**internet&worldwideweb**

**Internet or World Wide Web**

**A Data Communication Historical Series**

**By Bob Pollard**

**A basic overview of the operation and functions that make the Internet or World Wide Web possible (ca 2003):**

When a person makes a normal telephone call to someone on the other side of the country, the telephone system establishes a circuit between the calling telephone and the called telephone. The circuit might involve a half dozen or more steps through copper cables, telephone exchange switches, fiber optic cable, microwave and satellite. When the connection is established it will remain constant for the duration of the call. This circuit connection approach means that the quality of the line between the calling and the called device is consistent throughout the call, good or bad. But a problem with any portion of the circuit, some sort of a failure, can interrupt the call.

When a person sends an e-mail message, with or without an attachment, or accesses a Web Page or downloads a file from the other side of the country, a very different process is used.

It should be noted that up to the advent of the World Wide Web (WWW) / Internet most communication systems were a closed environment when considering unwanted outside interference or manipulation. In other words, it was very difficult for some outside source to enter a system computer and initiate a program revision, insert information or read and remove internal information.

What is the Internet? The Internet is a collection of millions of computers, all capable of being linked together on a computerized network. The network allows all of the computers to communicate with one another. A home computer may be linked to the Internet using a telephone line (dial-up), Digital Subscriber Line (DSL) or a cable MODEM that communicates with an Internet Service Provider (ISP). A computer in a business or some other large complex will usually have a Network Interface Card (NIC) that directly connects the Internet to the business Local Area Network (LAN).

The business may connect its LAN to an ISP using a high-speed phone line such as T1 line. A T1 line can handleapproximately 1.5 million bits per second (mbps), while a normal dial-up phone line using a MODEM can typically handle 14 kbps to 50 kbps (thousand bits per second), even when the MODEM specifies it can operate at 56.6 kbps. This reduction in kbps rate is caused by noise, distortion, bias and signal delay, which could be caused by the lines, equipment, distance (length of the circuit), etc.

Small Internet Service Providers (ISPs) connect to larger ISPs and the largest ISPs maintain common links (backbones) for a region or the entire nation. Backbone links around the world are connected through fiber-optic lines, undersea cables or satellite links. In this manner, every computer on the Internet can be connected to every other computer on the Internet.

Internet data travels over a Packet Switched network and the data in a message or file is broken up into Packets of 1,500 bytes or less. Each of these Packets is framed by information (leading and trailing) that includes the sender's address, the receiver's address, each Packet's place in the entire message, and error tests that allow the receiving computer to make sure that each Packet and the entire message (data) arrived intact. Each data Packet is sent off to its destination via the best available route. This route might be taken by all the other Packets in the message or by none of the other Packets in the message. This might seem very complicated compared to the direct connection approach used by the telephone system, but in a network designed for data there are two advantages to the Packet switching method:

* The network can balance the load across various pieces of equipment on a millisecond-by-millisecond basis.
* If there is a problem with one piece of equipment in the network while a message is being transferred, Packets can be routed around the problem, ensuring the delivery of the entire message.

Packet:All the information sent on the Internet involves Packets. For example: Web pages that you receive come as a series of Packets, and every e-mail you send or receive leaves or arrives as a series of Packets. Each Packet carries the information that is necessary to get it to the proper destination. For example: the sender's IP (Internet Protocol) address, the intended receiver's IP address, total number of Packets in the e-mail and a sequence number for each individual Packet. The Packets carry the data regulated by the protocols used by the Internet, Transmission Control Protocol/Internet Protocol (TCP/IP). Each Packet contains part of the body of your message. A typical packet contains 1,000 or 1,500 bytes.

As mentioned earlier each Packet is transmitted to its destination by the best available route. The network balances the load across various pieces of equipment on a millisecond-by-millisecond basis. If there is a problem with a piece of equipment in the network while a Packet is being transferred the Packet can be routed around the problem, ensuring the delivery of each Packet.

Depending on the type of network, Packets may be referred to by another name: such as: frame, block, cell or segment.

