

Offline Infrastructure and Data Processing

NOvA Readiness Review 2014

Craig Group

Overview

- The NOvA data set production paradigm has shifted over the last year to a more scalable file handling solution.
(*SAM*: Sequential data Access via Meta-data.)
- Data processing is going well -- recent addition of automated “keep up” processing for raw to root, calibration, and some reconstruction datasets.
- Support from the Computing Division has been excellent and data set production steps are performing well.
- New offline operations service from CD helping with job management tasks.
- The production group has continuity in the transition from the project to the collaboration era – this has always been a collaboration effort, and we have been producing production data sets for several years.

Demand is Large.

- More than 1 PB of NOvA files already written to tape -- More than 5M files.
- ~5,000 raw data files per day
- > 10M CPU hours used over the last year
- Plan to reprocess all data and generate new simulation ~2 times per year.
(We call this a production run)

The Computing Team is Strong

NOvA and CD Working closely together to improve data handling tools

NOvA

Computing Coordinator: Group

- Production Group: Tamsett
- Databases: Paley
- Code: J. Davies
- Other members:
 - G. Davies
 - Mayer
 - Backhouse
 - Rocco
 - Pandey
 - Sachdev

Computing Division

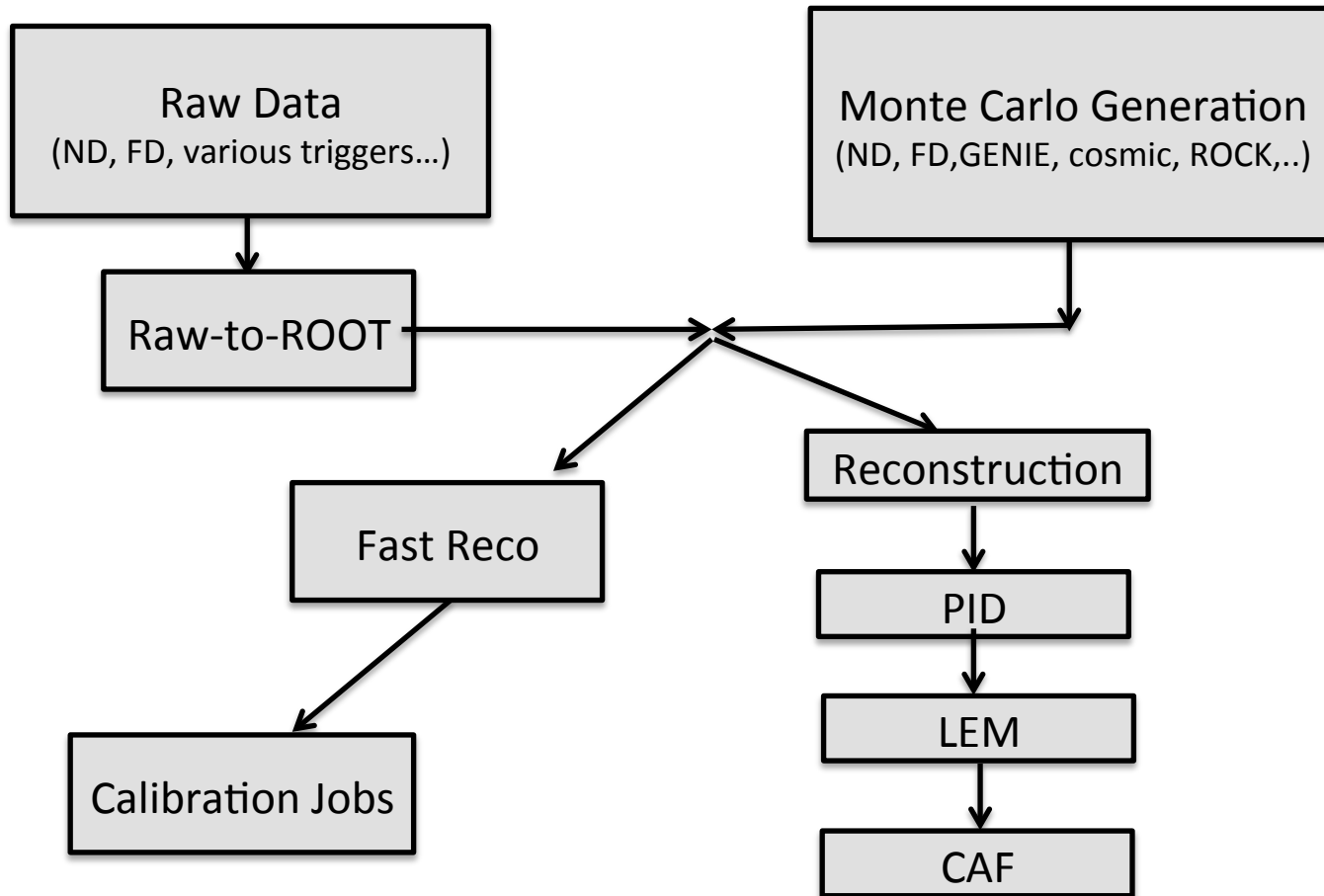
NOvA Liason: Norman

- Data production: Bitrigo, Sierra
- Databases: Mandrichenko
- SAM: Illingworth
- Other members:
 - Gheith
 - Bitrigo
 - Sierra
 - Mengel
 - Litvinse

