# A Longitudinal Evaluation of HTTP Traffic

Tom Callahan
Case Western Reserve University

March 22, 2012

# Introduction

- Over time, HTTP has evolved from being used solely to disseminate content formatted in HTML to a crucial building block of other applications
- Increasingly complex web applications has resulted in a dynamic development and server environment
- Myriad web browsers and associated versions can effect change in HTTP usage from the client side

# Introduction

- Over time, HTTP has evolved from being used solely to disseminate content formatted in HTML to a crucial building block of other applications
- Increasingly complex web applications has resulted in a dynamic development and server environment
- Myriad web browsers and associated versions can effect change in HTTP usage from the client side
- **We must frequently re-appraise the state of HTTP traffic on the Internet in order to understand how it is changing**

# Introduction (cont'd)

- We examine HTTP traffic logs from the border routers of two edge networks
  - International Computer Science Institute in Berkeley, CA
    - Over 5 years of data
  - Case Connection Zone in Cleveland, OH
    - Eight months of data

# Outline

- Overview of Datasets

# Outline

- Overview of Datasets
- HTTP Transactions
  - Request Volume, size, etc.

# Outline

- Overview of Datasets
- HTTP Transactions
  - Request Volume, size, etc.
- TCP Connections (HTTP)
  - Duration, parallelism, etc.

# Outline

- Overview of Datasets
- HTTP Transactions
  - Request Volume, size, etc.
- TCP Connections (HTTP)
  - Duration, parallelism, etc.
- Client Behavior
  - Popular domains, content types, caching, etc.

# Outline

- Overview of Datasets
- HTTP Transactions
  - Request Volume, size, etc.
- TCP Connections (HTTP)
  - Duration, parallelism, etc.
- Client Behavior
  - Popular domains, content types, caching, etc.
- Web Structure
  - IP/Hostname relationships, Content Delivery Networks

# Outline

- Overview of Datasets
- HTTP Transactions
  - Request Volume, size, etc.
- TCP Connections (HTTP)
  - Duration, parallelism, etc.
- Client Behavior
  - Popular domains, content types, caching, etc.
- Web Structure
  - IP/Hostname relationships, Content Delivery Networks
- Summary

# International Computer Science Institute (ICSI)

- Small, non-profit research institute in Berkeley, CA
- Most users are full-time CS researchers, along with support staff
- Dataset runs from January 2006 to September 2011
  - We analyze the 11th through 18th days of each month
- Mean of 167 distinct client IPs per month with a standard deviation of 16

# ICSI Dataset Overview

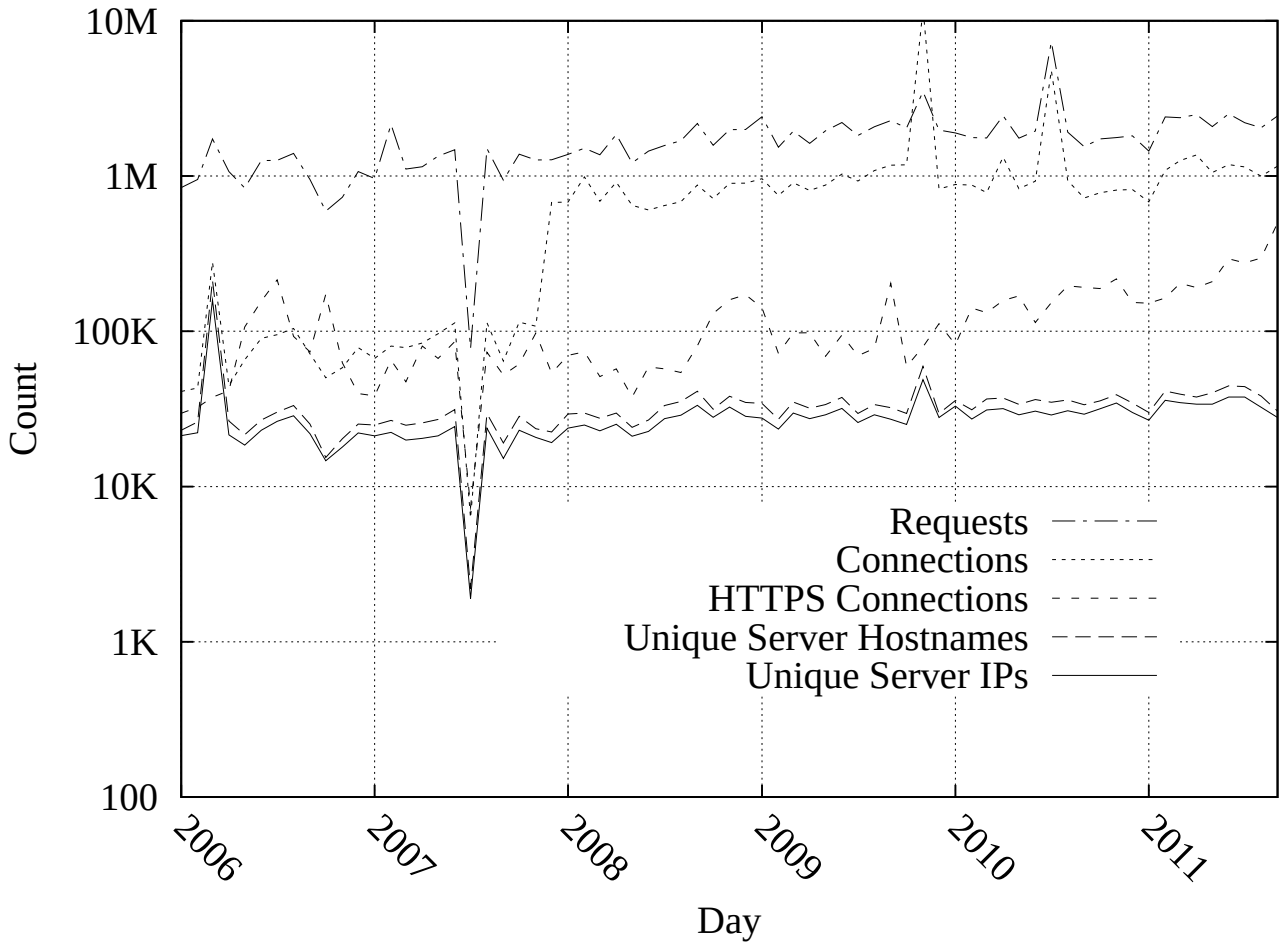


Figure: ICSI Dataset Summary

# Case Connection Zone (CCZ)

- Research project providing a neighborhood adjacent to Case Western with gigabit connectivity
- Approximately 100 houses, containing both students and non-students
- Dataset runs from February to September 2011
- Hardware provided to CCZ users known to employ Network Address Translation (NAT)
  - View of client IP addresses more likely to reflect number of housing units than number of users
  - Our HTTP Traffic logs provide no good way to approximate the real number of users
- Mean of 73 client IPs observed per month with a standard deviation of 8

