COMPSCI 111 / 111G

Mastering Cyberspace:
An introduction to practical computing

World Wide Web

Introduction

Andrew Luxton-Reilly
- Rm: 303.479
- Ext: 85654
- Email: andrew@cs.auckland.ac.nz
- Open door policy

Ground Rules
- Respect other people
- One person speaks at a time
- Questions welcome any time
- Answers expected when questions asked
- You are expected to think
- You are expected to work on exercises

Learning Outcomes

Students should be able to:
- describe the meaning of terms related to the WWW
- put a series of historical events into chronological order
- describe the difference between the WWW and the Internet
- describe the process that occurs when a user requests a web page
- list the different places that web pages are logged
- discuss some of the legal and ethical concerns that arise from our use of search engines to access the WWW

What is the World Wide Web?

Possible definitions:
- Information system using the Internet to access hypertext documents.

- A hypertext-based Internet service used for browsing Internet resources.

- Information available on the Internet that can be easily accessed with software usually called a "browser."

- A system of the Internet servers that support specifically formatted documents (hypertexts). The documents are written in HTML that supports links to other documents, graphics, audio and video files.

- A network within the Internet comprised of a constellation of networked resources. The Internet servers utilize HTTP to transfer documents and multimedia files formatted in hypertext markup language (HTML).

Hypertext

Hypertext
- Text with links
- Allows free association
- More creative learning
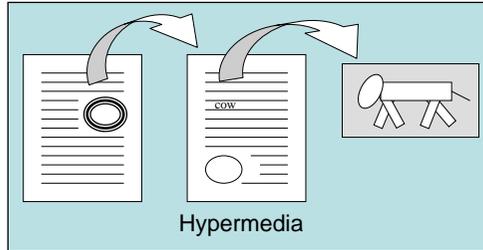
Hypermedia
- Extension of hypertext
- Different forms of media with links between them (Hyperlinks)
- Interactive

Multimedia
- The integration of many forms of media
- Text
- Images
- Sound

Hypertext

Vannevar Bush
- 1945 published idea of hypertext using MEMEX system

Ted Nelson
- 1960 Xanadu project
- First computer hypertext system
- 1965 Created the term Hypermedia

Tim Berners-Lee
- 1989 starts the WWW project at CERN

http://en.wikipedia.org/wiki/Www

The WWW project

Background: CERN
- Many networks existed
- Each network had many documents

Aims
- Access documents from any network in seamless manner
- World-Wide (distributed)
- Easy to add documents (dynamic)

Proposal
- Use Hypertext
- No intention to support hypermedia
- Research only

http://en.wikipedia.org/wiki/WorldWideWeb

Rapid Evolution

1989	**Tim Berners-Lee begins work on the WWW project**
1991	**WWW operational at CERN**
1992	**WWW goes public**
1993	**Mosaic created by Marc Andreessen (First GUI browser)**
1994	**US Senate allow commerce on Internet** **Netscape Communications formed, Yahoo! Formed**
1995	**Microsoft Internet Explorer**
1998	**Netscape became open-source, developed into Mozilla**
2003	**Mozilla Foundation**

http://en.wikipedia.org/wiki/Web_browser

Technical Details

HTTP

- Hypertext Transfer Protocol
- Language used to transfer Hypertext documents
- Client-Server Model

TCP/IP ———
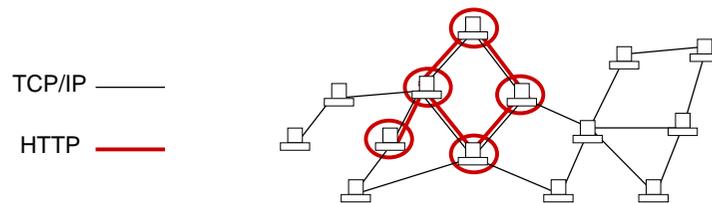
HTTP ———

Exercise

Explain how the WWW differs from the Internet

What does HTTP do?

