Recent Evolution of the Offline Computing Model of the NOνA Experiment

Talk #200

Craig Group & Alec Habig

CHEP 2015

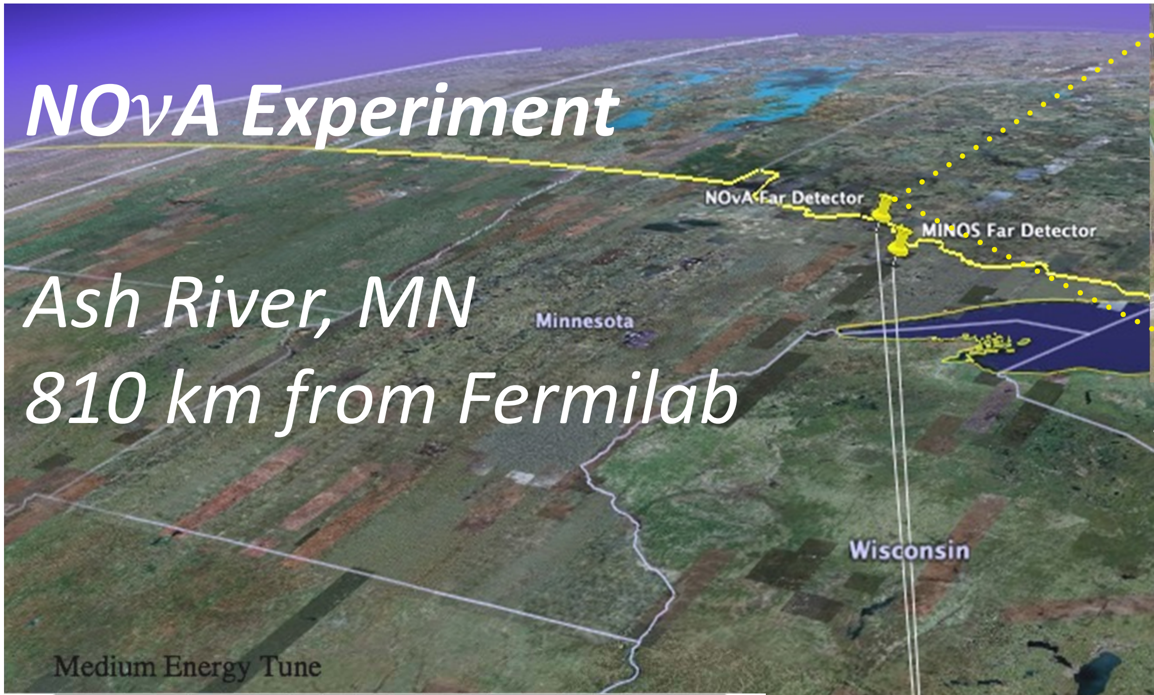
Okinawa, Japan

The NOνA Collaboration

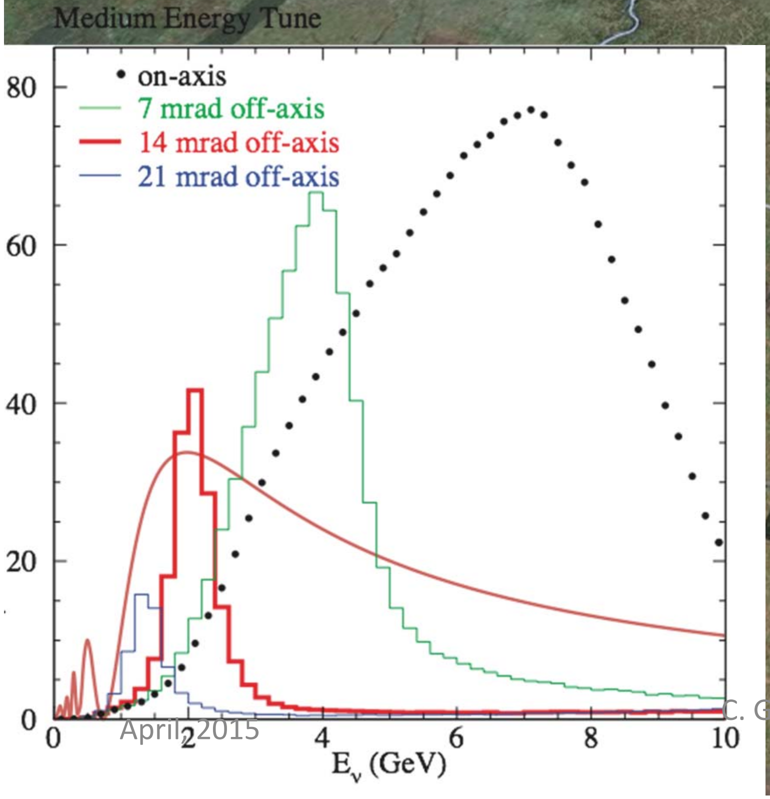
Over 200 scientists, students and engineers from 38 institutions and 7 countries.

NOνA Experiment

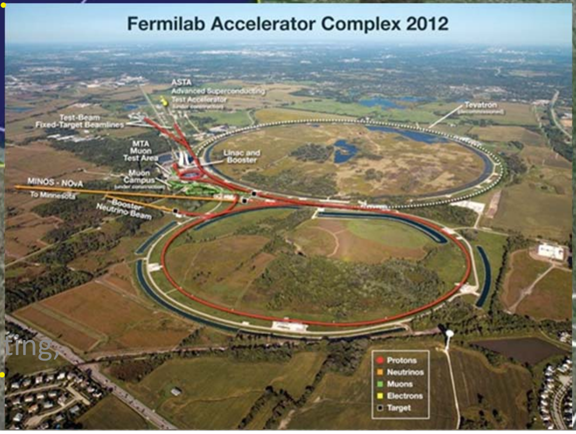
Ash River, MN
810 km from Fermilab



Far detector on the surface

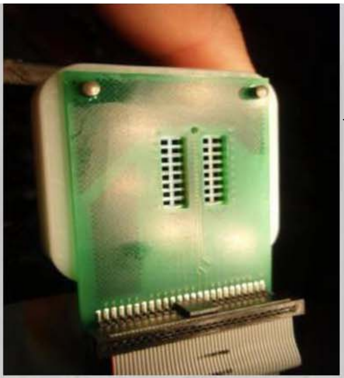


NuMI beam at 700 kW and Near detector underground

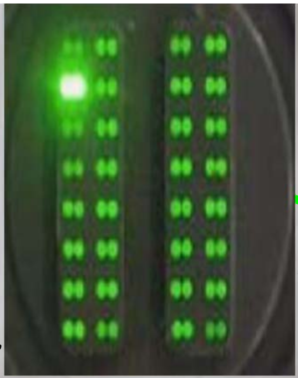


NOνA Detectors:

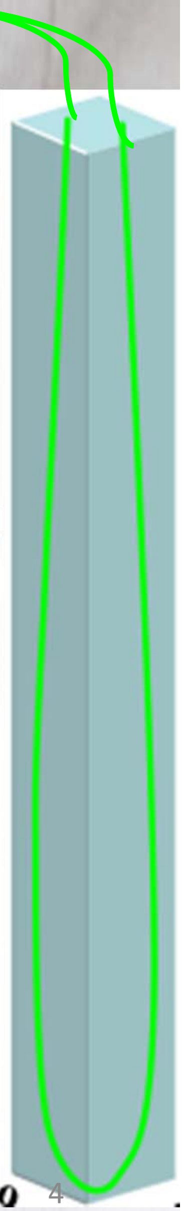
- Fine-grained, low-*Z*, highly-active tracking calorimeters
- 11 M liters of scintillator
- λ-shifting fiber and APDs

