The Challenges and Opportunities in Interfacing Hadoop with Condor

Miron Livny
Center for High Throughput Computing
Morgridge Institute for Research and
University of Wisconsin-Madison

MORGRIDGE INSTITUTE FOR RESEARCH AT THE UNIVERSITY OF WISCONSIN-MADISON

CENTER FOR HIGH THROUGHPUT COMPUTING

THE UNIVERSITY of WISCONSIN MADISON


Academia
Enterprise
IT Comp.
Technology
Adoption

Hadoop = MR + HDFS

where

MR « HDFS*

***6 Tier-2 sites of US-CMS are operating 2PB HDFS facilities each**

Condor-Users 11/11

I have some users who are interested in running Hadoop jobs on our Condor cluster … I'm wondering if anyone can point me to some more detailed information on how to get this going? … David Brodbeck, System Administrator, Linguistics University of Washington

I have started to use it. I think map reduce is irrelevant when you use condor. However, HDFS is extemly useful for streaming large data (filer bigger than 300mb). Rita rmorgan466@gmail.com

Its a phase. Now days everyone asks, Does it have Hadoop? people not even knowing what it is and does. Granted, HDFS is nice. Mag Gam magawake@gmail.com

'Condor' brings genome assembly down to Earth

July 19, 2010 | by Chris Barncard

A team of computer scientists from the University of Wisconsin-Madison and the University of Maryland recently assembled a full human genome from millions of pieces of data — stepping up from commonly assembled genomes several orders of magnitude less complex — and they did it without a big-ticket supercomputer. …

"It's two plus two equals five, if you will," Tannenbaum says. "**Condor integrated with Hadoop** is a software system powerful enough to tackle problems as complex as human genome assembly …

High Throughput Computing

We first introduced the distinction between High Performance Computing (HPC) and High Throughput Computing (**HTC**) in a seminar at the NASA Goddard Flight Center in July of **1996** and a month later at the European Laboratory for Particle Physics (CERN). In June of 1997 HPCWire published an interview on High Throughput Computing.

HIGH THROUGHPUT COMPUTING: AN INTERVIEW WITH MIRON LIVNY 06.27.97
by Alan Beck, editor in chief HPCwire
==========

This month, NCSA's (National Center for Supercomputing Applications) Advanced Computing Group (ACG) will begin testing Condor, a software system developed at the University of Wisconsin that promises to expand computing capabilities through efficient capture of cycles on idle machines. The software, operating within an HTC (High Throughput Computing) rather than a traditional HPC (High Performance Computing) paradigm, organizes machines

Why HTC?

For many experimental scientists, scientific progress and quality of research are strongly linked to computing throughput. In other words, they are less concerned about instantaneous computing power. Instead, what matters to them is the amount of computing they can harness over a month or a year --- they measure computing power in units of scenarios per day, wind patterns per week, instructions sets per month, or crystal configurations per year.

High Throughput Computing is a 24-7-365 activity

FLOPY ≠ *(60*60*24*7*52)*FLOPS*

Paper

Interfacing Condor and PVM to harness the cycles of workstation clusters

Jim Pruyne , , **Miron Livny**

Department of Computer Sciences, University of Wisconsin—Madison, Madison, WI 53 706, USA

Received 7 November 1995; Accepted 4 December 1995. Available online 16 February 1999.

Abstract

A continuing challenge to the scientific research and engineering communities is how to fully utilize computational hardware. In particular, the proliferation of clusters of high performance workstations has become an increasingly attractive source of compute power. Developments to take advantage of this environment have previously focused primarily on managing the resources, or on providing interfaces so that a

Map Reduce brought back (legitimized) distributed computing (dynamic number of processes) that we had with PVM and lost when MPI took over

We always believed in communicating via files!