Cyberspace Addresses

## Uniform Resource Locators (URL)

- Address used for any web resource

Protocol

- Name of the protocol used
- ftp://　　　http://

By default, web browsers will use http as the protocol.

If no file given, they will often default to index.html

Domain

- Name of a host computer (IP address or domain name)
- www.cs.auckland.ac.nz

File/ Resource

- Path of the file
- /andrew-l/index.html

ihug.co.nz　　**should be**

http://ihug.co.nz/index.html
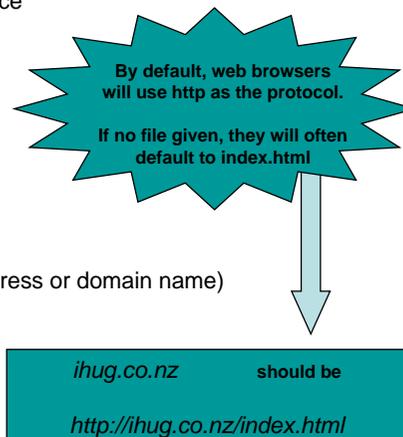
Terms

Web Site

- A collection of Web pages related to a single topic or theme. Normally designed and maintained by a single individual or organisation

Web Page

- A hypermedia document designed for the WWW

Web Browser

- Software used to access information on the World Wide Web
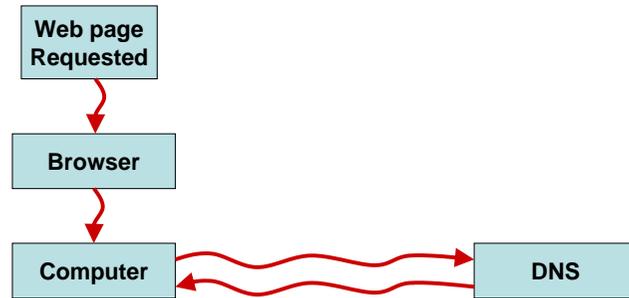- Sends requests to a web server
- Client

Web Server

- Software that makes local files available through the web
- Fulfils requests from a web browser
- Server

Accessing a web page (1)
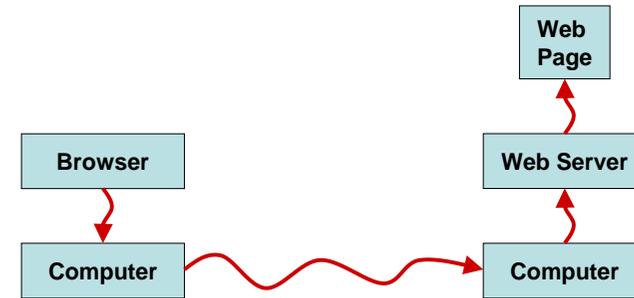
Client (Web Browser) runs on the local machine
- User requests a web page
- Client contacts the DNS to find the IP address

```
Web page Requested → Browser → Computer ⇄ DNS
```

Accessing a web page (2)

Web server runs on the destination machine
- Request sent to destination domain
- Web server accepts the request and finds the web page

```
Browser → Computer → Computer → Web Server → Web Page
```

Accessing a web page (3)

Web page is sent from the server to the client
- Client (web browser) displays the page

```
Web page displayed ← Browser ← Computer ← Computer ← Web Server ← Web Page
```
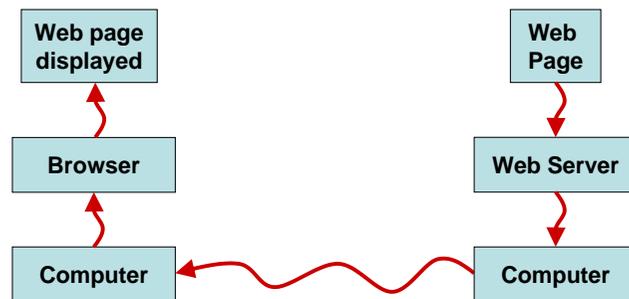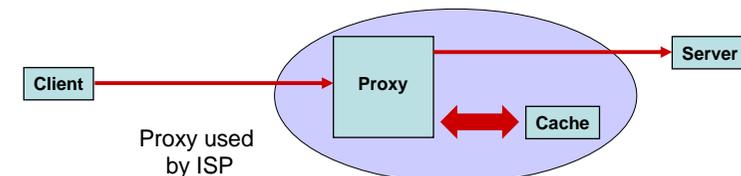
More Terms

