Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

FIFE Roadmap Workshop

FIFE Roadmap Workshop
17 January 2019

FABRIC FOR FRONTIER EXPERIMENTS

Fermilab

FIFE Team Changes

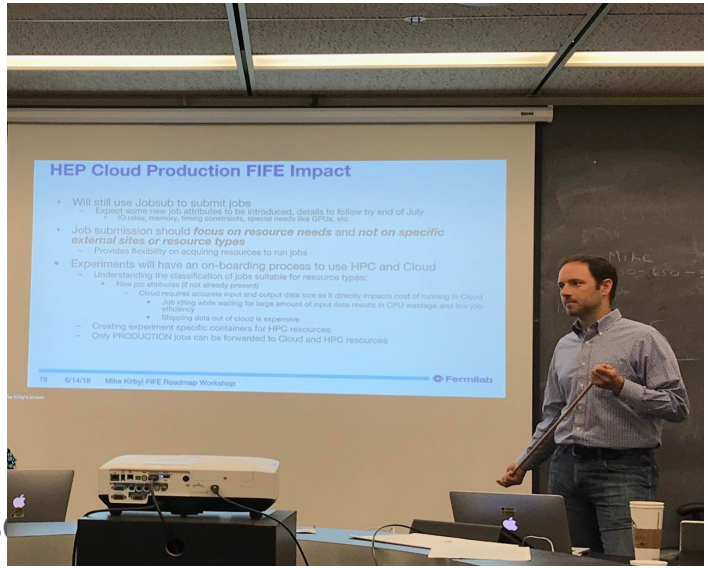
Ken Herner replacing Mike Kirby as FIFE Project Lead

Tanya Levshina remains co-lead

Many thanks to Mike for his leadership and wish him well in his new role in the DUNE Computing Consortium: Host Lab Technical Lead and Technical

Host Lab Technical Lead and Technical Board repres

Host Lab Technical Lead and Technical Board representative



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?

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

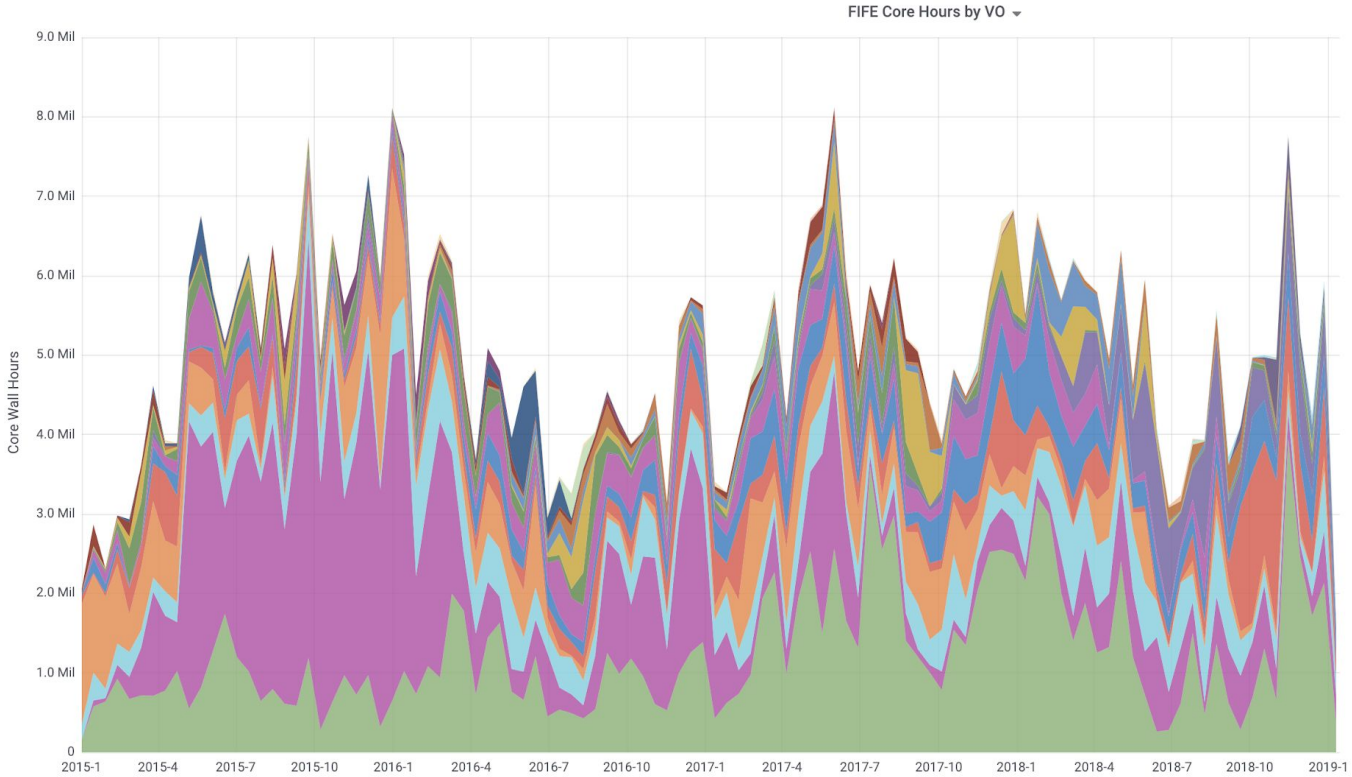
Expectations from the workshop

- Experiments and providers communicate about current services in a productive conversation
- Service providers present plans for the roadmap of their services and changes for the next 6-12 months
- Experiments help to guide the process of developing requirements for changes and new services (should not develop requirements during this workshop)
- Everyone has a better understanding of the direction of SCD services and experiment computing models
- Please feel free to contact FIFE to schedule working meetings with service providers - especially if coordination is required

Policy/procedure on killing expt workflows

- Occasionally an experiment will do something that impacts all other experiments/entire cluster.
- Working on a policy for when/how SCD will intervene in these cases.
  - Requires input from FIFE, experiments, service providers
- Requires development across:
  - Diagnostic tools (How do we decide there's a problem?)
  - Policy (What are the thresholds before which SCD takes action?)
  - Implementation (what will the actions be?)
- Policy will be generally informed by similar WLCG policies

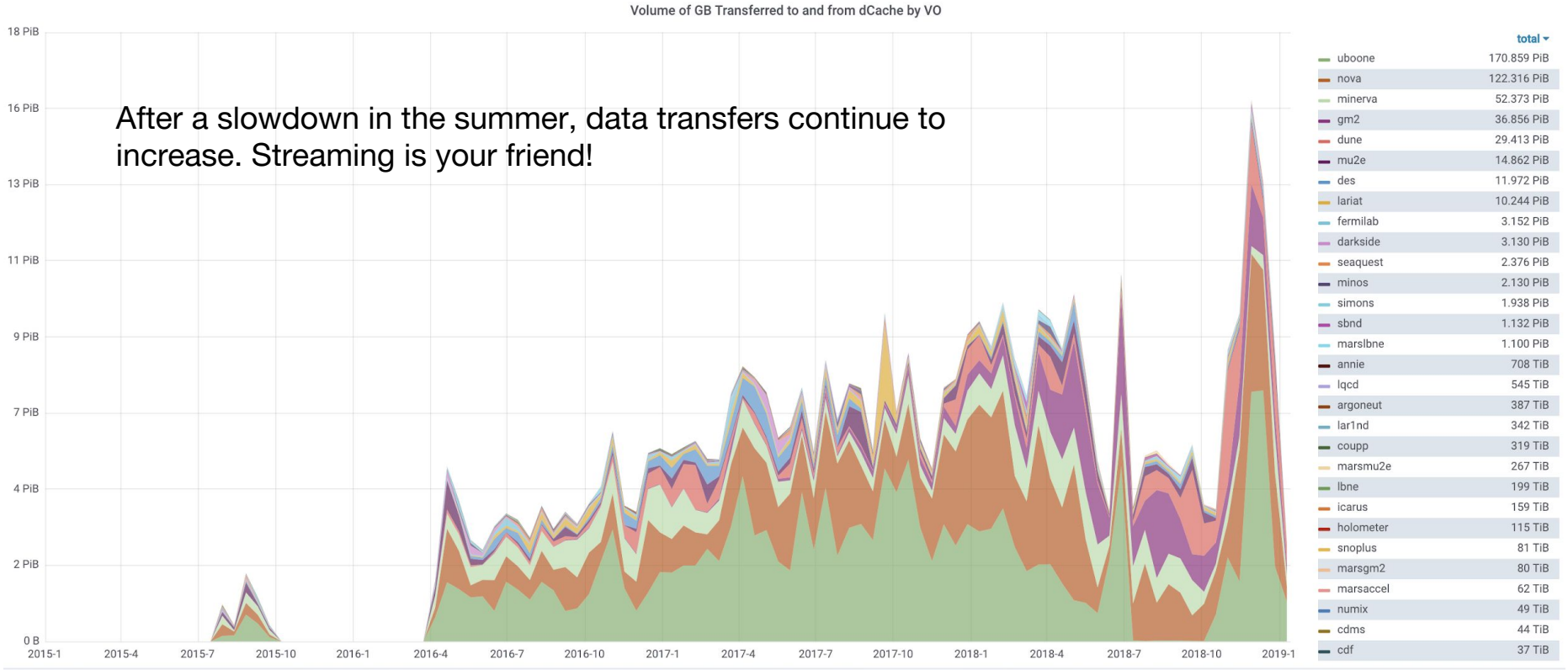
FIFE Trends - Core hours



	total
nova	131.489 Mil
mu2e	123.074 Mil
minerva	49.643 Mil
minos	46.793 Mil
dune	37.871 Mil
microboone	34.103 Mil
mars	26.469 Mil
gm2	19.509 Mil
cdf	16.574 Mil
seaquest	14.548 Mil
des	10.451 Mil
lariat	7.623 Mil
darkside	5.382 Mil
lar1nd	4.637 Mil
cdms	2.179 Mil
/dune/Role=MARS/Capability=NULL	1.665 Mil
sbnd	1.387 Mil
fermilab	1.222 Mil
icarus	642 K
annie	506 K

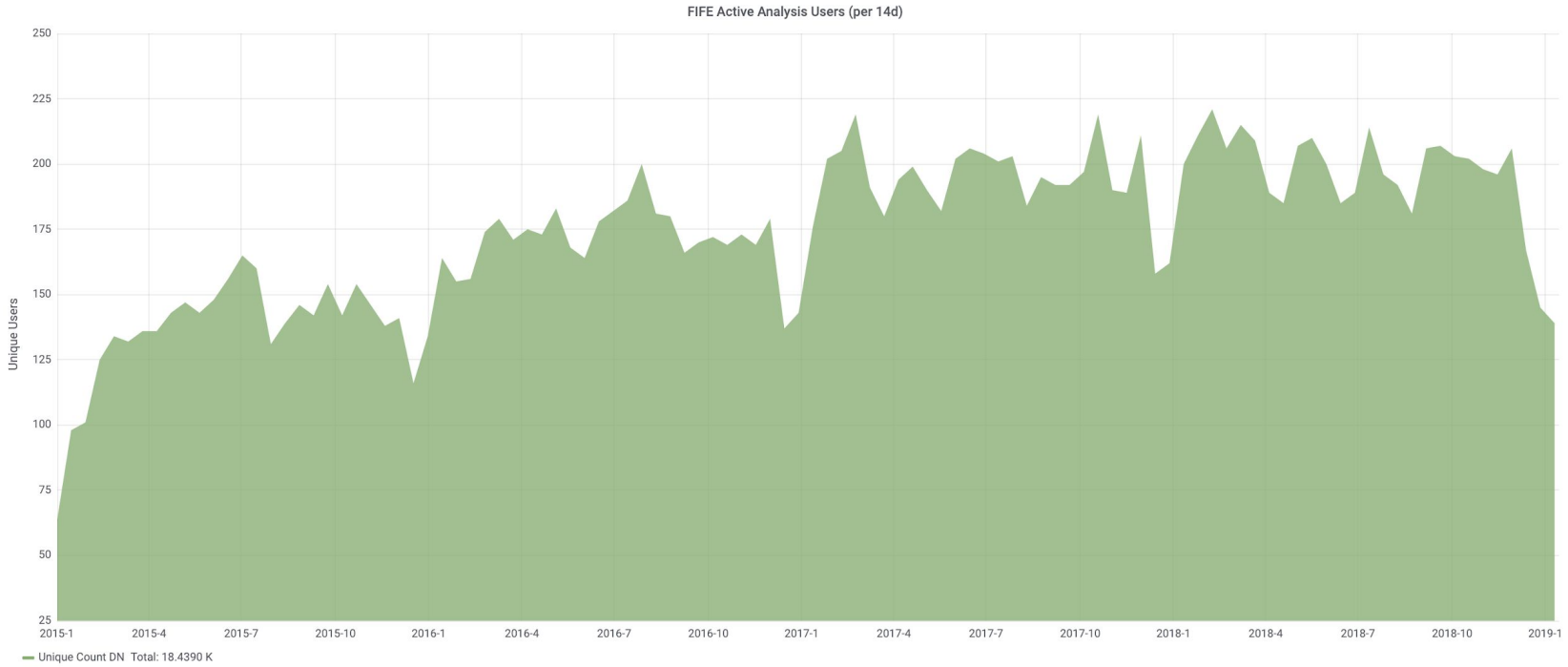
Total Core Hours continue to be “flat” across last 3 years

FIFE Trends - dCache Data Transfers



Unique active DNs

Growth has slowed a bit, but active users still increasing on average



Top-level strategic goals

- **HEPCloud** preparation- significant shift in the idea of resource availability and mapping to “cost” and shared monitoring resources

- **Container** service: Software and container distribution for user and Machine Learning software
- Generic access to GPUs and GPU/ML software matching to hardware
- Software packaging

- **Data management** improvements for both transfer efficiency and user interface
- Evaluation of dCache performance and understand scalability needs for usage

- Analyzer level **workflow management** tools improvement
- Anomaly detection and automatic blacklisting (resources, users,...) - Landscape expansion
- Improved monitoring of resource usage and feedback - Efficiency policy and suspension

Common solutions across the lab

Current Common Solutions:

- **HEPCloud** (access to wide range of resources, especially HPC)
- **Ferry** (fermilab computing authorization)
- **StashCache** (local access to remote data)
- **Landscape** (Fermilab computing monitoring)
- **GRACC** (OSG accounting)

Upcoming Common Solutions:

- **Rucio** (data management)
- **x509** certificate replacement

Agenda for the Workshop

- Policy updates and advice
- HEPCloud
- Container Build Service
- GPU update
- Jobsub and quota behavior
- User code distribution
- Storage
- Data Management
- Security
- FERRY
- Landscape
- POMS
- project.py replacement
- Striped data analysis
- SNOW tickets visibility
- Beyond the next 6 months

Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

HEPCloud

Tony Tiradani

FIFE Roadmap Workshop

January 17, 2019

FABRIC FOR FRONTIER EXPERIMENTS

HEPCloud Onboarding

- Go Live - early March. Should be completely transparent to users
- On-boarding to HPC and Commercial Cloud will be done in stages
  Experiments should submit a request. Priorities will be defined by SPPM
- Initially only production workflows in compliance with specific requirements will be accepted to run on HPC and Commercial Cloud. Requirements:
  - Need to do benchmarking and provides various metrics related to a job in order to be able to submit to HPC or Comercial Cloud
  - Limited data transfer in and out
- First available HPC resource is NERSC (shared FIFE alloc and/or CMS)
- Working on ways to integrate other HPC allocations
- What experiments? Those that have requested resources via SPPM and/or SCPMT (so far NOvA, DUNE, Mu2e, g-2)
- Responsibilities of production teams: close monitoring of job status and allocation utilization. Must be partners in developing resource requests

What resources will be available via HEPCloud?