Summary of Current Infrastructure

- VM
 - 10 virtual machines: novagpvm01 –novagpvm10)
- Blue Arc:
 - Interactive data storage for short term or small data sets
 - /nova/data (140 T),/nova/prod (100 T), /nova/ana (95 T)
- Tape:
 - Long term data storage
 - Files registered with SAM
 - 4 PB of cache disk available for IF experiments
 - File Transfer Service (FTS)
- Batch:
 - Local batch cluster: ~40 nodes
 - Grid slots at Fermilab for NOvA: 1300 node quota
(opportunistic slots also available)
 - Remote batch slots: thousands of additional slots
- Databases: Several, required for online and offline operations

Production Flow



Fall 2013 Workshop

				CUMULATIVE			PER TRIGGER		
	Exposure (<i>p.o.t.</i>)	triggers (<i>triggers</i>)	iteration per trigger	Tape (<i>TB</i>)	Disk (<i>TB</i>)	Time <i>kCPU-days</i>	Tape (<i>MB</i>)	Disk (<i>MB</i>)	Time (<i>CPU-sec</i>)
MC FD beam	2.5e24	8.3E+06	1	31	9	1.0	3.7	1.1	10.4
MC ND beam	1.2e21	2.4E+07	20	82	21	6.2	3.4	0.9	22.3
Data FD beam	-	-	-	-	-	-	-	-	-
Data ND beam	-	-	-	-	-	-	-	-	-
	(<i>seconds</i>)								
MC FD cosmics	2000	4.0E+06	50	50	14	0.4	12.5	3.5	8.6
MC ND cosmics	-	-	-	-	-	-	-	-	-
Data FD cosmics	10000	2.0E+07	50	79	26	2.1	4.0	1.3	9.1
Data ND cosmics	-	-	-	-	-	-	-	-	-
Totals				242	70	9.7	23.6	6.8	50.4

- Production goals:
 - The footprint for final output of a production run should be less than 100TB.
 - The production run should be possible to complete in a two week period.
- There was also a major effort to to understand resources and to streamline production tools . (Caveat – still validating these numbers...)
- Summarized in DocDB 10129
- Workshop: <http://nova-docdb.fnal.gov:8080/cgi-bin/DisplayMeeting?conferenceid=1281>

Fall 2013 Workshop

				CUMULATIVE			PER TRIGGER		
	Exposure (<i>p.o.t.</i>)	triggers (<i>triggers</i>)	iteration per trigger	Tape (<i>TB</i>)	Disk (<i>TB</i>)	Time <i>kCPU-days</i>	Tape (<i>MB</i>)	Disk (<i>MB</i>)	Time (<i>CPU-sec</i>)
MC FD beam	2.5e24	8.3E+06	1	31	9	1.0	3.7	1.1	10.4
MC ND beam	1.2e21	2.4E+07	20	82	21	6.2	3.4	0.9	22.3
Data FD beam	-	-	-	-	-	-	-	-	-
Data ND beam	-	-	-	-	-	-	-	-	-
	<i>(seconds)</i>								
MC FD cosmics	2000	4.0E+06	50	50	14	0.4	12.5	3.5	8.6
MC ND cosmics	-	-	-	-	-	-	-	-	-
Data FD cosmics	10000	2.0E+07	50	79	26	2.1	4.0	1.3	9.1
Data ND cosmics	-	-	-	-	-	-	-	-	-
Totals				242	70	9.7	23.6	6.8	50.4

- Production goals: **MC ND Beam drives CPU usage.**
 - The footprint for final output of a production run should be less than 100TB.
 - The production run should be possible to complete in a two week period.
- There was also a major effort to to understand resources and to streamline production tools . (Caveat – still validating these numbers...)
- Summarized in DocDB 10129
- Workshop: <http://nova-docdb.fnal.gov:8080/cgi-bin/DisplayMeeting?conferenceid=1281>