Proxy
- A computer which sits between the client and server, intercepts and processes requests

Cache
- Store of information for quick access
- (e.g. caching may be used by proxy servers to speed web use)

Firewall
- Prevents unauthorised access to or from a private network

```
Client → Proxy ⇄ Cache → Server
Proxy used by ISP
```

Logging web page access

Client keeps log
- History in web browser

Operating System keeps log
- Requests are logged by Windows on local machine

ISP keeps log
- Requests from "IP address" to "IP address" for "Page Name"
- Example, paradise.net has the logs available for users to check

Web server keeps log
- Gets requests from "IP address" for "Page Name"

http://en.wikipedia.org/wiki/Google_and_privacy_issues

Exercise

Is it possible to view pages in the WWW anonymously?

What privacy concerns are raised by viewing WWW pages?

Navigating

Finding information
- 45% users have problem finding new information
- 30% users have problems finding known information
- Web is very large, rapidly changing

Search Engines
- Automated
- Essential
- Our gateway to information

Search Engines

Companies
- Google
- Yahoo
- MSN Search

Automatically search every machine
- Archive the contents
- Index all the words
- Try to determine the relevance of the page

http://en.wikipedia.org/wiki/Search_engines

Can we trust the search engines?

Search Engines
- gateway to information
- pages are rated (how?)

Example: Veronica
- Camping search led to homosexuals
- Powerful groups lobbied
- Modified search engine (results filtered)

Publishers/ Advertisers
- Trick the search engines (repetition of words)
- Search engines tailor advertisements to searches
- Pay for higher rating?
- Who owns the search engines? How do they make their money?

http://en.wikipedia.org/wiki/Google_censorship

Recent Google News

Google sued for breach of copyright
- Claim fair use of images
- Appealing the decision

Google makes deal with Chinese Govt.
- Chinese Govt employs 30,000 people to censor the web
- Chinese version of Google is filtered

Brazilian Govt wants Google to help prevent online crime
- Google's social networking used to sell drugs

Online Books
- Google and UMich

http://en.wikipedia.org/wiki/History_of_Google

Searching

Searching Tips
- Learn how to use the advanced features of your search engine
- If the first page is not promising, choose different key words
- Use quote marks to search for phrases "Computer Science"

Finding useful sites
- Use specialist sites for specific searches
- Build a list of useful resources
- IMDB, New York Times, Ebay, BBC
- Trademe, NZ Herald, IRD, Woolworths, GOVT

Searching by URL
- Make a good guess
- Use knowledge of existing structure

Problems with the structure

Broken Links
- Pages which have been moved
- Referring page is not altered

No inherent security/ tracking/ accounting system
- Difficult to have layers of security
- Forces publishers to rely on advertising revenue

No inherent information indexing
- Much of the information is not accessed by search engines
- Information created on-the-fly from databases
- Information in other formats (postscript, pdf, archived)

Online innovations

Skype
- Voice over IP
- Talk to anyone in the world for free / cheap

Peer to Peer networks
- BitTorrent
- Swarming downloads
- Online Television

Google
- Acquired online word processor
- Rumours of alternate operating system
- Challenge Microsoft

Free Books
- http://digital.library.upenn.edu/books/
- http://books.google.com

Web-agents

Computer programs that operate on your behalf
- Tracks all your browsing habits
- Makes suggestions based on what you have read

Google Reader
- http://labs.google.com/

TiVo Suggestions
- Similar principle with television viewing

Future Directions

The only media of the future

The Internet is changing extremely rapidly
- Too fast for legislation to keep up
- Too fast to predict the future

Coming soon
- Wireless Connections
- Wearable PC's
- Integrated Media (Interactive T.V.)
- Household Appliance connections
- Crime