9-24-2013

Evaluating the SiteStory Transactional Web Archive with the ApacheBench Tool

Justin F. Brunelle
Old Dominion University

Michael L. Nelson
Old Dominion University, mnelson@odu.edu

Lyudmila Balakireva

Robert Sanderson

Herbert Van de Sampel

Evaluating the SiteStory Transactional Web Archive With the ApacheBench Tool

Justin F. Brunelle

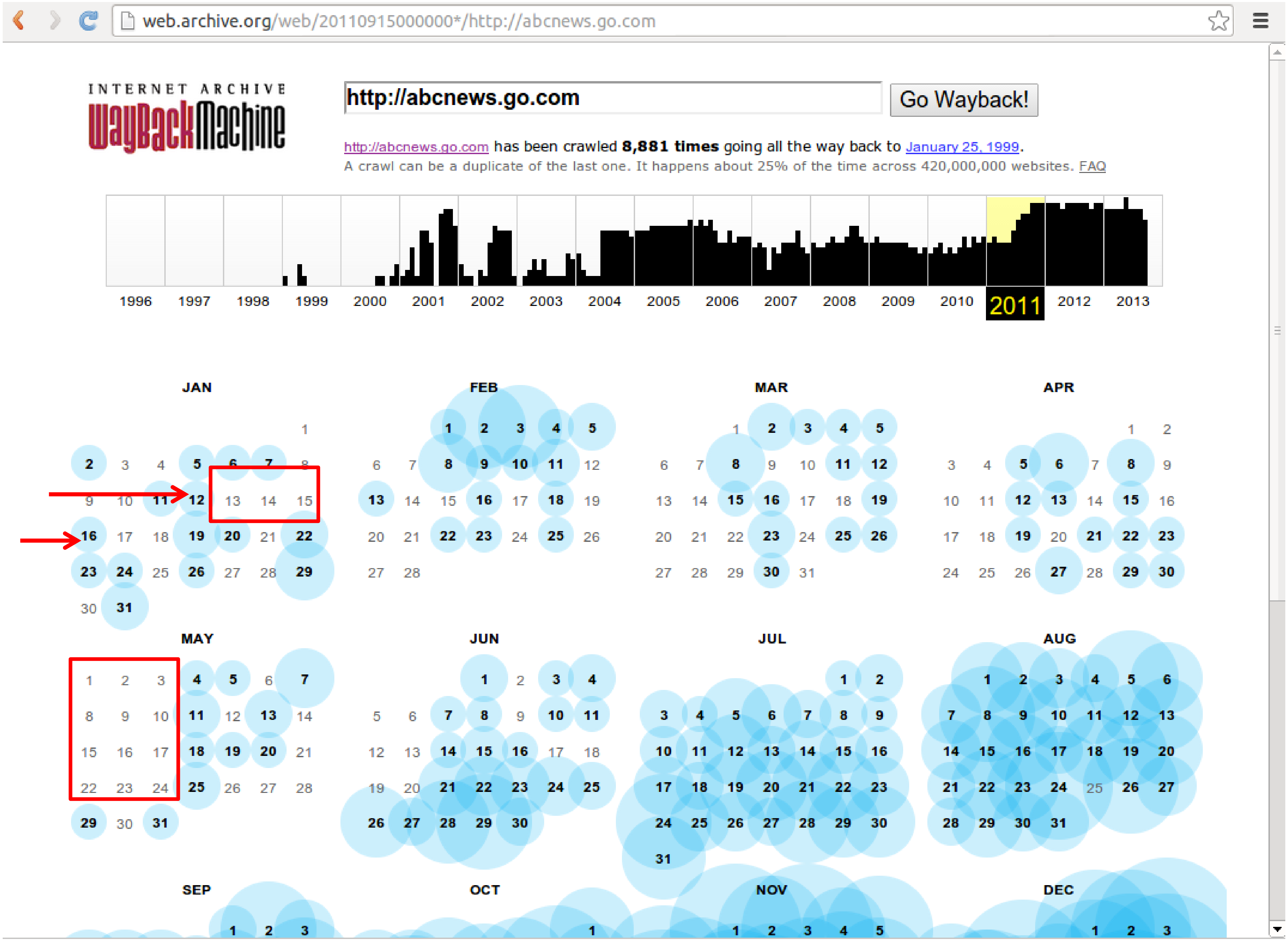
Michael L. Nelson

Lyudmila Balakireva

Robert Sanderson

Herbert Van de Sompel

TPDL 2013, Sept 24 2013



http://abcnews.go.com

http://abcnews.go.com has been crawled **8,881 times** going all the way back to January 25, 1999.

A crawl can be a duplicate of the last one. It happens about 25% of the time across 420,000,000 websites. FAQ

JAN
1
2 3 4 5 6 7 8
9 10 11 12 13 14 15
16 17 18 19 20 21 22
23 24 25 26 27 28 29
30 31

FEB
1 2 3 4 5
6 7 8 9 10 11 12
13 14 15 16 17 18 19
20 21 22 23 24 25 26
27 28

MAR
1 2 3 4 5
6 7 8 9 10 11 12
13 14 15 16 17 18 19
20 21 22 23 24 25 26
27 28 29 30 31

APR
1 2
3 4 5 6 7 8 9
10 11 12 13 14 15 16
17 18 19 20 21 22 23
24 25 26 27 28 29 30

MAY
1 2 3 4 5 6 7
8 9 10 11 12 13 14
15 16 17 18 19 20 21
22 23 24 25 26 27 28
29 30 31

JUN
1 2 3 4
5 6 7 8 9 10 11
12 13 14 15 16 17 18
19 20 21 22 23 24 25
26 27 28 29 30

JUL
1 2
3 4 5 6 7 8 9
10 11 12 13 14 15 16
17 18 19 20 21 22 23
24 25 26 27 28 29 30
31

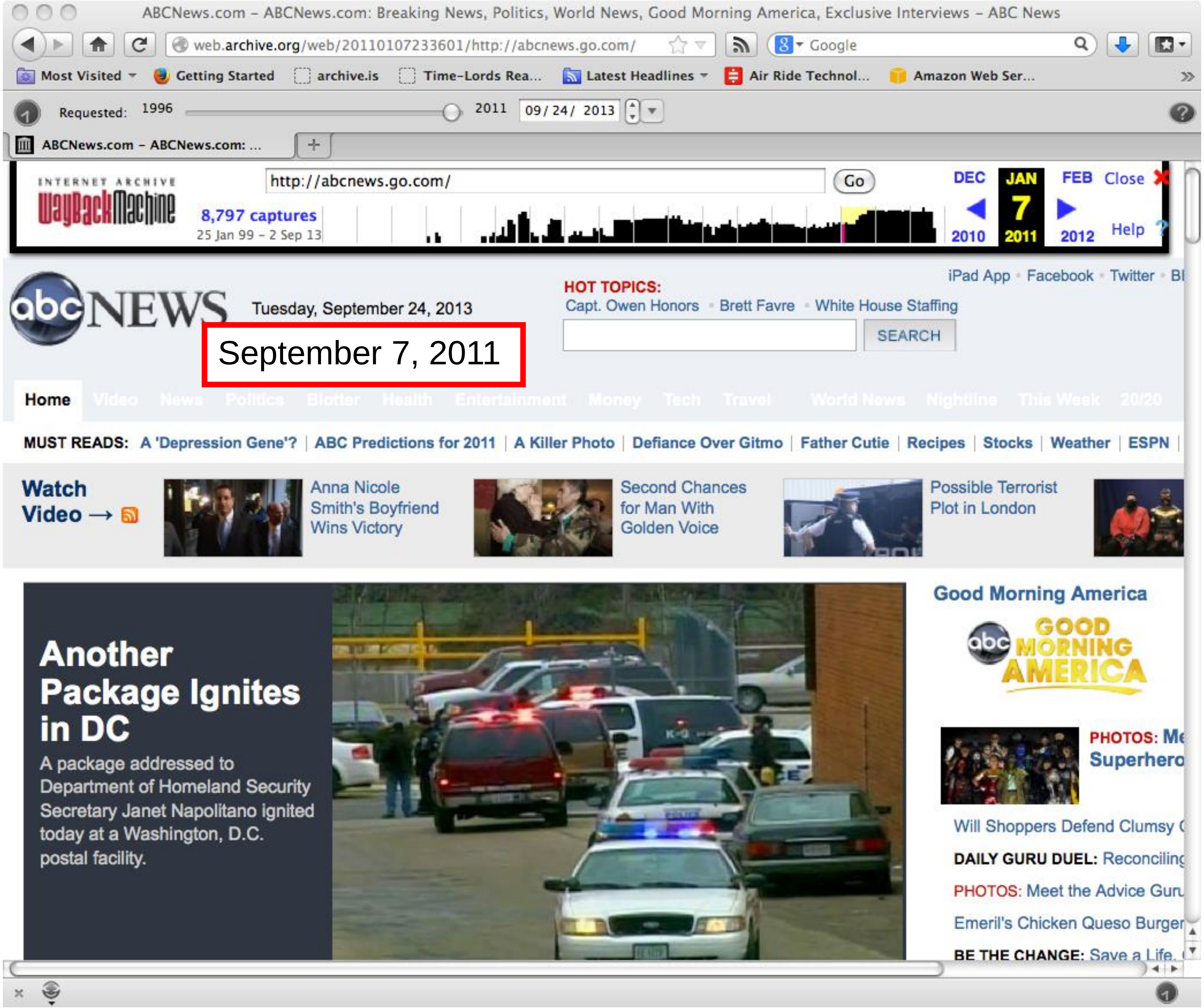
AUG
1 2 3 4 5 6
7 8 9 10 11 12 13
14 15 16 17 18 19 20
21 22 23 24 25 26 27
28 29 30 31

SEP
1 2 3

OCT
1

NOV
1 2 3 4 5

DEC
1 2 3



abc NEWS

Tuesday, September 24, 2013

September 7, 2011

HOT TOPICS: Capt. Owen Honors · Brett Favre · White House Staffing

MUST READS: A 'Depression Gene'? | ABC Predictions for 2011 | A Killer Photo | Defiance Over Gitmo | Father Cutie | Recipes | Stocks | Weather | ESPN |

Watch Video →

Anna Nicole Smith's Boyfriend Wins Victory

Second Chances for Man With Golden Voice

Possible Terrorist Plot in London

Another Package Ignites in DC

A package addressed to Department of Homeland Security Secretary Janet Napolitano ignited today at a Washington, D.C. postal facility.

