

# Offline Status and Plans

Rick and Rick

For the Offline Group

International Finance Committee  
30 Oct 2009

# Outline

- ▶ Highlights
- ▶ Who we are, were, will be and what we do
- ▶ Report on the subgroups: Production, Monte Carlo, Ntuples; Data Handling, Computing, Code Management.
- ▶ Dangers
- ▶ Budgets: In and Out of house
- ▶ Summary

## Highlights

- ▶ All CDF computing done on Grid
- ▶ Processing time down from 10-12 weeks to 9.1 weeks
- ▶ Significant shift of effort to CD
- ▶ Hardware good through 2011 running and beyond
- ▶ Staffed through 2011

## Offline Model Components

- ▶ Raw Data logging
- ▶ MC Production
- ▶ User analysis and data handling

Guiding principle for budgeting personnel:

The team for CAF and DH will form a core of experts that spend not too much time and aim for smooth operations. REX has been and continues to be a great defender of our operation: Special thanks to Margaret Votava for her proactive work!

Divide handling of this into groups of people.

## Offline Subgroups

### ► Operations

- **Production:** Raw data to reconstructed output, then to top and stn-tuples for analysis
- **Monte Carlo:** produced on sites where one must take the CDF software along and must access the database using Frontier. (Usually offsite: NAMGrid, PACCCaf, CNAF, LCGGrid)
- **Ntuples:** Top/Stntuples created for MC/Data
- **Analysis:** CDFGrid is 5500 processing slots where one can access data stored in various ways: managed cache with (DCache), without (Diskpool) tape, reading events on remote files (rootd); and there is access to the cdfsoftware server maintained by code management.

## ► Infrastructure

- **Data Handling:** Main Focus: keeping DCache and SAM working at Fermilab on CDFGrid. Includes loading of raw data to tape and keeping track of enough tapes in the silo. Also uses SAM for accessing data offsite. (CNAF, KISTI, Karlsruhe)
- **Code Management:** keep code working on available platforms: SL3, 4, 5! Handle requests for new packages. Provide code browser, CVS repository and non-cdf packages.
- **Hardware:** Replace/repair hardware servers (Grid left to take care of its own.)

## Who Are We?

- ▶ Leaders: Rick and Rick (St. Denis/**Snider**)
- ▶ Associates
  - Head of Operations: **R. Culbertson**
    - ◆ Team: **E. Gerchtein**, S. Golossanov, E. Lee, R. Lysak, **W. Sakumoto**, C. Vellidis, (M. Zvada), O. Tadevosyan, ntuplers, MC Producers
  - Head of Infrastructure: **S. Lammell**
    - ◆ Team: **J. Boyd**, D. Box, E. Wicklund, (A. Bellavance), **F. Moscato**, R. Illingworth, D. Onoprienko, M. Mengel, G. Compestella, D. Lucchesi, J. Bellinger, S. Rolli, **L. Garren**, D. Torretta, S. Pagan-Griso, SAM Shifters (11)

## The 2011 (and beyond) Computing Strategy

- ▶ Exploit infrastructure supported by Grid rather than CDF people while providing a CDF environment.
- ▶ Provide portal to a number of computing sites to minimize dependency on any given site: increase robustness.
- ▶ Minimize and monitor operational load : JIRA reporting  
→ Create short projects, request personnel from CD, CDF to get a net reduction in effort needed
- ▶ Stability, Stability, Stability: recovery from errors is expensive and absorbs physics output of postdocs and students.

- ▶ Minimize and do not tolerate interruptions during peak physics output periods (Summer/Winter conferences) while conceding during quieter periods (August, April).
- ▶ Exploit redundant data location opportunities: We welcome international data storage location and job resources on Grid. Most promising is KISTI in Korea.
  - Reduces need for FNAL to be up all the time.
- ▶ Positioning the Data and Job handling such that seamless offsite processing is a minor extension of the existing production model
  - Allows for improvements in algorithms that could bring us over discovery thresholds!

