Simple Description of the Internet and World Wide Web and How They Differ

Kelly Coffey; University of Alabama at Birmingham; Birmingham, AL

Keywords: World Wide Web, www, Internet, HTML, Hyper Text Markup Language, URL, IP address, browser, server, ISP, POP, NAP, T1, T3, OC-3, OC-12, OC-48, router, TCP/IP, email.

1. INTRODUCTION

Many people may think that the World Wide Web, Internet, and internet are all the same thing, but someone who knows a little bit about each of these subjects knows that they are not.

1.1. Definition of the World Wide Web

"World Wide Web - n. Abbr. www. The complete set of documents residing on all Internet servers that use the HTTP protocol, accessible to users via a simple point-and-click system [1]."

The Web has a mystical quality to those who do not fully understand the dictionary definition. This is completely understandable, since there is much more to be said. The Web is not the physical connection between the computers that log onto their AOL accounts. Rather, it is a means for making the Internet more user-friendly to the public by linking together software types and protocols in order to transfer documents [2].

1.2. Definition of the Internet

The Internet is a system of networks connecting computers around the world via the TCP/IP protocol [1]. The Internet involves a collection of computers and/or other devices that are physically connected by equipment allowing communication between them.

1.3. Definition of the Internet

According to the American Dictionary, internet and Internet are the same thing, but in actuality, they are not [1]. Internet is the term assigned to what used to be known as ARPAnet and NSFnet, but is now the network of networks that makes up the ability to have global access to information in a matter of minutes. The Internet, on the other hand, is a more generic term used to describe interconnected computers. This could mean an interconnection of two computers or an interconnection of two hundred computers. The word, internet, is just a generic term, while Internet, is a specific term. The Internet is just one internet. An internet could use any type protocol that the creators of the network wished to use, and the computers could be connected in any manner allowable [3].

2. ORIGINS

The need (backed by the money) first arose for a large-scale computer network in the 1960's when the Department of Defense (DoD) was interested in developing a way to communicate and share programs on defense-related subjects. The issue was to ensure that if one node was destroyed, the rest of the network could continue to function. The DoD's Advanced Research Project Agency (ARPA) funded the development of the first internet, which became known at that time as the ARPAnet. The first internet consisted of about twelve research laboratories and universities. The first node was created at UCLA in 1969.

Later the National Science Foundation (NSF) decided to use the technology to link universities together over a network. Naturally, what they created became known as the NSFnet. Eventually, people began to refer to them both as the Internet.

The original development was one network that proceeded to grow until the need arose to create a network of networks or interconnected networks, thus the name Internet.

The early Internet was completely text based and non-user friendly. It was a simple means of sending e-mails. There needed to be some changes made in order to allow color, pictures, video, and audio to be transmitted.

Different protocols were created to handle different particular functions. Soon, it became evident that there needed to be something to tie all these protocols together and create a more user-friendly environment. What was created was the World Wide Web. The World Wide Web's growth across the
globe was and is exponential, and now, there are enough computers connected to the Internet so that almost any information is available to anyone.

3. MECHANICS

3.1 Interconnected Networks

The word Internet comes from an interconnection of networks. People at home may use a dial up modem in order to access their Internet Service Provider (ISP). If a person works for a large enough business, they may be part of a Local Area Network (LAN) that probably still connects using a contracted ISP.

When a person connects to their ISP, they become part of their network. The ISP can connect to a larger network and hence become part of that network. The Internet is simply a network of networks [4]. Originally all computers were connected to one network, but this quickly became unfeasible due to growth [2].

3.2 POPs and NAPs

In each region there will be a Point of Presence (POP) that local users will use to access the network. Usually, this POP will be accessed through a dedicated line or local phone number. The POPs will be connected through Network Access Points (NAPs). NAPs will be located in various cities, which several Internet providers will connect to.

Figure 1 - Layout of how home computer connects to Internet [4].

