

Self-Similarity and Power Laws in the Web



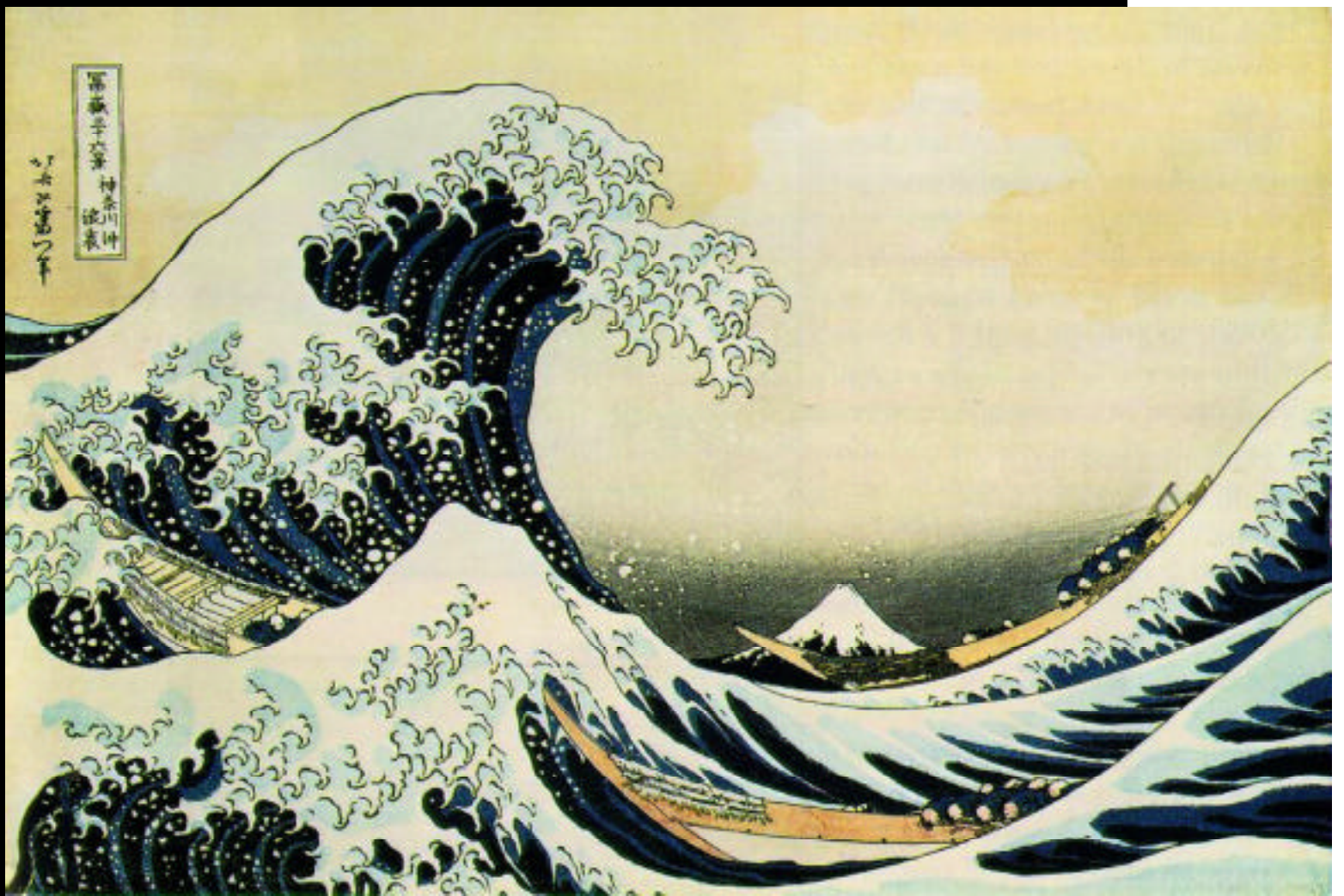
Boston University Computer Science

Mark Crovella

in collaboration with

Azer Bestavros, Murad Taqqu, Kihong Park,
Gi Tae Kim, Paul Barford, and Adam Bradley

The Great Wave (Hokusai)



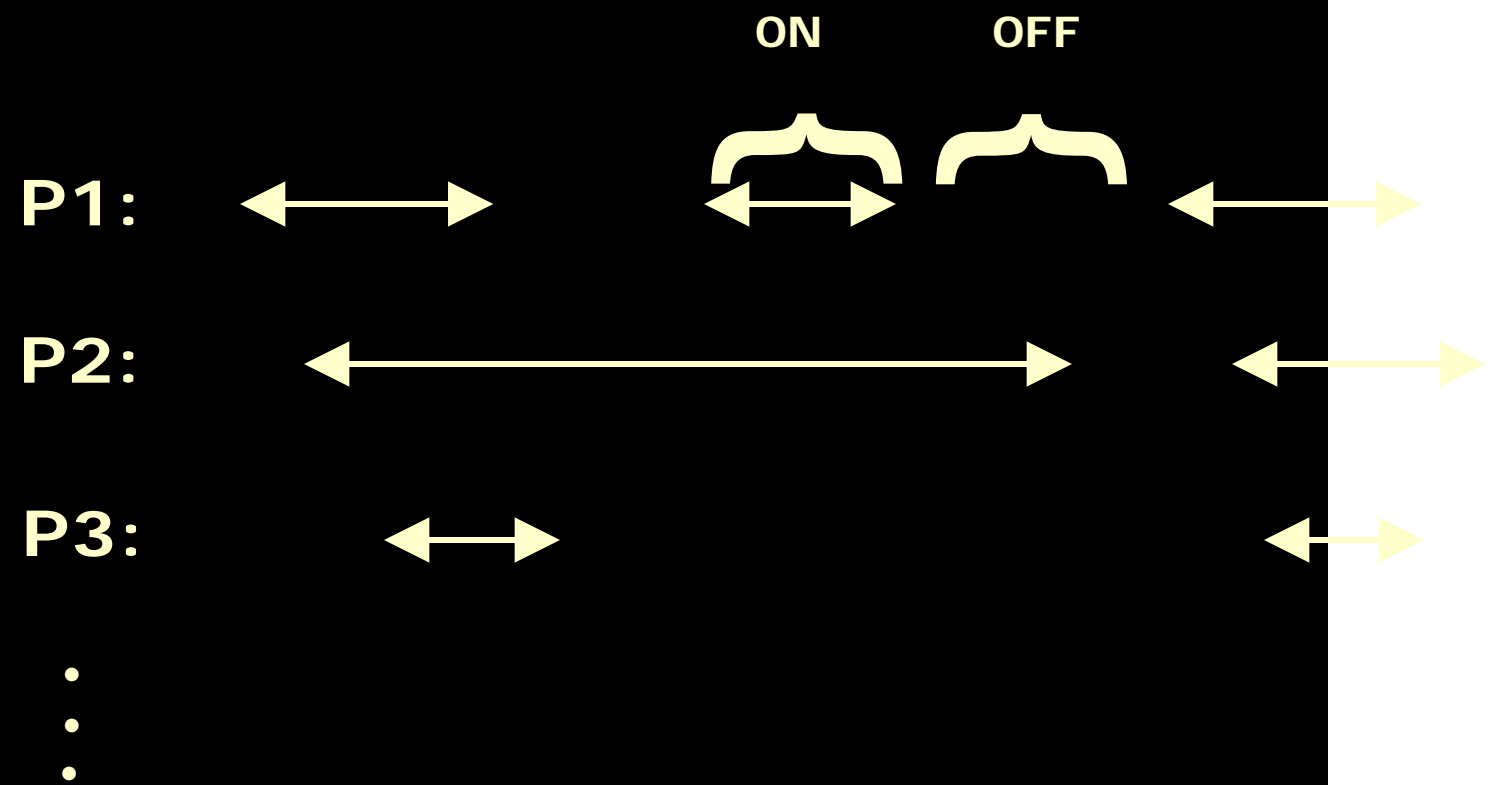
How Does Self-Similarity Arise?

