A Longitudinal View of HTTP Traffic

Tom Callahan[†], *Mark Allman*[‡], *Vern Paxson*[‡,¶]

[†]Case Western Reserve University,

[‡]International Computer Science Institute,

[¶]University of California, Berkeley

Passive and Active Measurement Conference – April 7-9, 2010

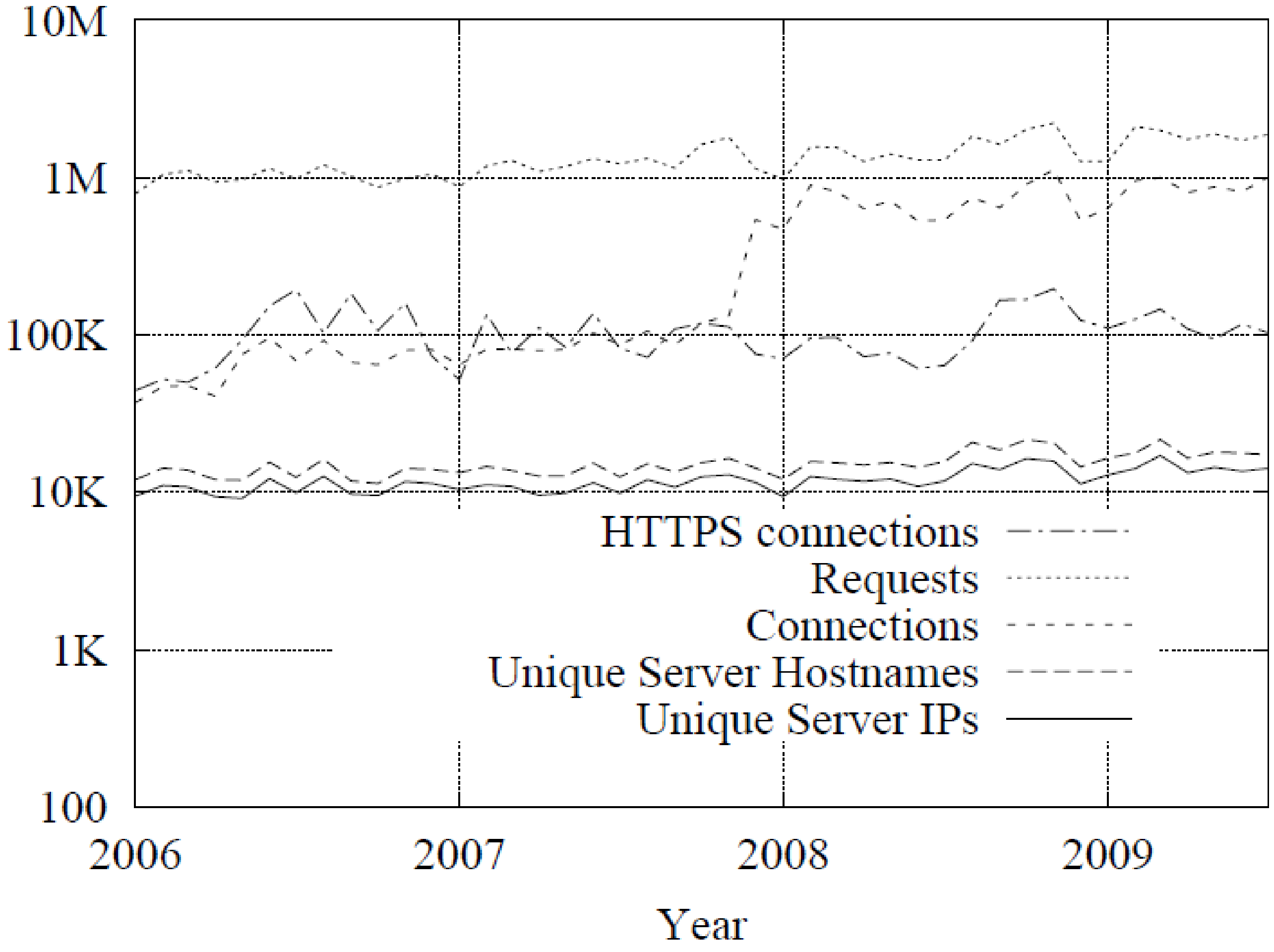
Objectives

- Follow a single user group and observe trends in HTTP usage over a long period of time
- Update view on HTTP usage patterns
- Study changes of client behavior, server behavior, and data behavior

