Poseidon House
Castle Park
Cambridge CB3 0RD
United Kingdom

TELEPHONE: **Cambridge (01223) 515010**
INTERNATIONAL: **+44 1223 515010**
FAX: **+44 1223 359779**
E-MAIL: **apm@ansa.co.uk**

Training

ANSAwise - Introduction to the WWW and Java

Mark Madsen

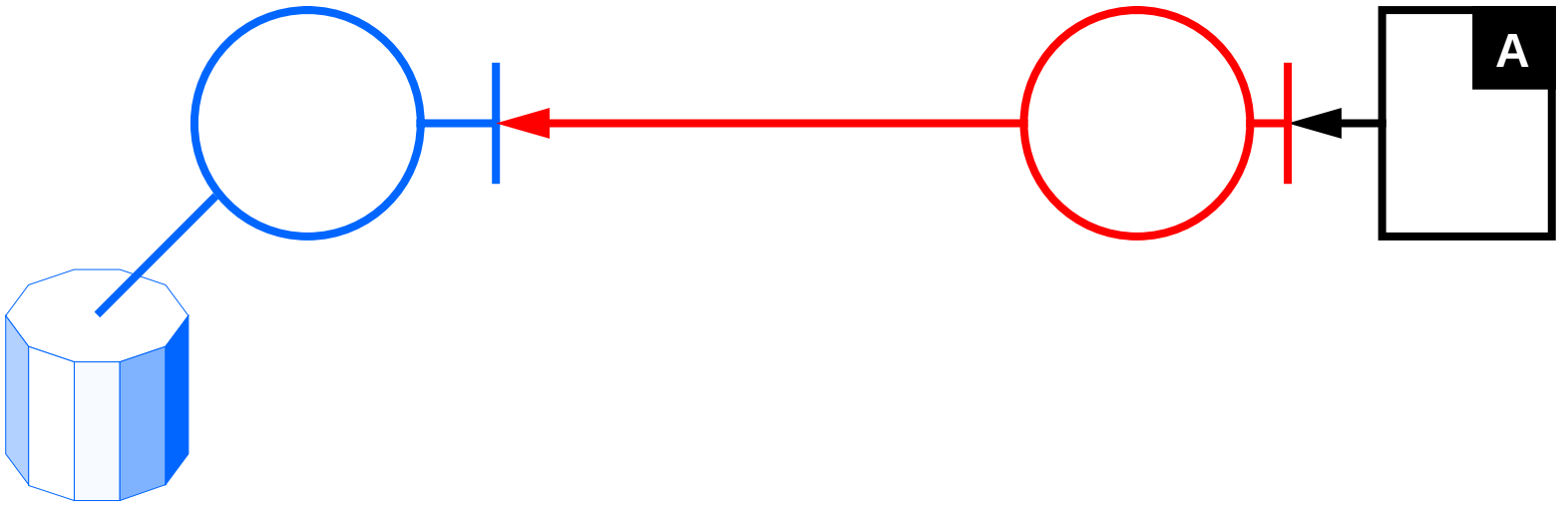
Abstract

The WWW has become the paradigmatic distributed information system. To understand how it works, one must take into account the functionality of the page description language HTML and the transport protocol HTTP used throughout the WWW. The WWW is also undergoing a period of revolutionary development with the introduction of Java as a programming language offering remote execution and security facilities.

This module introduces fundamental concepts of the WWW and Java, and shows how Java is designed for programming behaviour in a widely distributed system.

APM.1721.01 **Approved** 4th April 1996
Briefing Note

Distribution:
Supersedes:
Superseded by:



Introduction to the WWW and Java

In this session

- Describe the World Wide Web (WWW) and its essential components
  - HyperText Transfer Protocol (HTTP)
  - HyperText Markup Language (HTML)
  - Java and JavaScript
- How the WWW is affecting the development of distributed systems
- What effect this will have on electronic businesses

What is the WWW?

