Google Inc
All the World's Information

A Data Playground

…There's Google

Google’s mission is to …

Organize all the world’s information and make it universally accessible and useful

www.google.com/bangalore

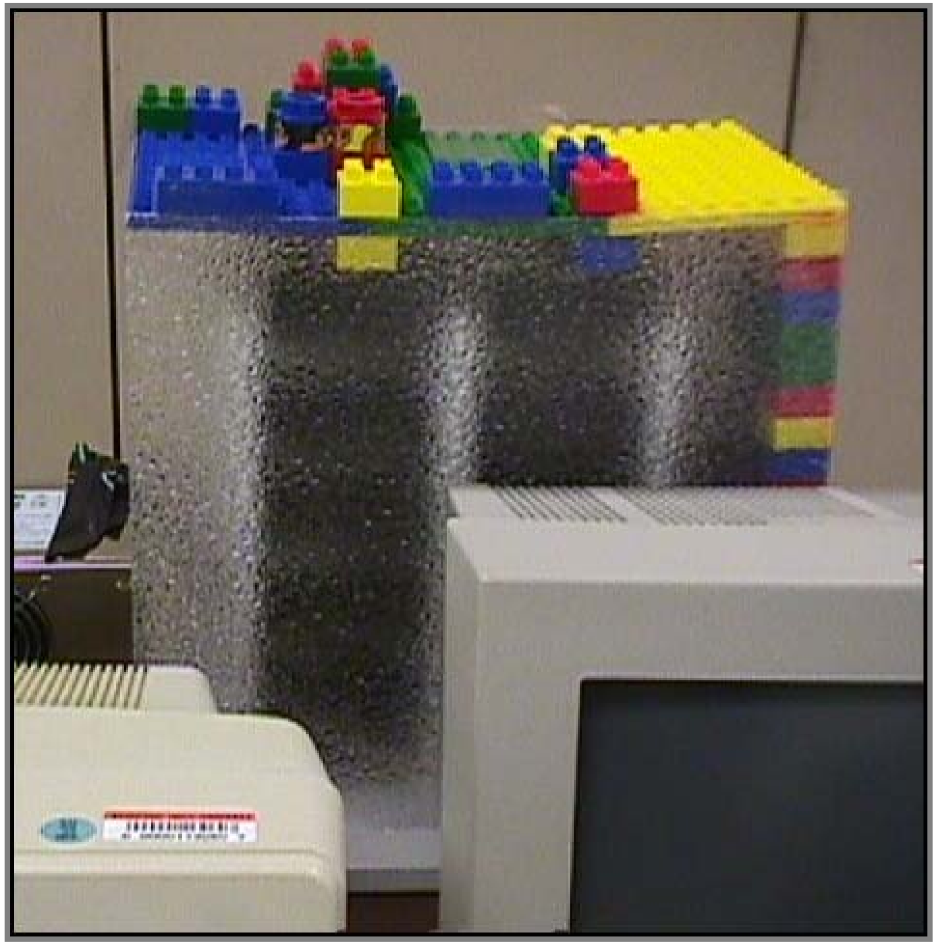
Google computing evolves…

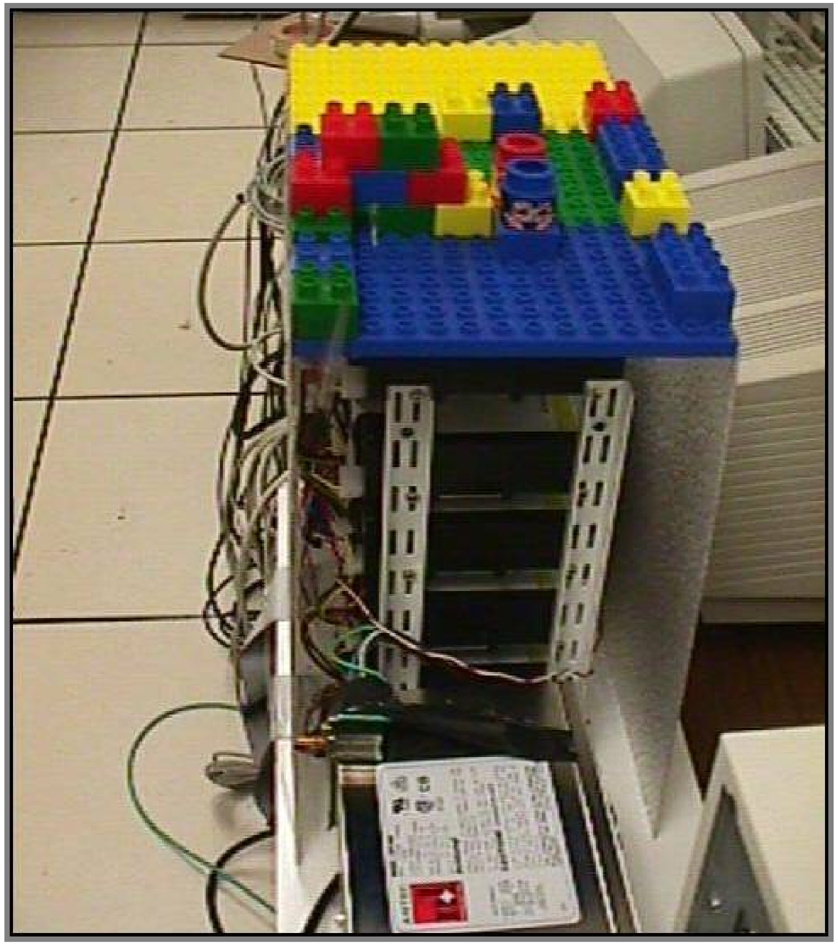
Stanford

Graduate student project

The Garage

Lego Disc Case (Version 0.1)

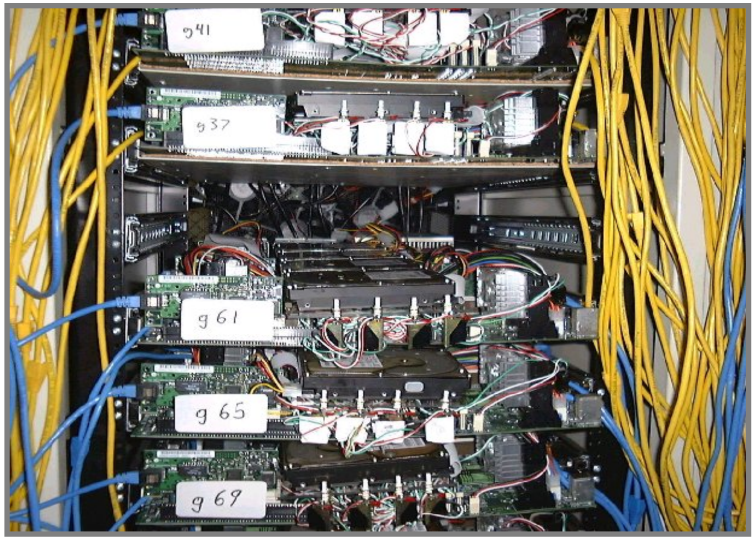




Two guys with a plan

Larry and Sergey built their own computers and everything that ran on them

Google - Version 0.1



Google - Version 1

Hardware Evolution: Spring 2000

Hardware Evolution: Late 2000

Three Days Later…

Google today

- Current Index: Billions of web pages,
  2 Billion images, 1 Billion usenet articles and other files
- Employees: >5,000
- Search and Content Partners: 1000s worldwide
  (including AOL, Disney, NEC, and The New York Times)
- Market Share: 55+ percent of Internet search referrals*
- Advertising: Thousands of advertisers. 80% of Internet users in the US are reached by Google’s ad network.
- Office Locations: More than 20 offices worldwide including Mountain View, New York, London, Tokyo, Zurich, Paris, Milan, and Bangalore
- International: 104 interface languages and 113 international domains

* ComScore, Oct. 2005.

•“Most Intelligent Agent on the Internet”

Lots of fun technology…

The Science of Spam...

Spam

Spamming Google's ranking is profitable

- 80+% of users use search engines to find sites
- 50+% of the world's searches come to Google
- Users follow search results; money follows users, which implies: Ranking high on Google makes you money

Do the math…

Spamming Google's ranking is profitable

500 million searches/day globally

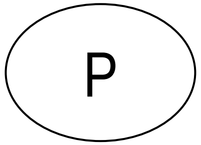
x 25% are commercially viable, say

x 5 cents/click

= $20 Billion a year / result click position

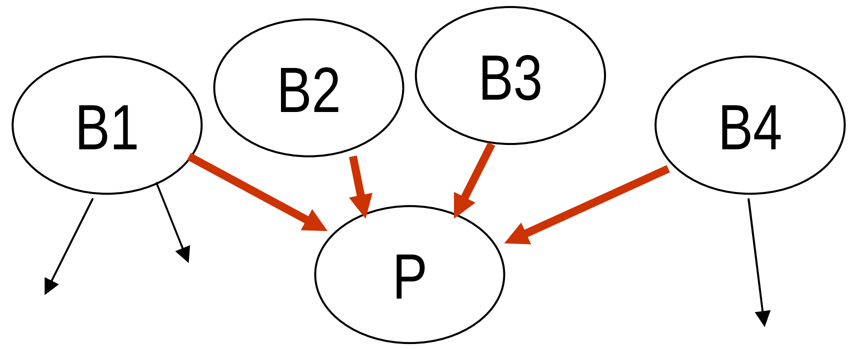
A new industry: Search Engine Optimization

Pagerank: Intuition



How good is page P?

Pagerank: Intuition

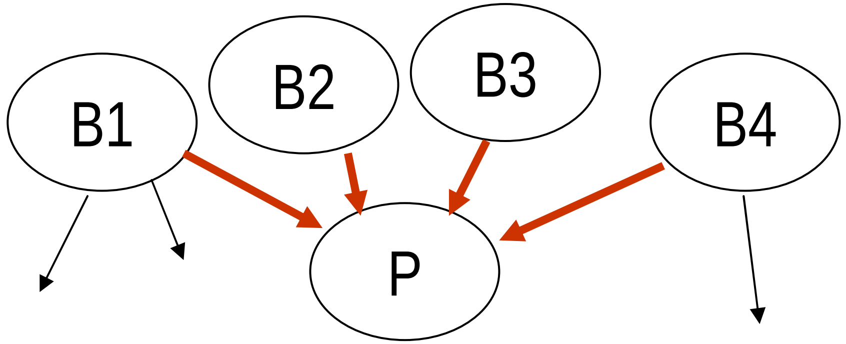


Intrinsic value of P

+

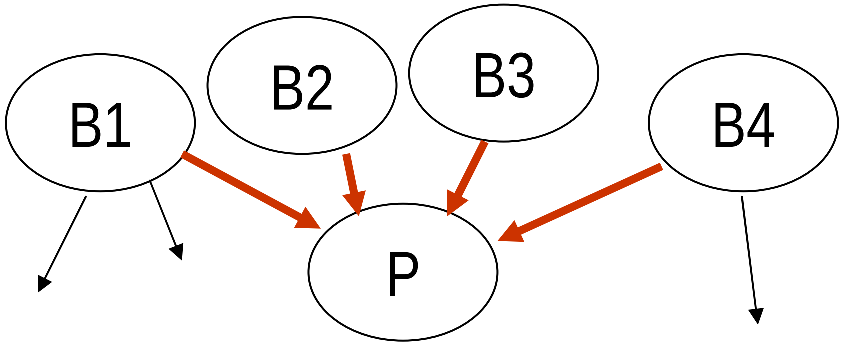
Referred value from pages that point to P

Measure value of page P



$$Value(P) = \overset{\text{Intrinsic Value of P}}{\downarrow \atop \alpha} + \underbrace{\beta \sum_{B \in BACK(P)} Value(B) / outdegree(B)}_{\text{Referred Value of P}}$$

Pagerank: Random Surfer Model



Probability of reaching P by a random jump

$$Pagerank(P) = \frac{1-\beta}{N} + \underbrace{\beta \sum_{B \in BACK(P)} Pagerank(B)/outdegree(B)}_{\text{Probability of surfing to P over a link}}$$

where N is the total number of pages on the web.

Mathematical interpretation

Consider the web graph as a matrix

- One row in matrix for each web page
- Order is 8 billion
- Entries denote transition probabilities

PageRank calculates the dominant eigenvector of the matrix

[Brin98] Sergey Brin and Larry Page. The anatomy of a large-scale hypertextual web search engine. *Proc. of 7th International WWW Conference*, pp. 107-117. 1998.

This is tough - Practical issues

- How do you represent 80B URLs?
- How do you sort 80B URL tuples?
- How do you distribute the PR vectors for iterations i and i+1?
- How do you distribute the link data?
- How to do this hourly (can we)?

The Science of Scale...