Fall 2013 Workshop

				CUMULATIVE			PER TRIGGER		
	Exposure (<i>p.o.t.</i>)	triggers (<i>triggers</i>)	iteration per trigger	Tape (<i>TB</i>)	Disk (<i>TB</i>)	Time <i>kCPU-days</i>	Tape (<i>MB</i>)	Disk (<i>MB</i>)	Time (<i>CPU-sec</i>)
MC FD beam	2.5e24	8.3E+06	1	31	9	1.0	3.7	1.1	10.4
MC ND beam	1.2e21	2.4E+07	20	82	21	6.2	3.4	0.9	22.3
Data FD beam	-	-	-	-	-	-	-	-	-
Data ND beam	-	-	-	-	-	-	-	-	-
	(<i>seconds</i>)								
MC FD cosmics	2000	4.0E+06	50	50	14	0.4	12.5	3.5	8.6
MC ND cosmics	-	-	-	-	-	-	-	-	-
Data FD cosmics	10000	2.0E+07	50	79	26	2.1	4.0	1.3	9.1
Data ND cosmics	-	-	-	-	-	-	-	-	-
Totals				242	70	9.7	23.6	6.8	50.4

- Production goals: **Almost 1 TB/hr !** (250 TB = IF+cosmics for full month)
 - The footprint for final output of a production run should be less than 100TB.
 - The production run should be possible to complete in a two week period.
- There was also a major effort to to understand resources and to streamline production tools . (Caveat – still validating these numbers...)
- Summarized in DocDB 10129
- Workshop: <http://nova-docdb.fnal.gov:8080/cgi-bin/DisplayMeeting?conferenceid=1281>

Fall 2013 Workshop

				CUMULATIVE			PER TRIGGER		
	Exposure (<i>p.o.t.</i>)	triggers (<i>triggers</i>)	iteration per trigger	Tape (<i>TB</i>)	Disk (<i>TB</i>)	Time <i>kCPU-days</i>	Tape (<i>MB</i>)	Disk (<i>MB</i>)	Time (<i>CPU-sec</i>)
MC FD beam	2.5e24	8.3E+06	1	31	9	1.0	3.7	1.1	10.4
MC ND beam	1.2e21	2.4E+07	20	82	21	6.2	3.4	0.9	22.3
Data FD beam	-	-	-	-	-	-	-	-	-
Data ND beam	-	-	-	-	-	-	-	-	-
	(<i>seconds</i>)								
MC FD cosmics	2000	4.0E+06	50	50	14	0.4	12.5	3.5	8.6
MC ND cosmics	-	-	-	-	-	-	-	-	-
Data FD cosmics	10000	2.0E+07	50	79	26	2.1	4.0	1.3	9.1
Data ND cosmics	-	-	-	-	-	-	-	-	-
Totals				242	70	9.7	23.6	6.8	50.4

About 1000 CPUs DC !

- Production goals:
 - The footprint for final output of a production run should be less than 100TB.
 - The production run should be possible to complete in a two week period.
- There was also a major effort to to understand resources and to streamline production tools . (Caveat – still validating these numbers...)
- Summarized in DocDB 10129
- Workshop: <http://nova-docdb.fnal.gov:8080/cgi-bin/DisplayMeeting?conferenceid=1281>