Good Morning America

GOOD MORNING AMERICA

PHOTOS: Me
Superhero

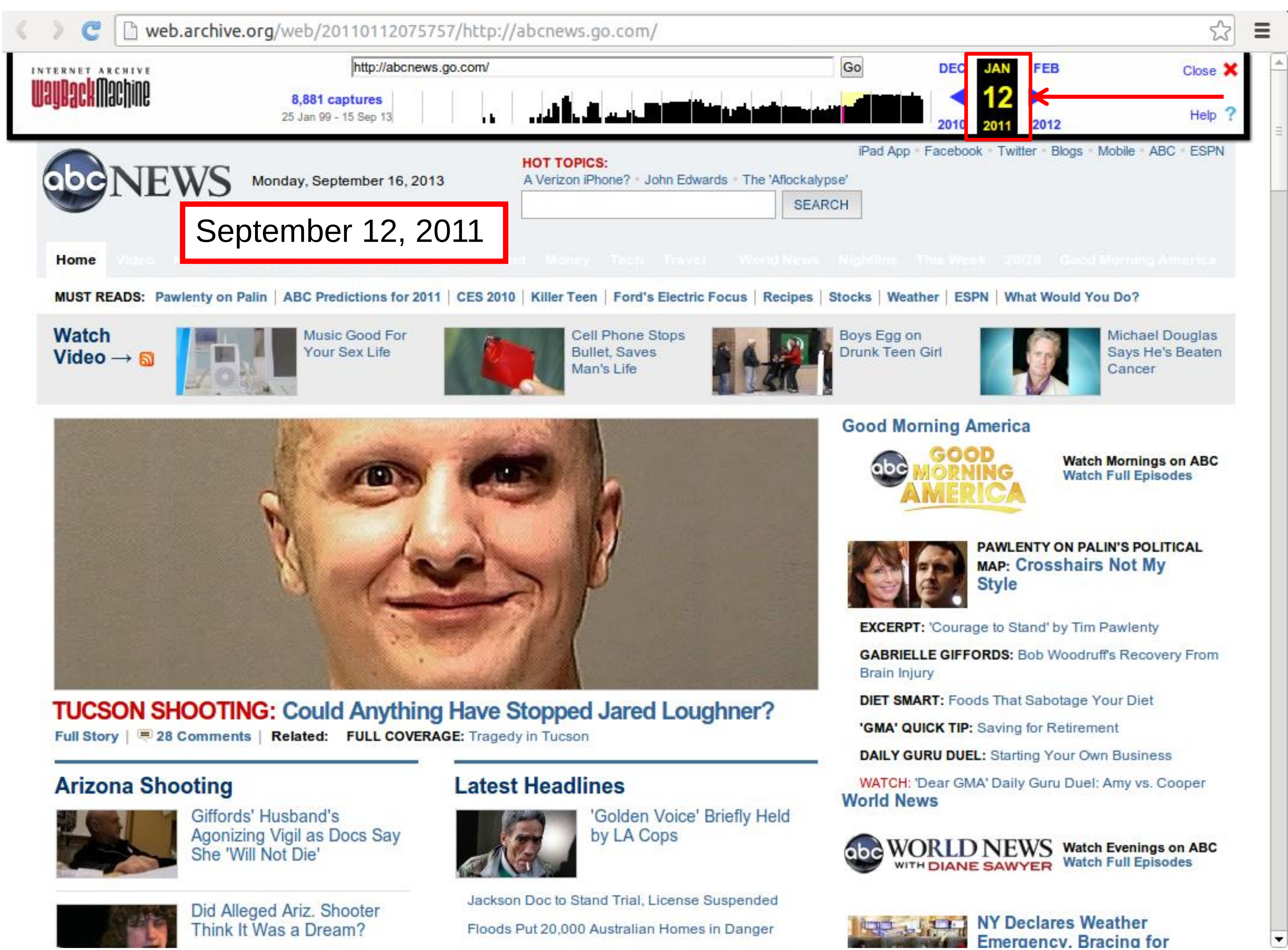
Will Shoppers Defend Clumsy (

DAILY GURU DUEL: Reconciling

PHOTOS: Meet the Advice Guru

Emeril's Chicken Queso Burger

BE THE CHANGE: Save a Life.



web.archive.org/web/20110112075757/http://abcnews.go.com/

ABC NEWS

Monday, September 16, 2013

September 12, 2011

HOT TOPICS: A Verizon iPhone? · John Edwards · The 'Aflockalypse'

SEARCH

iPad App · Facebook · Twitter · Blogs · Mobile · ABC · ESPN

Home Video Money Tech Travel World News Nightline This Week 20/20 Good Morning America

MUST READS: Pawlenty on Palin | ABC Predictions for 2011 | CES 2010 | Killer Teen | Ford's Electric Focus | Recipes | Stocks | Weather | ESPN | What Would You Do?

Watch Video →

Music Good For Your Sex Life

Cell Phone Stops Bullet, Saves Man's Life

Boys Egg on Drunk Teen Girl

Michael Douglas Says He's Beaten Cancer

TUCSON SHOOTING: Could Anything Have Stopped Jared Loughner?

Full Story | 28 Comments | Related: FULL COVERAGE: Tragedy in Tucson

Arizona Shooting

Giffords' Husband's Agonizing Vigil as Docs Say She 'Will Not Die'

Did Alleged Ariz. Shooter Think It Was a Dream?

Latest Headlines

'Golden Voice' Briefly Held by LA Cops

Jackson Doc to Stand Trial, License Suspended

Floods Put 20,000 Australian Homes in Danger

Good Morning America

GOOD MORNING AMERICA

Watch Mornings on ABC
Watch Full Episodes

PAWLENTY ON PALIN'S POLITICAL MAP: Crosshairs Not My Style

EXCERPT: 'Courage to Stand' by Tim Pawlenty

GABRIELLE GIFFORDS: Bob Woodruff's Recovery From Brain Injury

DIET SMART: Foods That Sabotage Your Diet

'GMA' QUICK TIP: Saving for Retirement

DAILY GURU DUEL: Starting Your Own Business

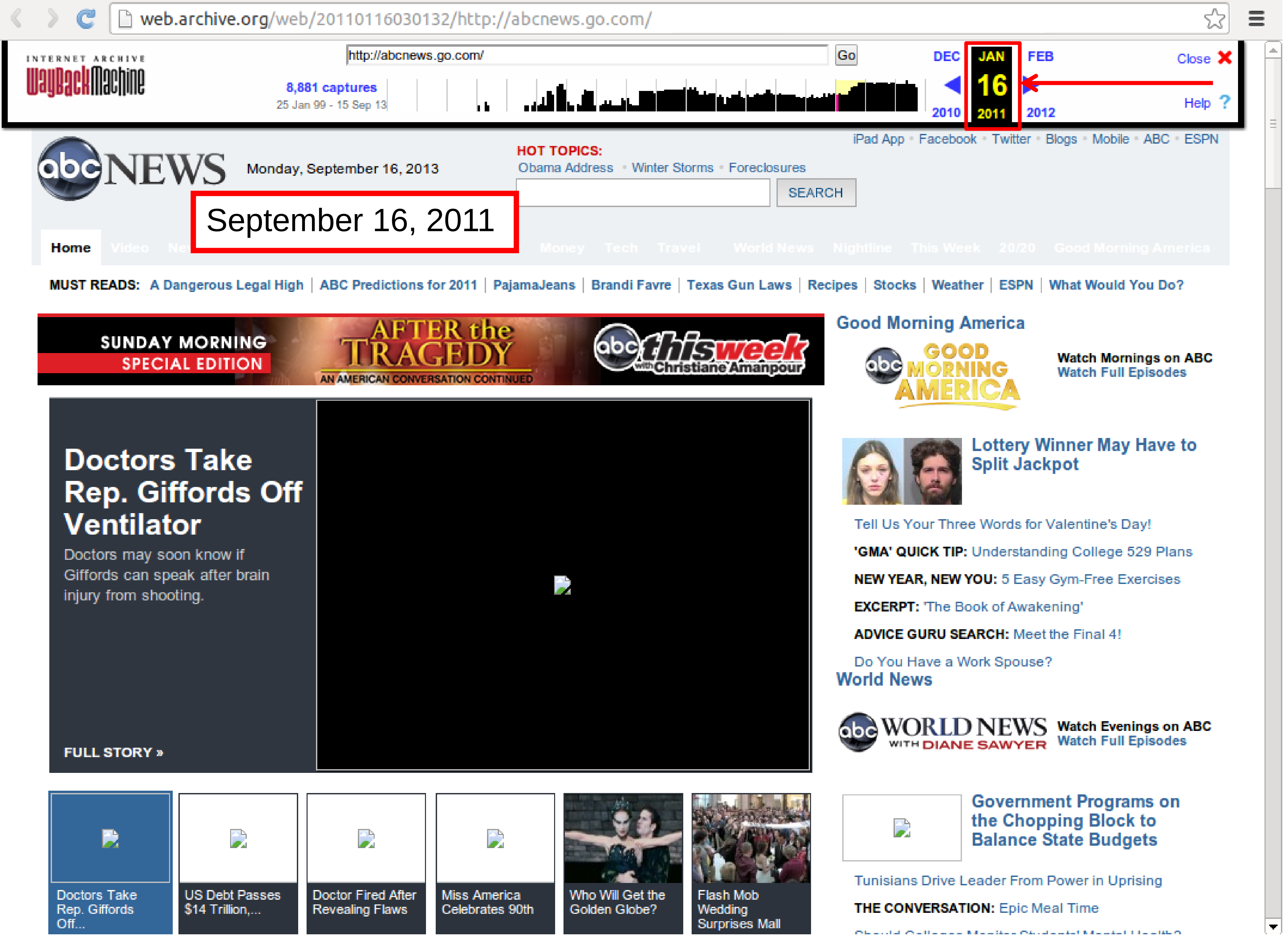
WATCH: 'Dear GMA' Daily Guru Duel: Amy vs. Cooper

World News

WORLD NEWS WITH DIANE SAWYER

Watch Evenings on ABC
Watch Full Episodes

NY Declares Weather Emergency, Bracing for



web.archive.org/web/20110116030132/http://abcnews.go.com/

INTERNET ARCHIVE WaybackMachine

http://abcnews.go.com/ Go

8,881 captures
25 Jan 99 - 15 Sep 13

DEC 2010 | JAN 16 2011 | FEB 2012

Close ✖
Help ?

iPad App • Facebook • Twitter • Blogs • Mobile • ABC • ESPN

abc NEWS Monday, September 16, 2013

HOT TOPICS:
Obama Address • Winter Storms • Foreclosures

SEARCH

September 16, 2011

Home Video Ne... Money Tech Travel World News Nightline This Week 20/20 Good Morning America

MUST READS: A Dangerous Legal High | ABC Predictions for 2011 | PajamaJeans | Brandi Favre | Texas Gun Laws | Recipes | Stocks | Weather | ESPN | What Would You Do?

SUNDAY MORNING SPECIAL EDITION
AFTER the TRAGEDY
AN AMERICAN CONVERSATION CONTINUED
abc thisweek with Christiane Amanpour

Doctors Take Rep. Giffords Off Ventilator

Doctors may soon know if Giffords can speak after brain injury from shooting.

