Managed by Fermi Research Alliance, LLC for the U.S. Department of Energy Office of Science

The Mu2e Experiment at Fermilab : Experience with OSG Opportunistic

Ray Culbertson

OSG All Hands Meeting 2016 (non-LHC session)

14 Mar 2016

This is a sub-talk

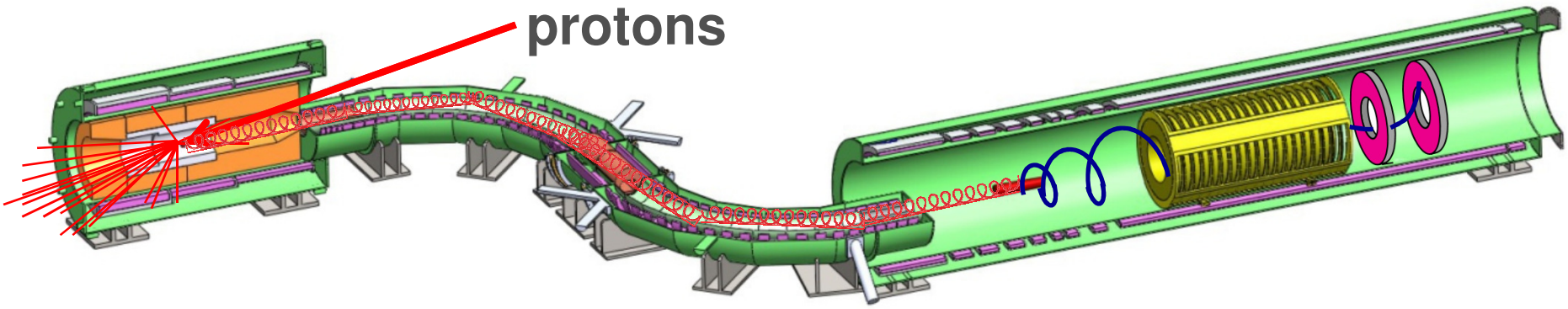
- I'll be giving an overview talk at the plenary on Wednesday
- Today, focus on some more details
  - Lessons learned
  - Requests
  - I'm open to discussions and suggestions

- Right up front, I'd like to thank Ken Herner (our main contact) and all the Fermilab and OSG support staff that made this project possible!

- OSG has enabled Mu2e to meet or exceed all our computing goals, and couldn't have been done without them!!

The Mu2e Experiment

- Building at Fermilab, commissioning in 2021
- Searches for very rare conversion of a muon into electron



- Looking for a few 105 MeV electrons in 3 years of running
- Requires extraordinary control of backgrounds, < 1 event
- Which requires lots of simulation!
- In the last year, a big push to prepare for DOE CD-3 review

OSG Solves the Mu2e CD-3 CPU Problem

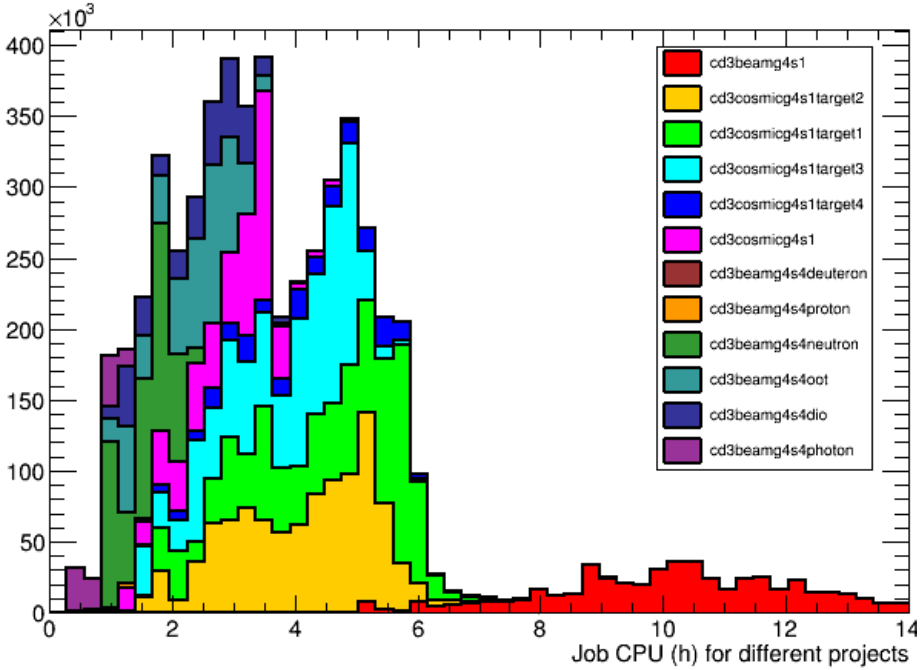
- Commissioning
  - Dec 2014 – started commissioning cvmfs
  - Feb 2015 started commissioning OSG
  - Apr 2015 running tests of 5K slots regularly, adding sites
  - Jun 2015 start production
  - Sep 2015 succeed making the baseline checkpoint goal
  - Production almost continuous, and only winding down now
- Met all basic goals (14 Mh) and went on to meet stretch goals!
- 60 Mh (wall) in the last year
- 10 M jobs submitted
- 33 M files (276 TB) uploaded to tape at FNAL
- 75 G events simulated