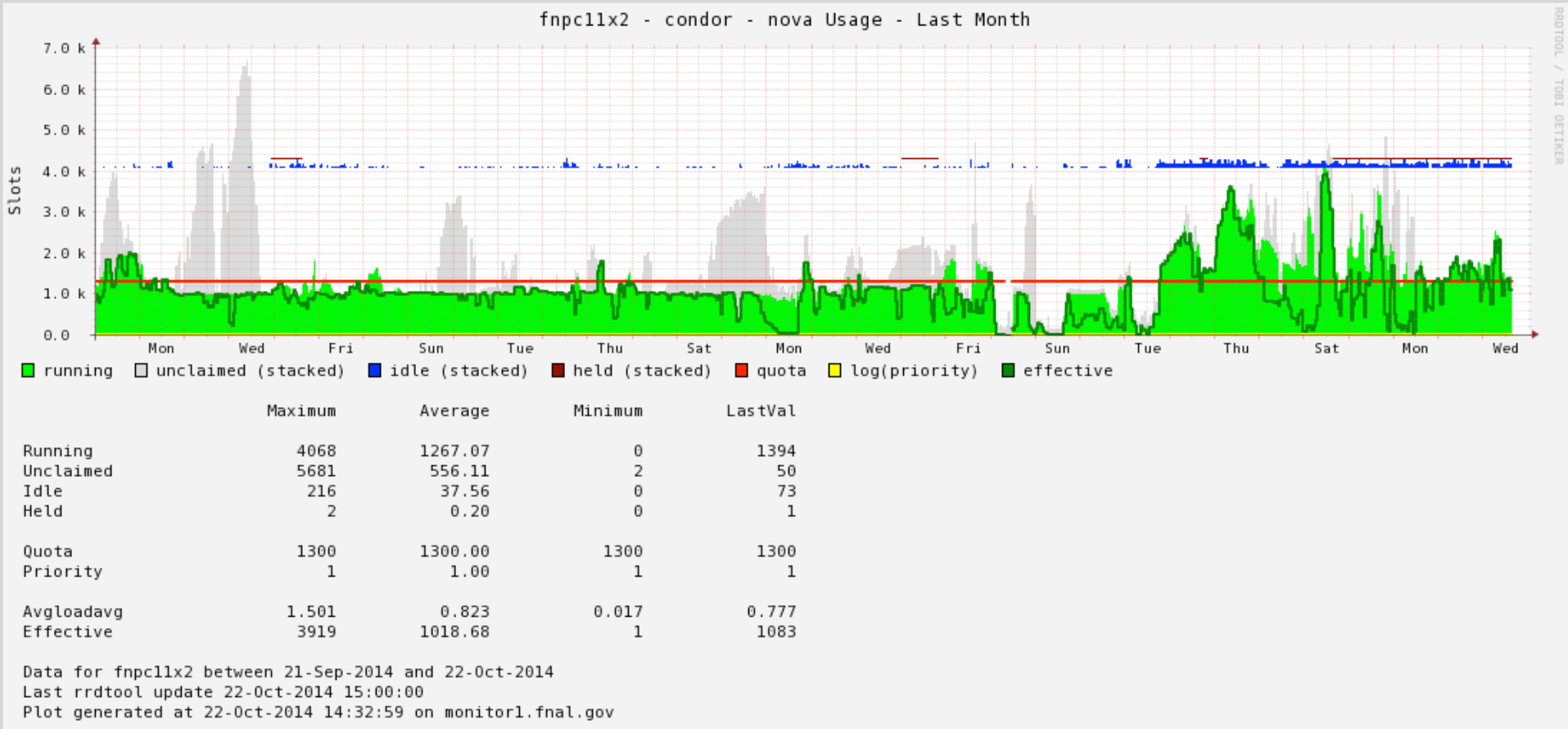
Fall 2013 Workshop

				CUMULATIVE			PER TRIGGER		
	Exposure (<i>p.o.t.</i>)	triggers (<i>triggers</i>)	iteration per trigger	Tape (<i>TB</i>)	Disk (<i>TB</i>)	Time <i>kCPU-days</i>	Tape (<i>MB</i>)	Disk (<i>MB</i>)	Time (<i>CPU-sec</i>)
MC FD beam	2.5e24	8.3E+06	1	31	9	1.0	3.7	1.1	10.4
MC ND beam	1.2e21	2.4E+07	20	82	21	6.2	3.4	0.9	22.3
Data FD beam	-	-	-	-	-	-	-	-	-
Data ND beam	-	-	-	-	-	-	-	-	-
	(<i>seconds</i>)								
MC FD cosmics	2000	4.0E+06							8.6
MC ND cosmics	-								-
Data FD cosmics	10000								9.1
Data ND cosmics									-
Totals							6.8		50.4

These were estimates.
Most validated in recent file
production runs.

- Production
 - The ...
 - The ... more than 100TB.
 - The ... two week period.
- There ... understand resources and to streamline production tools. (caveat = still validating these numbers...)
- Summarized in DocDB 10129
- Workshop: <http://nova-docdb.fnal.gov:8080/cgi-bin/DisplayMeeting?conferenceid=1281>

