at YAHOO!®

Owen O'Malley
Yahoo! Grid Team
owen@yahoo-inc.com

Who Am I?

- Yahoo! Architect on Hadoop Map/Reduce
  - Design, review, and implement features in Hadoop
  - Working on Hadoop full time since Feb 2006
  - Before Grid team, I worked on Yahoos' WebMap
  - PhD from UC Irvine in Software Testing.
- VP of Apache for Hadoop
  - Chair of Hadoop Program Management Committee
  - Responsible for interfacing to Apache Board

Problem

- How do you scale up applications?
  - 100's of terabytes of data
  - Takes 11 days to read on 1 computer
- Need lots of cheap computers
  - Fixes speed problem (15 minutes on 1000 computers), but…
  - Reliability problems
    - In large clusters, computers fail every day
    - Cluster size is not fixed
  - Tracking and reporting on status and errors
- Need common infrastructure
  - Must be efficient and reliable

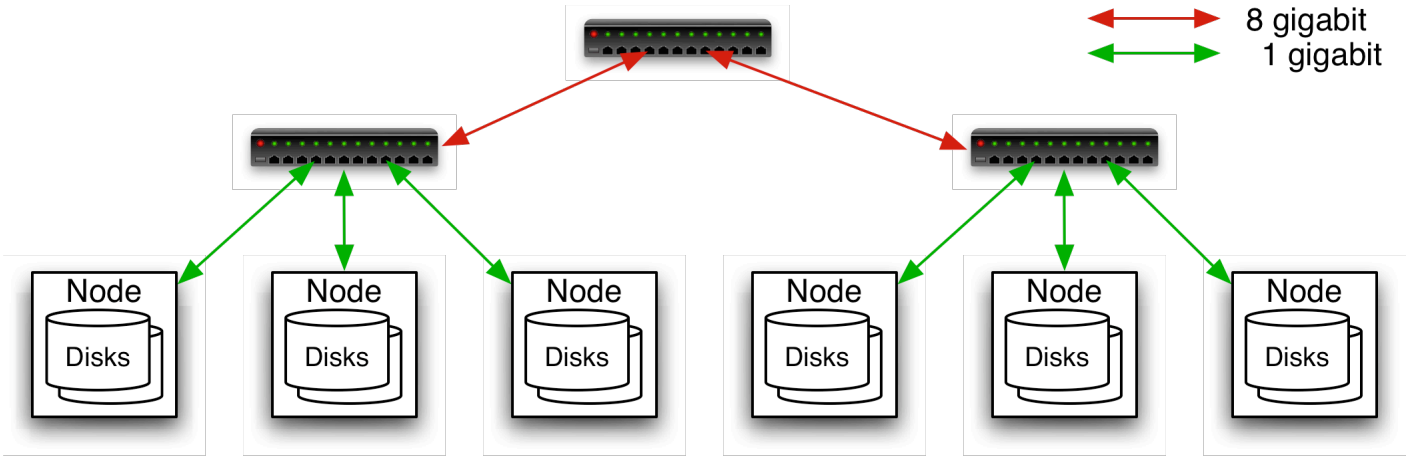
Solution

- Open Source Apache Project
- Hadoop Core includes:
  - Hadoop Distributed File System - distributed data
  - Map/Reduce – distributed application framework
- Started as distribution framework for Nutch
- Named after Doug's son's stuffed elephant.
- Written in Java and runs on
  - Linux, Mac OS/X, Windows, and Solaris

What is Hadoop NOT?

- Hadoop is aimed at moving large amounts of data efficiently.
- It is not aimed at doing real-time reads or updates.
- Hadoop moves data like a freight train, slow to start but very high bandwidth.
- Databases can answer queries quickly, but can't match the bandwidth.

