World Wide Web Networked Information

Schedule

Networked Computer Systems Overview (S&N has the details)

The World Wide Web

- History of the web
- Markup languages and the Web
- XML

Building a Web Site

- XHTML programming
- Style sheets and formatting
- Design, Accessibility and User Interface Issues
- Dynamic Web Sites
- "Web 2.0"

Networked Information Systems

A **computer network** is a collection of nodes which are:

- computers (called **hosts**)
- and other hardware (such as printers or disks)

The computers inter-communicate by sending data as **signals** through a carrier, such as an optical fibre, a cable or using radio waves

The nodes of network share:

- **resources** - such as data or printers
- **processes** - the programs running on the computers

The software may be either independent co-operating software or be a single collection of software called a **Distributed System**

In either case, there needs to be a configuration structure, a **naming structure** and a **communication protocol** defined over the network

Kinds of Network

There are four main kinds of network:

Local Area Networks (LAN) connect small numbers of relatively close computers usingcables or fibre optics. Example - the IT lab.

Wide Area Networks (WAN) connect large numbers of widely distributed computers.Example - JANET, the British University computer network.

Metropolitan Area Networks (MAN) are being constructed to transmit video, voice andother data over fibre optic cables laid down in cities or towns.

InterNetworks are networks of networks. Example - the Internet.

The Internet

The **Internet** is the InterNetwork which encompasses most of the world's computer sites

Given such a network, it become possible to:

- **send messages** between users
- **broadcast information** to a number of users
- **login** to another machine
- **transfer data** from another site
- **browse** amongst the data on other sites
- **search** for data on other sites

It is important that a transmitting computer be able to find the appropriate computer to send to. The Internet uses sockets which consist of:

- The **Internet Protocol (IP)** address - a **globally unique name** for each computer
  - which is a 32 bit address split into four 8-bit octets which identifies the network (1, 2 or 3 octets) and the host machine on that network (3, 2 or 1 octets)
- and a **port** address (a communication channel for transmitting and accepting data) on that computer

A History of the Internet

1959 - **ARPA** set up
1969 - **ARPAnet** starts
early 70s - start of **e-mail**
1979 - **USENET news** begins

1982 - **TCP/IP** established - universal low level protocols

1985 - **DNS** names begin

userlogin @ gateway . department . institution . country
rich @ www . dcs . gla.ac . uk

1986 – **SGML**

early 1990s - **WAIS** set up to help searching for data and **gopher** for accessing files

1991 - The World Wide Web set up, early **HTML**

1993 - The **Mosaic** browser - later becomes Netscape then Firefox

1996 - Microsoft brings out Internet Explorer
XML proposed, **Cascading Style Sheets**

1997 – **HTML Version 4**

2002 – **XHTML**

2005 – **Web 2.0, AJAX, Semantic Web**

World Internet Users and Population Statistics (30/9/2007)

(Source www.internetworldstats.com)

WORLD INTERNET USAGE AND POPULATION STATISTICS

World Regions	Population (2007 Est.)	Population % of World	Internet Usage, Latest Data	% Population Penetration	Usage % of World	Usage Growth 2000-2007
Africa	933,448,292	14.2 %	**43,995,700**	4.7 %	3.5 %	874.6 %
Asia	3,712,527,624	56.5 %	**459,476,825**	12.4 %	36.9 %	302.0 %
Europe	809,624,686	12.3 %	**337,878,613**	41.7 %	27.2%	221.5 %
Middle East	193,452,727	2.9 %	**33,510,500**	17.3 %	2.7 %	920.2 %
North America	334,538,018	5.1 %	**234,788,864**	70.2 %	18.9%	117.2 %
Latin America/ Caribbean	556,606,627	8.5 %	**115,759,709**	20.8 %	9.3 %	540.7 %
Oceania / Australia	34,468,443	0.5 %	**19,039,390**	55.2 %	1.5 %	149.9 %
WORLD TOTAL	6,574,666,417	100.0 %	**1,244,449,601**	18.9 %	100.0 %	244.7 %