Theoretical explanation:
superimposition of many **heavy-tailed ON/OFF processes**

ON/OFF process: one that alternates between sending at fixed rate, and being silent

heavy-tailed: having a distribution whose tail declines like a power law with low exponent:

ON/OFF Processes



Superimposing ON/OFF Sources

At each point in time, count how many processes are in the ON state

This forms the value of the aggregate process at that time instant

If either the ON or OFF periods are drawn from a heavy-tailed distribution, the aggregate process will show self-similarity with

How does this apply to the Internet?

ON/OFF processes could correspond to users transmitting data objects through the network

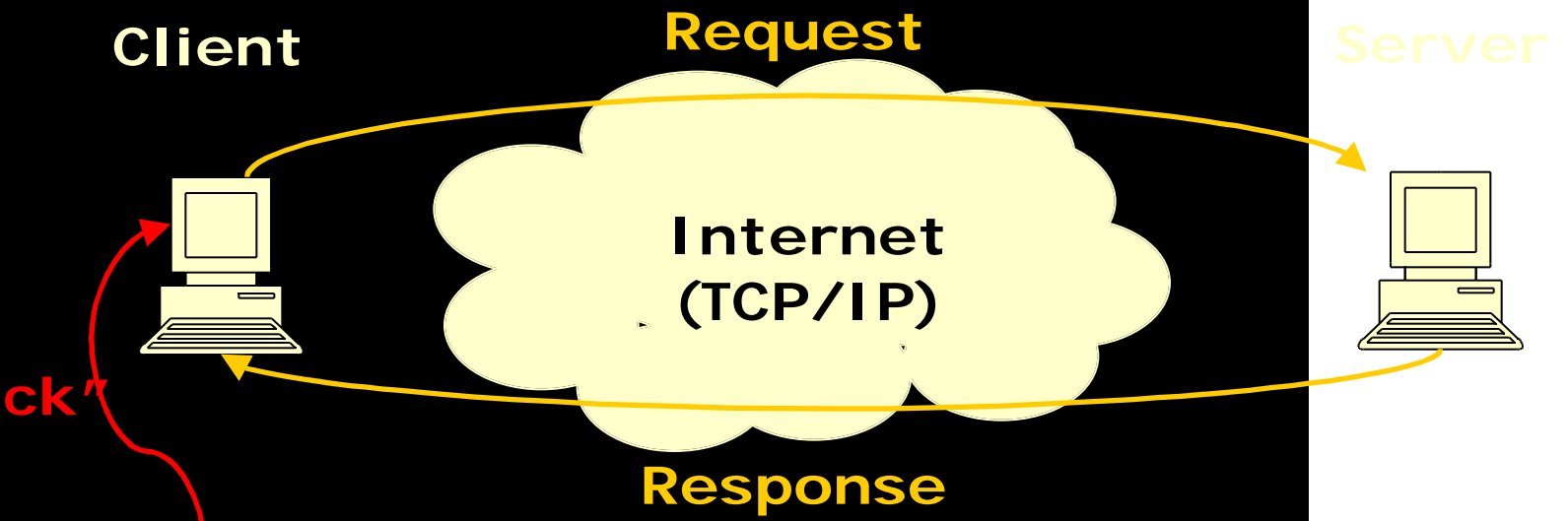
So we would like to understand if transmission times in the Internet are heavy tailed, and **why**

But the Internet supports a wide variety of applications

- Email, file transfer, multimedia, etc.

So we focus on the (currently) most popular application: the Web

The Web in action



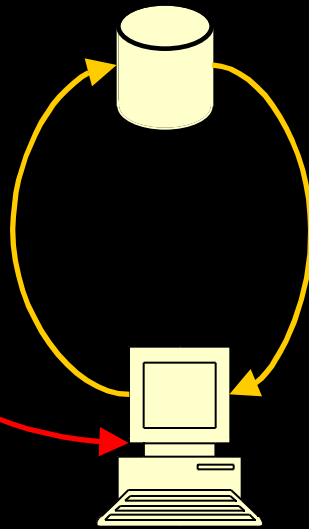
ck"



Client Caches

Local cache on disk

"click"



Client

Server



The ON/OFF Model in the Web

Traffic due to the Web alone is self-similar

Does the ON/OFF model apply?