FULL STORY »

Doctors Take Rep. Giffords Off...
US Debt Passes $14 Trillion,...
Doctor Fired After Revealing Flaws
Miss America Celebrates 90th
Who Will Get the Golden Globe?
Flash Mob Wedding Surprises Mall

Good Morning America

abc GOOD MORNING AMERICA

Watch Mornings on ABC
Watch Full Episodes

Lottery Winner May Have to Split Jackpot

Tell Us Your Three Words for Valentine's Day!

'GMA' QUICK TIP: Understanding College 529 Plans

NEW YEAR, NEW YOU: 5 Easy Gym-Free Exercises

EXCERPT: 'The Book of Awakening'

ADVICE GURU SEARCH: Meet the Final 4!

Do You Have a Work Spouse?

World News

abc WORLD NEWS WITH DIANE SAWYER

Watch Evenings on ABC
Watch Full Episodes

Government Programs on the Chopping Block to Balance State Budgets

Tunisians Drive Leader From Power in Uprising

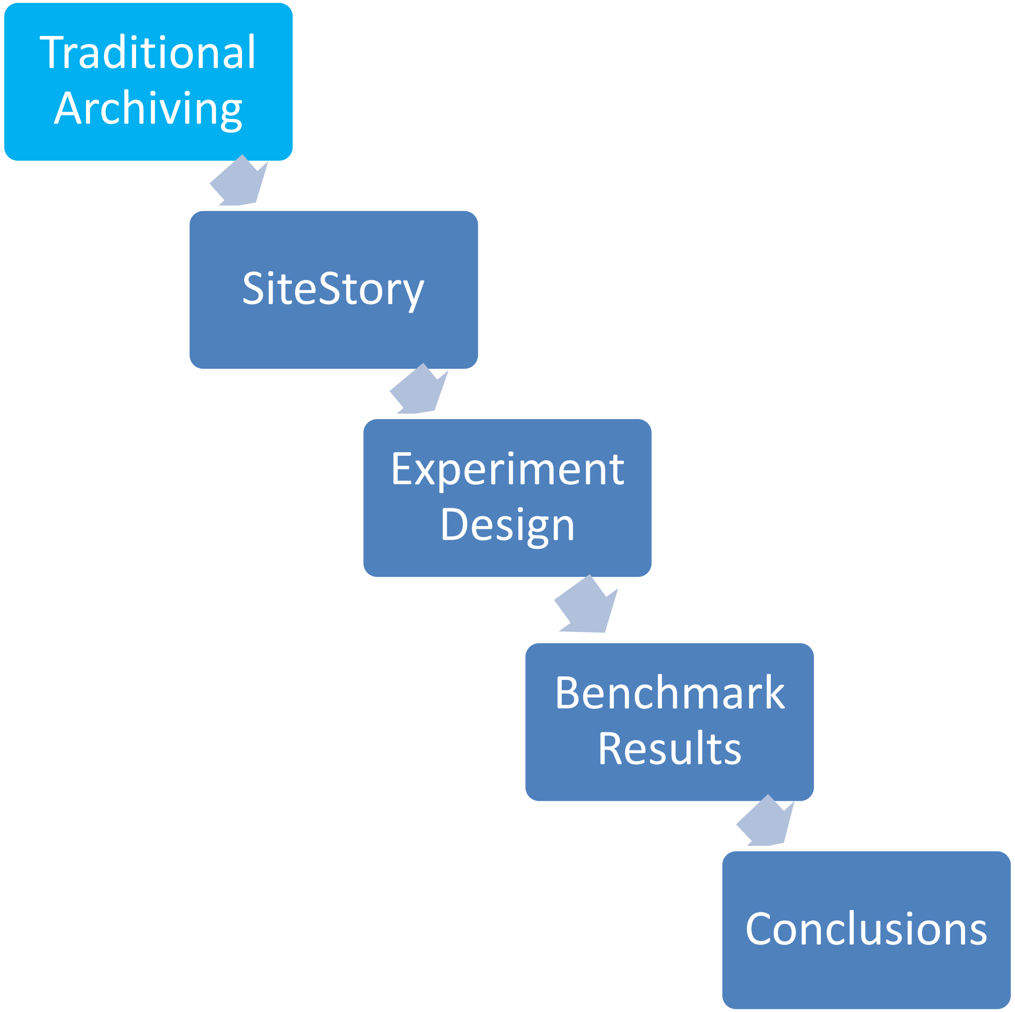
THE CONVERSATION: Epic Meal Time

Should Colleges Monitor Students' Mental Health?

Problem

- People view ABC News all the time
- No mementos for "all the time"
  - Stories missing or incomplete
- Possible solutions:
  - archive.org: crawl more often (how often is often enough?)
  - abcnews.com: install a ***Transactional Web Archive***

