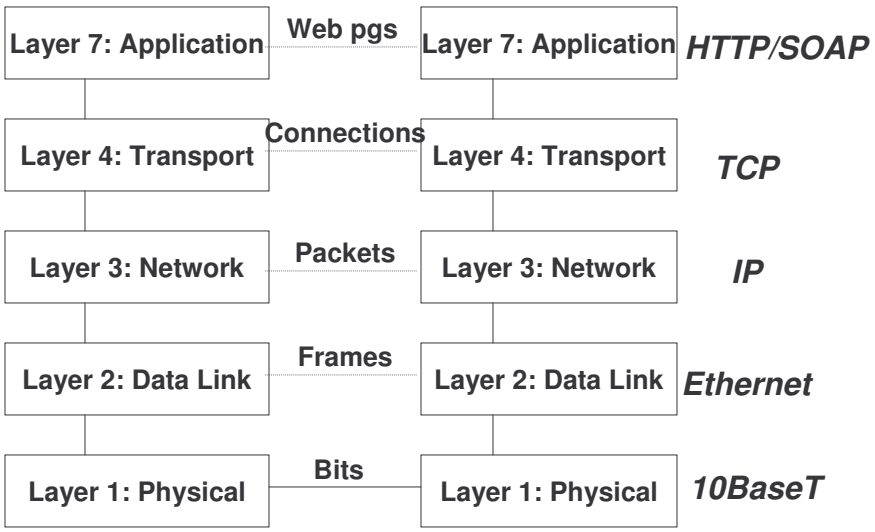
1.264 Lecture 9

Web fundamentals

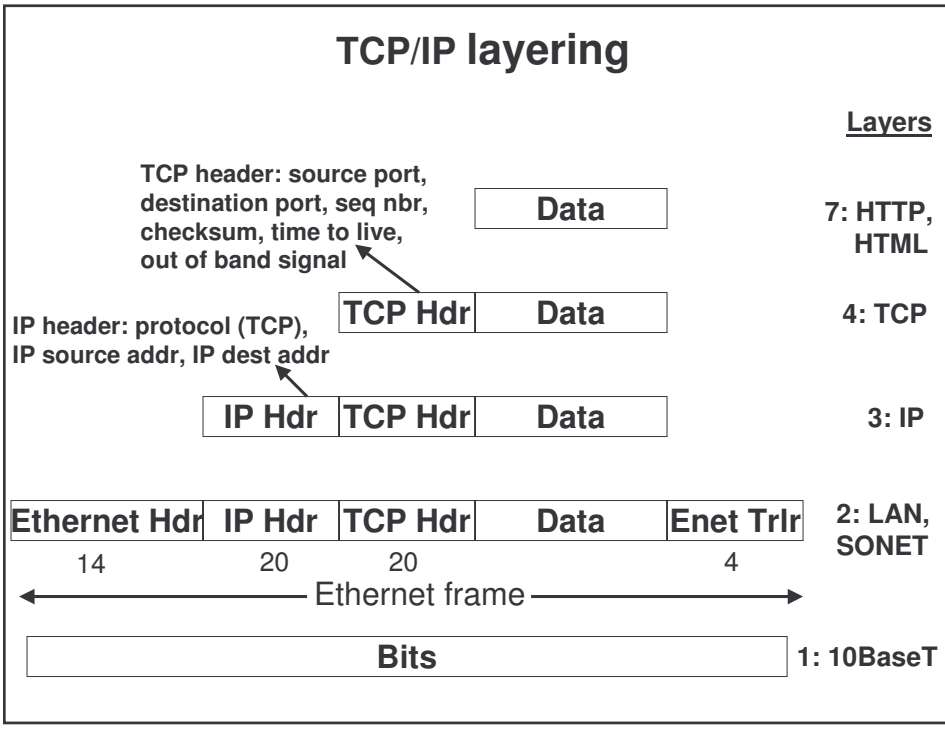
Web browsers and servers

- Internet is "just" a set of loosely interconnected networks
  - A set of local area networks connected via wide area networks
  - No centralized control or management
  - Network segments interconnect via routers
  - Routers are dedicated computers that manage packets of data
  - TCP/IP is the universal data protocol on the network
  - Actual format, content is left to higher-level protocols, like the Web
- TCP/IP connections
  - Client is typically a data consumer that sends short requests
    - On Web, client is a browser
  - Server is typically a data provider that sends long responses
    - Listen for requests and transmit desired data (often on port 80)
    - Can send static or dynamic data
  - Web connection is active only long enough to exchange data
    - Avoid overhead of many communication channels, but lose state

