

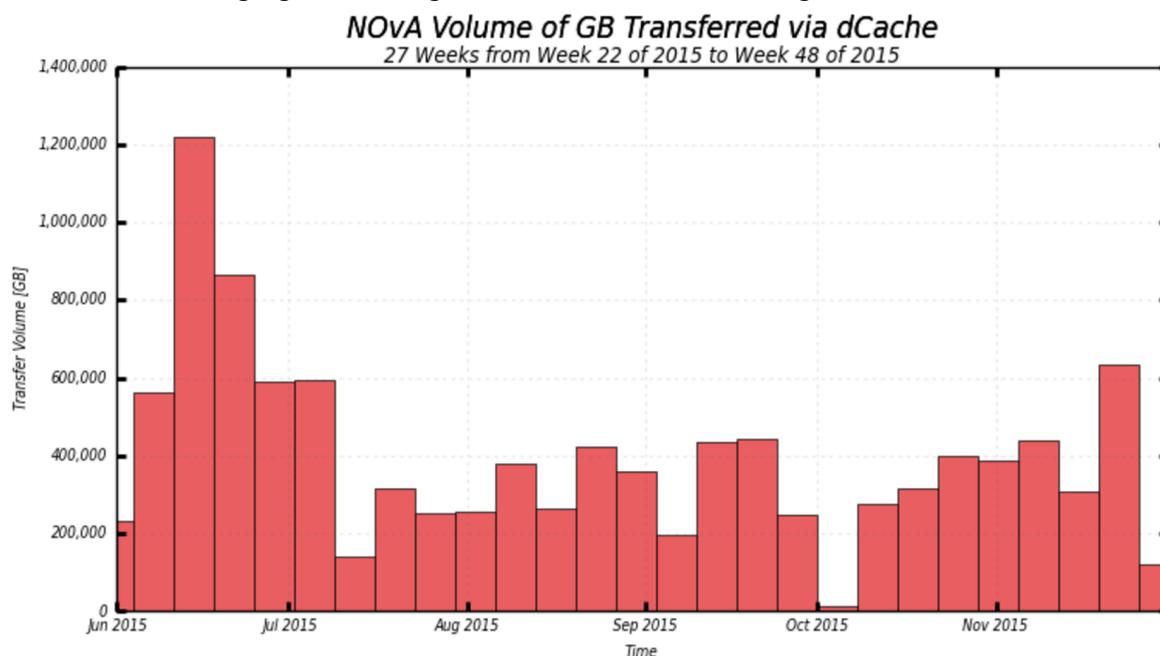
**FIFE Notes - December, 2015 News**

**for**

**Distributed Computing at Fermilab**

### NOvA success for first results

The NOvA experiment is the largest running experiment at Fermilab and this summer NOvA scientists presented their first neutrino oscillation results at the APS Division of Particles and Fields Meeting in August of 2015. Using a beam of neutrinos generated at Fermilab, NOvA studies the oscillation parameters that define the neutrinos transformation from one type to another type. The experiment compares the neutrino flavor composition observed in the 300 ton near detector and the 14,000 ton far detector in Minnesota. The computing challenges faced by the experiment involve not just processing the data, but transferring data from both the near and far detector, cataloging and storing the data, and then delivering that data to worker nodes.