- Browser sessions correspond to ON/OFF processes
- ON times correspond to transmission of Web files
- OFF times are user's idle periods

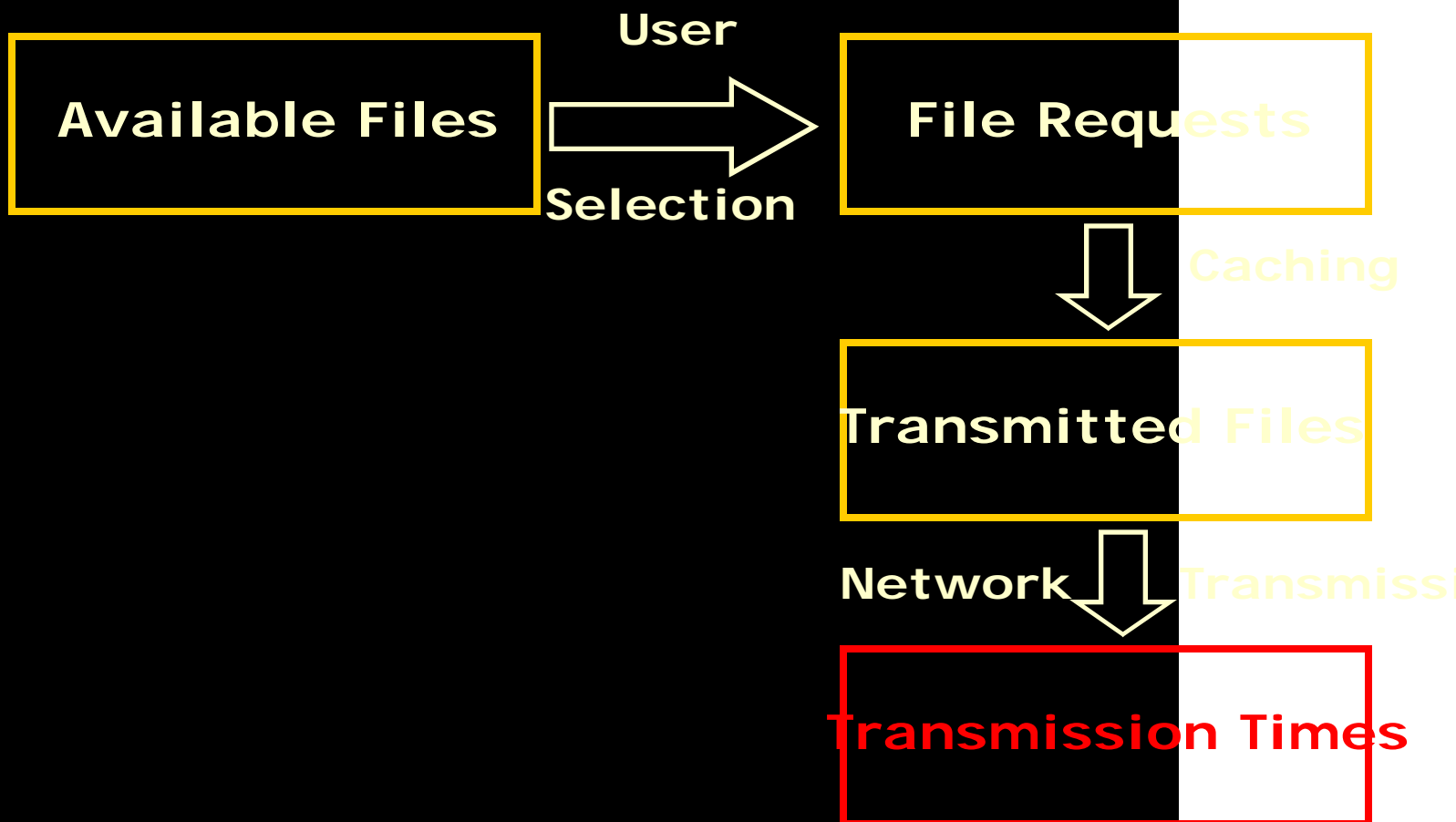
Questions:

- For the Web, are ON and OFF periods heavy-tailed?

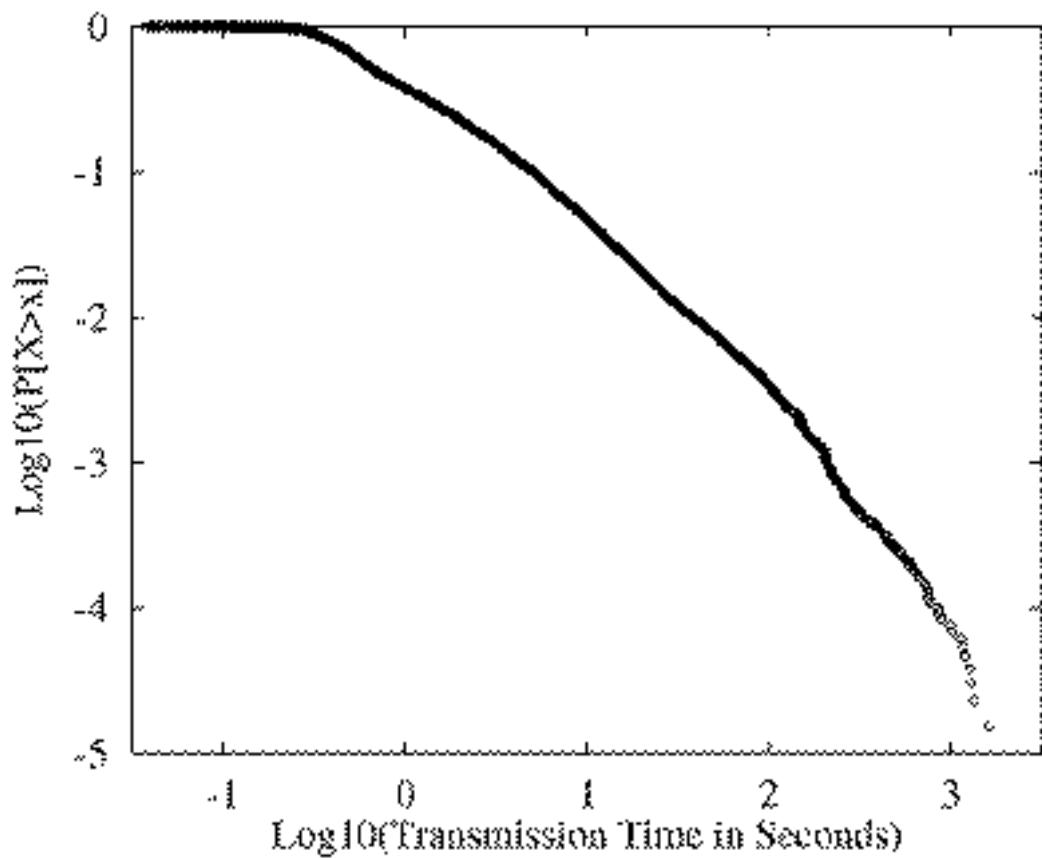
Capturing Web Traffic and User Activity

- Instrumented Mosaic and installed it as default
 - Records every URL requested by any user
 - Measures size of file, transfer time
- At the time, Mosaic was the dominant browser
 - January and February 1995
- Collected extensive data: 4700 Mosaic sessions
 - 591 users
 - URLs requested: 575,775