Dealing with scale

- Hardware, networking
  - Building a basic computing platform with low cost
- Distributed systems
  - Building reliable systems out of many individual computers
- Algorithms, data structures
  - Processing data efficiently, and in new and interesting ways
- Machine learning, information retrieval
  - Improving quality of search results by analyzing (lots of) data
- User interfaces
  - Designing effective interfaces for search and other products
- Many others…

Why use commodity PCs

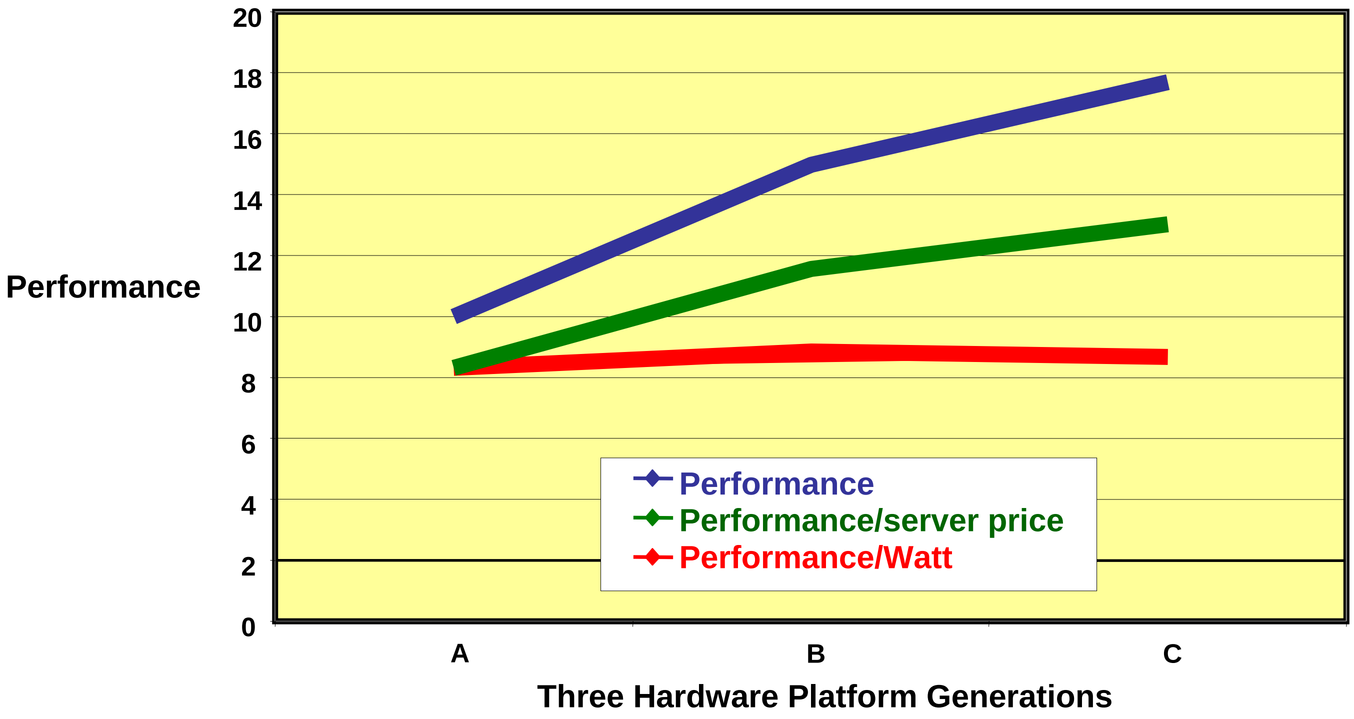
- Single high-end 8-way Intel server:
  - IBM eserver xSeries 440
  - 8 2-GHz Xeon, 64 GB RAM, 8 TB of disk
  - $758,000

- Commodity machines:
  - Rack of 88 machines
  - 176 2-GHz Xeons, 176 GB RAM, ~7 TB of disk
  - $278,000

- 1/3X price, 22X CPU, 3X RAM, 1X disk

Sources: racksaver.com, TPC-C performance results, both from late 2002

Power Trends: 3 Generations of Google Servers

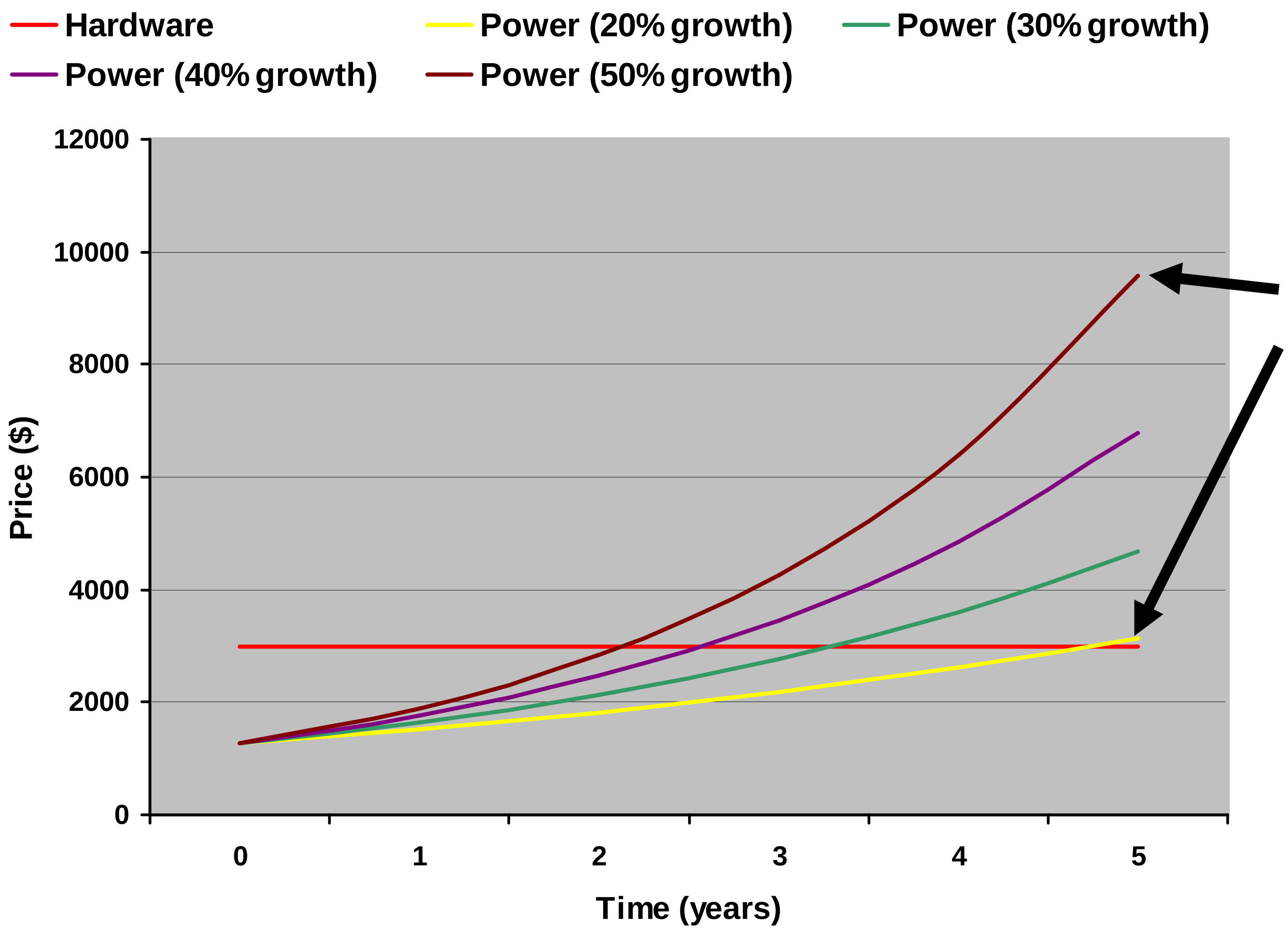


- Performance is up
- Performance/server price is up
- Performance/Watt is stagnant

Power vs Hardware costs today

- Example: high-volume dual-CPU Xeon server
  - System power ~250W
  - Cooling 1W takes about 1W ➡ ~500W
  - 4-year power cost >50% of hardware cost!
  - Ignoring:
    - Cost of power distribution/UPS/Backup generator equipment
    - Power distribution efficiencies
    - Forecasted increases in the cost of energy

Extrapolating: The next 5 years



Power costs dominate

The problem of utilization: Networking

- Cost of provisioning Gigabit networking
  - To a single server (NIC): $6
  - To a server rack (40 servers): ~$50/port
  - To a Google cluster (thousands of servers): priceless…
- Large gap in cost-efficiency improvements of servers and large networking switches
- Networking industry by enlarge is not motivated to address our requirements
- We are working on solutions that:
  - Provides tens of Terabits/sec bisection bandwidth for our clusters
  - Don't break the bank

What about failures?

Stuff breaks

- 1 computer: expect 3 year life
- 1000 computers: lose 1/day
- At Google scale, many machines will fail every day

Have to deal with failures in software

- Replication and redundancy
- Needed for capacity anyway

Fault-tolerant software, parallel makes cheap hardware practical

An Example: The Index

Similar to index in the back of a book (but big!)

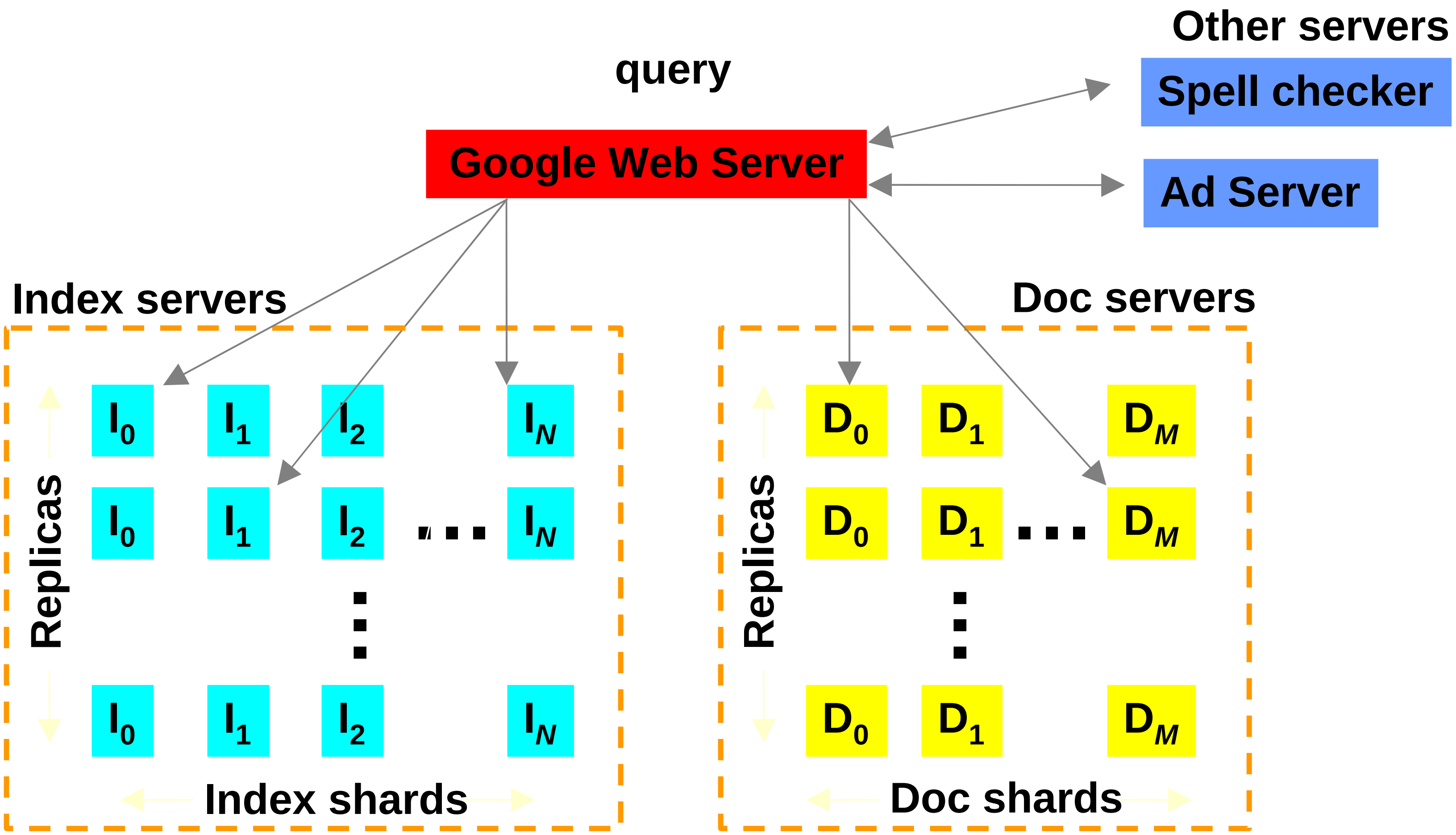
- Building takes several days on hundreds of machines
- Billions of web documents
- Images: 2000 M images
- File types: More than 35M non-HTML documents (PDF, Microsoft Word, etc.)
- Usenet: 1000M messages from >35K newsgroups

Structuring the Index

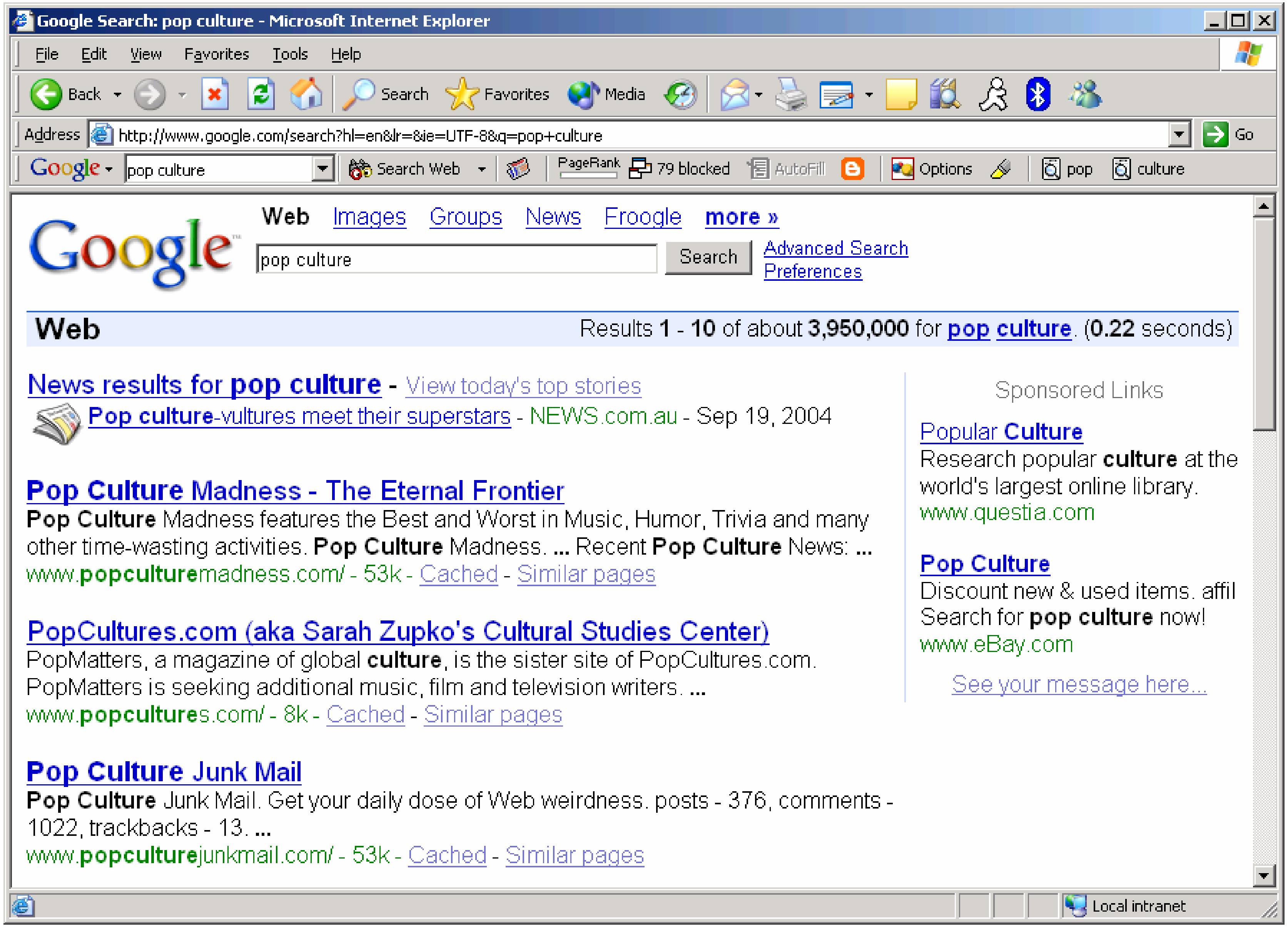
Too large for one machine, so...

