The Architecture of the World Wide Web

Nancy McCracken
NPAC
College of Engineering and Computer Science
Syracuse University
111 College Place
Syracuse NY 13244-4100

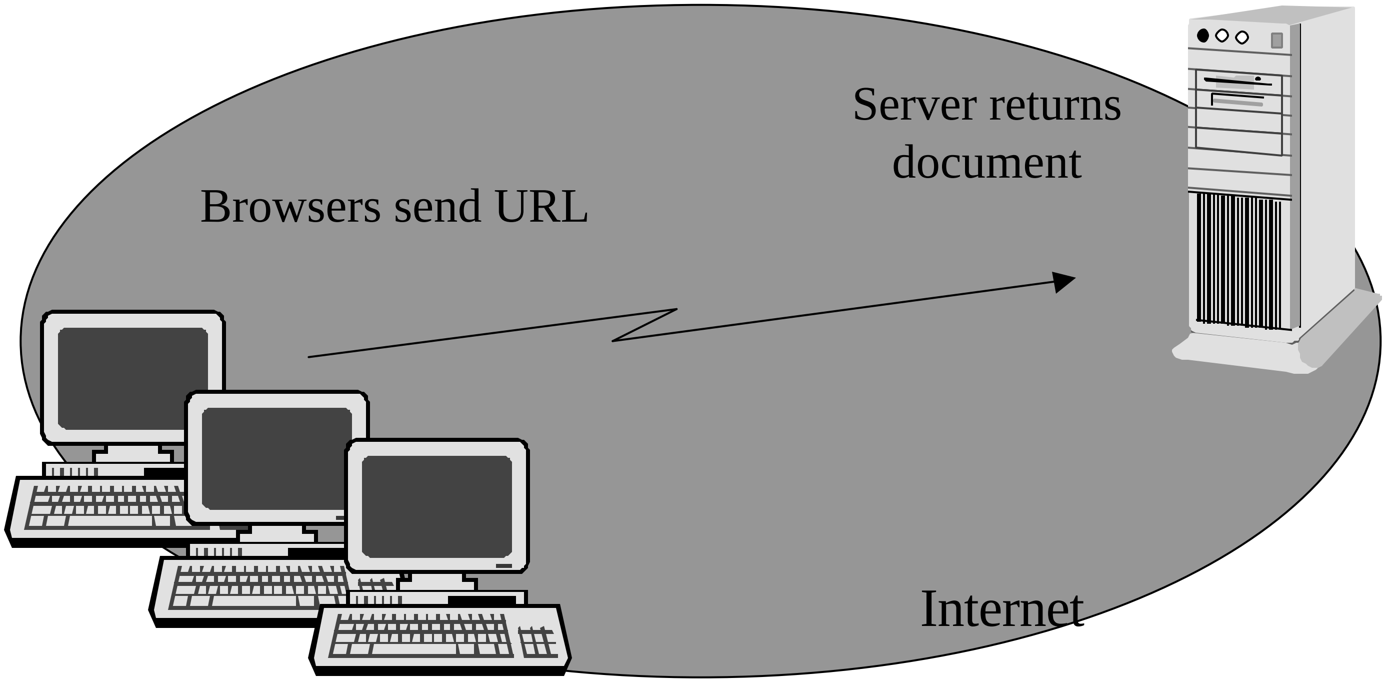
September 2, 1998

The Architecture of the World Wide Web

- The World Wide Web (WWW) (the Web) is a hyperlinked collection of documents and programs that reside on computers all over the world, linked by the Internet.
- This talk will show the underlying components and mechanisms that make the Web work.
  - Network protocols based on TCP/ IP and a common Domain Name Service
  - Message-passing protocols based on MIME
  - Web Server architecture based on the HTTP protocol
- This works on a world-wide basis is because these protocols are based on Open Standards which have been implemented by many vendors on a variety of machines. The Web software structure is strictly non-proprietary, while allowing proprietary pieces to fit in where needed.
- The same architecture and software that makes the Web work is also suitable for implementing distributed applications between hetereogeneous machines and networks. This makes the architecture attractive for the corporate Intranet as well.

