EPJ Web of Conferences **214**, 03059 (2019) https://doi.org/10.1051/epjconf/201921403059
CHEP 2018

Advances and enhancements in the FabrIc for Frontier Experiments project at Fermilab

Kenneth Herner[1,*], *Andres Felipe* Alba Hernandez[1], *Shreyas* Bhat[1], *Dennis* Box[1], *Joeseph* Boyd[1], *Bruno* Coimbra[1], *Vito* Di Benedetto[1], *Pengfei* Ding[1], *Dave* Dykstra[1], *Michele* Fattoruso[1], *Lisa* Giacchetti[1], *Michael* Kirby[1], *Tanya* Levshina[1], *Anna* Mazzacane[1], *Marc* Mengel[1], *Parag* Mhashilkar[1], *Nikolay* Kuropatkin[1], *Vladimir* Podstavkov[1], *Kevin* Retzke[1], and *Jeny* Teheran[1]

[1]Fermi National Accelerator Laboratory, PO Box 500, Batavia, 60510 USA

Abstract. The FabrIc for Frontier Experiments (FIFE) project within the Fermilab Scientific Computing Division is charged with integrating offline computing components into a common computing stack for the non-LHC Fermilab experiments, supporting experiment offline computing, and consulting on new, novel workflows. We will discuss the general FIFE onboarding strategy, the upgrades and enhancements in the FIFE toolset, and plans for the coming year. These enhancements include: expansion of opportunistic computing resources (including GPU and high-performance computing resources) available to experiments; assistance with commissioning computing resources at European sites for individual experiments; StashCache repositories for experiments; enhanced job monitoring tools; and a custom workflow management service. Additionally we have completed the first phase of a Federated Identity Management system to make it easier for FIFE users to access Fermilab computing resources.

1 Introduction

Fermilab has become the world's foremost laboratory for research in neutrino and precision muon physics and it also plays critical roles in experiments studying all physics drivers in high-energy physics today. The current and future precision muon and neutrino experiments require large amounts of computing resources, some similar in scale to LHC collider experiments, but the majority of them have one to two orders of magnitude fewer collaborators than LHC experiments. Thus they may may lack available effort to design a completely new analysis framework, job submission system, or batch cluster. The FabrIc for Frontier Experiments (FIFE) project [1–5] is a Fermilab Scientific Computing Division effort to meet these requirements. The project provides an interface through which experiments can adopt a modular toolkit that can cover the complete range of computing services required by a modern high energy physics experiment. The available services include job submission and monitoring tools, file delivery and cataloging, storage systems, event reconstruction and physics analysis frameworks, as well as collaboration services such as databases and document storage. Specific examples include the FIFE-Jobsub infrastructure [3], the SAM file delivery and metadata catalog service [6], the Intensity Frontier Data Handling Client for file transfer [7],

*e-mail: kherner@fnal.gov

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

and the ART software framework [8]. FIFE also provides a forum for experiments to closely interact with software developers and service providers in order to ensure that the services meet the experimenters' needs, and are adaptable to the full range of supported experiments, which often have very different physics goals, detector designs, and analysis strategies. The common, modular approach the FIFE project has chosen saves countless hours of otherwise duplicated experiment effort, makes it easy to work on multiple experiments simultaneously from a computing perspective, and reduces the overall support burden on laboratory computing personnel. It will also be easier for physicists to transition to the DUNE experiment in the future.

2 Advances in international resource integration

The FIFE experiments are all able to use the Fermilab General Purpose grid cluster (hereafter FermiGrid) for their computing work, but using opportunistic computing resources outside of Fermilab is typically required to meet peak demand. FermiGrid provides approximately 22,000 cores across the cluster, with approximately 2 GB of memory available per core. The Fermilab job submission infrastructure makes use of the GlideinWMS [9] tool to provision remote computing resources. The choice of GlideinWMS provides straightforward access to Open Science Grid (OSG) sites [10] to run opportunistic jobs. Adding opportunistic OSG computing resources can provide a severalfold increase over what might be available to a given experiment on FermiGrid alone (once experimental quotas are taken into account). There have been peaks of over 15,000 FermiGrid-equivalent CPUs available to FIFE experiments through OSG resources. To profit from such resource gain, all the experiments are required to do is ensure that their code contains no Fermilab-specific dependencies, and is available through CVMFS repositories [11].