Typical Hadoop Cluster



- Commodity hardware
  - Linux PCs with local 4 disks
- Typically in 2 level architecture
  - 40 nodes/rack
  - Uplink from rack is 8 gigabit
  - Rack-internal is 1 gigabit all-to-all

Distributed File System

- Single petabyte file system for entire cluster
  - Managed by a single *namenode*.
  - Files are written, read, renamed, deleted, but append-only.
  - Optimized for streaming reads of large files.
- Files are broken in to large blocks.
  - Transparent to the client
  - Blocks are typically 128 MB
  - Replicated to several *datanodes*, for reliability
- Client library talks to both namenode and datanodes
  - Data is not sent through the namenode.
  - Throughput of file system scales nearly linearly.
- Access from Java, C, or command line.

Block Placement

- Default is 3 replicas, but settable per file
- Blocks are placed (writes are pipelined):
  - On same node
  - On different rack
  - On the other rack
- Clients read from closest replica
- If the replication for a block drops below target, it is automatically re-replicated.

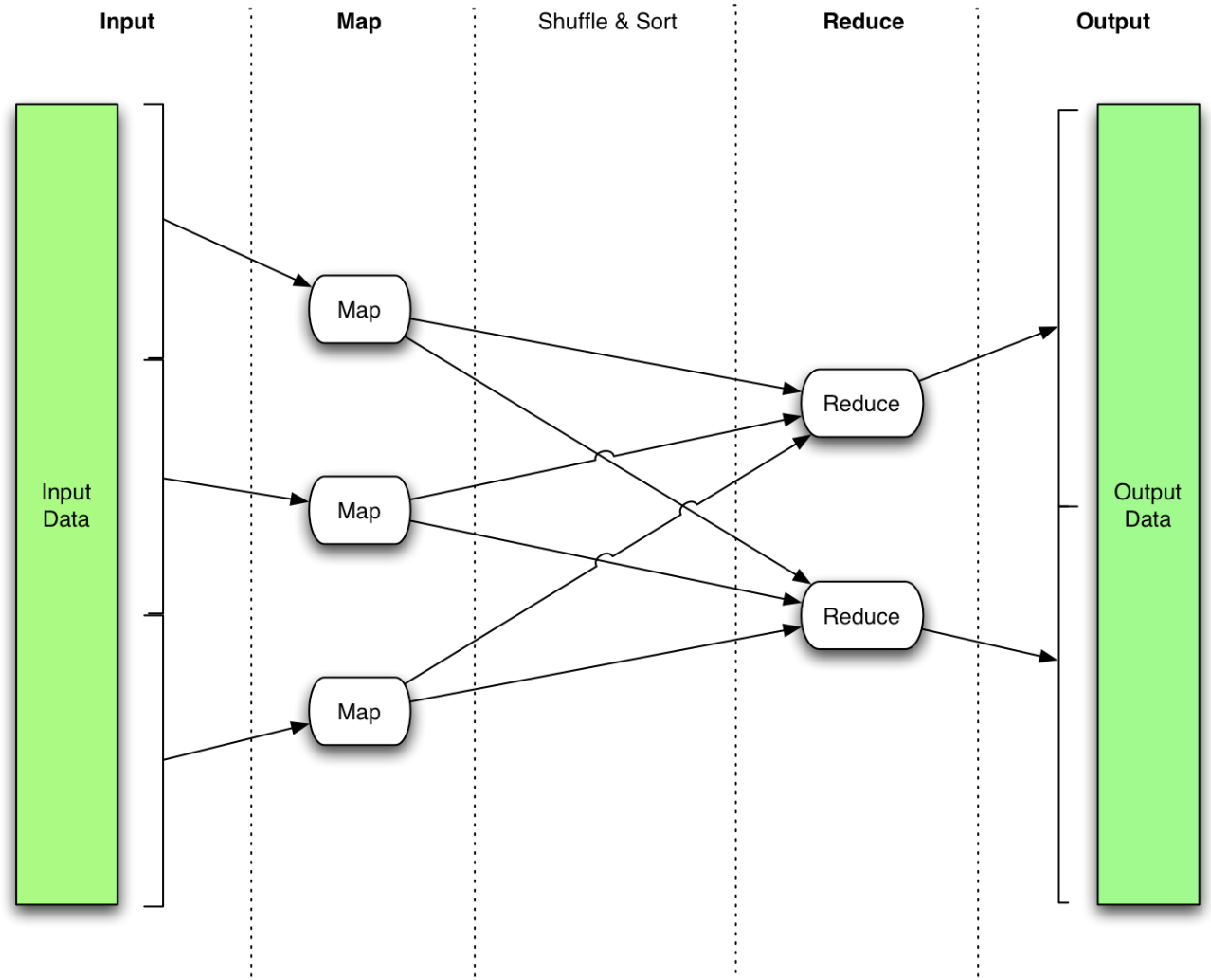
Data Correctness

- Data is checked with CRC32
- File Creation
  - Client computes checksum per 512 byte
  - DataNode stores the checksum
- File access
  - Client retrieves the data and checksum from DataNode
  - If Validation fails, Client tries other replicas
- Periodic validation by DataNode

Map/Reduce

- Map/Reduce is a programming model for efficient distributed computing
- It works like a Unix pipeline:
  - cat input | grep | sort | uniq -c | cat > output
  - **Input** | **Map** | Shuffle & Sort | **Reduce** | **Output**
- Efficiency from
  - Streaming through data, reducing seeks
  - Pipelining
- A good fit for a lot of applications
  - Log processing
  - Web index building
  - Data mining and machine learning