- Usual FermiGrid + opportunistic OSG
- NERSC
  - Separate allocations for CMS, FIFE, SciDAQ
  - FIFE alloc is 55 M hours (primarily Cori; mix of Haswell and KNL nodes), shared across all experiments
  - A number of advantages to working within FIFE umbrella to get allocations
- Commercial resources via Google and Amazon (constrained by budget)
- Budget and NERSC allocations controlled via SC-PMT (coming up in February) and SPPM WG (for out of band requests)
- Experiment-specific, non-NERSC HPC allocations will not be accessible at Go Live time (will be a process for incorporating these allocations; may take 1-2 months to integrate such allocations)

What do you need to do to access new resources

- All resources: make appropriate requests in your SC-PMT presentations (including resource type, expected hours, etc.)
- NERSC + commercial cloud: job profiling/monitoring (see the HEPCloud onboarding doc) requirements on data transfer out, failure rate, job length (preemption, etc.), memory. HEPCloud team will provide ways to calculate these things to see f you're compatible
  - Commercial resource requests MUST be presented to SPPM
- HEPCloud reserves the right to shut off access to HPC+commercial resources for a variety of reasons

Limitations and things to know

- Don’t wait too long to use HPC allocations, and be sure you use it all (or may not get as much next year).
- No “opportunistic” HPC use (i.e. the NERSC FIFE allocation is finite). If you think you’re going to exceed your request, need to go back to SPPM to request more (not guaranteed you’ll get it).
- Longer queue times at NERSC wrt FermiGrid
- Possible to make a reservation (as NOvA did) for massive parallel usage of NERSC cluster. Coordinate with SPPM.
- Getting help: Pretty much the same as now (start w/a ticket; will get reassigned as needed to solve the problem)

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

Containers

Glenn Cooper, Tony Tiradani

FIFE Roadmap Workshop

January 17, 2019

FIFE
FABRIC FOR FRONTIER EXPERIMENTS

Containers: Work for Common Solutions

- Proposing to organize a working group that will understand experiments' requirements for running "proprietary" containers on HEPCloud. We should address:
  - Initial software build cluster
  - Integration with CI
  - Repository (local repo available)
  - Registry
  - Policy (lightweight process for users, but that causes delays for security vetting, etc.)
  - GPU (libraries, drivers, etc.)
  - Specific OS
- Limit to 1-2 people per interested experiment, plus SCD personnel TBD
- Contact Ken Herner or Marco Mambelli if you are interested in getting involved

Container Build/Test/Register

We have a prototype infrastructure, ready for a few testers*, to do the following:

- Build a Docker container
- Test the container (CI)
- Store config info (docker file) in a (git) repository
- Register the container (local Docker registry available, or Docker Hub, etc.)

These are then typically converted to Singularity or Shifter containers to run on…

* *Contact Glenn Cooper and Ed Simmonds if you have someone available to start testing*

Type of Container Used

Fermigrid:	Singularity
OSG:	Singularity
NERSC (HPC):	Shifter
Cloud (AWS, Google):	Singularity

NOTE: HEPCloud already has a Shifter image at NERSC that should be suitable for most experiment workflows. This will be available to onboarded experiments after HEPCloud has transitioned to production.

GPU Resources and Containers

- As previously mentioned, GPU resources available on the OSG via jobsub
- Feedback in June was that it was too hard to get to non-Wilson Cluster resources going (too many issues building containers)
- Can be part of the working group effort
  - Step 1 is to collect requirements

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

JobSub and Quotas

Dennis Box, Joe Boyd

FIFE Roadmap Workshop

January 17, 2019

FABRIC FOR FRONTIER EXPERIMENTS

Jobsub Updates

- Delivered:
  - Tarball distribution via dCache/CVMFS (coming later today). Please use the feature! As documented HERE.
  - Job submission via config/text file (example here)
  - Pulling information from FERRY
  - SL7 containers now possible on FermiGrid (if SL7 is in your OS list)
- Planned:
  - Single sign-on integration: can pull logs without needing certs in your browser
  - Next release: v1_2_9. Expect a request to test your workflows in the next two weeks.

A word about some quota behavior

- Over the past few months we've had a couple of cases where quotas weren't working as expected
- December: experiments without hierarchical quotas set up were not getting their full quota despite other experiments being over (and getting new job starts)
  - Traced to the way HTCondor was sorting the subgroups before doling out slots. Expression changed; **should not be an issue anymore**
- January: Some subgroups (i.e. keepup) were not able to expand out past their fraction of the total relative to other jobs submitted without a special subgroup
  - Basically, it's hard now to give subset of users the ability to go before others short of changing their subgroup to take 100% of the full quota
  - Actively looking at some ways to support this kind of case; need to be tested more
  - Ability to adjust quotas more quickly via FERRY now somewhat mitigates this

Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

CVMFS update

Dave Dykstra

FIFE Roadmap Workshop

January 17, 2019

FABRIC FOR FRONTIER EXPERIMENTS

User Code distribution

- Reminder of the problem
  - Removing NAS mounts on worker nodes (necessary step for HEPCloud) required users to stage custom code to worker nodes via tar files
  - Most people started doing this via the scratch dCache pools, but dCache becomes overwhelmed when too many jobs try to access such tarballs at once (only one copy)
- Moving tarballs to resilient pools provided temporary help
  - Remember: files in resilient are replicated 20x and the space comes out of the shared r/w dCache pool.
- Very few experiments are using the jobsub feature that automatically manages tarballs in the resilient pools. No cleanup if you don't use the feature, and pointlessly overwrite an unchanged file sometimes.
- **Clear need for a better long-term solution**

FIFE rapid user code distribution in CVMFS

- Design has been approved for first version of distributing user code via CVMFS
  - Tarballs submitted through jobsub, as now
  - Jobsub client sends the tarballs directly to one of two dedicated CVMFS publishing machines, each with two repositories
    - Authentication accepted from all known Jobsub users
  - Very short (few minutes) update time configured on the repositories
  - Special fast update track on stratum 1s
  - Job wrapper finds the published tarball, waits if necessary
  - Monitor efficiency; could make jobsub server wait to submit jobs
- A prototype demonstrated that publishing could be done quickly enough after switching to gzip tarballs instead of bzip2
- Development has begun but proceeding at relatively low priority, expected availability around mid-year

Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

Storage

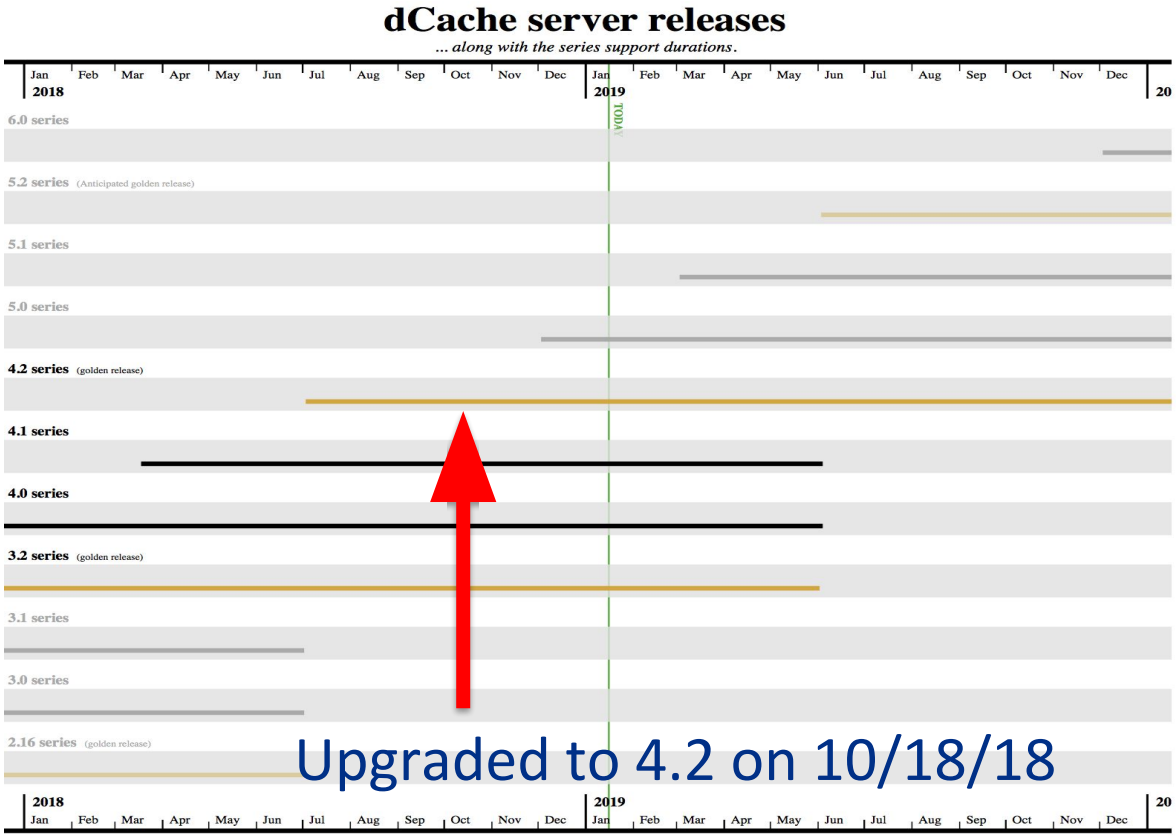
Dmitry Litvintsev

FIFE Roadmap Workshop

January 17, 2019

FIFE
FABRIC FOR FRONTIER EXPERIMENTS

dCache, releases

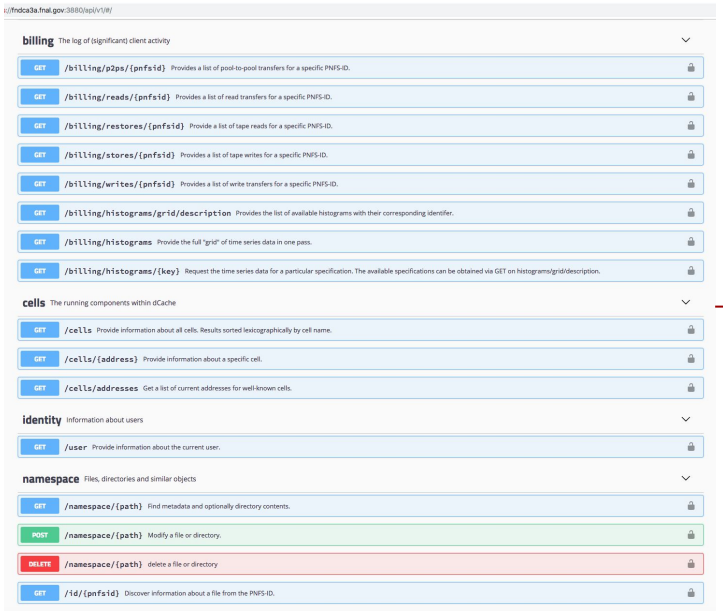


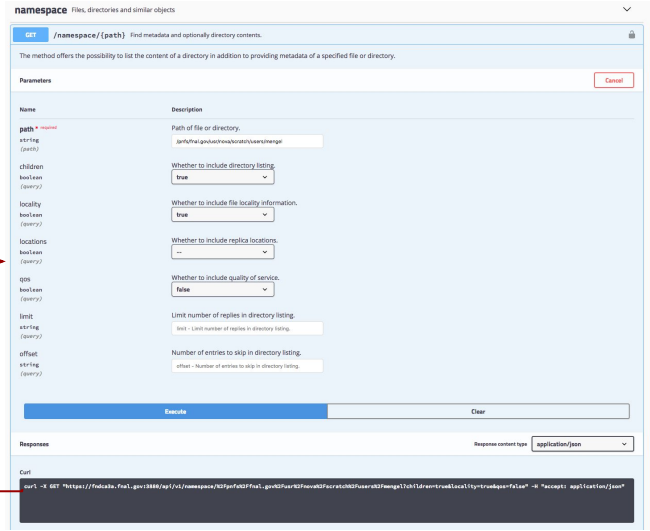
Fermilab

dCache 4.2: What's New

- Many improvements in dCache NFS server targeted at better handling misbehaving clients, overloaded pools and network glitches. **We believe NFS has been in better shape since 4.2 upgrade.**
- Experimental support for mutable files. It is possible to modify files in dCache (with some restrictions, not configured yet).
- 3rd party xrootd transfers.
- Improved 3rd party HTTP(s) transfers.
- Improved RESTFul monitoring interface (and web GUI):
  - allows easy integration with external monitoring tools/frameworks (like Grafana and thus FIFEMON).
- RESTful QoS and "data portal" API:
  - ability to pre-stage and query file status using RESTFul API
  - directory functions
  - upload/download and file sharing capability.

dCache 4.2: Restful API



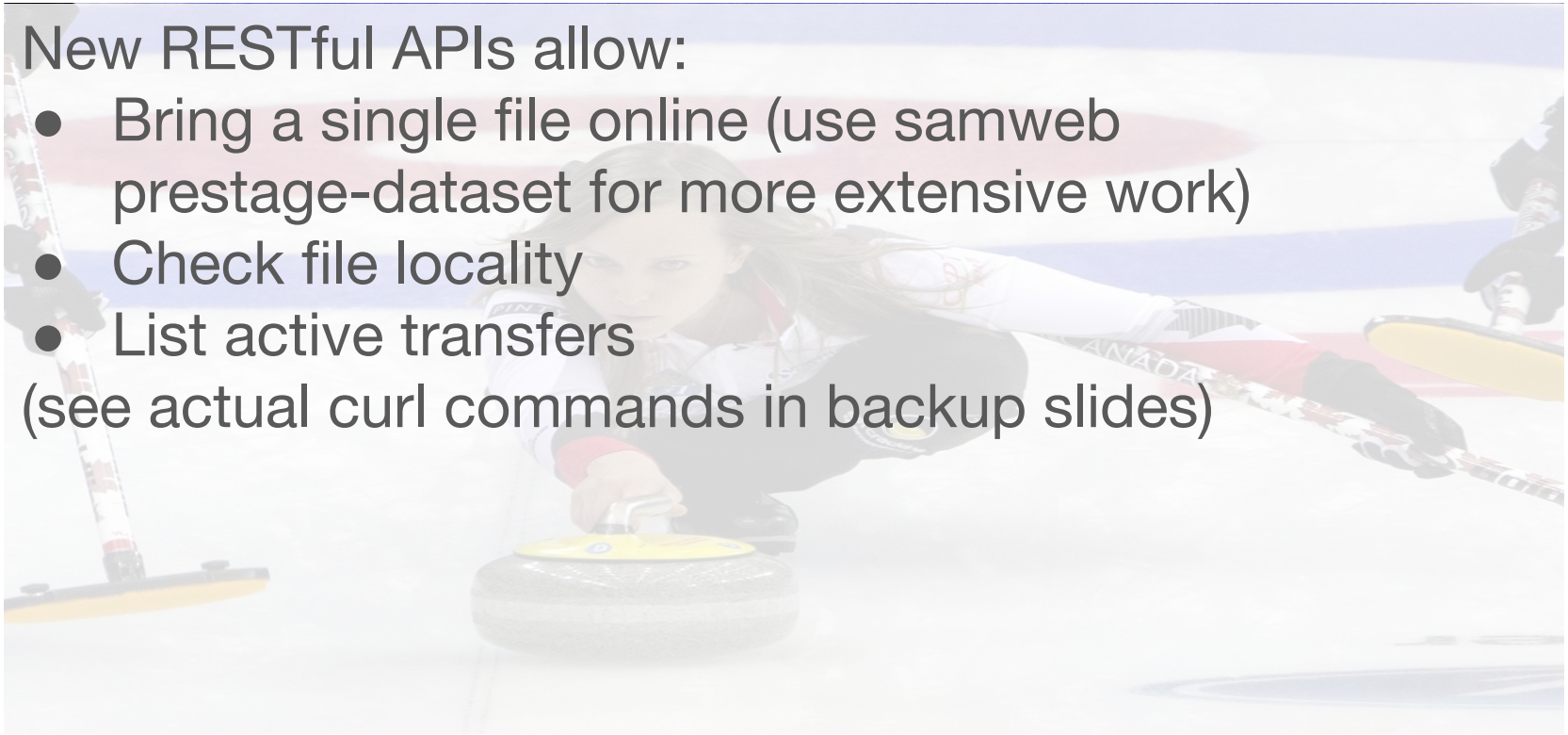


Fill out the form and it creates a curl command (see next slide)

