NOvA Computing

Gavin S. Davies

Iowa State University

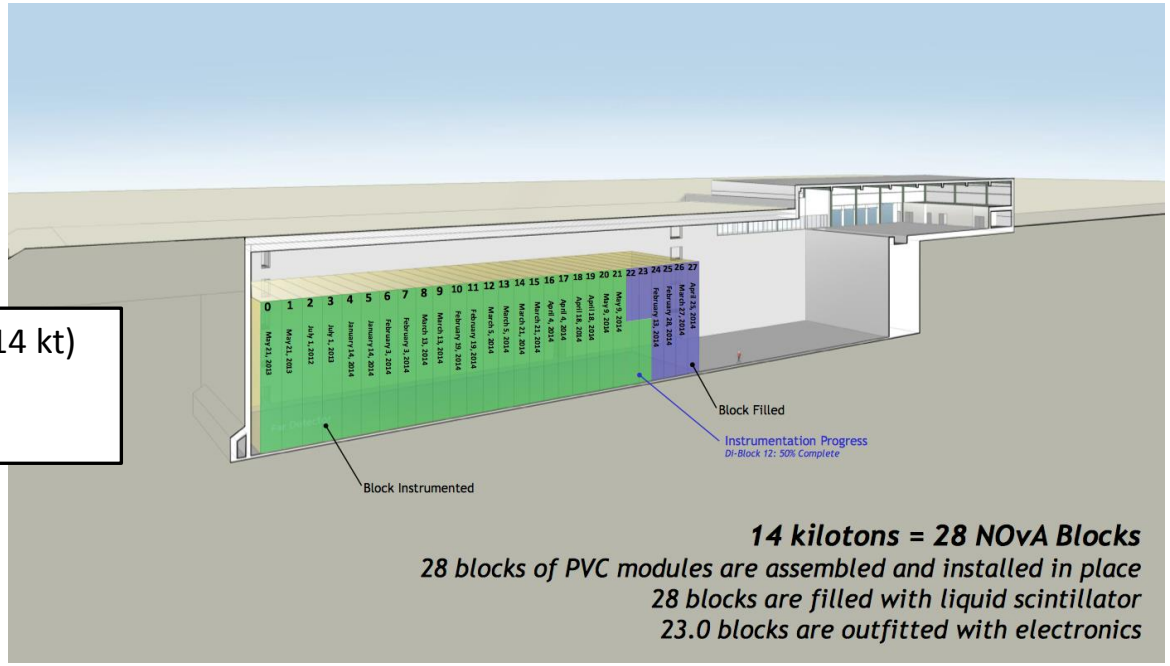
On behalf of the NOvA Collaboration

FIFE Workshop, June 16-17th 2014

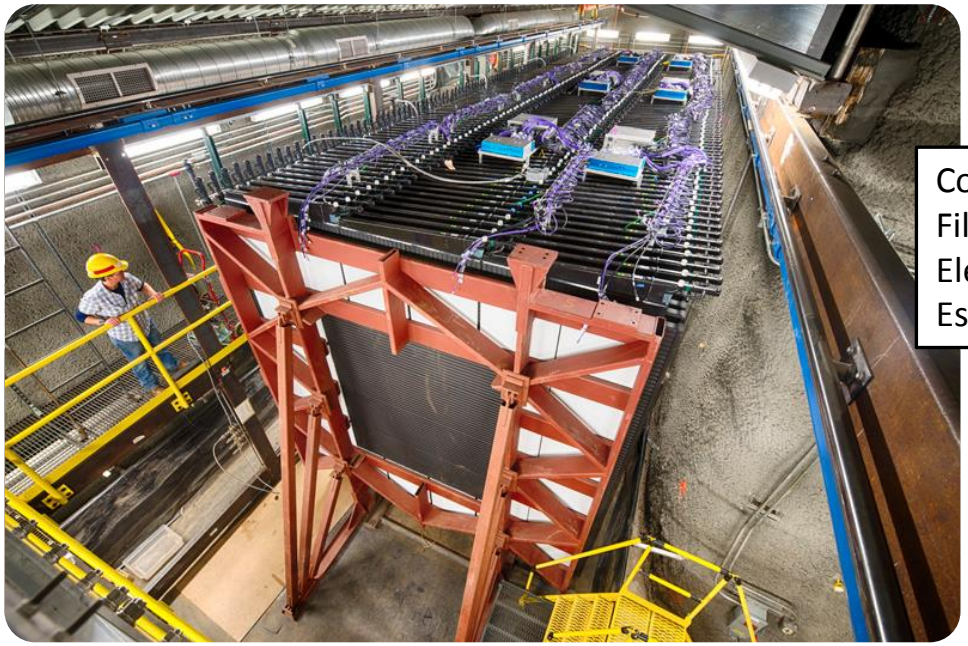
NOvA Progress

NOvA Far Detector

Construction Completed: Apr 25, 2014 (14 kt)
Electronics: 80% complete (11.25 kt)
Estimated Completion: July 2014



14 kilotons = 28 NOvA Blocks
28 blocks of PVC modules are assembled and installed in place
28 blocks are filled with liquid scintillator
23.0 blocks are outfitted with electronics



NOvA Near Detector

Construction Completed: Apr 25, 2014
Filling Status: 100%
Electronics: 16% complete
Estimated Completion: July 2014

NOvA recording data from two detectors and gearing up towards the First Physics Analysis Results late 2014 thus increasing our Scientific Computing needs going forward

Available Resources I

- **Virtual Machines** – User FNAL Computing gateway
  - 10 virtual machines: novagpvm01 –novagpvm10
  - Round-robin access through: "*ssh nova-offline.fnal.gov*"
- **BlueArc** - Interactive data storage

/nova/data (140 T), /nova/prod (100 T), /nova/ana (95 T)

- **Tape** - Long term data storage
  - 4 PB of cache disk available for IF experiments
- **Batch** – Data processing:
  - Local batch cluster: ~40 nodes
  - Grid slots at Fermilab for NOvA: 1300 nodes
  - Remote batch slots: Generation/simulation ready!
    - Off-site resources via novacfs and OSG oasis cvmfs servers

