Introduction

Managing Data on the World-Wide Web (236969)

The Internet and the Web
- The Internet (i.e., Inter-Network) is a network of networks
- The World-Wide Web is a collection of hypertext pages (HTML) available on the Internet
  - The Web is an application built on top of the Internet
  - Email, Telnet and FTP are some other applications built on top of the Internet

The Internet
- The main building block is TCP/IP
  - IP – The Internet Protocol
  - TCP – The transmission Control Protocol
- Many applications are built on top of TCP
  - Email, HTTP, Telnet, FTP, ...
  - And applications over IP
  - Streaming video, VOIP, ...

A computer connected to the Internet is called a host

The World-Wide Web
- The main building blocks (initially):
  - HTML and its variants (XHTML, DHTML)
  - HTTP
  - Web servers, Proxy servers, Browsers
  - Not just browsing HTML pages anymore
  - Web services
  - Semantic Web
  - Many new formats and technologies

HTML
- HTML stands for Hyper Text Markup Language
- An HTML file is a text file containing small markup tags
  - The tags tell the web browser how to structure the text and how to present it

Examples

```html
<html>
<head>
</head>
<body>
Hello world.
</body>
</html>
```
History

For a history of the Internet and the World-Wide Web, look at
http://www.isoc.org/internet/history/
A map of ARPANET in 1980
http://mappa.mundi.net/maps/maps_001/

Maps of the Arpanet (1980)

The Information Revolution

• Moving bits instead of atoms
  • Much faster
  • Much cheaper
• The world has become
  • More competitive?
  • More intimate?
  • More rapid?
  • More homogeneous?
  • More heterogeneous?
  • ...

Communication Networks

Measuring the Performance of Communication Networks

• Latency
  • Measures how long it takes to get the first bit
  • Equivalently, it is the cost (i.e., time) of sending a minimum-size message
• Bandwidth
  • Number of bits per time unit (second)

Improving the Performance

• Reduce latency
• Increase bandwidth
• It is harder to decrease the latency than to increase the bandwidth
• Usually, latency is the more important factor
  (see It's the Latency, Stupid)
• Send a jet full of DVDs from Tel-Aviv to NY – great bandwidth but lousy latency

What is the latency of an ordinary phone system?
The Effect of Latency

- Consider a 4-round protocol between a client in Israel and a server on the east cost of the USA:
  - connection request : agree reply : resource request : resource delivery
- The distance is approximately 9,200 km
- The speed of light is approximately 300,000 km/s
- It takes about 31 milliseconds for each round and at least 124 ms for the first bit of the requested resource to arrive to the client
- How does this affect "real-time applications"? (commerce, biddings, online games, ...)

Mbs vs. MBs

- Bandwidth is measured in terms of mega (kilo, giga) bits per seconds
  - Bits and not bytes
  - Divide by 8 to get the number of bytes per second
  - For example, using a 3 Mbs ADSL line, you can download a file at a rate of about 384 KBs

Local Area Network (LAN)

- A LAN connects computers by means of a particular communication protocol, such as
  - Ethernet
  - FDDI
  - Token Ring
  - ATM
- A LAN implements
  - The physical layer, i.e., translation of bits into electrical (or optical) signals and vice-versa
  - The data-link layer, i.e., one of the protocols on the left

Packets are sent using physical addresses, known as MAC (Media Access Control) addresses

Internetworking

- How different LANs can be connected together?
- Each LAN may use a different communication protocol
- Each host (i.e., computer) knows only about its own LAN
  - and can only send messages to other hosts on the same LAN

Sending Messages Across the Internet – The problems

- No central control or management
- Heterogeneous hardware and software
  - In particular, LANs use a variety of communication protocols
- Must share resources to reduce latency
  - In a phone system, one has to wait indefinitely if the line is busy
  - Call waiting reduces latency, but is not good enough for computer networks
  - In a computer network, many processes should share the resources concurrently
The Solution – Packet Switching

- Break a long message into many short datagrams
- Send each datagram independently
- Different datagrams of the same message need not follow the same route from the source to the destination
- The transmission, on the same data link, of datagrams from different messages can be interleaved

Circuit Switching vs. Packet Switching

- Traditional phone systems are based on circuit switching
- How long will it last?