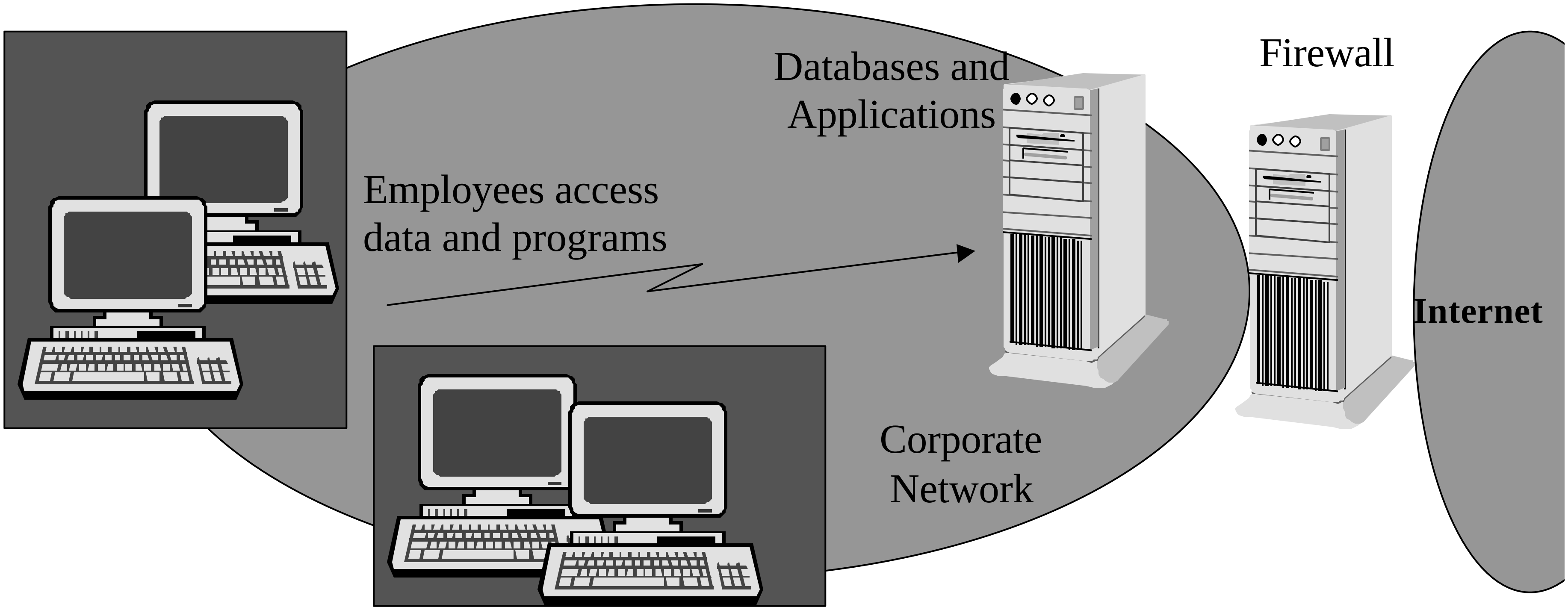
Top-level View of the World Wide Web

- All over the world, users can use browsers to access information stored in multimedia document collections of web server machines. Programs are also accessible through the Common Gateway Interface (CGI).



Top-level View of the Corporate Intranet

- All over the company, employees (and possibly affiliates and the public) can use browsers to access databases and use distributed applications stored on server machines, using web technology to interface to existing databases and applications.


Databases and Applications
Firewall
Employees access data and programs
Internet
Corporate Network

Networking Basics

The first section of this talk covers basic networking terminology, the OSI networking layers, the TCP/ IP protocol, and routing.

Background on the Internet

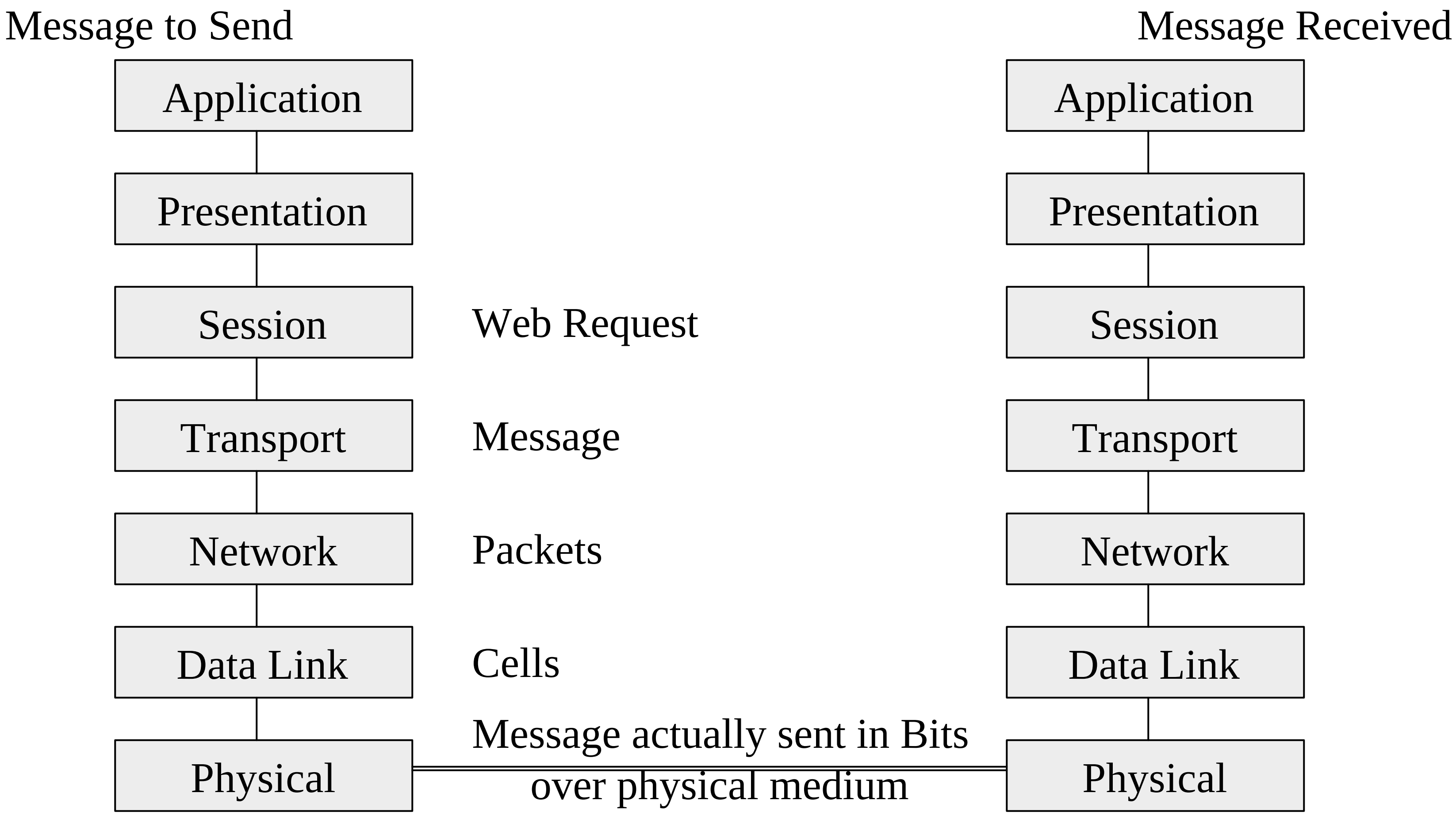
- The Internet is a loose federation of networks.
- Cooperative organization - no administration, no fees. Protocols and standards are evolved through the IETF, Internet Engineering Task Force.
- Most national and international networks are members: NSFNET, ESNET, ARPANET, BITNET
- All these networks are packet switched systems based on TCP/ IP. Together these protocols allow for communication over a wide variety of technologies. Machines called gateways connect the networks.
- Standard domain name system - names are looked up by name server to obtain routing information.
  - symbolic names: npac.syr.edu
  - internet addresses: 128.230.7.2

