CDF long term data preservation

- June 19, 2012 -

S. Amerio

(University of Padova)

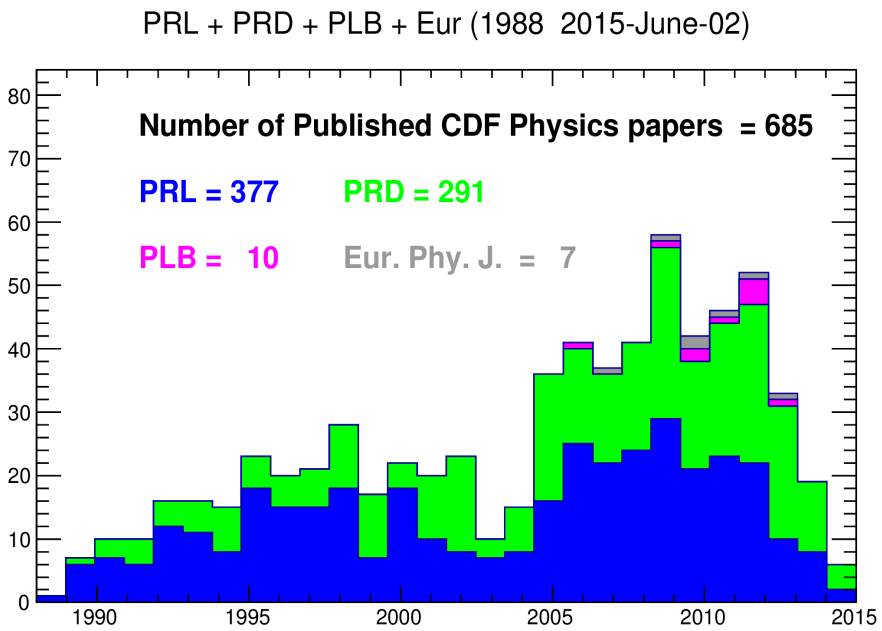
on behalf of CDF data preservation task force

Tevatron accelerator ceased operations 30 Sep 2011 (~12 fb-1 (~10 fb-1 recorded)/exp) but continued physics output:

- 2014: 20 papers
- 2015: 7 papers published/submitted
- Several analysis still ongoing (~ 10)

→ expect long tail.

- From 50 to 100 active collaborators



Tevatron Run II data are unique

- Unique initial state vs LHC
- Multiple energy collisions (300 GeV, 900 GeV, 1960 GeV)

It is very unlikely that a similar sample will ever be produced → need for data and software preservation.

