IIPC Memento Aggregator Experiment

Robert Sanderson
Herbert Van de Sompel
Michael L Nelson

http://www.mementoweb.org/

This research was funded by
the Library of Congress.

Towards Seamless Navigation of the Web of the Past

Summary

Project Overview

Participants

The Plan …
… and the Reality

Demo

Technical Explanation

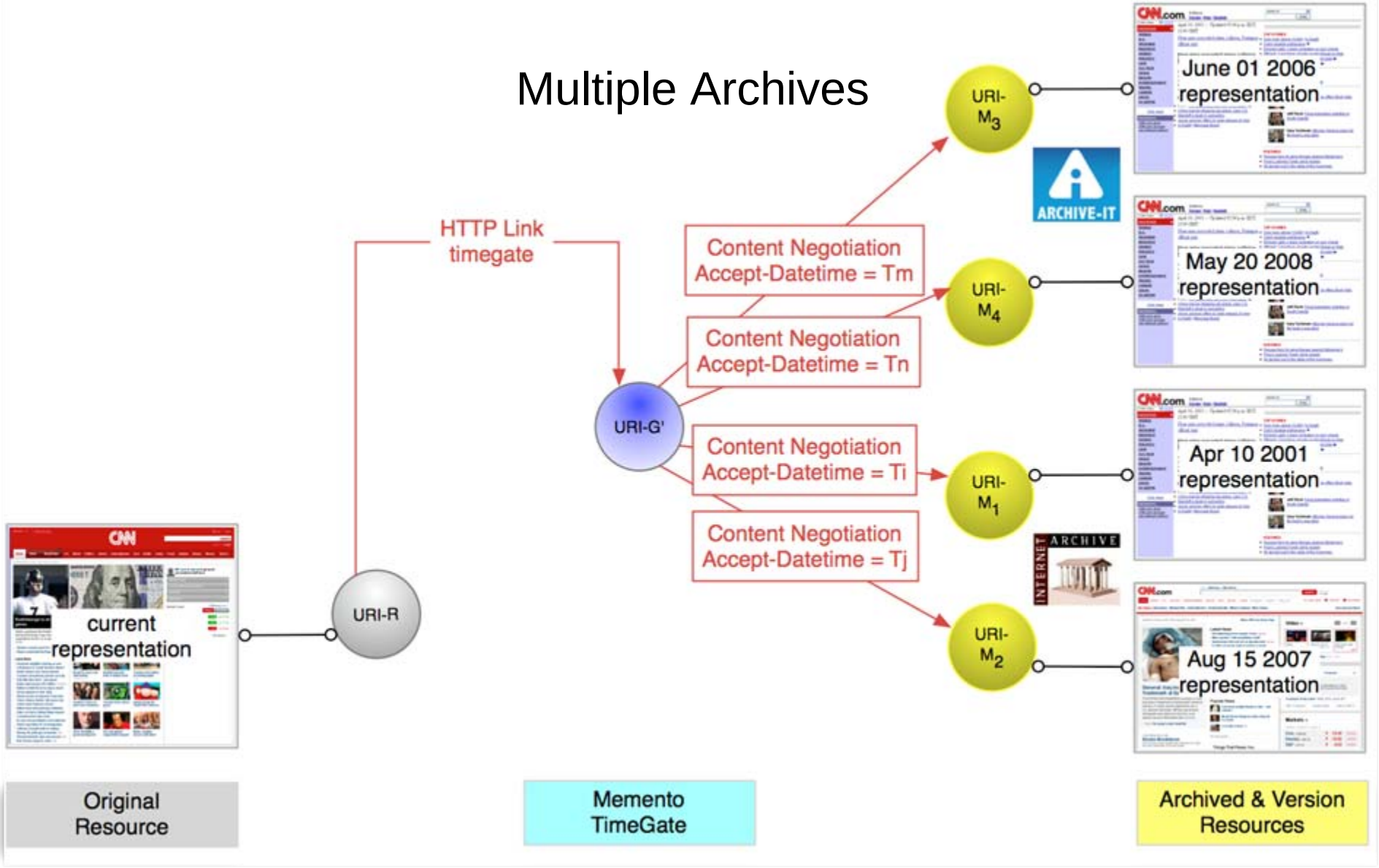
Next Steps

Project Overview

- Goal: To aggregate the metadata of the distributed archives of the IIPC, and
  - To provide Memento based access to the holdings of open archives
  - To provide knowledge of the holdings of restricted archives
  - To provide knowledge to IIPC members of the holdings of totally closed archives