# Production

( E. Gerchtein, S. Golossanov, E. Lee, R. Lysak, W. Sakumoto)

- ▶ All processing done in 200-400 fb<sup>-1</sup> sets: defines a data taking “Period”
- ▶ Last set was Period 25 and this completed.
- ▶ Ready for new data. 9.1 weeks from Period declaration to ntuples.
- ▶ Work on making things run in parallel and with easier startup of streams about done; also increased parallel concatenation.
- ▶ Model for computing is well known and we have analyzed the time for each step. (Backup on Calibration step):

Item	Time(weeks)
Average calibration time	3.7
Average production for high-priority datasets (using high pt electrons as benchmark)	3.8
Average time between finished high pt electrons to finished all streams	1.9
Total production	$3.8+1.9=5.7$
Average time between production done high-priority and ntupling done high-priority	1.7 (could be 0)
Total, average for high-prio	$3.7+3.8+1.7=9.1.$

- ▶ Calibrations: Careful study by Roman Lysak shows we can get the time down from the present 16-26 days to 6-11 days. Biggest bottleneck is waiting on people to sign off.

Data Type	Volume (TB)	No of Events (M)	No of Files
Raw Data	1398.5	9658.1	1609133
Production	1782.8	12988.5	1735186
MC	796.4	5618.5	907689
Stripped-Prd	76.8	712.1	75108
Stripped-MC	0.5	3.0	533
Ntuple	528.3	18033.4	473948
MC Ntuple	283.2	4138.3	257390
Total	4866.5	51151.9	5058987
Data Type	Yearly Volume Increase (TB)	Yearly Event Increase (M)	Yearly File Increase
Raw Data	286.4	1791.7	327598
Production	558.2	3395.9	466755
MC	196.8	1034.6	228153
Stripped-Prd	21.2	127.1	17066
Stripped-MC	0.0	0.0	0
Ntuple	187.3	5732.6	155129
MC Ntuple	86.0	1482.2	75954
Total	1335.8	13564.2	1270655

4.9 Petabytes total, 1.3 Petabytes in last year. 5.6 billion MC, 1 billion in last year. Ntuples: 528 TB, 187 TB in the last year.

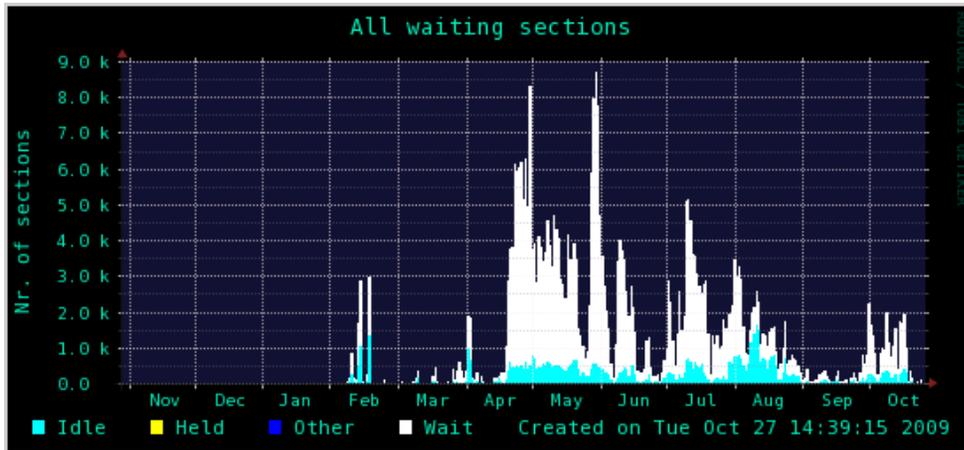
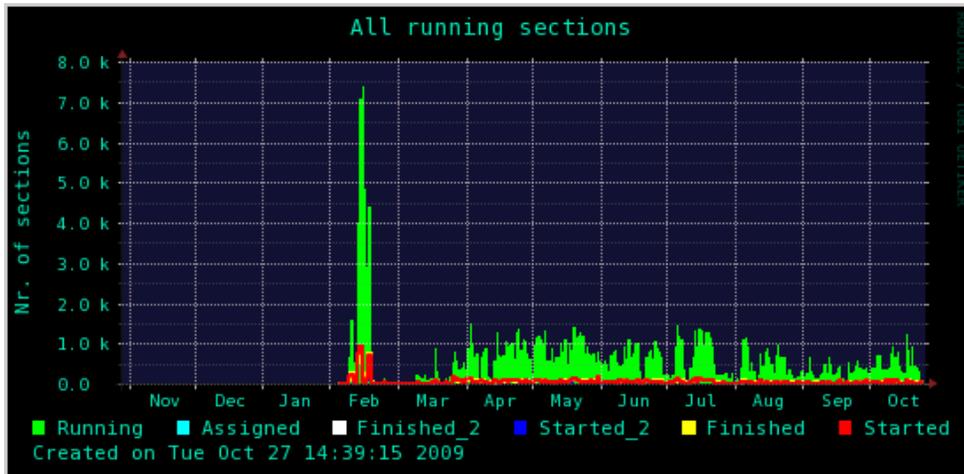
## Monte Carlo