Transmission Control Protocol/Internet Protocol TCP/IP



TCP/IP layering



Web organization: Web pages

- Web page is basic unit of organization
  - Simple pages retrieved in single operation
  - Compound pages (typically text and graphics) take multiple passes (multiple TCP/IP connections unless 'Keep-Alive is set)
  - Pages have hypertext
    - Links to other pages, implemented via embedded URLs
  - Pages are described using Hypertext Markup Language (html)
    - High level document description language
    - Specifies structure but not appearance of document
    - Defines sections, such as "level 1 header", "list", "emphasized text"
    - Browser handles rendering of page on client machine
  - HTML may be replaced by XHTML in the future
    - Subset of XML, more structured, verifiable than HTML

Web organization: MIME types

- Every document on Web has a type
  - Multipurpose Internet Mail Extensions (MIME) types, in http standard
    - Some types, such as html, text and some graphics, are displayed directly by browser
    - Others need helper or plug-in external programs to display type
  - Web type system is extensible. New types are easy to accommodate.
    - Define new MIME type in server
    - Browser will prompt for helper application if MIME type unrecognized
    - Browser can suggest Web site to download helper app
    - Download and install helper app, and use the new MIME type
  - Web server extensibility: programs to generate documents
    - URLs can point at programs as well as pages
    - Programs can be simple (time of day) or complex (database, analysis)
    - Java, perl, C++, Java Server Pages, Visual Basic, Active Server Pages used

MIME type examples

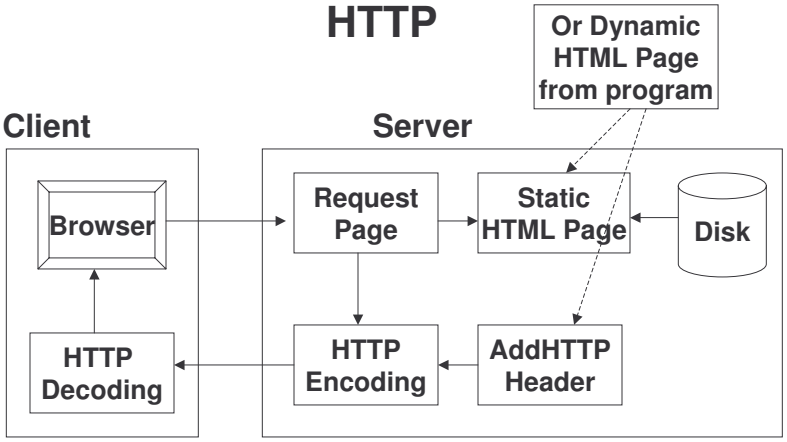
• application/msword	Word
• application/pdf	Acrobat
• application/vnd.ms-excel	Excel
• application/zip	Zip file
• audio/basic	.au, .snd
• audio/x-wav	MS audio
• image/gif	GIF
• image/jpeg	JPEG
• text/plain	Plain text
• text/html	HTML
• video/mpeg	Video

See http://www.rfc-editor.org for current list of HTTP, MIME, other Internet specifications

HTTP protocol

- **HTTP 1.1 is the current protocol**
  - **Based on ISO Latin1 alphabet (ASCII with extensions for European languages)**
  - **HTTP 1.1 adds persistent tcp/ip connection, partial document transfers over HTTP 1.0**
- **Four phases:**
  - **Open connection: Based on URL**
  - **Request: Browser opens connection to server and sends:**
    - **Request method, (and request data at bottom if POST or PUT request)**
    - **URL,**
    - **HTTP version number**
    - **Header information (informational, optional), terminated with blank line**
  - **Response: Server processes request and sends:**
    - **HTTP protocol version and status code**
    - **Header information, terminated by blank line**
    - **Text (data)**
  - **Close connection**

HTTP



- These transactions are stateless. The connection is closed after each page and re-established: Server can't connect successive requests from the client
- Dynamic content is produced in response to HTTP request by a server program.

