model, sometimes also described as the host-to-host layer, is mapped to OSI Layer 4 (Transport Layer), sometimes also including aspects of OSI Layer 5 (Session Layer) functionality. OSI's Application Layer, Presentation Layer, and the remaining functionality of the Session Layer are collapsed into TCP/IP's Application Layer. The argument is that these OSI layers do usually not exist as separate processes and protocols in Internet applications.

However, the Internet protocol stack has never been altered by the Internet Engineering Task Force from the four layers defined in RFC 1122. The IETF makes no effort to follow the OSI model although RFCs sometimes refer to it. The IETF has repeatedly stated that Internet protocol and architecture development is not intended to be OSI-compliant.

RFC 3439, addressing Internet architecture, contains a section entitled: "Layering Considered Harmful".

Implementations

Most computer operating systems in use today, including all consumer-targeted systems, include a TCP/IP implementation.

Minimally acceptable implementation includes implementation for (from most essential to the less essential) IP, ARP, ICMP, UDP, TCP and sometimes IGMP. It is in principle possible to support only one of transport protocols (i.e. simple UDP), but it is rarely done, as it limits usage of the whole implementation. IPv6, beyond own version of ARP (NBP), and ICMP (ICMPv6), and IGMP (IGMPv6) have some additional required functionalities, and often is accompanied with integrated IPSec security layer. Other protocols could be easily added later (often they can be implemented entirely in the userspace), for example DNS for resolving domain names to IP addresses or DHCP client for automatic configuration of network interfaces.

Most of the IP implementations are accessible to the programmers using socket abstraction (usable also with other protocols) and proper API for most of the operations. This interface is known as BSD sockets and was used initially in C.

Unique implementations include Lightweight TCP/IP, an open source stack designed for embedded systems and KA9Q NOS, a stack and associated protocols for amateur packet radio systems and personal computers connected via serial lines.

Chapter 5

Introduction to Cyberspace

Cyberspace is the electronic medium of computer networks, in which online communication takes place. It is the domain of electromagnetics readily identified with the interconnected information technology required to achieve the wide range of system capabilities associated with the transport of communication and control products and services. Current technology integrates a number of capabilities (sensors, signals, connections, transmissions, processors, and controllers) sufficient to generate a virtual interactive experience accessible regardless of a geographic location.

Cyberspace is the dynamic realization of electromagnetic energy through the application of communication and control technology. In pragmatic terms, operations within this global domain allow an interdependent network of information technology infrastructures (ITI), telecommunications networks, computer processing systems, integrated sensors, system control networks, embedded processors and controllers common to global control and communications across the electro-magnetic environment. As a social experience, individuals can interact, exchange ideas, share information, provide social support, conduct business, direct actions, create artistic media, play games, engage in political discussion, and so on. The term is rooted in the science of cybernetics from the Greek κυβερνήτης (kybernētēs, steersman, governor, pilot, or rudder) and Norbert Wiener's pioneering work in electronic communication and control science, a forerunner to current information theory and computer science.

The term "cyberspace" was first used by the cyberpunk science fiction author William Gibson. which he would later describe as an "evocative and essentially meaningless" buzzword that could serve as a cipher for all of his cybernetic musings. Now ubiquitous, the term has become a conventional means to describe anything associated with computers, information technology, the internet and the diverse internet culture. The United States government recognizes the interconnected information technology and the interdependent network of information technology infrastructures operating across this medium as part of the US National Critical Infrastructure.

According to Chip Morningstar and F. Randall Farmer, cyberspace is defined more by the social interactions involved rather than its technical implementation. The core characteristic is that must be an environment which consists of many participants with the ability to affect and influence each other. They derive this concept from the observation that people seek richness, complexity, and depth within a virtual world. Hence in

cyberspace, the computational medium is an augmentation of the communication channel between real people.

Origins of the term

The word "cyberspace" (from *cybernetics* and *space*) was coined by science fiction novelist and seminal cyberpunk author William Gibson in his 1982 story "Burning Chrome" and popularized by his 1984 novel *Neuromancer.* The portion of *Neuromancer* cited in this respect is usually the following:

Cyberspace. A consensual hallucination experienced daily by billions of legitimate operators, in every nation, by children being taught mathematical concepts... A graphic representation of data abstracted from the banks of every computer in the human system. Unthinkable complexity. Lines of light ranged in the nonspace of the mind, clusters and constellations of data. Like city lights, receding.

Despite its originally negative overtone, the term no longer implies a negative connotation.

Gibson later commented on the origin of the term in the 2000 documentary *No Maps for These Territories*:

All I knew about the word "cyberspace" when I coined it, was that it seemed like an effective buzzword. It seemed evocative and essentially meaningless. It was suggestive of something, but had no real semantic meaning, even for me, as I saw it emerge on the page.

Metaphorical

The metaphor used to describe the "sense of a social setting that exists purely within a space of representation and communication . . . it exists entirely within a computer space, distributed across increasingly complex and fluid networks." (Slater 2002, 355) The term "Cyberspace" started to become a de facto synonym for the internet, and later the World Wide Web, during the 1990s, especially in academic circles and activist communities. Author Bruce Sterling, who popularized this meaning, credits John Perry Barlow as the first to use it to refer to "the present-day nexus of computer and telecommunications networks." Barlow describes it thus in his essay to announce the formation of the Electronic Frontier Foundation (note the spatial metaphor) in June, 1990:

In this silent world, all conversation is typed. To enter it, one forsakes both body and place and becomes a thing of words alone. You can see what your neighbors are saying (or recently said), but not what either they or their physical surroundings look like. Town meetings are continuous and discussions rage on everything from sexual kinks to depreciation schedules.

Whether by one telephonic tendril or millions, they are all connected to one another. Collectively, they form what their inhabitants call the Net. It extends across that immense region of electron states, microwaves, magnetic fields, light pulses and thought which sci-fi writer William Gibson named Cyberspace.

—John Perry Barlow, *"Crime and Puzzlement," 1990-06-08*

As Barlow, and the EFF, continued public education efforts to promote the idea of "digital rights", the term was increasingly used during the internet boom of the late 1990s.

Virtual environments

In 1989, Autodesk, an American multinational corporation that focuses on 2D and 3D design software, developed a virtual design system called Cyberspace.

Cyberspace as an internet metaphor

While cyberspace should not be confused with the internet, the term is often used to refer to objects and identities that exist largely within the communication network itself, so that a website, for example, might be metaphorically said to "exist in cyberspace." According to this interpretation, events taking place on the internet are not happening in the locations where participants or servers are physically located, but "in cyberspace".

Firstly, cyberspace describes the flow of digital data through the network of interconnected computers: it is at once not "real", since one could not spatially locate it as a tangible object, and clearly "real" in its effects. Secondly, cyberspace is the site of computer-mediated communication (CMC), in which online relationships and alternative forms of online identity were enacted, raising important questions about the social psychology of internet use, the relationship between "online" and "offline" forms of life and interaction, and the relationship between the "real" and the virtual. Cyberspace draws attention to remediation of culture through new media technologies: it is not just a communication tool but a social destination, and is culturally significant in its own right. Finally, cyberspace can be seen as providing new opportunities to reshape society and culture through "hidden" identities, or it can be seen as borderless communication and culture.

Cyberspace is the "place" where a telephone conversation appears to occur. Not inside your actual phone, the plastic device on your desk. Not inside the other person's phone, in some other city. **The place between** the phones. [...] in the past twenty years, this electrical "space," which was once thin and dark and one-dimensional—little more than a narrow speaking-tube, stretching from phone to phone—has flung itself open like a gigantic jack-in-the-box. Light has flooded upon it, the eerie light of the glowing computer screen. This dark electric netherworld has become a vast flowering electronic landscape. Since the 1960s, the world of the telephone has cross-bred itself with computers and television, and though there is still no substance to cyberspace, nothing

you can handle, it has a strange kind of physicality now. It makes good sense today to talk of cyberspace as a place all its own.

– Bruce Sterling, Introduction to ***The Hacker Crackdown***

The "space" in cyberspace has more in common with the abstract, mathematical meanings of the term than physical space. It does not have the duality of positive and negative volume (while in physical space for example a room has the negative volume of usable space delineated by positive volume of walls, internet users cannot enter the screen and explore the unknown part of the internet as an extension of the space they are in), but spatial meaning can be attributed to the relationship between different pages (of books as well as webservers), considering the unturned pages to be somewhere "out there." The concept of cyberspace therefore refers not to the content being presented to the surfer, but rather to the possibility of surfing among different sites, with feedback loops between the user and the rest of the system creating the potential to always encounter something unknown or unexpected.

Videogames differ from text-based communication in that on-screen images are meant to be figures that actually occupy a space and the animation shows the movement of those figures. Images are supposed to form the positive volume that delineates the empty space. A game adopts the cyberspace metaphor by engaging more players in the game, and then figuratively representing them on the screen as avatars. Games do not have to stop at the avatar-player level, but current implementations aiming for more immersive playing space (i.e. Laser tag) take the form of augmented reality rather than cyberspace, fully immersive virtual realities remaining impractical.

Although the more radical consequences of the global communication network predicted by some cyberspace proponents (i.e. the diminishing of state influence envisioned by John Perry Barlow) failed to materialize and the word lost some of its novelty appeal, it remains current as of 2006.

Some virtual communities explicitly refer to the concept of cyberspace, for example Linden Lab calling their customers "Residents" of Second Life, while all such communities can be positioned "in cyberspace" for explanatory and comparative purposes (as did Sterling in *The Hacker Crackdown*, followed by many journalists), integrating the metaphor into a wider cyber-culture.

The metaphor has been useful in helping a new generation of thought leaders to reason through new military strategies around the world, led largely by the US Department of Defense (DoD). The use of cyberspace as a metaphor has had its limits, however, especially in areas where the metaphor becomes confused with physical infrastructure.

Alternate realities in philosophy and art

Predating computers

A forerunner of the modern ideas of cyberspace is the Cartesian notion that people might be deceived by an evil demon that feeds them a false reality. This argument is the direct predecessor of modern ideas of a brain in a vat and many popular conceptions of cyberspace take Descartes's ideas as their starting point.

Visual arts have a tradition, stretching back to antiquity, of artifacts meant to fool the eye and be mistaken for reality. This questioning of reality occasionally led some philosophers and especially theologians to distrust art as deceiving people into entering a world which was not real. The artistic challenge was resurrected with increasing ambition as art became more and more realistic with the invention of photography, film and immersive computer simulations.

Influenced by computers

Philosophy

American counterculture exponents like William S. Burroughs (whose literary influence on Gibson and cyberpunk in general is widely acknowledged) and Timothy Leary were among the first to extoll the potential of computers and computer networks for individual empowerment.

Some contemporary philosophers and scientists (i.e. David Deutsch in *The Fabric of Reality*) employ virtual reality in various thought experiments. For example Philip Zhai in *Get Real: A Philosophical Adventure in Virtual Reality* connects cyberspace to the platonic tradition:

Let us imagine a nation in which everyone is hooked up to a network of VR infrastructure. They have been so hooked up since they left their mother's wombs. Immersed in cyberspace and maintaining their life by teleoperation, they have never imagined that life could be any different from that. The first person that thinks of the possibility of an alternative world like ours would be ridiculed by the majority of these citizens, just like the few enlightened ones in Plato's allegory of the cave.

Note that this brain-in-a-vat argument conflates cyberspace with reality, while the more common descriptions of cyberspace contrast it with the "real world".

Art

Having originated among writers, the concept of cyberspace remains most popular in literature and film. Although artists working with other media have expressed interest in the concept, such as Roy Ascott, "cyberspace" in digital art is mostly used as a synonym for immersive virtual reality and remains more discussed than enacted. Indian epic

Mahabaratha written by sage Vyasar talks about concepts what is called today Virtual reality, Transportation in to matrix and web conferencing.

Computer crime

Cyberspace also brings together every service and facility imaginable to expedite money laundering. One can purchase anonymous credit cards, bank accounts, encrypted global mobile telephones, and false passports. From there one can pay professional advisors to set up IBCs (International Business Corporations, or corporations with anonymous ownership) or similar structures in OFCs (Offshore Financial Centers). Such advisors are loath to ask any penetrating questions about the wealth and activities of their clients, since the average fees criminals pay them to launder their money can be as much as 20 percent.

5-level model

A 5-level model has been designed recently in France. According to it, the cyberspace is composed of 5 layers based on information discoveries: language, writing, printing, Internet, etc. This original model links the world of information to telecommunication technologies.

Chapter 6

Cyberethics

Cyberethics is distinct from cyberlaw. Laws are formal written directives that apply to everyone, interpreted by the judicial system, and enforced by the police. Ethics is a broad philosophical concept that goes beyond simple right and wrong, and looks towards "the good life".

Privacy

In the late 18th century, the invention of cameras spurred similar ethical debates as the internet does today. During a Harvard Law Review seminal in 1890, Warren and Brandeis defined privacy from an ethical and moral point of view to be "central to dignity and individuality and personhood. Privacy is also indispensable to a sense of autonomy - to 'a feeling that there is an area of an individual's life that is totally under his or her control, an area that is free from outside intrusion.' The deprivation of privacy can even endanger a person's health." (Warren & Brandeis, 1890). Over 100 years later, the internet and proliferation of private data through governments and ecommerce is a phenomenon which requires a new round of ethical debate involving a person's privacy.

Privacy can be decomposed to the limitation of others' access to an individual with "three elements of secrecy, anonymity, and solitude" (Gavison, 1984). Anonymity refers to the individual's right to protection from undesired attention. Solitude refers to the lack of physical proximity of an individual to others. Secrecy refers to the protection of personalized information from being freely distributed.

Individuals surrender private information when conducting transactions and registering for services. Ethical business practice protects the privacy of their customers by securing information which may contribute to the loss of secrecy, anonymity, and solitude. Credit card information, social security numbers, phone numbers, mothers' maiden names, addresses and phone numbers freely collected and shared over the internet may lead to a loss of Privacy.

Fraud and impersonation are some of the malicious activities that occur due to the direct or indirect abuse of private information. Identity theft is rising rapidly due to the availability of private information in the internet. For instance, seven million Americans have fallen victim to identity theft in 2002, making it the fastest growing crime in the

United States (Latak, 2005). Public records search engines and databases are the main culprits contributing to the rise of cybercrime. Listed below are a few recommendations to restrict online databases from proliferating sensitive personnel information.

1. Exclude sensitive unique identifiers from database records such as social security numbers, birth dates, hometown and mothers' maiden names.
2. Exclude phone numbers that are normally unlisted.
3. Clear provision of a method which allows people to have their names removed from a database.
4. Banning the reverse social security number lookup services (Spinello, 2006).

Private Data Collection

Data warehouses are used today to collect and store huge amounts of personal data and consumer transactions. These facilities can preserve large volumes of consumer information for an indefinite amount of time. Some of the key architectures contributing to the erosion of privacy include databases, cookies and spyware.

Some may argue that data warehouses are supposed to stand alone and be protected. However, the fact is enough personal information can be gathered from corporate websites and social networking sites to initiate a reverse lookup. Therefore, is it not important to address some of the ethical issues regarding how protected data ends up in the public domain?

As a result, identity theft protection businesses are on the rise. Companies such as LifeLock and JPMorgan Chase have begun to capitalize on selling identity theft protection insurance.

Property

Ethical debate has long included the concept of property. This concept has created many clashes in the world of cyberethics. One philosophy of the internet is centered around the freedom of information. The controversy over ownership occurs when the property of information is infringed upon or uncertain.

Intellectual Property Rights

The ever-increasing speed of the internet and the emergence of compression technology, such as mp3 opened the doors to Peer-to-peer file sharing, a technology that allowed users to anonymously transfer files to each other, previously seen on programs such as Napster or now seen through communications protocol such as BitTorrent. Much of this, however, was copyrighted music and illegal to transfer to other users. Whether it is ethical to transfer copyrighted media is another question.

Proponents of unrestricted file sharing point out how file sharing has given people broader and faster access to media, has increased exposure to new artists, and has reduced

the costs of transferring media (including less environmental damage). Supporters of restrictions on file sharing argue that we must protect the income of our artists and other people who work to create our media. This argument is partially answered by pointing to the small proportion of money artists receive from the legitimate sale of media.

We also see a similar debate over intellectual property rights in respect to software ownership. The two opposing views are for closed source software distributed under restrictive licenses or for free and open source software (Freeman & Peace, 2004). The argument can be made that restrictions are required because companies would not invest weeks and months in development if there is no incentive for revenue generated from sales and licensing fees. Proponents for open source believe that all programs should be available to anyone who wants to study them.

Digital Rights Management (DRM)

With the introduction of Digital Rights Management software, new issues are raised over whether the subverting of DRM is ethical. Some champion the hackers of DRM as defenders of users' rights, allowing the blind to make audio books of PDFs they receive, allowing people to burn music they have legitimately bought to CD or to transfer it to a new computer. Others see this as nothing but simply a violation of the rights of the intellectual property holders, opening the door to uncompensated use of copyrighted media.

Security

Security has long been a topic of ethical debate. Is it better to protect the common good of the community or rather should we safeguard the rights of the individual? There is a continual dispute over the boundaries between the two and which compromises are right to make. As an ever increasing amount of people connect to the internet and more and more personal data is available online there is susceptibility to identity theft, cyber crimes and computer hacking. This also leads to the question of who has the right to regulate the internet in the interest of security?

Accuracy

Due to the ease of accessibility and sometimes collective nature of the internet we often come across issues of accuracy e.g. who is responsible for the authenticity and fidelity of the information available online? Ethically this includes debate over who should be allowed to contribute content and who should be held accountable if there are errors in the content or if it is false. This also brings up the question of how is the injured party, if any, to be made whole and under which jurisdiction does the offense lay?

Accessibility, Censorship and Filtering

Accessibility, censorship and filtering bring up many ethical issues that have several branches in cyberethics. Many questions have arisen which continue to challenge our understanding of privacy, security and our participation in society. Throughout the centuries mechanisms have been constructed in the name of protection and security. Today the applications are in the form of software that filters domains and content so that they may not be easily accessed or obtained without elaborate circumvention or on a personal and business level through free or content-control software. Internet censorship and filtering are used to control or suppress the publishing or accessing of information. The legal issues are similar to offline censorship and filtering. The same arguments that apply to offline censorship and filtering apply to online censorship and filtering; whether people are better off with free access to information or should be protected from what is considered by a governing body as harmful, indecent or illicit. The fear of access by minors drives much of the concern and many online advocate groups have sprung up to raise awareness and of controlling the accessibility of minors to the internet.

Censorship and filtering occurs on small to large scales, whether it be a company restricting their employees' access to cyberspace by blocking certain websites which are deemed as relevant only to personal usage and therefore damaging to productivity or on a larger scale where a government creates large firewalls which censor and filter access to certain information available online frequently from outside their country to their citizens and anyone within their borders. One of the most famous examples of a country controlling access is the Golden Shield Project, also referred to as the Great Firewall of China, a censorship and surveillance project set up and operated by the People's Republic of China. Another instance is the 2000 case of the League Against Racism and Antisemitism (LICRA), French Union of Jewish Students, vs. Yahoo! Inc (USA) and Yahoo! France, where the French Court declared that "access by French Internet users to the auction website containing Nazi objects constituted a contravention of French law and an offence to the 'collective memory' of the country and that the simple act of displaying such objects (e.g. exhibition of uniforms, insignia or emblems resembling those worn or displayed by the Nazis) in France constitutes a violation of the Article R645-1 of the Penal Code and is therefore considered as a threat to internal public order."(Akdeniz, 2001). Since the French judicial ruling many websites must abide by the rules of the countries in which they are accessible.

Freedom of Information

Freedom of information, that is the freedom of speech as well as the freedom to seek, obtain and impart information brings up the question of who or what, has the jurisdiction in cyberspace. The right of freedom of information is commonly subject to limitations dependant upon the country, society and culture concerned.

Generally there are three standpoints on the issue as it relates to the internet. First is the argument that the internet is a form of media, put out and accessed by citizens of governments and therefore should be regulated by each individual government within the

borders of their respective jurisdictions. Second, is that, "Governments of the Industrial World… have no sovereignty [over the internet] … We have no elected government, nor are we likely to have one, … You have no moral right to rule us nor do you possess any methods of enforcement we have true reason to fear." (Barlow, 1996). A third party believes that the internet supersedes all tangible borders such as the borders of countries, authority should be given to an international body since what is legal in one country may be against the law in the other.

Digital Divide

An issue specific to the ethical issues of the Freedom of Information is what is known as the digital divide. This refers to the unequal socio-economic divide between those who have access to digital and information technology such as cyberspace and those who have limited or no access at all. This gap of access between countries or regions of the world is called the global digital divide.

