THE FABRIC FOR FRONTIER EXPERIMENTS PROJECT AT FERMILAB: COMPUTING FOR EXPERIMENTS

K. Herner[a], M. Kirby, A. F. Alba Hernandez, S. Bhat, D. Box, J. Boyd, V. Di Benedetto, P. Ding, D. Dykstra, M. Fattoruso, G. Garzoglio, A. Kreymer, T. Levshina, A. Mazzacane, M. Mengel, P. Mhashilka, V. Podstavkov, K. Retzke, N. Sharma, and J. Teheran

Fermi National Accelerator Laboratory, Batavia, IL, 60510 USA

E-mail: [a] kherner@fnal.gov

The FabrIc for Frontier Experiments (FIFE) project is a major initiative within the Fermilab Scientific Computing Division designed to steer the computing model for non-LHC experiments at Fermilab. The FIFE project enables close collaboration between experimenters and computing professionals to serve high-energy physics experiments of differing scope and physics area of study. The project also tracks and provides feedback on the development of common tools for job submission, identity management, software and data distribution, job monitoring, and databases for project tracking. The computing needs of the experiments under the FIFE umbrella continue to increase, and present a complex list of requirements to their service providers. To meet these requirements, recent advances in the FIFE toolset include a new identity management infrastructure, significantly upgraded job monitoring tools, and a workflow management system. We have also upgraded existing tools to access remote computing resources such as GPU clusters and sites outside the United States. We will present these recent advances, highlight the nature of collaboration between the diverse set of experimenters and service providers, and discuss the project's future directions.

Keywords: FIFE, workflow, grid

1. Introduction

Fermi National Accelerator Laboratory (Fermilab) is the world's leading laboratory for particle physics research in neutrino physics, and is a key contributor on experiments covering the full range of physics drivers in high-energy physics today. The current and future precision muon and neutrino experiments have significant computing resource requirements, comparable to previous-generation collider experiments such as CDF and D0, which had collaborations that were several times larger. Many of these precision muon and neutrino experiments are one or two orders of magnitude smaller than LHC experiments, and may lack available effort to design a completely new analysis framework and job submission system. The FabrIc for Frontier Experiments (FIFE) Project [1, 2, 3, 4] within the Fermilab Scientific Computing Division aims to meet these requirements and challenges by working closely with experiments and service providers to develop a common, modular toolkit covering the complete range of necessary services. These services include job submission and monitoring, file delivery and cataloging, storage solutions, analysis and reconstruction frameworks, and collaboration services such as databases and document storage. Some examples of these tools include the jobsub infrastructure [3], the SAM file delivery and metadata catalog service [5], the Intensity Frontier Data Handling Client for file transfer [6], and the ART framework for reconstruction and analysis [7]. Figure 1 shows how these components can interoperate in the case of analysis or reconstruction jobs. Having these common tools that are adaptable to such a wide variety of experiments and detectors saves countless hours of otherwise duplicated experiment effort, and makes it easy to work on multiple experiments, as is common in neutrino physics.

The experiments using the JobSub infrastructure are now typically running between 15,000 and 30,000 combined jobs per day as shown in Fig. 2, and transferring approximately 1.8 PB per week in and out of Fermilab's public dCache [8] system, shown in Fig. 3. While these numbers do not equal those of the ATLAS or CMS experiments, , they are within a factor 6 to 7 of ATLAS or CMS, and these workflows require similar resources to LHC experiments as they continue to increase each year. They also must reach the same scalability and reliability levels that the LHC experiments have.