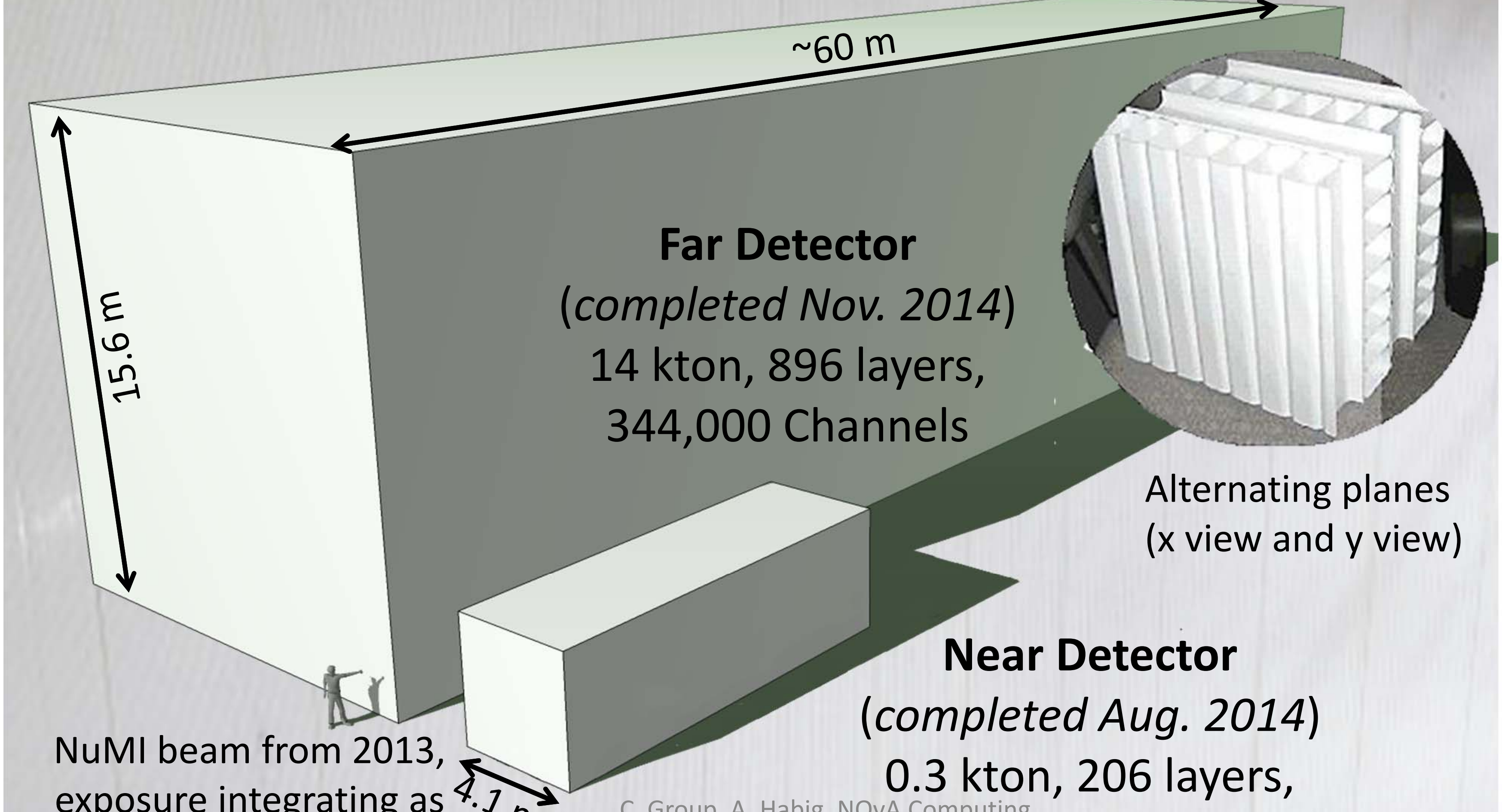


32-pixel APD



Fiber pairs from 32 cells





Far Detector
(*completed Nov. 2014*)
14 kton, 896 layers,
344,000 Channels

Alternating planes
(x view and y view)

Near Detector
(*completed Aug. 2014*)
0.3 kton, 206 layers,
18,000 Channels

NuMI beam from 2013, exposure integrating as detectors were completed

NOvA Demand is Large.

- Almost 2 PB of NOvA files already written to tape -- more than 5M individual files.
  - ~5,000 raw data files per day
  - > 15M CPU hours used over the last year
- Total dataset will be comparable to everything from the Tevatron
- Plan to reprocess all data and generate new simulation ~2 times per year.

(We call this a “production run”)

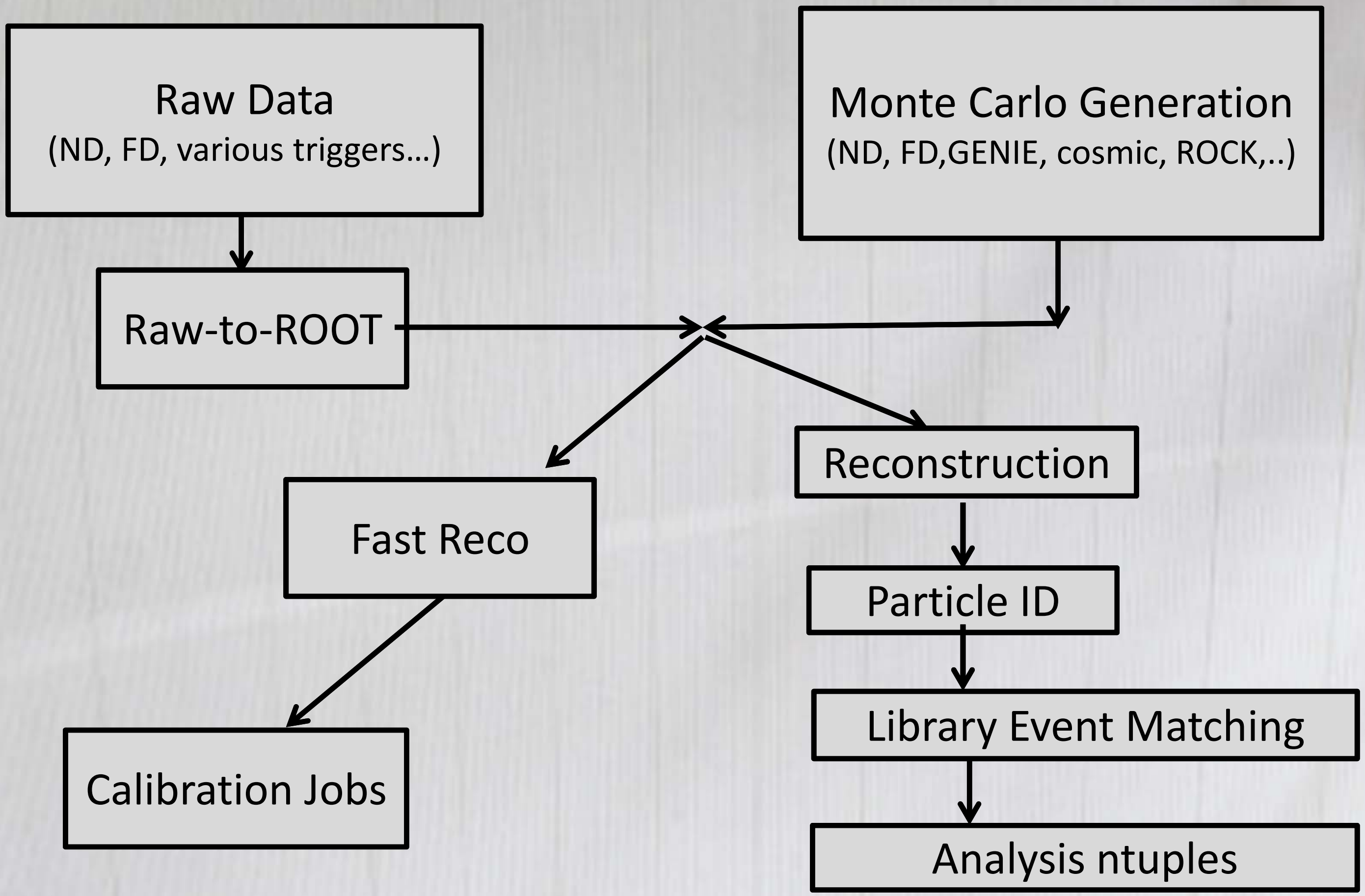
Transition of computing paradigm

- 3 years ago NOvA used a large networked disk (BlueArc) for all file storage and ran jobs locally on cores with direct access to the disk.
  - Not scalable!
- Transition to Sequential data Access via Metadata
  - Forerunner of LHC data management (CDF and D0)
  - Database with metadata catalog (file name, size, events, run info, luminosity, MC details, ...)
  - Dataset creation and queries
  - Coordinates and manages data movement to jobs (gridftp, dccp, future XRootD)
  - Cache management (now using dCache)
  - File consumption and success tracking, recovery
- All NOvA production efforts now use SAM for data handling, reading and writing files directly from tape-backed dCache.
- Capable of running on the Open Science Grid using CVMFS to make code releases available offsite.