Go to [this url], select action (some requires a certificate)


```
{
        "fileMimeType": "application/octet-stream",
        "fileLocality": "ONLINE",
        "pnfsId": "0000D491F98AA6EF4E0AA5CE058A23808620",
        "fileType": "REGULAR",
        "nlink": 1,
        "mtime": 1544049924791,
        "creationTime": 1544049924791,
        "size": 0
}
```

dCache status: Going curling



New RESTful APIs allow:

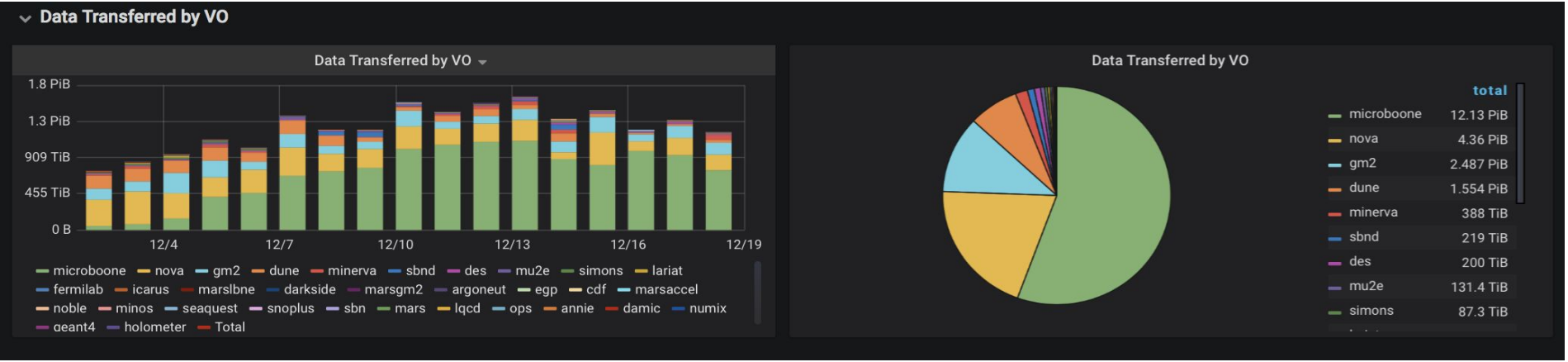
- Bring a single file online (use samweb prestage-dataset for more extensive work)
- Check file locality
- List active transfers

(see actual curl commands in backup slides)

dCache: other news

- Actively participating in CERN DOMA activity (Data Organization, Management and Access): realization that infinite disk expansion to meet LHC data requirements is not feasible focus is on efficient handling of tape and caching:
  - better integration on Storage and Experiment workflow/orchestration software (like Rucio).
  - propose use common language of QoS to define data lifecycles and transitions.
  - use publish/subscribe mechanism to handle storage events
- Post-globus world:
  - 3rd party copy with xrootd
  - 3rd party copy with HTTP(s)
  - federated identity
  - Access tokens

dCache: performance



Sustained well over 1PiB/day for > 2 weeks w/o issues.

Tape Migration and Retention

- SCD is currently phasing out its Oracle tape infrastructure
  - Currently writing all new data to IBM/LTO tapes since 4Q18
- Will begin migration of data on Oracle tapes to IBM tapes this year
- Aim to reduce existing footprint of Oracle libraries (5)
  - Down to **one** by November 2020
  - Down to **zero** end of 2021
- Asking experiments to identify data currently on T2 tapes (133 PB) as either
  - Only needed up to **November 2020**
  - Only needed up to **December 2021**
  - Retained **beyond 2021**
- A small amount of data (3.7PB) on even older LTO4 tapes to be migrated ASAP
  - Contacting experiments now about what can be deleted there

Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

NAS

Andy Romero

FIFE Roadmap Workshop

January 17, 2019

FABRIC FOR FRONTIER EXPERIMENTS

NAS

- POSIX-compliant areas known to be important; support for them will continue (technology underneath might change)
- Available NAS space very unlikely to expand (cost is over $1000/TB)
- Consider this information when preparing your SC-PMT requests for February
- Best thing you can do: clean up unneeded files/directories; gives a better idea of what the true resource needs are going forward
- Check the usage here: https://landscape.fnal.gov/monitor/d/0-5dQkQmk/nas-bluearc-usage?orgId=1
- Also on the wish list: move away from /grid/fermiapp volumes
  - Mostly used for building and/or staging to CVMFS. This could be done with /experiment/app areas with things otherwise staying the same

Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

Data Management

Robert Illingworth, Marc Mengel

FIFE Roadmap Workshop

January 17, 2019

FIFE
FABRIC FOR FRONTIER EXPERIMENTS

Data Management

- SAM core
  - Not many recent changes
  - Upcoming
    - Improved handling of preempted projects (see later)
    - Validation of checksum values
    - Fetch individual metadata values rather than everything for a file
- SAM client
  - V3_0 will become the default later this month
    - Python 3 compatible

New SAM station

- New features
  - Better handling of distributed storage elements
  - **Preemption**
    - If a SAM process detects it has been preempted then the whole of the old process will be marked as “cancelled” and the rerun job will be given the same files again.
    - Requires that jobs tell SAM their batch job id. There’s a new API for this, but it’s sufficient to pass it in the description field which some jobs already do (this is the POMS/fife_wrap default behavior)
  - Potentially greater scalability
    - Idle projects should use less resources, so we can afford longer timeouts
- New station installed on the dune-int SAM station now and is being tested with regular jobs

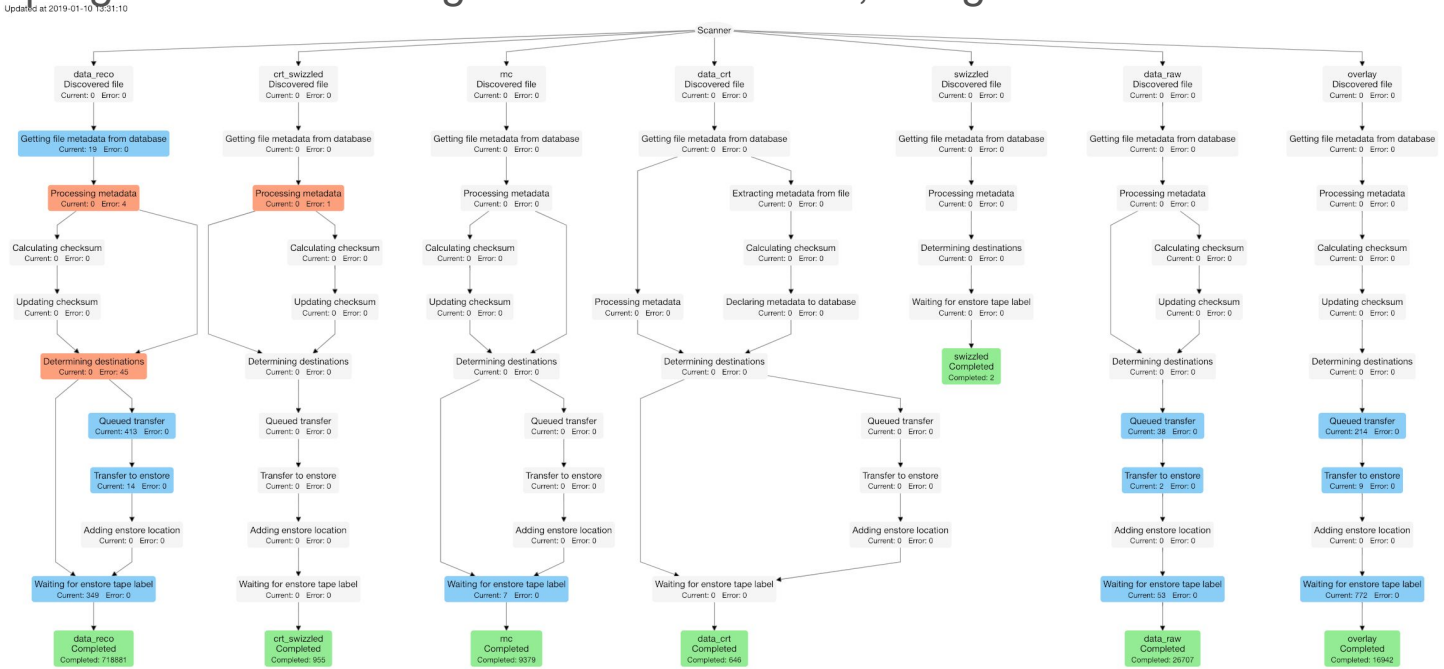
F-FTS

Version 6.2 recently released:

- Improvements to file transfer handling
  - More predictable limiting of transfers (limits set by destination)
  - Prioritization by file type (put priority = N in config file)
  - Improved status page info
    - Can now page, filter, and search the tables
  - Experimental new monitoring

New F-FTS Monitoring

- For example http://uboonesamgpvm01.fnal.gov:8787/fts/stategraph
- Shows progress of files through the different states, along with where the errors occurred



Fermilab

Rucio and the future of data management

- Future of HEP data management looks to be the Rucio project
  - Originally from ATLAS, now being adopted by CMS
  - Also, XENON-1T, LIGO, IceCube, SKA and others
- DUNE has a Rucio deployment with ProtoDUNE data
  - Not used as the primary data transfer from CERN to FNAL
  - But is used to copy data to various other European sites
  - Also used to manage space on CERN disk and cleanup unwanted data
- Running alongside existing SAM system
  - SAM is still the primary catalogue for most files

Plans for Rucio

- We expect to migrate towards Rucio for upcoming and some current experiments
  - “upcoming” and “current” are fuzzy terms here; aside from DUNE no experiment is committed to anything yet. However, long-term SAM support is not guaranteed.
- Plan is to create a hybrid SAM/Rucio system where we phase out the SAM location catalogue and use Rucio instead
  - For most users nothing will change; they’ll still interact with SAM much as now. Only those who want to manipulate SAM locations will see any difference.
  - Will then benefit from Rucio’s powerful data placement tools - it’s simple to transfer datasets around

DM Longer term plans

- Replace SAM metadata catalogue with something more Rucio integrated
- This will be much more disruptive to workflows and users; an “opt-in” choice for experiments
- Detailed design still needs to be done
  - Want to improve scalability of the metadata catalog
  - While benefiting from the experience with SAM

FIFE Roadmap Workshop

Security

Mine Altunay

January 2019

InCommon CA

- Take-home message: SCD will spend some time working with (obtaining, renewing and managing certs) new certs. Most of the work will be done by SCD for managed certs.
- Do you have any host certs that you are maintaining outside of the SCD provided infrastructure?
  - If so, we need to work with you.
- Background:
  - OSG CA was shutdown in April 2018.
  - Before the shutdown, we had a big campaign to renew all certs for another year. They will expire in April 2019
  - We acquired a new CA, InCommon CA, and we will start a new campaign for the new CA in February to provide you with new certs

ECL Local Account Policy

- Local accounts for people without Fermilab services accounts are not allowed.
- All parties using Fermilab resources must have Fermilab services accounts.
- This will ensure accounts are properly managed and terminated when they are no longer needed.
- If there are local accounts given to users without Fermilab services accounts:
  - These accounts must be terminated
  - Users must first get Fermilab services accounts
  - Then local accounts for the user can be set
  - We expect this process to finish by February 4th

Federated Identity News

- End goal: can we have seamless, easy access to compute resources (e.g. storage) without certificates?
- WLCG is very actively working on this
- A new WLCG working group, WLCG AuthZ WG, has been built and working very hard
  - Created several technical documents; conducted a survey of LHCB, Atlas, Alice, and CMS; created a summary document of different technologies
- Pilot projects are ongoing and a formal report to WLCG Mgmt Board will be given on Feb 19th
- Still in the early stages, but a lot of progress has been made

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

FERRY

Tanya Levshina

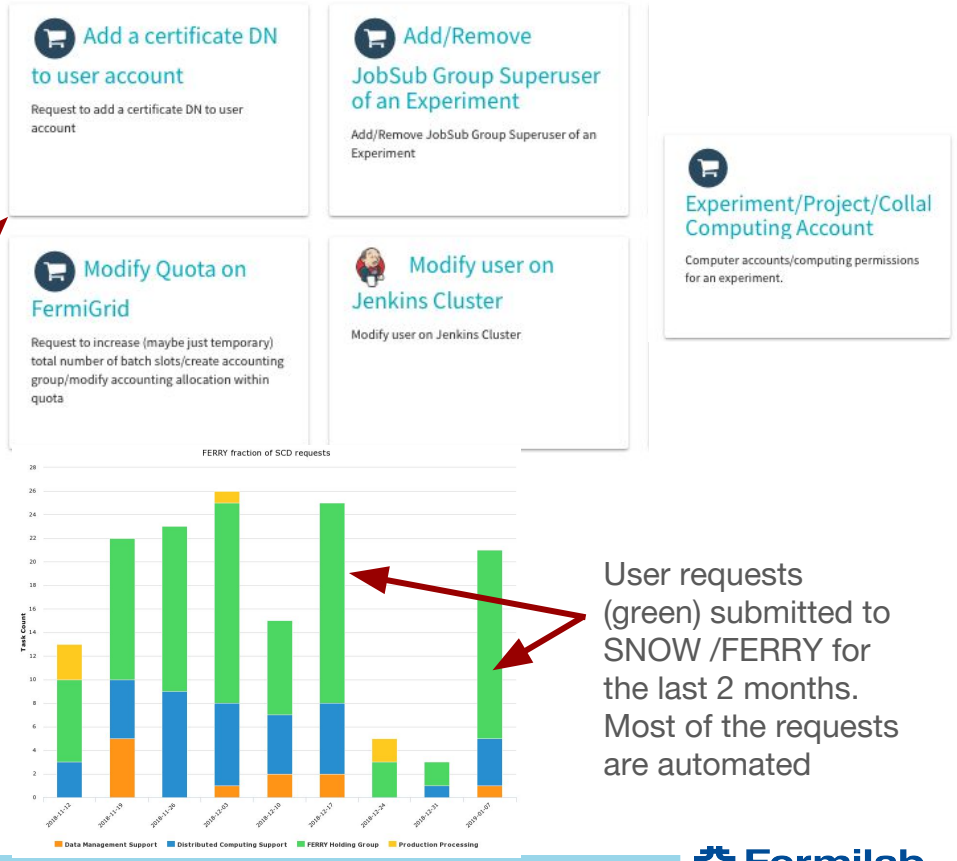
FIFE Roadmap Workshop

January 17, 2019

FIFE
FABRIC FOR FRONTIER EXPERIMENTS

FERRY/New SNOW Forms

- FERRY provides a central attribute repository which enables each reliant service to query for specific attribute values.
- In production since October, 2018
- Consolidate access to forms in SNOW related to distributed computing.Workflow includes:
  - Approval
  - FERRY Orchestration



