15.561
Information Technology Essentials

Session 8
Web Technologies

Acknowledgments:
Adapted from Chris Dellarocas, U. Md..

World-Wide-Web or The Triumph of Anarchy

- **Perhaps the most important human technological artifact that evolved more or less *ad-hoc***
- **Limited original vision of the WWW has *very little* to do with today's impressive reality**
- **Web Users have consistently innovated in figuring out new ways of leveraging this powerful medium**
- **Web architects then try to *catch up* by extending (read "patching") the Web infrastructure to support these new uses**

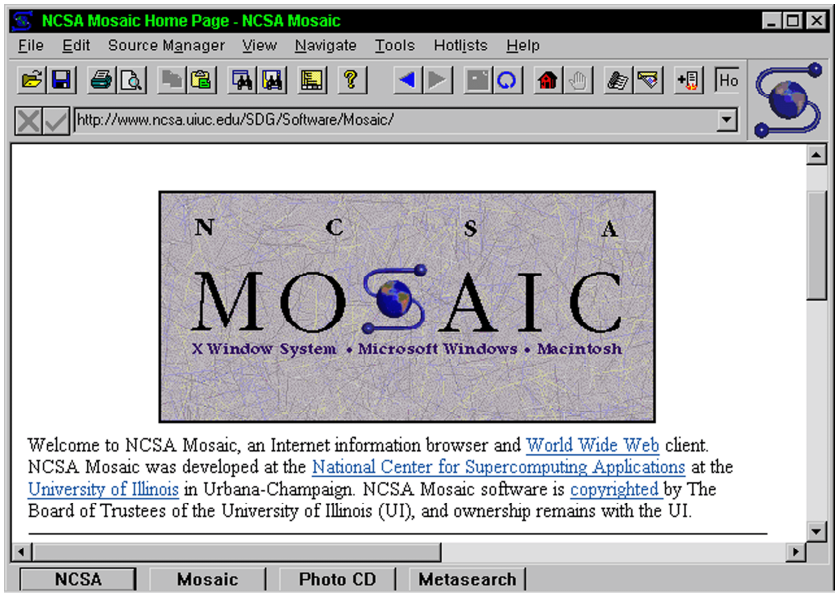
How it all started…
The Web as a Static Document Repository

- Tim Berners-Lee's original vision for the WWW (circa 1989)
- An easy way to access cross-linked static documents stored in a variety of servers around the world
- Initial specification defined:
  - A language for formatting such documents (HTML)
  - A simple protocol for communicating between browsers and servers (HTTP)

The turning point

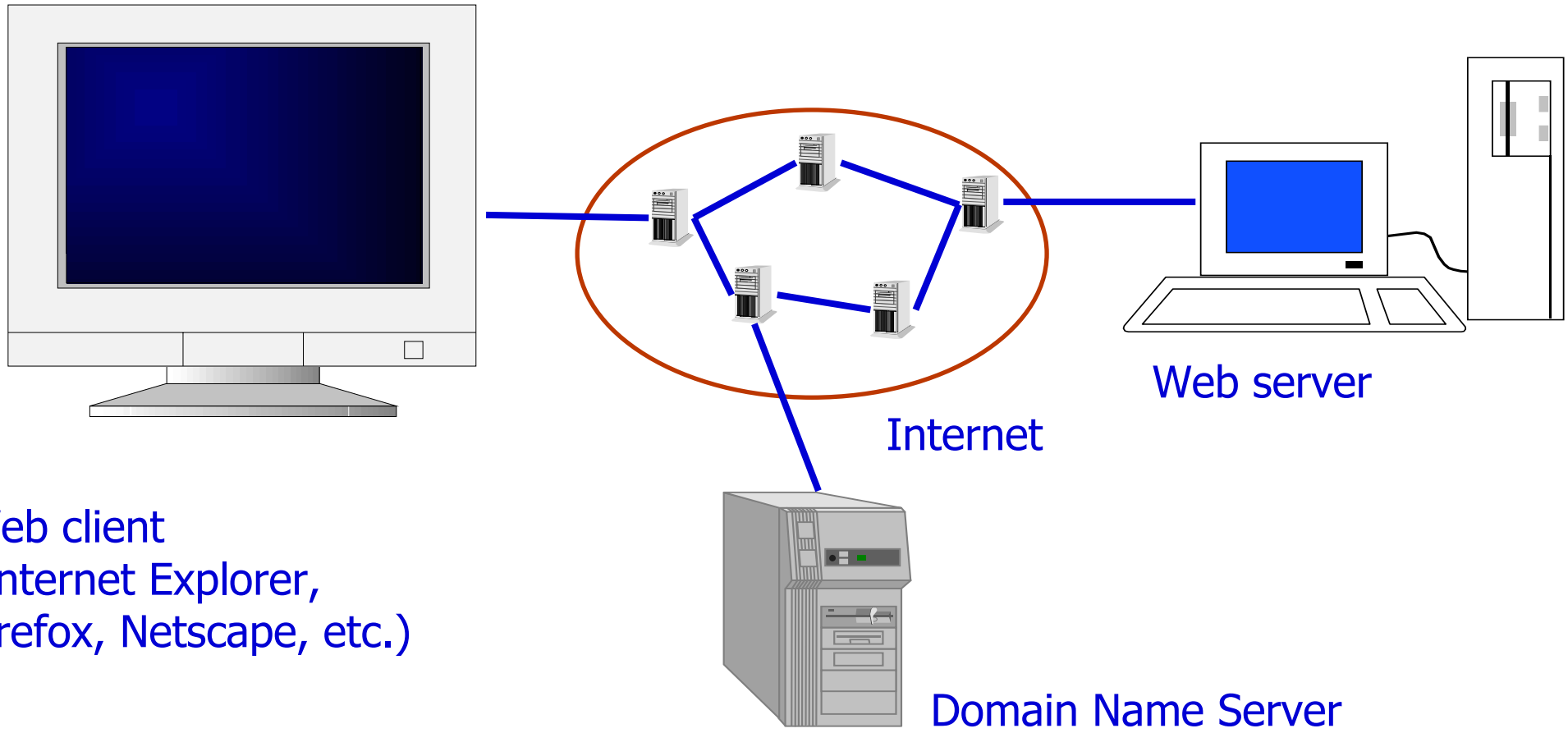
1993 - Marc Andreesen (student at UIUC) writes Mosaic – first graphical WWW browser