(C. Vellidis, (M. Zvada), O. Tadevosyan, MC Producers)

- ▶ Running on NAMGrid now. Some CDFGrid submitters direct, but...
- ▶ Using some of CDF grid but *through* NAMGrid: utilize our own resources to the maximum.
- ▶ Eliminated sources of failure which were due to the policies of individual sites. Review and monitor this to keep running robust. With many sites, don't rely on one being up. Adding KISTI as a huge powerful site: Great success in getting a test dataset, testing more.
- ▶ Need to understand LCGCaf as a facility: too hard to choose resources by hand? Why harder to transfer files from Europe than Korea? PacCaf also underutilized.

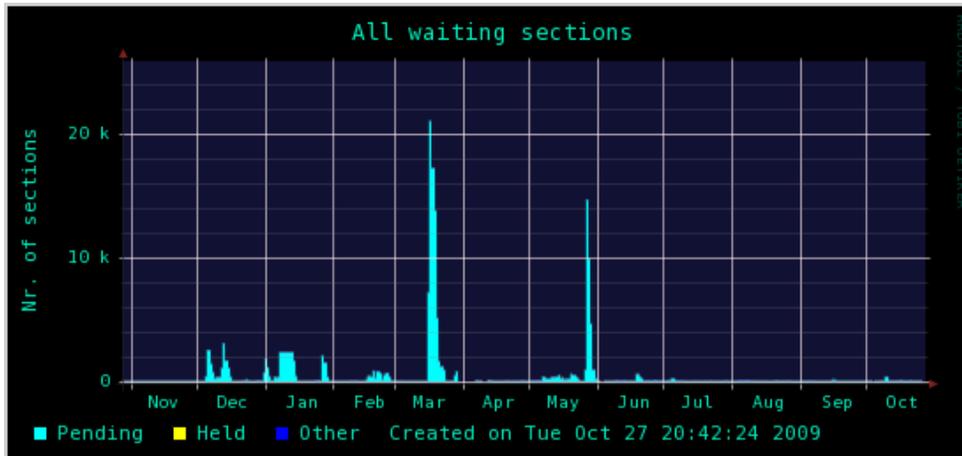
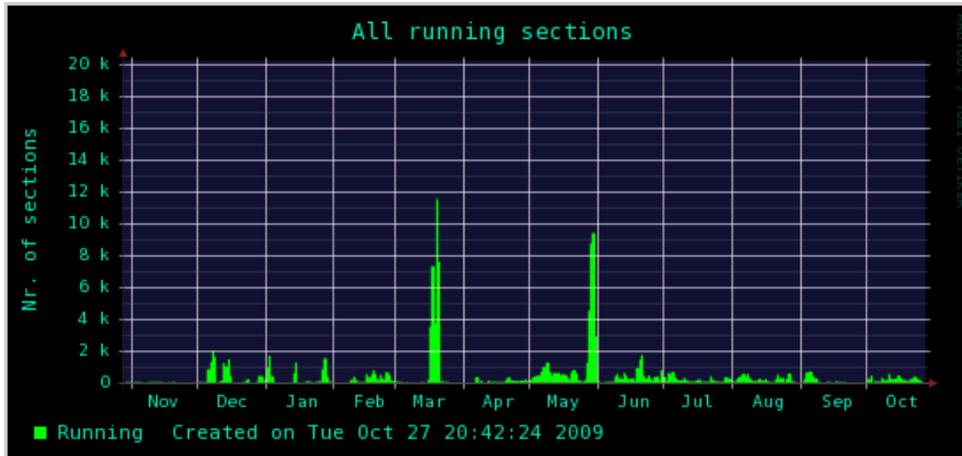
- ▶ Using both run-based and “lumi-based” (especially top group). Lumi-based allows the datasets to be generated ahead of time and once; the weight of the events is adjusted. Lumi is the main run-dependent quantity simulated. Nonetheless there is much discussion of whether this works: when 5 silicon ladders are recovered, one has to adjust acceptance.