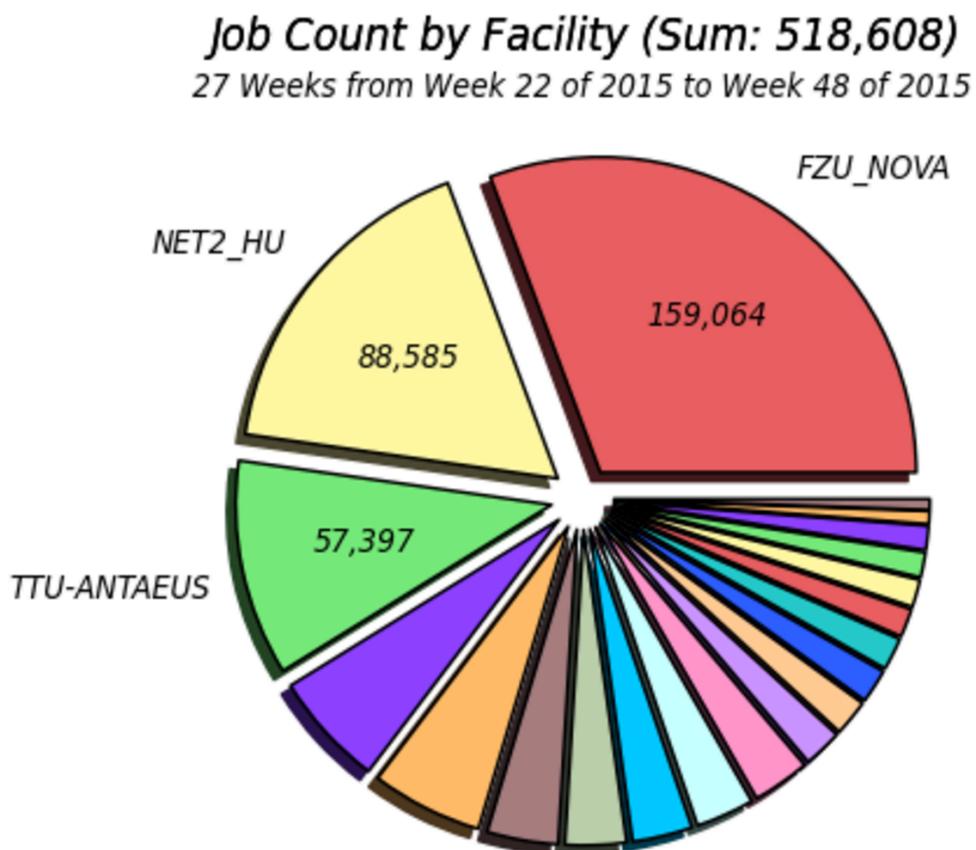


*Maximum: 1,221,329 GB, Minimum 10,249 GB, Average 395,130 GB*

To address their data handling needs, the NOvA experiment is using the File Transfer Service, SAM File Catalog and dCache/Enstore for data storage. The combination of both near and far detector transfer rate is approximately 1 Terabyte per day, but pales in comparison to the rate at which files must be delivered to worker nodes during peak analysis periods. During the preparation for DPF, dCache was delivering files at a rate of approximately 1 Terabyte per hour for analysis jobs. The utilization of FTS, SAM, and dCache allowed for complete integration into

essentially all workflows without customization by analyzers.

While data heavy processing was focused on workers nodes near (but not exclusively at) Fermilab, processing such as Monte Carlo generation was transitioned to offsite resources to increase the resources available for analyzers or processing requiring large input datasets. Three example sites were FZU, Harvard University, and Texas Tech University which provided the largest number of opportunistic jobs during the last six months to NOvA. All offsite opportunistic processing combined resulted in over 5 million CPU hours during that time and increased the average number of cores utilized by NOvA from 2200 cores onsite to 3250 cores total.



Numbers for NOvA running:

Total in the last 6 months:

14.8 Million CPU hours -> 546,000 hours per day -> 3250 cores DC  
offsite:

5.3 Million CPU hours -> 195,000 hours per day -> 1160 cores DC

Largest contributors to offsite running based on number of jobs:

- FZU in the Czech Republic
  - Harvard University
  - Texas Tech
- Mike Kirby & Ken Herner

## GENIE using OSG to improve neutrino interaction modeling

[GENIE \(Generates Events for Neutrino Interaction Experiments\)](#) is a key part of the simulation software stack for just about every neutrino experiment at Fermilab. To simulate an accelerator-based neutrino experiment, one typically breaks the software up into three parts - one which handles the production of the neutrino beam, one which handles the primary interaction of a neutrino with its target, and one to handle the propagation of the produced particles through the detector. GENIE is simulation software produced by an international collaboration of scientists that sits in the middle of that stack. It simulates both the dense nuclear matter of the target as well as the neutrino interaction dynamics.

While GENIE has been in existence since 2004, Fermilab only recently joined the collaboration, with an effort led by Gabriel Perdue that started in late 2013. One of the things that makes an institution like Fermilab an attractive member in a collaboration like GENIE is the laboratory's computing expertise and its ability to operate at scale.



*Photo courtesy Luanne O'Boyle. The Fermilab group consists of (from left to right), Robert Hatcher, Julia Yarba, Gabriel Perdue and Tomasz Golan, members of the Fermilab Physics and Detector Simulation group.*

From the beginning, one of the goals of the Fermilab GENIE group has been to move its validation processing to the Open Science Grid. Preparing a GENIE physics release involves intensive computation that is not practical in a desktop environment. For example, cross sections as a function of energy need to be computed on dozens of materials for almost a hundred different interaction models, and some models take more than a day to compute. But the work is largely "embarrassingly parallel," making it easy to spread out over the Grid and finish in a matter of hours what might otherwise take weeks. FIFE tools, jobsub in particular, but also data storage and data handling tools like ifdhcp make it straightforward to organize the complex production of these calculations.