Sexuality and Pornography

Sexuality in terms of sexual orientation, infidelity, sex with or between minors, public display and pornography have always stirred ethical controversy. These issues are reflected online to varying degrees. One of the largest cyberethical debates is over the regulation, distribution and accessibility of pornography online. Hardcore pornographic material is generally controlled by governments with laws regarding how old one has to be to obtain it and what forms are acceptable or not. The availability of pornography online calls into question jurisdiction as well as brings up the problem of regulation in particular over child pornography, which is illegal in most countries, as well as pornography involving violence or animals, which is restricted within most countries.

Gambling

Gambling is often a topic in ethical debate as some view it as inherently wrong and support prohibition while others support no legal interference at all. "Between these extremes lies a multitude of opinions on what types of gambling the government should permit and where it should be allowed to take place. Discussion of gambling forces public policy makers to deal with issues as diverse as addiction, tribal rights, taxation, senior living, professional and college sports, organized crime, neurobiology, suicide, divorce, and religion." (McGowan, 2007). Due to its controversy gambling is either banned or heavily controlled on local or national levels. The accessibility of the internet and its ability to cross geographic-borders have led to illegal online gambling, often offshore operations. Over the years online gambling, both legal and illegal, has grown exponentially which has led to difficulties in regulation. This enormous growth has even called into question by some the ethical place of gambling online.

Organizations Related to Cyberethics

The following organizations are of notable interest in the cyberethics debate:

- • International Federation for Information Processing (IFIP) IFIP
- • Association for Computer Machinery, Special Interest Group: Computers and Society (SIGCAS)
- • Ethical and Professional Issues in Computing (EPIC) Electronic Privacy Information Center
- • Electronic Frontier Foundation (EFF) Electronic Frontier Foundation
- • International Center for Information Ethics (ICIE)
- • Directions and Implications in Advanced Computing (DIAC)
- • The Centre for Computing and Social Responsibility (CCSR)
- • Cyber-Rights and Cyber-liberties

Codes of Ethics in Computing

Information Technology managers are required to establish a set of ethical standards common to their organization. There are many examples of ethical code currently published that can be tailored to fit any organization. Code of ethics is an instrument that establishes a common ethical framework for a large group of people. Four well known examples of Code of Ethics for IT professionals are listed below:

RFC 1087

In January 1989, the Internet Architecture Board (IAB) in RFC 1087 defines an activity as unethical and unacceptable if it:

1. Seeks to gain unauthorized access to the resources of the Internet.
2. Disrupts the intended use of the Internet.
3. Wastes resources (people, capacity, computer) through such actions.
4. Destroys the integrity of computer-based information, or
5. Compromises the privacy of users (RFC 1087, 1989).

The Code of Fair Information Practices

The Code of Fair Information Practices is based on five principles outlining the requirements for records keeping systems. This requirement was implemented in 1973 by the U.S. Department of Health, Education and Welfare.

1. There must be no personal data record-keeping systems whose very existence is secret.
2. There must be a way for a person to find out what information about the person is in a record and how it is used.

3. There must be a way for a person to prevent information about the person that was obtained for one purpose from being used or made available for other purposes without the person's consent.
4. There must be a way for a person to correct or amend a record of identifiable information about the person.
5. Any organization creating, maintaining, using, or disseminating records of identifiable personal data must assure the reliability of the data for their intended use and must take precautions to prevent misuses of the data (Harris, 2003).

Ten Commandments of Computer Ethics

The ethical values as defined in 1992 by the Computer Ethics Institute; a nonprofit organization whose mission is to advance technology by ethical means, lists these rules as a guide to computer ethics:

1. Thou shalt not use a computer to harm other people.
2. Thou shalt not interfere with other people's computer work.
3. Thou shalt not snoop around in other people's computer files.
4. Thou shalt not use a computer to steal.
5. Thou shalt not use a computer to bear false witness.
6. Thou shalt not copy or use proprietary software for which you have not paid.
7. Thou shalt not use other people's computer resources without authorization or proper compensation.
8. Thou shalt not appropriate other people's intellectual output.
9. Thou shalt think about the social consequences of the program you are writing or the system you are designing.
10. Thou shalt always use a computer in ways that ensure consideration and respect for your fellow humans (Computer Ethics Institute, 1992).

$(ISC)^2$ Code of Ethics

$(ISC)^2$ an organization committed to certification of computer security professional has further defined its own Code of Ethics generally as:

1. Act honestly, justly, responsibly, and legally, and protecting the commonwealth.
2. Work diligently and provide competent services and advance the security profession.
3. Encourage the growth of research – teach, mentor, and value the certification.
4. Discourage unsafe practices, and preserve and strengthen the integrity of public infrastructures.
5. Observe and abide by all contracts, expressed or implied, and give prudent advice.
6. Avoid any conflict of interest, respect the trust that others put in you, and take on only those jobs you are qualified to perform.
7. Stay current on skills, and do not become involved with activities that could injure the reputation of other security professionals (Harris, 2003).

Chapter 7

Web Search Engine

A **web search engine** is designed to search for information on the World Wide Web and FTP servers. The search results are generally presented in a list of results and are often called *hits*. The information may consist of web pages, images, information and other types of files. Some search engines also mine data available in databases or open directories. Unlike Web directories, which are maintained by human editors, search engines operate algorithmically or are a mixture of algorithmic and human input.

History

Timeline (full list)

Year	Engine	Event
	W3Catalog	Launch
1993	Aliweb	Launch
	JumpStation	Launch
	WebCrawler	Launch
1994	Infoseek	Launch
	Lycos	Launch
	AltaVista	Launch
	Daum	Founded
	Open Text Web Index	Launch
1995	Magellan	Launch
	Excite	Launch
	SAPO	Launch
	Yahoo!	Launch
	Dogpile	Launch
1996	Inktomi	Founded
	HotBot	Founded
	Ask Jeeves	Founded
1997	Northern Light	Launch
	Yandex	Launch

1998	Google	Launch
	AlltheWeb	Launch
	GenieKnows	Founded
1999	Naver	Launch
	Teoma	Founded
	Vivisimo	Founded
2000	Baidu	Founded
	Exalead	Founded
2003	Info.com	Launch
	Yahoo! Search	Final launch
2004	A9.com	Launch
	Sogou	Launch
	MSN Search	Final launch
2005	Ask.com	Launch
	GoodSearch	Launch
	SearchMe	Founded
	Quaero	Founded
2006	Ask.com	Launch
	Live Search	Launch
	ChaCha	Beta Launch
	Guruji.com	Beta Launch
2007	Sproose	Launched
	Blackle.com	Launched
	Powerset	Launched
	Picollator	Launched
	Viewzi	Launched
	Blueready.com	Launched
2008	Cuil	Launched
	Boogami	Launched
	LeapFish	Beta Launch
	Forestle	Launched
	VADLO	Launched
	Duck Duck Go	Launched
2009	Bing	Launched
	Yebol	Beta Launch

	Mugurdy	Closed due to a lack of funding
	Goby	Launched
	Yandex global (English)	Launched
2010	Cuil	Closed
	Blekko	Beta Launch
Zib Zoom	Launch	
Viewzi	Closed due to a lack of funding	

During the early development of the web, there was a list of webservers edited by Tim Berners-Lee and hosted on the CERN webserver. One historical snapshot from 1992 remains. As more webservers went online the central list could not keep up. On the NCSA site new servers were announced under the title "What's New!"

The very first tool used for searching on the Internet was Archie. The name stands for "archive" without the "v". It was created in 1990 by Alan Emtage, Bill Heelan and J. Peter Deutsch, computer science students at McGill University in Montreal. The program downloaded the directory listings of all the files located on public anonymous FTP (File Transfer Protocol) sites, creating a searchable database of file names; however, Archie did not index the contents of these sites since the amount of data was so limited it could be readily searched manually.

The rise of Gopher (created in 1991 by Mark McCahill at the University of Minnesota) led to two new search programs, Veronica and Jughead. Like Archie, they searched the file names and titles stored in Gopher index systems. Veronica (*V*ery *E*asy *R*odent-*O*riented *N*et-wide *I*ndex to *C*omputerized *A*rchives) provided a keyword search of most Gopher menu titles in the entire Gopher listings. Jughead (*J*onzy's *U*niversal *G*opher *H*ierarchy *E*xcavation *A*nd *D*isplay) was a tool for obtaining menu information from specific Gopher servers. While the name of the search engine "Archie" was not a reference to the Archie comic book series, "Veronica" and "Jughead" are characters in the series, thus referencing their predecessor.

In the summer of 1993, no search engine existed yet for the web, though numerous specialized catalogues were maintained by hand. Oscar Nierstrasz at the University of Geneva wrote a series of Perl scripts that would periodically mirror these pages and rewrite them into a standard format which formed the basis for W3Catalog, the web's first primitive search engine, released on September 2, 1993.

In June 1993, Matthew Gray, then at MIT, produced what was probably the first web robot, the Perl-based World Wide Web Wanderer, and used it to generate an index called 'Wandex'. The purpose of the Wanderer was to measure the size of the World Wide Web, which it did until late 1995. The web's second search engine Aliweb appeared in November 1993. Aliweb did not use a web robot, but instead depended on being notified by website administrators of the existence at each site of an index file in a particular format.

JumpStation (released in December 1993) used a web robot to find web pages and to build its index, and used a web form as the interface to its query program. It was thus the first WWW resource-discovery tool to combine the three essential features of a web search engine (crawling, indexing, and searching) as described below. Because of the limited resources available on the platform on which it ran, its indexing and hence searching were limited to the titles and headings found in the web pages the crawler encountered.

One of the first "full text" crawler-based search engines was WebCrawler, which came out in 1994. Unlike its predecessors, it let users search for any word in any webpage, which has become the standard for all major search engines since. It was also the first one to be widely known by the public. Also in 1994, Lycos (which started at Carnegie Mellon University) was launched and became a major commercial endeavor.

Soon after, many search engines appeared and vied for popularity. These included Magellan, Excite, Infoseek, Inktomi, Northern Light, and AltaVista. Yahoo! was among the most popular ways for people to find web pages of interest, but its search function operated on its web directory, rather than full-text copies of web pages. Information seekers could also browse the directory instead of doing a keyword-based search.

In 1996, Netscape was looking to give a single search engine an exclusive deal to be their featured search engine. There was so much interest that instead a deal was struck with Netscape by five of the major search engines, where for $5Million per year each search engine would be in a rotation on the Netscape search engine page. The five engines were Yahoo!, Magellan, Lycos, Infoseek, and Excite.

Search engines were also known as some of the brightest stars in the Internet investing frenzy that occurred in the late 1990s. Several companies entered the market spectacularly, receiving record gains during their initial public offerings. Some have taken down their public search engine, and are marketing enterprise-only editions, such as Northern Light. Many search engine companies were caught up in the dot-com bubble, a speculation-driven market boom that peaked in 1999 and ended in 2001.

Around 2000, the Google search engine rose to prominence. The company achieved better results for many searches with an innovation called PageRank. This iterative algorithm ranks web pages based on the number and PageRank of other web sites and pages that link there, on the premise that good or desirable pages are linked to more than others. Google also maintained a minimalist interface to its search engine. In contrast, many of its competitors embedded a search engine in a web portal.

By 2000, Yahoo was providing search services based on Inktomi's search engine. Yahoo! acquired Inktomi in 2002, and Overture (which owned AlltheWeb and AltaVista) in 2003. Yahoo! switched to Google's search engine until 2004, when it launched its own search engine based on the combined technologies of its acquisitions.

Microsoft first launched MSN Search in the fall of 1998 using search results from Inktomi. In early 1999 the site began to display listings from Looksmart blended with results from Inktomi except for a short time in 1999 when results from AltaVista were used instead. In 2004, Microsoft began a transition to its own search technology, powered by its own web crawler (called msnbot).

Microsoft's rebranded search engine, Bing, was launched on June 1, 2009. On July 29, 2009, Yahoo! and Microsoft finalized a deal in which Yahoo! Search would be powered by Microsoft Bing technology.

How web search engines work

1. Web crawler

A **Web crawler** is a computer program that browses the World Wide Web in a methodical, automated manner or in an orderly fashion. Other terms for Web crawlers are *ants*, *automatic indexers*, *bots*, or *Web spiders*, *Web robots*, or—especially in the FOAF community—*Web scutters*.

This process is called *Web crawling* or *spidering*. Many sites, in particular search engines, use spidering as a means of providing up-to-date data. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches. Crawlers can also be used for automating maintenance tasks on a Web site, such as checking links or validating HTML code. Also, crawlers can be used to gather specific types of information from Web pages, such as harvesting e-mail addresses (usually for spam).

A Web crawler is one type of bot, or software agent. In general, it starts with a list of URLs to visit, called the *seeds*. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the *crawl frontier*. URLs from the frontier are recursively visited according to a set of policies.

Crawling policies

There are important characteristics of the Web that make crawling very difficult:

- its large volume,
- its fast rate of change, and
- dynamic page generation.

The large volume implies that the crawler can only download a fraction of the Web pages within a given time, so it needs to prioritize its downloads. The high rate of change implies that by the time the crawler is downloading the last pages from a site, it is very likely that new pages have been added to the site, or that pages have already been updated or even deleted.

The number of possible crawlable URLs being generated by server-side software has also made it difficult for web crawlers to avoid retrieving duplicate content. Endless combinations of HTTP GET (URL-based) parameters exist, of which only a small selection will actually return unique content. For example, a simple online photo gallery may offer three options to users, as specified through HTTP GET parameters in the URL. If there exist four ways to sort images, three choices of thumbnail size, two file formats, and an option to disable user-provided content, then the same set of content can be accessed with 48 different URLs, all of which may be linked on the site. This mathematical combination creates a problem for crawlers, as they must sort through endless combinations of relatively minor scripted changes in order to retrieve unique content.

As Edwards *et al.* noted, "Given that the bandwidth for conducting crawls is neither infinite nor free, it is becoming essential to crawl the Web in not only a scalable, but efficient way, if some reasonable measure of quality or freshness is to be maintained." . A crawler must carefully choose at each step which pages to visit next.

The behavior of a Web crawler is the outcome of a combination of policies:

- a *selection policy* that states which pages to download,
- a *re-visit policy* that states when to check for changes to the pages,
- a *politeness policy* that states how to avoid overloading Web sites, and
- a *parallelization policy* that states how to coordinate distributed Web crawlers.

Selection policy

Given the current size of the Web, even large search engines cover only a portion of the publicly-available part. A 2005 study showed that large-scale search engines index no more than 40%-70% of the indexable Web; a previous study by Dr. Steve Lawrence and Lee Giles showed that no search engine indexed more than 16% of the Web in 1999. As a crawler always downloads just a fraction of the Web pages, it is highly desirable that the downloaded fraction contains the most relevant pages and not just a random sample of the Web.

This requires a metric of importance for prioritizing Web pages. The importance of a page is a function of its intrinsic quality, its popularity in terms of links or visits, and even of its URL (the latter is the case of vertical search engines restricted to a single top-level domain, or search engines restricted to a fixed Web site). Designing a good selection policy has an added difficulty: it must work with partial information, as the complete set of Web pages is not known during crawling.

Cho *et al.* made the first study on policies for crawling scheduling. Their data set was a 180,000-pages crawl from the `stanford.edu` domain, in which a crawling simulation was done with different strategies. The ordering metrics tested were breadth-first, backlink-count and partial Pagerank calculations. One of the conclusions was that if the crawler wants to download pages with high Pagerank early during the crawling process,

then the partial Pagerank strategy is the better, followed by breadth-first and backlink-count. However, these results are for just a single domain. Cho also wrote his Ph.D. dissertation at Stanford on web crawling.

Najork and Wiener performed an actual crawl on 328 million pages, using breadth-first ordering. They found that a breadth-first crawl captures pages with high Pagerank early in the crawl (but they did not compare this strategy against other strategies). The explanation given by the authors for this result is that "the most important pages have many links to them from numerous hosts, and those links will be found early, regardless of on which host or page the crawl originates".

Abiteboul designed a crawling strategy based on an algorithm called OPIC (On-line Page Importance Computation). In OPIC, each page is given an initial sum of "cash" that is distributed equally among the pages it points to. It is similar to a Pagerank computation, but it is faster and is only done in one step. An OPIC-driven crawler downloads first the pages in the crawling frontier with higher amounts of "cash". Experiments were carried in a 100,000-pages synthetic graph with a power-law distribution of in-links. However, there was no comparison with other strategies nor experiments in the real Web.

Boldi *et al.* used simulation on subsets of the Web of 40 million pages from the `.it` domain and 100 million pages from the WebBase crawl, testing breadth-first against depth-first, random ordering and an omniscient strategy. The comparison was based on how well PageRank computed on a partial crawl approximates the true PageRank value. Surprisingly, some visits that accumulate PageRank very quickly (most notably, breadth-first and the omniscent visit) provide very poor progressive approximations. Baeza-Yates *et al.* used simulation on two subsets of the Web of 3 million pages from the `.gr` and `.cl` domain, testing several crawling strategies. They showed that both the OPIC strategy and a strategy that uses the length of the per-site queues are better than breadth-first crawling, and that it is also very effective to use a previous crawl, when it is available, to guide the current one.

Daneshpajouh *et al.* designed a community based algorithm for discovering good seeds. Their method crawls web pages with high PageRank from different communities in less iteration in comparison with crawl starting from random seeds. One can extract good seed from a previously-crawled-Web graph using this new method. Using these seeds a new crawl can be very effective.

Focused crawling

The importance of a page for a crawler can also be expressed as a function of the similarity of a page to a given query. Web crawlers that attempt to download pages that are similar to each other are called **focused crawler** or **topical crawlers**. The concepts of topical and focused crawling were first introduced by Menczer and by Chakrabarti *et al.*

The main problem in focused crawling is that in the context of a Web crawler, we would like to be able to predict the similarity of the text of a given page to the query before

actually downloading the page. A possible predictor is the anchor text of links; this was the approach taken by Pinkerton in a crawler developed in the early days of the Web. Diligenti *et al.* propose to use the complete content of the pages already visited to infer the similarity between the driving query and the pages that have not been visited yet. The performance of a focused crawling depends mostly on the richness of links in the specific topic being searched, and a focused crawling usually relies on a general Web search engine for providing starting points.

Restricting followed links

A crawler may only want to seek out HTML pages and avoid all other MIME types. In order to request only HTML resources, a crawler may make an HTTP HEAD request to determine a Web resource's MIME type before requesting the entire resource with a GET request. To avoid making numerous HEAD requests, a crawler may examine the URL and only request a resource if the URL ends with certain characters such as .html, .htm, .asp, .aspx, .php, .jsp, .jspx or a slash. This strategy may cause numerous HTML Web resources to be unintentionally skipped.

Some crawlers may also avoid requesting any resources that have a "?" in them (are dynamically produced) in order to avoid spider traps that may cause the crawler to download an infinite number of URLs from a Web site. This strategy is unreliable if the site uses URL rewriting to simplify its URLs.

URL normalization

Crawlers usually perform some type of URL normalization in order to avoid crawling the same resource more than once. The term *URL normalization*, also called *URL canonicalization*, refers to the process of modifying and standardizing a URL in a consistent manner. There are several types of normalization that may be performed including conversion of URLs to lowercase, removal of "." and ".." segments, and adding trailing slashes to the non-empty path component.

Path-ascending crawling

Some crawlers intend to download as many resources as possible from a particular web site. So *path-ascending crawler* was introduced that would ascend to every path in each URL that it intends to crawl.

Many path-ascending crawlers are also known as Web harvesting software, because they're used to "harvest" or collect all the content — perhaps the collection of photos in a gallery — from a specific page or host.

Re-visit policy

The Web has a very dynamic nature, and crawling a fraction of the Web can take weeks or months. By the time a Web crawler has finished its crawl, many events could have happened, including creations, updates and deletions.

From the search engine's point of view, there is a cost associated with not detecting an event, and thus having an outdated copy of a resource. The most-used cost functions are freshness and age.

