

# HathiTrust Large Scale Search



[www.hathitrust.org](http://www.hathitrust.org)

**Tom Burton-West**  
**Information Retrieval Programmer**  
**Digital Library Production Service**  
**University of Michigan**

[www.hathitrust.org/blogs](http://www.hathitrust.org/blogs)

**October 7th 2010**





# HathiTrust

- **HathiTrust is a shared digital repository**
- **30+ member libraries**
- **Large Scale Search is one of many services built on top of the repository**
- **Currently about 6.5 million books**
- **250 Terabytes**
  - Preservation page images; jpeg 2000, tiff (244TB)
  - OCR and Metadata about (6TB)



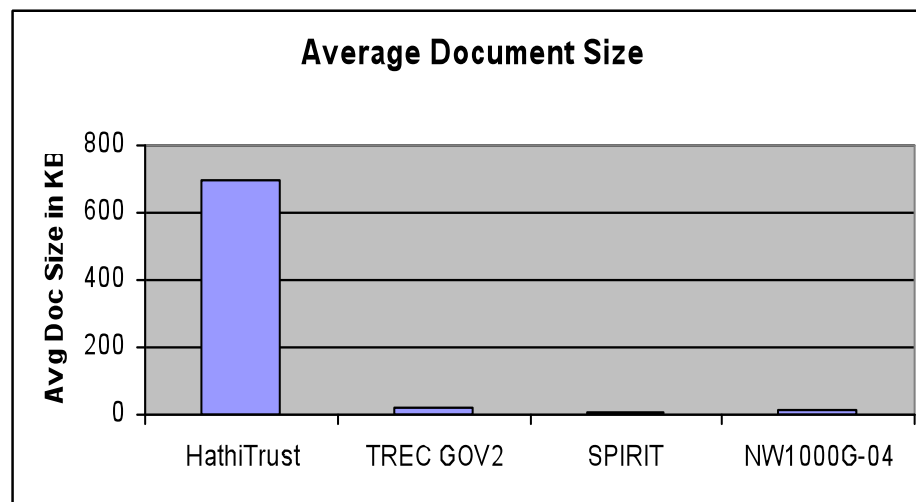
# Challenges

- **Goal: Design a system for full-text search that will scale to 7 million -20 million volumes (at a reasonable cost.)**
- **Challenges:**
  - Must scale to 20 million full-text volumes
  - Very long documents compared to most large-scale search applications
  - Multilingual collection (400+ languages)
  - OCR quality varies



# Long Documents

- **Average HathiTrust document is 700KB containing over 100,000 words.**
  - Estimated size of 7 million Document collection is 4.5TB.
- **Average HathiTrust document is about 38 times larger than the average document size of 18KB used in Large Research test collections**



Collection	Size	Documents	Average Doc size
HathiTrust	4.5 TB (projected)	7 million	700 KB
TREC GOV2	0.456 TB	25 million	18 KB
SPIRIT	1 TB	94 million	10 KB
NW1000G-04	1.3 TB*	100 million	16 KB