Most Packets are split into three parts:

1. Header: The header contains instructions for the data carried by the Packet; including:

 Length of Packet (this may be a fixed length, or as specified in the header)

 Synchronization bits (to maintain synchronization between the transmitter and receiver)

 Packet number (which Packet this is in a sequence of Packets)

 Protocol (on networks that carry multiple types of information, the protocol define what type of Packet is being transmitted, e-mail, Web page, video, etc.).

 Destination address (receiving device)

 Originating address (sending device)

2. Payload: Also called the ‘body’ or ‘data’ of a Packet. This is the actual data that the Packet is delivering to the destination. If a Packet is fixed-length, then the payload may be padded (filled) with blank information to make it the right size.

3. Trailer: The trailer, sometime called the ‘footer’, normally contains a couple of bits that identifies the end of the packet. It may also have of error checking bits. The most common error checking used in Packets is Cyclic Redundancy Check (CRC). One CRC method takes the sum of all the 1 bits in the payload and adds them together. The result is stored as a hexadecimal value in the trailer. The receiving device adds up all the 1 bits in the payload and compares the result to the value stored in the trailer. If the values match, the Packet is good. Mismatched values cause the receiving device to send a request to the originating device to resend the packet.



Example e-mail: A representative e-mail is about 3,500 bits (3.5 kilobits) in size. The connected network uses fixed-length Packets of 1,024 bits (1 kilobit). The header of each packet is 96 bits long and the trailer is 32 bits long, leaving 896 bits for the payload. Dividing the 3,500-bit message into the proper size Packets will result in four individual Packets (divide 3,500 by 896), three packets will contain 896 bits and the fourth will have 812 bits.

Each Packets header will contain the proper protocols, the originating address (IP address of the sending computer), the destination address (IP address of the receiving computer) and the Packet number (1, 2, 3 or 4 since there are 4 Packets). Routers in the network will look at the destination address in the header and compare it to their lookup table (configuration table) to determine where to send the Packet. Once the Packet arrives at its destination, the receiving computer will strip the header and trailer off each Packet and reassemble the e-mail based on the numbered sequence of the Packets.

Router:

The ‘Router(s)’, which are computers that communicate with each other and make up the main part of the Internet (network), can configure and when necessary reconfigure the paths that Packets take. This routing control is based on the information framing (header and trailer) of the data Packet and, in addition; line routing is based on equipment or line conditions. Rerouting will occur when there are delays in receiving and sending data (Packets) on various parts of the network.

The size and complexity of a Router (computer) is normally determined by the size of the Network and the workload requirements. For example:

 When the job function requires Internet connection sharing between two Windows 98-based computers, one of the computers (the computer with the Internet connection) will assume the role of a simple Router. In this instance the Router is simply looking at the data framing (address) to see which computer should receive the data. The Router program can operate in the background of the system without significantly affecting the running of other programs.

 Slightly larger Routers, the ones used to connect a small office network to the Internet, will perform more functions. These Routers frequently enforce rules concerning security for the office network. They also handle a higher data (message) volume so they are generally stand-alone computers rather than software running in the background on a server.

 The largest Routers, those used to handle large volumes of data at the major traffic points on the Internet, handle millions of data Packets every second. These stand-alone computers also configure the network for the most efficient data transmission routing. These large Router systems have far more in common with ‘supercomputers’ than they do with an office server.

Consider a medium size Router handling two networks, an office network with about 40 computers and devices, and the Internet. The office network of 40 computers connects to the Router through an Ethernet connection, a 100 base-T connection, which means the connection can operate at 100 megabits per second (mbps), and uses a twisted-pair cable; an 8-wire version of the cable that connects your telephone to the wall jack. There are two connections between the Router and the ISP (Internet Service Provider). One is a T-1 connection that supports 1.5 megabits per second and the second connection is an ISDN (Integrated Services Digital Network) that supports 128 kilobits per second (kbps). The configuration table in the Router determines that all out-bound Packets, to the Internet, are to use the T-1 line, unless it's unavailable for some reason. If the T-1 line can't be used then the outbound traffic is automatically transmitted on the ISDN line. The ISDN line is a back-up in case there is a problem with the faster T-1connection and the switch over from the T-1 line to the ISDN line is performed automatically, without manual intervention.

