UNIT 1
Fundamentals and Introduction to XHTML

1.1 Brief Introduction to the internet

- ✓ The internet is a collection of computers and other devices connected by equipment that allows to communicate with each other

1.1.1 Origins

- ✓ In 1960s, Advanced Research Projects Agency (ARPA) in the Department of Defense (DOD) found ARPANET to connect computers for sharing information.
- ✓ In 1967, Association for Computing Machinery (ACM) meeting, ARPA presented its ideas for ARPANET, a small network of connected computers. The idea was that each host computer would be attached to a specialized computer, called an Interface Message Processor (IMP).
- ✓ In 1969, ARPANET Connected following four nodes:
  - University of California, at Los Angles(UCLA)
  - The University of California at Santa Barbara (UCSB).
  - Stanford Research Institute(SRI)
  - University of UTA All these four nodes were connected via IMP to form a network and communication was through a software called Network Protocol.
- ✓ In 1973 Vint Cerf and Bob Khan found the concept of Transmission Control Protocol (TCP). Shorty thereafter, split TCP into two parts.
  - Transmission Control protocol (TCP)
  - Internetworking Protocol (IP).

1.1.2 What the Internet Is?

- ✓ Internet is a global system in which millions of computers are connected together. It is basically **network of network**
- ✓ Internet uses TCP/IP protocol and other common protocols for transmitting the data through various media.

1.2 Internet Protocol Addresses:

The individual host connected to Internet is uniquely identified by a 32-bit Internet protocol(IP address). Two version of the Internet Protocol(IP) are in use: IPv4(32-bit) and IPv6(128 bit). IPV4 address can be represented in the following two ways:

(1) Binary Notation

Example: 01110101 10010101 00011101 00000010

(2) Dotted-Decimal Notation

Example: 117.149.292.2

1.3 Domain Names:

1. The actual internet address of hosts on the internet is in numeric form (IP address). But it is very difficult to remember these numeric addresses. So, textual names are used to identify the hosts. These names begin with the name of the host machine, followed by progressively larger enclosing collections of machines, called **domains.**
2. There can be two, three or more domain names.
3. The first domain name, which appears immediately to the right of the hostname, is the domain of which the host is a part.
4. The second domain name gives the domain of which the first domain is part.
5. The last domain name identifies the type of organization in which the host resides, which is the largest domain in the sites name.

Example: **www.dte.kar.nic.in**

- Here **dte** is the host name
- The other parts **kar** is the local domain
- The **nic** is the domain which is part of **in**

The hostname and all of the domain names are together called a ***fully qualified domain name.***

Because IP addresses are the addresses used internally by the Internet, the fully qualified domain name of the destination for a message, which is what is given by a browser user, must be converted to an IP address before the message can be transmitted over the Internet to the destination. These conversions are done by software systems called **name servers**, which implement the Domain Name System (DNS).

Working of DNS:

Below figure shows how fully qualified domain names requested by a browser are translated into IPs before they are routed to the appropriate Web server.

All document requests from browsers are routed to the nearest name server. If the name server can convert the fully qualified domain name to an IP address, it does so. If it cannot, the name server sends the fully qualified domain name to another name server for conversion.

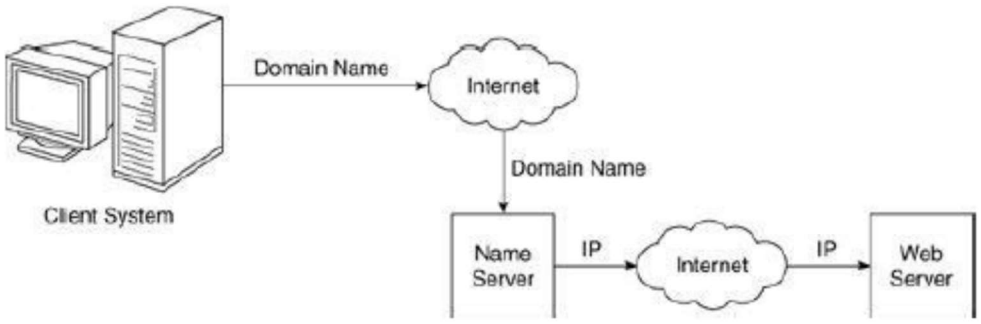


Fig: Domain Name Conversion

1.4 World Wide Web(W3):

The World Wide Web("WWW" or simply the "Web") is a global information space to which people can read and write via computers connected to the Internet.

1.4.1 Origin of WWW: Tim Berners-Lee at CERN proposed web in 1989.

Document form: hypertext

Hypermedia: images, sounds etc.

1.4.2 Web or Internet: The ***Internet*** is a collection of computers and other devices connected by equipment that allows them to communicate with each other. The ***Web*** is a collection of software and protocols that has been installed on most, if not all, of the computers on the Internet. Some of these computers run ***Web servers***, which provide

documents, but most run ***Web clients***, or browsers, which request documents from servers and display them to users..

