The University of North Carolina at Chapel Hill
Department of Computer Science

11th ACM/IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS)
Orlando, October 13th, 2003

Tracking the Evolution of Web Traffic: 1995-2003

Félix Hernández-Campos
Kevin Jeffay
F. Donelson Smith

http://www.cs.unc.edu/Research/dirt

Web Traffic Measurement and Analysis at UNC-Chapel Hill

- In 1997, *populating web traffic generators* for experimental networking research motivated a large-scale study of web traffic at UNC with three goals:

✓Develop a light-weight methodology
  – Based on passive measurement
  – *Easy* to maintain models up-to-date

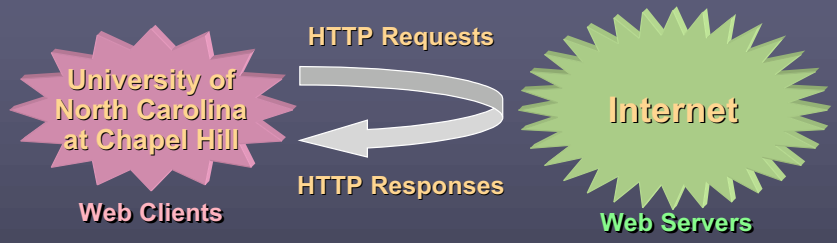
✓Replace smaller-scale, quickly aging models
  – Mah, 1995 data set
  – Crovella *et. al*, 1995 data set (revised with 1998 data)

✓Characterize the use of the HTTP protocol
  – *E.g.*, Use of persistent connections

Web Traffic Measurement and Analysis at UNC-Chapel Hill

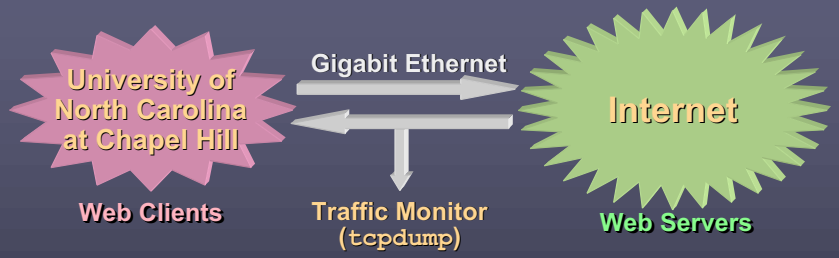
- Our methodology and first results were published in SIGMETRICS/Performance'01
  – *What TCP/IP Protocol Headers Can Tell Us About the Web*
- Modeling aspect explored in a series of papers
  – *E.g., Variable Heavy Tails in Internet Traffic* (with J.S. Marron)
    » (Part I: *Understanding Heavy Tails* published in MASCOTS'02)
- In this talk, I will describe our approach and our observation on the evolution of web traffic:
  – Three data sets: 1999, 2001 and 2003
  – Comparisons to Mah and Crovella *et al.*

Methodology
Study of Web Content Consumers



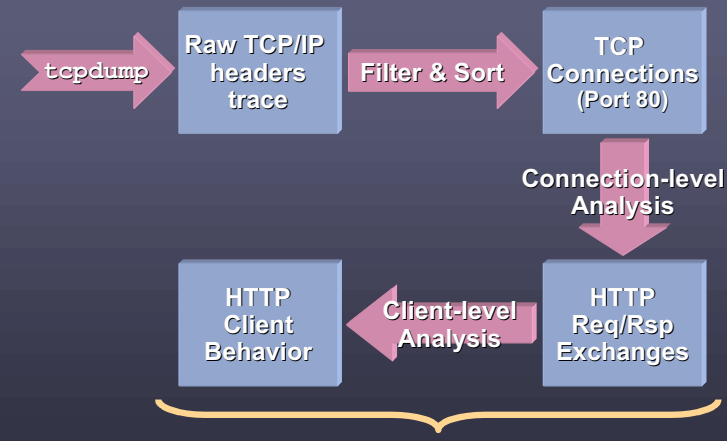
- We studied a large collection of users (~35,000) as *web content consumers*
- The only source of data for our study were packet header traces
  – Anonymized IP addresses
  – No HTTP headers