In addition to routing Packets from one point to another the Router also performs security functions and limits computer connections from outside the network. For example: The Router may use a mechanism called a ‘Subnet Mask’ to determine if a message should remain within the local office network or be routed to the Internet network. The Subnet mask looks like an IP (Internet Protocol) address and is used to determine if the message sender and the receiver share the first three groups of address numbers. If the numbers match then the message should be delivered within the local network and not routed to an outside network (Internet). For example: The computer at address 10.16.28.25 sends a message to the computer at 10.16.28.45. TheRouter, which sees all the Packets, matches the first three groups in the address of both sender and receiver (10.16.28) and causes the Packets to route within the local network. If the address, first three groups, did not match, which indicates the message should be routed to the Internet, the Router will process the message for delivery to the Internet.

Large Router: Backbone of the Internet:
In order to handle all the users on the Internet or a large private network, millions and millions of Packets must be processed simultaneously. Cisco Systems, Inc. a company that specializes in networking hardware manufactures some of the largest Router computer systems, such as, their Gigabit Switch Router 12000 series of Routers. These Router computers use the same design features found in some of the most powerful supercomputers, a design that allows the linking of many different processors together. The largest model in the 12000 series, the 12016, uses a series of switches that can handle up to 320 billion bits of information per second and, with the full complement of boards, can move up to 60 million packets of data every second.

Two prime functions a Router performs would include the following:

1. The Router ensures that information doesn't go where it’s not needed. This is crucial for keeping large volumes of data from clogging the connections of innocent users.

2. The Router ensures that information is delivered to the intended destination.

Figure 1: This view is not geographically accurate, but does provide a view of how routers are configured across the nation or worldwide. A message sent from # 1 to # 2 would be divided into Packets and each Packet would flow through the quickest and most reliable path, which would involve 8 or more routers. A request from # 1 for a Web page from # 2 would result in data (Packets) transmission in both directions.

****

**Figure 1**

Configuration Table:

The Router will scan the destination address and match that IP address against rules in the configuration table. The rules will determine that Packets in a particular group should go in a specific direction. Next the Router will check the reliability of the primary connection for the established direction against another set of rules. If the reliability of the connection is good, the Packet is transmitted and then the next Packet processed. If the connection is not performing up to expected parameters, then an alternate route is chosen and checked. When a reliable connection is found the Packet will be transmitted. All this activity happens in a fraction of a second and this sequence of events occur a millions times a second, 24 hours a day.

A configuration table is a collection of information, including:

 Information on which connections lead to particular groups of addresses

 Priorities for connections to be used

 Rules for handling both routine and special cases.

 A configuration table could involve half dozen lines in the smallest Routers, but can grow to massive size and complexity in the very large Routers, that handle the bulk of Internet messages.

A Router, then, has two separate but related jobs:

 The Router ensures that information doesn't go where it's not needed. This keeps large volumes of data from clogging the connections of Innocent users.

 The Router makes sure that information does make it to the intended destination.

In performing these two jobs, a Router is extremely useful in dealing with two separate computer networks. It joins the two networks, passing information from one to the other and, in some cases, performing translations of various protocols between the two networks. It also protects the networks from one another, preventing the traffic on one from unnecessarily spilling over to the other. As the number of networks attached to one another grows, the configuration table for handling traffic among them grows, and the processing power of the Router is increased. Since the Internet is one huge network made up of tens of thousands of smaller networks the use of Routers is a necessity.

E-mail:

For the majority of e-mail clients (users) the Internet e-mail system, LAN or ISP, will consist of two different server programs running on a server computer. One is called the SMTP (Simple Mail Transfer Protocol) Server that handles outgoing mail. The other is a POP3 (Post Office Protocol) Server that handles incoming mail.

Figure 2 provides a basic picture of the SMTP and POP3 server position in the E mail sending or receiving activity from the E mail client end of the process. The text files (receive) would contain information placed there by the Internet server (LAN or ISP). The STMP queue would contain information placed there by the E mail client awaiting transfer to the Internet.



