Nearline Web Archiving

Zhiwu Xie[1], Krati Nayyar[2], and Edward A. Fox[3]

[1]University Libraries, [2]Department of Electrical & Computer Engineering,
[3]Department of Computer Science;
Virginia Tech, Blacksburg, VA 24061

Terminology

Julien Masanès, “Web Archiving ”, Springer, 2006

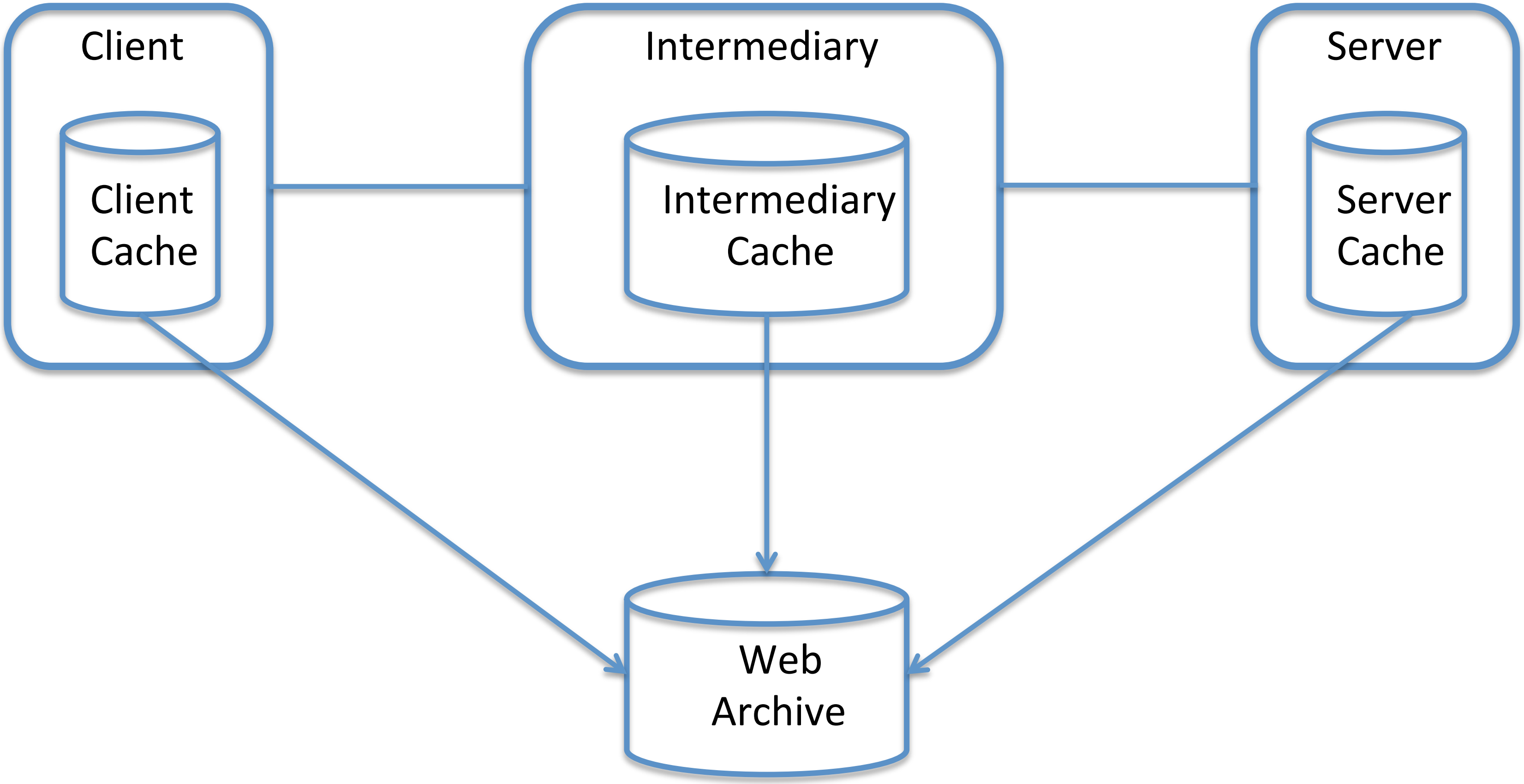
The gathering process can be done:

Client-side archiving … remotely as client, e.g., crawler, website copier, browser plug-ins
Server-side archiving … by direct access to the server’s files, e.g., regularly backup files under webroot and archive
Transactional archiving … close to the output of the server, based on recording transactions, e.g., SiteStory

Terminology

- Definitions are loose, not mutually exclusive
    - A 3-tier server, all content saved in MySQL that does not allow shell access, then we fire a web client, e.g., MySQLAdmin, send a request to either http://example.com/mysqladmin or http://localhost/mysqladmin, request a full db dump and extracts all content programmatically
    - Same as above but allows shell access to db, directly run a db dump, but from a remote ssh client?
    - On the server, run a crawler against localhost?
    - How “close” to the output?
    - A browser plug-in archives each request/response pair. Why isn’t this transactional archiving?
- Must not mix who, what, when, where, why, and how.

Archiving Cache: A New Method?



Online, Offline, Nearline

Online archiving Archiving is part of the web transaction and always adds to the server load, and can only archive one response at a time.

Offline archiving Archiving is NOT part of the web transaction. A separate process, therefore can batch archiving many responses at a time.

Nearline archiving Depends on the accumulation of web transactions, but as a separate process, can be batched, but in smaller granularity.

Archiving Apache Disk Cache: A Prototype

- Apache disk cache is not enabled by default but can be easily configured to enhance server performance
- Cached copies are stored as files on disk, and not automatically cleared even after expiration
- A separate tool, htcacheclean, may be invoked or set in deamon mode to remove cached files I regular intervals
- htcacheclean list and walk the cache dir, then first delete expired cache, then delete valid cache from old to new ones, until the cache directory is within the specified size limit
- htcacheclean may be (and is indeed usually) run in lower priority than the apache web server
- Before deleting files, we can first write them all to a warc file, then ship to the remote archive.

Archiving Apache Disk Cache: Demo

https://drive.google.com/file/d/0B1fNH27Z6Cu8dEpIOVR5dXlmZm8/view?usp=sharing

Nearline Archiving

- Like SiteStory style transactional archiving, can capture full history of a server, but needs site owners' coperation
- Unlike crawlers, does not incur additional server load
- Unlike SiteStory, archiving is done in small batches instead of one response at a time, much lower server side overhead
- Archiving is a separate process from the web server, will only run when the web server is not busy serving HTTP responses, therefore should not significantly affect the server performance and is more friendly to server operators
- Trade archival freshness for better performance

Future Work

- Better WARC support
- Expand to other cache, e.g., memcached, redis, browser cache
- Rigorous performance testing