Map/Reduce Dataflow



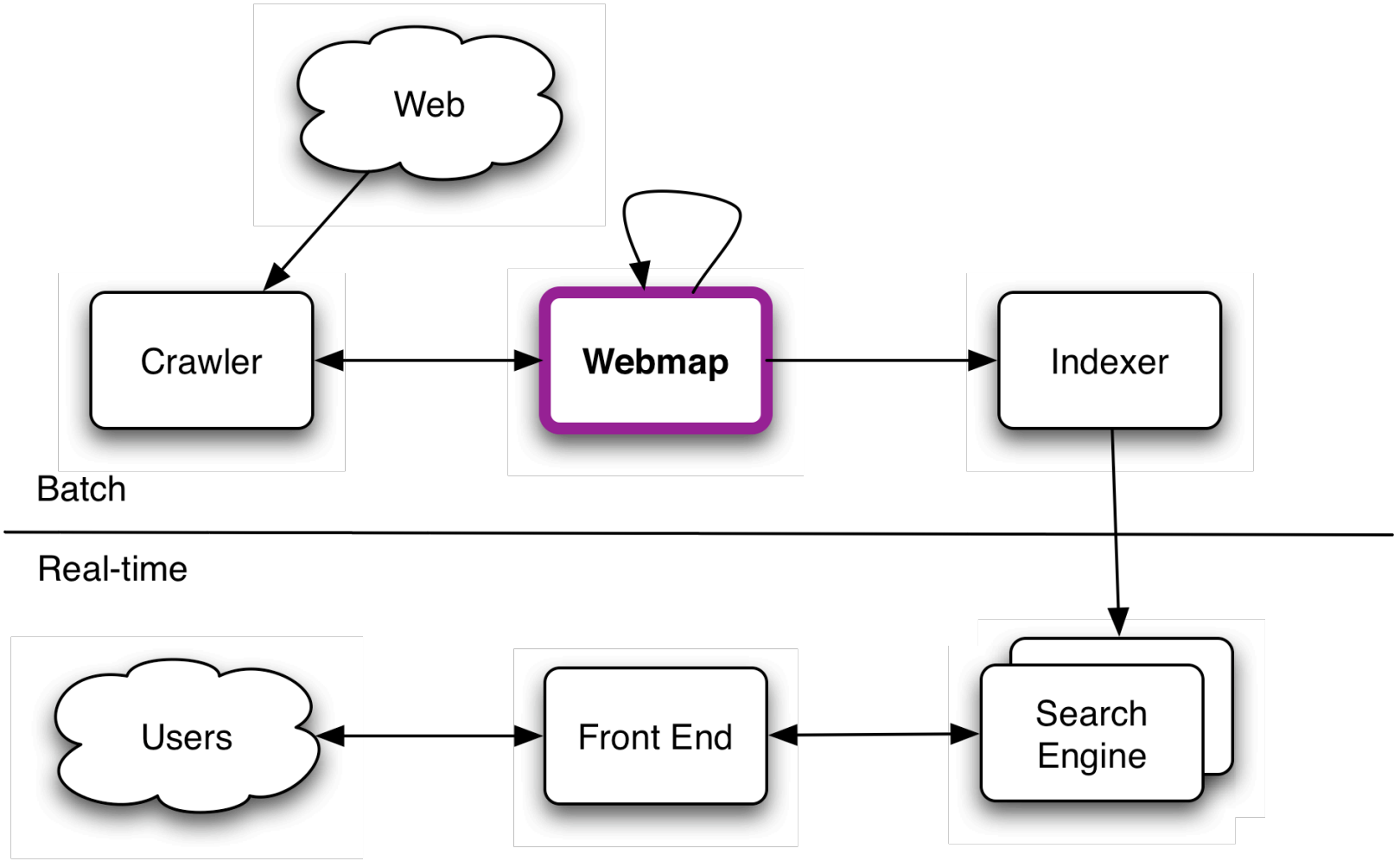
Map/Reduce features

- Java, C++, and text-based APIs
  - In Java use Objects and and C++ bytes
  - Text-based (streaming) great for scripting or legacy apps
  - Higher level interfaces: Pig, Hive, Jaql
- Automatic re-execution on failure
  - In a large cluster, some nodes are always slow or flaky
  - Framework re-executes failed tasks
- Locality optimizations
  - With large data, bandwidth to data is a problem
  - Map-Reduce queries HDFS for locations of input data
  - Map tasks are scheduled close to the inputs when possible

Why Yahoo! is investing in Hadoop

- We started with building better applications
  - Scale up web scale batch applications (search, ads, …)
  - Factor out common code from existing systems, so new applications will be easier to write
  - Manage the many clusters we have more easily
- The mission now includes research support
  - Build a **huge** data warehouse with many Yahoo! data sets
  - Couple it with a huge compute cluster and programming frameworks to make using the data easy
  - Provide this as a service to our researchers
  - We are seeing great results!
    - Experiments can be run much more quickly in this environment

Search Dataflow



Running the Production WebMap

- Search needs a graph of the “known” web
  - Invert edges, compute link text, whole graph heuristics
- Periodic batch job using Map/Reduce
  - Uses a chain of ~100 map/reduce jobs
- Scale
  - Largest known Hadoop application
  - 100 billion nodes and 1 trillion edges
  - Largest shuffle is 450 TB
  - Final output is 300 TB compressed
  - Runs on 10,000 cores
- Written mostly using Hadoop’s C++ interface