TCP/IP

- IP is the basis of internetworking
- It implements the network layer
- IP is capable of sending IP datagrams (IP packets) between two hosts (i.e., computers) that are either on the same LAN or on different LANs, each located anywhere in the world

Sending an IP Datagram Between Hosts

- If the hosts are on the same LAN, one only has to implement IP on top of the data-link layer (e.g., Ethernet, ATM, etc.)
- If the hosts are on different LANs, the IP datagram must be routed between the LANs
- When an IP datagram leaves the origin host, it does not know which route will lead it to its destination host

IP Addresses

- Each host on the Internet has a unique IP address
- A datagram specifies the IP address of the destination host
- An IP address has 32 bits and is usually written as a sequence of four integers separated by dots, e.g., 132.68.32.237
- Each integer is between 0 and 255
Subnet Mask

- A prefix consisting of the leftmost $n$ ($n \geq 8$) bits of an IP address determines the network (i.e., LAN) address
- The remaining bits determine the host address on that particular LAN
- Each host must know the value of $n$ for its own LAN
- The value of $n$ is given by the subnet mask

Subnetting

- All IP addresses that start with 132.68. are assigned to the Technion
- By choosing some $n > 16$, the Technion can divide its range of IP addresses into many LANs
- $n$ need not be the same for all LANs at Technion
- However, it is more complicated to divide a range of IP addresses into subnets if $n$ varies

Routing Messages Between LANs

- A router is a device that is connected to several LANs
- It has several IP addresses, one in each LAN
- If a host needs to send an IP datagram to another host that is on a different LAN, then it actually sends the datagram to a router that is connected to its own LAN

Hop-By-Hop Routing

- Each router sends the IP datagram to another router
- The two routers must be connected by a data link
- Eventually, the IP datagram gets to the LAN of the destination host
- IP routing does not guarantee delivery

Summary of IP

- IP routes datagrams across the Internet
- It implements the network layer
- It is connectionless, that is, datagrams are sent without first establishing connection with the destination
- It is unreliable
  - Packets may get out of order, garbled, duplicated
  - May not get there at all!

Transmission Control Protocol (TCP)

- TCP is implemented on top of IP
- TCP implements the transport layer
- In the origin host, TCP breaks a long message into a sequence of IP datagrams
- TCP uses IP to send the datagrams
- In the destination host, TCP assembles the datagrams together to generate the original message
Properties of TCP

- **Connection-Oriented**
  - First, it creates a connection (3-way handshake); hence, it has a slow start

- **Reliable**
  - TCP checks for errors and resends datagrams that are lost or garbled

- **Byte Stream**
  - It assembles datagrams in the right order, even if they don't arrive in that order; hence, it looks like a stream of bytes between two hosts

- **Flow Control**
  - Prevents congestion (i.e., exceeding network or destination-host capacity)

How is TCP/IP Being USED?

- When two windows (or tabs) of a browser present resources from the same host (server):
  - How is it guaranteed that the IP packets will reach the correct window?

Routers

- LAN switches are connected to routers (usually) by means of fiber optics
- Routers route IP packets across LANs
- A router is connected directly to two or more LANs and it can transmit IP packets between these LANs (local routing)
- Some routers are connected to each other via WANs (Wide-Area Networks) and do backbone routing

Hop-by-Hop Routing

- Suppose that an IP packet is sent from a LAN to another far-away LAN
- The message gets to the router that is directly connected to the source LAN
- The router sends it to the next hop, i.e.,
  - A router on the same LAN that is also connected to some other LANs, or
  - A router on the same WAN

Routing Tables

- Each router has routing table with prefixes of IP address
  - Each prefix has a router address for the router that handles that prefix
- Given an IP packet with some IP address, the next-hop router is determined by matching the longest prefix (of an IP address) from the routing table with the given IP address
- There is also (at least one) default entry that leads to a router on the backbone of the Internet
Updating the Routing Tables

- A routing table includes local information provided by the local network administrator.
- Routers periodically update their routing tables by exchanging information with their neighboring routers.
- Routing protocols: Distance Vector (Bellman-Ford), Open Shortest Path First (OSPF).

Hostnames, Domain Names and URLs

Hostnames and Domain Names

- In addition to an IP address, a host may also have a human-readable **hostname**.
- Some examples of hostnames:
  - www.cs.technion.ac.il
  - www.cnn.com
  - csd.cs.technion.ac.il
- The first part is the name of a particular **host** (i.e., computer).
- The rest is the **domain name**.