NUG30 Quadratic Assignment Problem

$$\min_{p \in \Pi} \sum_{i=1}^{30} \sum_{j=1}^{30} a_{ij} b_{p(i)p(j)}$$

Science AAAS

AAAS NEW

News Home ScienceNOW S

Home > News > ScienceNOW > July

Science Video Portal

Science LIVE
Every Thursday at 3:00 p.m. EST

Upcoming:

Mysteries of the Cell
01 December

RECENT ARTICLES

Science NOW UP TO THE MINUTE NEWS FROM SCIENCE

Harnessed Computers Crack Math Puzzle

by Jocelyn Kaiser on 25 July 2000, 5:00 PM | Permanent Link | 0 Comments

Email Print | More PREVIOUS ARTICL

Another devilishly hard computational problem has cracked under the coordinated assault of computers linked by the Internet. The alliance of 2500-some computer processors took just a week to solve "nug30," a problem that had resisted mathematicians' best efforts for 3 decades.

ENLARGE

Solution Characteristics.

Scientists	**4**
Workstations	**1**
Sites	**15**
Wall Clock	**6:22:04:31**
Avg. # CPUs	**653**
Max. # CPUs	**1007**
Total CPU time	**~11 Years**
Nodes	**11,892,208,412**
Linear Assignment Problems	**574,254,156,532**
Parallel Efficiency	**92%**

Communication between Master and workers only through files. Each worker had 4 file – 2 for sending work (input) and 2 for sending results (output). Each pair had a control (append) file and a data (rewind) file.

MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

Google, Inc.

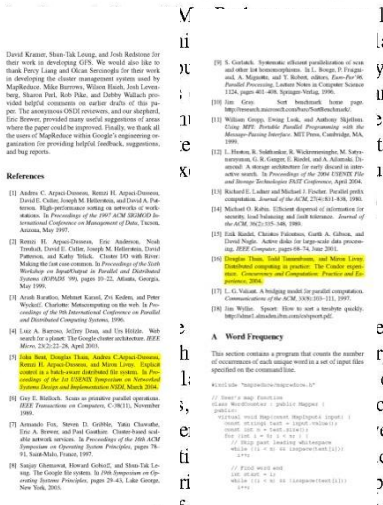
Abstract

MapReduce is a programming model
ated implementation for processing and
data sets. Users specify a *map* function
key/value pair to generate a set of interm
pairs, and a *reduce* function that merges
values associated with the same interme
real world tasks are expressible in this n
in the paper.

Programs written in this functional sty
cally parallelized and executed on a large
modity machines. The run-time system t
details of partitioning the input data, sch
gram's execution across a set of machine
chine failures, and managing the require
communication. This allows programmers without any
experience with parallel and distributed systems to eas-
ily utilize the resources of a large distributed system.

The MapReduce implementation relies on an in-house cluster management system that is responsible for distributing and running user tasks on a large collection of shared machines. Though not the focus of this paper, the cluster management system is similar in spirit to other systems such as Condor [16].

BAD-FS [5] has a very different programming model from MapReduce, and unlike MapReduce, is targeted to the execution of jobs across a wide-area network. However, there are two fundamental similarities. (1) Both systems use redundant execution to recover from data loss caused by failures. (2) Both use locality-aware scheduling to reduce the amount of data sent across congested network links.