(precursor of Netscape)



How the (original) Web works

Open Location: http://web.mit.edu/sloan/www/index.html



Anatomy of a URL

URL = Uniform Resource Locator

http://web.mit.edu/sloan/www/index.html

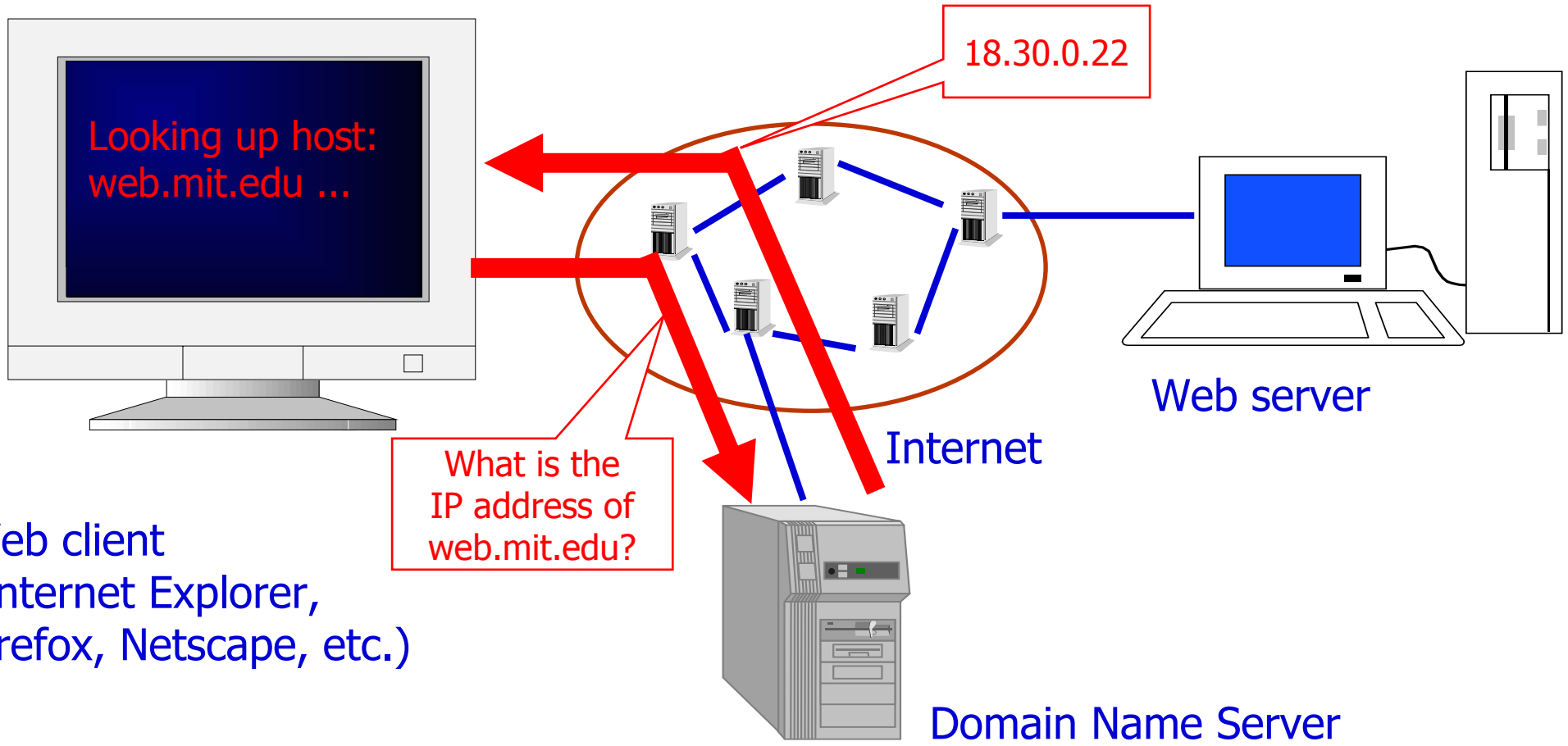
Name of protocol for communication with server (http is standard web protocol)

