

An Empirical Model of HTTP Network Traffic

Bruce A. Mah
bmah@ca.sandia.gov

University of California at Berkeley
and
Sandia National Laboratories



IEEE INFOCOM '97
10 April 1997

Motivation

HTTP dominates wide-area Internet traffic

Leading contributor to byte- and packet-count across NSFNET backbone as early as April 1995

A synthetic workload of this application is needed

Network simulators

Benchmarks

Synopsis

We have constructed a synthetic workload of HTTP network activity based on traffic traces.

This model is consistent with prior Web measurement studies.

Overview

Prior Work

Model Components

Methodology

Measurements

Prior Work

Measurement Studies

- Server logs Only measure activity at one server
- Client logs Require instrumented clients

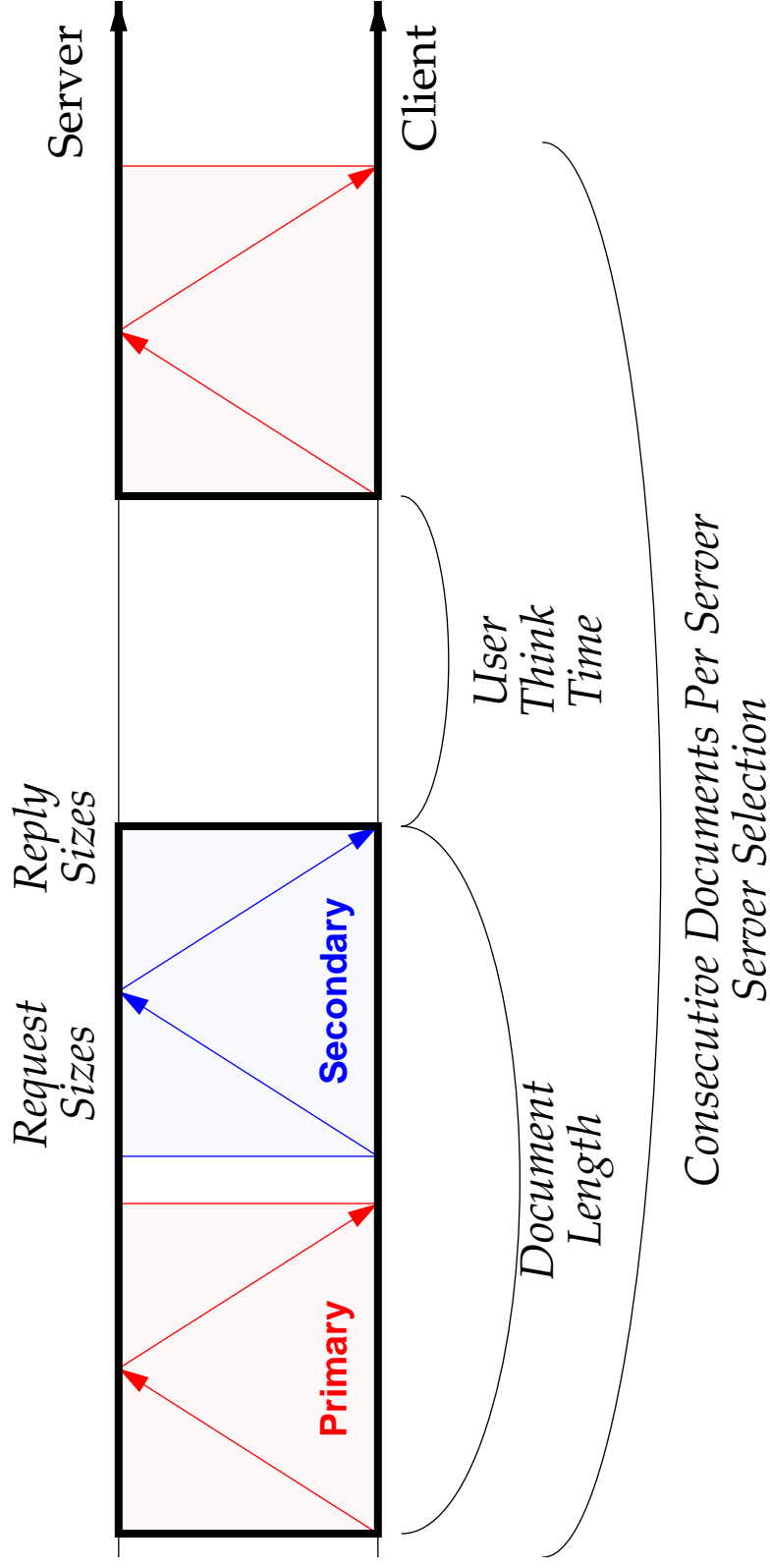
Static Document surveys

- Document indices Don't capture user reference patterns

Synthetic Workloads

- tcplib Predates the World Wide Web

Model Components



Each component defined by a probability distribution

Methodology

Packet Traces

- tcpdump on an Ethernet
- Easily trace many clients
- Capture protocol overheads
- Lose higher-level information (documents, cache behavior)

Filtering

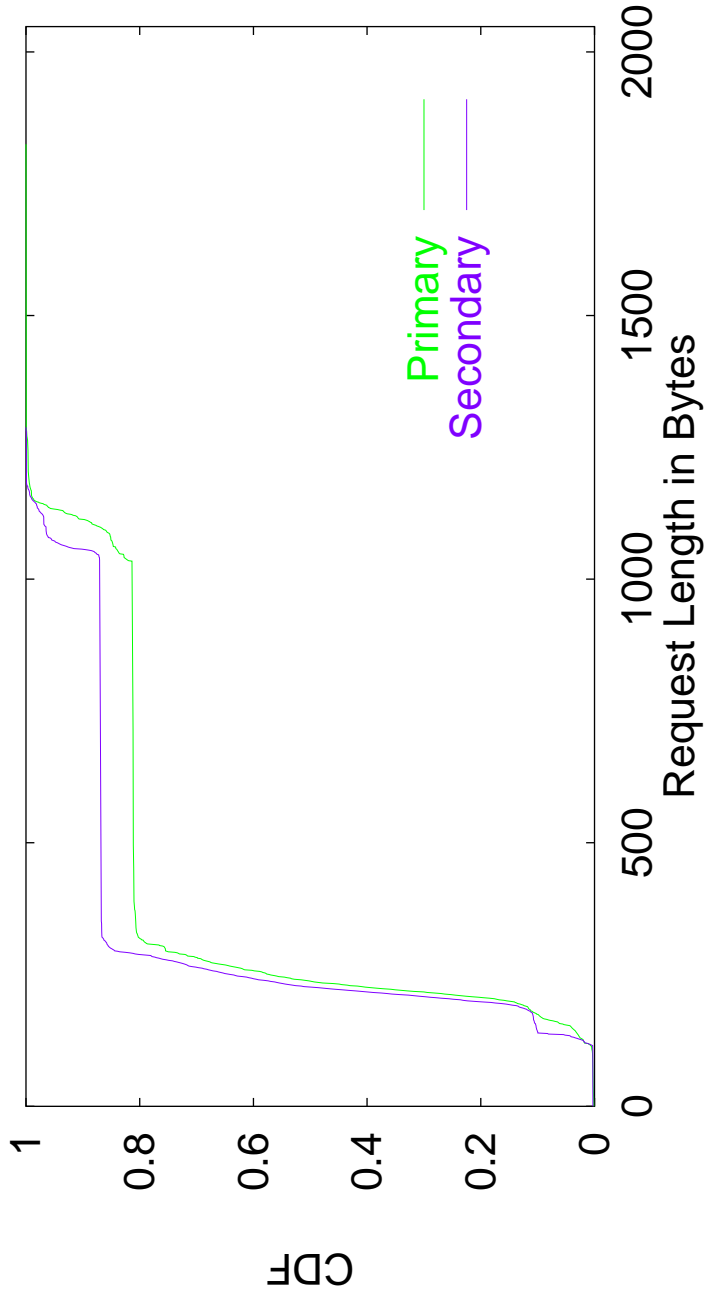
- Remove non-local clients
- Remove periodic retrievals