Available Resources II

ECL (Electronic Collaboration Logbook)

- Two logbooks currently in use for NOvA
  - Control Room - General DAQ and Operations
  - Ash River Construction - Assembly and Outfitting
  - Also utilise ECL as Shift Scheduler and other collaboration tools

Databases

- Online & Offline databases (development and production)
  - Improved monitoring tools requested (performance monitoring)
- Offline Conditions database access via web server
- NOvA Hardware Databases and applications
- IF Beams databases and applications layers (beam spill info)

Offsite Resources

- Off-site resources via novacfs and OSG oasis CVMFS servers
- NOvA can currently run batch jobs at multiple offsite farms
  - **SMU, OSC, Harvard, Prague**, Nebraska, San Diego, Indiana and U.Chicago
  - We use **NOvA-dedicated** sites for GENIE simulation generation
  - Successfully ran a first round of ND cosmics generation with Amazon EC2 (Elastic Cloud Computing) with lots of assistance from FNAL OSG group
    - Amazon spot-price charges – 1000 ND cosmics jobs:
      Cloud ~ $40, Data transfer ~ $27 ~230 GB
- Jobs can access files using SAM and write output to FNAL
- For undirected projects, FermiGrid will consume 75% of jobs
  - We need a job steering site-prioritisation system to maximise use of our dedicated sites
- MC Generation and Reconstruction have been run off-site successfully thus this is a viable use of resources