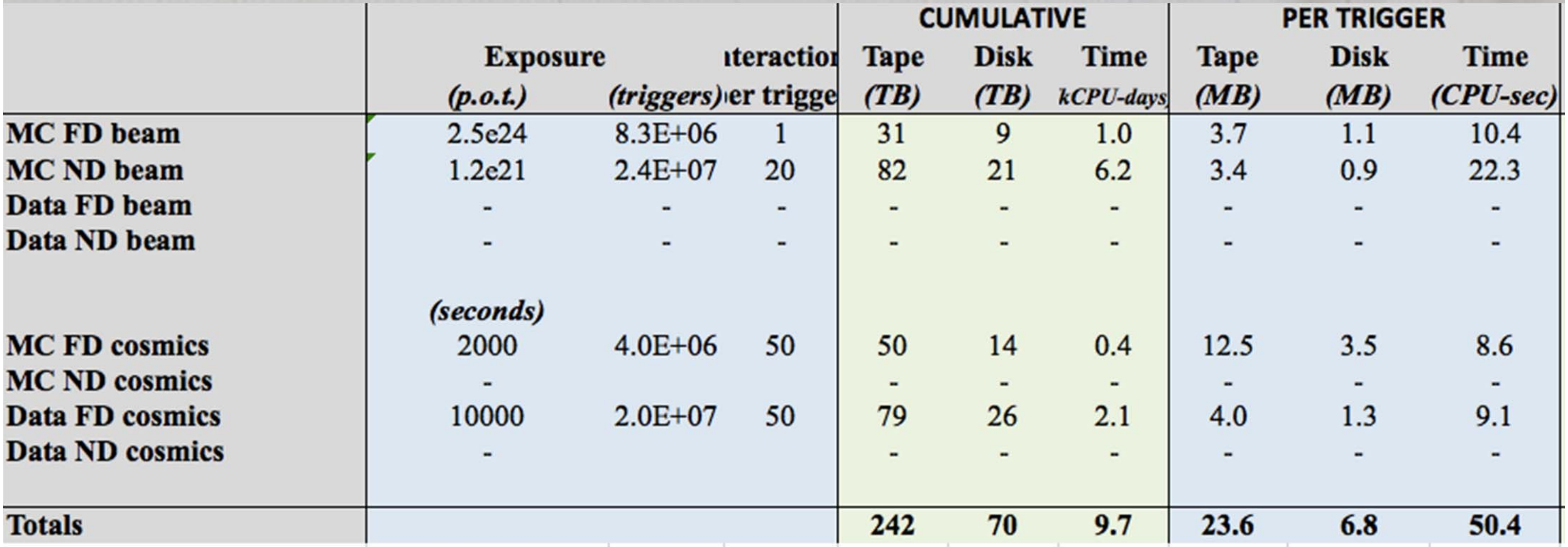
Summary of Current Infrastructure

- VM
  - 10 virtual machines for interactive use
- Blue Arc (*nfs-based NAS*):
  - 335 TB of interactive data storage for short term or small data sets
- Tape:
  - Long term data storage
  - Files registered with SAM
  - Frontend is 4 PB of dCache 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 queues: thousands of additional slots
- Databases: Several (PostgreSQL), required for online and offline operations
  - Accessed via http for ease of offsite usage

Production Flow



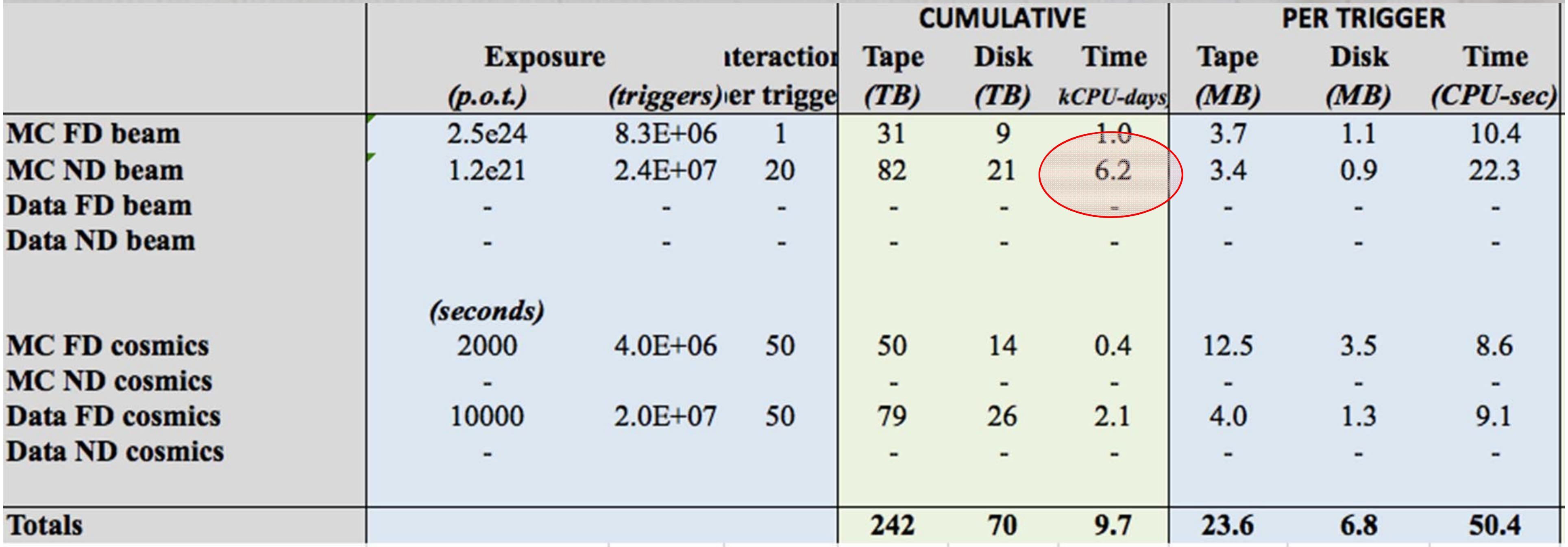
Goals set in Fall 2013



				CUMULATIVE			PER TRIGGER		
	Exposure		ıteractio	Tape	Disk	Time	Tape	Disk	Time
	(p.o.t.)	*(triggers)*	ıer trigge	*(TB)*	*(TB)*	*kCPU-days*	*(MB)*	*(MB)*	*(CPU-sec)*
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	-	-	-	-	-	-	-	-	-
	(seconds)								
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 a major effort to to understand resources and to streamline production tools in advance of doing big production runs

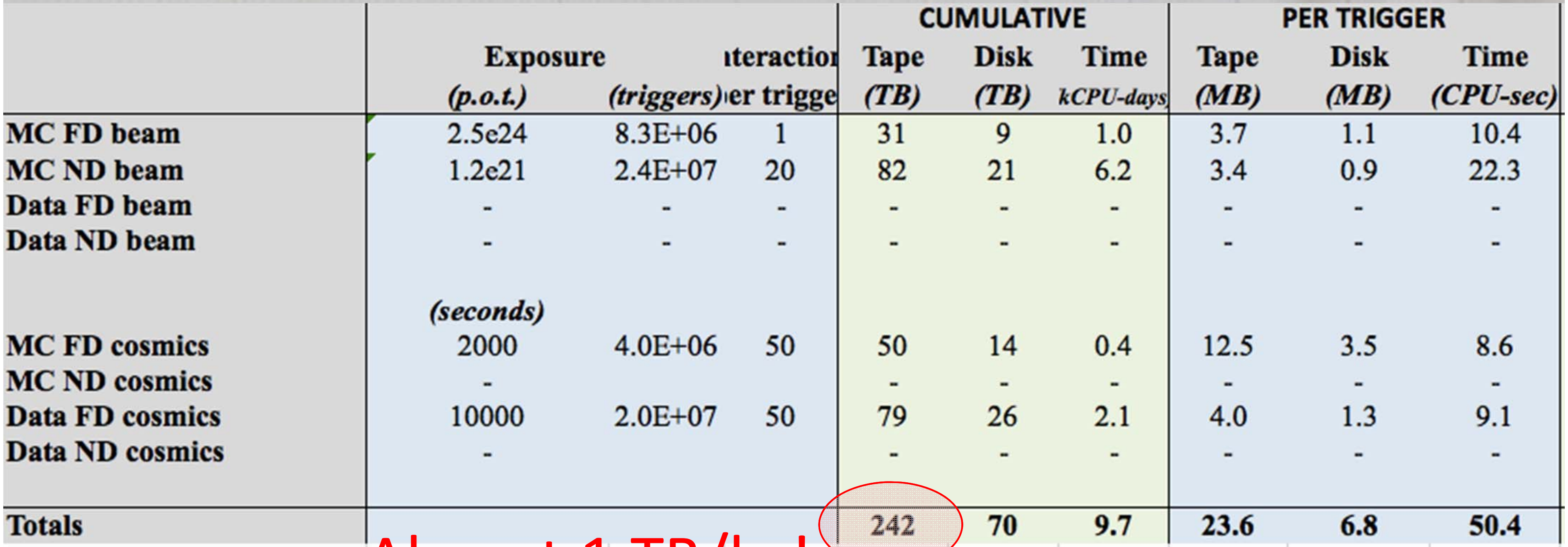
Goals set in Fall 2013



				CUMULATIVE			PER TRIGGER		
	Exposure		ıteractio	Tape	Disk	Time	Tape	Disk	Time
	(p.o.t.)	*(triggers)*	er trigge	*(TB)*	*(TB)*	*kCPU-days*	*(MB)*	*(MB)*	*(CPU-sec)*
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	-	-	-	-	-	-	-	-	-
	(seconds)								
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 a major effort to to understand resources and to streamline production tools in advance of doing big production runs