Project Lead, Apache Hadoop Distributed File system - a Condor Team Alumnus

Dhruba Borthakur

1st

Hadoop Engineer at Facebook

San Francisco Bay Area | Computer Software

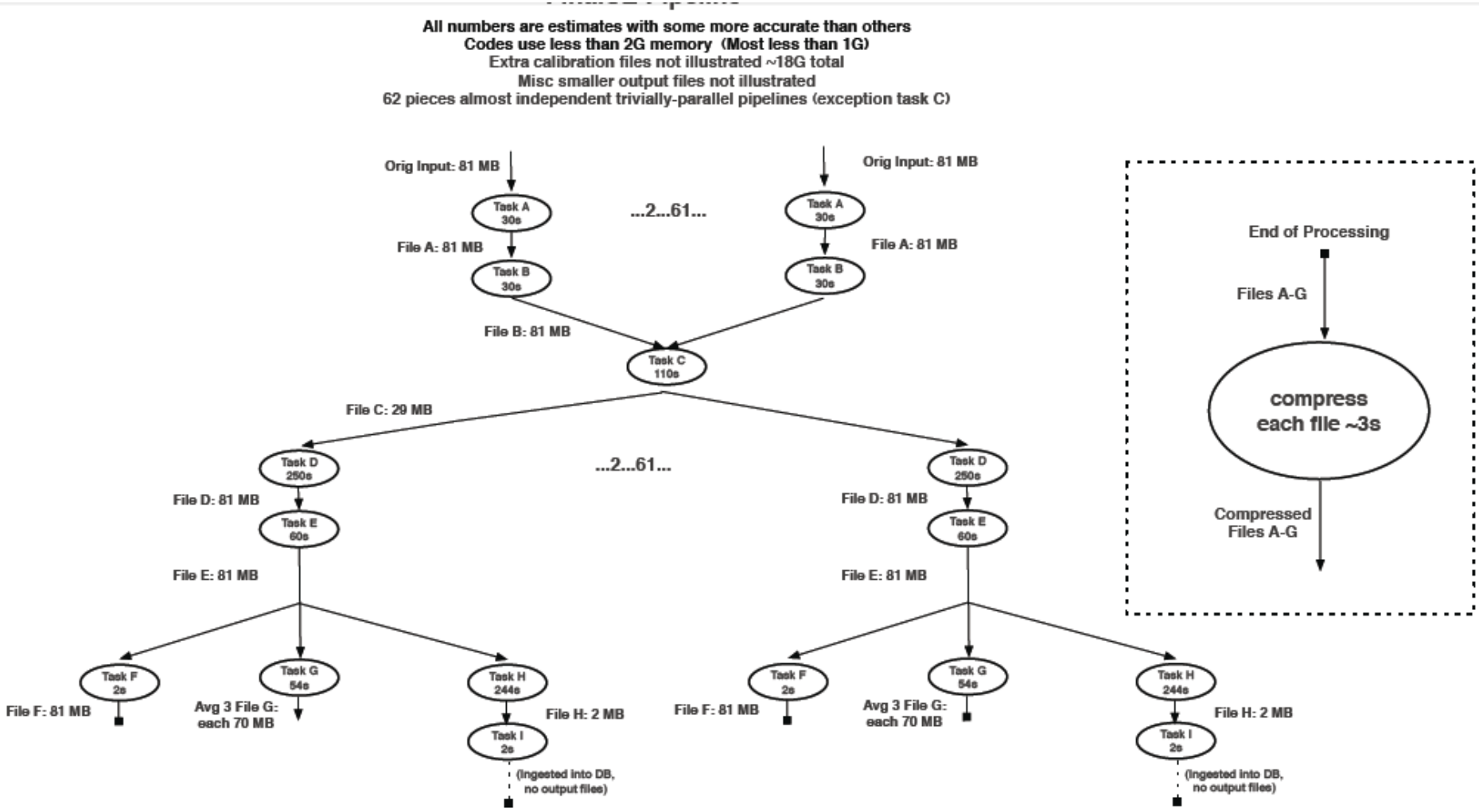
Current	**Software Engineer** at **Facebook**
Past	Principal Technical Yahoo at Yahoo
	Engineering Manager & Senior Software Engineer at Mendocino software
	Senior Member of Technical Staff at Veritas Software
	Chief Architect at Oreceipt.com
	Advanced Member of Technical Staff at IBM Transarc Labs
	Software Engineer at Center for Development of Telematics (CDOT)
Education	University of Wisconsin-Madison
	Birla Institute of Technology and Science

