

Tracking the Evolution of Web Traffic: 1999-2003

Felix Hernandez-Campos, Kevin Jeffay
F. Donelson Smith

*IEEE/ACM International Symposium on
Modeling, Analysis and Simulation of
Computer and Telecommunication Systems
(MASCOTS)*

Orlando, FL, October 2003



Introduction

- Since mid-1990, Web traffic dominant on Internet
 - Need to understand effects and technology
- Since 1990s, Web has evolved
 - HTTP delivers more than HTML pages
 - email news, IM, transactions
- Better understand Web as Internet traffic



This Paper

- Analysis of 1 terabyte of TCP/IP headers
 - From UNC during 1999, 2001, 2003
 - Compare to other researchers
- Contribute
 - Empirical models for simulation
 - Characterization of TCP using new HTTP 1.1
 - Characterization of TCP using new server loads, banner ads, etc



Outline

- Introduction (done)
- Related Work (next)
- Data Sets
- Analysis
- Comparison with Others
- Sampling Issues
- Conclusions



Related Work

- Web traffic typically based on two studies
 - Mah [10], gathered in 1995
 - Barford, Crovella + [2,3,7] in 1995, 1998
- But
 - Small users sets (students, single labs)
 - Small data sets (up to 1 million objects)
 - Before HTTP 1.1
 - Old
- This paper
 - 200 million objects, 35k users, in 2003




Outline

- Introduction (done)
- Related Work (done)
- Data Sets (next)
- Analysis
- Comparison with Others
- Sampling Issues
- Conclusions




Data Sets

- Gather TCP/IP headers from Web servers to clients
 - Sequence numbers, Ack numbers
- Data:
 - Fall 1999 (6 one-hour samples, over 7 days)
 - Spring 2001 (3 four-hour samples, 7 days)
 - 2003? (8 one-hour traces over 7 days)
- Network:
 - 1999 OC-3 (155 Mbps)
 - 2001 OC048 (2.4 Gbps)
 - 2003 (same?)
 - (WPI: <http://www.wpi.edu/Admin/Netops/MRTG/>)
 - 50 Mbps during term, 15 Mbps during summer)

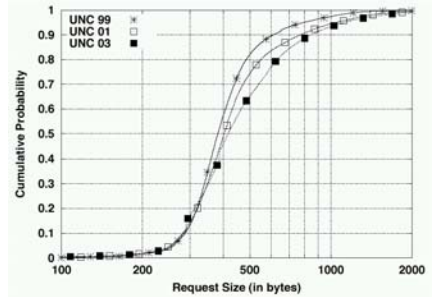


Outline


- Introduction (done)
- Related Work (done)
- Data Sets (done)
- Analysis (next)
- Comparison with Others
- Sampling Issues
- Conclusions



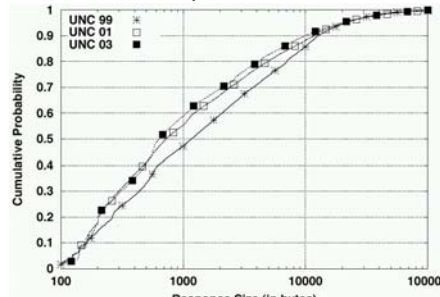
Request Sizes




- Requests becoming larger
- Still typically fit in one packet



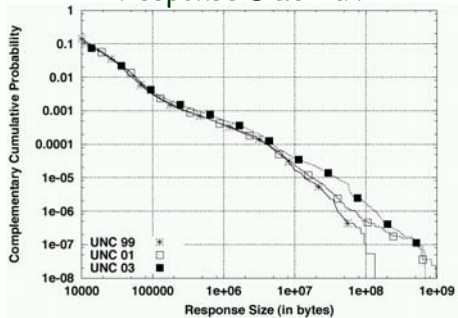
Response Sizes




- Responses becoming smaller
- Median fits in 1 packet (small!)