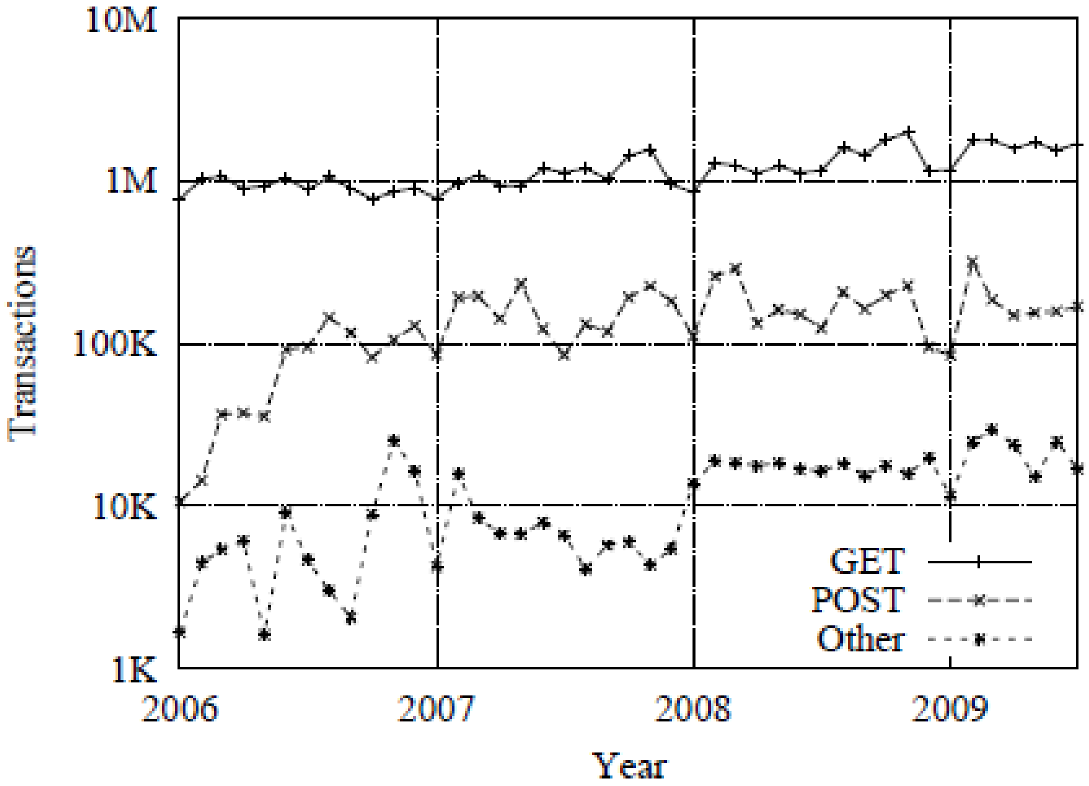
Dataset

- Used Bro IDS to reconstruct and log HTTP sessions from the real-time packet stream at the border connecting ICSI with its ISP
- Three and a half years of information about web connections and corresponding HTTP requests
  - Analyzed first week of every month from January 2006 to July 2009

Dataset (cont'd)