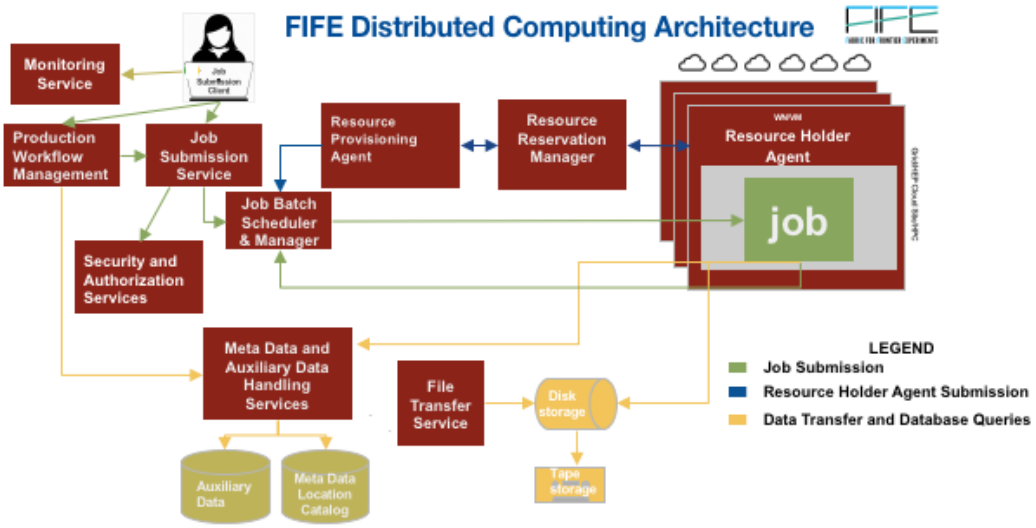

Figure 1. FIFE Architecture, including the job submission and monitoring tools, data transfer and storage options, and communication with remote sites. The GlideinWMS factory provisions job slots on computing resources both inside and outside of Fermilab, and then HTCondor and the VO Frontend match jobs to slots. User job output is normally sent to Fermilab public dCache as shown in the bottom right, which can also act as a high-speed frontend to tape storage. The SAM service provides file delivery and metadata cataloging (with the cataloguing via http), while the File Transfer Service (FTS) can automatically tag dCache and tape locations for new files in the SAM metadata catalog. Various WebUIs can provide other required information (e.g. detector or beam conditions) to

the jobs.

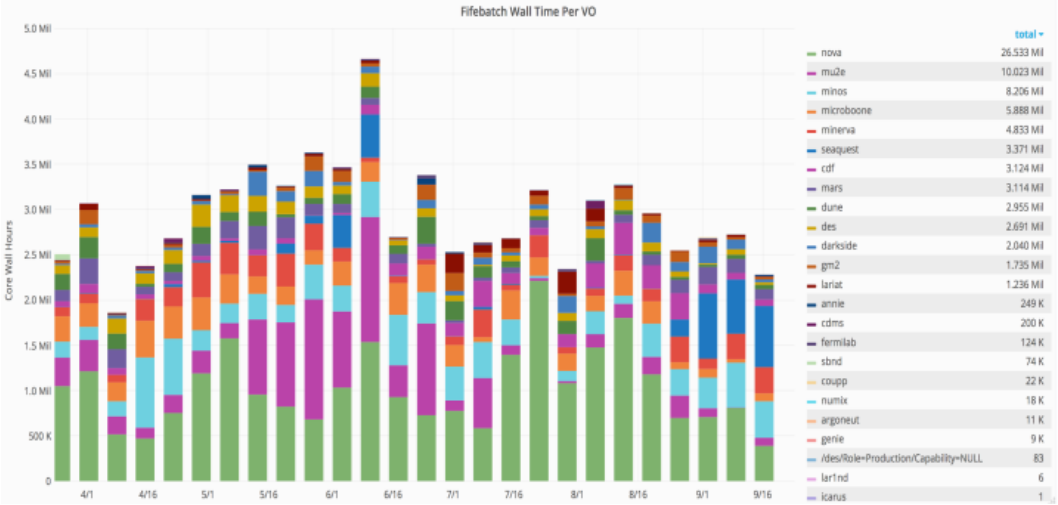

Figure 2: Weekly grid job wall hours for FIFE experiments, March - September 2017