# CCZ Dataset Overview

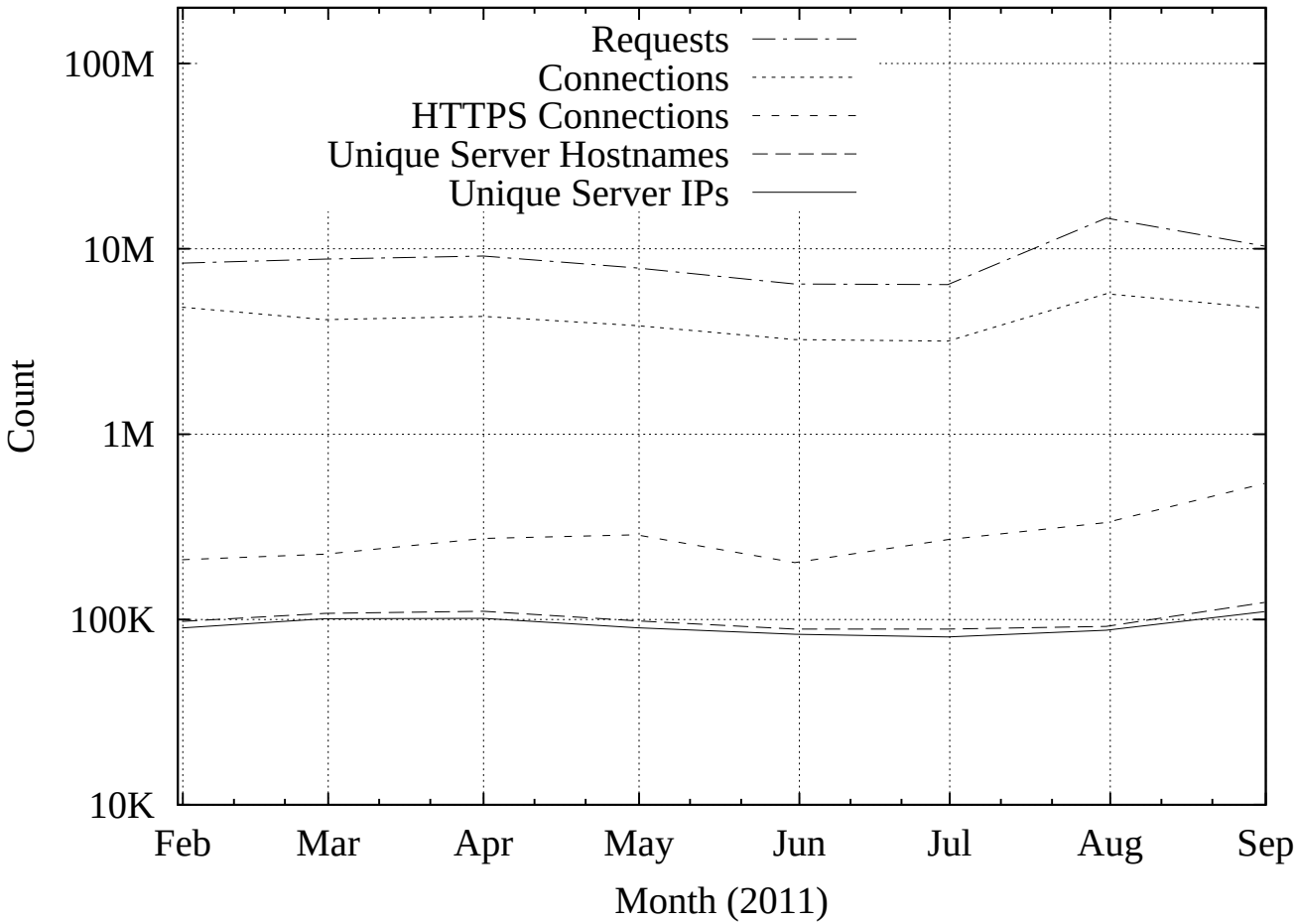


Figure: CCZ Dataset Summary

# HTTP Transactions

- Focus on individual HTTP transactions
  - Volume
  - Request Types
  - Sizes
- Susceptability of metrics to heavy hitters

