Web Cache Sharing

Kai Shen
Dept. of Computer Science, University of Rochester

Web Proxy Caching

- Cache is installed and shared by a group of users (university, company, residential ISP)
- Goal: satisfy client requests without involving the original server.
- user sets browser: accesses via Web cache
- browser sends all HTTP requests to Web cache
- object in cache: cache returns object
- otherwise cache requests object from the original server, then returns object to client

Benefits:
- reduce access latency; reduce long-haul traffic volume
- The key is cache hit ratio, which depends on:
  - request repetitiveness
  - cache size
  - replacement algorithm

A Group of Web Caches

origin servers

public Internet

UR

gateway

CS dept

Math dept

Web cache

ECE dept

Web cache

Yahoo! CNN
Web Cache Sharing

- Web cache sharing
  - in the event of a local cache miss, try to use a cooperative cache
  - advantage: increase overall hit ratio \(\Rightarrow\) reduce latency and reduce traffic to the wide-area Internet
- Potential benefit:
  - increase overall hit ratio
  - reduce latency
  - reduce traffic to the wide-area Internet

<table>
<thead>
<tr>
<th>Traces</th>
<th>DEC</th>
<th>UCB</th>
<th>UFisa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requests</td>
<td>3,543,968</td>
<td>1,907,762</td>
<td>2,833,624</td>
</tr>
<tr>
<td>Total size</td>
<td>28.6 GB</td>
<td>18.0 GB</td>
<td>20.7 GB</td>
</tr>
<tr>
<td>Hit ratio</td>
<td>49%</td>
<td>30%</td>
<td>49%</td>
</tr>
<tr>
<td>Client groups</td>
<td>10089</td>
<td>5780</td>
<td>2203</td>
</tr>
<tr>
<td>Hit with sharing</td>
<td>43%</td>
<td>27%</td>
<td>45%</td>
</tr>
<tr>
<td>Hit w/o sharing</td>
<td>21%</td>
<td>13%</td>
<td>27%</td>
</tr>
</tbody>
</table>

Web Cache Sharing - How?

- At the event of a local cache miss, a query message is broadcast to all cooperative caches (Internet Cache Protocol)
  - per-cache communication overhead and processing overhead increases linearly with the number of caches \(\Rightarrow\) unscalable
- Partition the URL space among caches
  - non-uniform network distances between caches and users
- Maintain local information about what others have
  - appears difficult since "what everyone has" changes quickly

Summary Cache

- Summary cache:
  - Proposed in [Li et al, ACM/IEEE TON 2000]
  - The basic idea is to maintain local information (called "summary cache") about what others have
  - At the event of a local miss, the local copy of summary caches for all cooperative caches will be examined. A query is sent to a specific cache if its summary cache reports a hit.
- Tolerated errors – they affect hit ratio, but not correctness
  - false misses
  - false hits
- Overhead
  - Reactive overhead to client HTTP requests
  - Proactive overhead for maintaining summary cache

Summary Cache Maintenance

- Each cache periodically calculates local summary cache – an inaccurate but compact data structure indicating what is in local cache
- Each cache then broadcasts the local summary cache to all other cooperative nodes
- The overhead can be controlled by the frequency of such calculation/broadcast
- Summary cache
  - its accuracy affects the overall hit ratio
  - its size affects the communication overhead and storage consumption
Bloom Filters

- Bloom filter - a compact data structure for representing a set \(\{a_1, a_2, ..., a_n\}\) to support membership queries.

![Bloom Filters Diagram]

Why is it compact?

- False positive?
- False negative?

Putting it together!

- Each cache periodically calculates local summary cache and then broadcasts the local summary cache to all other cooperative caches.
- At the event of a local miss, the local copy of summary caches for all cooperative caches will be examined. A query is sent to a specific cache if its summary cache reports a hit.

- Question:
  - Are there false hits? If so, how to control it?
  - Are there false misses? If so, how to control it?

Summary

- Web cache sharing can help improve overall cache hit ratio ⇒ reduce access latency; reduce wide-area long-haul traffic
- It is not straightforward to balance scalability and accuracy for Web cache sharing
- A solution: summary cache based on “Bloom filters”

Question: will a global replacement policy further improve the cache hit ratio?