NOvA Offline Software

- NOvA uses ART as its underlying framework
  - We attend and provide input to weekly ART stakeholders meetings
  - Fast turn around for features required for SAM interfacing (rollout of new ART releases)
- Relocatable ups system (/nusoft/app/externals)
  - External packages
    - ROOT/GEANT4/GENIE etc, ART-dependent (SCD-provided binaries)
  - nutools – IF experiment-common packages
  - novasoft (/grid/fermiapp/nova/novaart/novasvn/)
    - We maintain and is not a ups product
- All distributed for slf5 and slf6 via cvmfs
  - *oasis.opensciencegrid.org* and *novacfs.fnal.gov*
- Development environment based on SRT build system and svn repository
  - Proven to work with cmake build system also

Early Production setup

- Initially we had all Data and Monte Carlo files on BlueArc handled by hand, archived manually and/or deleted
- All simulation, reconstruction and particle identification previously processed onsite
  - With the exception of custom Library Event Matching (LEM) algorithm
  - LEM processed at Caltech
  - Requires large amount of total memory on machine (O(100G)) which breaks condor policy

- Everyone writes at will to all BlueArc disks
  - 100,000s of files
  - Organised by directory tree
  - No user-tools for archive storage and retrieval

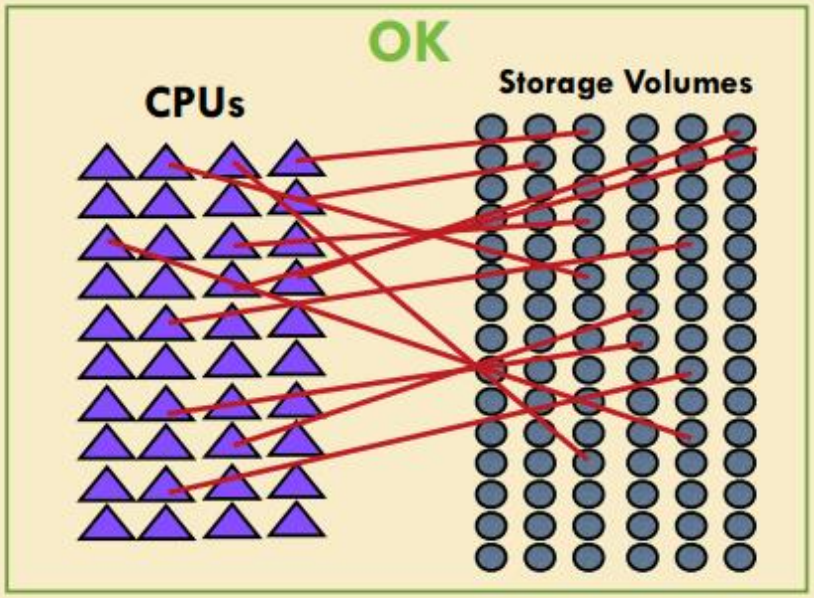
- A short-term solution before the transition to SAM
  - Hence the ongoing development of an updated system best suited for scalability

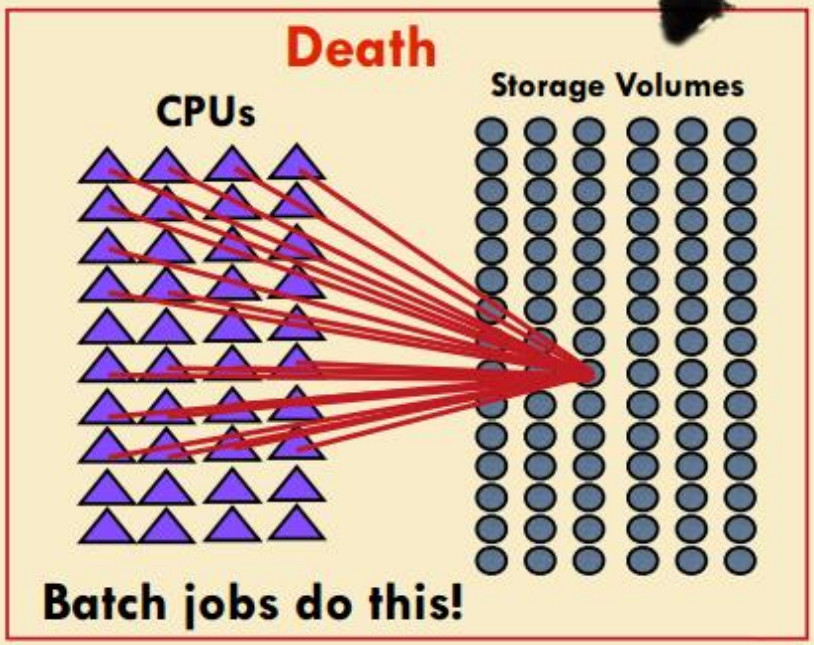
Scale problems begin
e.g. Uncontrolled Bluearc