Networking Basic Definitions

- A computer network is a communication system for connecting end-systems usually called hosts.
- A local area network, LAN, connects computer systems within a few kilometers, usually within a single building. A common technology is Ethernet, which operates at 10Mbps (million bits per second). Computers or workstations connect to the LAN via an interface card.
- A wide area network, WAN, connects computers in different cities or countries. A common technology is leased telephone lines operating between 9600 bps and 1.544 Mbps.
- Computers in a network use a set of protocols to communicate.

Networking Standards: OSI Layers

- Network communication protocols are usually described via a set of layering conventions from the International Standards Organization (ISO) known as the Open Systems Interconnection (OSI) Model.



Message to Send

Message Received

Application

Presentation

Session

Transport

Network

Data Link

Physical

Web Request

Message

Packets

Cells

Message actually sent in Bits over physical medium

Application

Presentation

Session

Transport

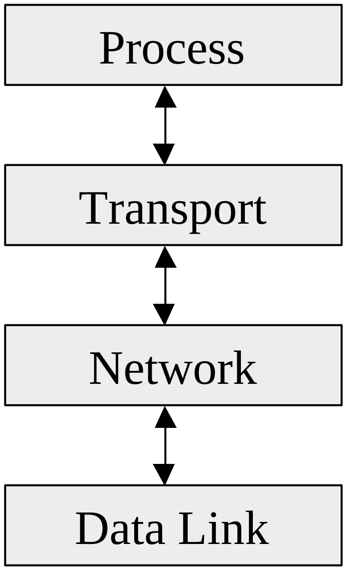
Network

Data Link

Physical

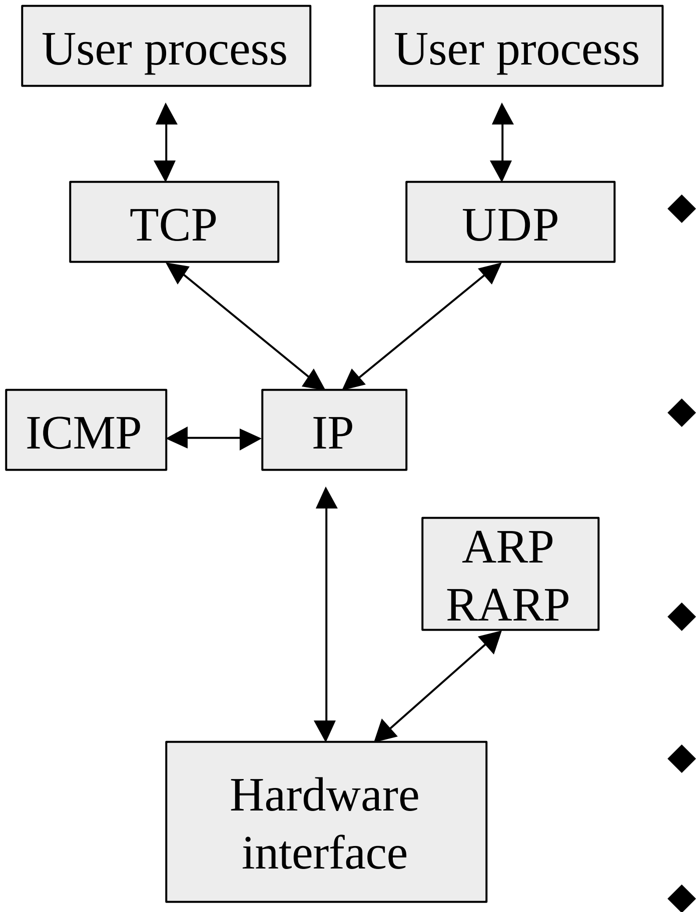
Simplified communication protocol model

- We simplify the model to the four lowest software layers - user applications use the process layer and the remaining three are usually included in the operating system, such as Unix, which has an OSI stack to process messages through the layers.



Process	User applications
Transport	Host-to-host layer: may establish virtual circuit for messages
Network	Determines route to transmit packets
Data Link	Transforms raw messages into one that appears free of transmission errors, transmits data cells and processes acknowledgement cells.