In addition to total cross section computations, GENIE recently began running large simulations to study the propagation of daughter particles from neutrino interactions through the nucleus. Very large numbers of events are required to study these simulations in detail, and Grid scale computing is an important enabling technology. Our goal is to move more and more of the validation onto the grid and automate the production of validation samples in order to free GENIE developers to focus on improving the physics of the simulation instead of organizing production. FIFE team members Neha Sharma, Ken Herner, and Tanya Levshina have been particularly helpful in our pursuit of this goal - so thanks!

- Gabriel Perdue

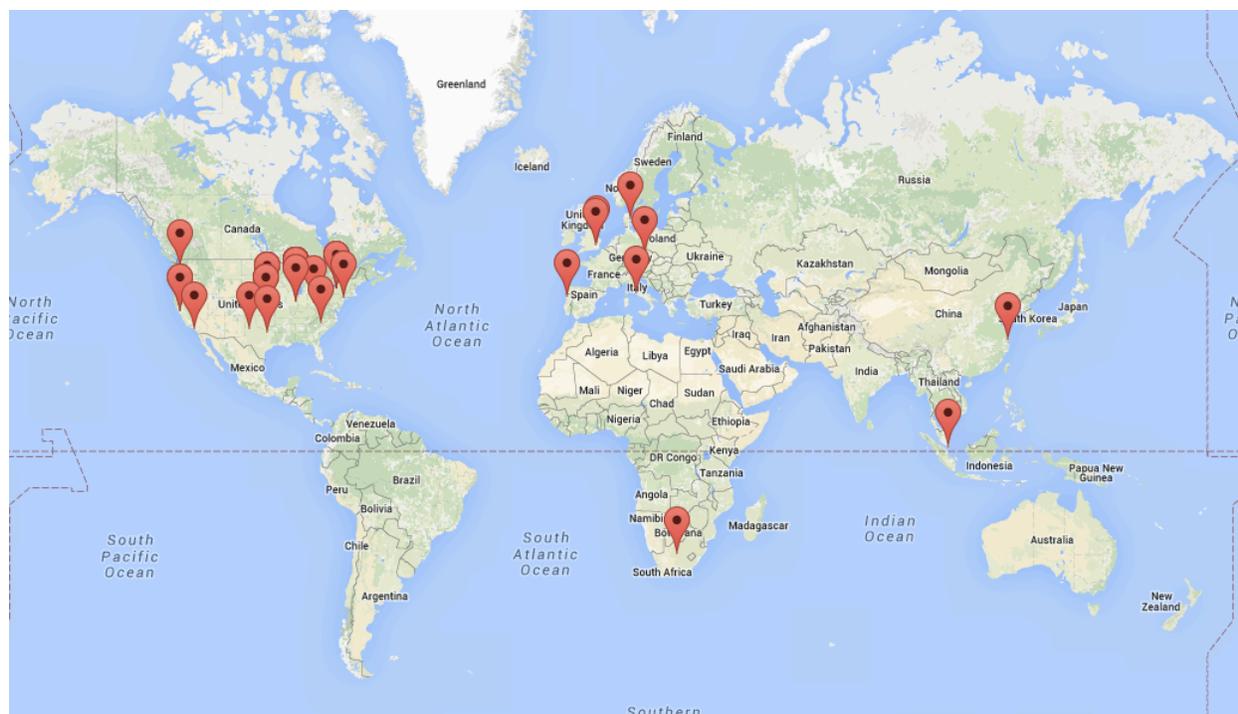
### **dCache : scaling out to new heights**

The dCache data storage system has been the dependable work horse of High Energy Physics experiments worldwide for the last 15 years. At Fermilab it was first adopted by the CDF Tevatron experiment and then became a backbone of regional CMS T1 data center storage. The traces of the Higgs boson particle in digitized form had been hidden in the haystack of more than hundreds of Petabytes of data delivered using dCache. These were reconstructed by analysis software to reveal a long sought missing piece of the Standard Model. An efficient data storage system underpins high data throughput scientific research. The dCache data storage system plays a very important role in helping to deliver this and many other major scientific results.

The public dCache instance at Fermilab has served the needs of a diverse community of customers for many years including neutrino experiments, astrophysics, lattice QCD and database group. Since the fall of 2013, the system has been actively used by intensity frontier experiments. It has been scaled out dramatically - from just over 100TB to over 5PB. Currently dCache routinely delivers (both reads and writes) more than 5M files per day and moves about 500TB of data per day. This combined level of performance puts it on par with ATLAS and CMS T1 sites in BNL and Fermilab.