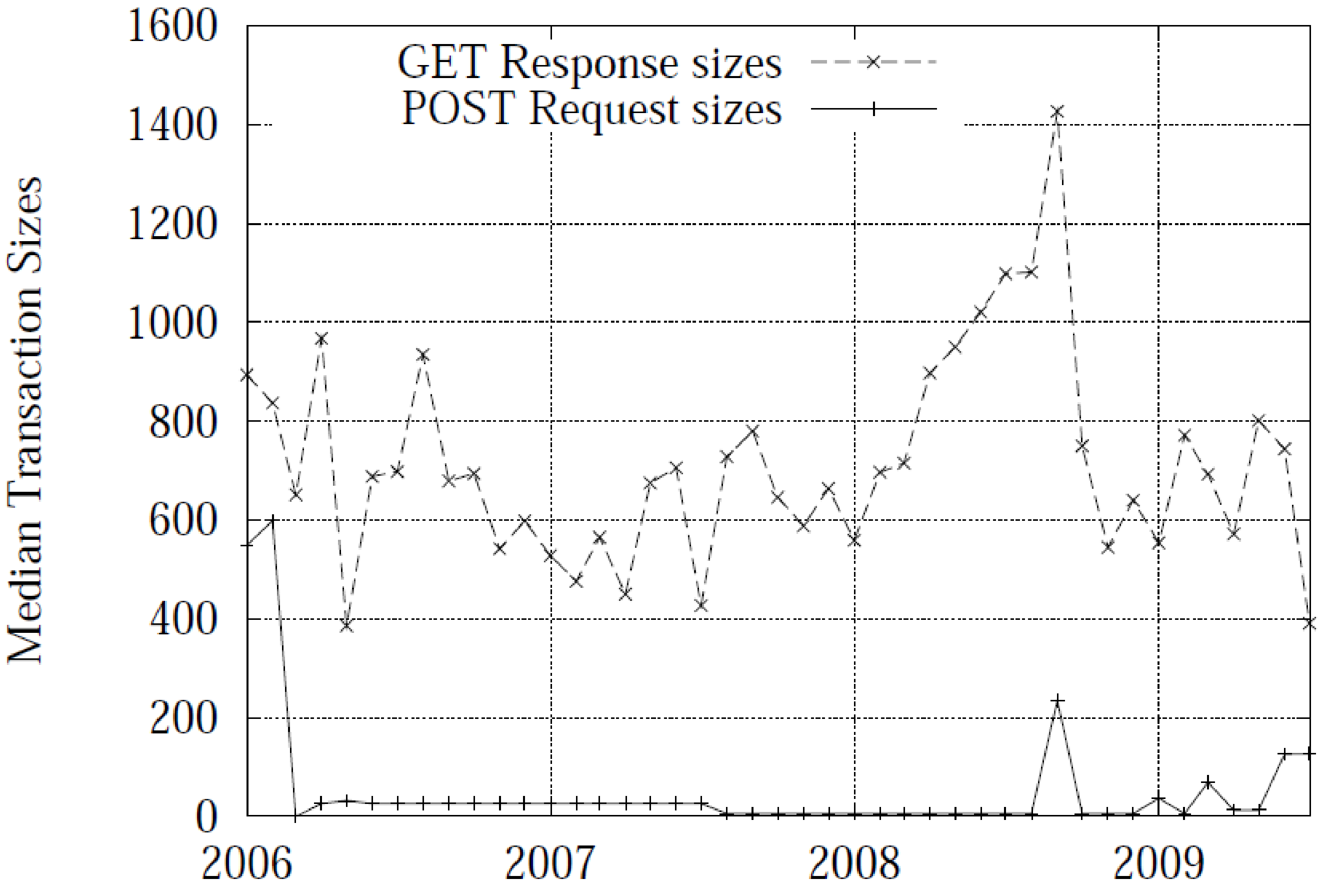


Transactions

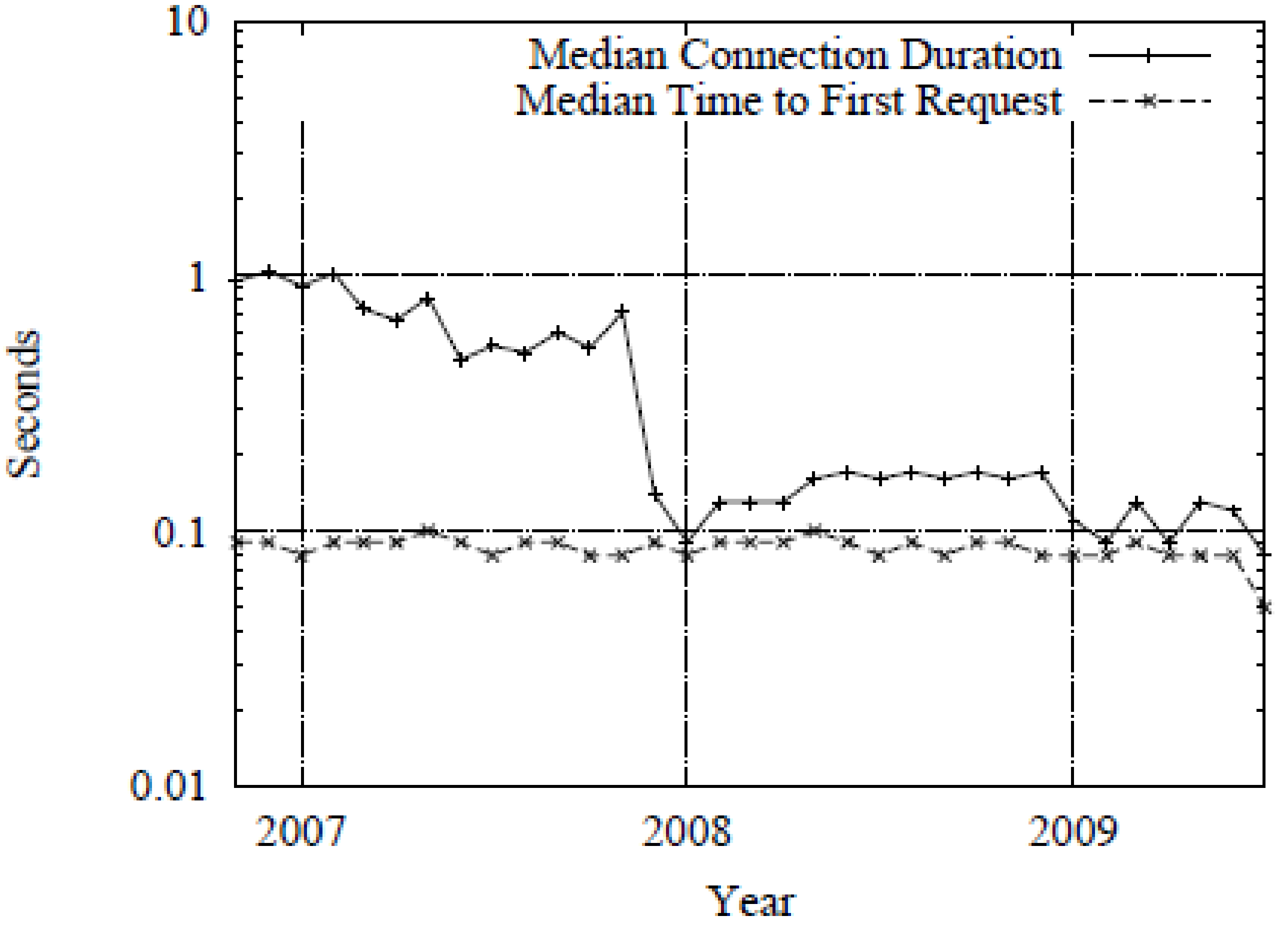


GET - ~90%
POST - <10%
Other - <1%

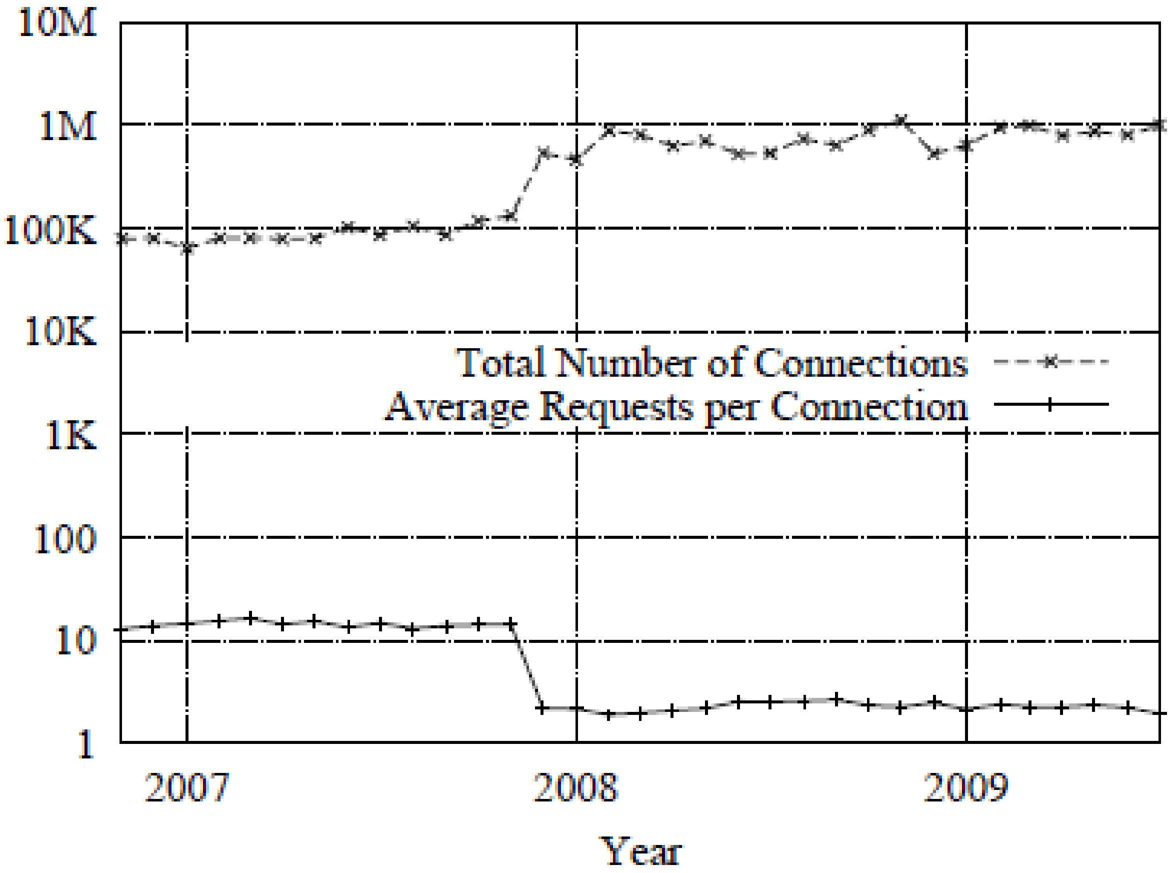
Transaction Sizes



Connection Timing



Connection Characteristics



Requests/Connections Artifact

- Confirmed that increase took place over nearly all popular servers
  - Must be due to a client-side change
- Seemingly benign network changes can have an immense effect on traffic patterns!
- Anyone running IDS systems, anomaly detectors, etc must be aware of these potential changes

Requests per Hostname

- Negligible year to year difference in distribution
- Median number of requests for a specific hostname less than 10 per year
- A few hosts accessed millions of times
- Only four stay in the "Top 10" throughout all years of analysis
  - Ad.doubleclick.net, graphics8.nytimes.com, www.google.com, mail.google.com

Requests per Object

- Again, negligible difference in distribution from year to year
- 90% of objects accessed only once
  - Distinct parameters = distinct object

