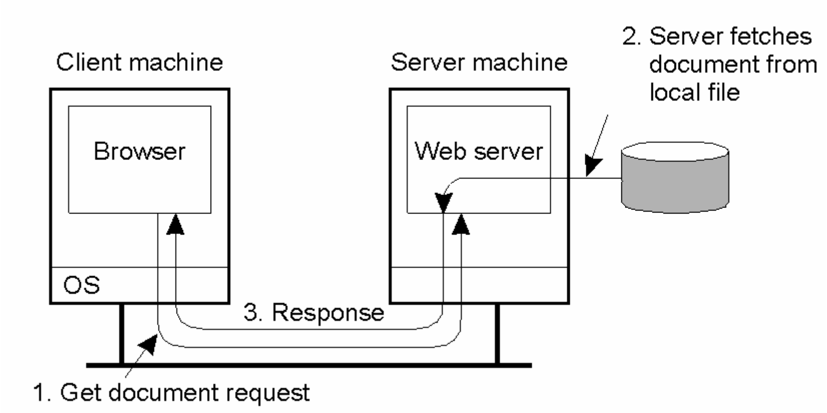
COP 6611 Advanced Operating System

Distributed Document-Based Systems

Chi Zhang

czhang@cs.fiu.edu

The World Wide Web



Overall organization of the Web.

HTML ⇒ HTTP ⇒ TCP

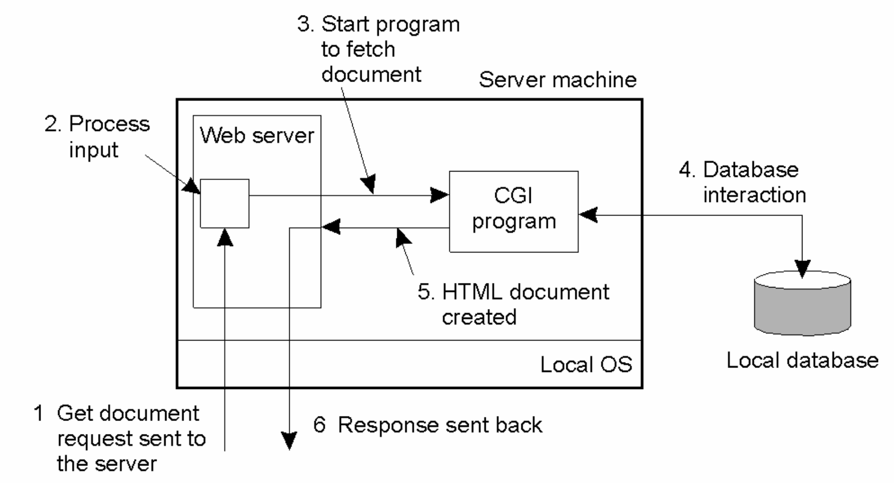
HTTP is a *stateless* application-layer protocol

Document Types

Type	Subtype	Description
Text	Plain	Unformatted text
	HTML	Text including HTML markup commands
	XML	Text including XML markup commands
Image	GIF	Still image in GIF format
	JPEG	Still image in JPEG format
Audio	Basic	Audio, 8-bit PCM sampled at 8000 Hz
	Tone	A specific audible tone
Video	MPEG	Movie in MPEG format
	Pointer	Representation of a pointer device for presentations
Application	Octet-stream	An uninterrupted byte sequence
	Postscript	A printable document in Postscript
	PDF	A printable document in PDF
Multipart	Mixed	Independent parts in the specified order
	Parallel	Parts must be viewed simultaneously

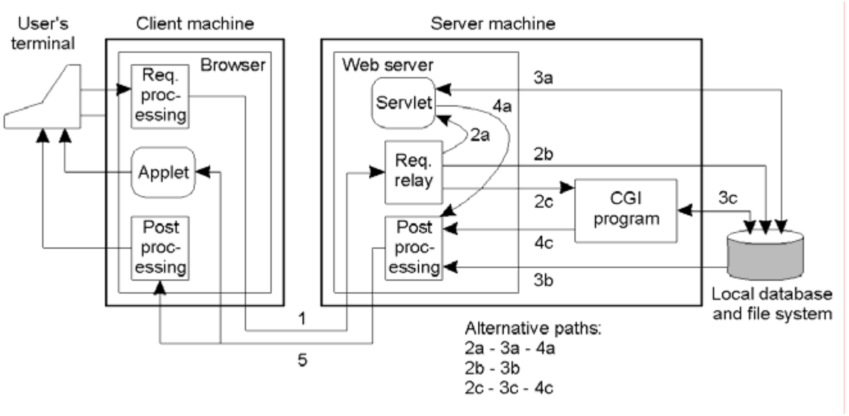
Six top-level MIME types and some common subtypes.
e.g. text/HTML, application/PDF