This fiscal year, we have made our first steps towards leveraging our expertise in storage to expand our user base by launching Active Archive Facility project (see <http://archive.fnal.gov>) that enables researches access to our storage facility through the Strategic Partnership Project (SPP) mechanism. Our first major customer is the Simons Foundation Genome Diversity Project whose participants actively use public dCache to store and access their data over the WAN.

The world map below shows distribution of dCache clients that have transferred at least 1TB of data in the last 3 months.



The successful operation of dCache instances at Fermilab is made possible in part by direct involvement of the Data Movement and Development group in dCache development within a framework of international collaboration between DESY, Fermilab and NDGF. Over the years, dCache software has been evolving towards embracing industry standards such as parallel NFS (pNFS) and WebDAV while maintaining and improving popular domain specific protocols like GFTP, XRootD, SRM and dCap. In fact, dCache provides its own, fully compliant, implementation of XRootD server in Java. We work closely with IFDH and SAM developers who provide a set of user friendly tools that hide protocol specifics from the end user.

Performance demands are always on the rise and the increased user load does not always come smoothly; there are always some issues to work on. The issues with pNFS Linux client, ripple effects caused by pool nodes going offline or certain corner use cases that result in non-anticipated behavior keep us occupied.

We look forward to new challenges with optimism. The system is capable of meeting ever increasing data throughput needs because it has been designed to be highly scalable and is able to adapt to changing load patterns.

Be sure to open a service desk ticket if you have problems or dCache does not work as expected. Your problem reports drive a process of continuous code improvement resulting in a better dCache product.

Dmitry Litvintsev & Gene Oleynik

## **OPOS – The importance of collaboration and cooperation**

Offline Production Operations Service (OPOS) assists experiments with running their large-scale production workflows and other large-scale offline production activities. With three computer scientists and two physicists, the group expands the human resources of the experiment's offline team while coordinating across multiple requests. The Service Level Agreement, [cd-DocDB-5563](#), gives the details of what is offered to the experiments as well as what is required to be eligible to request this service. Basically, the experiment must have production code, scripts and configuration files with ownership limited to authorized people. The experiment must train the OPOS team and make a request via a SNOW ticket to request service.

Once trained, OPOS team members schedule, according to the collaboration's deadlines, and run the large-scale production tasks. OPOS team members monitor the tasks and provide the experiment with status and progress information.

On November 30<sup>th</sup>, DUNE started using OPOS for MCC5.0 production which had requests from several physics working groups such as oscillations, supernova, proton decay and cosmogenics. DUNE had a training session and gave the OPOS team instructions on submitting and monitoring jobs and doing sam-related operations. An issue with account permission was quickly solved with help from SCD. By day two, OPOS was producing samples at full speed. Half the samples were finished after two and a half days. As Tingjun Yang said, “We are extremely impressed by their dedication and cooperation on the MC production. I think the OPOS group is doing a fantastic job and their contribution is very much appreciated by the DUNE collaboration.”

The OPOS group facilitates the transfer of tools and know-how among experiments by helping with the adoption of common tools. The goal is to help experiments increase their productivity while focusing on the final results: data analysis and papers.

