SWE 444
Internet and Web Application Development

Dr. Abdallah Al-Sukairi
sukairi@kfupm.edu.sa

Second Semester 2003 - 2004 (032)

King Fahd University of Petroleum & Minerals
Information & Computer Science Department

Course Outline (tentative)

1. Basic Internet Concepts
2. HTML
3. XHTML
4. CSS (Style Sheets)
5. Client-Side Scripting (JavaScript)
6. XML, XSL, XSLT, DTD, DOM, XSD, XPath, XForms
7. WAP (Wireless Application Protocol)
8. Server Side Scripting
9. Server Side Applications
10. Web Services
11. Web Security
12. Web Servers (Hosting)
What this course is not

… there is a difference between training and education. If computer science is a fundamental discipline, then university education in this field should emphasize enduring fundamental principles rather than transient current technology.


Grading

10% homework
10% project I
20% project II
15% exam I
20% exam II
25% final exam
Warning

- Demanding course
- No textbook
- Many different topics
- Large project component
- Field changes quickly
  - Each year is essentially a new course

Course Materials

- No textbooks
- WWW
  - http://www.ccse.kfupm.edu.sa/sukairi/swe444
  - Most content is on the web
  - Lectures are on the web
  - I’m setting up WebCT
Abbreviated History

1943  First electronic digital computer Harvard Mark I
1966  Design of ARPAnet
1970  ARPAnet spans country, has 5 nodes
1971  ARPAnet has 15 nodes
1972  First email programs, FTP spec
1973  Ethernet operation at Xerox PARC
1974  Intel launches 8080; TCP design
1975  Gates/Allen write Basic for Altair 8800
1976  Apple Computer formed by Jobs/Wozniak
1977  111 hosts on ARPAnet
1979  Visicalc
... Abbreviated History

1981  Microsoft has 40 employees; IBM PC
1982  Sun formed
1983  ARPAnet uses TCP/IP -> birth of internet
1983  Design of DNS
1984  launch of Macintosh; 1000 hosts on ARPAnet
1985  Symbolic.com first registered domain name
1989  100,000 hosts on Internet
1990  Cisco Systems goes public $288 M
      Tim Berners-Lee creates WWW at CERN

1993  Mosaic developed at UIUC
      Web grows by 341,000% in a year
1994  Netscape, Amazon, Archtext formed
1995  Netscape, Windows 95, MetaCrawler
1997  Amazon
2000  Internet “bubble” bursts

Jan 2003  171,638,297: Number of Hosts advertised in the DNS
        (Source: http://www.isc.org/)
Web Server Survey

- In the September 2003 survey Netcraft received responses from 43,144,374 sites (Source: http://news.netcraft.com/)

- Market Share for Top Servers

How Many Online?

- 605.60 million is an "educated guess" as to how many are online worldwide as of September 2002 (Source: http://www.nua.com/)

<table>
<thead>
<tr>
<th>Region</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>World Total</td>
<td>605.60 million</td>
</tr>
<tr>
<td>Europe</td>
<td>190.91 million</td>
</tr>
<tr>
<td>Asia/Pacific</td>
<td>187.24 million</td>
</tr>
<tr>
<td>Canada &amp; USA</td>
<td>182.67 million</td>
</tr>
<tr>
<td>Latin America</td>
<td>33.35 million</td>
</tr>
<tr>
<td>Africa</td>
<td>6.31 million</td>
</tr>
<tr>
<td>Middle East</td>
<td>5.12 million</td>
</tr>
</tbody>
</table>
How Many Online (by Language)

(Source: http://www.glreach.com/globstats/)

Web Content (by language)

- Source: http://www.vilaweb.com /

<table>
<thead>
<tr>
<th>Language</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>68.4%</td>
</tr>
<tr>
<td>Japanese</td>
<td>5.9%</td>
</tr>
<tr>
<td>German</td>
<td>5.8%</td>
</tr>
<tr>
<td>Chinese</td>
<td>3.9%</td>
</tr>
<tr>
<td>French</td>
<td>3.0%</td>
</tr>
<tr>
<td>Spanish</td>
<td>2.4%</td>
</tr>
<tr>
<td>Russian</td>
<td>1.9%</td>
</tr>
<tr>
<td>Italian</td>
<td>1.6%</td>
</tr>
<tr>
<td>Portuguese</td>
<td>1.4%</td>
</tr>
<tr>
<td>Korean</td>
<td>1.3%</td>
</tr>
<tr>
<td>Other</td>
<td>4.6%</td>
</tr>
<tr>
<td>Total Web pages:</td>
<td>3.30 B</td>
</tr>
</tbody>
</table>
Number of Internet Users in KSA