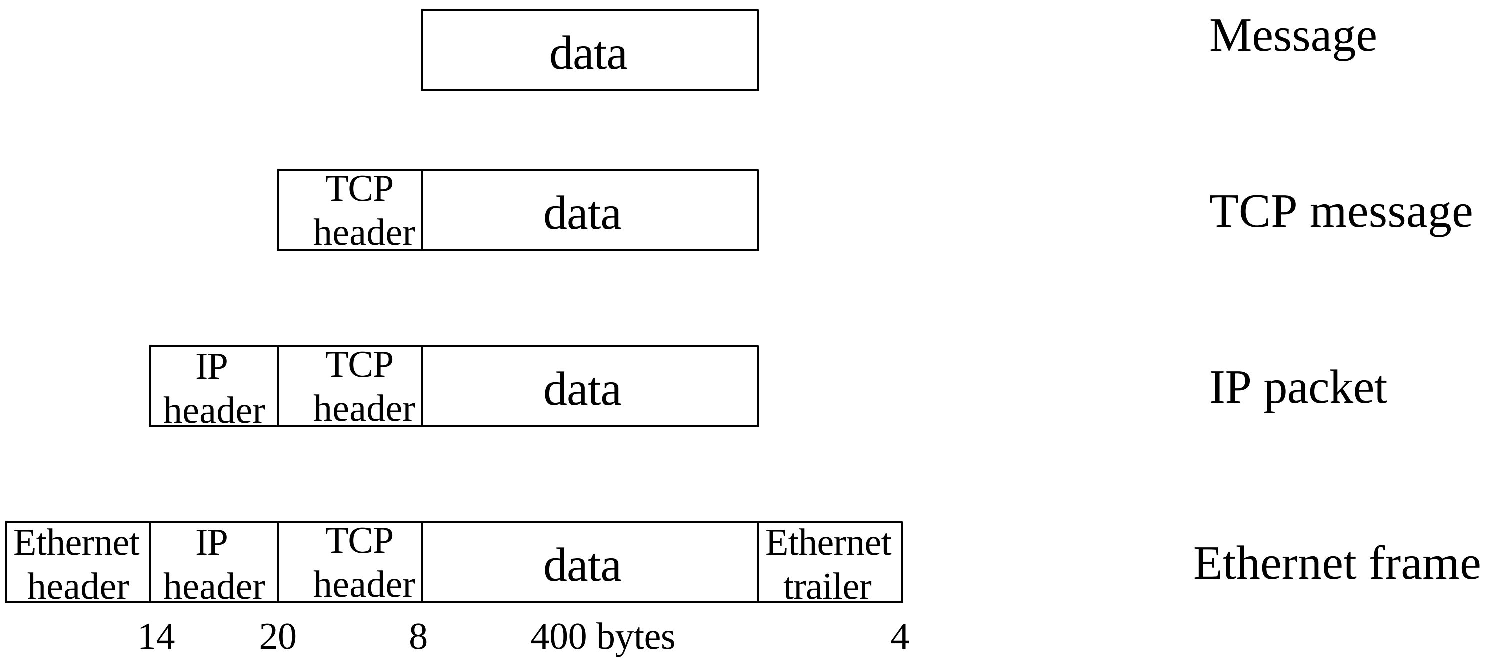
The TCP/ IP protocol suite



- TCP - Transmission Control Protocol. A connection-oriented protocol used by most Internet applications to provide a reliable, full-duplex, byte stream for a user process.
- UDP - User Datagram Protocol. A connectionless protocol for user processes. Also not reliable.
- ICMP - Internet Control Message Protocol. Handles error and control information between gateways and hosts.
- IP - Internet Protocol. Provides the packet delivery service for the upper layers.
- ARP - Address Resolution Protocol. Maps an Internet address into a hardware address.
- RARP - Reverse Address Resolution Protocol.

Typical message formats

- Each layer adds control information to the message - this process is called encapsulation.



data

Message

TCP header

data

TCP message

IP header

TCP header

data

IP packet

Ethernet header

IP header

TCP header

data

Ethernet trailer

Ethernet frame

14 20 8 400 bytes 4

Networking

- The Internet is a packet-switched network. Each message (or document) is broken up into a number of packets. Each packet has an address. A computer called a router sits on the local network and decides where to send it first on its way to its final address. Each computer along the network connection examines messages that come in and either keeps it or reroutes it along its way. The message is reassembled on the other end.

Communications Issues

- Multiplexing - Different protocols can be used to send different messages through the same network.
- Sequencing is the property that data is received by the receiver in the same order as transmitted by the sender, which is not true in a packet-switched network.
- Error control guarantees that error-free data is received by the application programs. Data can either get corrupted by the transmission medium or get lost. Checksums are added to the data and received data is acknowledged. If there is any problem, retransmission occurs.
- Flow control assures that the sender doesn¹t overwhelm the receiver by sending data at a faster rate than it can process.
- Error and flow control are handled on an end-to-end basis by TCP and on a hop-by-hop basis by IP. (A hop goes to only one intermediate machine on the network route.)

Networking Speeds

- Performance of network delivery depends on the size of the message, the capacity of the various pieces of network that the message may travel along and the congestion of the network.

Network Speed		Email	book	picture	audio	video
		(2.2KB)	(240KB)	(300KB)	(475KB)	(2.4MB - 1min)
Modem	14,400bps	1.22sec	2.22min	2.78min	4.40min	22.2min
ISDN	56,000bps	0.31sec	34.3sec	42.9sec	1.13min	5.71min
T1	1.54Mbps	.011sec	1.24sec	1.55sec	2.46sec	12.4sec
T3	45.0Mbps	.0004sec	0.04sec	0.05sec	0.08sec	0.42sec

Internet 2

- The current demand for applications involving transfer of multimedia and real time events has resulted in additional protocols for the Internet currently under development as the Integrated Services model
- RSVP - Reservation Protocol - a virtual circuit is established that can reserve a certain bandwidth for continuous transmission of packets.
  - Quality of service
  - Multicasting
- RTP and RTSP - Real Time Protocol and, more specificately, Real Time Streaming Protocol - for continuous multimedia transmission
- These protocols require network hosts which support them to save state regarding the virtual circuits.

