An Empirical Model of HTTP Network Traffic

Bruce A. Mah
bmah@CS.Berkeley.EDU

The Tenet Group
University of California at Berkeley

Daedalus Meeting
19 July 1996

Motivation

HTTP dominates wide-area Internet traffic

- Leading contributer 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

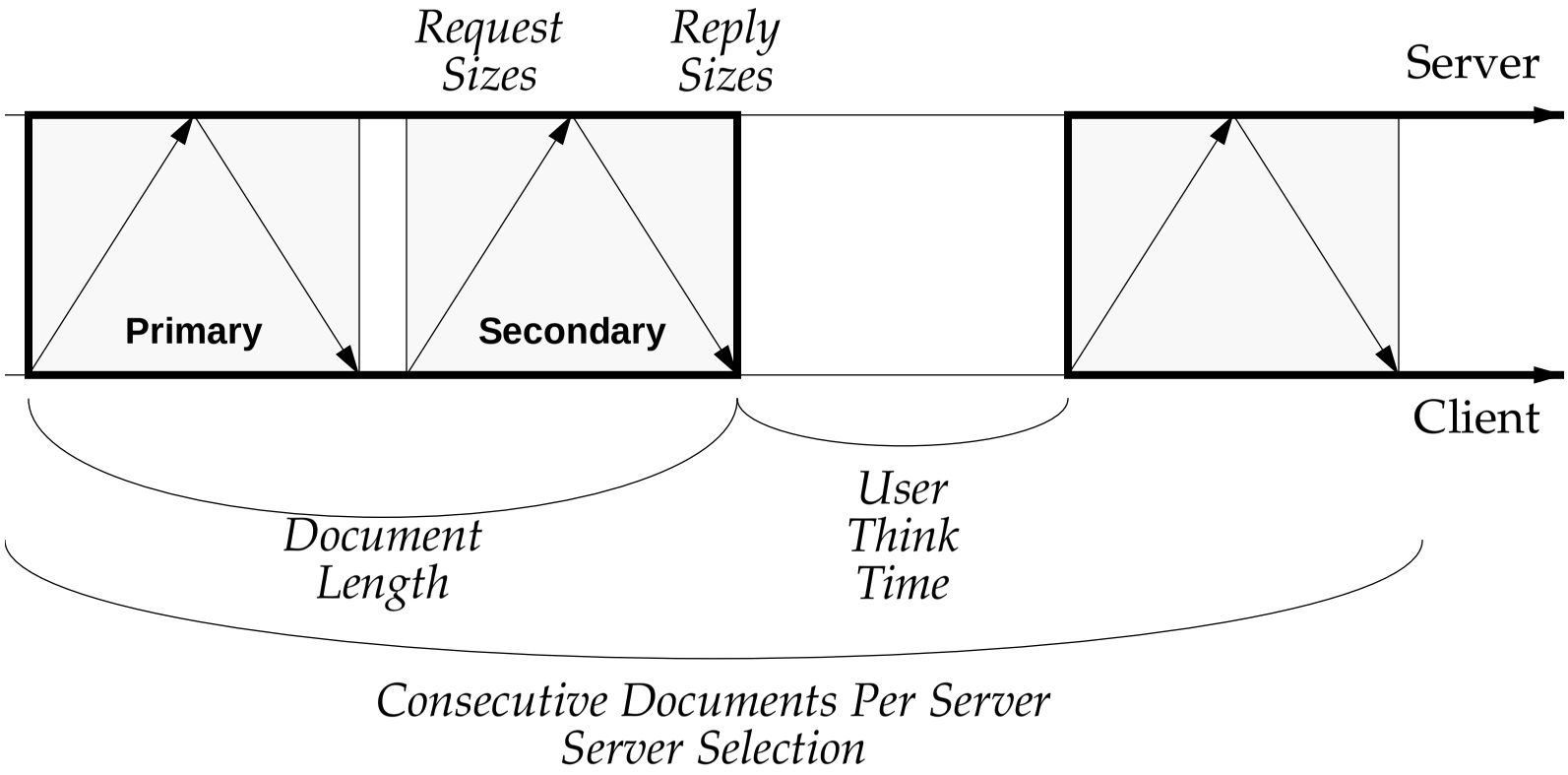
Static Document surveys

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

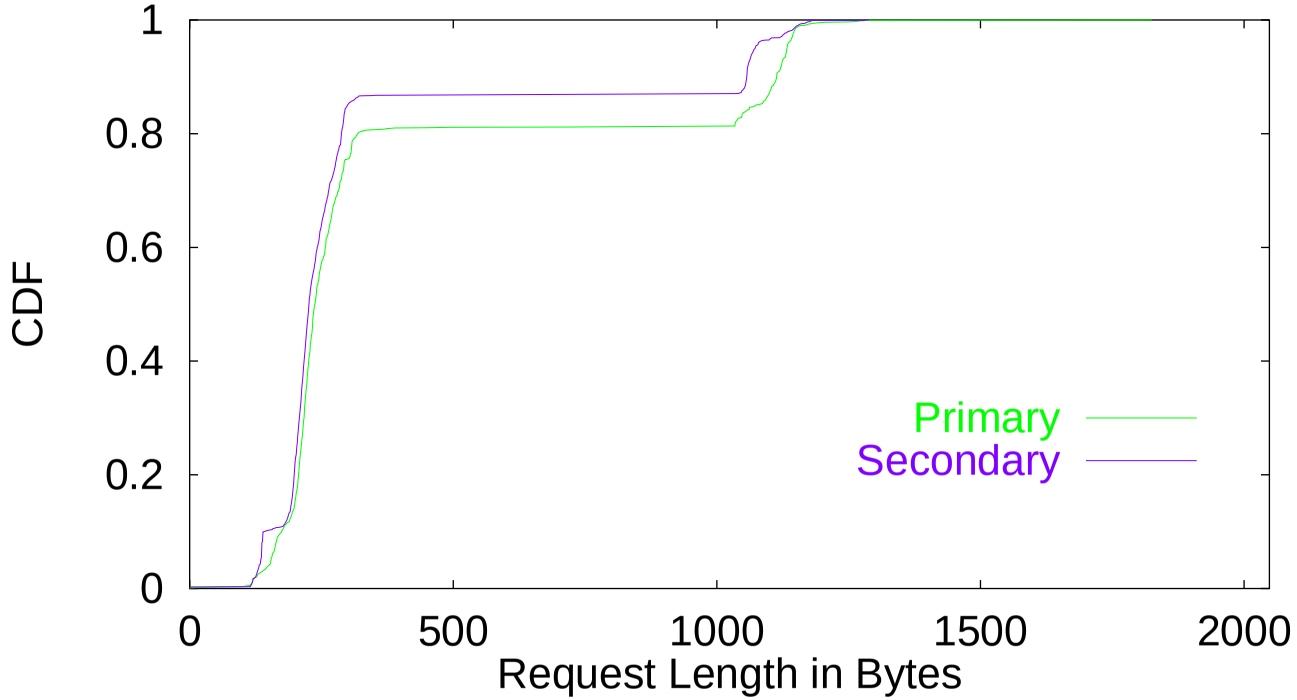
- Remove non-local clients
- Remove periodic retrievals

