# 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?

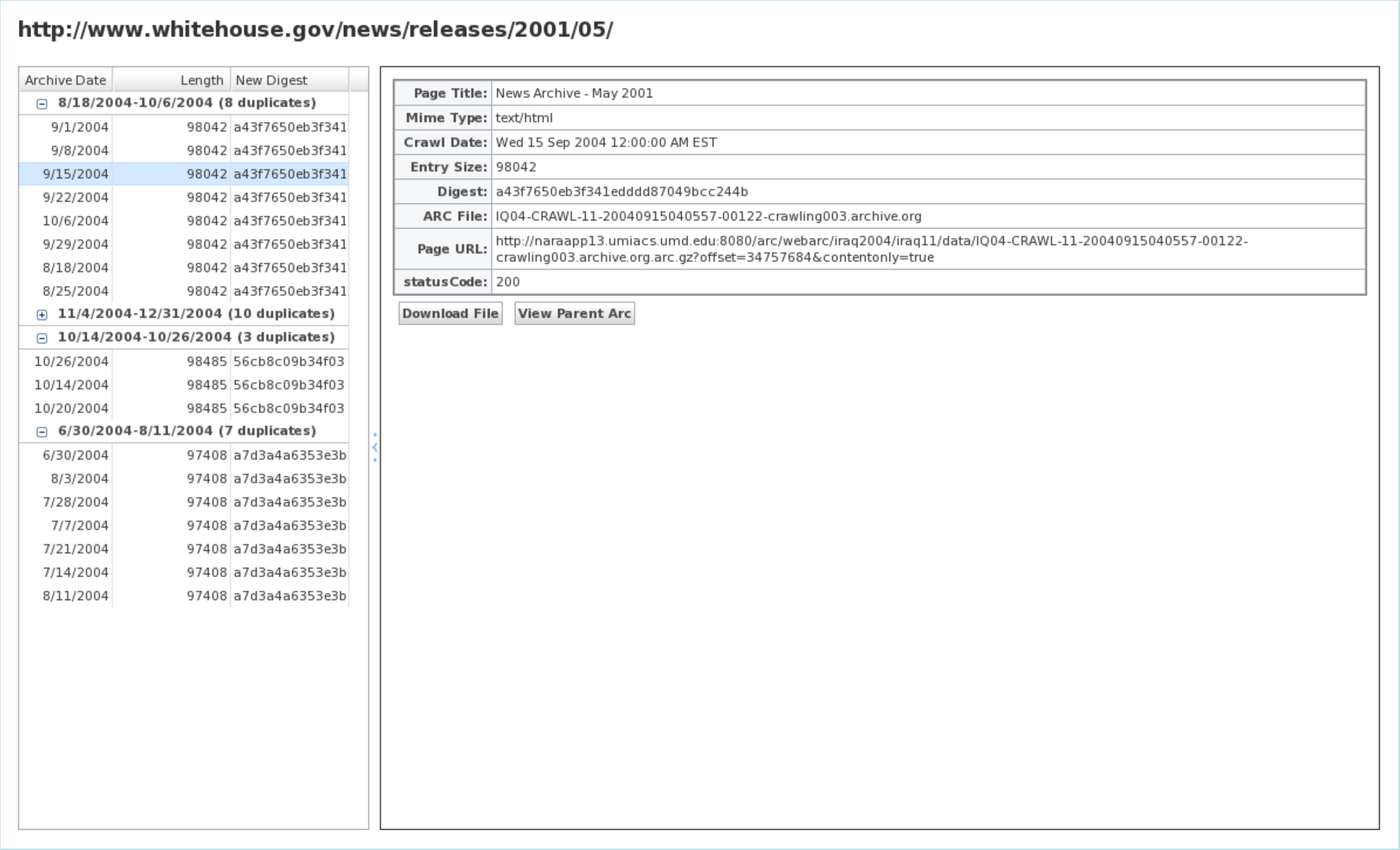
# Webarc Manager

- 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, 2200requests/s

# URL Example

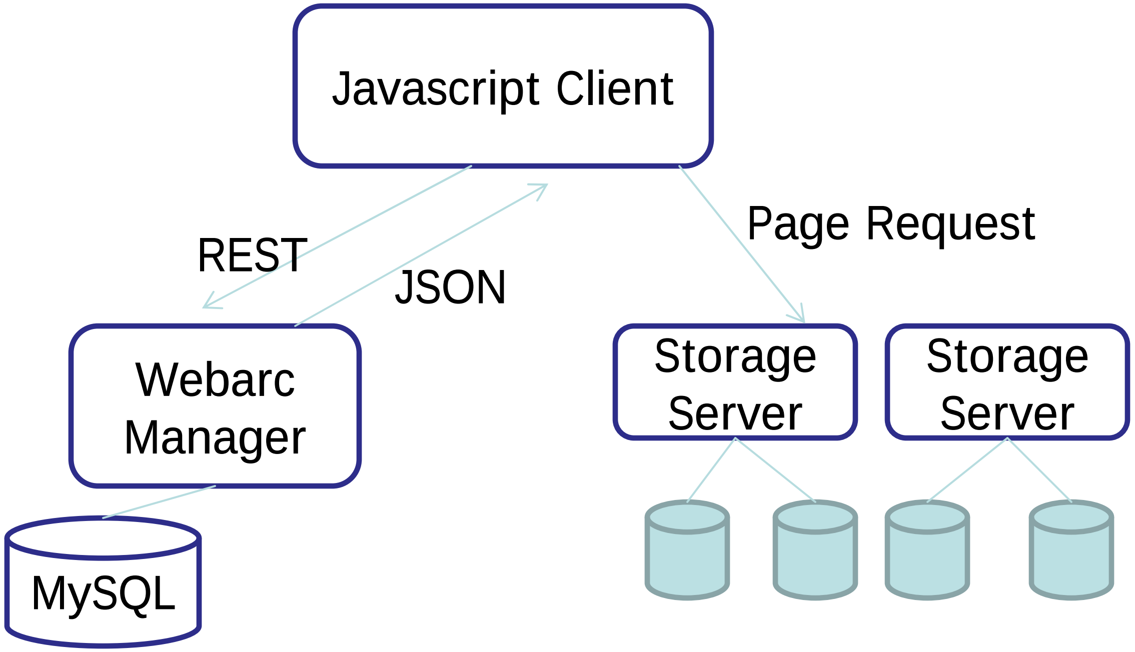


http://www.whitehouse.gov/news/releases/2001/05/

| Archive Date | Length | New Digest |
|---|---|---|
| ⊟ **8/18/2004-10/6/2004 (8 duplicates)** | | |
| 9/1/2004 | 98042 | a43f7650eb3f341 |
| 9/8/2004 | 98042 | a43f7650eb3f341 |
| 9/15/2004 | 98042 | a43f7650eb3f341 |
| 9/22/2004 | 98042 | a43f7650eb3f341 |
| 10/6/2004 | 98042 | a43f7650eb3f341 |
| 9/29/2004 | 98042 | a43f7650eb3f341 |
| 8/18/2004 | 98042 | a43f7650eb3f341 |
| 8/25/2004 | 98042 | a43f7650eb3f341 |
| ⊞ **11/4/2004-12/31/2004 (10 duplicates)** | | |
| ⊟ **10/14/2004-10/26/2004 (3 duplicates)** | | |