Figure 2

Whenever an individual (client) sends e-mail, the e-mail client (program) interacts with the SMTP server to handle the sending process. The SMTP server on your host, which could be a LAN or Internet Service Provider (ISP), may have conversations with other SMTP servers before delivering the e-mail. The term ‘client’ indicates a user program connection to a LAN or to an outside Internet Service Provider. The difference being that a telephone call connection, or some other form of connection, must be performed when connecting to the outside ISP. Also, the interaction may vary somewhat.

A typical simplified user e-mail transmission may function as follows:

The Simple Mail Transfer Protocol (SMTP) sending server will monitor computer port number 25, while POP3 (Post Office Protocol) receiving server will monitor port 110.

The client’s ID (identification) is eyesight and the e-mail account are maintained by seafoam.com, which is the LAN host or the ISP. The outgoing e-mail is addressed to rblack@fairhaven.com. For clarification the user (client) program is Outlook Express.

When the e-mail account was set up the information, mail.seafoam.com was provided to Outlook Express.

After the e-mail (message) has been composed and the send key is depressed the following will occur:

* Outlook Express connects to the SMTP server at mail. Seafoam.com using port 25.
* Outlook Express has an interaction with the SMTP server, giving the SMTP server the address of the sender and the address of the recipient, as well as the body of the message.
* The SMTP server takes the recipient (to) address, rblack@fairhaven.com, and breaks it into two parts:
	+ The recipient name (rblack)
	+ The domain name (fairhaven.com).

Note: Within a LAN if the to (recipient) address was another user at seafoam.com, the SMTP server would simply hand the message to the POP3 server for seafoam.com. Since the recipient is at another domain, a server on a different network, the SMTP needs to communicate with that domain.

The SMTP server interacts with a Domain Name Server (DNS), a separate server system, in order to obtains the Internet Protocol (IP) address of the server for fairhaven.com. The DNS replies with the one or more IP addresses for the SMTP server(s) that Fairhaven operates.

The SMTP server at seafoam.com connects with the SMTP server at Fairhaven via the Internet using port 25. The necessary interaction occurs and the message is delivered to the Fairhaven server from seafoam.

The Fairhaven server recognizes that the domain name for rblack is at Fairhaven so the message in posted to Fairhaven’s POP3 server, which puts the message in rblack’s mailbox.

Internet Protocol (IP):

A unique 32-bit IP Address (Internet Protocol Address) is assigned to every computer connected directly to the Internet and this unique identifying number would be a grouping of numbers (up to 12) such as, 192.1.26.2. Normally an e-mail originator is connected to a server or an Internet Service Provider (ISP). In this case the IP address will not be obvious to the e-mail originator. The IP address would be assigned to the server or ISP Internet connection. Also, this IP Address is not obvious when connecting to a Web site.

Normally an address such as, http://www.johndoe-connect.com/glossary is used by the person connecting to a selected Web site. This Web site address will be directly associated with a unique IP Address. The e-mail address assigned to an individual or company is based on an agreement between the individual and the ISP. For example, the e-mail address might appear as follows: Joeblow@seafoam.com, which identifies the individual or company to the local ISP and in turn to the Internet Routers through the assigned IP Address.

The four number groups (up to 3 numbers per group) in an IP Address are called ‘octets’ because they can have values between 0 and 255, which are 28 possibilities per octet.

A home or business machine (computer) that is dialing up through a MODEM often has an IP address that is assigned by the ISP when the machine dials in. That IP address is unique for that session since it is MODEM sensitive. The IP Address may be different the next time a user machine dials in because the ISP, with multiple MODEM connections to the Internet, uses a different IP Address for each MODEM it supports, rather than for each customer.

Domain Names / Name Servers:

Because the strings of numbers that make up IP Addresses are difficult to remember and because IP Addresses sometimes change, all servers on the Internet have normal readable names, called ‘domain’ names. For example, WWW seafoam.com is a permanent normal language readable name. It is easier to remember WWW seafoam.com than it is to remember 207.124.76.13. There can be a maximum of three numbers per group between the ‘dots’, xxx.xxx.xxx.xxx.