Running MR under Condor

- You can read all about it at https://condor-wiki.cs.wisc.edu/index.cgi/wiki?p=MapReduce
- Every thing, the Job-Tracker and the slots, are treated as Condor jobs. Slots can came-and-go as needed
- Each MR application is treated independently everything is encapsulated in a script
- At this point only mechanisms, no policies for how many slots to allocate to an MR application

More recently, we worked with Prof. Michael Sussman and George Phillips in Dept of Biochemistry at the University of Wisconsin-Madison on using **Map Reduce** with electric eel sequencing. We worked to obtain a complete sequence for the proteins encoded by the ca. 30,000 genes and comparing the transcriptome of the electric organs (weak electric organ which is used for navigating and communication and the strong voltage electric organ used for killing prey and enemies) with brain, muscle and other tissues.

Many scientists with large work/data flow applications!

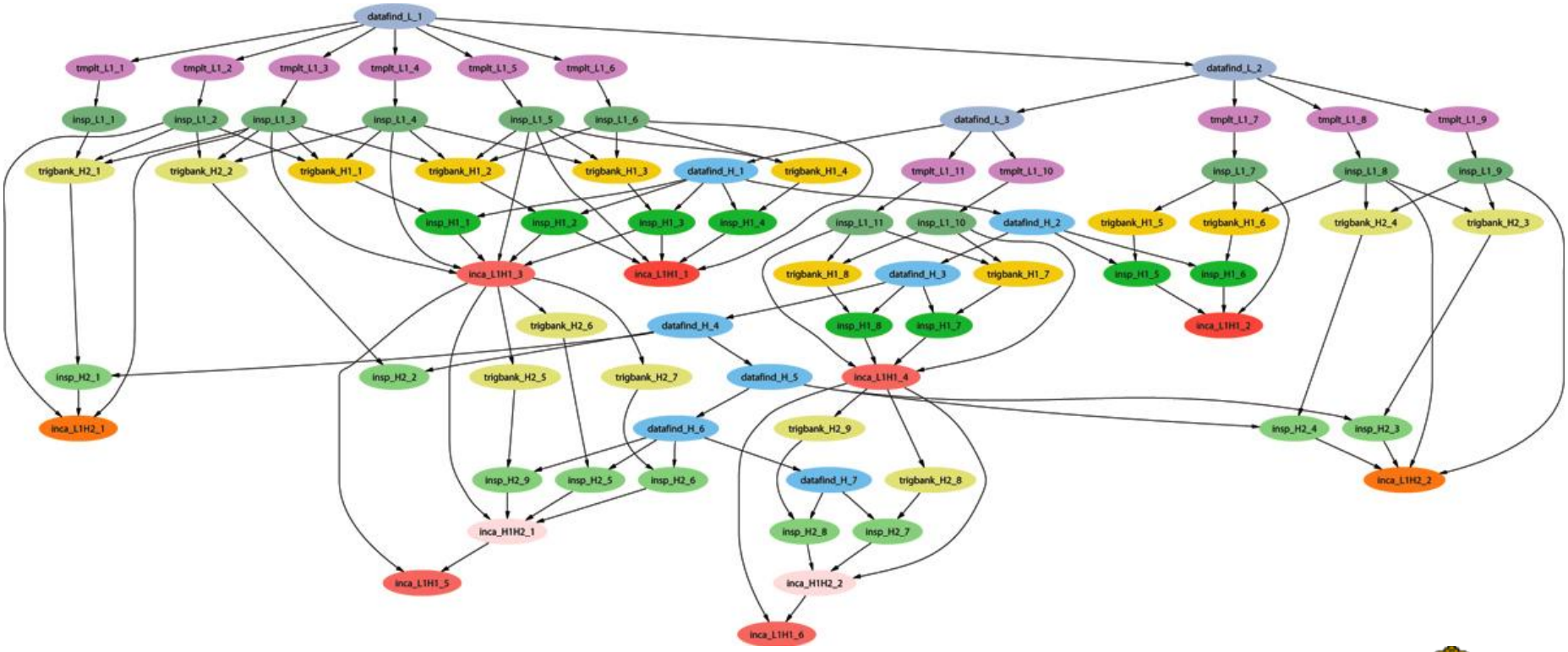


Managing Job Dependencies

15 years ago we introduced a simple language and a scheduler that use Directed Acyclic Graphs (DAGs) to capture and execute interdependent jobs. The scheduler (DAGMan) is a Condor job and interacts with the Condor job scheduler (SchedD) to run the jobs.

DAGMan has been adopted by the Laser Interferometer Gravitational Wave Observatory (LIGO) Scientific Collaboration (LSC).

Example of a LIGO Inspiral DAG



From: Stuart Anderson <anderson@ligo.caltech.edu>
Date: February 28, 2010 11:51:32 PM EST
To: Condor-LIGO mailing list <condorligo@aei.mpg.de>
Subject: [CondorLIGO] Largest LIGO workflow