Goals set in Fall 2013



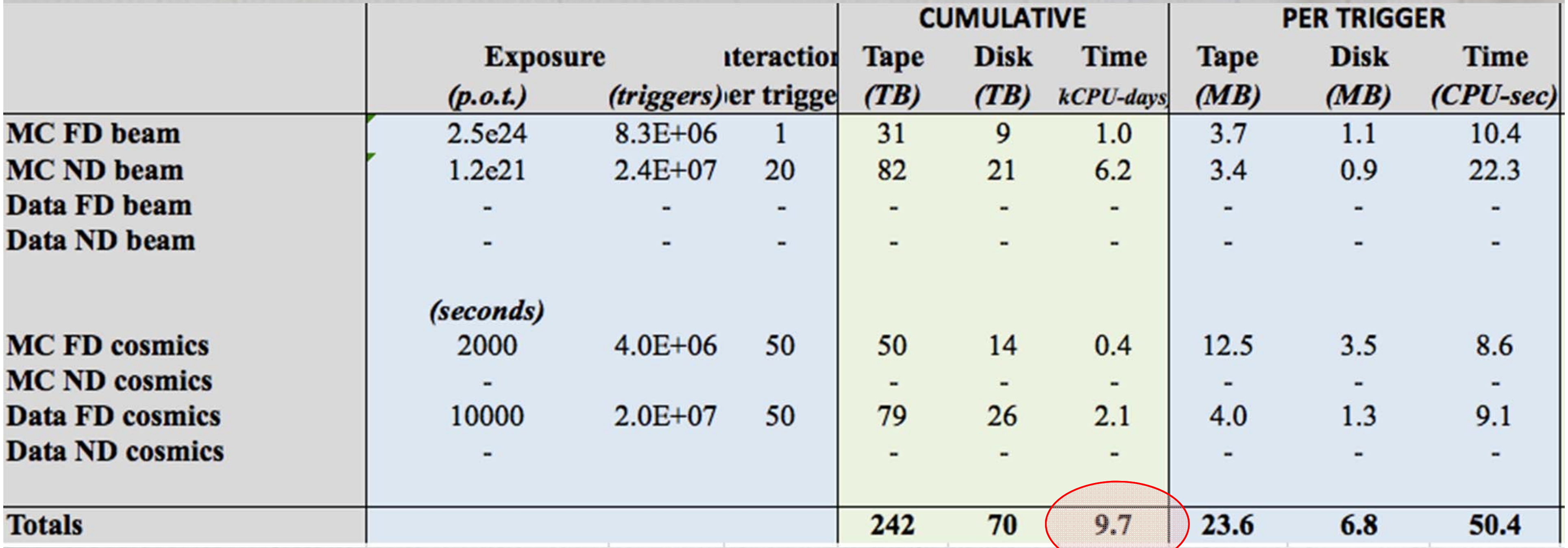
				CUMULATIVE			PER TRIGGER		
	Exposure		ıteractioı	Tape	Disk	Time	Tape	Disk	Time
	(p.o.t.)	*(triggers)*	ıer trigge	*(TB)*	*(TB)*	*kCPU-days*	*(MB)*	*(MB)*	*(CPU-sec)*
MC FD beam	2.5e24	8.3E+06	1	31	9	1.0	3.7	1.1	10.4
MC ND beam	1.2e21	2.4E+07	20	82	21	6.2	3.4	0.9	22.3
Data FD beam	-	-	-	-	-	-	-	-	-
Data ND beam	-	-	-	-	-	-	-	-	-
	(seconds)								
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

Almost 1 TB/hr!

(250 TB = IF+cosmics for full month)

- 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 a major effort to to understand resources and to streamline production tools in advance of doing big production runs

Goals set in Fall 2013

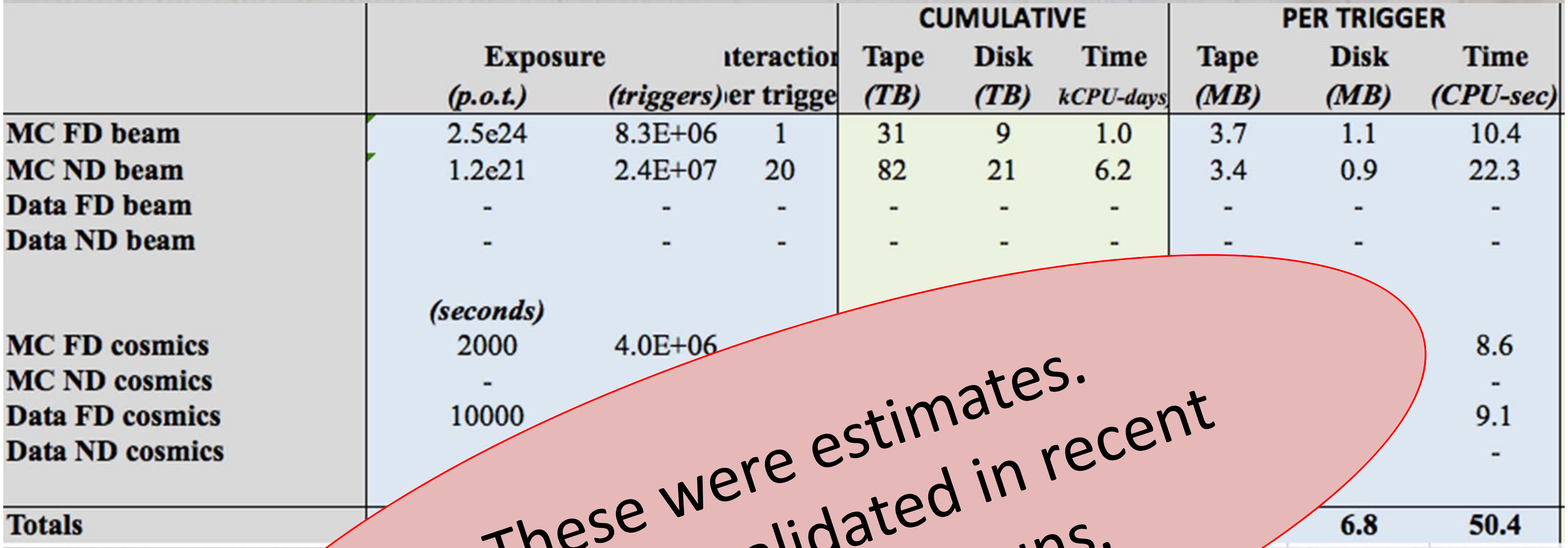


				CUMULATIVE			PER TRIGGER		
	Exposure		iteractio	Tape	Disk	Time	Tape	Disk	Time
	(p.o.t.)	*(triggers)*	er trigge	*(TB)*	*(TB)*	*kCPU-days*	*(MB)*	*(MB)*	*(CPU-sec)*
MC FD beam	2.5e24	8.3E+06	1	31	9	1.0	3.7	1.1	10.4
MC ND beam	1.2e21	2.4E+07	20	82	21	6.2	3.4	0.9	22.3
Data FD beam	-	-	-	-	-	-	-	-	-
Data ND beam	-	-	-	-	-	-	-	-	-
	(seconds)								
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 a major effort to to understand resources and to streamline production tools in advance of doing big production runs