Freshness: This is a binary measure that indicates whether the local copy is accurate or not. The freshness of a page *p* in the repository at time *t* is defined as:

$$F_p(t) = \begin{cases} 1 & \text{if } p \text{ is equal to the local copy at time } t \\ 0 & \text{otherwise} \end{cases}$$

Age: This is a measure that indicates how outdated the local copy is. The age of a page *p* in the repository, at time *t* is defined as:

$$A_p(t) = \begin{cases} 0 & \text{if } p \text{ is not modified at time } t \\ t - \text{modification time of } p & \text{otherwise} \end{cases}$$

Coffman *et al.* worked with a definition of the objective of a Web crawler that is equivalent to freshness, but use a different wording: they propose that a crawler must minimize the fraction of time pages remain outdated. They also noted that the problem of Web crawling can be modeled as a multiple-queue, single-server polling system, on which the Web crawler is the server and the Web sites are the queues. Page modifications are the arrival of the customers, and switch-over times are the interval between page accesses to a single Web site. Under this model, mean waiting time for a customer in the polling system is equivalent to the average age for the Web crawler.

The objective of the crawler is to keep the average freshness of pages in its collection as high as possible, or to keep the average age of pages as low as possible. These objectives are not equivalent: in the first case, the crawler is just concerned with how many pages are out-dated, while in the second case, the crawler is concerned with how old the local copies of pages are.

Two simple re-visiting policies were studied by Cho and Garcia-Molina:

Uniform policy: This involves re-visiting all pages in the collection with the same frequency, regardless of their rates of change.

Proportional policy: This involves re-visiting more often the pages that change more frequently. The visiting frequency is directly proportional to the (estimated) change frequency.

(In both cases, the repeated crawling order of pages can be done either in a random or a fixed order.)

Cho and Garcia-Molina proved the surprising result that, in terms of average freshness, the uniform policy outperforms the proportional policy in both a simulated Web and a real Web crawl. The explanation for this result comes from the fact that, when a page changes too often, the crawler will waste time by trying to re-crawl it too fast and still will not be able to keep its copy of the page fresh.

To improve freshness, the crawler should penalize the elements that change too often. The optimal re-visiting policy is neither the uniform policy nor the proportional policy. The optimal method for keeping average freshness high includes ignoring the pages that change too often, and the optimal for keeping average age low is to use access frequencies that monotonically (and sub-linearly) increase with the rate of change of each page. In both cases, the optimal is closer to the uniform policy than to the proportional policy: as Coffman *et al.* note, "in order to minimize the expected obsolescence time, the accesses to any particular page should be kept as evenly spaced as possible". Explicit formulas for the re-visit policy are not attainable in general, but they are obtained numerically, as they depend on the distribution of page changes. Cho and Garcia-Molina show that the exponential distribution is a good fit for describing page changes, while Ipeirotis *et al.* show how to use statistical tools to discover parameters that affect this distribution. Note that the re-visiting policies considered here regard all pages as homogeneous in terms of quality ("all pages on the Web are worth the same"), something that is not a realistic scenario, so further information about the Web page quality should be included to achieve a better crawling policy.

Politeness policy

Crawlers can retrieve data much quicker and in greater depth than human searchers, so they can have a crippling impact on the performance of a site. Needless to say, if a single crawler is performing multiple requests per second and/or downloading large files, a server would have a hard time keeping up with requests from multiple crawlers.

As noted by Koster, the use of Web crawlers is useful for a number of tasks, but comes with a price for the general community. The costs of using Web crawlers include:

- network resources, as crawlers require considerable bandwidth and operate with a high degree of parallelism during a long period of time;
- server overload, especially if the frequency of accesses to a given server is too high;
- poorly-written crawlers, which can crash servers or routers, or which download pages they cannot handle; and

- personal crawlers that, if deployed by too many users, can disrupt networks and Web servers.

A partial solution to these problems is the robots exclusion protocol, also known as the robots.txt protocol that is a standard for administrators to indicate which parts of their Web servers should not be accessed by crawlers. This standard does not include a suggestion for the interval of visits to the same server, even though this interval is the most effective way of avoiding server overload. Recently commercial search engines like Ask Jeeves, MSN and Yahoo are able to use an extra "Crawl-delay:" parameter in the robots.txt file to indicate the number of seconds to delay between requests.

The first proposal for the interval between connections was given in and was 60 seconds. However, if pages were downloaded at this rate from a website with more than 100,000 pages over a perfect connection with zero latency and infinite bandwidth, it would take more than 2 months to download only that entire Web site; also, only a fraction of the resources from that Web server would be used. This does not seem acceptable.

Cho uses 10 seconds as an interval for accesses, and the WIRE crawler uses 15 seconds as the default. The MercatorWeb crawler follows an adaptive politeness policy: if it took t seconds to download a document from a given server, the crawler waits for $10t$ seconds before downloading the next page. Dill *et al.* use 1 second.

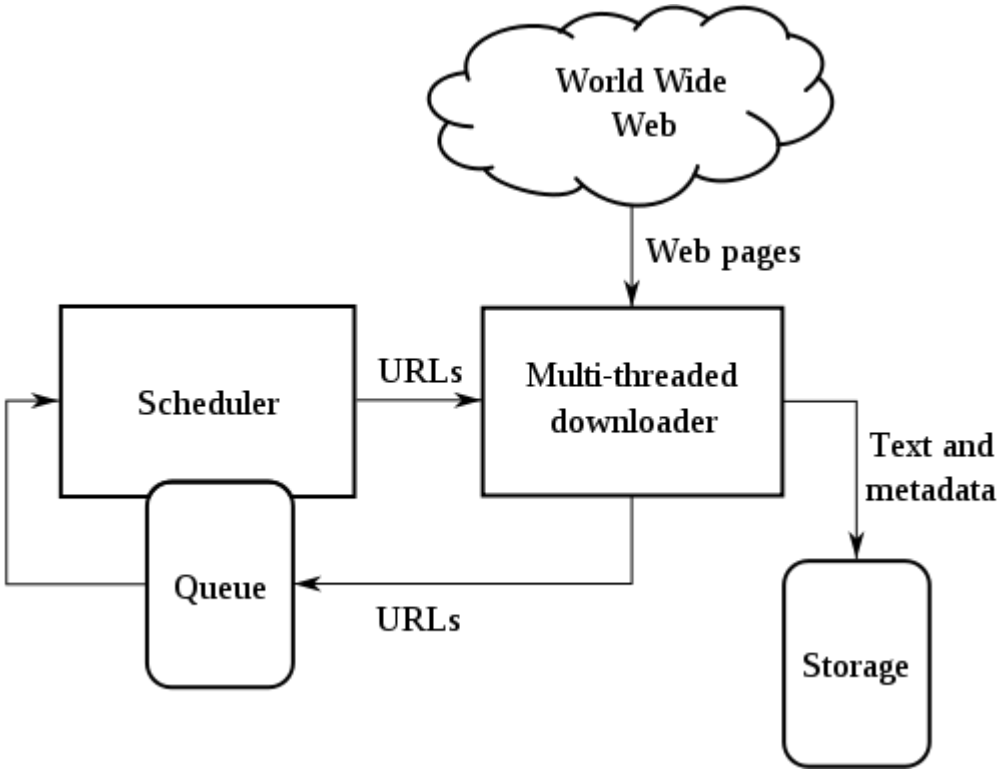
For those using Web crawlers for research purposes, a more detailed cost-benefit analysis is needed and ethical considerations should be taken into account when deciding where to crawl and how fast to crawl .

Anecdotal evidence from access logs shows that access intervals from known crawlers vary between 20 seconds and 3–4 minutes. It is worth noticing that even when being very polite, and taking all the safeguards to avoid overloading Web servers, some complaints from Web server administrators are received. Brin and Page note that: "... running a crawler which connects to more than half a million servers (...) generates a fair amount of e-mail and phone calls. Because of the vast number of people coming on line, there are always those who do not know what a crawler is, because this is the first one they have seen."

Parallelization policy

A parallel crawler is a crawler that runs multiple processes in parallel. The goal is to maximize the download rate while minimizing the overhead from parallelization and to avoid repeated downloads of the same page. To avoid downloading the same page more than once, the crawling system requires a policy for assigning the new URLs discovered during the crawling process, as the same URL can be found by two different crawling processes.

Web crawler architectures



High-level architecture of a standard Web crawler

A crawler must not only have a good crawling strategy, as noted in the previous sections, but it should also have a highly optimized architecture.

Shkapenyuk and Suel noted that: "While it is fairly easy to build a slow crawler that downloads a few pages per second for a short period of time, building a high-performance system that can download hundreds of millions of pages over several weeks presents a number of challenges in system design, I/O and network efficiency, and robustness and manageability."

Web crawlers are a central part of search engines, and details on their algorithms and architecture are kept as business secrets. When crawler designs are published, there is often an important lack of detail that prevents others from reproducing the work. There are also emerging concerns about "search engine spamming", which prevent major search engines from publishing their ranking algorithms.

Crawler identification

Web crawlers typically identify themselves to a Web server by using the User-agent field of an HTTP request. Web site administrators typically examine their Web servers' log and use the user agent field to determine which crawlers have visited the web server and

how often. The user agent field may include a URL where the Web site administrator may find out more information about the crawler. Spambots and other malicious Web crawlers are unlikely to place identifying information in the user agent field, or they may mask their identity as a browser or other well-known crawler.

It is important for Web crawlers to identify themselves so that Web site administrators can contact the owner if needed. In some cases, crawlers may be accidentally trapped in a crawler trap or they may be overloading a Web server with requests, and the owner needs to stop the crawler. Identification is also useful for administrators that are interested in knowing when they may expect their Web pages to be indexed by a particular search engine.

Examples of Web crawlers

The following is a list of published crawler architectures for general-purpose crawlers (excluding focused web crawlers), with a brief description that includes the names given to the different components and outstanding features:

- **Yahoo! Slurp** is the name of the Yahoo Search crawler.
- **Msnbot** is the name of Microsoft's Bing webcrawler.
- **FAST Crawler** is a distributed crawler, used by Fast Search & Transfer, and a general description of its architecture is available.
- **Googlebot** is described in some detail, but the reference is only about an early version of its architecture, which was based in C++ and Python. The crawler was integrated with the indexing process, because text parsing was done for full-text indexing and also for URL extraction. There is a URL server that sends lists of URLs to be fetched by several crawling processes. During parsing, the URLs found were passed to a URL server that checked if the URL have been previously seen. If not, the URL was added to the queue of the URL server.
- **Methabot** is a scriptable web crawler written in C, released under the ISC license.
- **PolyBot** is a distributed crawler written in C++ and Python, which is composed of a "crawl manager", one or more "downloaders" and one or more "DNS resolvers". Collected URLs are added to a queue on disk, and processed later to search for seen URLs in batch mode. The politeness policy considers both third and second level domains because third level domains are usually hosted by the same Web server.
- **RBSE** was the first published web crawler. It was based on two programs: the first program, "`spider`" maintains a queue in a relational database, and the second program "`mite`", is a modified `www` ASCII browser that downloads the pages from the Web.
- **WebCrawler** was used to build the first publicly-available full-text index of a subset of the Web. It was based on lib-WWW to download pages, and another program to parse and order URLs for breadth-first exploration of the Web graph. It also included a real-time crawler that followed links based on the similarity of the anchor text with the provided query.

- **World Wide Web Worm** was a crawler used to build a simple index of document titles and URLs. The index could be searched by using the `grep` Unix command.
- **WebFountain** is a distributed, modular crawler similar to Mercator but written in C++. It features a "controller" machine that coordinates a series of "ant" machines. After repeatedly downloading pages, a change rate is inferred for each page and a non-linear programming method must be used to solve the equation system for maximizing freshness. The authors recommend to use this crawling order in the early stages of the crawl, and then switch to a uniform crawling order, in which all pages are being visited with the same frequency.
- **WebRACE** is a crawling and caching module implemented in Java, and used as a part of a more generic system called eRACE. The system receives requests from users for downloading web pages, so the crawler acts in part as a smart proxy server. The system also handles requests for "subscriptions" to Web pages that must be monitored: when the pages change, they must be downloaded by the crawler and the subscriber must be notified. The most outstanding feature of WebRACE is that, while most crawlers start with a set of "seed" URLs, WebRACE is continuously receiving new starting URLs to crawl from.

In addition to the specific crawler architectures listed above, there are general crawler architectures published by Cho and Chakrabarti.

Open-source crawlers

- **Aspseek** is a crawler, indexer and a search engine written in C++ and licensed under the GPL
- **crawler4j** is a crawler written in Java and released under an Apache License. It can be configured in a few minutes and is suitable for educational purposes.
- **DataparkSearch** is a crawler and search engine released under the GNU General Public License.
- **GNU Wget** is a command-line-operated crawler written in C and released under the GPL. It is typically used to mirror Web and FTP sites.
- **Heritrix** is the Internet Archive's archival-quality crawler, designed for archiving periodic snapshots of a large portion of the Web. It was written in Java.
- **ht://Dig** includes a Web crawler in its indexing engine.
- **HTTrack** uses a Web crawler to create a mirror of a web site for off-line viewing. It is written in C and released under the GPL.
- **ICDL Crawler** is a cross-platform web crawler written in C++ and intended to crawl Web sites based on Web-site Parse Templates using computer's free CPU resources only.
- **mnoGoSearch** is a crawler, indexer and a search engine written in C and licensed under the GPL
- **Nutch** is a crawler written in Java and released under an Apache License. It can be used in conjunction with the Lucene text-indexing package.
- **Open Search Server** is a search engine and web crawler software release under the GPL.

- **Pavuk** is a command-line Web mirror tool with optional X11 GUI crawler and released under the GPL. It has bunch of advanced features compared to wget and httrack, e.g., regular expression based filtering and file creation rules.
- the **tkWWW Robot**, a crawler based on the tkWWW web browser (licensed under GPL).
- **YaCy**, a free distributed search engine, built on principles of peer-to-peer networks (licensed under GPL).

Crawling the Deep Web and Web Applications

Crawling the Deep Web

A vast amount of Web pages lie in the deep or invisible Web. These pages are typically only accessible by submitting queries to a database, and regular crawlers are unable to find these pages if there are no links that point to them. Google's Sitemap Protocol and mod oai are intended to allow discovery of these deep-Web resources.

Deep Web crawling also multiplies the number of Web links to be crawled. Some crawlers only take some of the `<a href="URL"`-shaped URLs. In some cases, such as the Googlebot, Web crawling is done on all text contained inside the hypertext content, tags, or text.

Crawling Web 2.0 Applications

- Sheeraj Shah provides insight into Crawling Ajax-driven Web 2.0 Applications.
- Interested readers might wish to read AJAXSearch: Crawling, Indexing and Searching Web 2.0 Applications.
- Making AJAX Applications Crawlable, from Google Code. It defines an agreement between web servers and search engine crawlers that allows for dynamically created content to be visible to crawlers. Google currently supports this agreement.

2. Index (search engine)

Search engine indexing collects, parses, and stores data to facilitate fast and accurate information retrieval. Index design incorporates interdisciplinary concepts from linguistics, cognitive psychology, mathematics, informatics, physics, and computer science. An alternate name for the process in the context of search engines designed to find web pages on the Internet is **Web indexing**.

Popular engines focus on the full-text indexing of online, natural language documents. Media types such as video and audio and graphics are also searchable.

Meta search engines reuse the indices of other services and do not store a local index, whereas cache-based search engines permanently store the index along with the corpus. Unlike full-text indices, partial-text services restrict the depth indexed to reduce index size. Larger services typically perform indexing at a predetermined time interval due to the required time and processing costs, while agent-based search engines index in real time.

Indexing

The purpose of storing an index is to optimize speed and performance in finding relevant documents for a search query. Without an index, the search engine would scan every document in the corpus, which would require considerable time and computing power. For example, while an index of 10,000 documents can be queried within milliseconds, a sequential scan of every word in 10,000 large documents could take hours. The additional computer storage required to store the index, as well as the considerable increase in the time required for an update to take place, are traded off for the time saved during information retrieval.

Index Design Factors

Major factors in designing a search engine's architecture include:

Merge factors
: How data enters the index, or how words or subject features are added to the index during text corpus traversal, and whether multiple indexers can work asynchronously. The indexer must first check whether it is updating old content or adding new content. Traversal typically correlates to the data collection policy. Search engine index merging is similar in concept to the SQL Merge command and other merge algorithms.

Storage techniques
: How to store the index data, that is, whether information should be data compressed or filtered.

Index size
: How much computer storage is required to support the index.

Lookup speed
: How quickly a word can be found in the inverted index. The speed of finding an entry in a data structure, compared with how quickly it can be updated or removed, is a central focus of computer science.

Maintenance
: How the index is maintained over time.

Fault tolerance
: How important it is for the service to be reliable. Issues include dealing with index corruption, determining whether bad data can be treated in isolation, dealing with bad hardware, partitioning, and schemes such as hash-based or composite partitioning, as well as replication.

Index Data Structures

Search engine architectures vary in the way indexing is performed and in methods of index storage to meet the various design factors. Types of indices include:

Suffix tree
: Figuratively structured like a tree, supports linear time lookup. Built by storing the suffixes of words. The suffix tree is a type of trie. Tries support extendable hashing, which is important for search engine indexing. Used for searching for patterns in DNA sequences and clustering. A major drawback is that the storage of a word in the tree may require more storage than storing the word itself. An alternate representation is a suffix array, which is considered to require less virtual memory and supports data compression such as the BWT algorithm.

Inverted index
: Stores a list of occurrences of each atomic search criterion, typically in the form of a hash table or binary tree.

Citation index
: Stores citations or hyperlinks between documents to support citation analysis, a subject of Bibliometrics.

Ngram index
: Stores sequences of length of data to support other types of retrieval or text mining.

Document-term matrix
: Used in latent semantic analysis, stores the occurrences of words in documents in a two-dimensional sparse matrix.

Challenges in Parallelism

A major challenge in the design of search engines is the management of serial computing processes. There are many opportunities for race conditions and coherent faults. For example, a new document is added to the corpus and the index must be updated, but the index simultaneously needs to continue responding to search queries. This is a collision between two competing tasks. Consider that authors are producers of information, and a web crawler is the consumer of this information, grabbing the text and storing it in a cache (or corpus). The forward index is the consumer of the information produced by the corpus, and the inverted index is the consumer of information produced by the forward index. This is commonly referred to as a **producer-consumer model**. The indexer is the producer of searchable information and users are the consumers that need to search. The challenge is magnified when working with distributed storage and distributed processing. In an effort to scale with larger amounts of indexed information, the search engine's architecture may involve distributed computing, where the search engine consists of several machines operating in unison. This increases the possibilities for incoherency and makes it more difficult to maintain a fully synchronized, distributed, parallel architecture.

Inverted indices

Many search engines incorporate an inverted index when evaluating a search query to quickly locate documents containing the words in a query and then rank these documents by relevance. Because the inverted index stores a list of the documents containing each word, the search engine can use direct access to find the documents associated with each word in the query in order to retrieve the matching documents quickly. The following is a simplified illustration of an inverted index:

Inverted Index

Word	Documents
the	Document 1, Document 3, Document 4, Document 5
cow	Document 2, Document 3, Document 4
says	Document 5
moo	Document 7

This index can only determine whether a word exists within a particular document, since it stores no information regarding the frequency and position of the word; it is therefore considered to be a boolean index. Such an index determines which documents match a query but does not rank matched documents. In some designs the index includes additional information such as the frequency of each word in each document or the positions of a word in each document. Position information enables the search algorithm to identify word proximity to support searching for phrases; frequency can be used to help in ranking the relevance of documents to the query. Such topics are the central research focus of information retrieval.

The inverted index is a sparse matrix, since not all words are present in each document. To reduce computer storage memory requirements, it is stored differently from a two dimensional array. The index is similar to the term document matrices employed by latent semantic analysis. The inverted index can be considered a form of a hash table. In some cases the index is a form of a binary tree, which requires additional storage but may reduce the lookup time. In larger indices the architecture is typically a distributed hash table.

Index Merging

The inverted index is filled via a merge or rebuild. A rebuild is similar to a merge but first deletes the contents of the inverted index. The architecture may be designed to support incremental indexing, where a merge identifies the document or documents to be added or updated and then parses each document into words. For technical accuracy, a merge conflates newly indexed documents, typically residing in virtual memory, with the index cache residing on one or more computer hard drives.

After parsing, the indexer adds the referenced document to the document list for the appropriate words. In a larger search engine, the process of finding each word in the

inverted index (in order to report that it occurred within a document) may be too time consuming, and so this process is commonly split up into two parts, the development of a forward index and a process which sorts the contents of the forward index into the inverted index. The inverted index is so named because it is an inversion of the forward index.