# Transactions by Type

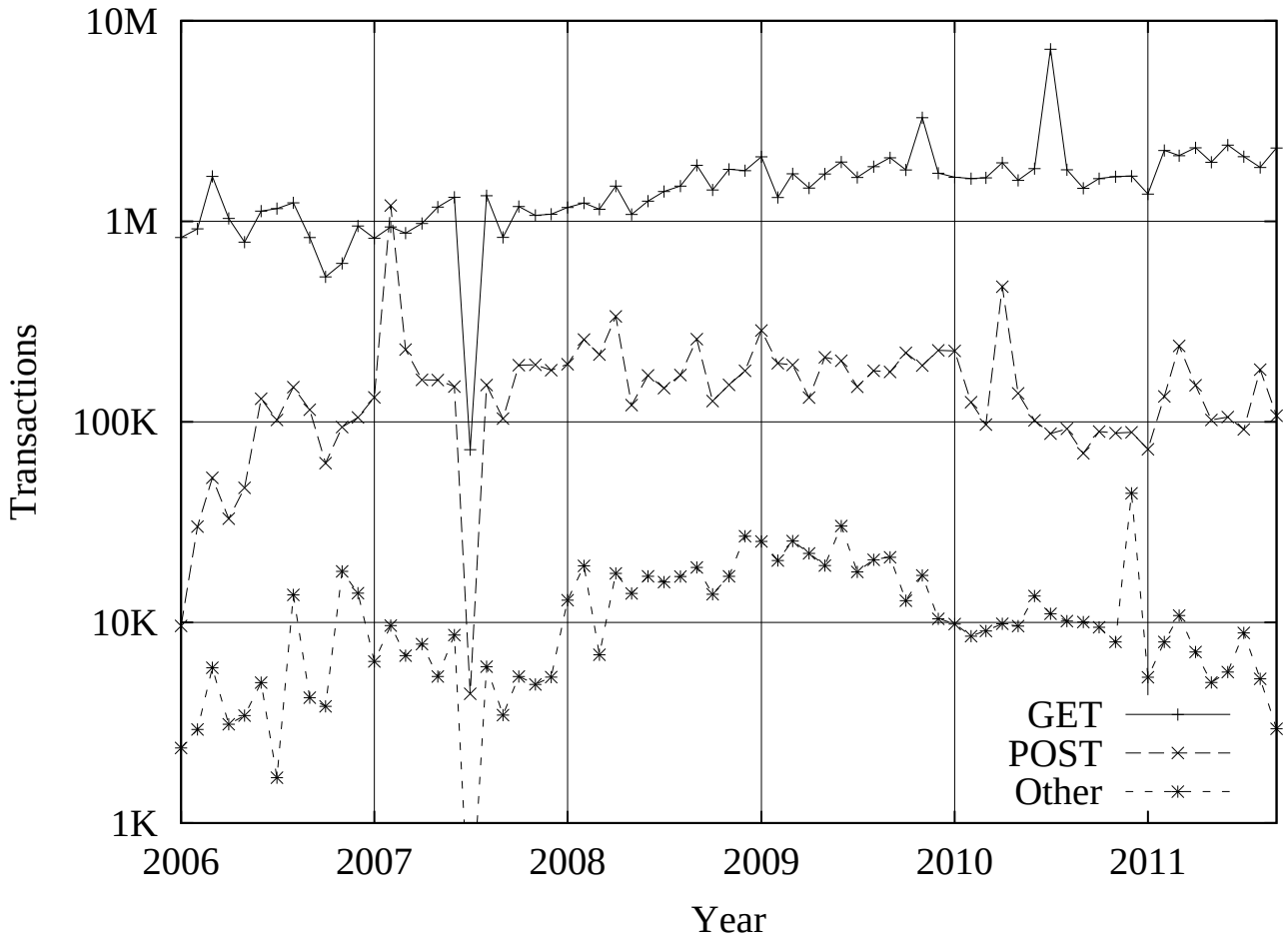


Figure: ICSI Transactions by Type

# GMail Influence

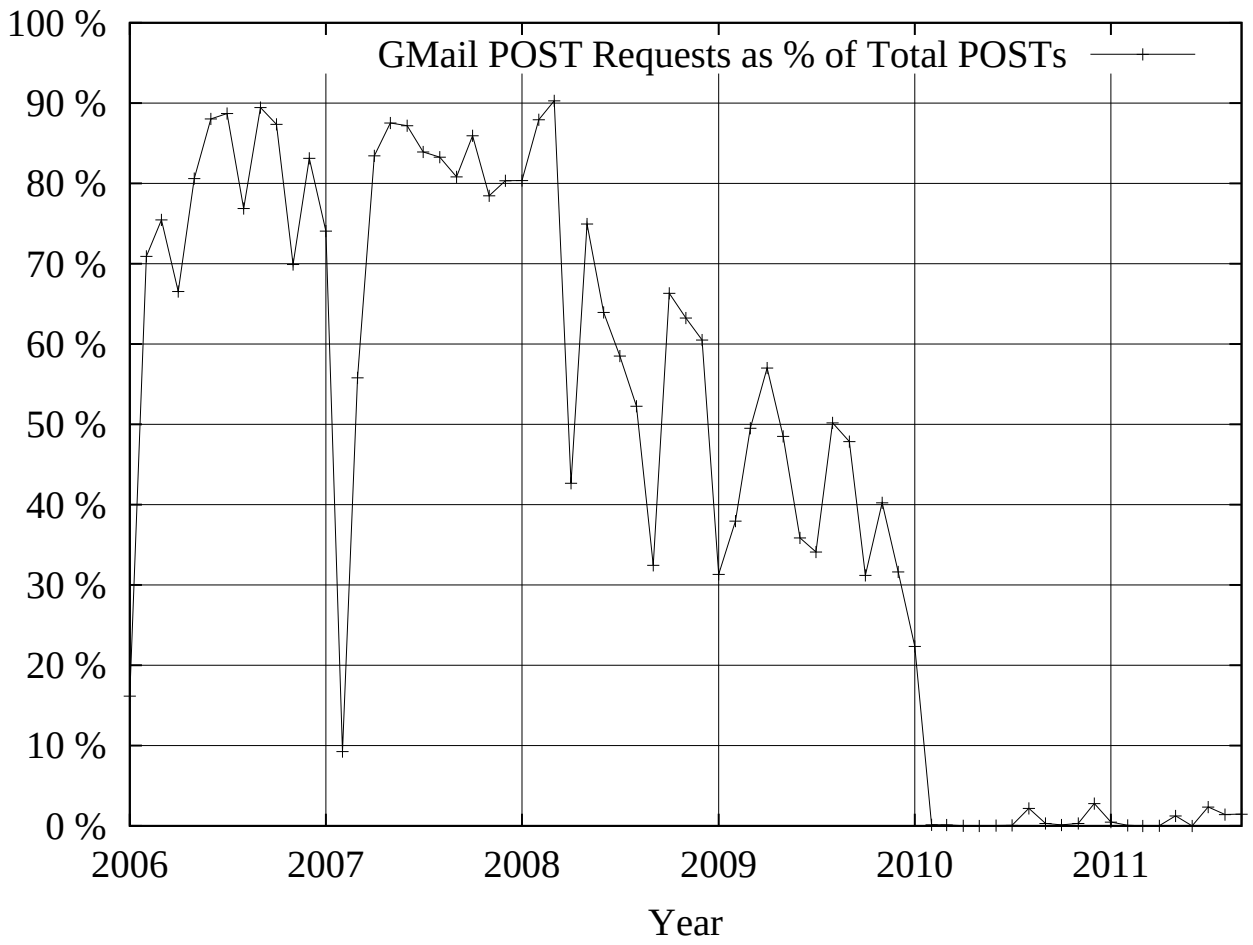


Figure: GMail POST Traffic

# Transaction Sizes

- Mean object size is increasing
  - ICSI mean object size has increased from 10K to 35K over the study
  - CCZ mean object size has remained above 50K throughout study
- Median GET response size at ICSI between 600 and 900 bytes throughout study
- Median GET response size at CCZ between 1600 and 2300 bytes throughout study
- CCZ saw a higher proportion of requests to Facebook, Youtube, and Netflix than ICSI

# Median Response Size by User Population

| | |
|---|---|
| Cunha (1994/1995) | 2,245 |
| Mah (1995) | 2,035 |
| Barford (1998) | 2,416 |
| UNC (1999) | 1,164 |
| UNC (2001) | 733 |
| UNC (2003) | 632 |
| ICSI (2006) | 895 |
| ICSI (2011) | 845 |
| CCZ (2011) | 1,977 |

Table: Median Response Sizes (bytes)

# HTTP Transaction Sizes (cont'd)

- From 2006 to 2011 at ICSI:
  - The largest 1% of requests have increased from being over 93KB to over 260KB
  - The largest 0.1% of requests have increased from being over 849KB to over 3.4MB
- Throughout all years at ICSI, between 15-25% of all requests have resulted in 0 content bytes
  - Typically HTTP 304 "Not Modified" or HTTP 302 "Redirect", among others

# Transaction Size Distribution

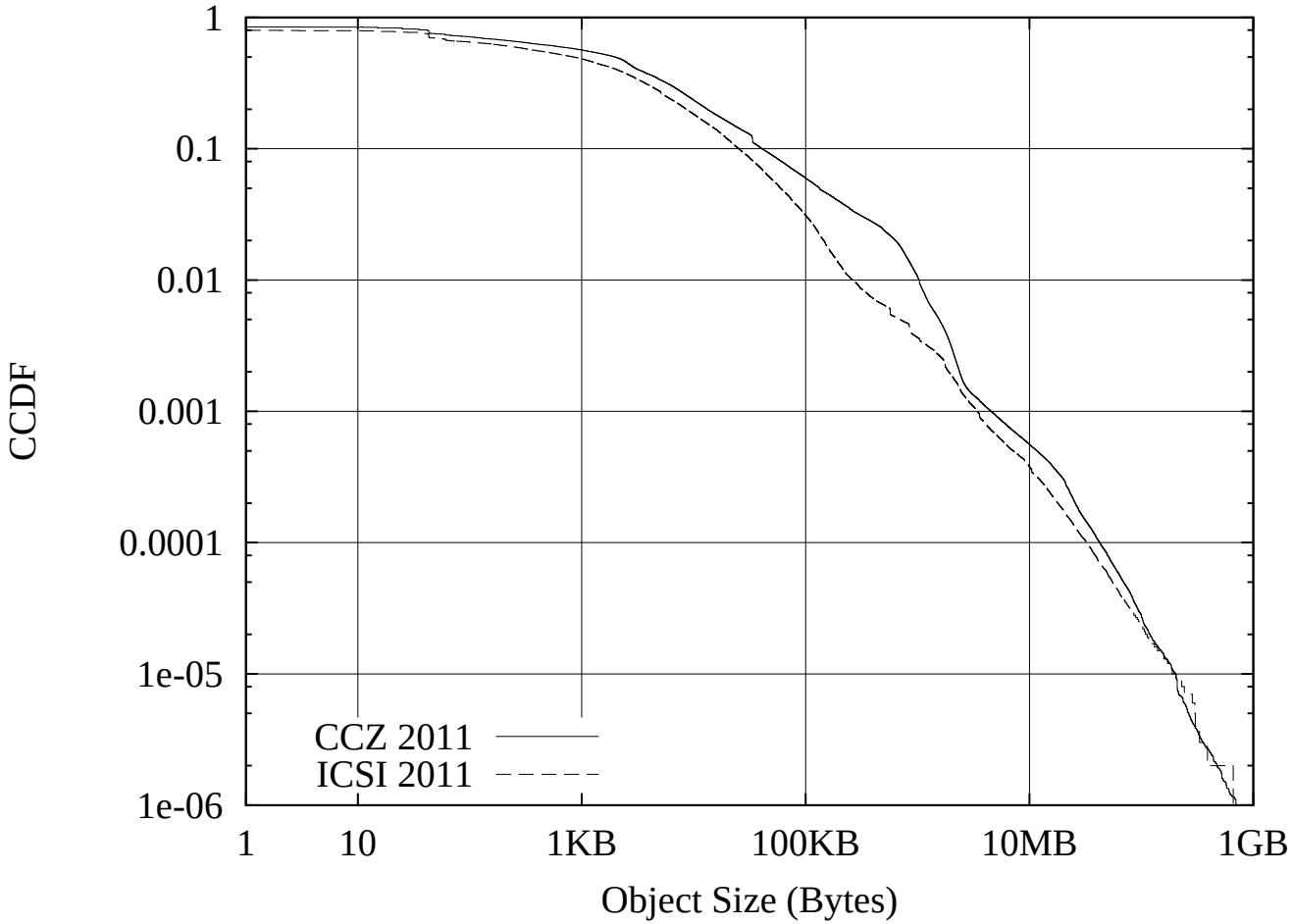


Figure: GET Response Distributions

# Transactions Summary

- While HTTP Traffic has grown in volume, the relative rankings of request types has stayed constant
- Some popular web applications (e.g. GMail) have the potential to strongly affect some metrics
- The shape of the distribution of HTTP transaction sizes has remained the same over time