1.5 WEB BROWSERS:

When two computers communicate over some network, in many cases one acts as a **client** and the other as a **serve**r. The client initiates the communication, which is often a request for information stored on the server, which then sends that information back to the client.

So, Web browsers are used to send request to the Web server for a particular static web page.

Web Browser sends the requests to web server in the form of URL(Uniform Resource Locator) using HTTP(Hypertext Transfer Protocol). The web server considers the URL and is in turn used to locate a particular file and the server sends it back to the browser.

- ✓ Types of Browsers
  1. Internet Explorer
  2. Google Chrome
  3. Mozilla Firebox
  4. Opera
  5. UC Browser
  6. Netscape Navigator
  7. Apple Safari

1.6 WEB SERVERS*:*

Web servers are programs that provide documents to requesting browsers. Servers are slave programs: They act only when requests are made to them by browsers running on other computers on the Internet.

Two leading Web Server are **Apache**, the most widely-installed Web server, and Microsoft's **Internet Information server(IIS).**

1.6.1 Web Server Operations:

- ➢ Web browsers initiate network communications with servers by sending them URLs. A URL can specify one of two different things: the address of a data file stored on the

server that is to be sent to the client, or a program stored on the server that the client wants executed, with the output of the program returned to the client.

- All the communications between a Web client and a Web server use the standard Web protocol, Hypertext Transfer Protocol (HTTP)
- When a Web server begins execution, it informs the operating system under which it is running that it is now ready to accept incoming network connections through a specific port on the machine.
- A Web client, or browser, opens a network connection to a Web server, sends information requests and possibly data to the server, receives information from the server, and closes the connection.
- The primary task of a Web server is to monitor a communications port on its host machine, accept HTTP commands through that port, and perform the operations specified by the commands
- All HTTP commands include a URL, which includes the specification of a host server machine. When the URL is received, it is translated into either a file name (in which case the file is returned to the requesting client) or a program name (in which case the program is run and its output is sent to the requesting client).

1.6.2 General Server Characteristics:

The file structure of a Web server has *two separate directories*.

1) **Document root:** This directory stores all the ***web documents*** to which the server has direct access and normally serves to clients.
2) **Server Root:**This directory, along with its descendant directories(such as *Conf,Logs, Cgi-bin subdirectories)* , stores the server and its support software
3) Normally server stores all documents outside the document root because its readable to its client. This is called as **Virtual Document tree.**
4) Some of the servers allow to access the web documents that are in the document root of the other machines such servers are called as **proxy servers**

Suppose the site name is www.gptbijapur.edu and the document root is name ***tutorials,*** and is stored in the ***/admin/web*** directory. So, /admin/web/tutorials is the document's directory address.

If a request URL is:

http://www.gptbijapur.edu/XML/xhtml.html

the server will search for the file with the given path

/admin/web/tutorials/XML/xhtml.html

Many servers allow part of the servable document collection to be stored outside the directory at the document root. The secondary areas from which documents can be served are called ***virtual document trees.***

1.6.3 Apache:

- ✓ Apache is an open-source and free web server software. The official name is **Apache HTTP Server** and it's maintained and developed by the **Apache Software Foundation**
- ✓ Apache is most widely used web server because it is **excellent server in terms of fast and reliable.**
- ✓ Apache is a **Unix-based systems** after that ported to windows and other network operating system.
- ✓ Apache also supports plug-in modules for extensibility
- ✓ Apache server includes the ability to support multiple programming languages, server-side scripting, authentication mechanism and database support

1.6.4 IIS –Internet Information Services:

- ✓ IIS server is most popular for **windows platform** and it's been developed by the Microsoft.
- ✓ IIS is used to host ASP.NET web applications and static websites. It can also be used as an FTP server, host WCF services, and be extended to host web applications built on other platforms such as PHP

1.7 Uniform Resource Locators(URLS):

Uniform (or universal) resource locators (URLs) are used to identify ***documents (resources***) on the Internet. There are many different kinds of resources, identified by different forms of URLs

1.7.1 URL format:

All URLs have the same general format: **scheme:object-address**

scheme: It refers to communication protocols such as: ***http, ftp, gopher, telnet, file, mailto, and news.***

For HTTP protocol, the object-address is in the following form:

http: **//fully-qualified-domain-name/path-to-document**

For ***File*** protocol, the object-address is in the following form:

file://path-to-document

The host name is the name of the server computer that stores the document (or provides access to it on some other computer). Messages to a host machine must be directed to the appropriate process running on the host for handling. Such processes are identified by their associated port numbers. The default port number of Web server processes is 80.