Methodology
One-Way Packet Header Traces

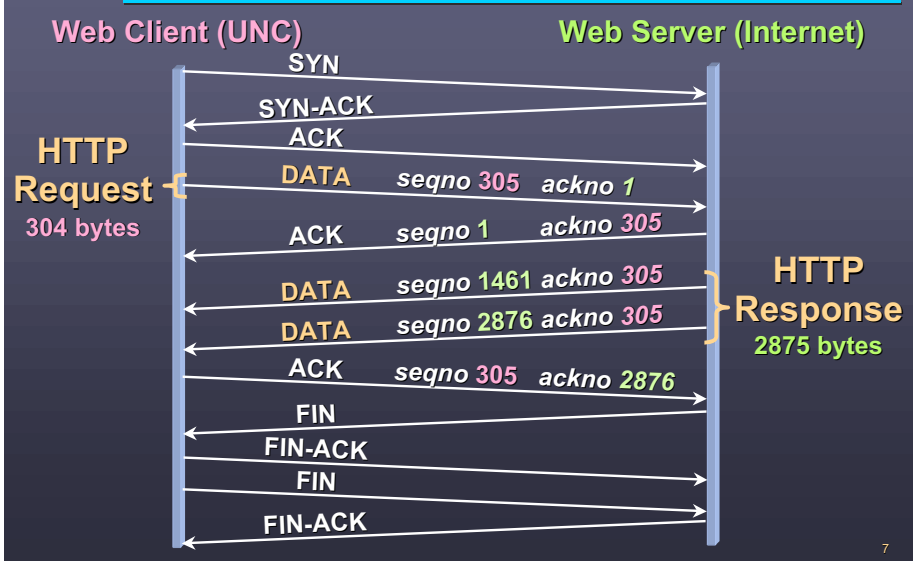


- Only inbound TCP/IP headers are captured
  - Eliminate synchronization and buffering issues on the NIC
  - Reduce trace size
- Trace collection: 2.7 TB of packet headers
  - ~40 billion packets ~16 TB of data transfers

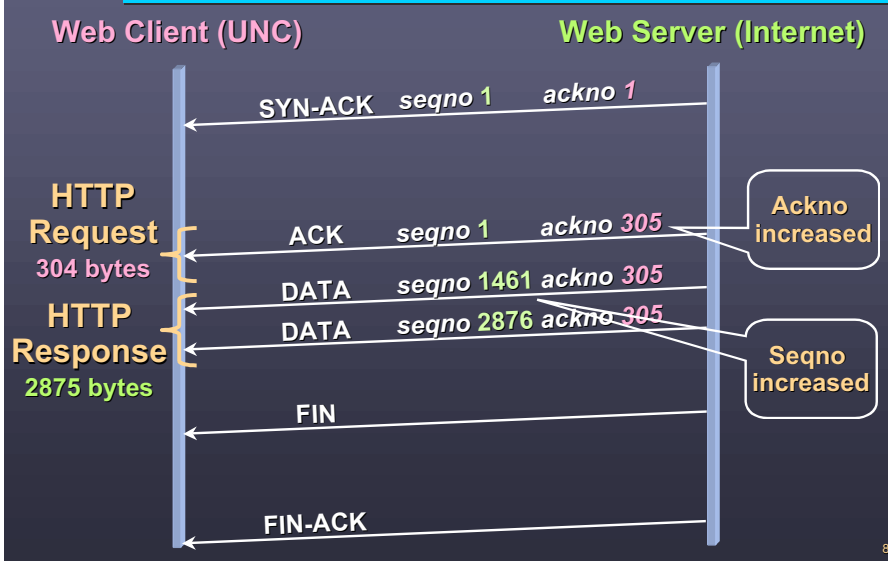
Methodology
Processing Sequence Overview