# HTTP Connections

- Shift from individual HTTP transactions to the TCP connections containing them
- Examine how many requests are carried by these connections
- Connection Timing
- Measures of simultaneous connections

# Connection Timing

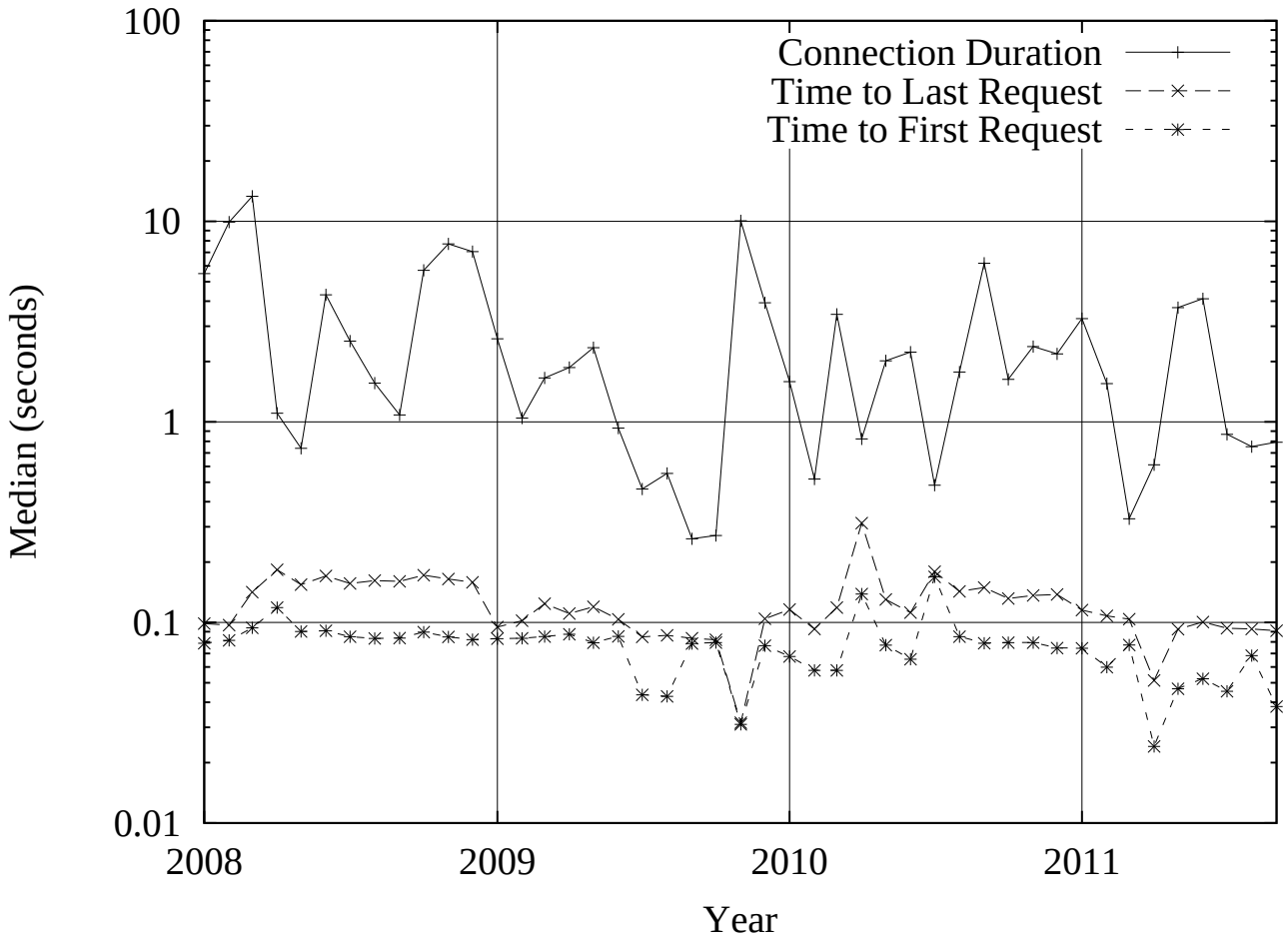


Figure: ICSI Connection Timing

# Connection : Request Ratio

- Avg. # of requests per connection has stayed constant over time for both datasets
  - Mean requests per connection always between 1.5 and 2.5
- 61% of all connections in ICSI dataset serve a single request
  - Median object size is less than the the payload of two typical TCP data packets
  - TCP handshake duration may be significant
- 21% of all connections in ICSI dataset carry no requests at all

# Parallelism

- Client IP:Server IP level of parallism consistently higher than ClientIP:Server Hostname parallelism
  - Possibly in part due to website operators "gaming" browser concurrency rules by using multiple hostnames
- Growth in concurrency behooves server operators to examine per-client connection limits
- In April 2008, Firefox increased default # of parallel connections to a single server from 2 to 6
- In March 2009, Internet Explorer made the same adjustment