The name WWW seafoam.com actually has three parts:

 The host name: WWW

 The domain (Web Site) name: seafoam

 The top-level domain name: com

The Domain names are managed by a company called VeriSign. VeriSign creates the top-level domain names and guarantees that all names within a top-level domain are unique. VeriSign also maintains contact information for each Web site and maintains an identification database. The company that hosts the domain (Web site) designates the host name. Since WWW is a very common host name, Web sites either omit it or replace it with a different host name that indicates a specific area of the Web site. For example, in encarta.msn.com, the domain name for Microsoft's Encarta encyclopedia is Encarta, which is designated as the host name instead of WWW.

A set of servers called Domain Name Servers (DNS) relates the normal readable names to the IP Addresses. These servers operate databases that correlate names to IP Addresses, and these servers are distributed all over the Internet. Most individual companies, ISPs and Universities, for instance, maintain small name servers to correlate host names to IP Addresses. There are also central name servers that use data supplied by VeriSign to correlate domain names to IP Addresses.

If you type the URL (Uniform Resource Locator / Universal Resource Locator) ‘http://www.groove.com/downloads/groove’ into your browser, your browser extracts the name ‘www.groove.com’ and passes it to a Domain name server, and the Domain name server returns the correct IP address for ‘www.groove.com’. A number of name servers may be involved in order to get the right IP Address.

The Internet is made up of millions of computers, each with a unique IP Address. Many of these computers are Server machines with multiple individual software servers, which imply they provide services to other computers on the Internet. Some of these Servers, among many, are the familiar ones, such as: E-mail servers, Web site servers, Gopher servers and Telnet servers.

Web page:

A Web page is a text file that contains not only text, but also a set of hypertext Markup Language (HTML) tags, which are instructions that tell the Web browser how the page should look when it is displayed. The Web browser interprets these tags to decide how to format the text onto the screen, such as change fonts, add colors, create headlines and embed graphics in a page.

The Web browser is a computer program, like Microsoft Internet Explorer or Netscape Navigator that performs two basic functions:

 A Web browser is designed to go to a Web server (site) on the Internet and request a page and transfer the page through the network into your machine.

 A Web browser can interpret the set of HTML tags within the page and display the page on your screen in the format created by the originator.

Web server:

A Web Server (Web site) is a combination of computer hardware and software that can respond to a Web browser request for a page, and deliver the page to the Web browser through the Internet. A Web server is kind of like a wall unit with many compartments (pigeonholes), with each compartment containing a Web page. These compartmentalized pages would be available for display and viewing by anyone all over the world. This is accomplished using the URL address, such as, http://www.Groove.com/downloads/more. The www.Groove selects the Web server and the /downloads selects one compartment (page) and the /more selects another different page. The / separates the URL parts and causes different pages to be displayed within the Groove.com site. Every day millions of Web servers deliver pages to millions of individual browser programs through the Internet.

Web Sites and Web pages located at the Web site are accessed through the URL (Uniform Resource Locator / Universal Resource Locator) code. An individual can set up a Web server (Web site) or contract with an established web server to install and make available to the public, web pages specified by the individual. This web server could be a service offered by the Internet Service Provider (ISP) or some other web server offering shared services. When a web page on the Internet is requested a URL, address is used and the addressed page can reside on any web server. Of course, it’s a little deeper than this simple explanation because there is a lot of coordination, feedback, design, links and statistical facts that must be considered. The Web server provider would be called your host, and the fee charged for the service is usually called a hosting charge

Clients and Servers:
In general, all of the machines (computers) on the Internet can be categorized as two types: Servers and Clients. Those machines that provide services (Web servers or FTP ‘File Transfer Protocol’ servers) to other machines are designated *servers*. And the machines that connect to those services are *clients*. When a person connects to the contracted (monthly fee) Internet Service Provider (ISP), which could be a local operation or a nationwide service, a server system is provided to service requests for information on the Internet. This server could be one machine (computer and/or an individual software system) or a cluster of very large machines.