# NAMGrid: North America



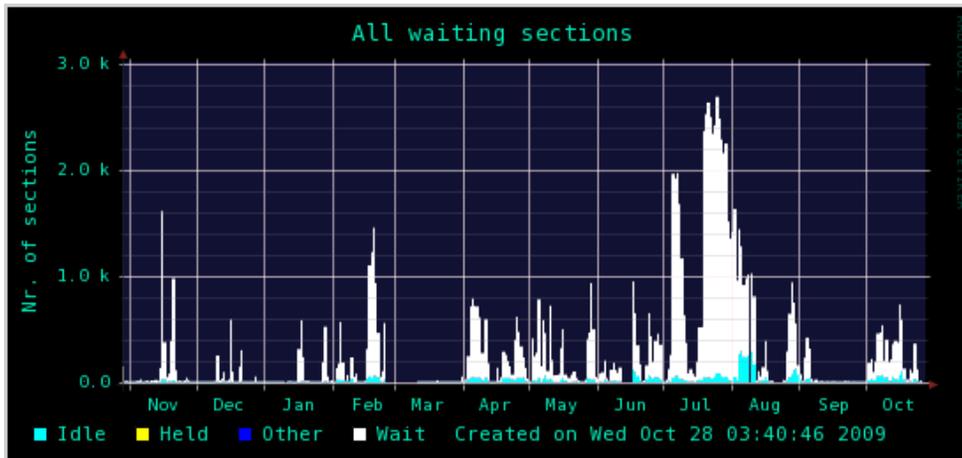
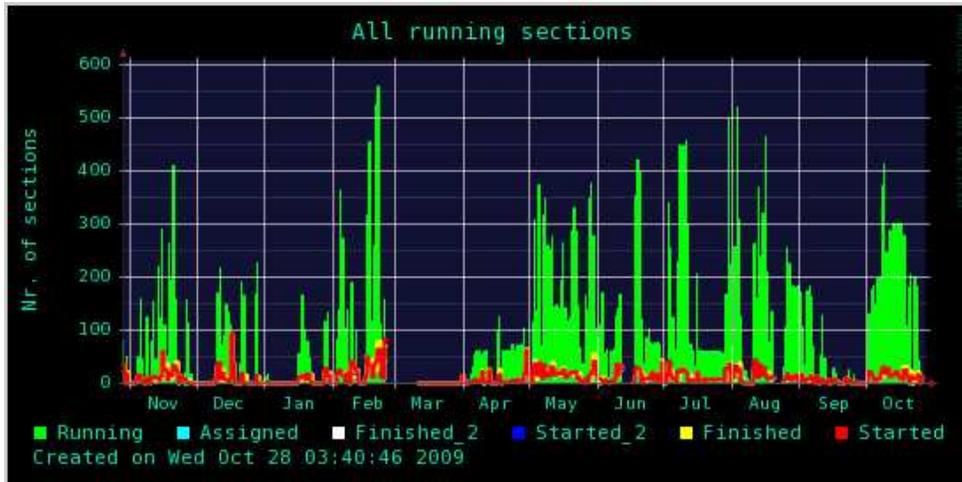
Few Resources now!: much coming back from using FNAL

## LCGCaf: Europe



Very light usage except when desperate (in August)

## PACCAf: Japan, Taiwan



Usage could be better

# Ntuples

(S. Golossanov, ntuplers)

- ▶ Need more help for the team since a majority of effort is in startup of MC Datasets.
- ▶ Running smoothly

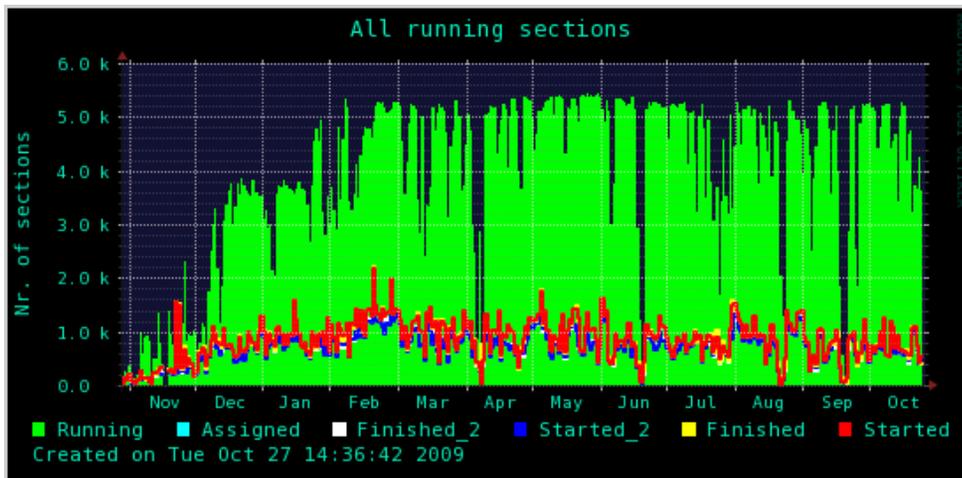
## Analysis: CDFGrid, CNAF

(J. Boyd, D. Box, F. Moscato, G. Compestella, D. Lucchesi, S. Pagan-Griso)

