

FIFE-Jobsub: a grid submission system for intensity frontier experiments at Fermilab

Dennis Box¹

Scientific Computing Division, Fermi National Accelerator Laboratory
PO Box 500, Batavia, IL USA 60510

E-mail: dbox@fnal.gov

Abstract. The Fermilab Intensity Frontier Experiments use an integrated submission system known as FIFE-jobsub, part of the FIFE (Fabric for Frontier Experiments) initiative, to submit batch jobs to the Open Science Grid. FIFE-jobsub eases the burden on experimenters by integrating data transfer and site selection details in an easy to use and well-documented format. FIFE-jobsub automates tedious details of maintaining grid proxies for the lifetime of the grid job. Data transfer is handled using the Intensity Frontier Data Handling Client (IFDHC) [1] tool suite, which facilitates selecting the appropriate data transfer method from many possibilities while protecting shared resources from overload. Chaining of job dependencies into Directed Acyclic Graphs (Condor DAGS) is well supported and made easier through the use of input flags and parameters.

1. Introduction

Numerous options and choices confront a new Open Science Grid user. There are different ways to accomplish the goal of Scientific Batch Computing, and large collaborations of experimenters have the resources to develop an optimal work flow for their needs. At Fermilab, there are small collaborations that desire to quickly manage all the details of authentication, data I/O, job scheduling and chaining, and retrieving output. For these groups we have developed the General Purpose Computing Facility (GPCF) [2] and recommend FIFE-jobsub, also known as jobsub_tools, to manage these details. Currently nine experiments at Fermilab are using jobsub_tools for their batch submission needs.

2. FIFE-Jobsub Philosophy and Approach

Our development goals were: Simplify job submission for users, provide sensible defaults, and give useful options for overriding these defaults.

The details of site selection, obtaining permission to run at these sites, I/O to the sites worker nodes, and monitoring job progress can be accomplished in many ways, but takes time. We assume that an experimenter with access to FIFE would prefer to spend more of their time doing actual science. New users are quickly productive with jobsub using default settings but can accomplish

¹ To whom any correspondence should be addressed.



complicated workflows as needed, accessing thousands of cores with high I/O by using the various options provided with the submission tools.

3. Current Architecture

The current iteration of FIFE-Jobsub was developed for the GPCF production environment at FNAL.