Agenda



Traditional Archiving

SiteStory

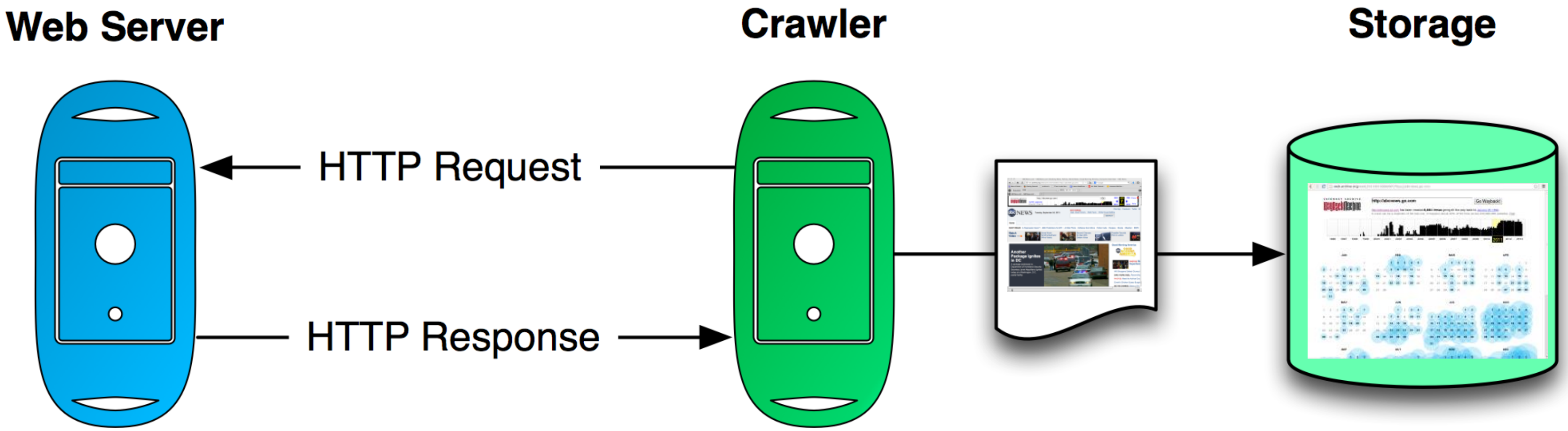
Experiment Design

Benchmark Results

Conclusions

Traditional Web Archiving

- Active crawling
- Heritrix



Web Server

Crawler

Storage

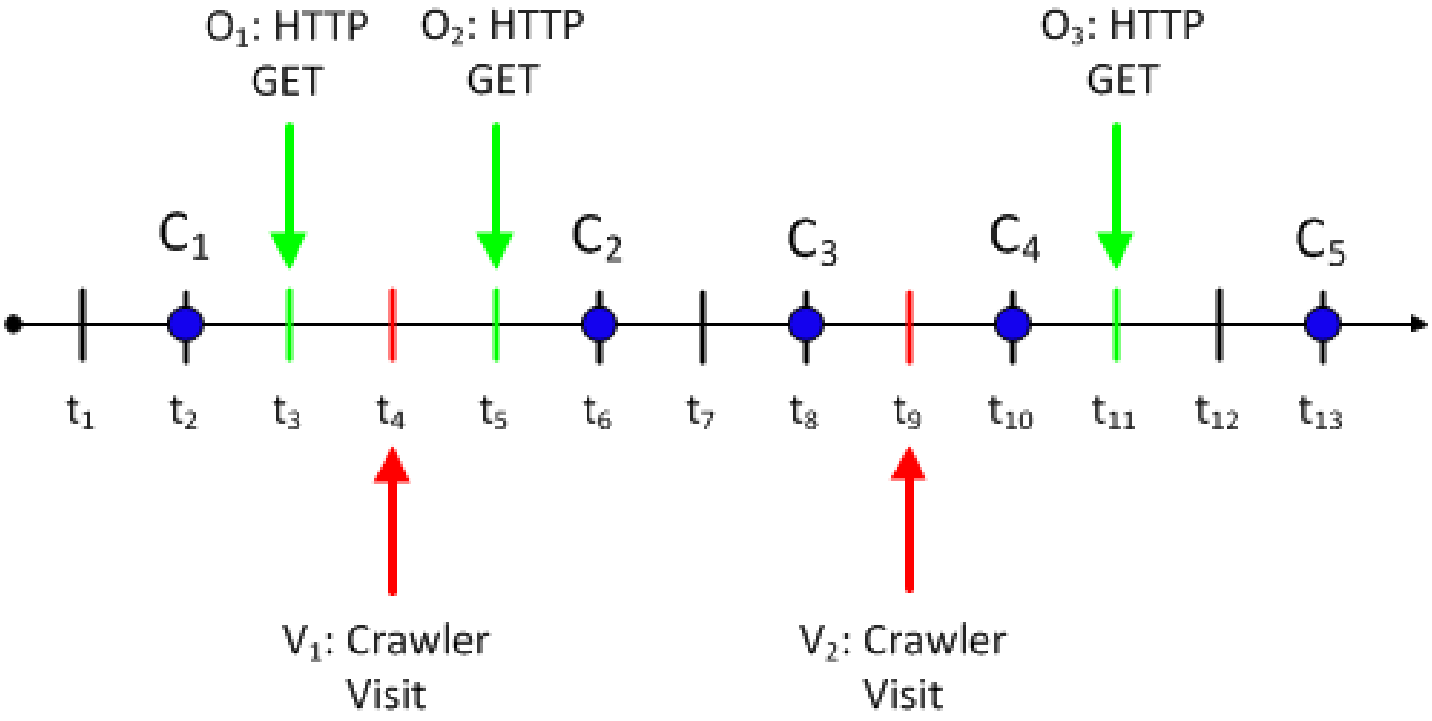
HTTP Request

HTTP Response

Issues with Traditional Web Archiving

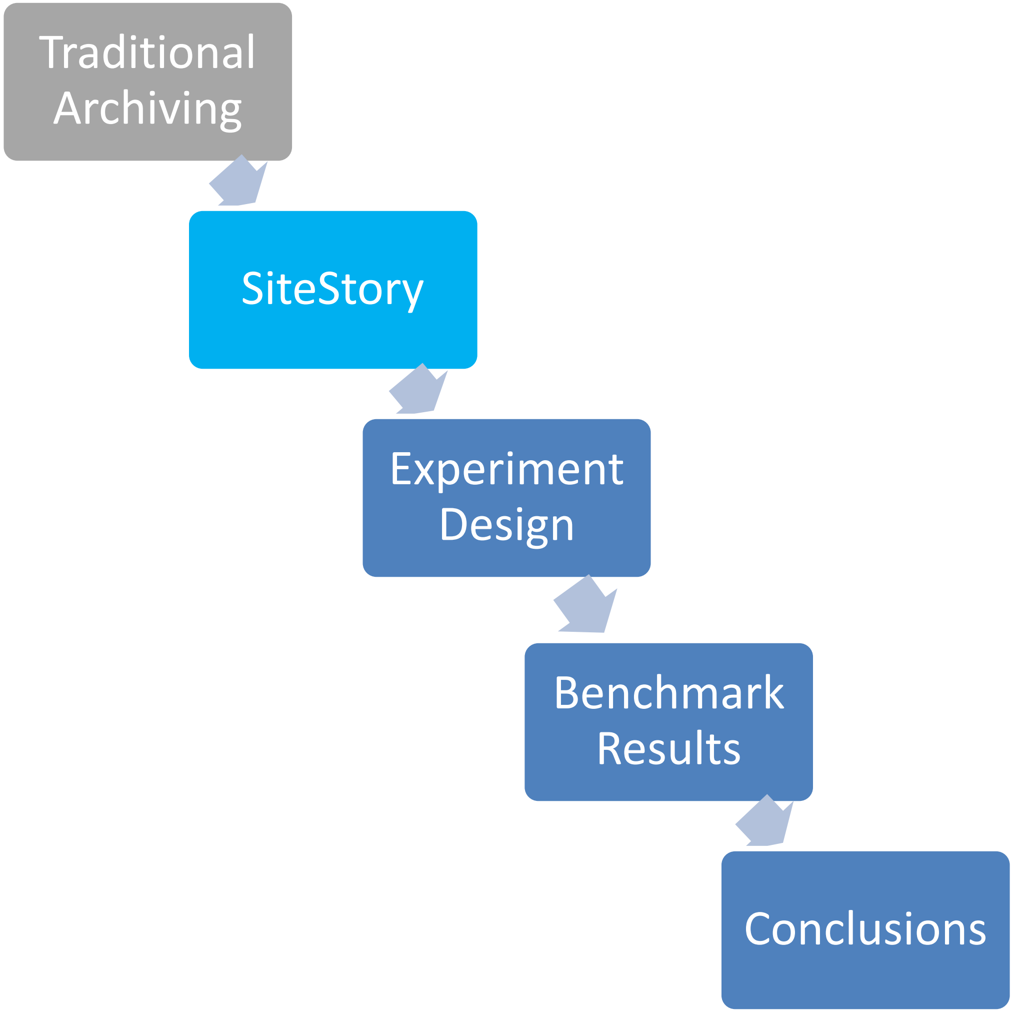
- Request can be rejected (robots.txt, user-agent, IP)
- Can be deceived (geo-location, user-agent)
- Can be trapped (crawl my calendar!)
- Resource-intense (bandwidth)
- Recrawl vs. change-rate