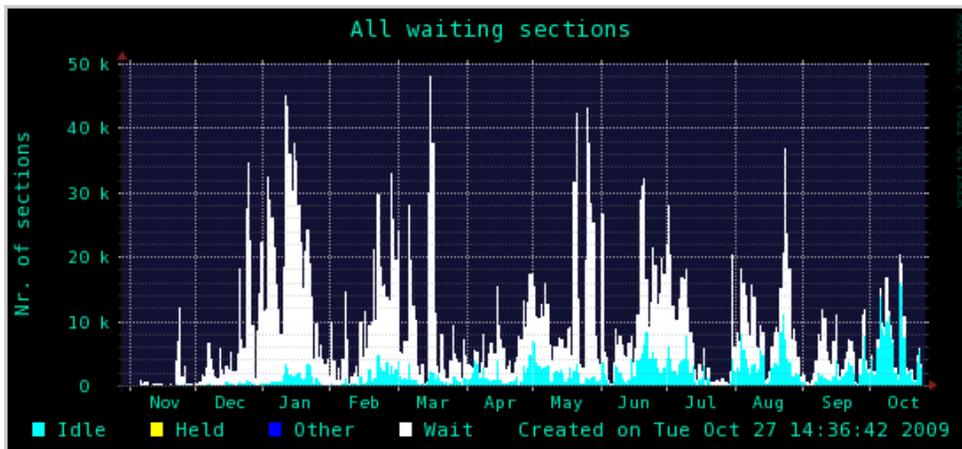
- ▶ Main problem in DH access: leaves computers idle. Studying all the use patterns and susceptibility to failure or overload.
- ▶ We only utilized 30% of the CDFGrid CPU when running on all slots in our last push to conference! (70% over the year) Also **underutilizing** CNAF but due to low demand.
- ▶ REX (running experiments = CD personnel) is working on prestaging of DCache. Also trying to get datasets to Kisti.