- According to Internet Services Unit (Source: http://www.isu.net.sa/)

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>April 2001</td>
<td>690,000 users</td>
</tr>
<tr>
<td>Dec 2001</td>
<td>900,000 users</td>
</tr>
<tr>
<td>July 2002</td>
<td>1,110,000 users</td>
</tr>
<tr>
<td>Dec 2002</td>
<td>1,453,000 users</td>
</tr>
</tbody>
</table>

- Assumptions
  - Estimated number of users per a 64kbps line is 20 users
  - User to dialup subscriber ratio is estimated at 2.5

Structure of the Internet

![Internet Structure Diagram](image)

SOURCE: CISCO SYSTEMS
Internet Backbone Structure

- Level 1 (interconnect level, NAPs)
  - billions of pages per day
- Level 2 (national backbone, MAE, FIX)
  - Federal Internet eXchange Points
  - Peering agreements: connect, share routing info
- Level 3 (regional providers, state level)
- Level 4 (local ISP)
- Level 5 (companies, individuals)
- Level 6 (routers)

The World Wide Web

- A way to access and share information
  - Technical papers, marketing materials, recipes, ...
- A huge network of computers: the Internet
- Graphical, not just textual
- Information is linked to other information
- Application development platform
  - Shop from home
  - Provide self-help applications for customers and partners
  - ...
WWW Architecture

Client/Server, Request/Response architecture

- You request a Web page
  - e.g. http://www.msn.com/default.asp
  - HTTP request
- The Web server responds with data in the form of a Web page
  - HTTP response
  - Web page is expressed as HTML
- Pages are identified as a Uniform Resource Locator (URL)
  - Protocol: http
  - Web server: www.msn.com
  - Web page: default.asp
  - Can also provide parameters: ?name=Leon
Web Standards

- Internet Engineering Task Force (IETF)
  - Founded 1986
  - A large open international community of network designers, operators, vendors, and researchers concerned with the evolution of the Internet architecture and the smooth operation of the Internet
  - It is open to any interested individual

- World Wide Web Consortium (W3C)
  - [http://www.w3.org](http://www.w3.org)
  - Founded 1994 by Tim Berners-Lee
  - An open forum of companies and organizations with the mission to lead the Web to its full potential
  - W3C has around 450 Member organizations from all over the world
  - Publishes technical reports and recommendations
  - The rule-making body of the Web is the W3C
  - W3C puts together specifications for Web standards
  - The most essential Web standards are HTML, CSS and XML

Web Design Principles

- **Interoperability**: Web languages and protocols must be compatible with one another independent of hardware and software

- **Evolution**: The Web must be able to accommodate future technologies. Encourages simplicity, modularity and extensibility

- **Decentralization**: Facilitates scalability and robustness
Hypertext Markup Language (HTML)

- The markup language used to represent Web pages for viewing by people
  - Designed to display data, not store/transfer data
- Rendered and viewed in a Web browser
- Can contain links to images, documents, and other pages
- Not extensible
- Derived from Standard Generalized Markup Language (SGML)
- HTML 3.2, 4.01, XHTML 1.0

HTML Forms

- Enables you to create interactive user interface elements
  - Buttons
  - Text boxes
  - Drop down lists
  - Check boxes
- User fills out the form and submits it
- Form data is sent to the Web server via HTTP when the form is submitted
Hypertext Transport Protocol (HTTP)

- The top-level protocol used to request and return data
  - e.g. HTML pages, GIFs, JPEGs, Microsoft Word documents, Adobe PDF documents, etc.

- Request/Response protocol

- Methods: GET, POST, HEAD, ...