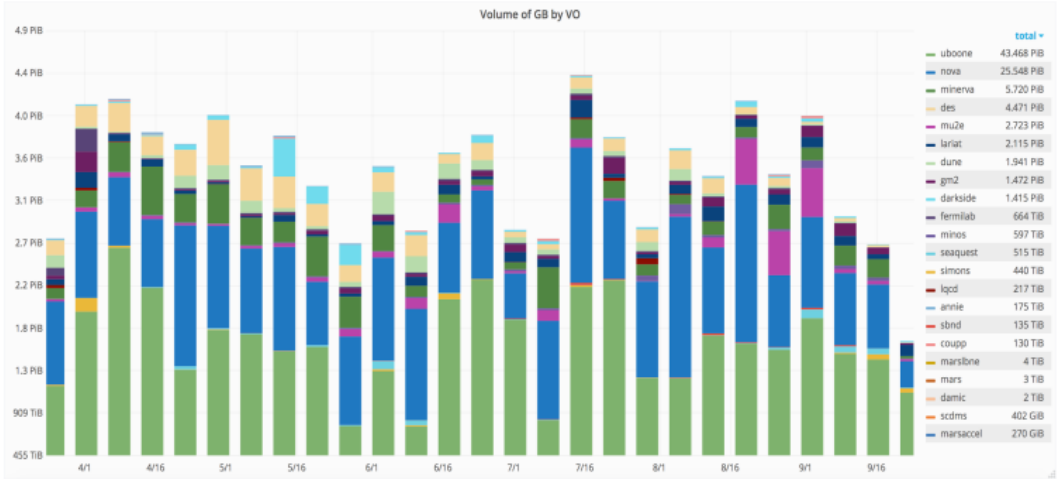

Figure 3. Volume of data transferred via Fermilab public dCache (in and out), March - September 2017

2. Expansion of remote computing resources

A major focus within FIFE is to increase the amount of experiment computing done opportunistically outside of Fermilab. Since the Fermilab job submission infrastructure utilizes GlideinWMS [9], it is straightforward to gain opportunistic access to Open Science Grid (OSG) sites. If experiments ensure that their code does not depend on any Fermilab-specific resources, and is available through CVMFS repositories [10], they can increase the resources available to them manyfold. The Mu2e experiment has been tremendously successful in their effort to use OSG computing, consuming over 50 million CPU hours since early 2015 at no cost to the experiment.

Nearly all experiments using the FIFE tools have collaborating institutions outside of the United States. These institutions may often have significant computing resources available to them, but integrating them into existing job submission systems can be extremely challenging. FIFE experts have made significant strides here in the past year by utilizing the GlideinWMS infrastructure, and following the OSG prescription for integrating new sites. The ability of GlideinWMS to interface with a variety of different local batch systems reduces the burden on local administrators. There are currently five sites in Europe that support at least one FIFE experiment; Figure 4 shows the total FIFE wall hours on these sites for mid-March to mid-September 2017. The total integration time for new sites has been steadily decreasing to an average of 1-2 weeks, and in some cases has been as little as one day.