IPs vs. Hostnames

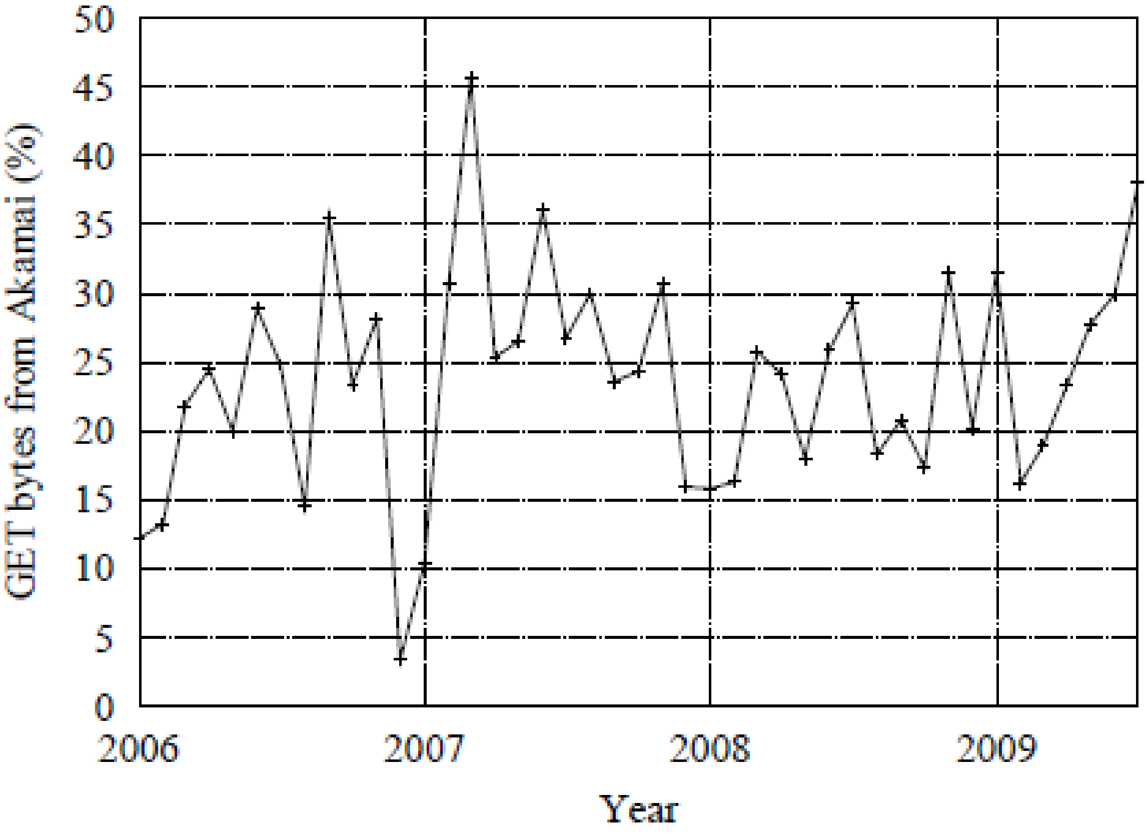
- Around 80% of IP addresses served data for a single hostname
- Around 10% of IPs served data for two
- A few IPs served traffic for hundreds of hostnames
- Alternatively, 80-90% of hostnames are served by one IP
- 5-10% by two, and ~5% by three or more

CDN Usage

- Attempt to establish amount of traffic from a major CDN – Akamai
- Checked historical DNS logs against a list of common Akamai suffixes, flagged all connections involving corresponding IPs as involving Akamai
- Flagged ~9200 unique IPs as Akamai
  - Definite undercount
    - See Sipat Triukose's SIGMETRICS 2009 poster

CDN Usage (cont'd)