Missed Updates



seen by humans: C_1, C_3, C_4; archived by crawler: C_1, C_3

Agenda



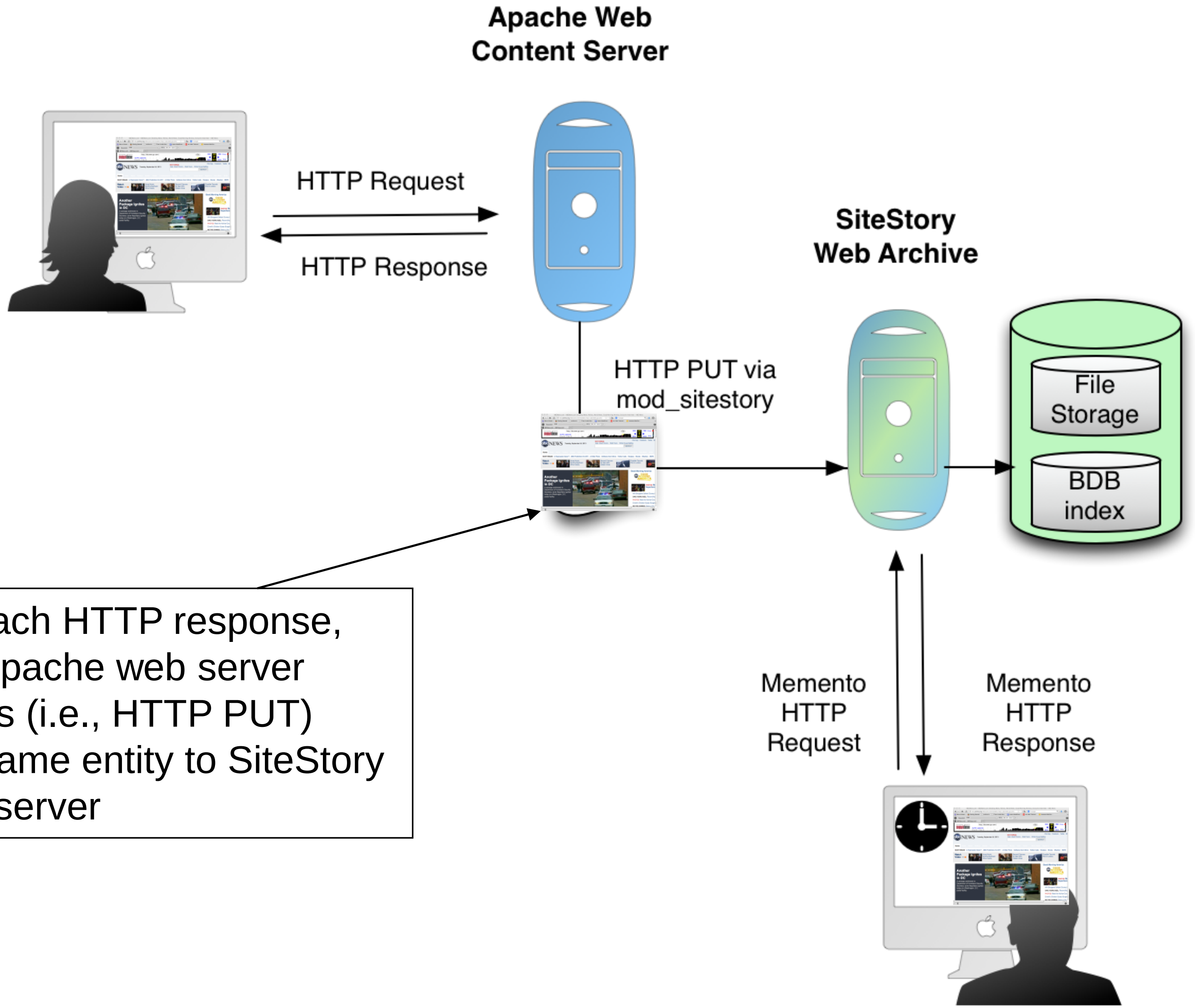
Traditional Archiving

SiteStory

Experiment Design

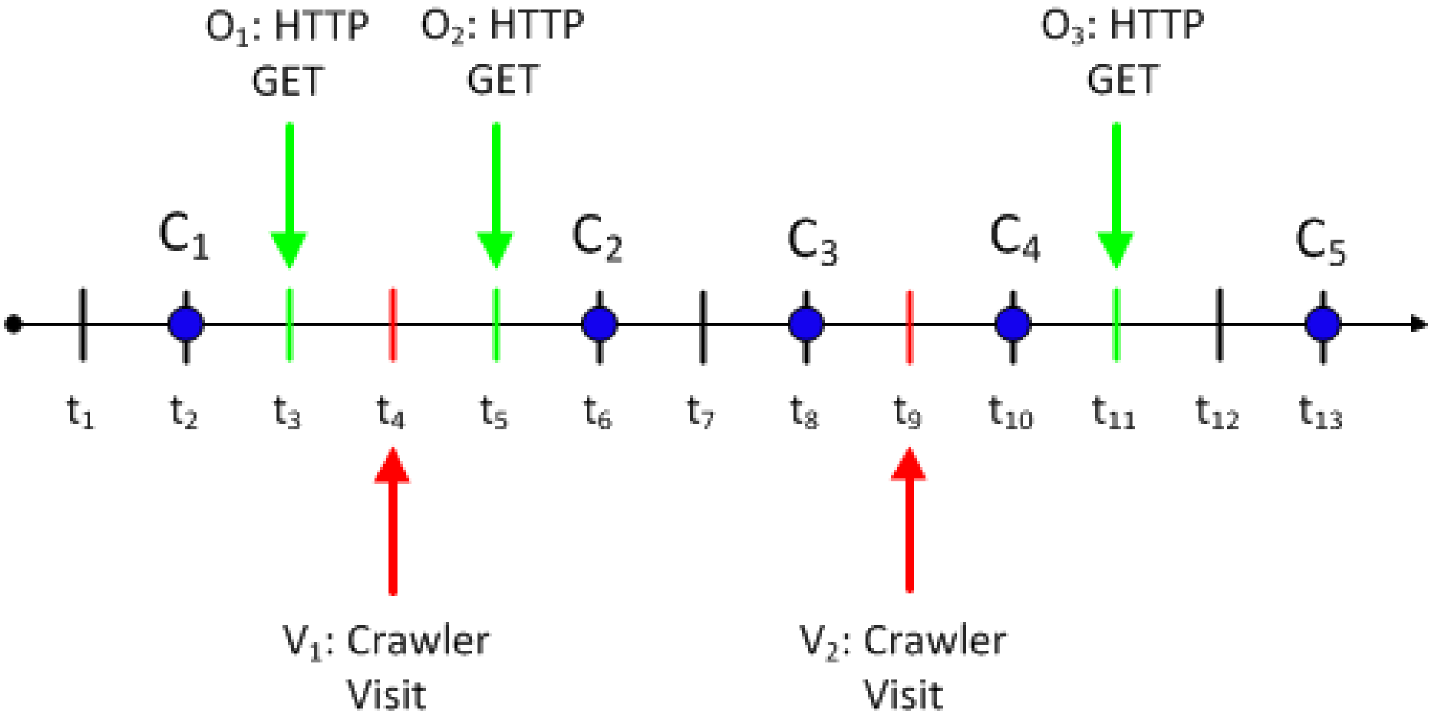
Benchmark Results

Conclusions



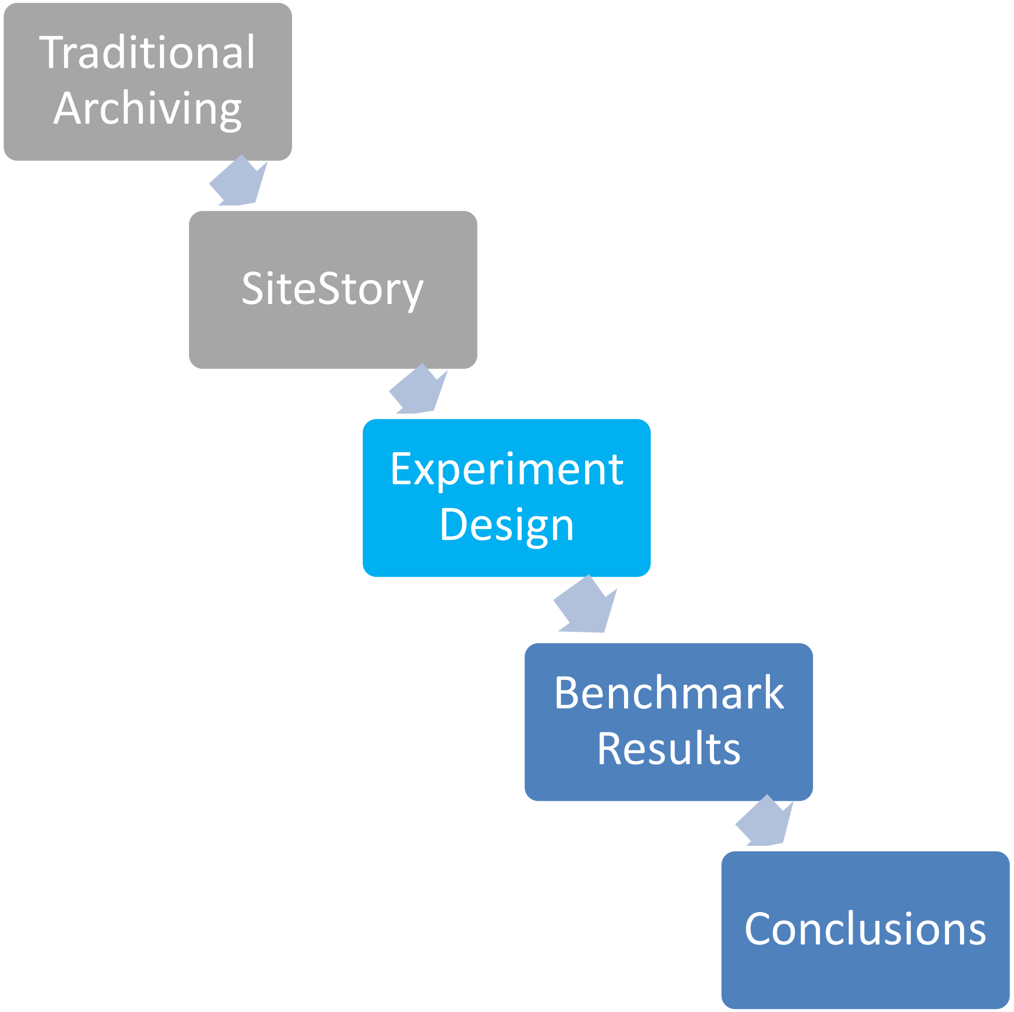
for each HTTP response,
the Apache web server
sends (i.e., HTTP PUT)
the same entity to SiteStory
web server

Now we have them all



seen by humans: C_1, C_3, C_4; archived by transactional archive: C_1, $C_{3,}$ C_4

Agenda



Traditional Archiving

SiteStory

Experiment Design

Benchmark Results

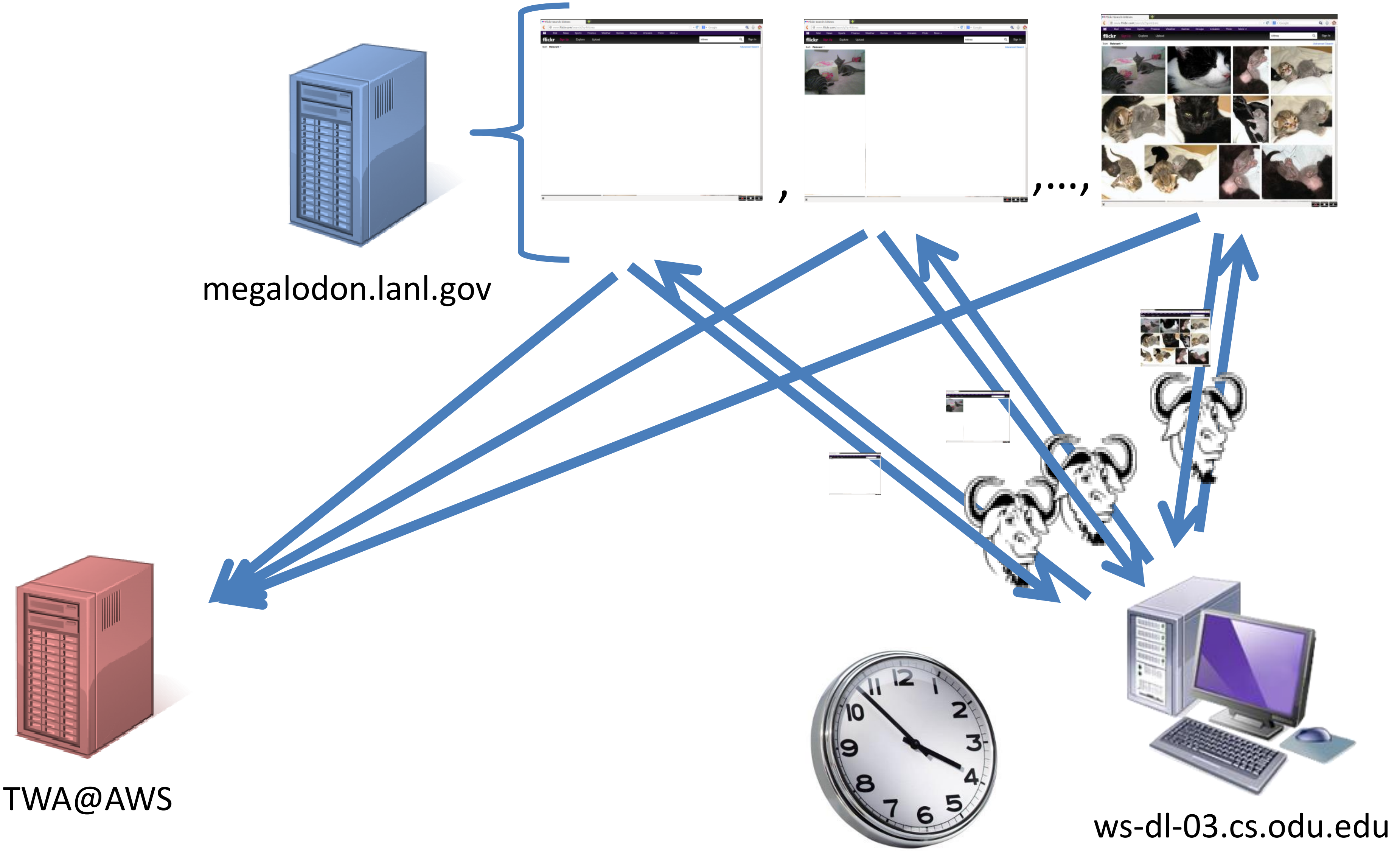
Conclusions

Benchmark with ab

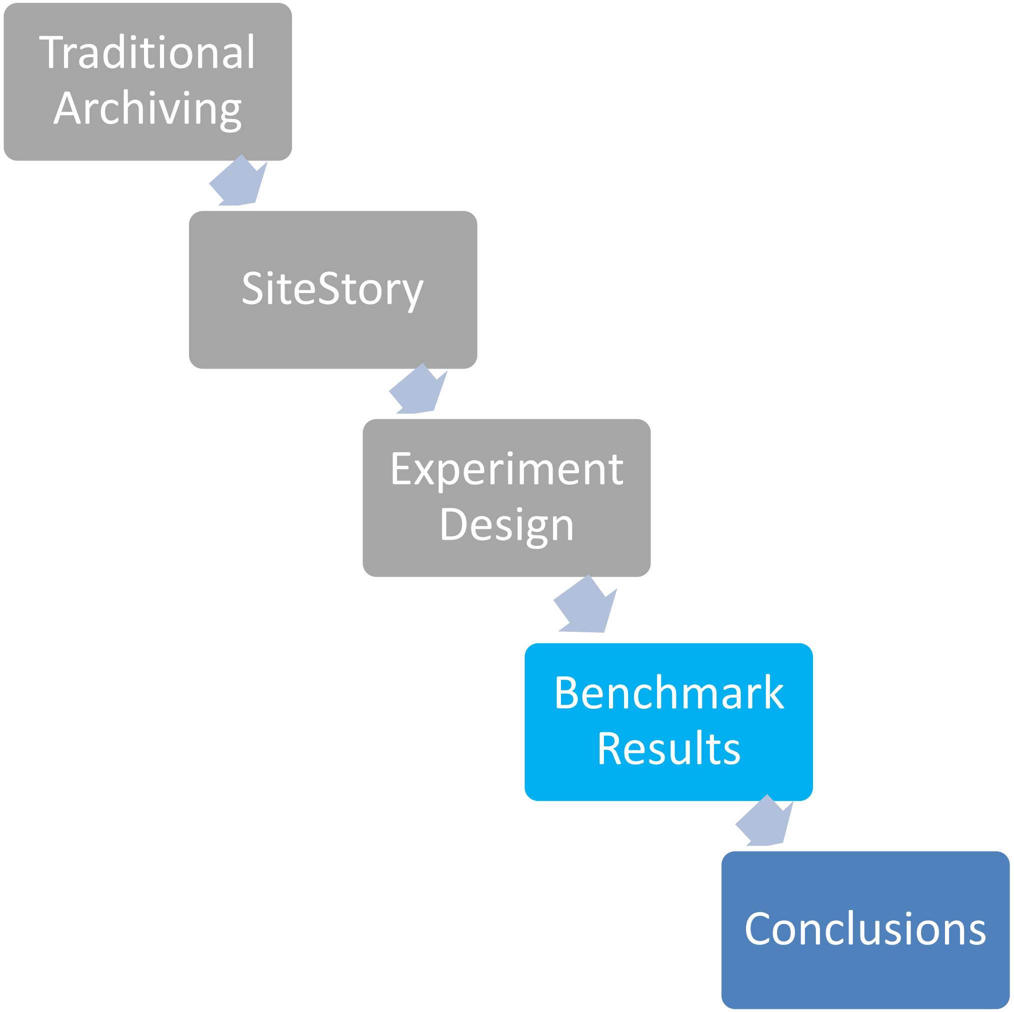
- ApacheBench: ab
  - -n [Number of Connections]
  - -c [Concurrency]
- Benchmarked with SiteStory on & off

