Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

# FIFE Roadmap Workshop

Mike Kirby

FIFE Roadmap Workshop

Dec 5, 2017

FABRIC FOR FRONTIER EXPERIMENTS

# FIFE Roadmap Workshop

- The goal of the roadmap discussion is to both inform experiments and gather feedback about strategic infrastructure changes and computing service modifications.
- SCD has budgetary, security, and effort constraints that must be met.
- Must also keep track of industry best practices and available tools.
- Constant large meetings between experiments and service providers are not productive to developing strategy.
- Feedback concerning specific enhancements, changes, and bugs are important on day-to-day basis.

Does it all seem confusing and directionless?

# FIFE Roadmap Workshop

- The goal of the roadmap discussion is to both inform experiments and gather feedback about strategic infrastructure changes and computing service modifications.
- SCD has budgetary, security, and effort constraints that must be met.
- Must also keep track of industry best practices and available tools.
- Constant large meetings between experiments and service providers are not productive to developing strategy.
- Feedback concerning specific enhancements, changes, and bugs are important on day-to-day basis.

Hopefully, at the end of the day at least a few of the signs will be more understood

# Outline of topics

- SCD Communications and Service Now
- SC-PMT Preparation
- Data Storage (dCache, Enstore)
- SAM Data Management
- HEPCloud
- FERRY
- GPGrid refactoring to FermiGrid
- CVMFS
- Monitoring
- Continuous Integration
- POMS
- GPU resources and OSG submission

Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

# Communication (SCD/SPPM)

Margaret Votava

FIFE Roadmap Workshop

Dec 5, 2017

FIFE FABRIC FOR FRONTIER EXPERIMENTS

# Communication with SCD

Feedback from one experiment:

*"This year, many changes has been deployed even though the change is small or big,and many of changes was made without enough announcement or notification to experiments. Is there any plan to make the communication more efficiently between SCD and experiments? (It seems the announcement of scheduled outage service is not enough to track down the changes)"*

- We apologize for any lack of communications and would be thrilled to work with experiments to improve the communication channels. What we have:
- We announce the overall plan at the FIFE workshop - but not all experiments [can] attend. Has been replaced with this semi annual meeting.
- Had biweekly liaison meetings in which many liaisons did not attend, sending out announcements to liaisons.
- Quarterly FIFE notes
- Outage Calendar - includes schedule downtimes

What will work for you?

# Suggestions for Service Now

The impression we have from experimenters is that users are not very familiar with the SNOW interface

Do we need tutorials? How do we make understanding the usefulness of SNOW easier?

Requested improvements to Service Now:

- *the ability to search for and see tickets you are not already a watcher on*
- *and the ability to add people to the watch list of a ticket*
- *have SNOW interact better with email, specifically*
    - *Shortening subject lines*
    - *Keeping the subject consistent across all ticket stages*
    - *Not sending duplicate emails*
    - *Not frequently re-duplicating the ticket history*

# Preparation for SCPMT Review

- SCPMT 2018 Prep underway.
- Similar format as last year.
- Expect requests for spreadsheets (all experiments need this) and slide decks (not all experiments need this) by the end of the year. Information back to the division mid January.
- Will include a physics component this year - more details to follow.
- Note that several experiments have mentioned the effect of dcache problems on running both production and analysis jobs. We are making compromises in quality of service to fit in the budget. We only have so many dollars. We use SCPMT for you (the experiments) to help us prioritize the spending, and you can use SCPMT to tell the laboratory the impact of that plan on the experiment.

Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

# Storage

Bo Jayatilaka and Dmitry Litvinsev

FIFE Roadmap Workshop

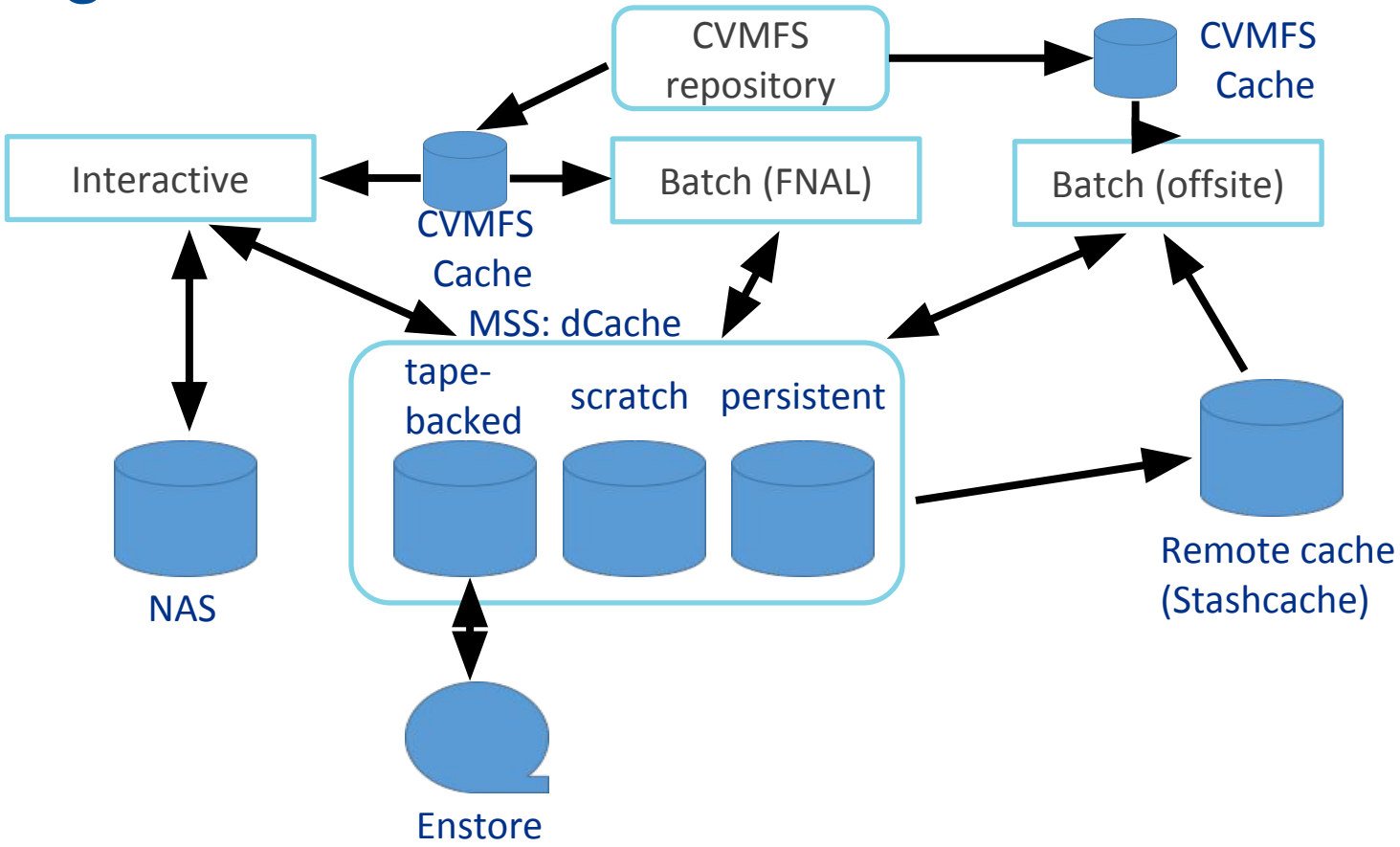
Dec 5, 2017

FIFE FABRIC FOR FRONTIER EXPERIMENTS

# Fermilab Scientific Storage Architecture

- Overall architecture for storage of scientific data and applications
- Some requirements are by necessity of scale
  - Removal of POSIX mounted storage (~~BlueArc~~ Network Attached Storage - NAS) from grid nodes - will remain on interactive nodes for now
- Assume three broad types of jobs
  - Production (centrally run), Analysis (user run), and Interactive (development and testing)
  - Production and Analysis jobs both assumed to be batch jobs
  - Both production and analysis jobs should be location-agnostic
- Define storage solutions: tape- and non-tape-backed mass storage, NAS, CVMFS
- Define data types: input experiment data, aux data (big or small), code, logfiles
- Still trying to understand impact of highly popular files that change rapidly
- Provide guidelines for each combination: (job type, data type) = solution
- SCD will provide assistance to experiments to adapt to recommended solutions where necessary

# What goes where



# Grapevine says: “~~BlueArc~~ NAS is going away”

- NAS is going away for **batch jobs**
  - All data areas are unmounted from GPGrid worker nodes and app areas are on the way
  - Why?
    - The system was never intended to be be hammered by thousands of grid jobs (and suffered from it frequently)
    - You are less tied to Fermilab-based computing
- Is NAS going away entirely?
  - Not for interactive use next year
  - **But**, SCD is actively exploring cheaper alternatives. This would **not** happen without a functional replacement
    - Same basic POSIX functionality for home dirs/interactive space

# Consequences of NAS unmounting on FermiGrid

- /exp/data already removed from mounts (soon for gridftp transfers (BestMan))
- /exp/app soon to be unmounted from worker nodes and gridftp
- Recognize there is effort required and will provide support to assist
- User built code distribution - need to get local area out to worker nodes
  - Larbatch already has a solution
  - Agreed to help NOvA modify workflows
  - Will work with other experiments as well
  - Onus is still on experiments to have portable libraries and understand build environment to allow for packaging
- Need to understand how to distribute tarballs without overloading dCache
- Nightly build testing on grid nodes - design in place for development of CVMFS with garbage collection - details later

# Message: dCache roadmap - now

Upgraded to 2.16 on Sep. Improved namespace database schema

New in 2.16:

- Support for HA dCache (core components). Not configured yet in production.
- File resilience service (we are not using it yet).
- Support for multiple type checksums for a file.
- Scalable SRM front-end service.
- Ability to query file locality (and other attributes) via WebDAV for individual files and directory listings which greatly reduced time it takes to determine what files are in cache.

# Message: dCache roadmap – 3.2

FABRIC FOR FRONTIER EXPERIMENTS

dCache is moving towards introducing SRM functionality (+ more) utilizing RESTful API:

- Directory functions.
- Recall from tape.
- Change file AccessLatency/RetentionPolict (e.g. from disk-only) to tape-backed.

New in 3.2:

- SRM-like RESTful API and GUI to interact with it (dCacheView).
- RESTful API for data transfer monitoring and GUI to interact with it.
- Experimental support for CEPH on pools.
- Improved hot pool detection (and mitigation).
- TLS encryption of internal messaging (optional).
- OpenId connect (OIDC) support for 3rd party WebDAV transfers.
- Macaroons support in WebDAV door - these are better cookies (I disagree with name), authentication tokens
- Improvements in NFS support.

We plan to upgrade to 3.2 this FY. Expect fast upgrade w/ short downtime

# Message: dCache roadmap – future

- Continue developing RESTful API expanding SRM-like features and introducing user selectable, policy driven data storage quality levels (QoS).
- NFS:
  - introduce **mutable files** for disk-only files. It will be possible to modify files in dCache.
  - NFS 4.2 protocol.
- Support Data Federations.
- Provide ‘cloud like’ storage ’sync & share’ , OIDC based authentication.

# What are the recommendations from SCD?

- For interactive access, utilize XRootD protocol for reading files from dCache - xrdcp, xrdfs (ls, mkdir, mv, rm, etc)
  - Recognize that this may necessitate new documentation and tutorials
  - Don’t yet know the TChain limit of XRootD - more on the next slide
- Minimize access over NFSv4 to include ls, cp, etc.
- When copying to and from dCache use ifdhc tools
  - These tools will be adapted as we improve understanding of dCache performance and tune doors
- Keep less than 10K files in a directory - 2K is optimal limit
- Please help us by educated your users about scaling, resource limitations, and a little bit of computer science
- Continue to ask questions and get recommendations

# How is dCache doing these days?

FABRIC FOR FRONTIER EXPERIMENTS

- Yesterday morning it wasn't great - dCache struggled for 20 minutes on Sunday - Not sure why (or if) it led to spectacular meltdown of NFS mounts on interactive nodes - was even XRootD
- Both NOvA and MicroBooNE commonly report difficulties with NFS mount points and these meltdowns can cause interactive nodes to hang - extremely disruptive
- Several options to prevent interactive nodes from crashing during dCache slow downs:
  - unmount NFS - dislike this option (CMS uses this model for EOS access - only xrootd and eos-ls - choice specifically made after POSIX access was not scalable)
  - Remove write capabilities - revert to NFSv3
  - Allow only ls and xrootd access on interactive nodes
  - Hope that NFSv4.2 alleviates problems with NFS access
- Continues to be a problem that users often do not consider scaling and load - people are very clever and yet not always very smart.
- XRootD strongly recommended - migration is not difficult (interactive, SAM project, etc)
- Again, XRootD is not a panacea, but also NFS mounts should not be used for heavy lifting
- SCD recognizes that we are once again a victim of our success - users heavily accessing dCache - NFSv4 mounts has allowed any one person to melt it down