- Lower bound for Akamai traffic



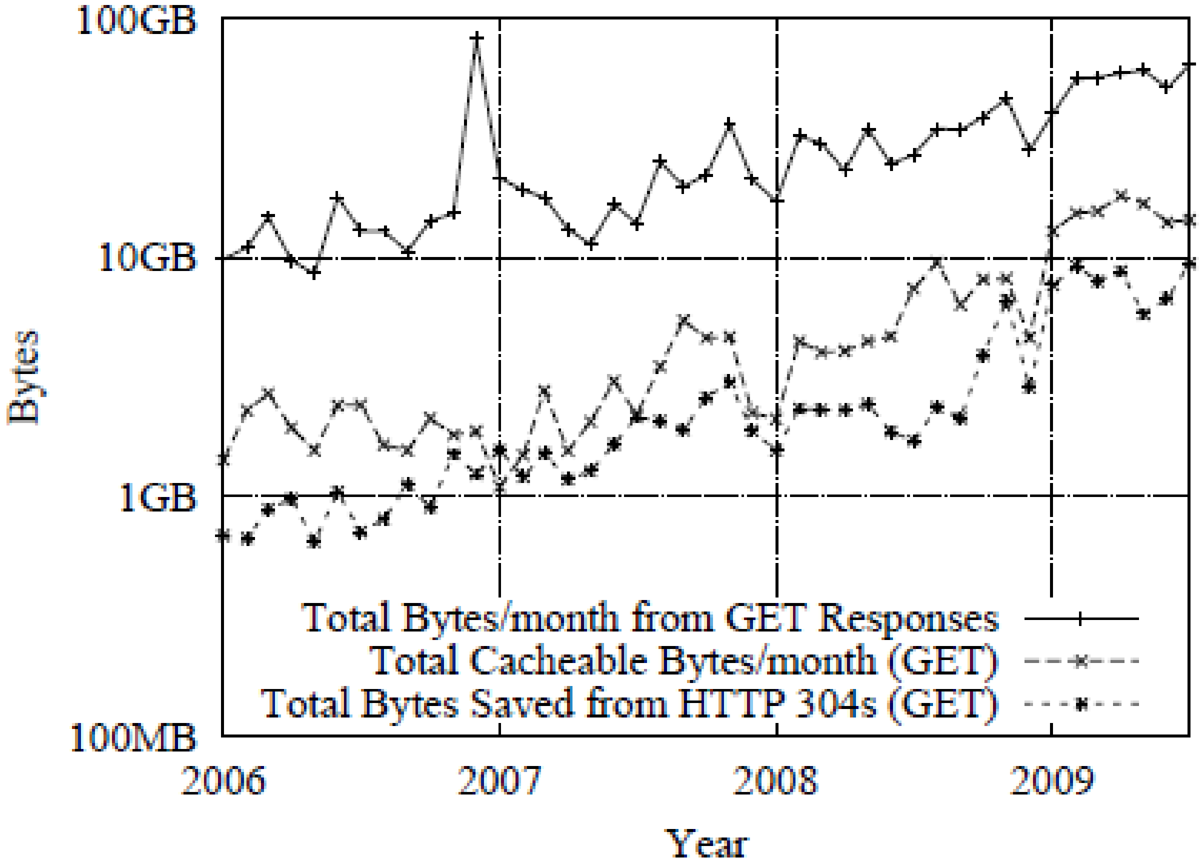
Current Usage

- Could not directly measure requests that were not made due to client caches
- Measured total bytes saved due to usage of HTTP 304 (Not Modified) messages
  - Total traffic would increase by approximately 10% without 304's

Caching Potential

- Simulated the presence of a border cache with unlimited capacity
- In any given month, 10-20% of traffic (in bytes) could be eliminated
  - Small user population could be reducing this number

Caching Potential (cont'd)