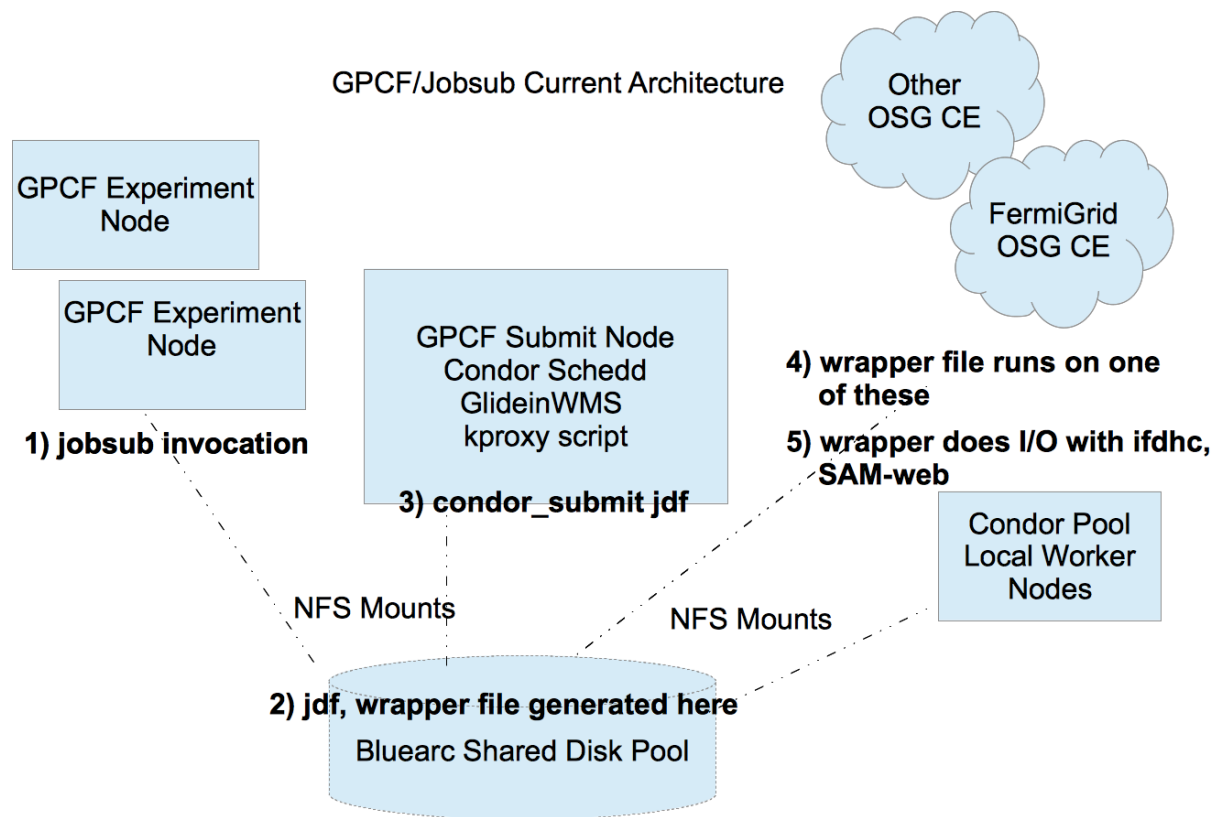


Figure 1. GPCF/Jobsub Current Architecture.

Users log on to and work from one of the GPCF Experiment Nodes, which are virtual machines provisioned as needed. Jobs are submitted from these using the jobsub_tools provided jobsub command to submit to the HTCondor schedd daemon on the GPCF Submit Node, which then is noticed by a second HTCondor Pool managed by GlideinWMS[3]. HTCondor matches the user job in the schedd with a worker node in either the local pool or one of the OSG CEs. A large shared disk pool is cross mounted between the Experiment Nodes, the Submit Node, and the worker Nodes in the local pool and the FermiGrid[4] OSG CE.

Users may submit jobs to the 'local pool' without obtaining any sort of grid identity or authorization. To access the OSG they need a grid or VOMS proxy. The FermiGrid CE requires a VOMS proxy identifying them as a member of the Fermilab VO attached to a specific experiment. The process of obtaining VO membership and proxies has been streamlined for the user with the jobsub_tools provided script 'request_robot_cert'. This script navigates VOMS and VOMRS web pages to request that a user proxy generated from a Fermilab kx509 cert is correctly formatted and associated with an experiment. Once the request has been made, VOMS administrators manually verify and approve the applicant. The local pool is available to new users while waiting for approval. Once approved, the jobsub_tools provided 'kproxy' script is run by a user to generate a VOMS proxy.

A cron job on the submit node should be set up by the user to periodically run the kproxy command to ensure that proxies do not expire for long running jobs.

When the 'jobsub' command is run on a GPCF Experiment Node a wrapper script and HTCondor Job Description File (JDF) are generated on the NFS file system shared by the Experiment Node, Submit Node, and Worker Nodes. The wrapper scripts purpose is to implement SAM, gridftp, and srmcp I/O with IFDHC commands at appropriate times and to execute the user job. The JDF sets up the execution environment and runs the wrapper script which in turn runs the user application on one of the worker nodes.

A user submitting with no modifying options sends jobs to the local HTCondor pool, which is a user friendly and accessible computing environment with 176 computing slots. Here users have local accounts and can log in and observe their jobs executing in real time. Input/output directories are mounted via NFS, stdout, stderr, and batch system logs end up in these directories easing the process of debugging.