Response Size Tail




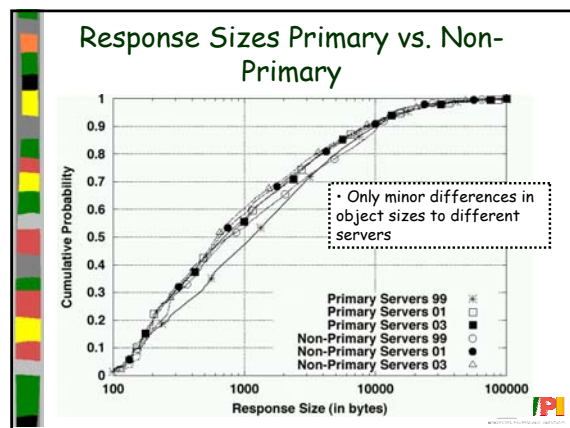
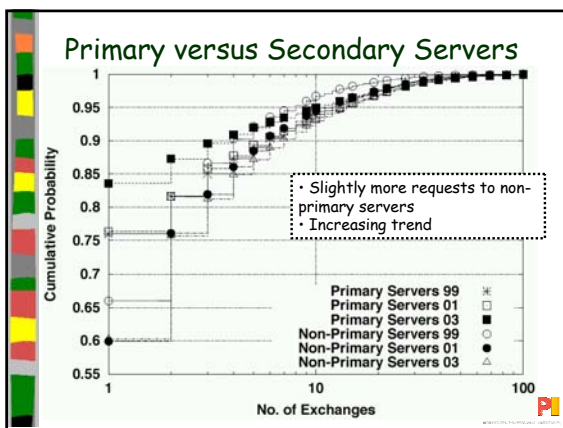
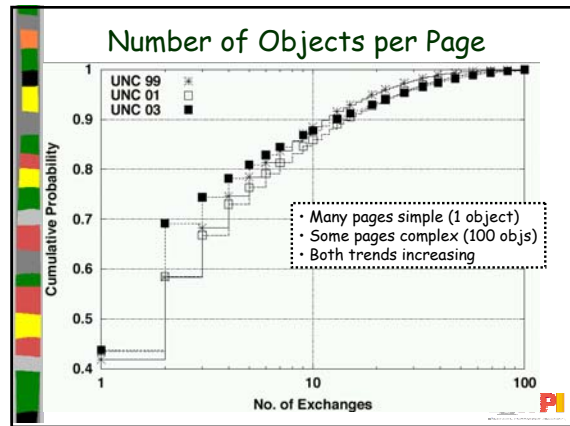
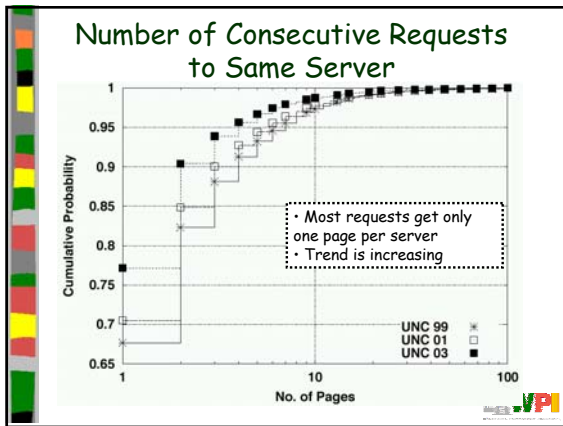
- Long-tailed
- Contributes to Self-Similarity



Inferring User and Browser Characteristics

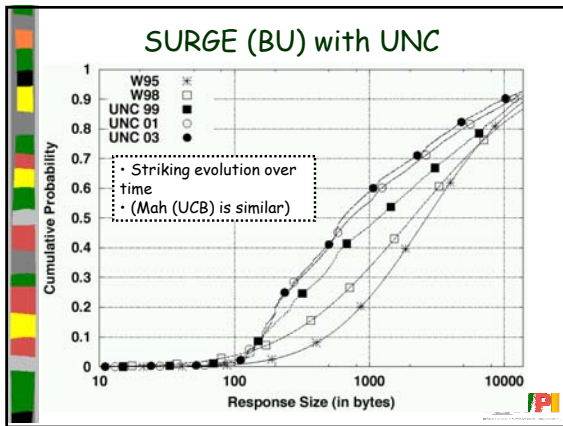
- Have TCP data, try to "infer" what user is doing at HTTP layer
- Time-sorted all flows
- Assume each IP is one user (fewer NATS on campus)
- First request is "page" and subsequent requests are "objects" in page
- If idle for more than 1 second, "think time"
- Note, does not include client cache