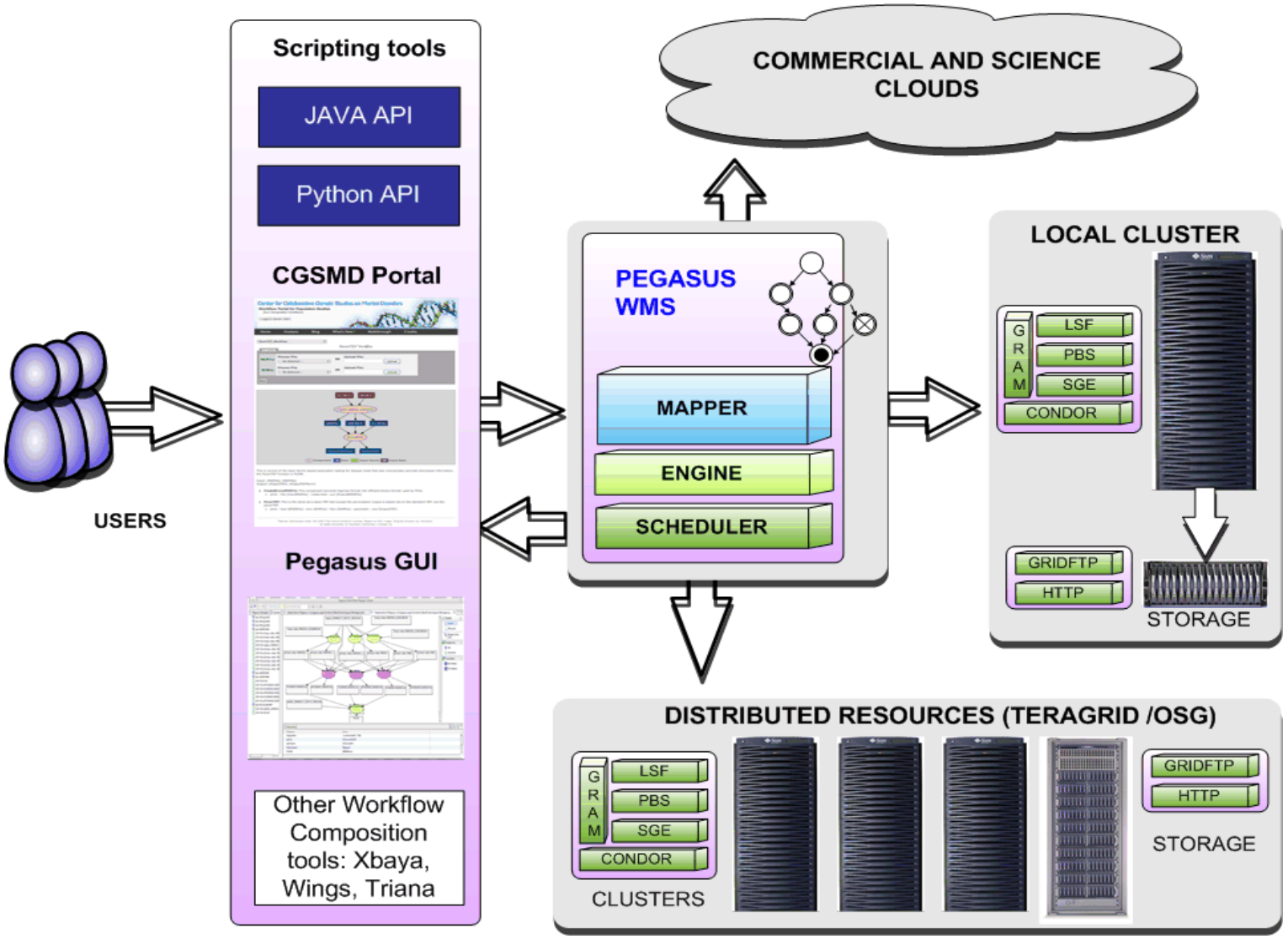
Pete,

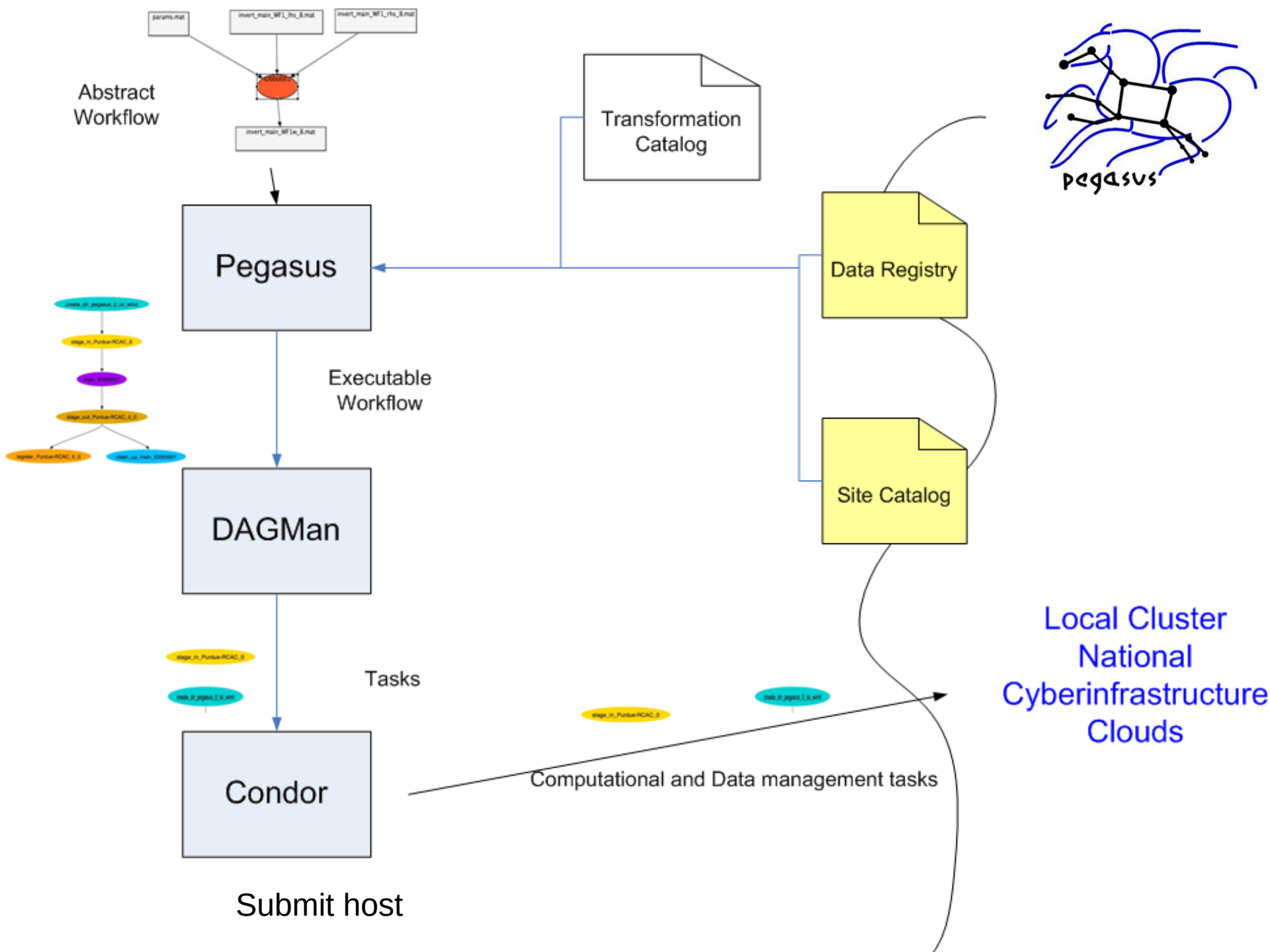
Here are some numbers you ask about for LIGO's use of DAGs to manage large data analysis tasks broken down by the largest number of jobs managed in different categories:

