Archiving, Indexing and Accessing Web Materials: Solutions for large amounts of data

David Minor
Bing Zhu
Reagan Moore
Charles Cowart

San Diego Supercomputer Center

The explosion of archived Internet materials presents challenges.

This is a growing issue everyone in this industry is facing.

Original methods don’t always work with large amounts of data.

SDSC has needed different solutions for different projects

We've come up with several methods to process our archived web crawls.

Library of Congress Pilot Project
"Building Trust"

National Science Digital Library
NSF Funded Digital Library

Library of Congress / SDSC Pilot Project: Building Trust

Election 2004

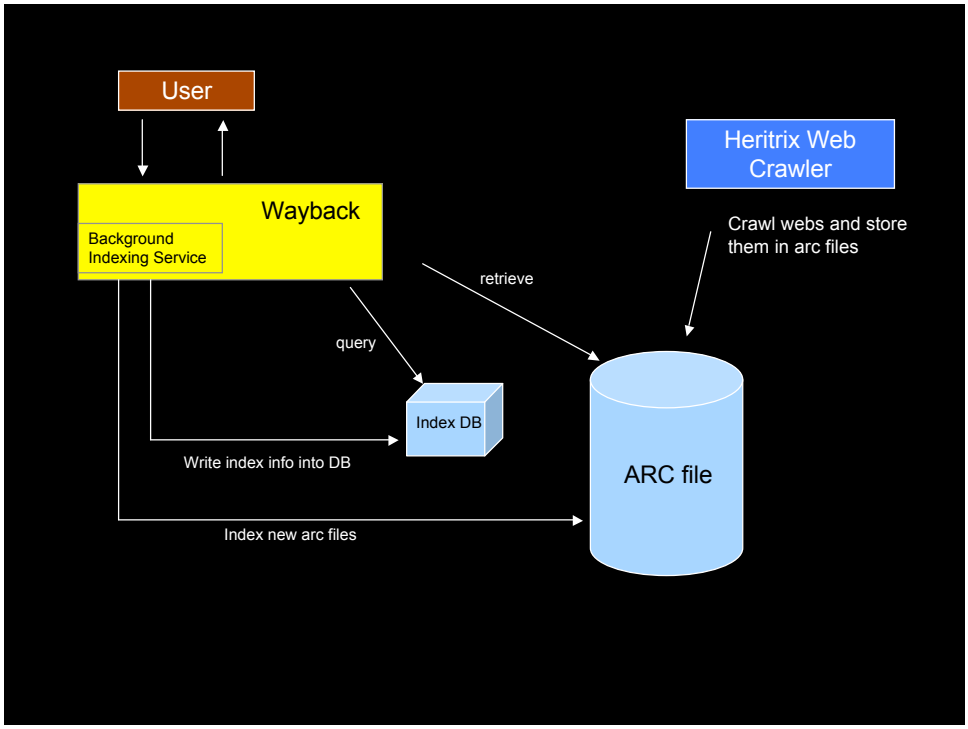
50,000 ARC Files

6 TB of data

Rapid deployment, custom solutions
But still had to "look like" LC environment

Version 1

Wayback 0.6.0
Default configuration.



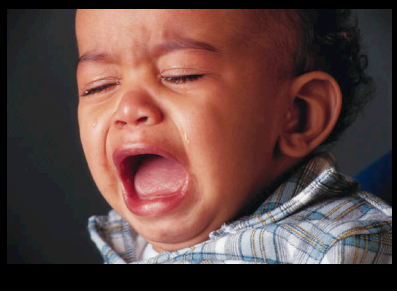
Procedurally this was straightforward

But …

Our indexing rate for a single Wayback instance is about 1000 files per day…

… we would need more than 42 days of constant computing to index entire collection.

This was over
our time budget.