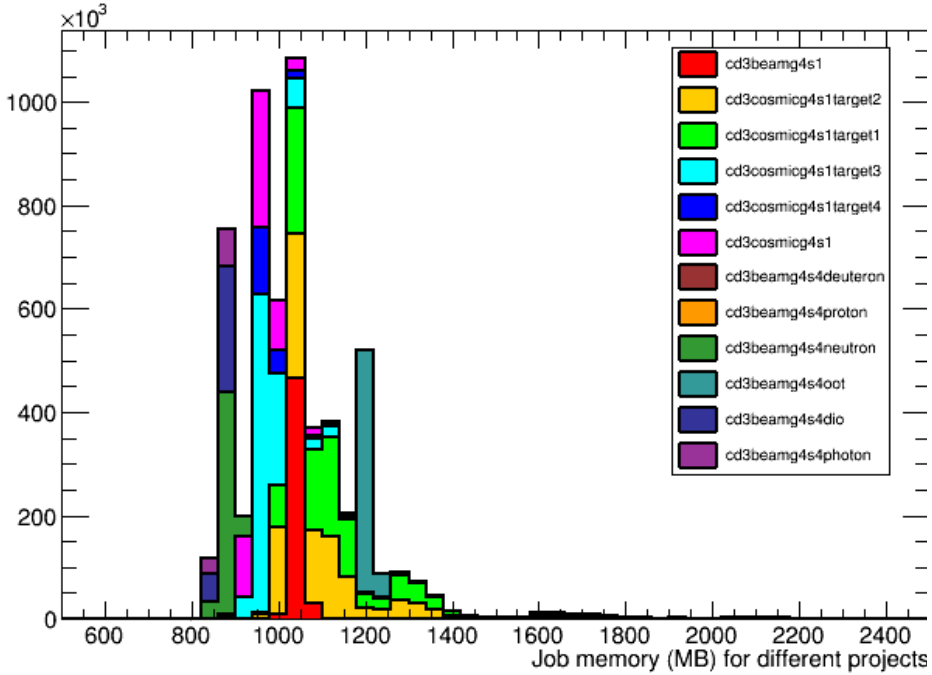
The Mu2e Simulation Job

- Submit via **jobsub** (FNAL interface to condor)
- Command file moved to worker node by condor
- FNAL infrastructure packages and Mu2e code are on **cvmfs**
- Copy in a small configuration file
  - via **ifdh** (FNAL interface to gridftp and other transport)
- Run the Mu2e simulation executable
- Copy back 5-10 small files (20 MB or less)
  - Also via **ifdh**
- Input and output to **dCache** (distributed disk at FNAL)
- Hand-run run scripts to validate job output
- Move to tape and final tape-backed dCache location
  - **FTS** (Fermilab procedure to upload files)