# Example decision to be made?

- dCache is a collection of disks that is organized into pools
- Pools can be configured numerous ways
- Recently, NOvA experienced a ~5 day for persistent dCache files starting Friday evening - cascade failure of worst kind
  - Support for persistent dCache is 8x5
  - Diagnosed as a failed RAID card - no spare
  - Ordered (vendor failed to deliver on time)
  - installed in ~36 hours
- Choice to be made by experiments - storage volume vs replication - there is no funding for additional disk
- Help us understand what you want and how you can make the choice that best suits you needs

Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

# Data Management

Robert Illingworth

FIFE Roadmap Workshop

Dec 5, 2017

FIFE FABRIC FOR FRONTIER EXPERIMENTS

# Continue to evolve SAM

We plan to continue to evolve the SAM system in response to user needs and requests

- Use FERRY for proxy -> account mappings (no need for separate registration of certificates) - implement 1 month after FERRY
- Better handling of preempted jobs
  - Phase 1: detect when a job has been restarted and flag the old one as being dead
  - Phase 2: “unwind” the processing history of restarted jobs and make the files available for processing again
- Better tools for transferring datasets around
- Better handling of source location priorities at different sites
- Performance improvements where we can find them
  - But we’ve probably got all the low hanging fruit by now…
- Note that SAM4Users considered in production - feature requests and bug fixes

# FTS

New version of FTS available for testing:

- Can now read dCache checksums directly from pnfs space; this means we can grab the checksum from scratch or persistent without rereading the file – no i/o overhead
  - dCache uses adler32 checksums, not the enstore algorithm
    - recommend migrating from enstore to adler32 generally as it's more standard. It is possible to combine the enstore checksum and the file size to get the adler32 value, so we can still check against the enstore database.
  - Not actually a new feature, but SAM can store multiple checksums for a file. So if you want crc32c, MD5, or even SHA-512 you can have it. The new FTS can compute many more algorithms itself
- FTS3 (the CERN data transfer service; I know it's confusing) support. Mostly of interest to those who need to move files over the WAN. This is still somewhat experimental and needs more large scale testing.
- Now uses standard python distribution tools – allows installation without using ups/upd.

# Future data management

- We're starting to think about the future of data management
  - No plans to force migration on existing experiments – SAM will still be supported
- But what are the needs of upcoming experiments?
  - What works well, what works badly?
  - What can we ~~steal~~ learn from outside FIFE?
  - How do we handle new storage paradigms (for example, event stores)?
    - note event based data models requested for R&D in Community White Paper
  - Can we benefit from even more use of common tools (ie with HL-LHC)?
- We'd welcome comments and feedback on this

Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

# HEPCloud: Progress and Roadmap

Parag Mhashilkar

FIFE Roadmap Workshop

Dec 5, 2017

FIFE FABRIC FOR FRONTIER EXPERIMENTS

# Changing Roles of HEP Facilities

## Drivers

- Elasticity: Experiments don't need all the resources all the time: Provisioning needs to be adaptable, providing facility "elasticity" to handle "burst usage"
- Capacity and Cost: Incorporate and manage "rental" and "allocated" resources in an efficient and cost effective manner.

Need to modernize and evolve the way we do computing

# HEPCloud: The Evolved Facility

## Vision Statement:

- HEPCloud is envisioned as a portal to an ecosystem of diverse computing resources commercial or academic
- Provides “complete solutions” to users, with agreed upon levels of service
- The Facility routes to local or remote resources based on workflow requirements, cost, and efficiency of accessing various resources
- Manages allocations of users to target compute engines

## Pilot project to explore feasibility, capability of HEPCloud:

- Goal of moving into production in 2018
- Seed money provided by industry in FY2016 and FY2017

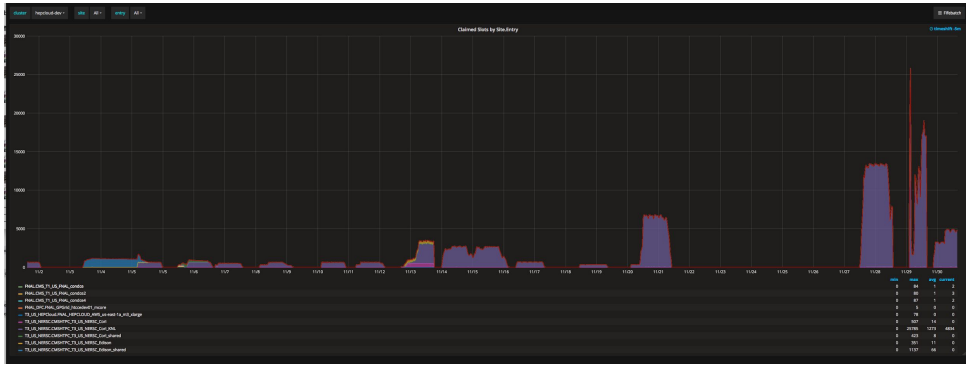
# HEPCloud Subprojects

- Security Controls
  - Security risk analysis, mitigations, documentation, training, […]
  - Automated monitoring probes and alarms on resource usage, finances, allocations, security, […]
  - This will include safeguards against spending all of your allocation at once on less important jobs
- Decision Engine
  - Intelligent decision support system
  - Uses information from several sources like job queues, monitoring, finances, allocations, etc. to make decisions on facility expansion

**Several other HEPCloud related projects and work that directly impacts users and experiments are covered in the GPGrid slides from Tony Tiradani**

# HEPCloud: Phase II

- Successful demo of Decision Engine in November: Proof of concept: 1400+ slots provisioned in AWS + NERSC to run CMS workflow




- Security Controls: Interim Security Controls developed and deployed
  - Includes critical security controls for AWS and NERSC
  - Approved by CS Board in December
- Phase 2 will be closed in December 2017

# HEPCloud Roadmap

- Support second Commercial Cloud Provider (Google)
- Leverage experience towards leadership computing facilities
  - Harden and test HTCondor-CE interface @ ALCF
- Elasticity Considerations: Understand how to integrate all of these components – always driven by the experiment needs
  - Consider network capacity of a site
  - Consider storage capacity and data locality
  - GPUs & Machine learning
  - Site specific security policies – applies to HPC facilities
- Singularity & Containers
- Phase 3 of the HEPCloud Project: Move to production at the end of 2018
  - Develop and deploy comprehensive security controls
  - Extending and hardening the Decision Engine
  - Acquire Authorization to Operate (ATO) from the Fermilab DOE Site Office

Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

# FERRY

Tanya Levshina/Mine Altunay

FIFE Roadmap Workshop

Dec 5, 2017

FABRIC FOR FRONTIER EXPERIMENTS

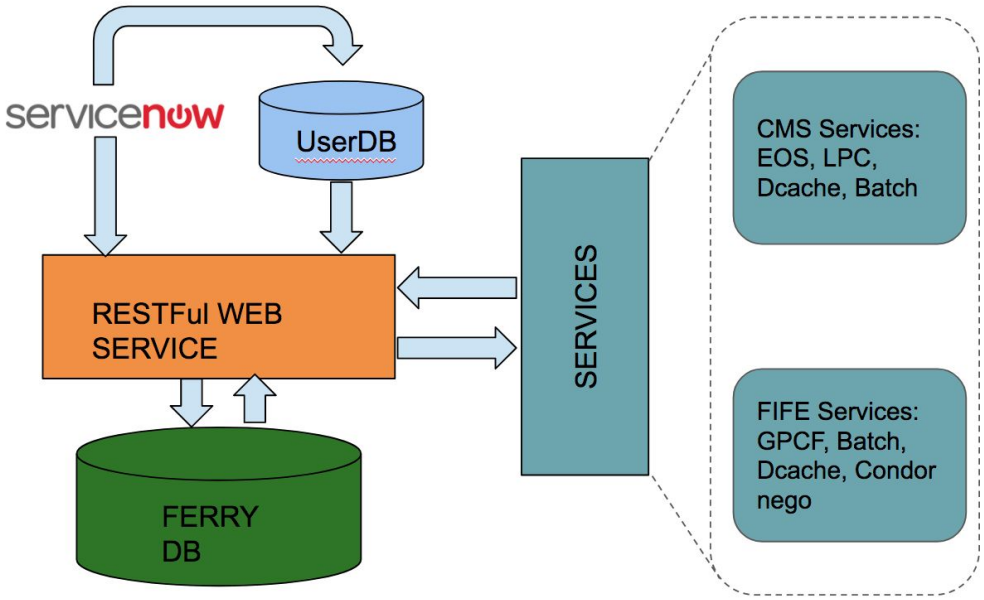
# Frontier Experiments RegistrY (FERRY)

- Creates account registry and quota management for Fermilab experimenters.
- Is “the source of truth” for various services, including FermiGrid Batch, dCache, EOS, NIS, POMS, SAM and some others.
- Provides info about users’ affiliations, unix groups, preferred shell, current quotas and relevant paths to home directory or storage.
- Stores all information in a central repository.
- Provides Restful API that will allow all existing services to pull this information from the central repository.
- Automates as much a possible the workflows that are initiated from Service Now as requests for an account, storage, and quota increase.

# Frontier Experiments RegistrY (FERRY)

Why Now:

- Too many sources of identity and quota attributes, no single source of truth.
- Out of control. There are a lot of obsolete users and groups on various clusters and in services databases, mismatches between various sources.
- VOMS-admin and GUMS are going away.
- Need to streamline flow from ticket generation to actions and have a record of the actions.
- Allow to save efforts on support on the long run.




Target production date: mid 2018.

Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

# GPGrid Refactor

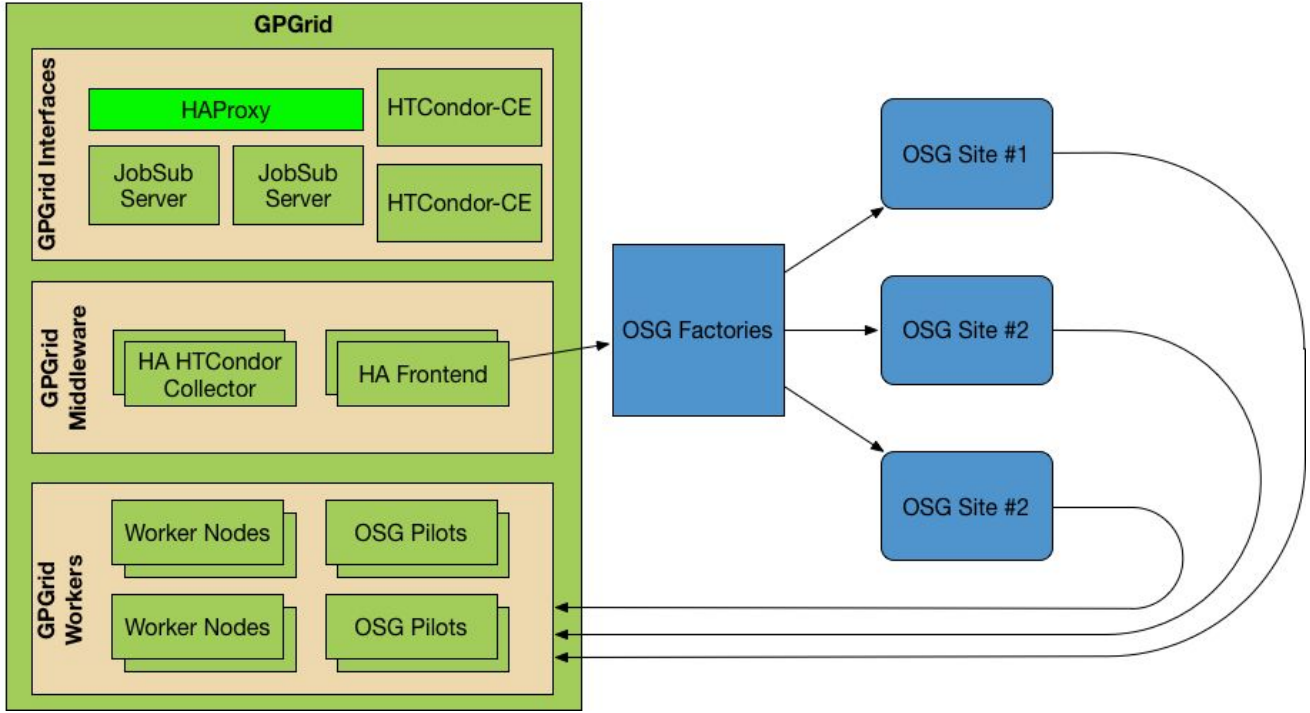
Anthony Tiradani

FIFE Roadmap Workshop

Dec 5, 2017

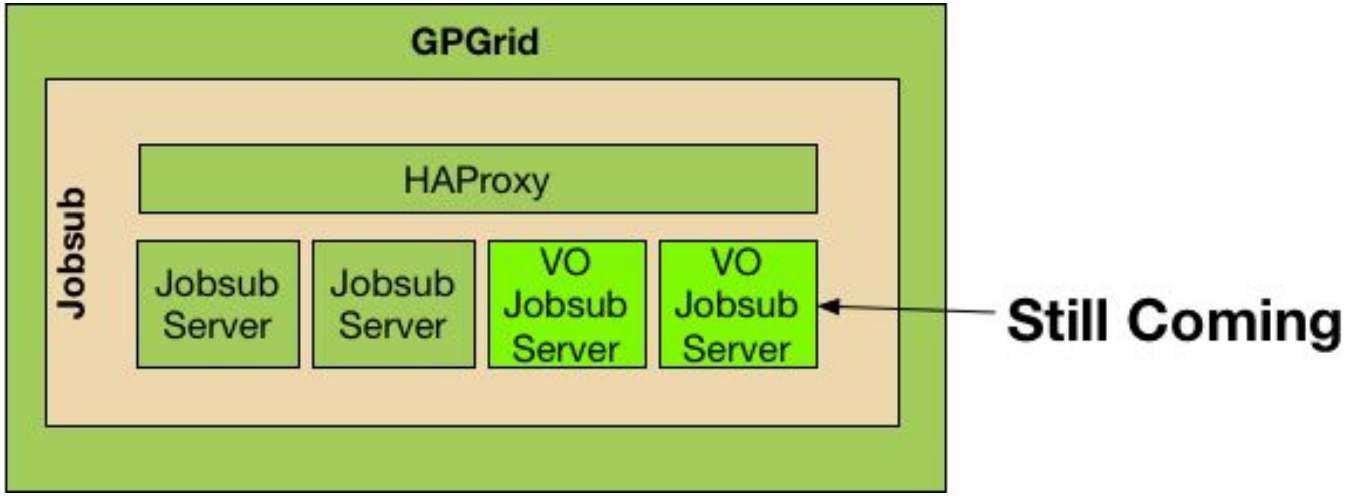
FIFE FABRIC FOR FRONTIER EXPERIMENTS

# GPGrid Refactor -> FermiGrid




The refactor is complete.

# GPGrid Refactor -> FermiGrid




- DNS Round Robin successfully replaced - avoids overloading one server
- VO Specific Jobsub servers still coming
  - this has been tested using VMs and waiting on dedicated hardware - Q3 FY2018
  - allows isolation of VO and possibility of more unique configurations (no promises)

# Worker Node Environment

- Docker is now the standard environment:
  - This means that you job now runs in a container
  - In the future, can have different flavors of containers
  - Right now only an SLF6 container is deployed
  - Created hiccups at the beginning of operations (libraries installed on interactive nodes not installed on worker nodes) - not enough testing? Communications breakdown?
  - Emphasize to users that custom installs of libraries adds complications
- New node, fnpctest1.fnal.gov, available for testing code in a worker node environment
  - Is OS is installed per a standard worker node (bare bones SLF7 kernel install)
  - Allows direct ssh access for testing code (you have to select OS between SLF6 and SLF7 - in future can install additional containers)
  - Is not connected to GPGrid, so it is not an actual worker node
  - Contact FIFE Support to obtain access

# Worker Node Environment (Continued)

- CentOS 6.9 is the standard container:
  - only used CentOS because SLF 6 Docker container wasn't available at the time of implementation
  - SLF 6 containers now available. Will migrate to SLF 6 in the future, but not a high priority right now
  - UPS_OVERRIDE is set due to kernel differences
- Coming soon: SLF7 containers:
  - Available now on fnpctest1.fnal.gov
  - UPS_OVERRIDE will not be set
  - Not yet ready for production use
  - The demand for SLF7 has not been high
- Support for 32-bit legacy software
  - There are no specific plans
  - Happy to provide platforms to test the libraries on, but not willing to make blanket statements about support

# Worker Node Environment (Continued)

- Singularity is now approved for use at Fermilab
  - Support for running Singularity images is coming, planned for first quarter of 2018
  - Docker allows for some privileged user actions - security concern so do not want to use untrusted docker images (e.g. an image made by an experiment)
  - Singularity isolates a container from the OS it is run on
  - Policies and effort for support for creating, publishing, and distributing Singularity images must be defined

Thanks to Brian Bockelman for the slide

How are Singularity images going to be distributed?
How are we to know what is experiment approved?
Who supports experiments when they need help building the Singularity images?
What are the tools that they will use?
Who supports those tools?

Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

# CVMFS Roadmap Update

Dave Dykstra

FIFE Roadmap Workshop

Dec 5, 2017

FIFE FABRIC FOR FRONTIER EXPERIMENTS

# Migration from BlueArc to CVMFS

- Many experiments are already using CVMFS – it's required for OSG usage – but now the remaining ones are expected to also migrate.
  - Two paths to take
    - Very small projects, too small to be a VO (and typically shared between experiments) go into fermilab.opensciencegrid.org
    - Larger projects, including most standalone experiments, get their own opensciencegrid.org repository
  - We could also consider fnal.gov repositories if they're only ever expected to be used at Fermilab
- No quotas on CVMFS repositories, but space is not unlimited so only publish files that are expected to be used on worker nodes - we will continue to monitor usage

# Nightly build repositories

Some experiments do nightly builds and would like to make them available to FermiGrid and possibly the OSG

- The main challenge for CVMFS is the churn; that is, files are frequently published & deleted, and normally CVMFS does not delete any files from the repository even when deleted by the user
- CVMFS garbage collection is now pretty reliable
  - Complex & time-consuming, however, so not recommended for normal software distribution repositories
- Request a separate repository with the VO name, a hyphen, and purpose, e.g. nova-nightlies.opensciencegrid.org
  - Request that garbage collection be enabled
- All 4 LHC experiments already have separate nightly build repos
- VOs will be responsible to run ‘cvmfs_server gc <reponame>’
  - Can’t run as same time as other transaction in the repo
  - Daily or even weekly is sufficient
  - Check return code
- Will work with experiments to meet their needs and provide documentation

# osgstorage.org data repositories

For data files that are accessed by many jobs but not all of them, (e.g. Genie flux files), use osgstorage.org repositories

- Provides convenient POSIX access
- Uses normal CVMFS mechanism for distributing metadata, but data files are read directly from high speed storage
  - FNAL workernodes read from FNAL dCache when that's the data's original source
  - For OSG sites, they files are cached on StashCache servers at some universities
  - Minimal caching on worker nodes (1GB)
- The process of scanning dCache for changes and publishing them to CVMFS is currently done at the University of Nebraska

# What's coming up for CVMFS?

FABRIC FOR FRONTIER EXPERIMENTS

- Faster times for published repository to show up on worker nodes
  - Stratum 1 delays 15 minutes -> 5 minutes (there's two stratum 1s between FNAL publish machine and worker nodes)
  - Squid cache delay 2 minutes -> 1 minute
  - cvmfs-2.4.x client delays 15 minutes -> 4 minutes
    - Also 4 minutes if repositories are published by >= cvmfs-2.4.2
- Client cache plugins, multi-level
  - E.g., could cache to local disk or RAM and via xrootd to dCache
- X.509 proxy protection on reading data
  - Currently used by LIGO, but doesn't work with StashCache servers
- X.509 proxy protection on reading filenames
  - Currently experimental. Doesn't work with squids.
- Ability for multiple people to publish to separate areas of same repository simultaneously. Will be in experimental release cvmfs-2.5 expected in January

Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

# FIFEMON

Kevin Retzke

FIFE Roadmap Workshop

Dec 5, 2017

FABRIC FOR FRONTIER EXPERIMENTS

# Monitoring Status - recent past

- Recently completed deployment of new monitoring hardware for better performance and reliability - old hardware out of warrantee at often at 100% utilization
- Much faster, have not yet measured improved speed
- Completely containerized (Docker) application stack to simplify development, deployment, and maintenance
- Upgraded servers allows for additional dashboards - make suggestions

**4 billion**
documents in Elasticsearch

**2 million**
metrics in Graphite

**25 thousand**
frontend requests per day

# Monitoring Roadmap (Near-tem)

- Dashboard improvements/refresh - review usage pattern of current dashboards and incorporate new data available
- Active alerting
  - Efficiency warnings
  - Threshold alarms (e.g. dCache space)
  - Anomaly detection (e.g. black hole site/node)
- Improve direct data access and documentation - hope to reduce the learning curve from several hours or sessions
- Offsite data collection (e.g. protoDUNE at CERN - update Experiment Overview to include data from both)
- Remember that alerts and alarms can be set to inform production group and/or offline coordinators

Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

# Continuous Integration

Vito Di Benedetto

FIFE Roadmap Workshop

Dec 5, 2017

FIFE FABRIC FOR FRONTIER EXPERIMENTS

# CI On-boarded Experiments/Projects

- LArSoft-based experiments:
  - uBooNE, DUNE/protoDUNE, LArIAT, ArgoNeuT and SBND
    - DUNE has 2 actively running CI validation workflows for 'Tracking efficiency' and 'Shower reconstruction efficiency'
    - uBooNE is working to to have a full set of CI validation workflows for MC, data and MC/data validation
- NOvA
- MINERvA
- GENIE (code testing in production, CI validation workflow is almost ready)
- g-2 (recently in production)
- GlideinWMS
- Art (recently in production)
- mu2e (under development)

# Core CI Development Status

- Provides a CI web application to monitor the historical status of the code.
- Support for GIT, SVN and CVS repositories.
- Support for an arbitrary build/test workflow using different build tools, the most common CI workflow includes:
  - code checkout, build, unit test, install, CI tests.
- Support for CPU intensive tests using the CI validation workflow to run test on the grid
  - provides access to validation plots/reports and job resource usage info through the CI web application.

# CI Development Roadmap

- Improve and extend existing functionality.
- Support for code memory/runtime profiling using dedicated tools.
- Simplify the required configuration keeping the flexibility.
- On-board new experiment/projects that make request of the CI service.

Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

# POMS

Anna Mazzacone/Marc Mengel

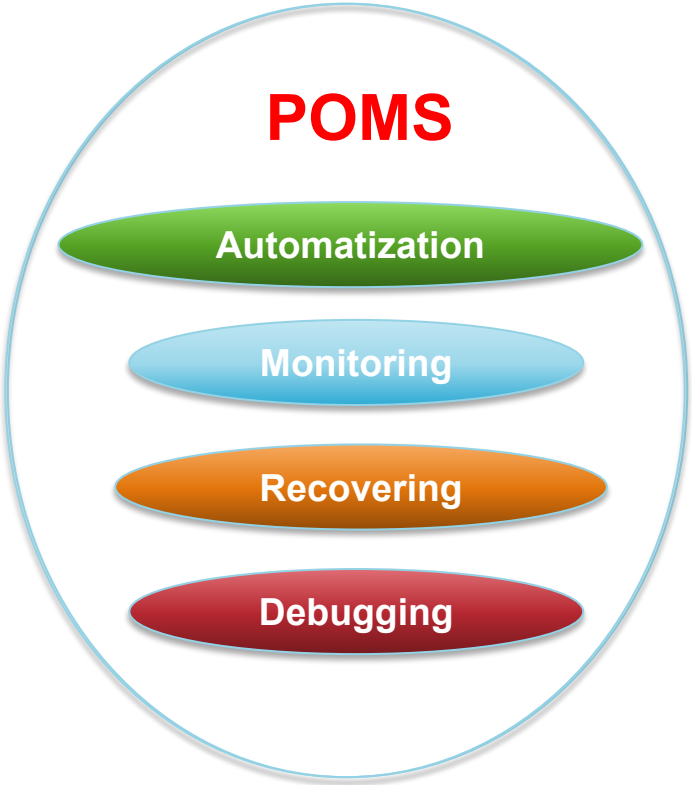
FIFE Roadmap Workshop

Dec 5, 2017

FIFE FABRIC FOR FRONTIER EXPERIMENTS

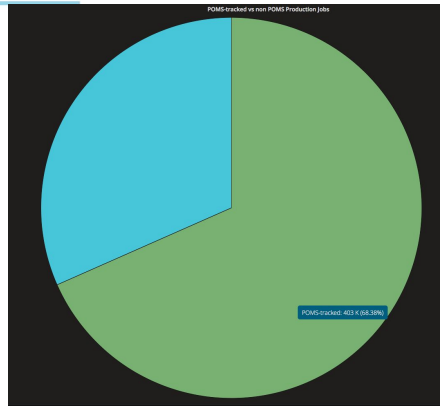
# POMS

- POMS goal is:
  - Minimize the effort needed for MC production and data processing
  - Increase the efficiency while optimizing resources
- POMS provides:
  - A dashboard to summarize services' status: are they up and running?
  - A calendar for scheduled downtime of services.
  - A web interface to create campaign stages
  - Job submissions and files delivery monitoring
  - Automatic launch job submissions at a scheduled time.
  - Recover of jobs for files that didn't process automatically.
  - Hold or kill submissions when critical issue are present
  - Dependent campaigns workflows.
  - Interface with SAM to track files processed in a campaign.
  - "Triage" interface for examining individual jobs/logs
  - Debugging failures.