Universities and laboratories outside of the United States often have significant computing resources, but it can be very challenging to integrate them all into a unified whole such that a user can reach all available resources with the same submission. There has been a concerted effort in the past year to involve more international computing resources, mostly in Europe, by utilizing the GlideinWMS infrastructure and following the OSG prescription for integrating new computing sites. In particular the DUNE collaboration has been integrating several sites in the United Kingdom in 2018, adding four from May to July with an additional five expected in 2018. As of July 2018 there are ten sites in Europe that can run jobs for one or more FIFE experiments. Figure 1 shows the weekly wall hours at non-US sites from January to July 2018. The total integration time for new sites is now less than one week, especially if the site is already supporting a FIFE or LHC experiment.

3 Improvements to auxiliary file delivery

The FIFE experiments often have workflows that require transferring a large amount of auxiliary input data into jobs, and/or a large amount of custom software outside of the experiment's official software release (often written by end users). In the auxiliary input data case, the total size to be transferred can be from several hundred MB to several GB per job, with file sizes in the tens to hundreds of MB. However, in these workflows, each job in a set will typically only sample a random subset of the overall dataset, leading to very low rates of shared file usage across jobs. This method therefore makes it difficult to take advantage of local caching, and the files are too large to be stored in a standard CVMFS repository (the low overlap rate between jobs quickly leads to thrashing the CVMFS cache on the worker node if several similar jobs happen to start at the same time).

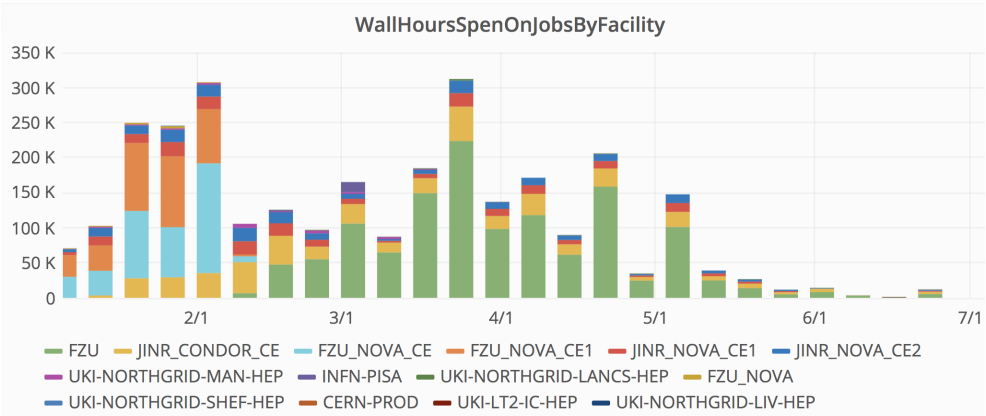

Figure 1. The weekly sum of wall-time hours of FIFE experiment jobs run on resources outside of the United States from January to July 2018. Several experiments were preparing results for the Neutrino 2018 conference that began in June. Thus the hourly totals drop sharply in the second half of May, coinciding with the completion of these results.

To solve this problem, several experiments have set up StashCache repositories [12] overlaid with a CVMFS repository to serve such files to jobs. The master copies of such files reside in Fermilab dCache [13], and the StashCache redirector will replicate the files as needed by pulling from Fermilab dCache over an xrootd connection [14]. By adding the CVMFS layer on top of the CVMFS repository, user jobs can simply access the files via a POSIX-like path. In this way, all details of transferring and caching are hidden from the user. The cache-trashing problem is also avoided as multiple stream can occur behind the scenes on each worker node.

Users may also make custom modifications to small parts of their experiment's software stack, and so may need custom libraries within their jobs in addition to the standard experiment software suite typically served over CVMFS. In some cases, however, these modifications can be more extensive and lead to a large (multi-GB) amount of data that must be shipped to each job. Normally that can be accomplished by storing a tarball in dCache and copying it at the beginning of each job, but dCache pools can quickly become overwhelmed if too many jobs attempt to access a given tarball file at the same time. To counter this problem, we have set up a designated area within dCache where user tarballs are automatically replicated across many pools, while in parallel we are setting up a separate fast-turnaround CVMFS repositories for user code. These repositories would have fast publishing times (less than five minutes) and revision numbers would be linked to jobs, so that the job scheduler would not start the corresponding jobs until the publishing step was complete. Inside the job, the user code would simply be available in another area apart from its usual area on interactive login nodes.