1) DAG Instance--one condor_dagman process: 196,862.
2) DAG Workflow--launched from a single condor_submit_dag but may include multiple automatic sub- or spliced DAGs: 1,120,659.
3) DAG Analysis--multiple instances of condor_submit_dag to analyze a common dataset with results combined into a single coherent scientific result: 6,200,000.
4) DAG Total--sum over all instances of condor dagman run: O(100M).

P.S. These are lower bounds as I did not perform an exhaustive survey/search, but they are probably close.

Pegasus Workload Management System



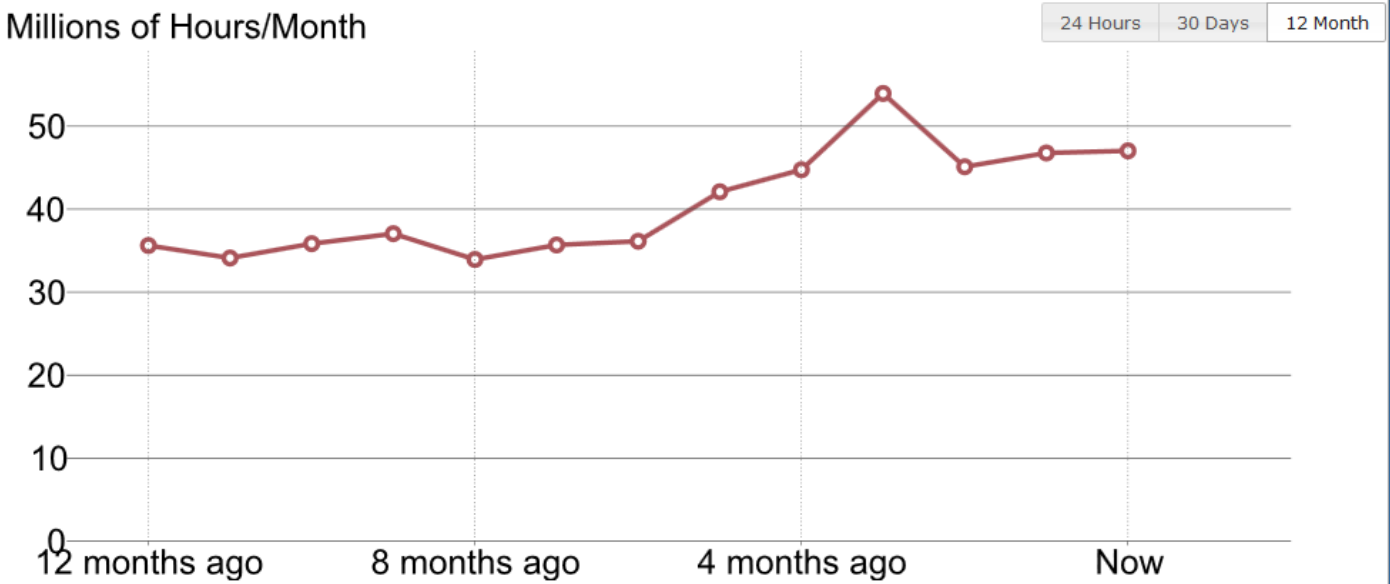

Abstract
Workflow
Transformation
Catalog
pegasus
Pegasus
Data Registry
Executable
Workflow
Site Catalog
DAGMan
Tasks
Local Cluster
National
Cyberinfrastructure
Clouds
Condor
Computational and Data management tasks
Submit host

Open Science Grid (OSG)
HTC at the National Level

Open Science Grid

A national, distributed computing partnership for data-intensive research

Jobs | CPU Hours | Transfers | TB Transferred | Status Map



CPU hours spent on an OSG resource are reported to the central accounting system. The above graph shows the number of CPU hours per month. A total of 528,164,000 CPU hours were spent.

OSG delivered across 106 sites

In the last 24 Hours	
676,000	Jobs
1,783,000	CPU Hours
1,228,000	Transfers
789	TB Transferred
In the last 30 Days	
14,990,000	Jobs
45,943,000	CPU Hours
30,041,000	Transfers
17,164	TB Transferred
In the last Year	
204,183,000	Jobs
528,164,000	CPU Hours
552,760,000	Transfers
286,193	TB Transferred

U.S. DEPARTMENT OF ENERGY

Status at **8:50 AM**
Data is loaded every 15 minutes

MORGRIDGE INSTITUTE FOR RESEARCH
AT THE UNIVERSITY OF WISCONSIN-MADISON

CENTER FOR HIGH THROUGHPUT COMPUTING

THE UNIVERSITY of WISCONSIN MADISON

HDFS and OSG