Architectural Overview (1)



The principle of using server-side CGI programs.

Architectural Overview (2)



Architectural details of a client and server in the Web.

Client-side script

```
<HTML>                                          <!- Start of HTML document     -->
<BODY>                                          <!- Start of the main body     -->
<H1>Hello World/H1>                             <!- Basic text to be displayed  -->
<P>                                             <!- Start of a new paragraph   -->
<SCRIPT type = "text/javascript">               <!- identify scripting language -->
  document.writeln ("<H1>Hello World</H1>;            // Write a line of text
</SCRIPT>                                       <!- End of scripting section   -->
</P>                                            <!- End of paragraph section -->
</BODY>                                         <!- End of main body          -->
</HTML>                                         <!- End of HTML section       -->
```

A simple Web page embedding a script written in JavaScript.
Also, client-side program: Java Applet.

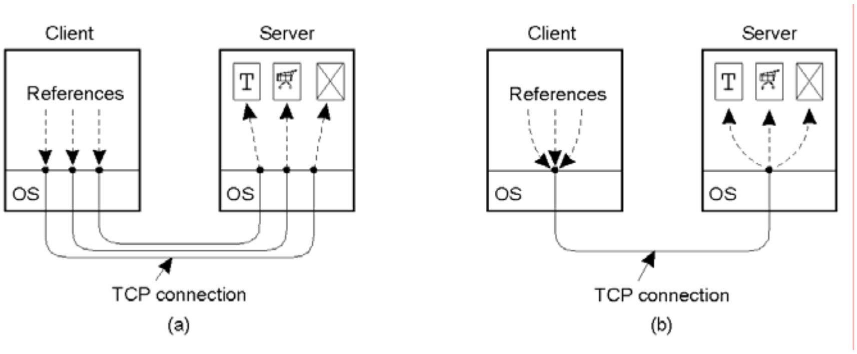
Server-side script

```
<HTML>
<BODY>
<P>The current content of <pre>/data/file.txt</PRE>is:</P>
<P>
<SERVER type = "text/javascript");
    clientFile = new File("/data/file.txt");
    if(clientFile.open("r")){
            while (!clientFile.eof())
                document.writeln(clientFile.readln());
            clientFile.close();
    }
</SERVER>
</P>
<P>Thank you for visiting this site.</P>
</BODY>
</HTML>
```

An HTML document containing a JavaScript to be executed by the server

Also, server-side application: servlet (servlets run as threads of the server, while CGI scripts run in separate processes)

HTTP Connections



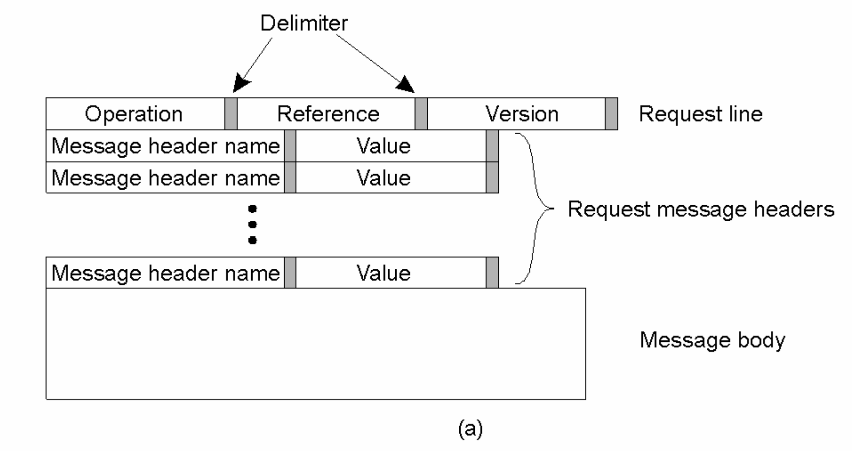
a) Using nonpersistent connections.
b) Using persistent connections (HTTP 1.1 or later)