The Forward Index

The forward index stores a list of words for each document. The following is a simplified form of the forward index:

Forward Index

Document	Words
Document 1	the,cow,says,moo
Document 2	the,cat,and,the,hat
Document 3	the,dish,ran,away,with,the,fork

The rationale behind developing a forward index is that as documents are parsing, it is better to immediately store the words per document. The delineation enables Asynchronous system processing, which partially circumvents the inverted index update bottleneck. The forward index is sorted to transform it to an inverted index. The forward index is essentially a list of pairs consisting of a document and a word, collated by the document. Converting the forward index to an inverted index is only a matter of sorting the pairs by the words. In this regard, the inverted index is a word-sorted forward index.

Compression

Generating or maintaining a large-scale search engine index represents a significant storage and processing challenge. Many search engines utilize a form of compression to reduce the size of the indices on disk. Consider the following scenario for a full text, Internet search engine.

- An estimated 2,000,000,000 different web pages exist as of the year 2000
- Suppose there are 250 words on each webpage (based on the assumption they are similar to the pages of a novel.
- It takes 8 bits (or 1 byte) to store a single character. Some encodings use 2 bytes per character
- The average number of characters in any given word on a page may be estimated at 5
- The average personal computer comes with 100 to 250 gigabytes of usable space

Given this scenario, an uncompressed index (assuming a non-conflated, simple, index) for 2 billion web pages would need to store 500 billion word entries. At 1 byte per character, or 5 bytes per word, this would require 2500 gigabytes of storage space alone, more than the average free disk space of 25 personal computers. This space requirement

may be even larger for a fault-tolerant distributed storage architecture. Depending on the compression technique chosen, the index can be reduced to a fraction of this size. The tradeoff is the time and processing power required to perform compression and decompression.

Notably, large scale search engine designs incorporate the cost of storage as well as the costs of electricity to power the storage. Thus compression is a measure of cost.

Document Parsing

Document parsing breaks apart the components (words) of a document or other form of media for insertion into the forward and inverted indices. The words found are called *tokens*, and so, in the context of search engine indexing and natural language processing, parsing is more commonly referred to as tokenization. It is also sometimes called word boundary disambiguation, tagging, text segmentation, content analysis, text analysis, text mining, concordance generation, speech segmentation, lexing, or lexical analysis. The terms 'indexing', 'parsing', and 'tokenization' are used interchangeably in corporate slang.

Natural language processing, as of 2006, is the subject of continuous research and technological improvement. Tokenization presents many challenges in extracting the necessary information from documents for indexing to support quality searching. Tokenization for indexing involves multiple technologies, the implementation of which are commonly kept as corporate secrets.

Challenges in Natural Language Processing

Word Boundary Ambiguity
: Native English speakers may at first consider tokenization to be a straightforward task, but this is not the case with designing a multilingual indexer. In digital form, the texts of other languages such as Chinese, Japanese or Arabic represent a greater challenge, as words are not clearly delineated by whitespace. The goal during tokenization is to identify words for which users will search. Language-specific logic is employed to properly identify the boundaries of words, which is often the rationale for designing a parser for each language supported (or for groups of languages with similar boundary markers and syntax).

Language Ambiguity
: To assist with properly ranking matching documents, many search engines collect additional information about each word, such as its language or lexical category (part of speech). These techniques are language-dependent, as the syntax varies among languages. Documents do not always clearly identify the language of the document or represent it accurately. In tokenizing the document, some search engines attempt to automatically identify the language of the document.

Diverse File Formats
: In order to correctly identify which bytes of a document represent characters, the file format must be correctly handled. Search engines which support multiple file

formats must be able to correctly open and access the document and be able to tokenize the characters of the document.

Faulty Storage

The quality of the natural language data may not always be perfect. An unspecified number of documents, particular on the Internet, do not closely obey proper file protocol. binary characters may be mistakenly encoded into various parts of a document. Without recognition of these characters and appropriate handling, the index quality or indexer performance could degrade.

Tokenization

Unlike literate humans, computers do not understand the structure of a natural language document and cannot automatically recognize words and sentences. To a computer, a document is only a sequence of bytes. Computers do not 'know' that a space character separates words in a document. Instead, humans must program the computer to identify what constitutes an individual or distinct word, referred to as a token. Such a program is commonly called a tokenizer or parser or lexer. Many search engines, as well as other natural language processing software, incorporate specialized programs for parsing, such as YACC or Lex.

During tokenization, the parser identifies sequences of characters which represent words and other elements, such as punctuation, which are represented by numeric codes, some of which are non-printing control characters. The parser can also identify entities such as email addresses, phone numbers, and URLs. When identifying each token, several characteristics may be stored, such as the token's case (upper, lower, mixed, proper), language or encoding, lexical category (part of speech, like 'noun' or 'verb'), position, sentence number, sentence position, length, and line number.

Language Recognition

If the search engine supports multiple languages, a common initial step during tokenization is to identify each document's language; many of the subsequent steps are language dependent (such as stemming and part of speech tagging). Language recognition is the process by which a computer program attempts to automatically identify, or categorize, the language of a document. Other names for language recognition include language classification, language analysis, language identification, and language tagging. Automated language recognition is the subject of ongoing research in natural language processing. Finding which language the words belongs to may involve the use of a language recognition chart.

Format Analysis

If the search engine supports multiple document formats, documents must be prepared for tokenization. The challenge is that many document formats contain formatting information in addition to textual content. For example, HTML documents contain HTML tags, which specify formatting information such as new line starts, **bold**

emphasis, and font size or style. If the search engine were to ignore the difference between content and 'markup', extraneous information would be included in the index, leading to poor search results. Format analysis is the identification and handling of the formatting content embedded within documents which controls the way the document is rendered on a computer screen or interpreted by a software program. Format analysis is also referred to as structure analysis, format parsing, tag stripping, format stripping, text normalization, text cleaning, and text preparation. The challenge of format analysis is further complicated by the intricacies of various file formats. Certain file formats are proprietary with very little information disclosed, while others are well documented. Common, well-documented file formats that many search engines support include:

- HTML
- ASCII text files (a text document without specific computer readable formatting)
- Adobe's Portable Document Format (PDF)
- PostScript (PS)
- LaTeX
- UseNet netnews server formats
- XML and derivatives like RSS
- SGML
- Multimedia meta data formats like ID3
- Microsoft Word
- Microsoft Excel
- Microsoft Powerpoint
- IBM Lotus Notes

Options for dealing with various formats include using a publicly available commercial parsing tool that is offered by the organization which developed, maintains, or owns the format, and writing a custom parser.

Some search engines support inspection of files that are stored in a compressed or encrypted file format. When working with a compressed format, the indexer first decompresses the document; this step may result in one or more files, each of which must be indexed separately. Commonly supported compressed file formats include:

- ZIP - Zip archive file
- RAR - Roshal ARchive File
- CAB - Microsoft Windows Cabinet File
- Gzip - File compressed with gzip
- BZIP - File compressed using bzip2
- Tape ARchive (TAR), Unix archive file, not (itself) compressed
- TAR.Z, TAR.GZ or TAR.BZ2 - Unix archive files compressed with Compress, GZIP or BZIP2

Format analysis can involve quality improvement methods to avoid including 'bad information' in the index. Content can manipulate the formatting information to include additional content. Examples of abusing document formatting for spamdexing:

- Including hundreds or thousands of words in a section which is hidden from view on the computer screen, but visible to the indexer, by use of formatting (e.g. hidden "div" tag in HTML, which may incorporate the use of CSS or Javascript to do so).
- Setting the foreground font color of words to the same as the background color, making words hidden on the computer screen to a person viewing the document, but not hidden to the indexer.

Section Recognition

Some search engines incorporate section recognition, the identification of major parts of a document, prior to tokenization. Not all the documents in a corpus read like a well-written book, divided into organized chapters and pages. Many documents on the web, such as newsletters and corporate reports, contain erroneous content and side-sections which do not contain primary material (that which the document is about). For example, this article displays a side menu with links to other web pages. Some file formats, like HTML or PDF, allow for content to be displayed in columns. Even though the content is displayed, or rendered, in different areas of the view, the raw markup content may store this information sequentially. Words that appear sequentially in the raw source content are indexed sequentially, even though these sentences and paragraphs are rendered in different parts of the computer screen. If search engines index this content as if it were normal content, the quality of the index and search quality may be degraded due to the mixed content and improper word proximity. Two primary problems are noted:

- Content in different sections is treated as related in the index, when in reality it is not
- Organizational 'side bar' content is included in the index, but the side bar content does not contribute to the meaning of the document, and the index is filled with a poor representation of its documents.

Section analysis may require the search engine to implement the rendering logic of each document, essentially an abstract representation of the actual document, and then index the representation instead. For example, some content on the Internet is rendered via Javascript. If the search engine does not render the page and evaluate the Javascript within the page, it would not 'see' this content in the same way and would index the document incorrectly. Given that some search engines do not bother with rendering issues, many web page designers avoid displaying content via Javascript or use the Noscript tag to ensure that the web page is indexed properly. At the same time, this fact can also be exploited to cause the search engine indexer to 'see' different content than the viewer.

Meta Tag Indexing

Specific documents often contain embedded meta information such as author, keywords, description, and language. For HTML pages, the meta tag contains keywords which are also included in the index. Earlier Internet search engine technology would only index the

keywords in the meta tags for the forward index; the full document would not be parsed. At that time full-text indexing was not as well established, nor was the hardware able to support such technology. The design of the HTML markup language initially included support for meta tags for the very purpose of being properly and easily indexed, without requiring tokenization.

As the Internet grew through the 1990s, many brick-and-mortar corporations went 'online' and established corporate websites. The keywords used to describe webpages (many of which were corporate-oriented webpages similar to product brochures) changed from descriptive to marketing-oriented keywords designed to drive sales by placing the webpage high in the search results for specific search queries. The fact that these keywords were subjectively specified was leading to spamdexing, which drove many search engines to adopt full-text indexing technologies in the 1990s. Search engine designers and companies could only place so many 'marketing keywords' into the content of a webpage before draining it of all interesting and useful information. Given that conflict of interest with the business goal of designing user-oriented websites which were 'sticky', the customer lifetime value equation was changed to incorporate more useful content into the website in hopes of retaining the visitor. In this sense, full-text indexing was more objective and increased the quality of search engine results, as it was one more step away from subjective control of search engine result placement, which in turn furthered research of full-text indexing technologies.

In Desktop search, many solutions incorporate meta tags to provide a way for authors to further customize how the search engine will index content from various files that is not evident from the file content. Desktop search is more under the control of the user, while Internet search engines must focus more on the full text index.

3. Web search query

A **web search query** is a query that a user enters into web search engine to satisfy his or her information needs. Web search queries are distinctive in that they are unstructured and often ambiguous; they vary greatly from standard query languages which are governed by strict syntax rules.

Types

There are four broad categories that cover most web search queries:

- **Informational queries** – Queries that cover a broad topic (e.g., *colorado* or *trucks*) for which there may be thousands of relevant results.

- **Navigational queries** – Queries that seek a single website or web page of a single entity (e.g., *youtube* or *delta airlines*).

- **Transactional queries** – Queries that reflect the intent of the user to perform a particular action, like purchasing a car or downloading a screen saver.

Search engines often support a fourth type of query that is used far less frequently:

- **Connectivity queries** – Queries that report on the connectivity of the indexed web graph (e.g., Which links point to this URL? and How many pages are indexed from this domain name?).

Characteristics

Most commercial web search engines do not disclose their search logs, so information about what users are searching for on the Web is difficult to come by. Nevertheless, a study in 2001 analyzed the queries from the Excite search engine showed some interesting characteristics of web search:

- The average length of a search query was 2.4 terms.
- About half of the users entered a single query while a little less than a third of users entered three or more unique queries.
- Close to half of the users examined only the first one or two pages of results (10 results per page).
- Less than 5% of users used advanced search features (e.g., Boolean operators like AND, OR, and NOT).
- The top four most frequently used terms were, *(empty search), and, of, and sex.*

A study of the same Excite query logs revealed that 19% of the queries contained a geographic term (e.g., place names, zip codes, geographic features, etc.).

A 2005 study of Yahoo's query logs revealed 33% of the queries from the same user were repeat queries and that 87% of the time the user would click on the same result. This suggests that many users use repeat queries to revisit or re-find information.

In addition, much research has shown that query term frequency distributions conform to the power law, or *long tail* distribution curves. That is, a small portion of the terms observed in a large query log (e.g. > 100 million queries) are used most often, while the remaining terms are used less often individually. This example of the Pareto principle (or *80-20 rule*) allows search engines to employ optimization techniques such as index or database partitioning, caching and pre-fetching.

Structured queries

With search engines that support Boolean operators and parentheses, a technique traditionally used by librarians can be applied. A user who is looking for documents that cover several topics or *facets* may want to describe each of them by a disjunction of characteristic words, such as `vehicles OR cars OR automobiles`. A *faceted query* is a

conjunction of such facets; e.g. a query such as `(electronic OR computerized OR DRE) AND (voting OR elections OR election OR balloting OR electoral)` is likely to find documents about electronic voting even if they omit one of the words "electronic" and "voting", or even both.

Market share and wars

According to Hitbox, Google's worldwide popularity peaked at 82.7% in December, 2008. July 2009 rankings showed Google (78.4%) losing traffic to Baidu (8.87%), and Bing (3.17%). The market share of Yahoo! Search (7.16%) and AOL (0.6%) were also declining.

In the United States, Google held a 63.2% market share in May 2009, according to Nielsen NetRatings. In the People's Republic of China, Baidu held a 61.6% market share for web search in July 2009.

Search engine bias

Although search engines are programmed to rank websites based on their popularity and relevancy, empirical studies indicate various political, economic, and social biases in the information they provide. These biases could be a direct result of economic and commercial processes (e.g., companies that advertise with a search engine can become also more popular in its organic search results), and political processes (e.g., the removal of search results in order to comply with local laws). Google Bombing is one example of an attempt to manipulate search results for political, social or commercial reasons.

Chapter 8

Web Page

A **web page** or **webpage** is a document or information resource that is suitable for the World Wide Web and can be accessed through a web browser and displayed on a monitor or mobile device. This information is usually in HTML or XHTML format, and may provide navigation to other web pages via hypertext links. Web pages frequently subsume other resources such as style sheets, scripts and images into their final presentation.

Web pages may be retrieved from a local computer or from a remote web server. The web server may restrict access only to a private network, e.g. a corporate intranet, or it may publish pages on the World Wide Web. Web pages are requested and served from web servers using Hypertext Transfer Protocol (HTTP).

Web pages may consist of files of static text and other content stored within the web server's file system (static web pages), or may be constructed by server-side software when they are requested (dynamic web pages). Client-side scripting can make web pages more responsive to user input once on the client browser.

Color, typography, illustration, and interaction

Web pages usually include information as to the colors of text and backgrounds and very often also contain links to images and sometimes other types of media to be included in the final view. Layout, typographic and color-scheme information is provided by Cascading Style Sheet (CSS) instructions, which can either be embedded in the HTML or can be provided by a separate file, which is referenced from within the HTML. The latter case is especially relevant where one lengthy stylesheet is relevant to a whole website: due to the way HTTP works, the browser will only download it once from the web server and use the cached copy for the whole site. Images are stored on the web server as separate files, but again HTTP allows for the fact that once a web page is downloaded to a browser, it is quite likely that related files such as images and stylesheets will be requested as it is processed. An HTTP 1.1 web server will maintain a connection with the browser until all related resources have been requested and provided. Web browsers usually render images along with the text and other material on the displayed web page.

Dynamic web page

A **dynamic web page** is a kind of web page that has been prepared with fresh information (content and/or layout), for each individual viewing. It is not static because it changes with the time (ex. a news content), the user (ex. preferences in a login session), the user interaction (ex. web page game), the context (parametric customization), or any combination of the foregoing.

Properties associated with dynamic web pages

Classical hypertext navigation occurs among "static" documents, and, for *web users*, this experience is reproduced using static web pages, meaning that a page retrieved by different users at different times is always the same, in the same form.

However, a web page can also provide a *live* user experience. Content (text, images, form fields, etc.) on a web page can change in response to different contexts or conditions. In dynamic sites, page content and page layout are created separately. The content is retrieved from a database and is placed on a web page only when needed or asked. This allows for quicker page loading, and it allows just about anyone with limited web design experience to update their own website via an administrative tool. This set-up is ideal for those who wish to make frequent changes to their websites including text and image updates, e.g. e-commerce.

Two types of dynamic web sites

Client-side scripting and content creation

Using client-side scripting to change interface behaviors *within* a specific web page, in response to mouse or keyboard actions or at specified timing events. In this case the dynamic behavior occurs within the presentation.

Such web pages use presentation technology called rich interfaced pages. Client-side scripting languages like JavaScript or ActionScript, used for Dynamic HTML (DHTML) and Flash technologies respectively, are frequently used to orchestrate media types (sound, animations, changing text, etc.) of the presentation. The scripting also allows use of remote scripting, a technique by which the DHTML page requests additional information from a server, using a hidden Frame, XMLHttpRequests, or a Web service.

The Client-side content is generated on the user's computer. The web browser retrieves a page from the server, then processes the code embedded in the page (often written in JavaScript) and displays the retrieved page's content to the user.

The innerHTML property (or write command) can illustrate the client-side dynamic page generation: two distinct pages, A and B, can be regenerated as `document.innerHTML =`

`A` and `document.innerHTML = B`; or "on load dynamic" by `document.write(A)` and `document.write(B)`.

There are also some utilities and frameworks for converting HTML files into JavaScript files. For example webJS uses innerHTML property for rendering pages from converted HTML on client-side.

The first "widespread used" version of JavaScript was 1996 (with Netscape 3 an ECMAscript standard).

Server-side scripting and content creation

Using server-side scripting to change the supplied page source *between* pages, adjusting the sequence or reload of the web pages or web content supplied to the browser. Server responses may be determined by such conditions as data in a posted HTML form, parameters in the URL, the type of browser being used, the passage of time, or a database or server state.

Such web pages are often created with the help of server-side languages such as PHP, Perl, ASP, ASP.NET, JSP, ColdFusion and other languages. These server-side languages typically use the Common Gateway Interface (CGI) to produce *dynamic web pages*. These kinds of pages can also use, on the client-side, the first kind (DHTML, etc.).

Server-side dynamic content is more complicated: (1) The client sends the server the request. (2) The server receives the request and processes the server-side script such as [PHP] based on the query string, HTTP POST data, cookies, etc.

The dynamic page generation was made possible by the Common Gateway Interface, stable in 1993. Then Server Side Includes pointed a more direct way to deal with server-side scripts, at the web servers.

Combining client and server side

Ajax is a web development technique for dynamically interchanging content with the server-side, without reloading the web page. Google Maps is an example of a web application that uses Ajax techniques and database.

Disadvantages

- Search engines work by creating indexes of published HTML web pages that were, initially, "static". With the advent of dynamic web pages, often created from a private database, the content is less visible. Unless this content is duplicated in some way (for example, as a series of extra static pages on the same site), a search may not find the information it is looking for. It is unreasonable to expect

generalized web search engines to be able to access complex database structures, some of which in any case may be secure.

History

It is difficult to be precise about "dynamic web page beginnings" or chronology, because the precise concept makes sense only after the "widespread development of web pages": HTTP protocol has been in use since 1990, HTML, as standard, since 1996. The web browsers explosion started with 1993's Mosaic. It is obvious, however, that the concept of dynamically driven websites predates the internet, and in fact HTML. For example, in 1990, before the general public use of the internet, a dynamically driven remotely accessed menu system was implimented by Susan Biddlecomb, at the University of Southern California BBS on a 16 line TBBS system with TDBS add-on.

Browsers

A web browser can have a Graphical User Interface, like Internet Explorer, Mozilla Firefox, Chrome and Opera, or can be text-based, like Lynx or Links.

Web users with disabilities often use assistive technologies and adaptive strategies to access web pages. Users may be color blind, may or may not want to use a mouse perhaps due to repetitive stress injury or motor-neurone problems, may be deaf and require audio to be captioned, may be blind and using a screen reader or braille display, may need screen magnification, etc.

Disabled and able-bodied users may disable the download and viewing of images and other media, to save time, network bandwidth or merely to simplify their browsing experience. Users of mobile devices often have restricted displays and bandwidth. Anyone may prefer not to use the fonts, font sizes, styles and color schemes selected by the web page designer and may apply their own CSS styling to the page.

The World Wide Web Consortium (W3C) and Web Accessibility Initiative (WAI) recommend that all web pages should be designed with all of these options in mind.

Elements

A *web page*, as an information set, can contain numerous types of information, which is able to be seen, heard or interact by the end user:

Perceived (rendered) information:

- *Textual information*: with diverse render variations.
- *Non-textual information*:
  - *Static images* may be raster graphics, typically GIF, JPEG or PNG; or vector formats such as SVG or Flash.