# Parallelism by server

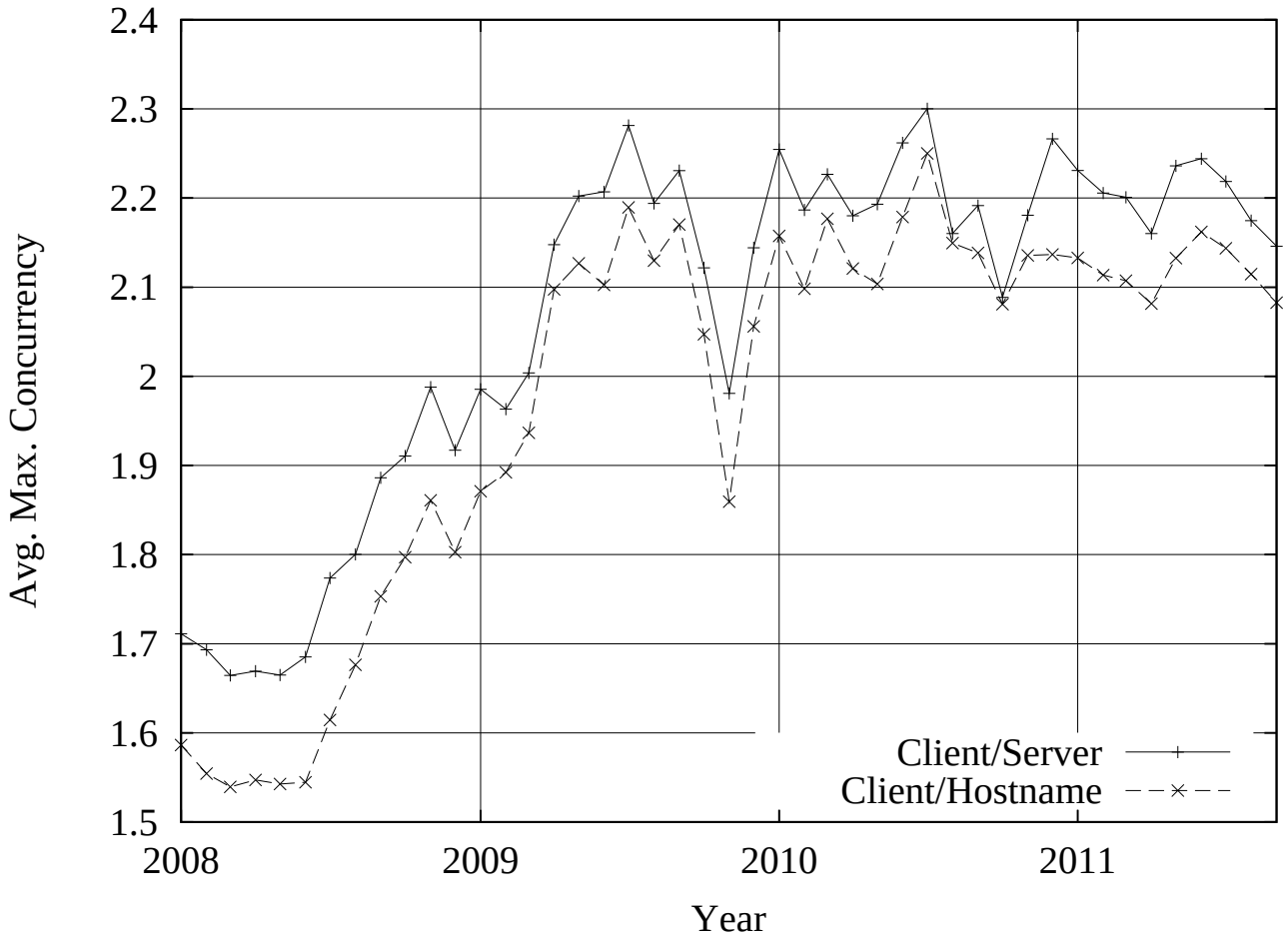


Figure: Pairwise Avg. Max. Concurrency

# Client Behavior

- Now let's move on to client-driven facets of HTTP

# Popular TLDs

- .com dominates with over 70% of requests at ICSI and over 60% of requests at CCZ
- .net consistently about 10% of requests at ICSI and 20% of requests at CCZ
- .org, .edu consistently within Top-10 at both sites
- Remaining popular domains include .gov and ccTLDs
  - Domains such as .mobi, .info have not yet reached widespread popularity in our user populations

# Popular File Extensions

- We use file extensions as a proxy for content type
  - Forced by our lack of complete header or payload data
- We exclude from consideration URLs with no extension (∼36% of all requests at ICSI)

# Popular File Extensions

- We use file extensions as a proxy for content type
  - Forced by our lack of complete header or payload data
- We exclude from consideration URLs with no extension ($\sim$36% of all requests at ICSI)
- Embedded content dominates request volume
  - .gif, .jpg, .png, .js, .css all popular throughout study

# Popular File Extensions

- We use file extensions as a proxy for content type
  - Forced by our lack of complete header or payload data
- We exclude from consideration URLs with no extension ($\sim$36% of all requests at ICSI)
- Embedded content dominates request volume
  - .gif, .jpg, .png, .js, .css all popular throughout study
- Noticeable decline in .gif usage (28% $\rightarrow$ 13%) over time at ICSI

# Popular File Extensions

- We use file extensions as a proxy for content type
  - Forced by our lack of complete header or payload data
- We exclude from consideration URLs with no extension ($\sim$36% of all requests at ICSI)
- Embedded content dominates request volume
  - .gif, .jpg, .png, .js, .css all popular throughout study
- Noticeable decline in .gif usage (28% $\rightarrow$ 13%) over time at ICSI
- Increase in .png use (1.6% $\rightarrow$ 6.9%) at ICSI

# Popular File Extensions

- We use file extensions as a proxy for content type
  - Forced by our lack of complete header or payload data
- We exclude from consideration URLs with no extension ($\sim$36% of all requests at ICSI)
- Embedded content dominates request volume
  - .gif, .jpg, .png, .js, .css all popular throughout study
- Noticeable decline in .gif usage (28% $\rightarrow$ 13%) over time at ICSI
- Increase in .png use (1.6% $\rightarrow$ 6.9%) at ICSI
- HTML usage has dropped in favor of increasing PHP usage

# Object Requests

- We consider the entire URL, with parameters, as an object
  - facebook.com/profile.php?id=XXXXX is not facebook.com/profile.php?id=YYYYY
- In every year and both datasets, ∼84-87% of objects were requested only once
- At ICSI in 2011:
  - The most popular 1% of objects accounted for 41% of GET requests and 20% of GET bytes
  - The most popular 0.1% of objects accounted for 11% of all GET bytes
  - Storing thse most popular 0.1% of objects would require 14GB of storage