Good news:
When used as designed, it works great

Bad news:
Can't handle concentrated access

December 19th & March 27th!

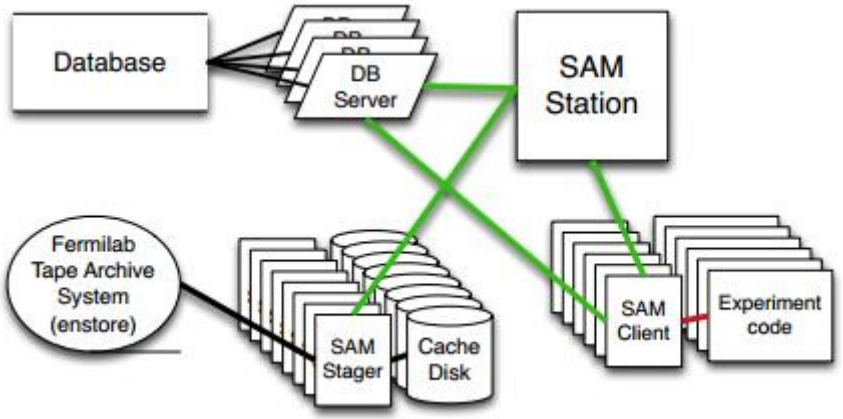




CHEP 2013 - Lyon

14

Transition to SAM



- Our detector and MC data is more than can be managed with BlueArc local disk
- Solution: Use SAM (worked for D0/CDF) for data set management interfaced with tape and large dCache disk
- 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

Transition to SAM: Goals

- Take advantage of off-site resources
- Move to SAM for data set management and delivery to reduce the dependence on local disk usage
  - Store only final stage data files and CAF (Common Analysis Files) on BlueArc
  - ~250 TB on tape, 70 TB on disk
  - Monte Carlo throughput matching estimates

- Streamline database usage
- Develop a scalable and robust production framework
- Define and document work flow for each production task
- Improve reporting of production effort and progress.

- Need a fast turnaround of Data and MC files processing for collaboration in lead up to conferences and collaboration meetings
  - Understanding lead times on each step is critical with first analysis results in mind

- We successfully implemented and demonstrated the above on NOvA!
  - Always room for improvement of course

Scientific Goals for Coming Year

- Goal: First physics results late 2014
  - ν_e appearance analysis 10^{20} POT
  - ν_μ disappearance analysis 10^{20} POT
  - Normalization to near detector
- Aggressive scale & schedule
  - Expect to run/rerun multiple versions of production chain due to immaturity of code base
  - Final phases of analysis currently being tested
  - Need development effort/support for the scale of data processing being processed, simulated, analyzed

Activities to meet goals

- Perform production data processing and MC generation (matched to detector configurations) about 2 times per year
  - 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 400+ TB)
- Need to store raw physics and calibration data sets from Far Detector as well as processed data from Far 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

System Status I

System	Tool	Status
Tape Storage	Enstore/dCache	Fully integrated with online/offline. Approaching 1 PB to tape
Framework	ART	Fully integrated with offline, simulation and triggering
Data Handling w/ Framework	SAM/IFDH-art	Fully integrated with offline. Used for project accounting, file delivery and copy back
Redesign of offline databases	SCD conditions DBI and art services	In use by experiment. Scaling issues.
General Data handling and file delivery	SAMweb/IFDH/ dCache/xrootd	Fully integrated with production activities. Started adoption by analysis users
Grid Resources	Fermigrid, OSG	Fully integrated with production activities. OSG onboard. Additional dedicated sites available through OSG interfaces