    - *Animated images* typically Animated GIF and SVG, but also may be Flash, Shockwave, or Java applet.
    - Audio, typically MP3, ogg or various proprietary formats.
    - Video, WMV (Windows), RM (Real Media), FLV (Flash Video), MPG, MOV (QuickTime)
- *Interactive information*
    - For "on page" interaction:
        - *Interactive text*
        - *Interactive illustrations*: ranging from "click to play" images to games, typically using *script orchestration*, Flash, Java applets, SVG, or Shockwave.
        - *Buttons*: forms providing alternative interface, typically for use with *script orchestration* and DHTML.
    - For "between pages" interaction:
        - *Hyperlinks*: standard "change page" reactivity.
        - *Forms*: providing more interaction with the server and server-side databases.

Internal (hidden) information:

- *Comments*
- *Linked Files through Hyperlink (Like DOC,XLS,PDF,etc).*
- *Metadata* with semantic meta-information, Charset information, Document Type Definition (DTD), etc.
- *Diagramation and style information*: information about rendered items (like image size attributes) and visual specifications, as Cascading Style Sheets (CSS).
- *Scripts*, usually JavaScript, complement interactivity and functionality.

Note: on server-side the web page may also have "Processing Instruction Information Items".

The web page can also contain dynamically adapted information elements, dependent upon the rendering browser or end-user location (through the use of IP address tracking and/or "cookie" information).

From a more general/wide point of view, some information (grouped) elements, like a navigation bar, are uniform for all website pages, like a standard. These kind of "website standard information" are supplied by technologies like web template systems.

Rendering

Web pages will often require more screen space than is available for a particular display resolution. Most modern browsers will place a scrollbar (a sliding tool at the side of the screen that allows the user to move the page up or down, or side-to-side) in the window to allow the user to see all content. Scrolling horizontally is less prevalent than vertical scrolling, not only because such pages often do not print properly, but because it

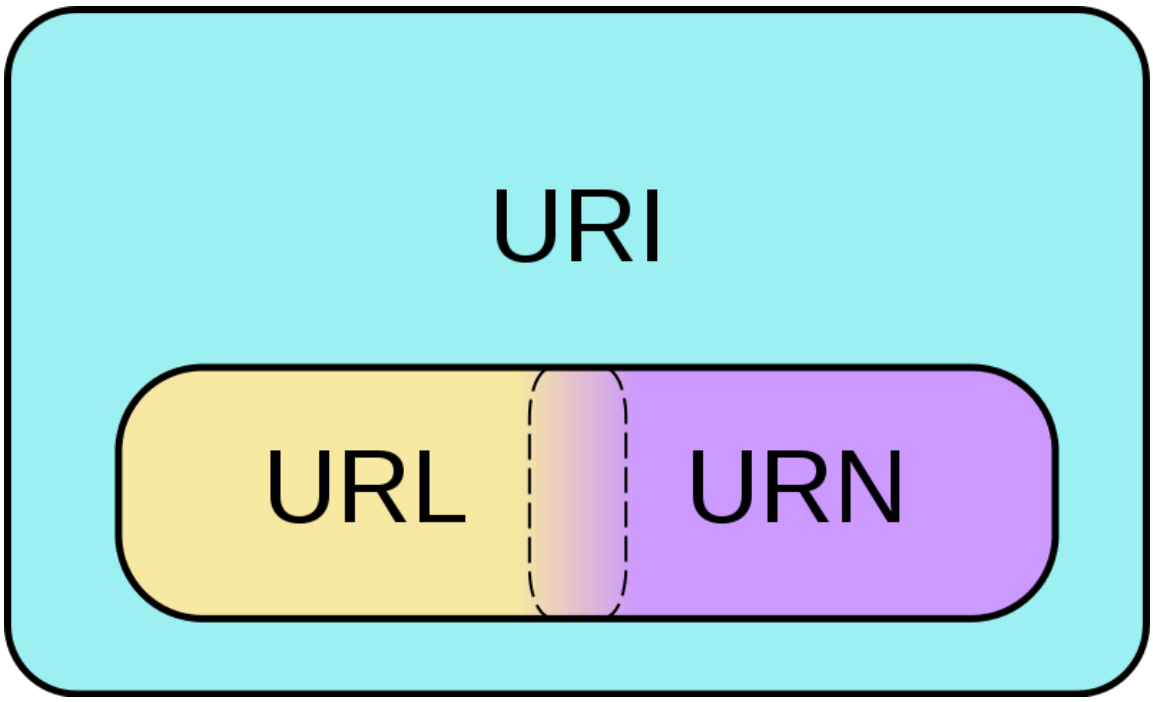
inconveniences the user more so than vertical scrolling would (because lines are horizontal; scrolling back and forth for every line is much more inconvenient than scrolling after reading a whole screen; also most computer keyboards have page up and down keys, and many computer mice have vertical scroll wheels, but the horizontal scrolling equivalents are rare).

When web pages are stored in a common directory of a web server, they become a website. A website will typically contain a group of web pages that are linked together, or have some other coherent method of navigation. The most important web page to have on a website is the index page. Depending on the web server settings, this index page can have many different names, but the most common is `index.html`. When a browser visits the homepage for a website, or any URL pointing to a directory rather than a specific file, the web server will serve the index page to the requesting browser. If no index page is defined in the configuration, or no such file exists on the server, either an error or directory listing will be served to the browser.

A web page can either be a single HTML file, or made up of several HTML files using frames or Server Side Includes (SSIs). Frames have been known to cause problems with web accessibility, copyright, navigation, printing and search engine rankings and are now less often used than they were in the 1990s. Both frames and SSIs allow certain content which appears on many pages, such as page navigation or page headers, to be repeated without duplicating the HTML in many files. Frames and the W3C recommended alternative of 2000, the <object> tag, also allow some content to remain in one place while other content can be scrolled using conventional scrollbars. Modern CSS and JavaScript client-side techniques can also achieve all of these goals and more.

When creating a web page, it is important to ensure it conforms to the World Wide Web Consortium (W3C) standards for HTML, CSS, XML and other standards. The W3C standards are in place to ensure all browsers which conform to their standards can display identical content without any special consideration for proprietary rendering techniques. A properly coded web page is going to be accessible to many different browsers old and new alike, display resolutions, as well as those users with audio or visual impairments.

Uniform Resource Locator



The relationship of URL to URI and URN

In computing, a **Uniform Resource Locator** (**URL**) is a Uniform Resource Identifier (URI) that specifies where an identified resource is available and the mechanism for retrieving it. In popular usage and in many technical documents and verbal discussions it is often incorrectly used as a synonym for URI. The best-known example of the use of URLs is for the *addresses* of web pages on the World Wide Web, such as `http://www.example.com/`.

History

The Uniform Resource Locator was created in 1994 by Tim Berners-Lee and the URI working group of the Internet Engineering Task Force. The format is based on Unix file path syntax, where forward slashes are used to separate directory or folder and file or resource names. Conventions already existed where server names could be prepended to complete file paths, preceded by a double-slash (//).

File formats may also be specified using a final dot suffix, so that requests for `file.html` or `file.txt` may be served directly whereas `file.php` needs to be sent to a PHP pre-processor before the processed result is served to the end user. The exposure of such implementation-specific details in public URLs is becoming less common; the necessary information can be better specified and exchanged using Internet media type identifiers, previously known as MIME types.

Berners-Lee later regretted the use of dots to separate the parts of the domain name within URIs, wishing he had used slashes throughout. For example, `http://www.example.com/path/to/name` would have been written `http:com/example/www/path/to/name`. Berners-Lee has also said that, given the colon following the URI scheme, the two forward slashes before the domain name were also unnecessary.

Syntax

Every URL consists of some of the following: the scheme name (commonly called protocol), followed by a colon, then, depending on scheme, a domain name (alternatively, IP address), a port number, the path of the resource to be fetched or the program to be run, then, for programs such as Common Gateway Interface (CGI) scripts, a query string, and an optional fragment identifier.

The syntax is
`scheme://domain:port/path?query_string#fragment_id`

- The scheme name defines the namespace, purpose, and the syntax of the remaining part of the URL. Software will try to process a URL according to its scheme and context. For example, a web browser will usually dereference the URL `http://example.org:80` by performing an HTTP request to the host at `example.org`, using port number 80. The URL `mailto:bob@example.com` may start an e-mail composer with the address `bob@example.com` in the To field.

Other examples of scheme names include https:, gopher:, wais:, ftp:. URLs with https as a scheme (such as `https://example.com/`) require that requests and responses will be made over a secure connection to the website. Some schemes that require authentication allow a username, and perhaps a password too, to be embedded in the URL, for example `ftp://asmith@ftp.example.org`. Passwords embedded in this way are not conducive to secure working, but the full possible syntax is
`scheme://username:password@domain:port/path?query_string#fragment_id`

- The domain name or IP address gives the destination location for the URL. The domain google.com, or its IP address `72.14.207.99`, is the address of Google's website.
- The domain name portion of a URL is not case sensitive since DNS ignores case: `http://en.example.org/` and `HTTP://EN.EXAMPLE.ORG/` both open the same page.
- The port number is optional; if omitted, the default for the scheme is used. For example, `http://vnc.example.com:5800` connects to port 5800 of vnc.example.com, which may be appropriate for a VNC remote control session. If the port number is omitted for an http: URL, the browser will connect on port 80, the default HTTP port. The default port for an https: request is 443.

- The query string contains data to be passed to software running on the server. It may contain name/value pairs separated by ampersands, for example `?first_name=John&last_name=Doe`.
- The fragment identifier, if present, specifies a part or a position within the overall resource or document. When used with HTTP, it usually specifies a section or location within the page, and the browser may scroll to display that part of the page.

Absolute vs relative URLs

According to RFC 1738, which defined URLs in 1994, when resources contain references to other resources, they can use relative links to define the location of the second resource as if to say, "in the same place as this one except with the following relative path". It went on to say that such relative URLs are dependent on the original URL containing a hierarchical structure against which the relative link is based, and that the `ftp`, `http`, and `file` URL schemes are examples of some that can be considered hierarchical, with the components of the hierarchy being separated by "/".

URLs as locators

A URL is a URI that, *"in addition to identifying a resource, provides a means of locating the resource by describing its primary access mechanism (e.g., its network location)"*.

Internet hostnames

On the Internet, a hostname is a domain name assigned to a host computer. This is usually a combination of the host's local name with its parent domain's name. For example, en.example.org consists of a local hostname (*en*) and the domain name *example.org*. The hostname is translated into an IP address via the local hosts file, or the Domain Name System (DNS) resolver. It is possible for a single host computer to have several hostnames; but generally the operating system of the host prefers to have one hostname that the host uses for itself.

Any domain name can also be a hostname, as long as the restrictions mentioned below are followed. For example, both "en.example.org" and "example.org" can be hostnames if they both have IP addresses assigned to them. The domain name "xyz.example.org" may not be a hostname if it does not have an IP address, but "aa.xyz.example.org" may still be a hostname. All hostnames are domain names, but not all domain names are hostnames.

Viewing

In order to graphically display a web page, a web browser is needed. This is a type of software that can retrieve web pages from the Internet. Most current web browsers

include the ability to view the source code. Viewing a web page in a text editor will also display the source code, not the visual product.

Creation

To create a web page, a text editor or a specialized HTML editor is needed. In order to upload the created web page to a web server, traditionally an FTP client is needed.

The design of a web page is highly personal. A design can be made according to one's own preference, or a premade web template can be used. Web templates let web page designers edit the content of a web page without having to worry about the overall aesthetics. Many people publish their own web pages using products like Tripod, or Angelfire. These web publishing tools offer free page creation and hosting up to a certain size limit.

Other ways of making a web page is to download specialized software, like a CMS, or forum. These options allow for quick and easy creation of a web page which is typically dynamic.

Saving

While one is viewing a web page, a copy of it is saved locally; this is what is being viewed. Depending on the browser settings, this copy may be deleted at any time, or stored indefinitely, sometimes without the user realizing it. Most GUI browsers provide options for saving a web page more permanently. These may include:

- Save the rendered text without formatting or images, with hyperlinks reduced to plain text
- Save the HTML as it was served — Overall structure preserved, but some links may be broken
- Save the HTML with relative links changed to absolute ones so that hyperlinks are preserved
- Save the entire web page — All images and other resources including stylesheets and scripts are downloaded and saved in a new folder alongside the HTML, with links to them altered to refer to the local copies. Other relative links changed to absolute
- Save the HTML as well as all images and other resources into a single MHTML file. This is supported by Internet Explorer and Opera. Other browsers may support this if a suitable plugin has been installed.

Most operating systems allow applications such as web browsers not only to print the currently viewed web page to a printer, but optionally to "print" to a file that can be viewed or printed later. Some web pages are designed, for example by use of CSS, so that hyperlinks, menus and other navigation items, which will be useless on paper, are rendered into print with this in mind. Sometimes, the destination addresses of hyperlinks

may be shown explicitly, either within the body of the page or listed at the end of the printed version. Web page designers may specify in CSS that non-functional menus, navigational blocks and other items may simply be absent from the printed version.

Chapter 9

Domain Name

A **domain name** is an identification label that defines a realm of administrative autonomy, authority, or control in the Internet. Domain names are also hostnames that identify Internet Protocol (IP) resources such as web sites. Domain names are formed by the rules and procedures of the Domain Name System (DNS).

Domain names are used in various networking contexts and application-specific naming and addressing purposes. They are organized in subordinate levels (subdomains) of the DNS root domain, which is nameless. The first-level set of domain names are the top-level domains (TLDs), including the generic top-level domains (gTLDs), such as the prominent domains `com`, `net` and `org`, and the country code top-level domains (ccTLDs). Below these top-level domains in the DNS hierarchy are the second-level and third-level domain names that are typically open for reservation by end-users that wish to connect local area networks to the Internet, run web sites, or create other publicly accessible Internet resources. The registration of these domain names is usually administered by domain name registrars who sell their services to the public.

Individual Internet host computers use domain names as host identifiers, or hostnames. Hostnames are the leaf labels in the domain name system usually without further subordinate domain name space. Hostnames appear as a component in Uniform Resource Locators (URLs) for Internet resources such as web sites.

Domain names are also used as simple identification labels to indicate ownership or control of a resource. Such examples are the realm identifiers used in the Session Initiation Protocol (SIP), the DomainKeys used to verify DNS domains in e-mail systems, and in many other Uniform Resource Identifiers (URIs).

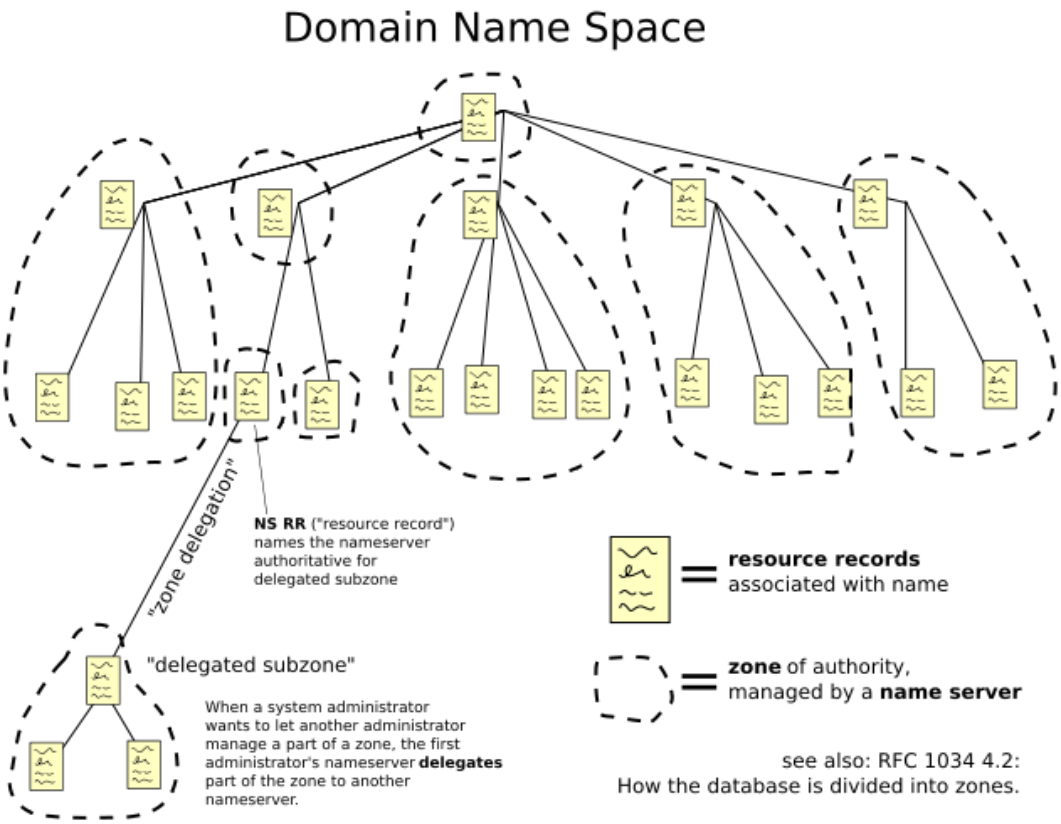
An important purpose of domain names is to provide easily recognizable and memorizable names to numerically addressed Internet resources. This abstraction allows any resource (e.g., website) to be moved to a different physical location in the address topology of the network, globally or locally in an intranet. Such a move usually requires changing the IP address of a resource and the corresponding translation of this IP address to and from its domain name.

Domain names are often referred to simply as *domains* and domain name registrants are frequently referred to as *domain owners*, although domain name registration with a

registrar does not confer any legal ownership of the domain name, only an exclusive right of use.

The Internet Corporation for Assigned Names and Numbers (ICANN) manages the top-level development and architecture of the Internet domain name space. It authorizes domain name registrars, through which domain names may be registered and reassigned. The use of domain names in commerce may subject strings in them to trademark law. In 2010, the number of active domains reached 196 million.

Domain name space



The hierarchical domain name system, organized into zones, each served by a name server

The domain name space consists of a tree of domain names. Each node in the tree holds information associated with the domain name. The tree sub-divides into *zones* beginning at the root zone.

Domain name syntax

A domain name consists of one or more parts, technically called *labels*, that are conventionally concatenated, and delimited by dots, such as *example.com*.

- The right-most label conveys the top-level domain; for example, the domain name *www.example.com* belongs to the top-level domain *com*.
- The hierarchy of domains descends from the right to the left label in the name; each label to the left specifies a subdivision, or subdomain of the domain to the right. For example: the label *example* specifies a subdomain of the *com* domain, and *www* is a subdomain of *example.com*. This tree of labels may consist of 127 levels. Each label may contain up to 63 ASCII characters. The full domain name may not exceed a total length of 253 characters. In practice, some domain registries may have shorter limits.
- A hostname is a domain name that has at least one associated IP address. For example, the domain names *www.example.com* and *example.com* are also hostnames, whereas the *com* domain is not. However, other top-level domains, particularly country code top-level domains, may indeed have an IP address, and if so, they are also hostnames.

Top-level domain

A **top-level domain** (TLD) is one of the domains at the highest level in the hierarchical Domain Name System of the Internet. The top-level domain names are installed in the root zone of the name space. For all domains in lower levels, it is the last part of the domain name, that is, the last label of a fully qualified domain name. For example, in the domain name www.example.com, the top-level domain is com, or COM, as domain names are not case-sensitive. Management of most top-level domains is delegated to responsible organizations by the Internet Corporation for Assigned Names and Numbers (ICANN), which operates the Internet Assigned Numbers Authority (IANA) and is in charge of maintaining the DNS root zone.

Originally, the top-level domain space was organized into three main groups, *Countries*, *Categories*, and *Multiorganizations*. An additional *temporary* group consisted only of the initial DNS domain, arpa, intended for transitional purposes toward the stabilization of the domain name system.

Countries are designated in the Domain Name System by their two-letter ISO country code; there are exceptions, however (e.g., .uk). This group of domains is therefore commonly known as country-code top-level domains (ccTLD). Since 2009, countries with non-Latin based alphabets or scripting systems may apply for internationalized country code top-level domain names, which are displayed in end-user applications in their language-native script or alphabet, but use a Punycode-translated ASCII domain name in the Domain Name System.

The *Categories* group has become known as the generic top-level domains. Initially this group consisted of `GOV`, `EDU`, `COM`, `MIL`, `ORG`, and `NET`.

In the growth of the Internet, it became desirable to create additional generic top-level domains. Some of the initial domains' purposes were also generalized, modified, or assigned for maintenance to special organizations affiliated with the intended purpose.

