The Web: Moving Data Around the World

LBSC 690: Jordan Boyd-Graber

University of Maryland

September 17, 2012

Adapted from Jimmy Lin’s Slides
Goals (Computer - Hardware / Computer - Computer)

- How data are stored
- How the web works
- Create your first webpage
- Learn how to transfer files
Outline

1. Storage
2. Protocols and the Internet
3. Making a Webpage
4. Discussion
5. Practice Problems
What are some kinds of storage?

- RAM
- Flash memory
- Magnetic (Hard Disk)
- Optical memory
RAM

- Lots of little electronic switches
- Jay Forrester (MIT): First practical RAM (1951)
- Little magnetic donuts; orientation could be switched / read by sending appropriate electric pulses
- Unlike tape, you could read anything at any time (random access)
- Volatile
RAM

- Lots of little electronic switches
- Jay Forrester (MIT): First practical RAM (1951)
- Little magnetic donuts; orientation could be switched / read by sending appropriate electric pulses
- Unlike tape, you could read anything at any time (random access)
- Volatile
- But don’t count on volatility for security
Flash

- Like RAM, lots of little electronic switches
- Retains memory when powered off
- Fairly cheap, getting denser
- Slower than RAM, faster than HDD
Flash

- Like RAM, lots of little electronic switches
- Retains memory when powered off
- Fairly cheap, getting denser
- Slower than RAM, faster than HDD
- Where can you find Flash memory?
Hard Drives

- Little magnetic flakes that get spun around
- Retains memory when powered off
- For consumers, cheapest per MB
- Relatively slow
- What made the iPod popular (in addition to its UI)
- RAID (Redundant Array of Inexpensive Disks)
  - Backup and speedup
  - Duplicated data across disks so the head doesn’t have to move as far on average
Optical

- Lasers detect little pits in media
- Retains memory when powered off
- Very cheap to produce
- Relatively slow
- Can be fairly durable
- (With some effort) Rewriteable
Cloud

- Physical storage doesn’t matter (you can’t see it)
- Follows you wherever you go
- Requires network access for update
- Not as cheap as buying a HD (backup costs?)
  - Google Docs
  - Dropbox
  - Mozy
How does your computer know where stuff is, physically, on your disk?

Examples: ZFS, ReiserFS, NTFS, FAT32, AFS, Ext3

The folder metaphor
How does your computer know where stuff is, physically, on your disk?

Examples: ZFS, ReiserFS, NTFS, FAT32, AFS, Ext3

The folder metaphor

- Hierarchically nested directories
- Absolute vs. relative paths (look out for this!)
  - ../index.html
  - c:/windows/index.html
- File extensions

Operating systems have their favorite file systems
Outline

1 Storage
2 Protocols and the Internet
3 Making a Webpage
4 Discussion
5 Practice Problems
The tubes of the Internets

Packet-based
- Each transmission is broken up into pieces and routed separately
- High network load results in long delays

Circuit-based
- Fixed connection between caller and called
- High network load results in busy signals
Packet Switching

- Break long messages into short “packets”
- Keeps one user from hogging a line
- Each packet is tagged with where it’s going
- Route each packet separately
- Each packet often takes a different route
- Packets often arrive out of order
- Receiver must reconstruct original message

Questions:
- How do packet-switched networks deal with continuous data?
- What happens when packets are lost?
Internet = collection of global networks
Web = particular way of accessing information on the Internet
Uses the HTTP protocol
Other ways of using the Internet
  - Usenet
  - FTP
  - email (SMTP, POP, IMAP, etc.)
  - Internet Relay Chat
The Internet is a Collection of Networks

What are Firewalls? Why can’t you do stuff behind them?
The Internet is a Collection of Networks

VPN = Virtual Private Network
The Web is Built on Standards

- Basic protocols for the Internet
  - TCP/IP (Transmission Control Protocol/Internet Protocol): basis for communication
  - DNS (Domain Name Service): basis for naming computers on the network

- Protocol for the Web
  - HTTP (HyperText Transfer Protocol): protocol for transferring Web pages

- Protocol for E-mail
  - SMTP, IMAP: broken?
    - privacy
    - spam
IP Address

- Every computer on the Internet is identified by an address.
- IP address = 32 bit number, divided into four “octets”.
- Example: go in your browser and type “http://128.8.237.26/”
- Also used for “geolocation” (which language Google uses, no Hulu for Canadians).
- Questions:
  - What’s the difference between static and dynamic IP?
  - Are there enough IP addresses to go around?
IP Address

- Every computer on the Internet is identified by an address
- IP address = 32 bit number, divided into four “octets”
- Example: go in your browser and type “http://128.8.237.26/”
- Also used for “geolocation” (which language Google uses, no Hulu for Canadians)

Questions:
  - What’s the difference between static and dynamic IP?
  - Are there enough IP addresses to go around?
  - Even with 4 billion, things are getting crowded
Every computer on the Internet is identified by an address.