# Experiment Production Campaigns

Experiments continue to onboard to POMS
Feedback seems to be positive
Good response time for development
Has been some requirement for the modification of experiment workflows




NOvA, uBooNE, G-2 and Dune production jobs (68% is tracked via POMS during last month)

## DUNE Data Catalog

**DUNE MCC9 Projects**

Updated 28-Nov-2017 18:01

| SAM Dataset Name | SAM Query Links | Comment |
|---|---|---|
| MUSUN_Cosmic_DUNEFDSP_mcc9.1 | describe, summary, files | MUSUN Cosmic-Ray Sample for a 10 kt FD module |
| prodgenie_atmnu_max_dune10kt_1x2x6_mcc9.2 | describe, summary, files | Atmospheric neutrinos using Bartol solar max flux |
| prodgenie_atmnu_min_dune10kt_1x2x6_mcc9.2 | describe, summary, files | Atmospheric neutrinos using Bartol solar min flux |
| prodgenie_atmnutau_max_dune10kt_1x2x6_mcc9.2 | describe, summary, files | Atmospheric neutrinos using Bartol solar max flux. Both muon neutrino and electron neutrino are converted to tau neutrino. |

# POMS

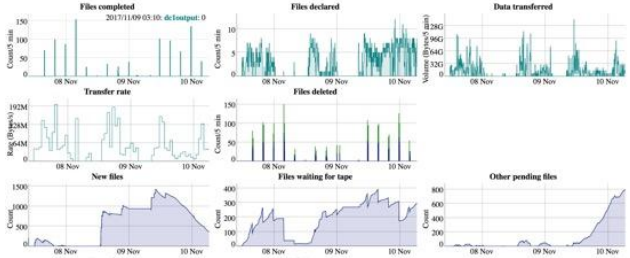
- The Production Operations Management System (POMS) is an SCD service which assists the production teams and the analysis groups of the experiments in their scientific computational work.
- The first Data Challenge (DC1) for protoDUNE took place the week of November 6th and a POMS Keepup campaign included end to end testing of the offline processing infrastructure.

File Info by Submission

2017-11-08 00:00 to 2017-11-11 00:00

<previous 3.0 days | next 3.0 days >

| jobsub_jobid | project | date | submit-ted | deliv-ered SAM | deliv-ered logs | out-put logs | con-sumed |
|---|---|---|---|---|---|---|---|
| 1036282.0@ jobsub01.fnal.gov | dunepro-null_lar.sh_20171110_200113_3048908 | 2017-11-10 20:01 | 53 | 0 | 0 | 0 | 0 |
| 1061948.0@ jobsub02.fnal.gov | dunepro-null_lar.sh_20171110_160111_645508 | 2017-11-10 16:01 | 30 | 0 | 0 | 0 | 0 |
| 1057008.0@ jobsub02.fnal.gov | dunepro-null_lar.sh_20171110_120111_4105258 | 2017-11-10 12:01 | 73 | 71 | 0 | 0 | 71 |
| 1024023.0@ jobsub01.fnal.gov | dunepro-null_lar.sh_20171110_080120_1773067 | 2017-11-10 08:01 | 112 | 112 | 0 | 0 | 112 |
| 1055090.0@ jobsub02.fnal.gov | dunepro-null_lar.sh_20171110_064340_3625154 | 2017-11-10 06:43 | 1 | 1 | 0 | 0 | 1 |
| 1054369.0@ jobsub02.fnal.gov | dunepro-null_lar.sh_20171110_040116_3425472 | 2017-11-10 04:01 | 98 | 98 | 0 | 0 | 98 |
| 1053235.0@ jobsub02.fnal.gov | dunepro-null_lar.sh_20171110_000118_3200816 | 2017-11-10 00:01 | 22 | 22 | 0 | 0 | 22 |




- A keepup processing was setup to run every 4 hours (6 submission/24 hours).
- At each submission, all new files coming from CERN were processed.
- The output files were sent to the offline dropbox where an FTS instance was configured to copy them on tape (@FNAL and @CERN).
- POMS was also configured to automatically recover data files that were not processed.

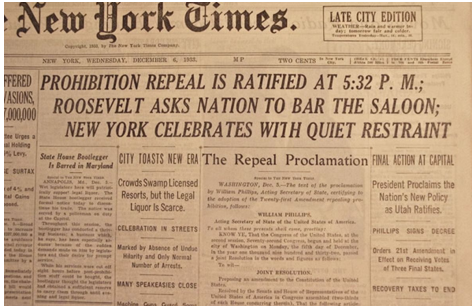
Fermilab | U.S. DEPARTMENT OF ENERGY | Office of Science

# GPU resources and OSG submission


Ken Herner


FIFE Roadmap Workshop

Dec 5, 2017



New York Times. LATE CITY EDITION

NEW YORK, WEDNESDAY, DECEMBER 6, 1933. M.P. TWO CENTS

PROHIBITION REPEAL IS RATIFIED AT 5:32 P. M.; ROOSEVELT ASKS NATION TO BAR THE SALOON; NEW YORK CELEBRATES WITH QUIET RESTRAINT

State House Bootlegger Is Barred in Maryland

CITY TOASTS NEW ERA

Crowds Swamp Licensed Resorts, but the Legal Liquor Is Scarce.

CELEBRATION IN STREETS

Marked by Absence of Undue Hilarity and Only Normal Number of Arrests.

MANY SPEAKEASIES CLOSE

The Repeal Proclamation

FINAL ACTION AT CAPITAL

President Proclaims the Nation's New Policy as Utah Ratifies.

PHILLIPS SIGNS DECREE

Orders 21st Amendment in Effect on Receiving Votes of Three Final States.

RECOVERY TAXES TO END

*Dec 5, 1933*

FIFE FABRIC FOR FRONTIER EXPERIMENTS

# GPU Access

- Some of you may be familiar with the Wilson Cluster at Fermilab; this requires a separate account
- Other GPU clusters available via OSG and reachable via jobsub, with some extra options
  - Today, jobs will likely end up at Nebraska or Syracuse; UCSD hopefully coming very soon. Unfortunately it tends to be slow going since this is somewhat new for all involved
- General GPU advice
  - Use the following slides to get information about the GPUs in the job
  - Don't confine yourself to only the latest and greatest models unless absolutely necessary. They have the longest wait.

# A GPU test job

- In general you just have to add **--lines='+RequestGPUs=1'** to your jobsub_submit command
  - The FNAL frontend will see this and steer to you an appropriate site
  - Can request >1 but no promises made about availability or fast start time. Could also add options for specific site if needed (not recommended)
- So try the following:

```
$ jobsub_submit -G your_expt -M --memory=1000MB --disk=1GB \
 --expected-lifetime=1h -N 8 \
--resource-provides=usage_model=OFFSITE --lines='+RequestGPUs=1' \
file:///nashome/k/kherner/basicscript_GPU.sh
```

# Getting GPU capabilities

- GPU information is now advertised in machine classad
- You can look for CUDA or OpenCL support
- NVIDIA driver info
- Attributes are advertised in the form of: <LANG><KEY>=<ATTR>
- To get all info about the machine, you can do "cat ${_CONDOR_JOB_IWD}/.machine.ad" somewhere in your executable. Then you can look for specific attributes