Goals set in Fall 2013

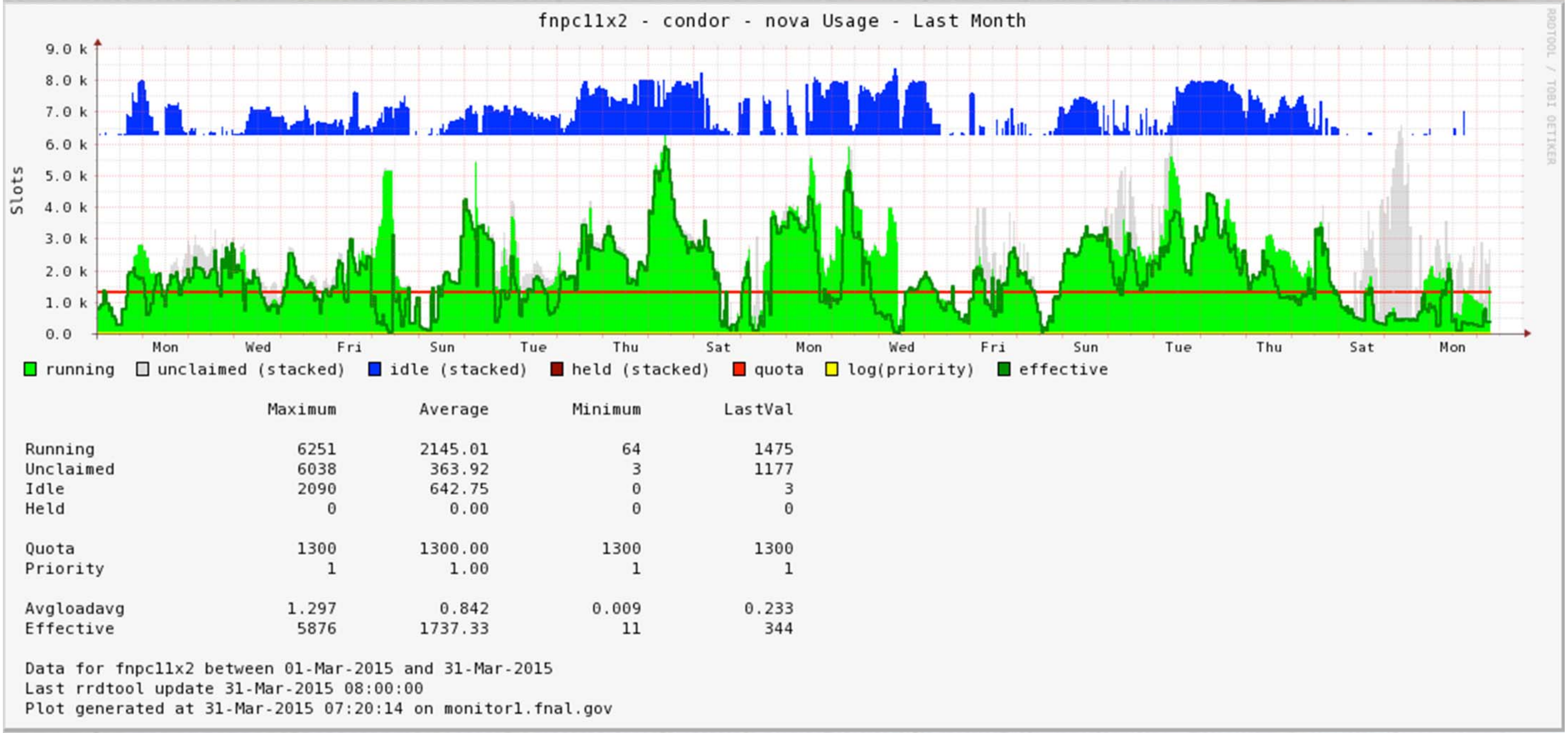


				CUMULATIVE			PER TRIGGER		
	Exposure		ıteractio	Tape	Disk	Time	Tape	Disk	Time
	(p.o.t.)	(triggers)	ıer trigge	(TB)	(TB)	kCPU-days	(MB)	(MB)	(CPU-sec)
MC FD beam	2.5e24	8.3E+06	1	31	9	1.0	3.7	1.1	10.4
MC ND beam	1.2e21	2.4E+07	20	82	21	6.2	3.4	0.9	22.3
Data FD beam	-	-	-	-	-	-	-	-	-
Data ND beam	-	-	-	-	-	-	-	-	-
	(seconds)								
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 were validated in recent file production runs.

- Producti
  - Th d be less than 100TB.
  - Tl plete in a two week period.
- There w to to understand resources and to streamline production tools in advance of doing big production runs

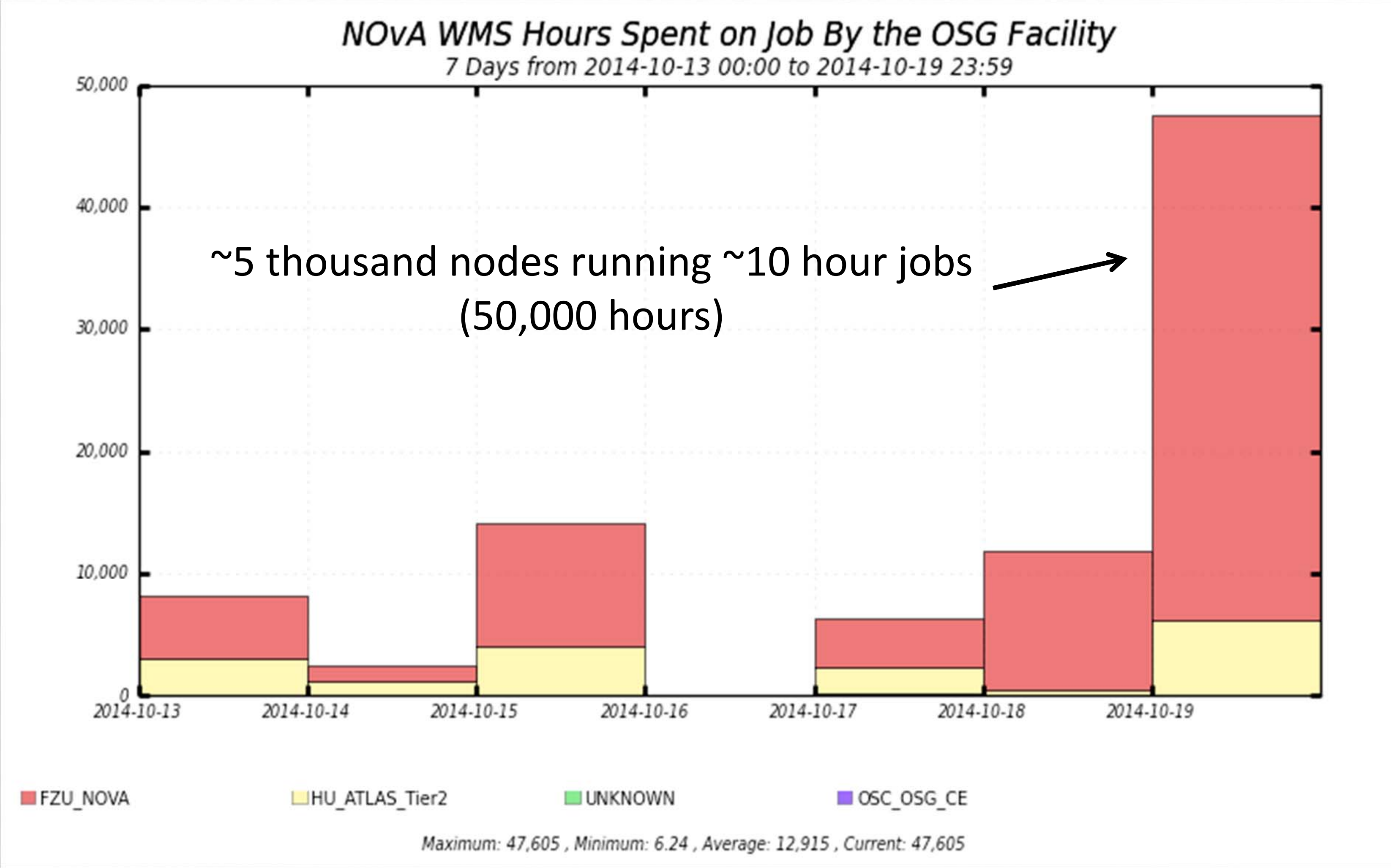
CPU (on site)



```
                    Maximum       Average       Minimum       LastVal

Running                6251       2145.01            64          1475
Unclaimed              6038        363.92             3          1177
Idle                   2090        642.75             0             3
Held                      0          0.00             0             0

Quota                  1300       1300.00          1300          1300
Priority                  1          1.00             1             1

Avgloadavg            1.297         0.842         0.009         0.233
Effective              5876       1737.33            11           344

Data for fnpc11x2 between 01-Mar-2015 and 31-Mar-2015
Last rrdtool update 31-Mar-2015 08:00:00
Plot generated at 31-Mar-2015 07:20:14 on monitor1.fnal.gov
```

CPU is has not been a limiting factor.