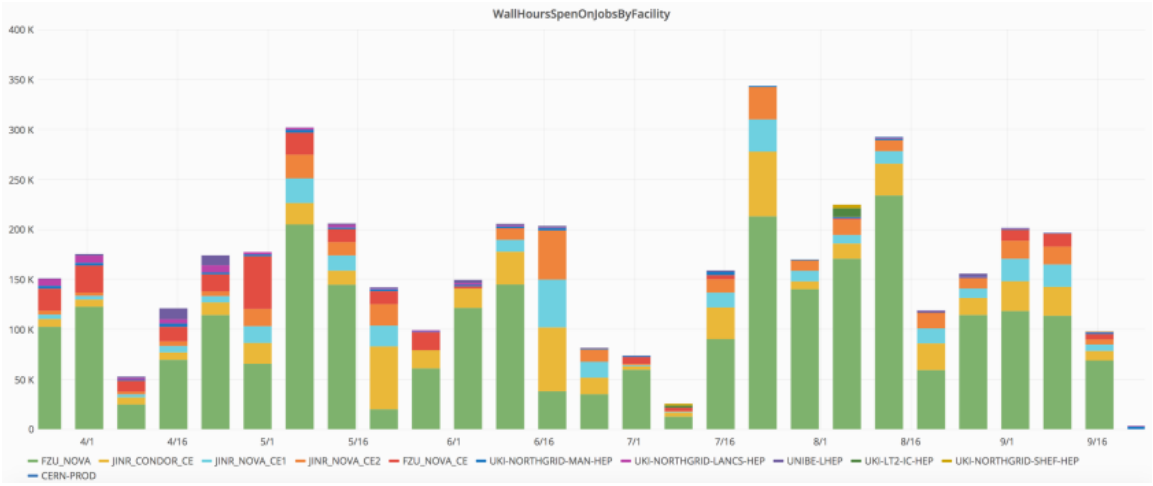

Figure 4. Wall hours from FIFE experiment jobs run outside of the United States, March - September 2017. Presently sites in the Czech Republic, Russia, Switzerland, and the United Kingdom support one or more FIFE experiments

3. Monitoring tools

A complex system such as the FIFE architecture requires robust and sophisticated monitoring tools. It is also imperative to design the tools such that they provide useful information to both system administrators and end users. The FIFE project has an integrated monitoring infrastructure called FIFEMON [11] based on common, open source tools such as Elastic Search [12], Graphite [13], Prometheus [14], and Grafana [15]. The system receives information from all grid jobs submitted via the JobSub tool, as well as system logs and events from machines that provide services, such as job submission and storage elements. By ingesting time series data into graphite, one can build a long-term picture of systems as a whole, and better understand their capabilities and how they interact with one another. Figure 5 shows the FIFEMON architecture, as an example of how one can go beyond simply tracking whether a given service is up or not, and build complete pictures of entire infrastructures over long timescales.

Based on FIFEMON's successful rollout, similar technologies are being used in the Open Science Grid's next-generation accounting tool, GRACC [14]. This tool has now replaced Gratia as the OSG's accounting tool. It provides the same information as Gratia, but offers an improved UI and a lower support burden for administrators.

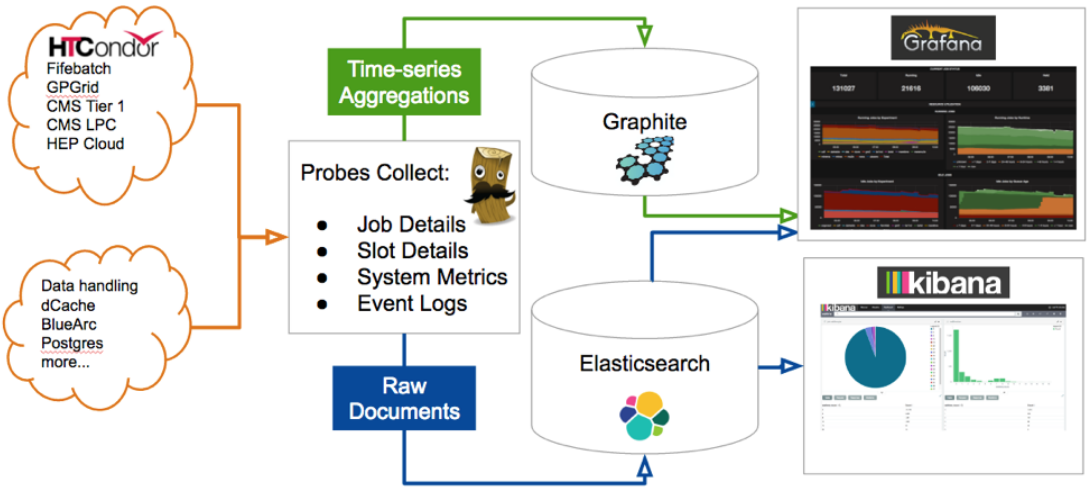

Figure 5. FIFEMON architecture. The system collects information from both jobs and services, displaying them in Grafana plots and also having a Kibana interface for detailed anayisis.