- Use PageRank as a total order
- Split it into pieces, called shards, small enough to have several per machine
- Replicate the shards, making more replicas of high PageRank shards
- Do the same for the documents
- Then replicate this whole structure within and across data centers

Query Serving Infrastructure



Elapsed time: 0.25s, machines involved: 1000+


Google Search: pop culture - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address
http://www.google.com/search?hl=en&lr=&ie=UTF-8&q=pop+culture
Go
Google
pop culture
Search Web
PageRank
79 blocked
AutoFill
Options
pop
culture
Web Images Groups News Froogle more »
pop culture
Search
Advanced Search
Preferences
Local intranet

Web Results **1** - **10** of about **3,950,000** for **pop culture**. (**0.22** seconds)

News results for **pop culture** - View today's top stories

Pop culture-vultures meet their superstars - NEWS.com.au - Sep 19, 2004

Pop Culture Madness - The Eternal Frontier

Pop Culture Madness features the Best and Worst in Music, Humor, Trivia and many other time-wasting activities. **Pop Culture** Madness. **...** Recent **Pop Culture** News: **...**

www.**popculture**madness.com/ - 53k - Cached - Similar pages

PopCultures.com (aka Sarah Zupko's Cultural Studies Center)

PopMatters, a magazine of global **culture**, is the sister site of PopCultures.com. PopMatters is seeking additional music, film and television writers. **...**

www.**popculture**s.com/ - 8k - Cached - Similar pages

Pop Culture Junk Mail

Pop Culture Junk Mail. Get your daily dose of Web weirdness. posts - 376, comments - 1022, trackbacks - 13. **...**

www.**popculture**junkmail.com/ - 53k - Cached - Similar pages

Sponsored Links

Popular **Culture**

Research popular **culture** at the world's largest online library.

www.questia.com

Pop Culture

Discount new & used items. affil Search for **pop culture** now!

www.eBay.com

See your message here...

The Google Computer – a playground for data

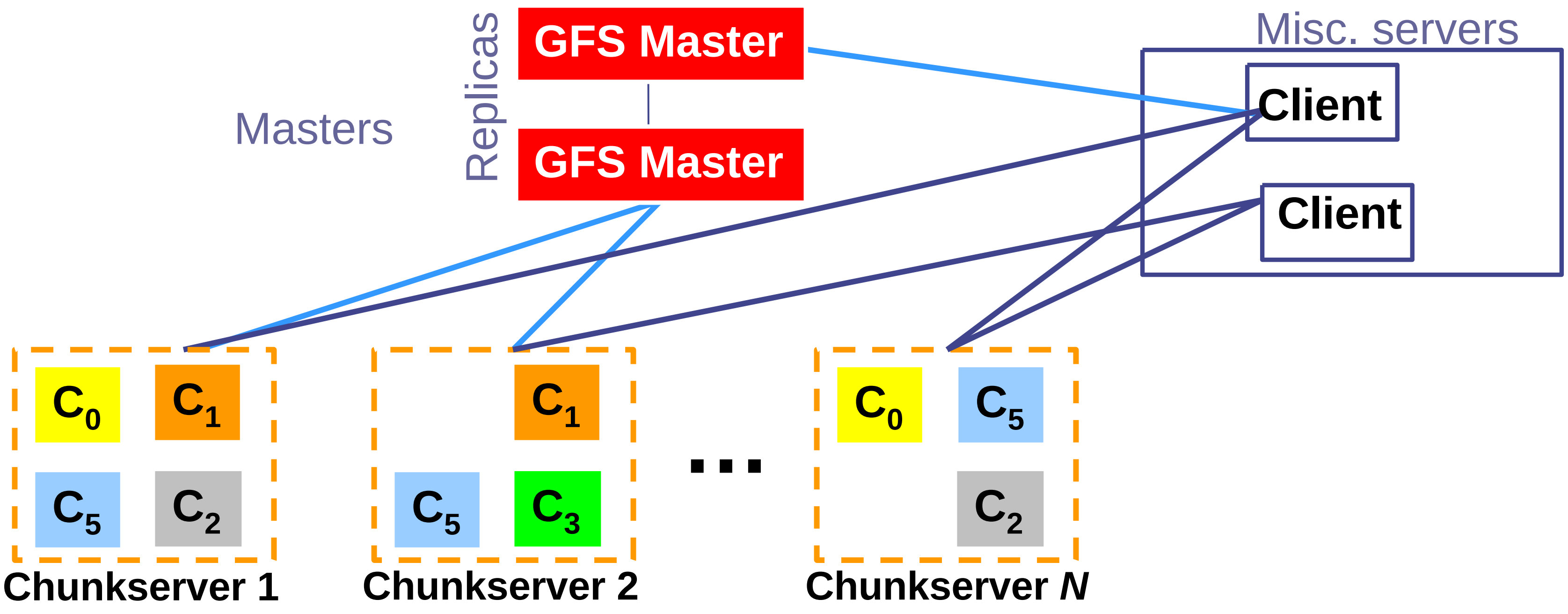
Our needs

- Store data reliably
- Run jobs on pools of machines
- Apply lots of computational resources to problems

In-house solutions

- Storage: Google File System (GFS)
- Job scheduling: Global Work Queue (GWQ)
- MapReduce: simplify large-scale data processing

Google File System



- Master manages metadata
- Data transfers happen directly between clients/chunkservers
- Files broken into chunks (typically 64 MB)
- Chunks triplicated across three machines for safety

GFS: Usage at Google

- 30+ Clusters
- Clusters as large as 2000+ chunkservers
- Petabyte-sized filesystems
- 2000+ MB/s sustained read/write load
- All in the presence of HW failures

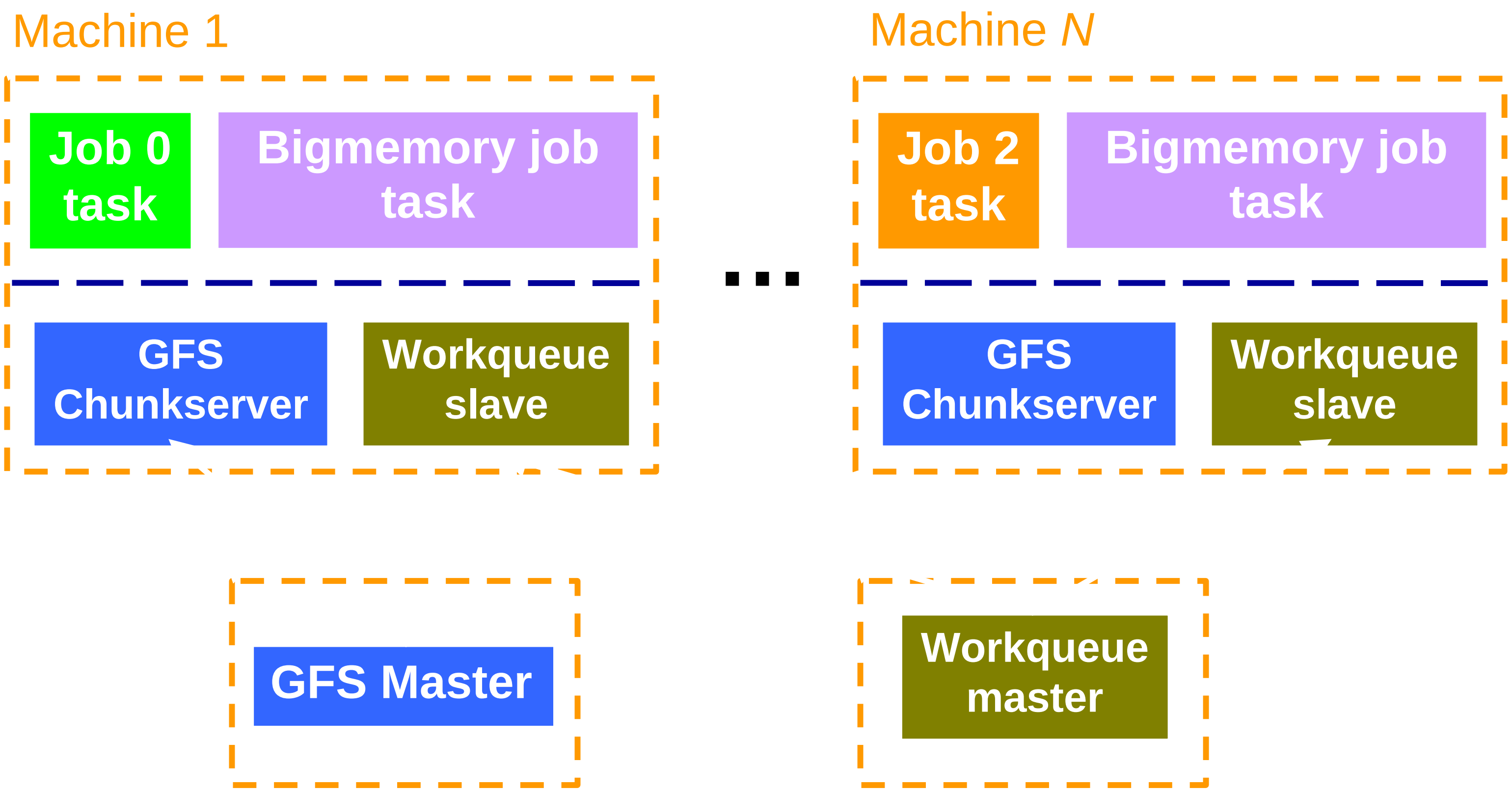
More information can be found in SOSP’03

Global Work Queue

- Workqueue master manages pool of slave machines
  - Slaves provide resources (memory, CPU, disk)
  - Users submit jobs to master (job is made up of tasks)
  - Tasks have resource requirements (mem, CPU, disk, etc.)
  - Each task is executed as a UNIX process
  - Task binaries stored in GFS, replicated onto slaves
  - System allows sharing of machines by many projects
  - Projects can use lots of CPUs when needed, but share with other projects when not needed

Timesharing on a large cluster of machines

Basic Computing Cluster



MapReduce: Easy-to-use Cycles

Many problems:

“Process lots of data to produce other data”

- Diverse inputs:

  e.g., document records, log files, sorted on-disk data structures

1. Want to use hundreds or thousands of CPUs
2. … but this needs to be easy to use

- MapReduce framework that provides
  (for certain classes of problems):
  - Automatic & efficient parallelization/distribution
  - Fault-tolerance
  - I/O scheduling
  - Status/monitoring

MapReduce: Programming Model

- Input is sequence of key/value pairs

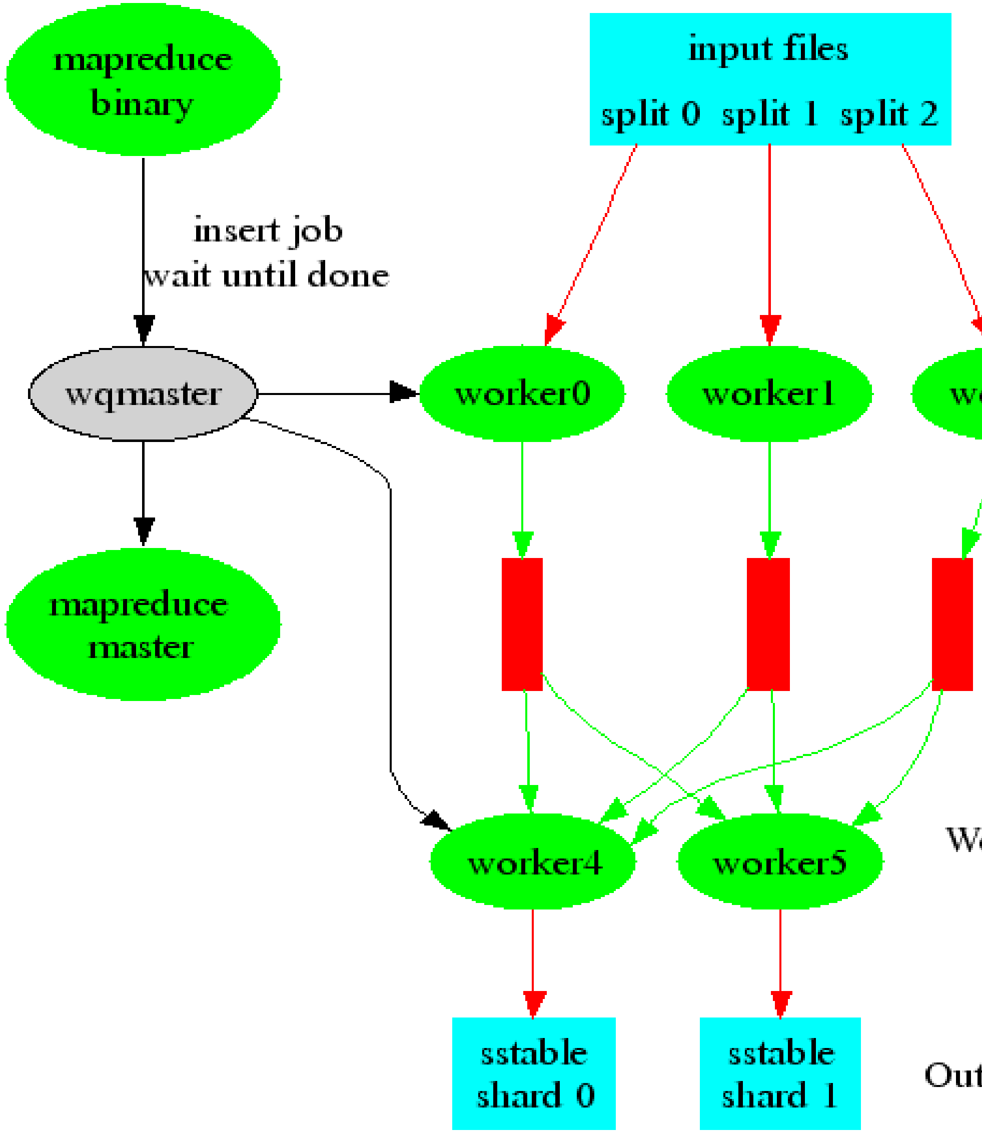
  e.g. url → document contents, docid → url, etc.