The ISP is providing an Internet server while the individual user (client) is probably providing no services to anyone else on the Internet. Therefore, the individual’s machine is referred to as the user or client machine, which could be a large operation or a simple PC (Personal Computer). It is possible for a machine to be both a server and a client, but for discussion purposes they will be treated separately.

A server machine may provide one or more services on the Internet. For example, a server machine might have different software running on it that allows it to act as a Web server, an E-mail server and a File Transfer Protocol (FTP) server. Clients that access a server machine do so with a specific request, so client’s requests are directed to a specific software server running on the overall server machine. For example, if the client is running a Web browser, it will most likely want to communicate with the Web server on the server machine. The E-mail application will communicate with the E-mail server software, etc.

Ports

Any server machine (computer) can make its services available to the Internet using numbered ‘ports’, one for each service that is available on the server. For example, if a server machine is running a Web server and an FTP (File Transfer Protocol) server, the Web server would typically be available on port 80, and the FTP server would be available on port 21. Clients (users) can connect to a service at a specific IP (Internet Protocol) address on a specific port, although this activity or specific connection occurs without the user’s knowledge. This port assignment becomes important when designing server software or web pages.

A few examples of services and port number would include the following:

 echo 7

 daytime 13

 qotd 17 (Quote of the Day)

 ftp 21

 telnet 23

 smtp 25 (Simple Mail Transfer Protocol, E-mail send)

 time 37

 nameserver 42

 nickname 43 (Who Is)

 gopher 70

 finger 79

 WWW 80 (web server)

 POP 3, 110 (E-mail receive)

If the server machine accepts connections on a port from the outside world a person can connect to the port from anywhere on the Internet and use the offered service. Note there isn’t any rule those forces a Web server to be available on port 80. If a person were to set up their machine and software system, an unassigned port could be specified for access to the Web server. For instance, port 918 could be used, which would require the URL (Uniform Resource Locator) for the Web server to include the number 918. If the server URL were http://aaa.bbb.ccc.com then in order to connect to the server on the Internet it would be necessary to use http://aaa.bbb.ccc.com:918. The ‘:918’ specifies the port number, and would have to be included in the address in order for a person to reach the server. When no port is specified, the user’s (client) browser simply assumes that the server is using the normal port 80.

Basic Web page access:

When a user (client) wants to access a Web page residing on a distant Web server a basic sequence of events or actions will occur. This sequence of events would be initiated and then continue to follow some basic steps, as listed below. A review of Figure 16-3 (next page) may be useful for a pictorial view of these events as they occur. The ‘hash’ marks, - - - - -, represent machines (computers) and the software systems that are variable in number and location.

First, assuming the Web connection is through a local Internet Service Provider (ISP), a connection to the ISP must be accomplished. Using a dial-up line a connection is made using the normal telephone system.

The user has an URL (Uniform Resource Locator) for a Web page, which is entered in the browser address line. Let’s say the URL is http://Autodin.net/alp/mhd.htm and after it is typed in the ‘go’ button or ‘enter’ is pressed. This Web site being accessed contains historical information on the AUTODIN communications system and the above URL will bring up the page containing information on a moving head disk.

The user’s browser connects to the Autodin Web server, through the example path illustrated in Figure 3, and the requested page will be displayed on the user’s screen (monitor).

The basic steps that occurred behind the scenes to cause this to happen would be as follows:

 The user’s browser broke the URL into three parts:

1. The protocol ("http")

2. The server name (www.autodin.net)

3. The file name (web-server.htm)

 The user’s browser communicated with a ‘name server’, through the ISP, in order to translate the web server name ‘www.autodin.net’ into an IP (Internet Protocol) address, which it uses to connect to the web server machine.

 The user’s browser then connects to the web server at that IP address on port 80.

 Following the HTTP (Hypertext Transfer Protocol) protocol, the user’s browser sends a request to the web server, asking for the file "http://autodin.net/alp/mhd.htm.

 The web server then sends the HTML text for the Web page to the user’s browser. Cookies may also be sent from web server to the user’s browser in the header for the requested page.

 The user’s browser reads the HTML tags and formats the page onto the screen.