TCP/IP Headers and HTTP
Request/response Exchange

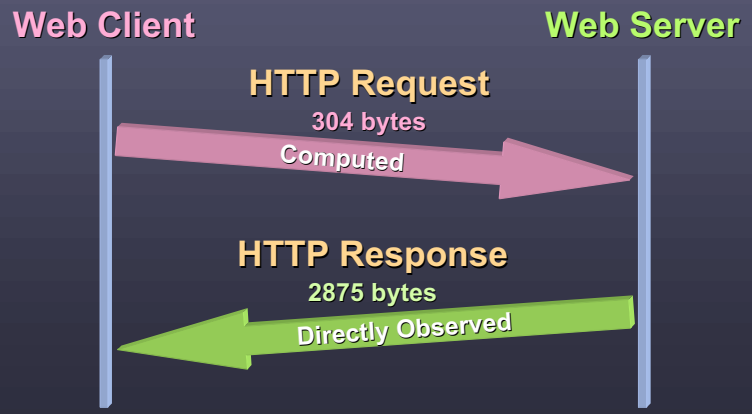


TCP/IP Headers and HTTP
Server-to-client Segments Only

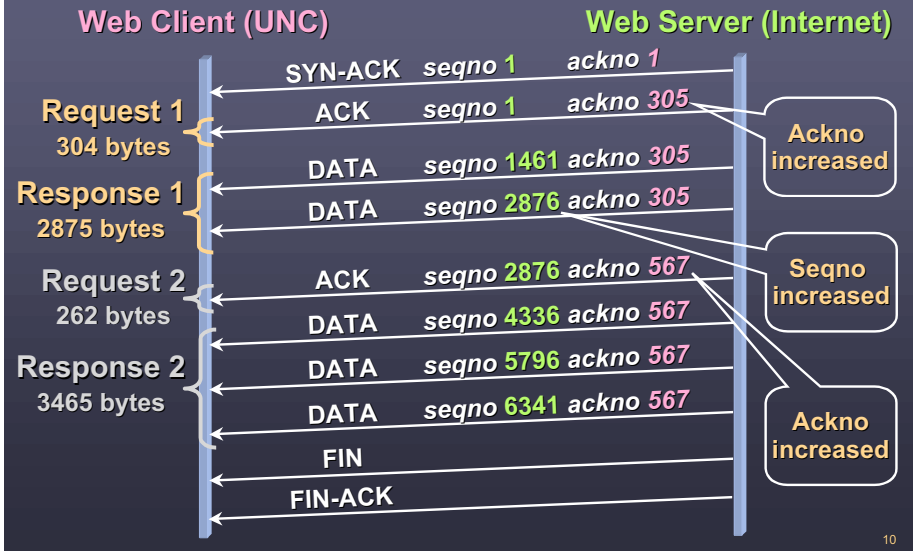


Methodology
Request/Response Traces

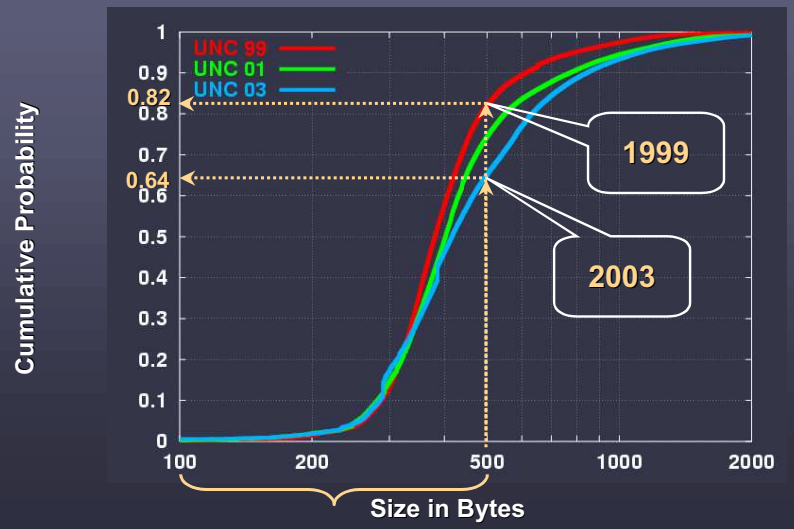
- Unidirectional TCP/IP header traces are sufficient for capturing application-level behavior