HTTP Methods

Operation	Description
Head	Request to return the header of a document
Get	Request to return a document to the client
Put	Request to store a document at a certain location
Post	Provide data that is to be put to a document (e.g. CGI script)
Delete	Request to delete a document

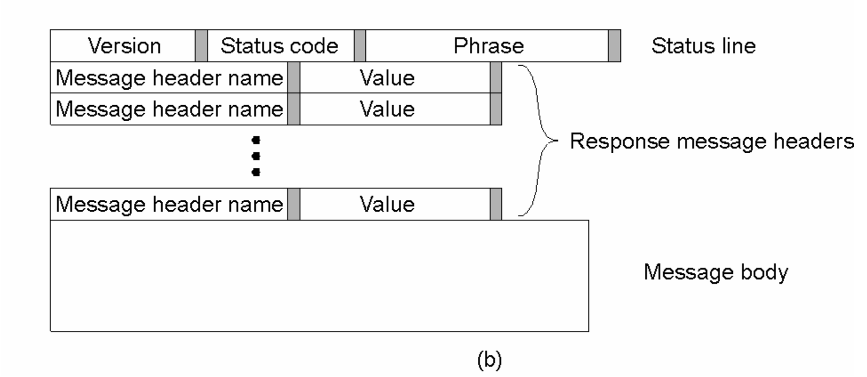
Request Operations supported by HTTP.

HTTP Messages (1)



HTTP request message
Reference: URL

HTTP Messages (2)



Version | Status code | Phrase — Status line

Message header name | Value

Message header name | Value

Message header name | Value

Response message headers

Message body

(b)

HTTP response message.

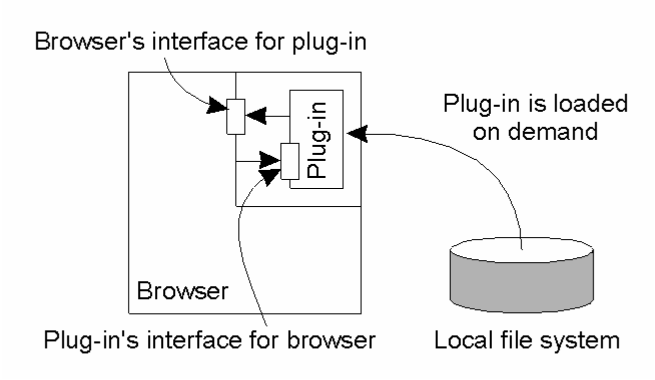
Status Code: the operation status.   Phrase: explain the status code.

HTTP Messages (3)

Header	Source	Contents
Accept-Language	Client	The natural language the client can handle
Expires	Server	The time how long the response remains valid
Host	Client	The TCP address of the document's server
Last-Modified	Server	The time the returned document was last modified
Location	Server	A document reference to which the client should redirect its request
Referer	Client	Refers to client's most recently requested document
Upgrade	Both	The application protocol the sender wants to switch to (maybe more secure SHTTP)

A request or response message may contain additional headers, indicating content type, length, encoding, time etc.

Clients (1)

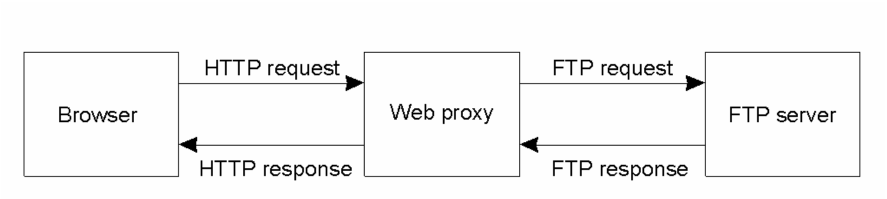

Browser's interface for plug-in
Plug-in
Plug-in is loaded on demand
Browser
Plug-in's interface for browser
Local file system

Using a plug-in in a Web browser.