x99

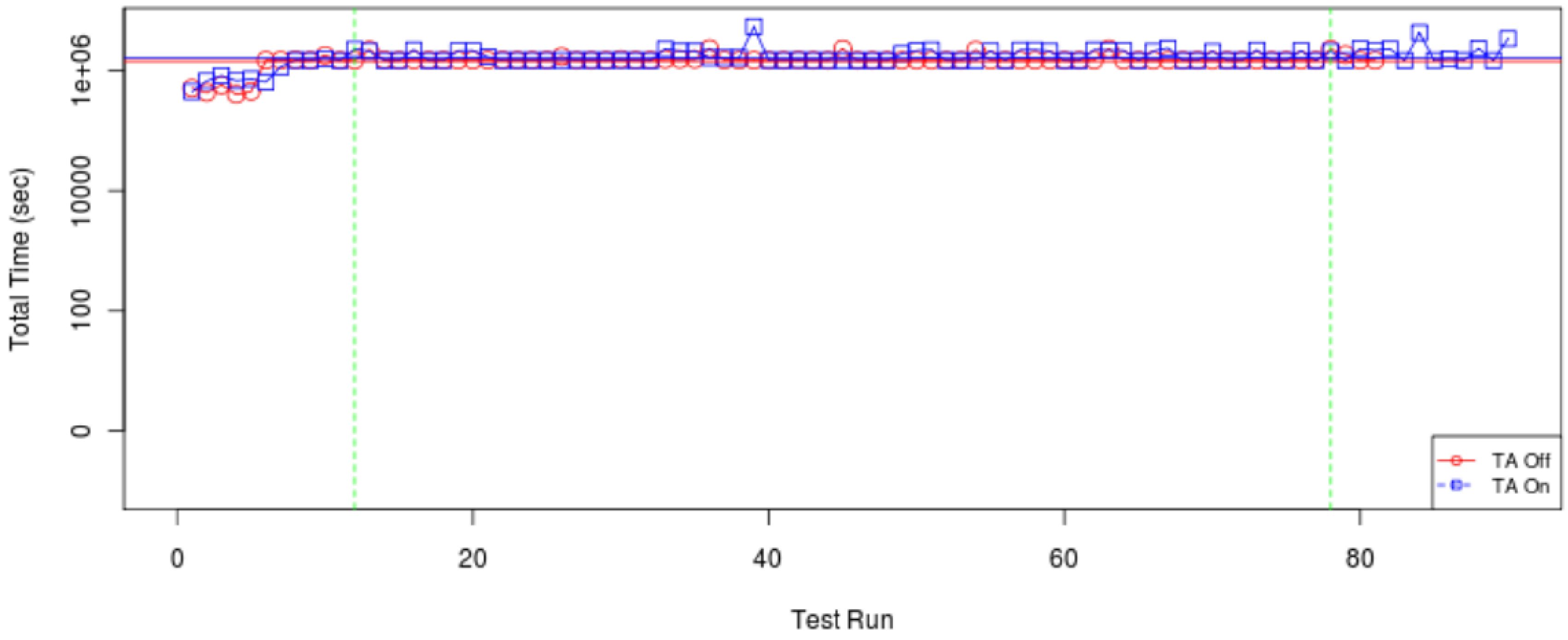
Benchmark with wget



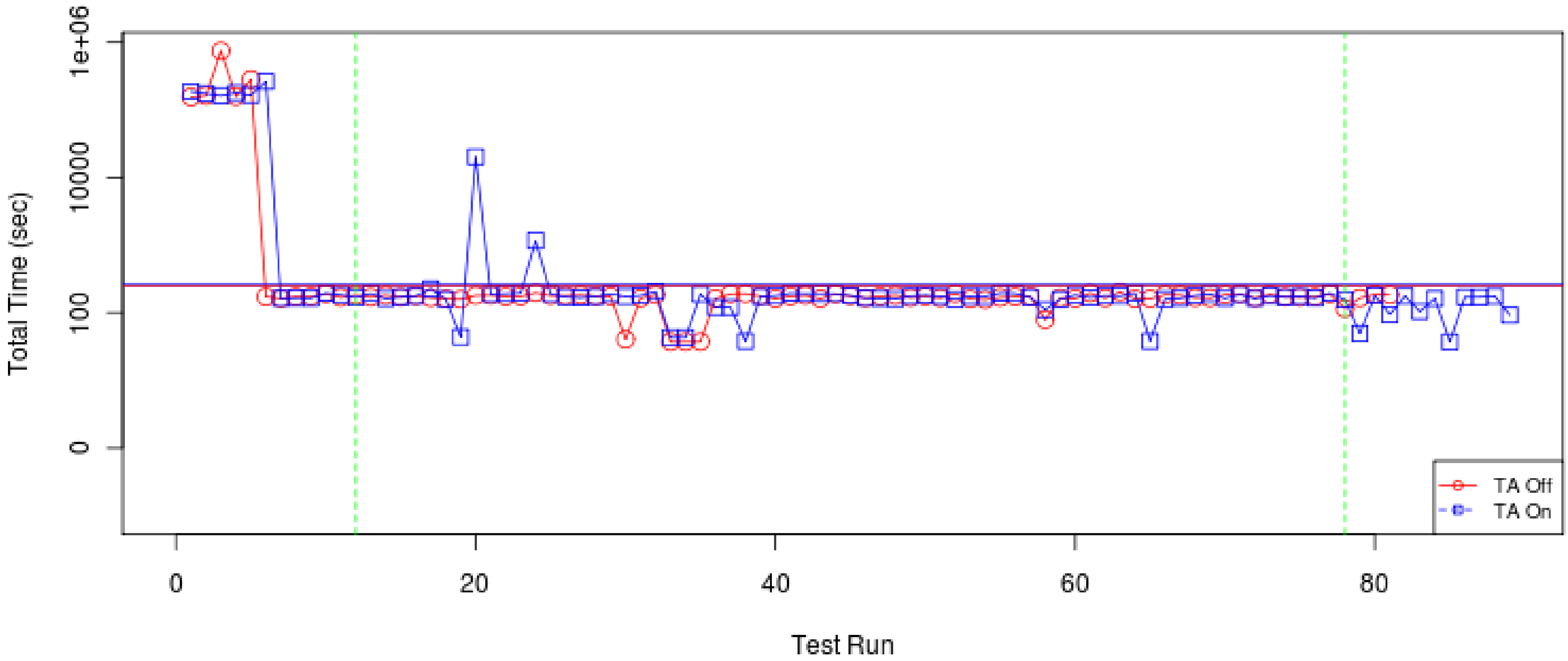
Agenda


Traditional Archiving
SiteStory
Experiment Design
Benchmark Results
Conclusions

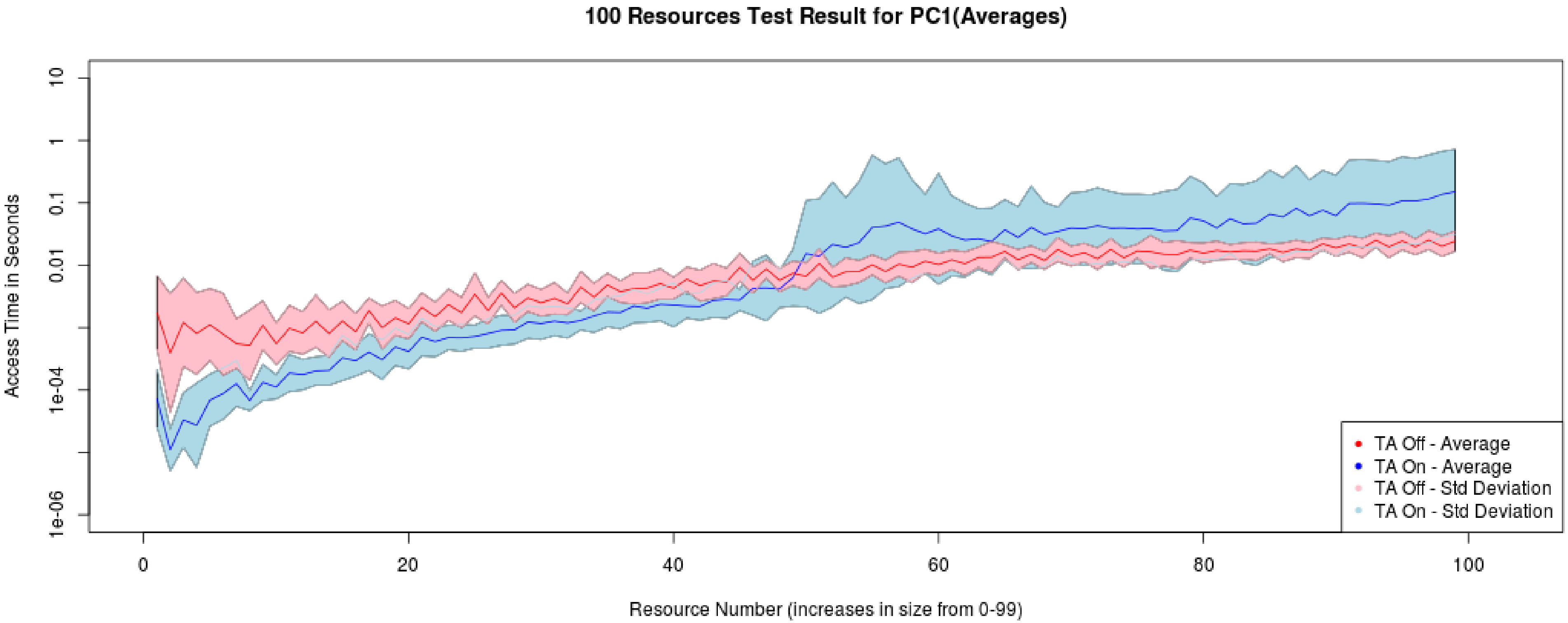
Testing LAN with ab



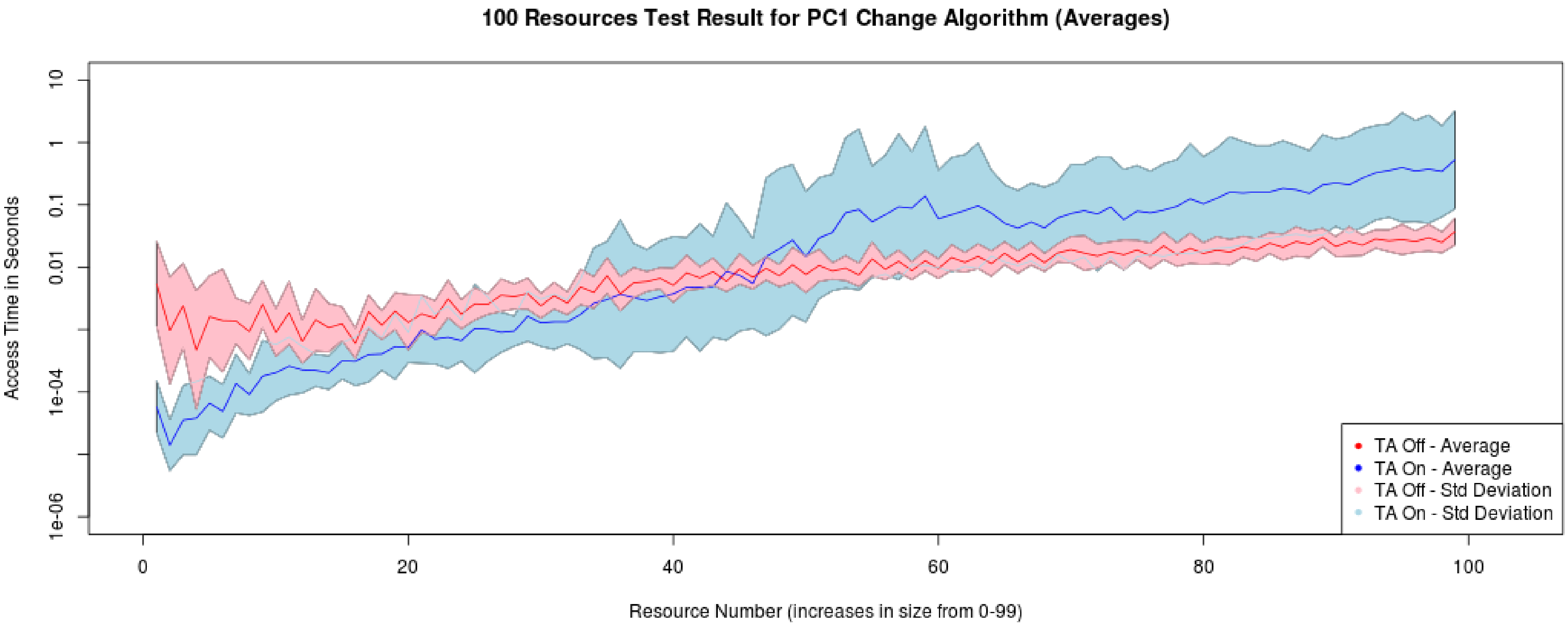