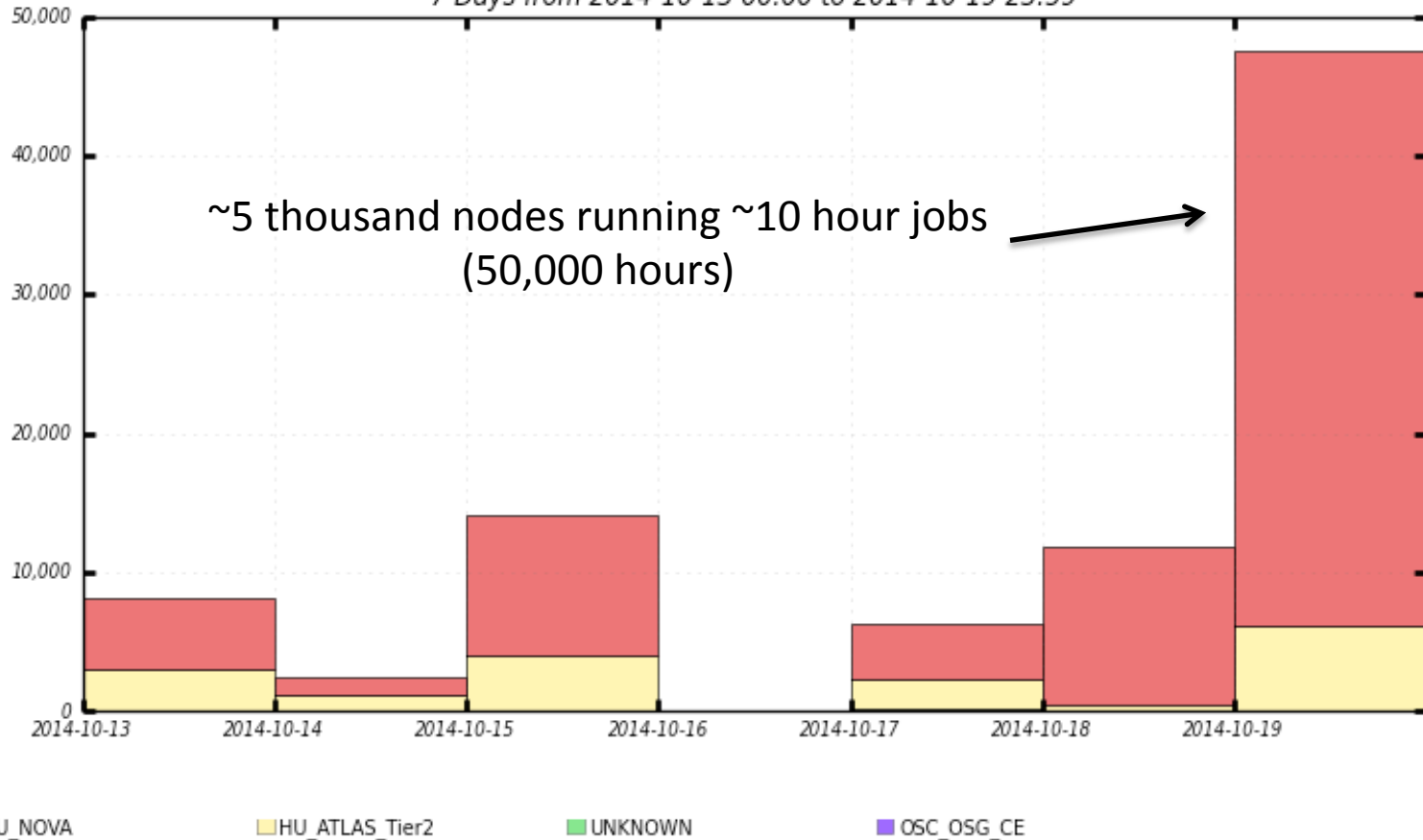
CPU (on site)



CPU is not currently a limiting factor.

CPU (off site)

NOvA WMS Hours Spent on Job By the OSG Facility
7 Days from 2014-10-13 00:00 to 2014-10-19 23:59