Domain name of web server where page is stored

Pathname of file within web server's local file system

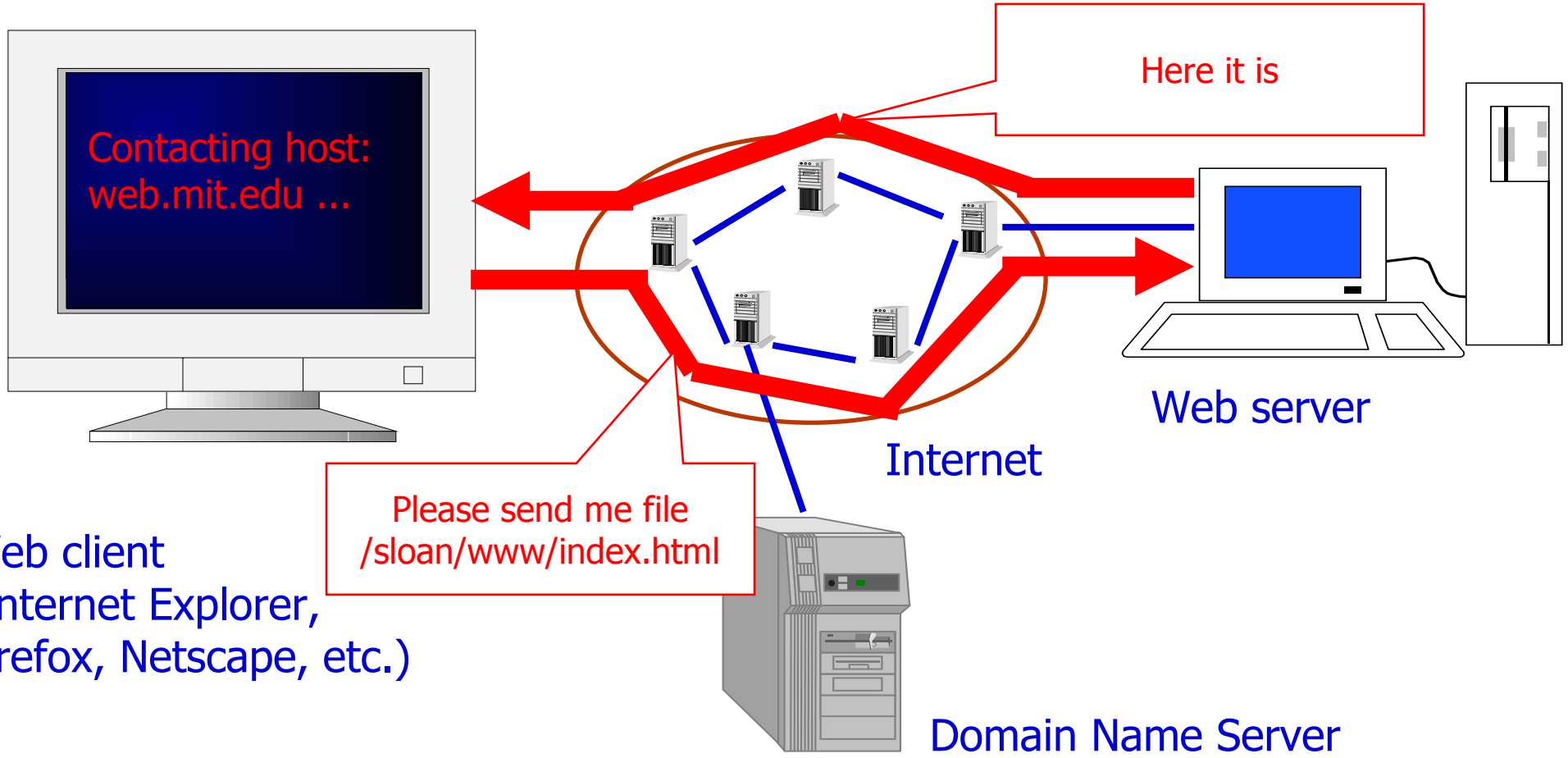
How the Web works

Open Location: http://web.mit.edu/sloan/www/index.html



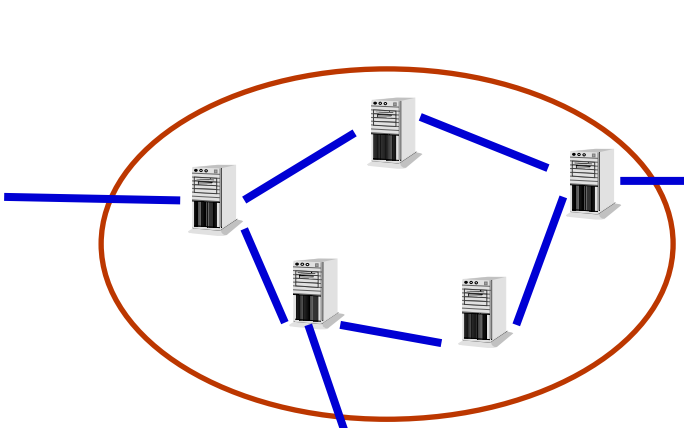
How the Web works

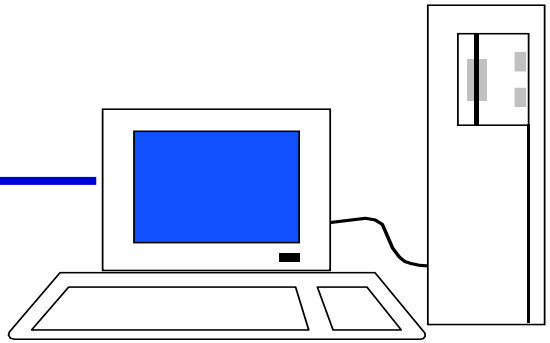
Open Location: http://web.mit.edu/sloan/www/index.html



How the Web works

Open Location: http://web.mit.edu/sloan/www/index.html

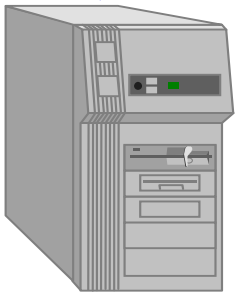




Web server

Internet

Web client
(Internet Explorer,
Firefox, Netscape, etc.)



Domain Name Server

WWW is a Client/Server System

- **Web Clients**
  - Use HTTP protocol to connect to servers
  - Request and display Web pages stored in servers
  - Typical clients: Web browsers
- **Web Servers**
  - Listen for incoming connections from clients
  - Use HTTP protocol to converse with clients
  - Store and transmit Web pages to clients

Summary: WWW The Original Concept (1989-1995)

- Human Readers ← → Interconnected Static Documents
- Main advantage:
  - Universality
- Main disadvantage
  - Lack of interactivity
  - Yet another mass broadcast medium
- Main business use
  - Awareness building

Evolution of the WWW

Business Drivers

- **Enable transactions**
- Allow interactivity between browser and server
- Facilitate personalization
- Support multiple browsing devices
- Better organize and retrieve Web content
- Support Business-to-Business applications

The Web as a transaction facilitator

- **Business Motivation: Low-cost front-end for allowing customers to connect to corporate computers**
  - **Customer registration/Address changes**
  - **Order tracking/Customer support**
  - **Online Transactions: eCommerce**

- **Problems of original Web concept**
  - **Static web pages**
  - **No interactivity**
  - **Stateless protocol: no support for multi-step transactions**
  - **Insecure communications**

Web Forms

- Pages which contain fields to be filled by user
- Usually contain a "Submit" button
- When user presses "Submit", server responds by sending a page containing information specific to the user-supplied parameters
- Examples:
  - Web search tools
  - Order forms in commercial web sites