HTTP request examples

Typical browser request: (Unix telnet web.mit.edu 80 to see this)

```
GET /about-mit.html HTTP/1.1
Host: web.mit.edu                    (required)
Accept: text/html, text/plain, image/jpeg, image/gif, */*    (optional)
                                     (blank line)
```

Typical server response:

```
HTTP/1.1 200 OK
Server: Apache/1.3.3 Ben-SSL/1.28 (Unix)
Content-Type: text/html
Content-Length: 8300
                                     (blank line)
<HTML>
<HEAD><TITLE>About MIT</TITLE></HEAD>
<BODY>The mission of MIT…</BODY>
</HTML>
```

Also try:
fedex.com
amazon.com

HTTP request phase (browser to server)

Command	Description
GET	Requests specified document (used to 'post' too)
HEAD	Requests only header of specified document
POST	Requests that server accept data from browser and generate dynamic content
OPTIONS	Get server and access options
TRACE	Used in debugging
PUT	Replace server document with data from browser
DELETE	Delete specified document on server

- GET is the most frequent request from browsers
- POST used for processing fill-out forms
- HEAD is used by search engines to check for live pages
- No security or authentication in HTTP ('Basic' sends password in clear).
  - Use Secure Socket Layer (SSL) to encrypt your Web exchanges

Common HTTP request headers

Header	Definition
Accept	MIME types accepted by client (multiple allowed)
Connection	Connection type for client (keep-alive, close)
Cookie	Return previous cookie to server (data exchange)
From	Email address of user (sent only by crawlers)
Host	Original host requested (forward, multiple names)
If-Modified-Since	Used to reduce fetching of docs browser has
Referer	URL of last document client displayed
User-Agent	Name and version of client software (browser)

HTTP response status codes

Status code	Definition
100-101	Informational response; client should respond with some other action (continue, new protocol)
200-206	Request was successful
300-307	Document has moved; indicate new address
400-417	Client error, such as unauthorized request
500-505	Server error

Examples:	
200	OK
404	Not found
500	Internal server error

HTTP response header examples

Header	Definition
Allow	Requests allowed, such as GET (400s)
Content-Length	Length in bytes of data to be returned (used for 'keep alive' connections)
Content-Type	MIME type of returned data
Expires	Date at which document expires
Last-Modified	Date at which document was last modified
Location	New document address (with 300 status)
Set-Cookie	Gives browser a 'cookie'

HTML

- Subset of Structured Generalized Markup Language (SGML), a document description language
  - SGML is ISO standard
- Current version of HTML is version 4.01
  - See www.w3.org for latest drafts, standards, discussions
  - New features in HTML 4.0 standard
    - Style sheets, internationalization, access for disabled, frames, improved tables and forms

Hypertext markup language (html) example

```
<HTML>
<HEAD>
<TITLE> Welcome to the Aircraft Parts System </TITLE>
</HEAD>
<BODY>
<H1> Welcome to the Aircraft Parts System </H1>
This system handles orders for aircraft and balloon parts.  We
    handle aircraft parts, and parts for all balloon manufacturers.
    We comply with the latest US regulations.
<P>
The use of this system is subject to <A HREF= MITRule.html>
    MIT rules and regulations. </A>
</BODY>
</HTML>
```

(Aircraft1.html)

HTML

- Tags (e.g. <TAG>) never display but direct the browser
- Often in pairs (e.g. <TAG> and </TAG>) to delimit section
- Some tags have attributes (e.g. <A HREF=abc.html> </A>)
- HTML document begins with <HTML>, ends with </HTML>
  - Two sections within document: HEAD and BODY
  - Head has identifying information not displayed
  - Body is displayed, with formatting:
    - Paragraph <P>
    - Header levels 1 through 6 <H1> through <H6>
    - Line break or carriage return
    - Anchor <A>, placed around text or graphics; used for hyperlinks

Tag examples

- Control appearance of page
  - Not as precise as MS Word or other editors
  - Intended for pages that can be viewed on machines with very different graphics capabilities. Tags discourage specific assumptions.