Apply Heuristics

- Attempt to recover some higher-level structures

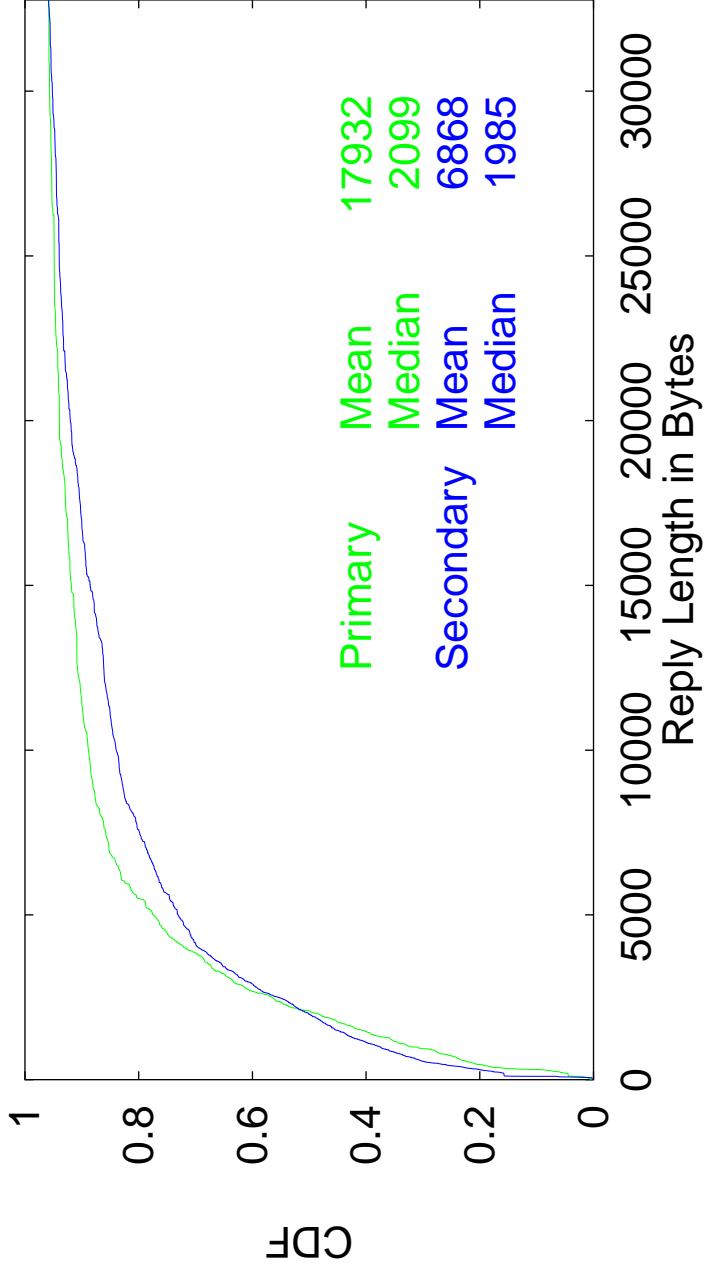
Construct Probability Distributions

Request Sizes



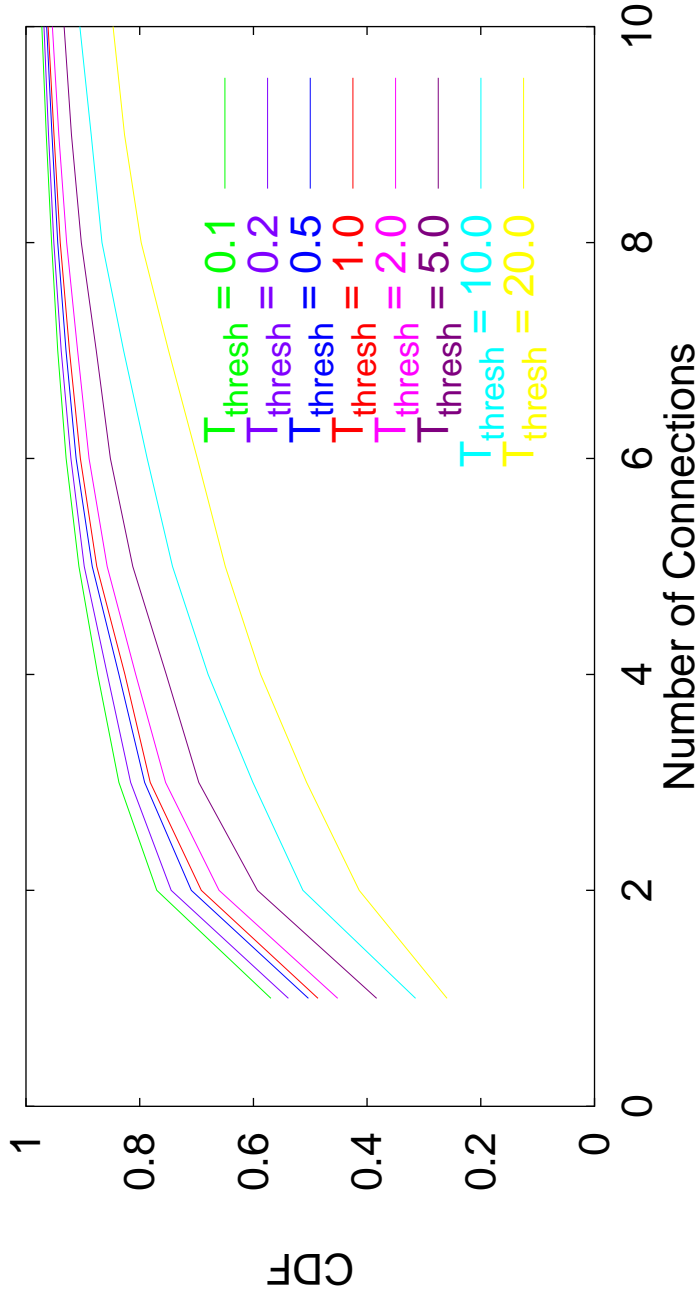
Request sizes have a bimodal distribution