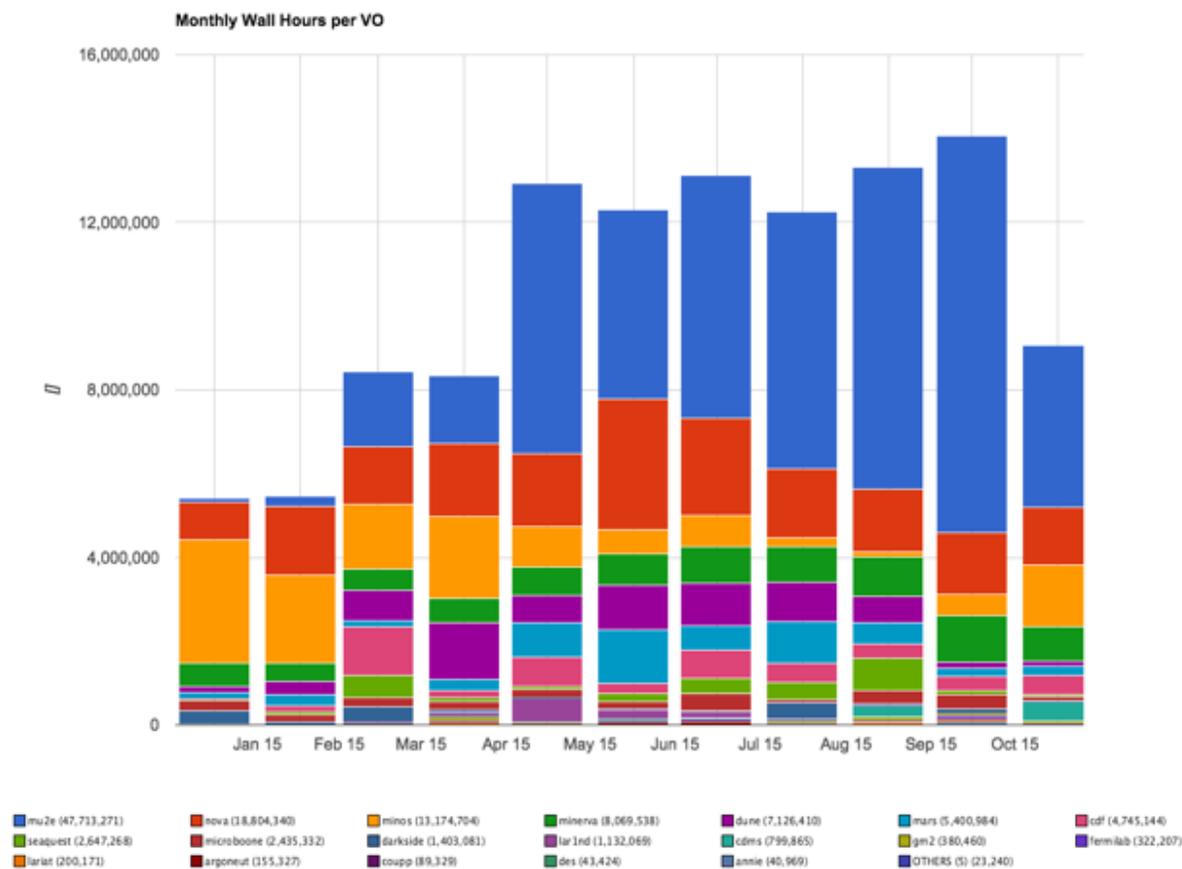
- Anna Mazzacane & Katherine Lato



### Jobsub Hints

Jobsub is a FIFE user's doorway into running jobs in computational grids, clouds and other HPC clusters. Jobsub provides a simple-to-use, scalable and reliable job submission abstraction layer for submitting scientific workflows that run on diverse computation resources. Jobsub's simplistic architecture and ease of adaptation by new and existing communities has contributed towards its success. Security is at the core of the Jobsub architecture.

Over the past several months, the Jobsub team has been working closely with the FIFE communities to define common interfaces by integrating complex grid tools and automating several mundane tasks during the job submission phase.



Since Jan 2015, users have consumed over 10M hours of computing cycle every month using Jobsub infrastructure. These numbers are expected to grow even further as more experiments start taking data and progress further into their life cycle.

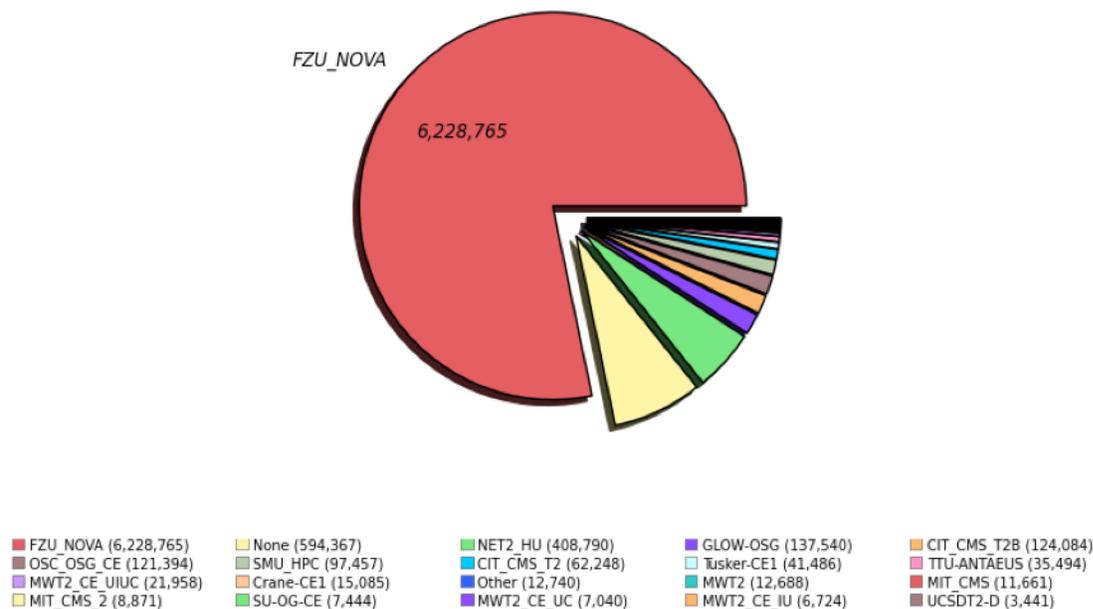
In the coming months, Jobsub will transition from using KCA to CILogon certificates for job submissions. The Jobsub team will be working with the Fermilab security team and the FIFE support group in SCD to make this transition as smooth and transparent as possible for the users. Meanwhile the Jobsub group will continue to make periodic releases to add useful features and fix bugs. More information about the Jobsub project, documentation and release plans are available at <https://cdcvs.fnal.gov/redmine/projects/jobsub/wiki> with a user guide available at <https://cdcvs.fnal.gov/redmine/projects/jobsub/wiki#Client-User-Guide>.