Testing LAN with ab



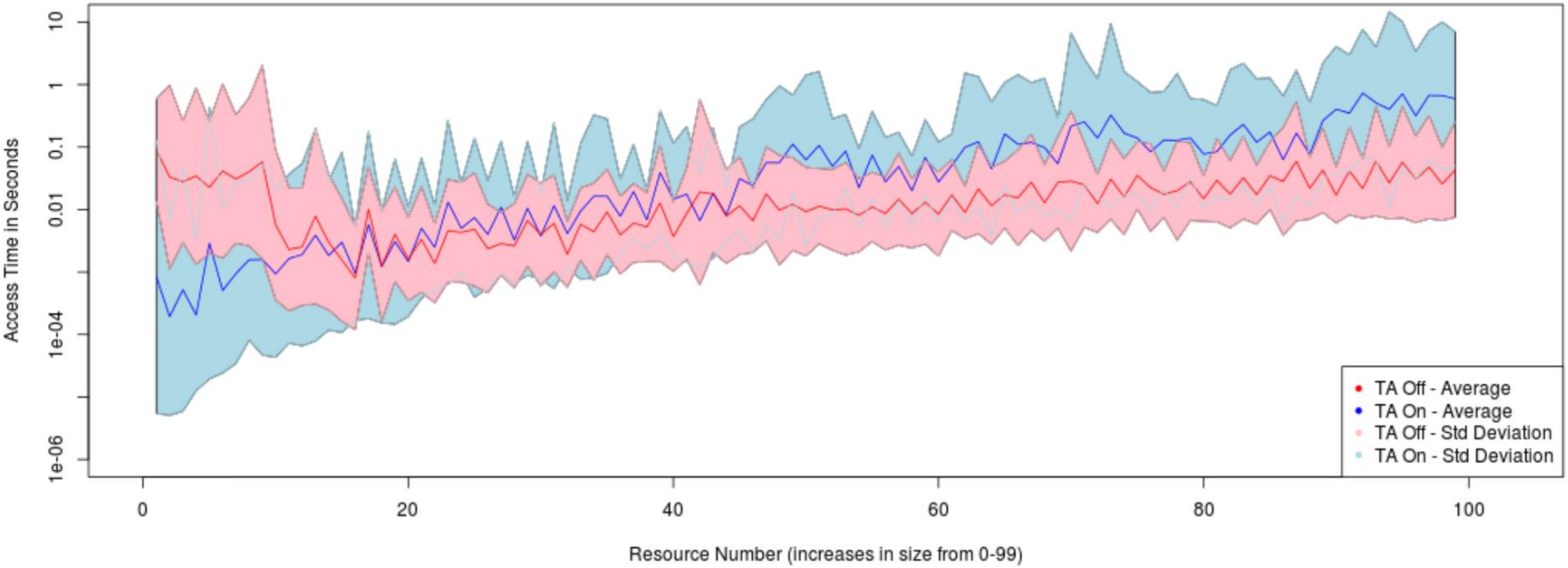
Benchmark with wget (unburdened)



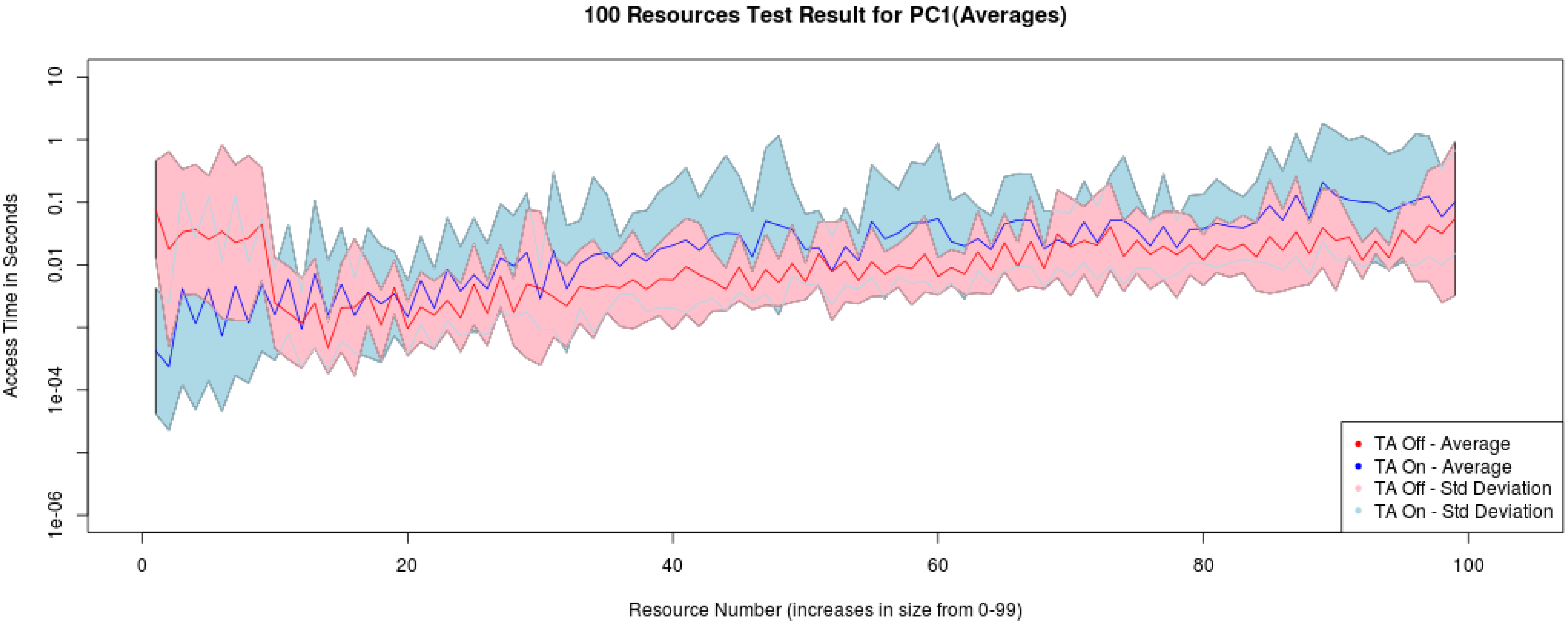
Benchmark with wget (unburdened)



Benchmark with wget (burdened)



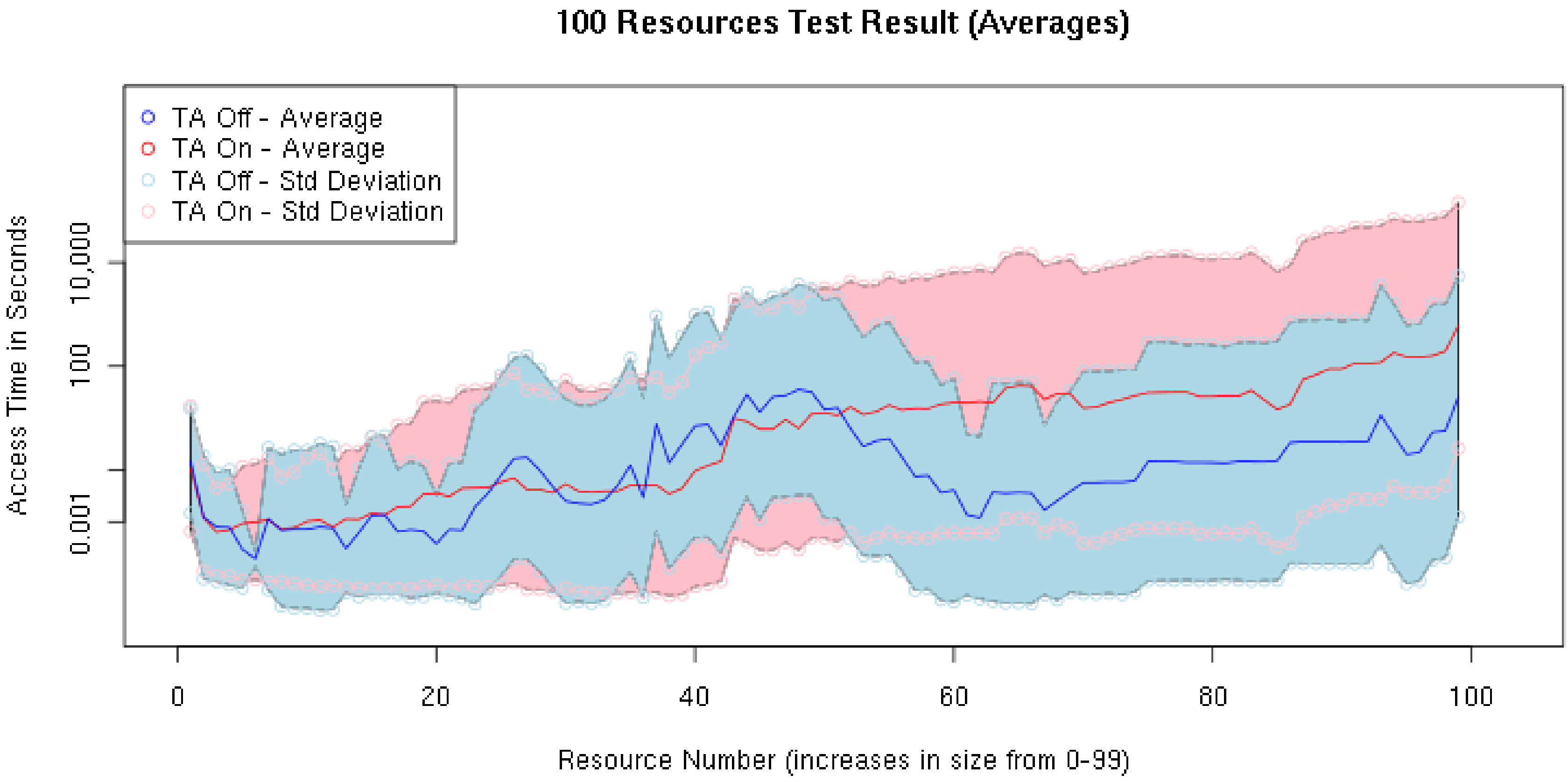
Benchmark with wget (burdened)