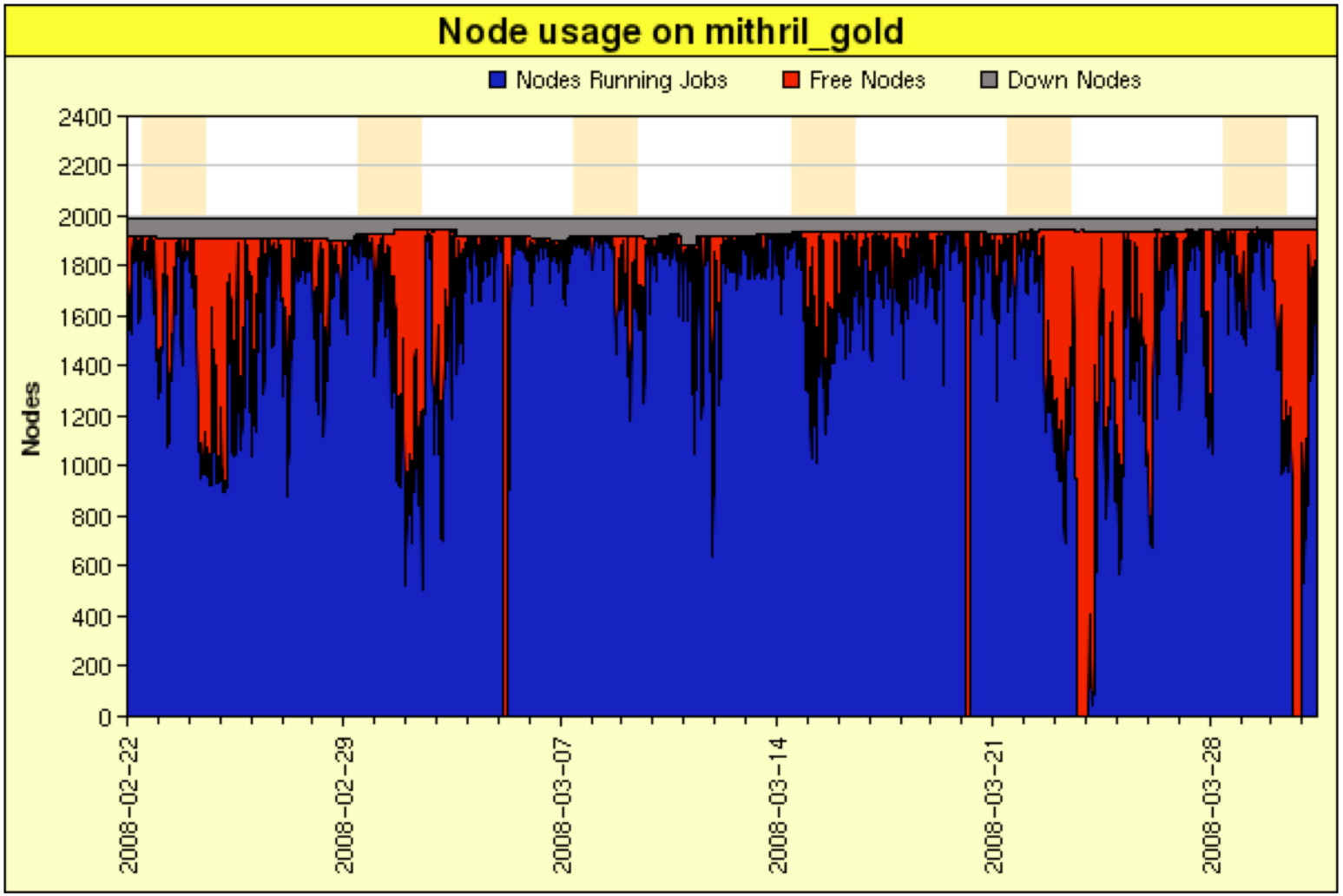
Research Clusters

- The grid team runs research clusters as a service to Yahoo researchers
  - Analytics as a Service
- Mostly data mining/machine learning jobs
- Most research jobs are *not* Java:
  - 42% Streaming
    - Uses Unix text processing to define map and reduce
  - 28% Pig