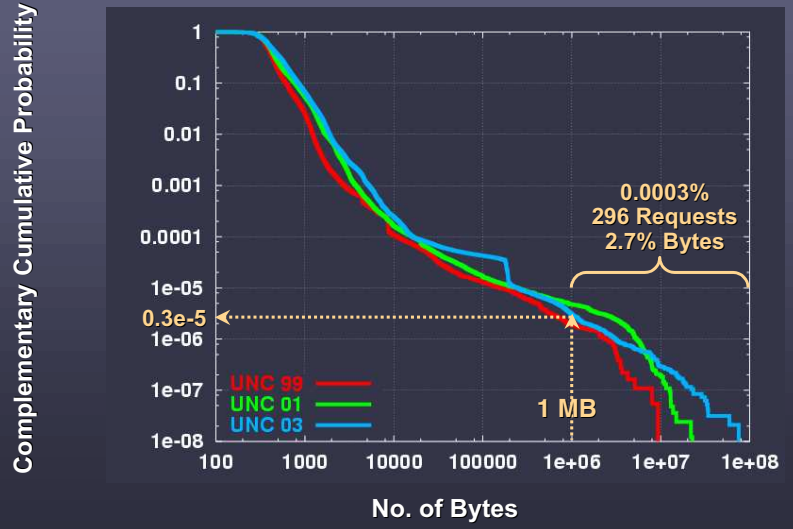
Persistent Connections in HTTP
Example – TCP/IP Headers