- The WWW is the largest information system ever created
- It is truly distributed
  - there is no centre
- The essentials are standardised by use and the IETF

Precise Definition Not Possible

- No narrow definition of the WWW entirely fits
- The WWW includes and simplifies access to
  - ftp
  - Gopher
  - Mail
  - News
- For the purposes of this presentation, the WWW is
  - A network of webservers and browsers...
  - ...using HTTP+HTML+Java

HTTP

- HTTP is the HyperText Transfer Protocol
- Defined by an ongoing series of internet-drafts
  - latest version is HTTP 1.1
- Built to on top of TCP/IP, and using TCP/IP transparently
  - TCP/IP is the Transport Control Protocol/Internet Protocol
- HTTP is *connectionless*
  - no state is stored between connections
  - alternatively: connections do not remember previous connections

HTML

- HTML is the HyperText Markup Language
  - originally defined as a subset of SGML
  - some vendors have made nonstandard extensions
- Defined by an ongoing series of internet-drafts
  - latest version is HTML 3.0
  - HTML 2.0 is currently heading for becoming a standards track RFC
- HTML does not define presentation of a document rigidly
  - exact presentation is under control of user

WWW Browsers

- The browser is a client in distributed systems terminology
- Responsible for retrieval and presentation of HTML documents
- Communicates with webserver via HTTP for retrieval
- The user's local machine resources are used for the presentation

Web Servers

- Webserver delivers documents in response to browser requests
- Server is a *process* that listens on a predefined port
- When it hears an incoming HTTP request it
  - forks a new child process to handle the request
  - hands the request and the connection to the child
  - resumes listening on its port
- The child process then
  - locates the requested document
  - returns it to the browser by HTTP
  - drops the connection and expires

The WWW Reference Library

- This is a suite of code for developing WWW applications
- Originally developed at CERN
  - where it was called the CERN library
  - has also been known as the Library of Common Code
- It is now owned and controlled by W3C (WWW Consortium)
- The library is supplied in source form
- Freely available by anonymous ftp from W3C

The Common Gateway Interface

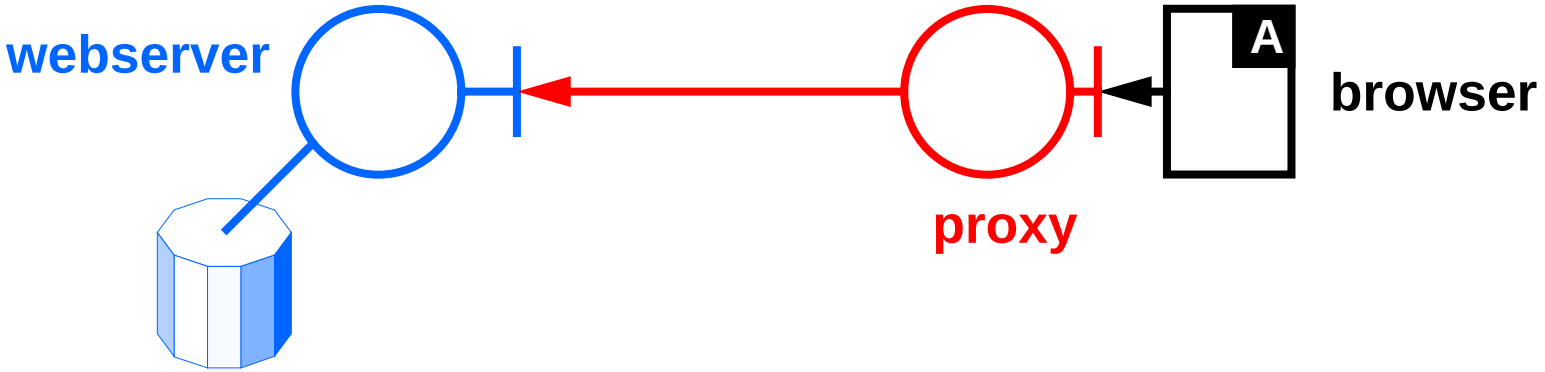
- Many services are based on delivery of information that must be
  - live: delivered in real time
  - customised on a per-use basis
- CGI was developed to allow the server to handle these cases
- CGI mechanism is standardised and controlled by NCSA

CGI Programs

- CGI programs are pieces of code runnable by the webserver
- They are specially privileged
  - the webserver can only run them from its cgi-bin directory
  - they must be extremely well trusted before installation
- They can be written in any programming language
  - the WWW community favours scripts
  - since these can easily be distributed and shared
  - and checked for security holes

The Proxy Mechanism

- Browsers usually communicate directly with webservers
- Security requires that Internet connections be controlled
- Implemented using the proxy mechanism