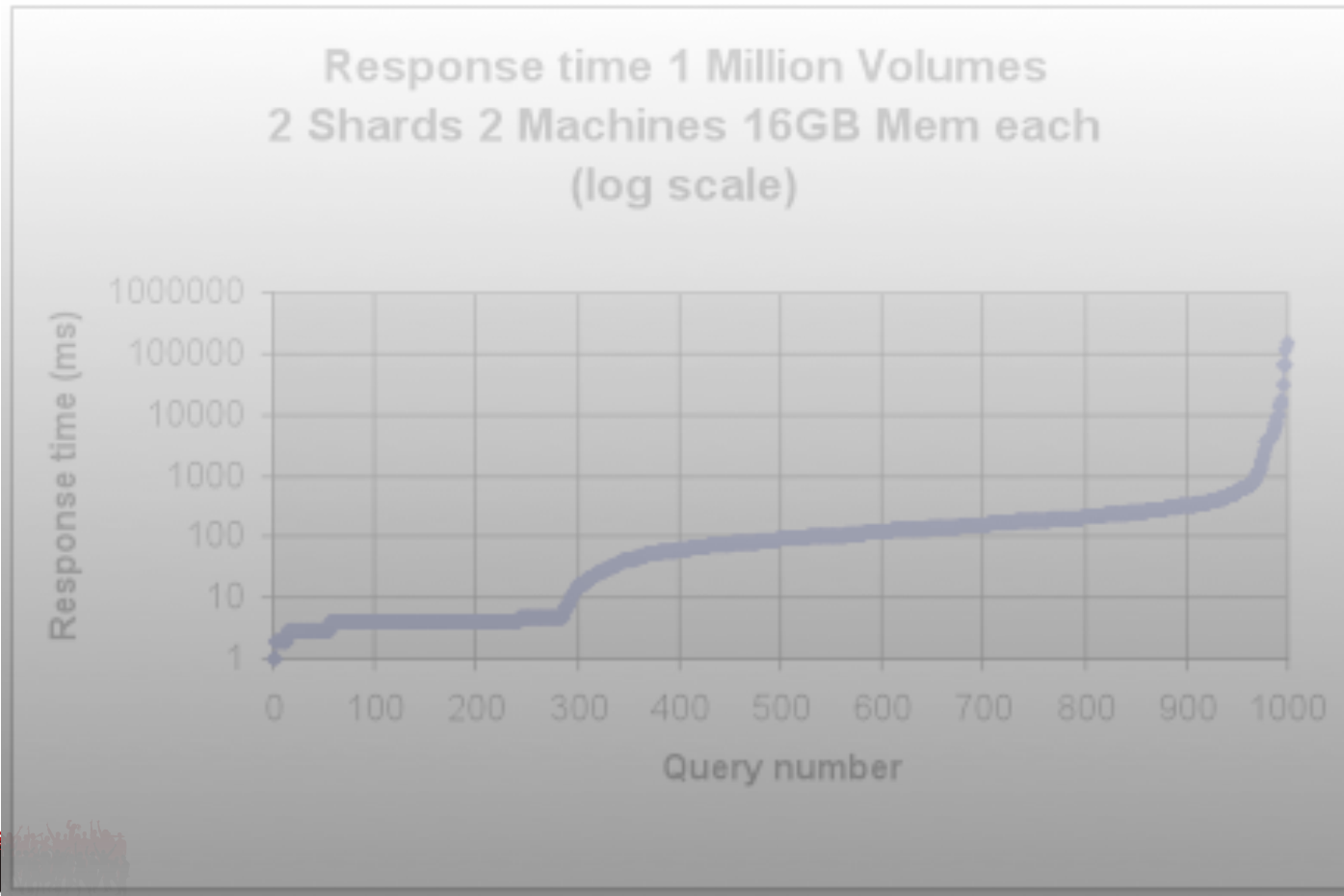


# Index Size, Caching, and Memory

- **Our 6 million document index is between 3 and 4 terabytes.**
  - About 450 GB per million documents
- **Large index means disk I/O is bottleneck**
- **Tradeoff JVM vs OS memory**
  - **Solr uses OS memory (disk I/O caching) for caching of postings**
  - **Memory available for disk I/O caching has most impact on response time (assuming adequate cache warming)**