System Status II

System	Tool	Status
Central Storage	Bluearc	• Exhausted capacity. • Does not scale under production load. • Wide spread outages. • Maintenance/cleanup nightmare
Databases (performance)	Conditions, Hardware, IFBeams	• Concurrency limits • Query response times • Caching issues (squids) • Monitoring tools
Job submission & Workflows with SAM	Jobsub/IFDH/SAM	• Complexity of workflows, wrappers, etc… • Not fully implemented for all workflows • Not fully implemented for final analysis layers
File staging (copy back)	dCache & Bluearc Fermi-FTS	• Performance bottleneck to entire production system • Access protocols available to validate files buggy

Production (FY14/FY15)

- **Goal**: ***Updated physics results ~2 times per year***
  - Production with current software over full data set + corresponding Monte Carlo.
  - Dominated by MC generation/simulation and off-beam (zero-bias) calibration data processing
  - Average *beam* event data is tiny (calibration data is 100x beam)
    - Near Det Size: ~6.7 KB
    - Far Det Size: ~247 KB (7.4 TB/yr)
  - Average Monte Carlo Event size
    - Near Det: 1 MB
    - Average generation time: 10 s/event
- **Estimated at ~ 1 M CPU hours / year**

Motivated NOvA to move a large fraction of Monte Carlo production to offsite facilities (collaborating institutions, OSG and even Amazon Cloud)

Production (FY14/FY15)

- *Resources we have requested are designed to complete NOvA's move to offsite MC generation and to ensure the success of the transition to large scale data handling and job management (w/ SAM)*
- **Storage Services**
  - Central Disk (BlueArc) production area is sized to accommodate production + data handling infrastructure
    - Retain ~300 TB of space
    - Will recover and reallocate space currently used by production for use as "project disks" for analysis
    - Utilize available space for final CAF ntuples and custom skims
  - Additional (non-tape backed) disk storage (dCache)
    - 100-200 TB Required for production activities. (current non-volatile 266 TB shared)
    - Temporary storage used for intermediate files, data validation, registration to SAM data catalog.
  - Tape storage will be significant for the coming year (~1.5 PB)

Production (FY14/FY15)

CVMFS (Production service)

- Required for distribution of code base, libs etc.. @ FNAL
- Enables offsite running and OSG integration
  - OSG provided Oasis servers requires additional engineering to meet NOvA requirements

Batch

- Require support and development of improved monitoring tools
  - Expansion of the FIFEmon suite to include user specified reporting
- Require support of IF-JOBSUB suite and development effort for requested features to support offsite computing resources
- Continued effort and development for integrating off-site MC generation with FNAL grid and OSG infrastructure
  - SMU, Ohio Supercomputing, Harvard and many other sites already demonstrated
- Recently generated MC on Amazon Cloud

Production (FY14/FY15)

Data handling

- Support for the SAM data management system is essential for NOvA production and analysis
- Support for the IF Data Handling (IFDH) and IF File Transfer Service (IF-FTS) are essential for production and analysis
- Continued support/development resources for further integration of SAM, IFDH and IF-FTS with the NOvA computing suite are requested