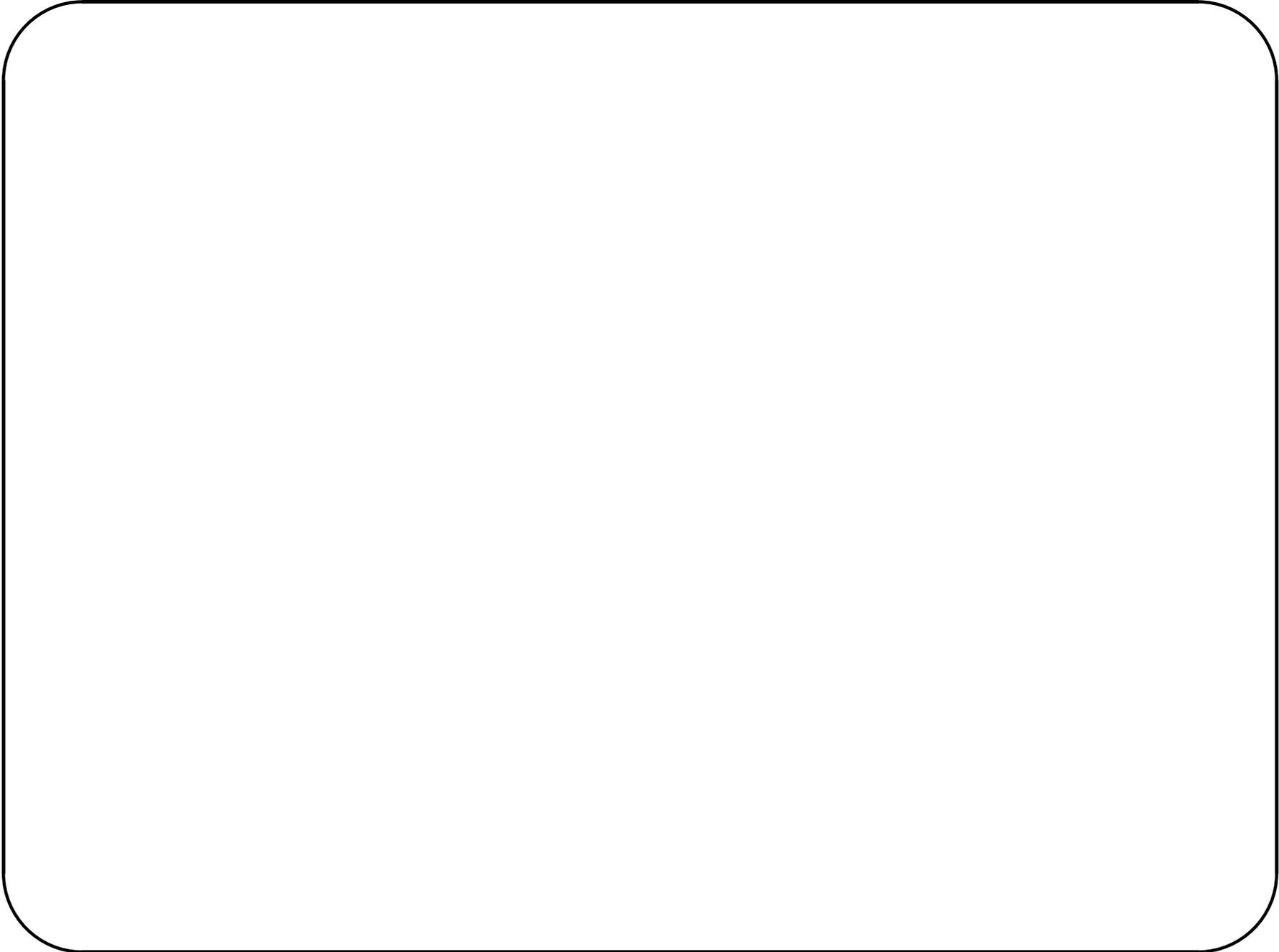
## CDFGrid: Fermilab

Gaps due to (scheduled) downtimes



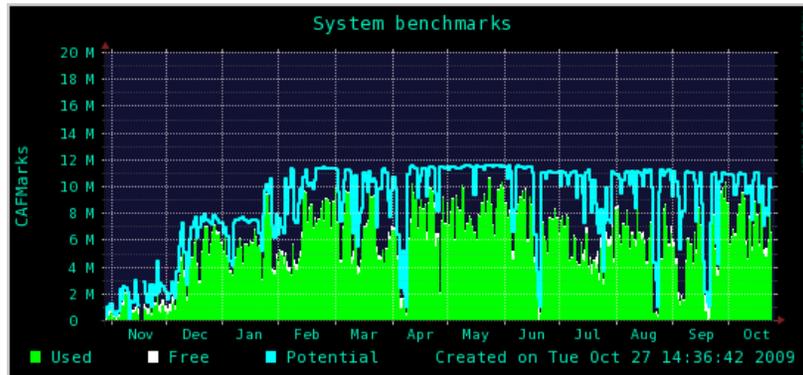
Enormous demand: try to bleed off to others