- Parag A. Mhashilkar

## FIFE computing at European sites

The experimental collaborations that make up the FIFE community have always been international with collaborators from dozens of countries. Now, computing resources that are used are coming from beyond Fermilab and other OSG sites in the United States. In the past year, the site outside of Fermilab providing the most computing resources to the NOvA experiment has been the Institute of Physics of the Czech Academy of Sciences (*Fyzikální Ústav AV ČR* or "FZU"). FZU, which in the past also provided computing resources to the DZero experiment, has been set up as an Open Science Grid (OSG) site which allows NOvA physicists to submit jobs via familiar FIFE tools. In the past year, NOvA has utilized over 6 million computational hours at FZU.

**Wall Hours by Facility (Sum: 7,959,277 Hours)**  
52 Weeks from Week 49 of 2014 to Week 49 of 2015



*Figure 1: NOvA computational hours at sites outside of Fermilab in the past 12 months*

When NOvA collaborators at the Joint Institute for Nuclear Research (JINR) in Russia were interested in providing computing resources to NOvA, FIFE and OSG staff followed the model set by FZU and set up access to a JINR computing cluster via an OSG site. A similar setup

is currently being established at the University of Bern in Switzerland for the MicroBooNE experiment.

- Bo Jayatilaka

## Intensity Frontier Data Handling (ifdh) usage helpful hints

**"use 'ifdh cleanup'"** - Use the cleanup call that gets rid of cached certificates, files pulled in with fetchInput, etc. If you use ifdh regularly, put it in your ~/.bash\_logout

**"use 'ifdh cp -f '"** - Make a list of files with 'ifdh ls' to get source/destination pairs instead of explicitly listing them.

**"use 'export IFDH\_DEBUG=1'"** - When something is not working, there are lots of ways to see what's going on such as IFDH\_DEBUG or other environment variables.

**"finding dcache errors"** - When you get an error on a copy to/from DCache on-site, look at [http://fndca.fnal.gov/cgi-bin/dcache\\_files.py](http://fndca.fnal.gov/cgi-bin/dcache_files.py) within 15 minutes to find your copy there.

**"use 'export IFDH\_STAGE\_VIA='"** - Stage output instead of copying files back which can bog down DCache. You can set IFDH\_STAGE\_VIA to a conditional syntax, based on hostname.

For details, please see <https://cdcv.s.fnal.gov/redmine/projects/ifdhc/wiki/TipsandTricks>