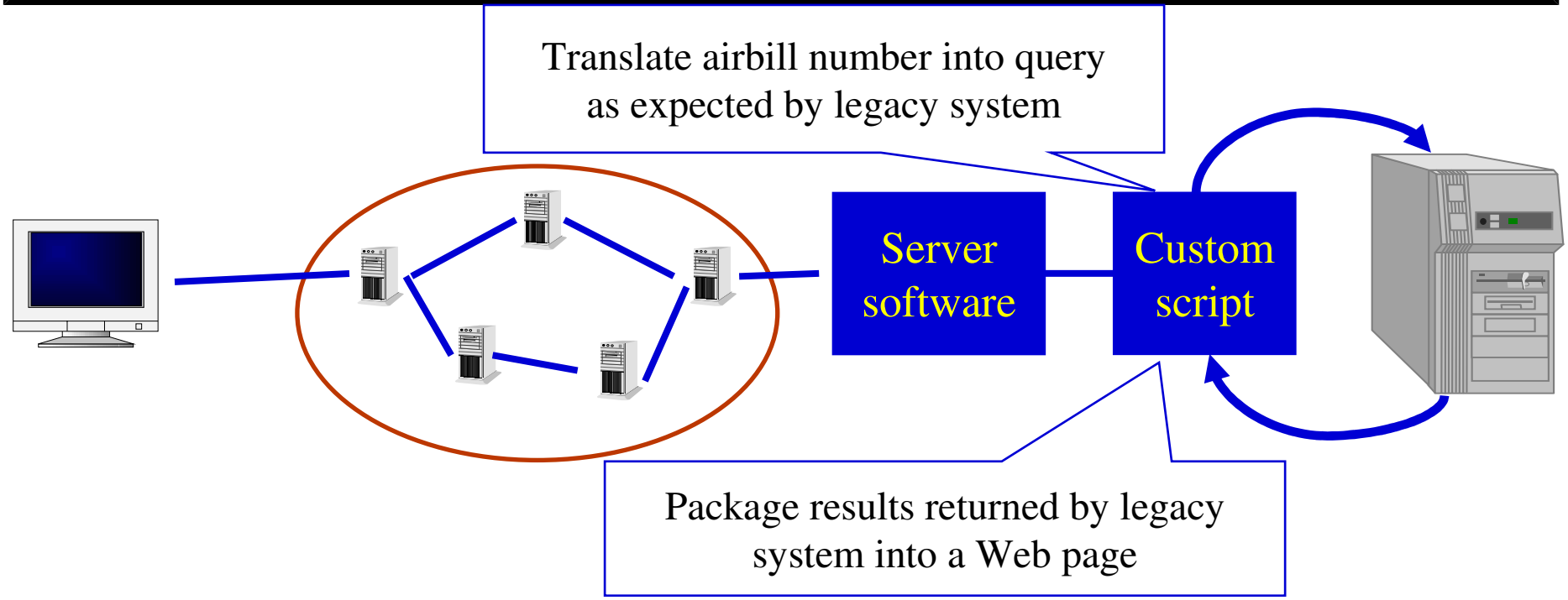
Web Forms Under the Hood

- Server sends original html page containing input fields
- User types info into fields and presses submit button
- Client establishes connection with handler script at server side (script filename contained in web page)
- Client collects user input into a long string and sends it along with an HTTP command back to server
  - POST customer=John+Doe&cardno=1234567890&expires=6/98& product=123&quantity=5
- Handler script at server reads parameter string and processes it, usually producing a new page as a result

CGI

- Common Gateway Interface
- Set of standards for writing handler scripts
- How it works
  - All URLs that refer to a special directory (e.g. /cgi) cause the execution of a corresponding script at the server (for example http://web.mit.edu/cgi/test)
  - Scripts typically translate parameters into SQL statements for a database and translate the query results into an HTML page

Example: FedEx



Microsoft Active Server Pages (ASP)

- Competing technology to CGI
  - Scripting Language is similar to Visual Basic
- MS Access can automatically convert database tables, queries and forms into ASP pages
- Requires Microsoft web server

Evolution of the WWW

Business Drivers

- Enable transactions
- **Allow interactivity between browser and server**
- Facilitate personalization
- Support multiple browsing devices
- Better organize and retrieve Web content
- Support Business-to-Business applications

