Web Archive Processing

Mike Smorul
ADAPT Group
University of Maryland, College Park
Web Archive Storage and Search

- Management tools
- Storage infrastructure
- Indexing, searching, compression experiments
Webarc Manager Motivation

• We are indexing large collections of crawled web sites:
  - What content do we have?
  - What are its characteristics?
  - How many URLs?
  - How much unique content? Duplicates?
Webarc Manager

• A tool to help manage webarc collections

• Show statistics of a series of crawls

• REST-API to easily query collection
  – List all copies of a page, etc
Manager Components

- WarcManager (server)
  - REST-based access to index
  - Index of DAT/ARC entries
  - URL Searching, ARC browsing,

- Javascript Client

- Simple Web-Accessible Preservation (SWAP)
  - Web-accessible distributed storage
  - ARC page retrieval
  - 1Gbps, 2200 requests/s
## URL Example

**http://www.whitehouse.gov/news/releases/2001/05/**

<table>
<thead>
<tr>
<th>Archive Date</th>
<th>Length</th>
<th>New Digest</th>
</tr>
</thead>
<tbody>
<tr>
<td>8/18/2004-10/6/2004 (8 duplicates)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9/1/2004</td>
<td>98042</td>
<td>a43f7650eb3f3f341</td>
</tr>
<tr>
<td>9/8/2004</td>
<td>98042</td>
<td>a43f7650eb3f3f341</td>
</tr>
<tr>
<td>9/15/2004</td>
<td>98042</td>
<td>a43f7650eb3f3f341</td>
</tr>
<tr>
<td>9/22/2004</td>
<td>98042</td>
<td>a43f7650eb3f3f341</td>
</tr>
<tr>
<td>10/6/2004</td>
<td>98042</td>
<td>a43f7650eb3f3f341</td>
</tr>
<tr>
<td>9/29/2004</td>
<td>98042</td>
<td>a43f7650eb3f3f341</td>
</tr>
<tr>
<td>8/18/2004</td>
<td>98042</td>
<td>a43f7650eb3f3f341</td>
</tr>
<tr>
<td>8/25/2004</td>
<td>98042</td>
<td>a43f7650eb3f3f341</td>
</tr>
<tr>
<td>10/14/2004-10/26/2004 (3 duplicates)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10/26/2004</td>
<td>98485</td>
<td>56cb8c99b34f03</td>
</tr>
<tr>
<td>10/14/2004</td>
<td>98485</td>
<td>56cb8c99b34f03</td>
</tr>
<tr>
<td>10/20/2004</td>
<td>98485</td>
<td>56cb8c99b34f03</td>
</tr>
<tr>
<td>6/30/2004</td>
<td>57408</td>
<td>a7d3a4e6353e3b</td>
</tr>
<tr>
<td>8/3/2004</td>
<td>57408</td>
<td>a7d3a4e6353e3b</td>
</tr>
<tr>
<td>7/28/2004</td>
<td>57408</td>
<td>a7d3a4e6353e3b</td>
</tr>
<tr>
<td>7/27/2004</td>
<td>57408</td>
<td>a7d3a4e6353e3b</td>
</tr>
<tr>
<td>7/21/2004</td>
<td>57408</td>
<td>a7d3a4e6353e3b</td>
</tr>
<tr>
<td>7/24/2004</td>
<td>57408</td>
<td>a7d3a4e6353e3b</td>
</tr>
<tr>
<td>8/11/2004</td>
<td>57408</td>
<td>a7d3a4e6353e3b</td>
</tr>
</tbody>
</table>

**Page Title:** News Archive - May 2001

**MIME Type:** text/html

**Crawl Date:** Wed 15 Sep 2004 12:00:00 AM EST

**Entry Size:** 98042

**Digest:** a43f7650eb3f3f341eddd8709b6c244b

**ARC File:** IQ04-CRAWL-11-20040915040557-00122-crawling003.archive.org


**Download File** | **View Parent Arc**
Manager Design

- Javascript Client
  - REST
  - JSON
- Webarc Manager
- MySQL
- Storage Server
- Storage Server
- Page Request
Storage Design

• SWAP – Simple, Web-Accessible Preservation
• Intelligent placement of files across multiple servers and disk partitions
• Simple HTTP access, PUT, GET, DELETE
• Use redirects to provide a uniform namespace
• Files organized into file groups
  – Each group resides on multiple partitions (slices)
  – Hash(file_path) % slices = partition
• No centralized catalog
How it works

Calculate hash(bag1/warc55.gz) % 4 = 3
Performance

- Good small file and large file performance
  - Over 2000 requests/s and 3000 redirects/
Time Machine for the Web

• Fast parallel indexer to handle large scale crawled web contents, coupled with a new compression scheme.

• Fast search of contents based on unstructured queries involving temporal specifications.

• Presentation of pertinent summary information in ranked order according to the temporal context.
Current and Future performance

• One node
  - Using GPUs, 300MB/s per node
• Scale out to 32, infiniband connected nodes
• Index writing, web page reading at a massive scale
Additional Information

• [http://adapt.umiacs.umd.edu](http://adapt.umiacs.umd.edu)
  - Papers, results, etc..

• E-mail: msmorul@umiacs.umd.edu