The Hierarchical Structure of Hostnames

- Example: www.cs.technion.ac.il
  - www is a name of a computer
  - That computer is in the CS Department
  - That dept. is at The Technion
  - That university is an Academic Campus (ac) in Israel (il)
- The rightmost name, il, is the main domain.
- As we move left, the sub-domains are more specific.

The First 7 Generic Domains

- **com** - commercial organizations (www.cocacola.com)
- **edu** - educational institutions (www.berkeley.edu)
- **gov** - U.S. governmental organizations (www.cia.gov)
- **int** - international organizations
- **mil** - U.S. military
- **net** - networks (InterNIC)
- **org** - other organizations (www.w3.org)
- More domains have been added in recent years.

Country Domains

- Generic domains usually refer to hosts inside the U.S.
- Other countries use two-letter country domains:
  - il - Israel
  - uk - United Kingdom
  - jp - Japan
  - se - Sweden
- These domains have sub-domains that correspond to the generic domains, for example:
  - co.il is the domain of all commercial organizations in Israel
  - ac.il is the domain of all academic institutions in Israel.
URLs

- Each information piece on the Web has a unique identifying address, called a URL (Uniform Resource Locator).
- A URL takes the following form:
  - `protocol` hostname file
- It has 3 parts: a `protocol` field, a `hostname` field, and a `file` field.

URL Fields

- The `protocol` field ("http" in the previous example) specifies the way in which the information should be accessed.
- The `hostname` field specifies the host on which the information is found.
- The `file` field specifies the particular location in the host's file system where the file is found.
- More complex forms of URLs are possible.

Using IP Addresses in URLs

- How does the browser know the IP address of the Web server?
- One possibility is that the user explicitly specifies the IP address of the server in the hostname field of the URL, for example:
  - `http://132.68.32.15/index.html`
- However, it is inconvenient for people to remember such addresses.

From Hostnames to IP Addresses

- When we address a host in the Internet, we usually use its hostname (e.g., using a hostname in a URL).
- The browser needs to map that hostname to the corresponding IP address of the given host.
- There is no algorithm for computing the IP address from the hostname.
- A lookup table provides the IP address of each hostname.

Where is the Translation Done?

- The translation of IP addresses to hostnames requires a lookup table.
- Since there are millions of hosts on the Internet, it is not feasible for the browser to hold a table that maps all hostnames to their IP-addresses.
- Moreover, new hosts are added to the Internet every day and hosts change their names.

DNS (Domain Name System)

- The browser (and other Internet applications) use a DNS Server to map hostnames to IP addresses.
- DNS is a hierarchical scheme for naming hosts.
  - DNS servers exchange information in order to update their tables.
  - The command `nslookup` gets an IP address and returns a hostname or vice-versa.
- It runs on clients and contacts a DNS server.
The HTTP Protocol

- Hypertext Transfer Protocol
- Used between Web clients (e.g., browsers) and Web servers (and proxies)
- Text based
- Built on top of TCP
- Stateless protocol (it doesn’t remember your previous requests)

Browsers Are Clients

- We use a browser to display HTML pages
- The browser is responsible for fetching the HTML pages and displaying their contents according to the HTML rules

Web Servers

- HTML pages are stored in file systems
- Some hosts, called Web servers, can access these HTML pages
- Each Web server runs an HTTP-daemon in order to make its HTML pages available to other hosts
- The term “Web server” refers to the software that implements the HTTP daemon, but sometimes it also refers to the host that runs that software

HTTP Daemons

- An HTTP-daemon is an application that constantly runs on a Web server, waiting for requests from remote hosts
- Technically, any host connected to the Internet can act as a Web server by running an HTTP-daemon application
- A Web client (e.g., browser) connects to a Web server through the HTTP protocol and requests an HTML page

Browser-HTTPD Interaction

- The file index.html is the default requested file
- Browser sends a request to the Web server at the URL http://www.google.com
- The Web server responds with the file index.html
- The file is then displayed in the browser
Browser-HTTPD Interaction
- The user requests http://www.cs.technion.ac.il/index.html
- The browser contacts the HTTP-daemon running on the host www.cs.technion.ac.il and requests the HTML page /index.html
- The HTTP-daemon translates the requested name to a specific file in its local file system
- The HTTP-daemon reads the file index.html from the disk and sends the content of the file to the browser
- The browser receives the HTML page, parses it according to the HTML rules and displays it