1.7.2 URL Paths:

The path to the document for the HTTP protocol is similar to a path to a file or directory in the file system of an operating system and is given by a sequence of directory names and a file name, all separated by whatever separator character the operating system uses. For UNIX servers, the path is specified with forward slashes;for Windows servers, it is specified with backward slashes.

Example: **http://www.dte.kar.nic.in/files/f99/store.html**

A path that includes all directories along the way is called a ***complete path.*** In most cases, the path to the document is relative to some ***base path*** that is specified in the configuration files of the server. Such paths are called ***partial paths***.

For example, if the server's configuration specifies that the root directory for files it can serve is **files/f99**, the above URL is specified as follows:

http://www.dte.kar.nic.in/store.html

If the specified document is a ***directory r***ather than a single document, the directory's name is followed immediately by a slash, as in the following:

http://www.dte.kar.nic.in/department/

1.8 MULTIPURPOSE INTERNET MAIL EXTENSIONS:

A browser need some way of determining the format of a document it receives from a web server. Without knowing the form of a document, the browser would be unable to render it. The forms of these documents are specified with the ***Multipurpose Internet Mail Extensions(MIME***).

1.8.1 Type Specifications:

MIME was developed to allow different kinds of documents(text, video, or sound data) to be sent using Internet mail

MIME specifications have the following form

Type/subtype

The most common MIME types are as follows

Type	Subtype	Description
Text	Plain Enriched	Unformatted text; may be ASCII or ISO 8859 Provides greater format Flexibility
Image	JPEG Gif	The image is in JPEG format, JFIF encoding The image is in GIF format
Video	Mpeg	MPEG format
Audio	Basic	Single-channel 8-bit ISDN mu-law encoding at sample rate of 8kHz

1.8.2 Experimental Document Type:

Experimental subtypes are sometimes used. The name of an experimental subtype begins with x-.

Ex: video/x-msvideo

1.9 HYPERTEXT TRANSFER PROTOCOL:

- ✓ HTTP is a communication protocol used between the web client (browser) and web server over the internet.

- ✓ HTTP protocol is follows the **request-response model**. The client makes a request for desired web page and server gives the response to the client.

HTTP consists of two phases: The request and the response. Each HTTP communication between a browser and a web server consists of two parts, a header and a body. The header contains information about the communication; the body contains data of the communication, if there is any.

1.9.1 THE REQUEST:

When a Web browser requests a web page, it sends a request message to a web Server. The message always includes a header and sometimes it also includes a body.
The header contains meta-information-information about the message and the body contains the content of the message.

The general form of HTTP request is as follows:

1. HTTP method domain part of URL HTTP version
2. Header fields
3. Blank line
4. Message body

1. HTTP Methods OR <Start Line>

- The <start line> consists of three parts which are separated by a single space.
- It has 3 parts
    1. Request Method
    2. Request URI
    3. HTTP Version
- **Example**

```
GET    /store.html    HTTP/1.1
```

- The following are the request methods are used to do the communication between web client and web server.

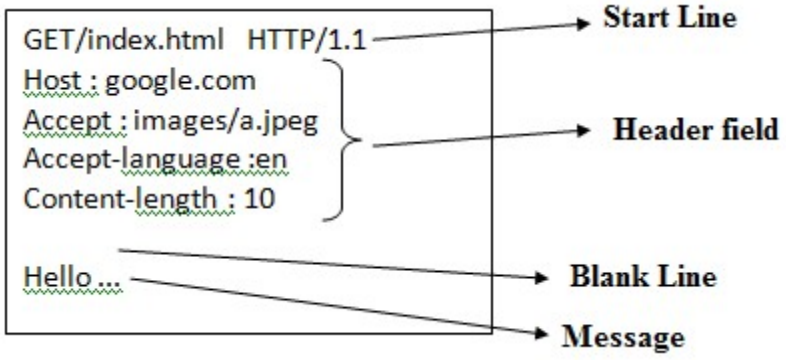
HTTP Method	Description
GET	Returns the contents of the specified document

POST	Executes the specified document, using the enclosed data
HEAD	Returns the header information for the specified document
PUT	Replaces the specified document with the enclosed data
DELETE	Deletes the specified document

- **Request URI →** The Uniform Resource Identifier is used to identify the names or resources on the internet.
- **HTTP Version →** It indicates the version of the HTTP protocol. The official version of HTTP is HTTP/1.1

2. **Header Fields**
   - This field is associated with the HTTP request. The header field contains field name and field value.
3. **Blank Line**
   - It indicates the space between the header fields and data.
4. **Message Body**
   - Actual data transmits between the web client and web server

✓ **Example**



1.9.2 THE RESPONSE PHASE:

✓ The general form of an HTTP response is

1. Status Line
2. Response Header fields
3. Blank Line
4. Response body

- ✓ The **status line** includes the HTTP Version and three digits status code for response and short textual explanation of the status code.
- ✓ **Example**