What Ran on Sites

- CPU and memory separated by project

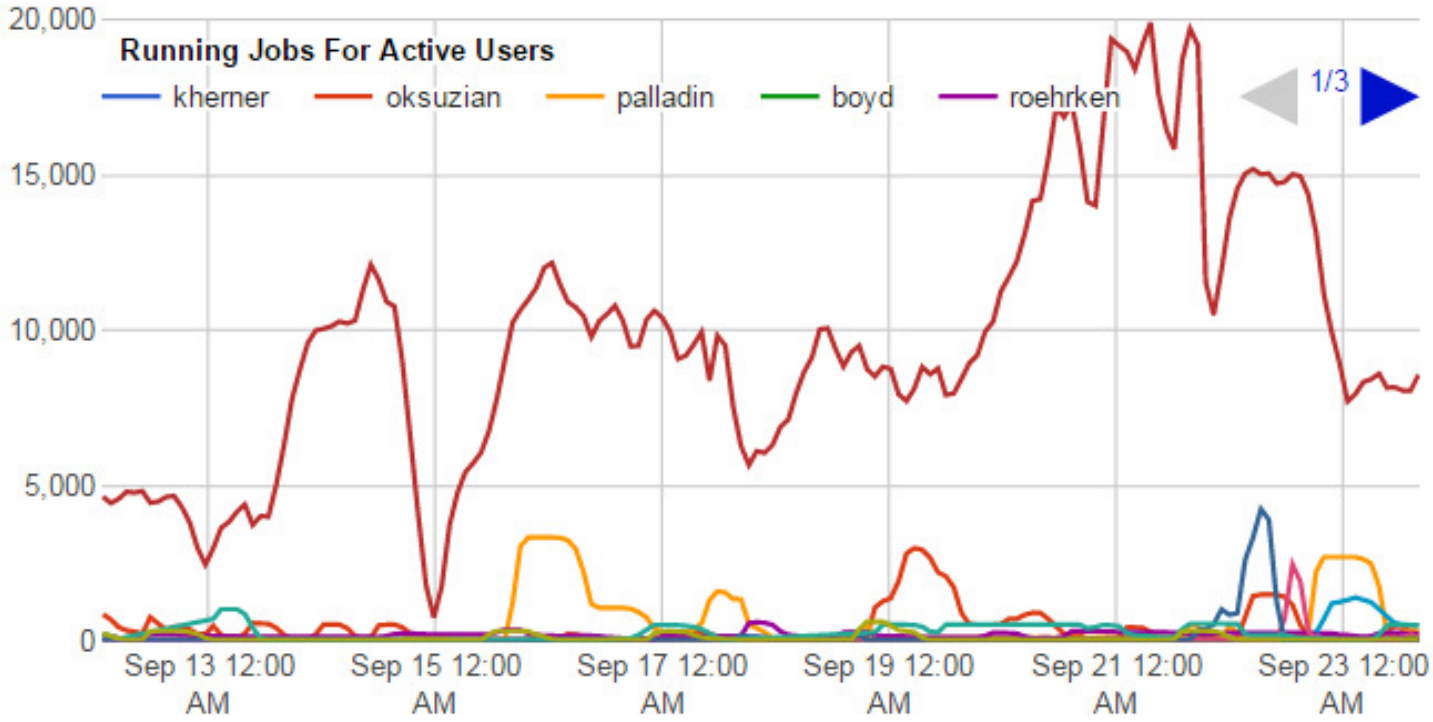




- CPU mostly 2-6h, one job longer, to 16h
- memory 0.8-1.8 GB

What Ran on Sites

- Typically running 8K jobs at a time, with large variations,
  … up to our record, 20K



Lesson #1 – need information!

In the process of running production and uploading files, we found the need to set up 11 monitoring processes

- OSG
  - Monitor each site, mine condor logs (2 ways)
- dCache
  - Response time, space, connections, tape activity by VO
- SAM file catalog
  - Dataset listings, response time
- File Upload
  - Upload area contents by dataset, processing rates
- Plus many scripts for collecting info, debugging
- Some existing monitors are “near-miss” to the info we want, and generally getting closer as time passes

Lesson #1 – need information!

