Welcome!

CSc 450: Web Technologies & Performance

Prof. Brian D. Davison
davison@cse.lehigh.edu

http://www.cse.lehigh.edu/~brian/csc450/
Course Description

- Research Seminar
- Unlike most seminars
  - Problem-oriented
  - Operate like a research group
- Topic: WWW Performance
  - Infrastructure of the Web
  - Measurement and Modeling
Today

- Personal introductions
- Course overview
- Outline of Web Performance
- Brainstorm
- Start reviewing some basics
Course Overview

- Blackboard information (add users)
  - discussion board
  - useful in other classes?
- Background textbook
  - Web Protocols and Practice (Krishnamurthy and Rexford)
- Relatively informal class
- Students take notes/do most presenting
Grading

- Individualized
- In-class participation
  - We must know your name!
- Presentations
- Reports/papers/project
- You earn your grade; I just record it
- Impress me!
Academic Honesty

• (relevant for all courses)
• Do your own work (unless explicit groups)
• Appropriately acknowledge all sources
  • Show your scholarship!
  • Both in presentations and in
  • Careful citations in papers
• Dishonesty will not be tolerated
Research on the Web

- Usually can be done "on the Web"
  - Search engines
    - Google, AltaVista, AllTheWeb (FAST)
  - Paper indices/repositories
    - CiteSeer (Science/ResearchIndex)
    - CORA, XXX.LNBL.GOV, WEBBIB
    - IEEE/ACM Digital Libraries
  - Standards organizations
    - W3C, IETF
For Next Week

- Academic home page
  - contact info, bio, interests, publications, research projects, etc.
- Set up individual meetings in my office
- Read Chapters 1-4 (and read ahead)
  - (student to present ch4)
- Bring something interesting to discuss, relating to Web performance (weekly)
- No class on Monday
Outline of Web Performance

- What is the Web?
- HTTP, HTML, DNS, TCP, Clients/Servers/Proxies
- What is wrong with the Web?
- What has been done to improve performance?
- Caching, new features of HTTP/1.1
- How else can Web performance be improved?
- Web performance measurement, analysis, and simulation
Introduction to the Web

• What is the Web?
  • Origin
    • Proposed by Tim Berners-Lee in 1989
    • Competed with FTP, Gopher, and WAIS
  • Now
    • Hundreds of millions of users
    • Billions of Web pages
    • Over a hundred million servers
What is the Web?

- Components of the Web
  - Uniform Resource Identifier (URI)
    - e.g. http:// www.cse.lehigh.edu/~brian/
  - Hypertext Markup Language (HTML)
    - what Web pages are made of
  - Hypertext Transfer Protocol (HTTP)
    - how Web pages are transferred
What is the Web?

- Major players
  - Content
    - Text, HTML, Images, PDF, Streaming Media, etc.
  - Software
    - Client browsers, Origin Web servers, Intermediaries
  - Network
    - Typically the Internet, using TCP over IP
  - Protocol
    - Usually HTTP, but includes others like FTP, NNTP
HTTP: Hypertext Transfer Protocol

- Versions .9, 1.0, 1.1 (+ a few extensions)
- For next week
  - Find which RFC describes HTTP/1.1 and the status of that RFC
  - Send answer by email, and tell me where you found the answer
- Uses headers (and footers) to describe content being requested or received
Example of HTTP Headers

GET http://www.perl.org/images/home601a.GIF HTTP/1.0
Host: www.perl.org
User-Agent: Mozilla/4.04 (en) (X11; I; SunOS 5.5.1 sun4u)
Referer: http://www.perl.org/

HTTP/1.0 200 OK
Content-Type: image/gif
Content-Length: 47018
Date: Thu, 18 Mar 1999 20:14:45 GMT
Server: Apache/1.3.4 (Unix) Debian/GNU
Last-Modified: Sun, 11 Jan 1998 21:46:12 GMT
ETag: "4844-b7aa-34b93da4"

19990318 20:14:45 TCP_MISS/200 47018 GET
http://www.perl.org/images/home601a.GIF
Retrieval process

• Given an HTTP URL:
  • Use DNS to resolve hostname
  • Connect to host
  • Send HTTP request
  • Receive HTTP response (with content)
Web Clients

- Browser
  - Seen as "The Web" by users
  - Integrates retrieval with rendering
  - Modern browsers parse various page description and rendering languages
    - HTML, JavaScript, Java, HTML-variants
  - Typically extensible with plug-ins
    - e.g., for PDF, RealAudio, Flash
Web Clients

• Spiders
  • Clients that automatically traverse links to much of the Web
  • e.g., for search engines, link checkers, or for any resource discovery need
  • May use a strategy similar to breadth-first (or possibly depth-first)
  • Sometimes concerned with permission (robots.txt) and dynamic resources
Web Clients

- Other specialized 'bots
  - "Intelligent Agents" -- specialized services
    - Meta-search engines
    - Auction watchers
    - Crawl a site for offline viewing
  - Shopping 'bots for price comparison
  - Change detectors
Web Proxies

- An intermediary that acts as both a server and as a client to forward requests and responses
- Portals through firewalls
- Content/access management
  - Filtering
  - Modifying
  - Virus checking
  - Anonymizing
  - Caching