

What's important about the web?

- huge economic impact
- uses lots of the things we talked about - TCP, self-sim, etc.

7

• people care about performance



13



What is the web?

• protocols

- formatting: HTML, now XML, also video, sound, etc.
- addressing: URLs

11d_web_overview: CSci551 SP2006 © John Heidemann

- data retrieval: HTTP, streaming?
- anything else?
 - · other thigns as the web via web services
 - security / encryption issues
 - · browsers and servers, proxies, search engines

11d_web_overview: CSci551 SP2006 © John Heidemann

22











HTTP/0.9

36

- original protocol
 client: "GET /path" (close)
 server: "data" (close)
- very, very simple
- still valid
- but no way to do control

HTTP/1.0 • informational rfc1945 (60 pages) • client – "GET /path HTTP/1.0", headers, body (close) – other operations (GET, HEAD, POST) • can send data (POST) • can check for changes (HEAD) • server – headers, data (close) – headers allow info (type, change time, etc.)

- uses MIME types, adds content negotiation, etc.
- uses fulling types, and content negotiation, e

37

HTTP/1.1

- rfc2068 (162 pages!)
- basically like 1.0, but adds:
 - persistent connections
 - identity of host (supports virtual hosts)
 - detailed caching models

11d_web_overview: CSci551 SP2006 © John Heidemann

HTTP Request—More Detail Request line — Method GET – return URI HEAD – return headers only of GET response POST – send data to the server (forms, etc.) URI E.g. http://www.isi.edu/~govindan/index.html with a proxy E.g. /index.html if no proxy HTTP version

11d_web_overview: CSci551 SP2006 © John Heidemann

HTTP Request

- · Request headers
 - Authorization authentication info
 - Acceptable document types/encodings
 - From user email
 - If-Modified-Since
 - Referer what caused this page to be requested
 - User-Agent client software
- Blank-line
- Body

11d_web_overview: CSci551 SP2006 © John Heidemann



42

38

40

HTTP Request Example

39

41

43

GET / HTTP/1.1 Accept: */* Accept-Language: en-us Accept-Encoding: gzip, deflate User-Agent: Mozilla/4.0 (compatible; MSIE 5.5; Windows NT 5.0) Host: www.isi.edu Connection: Keep-Alive

11d_web_overview: CSci551 SP2006 © John Heidemann

HTTP Response

• Headers

- Location for redirection
- Server server software
- WWW-Authenticate request for authentication
- Allow list of methods supported (get, head, etc)
- Content-Encoding E.g x-gzip
- Content-Length
- Content-Type
- Expires
- Last-Modified
- Blank-line
- Body

11d_web_overview: CSci551 SP2006 © John Heidemann

HTTP Response Example

HTTP/1.1 200 OK

Date: Tue, 27 Mar 2001 03:49:38 GMT Server: Apache/1.3.14 (Unix) (Red-Hat/Linux)
mod_ssl/2.7.1 OpenSSL/0.9.5a DAV/1.0.2 PHP/4.0.1pl2 mod_perl/1.24 Last-Modified: Mon, 29 Jan 2001 17:54:18 GMT ETag: "7allf-10ed-3a75ae4a" Accept-Ranges: bytes Content-Length: 4333 Keep-Alive: timeout=15, max=100 Connection: Keep-Alive Content-Type: text/html

<!DOCTYPE xxx.....

11d_web_overview: CSci551 SP2006 © John Heidemann

Caching • Why cache? - can save a *lot* of bandwidth • Risk: out-of-date data - file systems are usually *strongly cache* coherent, detecting when data is modified and *invalidating* cached copies - the web is provides only weak coherence, you can get old data

11d_web_overview: CSci551 SP2006 © John Heidemann

Web Cache Coherence • Assume page is good • Subsequent request: -if in valid period, replay until time *t* data out of cache (no -t could be in past request) · First request for http://foo/ w/0 valid period) -if beyond valid period, -returns data with valid issue a conditional period ("Expires: *t*") request -clients may estimate · GET with "If-Modifiedvalid period if none given

11d_web_overview: CSci551 SP2006 © John Heidemann

- -(server can prohibit this
 - Since" header Respose is either "304 Not modified" or "200 OK" + new data

44

Example Cache Check Request

45

47

GET / HTTP/1.1 Accept: */* Accept-Language: en-us Accept-Encoding: gzip, deflate If-Modified-Since: Mon, 29 Jan 2001 17:54:18 GMT If-None-Match: "7a11f-10ed-3a75ae4a" User-Agent: Mozilla/4.0 (compatible; MSIE 5.5; Windows NT 5.0) Connection: Keep-Alive

11d_web_overview: CSci551 SP2006 © John Heidemann

Example Cache Check Response

HTTP/1.1 304 Not Modified

Date: Tue, 27 Mar 2001 03:50:51 GMT Server: Apache/1.3.14 (Unix) (Red-Hat/Linux) mod_ssl/2.7.1 OpenSSL/0.9.5a DAV/1.0.2 PHP/4.0.1pl2 mod_perl/1.24 Connection: Keep-Alive

Keep-Alive: timeout=15, max=100

ETag: "7allf-10ed-3a75ae4a"

11d_web_overview: CSci551 SP2006 © John Heidemann

48



Web Workloads

- Users make connections and request pages made up of objects

 a structural model
- user arrival: Poisson
- connection duration, number of object per page, object size: all often heavy tailed (Pareto)

50

- P&M give object stats:
 - median: 1946 bytes
 - mean: 13767 bytes

11d_web_overview: CSci551 SP2006 © John Heidemann

Page Stats

- Popularity
 - Zipf distribution
 - requests for i^{th} most popular document ~ $i^{\cdot a}$
 - (popular are really popular, but long tail)

51

- Requests are *bursty*
 - recall self-similarity lecture

11d_web_overview: CSci551 SP2006 © John Heidemann