Database

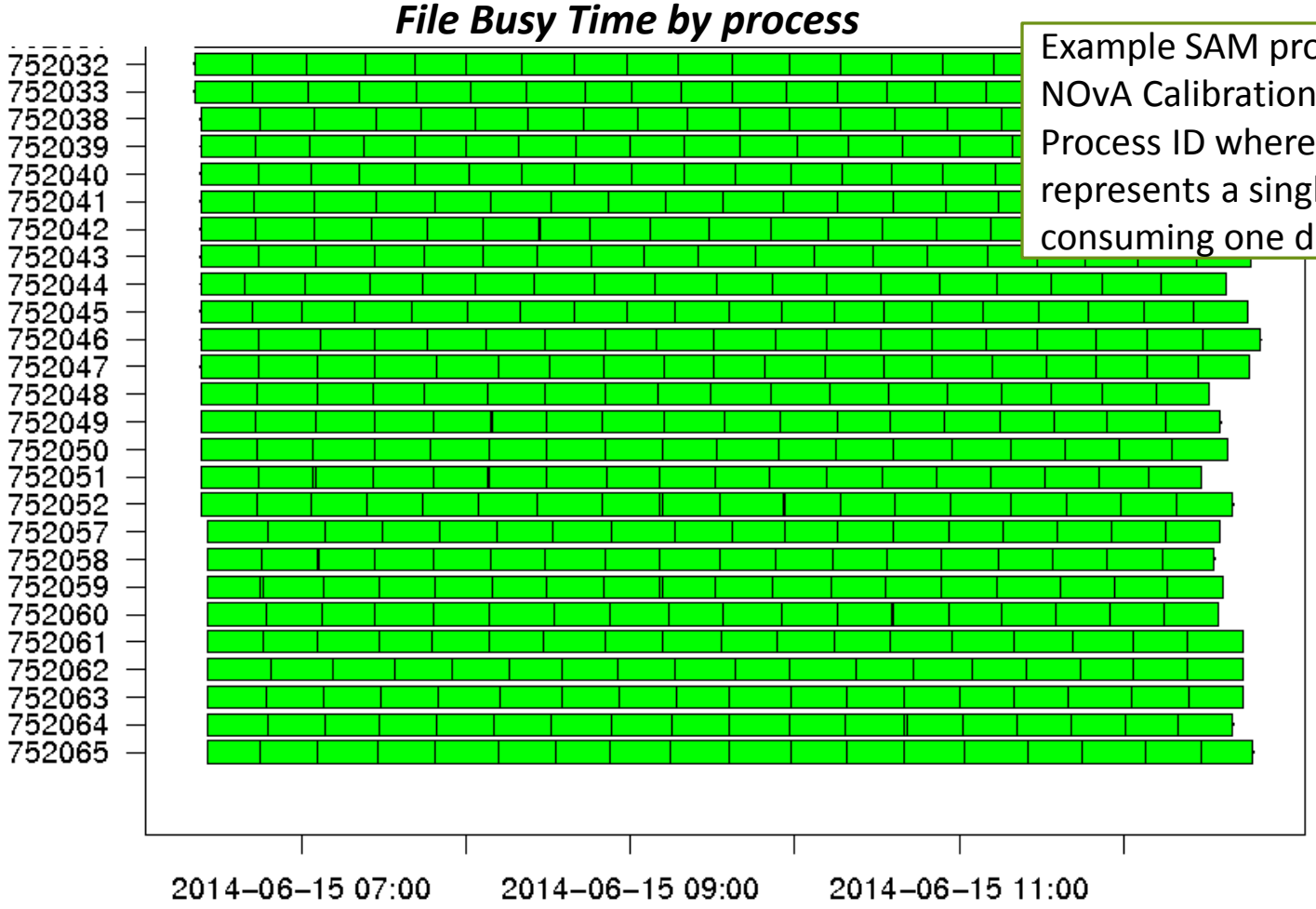
- Support of Database Interface (DBI) capable of handling NOvA peak usage and scaling
- Support for Development and integration of database resources with offsite Monte Carlo and production running
- Additional server infrastructure and improved monitoring may be required.

Development/consulting effort

- NOvA is in the early stages of the ramp up to full physics production
- Support for the batch system and consulting effort to further integrate the batch computing model with the NOvA software desired
- Dedicated SCD or PPD effort to manage production tasks are highly desired!