User requests (green) submitted to SNOW /FERRY for the last 2 months. Most of the requests are automated

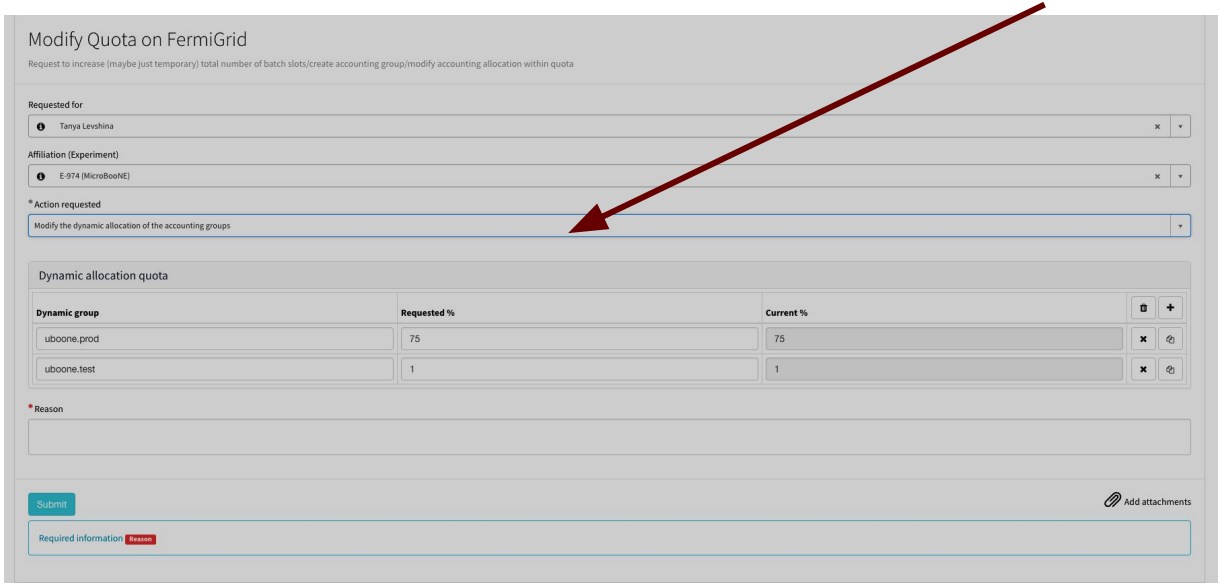
FERRY/New SNOW Forms

FERRY

Benefit to production team: working towards automatic, no human delays or typos after approval:

- VOMS update (registration, new Role, new DN)
- Login, home area creation on interactive cluster
- Superuser form
- Jenkins form
- Condor Quota form

Different set of approval based on action



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

Landscape

Kevin Retzke

FIFE Roadmap Workshop

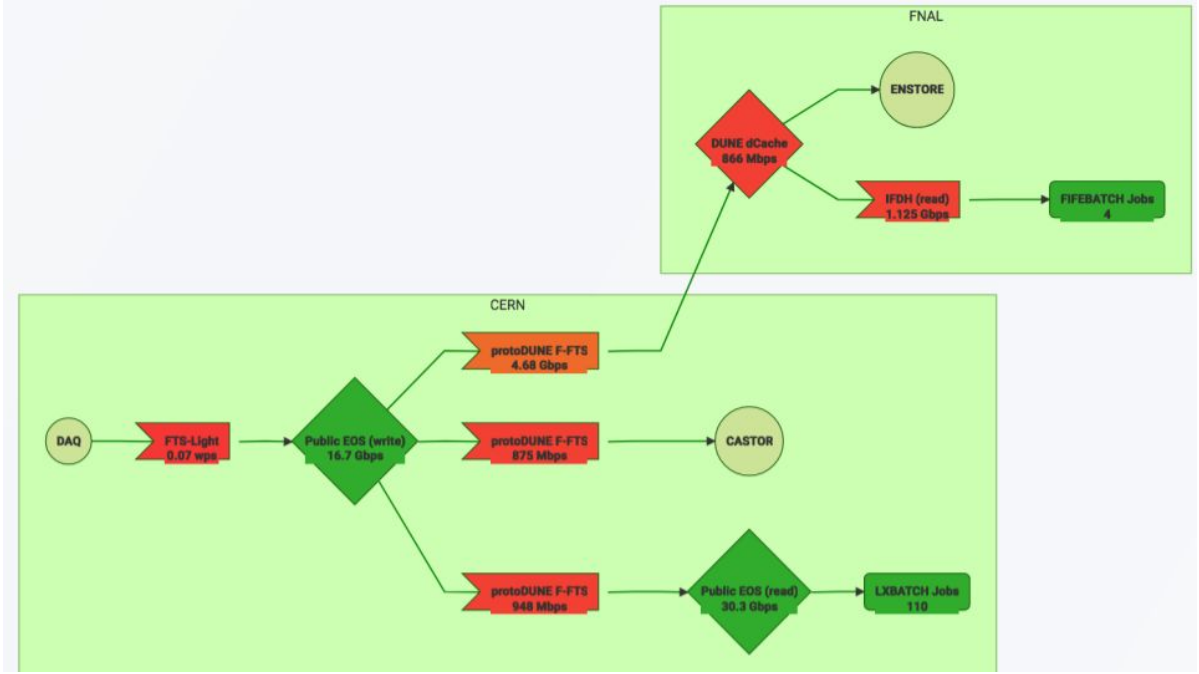
January 17, 2019

FIFE
FABRIC FOR FRONTIER EXPERIMENTS

Recent changes and additions

Landscape

Cross lab monitoring: big success for protoDUNE!



https://fifemon.fnal.gov/monitor/d/000000299/protodune?orgId=1

Recent additions: Production Shifter Dashboards

- Intended to be a central source of detailed monitoring information.
- Allows navigation from top level POMS campaign to a failed job log.
- **Customizable per experiment!** Simply open a SNOW ticket explaining what you would like to see.

https://fifemon.fnal.gov/monitor/d/HMj5hqVik/production-shifter?orgId=1

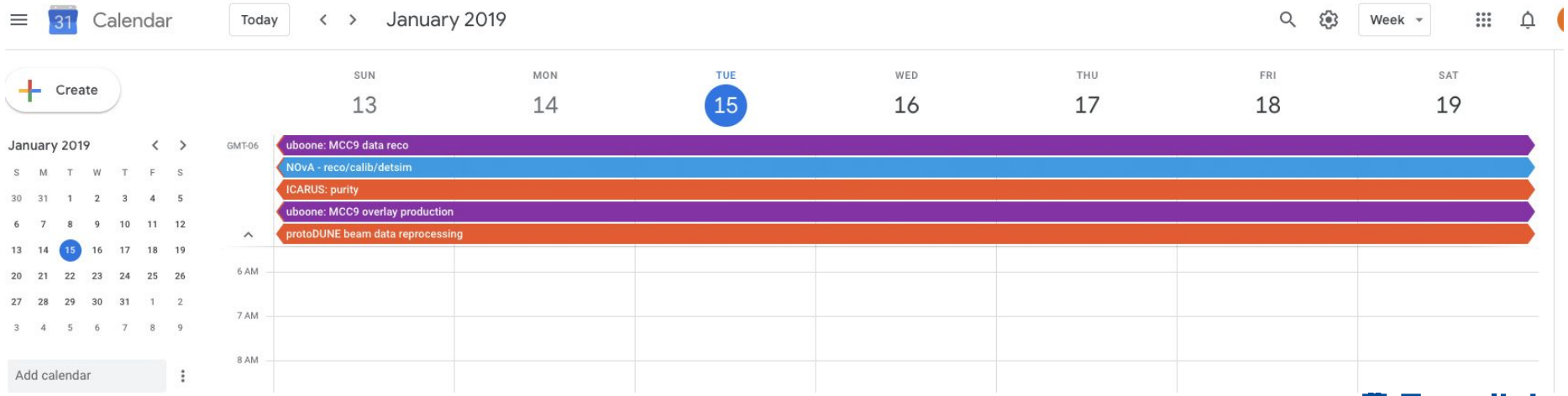
Production Campaigns Calendar

Landscape

Central place that experiments can write down their large-scale production processing campaigns (both data and MC)

Useful for managements to understand potential resource conflicts (both CPU and storage access). Viewable at
https://fifemon.fnal.gov/monitor/d/HQ2nGUamz/production-campaigns-calendar

Please start using if you aren't already. Open a ticket to USDC to request permission to edit



Future Landscape Work

Landscape

- Adding dCache detailed information, pool status
- NAS data per user dashboard
- Incorporate SAM project related info per job
- HEPCloud-related dashboards
- Better insights into data transfer problems and bottlenecks
- Improve automated “black hole” node detection

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

Continuous Integration Service

Vito Di Benedetto

FIFE Roadmap Workshop

January 17, 2019

FIFE
FABRIC FOR FRONTIER EXPERIMENTS

Continuous Integration News

- CI operations have been ramped down
  - Experiment code monitoring has been completely delegated to experiments.
    - LArSoft based experiments are monitoring their own CI instance.
  - The CI “user blame” feature has been improved to help experiment CI operation by sending a CI report to the user who broke the code.
- CI Validation improvements
  - Supports comparison of validation results vs reference results
    - Overlay normalized plots for *current* and *reference* data
    - Use of a chi square to evaluate the comparison agreement
  - Provides support for authentication to access the CI validation dashboard to prevent to disclose real data outside the collaboration
  - Provides Kibana links to access logs of failed jobs and to provide job HoldReason

Continuous Integration Roadmap

- Improve and extend existing functionality. CI users should keep providing feedback and feature requests

- Provide support for code memory/runtime profiling using a dedicated tool, we are considering “IgProf”. This will help experiments to keep their code resource requirements under control

Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

POMS

Steve White

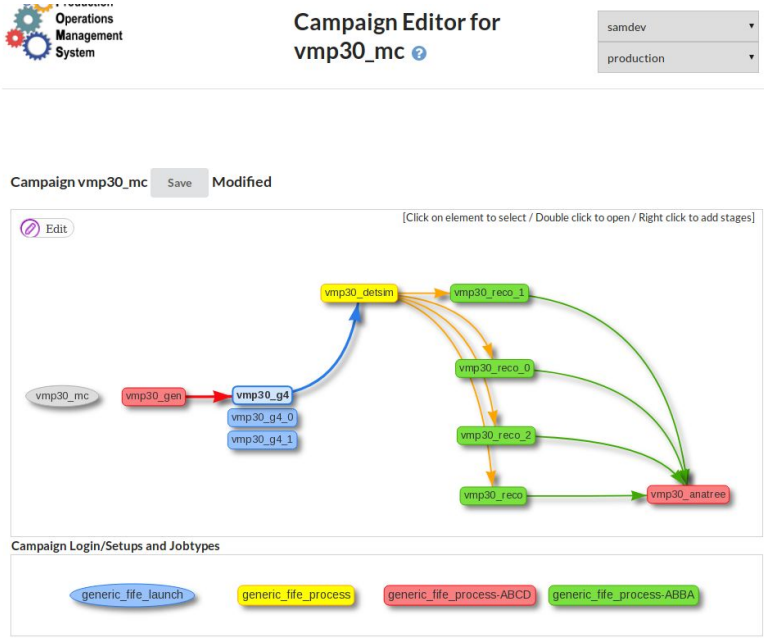
FIFE Roadmap Workshop

January 17, 2019

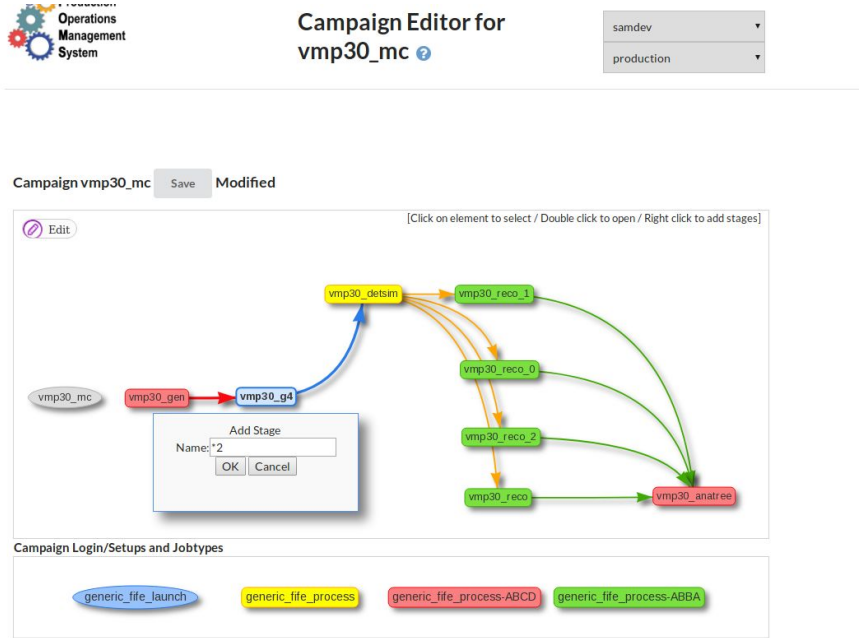
FABRIC FOR FRONTIER EXPERIMENTS

POMS v4.0 release

GUI campaign editor rework:



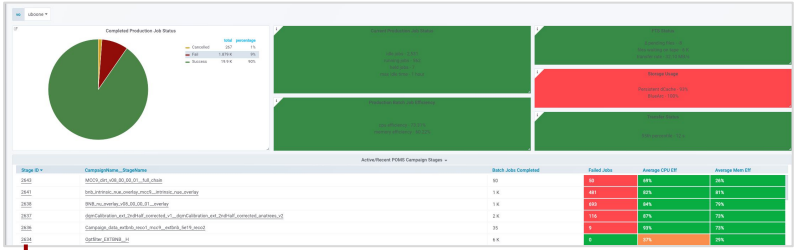
Now a user can create a custom layout. The color of the stage elements corresponds to the POMS job type used in the stage.



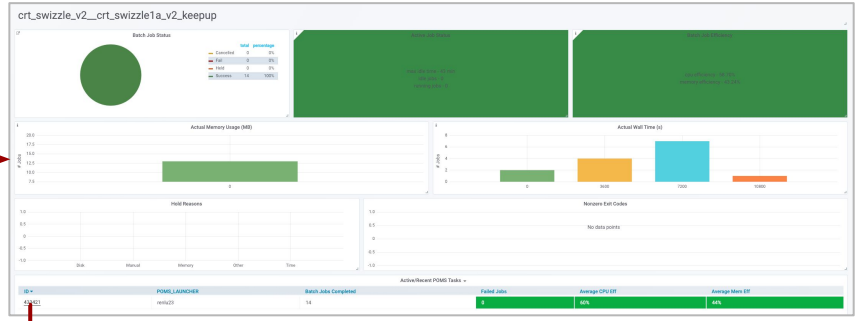
A user can clone an existing stage using right click and entering the number of desired clones in the form '*N', i.e. multiply by the number. Then the user can connect newly created stages the way one likes.