Once users are confident in the success of local batch jobs, they can access greater resources by running their jobs on Fermigrid or other OSG CEs. Adding a '-g' in the jobsub incantation sends the users jobs to the Fermigrid OSG CE where up to 5000 execution slots are available. Users cannot generally log into Fermigrid execution nodes, but their jobs are expected to work very nearly the same as on the 'local pool' as the NFS mounts with experiment applications and input/output directories are the same. The differences can include variations in installed versions of system library software and the fact that user jobs run under different UIDs in the two environments. 'Local pool' jobs run under the users UID, while Fermigrid jobs run under an experiment specific group ID. This can cause access problems for both applications and data I/O. These are typical problems that a user encounters when moving from one OSG CE to another, encountering them at this transition conditions them to think about these issues prior to moving to the greater OSG, which should result in more robust code.

Once the Fermigrid CE has been mastered, users can access even greater compute capacity by directing their jobs to other OSG CE's using the proper flag. In addition to software version and UID problems described above, it is common for experiment legacy code to encounter file system layout assumptions, expecting NFS mounts that exist on the local batch and Fermigrid but do not generally exist on other CE's. Many legacy applications expect experiment specific libraries to live on the NFS mounts, much of this code has been successfully modified to look in CVMFS instead.

Default I/O to the jobs has evolved over the years. Originally users took advantage of their direct access to NFS mounted files, reading and writing with standard unix utilities and libraries. While this worked, the file system and hence the entire submission system would lock up when a user accessed the same file from too many jobs at the same time. To prevent overloads, metered access to NFS mounted files using lockfiles was introduced. A campaign is now underway to enforce the use of OSG tools (gridftp, srmcp) while retaining metered access to the NFS mounts via the ifdh tool suite.

I/O of data files to the local disk worker node is trivially accomplished by selecting the proper flag. The wrapper files generated by jobsub use the ifdhc tools to correctly transfer these files with the most appropriate method when these flags are used.

More complicated input is possible by invoking jobsub options to generate diamond shaped DAGS using Sequential Access to Metadata via Web (SAM-Web)[5] projects as input. Naming a data set as input and N consumer processes to run will open the SAM dataset, spawn N grid jobs which will repeatedly use ifdh getnextfile consume the files in this dataset, then close the dataset after the last file is consumed.

4. Successes

Jobsub has been in continuous production use for over a year. Use of Agile development and release methodologies has resulted in version releases 38 times between v1_0 released 9/13/12 and the latest maintenance release of 10/25/13. In other words, a new release comes out approximately 3 times a month, resulting in rapid response to user requests for features or bug fixes.

During this time the MINERvA experiment has published at least 2 papers [6],[7] using jobsub to manage its job submission needs.

The NOvA experiment was recently able to use jobsub during a 2 week period to generate, retrieve, and archive 2TB (over 1 million events) of Monte Carlo data using 88k CPU hours of OSG and cloud resources at 5 OSG sites. Most of the input files were served via SAM-web using the tools DAG input flags.

The GPCF Submission Node is fed almost entirely by jobsub and usually has 2000-4000 running jobs as show in in Figure 2.