# Object Request Distribution

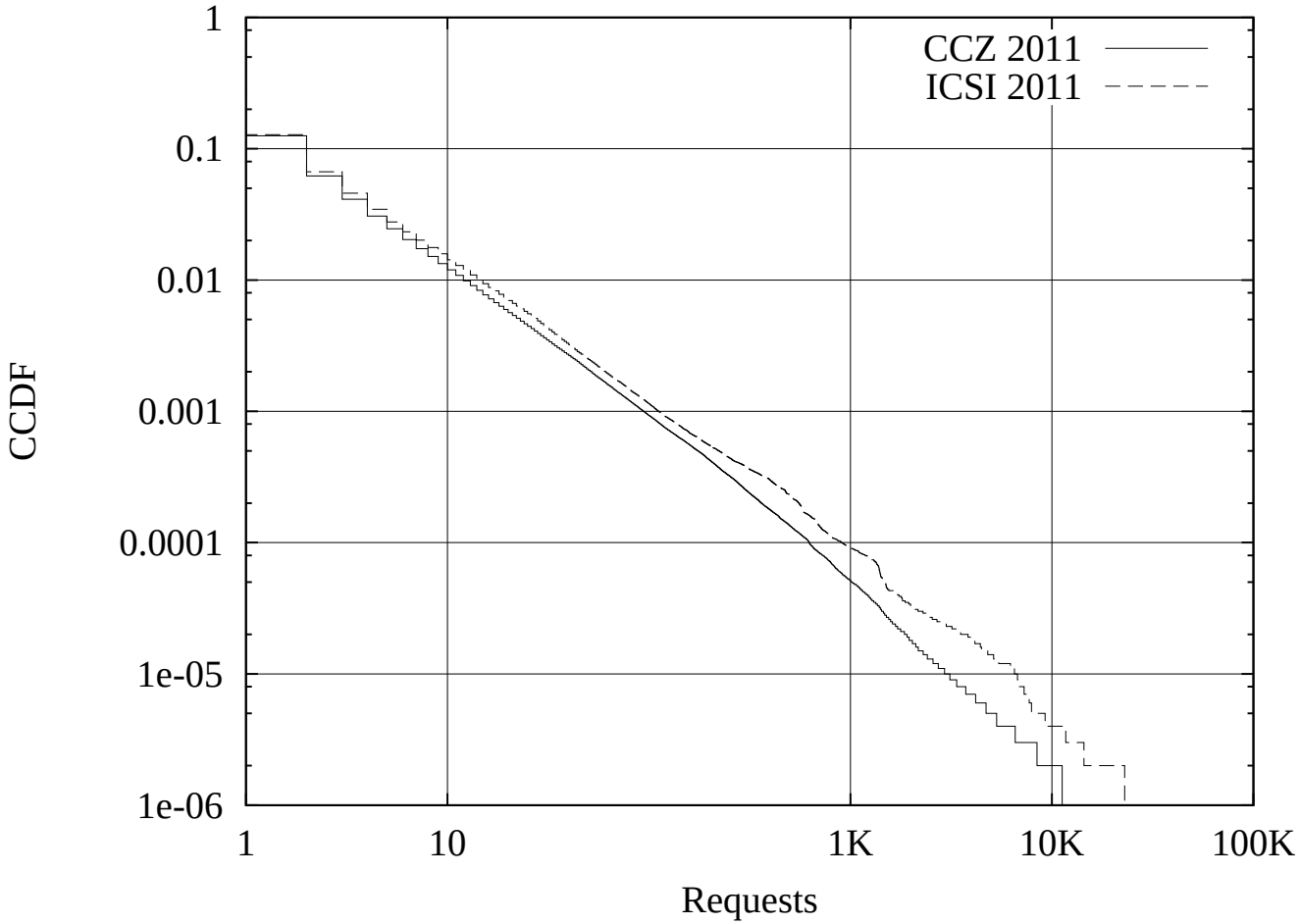


Figure: Requests per Object

# Hostname Requests

- Less one-time referencing than at the object level
  - From 2007 on at ICSI and at CCZ, we see between 19-24% of hostnames are visited only once
- The most popular 1% of hostnames account receive nearly 74% of total requests
- In 2001 at ICSI, users accessed over 8.7M unique objects, but only 87K hostnames

# Hostname Request Distribution

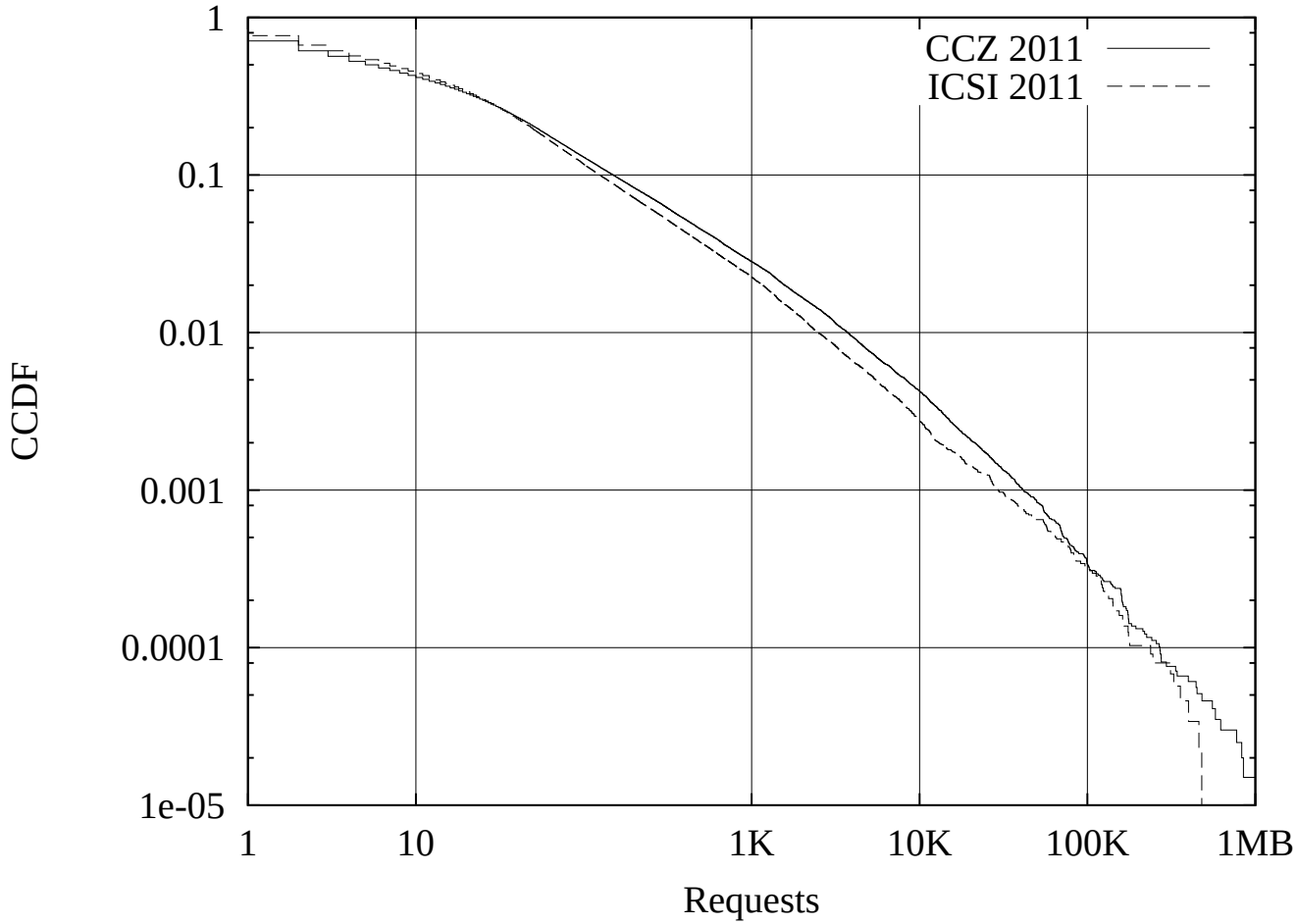


Figure: Requests per Hostname

# Caching

- We examine the usefulness of HTTP "304 Not Modified" messages
  - When we see a 304 response, we record a "savings" of bytes equivalent to last recorded size of that object in the same day. If the object was not previously requested that day, we record no savings
  - Therefore, we likely undercount 304 response savings
- We count all the bytes of an object transfer as "cacheable" if it was transferred previously in the same month and its size has not changed