Fermilab

POMS v4.0 release

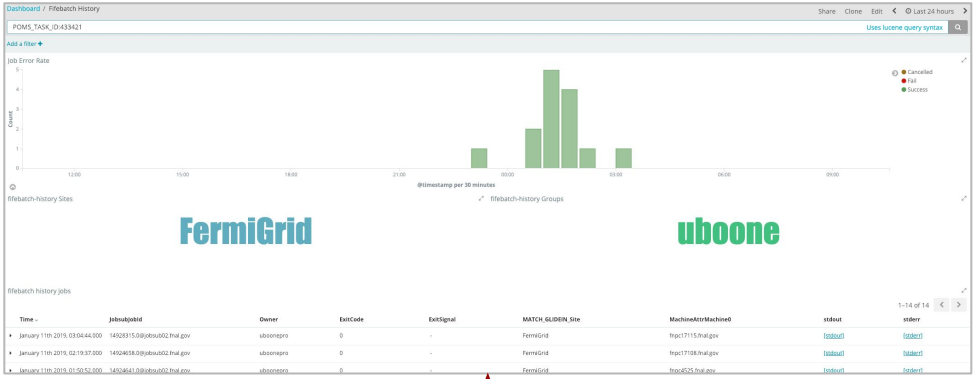


Shifter dashboard with POMS campaign names



Shifter dashboard for POMS task

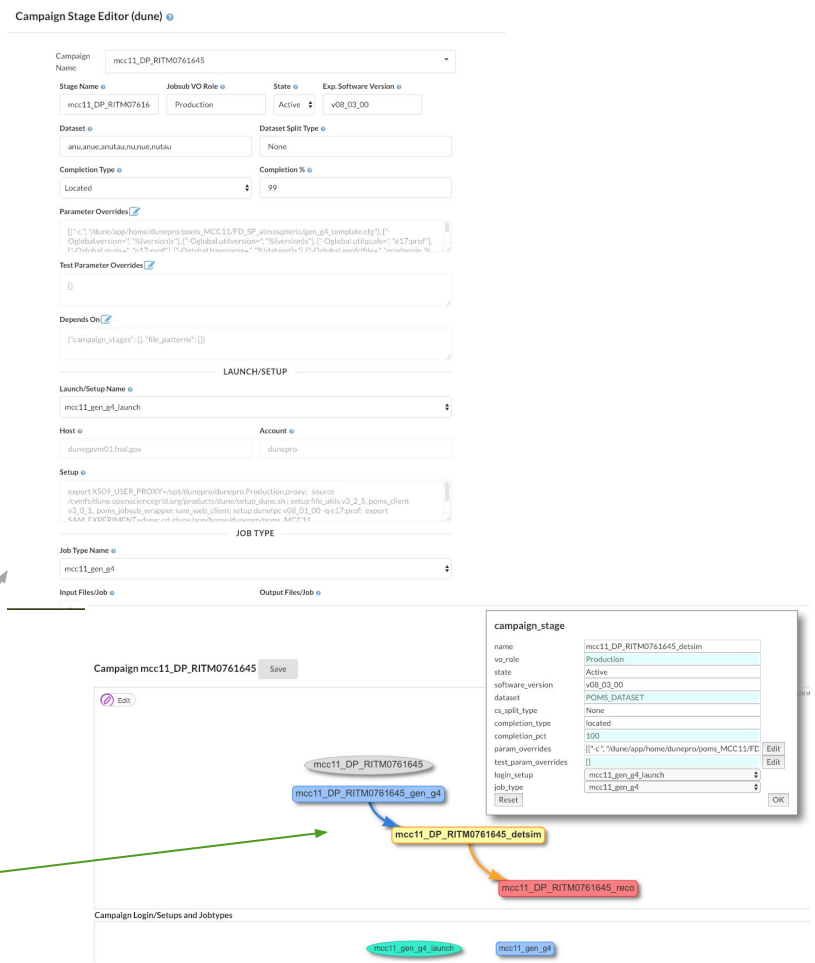
- Fifemon/Landscape integration
- Tracking submissions, not jobs
  - much lower database usage
  - Landscape has the details
- Much faster web page response now



Fermilab

POMS

- 4.1 release (February)
  - GUI campaign editor improvements
  - Job launch options
    - get launch commands for debugging
    - test launches w/ special options
    - hand-launch recovery jobs
  - Preliminary features for Analysis use
    - upload proxy for launch
    - upload (small) config files for launch
- 4.2 release (Summer)
  - Data handling stages
  - Full Analysis user support
  - Propose to remove old non-GUI editors in favor of editing via GUI



Fermilab

Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

Project-py

Yuyi Guo

FIFE Roadmap Workshop

January 17, 2019

FABRIC FOR FRONTIER EXPERIMENTS

New tools: Project-py

Project-py is a command line interface tool that makes use of the POMS system and API to schedule and submit jobs to the batch systems:

- a replacement for the current “project.py”; provides the same essential functions and user interfaces as “project.py” to configure, submit, report and book keep analysis jobs
- has a slimmer code stack, most of the functionality and interaction with the infrastructure is off loaded to the POMS and other FIFE tools (e.g. SAM, Jobsub, Fifemon, etc…)

Status:

- finished requirement documentation
- manually converted a 6-stage project.py campaign to a POMS campaign and ran successfully
- finished XML to ini Converter

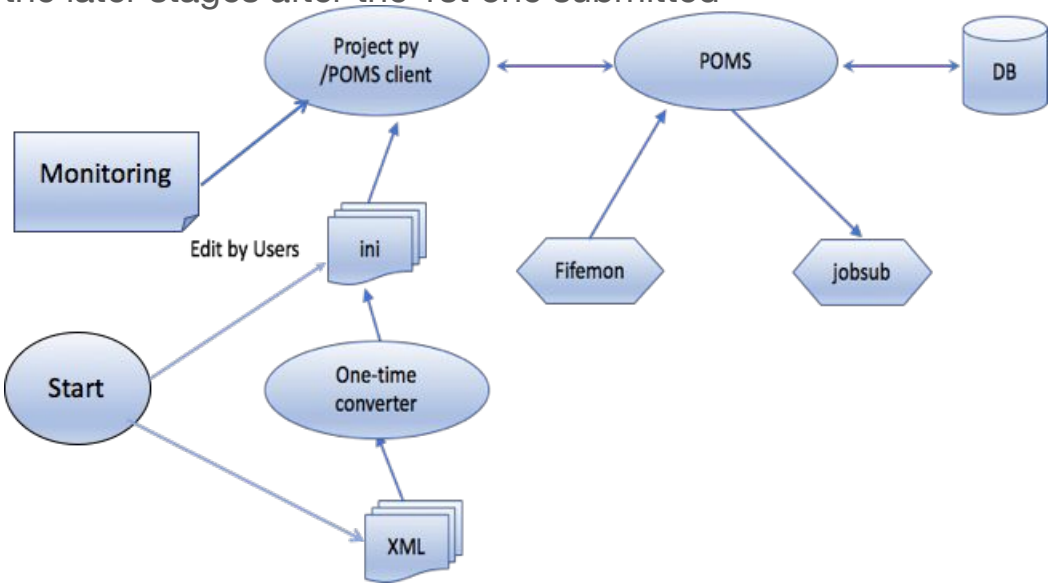
Contact: yuyi@fnal.gov

Call for testers

New tools: Project-py

Workflow:

- create a POMS ini file or convert a “project.py” XML config file into a POMS ini config file or do modification to provided template ini files
  - Support multiple stages in one ini file
  - POMS will take care the later stages after the 1st one submitted
- upload ini file to POMS
- submit job
- check job status



Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

SNOW Tickets Visibility

Mike Kirby/Tammy Whited

FIFE Roadmap Workshop

January 17, 2019

FABRIC FOR FRONTIER EXPERIMENTS

Use Cases

- Experiments want to be able to forward/post on slack a link to a ticket opened by somebody else but affecting their experiment as well. The existing solution -- individually adding users -- both creates a bunch of spam for those who aren't that interested and requires active effort on the part of those who can already see the ticket to extend access.
- Experiments want to include experiment support group to be added to a ticket so they could help/advice a user and monitor service provider response.
  - This is partially implemented. An experiment support team is added automatically when service provider replies to the ticket. The support group still wants be included at ticket submission time.

Update from SNOW team

- Visibility of SNOW tickets: In many cases could be handled via a knowledge article or using the Outage records. There are also other options available depending on the use cases. SNOW team needs more use cases.
- Watch lists in general:
  - There is a current issue when in the portal where the watch list is not available until submitting the ticket.
  - We do have the vendor providing an estimate to work on resolving issues with watchlists and having them do this work.
- People who have the ITIL role to support their experiment(s) run into some confusion in the portal versus the classic view. Tammy can organize a tutorial if people would like.

Fermilab U.S. DEPARTMENT OF ENERGY Office of Science

Striped Data Analysis Platform

Igor Mandrichenko

FIFE Roadmap Workshop

January 17, 2019

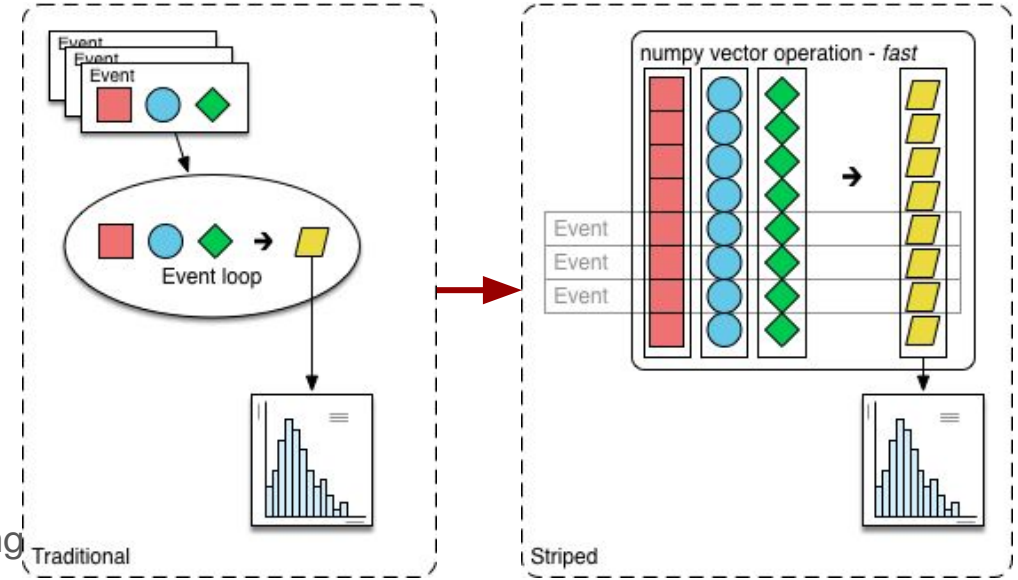
FABRIC FOR FRONTIER EXPERIMENTS

BigData LDRD Project

- LDRD FNAL-LDRD-2016-032 (2016-2018)
- Objective:
  - Introduce BigData philosophy and technologies into HEP analysis to speed up data analysis process and reduce time-to-insight
  - In particular, use NoSQL DB technologies to store event data and make them available for analysis
- What does BigData mean for HEP analysis:
  - Provide fast, efficient and scalable access to any event data at any time for analysis

Fundamental Ideas

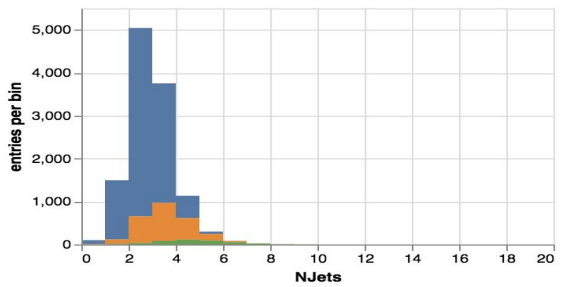
- Event storage:
  - Files --> Database
  - Use files only as backup, source of information
  - Eliminate file-based data massaging
    - skimming/slimming = data duplication
- Analysis model
  - Event loop --> Vectors
  - CPU -- > GPU
- Take advantage of the analysis process nature
  - Read the same data over and over, tweaking the analysis code
  - Computing/Data co-location: clustering --> caching
  - Elasticity of computing facility architecture

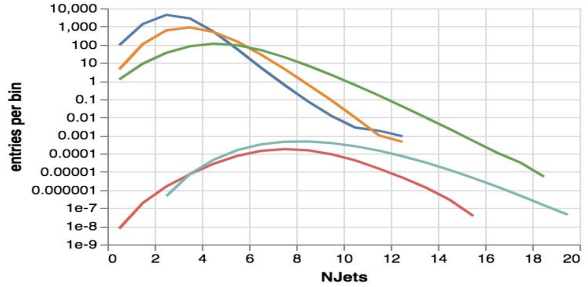


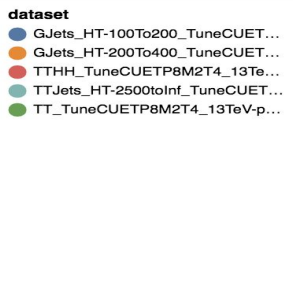
User Frontend: Jupiter/Python

In [9]:
```
display.init()
callback = Callback(disp)

for dataset_name in datasets:
    job = session.createJob(dataset_name,
        user_callback=callback,
        user_params = {"dataset":dataset_name},
        histograms=[njets_by_dataset, njets_combined, btags_by_dataset, btags_combined])
    job.run()
    runtime = job.TFinish - job.TStart
    nevents = job.EventsProcessed
    display.update()
    print "%-70s %7.3f M events, %7.3f M events/sec" % (dataset_name, float(nevents)/1000000, nevents/runtime/1000000
```








```
Summer16.TTHH_TuneCUETP8M2T4_13TeV-madgraph-pythia8                0.100 M events,   0.310 M events/sec
Summer16.TTJets_HT-2500toInf_TuneCUETP8M1_13TeV-madgraphMLM-pythia8   1.520 M events,   2.103 M events/sec
```

Results

- Hardware:
  - CPU Cluster of ~180 cores, 13 servers (old farm nodes), 64GB RAM each
  - 4 data web servers, SSD cache: 2x300 GB
  - 25 Database servers
- Performance (cached at workers):
  - Simple job (1 column): 30 M events/second
  - Realistic job (~20 columns): 9 M events/second
- With empty worker caches: 10-20 times slower

Under evaluation by:

- CMS
  - Multiple datasets from 5 to ~100 TB
  - Coffea analysis tool is being built on top of Striped framework
- MicroBooNe
  - Imported data, mix of reconstructed and raw hits
- NOvA
  - CAF data imported
- DES
  - Data processing (a step beyond the original scope)
  - Data analysis

Results

- Hardware:
  - CPU Cluster (
    farm nodes), (
  - 4 data web se
  - 25 Database
- Performance
  - Simple job (1
  - Realistic job (
    events/secon
- With empty workers cache: 10-20 times slower

Under evaluation by:

to ~100 TB

ing built on top of

onstructed and raw

- DES
  - Data processing (a step beyond the original scope)
  - Data analysis

Contact Igor Mandrichenko for details of getting started
Also, let FIFE know if you'd like to see a longer presentation at an upcoming meeting

FIFE Roadmap

Phase II Milestones - End of Q2 2019

- Solution for rapid user code distribution (CVMFS/JobSub)
- Clean up policy for experiment cvmfs repositories
- Ability to delete SAM DataSets (ability to identify what can be deleted; tools to organize deletion)
- Hybrid SAM (integration with RUCIO)
- Provide GPU solution for the experiments
- project.py for analyzers rewrite (integrate with POMS)
- Docker/singularity Build service in production
- Implement generic access to GPU resources for experiments
- Implement SBN file sharing based on collected requirement

Phase III Milestones - End of Q4 2019

- Investigate benefits for splitting resources between production and analyzers (FIFE/IF LPC)
- Unmount fermi app from interactive nodes
- Transition plan to SL7
- Landscape: Anomaly detection
- RUCIO with dataset catalogs
- UPS successor products available for evaluation with a new product(?)
- Visibility of SNOW tickets
- Federated IDs (no user certificates for job submission and storage)
- Python 2 -> 3 migration