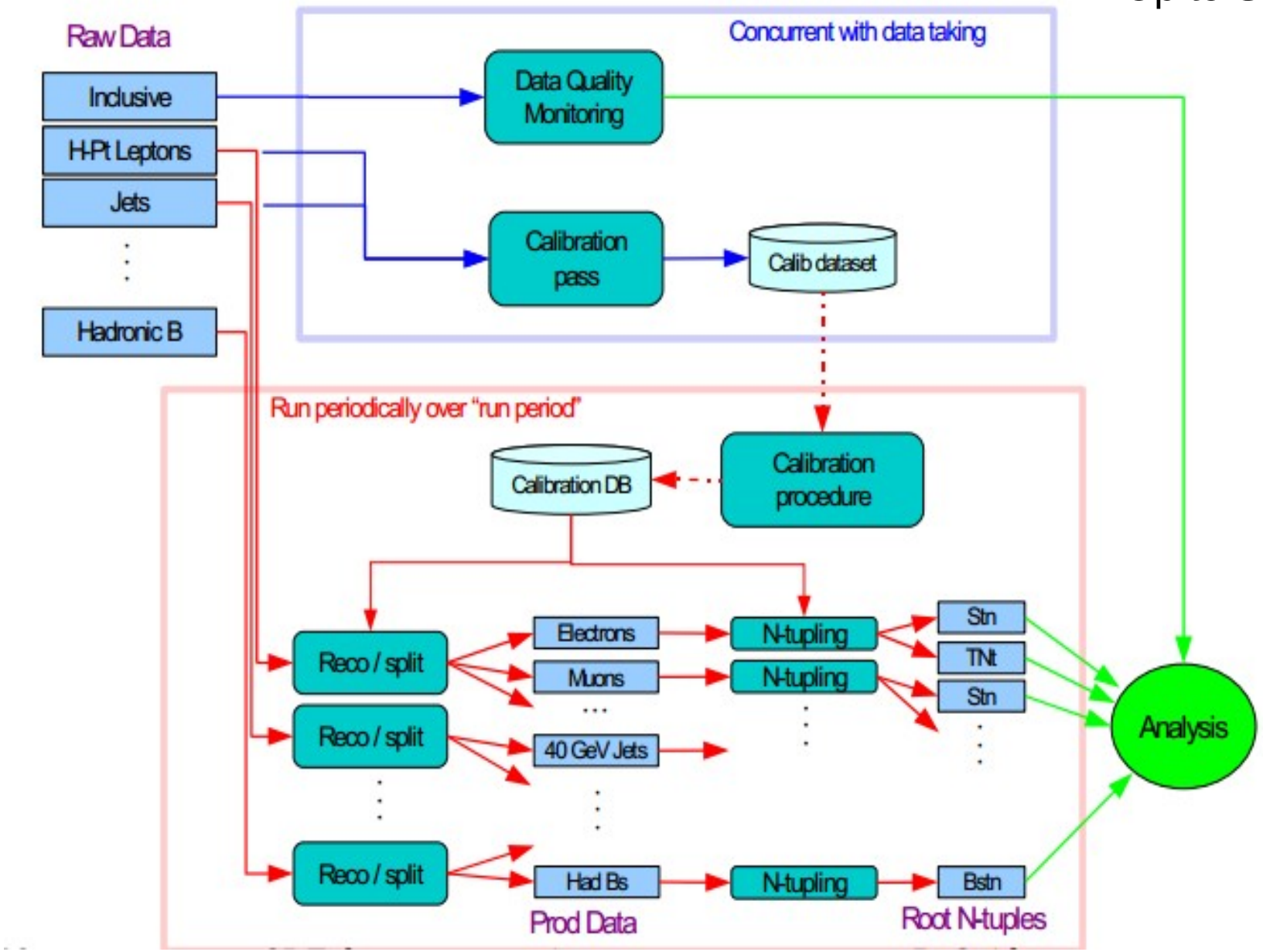
Goal: Complete analysis capability (DPHEP “level 4”) through Nov 2020 (SL6 EOL) *and beyond.*

- Includes ability to generate and simulate new MC
- All necessary documentation is preserved and accessible
- Collision data on tape remains accessible
- Computing environment for analysis is available and accessible

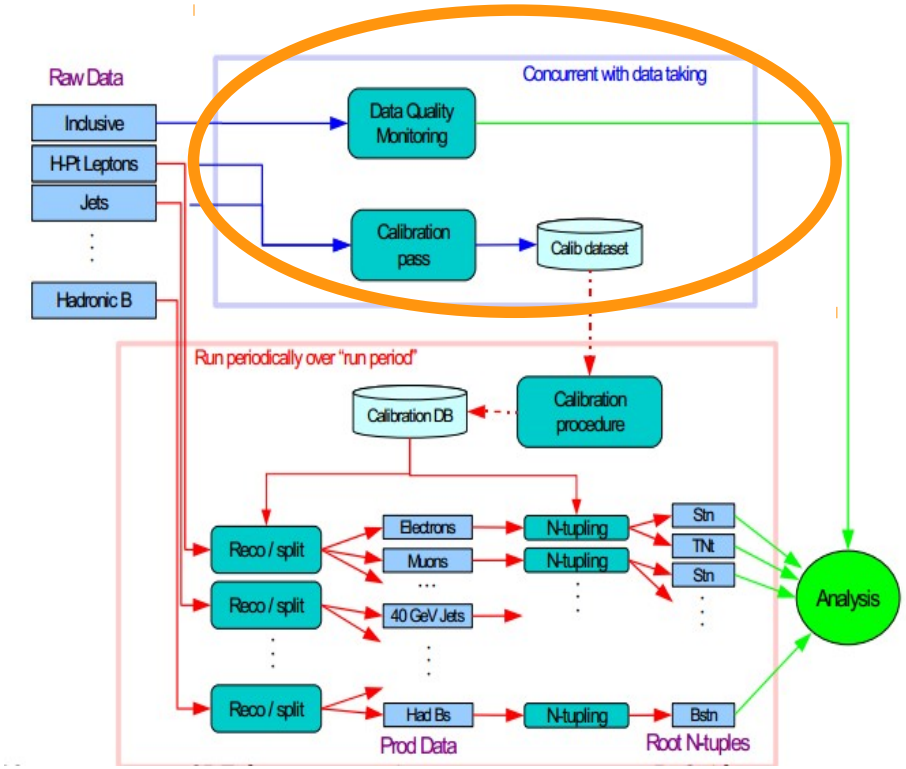
- It targets collaboration members (no discussion about open data yet)