As a result, IANA today distinguishes the following groups of top-level domains:

- country-code top-level domains (ccTLD): Two letter domains established for countries or territories. With some historical exceptions, the code for any territory is the same as its two-letter ISO 3166 code.
- internationalized country code top-level domains (IDN ccTLD).
- generic top-level domains (gTLD): Top-level domains with three or more characters
    - unsponsored top-level domains: domains that operate directly under policies established by ICANN processes for the global Internet community.
    - sponsored top-level domains (sTLD): These domains are proposed and sponsored by private agencies or organizations that establish and enforce rules restricting the eligibility to use the TLD. Use is based on community theme concepts.
- infrastructure top-level domain: This group consists of one domain, the Address and Routing Parameter Area (ARPA). It is managed by IANA on behalf of the Internet Engineering Task Force for various purposes specified in the Request for Comments publications.

In addition, a group of internationalized domain name (IDN) top-level domains has been installed under `test` for testing purposes in the IDN development process.

Internationalized country code TLDs

An internationalized country code top-level domain (IDN ccTLD) is a top-level domain with a specially encoded domain name that is displayed in an end user application, such as a web browser, in its language-native script or alphabet, such as the Arabic alphabet, or a non-alphabetic writing system, such as Chinese characters. IDN ccTLDs are an application of the internationalized domain name (IDN) system to top-level Internet domains assigned to countries, or independent geographic regions.

ICANN started to accept applications for IDN ccTLDs in November 2009, and installed the first set into the Domain Names System in May 2010. The first set was a group of Arabic names for the countries of Egypt, Saudi Arabia, and the United Arab Emirates. By May 2010, 21 countries had submitted applications to ICANN, representing 11 languages.

IDN test domains

In the process of testing internationalized top-level domains, ICANN implemented a set of IDN top-level domains that are a transliteration of the name *example.test* into each language's script.

DNS name	Link	Script	Language
xn--mgbh0fb.xn--kgbechtv	http://مثال.إختبار	Arabic	Arabic
xn--fsqu00a.xn--0zwm56d	http://例子.测试	Simplified Chinese	Chinese
xn--fsqu00a.xn--g6w251d	http://例子.測試	Traditional Chinese	Chinese
xn--hxajbheg2az3al.xn--jxalpdlp	http://παράδειγμα.δοκιμή	Greek	Greek
xn--r8jz45g.xn--zckzah	http://例え.テスト	Kanji, Hiragana, Katakana	Japanese
xn--9n2bp8q.xn--9t4b11yi5a	http://실례.테스트	Hangul	Korean
xn--mgbh0fb.xn--hgbk6aj7f53bba	http://مثال.آزمایشی	Perso-Arabic	Persian
xn--e1afmkfd.xn--80akhbyknj4f	http://пример.испытание	Cyrillic	Russian
xn--zkc6cc5bi7f6e.xn--hlcj6aya9esc7a	http://בדיקה.טסט	Hebrew	Yiddish

Infrastructure domain

The domain `arpa` was the first Internet top-level domain. It was intended to be used only temporarily, aiding in the transition of traditional ARPANET host names to the domain name system. However, after it had been used for reverse DNS lookup, it was found impractical to retire it, and is used today exclusively for Internet infrastructure purposes such as `in-addr.arpa` for IPv4 and `ip6.arpa` for IPv6 reverse DNS resolution, `uri.arpa` and `urn.arpa` for the Dynamic Delegation Discovery System, and `e164.arpa` for telephone number mapping based on NAPTR DNS records. For historical reasons, `arpa` is sometimes considered to be a generic top-level domain.

Reserved domains

RFC 2606 reserves the following four top-level domain names to avoid confusion and conflict. They may be used for various specific purposes however, with the intention that these should not occur in production networks within the global domain name system:

- `example`: reserved for use in examples
- `invalid`: reserved for use in obviously invalid domain names

- `localhost`: reserved to avoid conflict with the traditional use of `localhost` as a hostname
- `test`: reserved for use in tests

Historical domains

In the late 1980s InterNIC created the `nato` domain for use by NATO. NATO considered none of the then existing TLDs as adequately reflecting their status as an international organization. Soon after this addition, however, InterNIC also created the `int` TLD for the use by international organizations in general, and persuaded NATO to use the second level domain `nato.int` instead. The `nato` TLD, no longer used, was finally removed in July 1996.

Other historical TLDs are `cs` for Czechoslovakia (now `cz` for Czech Republic and `sk` for Slovak Republic), `dd` for East Germany (using `de` after reunification of Germany), `yu` for SFR Yugoslavia (now: `ba` for Bosnia and Herzegovina, `hr` for Croatia, `me` for Montenegro, `mk` for Macedonia, `rs` for Serbia and `si` for Slovenia), and `zr` for Zaire (now `cd` for Democratic Republic of the Congo). In contrast to these, the TLD `su` has remained active despite the demise of the Soviet Union that it represents.

Proposed domains

Around late 2000 when ICANN discussed and finally introduced `aero`, `biz`, `coop`, `info`, `museum`, `name`, and `pro` TLDs, site owners argued that a similar TLD should be made available for adult and pornographic websites to settle the dispute of obscene content on the Internet and the responsibility of US service providers under the US Communications Decency Act of 1996. Several options were proposed including `xxx`, `sex` and `adult`. As of June 2010, the .xxx TLD has received initial approval from the ICANN, based upon a proposal by the sponsoring agency for this TLD, a Florida-based company called ICM Registry.

An older proposal consisted of seven new gTLDs: `arts`, `firm`, `info`, `nom`, `rec`, `shop`, and `web`. Later `biz`, `info`, `museum`, and `name` covered most of these old proposals.

During the 32nd International Public ICANN Meeting in Paris in 2008, ICANN started a new process of TLD naming policy to take a *"significant step forward on the introduction of new generic top-level domains."* This program envisions the availability of many new or already proposed domains, as well a new application and implementation process. Observers believed that the new rules could result in hundreds of new gTLDs to be registered. Proposed TLDs include music, shop, berlin and nyc.

Alternative DNS roots

ICANN's slow progress in creating new generic top-level domains, and the high application costs associated with TLDs, contributed to the creation of alternate DNS roots with different sets of top-level domains. Such domains may be accessed by configuration of a computer with alternate or additional (forwarder) DNS servers or plugin modules for web browsers. Browser plugins detect alternate root domain requests and access an alternate domain name server for such requests.

Pseudo-domains

Several networks, such as BITNET, CSNET, UUCP or other networks, existed that were in widespread use among computer professionals and academic users, that were incompatible with the Internet and exchanged e-mail with the Internet via special e-mail gateways. For relaying purposes on the gateways, messages associated with these networks were labeled with suffixes such as `bitnet`, `oz`, `csnet`, or `uucp`, but these domains did not exist as top-level domains in the public Domain Name System of the Internet.

Most of these networks have long since ceased to exist, and although UUCP still gets significant use in parts of the world where Internet infrastructure has not yet become well-established, it subsequently transitioned to using Internet domain names, so pseudo-domains now largely survive as historical relics. One notable exception is the 2007 emergence of SWIFTNet Mail, which uses the `swift` pseudo-domain.

The top-level pseudo domain `local` is required by the Zeroconf protocol. It is also used by many organizations internally, which may become a problem for those users as Zeroconf becomes more popular. Both `site` and `internal` have been suggested for private usage, but no consensus has emerged.

The anonymity network Tor has a top-level pseudo-domain `onion`, which can only be reached with a Tor client because it uses the Tor-protocol (onion routing) to reach the hidden service to protect the anonymity of users.

Second-level and lower level domains

Below the top-level domains in the domain name hierarchy are the second-level domain (SLD) names. These are the names directly to the left of .com, .net, and the other top-level domains.

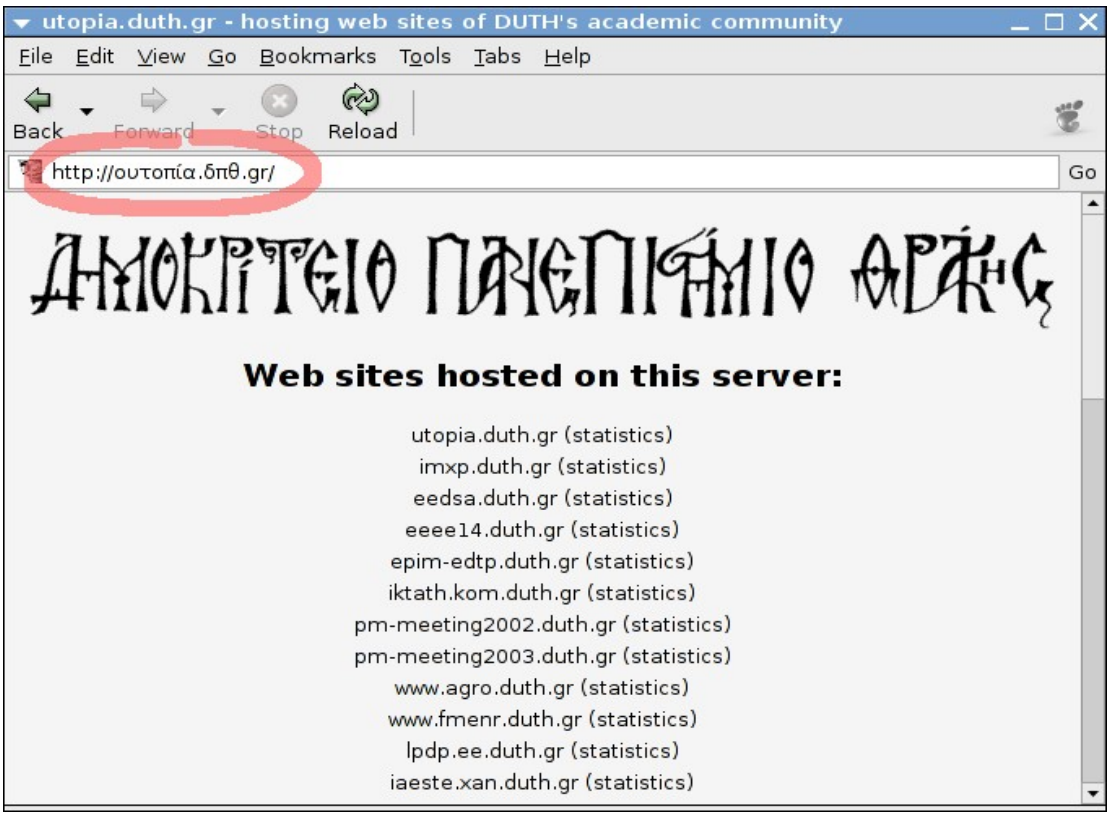
Next are third-level domains, which are written immediately to the left of a second-level domain. There can be fourth- and fifth-level domains, and so on, with virtually no limitation. An example of an operational domain name with four levels of domain labels is *www.sos.state.oh.us*. The www preceding the domains is the host name of the World-Wide Web server. Each label is separated by a full stop (dot). 'sos' is said to be a sub-

domain of 'state.oh.us', and 'state' a sub-domain of 'oh.us', etc. In general, subdomains are domains subordinate to their parent domain. An example of very deep levels of subdomain ordering are the IPv6 reverse resolution DNS zones, e.g., 1.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.ip6.arpa, which is the reverse DNS resolution domain name for the IP address of a loopback interface, or the localhost name.

Second-level (or lower-level, depending on the established parent hierarchy) domain names are often created based on the name of a company (e.g., *bbc*.co.uk), product or service. Below these levels, the next domain name component has been used to designate a particular host server. Modern technology allows multiple physical servers with either different (cf. load balancing) or even identical addresses (cf. anycast) to serve a single hostname or domain name, or multiple domain names to be served by a single computer. The latter is very popular in Web hosting service centers, where service providers host the websites of many organizations on just a few servers.

The hierarchical DNS labels or components of domain names are separated in a fully qualified name by the full stop (dot,.).

Internationalized domain name



Example of Greek IDN with country code top-level domain in non-Latin alphabet

An **internationalized domain name** (**IDN**) is an Internet domain name that contains at least one label that is displayed in software applications, in whole or in part, in a language-specific script or alphabet, such as Arabic, Chinese, Russian or the Latin alphabet-based characters with diacritics, such as French. These writing systems are encoded by computers in multi-byte Unicode. Internationalized domain names are stored in the Domain Name System as ASCII strings using Punycode transcription.

The Domain Name System, which performs a lookup service to translate user-friendly names into network addresses for locating Internet resources, is restricted in practice to the use of ASCII characters, a practical limitation that initially set the standard for acceptable domain names. The internationalization of domain names is a technical solution to translate names written in language-native scripts into an ASCII text representation that is compatible with the Domain Name System. Internationalized domain names can only be used with applications that are specifically designed for such use, and they require no changes in the infrastructure of the Internet.

IDN was originally proposed in December 1996 by Martin Dürst and implemented in 1998 by Tan Juay Kwang and Leong Kok Yong under the guidance of T.W. Tan. After much debate and many competing proposals, a system called *Internationalizing Domain Names in Applications* (IDNA) was adopted as a standard, and has been implemented in several top-level domains.

In IDNA, the term *internationalized domain name* means specifically any domain name consisting only of labels to which the IDNA ToASCII algorithm (see below) can be successfully applied. In March 2008, the IETF formed a new IDN working group to update the current IDNA protocol.

In October 2009, the Internet Corporation for Assigned Names and Numbers (ICANN) approved the creation of internationalized country code top-level domains (IDN ccTLDs) in the Internet that use the IDNA standard for native language scripts. In May 2010 the first IDN ccTLD were installed in the DNS root zone.

Internationalizing Domain Names in Applications

Internationalizing Domain Names in Applications (IDNA) is a mechanism defined in 2003 for handling internationalized domain names containing non-ASCII characters. These names either are Latin letters with diacritics (ñ, é) or are written in languages or scripts which do not use the Latin alphabet: Arabic, Hangul, Hiragana and Kanji for instance. Although the Domain Name System supports non-ASCII characters, applications such as e-mail and web browsers restrict the characters which can be used as domain names for purposes such as a hostname. Strictly speaking it is the network protocols these applications use that have restrictions on the characters which can be used in domain names, not the applications that have these limitations or the DNS itself. To retain backwards compatibility with the installed base the IETF IDNA Working Group decided that internationalized domain names should be converted to a suitable ASCII-based form that could be handled by web browsers and other user applications. IDNA

specifies how this conversion between names written in non-ASCII characters and their ASCII-based representation is performed.

An IDNA-enabled application is able to convert between the internationalized and ASCII representations of a domain name. It uses the ASCII form for DNS lookups but can present the internationalized form to users who presumably prefer to read and write domain names in non-ASCII scripts such as Arabic or Hiragana. Applications that do not support IDNA will not be able to handle domain names with non-ASCII characters, but will still be able to access such domains if given the (usually rather cryptic) ASCII equivalent.

ICANN issued guidelines for the use of IDNA in June 2003, and it was already possible to register .jp domains using this system in July 2003 and .info domains in March 2004. Several other top-level domain registries started accepting registrations in 2004 and 2005. IDN Guidelines were first created in June 2003, and have been updated to respond to phishing concerns in November 2005. An ICANN working group focused on country code domain names at the top level was formed in November 2007 and promoted jointly by the country code supporting organization and the Governmental Advisory Committee.

Mozilla 1.4, Netscape 7.1, Opera 7.11 were among the first applications to support IDNA. A browser plugin is available for Internet Explorer 6 to provide IDN support. Internet Explorer 7.0 and Windows Vista's URL APIs provide native support for IDN.

ToASCII and ToUnicode

The conversions between ASCII and non-ASCII forms of a domain name are accomplished by algorithms called ToASCII and ToUnicode. These algorithms are not applied to the domain name as a whole, but rather to individual labels. For example, if the domain name is www.example.com, then the labels are *www*, *example*, and *com*. ToASCII or ToUnicode are applied to each of these three separately.

The details of these two algorithms are complex, and are specified in RFC 3490. The following gives an overview of their function.

ToASCII leaves unchanged any ASCII label, but will fail if the label is unsuitable for the Domain Name System. If given a label containing at least one non-ASCII character, ToASCII will apply the Nameprep algorithm, which converts the label to lowercase and performs other normalization, and will then translate the result to ASCII using Punycode before prepending the four-character string "xn--". This four-character string is called the ASCII Compatible Encoding (*ACE*) prefix, and is used to distinguish Punycode encoded labels from ordinary ASCII labels. The ToASCII algorithm can fail in several ways; for example, the final string could exceed the 63-character limit of a DNS name. A label for which ToASCII fails cannot be used in an internationalized domain name.

The function ToUnicode reverses the action of ToASCII, stripping off the ACE prefix and applying the Punycode decode algorithm. It does not reverse the Nameprep

processing, since that is merely a normalization and is by nature irreversible. Unlike ToASCII, ToUnicode always succeeds, because it simply returns the original string if decoding fails. In particular, this means that ToUnicode has no effect on a string that does not begin with the ACE prefix.

Example of IDNA encoding

IDNA encoding may be illustrated using the example domain *Bücher.ch*. "Bücher" is German for "books", and .ch is the ccTLD of Switzerland. This domain name has two labels, *Bücher* and *ch*. The second label is pure ASCII, and is left unchanged. The first label is processed by Nameprep to give *bücher*, and then converted to Punycode to result in *bcher-kva*. It is then prepended with *xn--* to produce *xn--bcher-kva*. The resulting label suitable for use in the DNS is therefore *xn--bcher-kva.ch*.

Top-level domain implementation

In 2009, ICANN decided to implement a new class of top-level domains, assignable to countries and independent regions, similar to the rules for country code top-level domains. However, the domain names may be any desirable string of characters, symbols, or glyphs in the language-specific, non-Latin alphabet or script of the applicant's language, within certain guidelines to assure sufficient visual uniqueness.

The process of installing IDN country code domains began with a long period of testing in a set of subdomains in the `test` top-level domain. Eleven domains used language-native scripts or alphabets, such as δοκιμή, meaning *test* in Greek.

These efforts culminated in the creation of the first internationalized country code top-level domains (IDN ccTLDs) for production use in 2010.

In the Domain Name System, these domains use an ASCII representation consisting of the prefix `xn--` followed by the Punycode translation of the Unicode representation of the language-specific alphabet or script glyphs. For example, the Cyrillic name of Russia's IDN ccTLD is рф. In Punycode representation, this is `plai`, and its DNS name is `xn--plai`.

Non-IDNA or non-ICANN registries that support non-ASCII domain names

There are other registries that support non-ASCII domain names. The company ThaiURL.com in Thailand supports .com registrations via its own modified domain name system, ThaiURL. Because these companies, and other organizations that offer modified DNS systems, do not subject themselves to ICANN's control, they must be regarded as alternate DNS roots. Domains registered with them will therefore not be supported by most Internet service providers, and as a result most users will not be able to look up such domains without manually configuring their computers to use the alternate DNS.

ASCII spoofing concerns

The use of Unicode in domain names makes it potentially easier to spoof web sites visited by World Wide Web users as the visual representation of an IDN string in a web browser may appear identical to another, depending on the font used. For example, Unicode character U+0430, Cyrillic small letter a, can look identical to Unicode character U+0061, Latin small letter a, used in English.

Top-level domains accepting IDN registration

Many top-level domains have started to accept domain name registrations at the second or lower levels.

History

On 15 March 1985, the first commercial Internet domain name (.com) was registered under the name *Symbolics.com* by Symbolics Inc., a computer systems firm in Cambridge, Massachusetts.

By 1992 fewer than 15,000 dot.com domains were registered.

In December 2009 there were 192 million domain names. A big fraction of them are in the .com TLD, which as of March 15, 2010 had 84 million domain names, including 11.9 million online business and e-commerce sites, 4.3 million entertainment sites, 3.1 million finance related sites, and 1.8 million sports sites.

Domain name registration

The right to use a domain name is delegated by domain name registrars which are accredited by the Internet Corporation for Assigned Names and Numbers (ICANN), the organization charged with overseeing the name and number systems of the Internet. In addition to ICANN, each top-level domain (TLD) is maintained and serviced technically by an administrative organization operating a registry. A registry is responsible for maintaining the database of names registered within the TLD it administers. The registry receives registration information from each domain name registrar authorized to assign names in the corresponding TLD and publishes the information using a special service, the whois protocol.

Registries and registrars usually charge an annual fee for the service of delegating a domain name to a user and providing a default set of name servers. Often this transaction is termed a sale or lease of the domain name, and the registrant may sometimes be called an "owner", but no such legal relationship is actually associated with the transaction, only the exclusive right to use the domain name. More correctly, authorized users are known as "registrants" or as "domain holders".