Open Standards

All the network protocols just discussed are agreed on by various standards committees. The principal standards organization of the Internet is the Internet Engineering Task Force (IETF). The principal standards organization of the WWW is the World Wide Web Consortium (W3C).

Internet Documents: Drafts, Memos and Standards

- Some material presented here comes from Internet documents. Here is a summary of various document formats you may find.
- Internet Drafts
  - Working documents of the (IETF), its Area and Working Groups.
  - Other groups may also distribute Internet Drafts.
  - Some of these IDs are labelled by IETF-#.
  - IDs are valid for a maximum of 6 months and may be updated, replaced or made obsolete by other documents at any time.
- Internet Memos
  - Referred to as RFC-# (Request for Comments)
  - More formal and complete than Internet Drafts, usually represent standard proposals/ candidates.
  - Some RFCs become obsolete by subsequent RFCs, some others make it as standards
- Internet Standards
  - Labelled by STD-# and often associated with the RFC-# specs (e.g. Internet E-Mail is referred to as FRC-822 or STD-11)

Internet Documents - Examples

- Here are a few sample Internet documents relevant for Internet and WWW message-passing.
- RFC-822: Crocker, D., "Standard for the Format of ARPA Internet Text Messages", SRD 11, RFC 822, UDEL, 1982.
- RFC-1521: Borenstein, N. and Freed, N., "MIME (Multipurpose Internet Mail Extension) Part One: Mechanisms for Specifying and Describing the Format of Internet Message Bodies", RFC 1521, Bellcore, September 1993.
- RFC-1524: Borenstein, N. "A User Agent Configuration Mechanism for Multimedia Mail Format Information", RFC 1524, Bellcore, September 1993.
- Internet Draft: Tim Berners-Lee, "Basic HTTP", CERN, 1992/ 3.
- RFP-1890: H. Schulzrinne, RTP Profile for Audio and Video Conferences with Minimal Control, Jan. 1996.

Message-passing Protocols

Internet E-Mail (RFC-822)

- We all know and use it, but here is a formal specification.
- Each message is a stream of 7-bit ASCII chars which contains a header and optional (newline separated) body.
- Header consists of a set of entries with one entry per line given by a colon separated key:value pair.
- Key contains no spaces or tabs and cannot exceed 63 chars.
- Body is a fully unstructured sequence of ASCII chars.
- There is a finite set of standard keys and an extension mechanism via the "X"-prefix. The standard set (as used by MH) is:

◆ Date	Bcc	Resent-Date	Resent-Fcc
◆ From	Fcc	Resent-From	resent-
◆ Sender	Message-ID	Resent-To	Message-Id
◆ To	Subject	Resent-cc	Forwarded
◆ cc	In-Reply-To	Resent-Bcc	Replied

Multi-purpose Internet Mail Extension (MIME)

- Goals
  - Multimedia, multi-language, multi-component extension of RFC-822
  - Full backward compatibility with RFC-822
  - Open design to incorporate multiple well-known formats
  - Easy extension to new types and formats
- Retain RFC-822 header+body format
- Add new header fields
- Allow for multipart multimedia bodies
- Include media type and encoding information in new header fields such as: Content-Type, Content-Description, Content-Transfer-Encoding, Content-ID
- Retain 7-bit ASCII for all valid encoding schemes
- Implement multi-component bodies via a special 'magic type' Content-Type: multipart

MIME - "Content-Type" Header Field

- Two level hierarchical typing scheme adopted of the form: basetype/ subtype
- Seven base media types are defined this minimal set is enforced, i.e. all extensions must pass the whole ID->RFC->STD process.
- Allow for less restrictive subtyping the base types, for example:
  - Content-Type: text/ plain
  - Content-Type: text/ richtext
- Some standard subtypes are specified and many more are expected. New subtypes must be registered with the IANA (Internet Assigned Numbers Authority).
- Private experimental subtypes prefixed with "X-" may be used freely and without registration.
- Seven base types are: text, image, audio, video, multipart, message, application.

MIME - Base Content Types

- text
  - subtypes: plain (just ASCII) and richtext (a simple markup extension including <bold>, <italic> etc. tags)
  - character sets can be further specified in the header value field as follows:
    - » Content-type: text/ plain; charset=us-ascii
  - Other charsets can be used to support other languages such as iso-8859-1 (French) or iso-2022-JP (Japanese). These charsets need to be encoded in one of two encoding modes: base64 or quoted-printable. The latter retains ASCII subset and is more natural for non-ASCII extensions.
- image
  - Standard subtypes: gif, jpeg. Others expected.
- audio
  - Standard subtype: single-channel 8KHz u-law. Others expected.
- video
  - Standard subtype: mpeg. Others plausible.