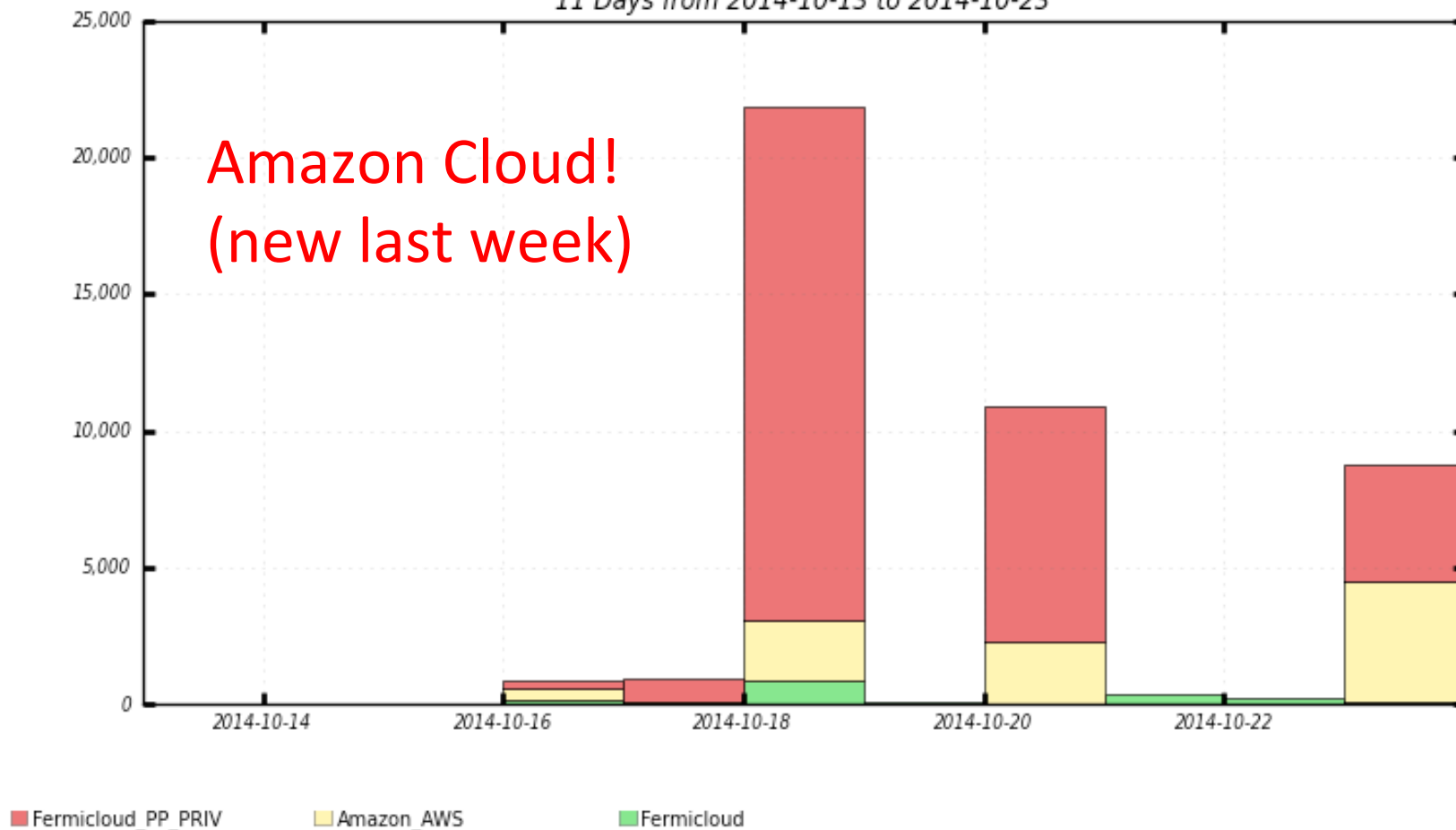


Maximum: 47,605 , Minimum: 6.24 , Average: 12,915 , Current: 47,605

Thousands of offsite CPU slots are also available to us.

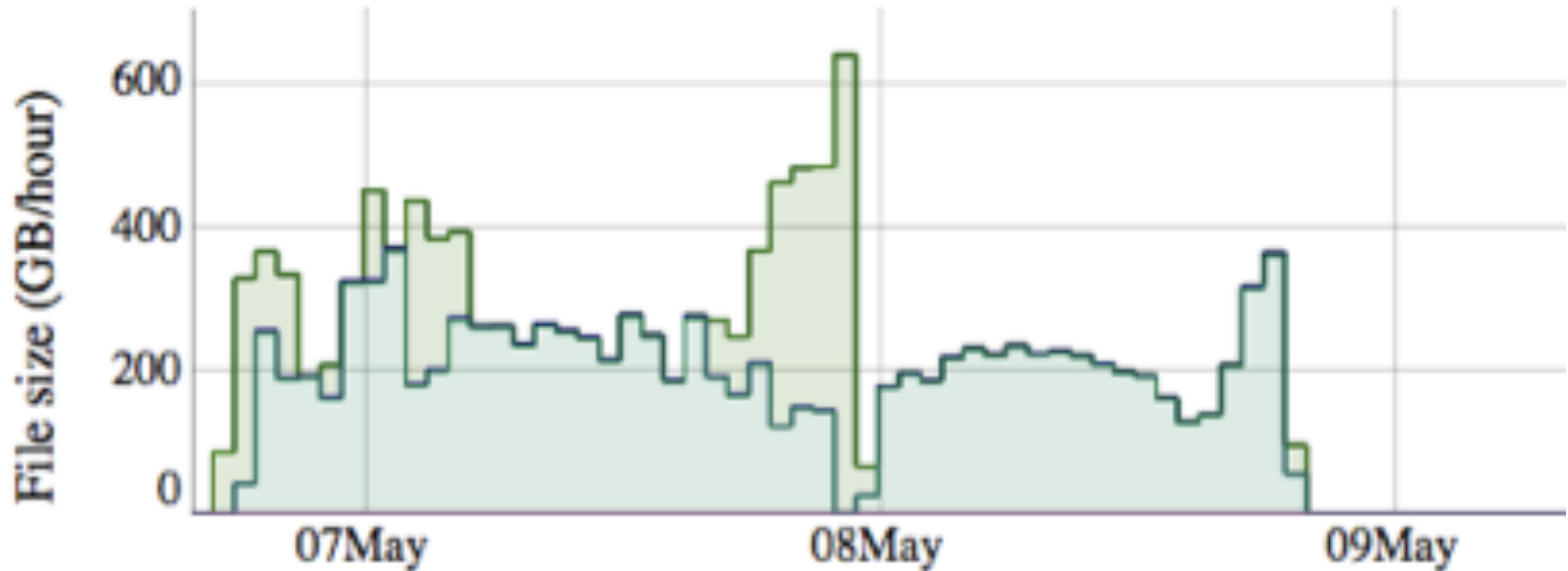
CPU (cloud)

NOvA WMS Hours Spent on Job By Cloud
11 Days from 2014-10-13 to 2014-10-23



Maximum: 21,873 , Minimum: 0.00 , Average: 3,997 , Current: 8,738

File Throughput



- Example from novasamgpvm02: sustained at >200 GB/hr)
- We have three FTS servers
- Excess of 1TB/hour total has been demonstrated

File Production: Spring 2014

- Spring 2014 production was a first in many respects:
 - First production run fully based on SAM datasets
 - First effort with a substantial FD data set
 - First effort since code was streamlined and footprint was reduced in the fall 2013 production workshop
- The SAM transition was far from smooth and we had ups and downs.
- In the end we ran all steps of production in time for Neutrino 2014 (some steps multiple times).