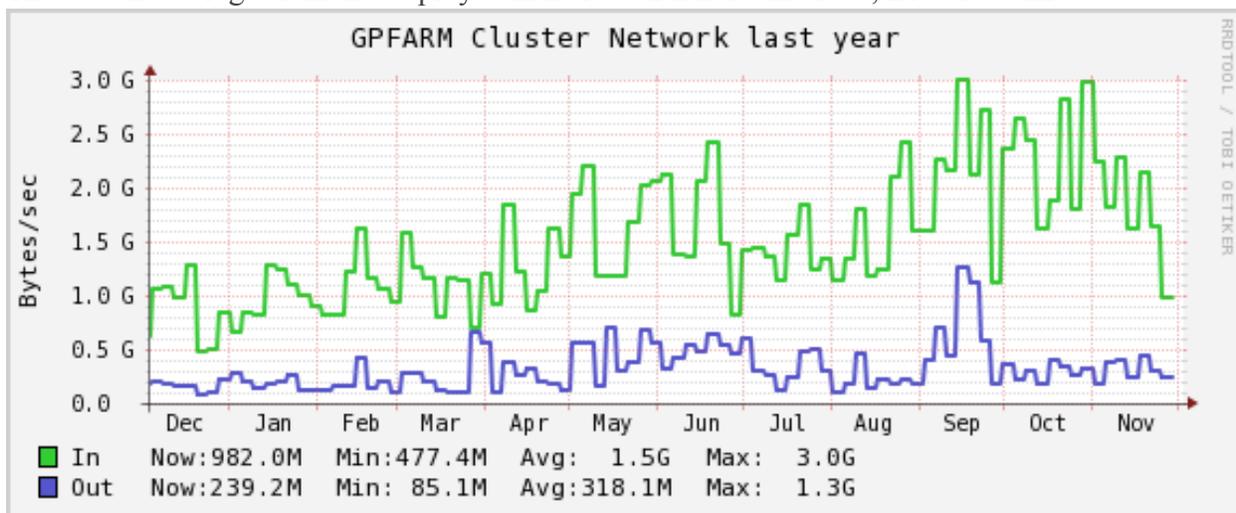
- Marc Mengel & Katherine Lato

## Bluearc unmounting from GPGrid nodes - So Long and Thanks for All the Files.

A long time ago in a cluster far, far away, it was a period of rebellion against the limitations of local batch clusters.

In 2009, the 3000 cores of the GP Grid Farm were a vast improvement over the 50 core FNALU batch system. GPGrid was connected to the then-new Bluearc data system with a 2 GBit network link. A simple lock system deployed in late 2009, still in use today, avoided head contention on the underlying Bluearc data system, improving uptime in 2010 from 97% to 99.9997%. Deployment of the ifdhc tools as new projects came on line kept uptime fairly good, 99.95% in 2014.

But there are new issues that locks cannot fix. Our Bluearc servers have about 1 GBytes/second service capacity. Single GPGrid worker nodes now have that much capacity. We are now running as many as 30,000 user processes on Fermigrad, sustaining over 3 GBytes/sec locally. The dCache storage elements deployed in 2015 can handle this load, Bluearc cannot.



### *GP Grid Networking throughput*

We need to proceed this year with the Bluearc Unmount process described in <http://cd-docdb.fnal.gov/cgi-bin/ShowDocument?docid=5522> and <https://cdcvs.fnal.gov/redmine/projects/fife/wiki/FermiGridBlue>. We need to go farther, removing even GridFTP access to Bluearc data. See <https://cdcvs.fnal.gov/redmine/projects/fife/wiki/FGB-DATASCHED>. We will be contacting Liaisons to schedule the data area unmounts.

The existing Bluearc data areas remain a valuable resource for interactive work, where full

Posix file access may be needed.