HTTP Transaction – Client
- Client request:
  - The request
    GET /index.html HTTP/1.0
  - Optional header information
    User-Agent: browser name
    Accept: formats the browser understands
    ...
  - A blank line (\n)
  - The client can also send data (e.g., the data that the user entered into an HTML form)

HTTP Transaction – Server
- Server response:
  - Status line
    HTTP/1.0 200 OK
  - Header information
    Content-type: text/html
    Content-length: 3022
    ...
  - A blank line (\n)
  - Document data

Proxy Servers
- A proxy server acts as a delegate of browsers for accessing the Web
- The browser transfers the request for a document to the Proxy
- The Proxy contacts the Web server and fetches the document on behalf of the browser

Advantages of Proxy Servers
- Proxy servers have several advantages over direct access:
  - They can be combined with a firewall to enable restricted access to the Internet
  - They enable caching of popular documents
  - They can extend the functionality of the browser by translating from one protocol to another (for example, from FTP to HTTP and vice-versa)
**Disadvantages of Proxy Servers**

- Delay the interactions
- Problematic for
  - Persistent connections
  - Secure connections
  - Using a cache may cause errors

**Responding to Clients’ Inputs**

- HTML pages are static documents
- Sometimes users supply input, for example, keywords submitted to a search engine
- The Web server has to react to this input
  - The output is an HTML page that is not known in advance
  - In order to react to the input, the Web server may have to use some applications (e.g., database queries)

**Server-Side Programming**

- Writing applications that react to clients’ inputs by creating HTML pages on the fly is known as **server-side programming**
- A client request will include, in addition to the URL of the service provider, a list of parameters, for example:
  
  ```
  http://www.google.com/search?q=search-word
  ```

- The response to the above request is a **dynamic HTML page** and generating it may involve interaction with other applications (e.g., database queries)

**Browser-HTTPD Interaction**

1. User requests
2. ```
   GET /search?q=en&q=me
   ```
3. Web server generates content
4. Browser shows the results

**Client-Side Programming**

- Certain parts of a Web application can be executed locally, in the client
- For example, some validity checks can be applied to the user’s input locally
- The user request is sent to the server only if the input is valid
- Java Script (not part of Java!) is an HTML-embedded scripting language for client-side programming

**JavaScript**

- JavaScript is a scripting language for generating dynamic HTML pages in the browser
- The script is written inside an HTML page and the browser runs the script and displays an ordinary HTML page
- There is some interaction of the script with the file system using cookies
- Cookies are small files that store personal information in the file system of the client
  - For example, a cookie may store your user name and password for accessing a particular site
Examples

```html
<html>
<body>
<script type="text/javascript">
    document.write("<h1>Hello World!</h1>";?></script>
</body>
</html>
```

Examples

```html
<html>
<head>
<script type="text/javascript">
    function hello() {
        alert("Hello world (called with the onload event); ")
    }
</script>
</head>
<body onload="hello()">
<p>Some content</p>
</body>
</html>
```

Style Sheets

- A file that is used for storing information about the way elements of HTML (or XML) should appear on the browser
- A style sheet increases the separation between content and presentation
  - Easier to generate large sites in which all the pages have the same style
  - It allows changing the look of many pages by changing a single file
  - May reduce network traffic

Common Style Languages

- CSS
  - Simple
  - Attach style properties to element types in a "cascading" manner
- XSL
  - Expressive
  - Can transform HTML and XML to any textual format
- It is possible to combine CSS and XSL

CSS Example

```html
<html>
<head>
<style type="text/css">
  h1 {text-decoration:overline;}
  h2 {text-decoration:line-through;}
  h3 {text-decoration:underline;}
  h4 {text-decoration:blink;}
</style>
</head>
<body>
<h1>Some content here</h1>
</body>
</html>
```

CSS Example

```html
<html>
<head>
<style type="text/css">
  p.normal {font-style: normal;}
  p.italic {font-style: italic;}
  p.oblique {font-style: oblique;}
</style>
</head>
<body>
<p class="normal">This is a paragraph, normal.</p>
<p class="italic">This is a paragraph, italic.</p>
<p class="oblique">This is a paragraph, oblique.</p>
</body>
</html>
```