If that was not already confusing enough, the connections between each of these are usually as follows: people from their homes will be connected to a POP by a dial up modem phone line, cable line, or with DSL; LANs will be connected to a POP by a T1 (1.5 Mbps) line; POPs will be connected to NAPs by a T3 (6.5 Mbps) line; and NAPs will be connected to each other by OC-3 (155.52 Mbps), OC-12 (622.06 Mbps) or OC-48 (2.488 Gbps) lines [3] [9, pp. 180-181]. These OC lines make up the backbone of the Internet, which are made up of fiber optic trunk lines designated for optical carrier [4].

3.3 Routers

Routers are used to determine which computer to send information to or which direction to send it in order to reach that destination. Routers will send information that is meant to be sent from one network to another and will block unnecessary information from spilling over [4]. The information arrives to the routers in packets, which will be discussed later [5].

4. ADDRESSING

4.1 IP Addresses

If a person or program is to send information from one computer to another, then there has to be a way of distinguishing between all the different computers connected to the Internet. Every computer has a specific Internet Protocol (IP) address. One example is

216.27.61.137

Computers communicate in binary, so the same IP address in binary would be

11011000.00011011.00111010.10001001

Binary IP addresses contain 32 bits (32 numbers each with two states: a 1 or 0), which means there are 2^32 possible combinations or 4, 294, 967, 296 different number values [4]. In 1998 a new IP standard was approved, IPv6, which changes the number of bits from 32 to 128. This change was made due to the rapidly diminishing number of unused IP addresses [2].

IP addresses like the example above are divided into octets and separated by periods. The octets divide the addresses into classes for a purpose [4]. An Internet routing computer uses each octet separately in order to decide where the information will go.

Organizations can be assigned a block of IP addresses with the size of the block depending on the size of that organization. The Department of Defense is assigned 16 million IP addresses from 12.0.0.0 up to 12.255.255.255 [1]. 0.0.0 is reserved for the default network, and 255.255.255.255 is reserved for broadcasting.

The first octet is referred to as the Net Section. This identifies the network. The last octet is the host section or node section. The Domain Name System (DNS) is responsible for converting text names into IP addresses for the computer to be able to read [4].

4.2 URLs

Some first level domain names include .com, .org, .edu, .net, and .gov. Some second level domain names included under the .com domain name are yahoo, AOL, and Microsoft. www is a host name. This designates a certain machine with a specific address [4]. Each host name must be unique to the domain. www could easily be replaced with any other combination of letters or numbers but would require the name to be mapped into the DNS.

The root DNS server knows the addresses of all DNS
servers associated with handling the 1st level domains. For example, there is a .com DNS server and an .edu DNS server, etc... There are multiple DNS servers of each type in each level so that if one fails, there is another to take over.

5. WEB SERVERS AND BROWSERS

5.1 Web Servers

Every computer on the Internet is either a server or a client. The server provides a service or information to the client. The client sends direct requests to a certain software server running on a server machine. Servers generally have static IP addresses while clients dial up and are randomly and temporarily assigned an IP number that will change every time they log on. The ISP does not need to reserve nearly as many IP addresses this way, which will save them a great deal of money in the long run [2].

6. HTTP

HTTP is the protocol to which web servers on the Internet conform [5].

6.2 TCP/IP

TCP/IP is the basic communication language of the Internet. It is a 2-layer program. The 1st layer is the Transmission Control Protocol (TCP), which turns messages into smaller packets to be transmitted over the Internet, and the second layer is the Internet Protocol (IP), which handles the address part of the packet. This address is checked at each router or gateway computer to ensure its arrival to the destination address. For this protocol to work properly, every machine that uses the Internet must have a TCP/IP [4].

6.3 Other Miscellaneous Protocols

Other protocols exist such as Telnet, which allows a person to use one computer in order to log onto and use another computer. File Transfer Protocol (FTP) is used to transfer files from one computer to another. Usenet is similar to an electronic bulletin board. Mailto allows messages to be sent from the user of one computer to another computer. File protocol means the document is residing on the machine running the browser. An example would be file://path-to-document [2].