- HTTP 1.0: simple

- HTTP 1.1: more complex

HTTP

- HTTP is a stateless protocol

- Each HTTP request is independent of previous and subsequent requests

- HTTP 1.1 introduced keep-alive for efficiency

- Statelessness has a big impact on how scalable applications are designed
HTTP Server Status Codes

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>OK</td>
</tr>
<tr>
<td>201</td>
<td>Created</td>
</tr>
<tr>
<td>301</td>
<td>Moved Permanently</td>
</tr>
<tr>
<td>302</td>
<td>Moved Temporarily</td>
</tr>
<tr>
<td>400</td>
<td>Bad Request – not understood</td>
</tr>
<tr>
<td>401</td>
<td>Unauthorized</td>
</tr>
<tr>
<td>403</td>
<td>Forbidden – not authorized</td>
</tr>
<tr>
<td>404</td>
<td>Not Found</td>
</tr>
<tr>
<td>500</td>
<td>Internal Server Error</td>
</tr>
</tbody>
</table>

What happens when you click?

- Suppose
  - You are at www.yahoo.com/index.html
  - You click on autos.yahoo.com
- Browser uses DNS => IP addr for autos.yahoo.com
- Opens TCP connection to that address
- Sends HTTP request
- Receives HTTP Response
- One click => several responses
- HTTP1.1: KeepAlive - several requests/connection
HTTP Request

Method | File | HTTP version | Headers
--- | --- | --- | ---
GET /default.asp | HTTP/1.0 | | Accept: image/gif, image/x-bitmap, image/jpeg, */*
Accept-Language: en
User-Agent: Mozilla/1.22 (compatible; MSIE 2.0; Windows 95)
Connection: Keep-Alive
If-Modified-Since: Sunday, 17-Apr-96 04:32:58 GMT

HTTP Response

HTTP version | Status code | Reason phrase | Headers
--- | --- | --- | ---
HTTP/1.0 | 200 | OK | Date: Sun, 21 Apr 1996 02:20:42 GMT
Server: Microsoft-Internet-Information-Server/5.0
Connection: keep-alive
Content-Type: text/html
Last-Modified: Thu, 18 Apr 1996 17:39:05 GMT
Content-Length: 2543

Data

<HTML> Some data... blah, blah, blah </HTML>
Cookies

- A mechanism to store a small amount of information (up to 4KB) on the client
- A cookie is associated with a specific web site
- Cookie is sent in HTTP header
- Cookie is sent with each HTTP request
- Can last for only one session (until browser is closed) or can persist across sessions
- Can expire some time in the future

HTTPS

- A secure version of HTTP
- Allows client and server to exchange data with confidence that the data was neither modified nor intercepted
- Uses Secure Sockets Layer (SSL)/Transport Layer Security (TLS)
URIs, URLs and URNs

- Uniform Resource Identifier (URI = URL or URN)
  - Generic term for all textual names/addresses

- Uniform Resource Locator (URL)
  - The set of URI schemes that have explicit instructions on how to access the resource over the Internet, e.g. http, ftp, gopher

- Uniform Resource Name (URN)
  - A URI that has an institutional commitment to availability, etc.
  - A particular scheme intended to identify resources e.g. urn:schemas:httpmail:subject

Multipurpose Internet Mail Extensions (MIME)

- Defines types of data/documents
  - text/plain
  - text/html
  - image/gif
  - image/jpg
  - audio/x-pn-realaudio
  - audio/x-ms-wma
  - video/x-ms-asf
  - application/octet-stream
**MIME**

- Specifies character sets, e.g. ASCII
- Supports multi-part messages
- Originally designed for email, but also used in other places, such as HTTP

**Browsers**

- Client-side application
- Requests HTML from Web server and renders it
- Popular browsers:
  - Internet Explorer
  - Netscape
  - Opera
  - others
- Also known as a User Agent
Clients & Servers

- **Clients**
  - Generally supports a single user
  - Optimized for responsiveness to user
  - User interface, graphics

- **Servers**
  - Supports multiple users
  - Optimized for throughput
  - More: CPUs (SMP), memory, disks (SANs), I/O
  - Provide services (e.g. Web, file, print, database, e-mail, fax, transaction, telnet, directory)

Proxy Servers & Firewalls

- **Proxy Server**
  - A server that sits between a client (running a browser) and the Internet
  - Improves performance by caching commonly used Web pages
  - Can filter requests to prevent users from accessing certain Web sites

- **Firewall**
  - A server that sits between a network and the Internet to prevent unauthorized access to the network from the Internet
Networks