Recent Successes

- Feedback from the collaboration of the utility of SAM is very positive
- Scalability is proven in IF experiment environment
- Successfully processed required full Data and MC datasets in time for Neutrino 2014



Example SAM project processing NOvA Calibration data: Time vs Process ID where each green box represents a single "nova" job consuming one data file

Summary

- NOvA Production group provided with a large amount of coordination, cooperation and support from the SCD Data Handling group in transition to SAM
  - In lead up to Neutrino 2014 we fully utilised the SAM system for production purposes with great success and fast turnaround
- Off-site resources performing well – CPU is not a limiting factor
  - Site prioritisation schema requested
  - Request improved feedback for completion of pushing software updates
- NOvA SAM Tutorial to help educate the collaboration was successful
  - Collaborators utilising the SAM system with limited cases of difficulty
- Crucial efforts to streamline NOvA require future SCD personnel
  - FTS + SAM/ART streamlining
  - Database development/support
  - Offsite MC
- NOvA look forward to understanding the future role of FIFE in production job submission/monitoring

Backup

Production Overview

	Exposure		CUMULATIVE			PER TRIGGER		
	(p.o.t.)	*(triggers)*	Tape *(TB)*	Disk *(TB)*	Time *kCPU-days*	Tape *(MB)*	Disk *(MB)*	Time *(CPU-sec)*
MC FD beam	2.5e24	8.3E+06	31	9	1.0	3.7	1.1	10.4
MC ND beam	1.2e21	2.4E+07	82	21	6.2	3.4	0.9	22.3
Data FD beam	-	-	-	-	-	-	-	-
Data ND beam	-	-	-	-	-	-	-	-
	(seconds)							
MC FD cosmics	2000	4.0E+06	50	14	3	12.5	3.5	64.8
MC ND cosmics	-		-	-	-	-	-	-
Data FD cosmics	10000	2.0E+07	79	26	5.0	4.0	1.3	21.6
Data ND cosmics	-		-	-	-	-	-	-
Totals			**242**	**70**	**15.2**	**23.6**	**6.8**	**119.1**

- Output from workshop last fall to estimate our needs for production resulting in the following 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 understand resources and to streamline production tools . (Caveat – still validating these numbers from latest production round, initial accounting appears to match estimates)

FY14 Resource Request Summary

Resource	Type	FY14 Addition	Total
Central Disk	Bluearc	0 TB	335 TB
Archival Storage	Enstore tape	1.5 PB	> 2 PB
Cache Disk	DCACHE	0.5 PB	0.5 PB
Batch Processing	Grid Slots	--	1300*
	Local Batch	--	40
Interactive Login	SLF VMs	--	10/10

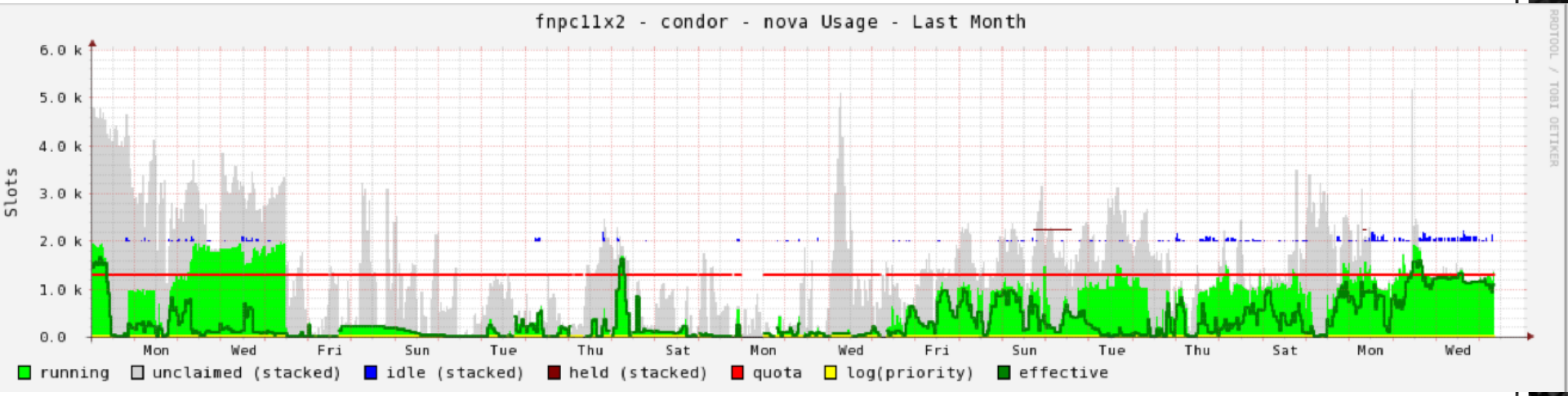
*Does not include opportunistic usage, offsite resources and requested “burst” capabilities