| 10/26/2004 | 98485 | 56cb8c09b34f03 |
| 10/14/2004 | 98485 | 56cb8c09b34f03 |
| 10/20/2004 | 98485 | 56cb8c09b34f03 |
| ⊟ **6/30/2004-8/11/2004 (7 duplicates)** | | |
| 6/30/2004 | 97408 | a7d3a4a6353e3b |
| 8/3/2004 | 97408 | a7d3a4a6353e3b |
| 7/28/2004 | 97408 | a7d3a4a6353e3b |
| 7/7/2004 | 97408 | a7d3a4a6353e3b |
| 7/21/2004 | 97408 | a7d3a4a6353e3b |
| 7/14/2004 | 97408 | a7d3a4a6353e3b |
| 8/11/2004 | 97408 | a7d3a4a6353e3b |

| | |
|---|---|
| **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:** | a43f7650eb3f341edddd87049bcc244b |
| **ARC File:** | IQ04-CRAWL-11-20040915040557-00122-crawling003.archive.org |
| **Page URL:** | http://naraapp13.umiacs.umd.edu:8080/arc/webarc/iraq2004/iraq11/data/IQ04-CRAWL-11-20040915040557-00122-crawling003.archive.org.arc.gz?offset=34757684&contentonly=true |
| **statusCode:** | 200 |