ICANN publishes the complete list of TLD registries and domain name registrars. Registrant information associated with domain names is maintained in an online database accessible with the WHOIS service. For most of the 250 country code top-level domains (ccTLDs), the domain registries maintain the WHOIS (Registrant, name servers, expiration dates, etc.) information.

Some domain name registries, often called *network information centers* (NIC), also function as registrars to end-users. The major generic top-level domain registries, such as for the `COM`, `NET`, `ORG`, `INFO` domains and others, use a registry-registrar model consisting of hundreds of domain name registrars. In this method of management, the registry only manages the domain name database and the relationship with the registrars. The *registrants* (users of a domain name) are customers of the registrar, in some cases through additional layers of resellers.

In the process of registering a domain name and maintaining authority over the new name space created, registrars use several key pieces of information connected with a domain:

- *Administrative contact*. A registrant usually designates an administrative contact to manage the domain name. The administrative contact usually has the highest level of control over a domain. Management functions delegated to the administrative contacts may include management of all business information, such as name of record, postal address, and contact information of the official registrant of the domain and the obligation to conform to the requirements of the domain registry in order to retain the right to use a domain name. Furthermore the administrative contact installs additional contact information for technical and billing functions.
- *Technical contact*. The technical contact manages the name servers of a domain name. The functions of a technical contact include assuring conformance of the configurations of the domain name with the requirements of the domain registry, maintaining the domain zone records, and providing continuous functionality of the name servers (that leads to the accessibility of the domain name).
- *Billing contact*. The party responsible for receiving billing invoices from the domain name registrar and paying applicable fees.
- *Name servers*. Most registrars provide two or more name servers as part of the registration service. However, a registrant may specify its own authoritative name servers to host a domain's resource records. The registrar's policies govern the number of servers and the type of server information required. Some providers require a hostname and the corresponding IP address or just the hostname, which must be resolvable either in the new domain, or exist elsewhere. Based on traditional requirements (RFC 1034), typically a minimum of two servers is required.

Domain names are often seen in analogy to real estate in that (1) domain names are foundations on which a website (like a house or commercial building) can be built and (2) the highest "quality" domain names, like sought-after real estate, tend to carry

significant value, usually due to their online brand-building potential, use in advertising, search engine optimization, and many other criteria.

A few companies have offered low-cost, below-cost or even cost-free domain registrations with a variety of models adopted to recoup the costs to the provider. These usually require that domains be hosted on their website within a framework or portal that includes advertising wrapped around the domain holder's content, revenue from which allows the provider to recoup the costs. Domain registrations were free of charge when the DNS was new. A domain holder can give away or sell infinite number of subdomains under their domain name. For example, the owner of *example.org* could provide subdomains such as *foo.example.org* and *foo.bar.example.org* to interested parties.

Because of the popularity of the Internet, many desirable domain names are already assigned and users must search for other acceptable names, using Web-based search features, or WHOIS and dig operating system tools. Many registrars have implemented **Domain name suggestion** tools which search domain name databases and suggest available alternative domain names related to keywords provided by the user.

Resale of domain names

The business of resale of registered domain names is known as the domain aftermarket. Various factors influence the perceived value or market value of a domain name.

As of 2004, according to Guinness World Records and MSNBC, the most expensive domain name sales on record were:

- Business.com resold for $350 million in July 2007
- Business.com for $7.5 million in December 1999
- AsSeenOnTv.com for $5.1 million in January 2000
- Altavista.com for $3.3 million in August 1998
- Wine.com for $2.9 million in September 1999
- CreditCards.com for $2.75 million in July 2004
- Autos.com for $2.2 million in December 1999
- Sex.com for $13 million
- Toys.com: Toys 'R' Us by auction for $5.1 million in 2009

Domain name confusion

Intercapping is often used to emphasize the meaning of a domain name. However, DNS names are case-insensitive, and some names may be misinterpreted in certain uses of capitalization, creating slurls. For example: *Who Represents*, a database of artists and agents, chose `whorepresents.com`, which can be misread as *whore presents*. Similarly, a therapists' network is named `therapistfinder.com`. In such situations, the proper meaning may be clarified by use of hyphens in the domain name. For instance, Experts

Exchange, a programmers' discussion site, for a long time used `expertsexchange.com`, but ultimately changed the name to `experts-exchange.com`.

Intellectual property entrepreneur Leo Stoller threatened to sue the owners of StealThisEmail.com on the basis that, when read as `stealthisemail.com`, it infringed on claimed (but invalid) trademark rights to the word "stealth".

Use in web site hosting

A domain name is a component of a Uniform Resource Locator (URL) used to access web sites, for example:

> URL: `http://www.example.net/index.html`
> Top-level domain name: `.net`
> Second-level domain name: `example.net`
> Host name: `www.example.net`

A domain name may point to multiple IP addresses to provide server redundancy for the services delivered. This is used for large, popular web sites. More commonly, however, one server at a given IP address may also host multiple web sites in different domains. Such address overloading enables virtual web hosting commonly used by large web hosting services to conserve IP address space. It is possible through a feature in the HTTP version 1.1 protocol, but not in HTTP 1.0, which requires that a request identifies the domain name being referenced.

Abuse and regulation

Critics often claim abuse of administrative power over domain names. Particularly noteworthy was the VeriSign Site Finder system which redirected all unregistered .com and .net domains to a VeriSign webpage. For example, at a public meeting with VeriSign to air technical concerns about SiteFinder, numerous people, active in the IETF and other technical bodies, explained how they were surprised by VeriSign's changing the fundamental behavior of a major component of Internet infrastructure, not having obtained the customary consensus. SiteFinder, at first, assumed every Internet query was for a website, and it monetized queries for incorrect domain names, taking the user to VeriSign's search site. Unfortunately, other applications, such as many implementations of email, treat a lack of response to a domain name query as an indication that the domain does not exist, and that the message can be treated as undeliverable. The original VeriSign implementation broke this assumption for mail, because it would always resolve an erroneous domain name to that of SiteFinder. While VeriSign later changed SiteFinder's behaviour with regard to email, there was still widespread protest about VeriSign's action being more in its financial interest than in the interest of the Internet infrastructure component for which VeriSign was the steward.

Despite widespread criticism, VeriSign only reluctantly removed it after the Internet Corporation for Assigned Names and Numbers (ICANN) threatened to revoke its contract to administer the root name servers. ICANN published the extensive set of letters exchanged, committee reports, and ICANN decisions.

There is also significant disquiet regarding the United States' political influence over ICANN. This was a significant issue in the attempt to create a .xxx top-level domain and sparked greater interest in alternative DNS roots that would be beyond the control of any single country.

Additionally, there are numerous accusations of domain name front running, whereby registrars, when given whois queries, automatically register the domain name for themselves. Recently, Network Solutions has been accused of this.

Truth in Domain Names Act

In the United States, the Truth in Domain Names Act of 2003, in combination with the PROTECT Act of 2003, forbids the use of a misleading domain name with the intention of attracting Internet users into visiting Internet pornography sites.

The Truth in Domain Names Act follows the more general Anticybersquatting Consumer Protection Act passed in 1999 aimed at preventing typosquatting and deceptive use of names and trademarks in domain names.

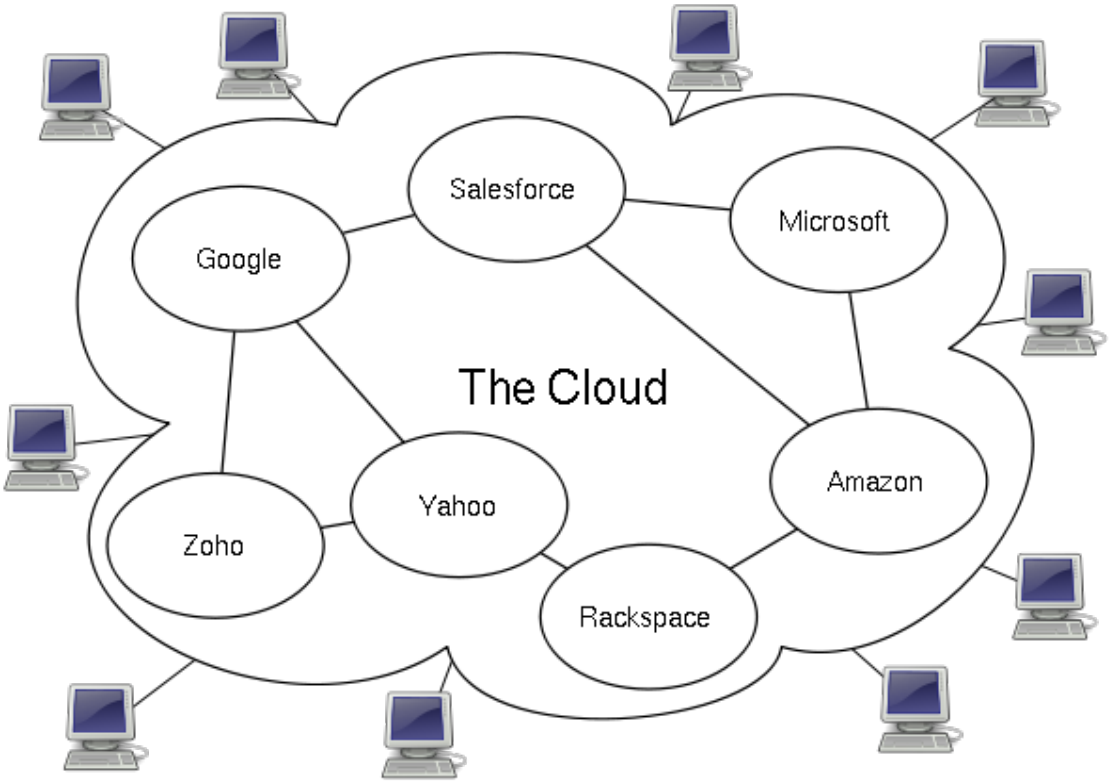
Fictitious domain name

A *fictitious domain name* is a domain name used in a work of fiction or popular culture to refer to a domain that does not actually exist.

Domain names used in works of fiction have often been registered in the DNS, either by their creators or by cybersquatters attempting to profit from it. This phenomenon prompted NBC to purchase the domain name Hornymanatee.com after talk-show host Conan O'Brien spoke the name while ad-libbing on his show. O'Brien subsequently created a website based on the concept and used it as a running gag on the show.

Chapter 10

Cloud Computing



Cloud computing conceptual diagram

Cloud computing is Internet-based computing, whereby shared servers provide resources, software, and data to computers and other devices on demand, as with the electricity grid. Cloud computing is a natural evolution of the widespread adoption of virtualization, service-oriented architecture and utility computing. Details are abstracted from consumers, who no longer have need for expertise in, or control over, the technology infrastructure "in the cloud" that supports them.

Cloud computing describes a new supplement, consumption, and delivery model for IT services based on the Internet, and it typically involves over-the-Internet provision of dynamically scalable and often virtualized resources. It is a byproduct and consequence

of the ease-of-access to remote computing sites provided by the Internet. This frequently takes the form of web-based tools or applications that users can access and use through a web browser as if it were a program installed locally on their own computer.

The National Institute of Standards and Technology (NIST) provides a somewhat more objective and specific definition here. The term "cloud" is used as a metaphor for the Internet, based on the cloud drawing used in the past to represent the telephone network, and later to depict the Internet in computer network diagrams as an abstraction of the underlying infrastructure it represents. Typical cloud computing providers deliver common business applications online that are accessed from another Web service or software like a Web browser, while the software and data are stored on servers.

Most cloud computing infrastructures consist of services delivered through common centers and built on servers. Clouds often appear as single points of access for consumers' computing needs. Commercial offerings are generally expected to meet quality of service (QoS) requirements of customers, and typically include service level agreements (SLAs). The major cloud service providers include Amazon, Rackspace Cloud, Salesforce, Skytap Cloud, Microsoft and Google. Some of the larger IT firms that are actively involved in cloud computing are Fujitsu, Dell, Red Hat, Hewlett Packard, IBM, VMware and NetApp.

Overview

Comparisons

Cloud computing derives characteristics from, but should not be confused with:

1. Autonomic computing — "computer systems capable of self-management"
2. Client–server model – *client–server computing* refers broadly to any distributed application that distinguishes between service providers (servers) and service requesters (clients)
3. Grid computing — "a form of distributed computing and parallel computing, whereby a 'super and virtual computer' is composed of a cluster of networked, loosely coupled computers acting in concert to perform very large tasks"
4. Mainframe computer — powerful computers used mainly by large organizations for critical applications, typically bulk data-processing such as census, industry and consumer statistics, enterprise resource planning, and financial transaction processing.
5. Utility computing — the "packaging of computing resources, such as computation and storage, as a metered service similar to a traditional public utility, such as electricity";
6. Peer-to-peer – distributed architecture without the need for central coordination, with participants being at the same time both suppliers and consumers of resources (in contrast to the traditional client–server model)

7. Service-oriented computing – Cloud computing provides services related to computing while, in a reciprocal manner, service-oriented computing consists of the computing techniques that operate on software-as-a-service.

Characteristics

The fundamental concept of cloud computing is that the computing is "in the cloud" i.e. that the processing (and the related data) is not in a specified, known or the same place(s). This is in opposition to where the processing takes place in one or more specific servers that are known. All the other concepts mentioned are supplementary or complementary to this concept.

Generally, cloud computing customers do not own the physical infrastructure, instead avoiding capital expenditure by renting usage from a third-party provider. They consume resources as a service and pay only for resources that they use. Many cloud-computing offerings employ the utility computing model, which is analogous to how traditional utility services (such as electricity) are consumed, whereas others bill on a subscription basis. Sharing "perishable and intangible" computing power among multiple tenants can improve utilization rates, as servers are not unnecessarily left idle, which can reduce costs significantly while increasing the speed of application development. A side-effect of this approach is that overall computer usage rises dramatically, as customers do not have to engineer for peak load limits. In addition, "increased high-speed bandwidth" makes it possible to receive the same. The cloud is becoming increasingly associated with small and medium enterprises (SMEs) as in many cases they cannot justify or afford the large capital expenditure of traditional IT. SMEs also typically have less existing infrastructure, less bureaucracy, more flexibility, and smaller capital budgets for purchasing in-house technology. Similarly, SMEs in emerging markets are typically unburdened by established legacy infrastructures, thus reducing the complexity of deploying cloud solutions.

Economics

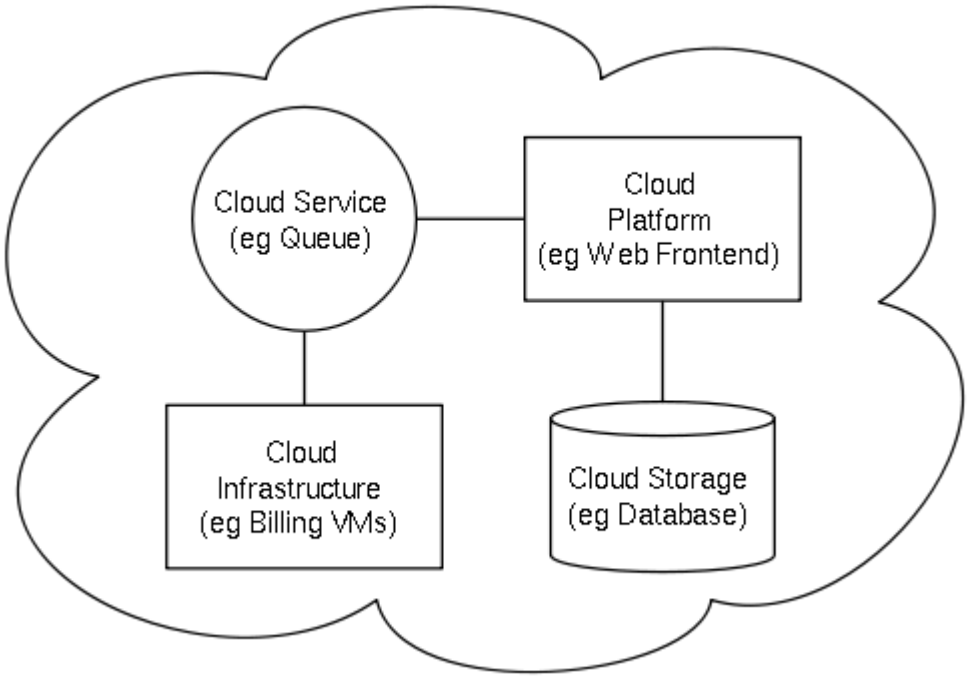
Cloud computing users avoid capital expenditure (CapEx) on hardware, software, and services when they pay a provider only for what they use. Consumption is usually billed on a utility (resources consumed, like electricity) or subscription (time-based, like a newspaper) basis with little or no upfront cost. Other benefits of this approach are low barriers to entry, shared infrastructure and costs, low management overhead, and immediate access to a broad range of applications. In general, users can terminate the contract at any time (thereby avoiding return on investment risk and uncertainty), and the services are often covered by service level agreements (SLAs) with financial penalties.

According to Nicholas Carr, the strategic importance of information technology is diminishing as it becomes standardized and less expensive. He argues that the cloud computing paradigm shift is similar to the displacement of frozen water trade by electricity generators early in the 20th century.

Although companies might be able to save on upfront capital expenditures, they might not save much and might actually pay more for operating expenses. In situations where the capital expense would be relatively small, or where the organization has more flexibility in their capital budget than their operating budget, the cloud model might not make great fiscal sense. Other factors having an impact on the scale of potential cost savings include the efficiency of a company's data center as compared to the cloud vendor's, the company's existing operating costs, the level of adoption of cloud computing, and the type of functionality being hosted in the cloud.

Among the items that some cloud hosts charge for are instances (often with extra charges for high-memory or high-CPU instances), data transfer in and out, storage (measured by the GB-month), I/O requests, PUT requests and GET requests, IP addresses, and load balancing. In some cases, users can bid on instances, with pricing dependent on demand for available instances.

Architecture



Cloud computing sample architecture

Cloud architecture, the systems architecture of the software systems involved in the delivery of cloud computing, typically involves multiple *cloud components* communicating with each other over application programming interfaces, usually web services. This resembles the Unix philosophy of having multiple programs each doing one thing well and working together over universal interfaces. Complexity is controlled and the resulting systems are more manageable than their monolithic counterparts.

The two most significant components of cloud computing architecture are known as the front end and the back end. The front end is the part seen by the client, i.e. the computer user. This includes the client's network (or computer) and the applications used to access the cloud via a user interface such as a web browser. The back end of the cloud computing architecture is the 'cloud' itself, comprising various computers, servers and data storage devices.

History

The underlying concept of cloud computing dates back to the 1960s, when John McCarthy opined that "computation may someday be organized as a public utility." Almost all the modern-day characteristics of cloud computing (elastic provision, provided as a utility, online, illusion of infinite supply), the comparison to the electricity industry and the use of public, private, government and community forms was thoroughly explored in Douglas Parkhill's 1966 book, *The Challenge of the Computer Utility*.

The actual term "cloud" borrows from telephony in that telecommunications companies, who until the 1990s primarily offered dedicated point-to-point data circuits, began offering Virtual Private Network (VPN) services with comparable quality of service but at a much lower cost. By switching traffic to balance utilization as they saw fit, they were able to utilize their overall network bandwidth more effectively. The cloud symbol was used to denote the demarcation point between that which was the responsibility of the provider from that of the user. Cloud computing extends this boundary to cover servers as well as the network infrastructure. The first scholarly use of the term "cloud computing" was in a 1997 lecture by Ramnath Chellappa.

Amazon played a key role in the development of cloud computing by modernizing their data centers after the dot-com bubble, which, like most computer networks, were using as little as 10% of their capacity at any one time, just to leave room for occasional spikes. Having found that the new cloud architecture resulted in significant internal efficiency improvements whereby small, fast-moving "two-pizza teams" could add new features faster and more easily, Amazon initiated a new product development effort to provide cloud computing to external customers, and launched Amazon Web Service (AWS) on a utility computing basis in 2006.

In 2007, Google, IBM and a number of universities embarked on a large scale cloud computing research project. In early 2008, Eucalyptus became the first open source AWS API compatible platform for deploying private clouds. In early 2008, OpenNebula, enhanced in the RESERVOIR European Commission funded project, became the first open source software for deploying private and hybrid clouds and for the federation of clouds . By mid-2008, Gartner saw an opportunity for cloud computing "to shape the relationship among consumers of IT services, those who use IT services and those who sell them" and observed that "[o]rganisations are switching from company-owned hardware and software assets to per-use service-based models" so that the "projected shift to cloud computing ... will result in dramatic growth in IT products in some areas and significant reductions in other areas."