4. Recent FIFE toolset additions

4.1 Continuous integration

The FIFE project also connects users with a Continuous Integration (CI) system. The system, based on the Jenkins toolkit, detects commits to monitored repositories and can automatically build and runs tests on the new software code. It also provides reports on current and past builds, along with information on pre- and post-install tests of the software. The currently platforms are supported platforms are Scientific Linux Fermi 6, Scientific Linux 7, Ubuntu, and OSX. Packages common to several experiments are typically built in all of the Scientific Linux flavors, plus Ubuntu and OSX on request. The experiments are free to choose any or all OS flavors when building their own software.

An important feature of the CI suite is the ability to run so-called "physics validation" jobs. These jobs run on standard grid resources and use a well-understood set of input data with reference outputs. When an experiment builds a new software release, these jobs use the new release to process these standard input files, and generate a set of plots and tuples for automatic comparison to the reference outputs. If the difference between the outputs from the validation jobs and the reference outputs is greater than the specified tolerance, the system can automatically send alerts to the appropriate release managers, saving them significant effort.

4.2 Workflow management system

The FIFE toolset offers a meta-workflow management system that enables fully integrated job submission, file delivery, job tracking, and automated failure recovery. This service, the Production Operations Management Service (POMS), is aimed at experiment's large-scale production teams, but can perform user analysis tasks as well should experiments request it. The production team or end user defines a "campaign" structure that describes the input dataset(s), types of jobs to be run (including any dependencies on other jobs in the workflow structure) and POMS automatically prepares and submits the proper job types, including creating HTCondor DAGs if required by the job dependencies. POMS also tracks the status of every job, and can submit recovery jobs or DAGs without user intervention if there are failures in the processing chain. The user interface is either web-based via a REST API, or via a suite of command line tools.

A database stores information about every job's configuration, making it easy to resubmit certain stages of a workflow with different settings. POMS also has a mode where the user can retain full control over job submission, and use POMS only for jobs monitoring and campaign progress tracking. We expect that POMS will greatly improve experiments' productivity as they adopt it into their standard operations.

5. Future Directions

The FIFE Project's future focus will cover three main areas: helping to shape the HEP computing model of the future, lowering access barriers to computing resources, and improving our existing services. The HEP computing model of the future will likely include increased use of High Performance Computing (HPC) resources and cloud resources, both in terms of job processing and perhaps storage, and increased use of multithreaded programs. The FIFE job submission infrastructure can now allow experiments to run on allocation-based HPC resources.

As the HEP computing model incorporates more cloud-based resources, especially for meeting large, short-term computing power demand, it will be critical to ensure that users can seamlessly access these resources via familiar tools. The FIFE Project will work very closely with the Fermilab HEPCloud project, one of the leaders in the field in this emerging area [17]. One of HEPCloud's main components is a Decision Engine that can steer jobs to dedicated experiment resources, general opportunistic resources, or commercial clouds as appropriate to each experiment's constraints. Each project or experiment may have different resources, both computational and financial, available to it, and experiments may choose to prioritize different types of work (e.g. reconstruction vs. analysis) at different times. The Decision Engine can optimally match the work type to a resource, subject to any rules that the experiment may impose (e.g. a certain work type may access commercial resources and another may not), in order to maximize throughput in the most cost-effective way.

While most software currently used by FIFE experiments is single-threaded, memory and speed requirements will make multithreading a necessity. While a number of common tools are already multithreaded, more work needs to be done in this area on experiment-specific software.
FIFE experts will work closely with developers on the various experiments to teach them proper multithreading techniques as needed. Applications wishing to efficiently utilize HPC platforms must now include multithreading as a core design principle.