CDF computing model

Up to September 2011



CDF computing model



All online webpages and code archived, still accessible from CDF webpages.

	Online				
Detector Operations and Shift Tools (Archived)	B0 Home				
Detector	**Detector Groups**				**Upgrades**
	Silicon / COT Rad Monitoring	Calorimeter Muon	CLC TOF	Forward Detectors	Run IIa / Run IIb
Trigger	Trigger Home	L2 / L3	Trigger WG	B Trigger	Exotics Trigger
Data quality	DQM Home	Goodrun lists	Consumer Slides	Consumer Home	Physmon

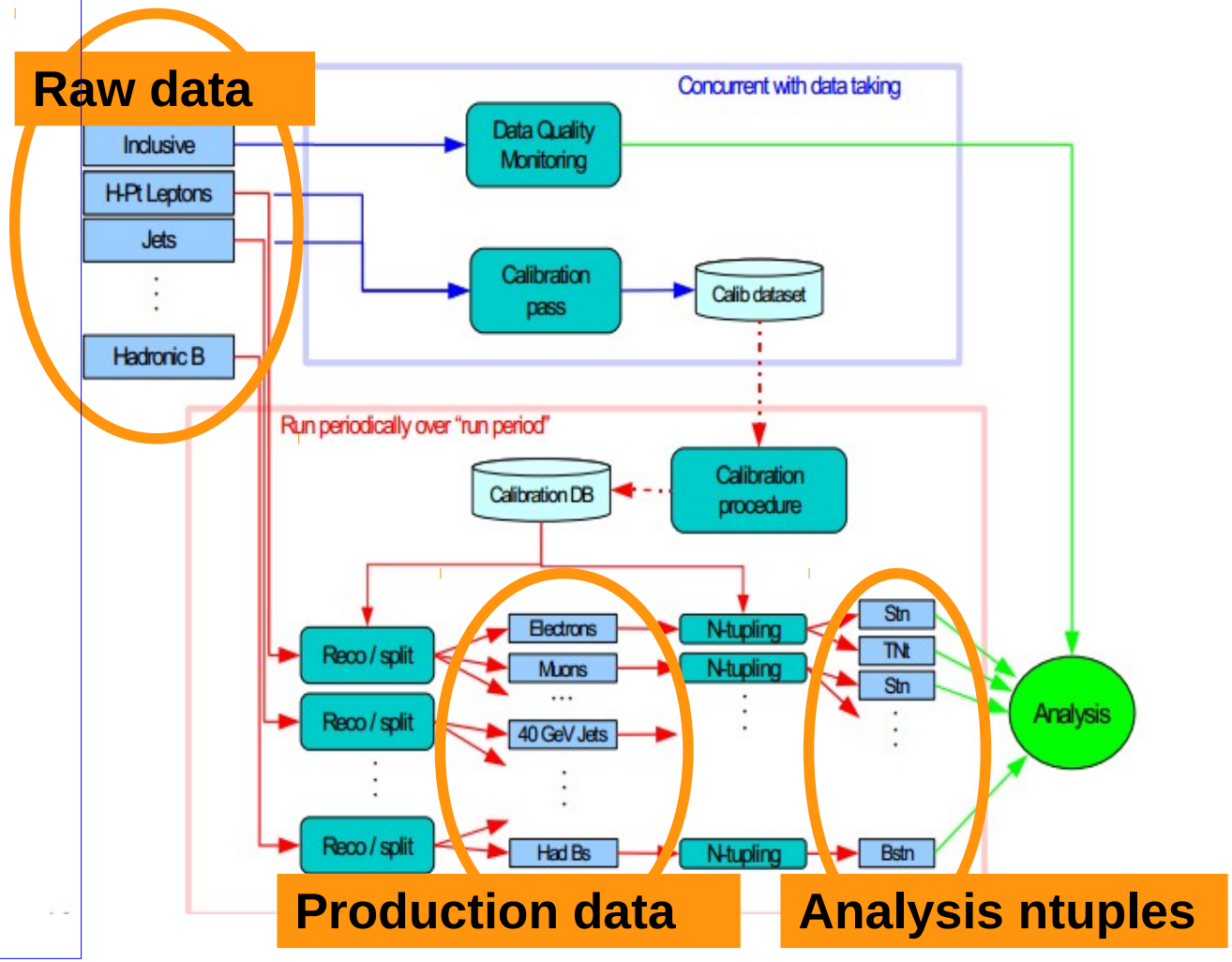
Data (bit) preservation

All raw data reprocessed with the latest version of CDF software.