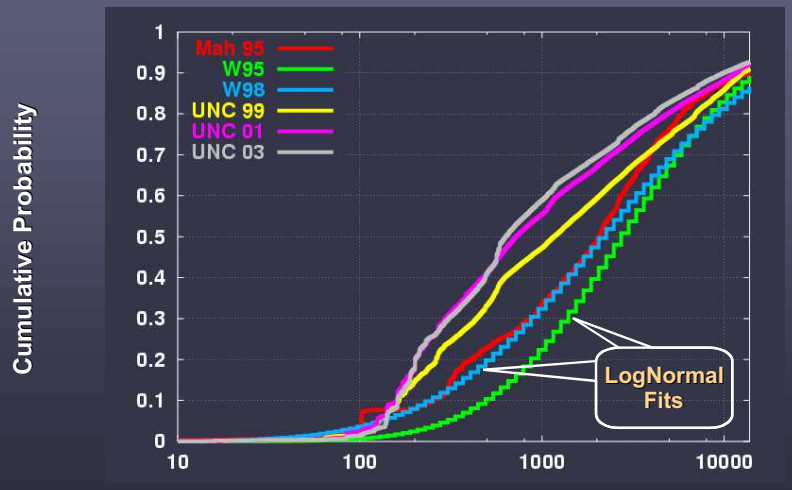
Sizes of HTTP Requests
Empirical CDFs



Sizes of HTTP Requests
Empirical CCDFs

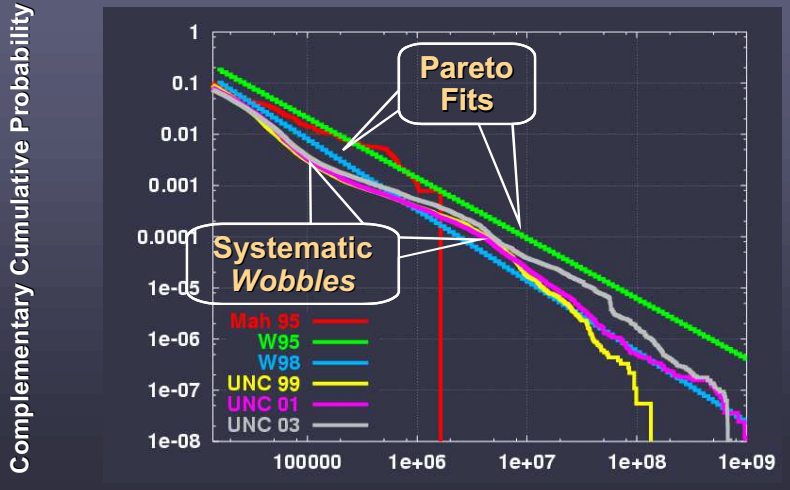


Response Sizes
Comparison with Earlier Studies



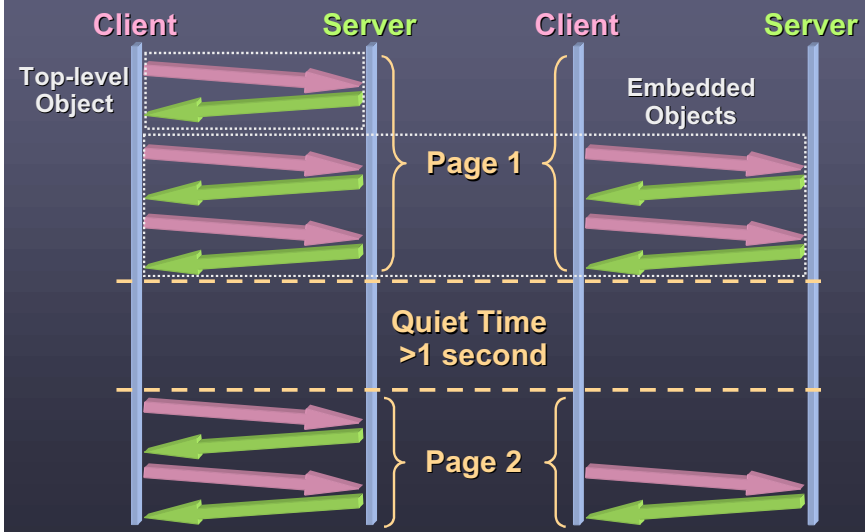
13

Response Sizes
Comparison with Earlier Studies



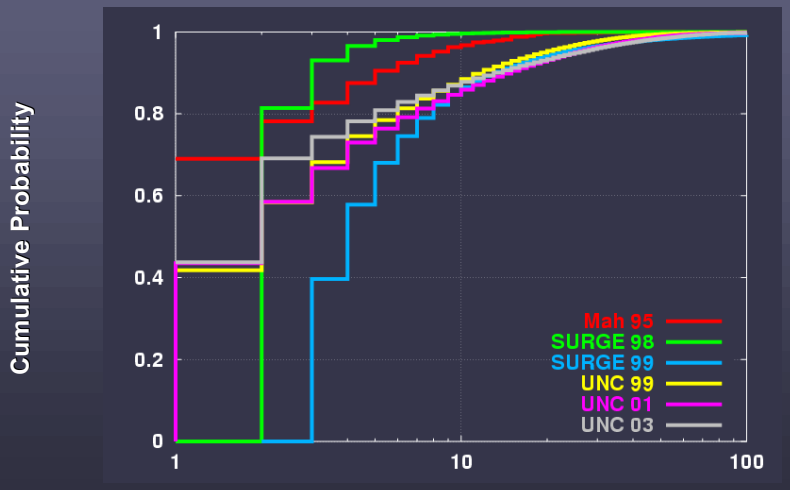
14

Page Identification Heuristic
Two TCP Connections Example



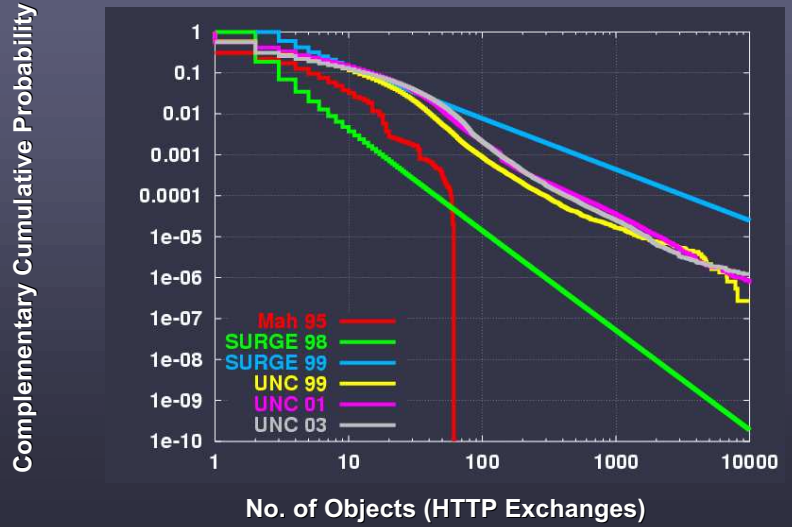
15

Objects Per Page
Comparison with Earlier Studies

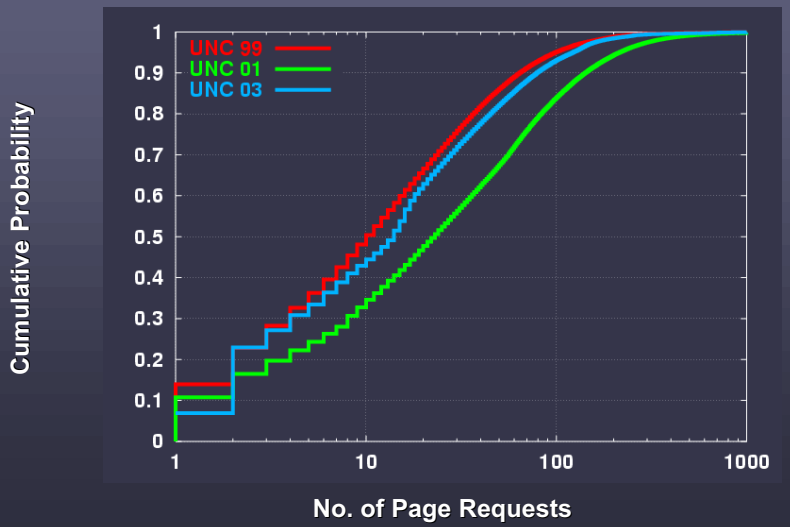


16

Objects Per Page
Comparison with Earlier Studies



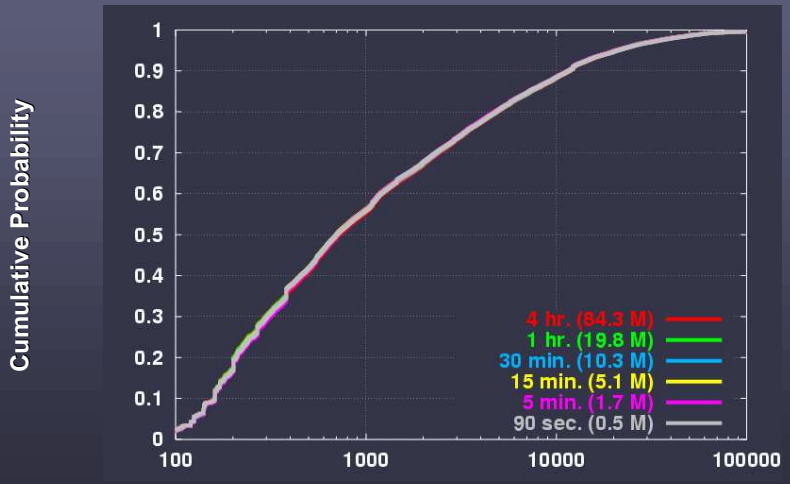