    - Higher level dataflow scripting language
  - 28% Java
  - 2% C++

Hadoop clusters

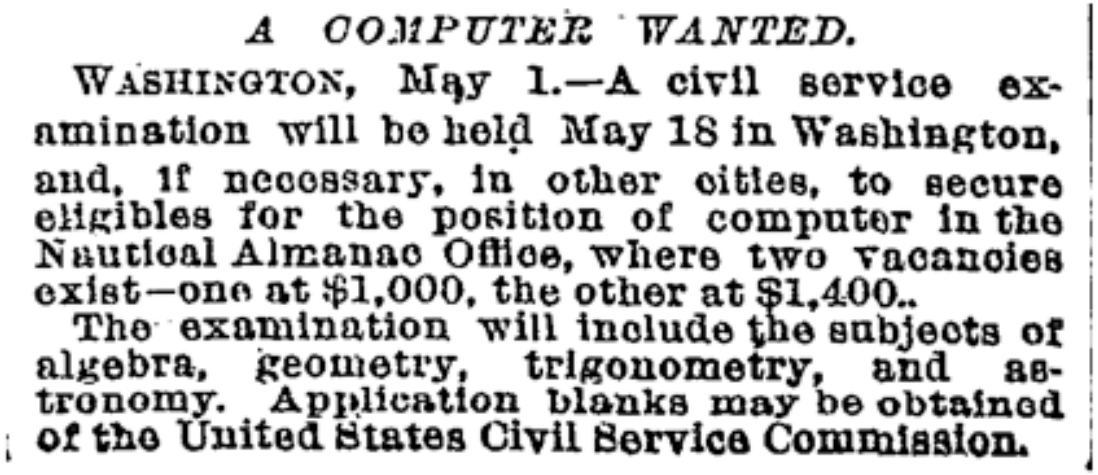
- We have ~24,000 machines in 17 clusters running Hadoop
- Our largest clusters are currently 2000-3000 nodes
- More than 10 petabytes of user data (compressed, unreplicated)
- We run hundreds of thousands of jobs every month