- Submit probe test job on all sites
  - Test the whole infrastructure chain *for Mu2e needs*
- Submit twice a day, kill it if it doesn't run in 12 hours
- Summarize all results
  - A "5" is complete success, lower means some step failed

Time	DEDI	MWT2	Nebr	OPPO	Corn	Omah	SU-O	UCSD	UChi	Wisc	Calt	Notr	FNAL	MIT	BNL	Clem	Mich	TTU	Hyak
2016-02-22 06:00	5	5	5	5	0	5	5	5	5	5	5	5	5	0	5	4	5	0	5
2016-02-21 18:00	5	5	5	5	0	5	5	5	5	5	0	5	5	0	5	0	5	0	5
2016-02-21 06:00	5	5	5	5	0	5	5	5	5	5	5	5	5	0	5	0	5	0	5
2016-02-20 18:00	5	5	5	5	0	5	0	5	5	5	5	5	5	0	5	0	5	0	5
2016-02-20 06:00	5	5	5	5	0	5	0	0	5	5	5	5	5	0	5	0	5	0	5
2016-02-19 18:00	5	5	5	5	0	5	5	0	5	5	5	5	5	0	5	0	5	0	5
2016-02-19 06:00	5	5	5	5	5	5	5	0	5	5	5	5	5	0	5	4	5	0	5
2016-02-18 18:00	5	5	5	5	5	5	5	4	5	0	5	5	5	0	5	4	5	0	4
2016-02-18 06:00	5	5	5	5	0	5	0	0	5	5	5	5	5	0	5	0	5	0	5
2016-02-17 18:00	5	5	5	5	0	5	5	0	5	5	5	5	5	0	5	0	5	0	5
2016-02-17 06:00	5	5	5	5	0	5	5	0	5	5	5	5	5	0	5	0	5	0	0
2016-02-16 18:00	5	5	5	5	0	5	5	0	5	5	5	5	5	0	5	0	5	0	5
2016-02-16 06:00	5	5	5	5	0	5	0	0	5	5	0	5	5	0	5	4	5	0	5
2016-02-15 18:00	5	5	5	5	0	5	5	5	5	5	0	5	5	0	5	0	5	0	5

Lesson #2 – small files can create a major problem

- 2-16h CPU job produces 5-10 small files, 3-5 to be uploaded
- While this style was generally agreed to early on, I don't anyone realized how much trouble it could be
- We really made it worse
  - Uploading an individual control file per job – could be scripted
  - jobs could be chunked more logically, pipelined
  - Uploaded individual log files – could be tarred
  - Split output files for convenience – could be redesigned
  - Reducing by x2 would bring it in reasonable range
- The fundamental problem of one 8h jobs producing ~20MB output can't be avoided – it's the physics and detector

Lesson #2 – small files created a major problem

Small files cause problems at every stage

- Waiting for 1 M job control files to upload was the single largest cause of delays
  - The upload system could become overwhelmed, and only crude control of upload priorities
- Gridftp can become overloaded with the number requests
- Gridftp and dCache have significant per-file overhead
- Every “find” in processing takes minutes to hours
- Slow processes require them to be “kept up” adding to maintenance and monitoring/debugging can be painful!
- Production procedures are evolving
- Improved matching procedures to services is underway

Lesson #3 – understanding friction

- Friction is ongoing and requires constant vigilance and maintenance
  - sites may come or go for whatever reasons
  - dCache issues pop up in various forms at any time
  - projects differ and create different issues
- For foreseeable future, always need automated resubmission

However,

- Infrastructure failure rates have gone down continuously
- At start they were 10-20%, now almost always under 1%
  - Birthing pains are gone
  - Sites seem more consistent, stable
  - cvmfs errors much more rare
  - Many more alarms, checks, and automatic retries, etc.

