



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

Recent infrastructure and Resource Needs

Ray Culbertson

Mu2e Computing Review

5 Mar 2015

Overview

Recent Infrastructure

- Sam
- File Upload
- CVMFS
- OSG
- Resources for CD3 processing
 - CPU, disk, tape
 - dCache and FTS
 - People
- Summary

Recent Infrastructure Advances

All these were enabled by SCD
infrastructure and expert support !

SAM

- Defined file names
data_tier.owner.description.configuration.sequencer.file_format
sim.mu2e.tdr-cosmic-g4s2.1613a_1613a.15799263_000715.art
- Defined Enstore File Families
3 for group: phy-sim, phy-nts, phy-etc, 3 for users, 1 for test beam
- Defined SAM metadata
MC generator type and simulation stage
Run/event/subrun for finding data
Datasets and bookkeeping
- Documented, with user scripts
- Running well, modulo issues:
 - Two files have 250K subruns listed – access can occasionally fail
 - MC production/user read workflows not updated for SAM – but should be straightforward

File Upload - FTS

- Collaboration code: `jsonMaker.py`
 - Input: the file (runs code), plus a little metadata
 - Output: json file of complete SAM metadata
 - Can also rename, move files to FTS
 - Web documented
- File Transfer Service (FTS) commissioned
 - A directory for each file family
 - Only production user can write to production FF
 - Waits for file and its json metadata file
 - Creates SAM record
 - Moves file to tape-backed dCache (auto to tape)
- We have uploaded 1M files in the last 4 months (more later)

CVMFS

- mu2e CVMFS is available
 - Created in December: /cvmfs/mu2e.opensciencegrid.org
 - Parallel to the other experiments
 - Following all known best practices – catalogs, limitations
- Commissioned
 - Mounted on all interactive nodes
 - Loaded with releases: ~2.5G/8K files each flavor, products: 8GB, 207K files, B field data: 4GB
 - Setups modified, web documented
 - Tested interactive use, faster than current bluearc code disk
- Offsite
 - OK on four US sites, missing on other
 - International mounts in progress in GOC ticket

OSG

- 97Mh/y opportunistic jobs run, 50Mh/y probably still idle
- Mu2e group in Fermilab VO
- Have been submitting for one month
 - Wisconsin - runs OK
 - Omaha – usually runs OK
 - SU-OG - missing rpm, (RLC)
 - MWT2 – runs sometimes, checking CVMFS (RLC)
 - OSC - CVMFS not mounted (Ken)
 - Nebraska – won't start (Ken)
 - UCSD – won't start (Ken)
 - UChicago – won't start (Ken)
- Possible additional sites being added
- Reliability, load test starting: ~1 FTE week physicist's time (RLC)

Resources for CD3 MC Production

Overview of Jobs

- From a survey of the collaboration

CPU (Mh)	Tape (TB)	Notes
2.0	2	Beam stage 1 (Project 1)
1.3	18	Beam, other stages (Project 1)
0.4	12	Cosmic general (Project 2)
4.8	92	Cosmic targeted (Project 2)
0.5	5	Neutron task force (Project 3)
0.8	30	Stopping Target optimization
0.3	3	Pbar studies
1.0	20	Analysis, includes smaller MC projects
0.02	2	Pi+ decay calibration
2.2	-	20% contingency
14	184	TOTAL
2.25	4	MARS

Job Types

- MARS jobs will be performed by MARS experts in their established workflow, in their dedicated slots
- Bulk jobs
 - Many submissions with only different random seeds
 - Most of projects 1,2,3 in talks from Andrei, Yuri and Ralf
- Specialty
 - Requires much more attention to detail
 - Will be organized and run by physicists
 - Contain some special needs such as large-memory VM
 - Most have been exercised in TDR sample production
 - Analysis

CPU

- Totals
 - 11Mh = 460 k CPU-days
 - Adding 20% for inefficiencies = 14Mh
 - To complete Apr1-Sep1, we need to attempt ~4000 slots DC
- Supply
 - Fermigrid mu2e quota=1000 (typically 500 in use),
 - Previous experience: 2000-3000 on Fermigrid
 - Therefore need 1000-2000 slots, DC, from OSG
- Risks
 - Competition on Fermigrid
 - Consistency of people and systems
 - offsite grid availability

Disk and Tape

- Disk
 - Assume workflow with data files grid to scratch dCache
 - Remainder 100KB per job sent to /mu2e/data bluearc
 - 11Mh at 8h per job = 1.4M jobs = ~ 220
 - realistically, some smaller samples will be kept on disk
 - TDR data complete by Apr 1
 - trivial size and load on /mu2e/data
- Tape
 - TDR sample was 28TB, CD3 is 184TB or ~40 tapes
- Processing
 - Must average 8s/ job to receive and check output

dCache and FTS

- dCache has greatly reduced bluearc overloads!
- Observations from TDR upload
 - Max sustained transfer rate out of FTS to tape-backed dCache (from monitor): 1Gb link
 - Max sustained files per hour: 4K
- Needs
 - Data files: ifdh grid node to scratch dCache, mv'ed into FTS
 - TDR Sample was 1M files, CD3 will be order 4M
 - = 1.1 k files per hour
 - = 0.2Gb/s
- Conclusions
 - Only a margin of x2 in the FTS, but it should average well
 - Can always open up another FTS (~1-2 weeks)

People

- Contacts, organizers
 - Andrei, Yuri and Ralf for job definition and output validation
 - Andrei for workflows
 - RLC (operator liaison) for operations
- Operators
 - Excellent performance in TDR sample upload!
 - ~6h training for TDR job, CD3 should be simpler
 - They do not work weekends and have competing demands
- Requirements
 - Submit and receive 10K jobs/day – 0.5 FTE from operators
 - Need to submit from mu2epro account – use jobsub_client with service cert – should be OK

Summary

- We believe the CD3 challenge will be met!
- Potential risks
 - OSG commissioning and load testing - RLC
 - Could use a little more bandwidth on FTS, but contingency is available
 - Workflow must process job output $\ll 8s$
- Requests
 - ~0.5 FTE from operators, with high priority
 - OSG commissioning and maintenance with high priority