File Production: Fall 2014

- This is the data set production effort for first physics results.
- Many first-time requests: new keep-up data sets, calibration requests, systematic samples...
- The SAM paradigm is functioning well.

Validation of File Production Tools

- New tool available to check all data processing steps for every new software release.
- Reports any failure of a file production step.
- Metrics of each step compared between new and past releases:
 - Output file sizes
 - Memory Usage
 - CPU usage
- All info published to the web
- Easy to check for major changes in file production chain.

Validation of File Production Tools









FA14-09-23 10:13:54 23/09/2014

The [projections section](#) interpretes any results displayed here.







Test parameters

- **Time:** 2014-09-23 10:14:28
- **Release:** FA14-09-23
- **Message:** Full test of FA14-09-23.

Chains

FD_data_cosmics:raw2root:	cmd, log, out, err	 
FD_data_NuMI:raw2root:	cmd, log, out, err	
FD_cosmics	cmd, log, out, err	 
FD_genie_FHC_nonswap	cmd, log, out, err	
FD_genie_FHC_swap	cmd, log, out, err	
FD_genie_FHC_tau	cmd, log, out, err	
FD_genie_RHC_nonswap	cmd, log, out, err	 
FD_genie_RHC_swap	cmd, log, out, err	 

Batch job status keys

-  job ended successfully
-  job ongoing
-  STDERR not empty
-  job was killed by batch robots
-  error in run tier
-  no pkl file for a completed chain

Validation of File Production Tools










FA14-09-23 10:13:54 23/09/2014

The [projections section](#) interpretes any results displayed here.







Test parameters

- **Time:** 2014-09-23 10:14:28
- **Release:** FA14-09-23
- **Message:** Full test of FA14-09-23.

Chains

FD_data_cosmics:raw2root:	cmd, log, out, err	 
FD_data_NuMl:raw2root:	cmd, log, out, err	
FD_cosmics	cmd, log, out, err	 
FD_genie_FHC_nonswap	cmd, log, out, err	
FD_genie_FHC_swap	cmd, log, out, err	
FD_genie_FHC_tau	cmd, log, out, err	
FD_genie_RHC_nonswap	cmd, log, out, err	 
FD_genie_RHC_swap	cmd, log, out, err	 

Batch job status keys

-  job ended successfully
-  job ongoing
-  STDERR not empty
-  job was killed by batch robots
-  error in run tier
-  no pkl file for a completed chain

Validation of File Production Tools

Production Testing Configurations Results Projections FA14-09-23 10:13:54 23/09/2014

FA14-09-23 10:13:54 23/09/2014

The projections section in the test results is displayed here

Production Testing Configurations Results Projections FA14-09-23 10:13:54 23/09/2014

Test p

This test was run using:

- Time: 2014-09-23 14:27:19
- USER: novagli
- HOSTNAME: fnpc3066.fnal.gov
- SRT_BASE_RELEASE: FA14-09-23
- SRT_QUAL: maxopt
- SRT_PUBLIC_CONTEXT: /nova/app/home/novasoft/slf6/novasoft/releases/FA14-09-23
- SRT_PRIVATE_CONTEXT: /local/stage1/disk4/dir_4840/glide_o3w2GR/execute/dir_8777/no_xfer/rel

Tier	In evt	User CPU (/in evt) [s]	Memory	DB queries	Query time [s]	Child	Events (efficiency)	Size (/out evt)
cry log, fcl, metrics	200	11935.56 (59.68)	626.31 MB	0	0.00	osmics_gen.root	200 (100 [%])	562.76 MB (2.81 MB)
pchits log, fcl, metrics	200	370.11 (1.85)	443.5 MB	6	0.17	clist_reco.root	200 (100 [%])	55.84 MB (285.92 KB)
"	"	"	"	"	"	tstop_reco.root	194 (97 [%])	994.53 KB (5.13 KB)
attenprof log, fcl, metrics	200	9.79 (0.05)	389.25 MB	0	0.00	.attenprof.root	1 (0 [%])	1.29 MB (1.29 MB)

Recent Progress

- Now taking advantage of offsite CPU resources.
- Move to SAM for data set management and delivery to reduce the dependence on local disk.
- Database performance has been recently stable.
- Demonstrated production framework, and measure/document resource requirements.
- New production validation framework is very useful.
- On going: producing a full set of production files for analysis groups and first physics.

