Behind your browser

What goes on behind the scenes when you surf the web?

Outline:

- URLs, IP addresses and the network
- Static web page content and HTML
- Dynamic web page content and browser "state"
- User-generated web page content and "Web 2.0"

Universal Resource Locators (URLs)

URLs, eg `http://www.csse.unimelb.edu.au/~lee/ims/demo.html` specify which web page you want to view (and how to get it)

The "`http://`" part tells your computer to communicate with the remote computer which hosts the web site using the Hyper-Text Transfer Protocol

The "`www.csse.unimelb.edu.au`" part is the *domain name* of the remote computer hosting the web site

The "`~lee/ims/demo.html`" tells that computer where on its file system to look for the web page content

The "`.html`" extension tells your computer the content is HyperText Markup Language

IP addresses

Each computer connected to the internet has an IP (Internet Protocol) number (or "address"), normally written as four numbers between 0 and 255, eg 202.29.151.3 (somewhat more readable than 11001010000111011001011100000011)

There is a hierarchy of Domain Name Servers (DNS) which convert domain names (eg "`www.google.com`") to IP addresses

Any computer on the web knows where to find a DNS. To contact `www.google.com` it first contacts a DNS to convert it to an IP address, 202.29.151.3 and this number is used for subsequent communication

If that DNS doesn't know about the domain it asks another (more authoritative) DNS

Getting a web page

To get a web page, your computer sends a message to a particular IP address, asking for a particular page (eg `~lee/ims/demo.html`)

As part of the message, your computer includes its own IP address, so the remote computer knows *where* to send the page contents

The message also contains other information, such as your browser and operating system and the URL the page you are currently viewing (its like sending a letter, including your return address etc, rather than making a phone call)

Once the request has been sent, your computer waits for a reply, which (hopefully) arrives promptly with the web page content, which is displays appropriately in your browser

The *web server* on the remote computer keeps an eye out for requests arriving and for each one, sends out a reply

Transmission Control Protocol/Internet Protocol

Sending web page requests and replies (etc) between computers on the web is quite complex

All the messages are broken down into chunks when they are sent, and reassembled when all chunks have arrived (its more like sending a bunch of post cards than a single long letter)

Each chunk has to find a route through the network from the source computer, often via numerous other computers, to the destination computer (typically the source and destination computers do not have a direct connection)

The individual computers on the network are not completely reliable, so the network is constantly changing, and its not guaranteed the message will get through, but it nearly always does

User Datagram Protocol is used for streaming audio/video etc and all these protocols rely on lower level protocols

Static web pages — HTML

HyperText Markup Language (HTML) is a simple language (with many different versions) designed for web page content

It supports plain text and "markup" which describes the text

For example, this is a heading, this is the start of a numbered list, this is an item in the list, this should be emphasised

It can also specify details of layout, fonts etc — this is *widely* miss-used, eg `<p class=MsoNormal align=center style='text-align:center'><span style='font-size:10.0pt;font-family:Arial'><o:p> </o:p></span></p>`

It also supports *links* to other web pages

It also supports other media, such as images (by having a link to URL containing the image and markup saying its an image)

Thus a single web page displayed in your browser may contain content from multiple URLs anywhere on the web

Dynamically created web pages

HTML and its various extensions allow you to create files with (relatively) static content

The files can be manually updated occasionally (eg what subjects I'm teaching this year), or even automatically updated regularly (eg, the Melbourne weather forecast)

But Google can't possibly maintain a web page for each possible search anyone is likely to do

One solution is to encode information in the "file name" part of the URL and pass that information to a *program* running on the web server which will generate HTML whenever it is run

Eg, `http://` `www.google.com` `/search?q=tank+man&...`

Dynamically created web pages

One way of generating content dynamically is by running a program on the "server side" — the remote computer which runs the web server

Another way is to run a program on *your* computer ("client-side")

Programs written in languages such as Javascript and Java can be embedded in web pages (with markup) and then run by your browser

These programs can process your input and produce animations etc without (necessarily) communicating with the server

Its important that such programs have only restricted access to your computer — you don't want them to trash your hard drive or send your banking details to any hackers!

More on security

Transmitting sensitive information (eg, your credit card number) between your computer and another is also problematic

The information may go via many other computers, many of which are untrustworthy

Some URLs start with "`https://`", which means Secure HTTP

The data communicated will be encrypted with a special *key* (basically a random number) so eavesdropping is no use

Your browser invents a key and communicates it to the remote web server in a secure way (relying on some very neat mathematics)

Maintaining browser "state"

So far, we have seen a (more or less) one to one correspondence between a URL and the web page content

Often its beneficial to customise web page content depending on who is requesting the page, and when

For example, in a discussion forum you typically want to see only the posts you have not seen already (and you don't want to use a different URL for each visit)

As you browse an online shop you might want to keep track of a "shopping cart" with potential purchases

You might like a web site to remember your preferences

Or someone else might like to track your browsing...

Cookies

Cookies allow browsers to maintain state by storing a small quantity of information on your computer's hard drive

Cookies are associated with particular domains, eg `unimelb.edu.au`

Whenever you visit a page in the domain your browser will send the associated cookies to the web server along with the page request

When the page contents is sent back, the web server can also send back new cookies for your browser to save

Cookies may be automatically deleted when you close your browser or after some particular time (or you can manually delete them or tell your browser not to accept them)

There are now a variety of other ways state is stored. Sometimes it can be hard to delete

Cookies and privacy

Cookies have important privacy implications (see later lecture), in part because web pages can have content from many different URLS (some of which may not be obvious)

When you view a `unimelb.edu.au` web page, your browser may be exchanging cookies with other web sites, so those web sites will know what web pages you are viewing

If they know who you are and have content on enough web sites, they may be able to build up a good picture of your interests and habits

Doubleclick (with advertisements), Google (with search, e-mail, groups, "analytics" etc), Facebook and Twitter are companies who specialise in gathering such information

User-generated web page content and "Web 2.0"

The web has always been filled with user-generated content — thats always been the vision

In the early days you needed to know HTML to produce web pages

Now there are many tools for creating creating content — its much easier than before

There is also more emphasis on collaboration — tools to help multiple people cooperate in building content (Wikis etc)

There is also more "real time" content (feeds etc)

But ease of participation has come with a price

A New Yorker comic from 1993 had the caption "On the internet no-one knows you are a dog" — nowadays Google, Facebook, the NSA etc most likely know you are a dog, whether you have fleas, and other personal information

Summary

The next time you hit on a link and a new page appears, you will have more of an idea of what happens

Your computer will send information to possibly several other computers around the world, and the information will travel via many more

On those computers, files will be accessed and programs may be run to access databases and compute information

The information will be sent back to your computer and may contain programs which will run on your computer

Its a technological marvel

And how it impacts on us personally and as a society is even more complex