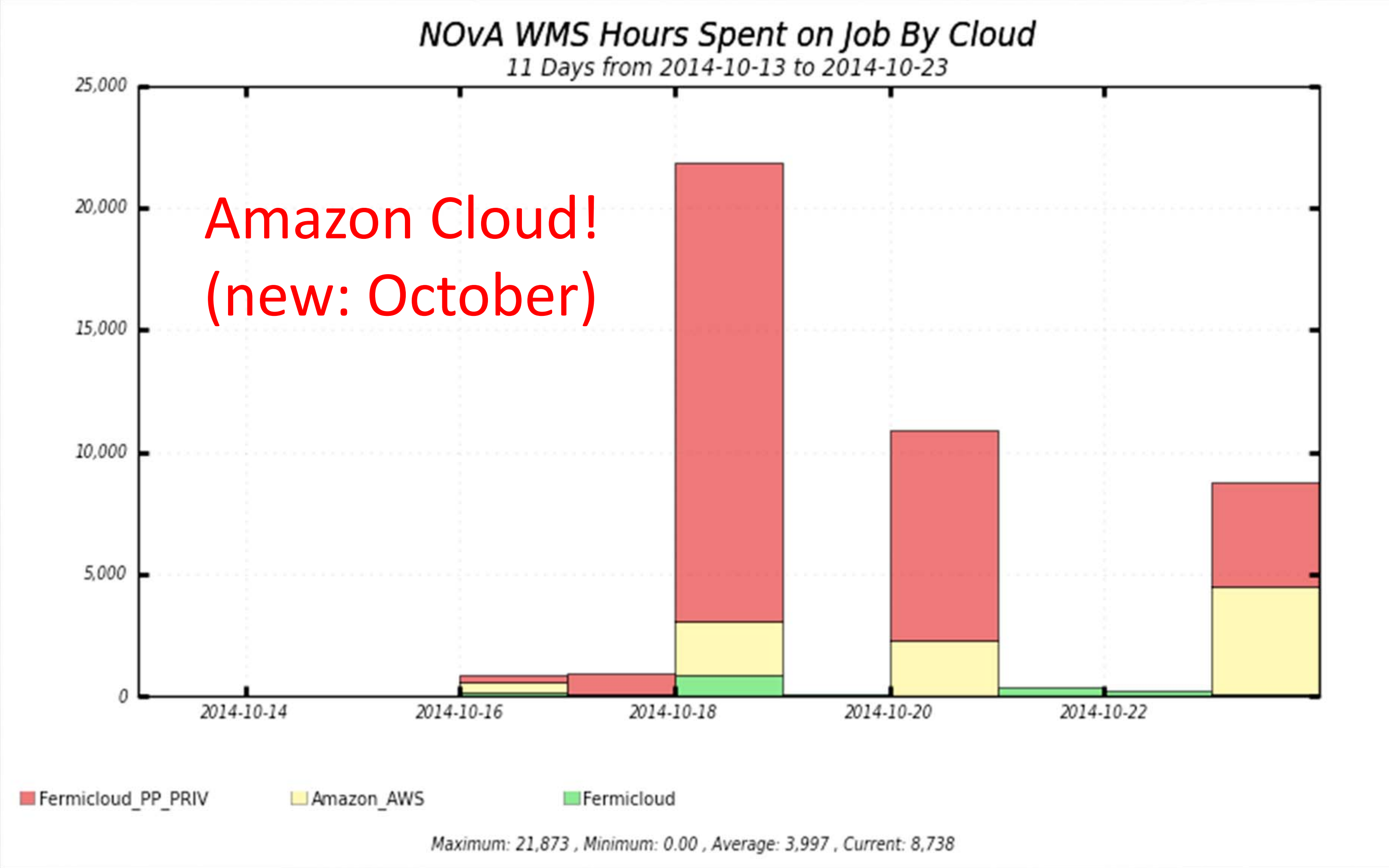
CPU (off site)



~5 thousand nodes running ~10 hour jobs
(50,000 hours)

Thousands of offsite CPU slots are also available to us.

CPU (cloud)

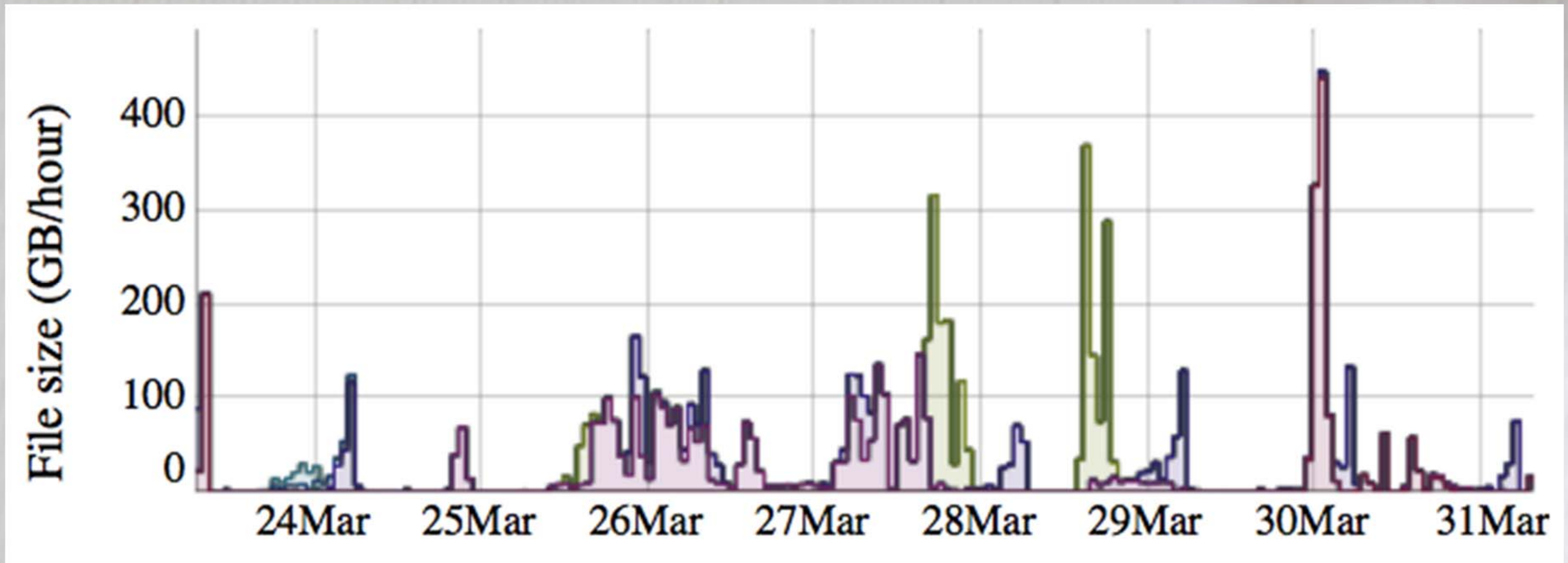


Amazon Cloud!
(new: October)

Recently received funding for significant Amazon Cloud running for production data sets.

File Throughput

(to obtain dCache location and register in SAM)



- Example file transfer from last week of March (*1 of 3 servers*)
- Often have sustained throughput >200 GB/hr on each server
- We have three FTS servers
- More than 1TB/hour total has been demonstrated