- Initial demo for participants, then IIPC
- No access provided to restricted archives (of course)

Multiple Archives



Experiment Participants

- Austrian National Library
- Bibliothèque Nationale de France
- British Library
- Institut National de l'Audiovisuel
- Internet Archive
- Koninklijke Bibliotheek
- Library of Congress
- Netarchive.dk
- Swiss National Library
- University of North Texas

- Los Alamos National Laboratory
- Old Dominion University

The Plan…

- To provide fast access to distributed archives, LANL would merge the indexes of the holdings of multiple archives and provide Memento based access

- Step 1: Library of Congress gathers CDX files
  Step 2: LANL indexes (…)
  Step 3: Profit

- Data: 5T of gzipped CDX files (mostly from IA)
  - Shipped on hard drives
- Computing: 210 node cluster at LANL
  - 2x 2ghz processors, 2x 2T HDD, 8G RAM

… and the Reality

- Hardware failure killed one of the drives en route
  - Transferred remaining files via BagIt from LoC
- Compute cluster has very restricted access:
  - Had to transfer data over infranet
  - 2 weeks to sync (5Mb/sec)
  - And then 2 weeks to get the processed results off
- Compute cluster has faulty switch, unreliable nodes:
  - Ran original processing 15 times without success due to hardware failures

Demo

Processing Design

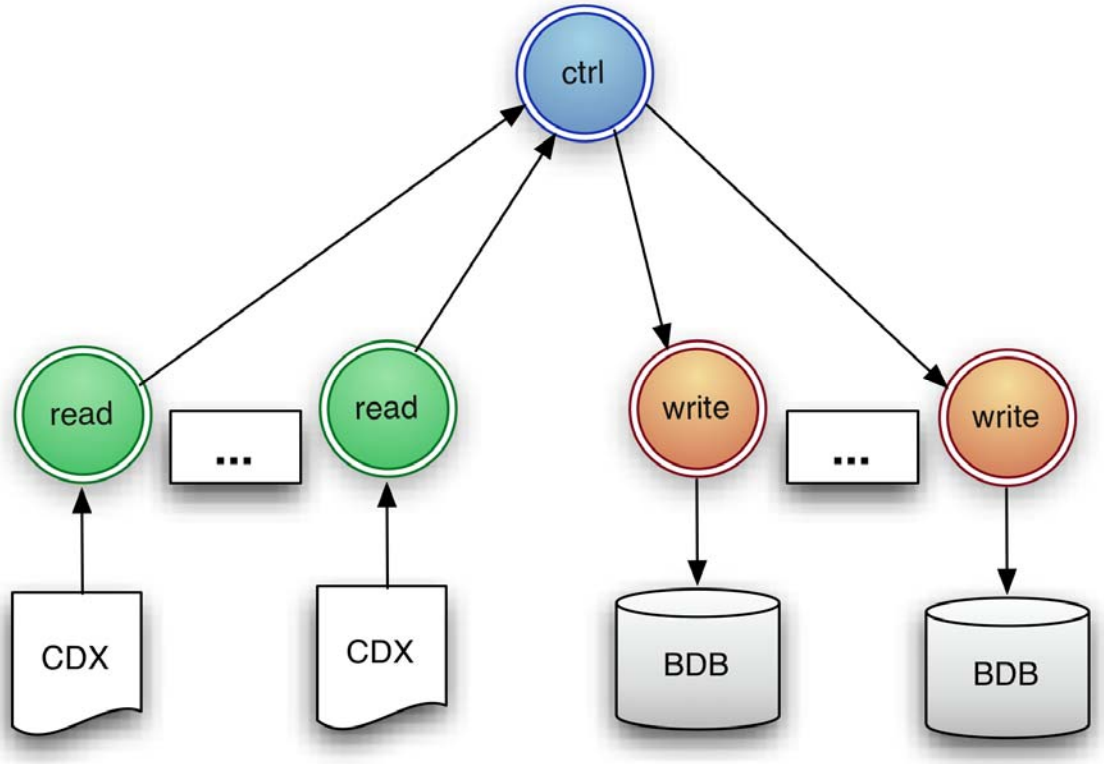
- For each CDX file,
  - For each URI + timestamps,
    - Map URI to an appropriate database slice
    - Merge timestamps with those of previous CDXs