**IP address** = 32 bit number, divided into four “octets”

Example: go in your browser and type “http://128.8.237.26/”

Also used for “geolocation” (which language Google uses, no Hulu for Canadians)

Questions:
- What’s the difference between static and dynamic IP?
- Are there enough IP addresses to go around?
- Even with 4 billion, things are getting crowded

**IPv6** - 128 bits long ($5 \times 10^{28}$ IP Addresses per person)

**Network Address Translation** - Not everybody gets a private IP
Historical Bias of IPv4
IPv6

- Written as eight 4-digit hexadecimal numbers (base 16)
- Plenty of room!
- Harder to write down
- e.g. Google: 2001:4860:4860::8888
- Some technical advantages
  - “ephemeral” addressed for privacy
  - multicast
### Hexadecimal

<table>
<thead>
<tr>
<th>Hexadecimal</th>
<th>Binary</th>
<th>Decimal</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0000</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0001</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0010</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>0011</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>0100</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>0101</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>0110</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>0111</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>1000</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>1001</td>
<td>9</td>
</tr>
<tr>
<td>A</td>
<td>1010</td>
<td>10</td>
</tr>
<tr>
<td>B</td>
<td>1011</td>
<td>11</td>
</tr>
<tr>
<td>C</td>
<td>1100</td>
<td>12</td>
</tr>
<tr>
<td>D</td>
<td>1101</td>
<td>13</td>
</tr>
<tr>
<td>E</td>
<td>1110</td>
<td>14</td>
</tr>
<tr>
<td>F</td>
<td>1111</td>
<td>15</td>
</tr>
</tbody>
</table>
**Hexadecimal**

<table>
<thead>
<tr>
<th>Hexadecimal</th>
<th>Binary</th>
<th>Decimal</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0000</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0001</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0010</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>0011</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>0100</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>0101</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>0110</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>0111</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>1000</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>1001</td>
<td>9</td>
</tr>
<tr>
<td>E</td>
<td>1110</td>
<td>14</td>
</tr>
<tr>
<td>F</td>
<td>1111</td>
<td>15</td>
</tr>
</tbody>
</table>

**Huh?**

More when we do HTML colors!
“Domain names” improve usability
- Easier to remember than numeric IP addresses
- DNS converts between names and numbers
- Written like a postal address: specific-to-general

Each name server knows one level of names
- “Top level” name server knows .edu, .com, .mil, ...
- .edu name server knows umd, caltech, mit, stanford, princeton, ...
- .umd.edu name server knows ischool, wam, ...

Recent developments
- New TLDs
- Non-Latin addresses
Transport Control Protocol specifies how data moves across the Internet.

Each node has address and ports:
- Loopback: 127.0.0.1
- Local: 10.x.x.x, 192.168.x.x (What does it mean if this is your IP address?)
Transport Control Protocol specifies how data moves across the Internet.

Each node has address and ports:
- Loopback: 127.0.0.1
- Local: 10.x.x.x, 192.168.x.x (What does it mean if this is your IP address?)

A port is a number to channel traffic:
- 20 FTP
- 22 SSH
- 25 SMTP
- 80 HTTP
- 2710 Bittorrent tracker

Uses:
- Block applications
- Have computers specialize (e.g. behind NAT)
- Security (Firewall only opens port 80)
(Quite simplified) Routing table for 4.8.15.2

<table>
<thead>
<tr>
<th>Destination</th>
<th>Next Hop</th>
</tr>
</thead>
<tbody>
<tr>
<td>52.55.***</td>
<td>63.6.9.12</td>
</tr>
<tr>
<td>18.1.***</td>
<td>52.55.64.2 or 63.6.9.12</td>
</tr>
<tr>
<td>4.***</td>
<td>225.2.55.1</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

Can also include:
- Cost
- Quality
- Filtering
TCP is **how**, IP is **what**

Fundamental unit of IP communication is the packet

IP Provides support for:
- Missing data
- Repeated arrivals
- Out of order arrival
- Data corruption
IP is just a way of breaking up data

Doesn’t even have to be on computers

Pigeons: 1 hr latency, 55% packet loss

This is why the Internet is in so many places on so many devices
Fiber Optics
Ethernet
  - Hub - Everyone talks at once, shuts up if they conflict
  - Router - There’s a moderator
IEEE 802.11(a/g) (Wireless) - Radio in your building
EDGE (Enhanced Data rates for GSM Evolution) - Radio to your phone

Takeaway
To improve connectivity, focus on the weakest link. In a crowded dorm, don’t upgrade the T1 if the wireless is saturated. In rural Iowa, don’t install fiber optic cable to every room.
Why Code HTML by Hand?