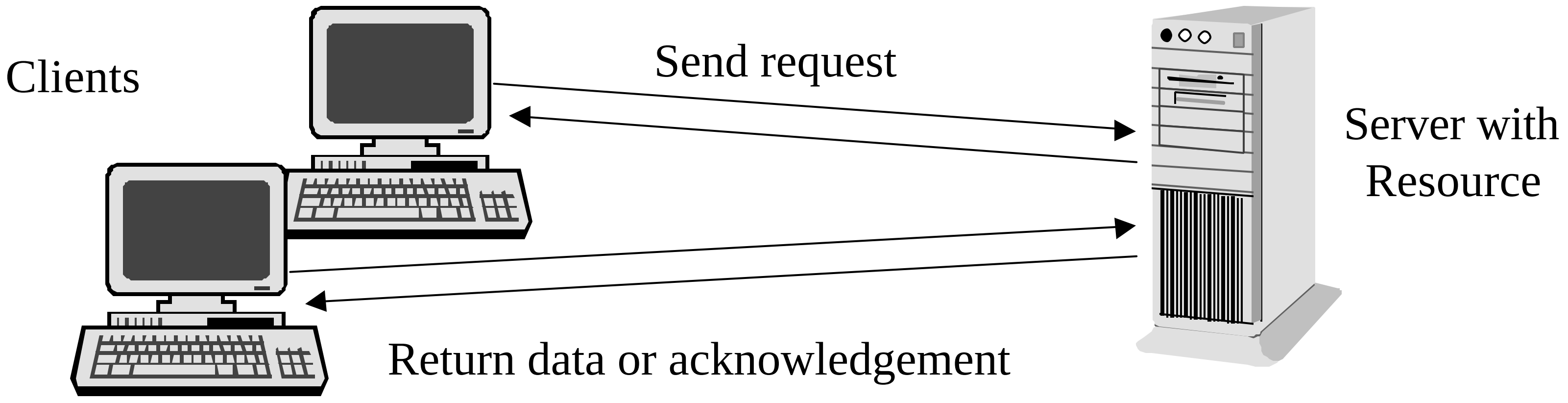
MIME - Base Content Types, continued

- multipart
  - Specifies a MIME message composed of several parts with possible different Content-Type fields.
  - Parts are separated by a boundary string, specified in the multipart header entry
  - Subtypes: mixed (serial combination of media), parallel (for parallel presentation if possible), alternative (multiple representations of the same data) and digest (all parts are messages)
- message
  - Subtypes: rfc822 (standard ARPA e-mail format), partial (a single chunk of a larger message, chopped into pieces for transmission and then reassembled), external-body (pointer to a remote data - similar to typerlink/ URL but different representation)
- application
  - Current subtypes: postscript, ODA
  - Placeholder for "anything else" - several interactive/ custom/ creative extensions expected here
  - Already registered: Andrew-inset,t ATOMICMAIL (Bellcore)

Web Services - HTTP Protocol

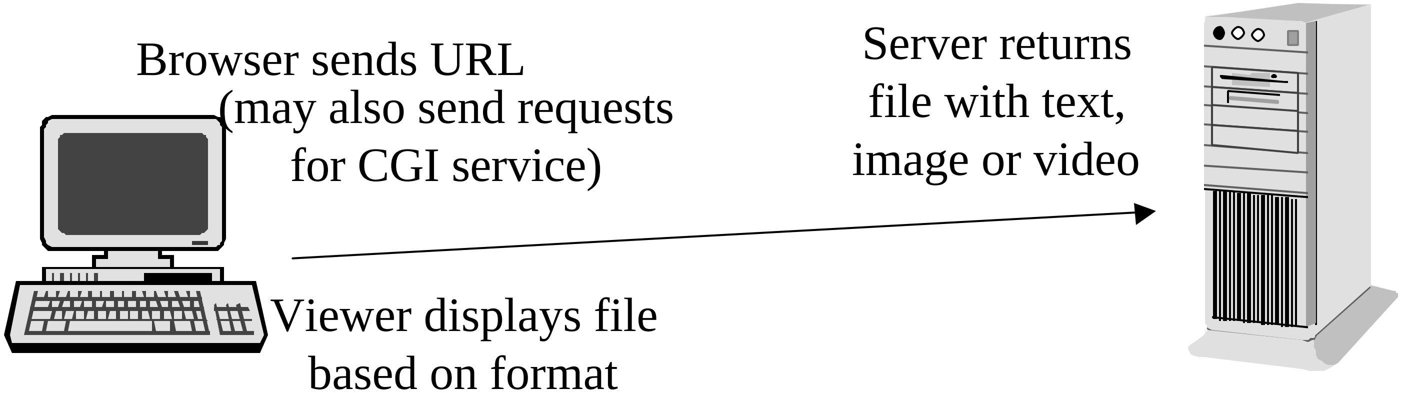
Applications based on information services typically use a Client/ Server Architecture

- Server: A program in charge of a resource or information.
  - Operation is defined by list of services.
  - Normal mode is to listen for requests, stopping to fulfill a request when it arrives.
- Client: Any program that makes a request for service from the server.



The World Wide Web is a collection of clients and servers called browsers and Web sites

- Web servers provide access to a collection of files containing hyperlinked information
  - primary service is to send text files, images, digitized video
  - can also provide customized services through the form/ CGI script interface
- Browsers provide an easy graphical interface for users to request information. The client machine also provides viewers for a standard set of image and video formats.
- The interface is kept very simple to run on all networks and most machines.