HTTP/1.1 200 OK

- ✓ The following are the HTTP status codes

First digit	Category
1XX	Informational
2XX	Success
3XX	Redirection
4XX	Client Error
5XX	Server Error

- ✓ Example

```
HTTP/1.1      Document follows
MIME-version: 1.0
Server: CERN/3.0
Date: Sundat, 10-Nov-96 06:54:46 GMT
Content-Type: text/html
Content-Length: 4531
Last-Modified: Wednesday , 16-Oct-96  10:14:01  GMT
 <head>
 <Title> Sample  Document </title>
</head>
<h1>  Sample Document </h1>
</body>
</html>
```

1.10 SECURITY:

- ✓ Web security is the protection of information, that can accessed between web server and web client.
- ✓ There are four security issues
  1. **Privacy**
  2. **Authentication**
  3. **Non-repudiation**
  4. **Integrity**
- ✓ Consider the simplest case like transmitting a **credit card number to a company from which a purchase is being made**. The security **issues** for this transaction are as follows:
  1. **Privacy** →It must not be possible for the credit card number to be stolen on its way to the company's server.

2. **Integrity→** It must not be possible for the credit card number to be modified on its way to the company's server.
3. **Authentication→** It must be possible for both the purchaser and the seller to be certain of each other's identity.
4. **Nonrepudiation→** It must be possible to prove legally that the message was actually sent and received.

The basic tool to support confidentiality and integrity is encryption. It can conventional encryption(With single secret key) or Public key encryption(using private key and public key).

1.11 WEB PROGRAMMER's TOOL BOX:

- ✓ There are some tools used for programming the web applications. These tools are nothing but programming languages and markup/scripting languages.
- ✓ Web programs and scripts are divided into two categories
  1. Client-side scripting language
     - ✓ Example : XHTML, XML, JavaScript, VB script, Ajax
  2. Server-side scripting language
     - ✓ Example : PHP, Ruby, Perl etc

1.11.1 Tools for creating HTML documents

XHTML documents can be created with a general-purpose text editor. There are two kinds of tools that can simplify this task: **XHTML editors** and what-you-see-iswhat-you-get (**WYSIWYG,** pronounced wizzy-wig) XHTML editors.

Two examples of WYSIWYG XHTML editors are **Microsoft FrontPage** and **Adobe Dreamweaver**

1.11.2 Plug ins and filters

Plug ins

Two different kinds of converters can be used to create XHTML documents. Plug-ins are programs that can be integrated together with a word processor. Plug-ins add new capabilities to the word processor, such as toolbar buttons and menu elements that provide convenient ways to insert XHTML into the document being

Filters:

Filter converts an existing document in some form, such as LaTeX or Microsoft Word, to XHTML. The disadvantage of filters is that creating XHTML documents with a filter is a two-step process: First you create the document, and then you use a filter to convert it to XHTML

The two advantages of both **plug-ins and filters**:

1) Existing documents produced with word processors can be easily converted to XHTML
2) Users can use a word processor with which they are familiar to produce XHTML documents

1.11.3 XML

- A meta-markup language
- Used to create a new markup language for a particular purpose or area
- Because the tags are designed for a specific area, they can be meaningful
- No presentation details
- A simple and universal way of representing data of any textual kind

1.11.4 JavaScript

- A client-side HTML-embedded scripting language
- Only related to Java through syntax
- Dynamically typed and not object-oriented
- Provides a way to access elements of HTML documents and dynamically change them

1.11.5 PHP

- A server-side scripting language
- An alternative to CGI
- Similar to JavaScript
- Great for form processing and database access through the Web

1.11.6 Overview of Ruby: Ruby is a scripting language designed by Yukihiro Matsumoto, also known as Matz. Ruby is a pure object oriented programming language and runs on different platforms.

1.11.7 Overview of Rails: Rails is a web application development framework that increases the speed and ease of web development. A framework is a set of software tools and libraries that make it easier to create web applications.

1.11.8 Overview of Ajax: Ajax stands for Asynchronous Javascript And XML. Ajax is widely used to create faster, user-friendly, interactive, dynamic and better web applications. The main purpose of using AJAX is to enable web applications to extract data from various servers in an asynchronous manner.

Differences between HTML and XHTML

XHMTL	HTML
XHTML uses an XML syntax	HTML uses a pseudo-SGML(Standard Generalized Markup Language)
XHTML elements must always be closed	HTML syntax permits some elements to be unclosed.
XHTML is case-sensitive	Case-insensitive
XHTML elements must be properly nested	It can be or can not be nested
All Attributes value must be double quoted	Attributes value may or may not be quoted
Attributes value must be explicitly specified	Attributes values are implicit
For referring elements, It is better to use “id” attribute	It uses ‘name’ attribute for referring elements
All elements and attribute names must be in lowercase	All elements and attributes names may be in lowercase or uppercase