- The only way to learn is by doing
- WSIWYG editors . . .
  - Often generate unreadable code
  - Ties you down to that particular editor
  - Cannot help you connect to backend databases
- Hand coding HTML allows you to have finer-grained control
- HTML is merely demonstrative of other important concepts:
  - Structured documents
  - Metadata
Editing Plaintext

- Used to be the norm!
- Stuff you already have:
  - Notepad (Windows)
  - TextEdit (Mac)
  - pico (Linux)
- Good options:
  - TextWrangler (Mac)
  - Editpad (Windows)
  - VI, Emacs, gedit (Linux)
- One-to-one correspondence between characters and ASCII written to disk
Hello World

<html>
<head>
<title>Hello World!</title>
</head>
<body>
<p>Hello world! This is my first webpage!</p>
</body>
</html>
Brian Kernighan: engineer at AT&T who helped create UNIX, C, AWK, AMPL, other programming languages. Created an example program that printed “hello world” and nothing else to show off C. Now everybody does it.
Tips

- Edit files on your own machine, upload when you're happy
- Save early, save often, just save!
- Reload browser
- File naming
  - Don’t use spaces!
  - Punctuation matters!
Uploading Your Page

- Connect to "terpconnect.umd.edu"
- Change directory to "public_html" (Assignment 0)
- Upload files
- Your very own home page at:

http://terpconnect.umd.edu/~USERID/
WinSCP
The server’s host key was not found in the cache. You have no guarantee that the server is the computer you think it is.

The server’s rsa2 key fingerprint is:

If you trust this host, press Yes. To connect without adding host key to the cache, press No. To abandon the connection press Cancel.

Continue connecting and add host key to the cache?

Yes  No  Cancel  Copy Key  Help
WinSCP
Fetch
Fetch
### Fetch

A screenshot of a file management interface showing a local directory on a computer. The directory contains various folders and files:

- **documents** (Folder)
- **dyn-documents** (Folder)
- **dyn-pub-quiz** (Folder)
- **index.html** (HTML Document, 2.9 KB, 09.01.10)
- **pub_quiz** (Folder)
- **random** (Folder)
- **restricted** (Folder)
- **static** (Folder)
- **style.css** (CSS style sheet, 1.7 KB, 09.01.10)

The directory is connected to a server named `terpconnect.umd.edu`. The interface includes options for Back, Path, Recent, Get, Put, Quick Look, Edit, Get Info, WebView, New Folder, and Delete.
Outline

1. Storage
2. Protocols and the Internet
3. Making a Webpage
4. Discussion
5. Practice Problems
What’s wrong with this picture?
This week’s discussion

As part of your school’s technology committee, you need to plan the networking hardware purchases. Describe what hardware components you might need in your school to connect all of your classrooms to the school network and the Internet (server, wireless access points, switches, storage, cables etc.). How will you handle addressing the computers; what use cases would change your decision? Context: Your school has a special room for your server(s) with the outside T1 connection to your Internet Service Provider (ISP); it receives a single static IP. The school is also wired with a single 10Mbs ethernet connector into each classroom from the server room. All computers connect to a DHCP server that gives it a 192.168.1.X address.
This week’s discussion

- Your vendor wants you to upgrade your wiring. Is it worth it?
This week’s discussion

- Your vendor wants you to upgrade your wiring. Is it worth it?
- A teacher wants to use a classroom computer as a webserver. Who can see what webpages its serving?
This week’s discussion

- Your vendor wants you to upgrade your wiring. Is it worth it?
- A teacher wants to use a classroom computer as a webserver. Who can see what webpages its serving?
- Students are going to be allowed to bring in their personal laptops. How might you change the way your system is set up?
This week’s discussion

- Your vendor wants you to upgrade your wiring. Is it worth it?
- A teacher wants to use a classroom computer as a webserver. Who can see what webpages its serving?
- Students are going to be allowed to bring in their personal laptops. How might you change the way your system is set up?
- Disney caught one of the computers on your network serving a bittorrent of a popular film. How did they know it was your school? How can you prevent this from happening?
As a rule of thumb, MP3-encoded sound takes about 1 MB/minute of storage. How big a disk would be required to record everything you have ever heard in your life so far in MP3?
As a rule of thumb, MP3-encoded sound takes about 1 MB/minute of storage. How big a disk would be required to record everything you have ever heard in your life so far in MP3?