Five of the US-CMS Tier2 sites use HDFS to manage their StorgaeElement with a total capacity of more than 10PB and ~5PB of data.

- Replication is used for availability not performance
- Most of the data is cashed data
- No writes to HDFS files
- Evaluation started 10/08. First site operational 03/09.

A typical CMS Tier2 (UNL)

1M files, 2 replicas, and average about 10-20TB of data delivered to worker nodes a day. The WAN traffic averages 1-5TB, but bursts much higher upon demand. Anything that is unique to the site has 3 copies.

HDFS and Condor

- Condor deploys and manages the operation of the NameNode(s) and the DataNodes. Two types of deployment – permanent and on-the-fly. In both cases nodes come-and-go
- HDFS provides Condor with a distributed file system for input/output sandboxes, execution files, and checkpoint files. In many cases replication is driven by availability

HDFS at UW-CHTC

As part of the Grid Laboratory of Wisconsin the Center for High Throughput Computing (CHTC) is operating a Condor pool with 1.8K cores and 130TB managed by HDFS

We used HDFS plus Condor to assist Prof Chris Re in Computer Sciences to process a **half-billion web** pages to build a Web-scale demo of knowledge base construction using the statistical reasoning technologies.

Challenges/opportunities

- Monitor the health of the HDFS components
- Notifications
- Control over locations of replicas
- Storage allocation (especially for intermediate data)
- Authentication and Authorization

The words of Koheleth son of David, king in Jerusalem ~ 200 A.D.

Only that shall happen
Which has happened,
Only that occur
Which has occurred;
There is nothing new
Beneath the sun!

Ecclesiastes, (,קֹהֶלֶת*Kohelet*, "son of David, and king in Jerusalem" alias Solomon, Wood engraving Gustave Dor é(1832–1883)

Ecclesiastes Chapter 1 verse 9