Total data 10 PB (Raw + Production + Analysis data)

Bit preservation:
• all data migrated to T10k technology (2 ½ years).

Data integrity checks
- After each copy during migration
- Periodic reads from each tape



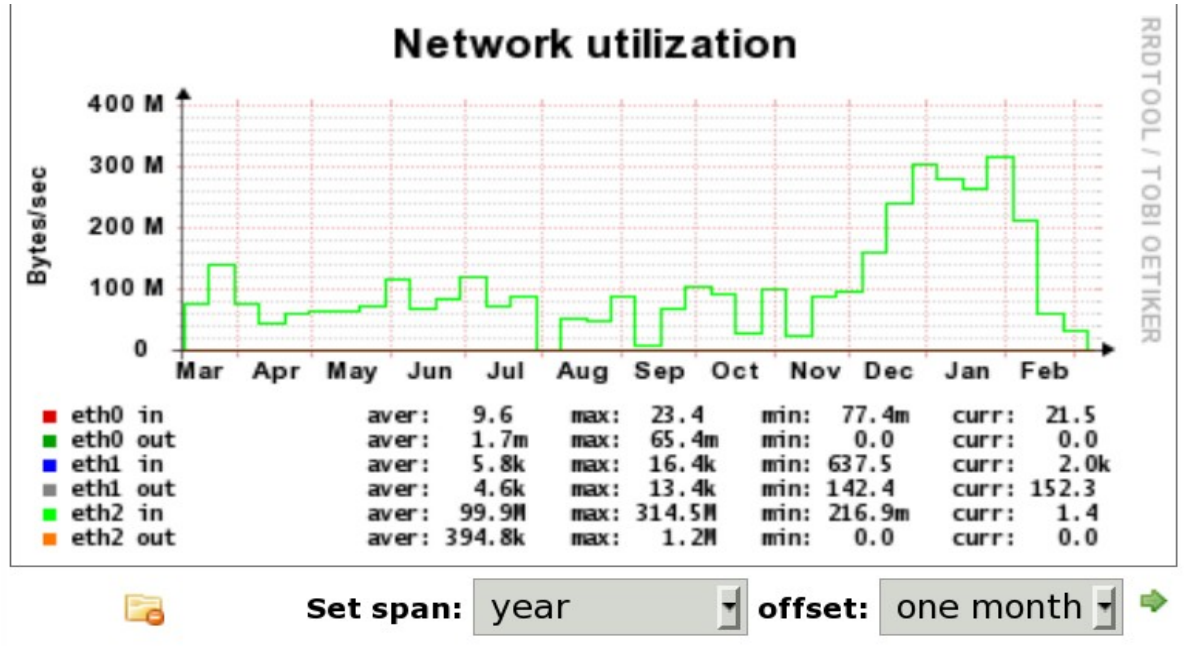
Data recovery:
- two copies of raw data at FNAL, in different locations
- In case of damage/loss analysis ntuples can be reproduced
- eventually recovered from CNAF.

CDF bit preservation @ CNAF

Long term future preservation of CDF data at INFN-CNAF, developed in collaboration with CDF and FNAL SCD.
First INFN funded project on data preservation.