Browser sends URL (may also send requests for CGI service)

Server returns file with text, image or video

May also use CGI interface to access other services

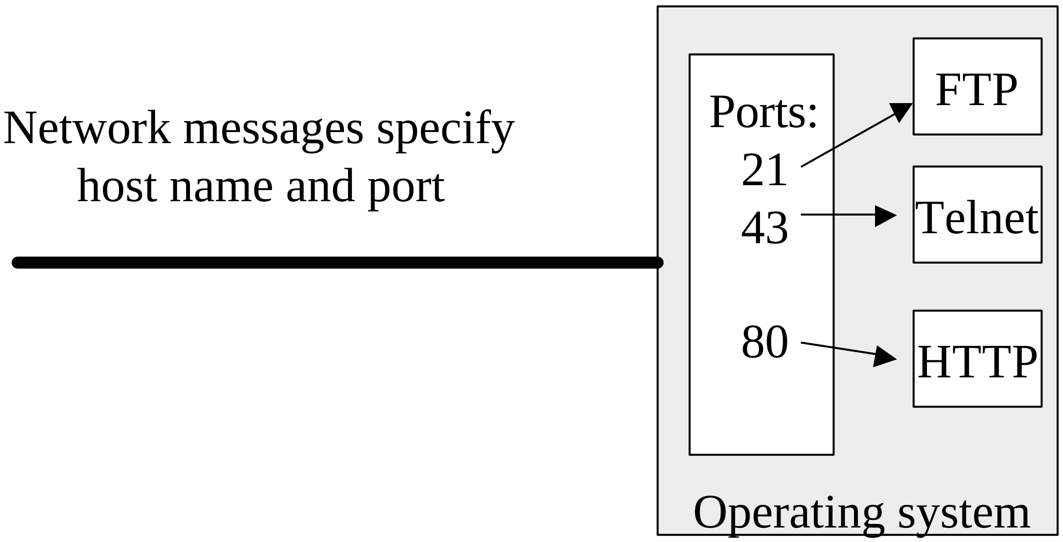
Viewer displays file based on format

HTTP - Hypertext Transport Protocol

- HTTP provides an upper level to the Internet, that is, it is built on top of a back-bone network with all the packets flowing from client to server and vice versa using the standard TCP/ IP protocol.
- It uses MIME formats and concepts, but does not fully conform to MIME as the WWW is not a mail system.
- HTTP protocol is compatible with other network services such as FTP (File Transfer Protocol), NNTP (Network News Transport Protocol).
  - On a UNIX-based machine, the basic services are enumerated in the file / etc/ services. Each service cooresponds to a standard port. For example, telnet is mapped to port 43, and FTP is mapped to port 21. All ports below 1024 are privileged - only the system administrator can determine port use.
- The HTTP service is standardly assigned to port 80 - it provides a much shorter service connection than the other services.

HTTPD - HTTP Daemon

- The HTTP daemon is the server which responds to the Internet service requests on standard port 80 (or on another custom port). The server program is available from NCSA and is easily installed by editing a set of configuration files which give directory locations for documents, cgi scripts, error messages and icons, and which allows for options regarding path names, domain access, and so on.



URL - Uniform Resource Locator

- A URL has the standard form
  - service:/ / machine:port/ file.file-extension
- HTML hyperlinks typically use the service http for linking to other documents and media files. Some other internet services can also be used such as
  - ftp:/ / machine/ file.file-extension.
- In this way, a Web server can provide other Internet services through the browser interface.
- The machine is an Internet address and can either be a symbolic name provided by the Domain Name Service (DNS) or the IP numbers.
- If the port is not specified, it defaults to 80.
- The file.file-extension is given by any Unix path name starting from the directory known to the server as "document root". Which path names are valid is one of the options of the server - whether "public_html" is automatically put into the path name and whether paths starting with "~username" are allowed.
- In the http service, the file-extension is used to tell the browser what helper application to use to view the file. Typical file extensions are html, gif, jpeg, mpeg, au, ram, etc.

Web Links can go to other Internet Services

- For other services, the Web server transfers the connection to the appropriate server.

Protocol	**URL identifier**	**Example**
E-mail	mailto:	mailto:njm@npac.syr.edu
FTP	ftp://	ftp://ftp.npac.syr.edu
Telnet	telnet://	telnet://gamera.syr.edu
Usenet News (NNTP)	news:	news:comp.infosystems.www
WWW(HTTP)	http://	http://www.npac.syr.edu

HTTP - How does it work?

- On each hyperlink click, the browser (client) initiates a connection with the server at the "machine" (e.g. using UNIX BSD connect call on the default port 80, or a custom user-defined port)
- A request is sent to the server, formatted as a MIME-like message.
- The server replies with another MIME-like message which is received by the browser and either formatted in the browser window or viewed with a helper application.
- The connection is closed on both sides. (The exception to this is the "server push" connection.)

HTTP - GET Request Example

- GET / document.html HTTP/ 1.0
  Accept: www/ source
  Accept: text/ html
  Accept: image/ gif
  User-Agent: Lynx/ 2.2 libww/ 2.14
  From: mnotulli@ukonaix.cc.ukans.edu
  -- blank-line-terminating-the-request --
- First line syntax is always: METHOD URL ProtocolVersion
- The following lines form a header of an (extended) MIME message
- "User-Agent" specifies the browser type
- "Accept" specifies MIME types recognized by the browser
- The server is expected to provide the requested data in one of these acceptable formats.