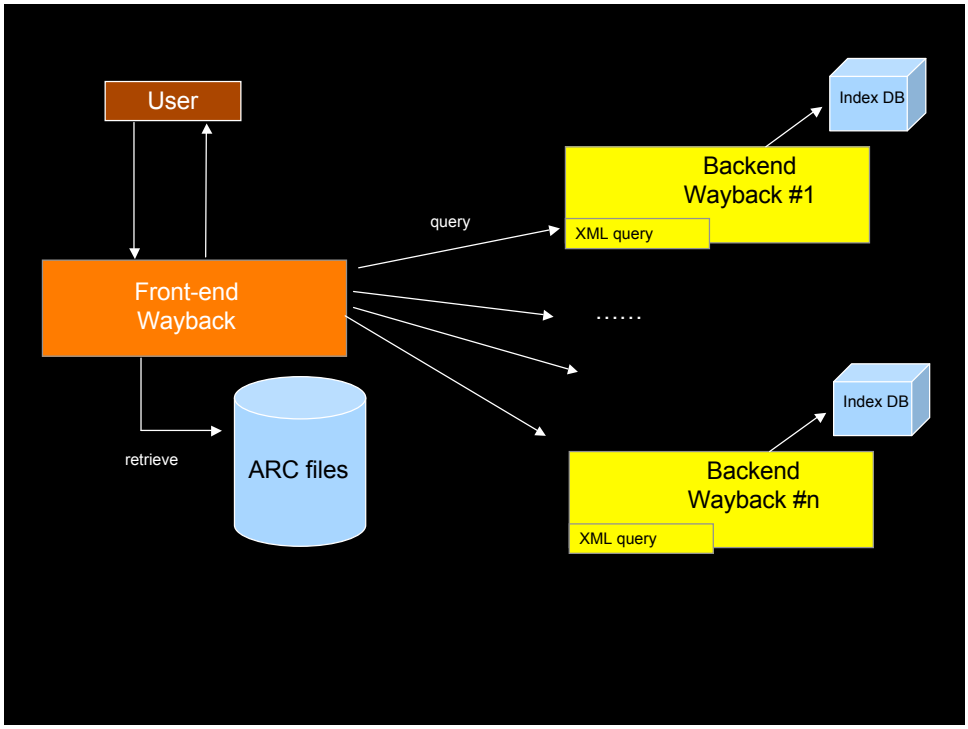


We ran 18 indexing instances – reduced processing to a week

18 Seemed to be a “magic number” for us – any more or less and we lost processing efficiency

Version 2

A master Wayback was developed that virtualizes other Wayback instances as sub-collections.



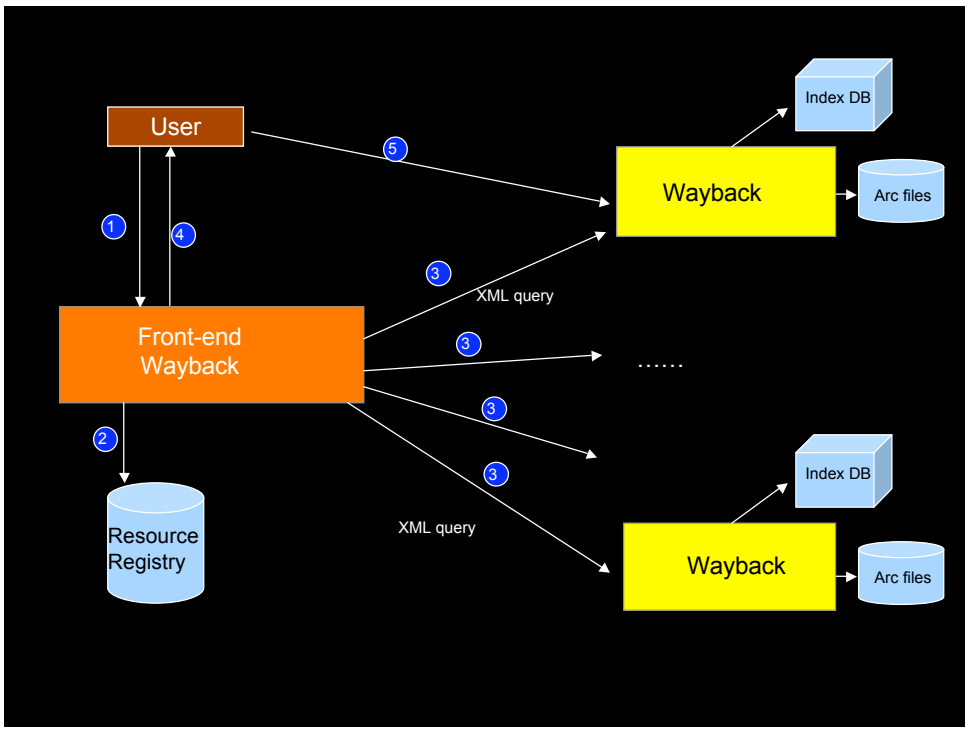
This worked well but generated large amounts of queries.