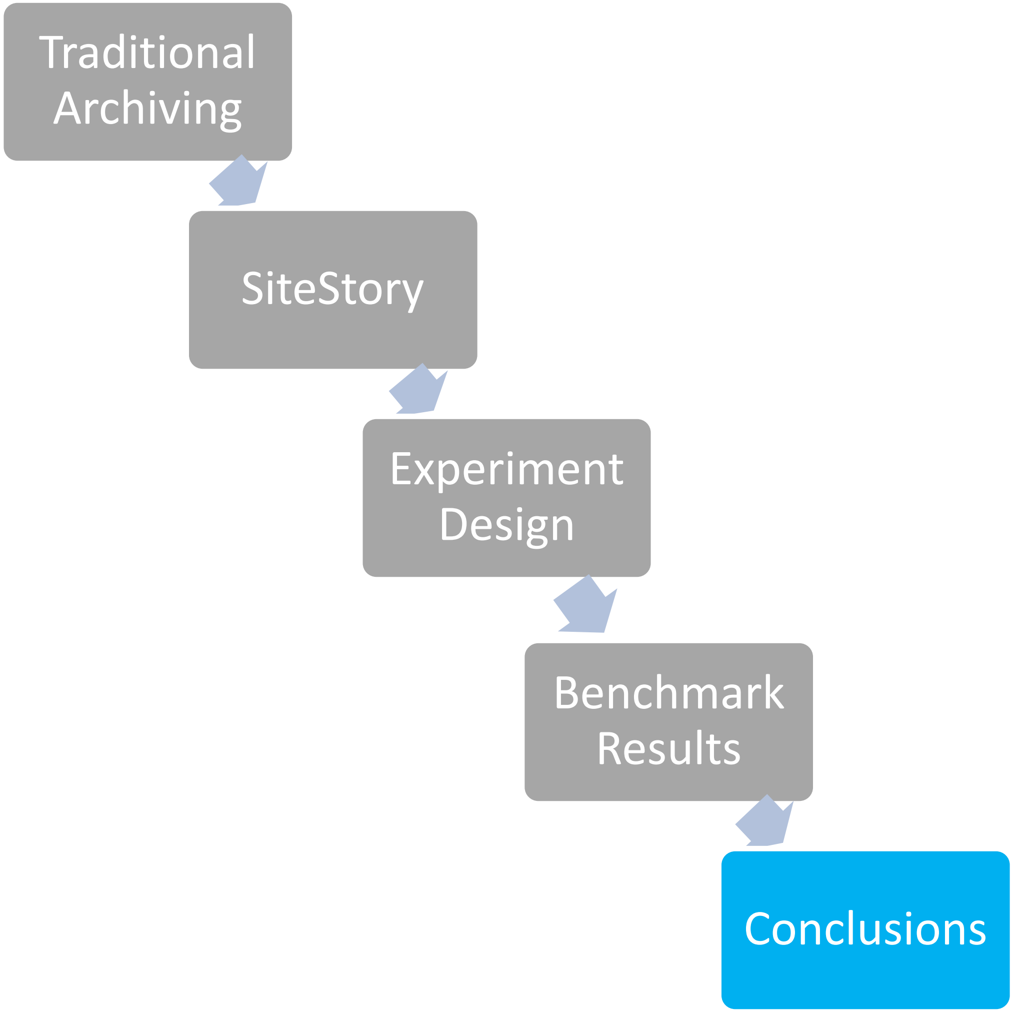
Results

- Negligible difference SiteStory On vs Off
- Limited to local LAN
- Performance over WAN?

WAN Testbed Performance



Agenda



Traditional Archiving

SiteStory

Experiment Design

Benchmark Results

Conclusions

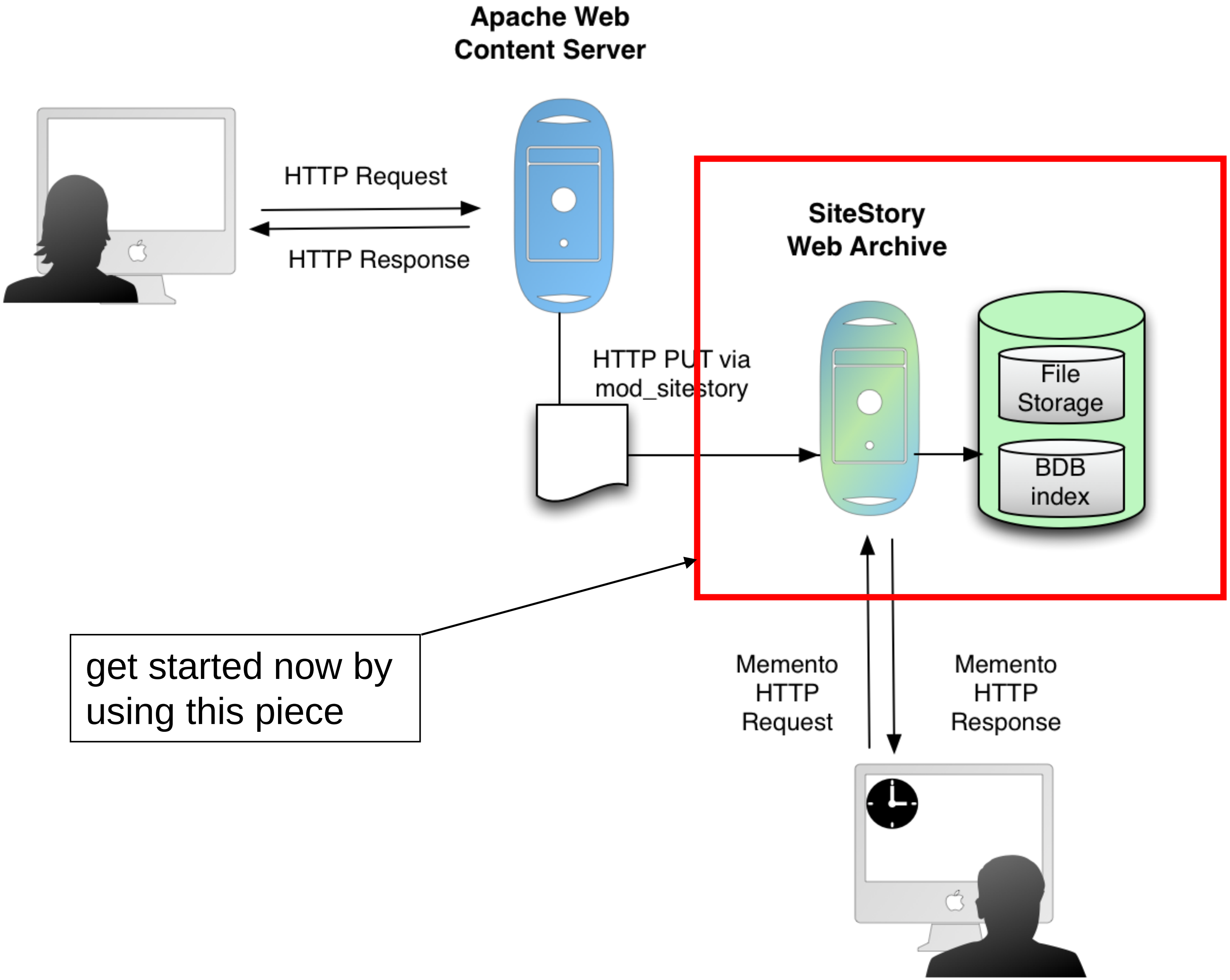
Results

- Distributed: Higher variance
- Increased delay due to network
- On vs. Off Comparison still comparable

Conclusions

- Small performance difference
- No gaps in coverage -- archives *every* HTTP response sent (optimizations possible)

http://mementoweb.github.io/SiteStory/



get started now by using this piece

SiteStory Testbed

- Use *our* SiteStory web archive on *your* server!

1. Install and configure *mod_sitestory* on your Apache Server
2. Send an email containing:
   1. Your contact info
   2. Web server IP address
   3. Web server domain name
3. Happy Sitestory’ing!

- mailto: SiteStory-Testbed@googlegroups.com

Backups

Sample ab output

```
$ ab -n 10 -c 2 "http://www.cs.odu.edu/"
This is ApacheBench, Version 2.3 <$Revision: 655654 $>
…

Server Software:       Apache/2.2.17
Server Hostname:       www.cs.odu.edu
Server Port:           80

Document Path:         /
Document Length:       62289 bytes

Concurrency Level:     2
Time taken for tests:  0.213 seconds
Complete requests:     10
Failed requests:       0
Write errors:          0
Total transferred:     624810 bytes
HTML transferred:      622890 bytes
Requests per second:   47.01 [#/sec] (mean)
Time per request:      42.540 [ms] (mean)
Time per request:      21.270 [ms] (mean, across all concurrent requests)
Transfer rate:         2868.66 [Kbytes/sec] received
```

```
…
Connection Times (ms)
              min  mean[+/-sd] median   max
Connect:        0    1   0.0      1       1
Processing:    27   41  10.8     45      62
Waiting:        3    3   0.4      4       4
Total:         27   41  10.8     45      63

Percentage of the requests served within a certain time (ms)
  50%     45
  66%     46
  75%     46
  80%     46
  90%     63
  95%     63
  98%     63
  99%     63
 100%     63 (longest request)
```