# GPU attribute examples

CUDAGlobalMemoryMb = 12194
CUDARuntimeVersion = 6.0
CUDACapability = 6.0
CUDAComputeUnits = 56
CUDAClockMhz = 1328.5
CUDACoresPerCU = 192
CUDADriverVersion = 8.0
CUDAECCEnabled = false
CUDADeviceName = "Tesla P100-PCIE-12GB"
NV_DRIVER = 375.66

Example of machine.ad dump on a CUDA-installed node

OCLOpenCLVersion = 1.2
OCLGlobalMemoryMb = 2001
OCLComputeUnits = 5
OCLECCEnabled = false
OCLDeviceName = "GeForce GTX 750 Ti"
OCLClockMhz = 1084
NV_DRIVER = 375.66

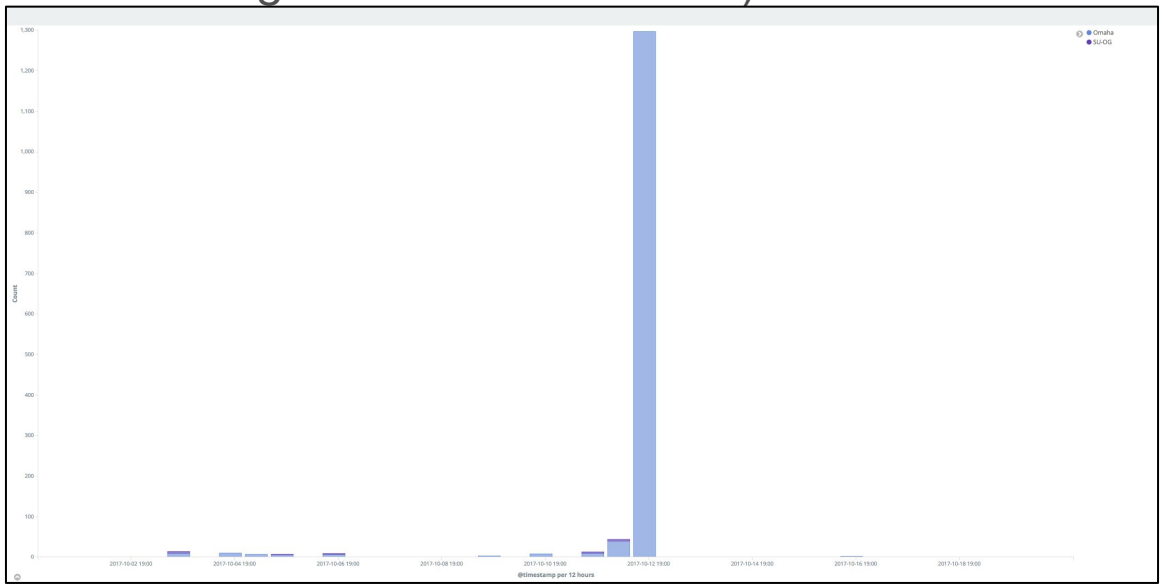
Example of machine.ad dump on a OpenCL-installed node

# Using these variables in job requirements

- Using jobsub
- Example where we want to get a Tesla P100 GPU:
  - jobsub_submit ... --lines='+RequestGPUs=1' --append_condor_requirements='(regexp(\"^Tesla\ P100.+$\",TARGET.CUDADeviceName,\"i\")||regexp(\"^Tesla\ P100.+$\",TARGET.OCLDeviceName,\"i\"))'
  - Be careful about cutting/pasting this -- unicode/ascii issues with python 2!
- Find OPENCL support:
  - --lines='+RequestGPUs=1'--append_condor_requirements=’isUndefined(TARGET.OCLDeviceName)==FALSE’
- Finding CUDA support is similar to above (OCLDeviceName → CUDADeviceName)
- Any other combination of vars on previous slide possible. Need to be careful about escaping quotes and spaces with the backslash (\) though.

# GPUs in action

Wes Gohn recently completed a CUDA-based simulation study using OSG GPUs. (Will appear in a forthcoming version of FIFENotes)




G-2 jobs ran by Wes using GPUs on Omaha and Syracuse clusters

# OSG Submission

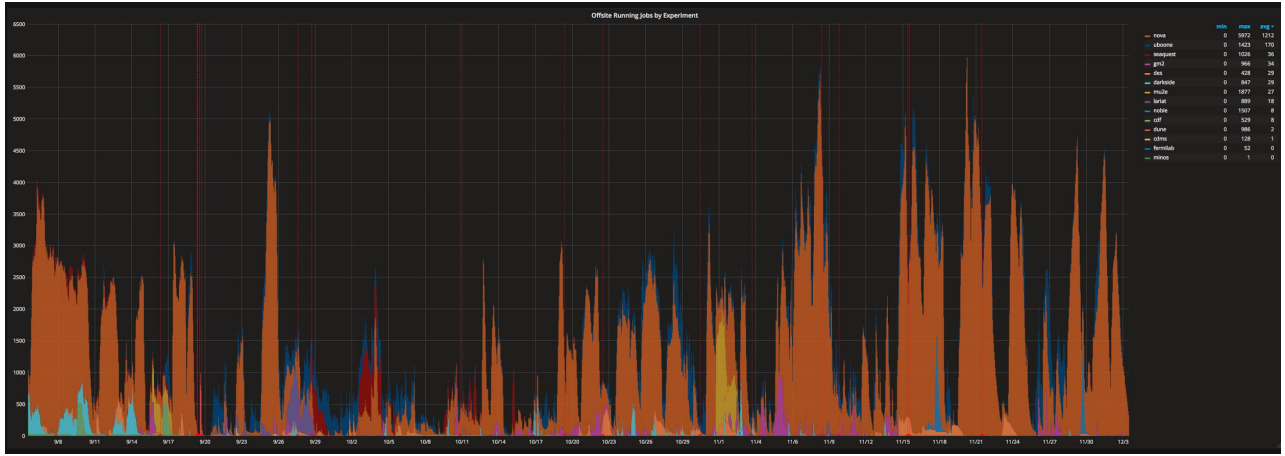
- Not much is changing in the short term

  Since BlueArc is going away, there is nothing special about GPGrid, so there's no reason not to take advantage of free slots
- Standard advice continues to apply: submit to all available sites and let the system pick
- *Do alert us ASAP if you find problems on sites*
- Remember to allow sufficient data transfer times in your job time requests.

**Important note: the UPS_OVERRIDE variable is NOT set by default offsite. You will need to do that in the job (or pass the variable with -e in your jobsub_submit command) if you need it set.**

# OSG Submission

- We have new experiments that are easily moved to OSG: g-2, seaquest, noble, ICARUS




- NOvA is fully utilizing OSG sites; successfully running on more than 17.
- We are adding more and more dedicated sites from Europe (protoDune on CERN Tier-0, g-2 at Pisa and CNAF, Mu2e on JINR).