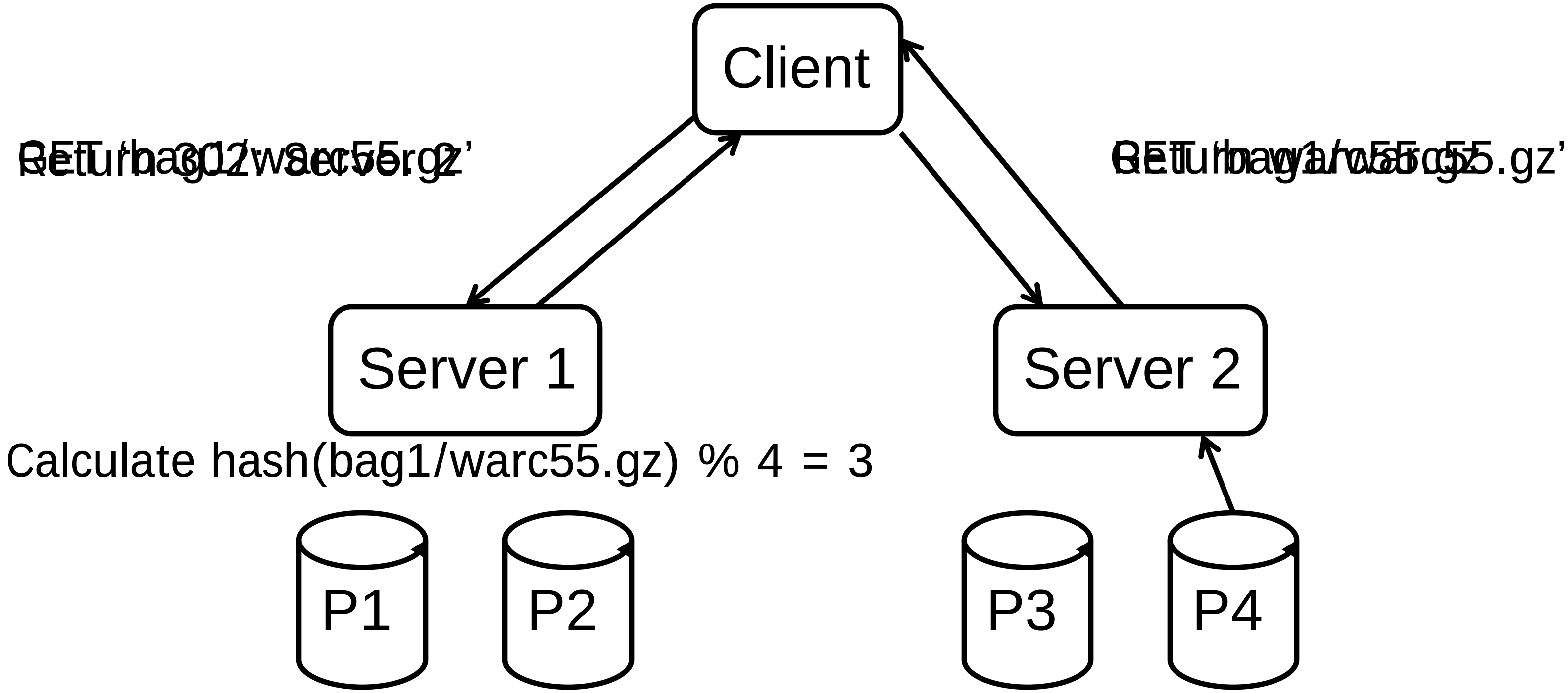
# Manager Design



# Storage Design

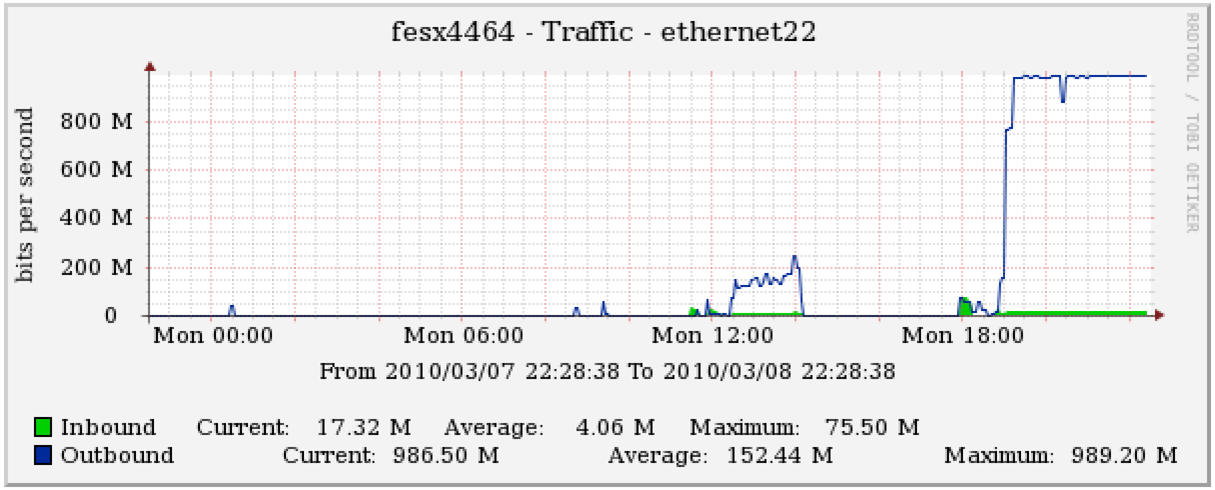
- SWAP – Simple, Web-Accessible Preservation
- Intelligent placement of files across multiple servers and disk partitions
- Simple HTTP access, PUT, GET, DELETE
- Use redirects to provide a uniform namespace
- Files organized into file groups
  - Each group resides on multiple partitions (slices)
  -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

# How it works


Client
GET 'bag1/warc55.gz'
Return 302: Server 2
GET 'bag1/warc55.gz'
Return warc55.gz
Server 1
Server 2
Calculate hash(bag1/warc55.gz) % 4 = 3
P1
P2
P3
P4

# 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
  - – Papers, results, etc..
- E-mail: msmorul@umiacs.umd.edu