HTTP - Reply Example

- HTTP/ 1.0 200 OK
  Date: Wednesday, 02-Feb-95 23:04:12 GMT
  Server: NCSA/ 1.1
  MIME-version: 1.0
  Last-modified: Monday, 15-Nov-94 23:33:16 GMT
  Content-type: text/ html
  Content-length: 2345
     ---- blank-line-separating-header-and-body--<HTML><HEAD>
  <TITLE> Document Title </ TITLE>. . .
- This message contains both header and body
- Some replies contain only header (e.g. error reports, such as HTTP/ 1.0 404 Not Found)
- GET request also contained header only, whereas POST request (see next example) contains both header and body

HTTP - POST Request Example

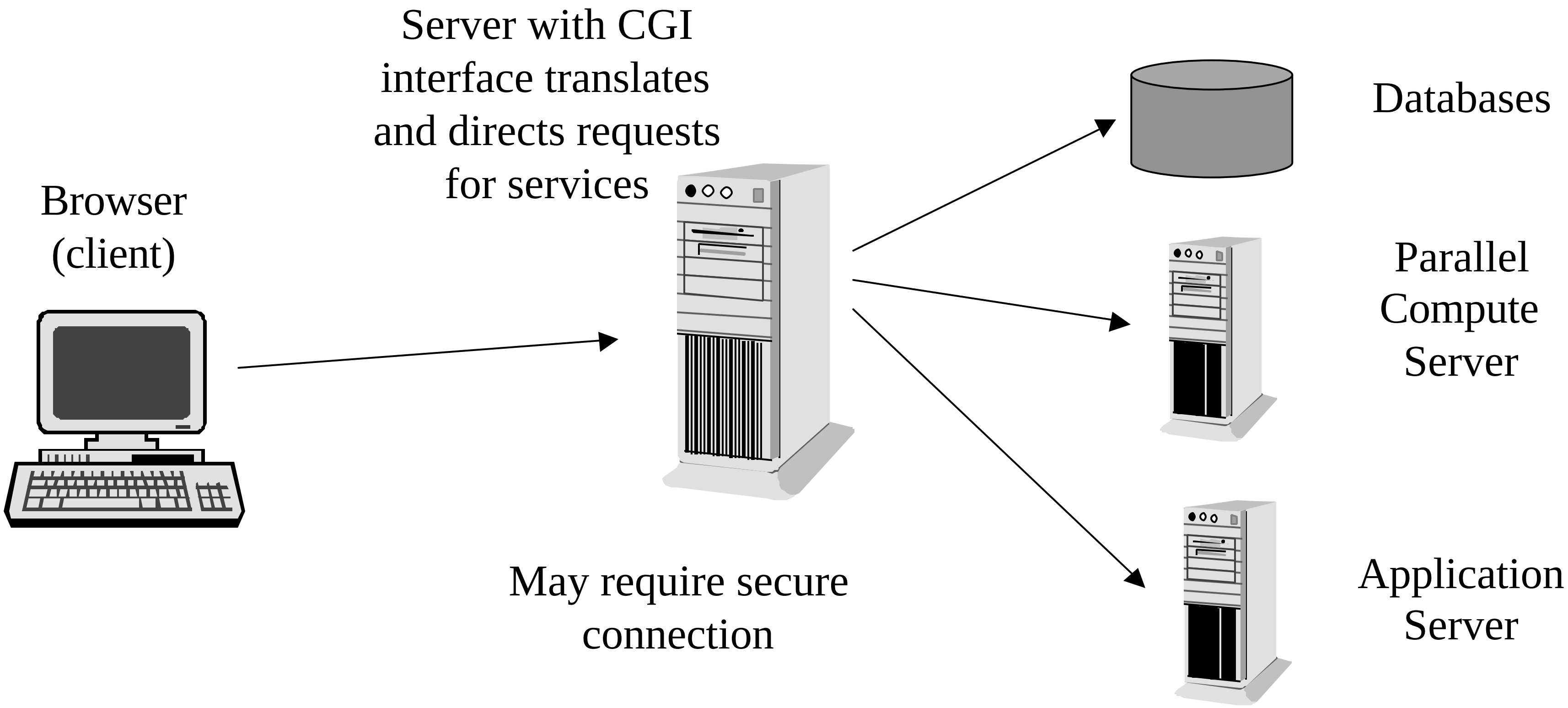
-
  ```
  POST / cgi-bin/ post-query HTTP/ 1.0
  Accept:  www/ sourceAccept:  text/ html
  Accept:  video/ mpeg
  Accept:  image/ x-rgb
  Accept:  application/ postscript
  User-Agent:  Lynx/ 2.2 libwww/ 2.14
  From:  grobe@unanaix.cc.ukans.edu
  Content-type:  application/ x-www-form-urlencoded
  Content-length:  150
          --blank-line-separating-header-and-body---
  org=Academic%20Computing%20Services
  &users=10000
  &browser=lynx
  &contact=Michael%20Grobe%20grobe@kuhbuh.cc.ukans.edu
  ```
- Both header and body present in POST requests - the body is typically used to pass a form contents to the server.

Common Gateway Interface (CGI) - an introduction

- CGI is an interface for running programs on the server at the request of the client.
- When the user clicks on a CGI link, the server calls the corresponding process and returns its output, not the data/ file/ code associated with the process.
- Typical Applications
  - Support for dynamic generation of HTML documents, such as on-the-fly conversions from other formats.
  - Interface to and integration with Forms/ GUI area of HTML - submitted forms are handled by suitable CGI processes.
  - Interfacing with other (non-HTTP) remote services such as databases, video-on-demand, simulations, etc.
    - This is current area of major development of the web.
- Look at a simple example of an HTML form with its CGI Perl program.

Three-Tier Web Architecture

- For most major web applications, the web server layer is just the "middleware" to support access to applications.



Server with CGI interface translates and directs requests for services

Browser (client)

Databases

Parallel Compute Server

Application Server

May require secure connection