- Users write two simple functions:
  - *Map*: takes input key/value and produces set of intermediate key/value pairs

    e.g., map(url, contents) ➔ hostname → "1"
  - *Reduce* takes intermediate key and all intermediate values for that key, combines to produce output key/value

    e.g., reduce(hostname → {“1”,”1”,”1”,”1”}) ➔ hostname → “4"
- key+combined value are emitted to output file

MapReduce: System Structure



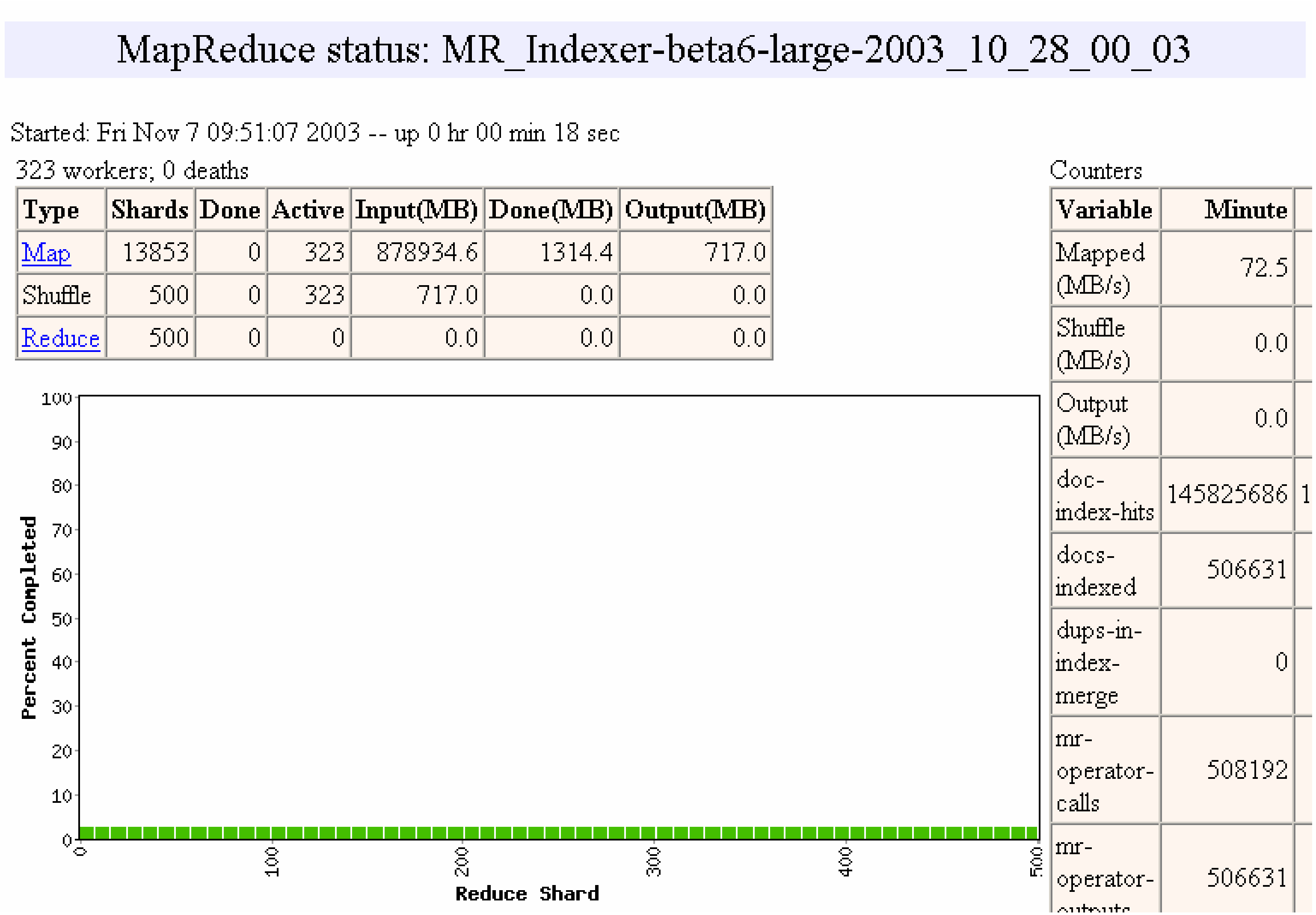
Input files split into chunks

Map phase: Generate output keys/values

Intermediate files (on local disks)

Workers assigned to reduce phase work
Sort+Merge duplicate records

Output files (written into GFS)



MapReduce status: MR_Indexer-beta6-large-2003_10_28_00_03

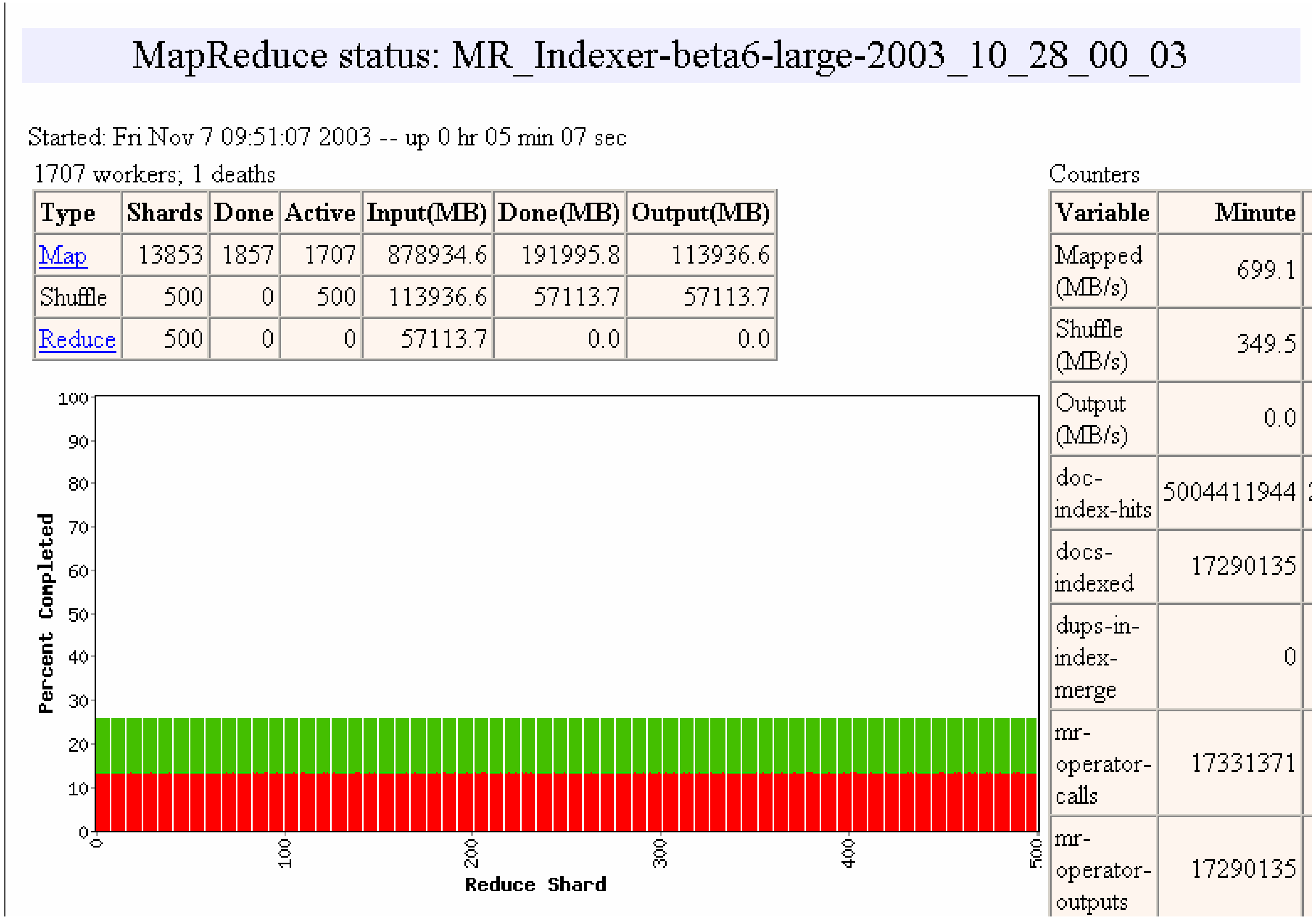
Started: Fri Nov 7 09:51:07 2003 -- up 0 hr 00 min 18 sec

323 workers; 0 deaths

Type	Shards	Done	Active	Input(MB)	Done(MB)	Output(MB)
Map	13853	0	323	878934.6	1314.4	717.0
Shuffle	500	0	323	717.0	0.0	0.0
Reduce	500	0	0	0.0	0.0	0.0

Counters

Variable	Minute	
Mapped (MB/s)	72.5	
Shuffle (MB/s)	0.0	
Output (MB/s)	0.0	
doc-index-hits	145825686	1
docs-indexed	506631	
dups-in-index-merge	0	
mr-operator-calls	508192	
mr-operator-outputs	506631	



MapReduce status: MR_Indexer-beta6-large-2003_10_28_00_03

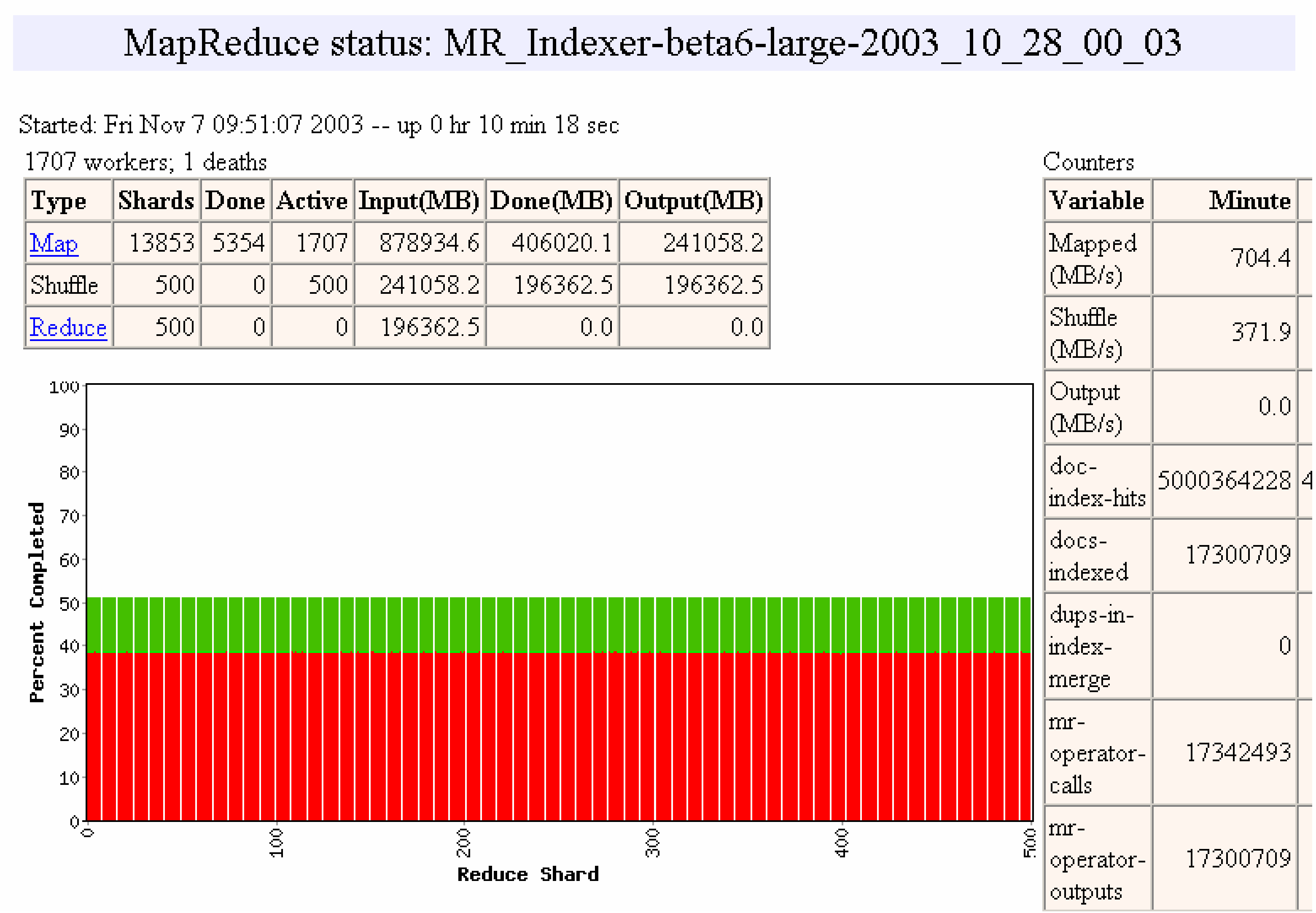
Started: Fri Nov 7 09:51:07 2003 -- up 0 hr 05 min 07 sec