- The diagram shows this in an object-based view of the WWW

How Proxies Work

- The normal scenario is transparent to the browser
  - browser passes request to proxy
  - proxy requests document
  - server returns document to proxy
  - proxy returns document to browser
- Advantages of proxying are
  - many browsers can share one proxy
  - proxy comprises single point of control for security
  - proxy can cache heavily used documents effectively

Introduction to Java

- Java is a normal programming language
- Basically a lot like C++
  - with a few improvements added...
  - ...and a few undesirable features removed
- Java is special with regard to the WWW because
  - it runs on a virtual machine, so is platform independent
  - bytecodes for the virtual machine can be moved across the WWW...
  - ...and run remotely, in the user's browser

Java Language Design

- The Java language specification is public
  - although owned and controlled by Sun
- Java is object-oriented
  - uses C++ based ideas and syntax
- Java omits C++ assumptions about memory space structure
  - no pointer arithmetic or conversions
- Java is designed for distributed systems technology

Java Language Properties

- Simple
  - Object-orientation
  - Distributed
- Robust
  - Secure
- Architecture Neutral
  - Portable
- Interpreted
  - High Performance
  - Multithreaded
  - Dynamic

Java Application Anatomy

```
import java.util.Date;
        class DateApp {
            public static void main (String args[]) {
                Date today = new Date();
                System.out.println(today);
            }
        }
```

- This example, DateApp, is from Sun's Java Tutorial
- Main concepts shown are
  - class membership hierarchy (“.” is membership operator)
  - type hierarchy (DateApp inherits “Date” type from java.util)

Java Virtual Machine and Byte Codes

- Java compiler generates byte code for a virtual machine
- This virtual machine is the Java byte code interpreter
- The interpreter is architecture neutral
  - widely ported (mostly by third parties)
  - no implementation dependencies permitted
- Only byte codes are shipped from webserver to browser
  - first verified by compiler
  - source code is not distributed

JavaScript

- Purely interpreted version of the Java language
  - no precompilation to bytecodes
  - slower performance
  - simpler to prototype applets
- Developed jointly by Netscape and Sun
- Actual language syntax not identical with Java
- Intended for WWW authors to create pages that use Java applications

Verification and Security in Java

- Java is intended for networked distributed environments
- Language restricted to disable most illegal memory accesses
- Byte codes verified for safety during compilation
- Byte codes authenticated using public key encryption techniques
- Byte codes executed inside a safe interpreter
  - client can restrict access to local filesystem and communications

WWW Security

- Web sites must guard against security compromise
- Generic problems
  - remote filesystem accesses
  - remote client executing privileged programs on webserver
- Wide range of holes must be plugged

Known Risks

- **Server configuration**
  - server should not run with privileges
  - should not have filesystem access outside document tree
- **CGI mechanism**
  - scripts are run with server privileges
  - scripts should be checked before installation
  - scripts must not use pipelines
- **Java/JavaScript**
  - interpreter must not have access to filesystem
  - interpreter must not have access to communications system

Firewall Defences

- Firewalls are the standard defence for webservers
- Two-firewall configurations are favoured
- Webserver is placed inside outer firewall
  - protected by packet filters
- Corporate Intranet is behind an inside firewall
  - protected with a huge array of defences

Summary

- The WWW is a global distributed system consisting of
  - webservers
  - browsers
  - proxies
  - communicating using HTTP
- The WWW is a global information system consisting of
  - HTML documents
  - CGI scripts
- The WWW is programmable in Java
  - Java code is mobile
  - Java programs are handled as if they were web documents

More information?

- For more general information on WWW searching
  - <URL:http://www.altavista.digital.com/>
  - <URL:http://www.yahoo.com/>
- For more on HTTP
  - <URL:http://www.w3.org/pub/WWW/Protocols/>
- For more on HTML
  - <URL:http://www.w3.org/pub/WWW/MarkUp/>
- For more on CGI
  - <URL:http://hoohoo.ncsa.uiuc.edu/cgi/>
- For more on the Java language
  - <URL:http://java.sun.com/>
  - <URL:http://java.sun.com/tutorial/index.html>

Even More Information

- **For more on Firewalls**
  - **see *Firewalls and Internet Security* by Cheswick & Bellovin (Addison-Wesley 1994).**
- **For more on relevant IETF standards activities**
  - **<URL:http://www.ietf.org/>**
- **For more on W3C activities**
  - **<URL:http://www.w3.org/>**