World Internet Users and Population Statistics (7/10/2008)

(Source www.internetworldstats.com)

WORLD INTERNET USAGE AND POPULATION STATISTICS

World Regions	Population (2008 Est.)	Internet Users Dec/31, 2000	Internet Usage, Latest Data	% Population (Penetration)	Usage % of World	Usage Growth 2000-2008
Africa	955,206,348	4,514,400	**51,065,630**	5.3 %	3.5 %	1,031.2 %
Asia	3,776,181,949	114,304,000	**578,538,257**	15.3 %	39.5 %	406.1 %
Europe	800,401,065	105,096,093	**384,633,765**	48.1 %	26.3 %	266.0 %
Middle East	197,090,443	3,284,800	**41,939,200**	21.3 %	2.9 %	1,176.8 %
North America	337,167,248	108,096,800	**248,241,969**	73.6 %	17.0 %	129.6 %
Latin America/Caribbean	576,091,673	18,068,919	**139,009,209**	24.1 %	9.5 %	669.3 %
Oceania / Australia	33,981,562	7,620,480	**20,204,331**	59.5 %	1.4 %	165.1 %
WORLD TOTAL	6,676,120,288	360,985,492	**1,463,632,361**	21.9 %	100.0 %	305.5 %

Domain Names

A domain name, such as *catalog.com* or *dcs.gla.ac.uk* is a more human way to identify a web site.

The **Domain Name Service** transforms these into IP addresses.

The domain name includes:

a country identifier at the end (e.g. .uk)

the top level domain it is (e.g. .com or .org)

the specific name, which must be less than 63 characters and use only letters, numbers or hyphens ("-") and not begin or end with a hyphen

A growing number of top level domains are identified:

.com - for commercial and personal sites

.org - recommended for not-for-profit organizations

.net - recommended for companies involved in Internet infrastructure

.biz, .info, .name, .co-op, .pro, .museum, .aero, .tv are other extensions

81,733 top level domains created yesterday in the USA

Electronic Mail

E-mail is now a standard communication medium for sending messages.

A mail program will also have other facilities:

- aliases and mailing lists
- mail boxes
- kill files and other filters
- text encodes and attachments

E-mail requires a new way of communicating.

- It is depersonalised and can easily cause offence
- "Netiquette" - polite ways to behave
- Embedded affect - e.g. stars to embolden, smileys;

Sending multimedia documents

- mail was originally purely textual – now can send “attachments”
- attachments are multimedia files encoded as text for transmission

Encoding Attachments

The important feature is that binary information must be turned into ASCII characters so that these can be sent as if they were ordinary textual messages -

because transfer was traditionally designed for 7-bit characters.

Several techniques exist for this:

- **Rich Text Format** - Microsoft’s way of encoding Word and other files
- **Uuencode** - the UNIX utility which turns binary files into text
- **BinHex** - the Macintosh equivalent to Uuencode
- **MIME** - a set of formats for encoding Multimedia data

Internet Media Types (MIME Types)

MIME (the old name) stands for Multipurpose Internet Mail Extensions - a set of formats for encoding Multimedia data:

this starts with a **header** which describes the contents of the file as:

- **content-type/sub-type** in which the type identifies a multimedia type and the sub-type identifies a particular file format, e.g.:
  e.g. text/plain, image/gif, video/mpeg or audio/wav;
- **content-transfer-encoding** - 7 bit (for text), base64 (just send the bits verbatim), etc.
- **content-id** - unique
- **content-description** - human readable description of the message

it also supports multi-part messages – the content type is "multi-part" and the rest of the message is a tree-like structure for structured documents - attachments become parts in such a document

The legitimate types are registered with IANA – The Internet Assigned Numbers Authority

