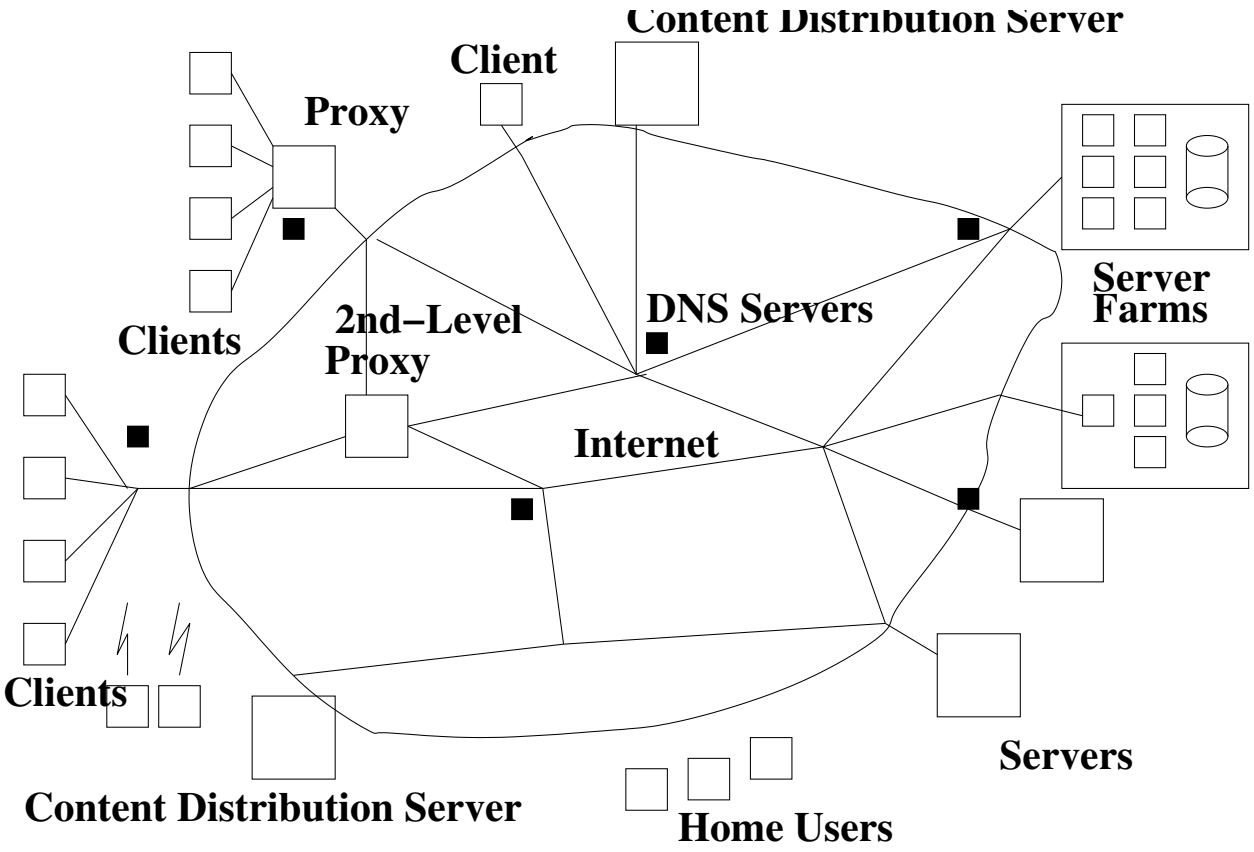
Web as a Distributed System

The World Wide Web is a large distributed system. In 1998 comprises 70-75% of Internet traffic. With large transfers of streaming media and p2p, no longer a majority of bytes, but is in terms of flow.

Active area of personal research. Look at my home page for online papers.



E2E Factors and Relevance to Distributed Systems

Build up end-to-end performance picture. Basic picture of clients and servers. How DNS and HTTP are used for each object.

Static Objects

Served from a static file. Similarities with a distributed file system, but differences.

- much larger scale (number of clients)
- heterogeneous clients (machine type/operating system)
- Web generally has single-writer/multiple-reader versus a DFS is concerned about multiple writers.
- no central administration

Composite Objects

Pages with embedded objects (images). In the simplest case just make subsequent requests to the same server over separate TCP connections.

Browsers use parallel connections to speed up delivery speed.

Increasingly obtain objects for a page from a variety of different servers—specialized, ad, CDNs

Network Protocols

Use of HTTP/1.1 for persistent connections (retrieval of multiple objects over the same connection) and pipelining (multiple requests before receiving any replies).

When persistence and pipelining works it is better than parallel connections.

Caches

- Browser vs. Proxy (depends on shared interests of users).
- Cache Replacement. LRU, GreedyDual-Size (value divided by size with credit for recent accesses)
- Cache Coherency. Strong consistency (validate each access, server invalidation like AFS). Weak consistency using a heuristic TTL.
- Hierarchy of caches. How to should caches work together?
- DNS Caches—local DNS servers handle client requests and recursively resolve names. Cache results based on TTL of response.
- Server Caches—cache content in memory for faster response.

Dynamic Content

Retrieval is no longer a file retrieval, but a remote procedure call with one or more parameters.

Query to a search engine where key words are parameters.

Use of cookies, which are pieces of information sent by browser on each request for server to track requests and customize content.

Server-side computation is being performed (CGI, JavaScript, ASP (VBScript). May include database access.

Server Clusters

A single server is not enough. Have a farm (cluster) of servers. How to organize? Make it visible to the world or not?

- HTTP redirection
- multiple IP addresses for the same server name.
- a single high-speed switch in front of a group of web servers. Transparent to clients. L4 or L7.

Content Distribution Networks

- How to use them? Talk about Akamai use of DNS.
- Problems of using DNS to determine client location.
- What content to store there?
- How to update?

Other Factors

- Client-side scripts (java applets)—distriblet use
- Secure transactions (SSL)
- Streaming audio/video content

Interaction Between Factors

Impact of one factor may be more or less depending on another factor.

For example, impact of caching or content distribution network may depend on the HTTP protocol option being used.

Secure Sockets Layer

With increased use of electronic commerce, secure transactions are important. SSL protocol developed by Netscape for encrypted communication between clients and servers. Now it is called Transport Layer Security (TLS)

TSL sits between TCP/IP and HTTP layers. Invoked by browsers using the protocol "https".

Uses public-key cryptography. Features:

- Clients can authenticate servers through a certificate authority. Important as a client should not send credit card information to an unknown server.
- Servers can authenticate clients in the same way.
- Encrypted connection (no plain text).

Client initially generates a random number and encrypts it with server's public key (obtained from authentication server).

Exchange of messages does the following:

- authenticate the server to the client
- allow client and server to select the cryptographic algorithms, or ciphers, they both support. Choices include DES, Digital signature algorithm (DSA), MD5, RSA, RSA key exchange. Most common is RSA key exchange developed for SSL.
- Optionally authenticate client to the server.
- Use public-key encryption techniques to generate shared secrets.
- Establish encrypted SSL connection.

Additional Directions

- extraneous content—ads
- web registration
- measurement of e2e performance—potential projects
- p2p contrast
- virtualization of servers