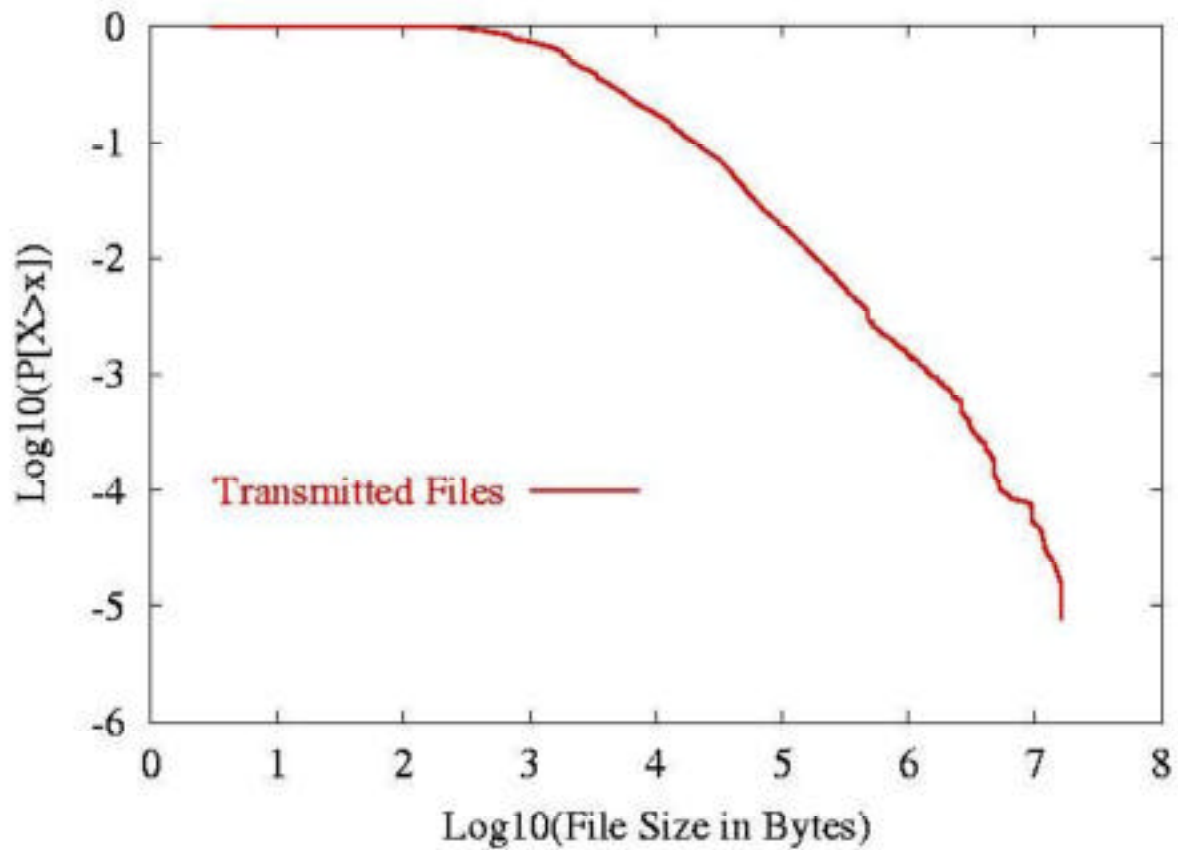
Relationships Among Datasets



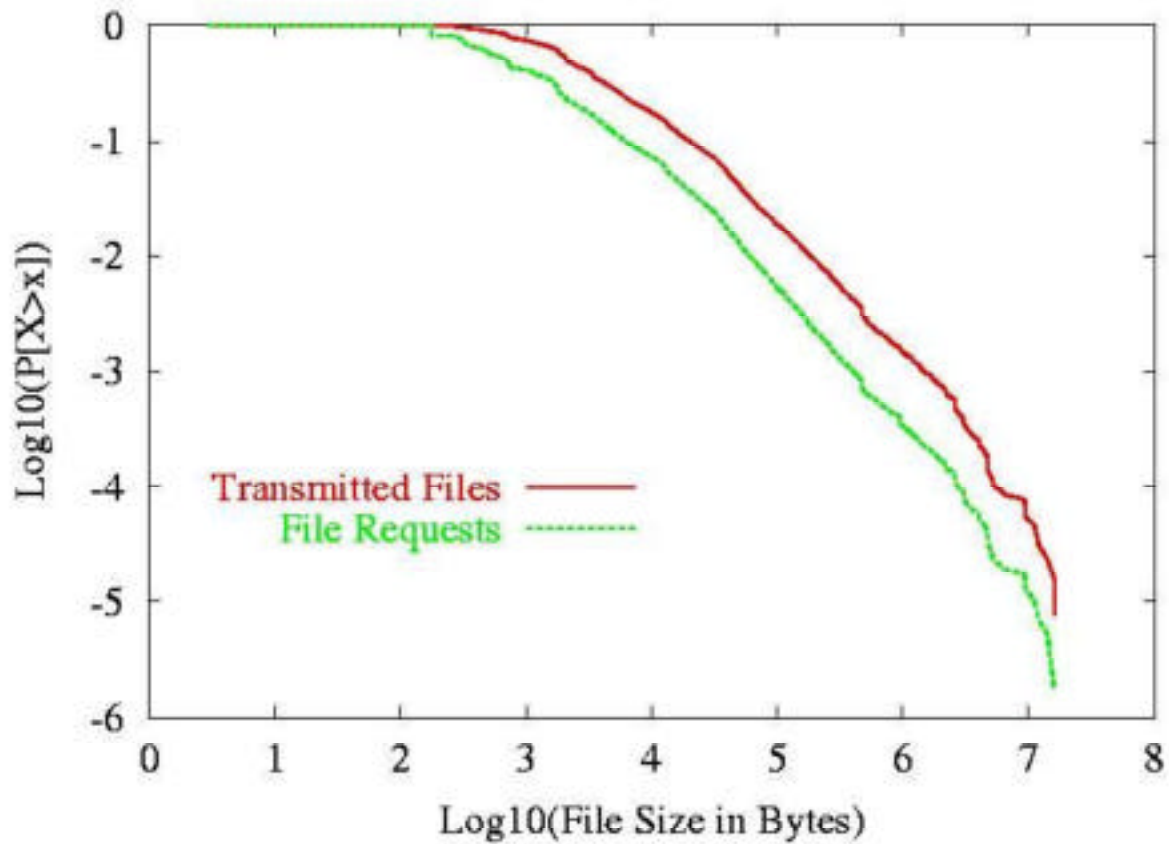
Transmission Times are Heavy Tailed