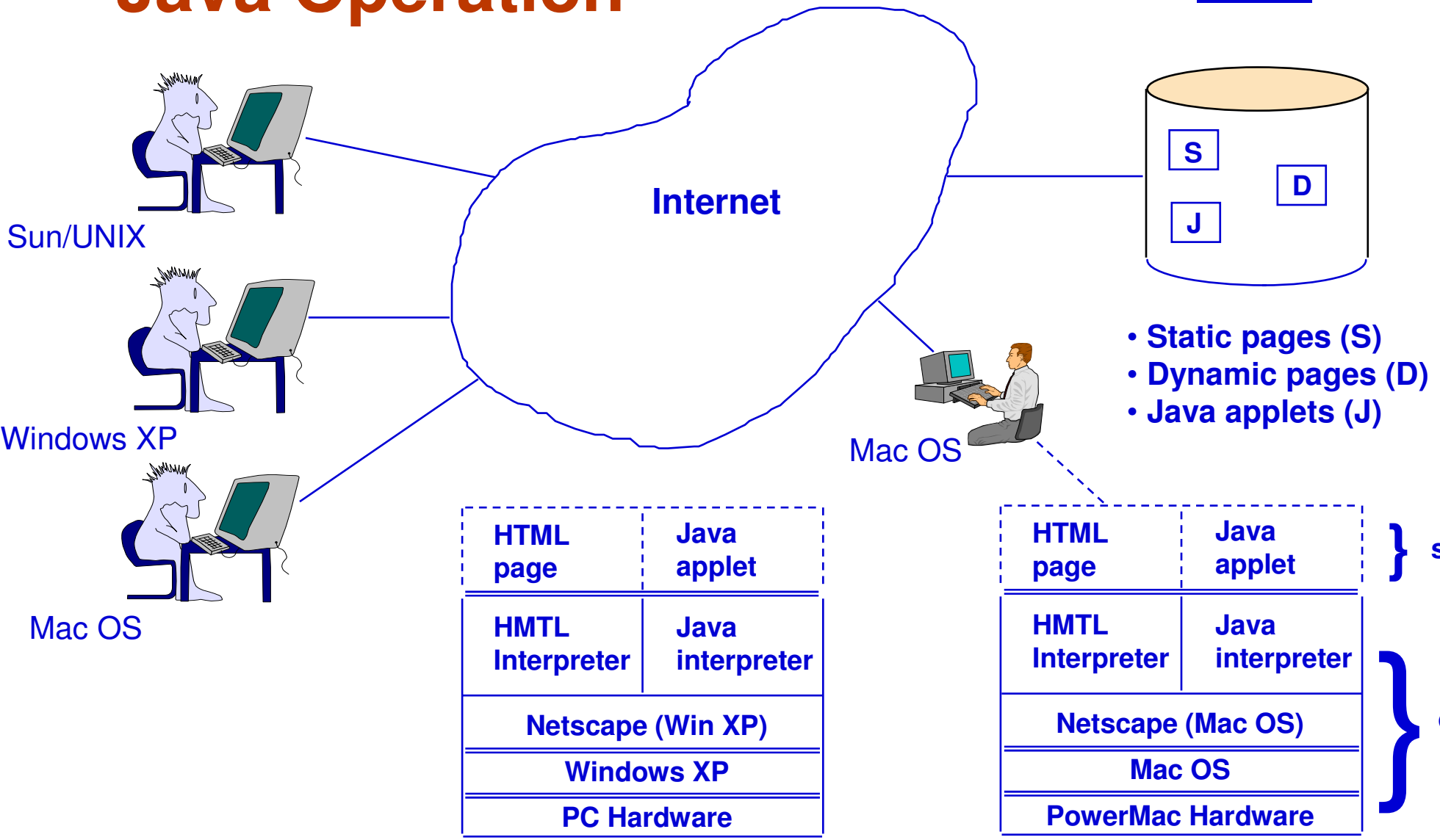
The Interactive Web

- **Business Motivation:**
  - Allow complex interaction between user browser and corporate server
  - Web becomes an extension of the user's PC
  - Browser becomes a window to a variety of corporate applications
- **Problems with Web Forms/CGI/ASP**
  - All processing done at server side
  - Rapid user interaction with Web page not possible
  - Need local processing to create highly interactive Web pages

Enter Java Applets

- Programming language to enable interactive Web pages
- Developed by Sun Microsystems
  - originally for programming intelligent microwave ovens!!!
- Java programs are called applets
- Applets are platform-independent
  - They can run equally well on Windows, Macs, Unix, etc.
  - Require special browsers that can support Java though

Java Operation



Adapted from Stuart Madnick, MIT

Evolution of the WWW

Business Drivers

- Enable transactions
- Allow interactivity between browser and server
- **Facilitate personalization**
- Support multiple browsing devices
- Better organize and retrieve Web content
- Support Business-to-Business applications