Lowering barriers to computing resources includes making it easier for end users, especially those not based at Fermilab, to easily access all available resources. One initiative underway now is to the Distributed Computing Access via Federated Identities (DCAFI) project [18]. The eventual goal of the project is to enable users to access Fermilab resources using their own institutional credentials, if their institution is part of the federated trust realm. This change will reduce the burden on collaborators who are not based at Fermilab; a large number of such people will work on the DUNE collaboration in the coming years. In this context, the term lowering barriers also encompasses improvements to existing middleware such as JobSub to fully interface with all federated credentials.

Several other enhancements to the existing FIFE toolset are underway. Modifications to SAM include a new tool suite that will enable general end users to more easily group analysis files according to general criteria, easily define and remove datasets, and bring SAM's advanced file delivery capabilities to small-scale analysis jobs. Another improvement underway for SAM is to begin allowing for a "send the jobs where the data are" model, in addition to the traditional "send the data to the jobs" model. Improvements to POMS include a more robust set of command-line tools and options to automatically resubmit failed jobs with different resource requests. As always, FIFE will work to keep close interaction between developers and experiment liaisons to ensure that the tools are developed to match the experiments' requirements as closely as possible.

6. Conclusion

The FIFE Project aims to lead the computing model for non-LHC experiments at Fermilab in all areas of high energy physics and astrophysics. FIFE provides a comprehensive toolset to experiments, and offers it in a modular design. The tools are designed for users of all experience levels, and provide a common structure across experiments, something that is very important to physicists participating in more than one experiment. The combined computing demands of experiments using the FIFE tools continues to grow, and now approaches the scale of a single LHC experiment. The FIFE tools include job submission and monitoring, complete workflow management, continuous integration, seamless access to both Fermilab and remote computing resources, including HPC sites, commercial clouds, and GPU clusters. Throughout the foreseeable future, the FIFE project will continue to play a major role in the evolution of the computing model for non-LHC experiments at Fermilab.

Acknowledgements

Fermilab is operated by Fermi Research Alliance, LLC under contract number DE-AC02-07CH11359 with the United States Department of Energy.

References

[1] M. Kirby, J. Phys. Conf. Ser 513 no. 3, p. 032049. IOP Publishing, 2014.
[2] D. Box *et al.*, J. Phys.Conf. Ser. 664, no. 6, p. 062040. IOP Publishing, 2015.
[3] D. Box, J. Phys. Conf. Ser 513, 032010 (2014).
[4] D. Box *et al.*, PoS(ICHEP2016) 176 (2017).
[5] R. A. Illingworth, J. Phys. Conf. Ser 513, 032045 (2014).
[6] A.L. Lyon and M.W. Mengel, J. Phys. Conf. Ser. 513, 032068 (2014).
[7] C. Green *et al.*, J. Phys. Conf. Ser. 396, 022020 (2012).

[8] P. Fuhrman and V. Gulzow, 2006 Euro-Par Parallel Processing (Springer) pp 1106-1113.
[9] I. Sfiligoi *et al.*, 2009 WRI World Congress on Computer Science and Information Engineering (CSIE2009), vol. 02, pp. 428-432, IEEE, 2009.
[10] J. Blomer *et al.*, J. Phys. Conf. Ser. 331, 042003 (2011).
[11] "FIFEMON", https://fifemon.github.io/
[12] "Open Source Search & Analytics: Elasticsearch", https://www.elastic.co/
[13] "Graphite", https://graphiteapp.org/
[14] "Prometheus – Monitoring system & time series database", https://prometheus.io/
[15] "Grafana - The open platform for analytics and monitoring", https://grafana.com/
[16] "GRACC documentation", https://opensciencegrid.github.io/gracc/
[17] S. Fuess, G Garzoglio, B Holzman, R Kennedy, A Norman, S Timm, and A Tiradani, FERMILAB-CONF-16-643-CD (2016).
[18] J. Teheran, D. Dykstra, M. Altunay, FERMILAB-CONF-16-047-CD (2016).