The Eight Content Types

Type		Sub-Type
Application	The content is data which must be processed by an application	pdf javascript x-shockwave-flash x-www-form-urlencoded
Audio	The content is an audio file	mpeg x-wav
Image	The content is an image file	jpeg gif, png
Message	Used for e-mail and newsgroup messages	
Model	The content is a 3D model	VRML
Multipart	The content has several parts	mixed – for e-mail
Text	The content is human readable text	plain html css xml
Video	The content is a video file	mpeg mp4 quicktime

Network News

Replacing the active sending of messages to a mailing list by

- the sending of messages to a passive repository (bulletin board) where they can be read

There got be a huge number of such newsgroups

71,618 in 1997

Moderation necessary if they are to remain focused and inoffensive

Remote Login

Facility for gaining access to a remote machine - e.g. telnet

File Transfer

File Transfer Protocol (FTP) allows you to go to another site and download files

Often you need an anonymous login

Browsing and Searching

Computers offer two ways of locating information:

browsing is the technique of going from one piece of data to a related one

searching is the technique of requesting data which fits certain criteria

Browsing on the Internet is achieved through the hypermedia structure

Searching requires **Information Retrieval** or **Database** techniques:

IR techniques match an input phrase with file contents

e.g. "The data about Oldham Athletic and hill walking" turns into

"oldham" "athletic" "hill" "walk" since common words are dropped and other words are "stemmed" – "walking" -> "walk"

I would have done done better to enter "… about 'Oldham Athletic' and …." to get a search as just one word

Database techniques use the file structure.

find me the *Football Club* whose *name* is "Oldham Athletic"

The World Wide Web

WWW is a tool for viewing the Internet as a hyperlinked document, but:

- only part of the data files are made available - the **web site**
- the site is accessed via a gateway called a **web server**
- the web pages require a special format - initially HTML
- there is a universal **addressing mechanism** for pages - the URL
- access must be via a **web browser** - Internet Explorer, Opera, Firefox

Hyperlinks

The web functions as a unified whole by the use of hyperlinks.

Each web page will typically have links to other web pages. Links are distinguishable on a web page:

- by being in a different colour or being underlined;
- because the cursor changes shape when over a link.

Links are specified in a web document as **Uniform Resource Locators** (URLs)

Uniform Resource Locators

A URL typically looks like:

aaaa: //bbb.bbb.bbb /ccc/ccc/ccc ?ddd

and consists of four parts:

i) The **Protocol**

The *aaaa*: shown above indicates the message structure

ii) The **Server Address** (Internet Node Address)

This is the domain name and is shown as *//bbb.bbb.bbb*

iii) The **File Path**

The */ccc/ccc/ccc* identifies the desired file on the server

iv) **Parameters**

This is shown as ?ddd and is used to send information such as form data to a program responding to the link.

The term **Uniform Resource Indicator** (URI) is used to mean the address of page which may be either a URL or a local file path.

URL Encoding

The only characters allowed in a URL are

- letters, digits and - _ . + ! ~ * ‘ ( )

Other characters have a special purpose and are **reserved**:

- % $ & + , / : ; = ? @

So any other character **(e.g. space)** needs to be treated specially

- They are replaced by % followed by their ASCII value as two hexadecimal digits
- e.g. space is %20 while % itself is %25

Turning a URL into this version is called **URL encoding**

- and most programming systems have a function for doing this

Internet Protocols

There are a number of protocols which can be used over the internet, among the most important being:

http: - This is the most common protocol and indicates a file that a web browser can format and display - an HTML file, image file, sound file, etc.

file: - this indicates a file which is not in a recognised web format and will be displayed as text

ftp: - file transfer protocol is used to refer to web sites from which files can be extracted and downloaded to the client machines

mailto: - if selected such a link generates a form in which an e-mail message can be constructed and sent to a designated user

news: - the resource is a news group or article

telnet: - generate a telnet session to this server or

WAIS - a Wide Area Information Server, an old free text search system

HTTP – Hypertext Transfer Protocol

When you summon a web page by selecting a link, etc., you:

- send a **request message** to the server
- and get a **response message** back
- the format of these two messages is the HTTP protocol

The sequence of events is:

- turn the URI into an IP address
- send a message to that address on port 80

```
GET nextpage.html HTTP/1.1      for retrieving data, PUT for sending
Connection: close
Host: www.dcs.gla.ac.uk
User-Agent: Mozilla/4.0         if it’s Firefox
```

- gets back:

```
HTTP/1.1 200 OK                 3xx for redirect, 4xx for site error
Date: 1/10/2007                         5xx for server error
Content-Type: text/html
Content-Length: 1234
then the HTML
```


Statelessness

The main problem with HTTP is that it deals with each request **in isolation**

- which is fine if all you are doing is retrieving information
- but no good if you think you are entering into a conversation such an e-commerce transaction

Programming a web site which supports a conversation means

- Creating a **session** to manage the conversation
- Passing the session identifier to and fro
  - by adding it to the URL in the second and subsequent pages
  - by adding a **hidden field** to a form
  - by using **cookies** which are small text files held on the user’s computer and sent back with every request

User Agents

A user agent is a program which runs on a client device and manages the interaction between the user and the server

- for instance a **web browser**

Other user agents are:

- screen readers
- mobile phone interfaces
- web crawlers

Web Browsers

These render a formatted page and allow the user to follow hyperlinks, but also provide:

- a **history**, enabling the user to step forward and backward buttons
- **bookmarks** - files of frequently accessed URLs
- a **cache** of recently visited web pages and used to cut down on network traffic
- integrated **search engines**
- integrated **news readers** and **mail readers**
- **plug-ins** - software components which allow the browser to render other kinds of data
- **programming extensions**, through the use of:
  - **scripting languages** such as JavaScript or VBScript
  - **Java applets** - programs which are loaded onto the client machine and executed there

Search Engines

There a number of web sites (Google, Yahoo, Altavista, etc.) which exist to provide searchable indexes of the web

They employ a program (called a **robot**, **spider** or **web crawler**) which crawls around the web and sends back the files it finds for indexing

The index created sorts searchable words (search terms) into order and lists the set of web pages in which that term occurs

When the user searches for a series of terms, the engine returns the pages in which all (or most) of those terms are found

Ordering of the output is achieved by giving higher priority to rarer terms

The Structure of a Web Site

Each web server manages a web site - a set of files which typically contain:

- **web pages** - directories and files holding displayable web documents
- **files available for transfer** - files which may be uploaded from the web site by remote users using ftp
- **programs** executable on the server by remote invocation over the web, which return dynamically created displayable web documents:
- **data sources** (e.g. databases) used by these programs to hold the content of web pages

The server side programs can use a number of different techniques:

- HTML with embedded program scripts – e.g. ASP, PHP, JSP or ColdFusion
- **Programs which return HTML** as their output, e.g. CGI – the Common Gateway Interface, PHP and Java Servlets

Web 2.0

Web 2.0 is a buzz word which refers to a collection of independent techniques and social uses of the web:

- The techniques include:
  - **AJAX** – the ability to update only part of the page
  - **RSS** and **Podcasts** – to disseminate changing information
  - **Ontologies** – taxonomies to describe the meaning of web data
- The social uses include:
  - **Weblogs** and **Wikis**
  - **Folksonomies** – community developed ontologies

Mash-ups are web programs which combine a number of components provided separately

- e.g. Chicago Crime Map
- using Google Maps as a background is typical

These techniques jointly make web sites seem like standard desktop applications – e.g. MLB Gameday