1707 workers; 1 deaths

Type	Shards	Done	Active	Input(MB)	Done(MB)	Output(MB)
Map	13853	1857	1707	878934.6	191995.8	113936.6
Shuffle	500	0	500	113936.6	57113.7	57113.7
Reduce	500	0	0	57113.7	0.0	0.0

Counters

Variable	Minute
Mapped (MB/s)	699.1
Shuffle (MB/s)	349.5
Output (MB/s)	0.0
doc-index-hits	5004411944
docs-indexed	17290135
dups-in-index-merge	0
mr-operator-calls	17331371
mr-operator-outputs	17290135



MapReduce status: MR_Indexer-beta6-large-2003_10_28_00_03

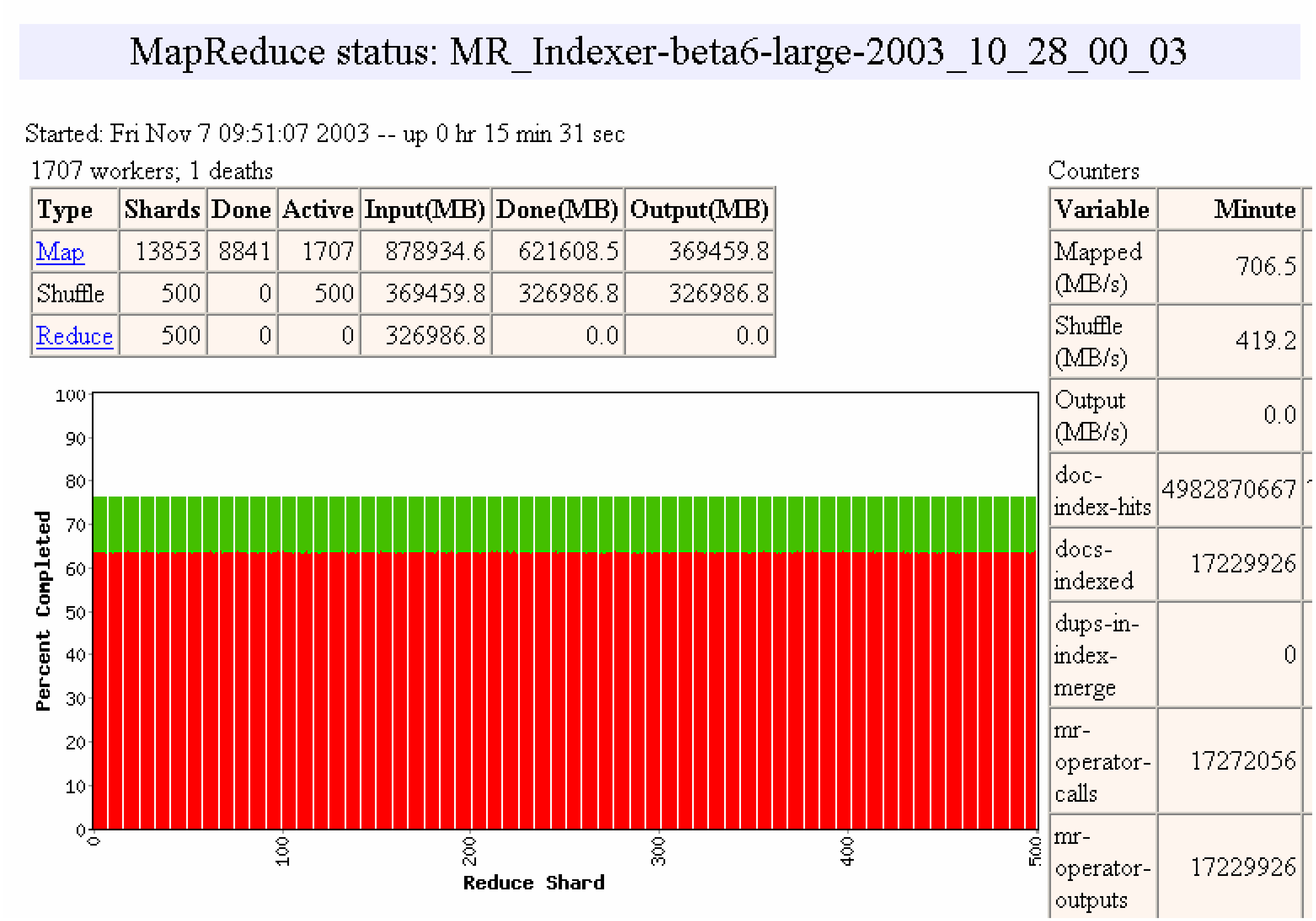
Started: Fri Nov 7 09:51:07 2003 -- up 0 hr 10 min 18 sec

1707 workers; 1 deaths

Type	Shards	Done	Active	Input(MB)	Done(MB)	Output(MB)
Map	13853	5354	1707	878934.6	406020.1	241058.2
Shuffle	500	0	500	241058.2	196362.5	196362.5
Reduce	500	0	0	196362.5	0.0	0.0

Counters

Variable	Minute	
Mapped (MB/s)	704.4	
Shuffle (MB/s)	371.9	
Output (MB/s)	0.0	
doc-index-hits	5000364228	4
docs-indexed	17300709	
dups-in-index-merge	0	
mr-operator-calls	17342493	
mr-operator-outputs	17300709	



MapReduce status: MR_Indexer-beta6-large-2003_10_28_00_03

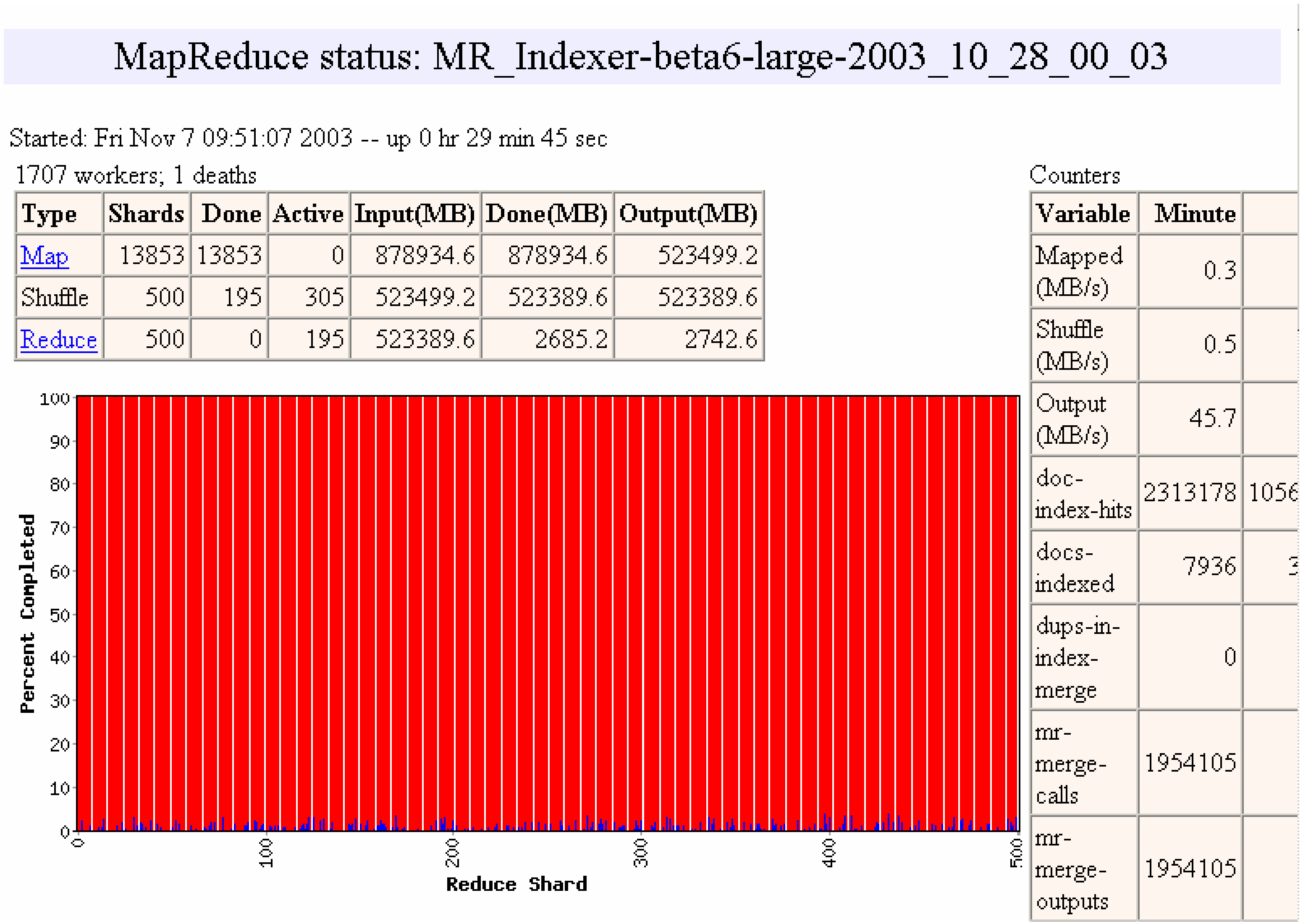
Started: Fri Nov 7 09:51:07 2003 -- up 0 hr 15 min 31 sec

1707 workers; 1 deaths

Type	Shards	Done	Active	Input(MB)	Done(MB)	Output(MB)
Map	13853	8841	1707	878934.6	621608.5	369459.8
Shuffle	500	0	500	369459.8	326986.8	326986.8
Reduce	500	0	0	326986.8	0.0	0.0

Counters

Variable	Minute
Mapped (MB/s)	706.5
Shuffle (MB/s)	419.2
Output (MB/s)	0.0
doc-index-hits	4982870667
docs-indexed	17229926
dups-in-index-merge	0
mr-operator-calls	17272056
mr-operator-outputs	17229926



MapReduce status: MR_Indexer-beta6-large-2003_10_28_00_03

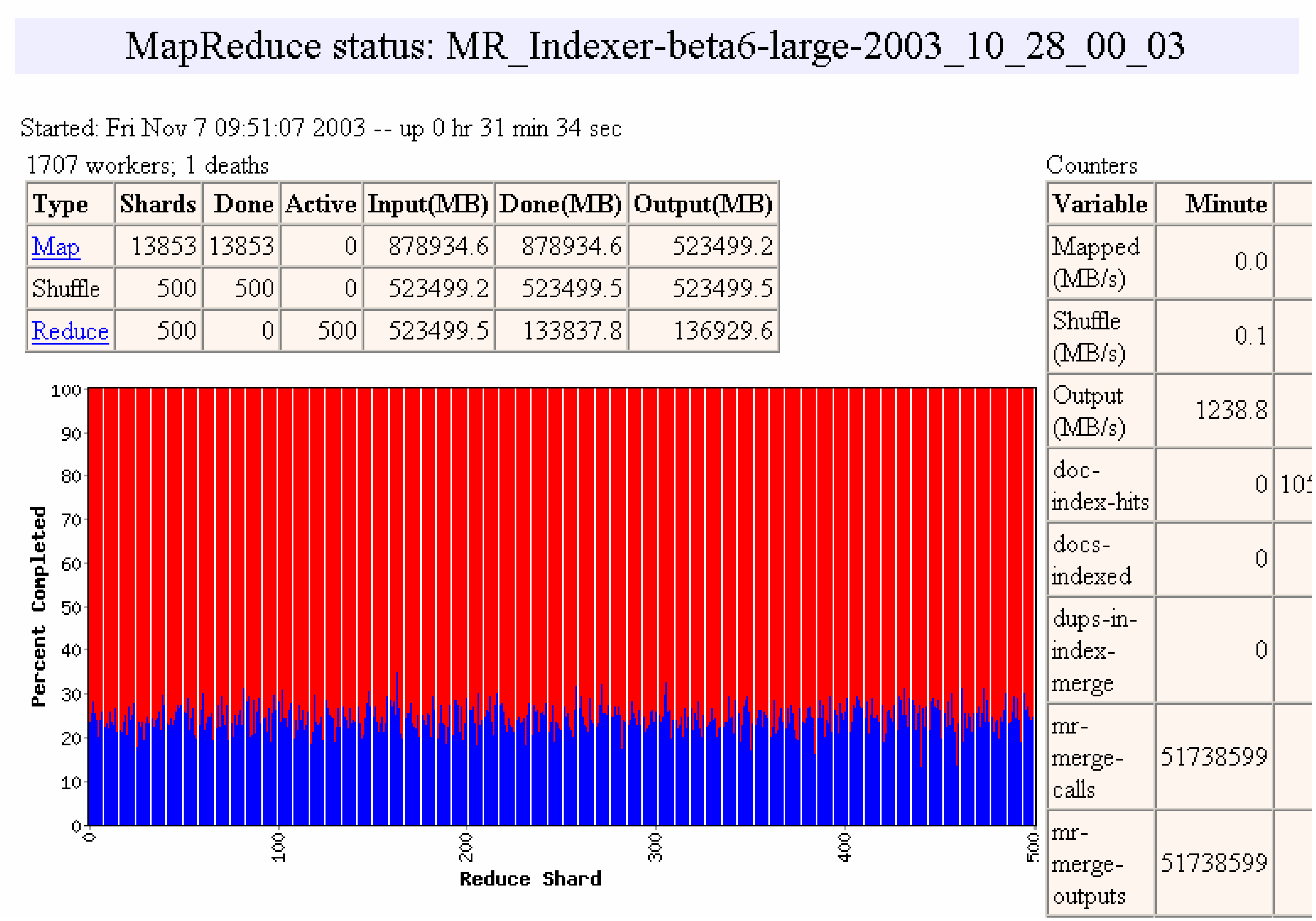
Started: Fri Nov 7 09:51:07 2003 -- up 0 hr 29 min 45 sec

1707 workers; 1 deaths

Type	Shards	Done	Active	Input(MB)	Done(MB)	Output(MB)
Map	13853	13853	0	878934.6	878934.6	523499.2
Shuffle	500	195	305	523499.2	523389.6	523389.6
Reduce	500	0	195	523389.6	2685.2	2742.6

Counters

Variable	Minute	
Mapped (MB/s)	0.3	
Shuffle (MB/s)	0.5	
Output (MB/s)	45.7	
doc-index-hits	2313178	1056
docs-indexed	7936	3
dups-in-index-merge	0	
mr-merge-calls	1954105	
mr-merge-outputs	1954105	