4 Enhancements to job and infrastructure monitoring

The FIFEMON monitoring suite [5] is now part of Fermilab's Landscape project [15], a lab-wide suite of monitoring tools for HTCondor and related infrastructure. FIFEMON has been available for FIFE experiments since 2016, and it allows both service providers and end users to access a rich set of information regarding job status, storage utilization, and overall computing infrastructure health. The past year has seen two significant enhancements

to FIFEMON: the addition of experiment production shifter dashboards and cross-laboratory monitoring.

The production shifter dashboards allow experiment personnel involved in large-scale production workflows to easily see all relevant information on a single page, with colored annunciators to quickly identify if there are any potential issues that need to be addressed. Figure 2 shows the shifter dashboard for the MicroBooNE experiment as an example. Every panel on the dashboard combines multiple related key service metrics and provides links to more detailed dashboards for those services.

The cross-laboratory monitoring page, shown in part in Figure 3, provides a single, convenient location to view information collected on computing services located at both Fermilab and CERN, leveraging the local monitoring infrastructure at each lab. This dashboard has proven especially useful for the ProtoDUNE experiment located at CERN. ProtoDUNE utilizes significant portions of the CERN computing infrastructure as well as Fermilab's. The DAQ, initial raw output, and the prompt processing for data quality are all located at CERN, while Fermilab infrastructure handles the submission of full reconstruction jobs. A single interface through which one can obtain a comprehensive view of dataflow from DAQ to storage, as well as reconstruction processing queues, allows experimenters to quickly absorb information needed to troubleshoot problems instead of wasting valuable time tracking down disparate pieces of information.

Figure 2. Production shifter dashboard for the MicroBooNE experiment. Shifters have a single place where they can view metrics such as job outcome rates (pie chart, upper left), job counts, job CPU efficiency, file transfer times (upper center), file archiving status, and overall utilization of selected storage elements (top right). It is also possible to quickly identify under-performing job groups (bottom panel) and drill down into their details.

5 Workflow management enhancements

The FIFE toolset includes the Production Operations Management Service (POMS), a meta-workflow management system that integrates job submission, file delivery, job tracking, and automated job failure recovery, with user-defined custom recovery options available. The primary service customers are those involved in the experiments' large-scale production teams,

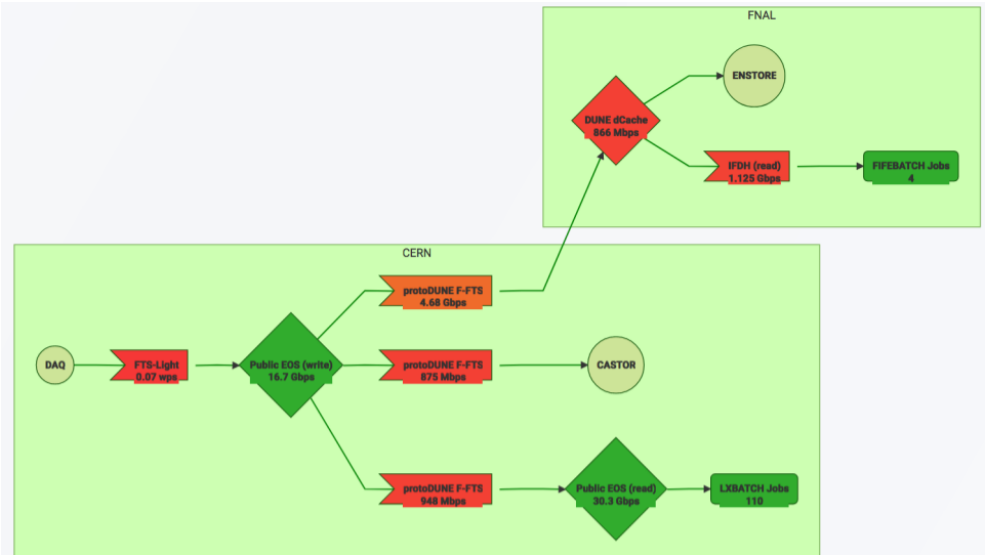

Figure 3. Screen capture of the FIFEMON cross-laboratory monitoring page. Here one can view the overall dataflow and system health information for both the Fermilab and CERN infrastructures in a single location.

but individual user analysis workflows can also run if an experiment chooses to allow it. The end user or team creates a "campaign" structure, describing types of jobs to be run (including any job dependencies), the input datasets, if any, and POMS automatically prepares and submits the proper job types, shielding the users from the intricacies of setting up job dependencies such as HTCondor DAGs [16]. POMS also tracks the status of every submission through our Landscape monitoring system, and can submit recovery jobs or DAGs without user intervention if there are failures in the processing chain. The user interfaces include a web-based REST API and a suite of command line tools. A database stores information about every job's configuration. Users can describe the jobs to be run via an *ini*-style configuration file, with string substitutions possible from either the web API or command-line tools. This setup makes it easy to resubmit certain stages of a workflow with different settings, or to create multiple submissions, iterating through a batch of settings or input datasets each time. It is also now possible to edit campaigns and dependencies via an interactive GUI (see Figure 4). As of July 2018 nine experiments are using POMS for their large-scale submissions, with more planned in the coming months.