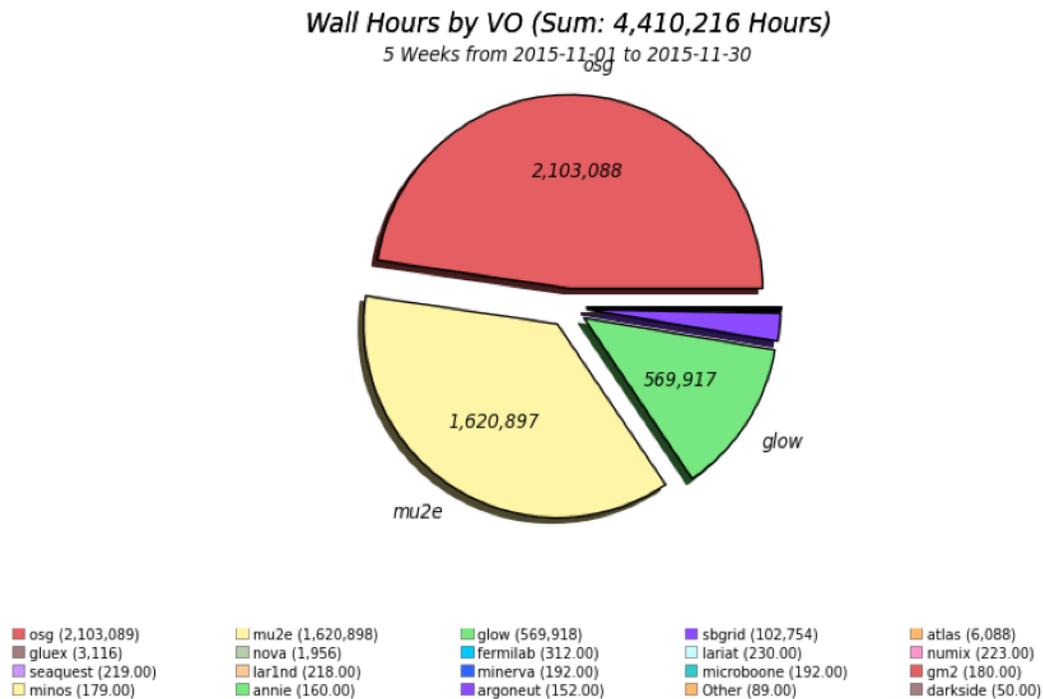
- Arthur E. Kreymer

## Most opportunistic cycles for a site

In the month of November, sites on the Open Science Grid (OSG) provided 13.8 million computational hours to opportunistic users—users who do not own any of the resources on the site. The single largest source of these resources was the Syracuse University OrangeGrid (SU-OG), which provided over 4.4 million hours. SU-OG, developed in 2012, harvests idle computing power in computers across the Syracuse University campus for research computing and has now grown to include over 12,000 CPU cores. In November 2014 SU-OG was added as an OSG site, allowing capacity unused by Syracuse users to go to users and virtual organizations (VOs) of the OSG. In the month of November, the FIFE experiment harvesting the most cycles from SU-OG was Mu2e, which obtained 1.6 million computational hours.

More about SU-OG: <http://researchcomputing.syr.edu/resources/orange-grid/>

OSG Accounting: <http://gratiaweb.grid.iu.edu/>



- Bo Jayatilaka

## Most Efficient Experiment

Most Efficient Experiment on FermiGrid that used more than 500,000 hours since October 1st — Minos (98.24%) and MU2E (98%)



- Tanya Levshina

**Most efficient big non-production user**

Most efficient big, but not Production, users on FermiGrid who used more than 100,000 hours since August 1st was Ashley M. Timmons from Minos with 98.6% efficiency.

<b>Experiment</b>	<b>User</b>	<b>Wall Hours</b>	<b>Efficiency</b>
Minos	Ashley M. Timmons	583,730	98.6%
Minos	Adam J. Aurisano	1,266,091	98.0%
Mu2e	Anthony Palladino	578,438	98.0%

- Tanya Levshina

### Experiment with the most opportunistic hours

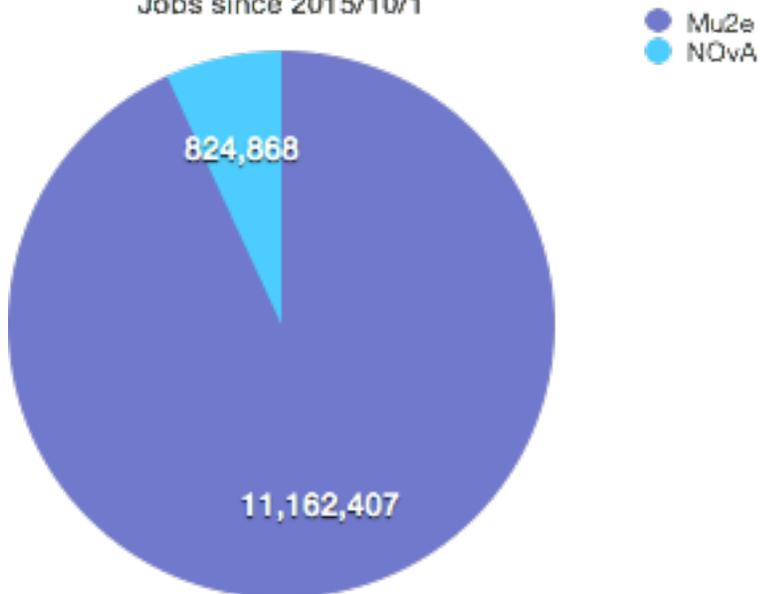
The experiment with the most opportunistic hours on OSG between October 1st and November 30 was Mu2e with 11,162,407 hours.

Top two:

11,162,407 - Mu2e

824,868 - NOvA

Opportunistic Cycles (Hours) Spent on  
Jobs since 2015/10/1



- Tanya Levshina

This newsletter is brought to you by:

- Ken Herner
- Bo Jayatilaka
- Mike Kirby
- Arthur E. Kreymer
- Katherine Lato
- Marc Mengel
- Parag A. Mhashilkar
- Tanya Levshina
- Dmitry Litvintsev
- Anna Mazzacane
- Gene Oleynik
- Gabriel Perdue

## **Feedback**

This is the second in a series of newsletters to the community. We welcome articles you might want to submit or feedback. Please email [fife-group@fnal.gov](mailto:fife-group@fnal.gov).