One Year of Servicedesk Tickets

A very rough sorting/analysis of 179 tickets related to Mu2e production in the last year…

- 51 submission infrastructure
  - Jobsub, fifebatch, condor, glideins, monitoring, condor logs
- 37 dCache
- 35 file handling at FNAL (SAM, FTS and other)
- 26 ifdh and gridftp
- 18 CMVFS
  - Many of these are problems at specific OSG sites
- 12 OSG site issues

Scale is 2/week for lab infrastructure, 1/week for OSG sites, plus numerous email threads, conversations

Some Typical Issues local to FNAL

- Fifebatch (condor servers) overloaded or crashed
- “Sandbox”
  - No disk space
  - Cant change ownership
- Monitoring down or incorrect
- Gridftp servers
  - Overloaded or hung
  - rejecting authentication
- dCache
  - Overloaded or hung, not responding
  - Components crashed
  - Missing directory entries

Site Issues I

- Local software
  - uberftp not at latest version - triggers knowns bugs
  - /usr/bin/time command not installed (reports memory)
    - eventually we customized it and put on cvmfs
    - pass all signals, memory incorrect by a factor of 4
- Kernels
  - Request SL6, see 99% 2.6 1% 3.x 0.1% 4.x
  - This issue mostly manifested in FNAL “setup” infrastructure
- Optional libraries
  - Some sites do not install X11 display libraries
  - We now provide them on cvmfs and include them in library path
  - Developing graphics-free builds, but would rather it just works

Site Issues II

- Authentication failing
  - By VO or user
  - Hard to differentiate from no slots available and ad mismatch
- CMVFS
  - Not mounted, wrong version
  - Cache not up to date
  - Corrupt, causes seg faults and missing files
- Single–node black holes
  - Often CVMFS errors
  - Hardware errors: seg fault, bus error, input/output error, disk full
- Job restarts (see next)

Lesson #4, Restarts are major factor

	Job Stats	Disconnect Prob (%)	Eviction Prob(%)
BNL	64556	0	0
Caltech	758339	3	0
Clemson	1371	0	3
Cornell	295	16	3
FNAL	383615	1	0
Fermigrid	4861	0	0
Hyak_CE	9319	54	5
MIT	5205	73	0
MWT2	203678	1	4
Michigan	95476	0	1
Nebraska	462291	44	0
NotreDame	282337	16	0
Omaha	517470	34	0
SU-OG	2003389	10	0
UCSD	45546	2	0
Wisconsin	602861	36	0

- Counted by condor log status 10/2015-2/2016

Job restart issues and questions

- Overall, 17% of our jobs get disconnected or evicted
- Every day 1000's of jobs are disconnected
- This leads to a long tail in processing any large submissions
- Many cases where we are getting N slots continuously on a site, but also continuously disconnecting at very high rates - 30% or more – why?
- I'm told some evictions are reported as disconnects – why?
- Sites show a huge variation in this metric, clearly completely different procedures or policies

OK, now some thoughts

I have no significant knowledge of OSG history and ongoing discussions, known roadblocks, etc.

But, under the assumption that you want to hear what users want, I will attempt some detailed feedback

Hopefully, if nothing else, it reinforces existing priorities, or provides ammunition in future negotiations

Thoughts #1 – reliability

Make opportunistic access more reliable

- *Guarantee at least 8 h running time before evicting a job*
  - Or at least 2 h !?
- This is the single most important request I have!
- Make this uniform on all sites
- Also
  - Label evictions as evictions, and hold as hold
  - Reduce disconnect timeout
  - Debug/reduce any other disconnect issues
- This will virtually eliminate CPU waste and reduce long tails

- Of course, always add sites and open more sites to all VO's!
- Can we get into non-U.S. sites?

Thoughts #2 – transparency

- Announce capabilities in detail
  - Hardware (OS, memory, disk)
  - Slot counts for each
- Announce site policies
  - How quotas and priorities are set
  - How much is open to opportunistic? Any VO limits?
  - When and how are jobs evicted?
- Announce opportunistic availability continuously
  - Ideally slots with their characteristics
  - When our jobs don't land on a site, I can't tell if that's an error, a mismatch, or simply reduced opportunistic slots
- I expect some of this is available now, but why don't I, as a power-user, know about it?

Thoughts #3 – monitoring and debugging

- Access to an example node for commissioning and debugging
- Some access to ongoing problems on all nodes – top, tail, ls or return log files of jobs which go over resource limits. Some core files on demand?