- Possible because:
  - No need to do truncated search
  - No need to walk through URIs in order
  - No need for time based access, only URI

- Problem is “Embarrassingly Parallel”

Approach 1: Online Messaging

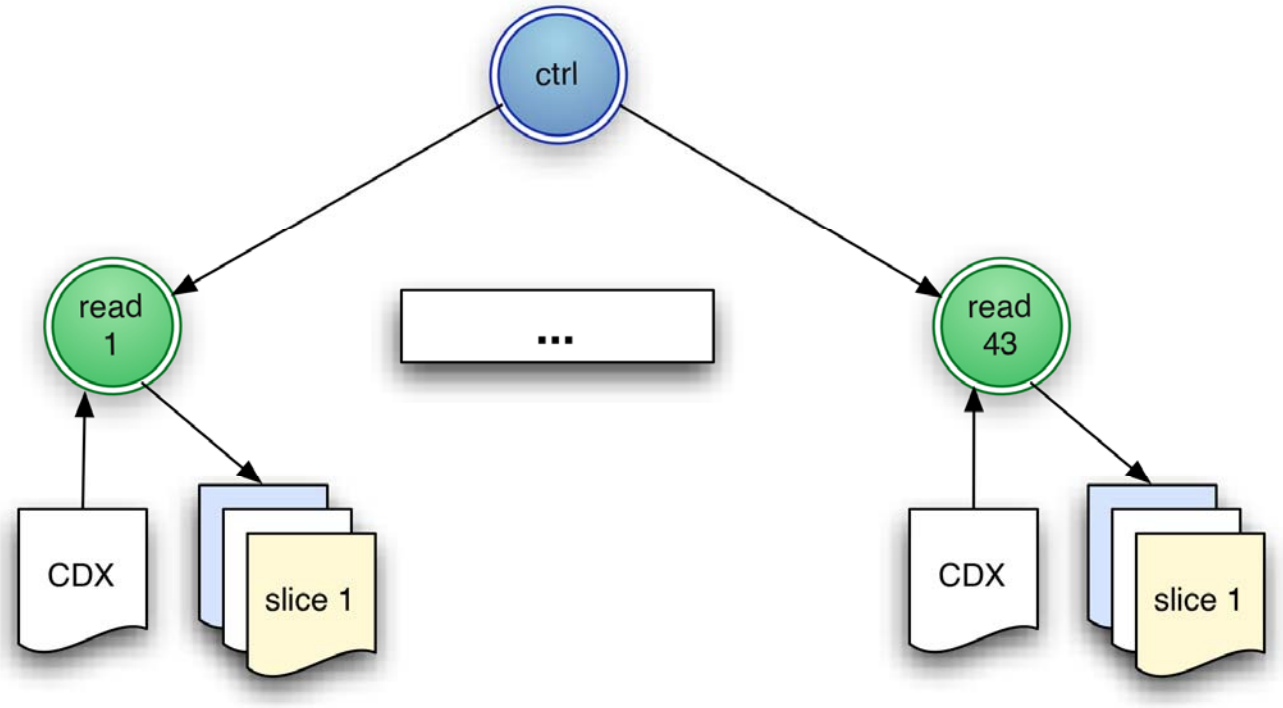
- 25 read nodes, 1 control node, 150 write nodes
- Messages (1000 URIs) sent via control node to write