- ### Limitations of Methodology
- TCP analysis solid (harder to mistake number of packets, flows, etc)
 - HTTP analysis less certain
 - Pipelined exchanges (look like one)
 - User/browser interactions (stop, reload)
 - Browser and proxy caches
 - TCP processing to deal with loss, duplication, re-ordering

- ### Outline
- Introduction (done)
 - Related Work (done)
 - Data Sets (done)
 - Analysis (done)
 - Comparison with Others (next)
 - Sampling Issues
 - Conclusions



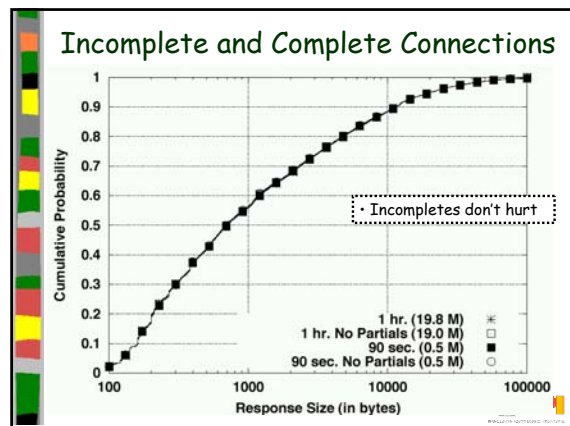
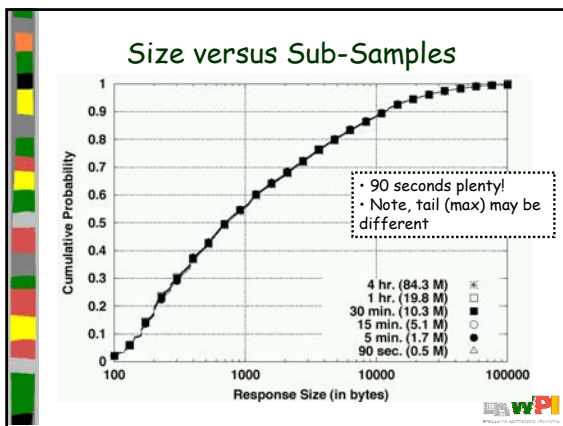
Summary Data

Data Set	Sample Size (Number of responses)	Min Response Size	Max Response Size	Mean Response Size	Median Response Size
W95	269,811	3	20,135,435	14,826	2,245
W98	66,988	1	4,092,928	7,247	2,416
Mah 95	5,300	62	8,146,796	10,664	2,035
UNC99	18,526,201	1	135,294,044	6,734	1,164
UNC01	84,343,238	1	984,871,070	6,397	722
UNC03	96,836,703	1	718,067,386	7,296	632

- Notice trend in median sizes
- Largest sizes are because of larger samples

- ### Outline
- Introduction (done)
 - Related Work (done)
 - Data Sets (done)
 - Analysis (done)
 - Comparison with Others (done)
 - Sampling Issues (next)
 - Conclusions

- ### Network Traces
- Size is a concern
 - 1 hour, headers (68 bytes) only, consumes 30 Gigabytes of storage
 - Processing takes *hours*
 - Capturing can slow down routers
 - So:
 - Do lengths of traces affect distribution shape?
 - Do incomplete TCP connections affect distribution shapes?



Conclusions

- Captured and Analyzed Web traffic for 35,000 UNC people, 3 sets from 3 years
- Find:
 - HTTP request sizes are increasing
 - HTTP response sizes are decreasing
 - Largest HTTP responses are increasing
 - Web pages complexity is increasing (more servers per page)



Future Work

- Effects of persistent connections and pipelining?
- What about other (non-port 80) traffic over HTTP?
 - About $\frac{1}{2}$ of all TCP traffic "other"
- Are all objects Web objects?
 - As opposed to re-direction requests, error messages
 - This may help understand Web structure