Not sure what the following would actually look like

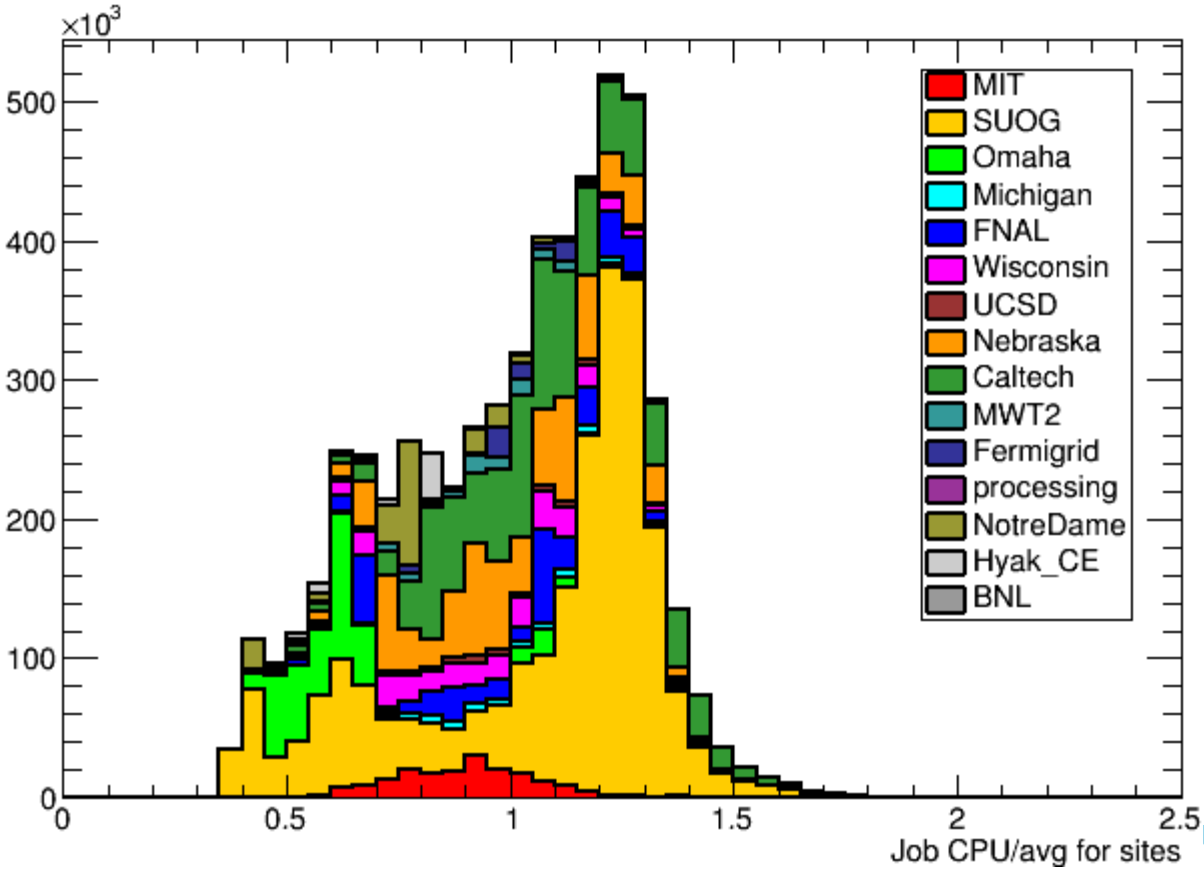
- Monitor cvmfs on each node and blacklist any that show errors
- Mitigate black holes – detect, throttle/blacklist, and report nodes which show anomalous job lifetime? On request only..

More uniformity:

- Uniform software
- Uniform, optimal /usr/bin/time command

Thought #3, CPU power normalization

- CPU variation between jobs in a project is order 2%
- CPU/(avg CPU for project) shows factor 3 variation
- Without normalization, makes planning harder..



Fermilab

Thank you for these resources!

- The OSG resources were critical to the Mu2e success for CD-3 review and will be critical for our future efforts
- Overall, a huge success using a fantastic resource!
- Effort was very manageable
  - Order 1 FTE-month to get going on 10-15 sites
  - Maintenance is order 5-10% FTE

The Mu2e experiment would like to thank all those involved from the OSG, the sites, and the lab infrastructure support. You made this success possible!