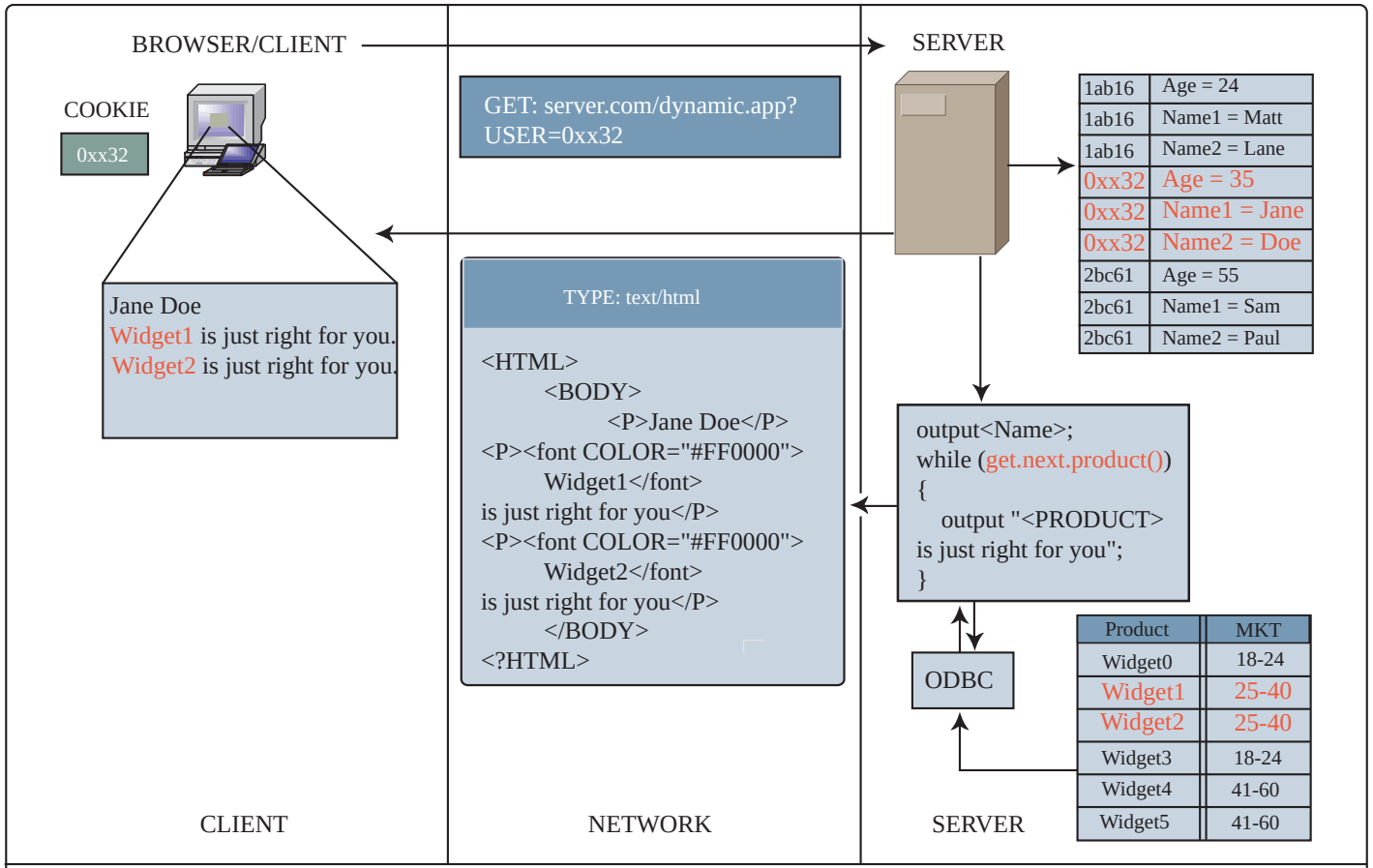
Personalized Interaction

- **Business motivation:**
  - Low cost medium for gathering information from customers to allow
    - » Personalized service
    - » Targeted advertising
- **Problems with current model**
  - Does not allow easy identification of distinct customers

Cookies

- A method for identifying web users and delivering customized web sites
  - First time user connects to a web site, s/he is asked to fill in personal information form
  - Server packages information into a “cookie” file and sends cookie to browser
  - Browser stores cookie in local file system
  - Each subsequent time browser visits site, it sends cookie back to server
  - Server uses information stored in cookie to identify user and possibly customize the supplied web pages
- Privacy implications?