Page Requests Per IP Address



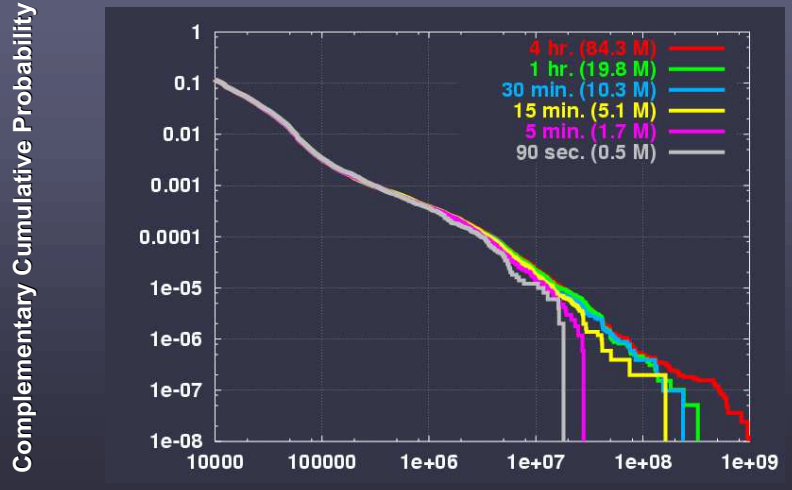
Sampling Issues

- Questions:
  - Can we obtain a *sufficiently large sample* with a small number of short traces?
  - How does the *length of the tracing interval* affect the overall empirical distribution shapes?
  - Should we include in the empirical distributions the data from *incomplete TCP connections*?
- Approach:
  - Examine a wide range of trace lengths
    - » 4 h., 2 h., 1h., 30 min., 15 min., 5 min. and 90 sec.
  - Construct datasets by sub-sampling the 21 4-hour-long traces collected in 2001
  - *E.g.*, remove first and last hour of each trace to produce 21 2-hour-long traces