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

State of Resources

Stu Fuess

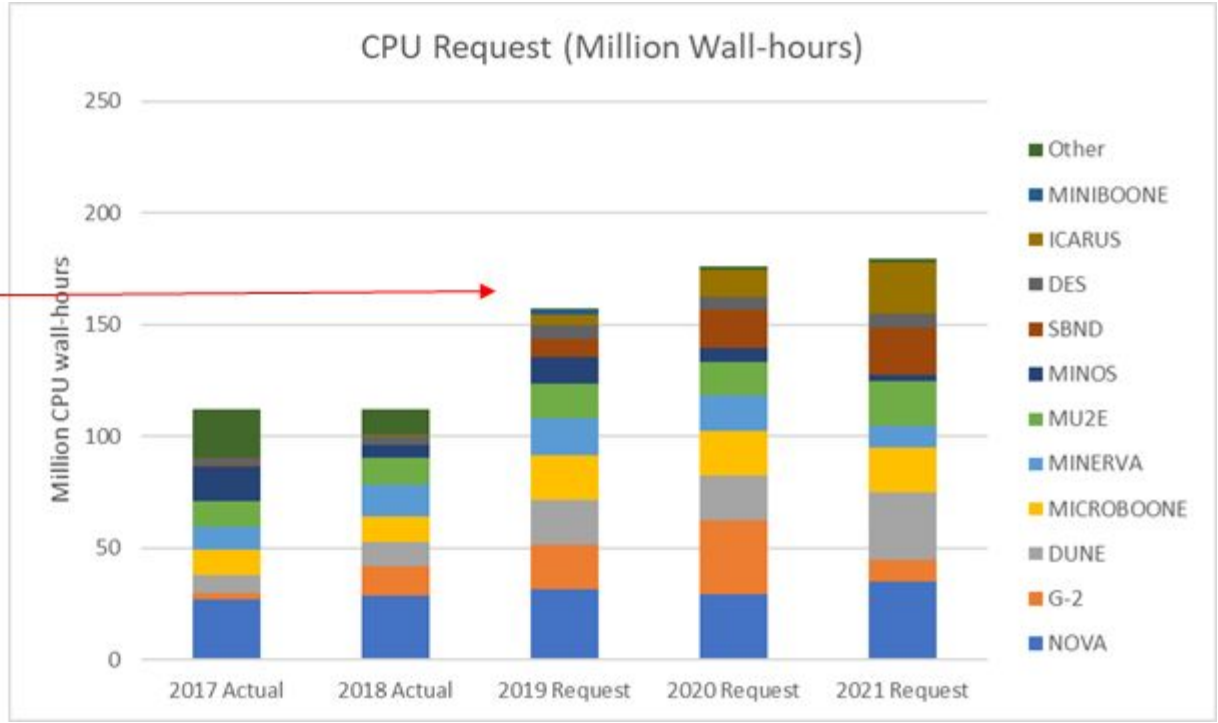
FIFE Roadmap Workshop

January 17, 2019

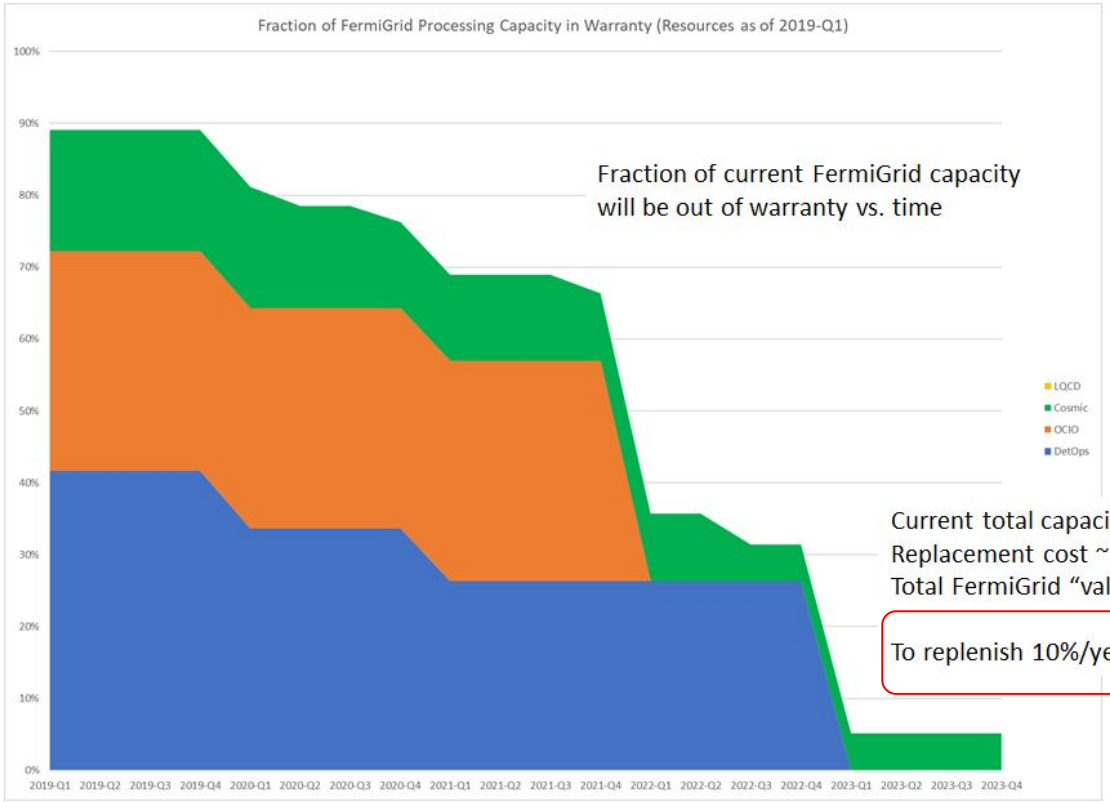
FABRIC FOR FRONTIER EXPERIMENTS

SCPMT - CPU requests

Current capacity
160 M hours/year



FermiGrid warranty expiration



Fraction of current FermiGrid capacity will be out of warranty vs. time

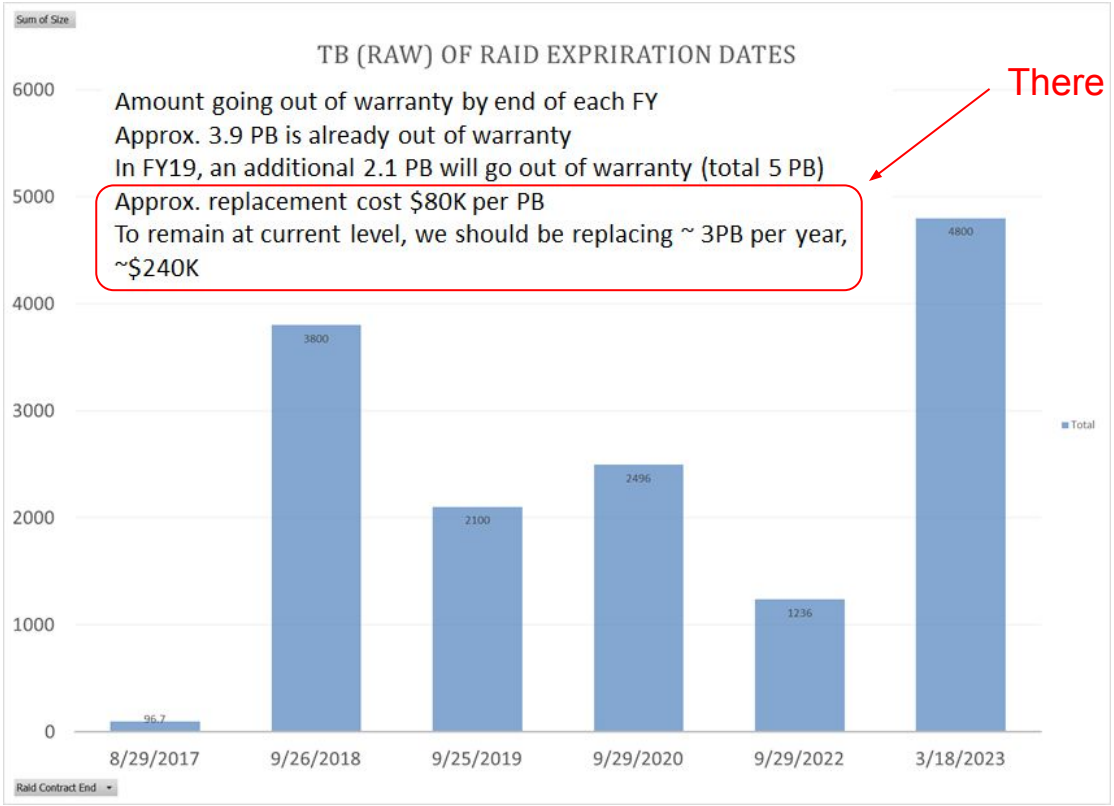
Current total capacity ~ 200 kHS06
Replacement cost ~ $9K / kHS06
Total FermiGrid "value" ~ $1.8M