A plug-in is a small program that can be dynamically loaded into a browser for handling a specific document (MIME) type.

The interfaces are standardized.

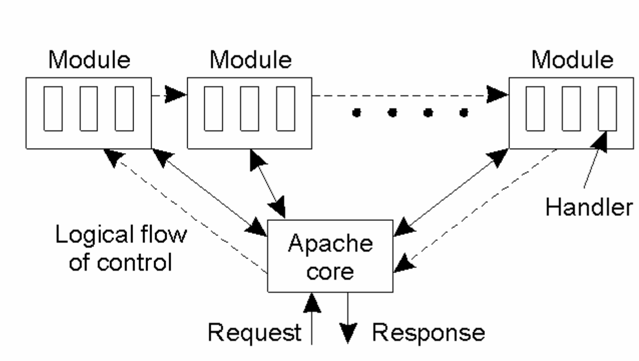
Clients (2)



Using a Web proxy when the browser does not speak FTP.

A Web proxy can be shared by a number of browsers.

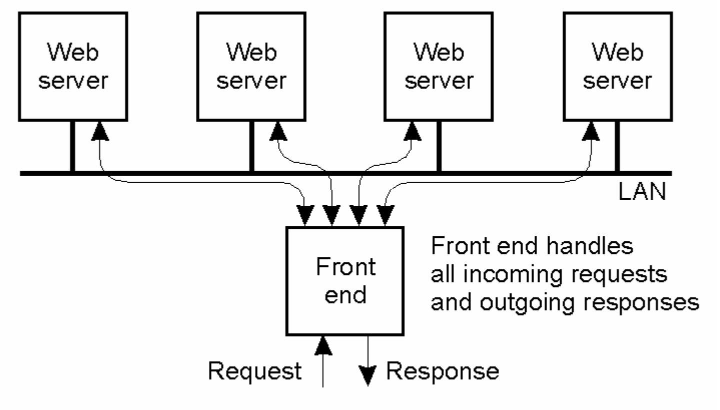
Servers



General organization of the Apache Web server.

Apache servers are highly configurable: modules can be incorporated. Each module can provide one or more handlers that can assist in processing an incoming HTTP request.

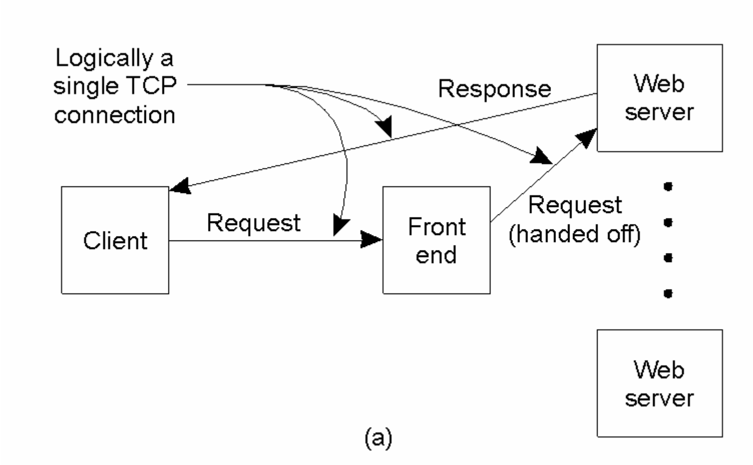
Server Clusters (1)



A transport-layer switch passes the data of a TCP connection to one of the servers, depending on some measurement of the server’s load.

With content-aware distribution, the front end also distributes the HTTP request based also its content.

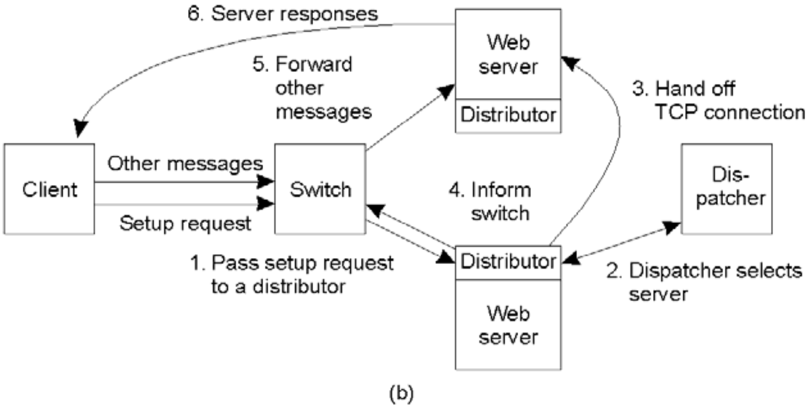
Server Clusters (2)



(a) The principle of TCP handoff.