Apply Heuristics

- Attempt to recover some higher-level structures

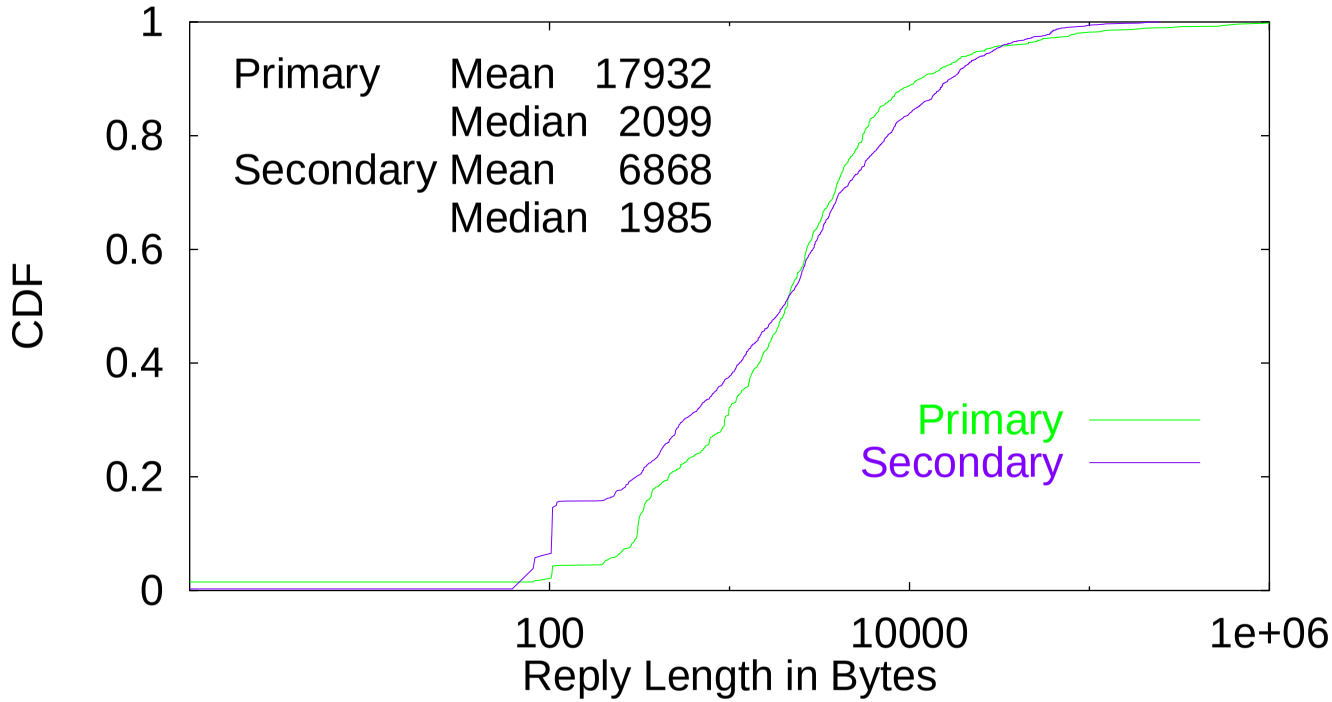
Construct Probability Distributions

Request Sizes



Request sizes have a bimodal distribution

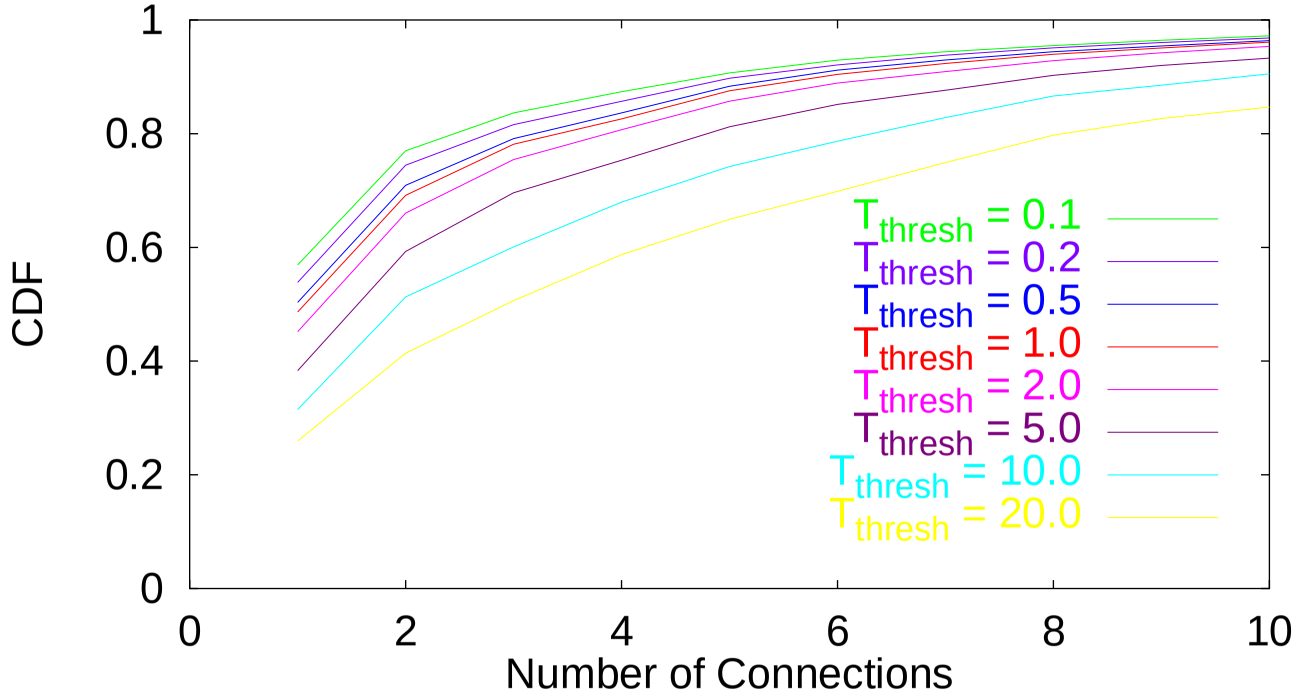
Reply Sizes



HTTP reply sizes have a heavy-tailed distribution

Primary replies tend to be longer than secondary replies