MapReduce status: MR_Indexer-beta6-large-2003_10_28_00_03

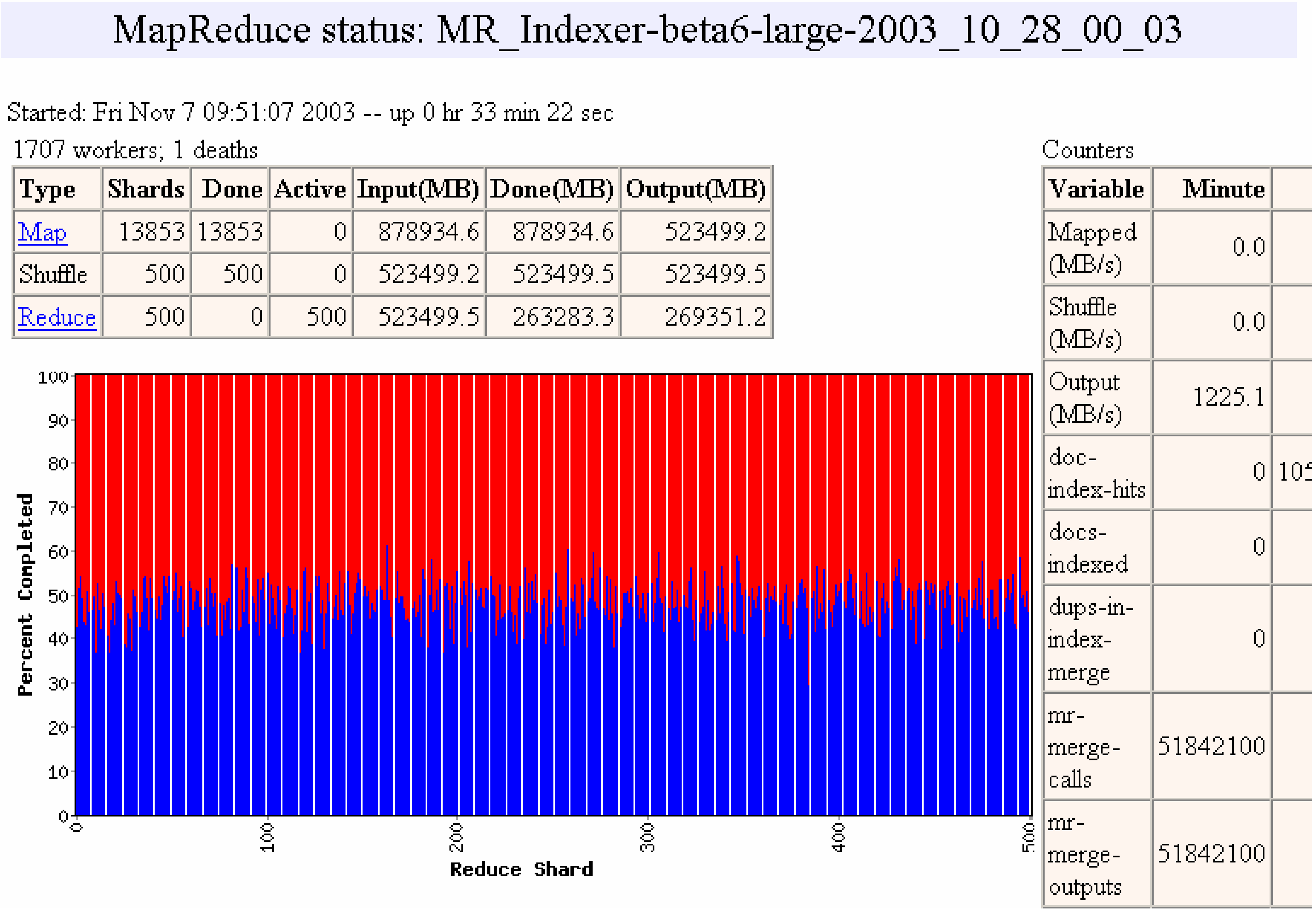
Started: Fri Nov 7 09:51:07 2003 -- up 0 hr 31 min 34 sec

1707 workers; 1 deaths

Type	Shards	Done	Active	Input(MB)	Done(MB)	Output(MB)
Map	13853	13853	0	878934.6	878934.6	523499.2
Shuffle	500	500	0	523499.2	523499.5	523499.5
Reduce	500	0	500	523499.5	133837.8	136929.6

Counters

Variable	Minute	
Mapped (MB/s)	0.0	
Shuffle (MB/s)	0.1	
Output (MB/s)	1238.8	
doc-index-hits	0	10
docs-indexed	0	
dups-in-index-merge	0	
mr-merge-calls	51738599	
mr-merge-outputs	51738599	



MapReduce status: MR_Indexer-beta6-large-2003_10_28_00_03

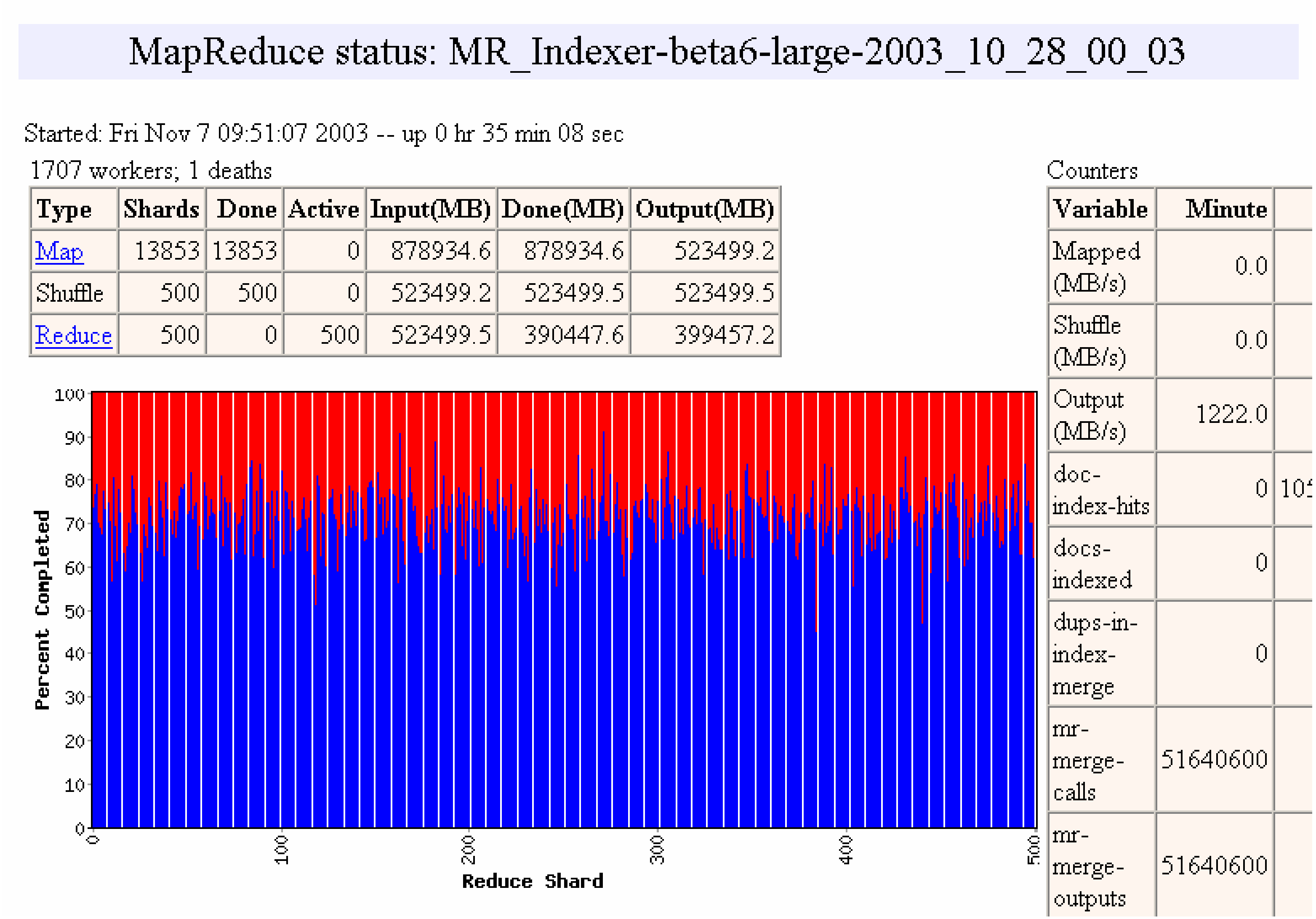
Started: Fri Nov 7 09:51:07 2003 -- up 0 hr 33 min 22 sec

1707 workers; 1 deaths

Type	Shards	Done	Active	Input(MB)	Done(MB)	Output(MB)
Map	13853	13853	0	878934.6	878934.6	523499.2
Shuffle	500	500	0	523499.2	523499.5	523499.5
Reduce	500	0	500	523499.5	263283.3	269351.2

Counters

Variable	Minute	
Mapped (MB/s)	0.0	
Shuffle (MB/s)	0.0	
Output (MB/s)	1225.1	
doc-index-hits	0	105
docs-indexed	0	
dups-in-index-merge	0	
mr-merge-calls	51842100	
mr-merge-outputs	51842100	



MapReduce status: MR_Indexer-beta6-large-2003_10_28_00_03

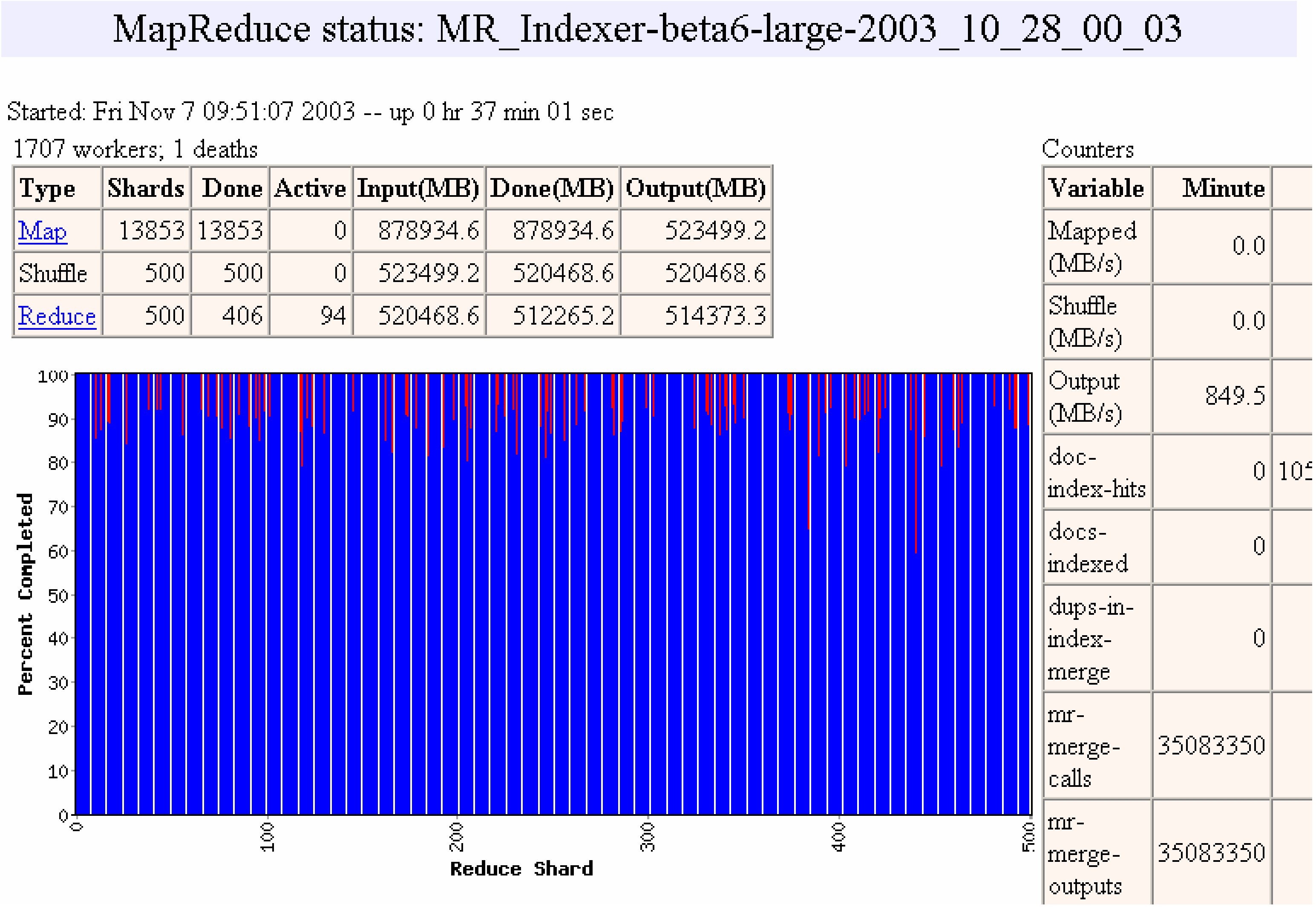
Started: Fri Nov 7 09:51:07 2003 -- up 0 hr 35 min 08 sec

1707 workers; 1 deaths

Type	Shards	Done	Active	Input(MB)	Done(MB)	Output(MB)
Map	13853	13853	0	878934.6	878934.6	523499.2
Shuffle	500	500	0	523499.2	523499.5	523499.5
Reduce	500	0	500	523499.5	390447.6	399457.2

Counters

Variable	Minute	
Mapped (MB/s)	0.0	
Shuffle (MB/s)	0.0	
Output (MB/s)	1222.0	
doc-index-hits	0	10
docs-indexed	0	
dups-in-index-merge	0	
mr-merge-calls	51640600	
mr-merge-outputs	51640600	