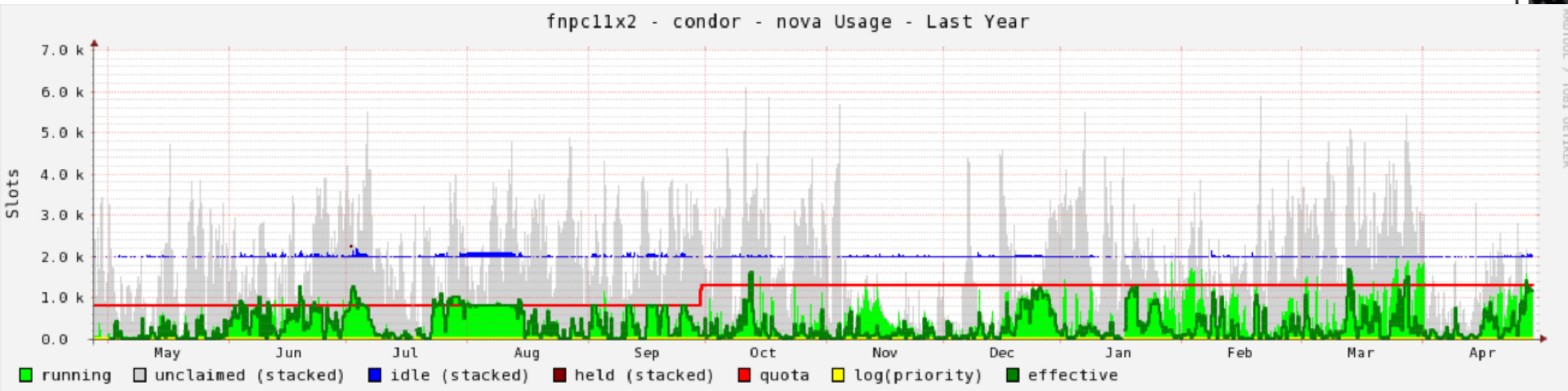
FY15 Resource Request Summary

Resource	Type	FY15 Addition	Total
Central Disk	Bluearc	0 TB	335 TB
Archival Storage	Enstore tape	1.5 PB	>3.5 PB
Cache Disk	DCACHE	0.5	1 PB
Batch Processing	Grid Slots	500*	1800*
	Local Batch	--	40
Interactive Login	SLF VMs	--	10/10

*Assumes that we can demonstrate need for increased quota
Note: Does not include opportunistic usage, offsite resources and requested “burst” capabilities

Fermigrid CPU usage over last month and year.





Database bottleneck

- While integrating the SAM system into our production model, the database was identified as a bottleneck
  - February 7th meeting with NOvA and SCD key personnel
- Several improvements implemented
  - Replica server, more slots in queues, “try n times”
- Also many recent updates to IFBeam DB
  - Where we get our Protons-on-target count
- Our stress tests have proven that >1500 jobs can now run concurrently
  - Areas for improvement identified and implemented

Database Outlook

- Tests of the system are ongoing
- Require more resources, but we need to understand our needs and **specify our requirements**.
- Final assessment of recent production run and new database metrics in our jobs will go a long way to help here
- **We need improved Database monitoring tools**
  - E.g. We appear to have a short-lived memory spike that pushes some jobs over the condor memory threshold (4GB) – debugging is proving very difficult