Research Cluster Usage



NY Times

- Needed offline conversion of public domain articles from 1851-1922.
- Used Hadoop to convert scanned images to PDF
- Ran 100 Amazon EC2 instances for around 24 hours
- 4 TB of input
- 1.5 TB of output



A COMPUTER WANTED.

WASHINGTON, May 1.—A civil service examination will be held May 18 in Washington, and, if necessary, in other cities, to secure eligibles for the position of computer in the Nautical Almanac Office, where two vacancies exist—one at $1,000, the other at $1,400.

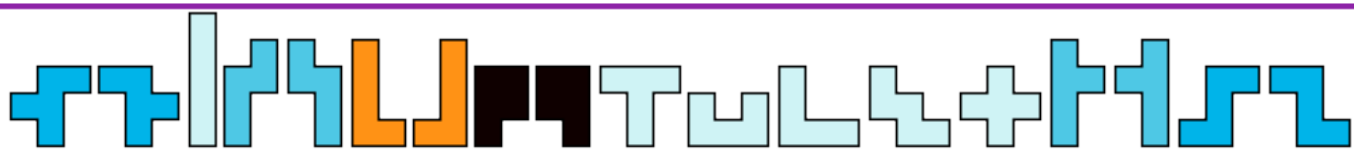
The examination will include the subjects of algebra, geometry, trigonometry, and astronomy. Application blanks may be obtained of the United States Civil Service Commission.

Published 1892, copyright New York Times

Terabyte Sort Benchmark

- Started by Jim Gray at Microsoft in 1998
- Sorting 10 billion 100 byte records
- Hadoop won general category in 209 seconds (prev was 297)
  - 910 nodes
  - 2 quad-core Xeons @ 2.0Ghz / node
  - 4 SATA disks / node
  - 8 GB ram / node
  - 1 gb ethernet / node and 8 gb ethernet uplink / rack
  - 40 nodes / rack
- Only hard parts were:
  - Getting a total order
  - Converting the data generator to map/reduce
- http://developer.yahoo.net/blogs/hadoop/2008/07