- Failed 15 times due to hardware issues

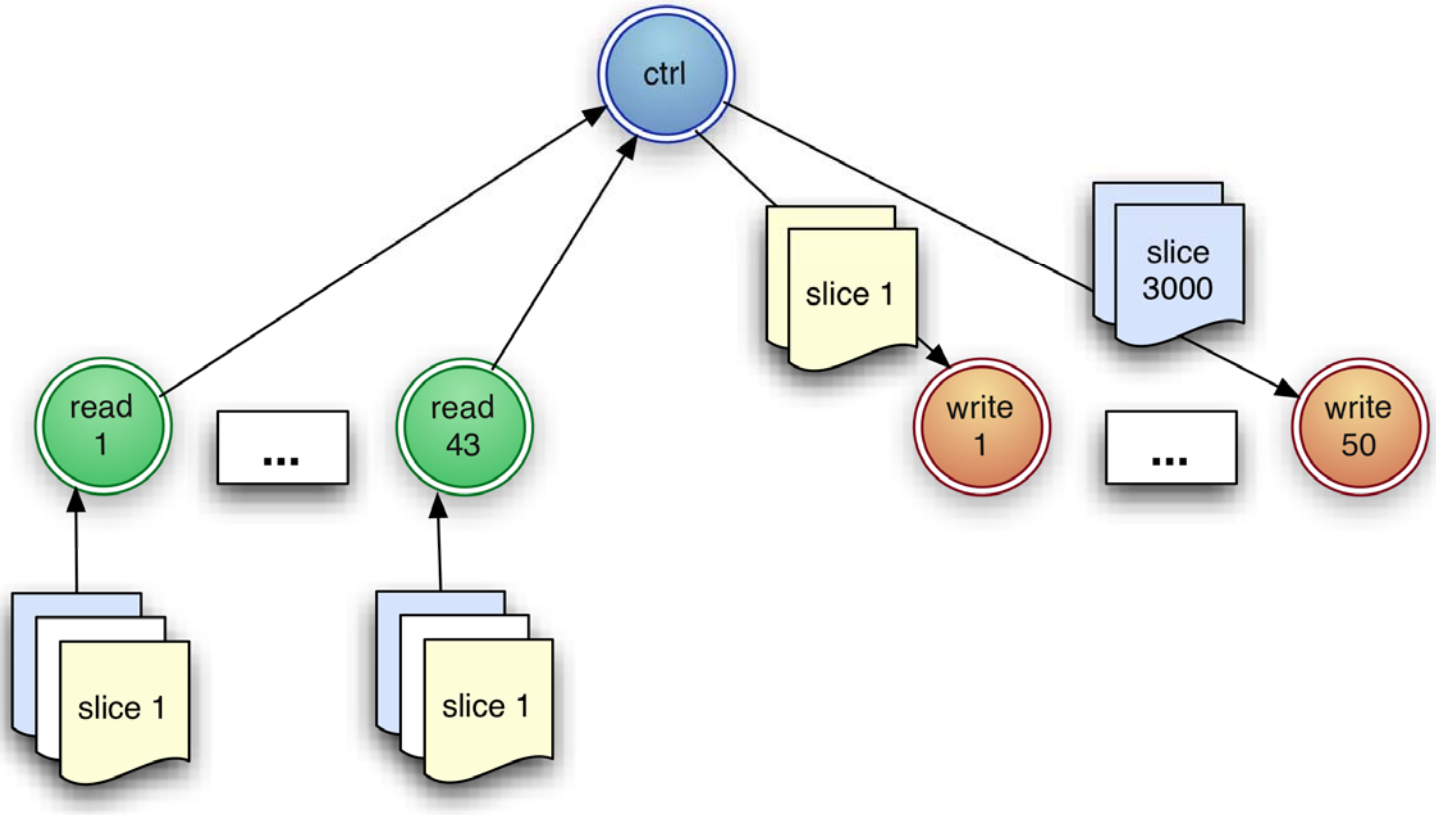
Approach 2: No Interaction

- 43 read/split nodes
- Phase 1: Read nodes split CDX files to 3000 slices



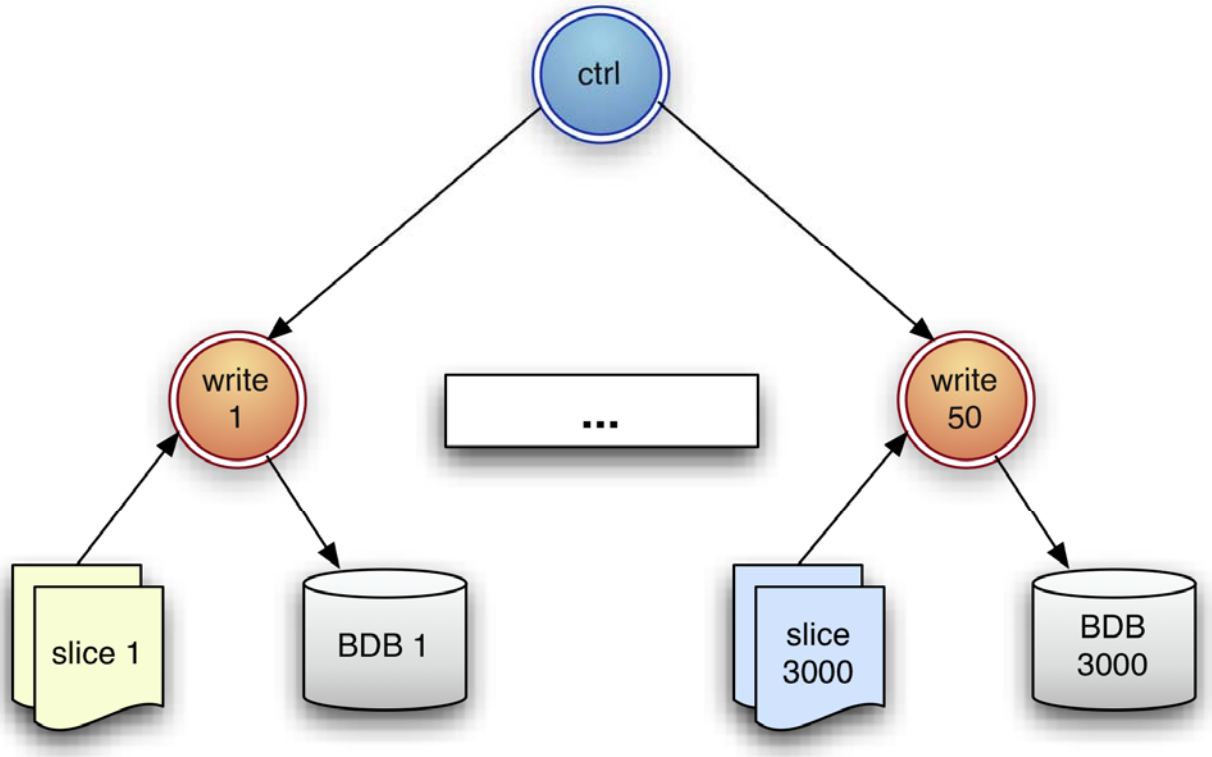
Approach 2: No Interaction

- Phase 2a: Transfer CDX slices to Control node
- Phase 2b: Transfer CDX slices to Write nodes



Approach 2: No Interaction

- 50 write nodes (* 60 slices each = 3000 slices)
- Phase 3: Merge slices from nodes to BerkeleyDBs



Next Steps

- New partners!
  - Please let us have your CDX files :)
- Re-index, using now verified approach
  - All existing and new partners
  - Currently only 1/3 the IA data is indexed
  - Transfer DBs to IIPC to run aggregator service
- Provide online synchronization for new crawls
  - Remote update of the indexes, not CDX transfer
- Distribution analysis, data mining
  - To enable inter-archive crawl prioritization

Memento

http://mementoweb.org/

Herbert Van de Sompel
Robert Sanderson
Michael L. Nelson