To replenish 10%/year need ~ $180K

There is no budget for this

Fermilab

Public dCache warranty expiration

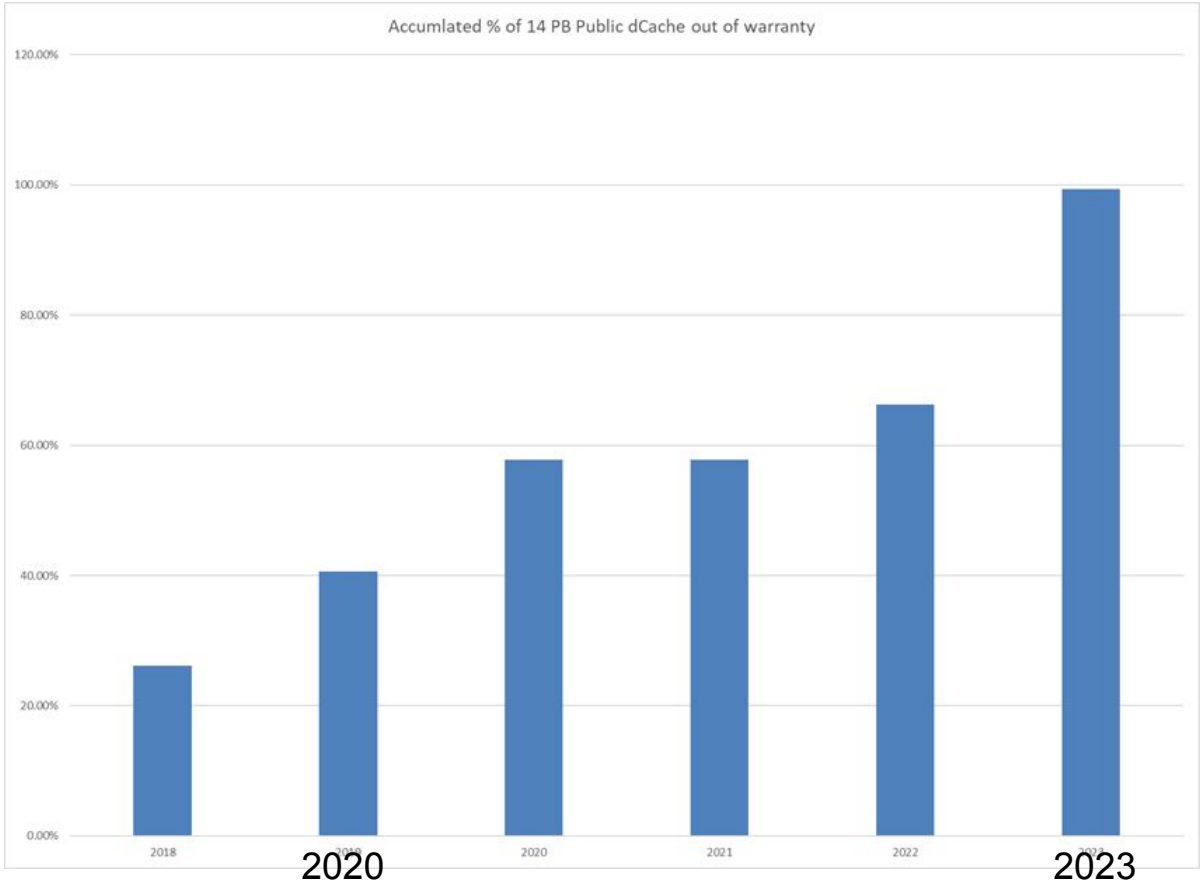


There is no budget for this

Fermilab

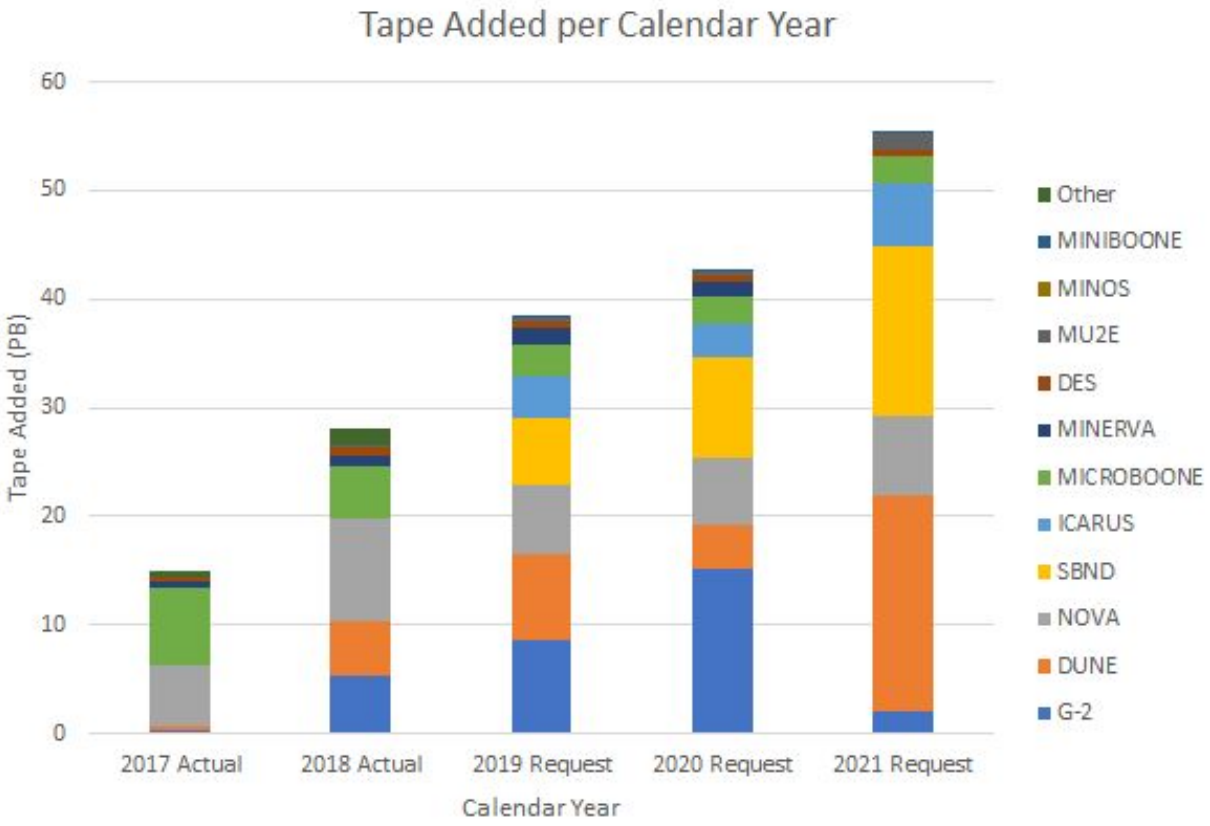
Public dCache warranty expiration

100%



Fermilab

SCPMT - Tape requests



Also need to migrate ~50 PB per year from old media

Budgeted FY19 to make this happen - but not so clear following years

Fermilab

Resource Summary

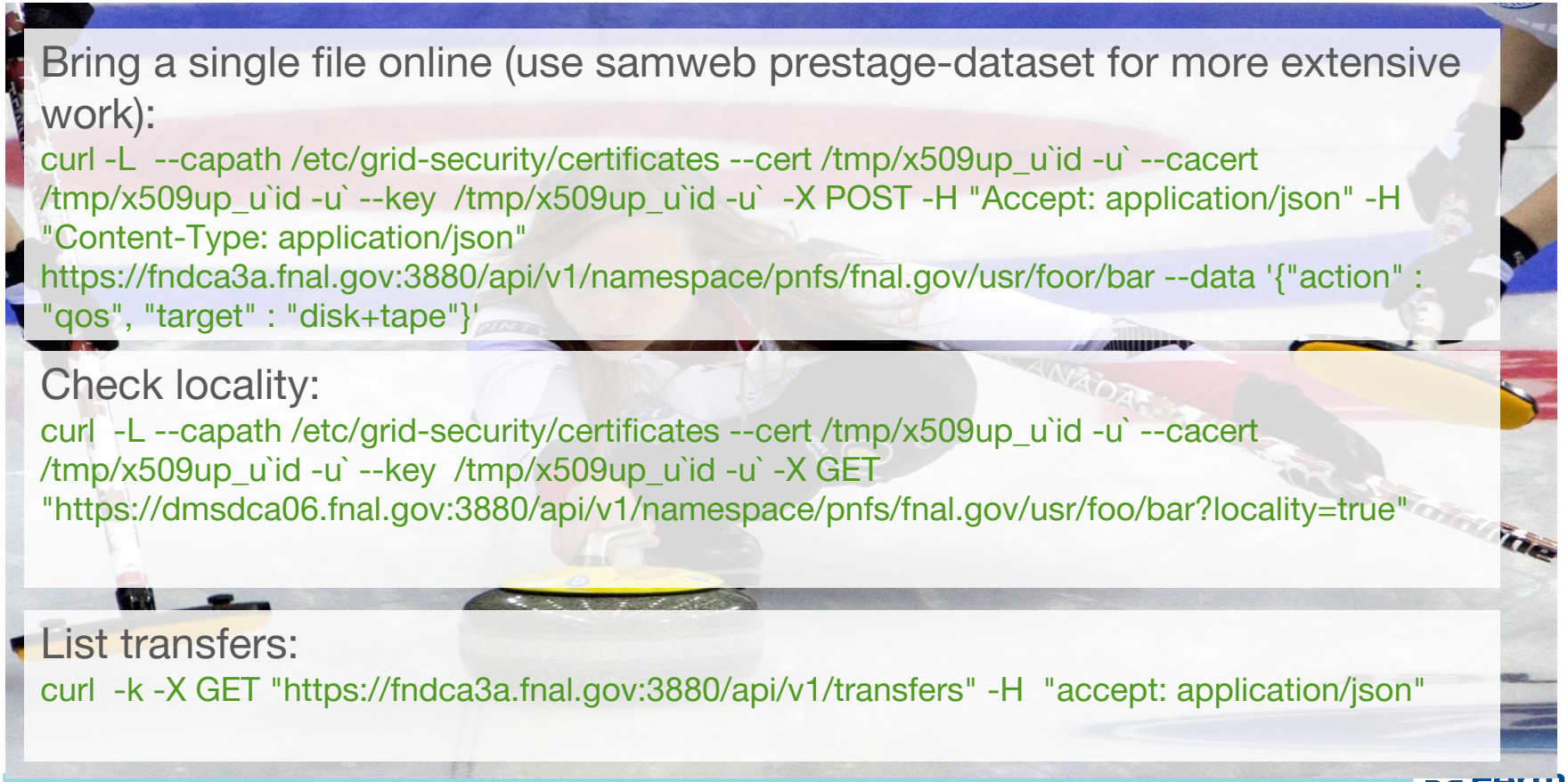
- There is no budget to update FermiGrid in FY19
  - Will argue to keep at same level in future as systems go out of warranty
  - Must learn to use HEPCloud (OSG, HPC, Cloud)
- There is no budget to update dCache in FY19
  - Storage must be used wisely
- Substantial needs for tape media may exceed funds
  - We need to eliminate migration of worthless data

Summary

- FIFE Project continues to provide integrated solutions for the Fermilab scientific stakeholders. We have had several major successes the past several months in monitoring, international resource expansion, automation of SNOW requests, and POMS improvements
- The next big step: the HEPCloud Facility, including possibility to execute jobs on HPC and Commercial clouds. The HEPCloud team helped NOvA use NERSC with great success last year; looking forward to onboarding others. FIFE stands ready to work with experiments when transitioning workflows to these new resources
- We are working to find common solutions whenever it is possible: promoting a common monitoring framework, adapting Rucio, or contributing to the SciToken project
- We are starting several new projects that include:
  - Rapid user code distribution
  - Container Build Service
  - Project-py rework
- We are proposing an exciting roadmap for the next year that will help experiments to improve code building and distribution frameworks, access to GPUs, address anomaly detection problems, and transition to SL7
- **See you at the next Roadmap workshop in the summer!**

BACKUP

dCache status: Going curling



Bring a single file online (use samweb prestage-dataset for more extensive work):

```
curl -L  --capath /etc/grid-security/certificates --cert /tmp/x509up_u`id -u` --cacert
/tmp/x509up_u`id -u` --key  /tmp/x509up_u`id -u`  -X POST -H "Accept: application/json" -H
"Content-Type: application/json"
https://fndca3a.fnal.gov:3880/api/v1/namespace/pnfs/fnal.gov/usr/foor/bar --data '{"action" :
"qos", "target" : "disk+tape"}'
```

Check locality:

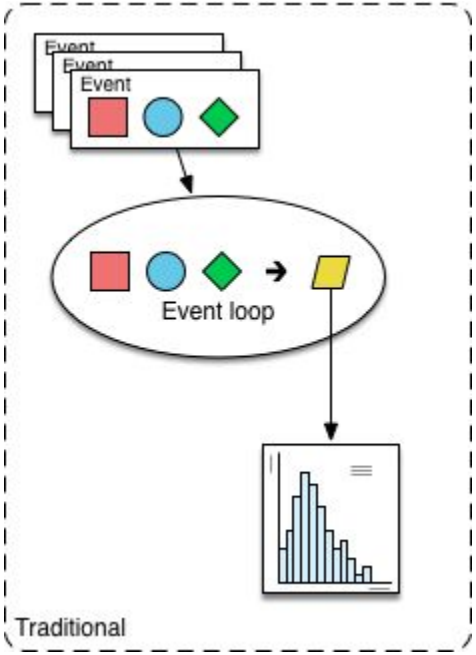
```
curl  -L --capath /etc/grid-security/certificates --cert /tmp/x509up_u`id -u` --cacert
/tmp/x509up_u`id -u` --key  /tmp/x509up_u`id -u` -X GET
"https://dmsdca06.fnal.gov:3880/api/v1/namespace/pnfs/fnal.gov/usr/foo/bar?locality=true"
```

List transfers:

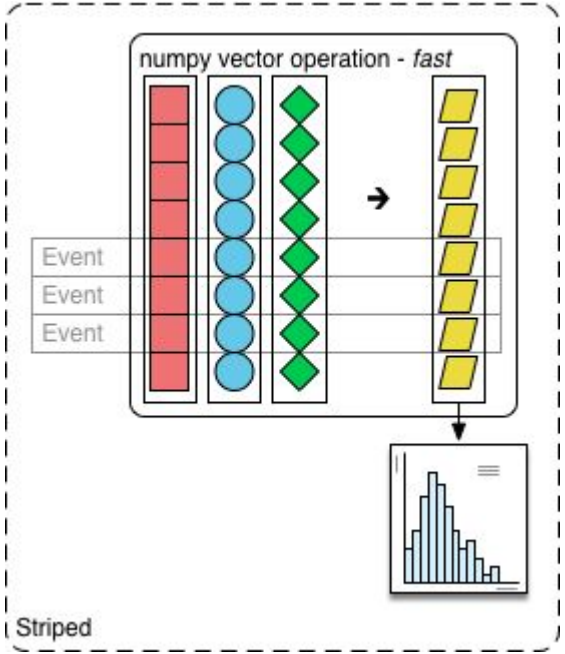
```
curl  -k -X GET "https://fndca3a.fnal.gov:3880/api/v1/transfers" -H  "accept: application/json"
```

New analysis model

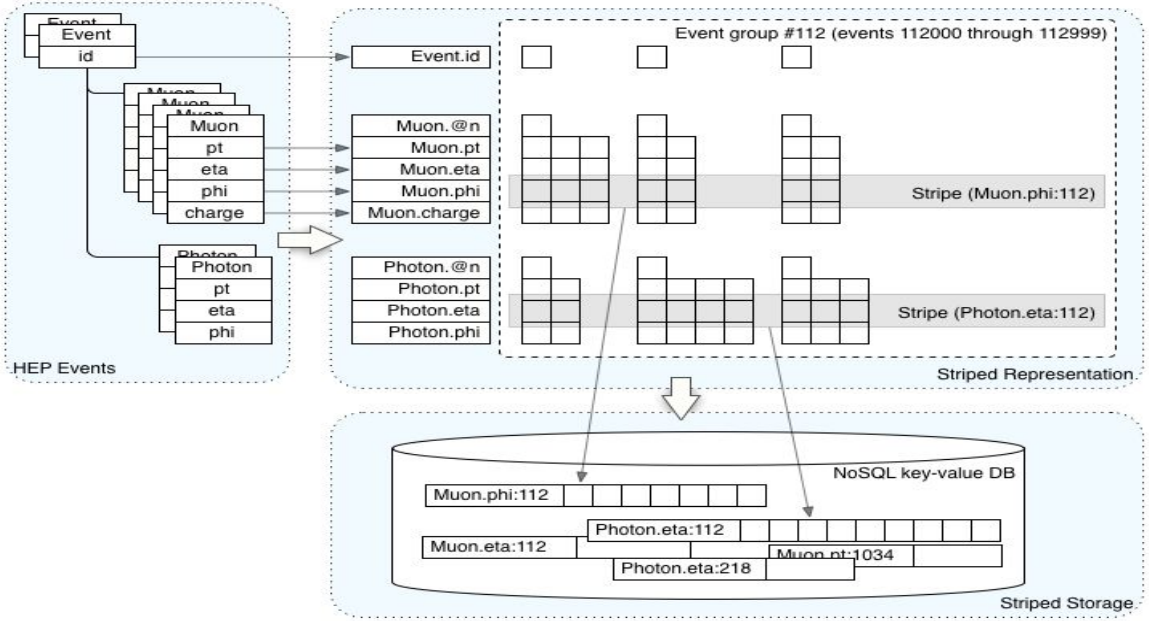
From Event Loop



To Vector Computations

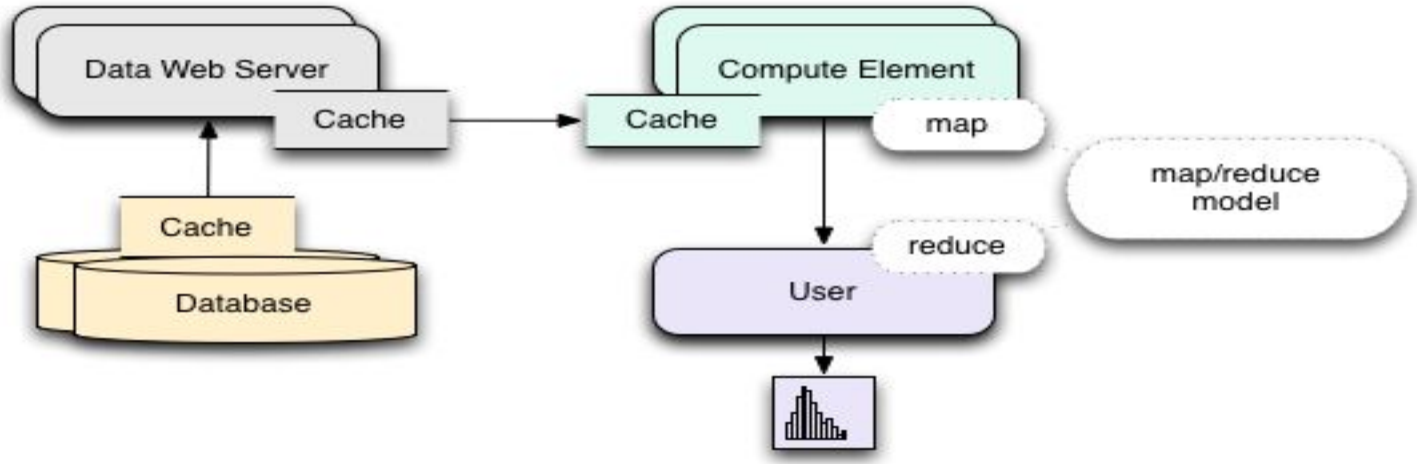


Storage: Striped Data Representation

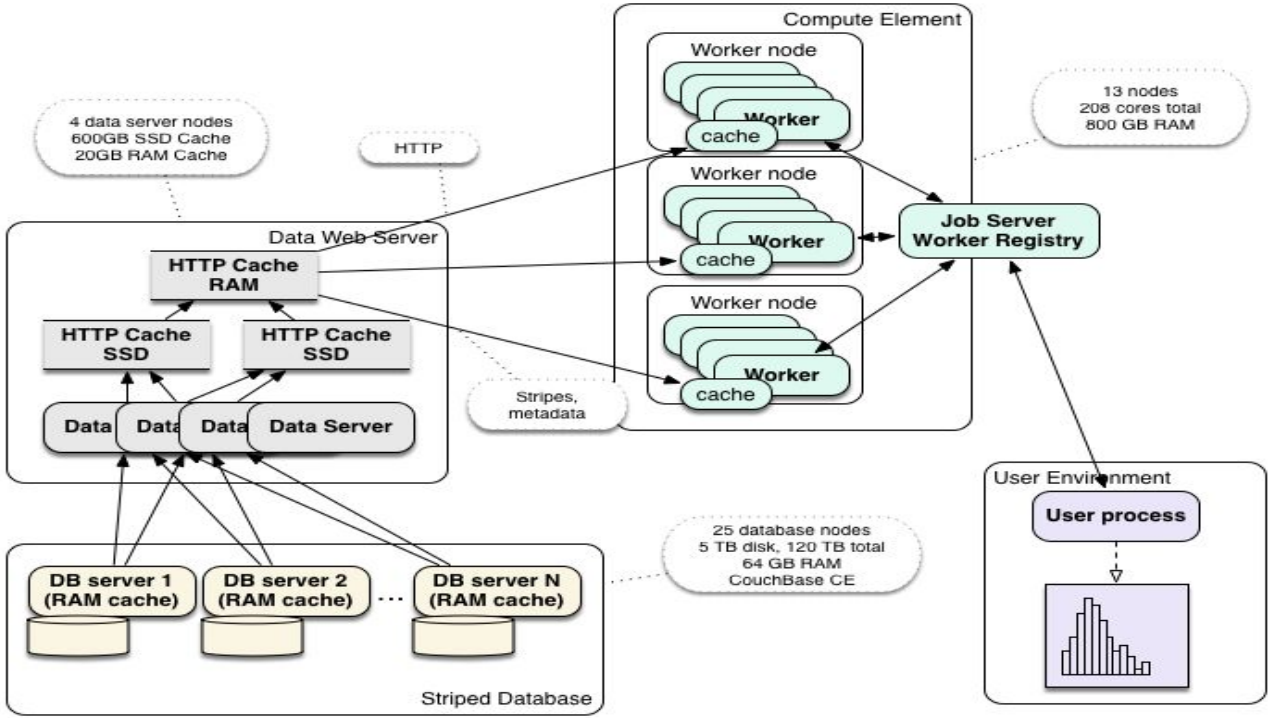


- Columnar representation
- Unit of representation is a Stripe
  - one column for 10-100K events
- Internal format is numpy array
  - Vector-ready data representaiton
- Any key-value storage can be used as the backend
- Distributed NoSQL backend for scalability
- Stripes are perfectly cacheable
- Scales easily to 1000s columns and beyond