MapReduce status: MR_Indexer-beta6-large-2003_10_28_00_03

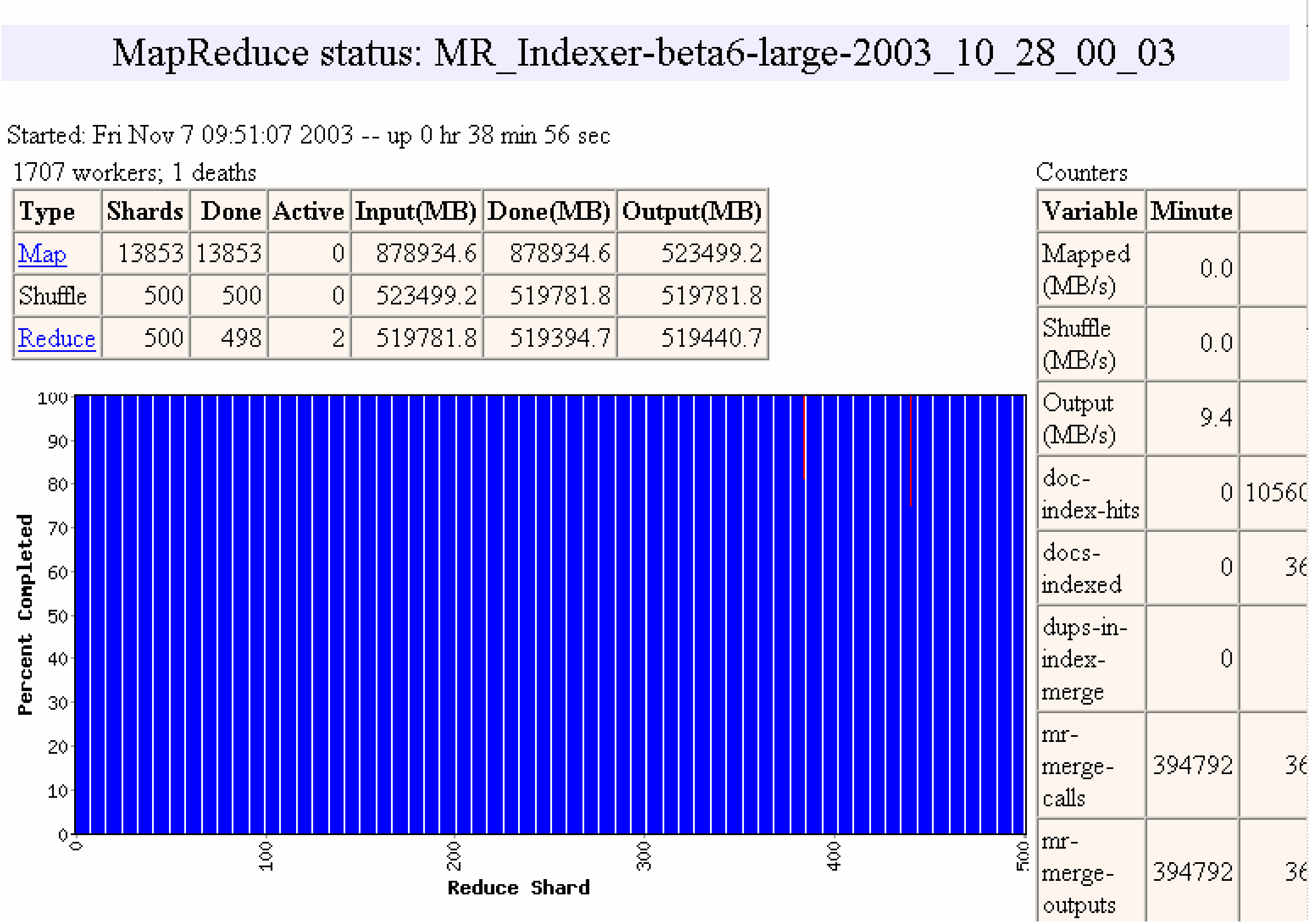
Started: Fri Nov 7 09:51:07 2003 -- up 0 hr 37 min 01 sec

1707 workers; 1 deaths

Type	Shards	Done	Active	Input(MB)	Done(MB)	Output(MB)
Map	13853	13853	0	878934.6	878934.6	523499.2
Shuffle	500	500	0	523499.2	520468.6	520468.6
Reduce	500	406	94	520468.6	512265.2	514373.3

Counters

Variable	Minute	
Mapped (MB/s)	0.0	
Shuffle (MB/s)	0.0	
Output (MB/s)	849.5	
doc-index-hits	0	105
docs-indexed	0	
dups-in-index-merge	0	
mr-merge-calls	35083350	
mr-merge-outputs	35083350	



MapReduce status: MR_Indexer-beta6-large-2003_10_28_00_03

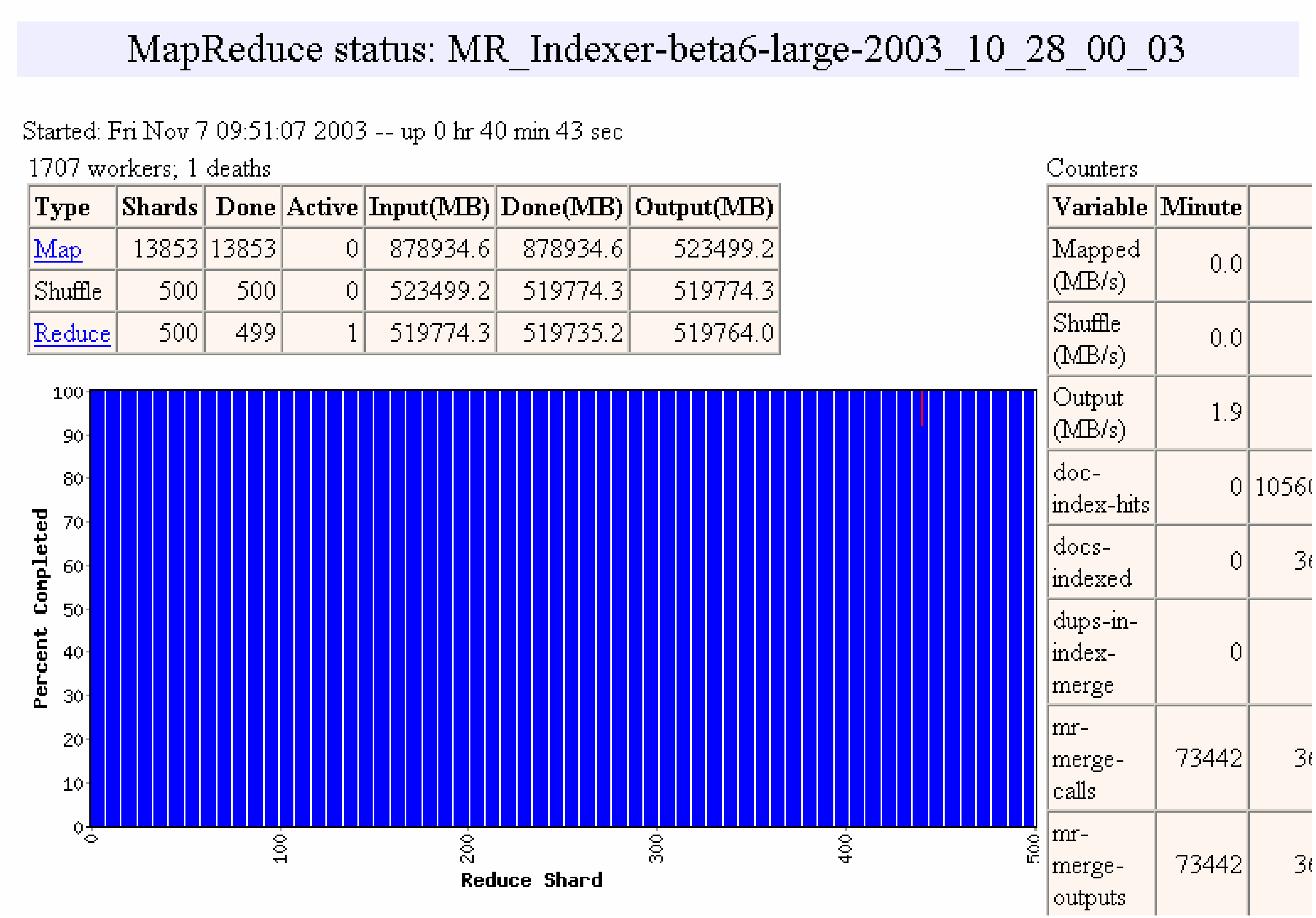
Started: Fri Nov 7 09:51:07 2003 -- up 0 hr 38 min 56 sec

1707 workers; 1 deaths

Type	Shards	Done	Active	Input(MB)	Done(MB)	Output(MB)
Map	13853	13853	0	878934.6	878934.6	523499.2
Shuffle	500	500	0	523499.2	519781.8	519781.8
Reduce	500	498	2	519781.8	519394.7	519440.7

Counters

Variable	Minute	
Mapped (MB/s)	0.0	
Shuffle (MB/s)	0.0	
Output (MB/s)	9.4	
doc-index-hits	0	10560
docs-indexed	0	36
dups-in-index-merge	0	
mr-merge-calls	394792	36
mr-merge-outputs	394792	36



MapReduce status: MR_Indexer-beta6-large-2003_10_28_00_03

Started: Fri Nov 7 09:51:07 2003 -- up 0 hr 40 min 43 sec

1707 workers; 1 deaths

Type	Shards	Done	Active	Input(MB)	Done(MB)	Output(MB)
Map	13853	13853	0	878934.6	878934.6	523499.2
Shuffle	500	500	0	523499.2	519774.3	519774.3
Reduce	500	499	1	519774.3	519735.2	519764.0

Counters

Variable	Minute	
Mapped (MB/s)	0.0	
Shuffle (MB/s)	0.0	
Output (MB/s)	1.9	
doc-index-hits	0	1056
docs-indexed	0	3
dups-in-index-merge	0	
mr-merge-calls	73442	3
mr-merge-outputs	73442	3

MapReduce: Uses at Google

Broad applicability has been a pleasant surprise

- Quality experiments, log analysis, machine translation, ad-hoc data processing, …
- Production indexing system: rewritten w/ MapReduce

  ~10 MapReductions, *much* simpler than old code

Two week period in Aug 2004:

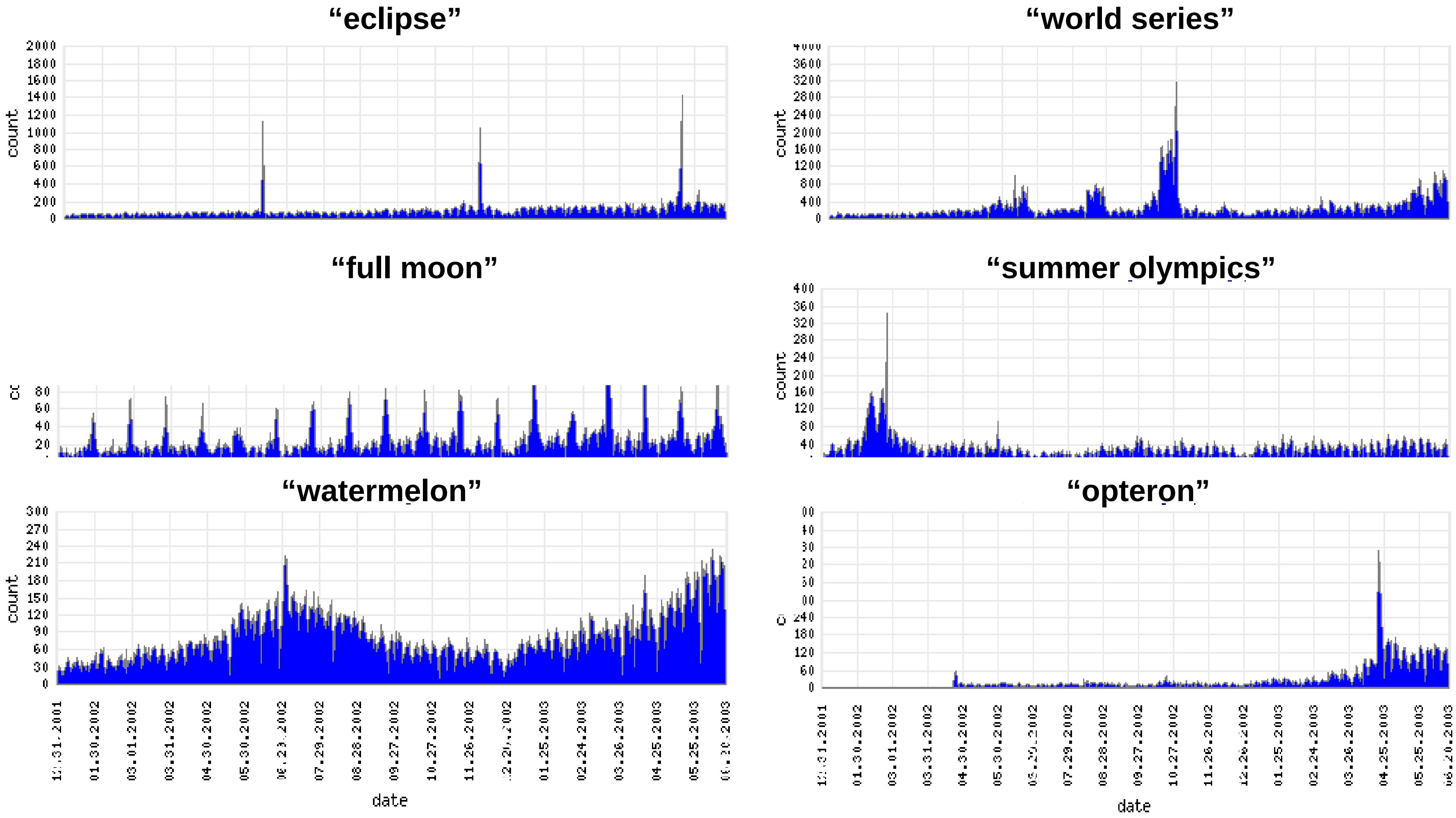
- ~8,000 MapReduce jobs, >450 different MR operations
- Read ~1500 TB of input to produce ~150 TB of output
- ~36,000 machine days, >26,000 worker deaths

"MapReduce: Simplified Data Processing on Large Clusters" to appear in OSDI'04

Data + CPUs = Playground

- Substantial fraction of Internet available for processing
- Easy-to-use teraflops and petabytes
- High-level abstractions, lots of reusable code
- Cool problems, great colleagues

Query Frequency Over Time



“eclipse”

“world series”