Opening tag	Closing tag	Definition
<EM>	</EM>	Emphasis (often italic)
<STRONG>	</STRONG>	Strong emphasis (often bold)
<VAR>	</VAR>	Variable
<CITE>	</CITE<	Citation (of an article or book)
<ADDRESS>	</ADDRESS>	Address (identifying info at bottom of doc)
<CODE>	</CODE>	Computer code fragment (tags ignored)
<I>	</I>	Italic (as well as other *physical* styles)

(Aircraft2.html

Tag examples, cont

- Font size and color: <FONT>, <FONT COLOR= name>
- Special characters: > is >
- Paragraph format: <H2 ALIGN=CENTER> text </H2>
- Preformatted text: <PRE> (lines up columns exactly, etc.)
- Horizontal lines: <HR WIDTH=80% ALIGN= LEFT>
- Lists: can be multi-level, etc.
  - Ordered (numbered) <OL>
  - Unordered (bulleted) <UL>
  - Definition (indented) <DL>
- Tables: <TABLE>, <CAPTION>, <TR> (row), etc.

(Aircraft3.html)

HTML Filenames

- Relative names (UNIX conventions)
  - MITRule1.html — same folder (directory) as current doc
  - Laws/MITRule2.html — in subfolder (subdirectory) of current doc
  - ../MITRule3.html — in folder above current doc
- Absolute names (also UNIX)
  - /MITRule4.html — in root folder (Web root or local root)
- Document sections
  - MITRule1.html#sec1 — points to target tag (sec1) in same doc
    - Sections defined by <A NAME=sec1> </A>
- Documents on other sites
  - <A HREF=http://www.mit.edu/index.html>

(Next page: Aircraft4.html)

Images

```
<H1> Welcome to the Aircraft Distributor system </H1>

<A HREF=aircraft.gif>
<IMG SRC=thumbnail_aircraft.gif ALT="Aircraft parts
   services">
</A>
Welcome to the Aircraft Parts System
```

- This shows a small picture of the aircraft parts but allows the user to download the large picture
- The HREF can point to other documents about aircraft, carriers, etc.

(Aircraft4.html)

Frames

- Subdivide main browser window into several panels
  - Developed by Netscape, supported by Microsoft IE also
  - Framesets assign initial URLs to each of the panels
- Example of 2 equal, side-by-side panels

```
<HTML><HEAD>
<TITLE> Two frames side by side </TITLE>
</HEAD>
<FRAMESET COLS=50%,50%>
   <FRAME SRC=Aircraft5.html NAME=graphics>
   <FRAME SRC= MITRule.html NAME=text>
</FRAMESET>
<NOFRAME>
   This document contains frames.  You need a frame-enabled
   browser to view it.
</NOFRAME>
</HTML>
```

(Aircraft5.html)

HTTP and HTML

- **HTTP**
  - **Is only direct form of interaction between browser and server**
  - **Was an extremely perceptive extension of email, ftp protocols by Tim Berners-Lee to enable Web browsers**
  - **Request-response paradigm**
  - **Connection made for each request/response pair**
  - **Most popular protocol on Internet, fairly stable definition**
- **HTML**
  - **Text description language, based on tags**
  - **High level description rather than specific formatting**
  - **HTML is static, which is not sufficient for Web applications**
  - **Many extensions and changes (scripting, Java) for dynamic content**
  - **HTML has forms, which are the basis for submitting data to databases, covered next time. XML is also used for data exchange, along with HTTP extensions, also covered next**