7. MISCELLANEOUS TOPICS

7.1 E-mail

The E-mail system is not very complex. E-mails are nothing more than text messages. If a person wants more than just text in their message, it must be in the form of an attachment. Attachments are converted to and from binary using uudecode and uudecode. Modern e-mail systems perform these conversions without the user ever knowing anything took place at all.

There are many e-mail servers running on the Internet at any given moment with certain ports on the machine waiting to be attached to by certain ports on the client machines sending the message. The e-mail client from whom the sender is sending his/her message (ex/ Microsoft Outlook) has a list of account holders with a single text file for each of these accounts, as does the receiving e-mail server. The person that was sending the message would press send, and after they pressed send, the message would be sent from the e-mail client to the e-mail server and the received message would be saved as a text file under that recipient’s account; e.g.,

From: coffek74
To: raygan
Professor,
May I have an “A” on this paper, please?
Thanks,
Kelly

The simplest possible breakdown of the steps to sending an e-mail is as follows: The client would ask the server to send a copy of the text file, then the client would ask the server to erase and reset the text file for that account, then the client would save the text file onto the local machine, and finally the...
file would be separated into individual messages being
separated by the “From:” (such as in the example above) and
displayed with the headers only and not the body.

The e-mail system consists of two major servers: the Simple
Mail Transfer Protocol (SMTP), which handles outgoing mail,
and the Post Office Protocol version 3 (POP3), which handles
incoming mail. The e-mail server, such as Outlook Express,
connects to the SMTP server who in turn contacts the Domain
Name Server (DNS). The DNS will return the IP address of
the SMTP for the domain of the recipient. If
somebody@aol.com is who the message is being sent to,
somebody would be the recipient name while aol.com would
be the domain name. The DNS locates the IP address, in this
case, for aol.com. When the message is sent to the aol.com
server, the message will be sent to the POP3 server using a
delivery agent program. If the message cannot be
immediately received, then the message will be placed in a
queue [7].

7.2 Search Engines

An Internet search engine is nothing more than a program
running on a computer operating to help people find
information on different websites. Internet search engine
usually is really referring to a Web search engine. Search
engines search certain parts of the Web using key words. They
store the words they find and where they found them, then
they allow users to use that information that is stored in that
search engine to locate information that may be of use to
them.

Search engines have evolved from using a few hundred
thousand pages of stored information to using millions of

pages. They have gone from performing a couple of thousand
inquiries a day to tens of millions of inquiries a day. When a
person makes an inquiry on a search engine, the search engine
must first locate the document. To do this, the search engine
uses what are called spiders. A spider is a special software
robot designed to search Web pages and gather a list of key
words from each of the Web pages visited. This process of
searching the Web sites is known as Web crawling.

Spiders begin their journey to searching pages by starting
with the most widely used servers and searching their most
popular pages. The spider will follow every link from the web
page that was just indexed. A typical large search engine,
such as google.com, will usually have approximately three
spiders working at any given time. Each spider can have over
hundred connections to web pages at once. This means
that the spiders can crawl over right under one hundred pages
per second at peak performance.

The way some spiders work, is that they will make a list of
important words such as those found in the title, subheadings,
links, and the top one hundred repeated words in the body
excluding words such as and, or, but, an, and the. Other
spiders will index every single word used on the page in order
to have a more complete searching approach. Different spiders
use different approaches.

Owners of web pages can choose to include meta tags on
their pages, which will include a list of key words. Doing this
can help the spiders create indexes of the information on the
page. However, as a safety precaution, the spiders will attempt
to correlate the keywords with the actual words used in the
paper in order to prevent useless index words from being
associated with certain web pages. Unless the safety
precautions were used, unscrupulous owners could place
keywords in their meta tag that have little or nothing to do
with the web page itself.

The robot exclusion protocol was developed to prevent a
spider from indexing any words on the page or following any
of its links.