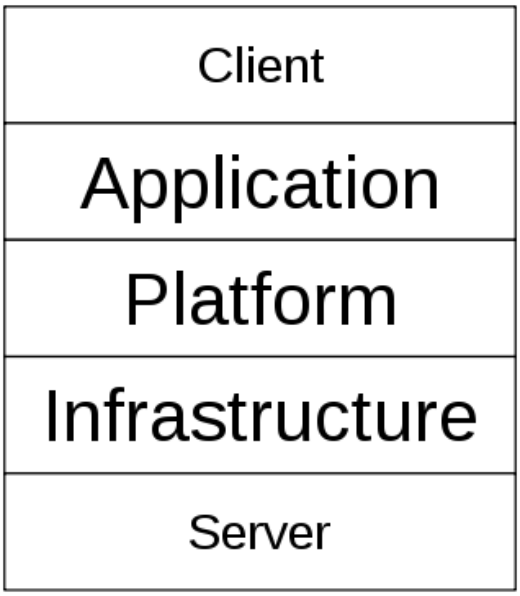
Key features

- **Agility** improves with users' ability to rapidly and inexpensively re-provision technological infrastructure resources.
- **Application Programming Interface** (API) accessibility to software that enables machines to interact with cloud software in the same way the user interface facilitates interaction between humans and computers. Cloud Computing systems typically use REST based APIs.
- **Cost** is claimed to be greatly reduced and capital expenditure is converted to operational expenditure. This ostensibly lowers barriers to entry, as infrastructure is typically provided by a third-party and does not need to be purchased for one-time or infrequent intensive computing tasks. Pricing on a utility computing basis is fine-grained with usage-based options and fewer IT skills are required for implementation (in-house).
- **Device and location independence** enable users to access systems using a web browser regardless of their location or what device they are using (e.g., PC, mobile). As infrastructure is off-site (typically provided by a third-party) and accessed via the Internet, users can connect from anywhere.
- **Multi-tenancy** enables sharing of resources and costs across a large pool of users thus allowing for:
    - **Centralization** of infrastructure in locations with lower costs (such as real estate, electricity, etc.)
    - **Peak-load capacity** increases (users need not engineer for highest possible load-levels)
    - **Utilization and efficiency** improvements for systems that are often only 10–20% utilized.
- **Reliability** is improved if multiple redundant sites are used, which makes well designed cloud computing suitable for business continuity and disaster recovery. Nonetheless, many major cloud computing services have suffered outages, and IT and business managers can at times do little when they are affected.
- **Scalability** via dynamic ("on-demand") provisioning of resources on a fine-grained, self-service basis near real-time, without users having to engineer for peak loads. Performance is monitored, and consistent and loosely coupled architectures are constructed using web services as the system interface. One of the most important new methods for overcoming performance bottlenecks for a large class of applications is data parallel programming on a distributed data grid.
- **Security** could improve due to centralization of data, increased security-focused resources, etc., but concerns can persist about loss of control over certain sensitive data, and the lack of security for stored kernels. Security is often as good as or better than under traditional systems, in part because providers are able to devote resources to solving security issues that many customers cannot afford. Providers typically log accesses, but accessing the audit logs themselves can be difficult or impossible. Furthermore, the complexity of security is greatly increased when data is distributed over a wider area and / or number of devices.

- **Maintenance** of cloud computing applications is easier, since they don't have to be installed on each user's computer. They are easier to support and to improve since the changes reach the clients instantly.
- **Metering** means that cloud computing resources usage should be measurable and should be metered per client and application on a daily, weekly, monthly, and yearly basis.

Layers

The Internet functions through a series of network protocols that form a stack of layers, as shown in the figure (or as described in more detail in the OSI model). Once an Internet connection is established among several computers, it is possible to share services within any one of the following layers.



Client

A *cloud client* consists of computer hardware and/or computer software that relies on cloud computing for application delivery, or that is specifically designed for delivery of cloud services and that, in either case, is essentially useless without it. Examples include some computers, phones and other devices, operating systems and browsers.

Application

Cloud application services or "*Software as a Service (SaaS)*" deliver software as a service over the Internet, eliminating the need to install and run the application on the customer's own computers and simplifying maintenance and support. People tend to use the terms ‘SaaS’ and ‘cloud’ interchangeably, when in fact they are two different things. Key characteristics include:

- Network-based access to, and management of, commercially available (i.e., not custom) software
- Activities that are managed from central locations rather than at each customer's site, enabling customers to access applications remotely via the Web
- Application delivery that typically is closer to a one-to-many model (single instance, multi-tenant architecture) than to a one-to-one model, including architecture, pricing, partnering, and management characteristics
- Centralized feature updating, which obviates the need for downloadable patches and upgrades.

Platform

Cloud platform services or "*Platform as a Service (PaaS)*" deliver a computing platform and/or solution stack as a service, often consuming *cloud infrastructure* and sustaining *cloud applications*. It facilitates deployment of applications without the cost and complexity of buying and managing the underlying hardware and software layers.

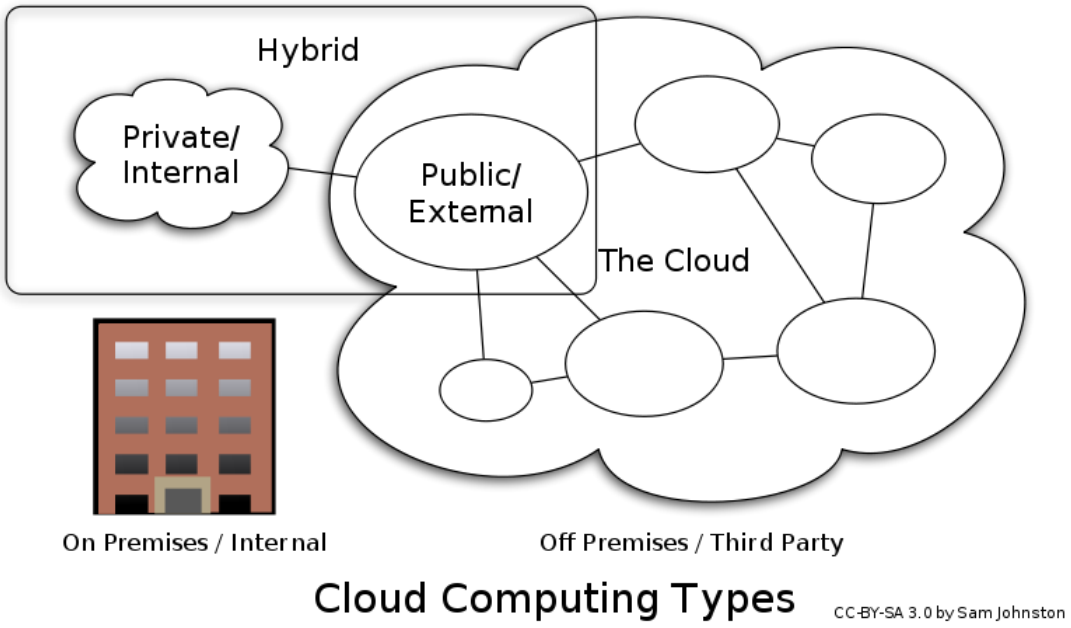
Infrastructure

Cloud infrastructure services, also known as "*Infrastructure as a Service (IaaS)*", delivers computer infrastructure - typically a platform virtualization environment - as a service. Rather than purchasing servers, software, data-center space or network equipment, clients instead buy those resources as a fully outsourced service. Suppliers typically bill such services on a utility computing basis and amount of resources consumed (and therefore the cost) will typically reflect the level of activity. IaaS evolved from virtual private server offerings.

Server

The *servers* layer consists of computer hardware and/or computer software products that are specifically designed for the delivery of cloud services, including multi-core processors, cloud-specific operating systems and combined offerings.

Deployment models



Cloud computing types

Public cloud

Public cloud or *external cloud* describes cloud computing in the traditional main stream sense, whereby resources are dynamically provisioned on a fine-grained, self-service basis over the Internet, via web applications/web services, from an off-site third-party provider who bills on a fine-grained utility computing basis.

Community cloud

A *community cloud* may be established where several organizations have similar requirements and seek to share infrastructure so as to realize some of the benefits of cloud computing. With the costs spread over fewer users than a *public cloud* (but more than a single tenant) this option is more expensive but may offer a higher level of privacy, security and/or policy compliance. Examples of *community cloud* include Google's "Gov Cloud".

Hybrid cloud

There is some confusion over the term "Hybrid" when applied to the cloud - a standard definition of the term "Hybrid Cloud" has not yet emerged. The term "Hybrid Cloud" has been used to mean either two separate clouds joined together (public, private, internal or external), or a combination of virtualized cloud server instances used together with real

physical hardware. The most correct definition of the term "Hybrid Cloud" is probably the use of physical hardware and virtualized cloud server instances together to provide a single common service. Two clouds that have been joined together are more correctly called a "combined cloud".

A *combined cloud* environment consisting of multiple internal and/or external providers "will be typical for most enterprises". By integrating multiple cloud services users may be able to ease the transition to *public cloud* services while avoiding issues such as PCI compliance.

Another perspective on deploying a web application in the cloud is using Hybrid Web Hosting, where the hosting infrastructure is a mix between Cloud Hosting and Managed dedicated servers - this is most commonly achieved as part of a web cluster in which some of the nodes are running on real physical hardware and some are running on cloud server instances.

A hybrid storage cloud uses a combination of public and private storage clouds. Hybrid storage clouds are often useful for archiving and backup functions, allowing local data to be replicated to a public cloud.

Private cloud

Douglas Parkhill first described the concept of a "Private Computer Utility" in his 1966 book *The Challenge of the Computer Utility*. The idea was based upon direct comparison with other industries (e.g. the electricity industry) and the extensive use of hybrid supply models to balance and mitigate risks.

Private cloud and *internal cloud* have been described as neologisms, however the concepts themselves pre-date the term *cloud* by 40 years. Even within modern utility industries, hybrid models still exist despite the formation of reasonably well-functioning markets and the ability to combine multiple providers.

Some vendors have used the terms to describe offerings that emulate cloud computing on private networks. These (typically virtualization automation) products offer the ability to host applications or virtual machines in a company's own set of hosts. These provide the benefits of utility computing -shared hardware costs, the ability to recover from failure, and the ability to scale up or down depending upon demand.

Private clouds have attracted criticism because users "still have to buy, build, and manage them" and thus do not benefit from lower up-front capital costs and less hands-on management, essentially "[lacking] the economic model that makes cloud computing such an intriguing concept".

Cloud Engineering

Cloud Engineering is the application of a systematic, disciplined, quantifiable, and interdisciplinary approach to the ideation, conceptualization, development, operation, and maintenance of Cloud Computing, as well as the study and applied research of the approach, i.e., the application of engineering to Cloud. It is a maturing and evolving discipline to facilitate the adoption, strategization, operationalization, industrialization, standardization, productization, commoditization, and governance of Cloud solutions, leading towards a Cloud ecosystem. Cloud engineering is also known as Cloud service engineering.

Cloud storage

Cloud Storage is a model of networked computer data storage where data is stored on multiple virtual servers, generally hosted by third parties, rather than being hosted on dedicated servers. Hosting companies operate large data centers; and people who require their data to be hosted buy or lease storage capacity from them and use it for their storage needs. The data center operators, in the background, virtualize the resources according to the requirements of the customer and expose them as virtual servers, which the customers can themselves manage. Physically, the resource may span across multiple servers.

The Intercloud

The Intercloud is an interconnected global "cloud of clouds" and an extension of the Internet "network of networks" on which it is based. The term was first used in the context of cloud computing in 2007 when Kevin Kelly stated that "eventually we'll have the intercloud, the cloud of clouds. This Intercloud will have the dimensions of one machine comprising all servers and attendant cloudbooks on the planet.". It became popular in 2009 and has also been used to describe the datacenter of the future.

The Intercloud scenario is based on the key concept that each single cloud does not have infinite physical resources. If a cloud saturates the computational and storage resources of its virtualization infrastructure, it could not be able to satisfy further requests for service allocations sent from its clients. The Intercloud scenario aims to address such situation, and in theory, each cloud can use the computational and storage resources of the virtualization infrastructures of other clouds. Such form of pay-for-use may introduce new business opportunities among cloud providers if they manage to go beyond theoretical framework. Nevertheless, the Intercloud raises many more challenges than solutions concerning cloud federation, security, interoperability, QoS, vendor's lock-ins, trust, legal issues, monitoring and billing.

The concept of a competitive utility computing market which combined many computer utilities together was originally described by Douglas Parkhill in his 1966 book, the "Challenge of the Computer Utility". This concept has been subsequently used many times over the last 40 years and is identical to the Intercloud.

Issues

Privacy

The Cloud model has been criticized by privacy advocates for the greater ease in which the companies hosting the Cloud services control, and thus, can monitor at will, lawfully or unlawfully, the communication and data stored between the user and the host company. Instances such as the secret NSA program, working with AT&T, and Verizon, which recorded over 10 million phone calls between American citizens, causes uncertainty among privacy advocates, and the greater powers it gives to telecommunication companies to monitor user activity. While there have been efforts (such as US-EU Safe Harbor) to "harmonize" the legal environment, providers such as Amazon still cater to major markets (typically the United States and the European Union) by deploying local infrastructure and allowing customers to select "availability zones."

Compliance

In order to obtain compliance with regulations including FISMA, HIPAA and SOX in the United States, the Data Protection Directive in the EU and the credit card industry's PCI DSS, users may have to adopt *community* or *hybrid* deployment modes which are typically more expensive and may offer restricted benefits. This is how Google is able to "manage and meet additional government policy requirements beyond FISMA" and Rackspace Cloud are able to claim PCI compliance. Customers in the EU contracting with Cloud Providers established outside the EU/EEA have to adhere to the EU regulations on export of personal data.

Many providers also obtain SAS 70 Type II certification (e.g. Amazon, Salesforce.com, Google and Microsoft), but this has been criticised on the grounds that the hand-picked set of goals and standards determined by the auditor and the auditee are often not disclosed and can vary widely. Providers typically make this information available on request, under non-disclosure agreement.

Legal

In March 2007, Dell applied to trademark the term "cloud computing" (U.S. Trademark 77,139,082) in the United States. The "Notice of Allowance" the company received in July 2008 was canceled in August, resulting in a formal rejection of the trademark application less than a week later. Since 2007, the number of trademark filings covering cloud computing brands, goods and services has increased at an almost exponential rate. As companies sought to better position themselves for cloud computing branding and marketing efforts, cloud computing trademark filings increased by 483% between 2008 and 2009. In 2009, 116 cloud computing trademarks were filed, and trademark analysts predict that over 500 such marks could be filed during 2010.

Other legal cases may shape the use of cloud computing by the public sector. On October 29, 2010, Google filed a lawsuit against the U.S. Department of Interior, which opened up a bid for software that required that bidders use Microsoft's Business Productivity Online Suite. Google sued, calling the requirement "unduly restrictive of competition." Scholars have pointed out that, beginning in 2005, the prevalence of open standards and open source may have an impact on the way that public entities choose to select vendors.

Open source

Open source software has provided the foundation for many cloud computing implementations. In November 2007, the Free Software Foundation released the Affero General Public License, a version of GPLv3 intended to close a perceived legal loophole associated with free software designed to be run over a network.

Open standards

Most cloud providers expose APIs which are typically well-documented (often under a Creative Commons license) but also unique to their implementation and thus not interoperable. Some vendors have adopted others' APIs and there are a number of open standards under development, including the OGF's Open Cloud Computing Interface. The Open Cloud Consortium (OCC) is working to develop consensus on early cloud computing standards and practices.

Security

The relative security of cloud computing services is a contentious issue which may be delaying its adoption. Issues barring the adoption of cloud computing is due in large part to the private and public sectors unease surrounding the external management of security based services. It is the very nature of cloud computing based services, private or public, that promote external management of provided services. This delivers great incentive amongst cloud computing service providers in producing a priority in building and maintaining strong management of secure services.

Organizations have been formed in order to provide standards for a better future in cloud computing services. One organization in particular,the Cloud Security Alliance is a non-profit organization formed to promote the use of best practices for providing security assurance within Cloud Computing.

Availability and performance

In addition to concerns about security, businesses are also worried about acceptable levels of availability and performance of applications hosted in the cloud.

There are also concerns about a cloud provider shutting down for financial or legal reasons, which has happened in a number of cases.

Sustainability and siting

Although cloud computing is often assumed to be a form of "green computing", there is as of yet no published study to substantiate this assumption. Siting the servers affects the environmental effects of cloud computing. In areas where climate favors natural cooling and renewable electricity is readily available, the environmental effects will be more moderate. Thus countries with favorable conditions, such as Finland, Sweden and Switzerland, are trying to attract cloud computing data centers.

SmartBay, marine research infrastructure of sensors and computational technology, is being developed using Cloud computing, an emerging approach to shared infrastructure in which large pools of systems are linked together to provide IT services.

Research

A number of universities, vendors and government organizations are investing in research around the topic of cloud computing. Academic institutions include University of Melbourne (Australia), Georgia Tech, Yale, Wayne State, Virginia Tech, University of Wisconsin–Madison, Carnegie Mellon, MIT, Indiana University, University of Massachusetts, University of Maryland, North Carolina State University, Purdue University, University of California, University of Washington, University of Virginia, University of Utah, University of Minnesota, among others.

Joint government, academic and vendor collaborative research projects include the IBM/Google Academic Cloud Computing Initiative (ACCI). In October 2007 IBM and Google announced the multi- university project designed to enhance students' technical knowledge to address the challenges of cloud computing. In April 2009, the National Science Foundation joined the ACCI and awarded approximately $5 million in grants to 14 academic institutions.

In July 2008, HP, Intel Corporation and Yahoo! announced the creation of a global, multi-data center, open source test bed, called Open Cirrus, designed to encourage research into all aspects of cloud computing, service and data center management. Open Cirrus partners include the NSF, the University of Illinois (UIUC), Karlsruhe Institute of Technology, the Infocomm Development Authority (IDA) of Singapore, the Electronics and Telecommunications Research Institute (ETRI) in Korea, the Malaysian Institute for Microelectronic Systems (MIMOS), and the Institute for System Programming at the Russian Academy of Sciences (ISPRAS). In Sept. 2010, more researchers joined the HP/Intel/Yahoo Open Cirrus project for cloud computing research. The new researchers are China Mobile Research Institute (CMRI), Spain's Supercomputing Center of Galicia (CESGA by its Spanish acronym), Georgia Tech's Center for Experimental Research in Computer Systems (CERCS) and China Telecom.

In July 2010, HP Labs India announced a new cloud-based technology designed to simplify taking content and making it mobile-enabled, even from low-end devices. Called SiteonMobile, the new technology is designed for emerging markets where people are

more likely to access the internet via mobile phones rather than computers. In Nov. 2010, HP formally opened its Government Cloud Theatre, located at the HP Labs site in Bristol, England. The demonstration facility highlights high-security, highly flexible cloud computing based on intellectual property developed at HP Labs. The aim of the facility is to lessen fears about the security of the cloud. HP Labs Bristol is HP's second-largest central research location and currently is responsible for researching cloud computing and security.

The IEEE Technical Committee on Services Computing in IEEE Computer Society sponsors the IEEE International Conference on Cloud Computing (CLOUD). CLOUD 2010 was held on July 5–10, 2010 in Miami, Florida

Criticism of the term

During a video interview, Forrester Research VP Frank Gillett expresses criticism about the nature of and motivations behind the push for cloud computing. He describes what he calls "cloud washing" in the industry whereby companies relabel their products as cloud computing resulting in a lot of marketing innovation on top of real innovation. The result is a lot of overblown hype surrounding cloud computing. Gillett sees cloud computing as revolutionary in the long term but over-hyped and misunderstood in the short term, representing more of a gradual shift in our thinking about computer systems and not a sudden transformational change.

Larry Ellison, CEO of Oracle Corporation has stated that cloud computing has been defined as "everything that we already do" and that it will have no effect except to "change the wording on some of our ads". Oracle Corporation has since launched a cloud computing center and worldwide tour. Forrester Research Principal Analyst John Rymer dismisses Ellison's remarks by stating that his "comments are complete nonsense and he knows it".

Richard Stallman said that cloud computing was simply a trap aimed at forcing more people to buy into locked, proprietary systems that would cost them more and more over time. "It's stupidity. It's worse than stupidity: it's a marketing hype campaign", he told The Guardian. "Somebody is saying this is inevitable – and whenever you hear somebody saying that, it's very likely to be a set of businesses campaigning to make it true."