6 Future directions

The FIFE project's future involves three main thrusts: helping experiments navigate changes to HEP computing models in the future, simplifying access to computing resources, and improving our existing services. Future HEP computing models will likely include heavy use of High Performance Computing (HPC) and commercial cloud resources, both in terms of job processing and perhaps storage, necessitating increased use of multi-threaded software. The FIFE job submission infrastructure currently allows experiments to run on allocation-based HPC resources. We expect future changes in this area to be tightly coupled with the

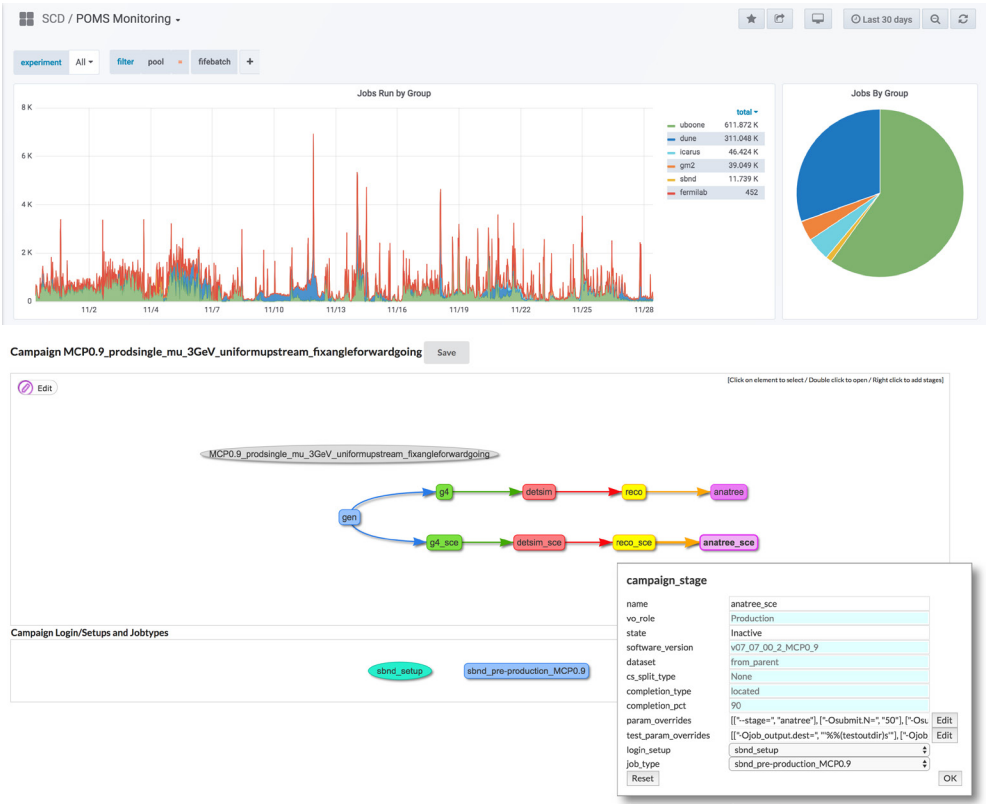

Figure 4. Top panel: Top-level POMS submission monitoring page. Each experiment is a separate color and can be viewed individually. Bottom panel: POMS campaign editor. Job dependencies (indicated by arrows) can be modified within the graphical portion, and the pop-up text box shown allows for quick edits to override input parameters to each type of job, including input dataset, software release, runtime parameters, and options for automated recovery from job failures.

HEPCloud project [17, 18]. FIFE will provide a forum for physicists and service providers to work closely together to adjust to the changing landscape.