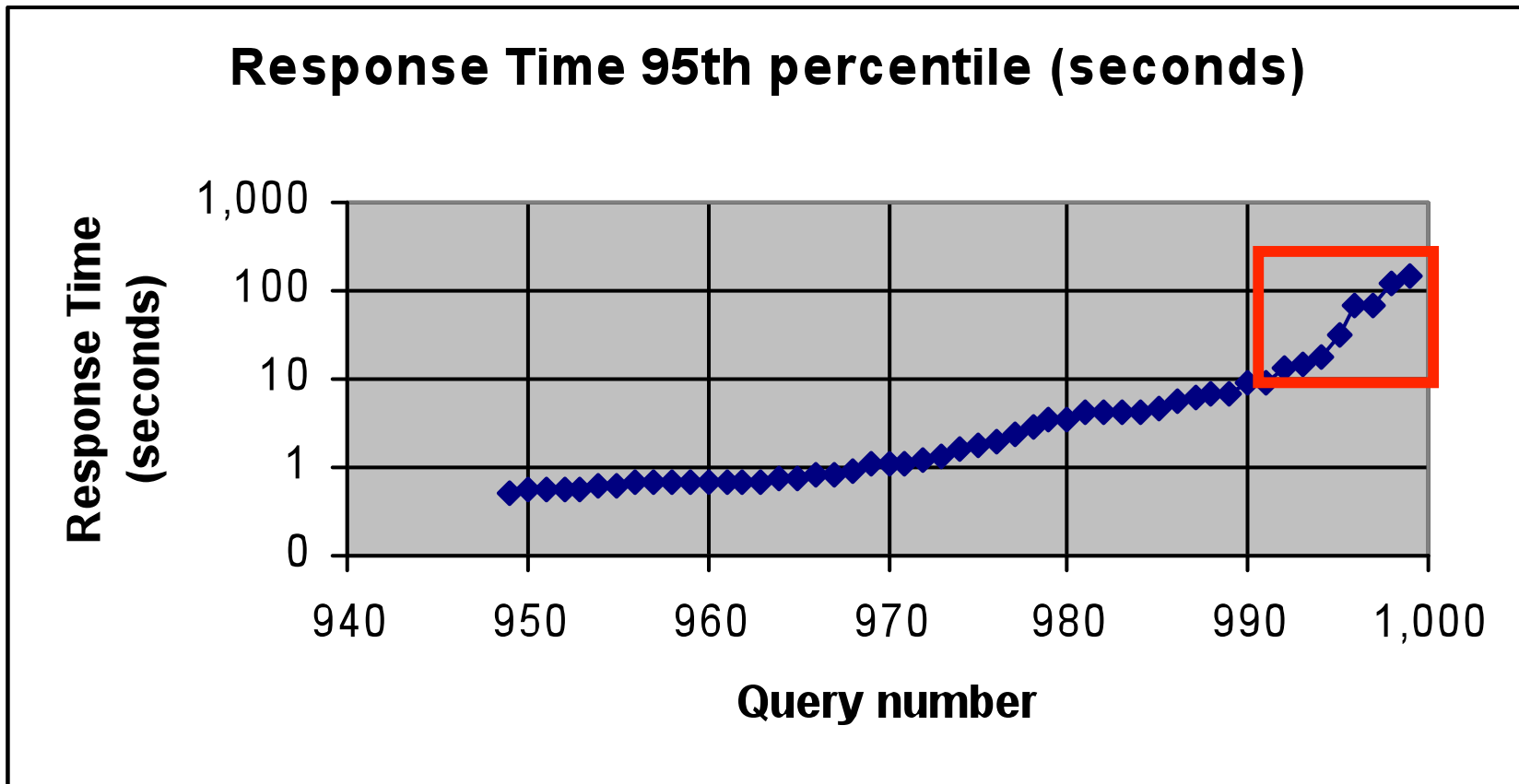


# Response Time Varies with Query



Average:	673
Median:	91
90 <sup>th</sup> :	328
99 <sup>th</sup> :	7,504

# Slowest 5% of queries



# Standard index vs. CommonGrams

## Standard Index

WORD	TOTAL OCCURRENCES IN CORPUS (MILLIONS)	NUMBER OF DOCS (THOUSANDS)
the	2,013	386
of	1,299	440
and	855	376
literature	4	210
lives	2	194
generation	2	199
beat	0.6	130
<b>TOTAL</b>	<b>4,176</b>	

## Common Grams

TERM	TOTAL OCCURRENCES IN CORPUS (MILLIONS)	NUMBER OF DOCS (THOUSANDS)
of-the	446	396
generation	2.42	262
the-lives	0.36	128
literature-of	0.35	103
lives-and	0.25	115
and-literature	0.24	77
the-beat	0.06	26
<b>TOTAL</b>	<b>450</b>	



# CommonGrams

## Comparison of Response time (ms)

	average	median	90th	99th	slowest query
Standard Index	459	32	146	6,784	120,595
Common Grams	68	3	71	2,226	7,800



# Thank You !



[www.hathitrust.org](http://www.hathitrust.org)

Tom Burton-West  
tburtonw@umich.edu

[www.hathitrust.org/blogs/large-scale-search](http://www.hathitrust.org/blogs/large-scale-search)



# How SLIP is Distributed Over Network & Hardware