Summary

- Offline computing isn't significantly affected by the transition from project to operations.
- There has been a recent transition to a more scalable file handling system similar to what was employed by CDF and D0.
- We have the resources we need and CD is working closely with us to solve issues as they arrive.
- Computing resources are sufficient and we are ready to serve the data sets required by the collaboration for physics.

Extra slides follow...

Large scale computing activities:

- Perform production data processing and MC generation (matched to detector config) about 2 times per year
 - Major productions scheduled in advance of NOvA collaboration meetings and/or Summer/Winter conferences
- Full Reprocessing of raw data 1 time per year (current data volume 500+ TB)
- Need to store raw and processed data, calibration data sets from far and near detector.
- Need to store Monte Carlo sets corresponding to the near & far detectors matched to current production
- Need to store data sets processed with multiple versions of reconstruction.
- Need 2000+ slots dedicated to production efforts during prod/reprocessing peak to complete simulation/analysis chains within a few weeks.

Database Overview

Summary of NOvA Databases				
Database Name	Schema(s)	Hostname	Port	Comment
nova_hardware	public, module_factory, ashriverprod_factory	ifdbprod	5432	FNAL Hardware DB and copies of the UMN and Ash River DBs
nova_prod	fardet	novadaq-far-db-03	5432	FarDet DAQ DB
nova_prod	ndos, neardet	novadaq-ctrl-db-01	5432	NDOS DAQ/DCS DB NearDet DAQ DB
nova_prod	fardet	novadcs-far-logger	5432	FarDet DCS DB
nova_prod	fardet, ndos, neardet	ifdbprod	5433	Offline DB for FarDet tables. Contains replicated DAQ and DCS tables
nova_dev	fardet, ndos, fccdaq, neardet	ifdbdev	5433	Offline and Online sandbox DBs.

DAQ/DCS databases are critical for online operations: 24/7 support
(Database executive summary available in DocDB 10602)

Offsite Resources

- CVMFS makes NOvA code releases available offsite.
- We can currently run NOvA art jobs at many offsite farms.
 - SMU, OSC, Harvard, Nebraska, San Diego, Indiana and U. Chicago
 - Prague is about to come online.
 - Even a recent successful test run with some jobs on Amazon Cloud
- Jobs can access files using SAM and write output to FNAL

SAM transition

- Our detector and MC data is more than can be managed with local disk (Bluearc).
- Solution: Use SAM for data set management interfaced with tape and large CACHE disc.
- Each file declared to SAM must have metadata associated with it that can be used to define datasets.
- SAM alleviates the need to store large datasets on local disk storage and helps ensure that all data sets are archived to tape.

Successes

- Running MC offsite works.
- File handling with SAM works well for production jobs.
- Recent SAM tutorial was well received and users seem to be having positive experience with SAM.
- Production tools are more robust and scalable.

What drives resource requirements?

- CPU – ND Beam simulation
- Disk:
 - FD Raw data – large calibration sample required
 - Many stages of processing each produce data copies (important for intermediate validation steps)

Production: CPU Requirements

CPU Requirements: ND Event MC dominates ~60% of production

- Driven by generation speed: Order 10 seconds per event
- Driven by quantity of events (MC to data ratio)
 - ND crucial for: tuning simulation, evaluating efficiencies, estimating background rates, and controlling systematics.
 - Minimal ND data set for first NOvA analyses is $1e20$ protons-on-target (2 Months of ND data)
 - MC samples need to be a few times larger than this to keep their statistical uncertainties from playing a significant role
 - Additionally, both nominal and systematically varied samples are needed.
 - So, our estimate is based on $1.2e21$ p.o.t.
- 2014/2015 estimates based on 3 production runs:
 - **1 M CPU hours** (.35 M per production run)
 - This manageable with our current grid quota and offsite resources.

(Note: This only includes production efforts (no analysis, calibration, ...)

FD Data rate

As an upper limit consider the current data transfer limit from Ash River to Fermilab of 60 MB/s.

- This is about 10% of FD data.
- 5 TB / day (seems possible data rate to transfer to tape)
- 1.8 PB/year (Full set of Tevatron datasets ~ 20 PB)
- Only Raw data – gain about 4x from full production steps
- Could be 10 PB/year, but we won't process all of that.
- Assuming 100 us for beam spill, <0.07 MB/s
- Cosmic Pulsar, < 4 MB/s (currently ~2% of live time)
- Calibration and other triggers (DDT) fill in ~ 50 MB/s.

GOAL: Tape storage should not limit the physics potential of the experiment!