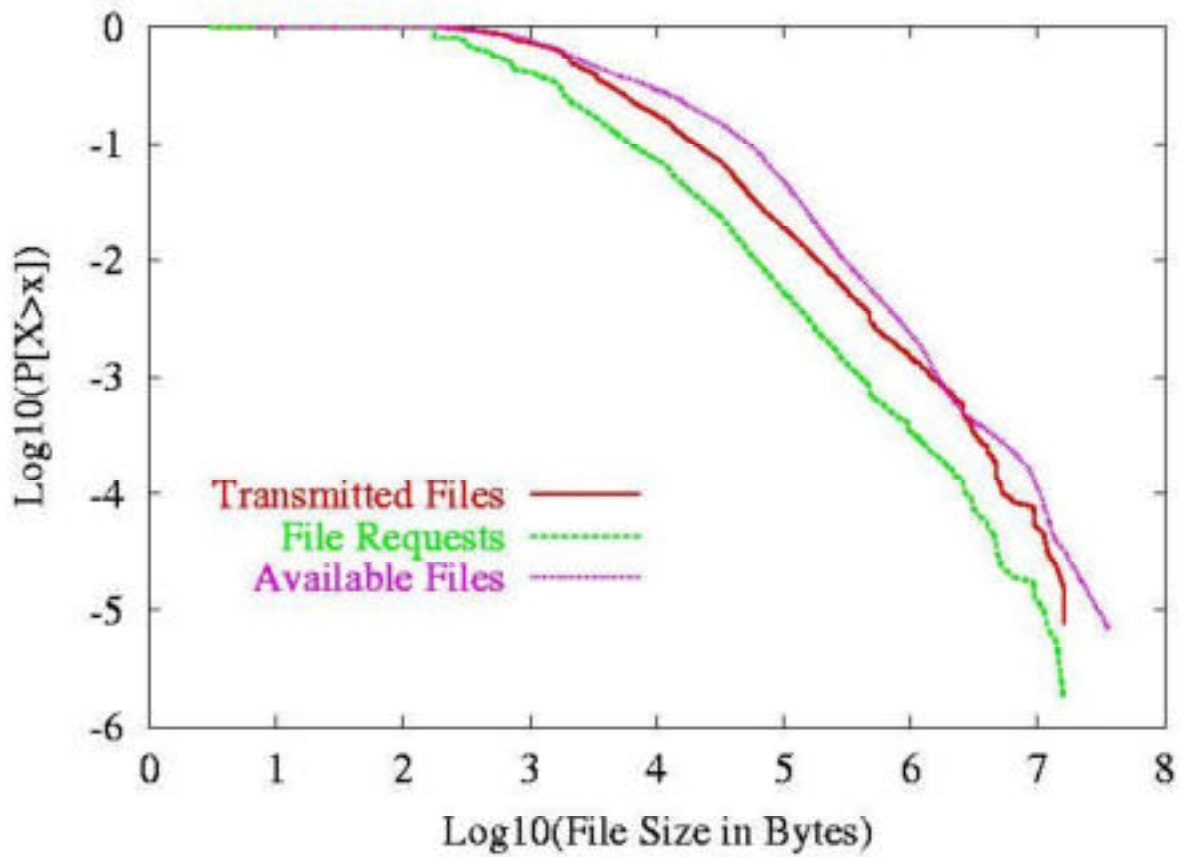
...as are Transmitted Files ...



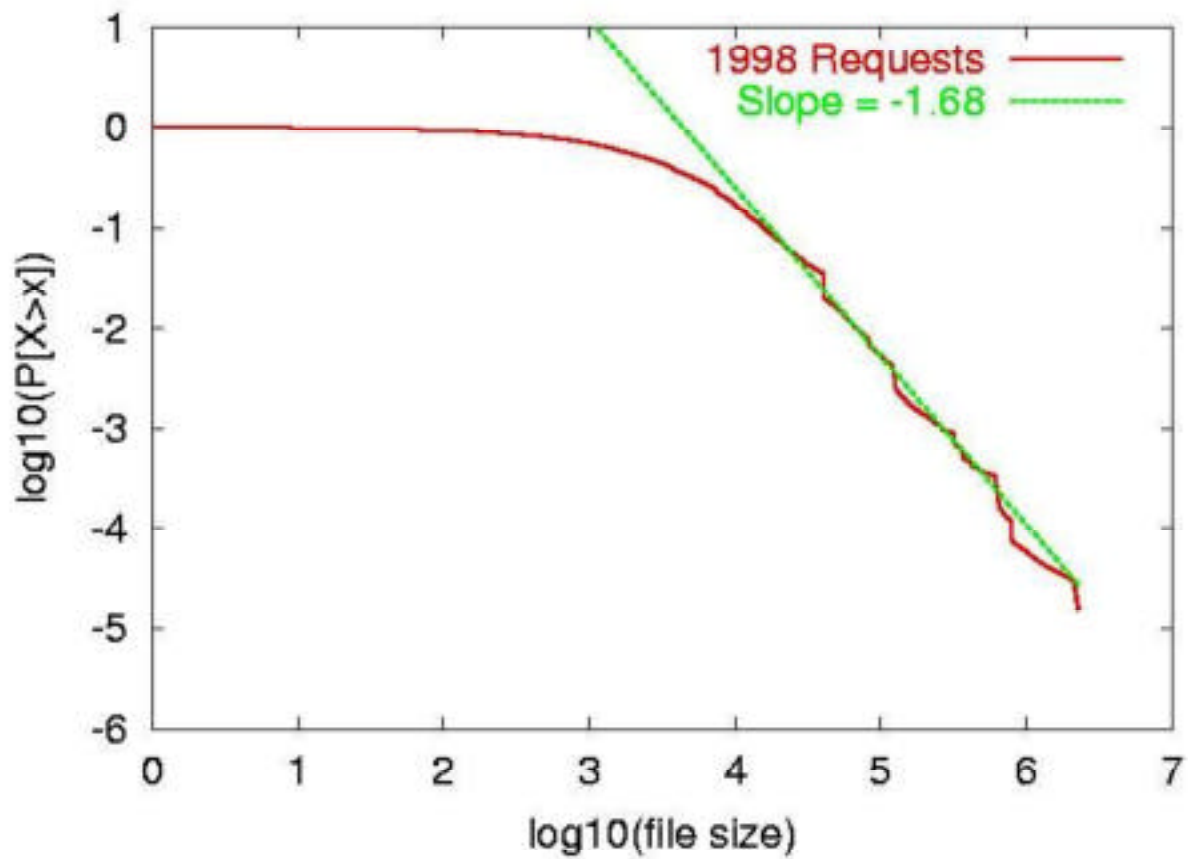
...as are Requested Files...