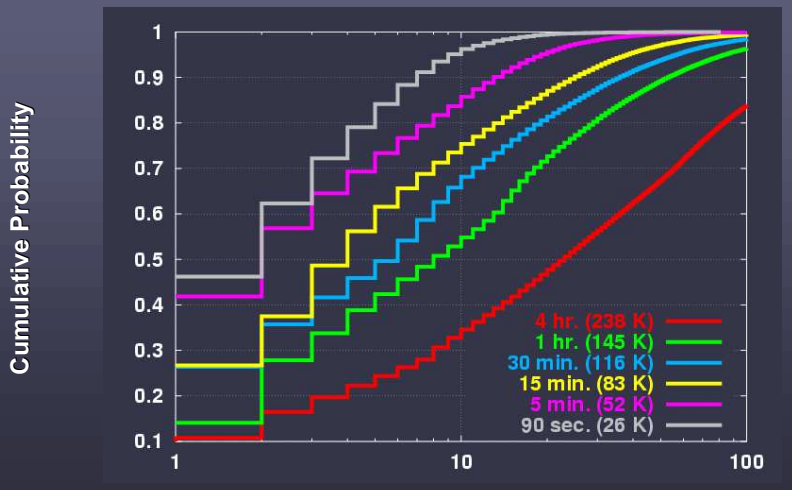
Sampling Issues
Impact of Tracing Interval Length



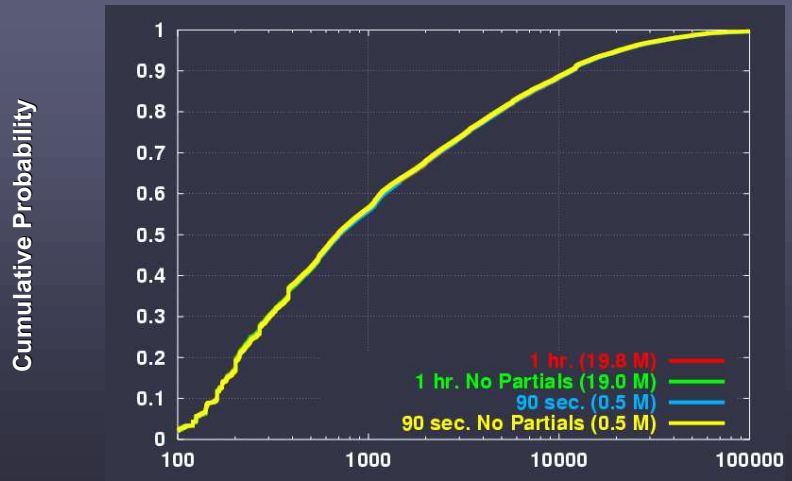
Sampling Issues
Impact of Tracing Interval Length



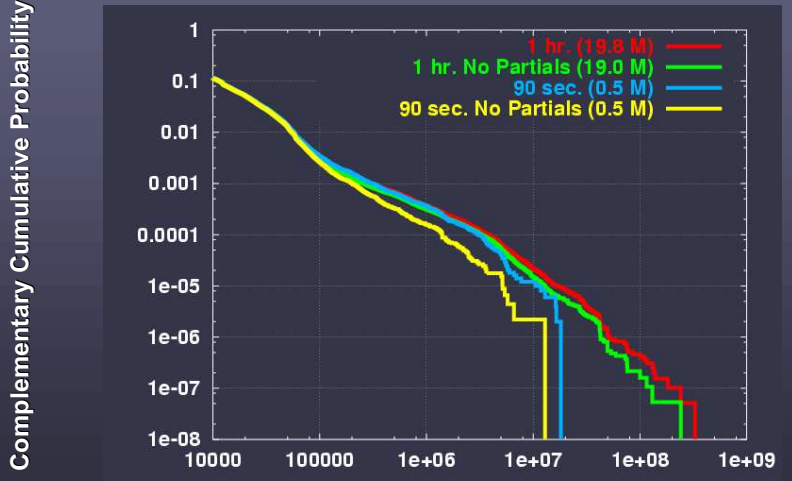
Sampling Issues
Impact of Tracing Interval Length



Sampling Issues
Impact of Partially-Captured Objects



Sampling Issues
Impact of Partially-Captured Objects



Summary and Conclusions
Web Traffic Characterization

- New data to populate traffic generators
  - Request sizes
  - Response sizes
  - Use of persistent connections
  - ...
- 1-hour long traces are sufficient to capture application-level behavior
  - Short traces cut off large objects, which skews the tails of the distributions
- Persistent Connections:
  - ~15% of all the HTTP connections
  - 40-50% of all the transferred HTTP bytes