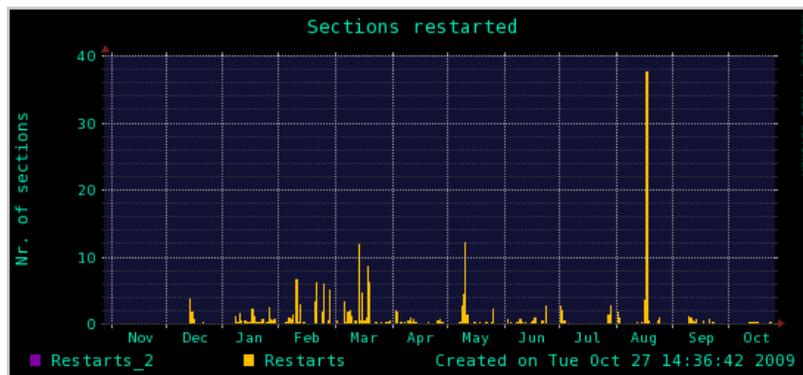


# CDFGrid: Fermilab

Gap between line and shaded: wasted cpu

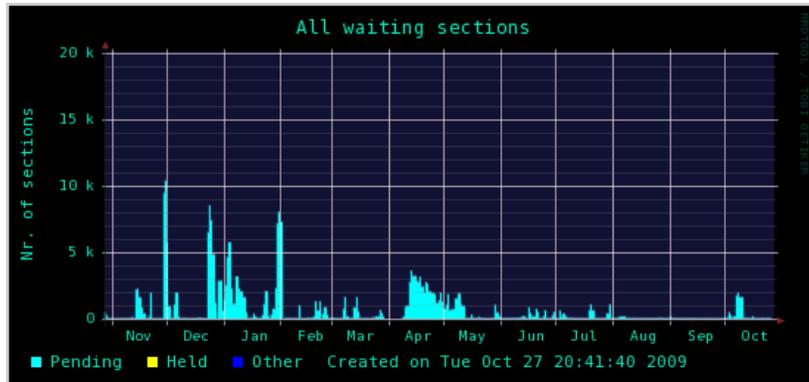
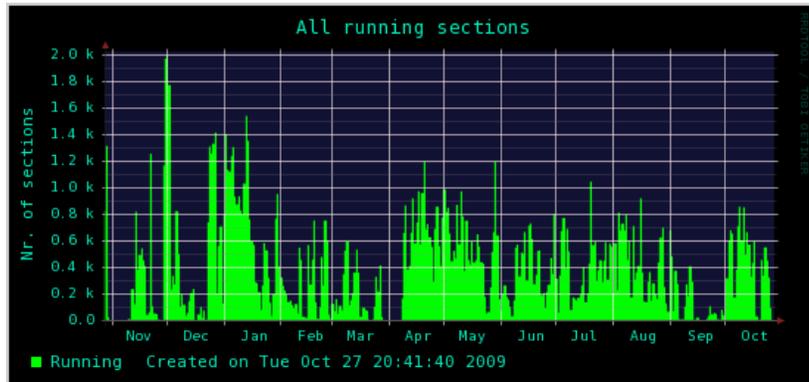


Every restart potential pain to users



# CNAF: Italy

Low Demand: hard for users to know when to use



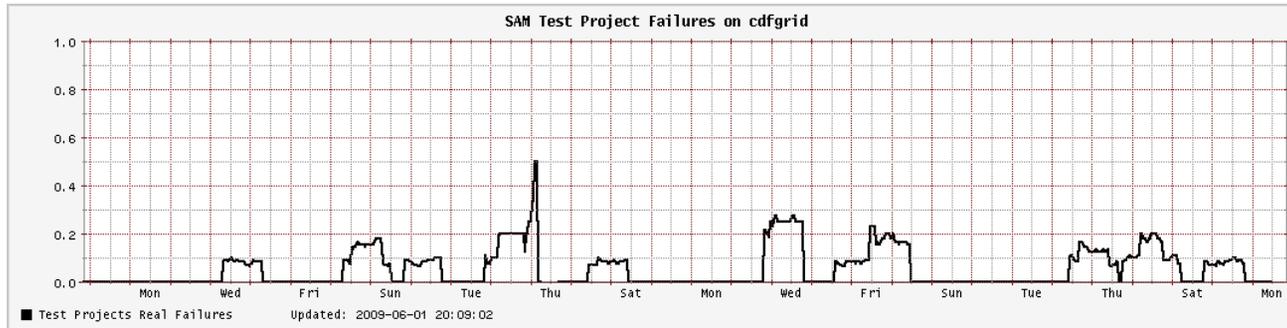
## Data Handling:

(E. Wicklund, (A. Bellavance), D. Onoprienko,

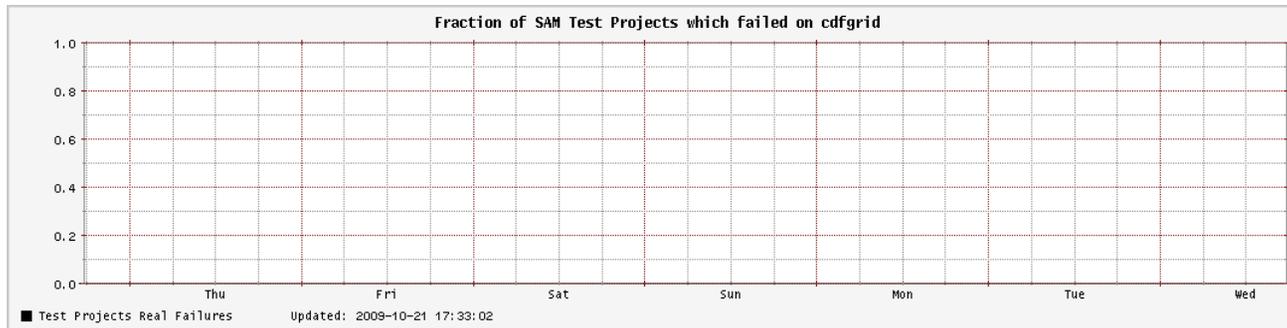
M. Mengel, R. Illingworth, SAM Shifters)