Some browsers in particular struggled with this solution.

Browser	Test One	Test Two	Test Three	General User Perception
IE	5	10	10	Good
Firefox	129	134	263	Slow

Version 3

To improve the performance, we tried another idea: re-directing page requests to specific backend Wayback instances.



So we ended up with …

50k Files indexed in a week

A system which “acts like” a traditionally-configured system

Total process took 3+ weeks

National Science Digital Library project at SDSC

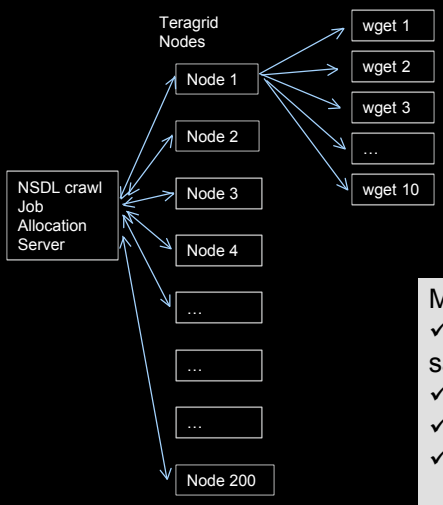
1.65 million URLs in 2007

Each URL is crawled up to a depth of 10.

Use 100-200 compute nodes to conduct parallelized wget web crawls

Created 2TB+ “good” data from the first pass of web crawl

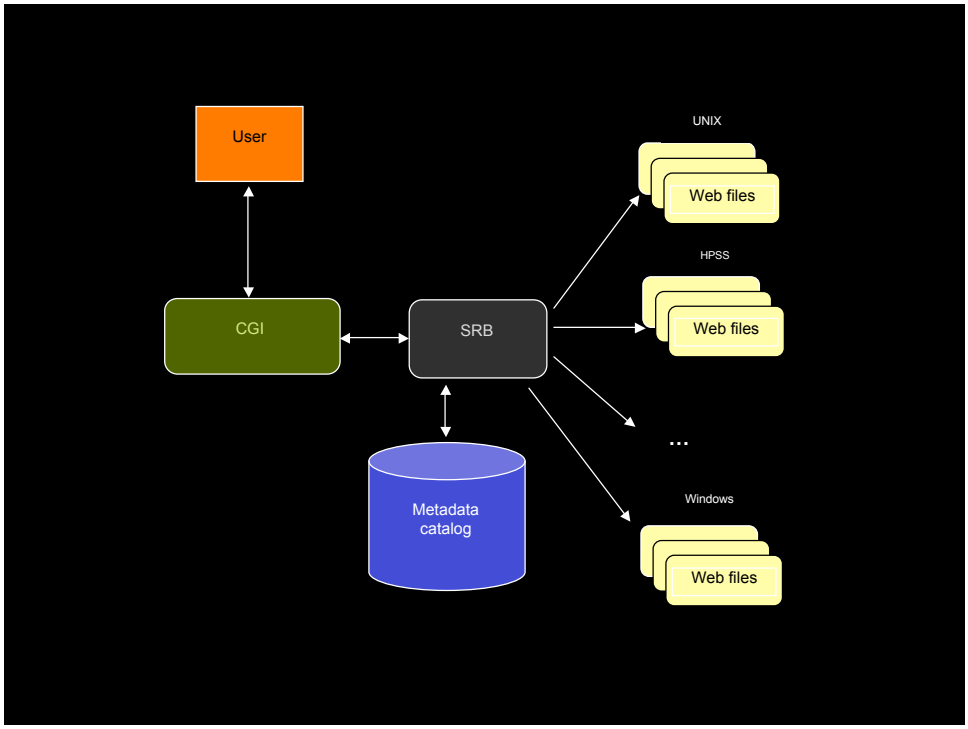
Team uses “wget” to crawl websites.



Main options in using ‘wget’:
- ✓ “-nd”: Web files for a given URL are saved in one directory rather than a tree.
- ✓ “-l 10”: depth 10 crawl.
- ✓ “-k”: convert links to relative path.
- ✓ “-r”: recursive crawl.