Lowering access barriers consists of such efforts as creating a smoother process to on-board users and experiments, and working towards a federated identity structure. The FERRY project [19] is an important piece of streamlining the onboarding process for new users and experiments. It provides an integrated way to perform user and quota management, and replaces soon-to-be-abandoned grid authentication and authorization middleware. The single interface saves users and experiments from having to go through multiple steps to register, and decreases the support load on service providers. In a federated identity structure, identities issued by institutions that are members of a trust federation can be used to access computing resources owned by other members of the federation. In practice it will enable international collaborators to use their own institutional credentials to access Fermilab resources. Such a capability will make it easier for collaborators who rarely travel to Fermilab to access the resources necessary to their work and will reduce the burden on Fermilab support staff. The first phase of this work, the DCAFI project [20], was completed in late 2016.

Planned improvements to the FIFE toolset include: a more robust SAM service, Rucio [21] evaluation, and POMS improvements. The SAM improvements include modifications to allow interoperability with additional file replica catalog services such as Rucio. Service providers are currently evaluating Rucio for use with a number of additional HEP experiments, including DUNE. POMS is also developing a number of backend improvements to allow for additional failure recovery options. 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.

7 Conclusions

The FIFE project is a Fermilab-based effort to support computing model for non-LHC experiments in high energy physics. FIFE provides a complete, modular set of tools to experiments, including tools for job submission, storage and complete workflow management. The commonality across experiments found within the toolset is an extremely helpful feature for physicists participating in more than one experiment. Recent advances in the FIFE toolset include increased access to international computing resources, additional options for auxiliary file delivery to jobs, enhanced monitoring tools for large-scale production workflows and for combining information across laboratories, and improved workflow management systems. In the future the FIFE project will continue to shape the computing model for non-LHC experiments at Fermilab. It will also provide an important forum for experimentalists and service providers to come together to ensure that the available computing resources will enable the experiments to reach their science goals.

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. Government purposes.

This research was done using resources provided by the Open Science Grid [9, 10], which is supported by the National Science Foundation award 1148698, and the U.S. Department of Energy's Office of Science.

References

[1] M. Kirby, J. Phys. Conf. Ser **513**, 032049 (2014).
[2] D. Box *et al.*, J. Phys. Conf. Ser. **664**, 062040 (2015).
[3] D. Box, J. Phys. Conf. Ser **513**, 032010 (2014).
[4] D. Box *et al.*, PoS(ICHEP2016) **176** (2017).
[5] K. Herner *et al.*, J. Phys. Conf. Ser. **898**, 052026 (2017).
[6] R. A. Illingworth, J. Phys. Conf. Ser **513**, 032045 (2014).
[7] A.L. Lyon and M.W. Mengel, J. Phys. Conf. Ser. **513**, 032068 (2014).
[8] C. Green *et al.*, J. Phys. Conf. Ser. **396**, 022020 (2012).
[9] I. Sfiligoi *et al.*, *2009 WRI World Congress on Computer Science and Information Engineering (CSIE2009)* (IEEE, 2009) 428-432.
[10] R. Pordes *et al.*, J. Phys. Conf. Ser. **78**, 012057 (2007).
[11] J. Blomer *et al.*, J. Phys. Conf. Ser. **331**, 042003 (2011).
[12] D. Weitzel *et al.*, *opensciencegrid/StashCache: Multi-Origin Support* (Zenodo, 2017)

[13] P. Fuhrman and V. Gulzow, *2006 Euro-Par Parallel Processing* (Springer, 2006) 1106-1113.

[14] A. Dorigo, P. Elmer, F. Furanoand and A. Hanushevsky, *Proceedings of the 4th WSEAS International Conference on Telecommunications and Informatics* (WSEAS, Stevens Point, 2005) 46.

[15] Landscape Project, "Landscape" [software], version 1.0.3, 2016. Available from https://github.com/fifemon/probes/releases/tag/v1.0.3[accessed 2019-01-27]

[16] P. Couvares, T. Kosar, A. Roy, J. Weber and K. Wenger, *Workflows for e-Science* (Springer Press, 2007) 357-375.

[17] B. Holzman, L.A.T. Bauerdick, B. Bockelman *et al.*, Comput. Softw. Big Sci. **1**, 1 (2017).

[18] P. Mhashilkar *et al.*, *HEPCloud, an Elastic Hybrid HEP Facility using an Intelligent Decision Support System*, this conference, see https://indico.cern.ch/event/587955/contributions/2937279/

[19] M. Altunay *et al.*, *FERRY: Access Control and Quota Management Service*, this conference, see https://indico.cern.ch/event/587955/contributions/2937426/

[20] J. Teheran, D. Dykstra, and M. Altunay, FERMILAB-CONF-16-047-CD (2016).

[21] C. Serfon *et al.*, Nuclear and Particle Physics Proceedings **273-275**, 969 (2016).