Tiling Pentominos

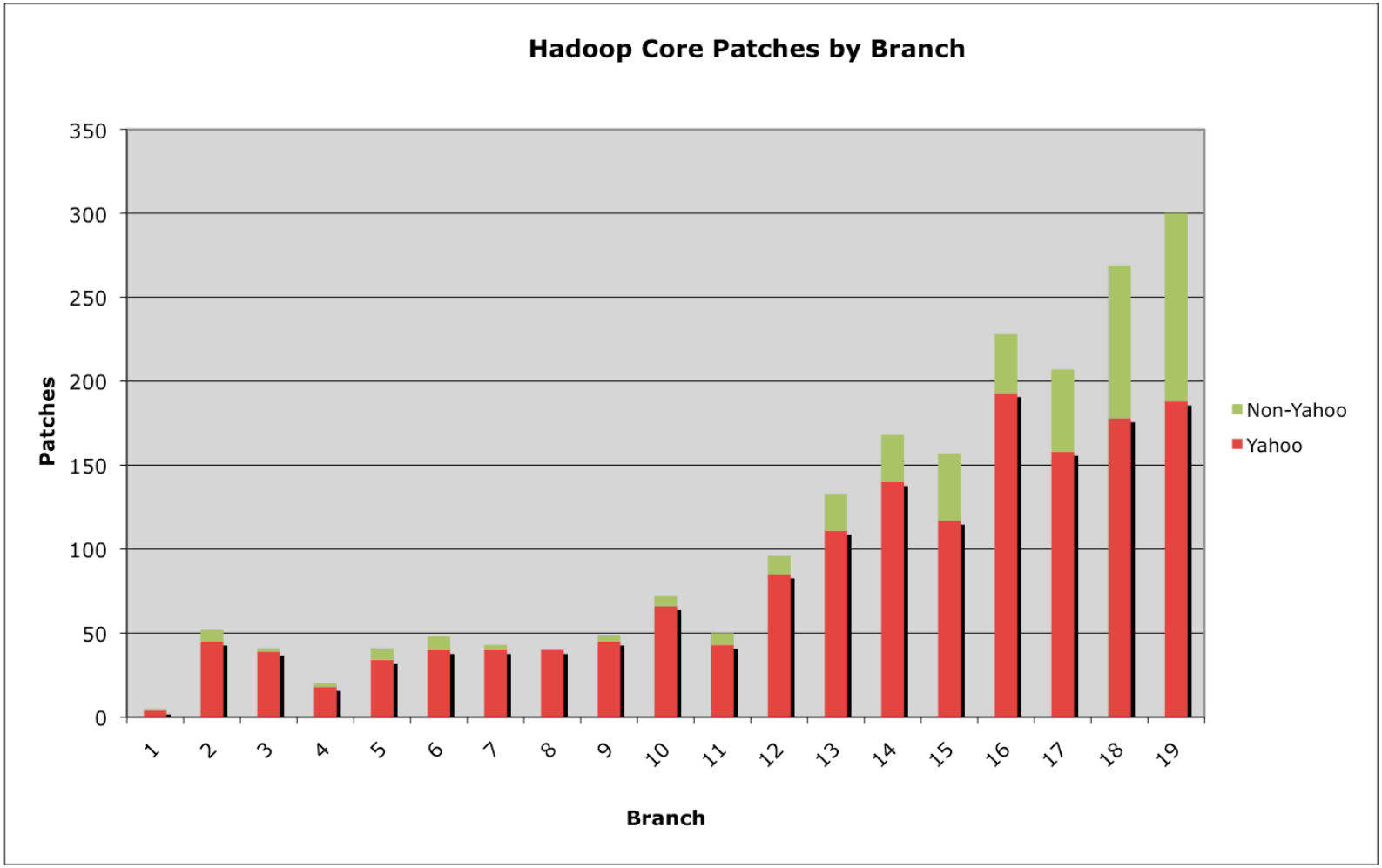


- Use the one-sided pentominos to fill a box.
- Find all possible solutions using back-tracking, just needs lots of cpu time.
- Knuth tried, but didn't want to spend months finding all 9x10 solutions.
- With Hadoop on an old small cluster (12 nodes), it ran in 9 hours.
- Generate all moves to a depth of 5 and split between maps.