- ▶ Runs very smoothly but is serious bottleneck as mentioned.
- ▶ Fantastic success in eliminating failures
- ▶ Work now to unify Data and Job handling by combining to one subgroup.
- ▶ Disk Pool
  - B Pools: 7.2TB across 4 pools
  - Raw Pools: 22.8TB across 10 pools
  - read Pools: 330.6TB across 104 pools

June:



Now:



# Code Management

(J. Bellinger, S. Rolli, L. Garren, D. Torretta)

- ▶ Biggest concern getting SL4 and SL5 working
  - Can run SL5 on executables built on SL3/4
  - In final steps to build SL5.

## Dangers

- ▶ Reboots for Linux security kernel patches: ongoing work to minimize disruption.
- ▶ Kerberos Certificate change: Should be ok for Mid November and a drop dead date of December 1.

## Budget Strategy

- ▶ Buy needed tape: easy to budget for raw data, production, ntuples. MC is less predictable: watch demands with Physics coordinator, MC rep group.
- ▶ Disk and server upgrades complete through 2011.

## Major FY2009 Fermilab Expenditures

Estimate from last year in ( )

### ▶ Equipment

1. CPU: \$120(130)k to replace retirements
2. Disk: \$395(390)k to replace retirements (small increase in volume)  
→ (\$140)k cache + (\$210)k user scratch and project disk + (\$40)k ntuple servers
3. Servers: (\$160)k to replace retirements  
→ Includes \$(50k) for production DB server + \$(70k) for service node replacements

### ▶ Operating

1. Tapes \$270(240)k including \$((82)K from KEK) → \$(180)k for migration from 9940B tapes (2PB)

2. Data storage: Covered from 2008 purchase ( \$(130k) for robot maintenance) → \$(68k) went away with retirement of 9940B tape library

▶ Total computing expenditures: ~\$1.3(1.4)M

## Major Items in FY2010 Fermilab Budget Request

### ▶ Equipment

1. CPU: \$496k to replace retirements + \$19k for server replacements
2. Disk: \$60k to replace retiring cache disk → \$140k cache + \$303k to replace retiring project disk while providing for a modest expansion of capacity to cover growth in the volume of data
3. Servers: \$39k for production, \$10k for Database Replicas
4. Data storage: \$128k tape drives + slots in tape library

▶ Operating

1. Tapes \$168k Includes 1.5 PB/year + completion of 9940B to LTO4 migration + start of LTO3 migration
2. Data storage:\$144k tape library maintenance, data migration operating.

- ▶ Total computing request: ~\$1.6M, \$200K more than last year due to need to replace 420 nodes that retire this year.

## Request to IFC

- ▶ Maintain expand offsite commitment, confirm and implement MOU.
  - KISTI: 10Gps, 120/420/600TB cache, 500/700/1000 kSi2k in 09/10/11, 1.2 FTE tech, 2 FTE physicist during implementation
  - LYON: 479 KSI (09) Need renewal
  - CNAF: Up to 2000 "slots"
  - MIT: 100-200 slots
  - Taiwan, Japan: PACCAF slots of order 500
  - GridKA, Karlsruhe, Germany: of order 500 slots.

## Summary

- ▶ Running smoothly: success of the physics program depends on CD and IFC support
- ▶ Almost all infrastructure good through 2011 (or more).
- ▶ Focus on DH and CAF integration and efficiency: need to exploit the underutilized resources abroad.
- ▶ Keep expert pool large, burden on experts low
- ▶ Minimize our analysis disruption, retries, shutdowns: offsite computing key.
- ▶ REX and CD providing enormous support and help solve security and other problems as our strong advocate.

# Backup

## They come and go

- ▶ Sadly:
  - Angela Bellavance has moved on after being a great hero in data handling!
  - Marian Zvada a key person in Grid helped get the B MC going before moving on
- ▶ Happily: we welcome Mark Mengel, Dima Onoprienko, Oksana Tadevosyan, Donetalla Torretta (part time, with a small component of Jonathan Lewis) and Jim Bellinger!