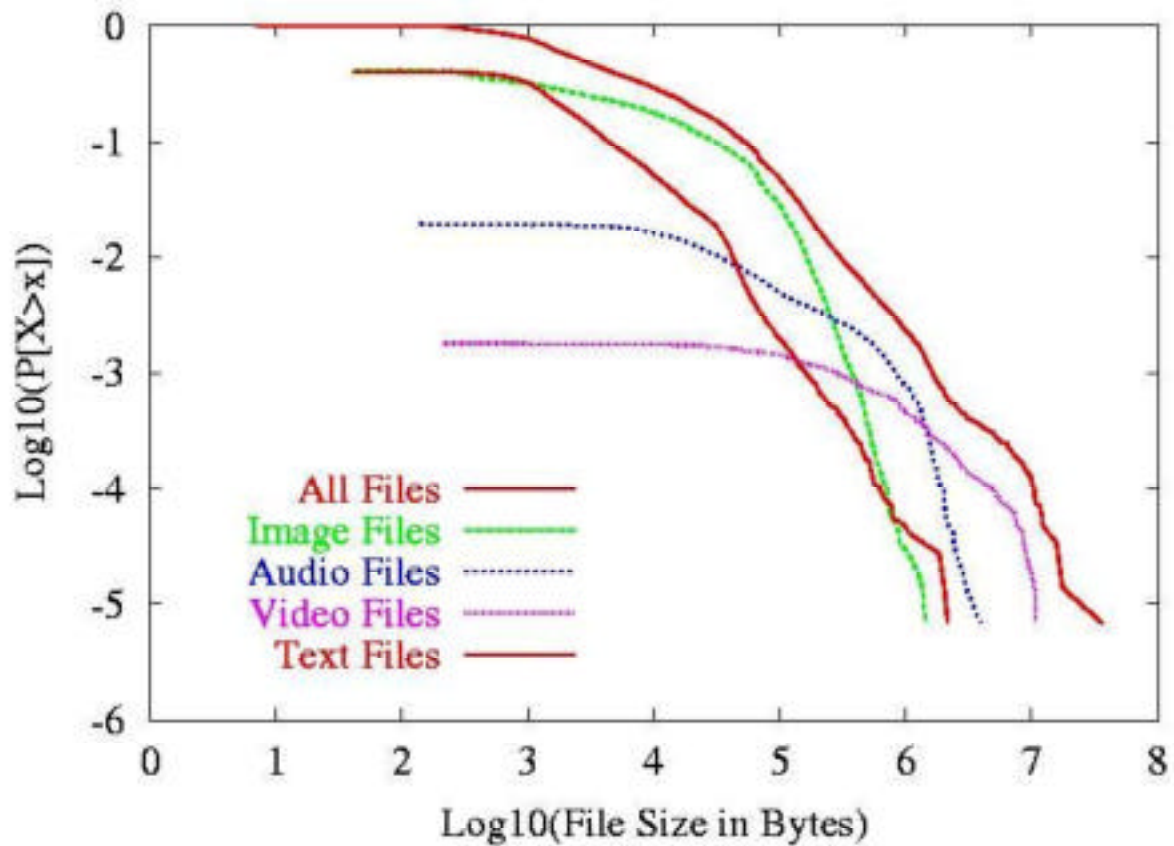
...as are Files on Servers



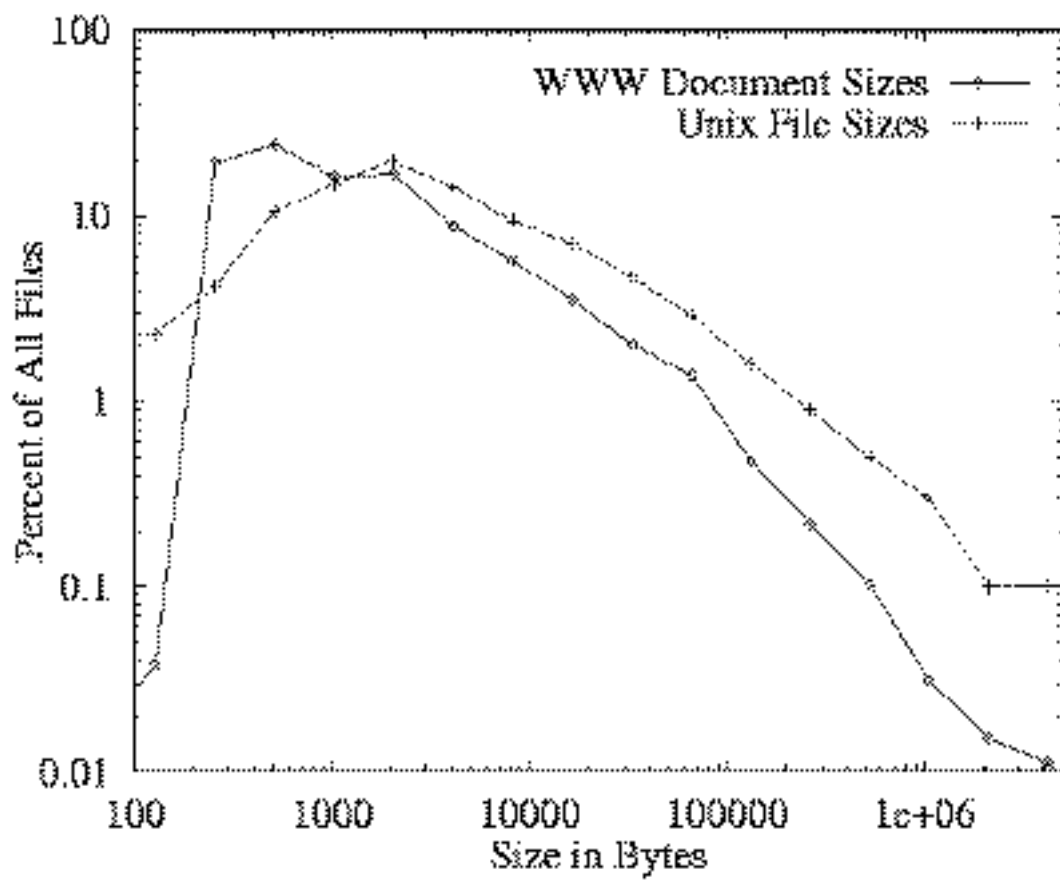
Same Lab, 3.5 Years Later