The files are stored directly in SRB containers rather than as ARC files

```
C-/home/nsdl.sdsc/2005-01-24T11:04:26Z/FD46621945723C30B88EC6224888F10C
C-/home/nsdl.sdsc/2005-01-24T11:04:26Z/FD4A90EAA247C30E0FFF1A43D786CBDA
C-/home/nsdl.sdsc/2005-01-24T11:04:26Z/FD5996C0321B621C88D483A6564DB7F5
C-/home/nsdl.sdsc/2005-01-24T11:04:26Z/FD90952B0513B0F51E8AA4076270EF25
C-/home/nsdl.sdsc/2005-01-24T11:04:26Z/FD9872E8E50C6959018094EEEA573CDB
C-/home/nsdl.sdsc/2005-01-24T11:04:26Z/FD98CE01D819357416D9F9BF6ACF9925
C-/home/nsdl.sdsc/2005-01-24T11:04:26Z/FDEA26941F9E81A9E85ED8D98E4D8BC1
C-/home/nsdl.sdsc/2005-01-24T11:04:26Z/FDEFDC74180BAB1EB234F9FDF95A1AA2
C-/home/nsdl.sdsc/2005-01-24T11:04:26Z/FE3296F8C77B7283784C08A0A7E8F462
C-/home/nsdl.sdsc/2005-01-24T11:04:26Z/FE74E6752FD0977260744598 9A89F04A
C-/home/nsdl.sdsc/2005-01-24T11:04:26Z/FEAAA0EA0F3A45E5FE78A719AB89CC83
C-/home/nsdl.sdsc/2005-01-24T11:04:26Z/FED6ABD8A445F53F6F91C2300DE539D9
C-/home/nsdl.sdsc/2005-01-24T11:04:26Z/FEEB38FFFD935DB44926441C6D1BF961
C-/home/nsdl.sdsc/2005-01-24T11:04:26Z/FEFCD82EED4100F5AC0B35D21C2599E0
C-/home/nsdl.sdsc/2005-01-24T11:04:26Z/FF079816EECE7CECF75B0A0E3C6E2090
C-/home/nsdl.sdsc/2005-01-24T11:04:26Z/FF362A77A28ADD22BBD71183D7DDC2B0
C-/home/nsdl.sdsc/2005-01-24T11:04:26Z/FF684170916940BCAA188B08CEDF405E
C-/home/nsdl.sdsc/2005-01-24T11:04:26Z/FF7E62ACA0FE2EC41C97D3523923F814
C-/home/nsdl.sdsc/2005-01-24T11:04:26Z/FF81240E118D5E9EF55399420F17D9E2
C-/home/nsdl.sdsc/2005-01-24T11:04:26Z/FFC5957E9775ADEFB06CC678E6AD8945
C-/home/nsdl.sdsc/2005-01-24T11:04:26Z/FFD1DC987C4EABA2A4861EF73DB9C428
C-/home/nsdl.sdsc/2005-01-24T11:04:26Z/FFD3812AD5CCDE4C04C022F4DC2EA207
C-/home/nsdl.sdsc/2005-01-24T11:04:26Z/FFE288D822BA2A10FB817DDC0211FED7
```



Comparison of methods

	Wayback	NSDL@SDSC
Web Crawler	Heritrix	Parallelized 'Wget'
Archive Format	ARC format	SRB Container
Index Data	Inside each ARC file	SRB MCAT
Index Key	URL, Capture Date	URL, Capture Date
Index Store	Berkeley DB, CDX files	Oracle, PostgreSQL, DB2, etc.
Access software	Wayback	CGI script, Windows browser, workflow
Data Compression	Yes	Hardware IDRC

Using SRB as the backend allows for robust collection management.

The SRB supports direct read and write of web pages from the containers.

Containers may be replicated.

Metadata can be linked to each file within the container.

Checksums are managed for each file.

Replicated containers may be synchronized along with validation of the checksums.

Access controls can be assigned to individual web pages.

It is possible to move containers between storage systems without losing either the access controls or links to metadata.

So …

What has SDSC learned for our organization's web archiving projects?

Different scenarios can raise unique issues and demand unique solutions

Sometimes issues are driven by customer demands

Sometimes they are based on resource or time constraints

Sometimes it's just illustrative to try something different!

But, in general

Our organization needs to be flexible in our web-archiving processes

For more information …

David Minor – minor@sdsc.edu

Bing Zhu - bzhu@sdsc.edu
(technical contact)