The server's response is sent directly to the client, without the intervention of the front end.

Server Clusters (3)



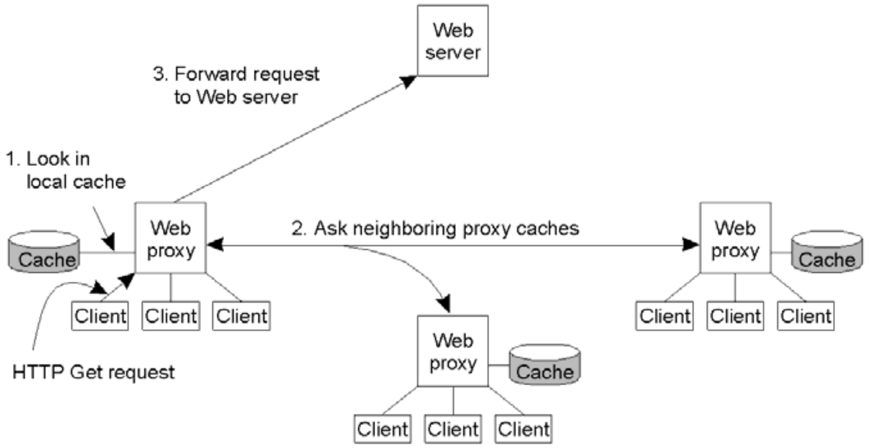
(b) A scalable content-aware cluster of Web servers.

Switch + Distributor + Dispatcher = Front End

Caching and Proxy

- A proxy send a *conditional* HTTP request (with header *If-Modified-Since*) to a server.
- To improve performance at the cost of weak consistency, Squid Web Proxy assigns $T_{expire} = \alpha (T_{cached} - T_{last\text{-}modified}) + T_{cached}$
- Push-based mechanism and Leases
- Active cache: In some cases, it is possible to shift generation of the document from the server to the proxy.

Cooperative Caching

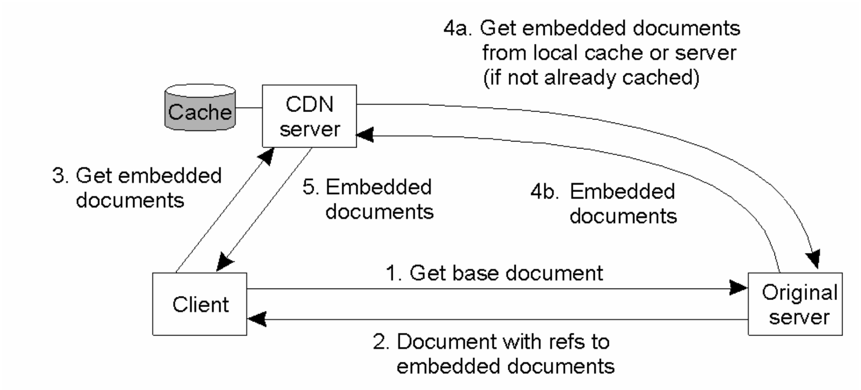


The principle of cooperative caching

Akamai CDN (1)

- A main HTML may contain several other documents such as images, video, and audio.
  - Embedded documents are large
  - Embedded documents rarely change
  - Cache the embedded documents
- In the main HTML, URLs to the embedded documents actually refer to the pages cached in CDN.
- The CDN DNS returns the IP address of the CDN server closest to the client, or with less load.
  - Alternative: assign the same IP address to several servers, and let the network layer direct the request to the nearest server.

Akamai CDN (2)


4a. Get embedded documents from local cache or server (if not already cached)
Cache
CDN server
3. Get embedded documents
5. Embedded documents
4b. Embedded documents
1. Get base document
Client
Original server
2. Document with refs to embedded documents

The principle working of the Akami CDN.