Influence of Multimedia on Tail Weight



Comparison to Unix Files



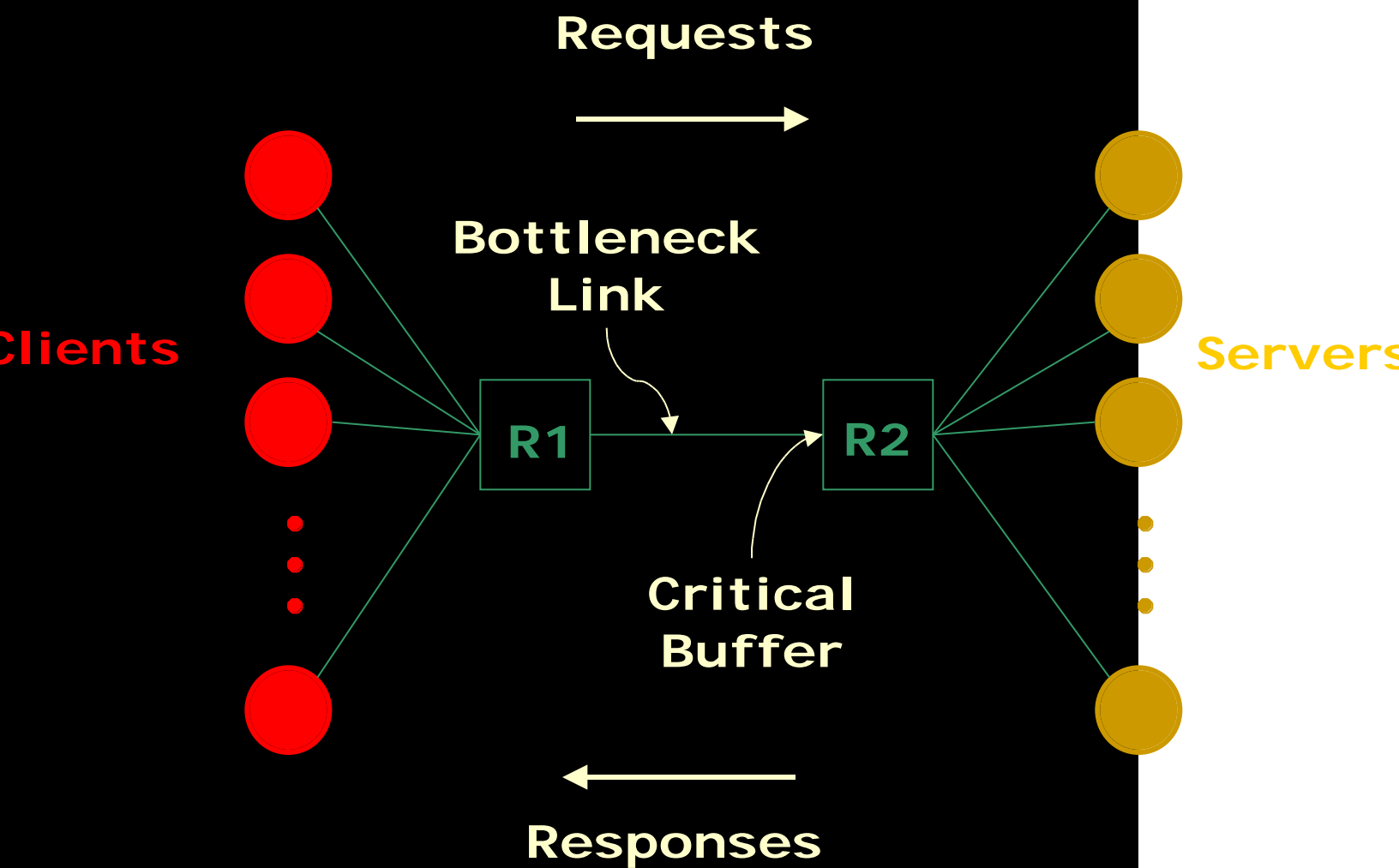
Closing the Loop: Are Heavy Tailed Files Sufficient?