- Network scope
  - Internet: a specific world-wide network based on TCP/IP, used to connect companies, universities, governments, organizations and individuals
  - Intranet: a network based on Internet technologies that is internal to a company or organization
  - Extranet: a network based on Internet technologies that connects one company or organization to another

- Network technology
  - Broadcasting
    - Packets of data are sent from one machine and received by all computers on the network
    - Multicast: packets are received by a subset of the machines on a network
  - Point-to-point
    - Packets have to be routed from one machine to another; there may be many paths
  - In general, geographically localized networks use broadcasting, while disperse networks use point-to-point
Network Protocol Stack

Networks - Internet Layer

- Internet Protocol (IP)
  - Responsible for getting packets from source to destination across multiple hops
  - Not reliable
  - IP address: 32 bit value usually written in dotted decimal notation as four 8-bit numbers (0 to 255); e.g. 130.50.12.4
Networks - Transport Layer

- Provides efficient, reliable and cost-effective service
- Uses the Sockets programming model
- Ports identify application
  - Well-known ports identify standard services (e.g. HTTP uses port 80, SMTP uses port 25)
- Transmission Control Protocol (TCP)
  - Provides reliable, connection-oriented byte stream
- UDP
  - Connectionless, unreliable

Networks - Application Layer

- Telnet: Remote sessions
- File Transfer Protocol (FTP)
- Network News Transfer Protocol (NNTP)
- Simple Network Management Protocol (SNMP)
- Simple Mail Transfer Protocol (SMTP)
- Post Office Protocol (POP3)
- Interactive Mail Access Protocol (IMAP)
Networks - Domain Name System (DNS)

- Provides user-friendly domain names, e.g. www.msn.com

- Hierarchical name space with limited root names
  
  - .com
  - .net
  - .gov
  - .edu
  - .org
  - .mil
  - .jp
  - .sa

- DNS servers map domain names to IP addresses

Extensible Markup Language (XML)

- Represents hierarchical data

- A meta-language: a language for defining other languages

- Extensible

- Useful for data exchange and transformation

- Simplified version of SGML
Client-Side Code

- What is client-side code?
  - Software that is downloaded from Web server to browser and then executes on the client

- Why client-side code?
  - Better scalability: less work done on server
  - Better performance/user experience
  - Create UI constructs not inherent in HTML
    - Drop-down and pull-out menus
    - Tabbed dialogs
  - Cool effects, e.g. animation
  - Data validation

Client-Side Technologies

- DHTML/JavaScript
- COM
  - ActiveX controls
  - COM components
  - Remote Data Services (RDS)
- Java
- Plug-ins
- Helpers
- Remote Scripting
Server-Side Code

- What is server-side code?
  - Software that runs on the server, not the client
  - Receives input from
    - URL parameters
    - HTML form data
    - Cookies
    - HTTP headers
  - Can access server-side databases, e-mail servers, files, mainframes, etc.
  - Dynamically builds a custom HTML response for a client

Server-Side Code

- Why server-side code?
  - Accessibility
    - You can reach the Internet from any browser, any device, any time, anywhere
  - Manageability
    - Does not require distribution of application code
    - Easy to change code
  - Security
    - Source code is not exposed
    - Once user is authenticated, can only allow certain actions
  - Scalability
    - Web-based 3-tier architecture can scale out
Server-Side Technologies

- Common Gateway Interface (CGI)
- Internet Server API (ISAPI)
- Netscape Server API (NSAPI)
- Active Server Pages (ASP)
- Java Server Pages (JSP)
- Personal Home Page (PHP)
- Cold Fusion (CFM)
- ASP.NET

Web Services

- A programmable application component accessible via standard Web protocols
- The center of the .NET architecture
- Exposes functionality over the Web
- Built on existing and emerging standards
  - HTTP, XML, SOAP, UDDI, WSDL, …
Evolution of the Web

- Generation 1: Static HTML
  - HTML

- Generation 2: Web Applications
  - HTML, XML

- Generation 3: Web Services
  - HTML, XML

Search Engine vs Directory vs...

- How do you find information on the Web?
  - Google
  - Teoma
  - altheweb
  - altavista
  - ??????
Standard Web Search Engine Architecture

- Crawl the web
- Store documents, check for duplicates, extract links
- Create an inverted index
- Inverted index

Reading List

- History of the WWW
- Glossary of Internet Terms
- HTTP Made Really Easy
- Internet Search Techniques