Cookie applications



The user's browser passes the contents of the cookie to the Web server. The server uses this cookie to look up information about the user (this information may have been gained from a previous registration process). The information (specifically, the user's age) is used to query a database via ODBC to determine which products should be recommended to this user. The resulting list of products is inserted into a dynamically created HTML page and sent to the user's browser for display.

Figure by MIT OCW.

Summary: WWW The Current Concept (1995-today)

- Human Users ← → Documents and Applications
- Main business use
  - B2C Transactions
  - Customer Support

Evolution of the WWW

Business Drivers

- Enable transactions
- Allow interactivity between browser and server
- Facilitate personalization
- **Support multiple browsing devices**
- **Better organize and retrieve Web content**
- Support Business-to-Business applications

Multiple Delivery Devices

- **Business motivation:**
  - Allow users to access web content from a variety of devices
    - » PC Browsers
    - » PDAs (e.g. Palm Pilots)
    - » Mobile Phones
    - » Telephones (via voice interface)
    - » …
- **Problems of current Web model**
  - Each access device has different look-and-feel requirements
  - HTML specifies formatting for PC browsers only

Organize and Index Web Content

- Web is useless unless we can easily locate relevant resources
- Current solution: Search Engines
  - Index the Web by automatically "discovering" web pages and organizing them around keywords found in their text

How does Google work?

- Before you ever enter a query:
  - Programs (called "web crawlers" or "spiders") follow links from one page to another all over the web.
  - The programs construct indexes of which words appear on which pages and save the indexes (and often copies of the pages) on massive "server farms" maintained by Google.
  - Each page is also assigned a "page rank" based on the number of other pages that link to it. Links from pages that, in turn, have lots of other pages linking to them are weighted more heavily.

Google's page rank formula

- PR(A) = (1-d) + d(PR(t1)/C(t1) + ... + PR(tn)/C(tn))
  - Where
    - » t1 - tn are pages linking to page A,
    - » C is the number of outbound links that a page has
    - » d is a damping factor, usually set to 0.85.

- In words:
  - a page's PageRank = 0.15 + 0.85 * (a "share" of the PageRank of every page that links to it)
  - "share" = the linking page's PageRank divided by the number of outbound links on the page.

How does Google work? (cont.)

- After you enter a query:
  - Programs check the indexes to determine which pages contain the combination of words you entered.
  - Google provides a list of these pages in order of page rank (order is probably affected by other factors, too)

Google numbers (as of December 2004)

- Over four billion Web pages, each an average of 10KB, all fully indexed.
- Up to 2,000 PCs in a cluster.
- Over 30 clusters.
- One petabyte of data in a cluster -- so much that hard disk error rates of 10^{-15} begin to be a real issue.
- Sustained transfer rates of 2Gbps in a cluster.
- An expectation that two machines will fail every day in each of the larger clusters.
- No complete system failure since February 2000.

Source: ZDNet – UK <http://www.zdnet.com.au/insight/software/0,39023769,39168647,00.htm>

Problems with today's searches…

- **Text keywords are misleading…**
- **HTML does not give any clues as to the true meaning of the data**
- **Example:**
  - *Desperately seeking Wendy Cook*

Evolution of the WWW

Business Drivers

- Enable transactions
- Allow interactivity between browser and server
- Facilitate personalization
- Support multiple browsing devices
- Better organize and retrieve Web content
- **Support Business-to-Business applications**

Support for B2B applications

- Original Web was conceived as a communication medium between computers and humans
- Amazing new applications will become possible if computers can automatically read and understand Web pages
  - Electronic purchasing
  - Intelligence gathering
  - ….
- Problem:
  - HTML pages are unstructured
  - HTML only provides information about presentation, not meaning

What is the underlying issue?

- When storing documents on the web, specify not only their appearance, but also their semantics (i.e. their meaning!)

Enter: The Semantic Web

- The "Next Generation Web" with well-established infrastructure for expressing information in a
  - Precise, Human-readable, and Machine-interpretable form.
- Enable syntactic and semantic interoperability among independently-developed Web applications, allowing them to efficiently perform sophisticated tasks for humans.
- Enable Web resources to be accessible by their semantics rather than by keywords and syntactic forms.
- Enable inferencing:
  - Chris is an associate professor at MIT.
  - Associate professors are permanent employees.
  - Chris is a permanent employee of MIT.

Evolution Of The Web

Presence	Transactions	Business
Publish Info	**Process Transactions**	**Digital Economy**
Web sites	Web-enable existing systems	Business transformation
Pages	Transactions	Business processes
Islands	Islands	Constellations
Eyeballs	Revenue	Profits