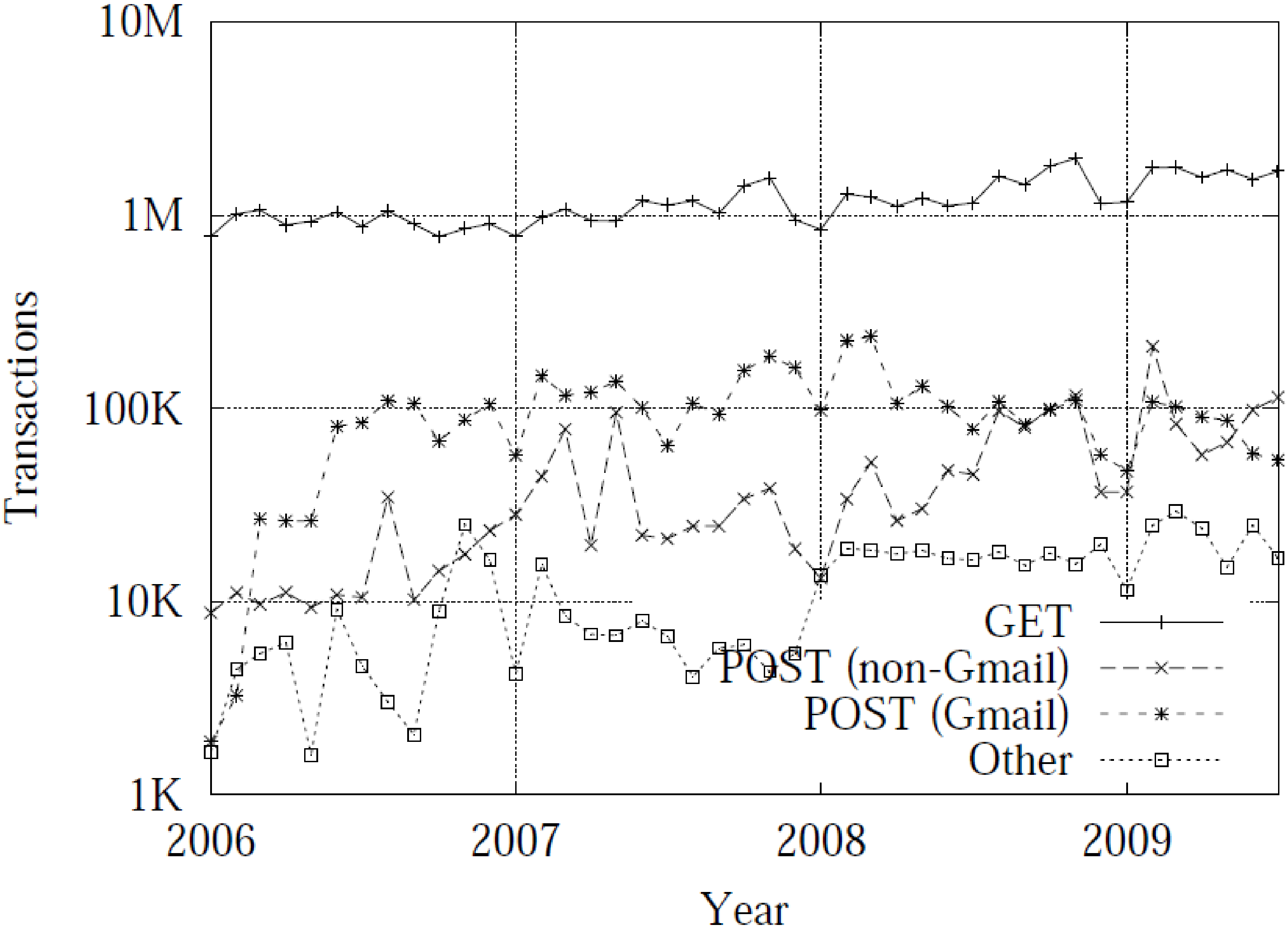
Summary

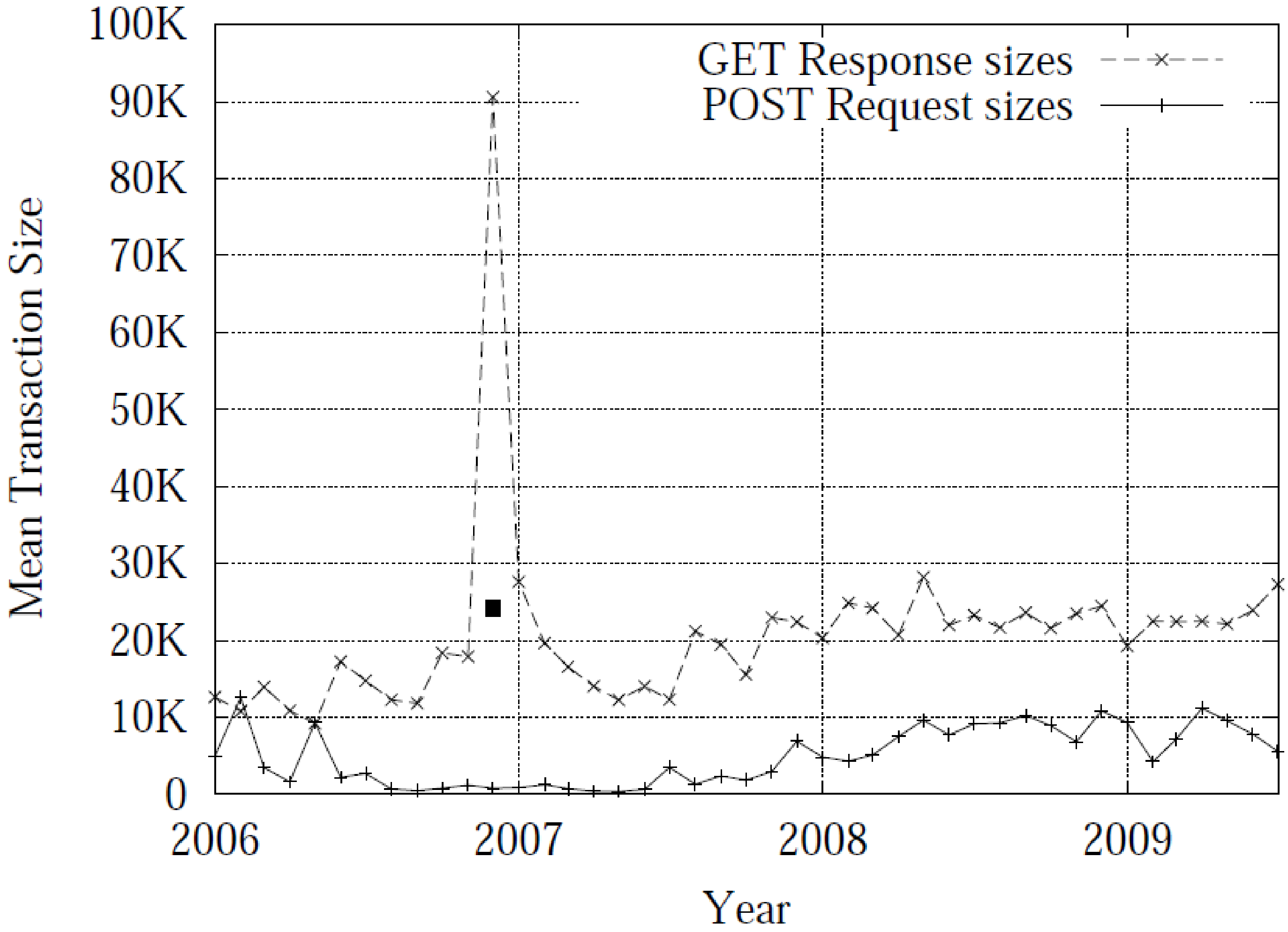
- Studied attributes of HTTP traffic as a function of time over three and a half years
- AJAX/Gmail has had a profound effect on HTTP characteristics
- Reexamined the impact and potential of caching at both the client and the ISP border