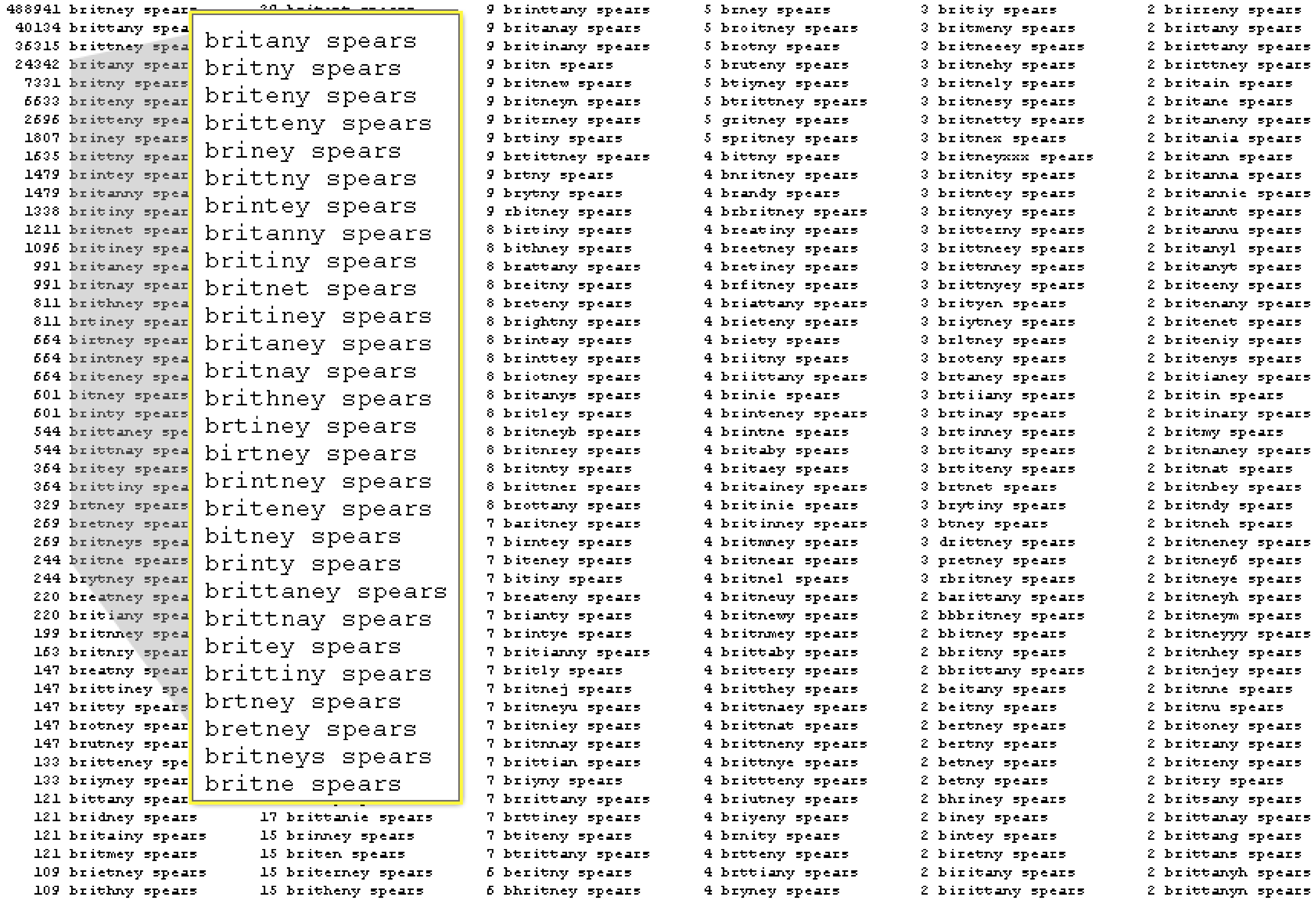
“full moon”

“summer olympics”

“watermelon”

“opteron”

Searching for Britney Spears



```
488941 britney spears
 40134 brittany spea
 36315 brittney spea
 24342 britany spear
  7331 britny spears
  6633 briteny spear
  2696 britteny spea
  1807 briney spears
  1635 brittny spear
  1479 brintey spear
  1479 britanny spea
  1338 britiny spear
  1211 britnet spear
  1096 britiney spea
   991 britaney spea
   991 britnay spear
   811 brithney spea
   811 brtiney spear
   664 birtney spear
   664 brintney spea
   664 briteney spea
   601 bitney spears
   601 brinty spears
   544 brittaney spe
   544 brittnay spea
   364 britey spears
   364 brittiny spea
   329 brtney spears
   269 bretney spear
   269 britneys spea
   244 britne spears
   244 brytney spear
   220 breatney spea
   220 britiany spea
   199 britnney spea
   163 britnry spear
   147 breatny spear
   147 brittiney spe
   147 britty spears
   147 brotney spear
   147 brutney spear
   133 britteney spe
   133 briyney spear
   121 bittany spear
   121 bridney spears
   121 britainy spears
   121 britmey spears
   109 brietney spears
   109 brithny spears
```

```
britany spears
britny spears
briteny spears
britteny spears
briney spears
brittny spears
brintey spears
britanny spears
britiny spears
britnet spears
britiney spears
britaney spears
britnay spears
brithney spears
brtiney spears
birtney spears
brintney spears
briteney spears
bitney spears
brinty spears
brittaney spears
brittnay spears
britey spears
brittiny spears
brtney spears
bretney spears
britneys spears
britne spears
```

```
17 brittanie spears
15 brinney spears
15 briten spears
15 briterney spears
15 britheny spears
```

```
9 brinttany spears
9 britanay spears
9 britinany spears
9 britn spears
9 britnew spears
9 britneyn spears
9 britrney spears
9 brtiny spears
9 brtittney spears
9 brtny spears
9 brytny spears
9 rbitney spears
8 birtiny spears
8 bithney spears
8 brattany spears
8 breitny spears
8 breteny spears
8 brightny spears
8 brintay spears
8 brinttey spears
8 briotney spears
8 britanys spears
8 britley spears
8 britneyb spears
8 britnrey spears
8 britnty spears
8 brittner spears
8 brottany spears
7 baritney spears
7 birntey spears
7 biteney spears
7 bitiny spears
7 breateny spears
7 brianty spears
7 brintye spears
7 britianny spears
7 britly spears
7 britnej spears
7 britneyu spears
7 britniey spears
7 britnnay spears
7 brittian spears
7 briyny spears
7 brrittany spears
7 brttiney spears
7 btiteny spears
7 btrittany spears
6 beritny spears
6 bhritney spears
```

```
5 brney spears
5 broitney spears
5 brotny spears
5 bruteny spears
5 btiyney spears
5 btrittney spears
5 gritney spears
5 spritney spears
4 bittny spears
4 bnritney spears
4 brandy spears
4 brbritney spears
4 breatiny spears
4 breetney spears
4 bretiney spears
4 brfitney spears
4 briattany spears
4 brieteny spears
4 briety spears
4 briitny spears
4 briittany spears
4 brinie spears
4 brinteney spears
4 brintne spears
4 britaby spears
4 britaey spears
4 britainey spears
4 britinie spears
4 britinney spears
4 britmney spears
4 britnear spears
4 britnel spears
4 britneuy spears
4 britnewy spears
4 britnmey spears
4 brittaby spears
4 brittery spears
4 britthey spears
4 brittnaey spears
4 brittnat spears
4 brittneny spears
4 brittnye spears
4 britttteny spears
4 briutney spears
4 briyeny spears
4 brnity spears
4 brtteny spears
4 brttiany spears
4 bryney spears
```

```
3 britiy spears
3 britmeny spears
3 britneeey spears
3 britnehy spears
3 britnely spears
3 britnesy spears
3 britnetty spears
3 britnex spears
3 britneyxxx spears
3 britnity spears
3 britntey spears
3 britnyey spears
3 britterny spears
3 brittneey spears
3 brittnney spears
3 brittnyey spears
3 brityen spears
3 briytney spears
3 brltney spears
3 broteny spears
3 brtaney spears
3 brtiiany spears
3 brtinay spears
3 brtinney spears
3 brtitany spears
3 brtiteny spears
3 brtnet spears
3 brytiny spears
3 btney spears
3 drittney spears
3 pretney spears
3 rbritney spears
2 barittany spears
2 bbbritney spears
2 bbitney spears
2 bbritny spears
2 bbrittany spears
2 beitany spears
2 beitny spears
2 bertney spears
2 bertny spears
2 betney spears
2 betny spears
2 bhriney spears
2 biney spears
2 bintey spears
2 biretny spears
2 biritany spears
2 birittany spears
```

```
2 brirreny spears
2 brirtany spears
2 brirttany spears
2 brirttney spears
2 britain spears
2 britane spears
2 britaneny spears
2 britania spears
2 britann spears
2 britanna spears
2 britannie spears
2 britannt spears
2 britannu spears
2 britanyl spears
2 britanyt spears
2 briteeny spears
2 britenany spears
2 britenet spears
2 briteniy spears
2 britenys spears
2 britianey spears
2 britin spears
2 britinary spears
2 britmy spears
2 britnaney spears
2 britnat spears
2 britnbey spears
2 britndy spears
2 britneh spears
2 britneney spears
2 britney6 spears
2 britneye spears
2 britneyh spears
2 britneym spears
2 britneyyy spears
2 britnhey spears
2 britnjey spears
2 britnne spears
2 britnu spears
2 britoney spears
2 britrany spears
2 britreny spears
2 britry spears
2 britsany spears
2 brittanay spears
2 brittang spears
2 brittans spears
2 brittanyh spears
2 brittanyn spears
```

Enough Data to Learn

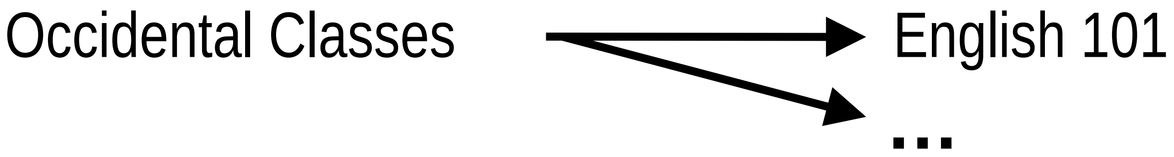
Goal: Better conceptual understanding

Query: [Pasadena english courses]

Should match:

Pasadena City College Night Class
“American Literature”

Caltech Humanities Course
“Creative Writing: Short Stories”



Occidental Classes

English 101

...

Correlation Clustering of Words

Model trained on millions of documents

Completely unsupervised learning

Learning uses many CPU years

Learned ~500K clusters: some tiny, some huge

Clusters named automatically

How much information is out there?

- How large is the Web?
  - Tens of billions of documents? Hundreds?
  - ~10KB/doc => 100s of Terabytes
- Then there’s everything else
  - Email, personal files, closed databases, broadcast media, print, etc.
- Estimated 5 Exabytes/year (growing at 30%)*
- Web is just a tiny starting point

Source: How much information 2003

Google takes it’s mission seriously

- Started with the Web (html)
- Added various document formats
- Images
- Commercial data: ads and shopping (Froogle)
- Enterprise (corporate data)
- News
- Email (Gmail)
- Scholarly publications (http://scholar.google.com)
- Local information
- Maps
- Yellow pages
- Satellite images
- Instant messaging and VoIP
- Communities (Orkut)
- Printed media
- Classified ads
- …

The other datacenter: your home

Data growing at 800 MB/year/person (~8 Petabytes/yr)

As the organization is automated, horizon moves back

Internet users growing at ~20%/year

Bandwidth increases triggers storage increase

…

Our reliance on this information increases

Availability, reliability, security needs ~corporate needs

Emergence of commodity devices and services awaited

Who Does All This?

googler = designer & computer scientist & programmer & entrepreneur

- Talented, motivated people

  … working in small teams (3-5 people)

  … on problems that matter

  … with freedom to explore their ideas

  “20% rule”, access to computational resources

- It’s not just search! Google has experts in…

  Hardware, networking, distributed systems, fault tolerance, data structures, algorithms, machine learning, information retrieval, AI, user interfaces, compilers, programming languages, statistics, product design, mechanical eng., …

Engineering culture – Hire Carefully

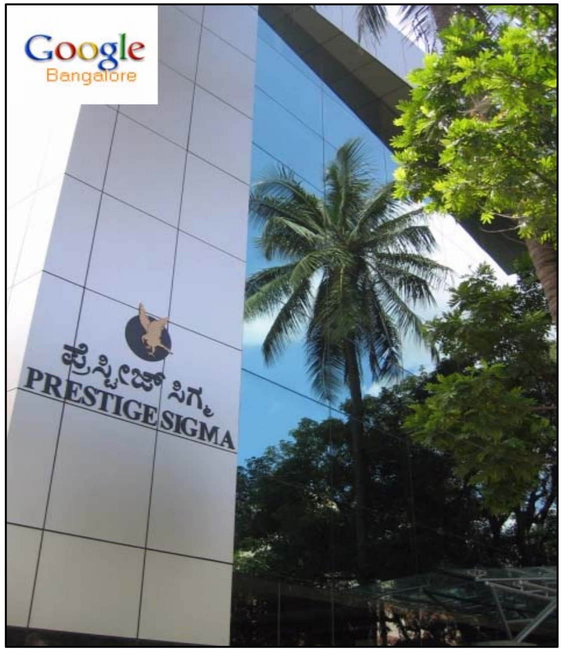
- Computer Scientists: Understand how
- Experts: Know the state of the art
- Builders: Can translate ideas to reality
- Tinkerers: Ask why not
- Diverse: CS, EE, Hardware, Neuro Surgeons, Robotics, …

Engineering culture – Everyone Innovates

- 20% Time: Management does not know best
- Small Teams: If it can be done, can be done by a few
- Take Risks: Projects with high risk and high impact
- Prepare to fail: No stigma, experiment rapidly
- Blur Roles: Engineering has more PhDs than Research

Engineering culture – User Focused Research

- Singular focus on the user
- Engineering does not worry about money
- Entrepreneurship encouraged
- Roll baby roll



About Google India

Charter to Innovate

Google Bangalore is building future Google products

Conceive locally…
Implement locally…
Deploy globally

Google™

भारत