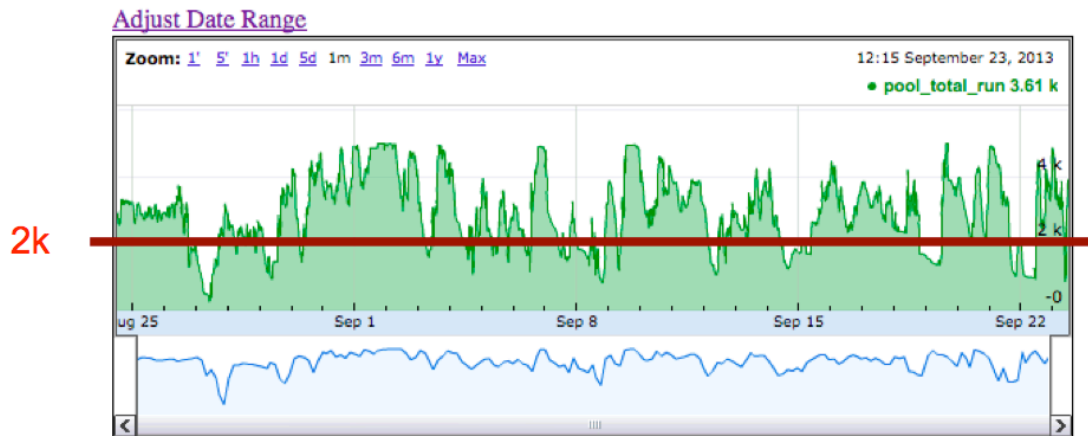


Figure 2. GPCF Running Jobs September 2013

5. Lessons Learned

Adding a new experiment to GPCF and jobsub is more cumbersome than it should be. Many of the jobsub version releases during the previous year were solely to accommodate new experiments that wished to join in on all the fun.

The philosophy of making things easy for the new user has possibly been taken too far in the case of the cross mounted NFS directories in particular. Users write applications wholly dependent on the existence of these directories, run on the FermiGrid CE where they exist, and think they have a 'Grid Application'. Much work and user education is currently required to transform a FermiGrid application into a 'real' grid application. The NFS mounts have proved to be problematic for other reasons, a single user can run (say) a thousand jobs that all write to the same NFS directory locking up the entire system until said user is found and educated as to why this is a bad thing.

6. Future Plans

In light of experience and lessons learned we plan the following:

- A redesign of the way VOMS proxies are generated will eliminate the need for users to have logins on the Submit Node.
- A RESTful API is being designed which separates client and server side functionality more cleanly.
- The shared NFS mounts are going to be eliminated, users will submit jobs and locations of input through the client API and retrieve their data files back the same way.
- Experiment specific information is being moved out of code and into configuration files, which will allow easier addition of new experiments.
- All of these changes will help jobsub move from a GPCF-centric to a more general OSG submission tool.

7. Conclusions

Intensity Frontier experiments at Fermilab have successfully used FIFE and jobsub to quickly ramp up their use of the Open Science Grid. Complex work flows have been developed and are currently being used by multiple experiments. NOvA has successfully used these tools to submit jobs to the Open Science Grid to generate more than 1 million simulation events in less than two weeks. Nine other experiments are actively using FIFE and jobsub tools to submit jobs to FermiGrid resources and will soon be integrated to use OSG resources as well.

Acknowledgements

Fermilab is operated by Fermi Research Alliance LLC under Contract No. DE-AC02-07CH11359 with the United States Department of Energy

References

- [1] M. Mengel, A. Lyon The “Last Mile” of Data Handling – Fermilabs IFDH tools. *In these proceedings*
- [2] E Berman et al. General Physics Computing Facility (GPCF) development docs. Fermilab CD Document 3453 <https://cd-docdb.fnal.gov:440/cgi-bin/ShowDocument?docid=3453>
- [3] GlideinWMS home page accessed October 31 2013
<http://www.uscms.org/SoftwareComputing/Grid/WMS/glideinWMS/doc.prd/index.html>
- [4] K. Chadwick Chep 2012 Fermigrid Fermilab CD Document 4494
<https://cd-docdb.fnal.gov:440/cgi-bin/ShowDocument?docid=4494>
- [5] R. Illingworth. A data handing system for modern and future Fermilab experiments *In these proceedings*
- [6] G. A. Fiorentini, D. W. Schmitz, P. A. Rodrigues et al. (MINERvA Collaboration) Measurements of $d\sigma/dQ^2$ and Final State Nucleons in Muon Neutrino Quasi-Elastic Scattering on a Hydrocarbon Target, Phys. Rev. Lett. 111, 022502 (2013) <http://arxiv.org/abs/1305.2243>
- [7] L. Fields, J. Chvojka et al. (MINERvA Collaboration) Measurement of $d\sigma/dQ^2$ in Muon Anti-Neutrino Quasi-Elastic Scattering on a Hydrocarbon Target Phys. Rev. Lett. 111, 022501 (2013)
<http://arxiv.org/abs/1305.2234>