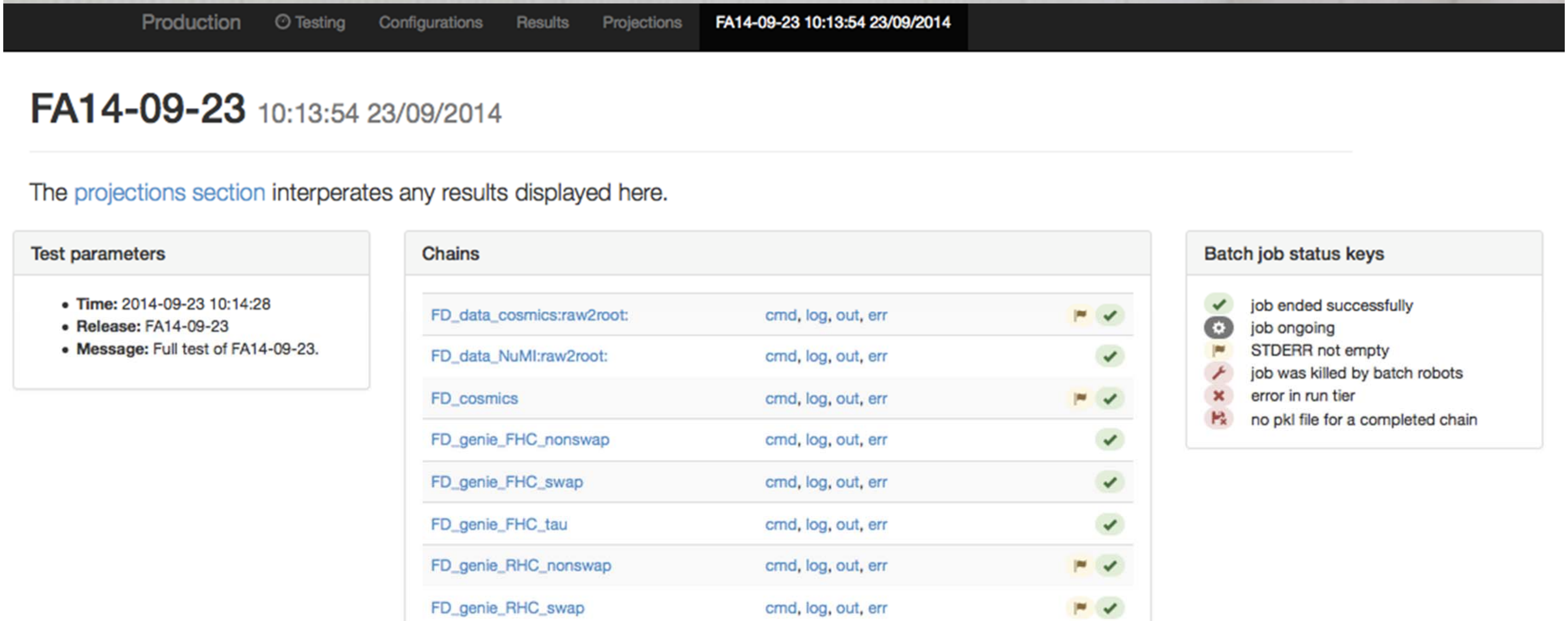
Two Example Production Runs

- Spring 2014 production: 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, we had ups and downs, learned a lot
  - In the end we ran all steps of production in time for Neutrino 2014 (some steps multiple times)
- Winter 2015 production:
  - The data set production effort for first physics results
    - Includes completed detector data
  - Many first-time requests: new keep-up data sets, calibration requests, systematic samples…
  - The SAM paradigm is functioning well.
- Earlier estimates and resource predictions right on the mark

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 interperates 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

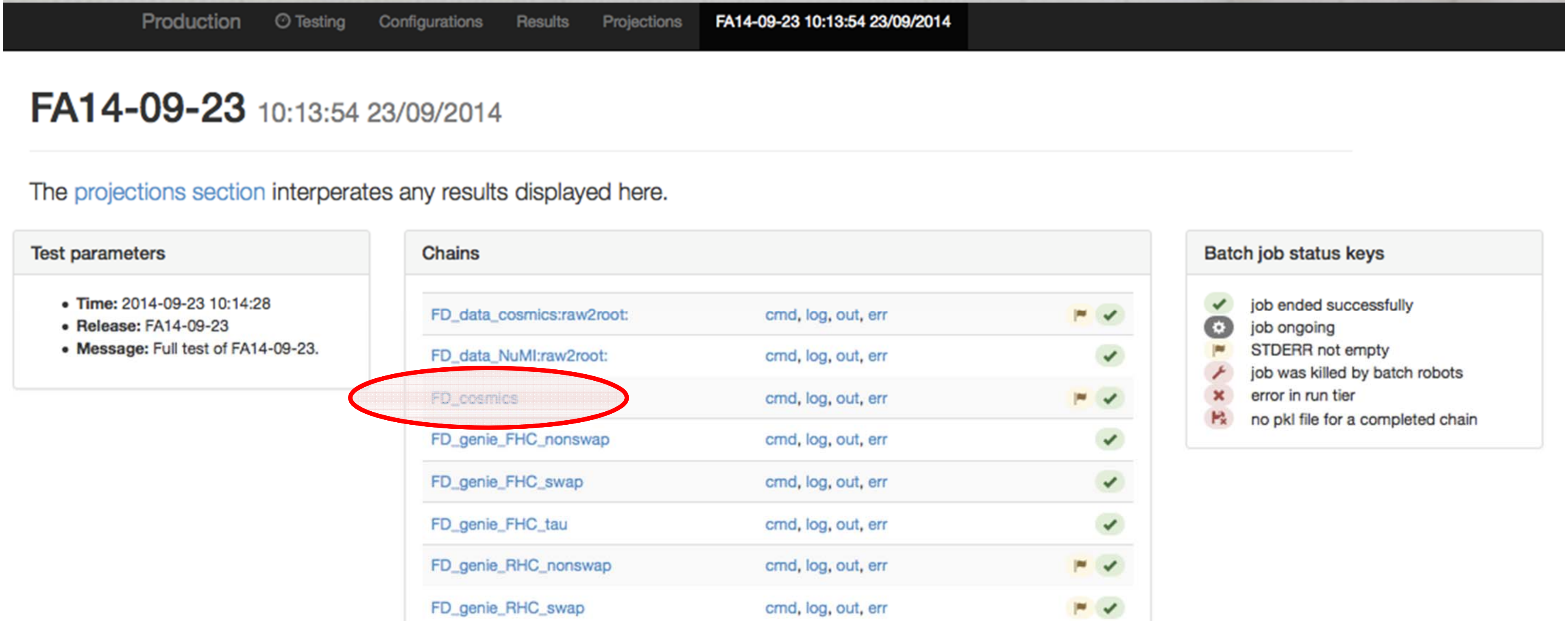
FD_data_cosmics:raw2root:	cmd, log, out, err	⚑ ✓
FD_data_NuMI:raw2root:	cmd, log, out, err	✓
FD_cosmics	cmd, log, out, err	⚑ ✓
FD_genie_FHC_nonswap	cmd, log, out, err	✓
FD_genie_FHC_swap	cmd, log, out, err	✓
FD_genie_FHC_tau	cmd, log, out, err	✓
FD_genie_RHC_nonswap	cmd, log, out, err	⚑ ✓
FD_genie_RHC_swap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

See M.Tamsett and R.Group's poster #201
"Software framework testing at the Intensity Frontier"

Validation of File Production Tools



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

The projections section interperates 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

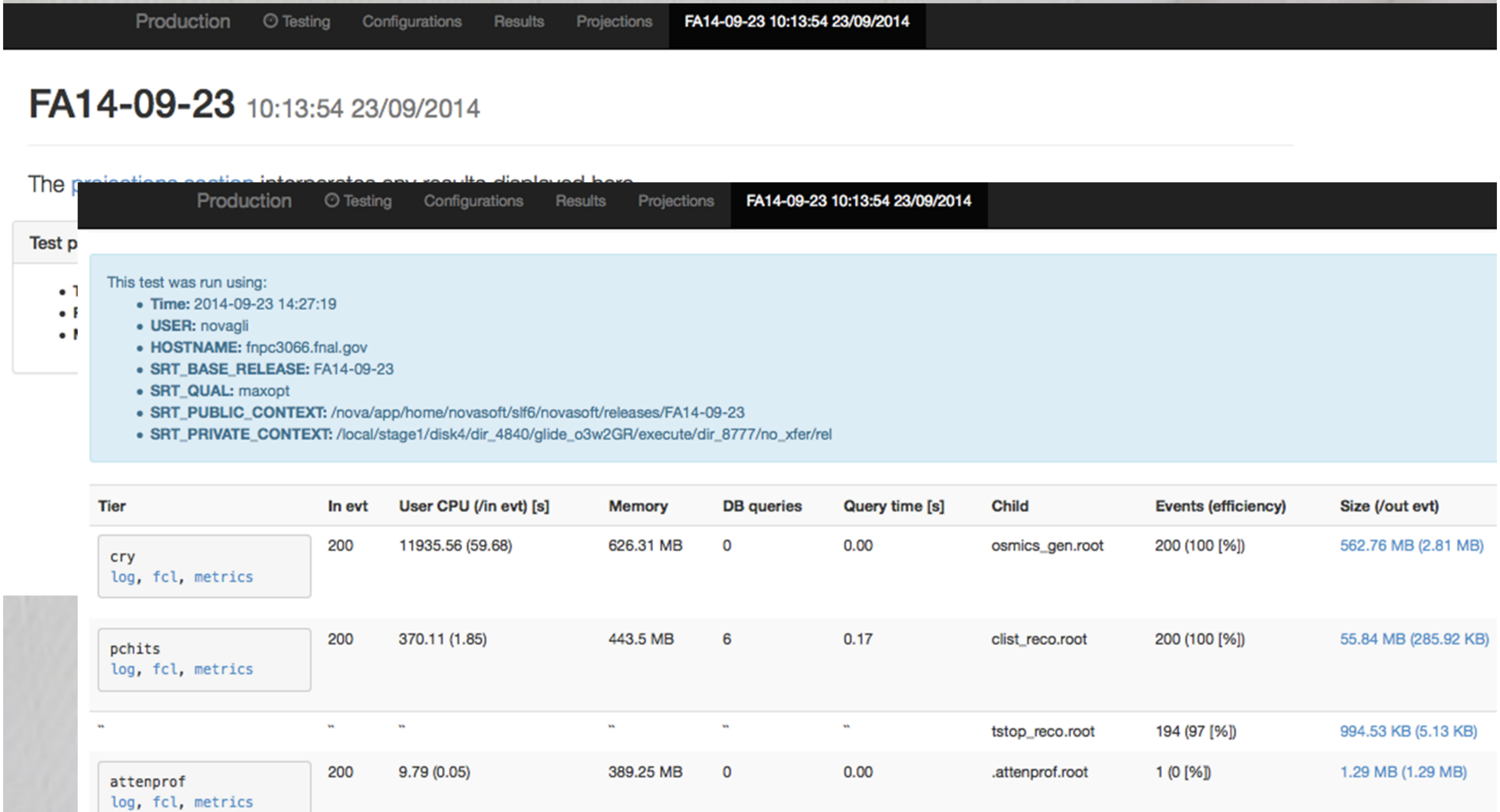
FD_data_cosmics:raw2root:	cmd, log, out, err	⚑ ✓
FD_data_NuMI:raw2root:	cmd, log, out, err	✓
FD_cosmics	cmd, log, out, err	⚑ ✓
FD_genie_FHC_nonswap	cmd, log, out, err	✓
FD_genie_FHC_swap	cmd, log, out, err	✓
FD_genie_FHC_tau	cmd, log, out, err	✓
FD_genie_RHC_nonswap	cmd, log, out, err	⚑ ✓
FD_genie_RHC_swap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