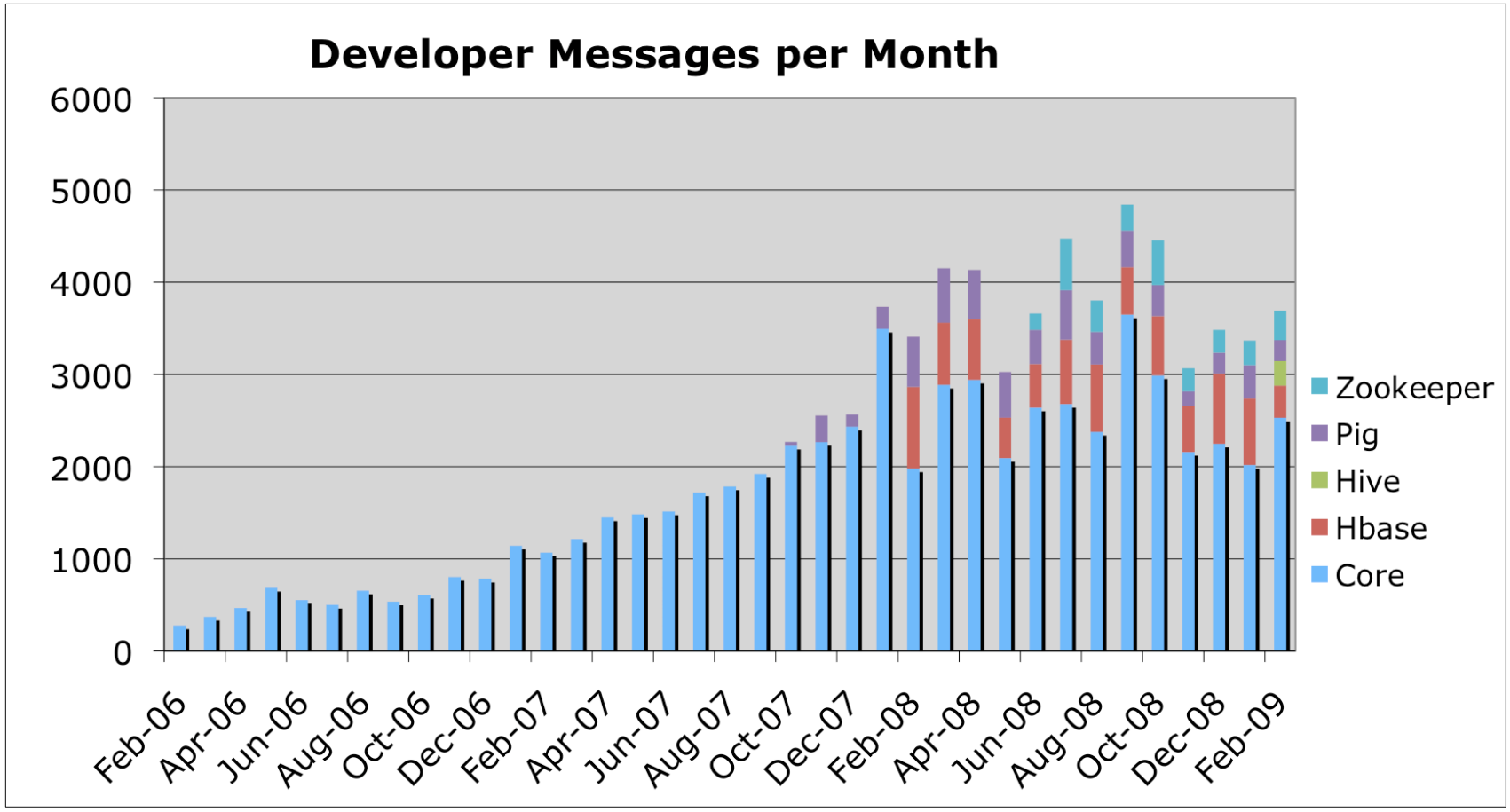
Hadoop Community

- Apache is focused on project communities
  - Users
  - Contributors
    - write patches
  - Committers
    - can commit patches **too**
  - Project Management Committee
    - vote on new committers and releases **too**
- Apache is a meritocracy
- Use, contribution, and diversity is growing
  - But we need and want more!

Size of Releases



Size of Developer Community



Who Uses Hadoop?

- Amazon/A9
- AOL
- Baidu
- Facebook
- IBM
- Joost
- Last.fm
- New York Times
- PowerSet (now Microsoft)
- Quantcast
- Universities
- Veoh
- Yahoo!
- More at http://wiki.apache.org/hadoop/PoweredBy

What's Next?

- 0.20
  - New map/reduce API
  - Better scheduling for sharing between groups
- Moving toward Hadoop 1.0
  - RPC and data format versioning support
    - Server and clients may be different versions
    - Master and slaves are still consistent versions.
  - HDFS and Map/Reduce security
  - Upward API compatibility from 1.0 until 2.0

Hadoop Subprojects

- Chukwa – Cluster monitoring
- Core – Common infrastructure
  - HDFS
  - Map/Reduce
- HBase – BigTable
- Hive – SQL-like queries converted into Map/Reduce
- Pig – High level scripting language into Map/Reduce
- Zookeeper – Distributed coordination

Q&A

- For more information:
  - Website: http://hadoop.apache.org/core
  - Mailing lists:
    - $project-dev@hadoop.apache.org
    - $project-user@hadoop.apache.org
    - general@hadoop.apache.org
  - IRC: #hadoop on irc.freenode.org