# Caching Use and Potential

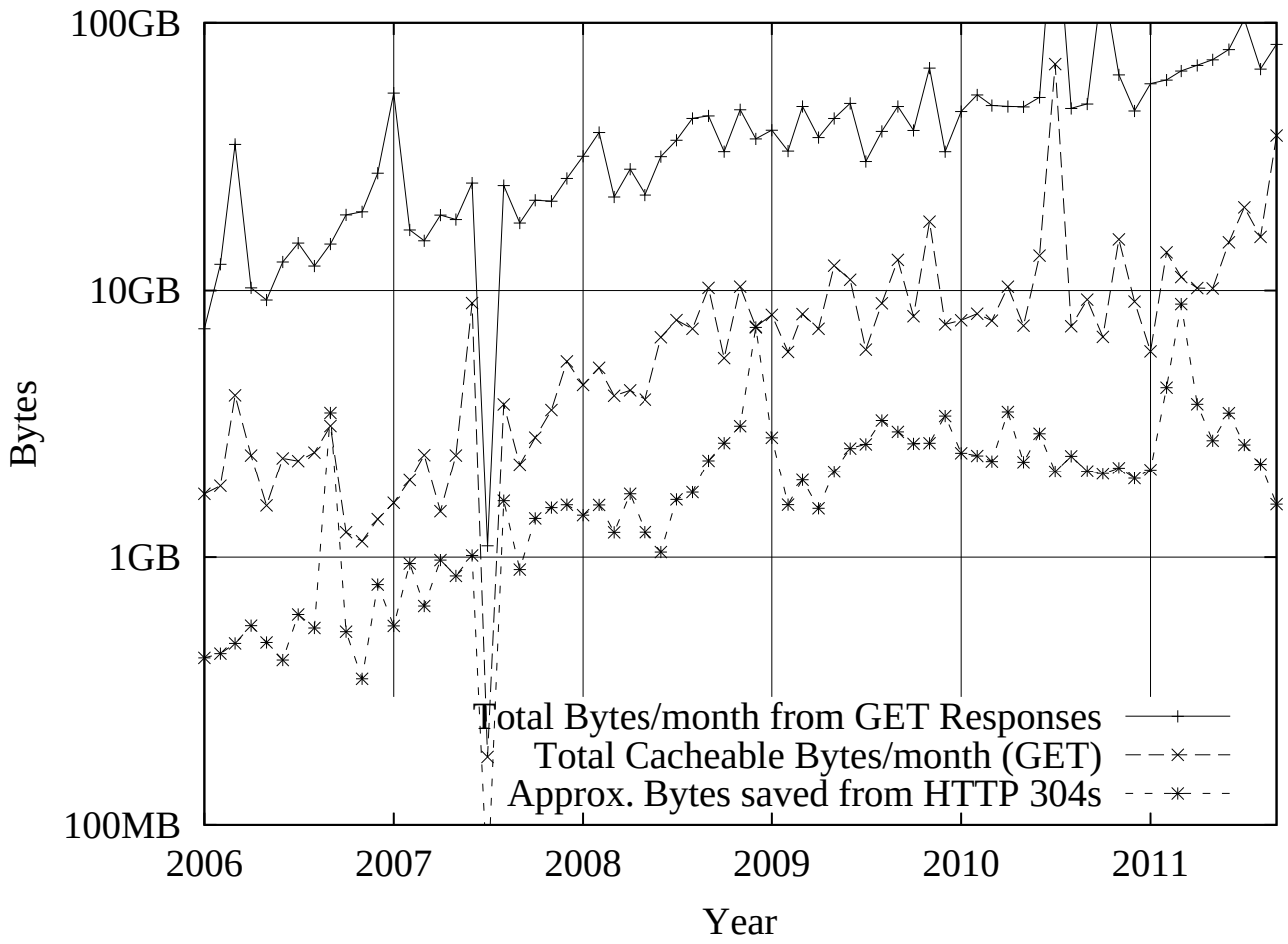


Figure: ICSI Caching

# Caching Potential as Traffic Percentage

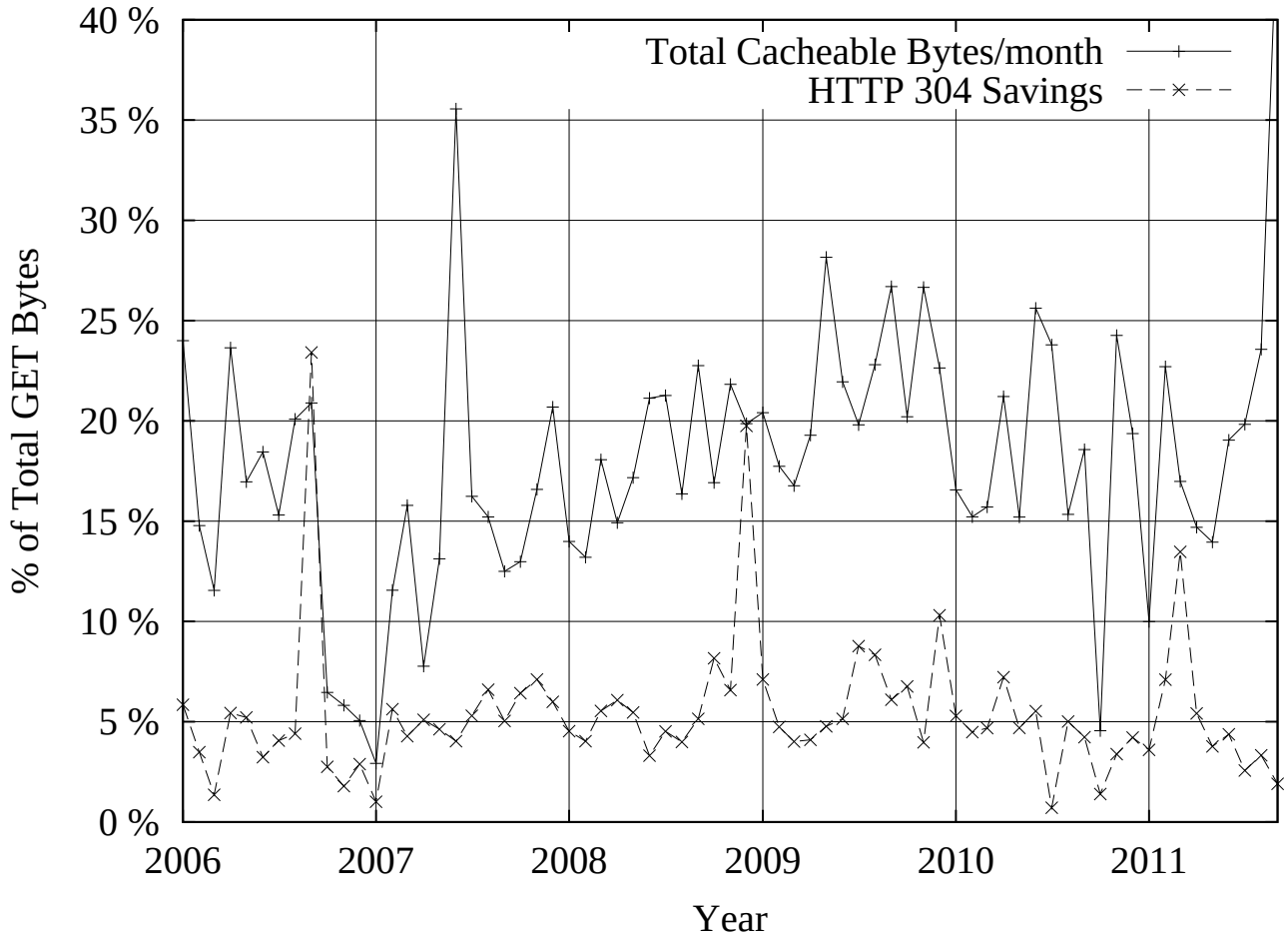


Figure: ICSI Caching Ratio

# Server Structure

- Relationship between IP addresses and hostnames
- Distribution of objects served by each website
- Bytes provided by a popular Content Delivery Network (CDN)

# Objects per Hostname

- At ICSI from 2007 on, $\sim$ 30% of sites serve only a single object to our users
- Likewise, 72% of sites serve ICSI's users 10 or fewer objects during this timeframe