Architecture: high level view



120TB Striped Analysis Facility



Framework Software Stack

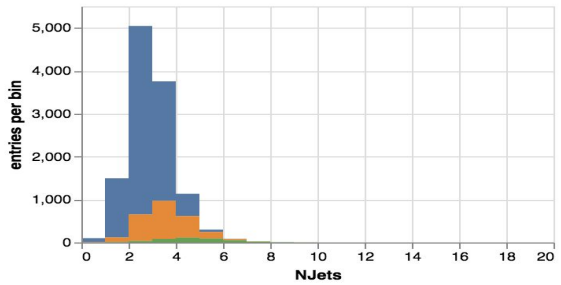
- Python (2.7 for now), WSGI
  - +1 C/Python module
- Numpy – vector computations
- Histbook, Matplotlib - visualization
- Nginx, uwsgi – web server, cache
- CouchBase – database (can use any key/value storage)
- Jupyter/IPython – user analysis frontend
  - Can be used in command line mode, batch mode
- Docker – client, worker virtualization

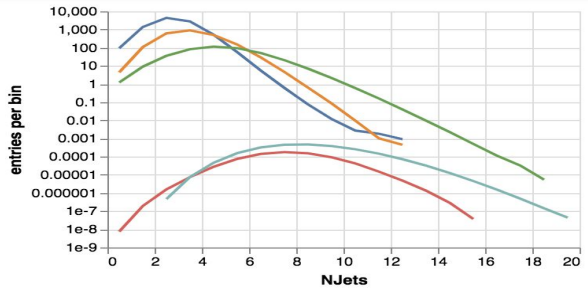
Jupyter/IPython user frontend

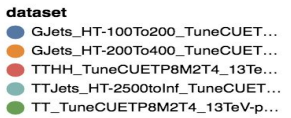
In [9]:

```
display.init()
callback = Callback(disp)

for dataset_name in datasets:
    job = session.createJob(dataset_name,
        user_callback=callback,
        user_params = {"dataset":dataset_name},
        histograms=[njets_by_dataset, njets_combined, btags_by_dataset, btags_combined])
    job.run()
    runtime = job.TFinish - job.TStart
    nevents = job.EventsProcessed
    display.update()
    print "%-70s %7.3f M events, %7.3f M events/sec" % (dataset_name, float(nevents)/1000000, nevents/runtime/1000000
```





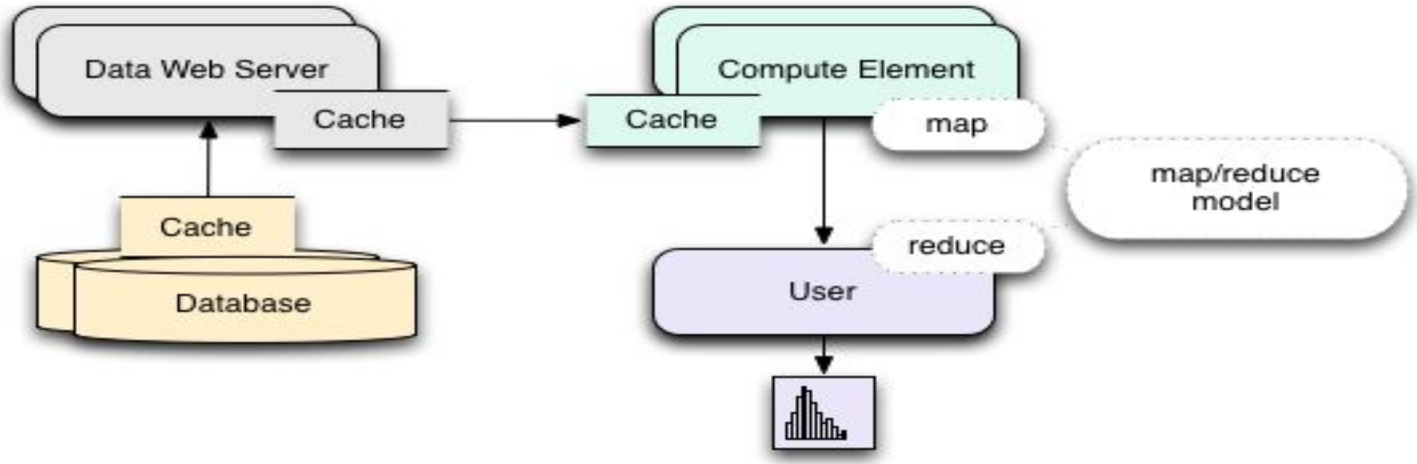



```
Summer16.TTHH_TuneCUETP8M2T4_13TeV-madgraph-pythia8                0.100 M events,    0.310 M events/sec
Summer16.TTJets_HT-2500toInf_TuneCUETP8M1_13TeV-madgraphMLM-pythia8      1.520 M events,    2.103 M events/sec
```

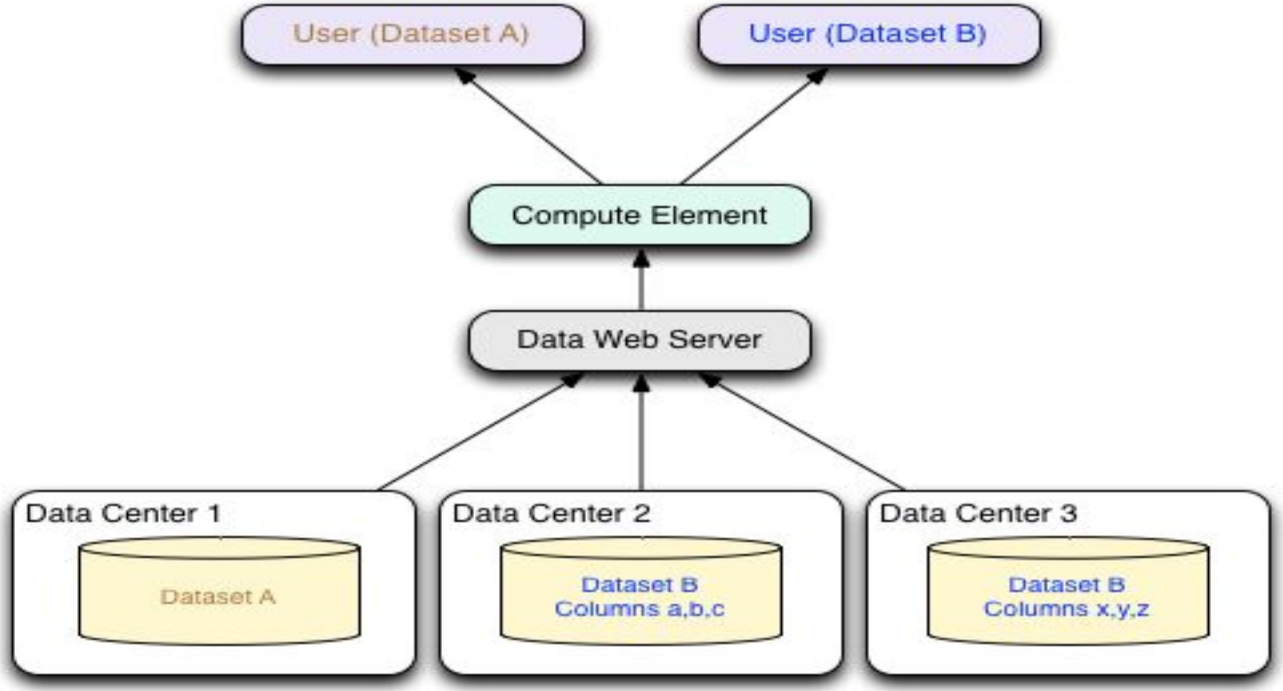
Performance

- CPU Cluster of ~180 cores, 13 servers (old farm nodes), 64GB RAM each
- 4 data web servers, SSD cache: 2x300 GB
- 25 Database servers
- Performance (cached at workers):
  - Simple job (1 column): 30 M events/second
  - Realistic job (~20 columns): 9 M events/second
- With empty workers cache: 10-20 times slower

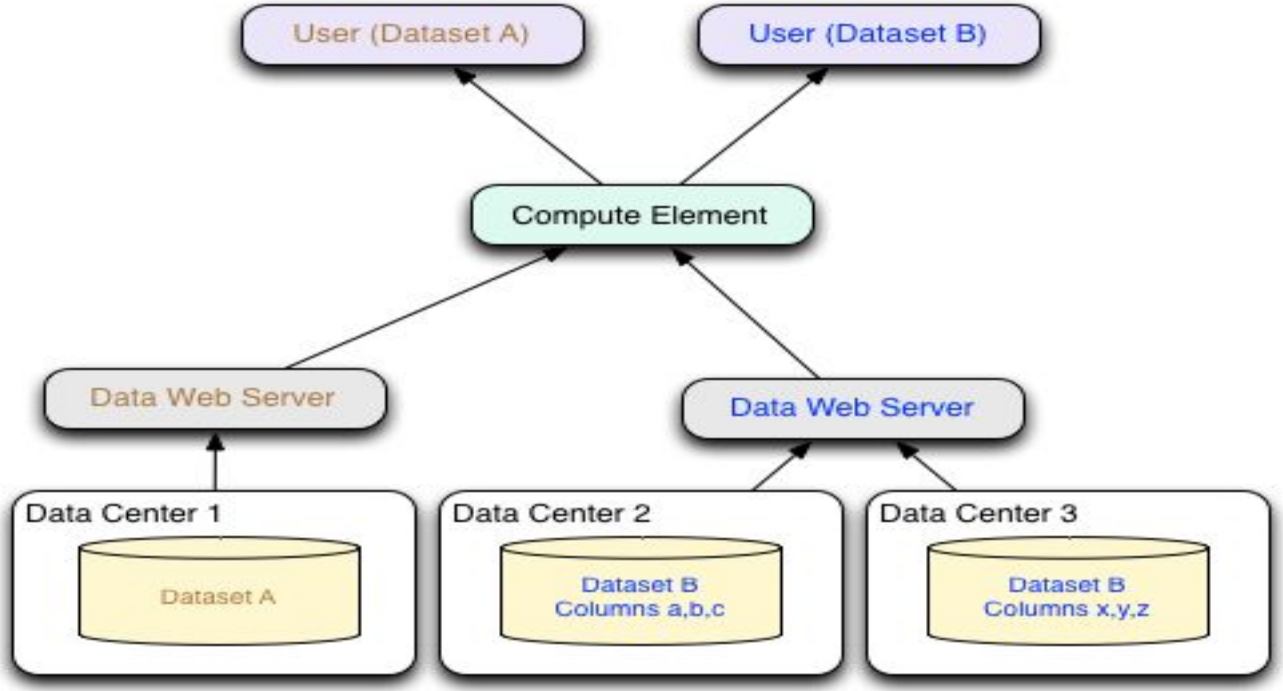
Flexible Architecture



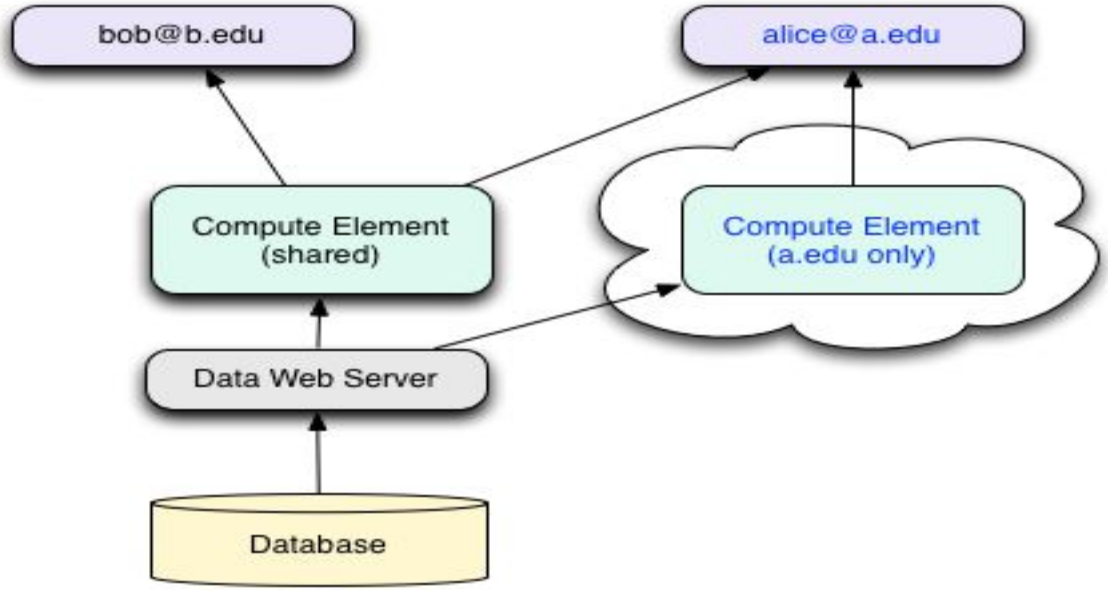
Storage Flexibility



Storage Flexibility 2



Compute Flexibility



Virtualization

- Docker images for worker and client
- Successful attempt to run a job on a mix of workers running on:
  - Real computers @FNAL
  - Docker containers @AWS
  - Docker container on my iMac desktop

100 Terabyte System

System capacity	10 TB	100 TB
Events	1B	10B
Events/dataset	10M	10M
Frequently analyzed data	0.1TB	0.5TB
Worker Cores	200	200
Worker RAM	1TB	1TB
Web cache	1TB	1TB
Web servers	4	4
Database nodes, TB/node	1	40
Database nodes	14	30
Database nodes RAM	64 GB	2TB
Total network throughput	20 Gbit/s	20 Gbit/s

1 Petabyte System

System capacity	100 TB	1 PB
Events	10B	100B
Events/dataset	10M	50M
Frequently Analyzed data	0.5TB	2TB
Worker Cores	200	2000
Worker RAM	1TB	10TB
Web cache	1TB	3TB
Web servers	4	10
Database nodes, TB/node	40	200
Database nodes	30	60
Database nodes RAM	2TB	4TB
Total network throughput	20 Gbit/s	200 Gbit/s

Computer types needed

- Worker
  - RAM, cores, network
  - GPU
- Web cache
  - RAM, SSD, network
- Web server
  - Cores
- Database
  - RAM, cores, fast (not big) disk, SSD is better

InCommon CA: for international science users

- FYI for DUNE and others: the InCommon CA provides user certificates, too. Standard FNAL certs are CILogon Basic CA (not accepted at some EGI sites).
- If you have need other certs to work with your collaborators, e.g. CERN, these InCommon certs may be very useful to you
- These certificates are not yet integrated into jobsub or other command line client tools. They can be useful if all you want is a browser-based certificate. We can talk more if you think you can use it