Reply Sizes



HTTP reply sizes have a heavy-tailed distribution
Primary replies tend to be longer than secondary replies

Document Length

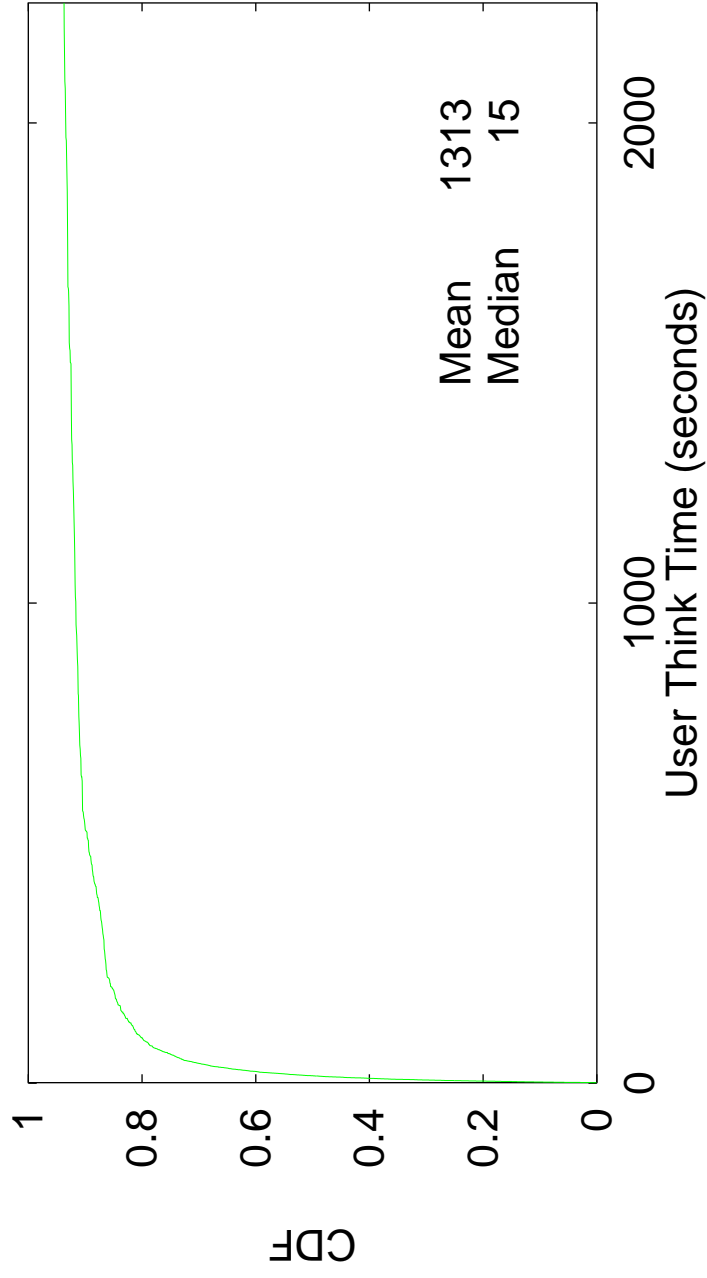


A timing heuristic can discriminate between documents

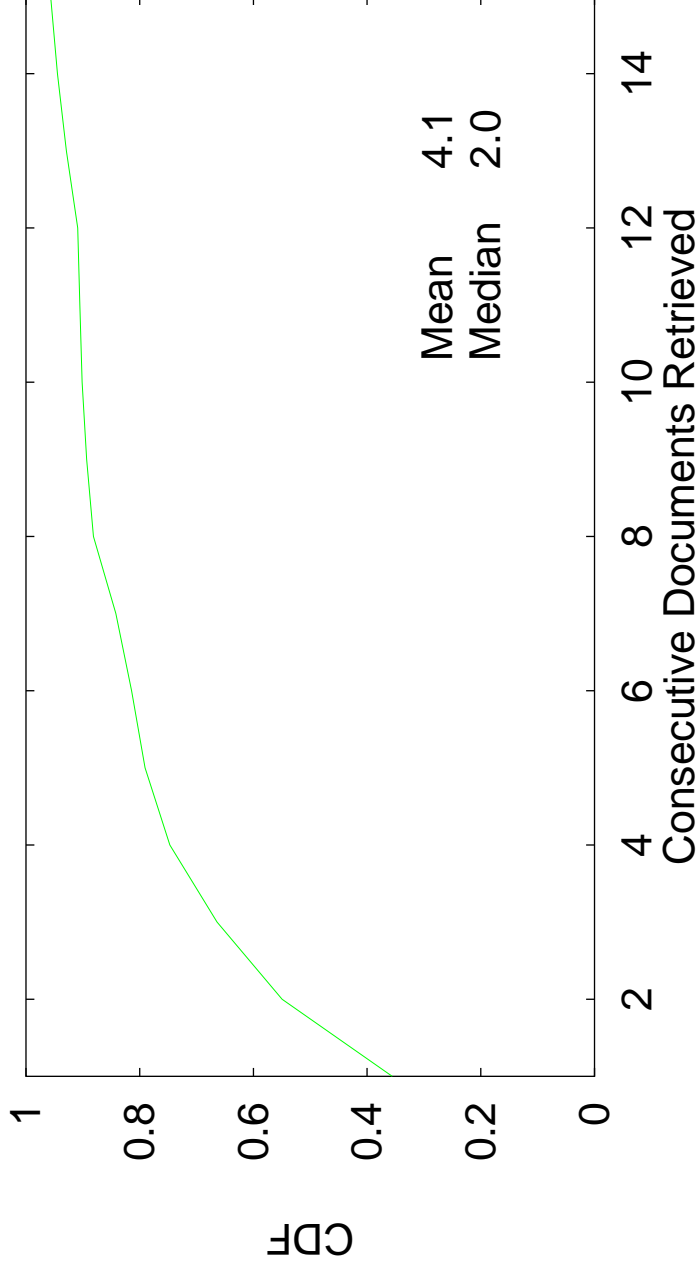
80% of documents require fewer than four file transfers

Picked idle threshold $T_{\text{thresh}} = 1.0$ seconds

User Think Time



Consecutive Documents per Server



80% of visits to a server retrieve fewer than six documents

Server Selection

Rank	#	Location
1	43	* .cs.berkeley.edu
2	11	* .berkeley.edu
3	8	* *.com
4	7	* *.com
5	6	* .cs.berkeley.edu
6	6	* *.com
7	6	www4.*.net
8	6	* .cs.berkeley.edu
9	5	www5.*.com
10	5	www.*.com

Four hosts in the top ten ranking are local

Sample size seems small, Zipf's Law approximation

Areas for Future Work

Better Server Selection Distribution

- Filter effects caused by local servers

- Larger sample size

Persistent-Connection HTTP

- Can't use TCP connection sizes to determine request/reply sizes

Correlation between model distributions

- Do users retrieve more or fewer consecutive documents from “popular” sites?

Newer datasets

Summary

Packet traces can help to build a model of HTTP traffic

Characterized HTTP network traffic to build a synthetic workload

Results consistent with prior Web measurement studies

C++ source code available for simulators

For more information and model data:

<http://http.cs.berkeley.edu/~bmah/Software/HttpModel>