# Distribution of Objects per Hostname

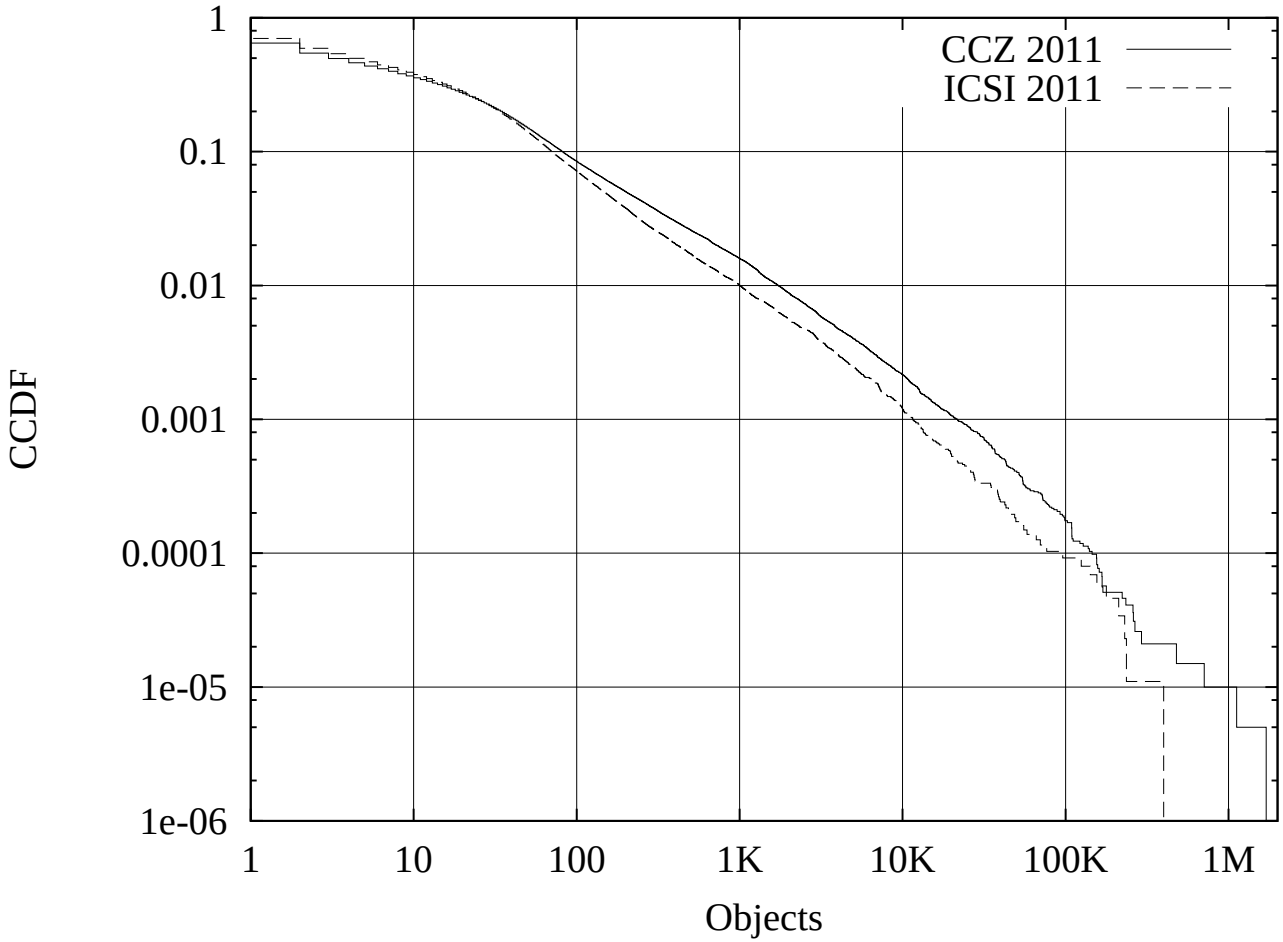


Figure: Objects per Hostname

# IP/Hostname Topology

- In our user populations, 20-30% of IP addresses serve multiple hostnames
  - Shared hosting
  - CDN Usage
- At the top 1% of IPs, at least 35 hostnames are served
- In both populations, $\sim$ 17% of hostnames are served by multiple IP addresses
  - Server clusters
  - CDNs
- Approximately 8% of sites are seen on 2 IPs, another 5% on 3-4, and 4% on 5 or more

# Content Delivery Networks

- CDNs composed of a large number of servers (or clusters) distributed around the globe
- Websites pay CDNs to serve their content to clients, typically from a CDN server close to the client
- Often easy to spot this traffic as DNS resolutions of affected hostnames contain obvious cues
  - akamai.net, edgesuite.net, etc
- One popular CDN, Akamai, claims to send between 15-30% of bytes on the Internet

# Akamai Traffic

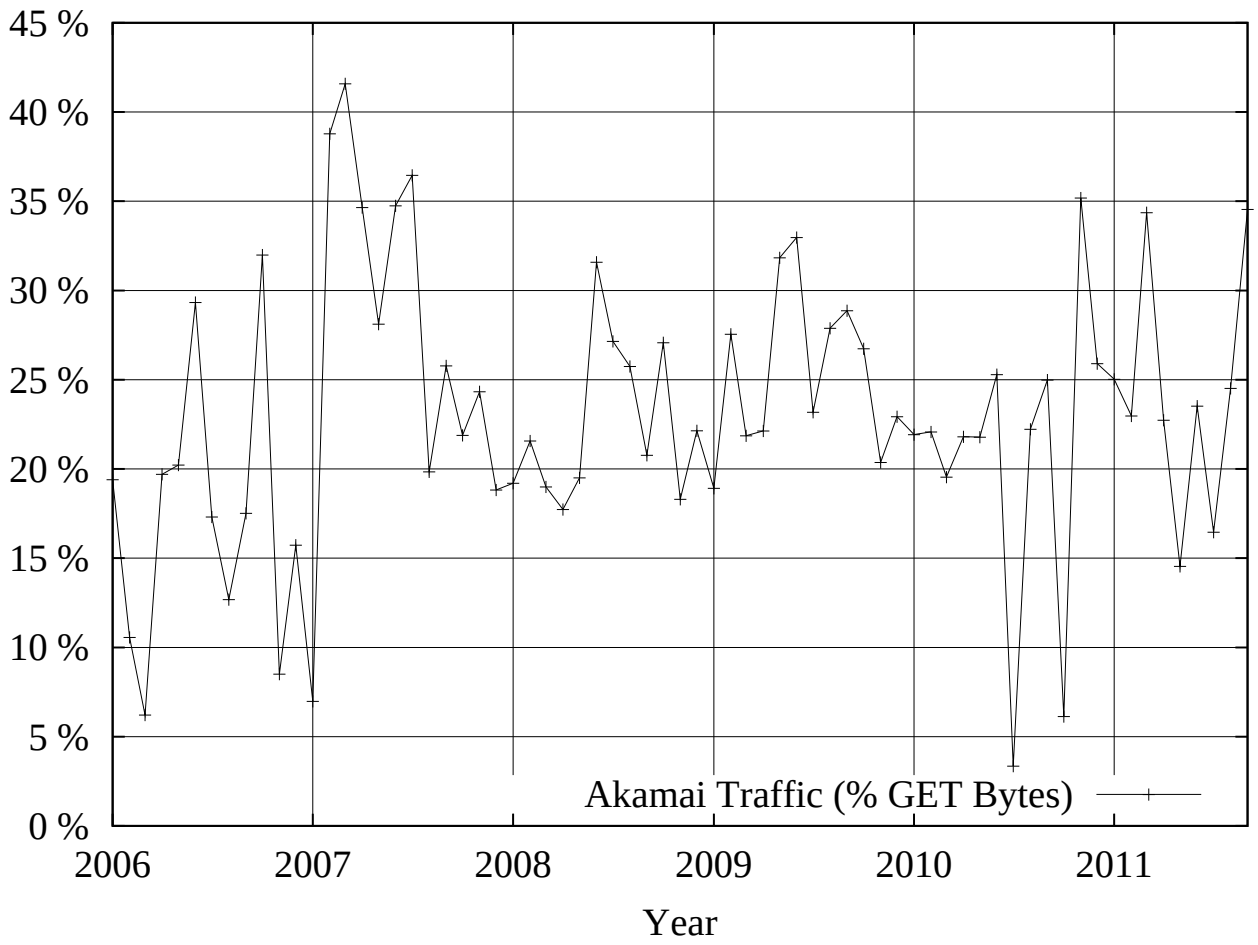


Figure: Akamai Traffic

# Summary

Throughout this thesis, we strove to update the community's mental models regarding the use of HTTP in the wild. Some elements that we have observed to change over time include:

- An increase in nearly all raw counts such as connections, requests, bytes, etc.
- A large increase in HTTPS connections in 2010 and 2011
- An increase in average transaction sizes with no apparent effect on median transaction sizes
- Increases in connection parallelism due to widespread browser changes
- A trend towards dynamic content and newer image formats

# Summary (cont'd)

We have also found many properties with little change:

- The shape of the distribution of object sizes
- Average number of requests in an HTTP connection
- The original Internet TLDs remain the most popular, .com and .net in particular
- Images remain the most commonly requested file type
- The shapes of the distributions of object and hostname popularity remain static

# Future Work

- Distinguishing characteristics of primary objects (pages visited) versus secondary objects (embedded within pages visited)
- Distinction between automated and user-generated requests
- Page load times
- Analysis of multiple CDNs

# That's all, folks!

Questions?