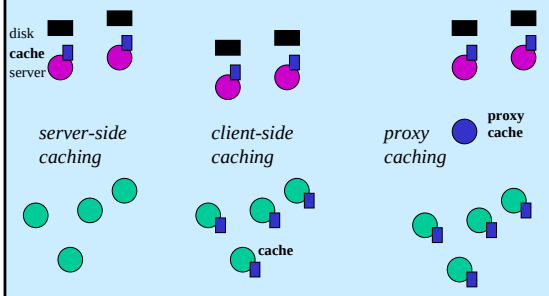
Cooperative Web Caching: Wolman, Voelker, Sharma, Cardwell, Karlin, Levey [Wolman99a]

CSci551: Computer Networks
SP2006 Thursday Section
John Heidemann

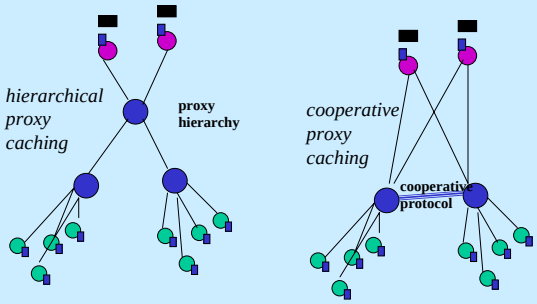
Key ideas

- cooperative proxy caching
- looks at performance of existing proxy caching
- looks at population size served by the proxy

Simple Web Caching



Distributed Web Caching



Caching Questions

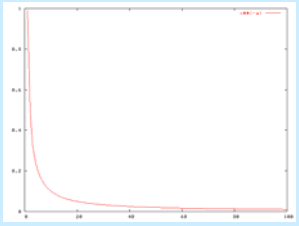
- Does {server, client, proxy, hierarchical, cooperative} help?
  - relative benefits?
    - latency
    - bandwidth
  - relative costs?
    - latency and bandwidth overheads
  - do they vary as a function of population?
    - size
    - homogeneity

Cooperative Caching: Internet Cache Protocol

- Upon cache miss
  - … proxy sends request to all other peer proxies using UDP
  - peers respond if they have a hit
- Key result
  - Number of UDP messages with ICP increases by 2 orders of magnitude
  - … but ICP can increase hit rate
- Question: what are the trade-offs?

Web Page Popularity

- Access to Web pages follows Zipf's law
  - The number of times the ith most popular document
  - … is accessed is proportional to i-a
  - (classic heavy tail: linear on a log-log plot)



Caching Performance?

- Access to Web pages follows Zipf's law
- The hit ratio of large caches
  - Grows logarithmically with the client population
  - or with the number of requests seen by the cache
- Should nearby caches exchange their contents to improve performance?
  - What are the limits to this performance improvement?

Methodology

- Collect Web traces simultaneously from two large organizations *simultaneously*
  - University site (U-Washington)
  - Large corporation (Microsoft)
- Can study the benefits of cache cooperation at two levels of granularity
  - Between departments at the university
  - Between the large organizations
- Traces collected in the absence of caching
  - But can use these to infer the impact of caching
- Methodological issues
  - Infinite cache size assumptions
  - Practical (conforms to HTTP control directives) vs. ideal (assuming all documents are cacheable) caching

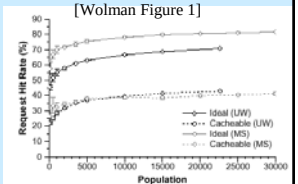
Key Ideas, reviewed

- XXX

Impact of Client Population

- Higher ideal hit rate
  - Due to a more homogeneous accesses from MS compared to UW
- But, comparable hit rates with only cacheable documents
  - Current cacheability constraints limit performance
  - Many images marked uncacheable (why?)
- Knee of performance curve
  - where? 2500 people
  - why?

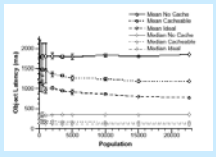
[Wolman Figure 1]

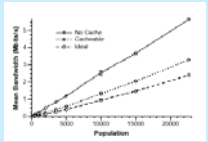


• incremental benefit of caching for more than 2500 people is minmal
• therefore, collaborative caching isn't necessary for >2500 people

Latency and Bandwidth

- Latency
  - Caching can reduce latency
  - Difference between practical and ideal
- Bandwidth
  - Caching reduces bandwidth
  - But the bandwidth reduction is insensitive to client population



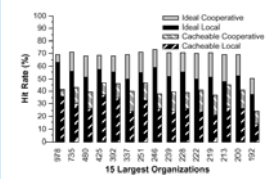


Why?

- xxx

Cooperative Caching by Group

- Can provide some benefits
  - how much? xxx
- There exists little affinity of access within an organization
- For large populations, not much increased benefit to cooperative caching



Analytic Model

- to complement the trace analysis, they develop an analytic model
  - approximate hit rate, latency, bandwidth savings
  - set parameters from traces, then explore alternatives
- results similar to trace results
  - big win (hit rate, latency) to caching up to 10^5-10^6 users; little benefit beyond that

Summary

- Most benefit from small populations
  - But this can be served by a single proxy, so little need for cooperative caching
- Some benefit for collections of small, diverse populations
- Greater benefits obtainable from increasing the cacheability of documents

Other questions/observations?

- use caching in your web browsing?
  - mostly noone?
- if bandwidth keeps going up (faster than processing), why bother caching?
  - (and in fact the realative penalty is growing)
  - but bandwidth is shared and processors are dedicated
  - and what about latency?