See M.Tamsett and R.Group's poster #201
"Software framework testing at the Intensity Frontier"

Validation of File Production Tools



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

The projections section interperates any results displayed here.

Test p

- T
- F
- N

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)

Summary

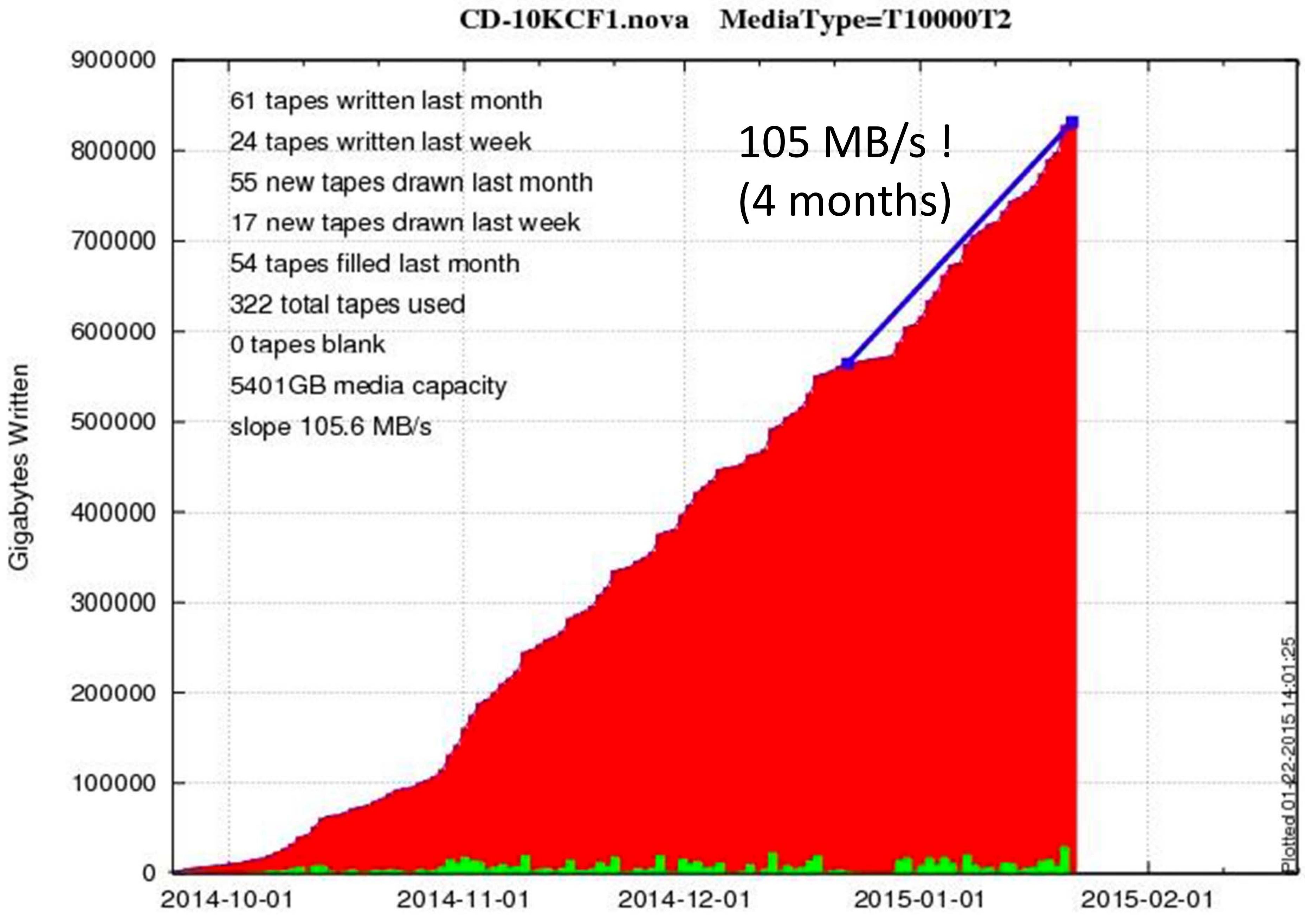
- There has been a recent transition to ascalable file handling system similar to what was employed by CDF and D0
- Computing resources are sufficient and we are ready to serve the data sets required by the collaboration for physics
  - CD is working closely with is to solve issues as they arrive
- Now taking advantage of offsite CPU resources (CVMFS works great!)
- Demonstrated production framework, and measured/documented resource requirements
- New production validation framework is very useful
- Now producing a full set of production files for analysis groups and first physics

Other relevant parallel talks and posters on NOvA computing…

- “Software framework testing at the Intensity Frontier”
  - M.Tamsett & R.Group, poster #201
- “Large Scale Monte Carlo Simulation of neutrino interactions using the Open Science Grid and Commercial Clouds”
  - A.Norman, poster #465
- “Data Handling with SAM and ART at the NOvA Experiment”
  - A.Aurisano, poster #214
- “The NOvA Simulation Chain”
  - A.Aurisano, talk #213
- “Software Management for the NOvA Experiment”
  - G.Davies, R.Group, poster #293
- “A Data Summary File Structure and Analysis Tools for Neutrino Oscillation Analysis at the NOvA Experiment”
  - D.Rocco & C.Backhouse, poster #453

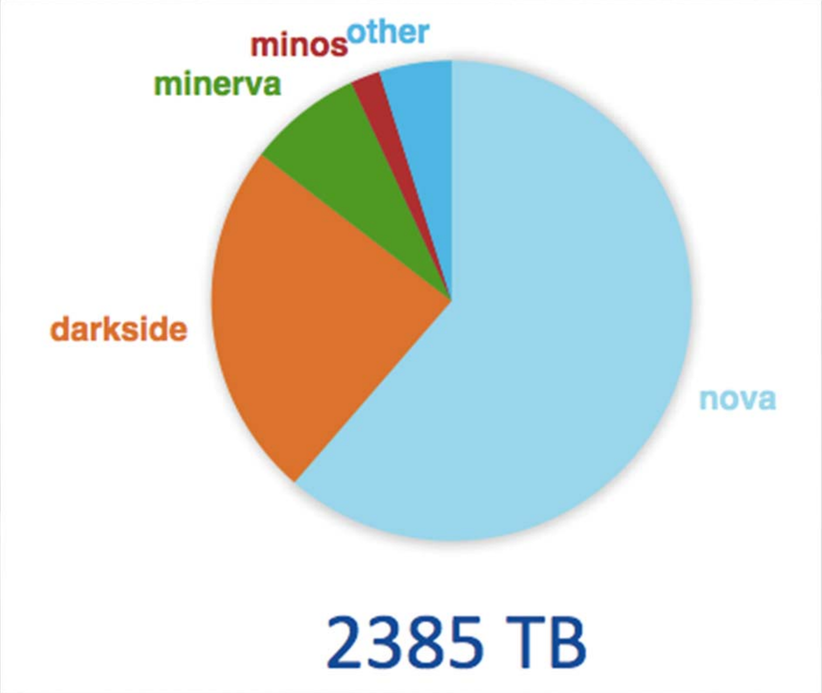
Extra slides follow…

Recent tape usage



105 MB/s !
(4 months)

dCache usage by experiment



dCache

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 in 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 date 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.

- UPPER LIMIT: online triggering used to save much less data

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