Ubiquitous Computing

The web was designed to be used with desktop computers

Now computing capability is built into a range of devices by a range of users

- PDAs, Interactive TV, Mobile Phones and Touchtone phones
- **Sensors** recording information about an environment – location, temperature
- **Actuators** or **effectors** which changing device settings – e.g. a thermostat
- Use by users with hearing or site impairment

These can all be programmed in a similar way to the web although:

- the user interface varies and can be
  - a smaller screen or audio only
- will use different (but similar) protocols
  - WML for mobile phones, VoiceXML for touchtone phones

There is a need to create a disciplined multi-device ubiquitous development environment

The World Wide Web and Mark-up Languages

The World Wide Web is a **distributed hypermedia application** running on the **Internet** using **mark-up** to achieve the **hyperlinking**

- Distributed Application - one program running on a number of computers
- The Internet - the inter-network connecting most of the world's computers
- Hypermedia Application - one in which the user can navigate around a document using hyperlinks
- Hyperlink - a selectable fragment of a document which if selected moves the point of view to another document or another point in the same document
- Mark-up - the ability to identify fragments of a document as having a particular meaning

For more information look in the web site of the organisation which controls web development - **www.w3c.org**

SGML

The Standard Generalised Markup Language (SGML) is somewhat misnamed, because it is not itself a mark-up language, but rather a **meta-language** with which mark-up languages can be described

All SGML described **mark-up languages** have a similar structure:

Documents are **tagged**, i.e. subsets of the text are delimited by tags (a word surrounded by angle brackets) which indicates its structure

The whole text then consists of **hierarchically organised** tagged sub-sections - e.g. book-chapter-section-sub-section-paragraph

The whole document is the outermost element of the hierarchy

The tagged section are called **elements**

The kinds of element allowed are called **element types**

The elements can be parameterised by the use of **attributes**

e.g.

which introduces an image into a web page and has the name of its file as an attribute

An Example

Here is a short piece of XHTML to illustrate this:

```
<p>This is a <em>paragraph</em> in which the word "paragraph" has been emphasised</p>
```

In this example, the word paragraph has been tagged with an indication that is to be emphasised and the whole structure is tagged as being a paragraph:

- The tags "p" and "em" are the element types
- The whole text and paragraph are the elements

The full definition of the mark-up language is called the Document Type Definition (DTD) for the language

Thus the DTD for HTML includes definitions for <p> and

Examples from the XHTML Strict DTD

An HTML element has a head and body, might have an ID attribute and is tied to a central namespace

```
<!ELEMENT html (head, body) >
    <!ATTLIST html
          id ID #IMPLIED
          xmlns %URI; #FIXED 'http://www.w3.org/1999/xhtml' >
```

```
<!ELEMENT ul (li)+>
    <!ATTLIST ul
          %attrs; >
    <!ENTITY % attrs
          "id ID;                #IMPLIED
          class CDATA;           #IMPLIED
          style %StyleSheet;     #IMPLIED
          title  %Text;          #IMPLIED" >
```

 is made up of one or more 's and has the standard set of attributes defined as an entity so it can be referenced in many places

```
<!ELEMENT img EMPTY>
<!ATTLIST img
          %attrs;
          src %URI; #REQUIRED
          alt %Text; #REQUIRED
          longdesc %URI; #IMPLIED
          height %Length; #IMPLIED
          width %Length; #IMPLIED
          usemap %URI; #IMPLIED>
```

<img> has no content but must have src and alt attributes

HTML

The principal language for structuring web pages is a family of languages instantiated from SGML called HTML - HyperText Markup Language.

HTML is primarily designed to specify the structure of the page and not to be explicit about how it should look.

This means that different browsers can display the same file in different ways - i.e. the language is hardware and system independent.

Like most computer languages, HTML has gone through a number of versions:

HTML 2 - was the first important standard which was fixed in 1994

HTML 3 - was an upgraded version created in 1996

HTML 4 - is the current version released in 1997

XHTML is a reformulation of HTML in XML - you must use this for your coursework

A Typical HTML Page