The page displayed on the screen is now available to be saved or copied for whatever purpose the user may have in mind, although not all web sites allow the page to be directly copied. The displayed page may have links to other pages and, if clicked, will cause the previous step-by-step process to be accomplished for that requested page. A ‘link’ is usually identified by an underline (\_\_\_\_\_), italics, a different color, named buttons or a list of pages available within the web server. When the mouse arrow is pointed to the link the pointer usually changes to the finger pointing hand.

Also, the speed at which the pages are downloaded from the web server to the user’s screen is dictated by the speed of the various communication links between the user and the web site. Usually the slowest bit per second (bps) rate is between the user and the local ISP when a dial up telephone line connection is used. Even though a 56.6 Kbps MODEM is used the actual transfer rate is around 14 to 50 Kbps. This reduction in the bps rate is due to noise (distortion) and electrical interference caused by the line and the various components (equipment) required for a connection between the user and the ISP. This is the reason user’s switch to the broadband digital or cable facilities to increase the bps rate and improve download time.



**Figure 3**

Internet Search Engine:

Search Engines are Web sites on the Internet that help people find information stored on World Wide Web sites (Internet). There are various differences in the way search engines work, but they all perform three basic tasks:

 They search the World Wide Web (WWW), or pieces of the WWW (Internet), in order to find individual Web Site identification information; key words that identify individual Web Sites.

 They maintain an index of the key I.D. words they find and where they found them.

 They allow users to find information based on the words or combinations of words found in the index.

The early search engines maintained an index of a few hundred thousand Web pages and documents, and received maybe two thousand inquiries a day. Today, a top search engine will index many millions of pages, and respond to many millions of queries per day.

To find information on the millions of Web pages that exist, a search engine employs special software searching tools called ‘spiders’ to construct the list of key words that identify web sites (home page(s)) and the stored pages. This list building process is referred to as ‘Web crawling’ and in order to build and maintain a useful list of words a search engine's spiders have to look at millions of pages. Initially the spider will begin with the most popular sites, indexing the key words on its pages and following every link found within the site. This is followed by a continued search of all Web sites and then continuing to search new sites as they come on line. Multiple spider systems may be used to speed up the process. Each spider system could maintain up to 300 connections to Web pages at any given time and view about 25 pages per second.

Once the spiders have accumulated information on Web pages the search engine stores the information for later use. Each commercial search engine has a different encoding formula for assigning priority weight to the information in its index. This is one of the reasons that a search for the same word (request) on different search engines will produce different lists, with the pages presented in a different order.

A couple hundred (or more) search engines exist, some are very popular and others are seldom used. Some search engines cross check information and coordinate with each other.

Comments:

Comparing the Message formats of the 1970’s to the present-day formats one can see that the earlier formats were simple and straightforward where the Internet has a multitude or varying formats. Formats vary based on what particular activity is initiated, such as sending and receiving E-mail or accessing the internal pages of a Web site. The Address continues to grow longer the deeper one delves into a Web site.

For example, refer to the URL used previously for discussion:

http://www.groove.com/downloads/more. The normal access to the Web site home page would be: http://www.groove.com, while the remainder of the Address: /downloads/more were pages one and two levels past the home page. As more page levels are accessed within a Web site the longer the address grows. Also, other miscellaneous symbols are added to the address.

When a comparison is made between the Internet and the pre 1990 computerized switching systems some similarities are present in the new and the old. The Internet Service Provider (ISP) is basically a combination of a store and forward concentrator with multiplexing features and a Front-end computer system. Local Area Networks (LAN) have been in use since the ‘server’ was invented and implemented in the 80’s. The ‘router(s)’ functions are similar to the ‘packet’ switching systems implemented in the 1960’s. The Web sites started out as intelligent terminals, which were in the mini-computer class.

Of course, today there are millions of these various system components and they are faster, have tons of memory storage (memory and disk), are more efficient, have a much greater processing capability, and so many protocols (changing daily), that it’s difficult to keep up with them. Cost of all communications equipment has decreased dramatically since the 1960’s to the point where an individual can spend less than a thousand dollars and have a computer / software system more powerful than the 2 million-dollar 1970 computer / software system.