The easiest approach to using the information that is
gathered by the spiders would be to just store the word and the
URL where it was found. Obviously, if this were the method
actually used, then there would be no way to tell the difference
between a significant word used throughout the page, and an
insignificant word used only once. There has to be a way to
assign a weight to each word in the page. Some ways of doing
this are to assign a greater value to words appearing in the
title, subheadings, links, meta tags, and towards the top
portion of the page. Each search engine assigns “weights”
differently.

There are several different ways to index information. One
way is to build a hash table [7]. The advantage of using
hashing is that it makes the process of retrieving information
much faster. For example, for a person to look up a name in
the phone book starting with the letter ‘S’ would take longer
than a letter starting with ‘Z,’ since more last names start with
an ‘S’ than they do a ‘Z.’ Hashing evens out the search time
[8].
7.3 Web Pages

A Web page is a text file that contains tags that are suggestions to how the browser should make the page appear on the screen. The tags are usually HTML tags. The tags are simple instructions that can tell the browser to change a words color, size, font, or position on the page.

The simplest tool possible to use to build a Web page is the tool that is already on every computer: Notepad (SimpleText on Macintosh systems). To use notepad, a person needs to know a little bit about HTML tags and how to use them; e.g.,

```
<html>
  <head>
    <title>This is the title</title>
  </head>
  <body>
    Hello World!
  </body>
</html>
```

Then the file must be saved on the computer with an html extension (example/ KellysPage.html). A person can create a Web page and view the page on his/her own computer and stop there, but if the creator wants the page to be viewable to the world, than he/she must place the page on a Web server.

There are several ISPs (such as aol) and universities that will provide space to their users or students for their own Web page for free. Geocities, tripod, and homestead also offer free Web space, but they require an advertisement banner to appear at the top of the page.

Another option is to pay a small fee to a hosting service for Web space. If the route chosen is through a hosting service, then the owner has to option of choosing a unique domain name. Most professional Web pages choose to create their Web sites this way.

If a person wants their own domain name, they must check with the Accredited Registrars at ICANN.org. This agency is responsible for maintaining IP addresses. If the owner already has a hosting service, then the price of reserving a domain name will be less. Most hosting services will handle all the registration.

One way to promote a site so that people on the Web can locate the page is to register the site on different search engines. On the search engine Web site there is usually a “submit a site” link or something very similar. Each individual page must be submitted for the spider to crawl over [9].

7.4 HTML

Hyper Text Markup Language (HTML) is a computer language used to provide a guide to how a page should appear aside from being just plain black and white text appearing across the screen. HTML is not a programming language though. Calculations cannot be performed, and animation cannot be provided.

HTML can be used to create formatted words, links to other pages, links to email, inserted objects, lists, tables, background color, etc…

HTML also has to capability of using frames. Frames make one portion of a page remain stationary while the rest of the page changes according to the links used. If a frame is being used and there is one stationary portion of the page and one portion that is being changed, then there are two HTML documents displaying at once on the screen [9].

8. CONCLUSION

As has been seen in the paper, the main ingredients to a working Internet are networks, routers, NAPs, ISPs, DNS, and powerful servers. Information can easily, readily, and seemingly instantaneously be sent around the world. Everyday, new technologies are being developed to handle increased speeds on the Internet. And to think, all this was achieved in a matter of two decades. Who knows what might be possible in two more decades.

Hopefully, after reading this paper, a person who did not already completely understand how the Internet and/or World Wide Web worked will now understand a little bit better. All the major points were touched on and this paper should serve as a good reference for future research on a variety of topics.

REFERENCES


BIOGRAPHIES

Kelly L. Coffey
Computer and Electrical Engineering Student
University of Alabama at Birmingham
6121 Barrington Parkway
Moody, AL 35004 USA
E-mail: coffek74@eng.ua.edu

Kelly Coffey is a Computer and Electrical Engineering student at the University of Alabama at Birmingham where she will graduate in May of 2003. Before attending UAB, she was a student at Jefferson State Community College in Birmingham, Alabama where she earned her associates in applied science.