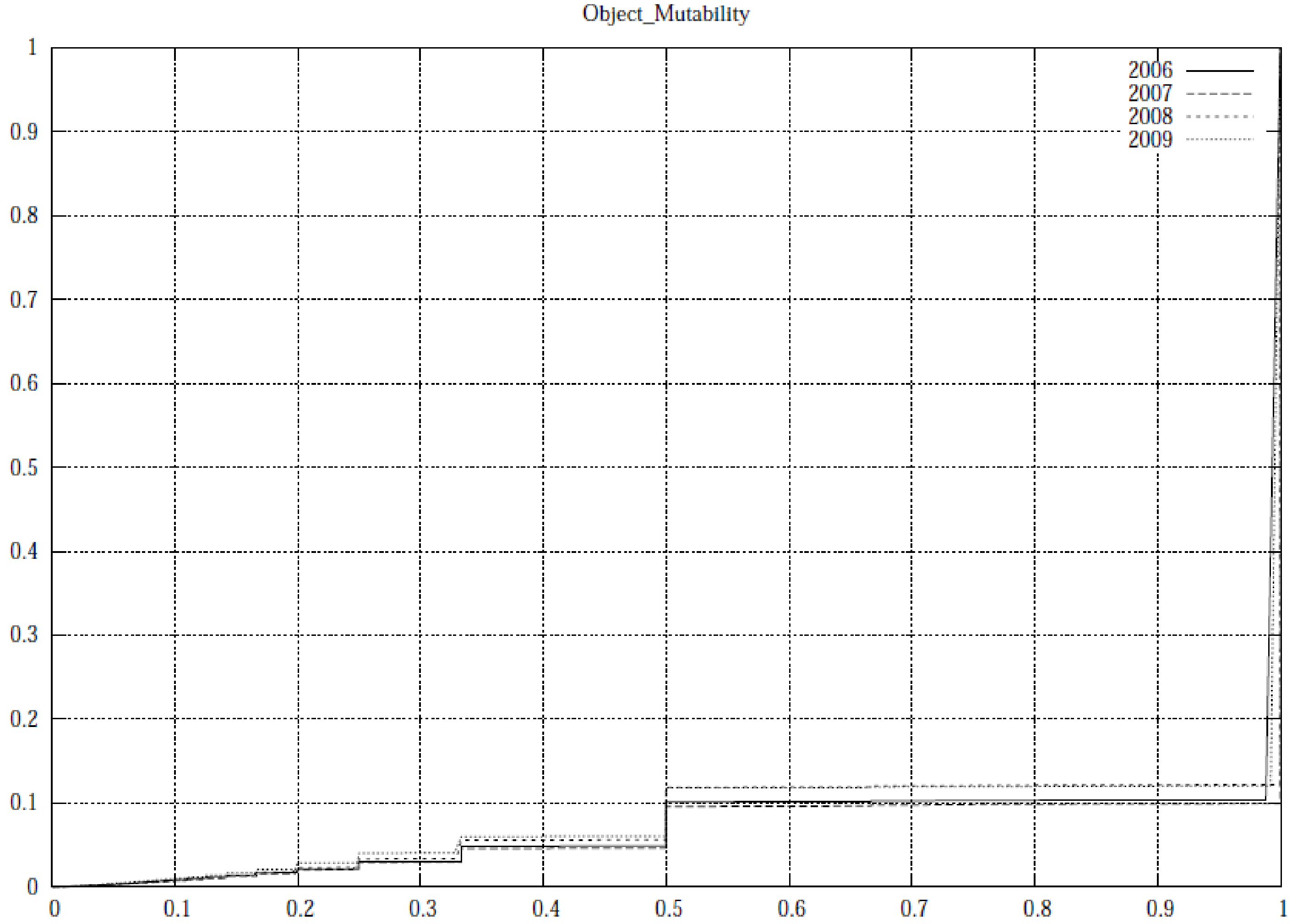
Thank you!

Questions?

Other


10M
1M
100K
10K
1K
Transactions
2006
2007
2008
2009
Year
GET
POST (non-Gmail)
POST (Gmail)
Other


GET Response sizes
POST Request sizes
Mean Transaction Size
0
10K
20K
30K
40K
50K
60K
70K
80K
90K
100K
2006
2007
2008
2009
Year


Object_Mutability
2006
2007
2008
2009
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1