Measurement results are suggestive, but don't directly demonstrate causality

However, we can now test the file size hypothesis directly in simulation

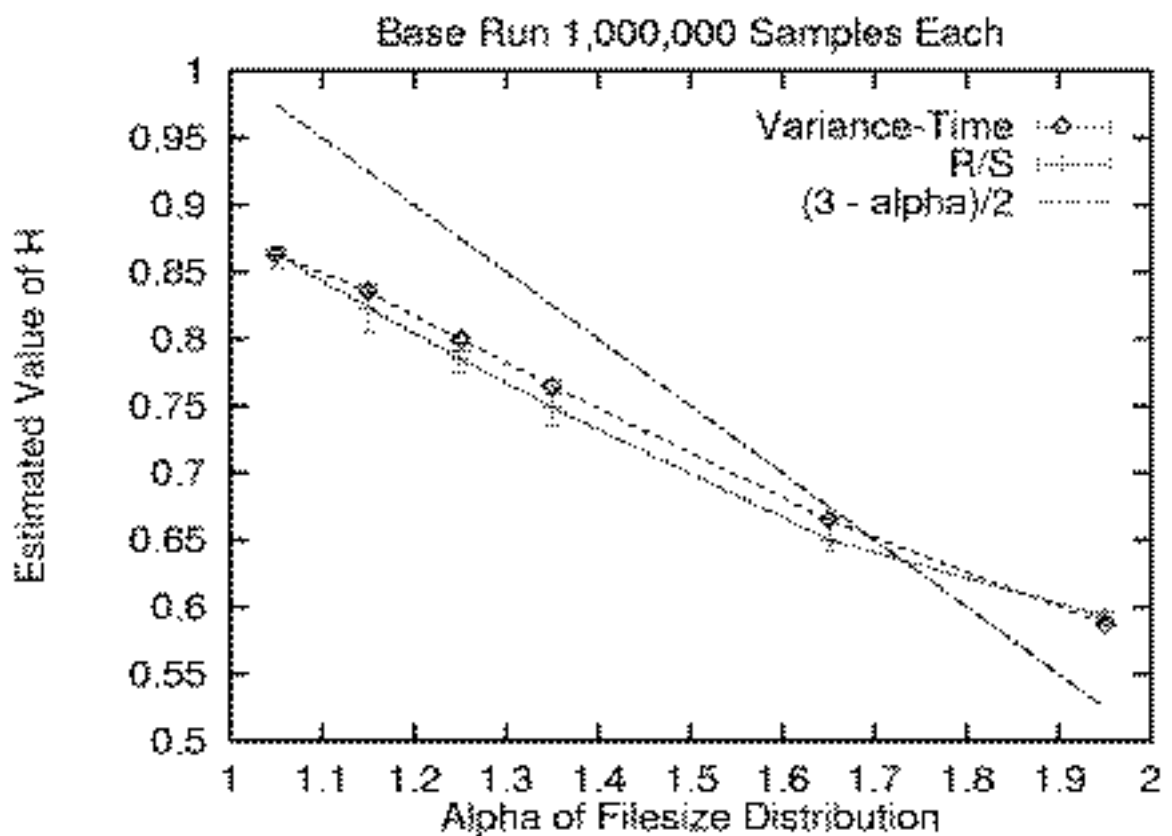
- Simulated network with multiple clients, servers
- Clients alternate between requests, idle times
- **Files drawn from heavy-tailed distribution**
 - Vary to vary self-similarity of traffic
- Each request is simulated at packet level,

Simulated Topology



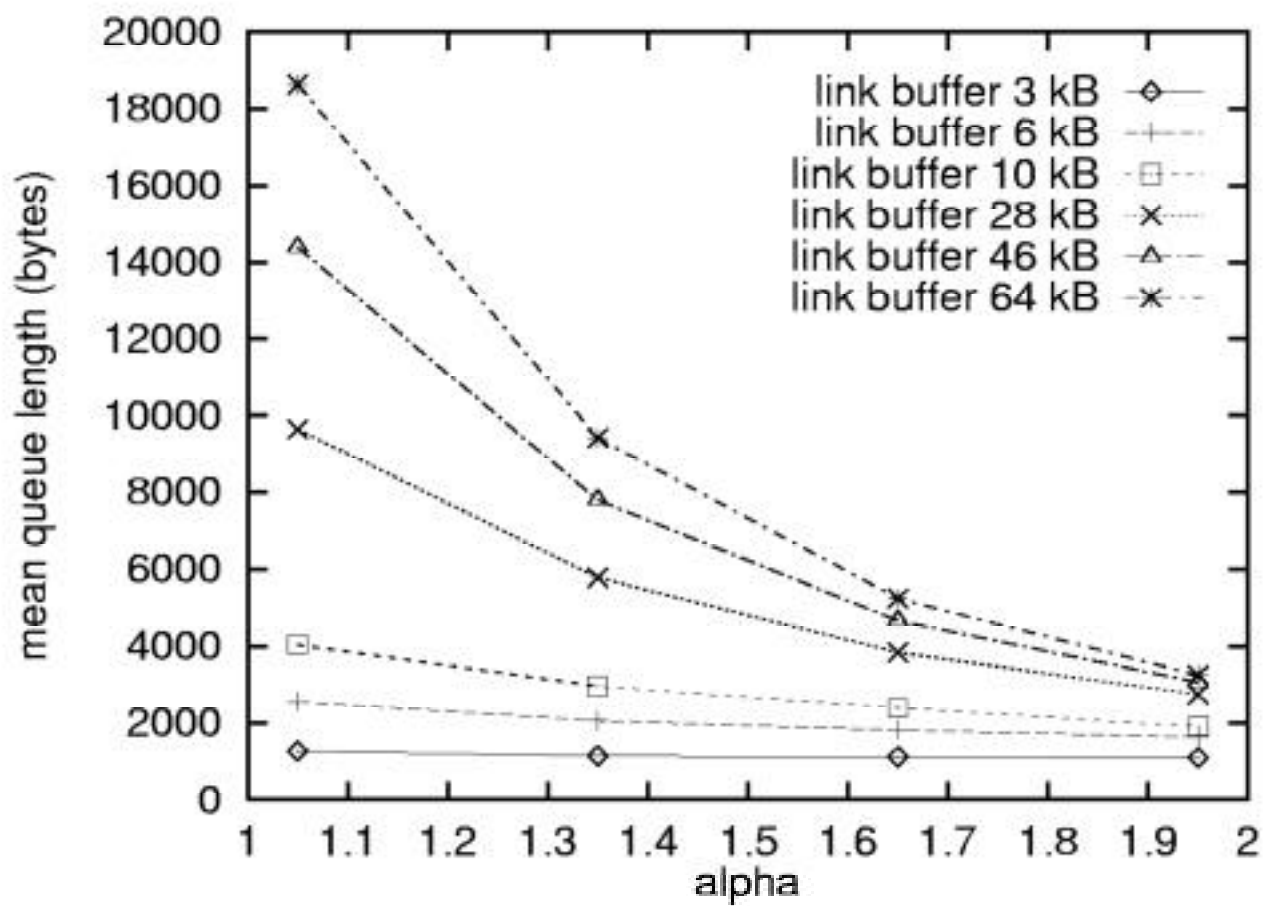
Resulting Traffic Characteristics

Self-similarity varies smoothly as
function of α



Severity of Packet Delay

Buffer utilization is strongly dependent on H



Conclusions

Self-similarity is present in Web traffic

- The Internet's most popular application
- For the Web, the causes of s.s. can be traced to the heavy-tailed distribution of Web file sizes
- Caching doesn't seem to affect things much
- Multimedia tends to increase tail weight of Web files
- But, even text files alone appear to be

Future Directions:

From to

Measurements show that traffic also shows interesting scaling properties on small scales

However, instantaneous scaling properties vary with time: traffic appears to be **multifractal**

Nonetheless, the simulation approach can shed light on causes and effects of multifractal behavior

Questions:

- Are small local scaling exponents associated with packet loss?