\[
\frac{30\text{years}}{1} \times \frac{1440\text{minutes}}{1\text{day}} \times \frac{365.25\text{days}}{1\text{year}} \times \frac{1\text{MB}}{\text{minute}} \approx 16 \cdot 10^6\text{MB}
\] (1)
As a rule of thumb, MP3-encoded sound takes about 1 MB/minute of storage. How big a disk would be required to record everything you have ever heard in your life so far in MP3?

\[
\frac{30\text{years}}{1} \cdot \frac{1440\text{minutes}}{1\text{day}} \cdot \frac{365.25\text{days}}{1\text{year}} \cdot \frac{1\text{MB}}{1\text{minute}} \approx 16 \cdot 10^6\text{MB} \quad (1)
\]

\[
16 \cdot 10^6\text{MB} \cdot \frac{10^6\text{bytes}}{\text{MB}} \approx 16 \cdot 10^{12}\text{bytes} = 16\text{TB} \quad (2)
\]
A New York Times article on 6/9/04 says that it can take “days” to download a high quality movie over a DSL line. Suppose that the DSL line is 1 Mbps, and that a standard movie DVD is about 5 GB. How long does the download take under these assumptions?
A New York Times article on 6/9/04 says that it can take “days” to download a high quality movie over a DSL line. Suppose that the DSL line is 1 Mbps, and that a standard movie DVD is about 5 GB. How long does the download take under these assumptions?

\[
5 \text{GB} \cdot \frac{1 \text{s}}{\text{Mbit}} \cdot \frac{10^3 \text{MB}}{\text{GB}} \cdot \frac{8 \text{bit}}{\text{byte}} \approx 40 \cdot 10^3 \text{s}
\]
A New York Times article on 6/9/04 says that it can take “days” to download a high quality movie over a DSL line. Suppose that the DSL line is 1 Mbps, and that a standard movie DVD is about 5 GB. How long does the download take under these assumptions?

\[
5 \text{GB} \cdot \frac{1 \text{s}}{1 \text{Mbit}} \cdot \frac{10^3 \text{MB}}{\text{GB}} \cdot \frac{8 \text{bit}}{\text{byte}} \approx 40 \cdot 10^3 \text{s}
\]

\[
40 \cdot 10^3 \text{s} \cdot \frac{1 \text{hour}}{3600 \text{s}} \approx 11 \text{hours}
\]
How many bits are needed to represent monetary values of up to twenty dollars to the nearest penny?
How many bits are needed to represent monetary values of up to twenty dollars to the nearest penny?

If we have \( n \) bits, we can represent \( 2^n \) values. There are a total of 2000 pennies in twenty bucks, so we need at least 2000 unique values. Everybody should know that

\[
2^{10} = 1024, \quad (5)
\]

which is too small, so

\[
2^{11} = 2048, \quad (6)
\]

should do it.
Compute the number of bits stored per square inch of recording surface for a CD-ROM.
Compute the number of bits stored per square inch of recording surface for a CD-ROM.

\[
\frac{750 \text{MB}}{CD} \cdot \frac{CD}{((120\text{mm})^2 - (15\text{mm})^2)\pi} \cdot \frac{645.16\text{mm}^2}{8\text{bit} \cdot 2^{20}\text{bytes}} \cdot \frac{8\text{bit}}{\text{byte}} \cdot \frac{1 \text{in}^2}{\text{MB}} \quad (7)
\]
At Google, somewhere they store the satellite views of the earth displayed at maps.google.com. Suppose the finest resolution is 1 meter (that is, they store one pixel for each 1 meter by 1 meter square of the earth's surface). How many pixels are there if you ignore compression? To save you a trip to Google, the surface of a sphere is $4\pi r^2$, and the radius of the earth is 6000 kilometers.
At Google, somewhere they store the satellite views of the earth displayed at maps.google.com. Suppose the finest resolution is 1 meter (that is, they store one pixel for each 1 meter by 1 meter square of the earth's surface). How many pixels are there if you ignore compression? To save you a trip to Google, the surface of a sphere is \(4\pi r^2\), and the radius of the earth is 6000 kilometers.

\[
\frac{1\text{pixel}}{m^2} \cdot \left(\frac{10^3\text{m}}{1\text{km}}\right)^2 \cdot 4\pi (6 \cdot 10^3\text{km})^2
\]

\[
\frac{10^6\text{pixel}}{km^2} \cdot 450 \cdot 10^6 \approx 4.5 \cdot 10^{14}
\]