Document Length

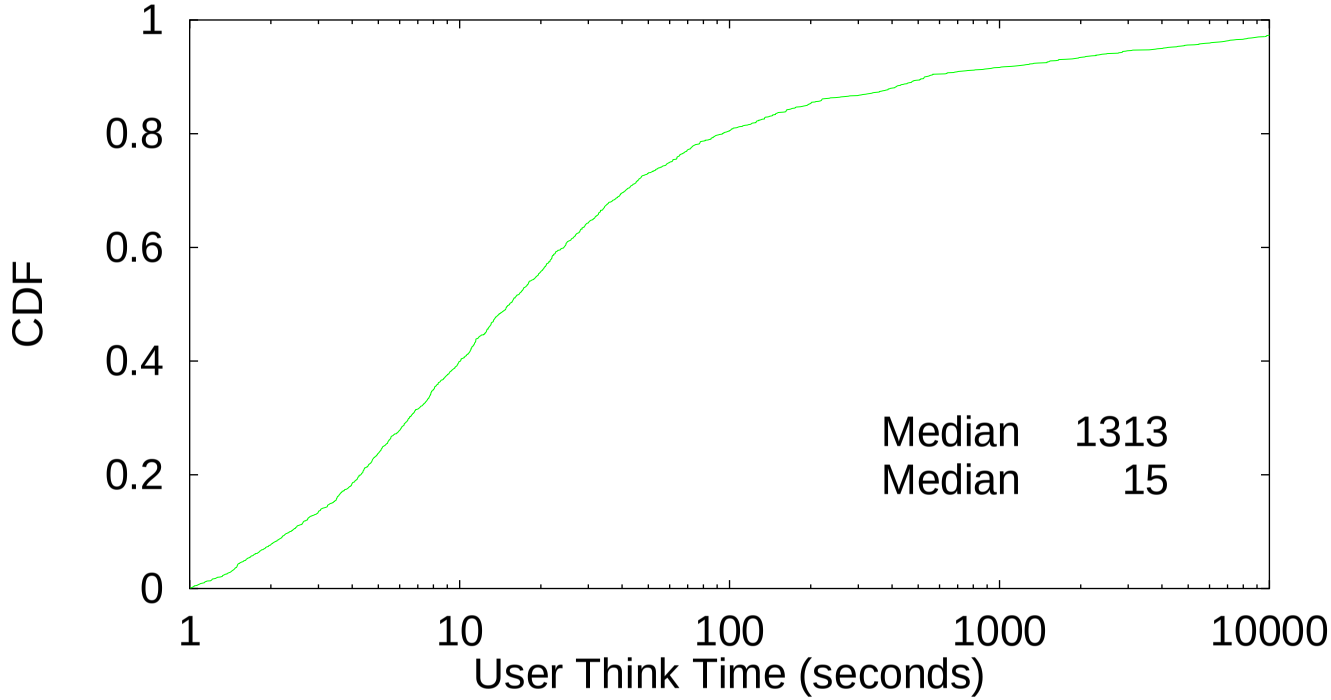


A timing heuristic can discriminate between documents

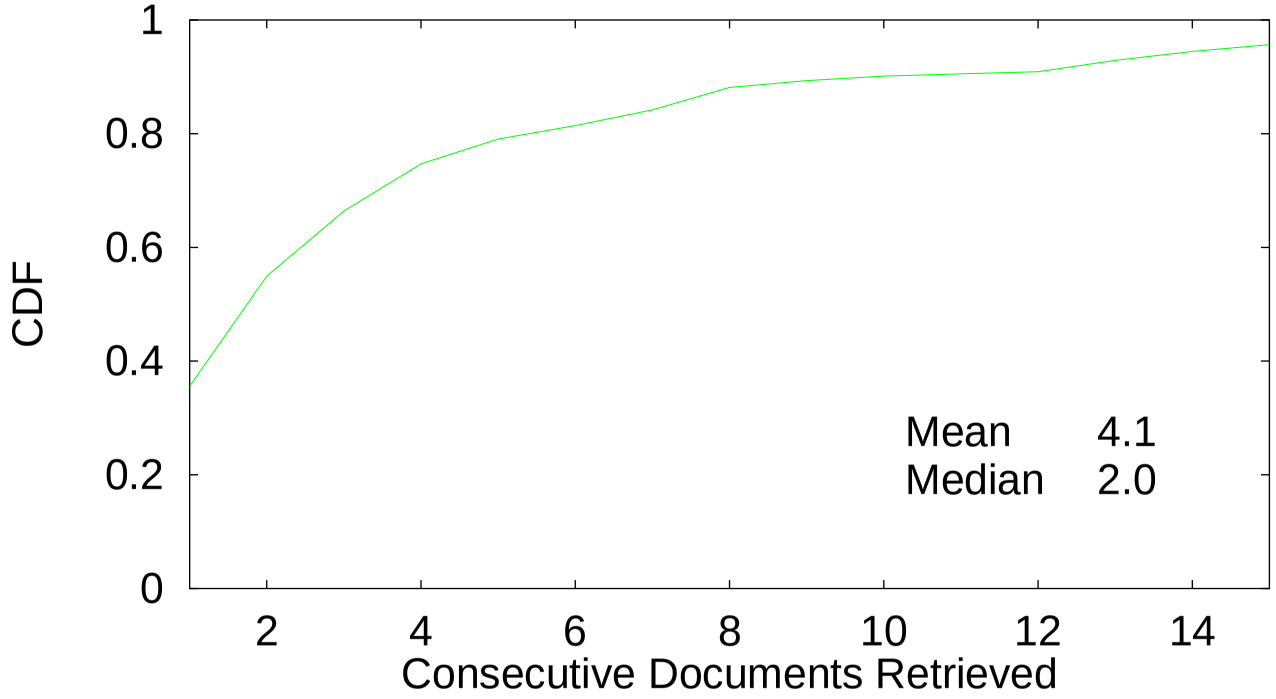
80% of documents require fewer than four file transfers

Picked idle threshhold T_{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	#	Host
1	43	**kholer.cs.berkeley.edu**
2	11	**inktomi.berkeley.edu**
3	8	zcias4.ziff.com
4	7	networth.galt.com
5	6	**daedalus.cs.berkeley.edu**
6	6	farstar.secapl.com
7	6	www4.nando.net
8	6	**now.cs.berkeley.edu**
9	5	www5.yahoo.com
10	5	www.triangle-st.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?

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 (e.g. INSANE)

For more information and model data:

`http://http.cs.berkeley.edu/~bmah/Software/HttpModel`