A complete copy of *CDF raw data and analysis level ntuples* (4PB) is now at CNAF.

Dedicated system to transfer data; up to 5 Gb/s copy rate.
Data copy almost completed, final checks ongoing.



Data integrity and recovery: regular checks are foreseen, with dedicated resources (tape drive, disk cache).

CDF data handling

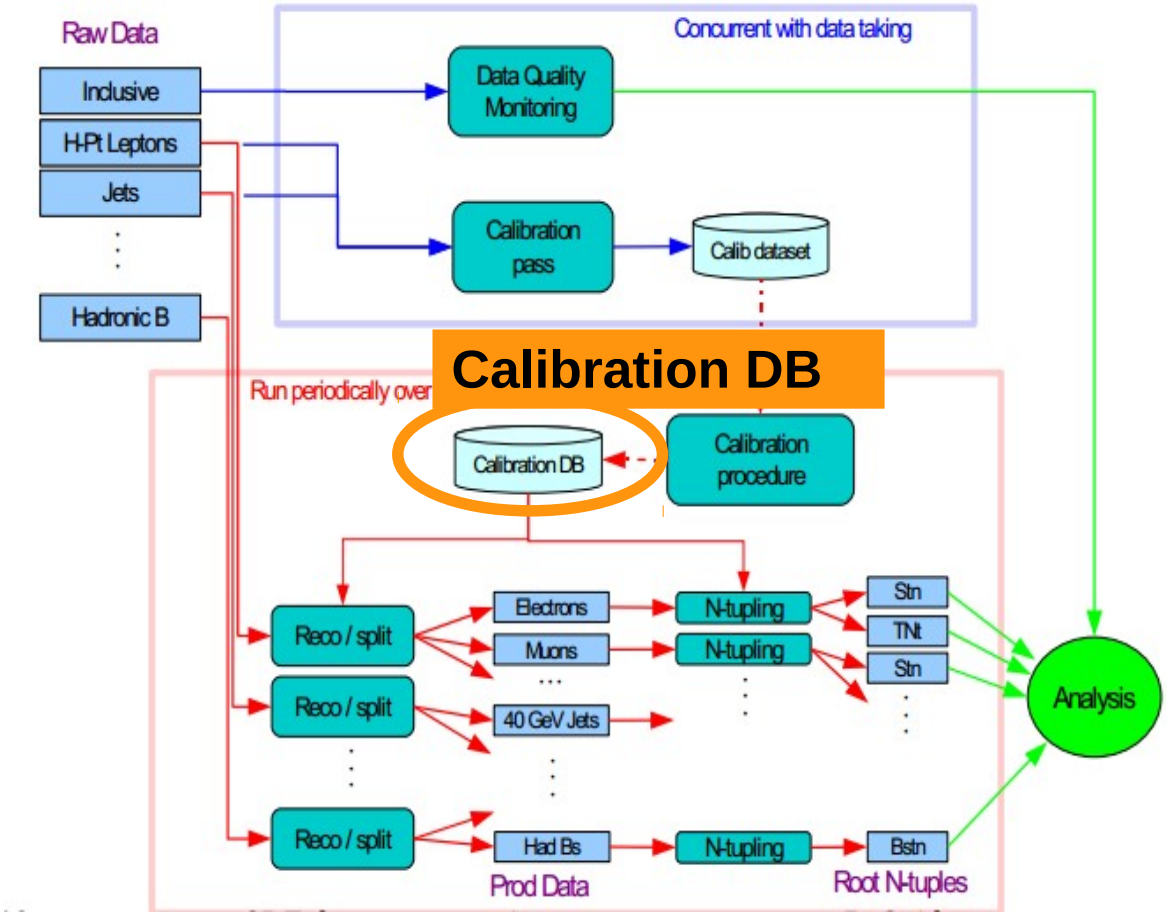
CDF data handling based on

- SAM (Sequential Access via Metadata), developed at FNAL
  - Run2 SAM version was based on CORBA software. To reduce support requirements, for new experiments, FNAL re-implemented SAM using HTTP interfaces (SAMWeb)
  - CDF code modified to interface with SAMWeb
  - Changes transparent for the users
- dCache

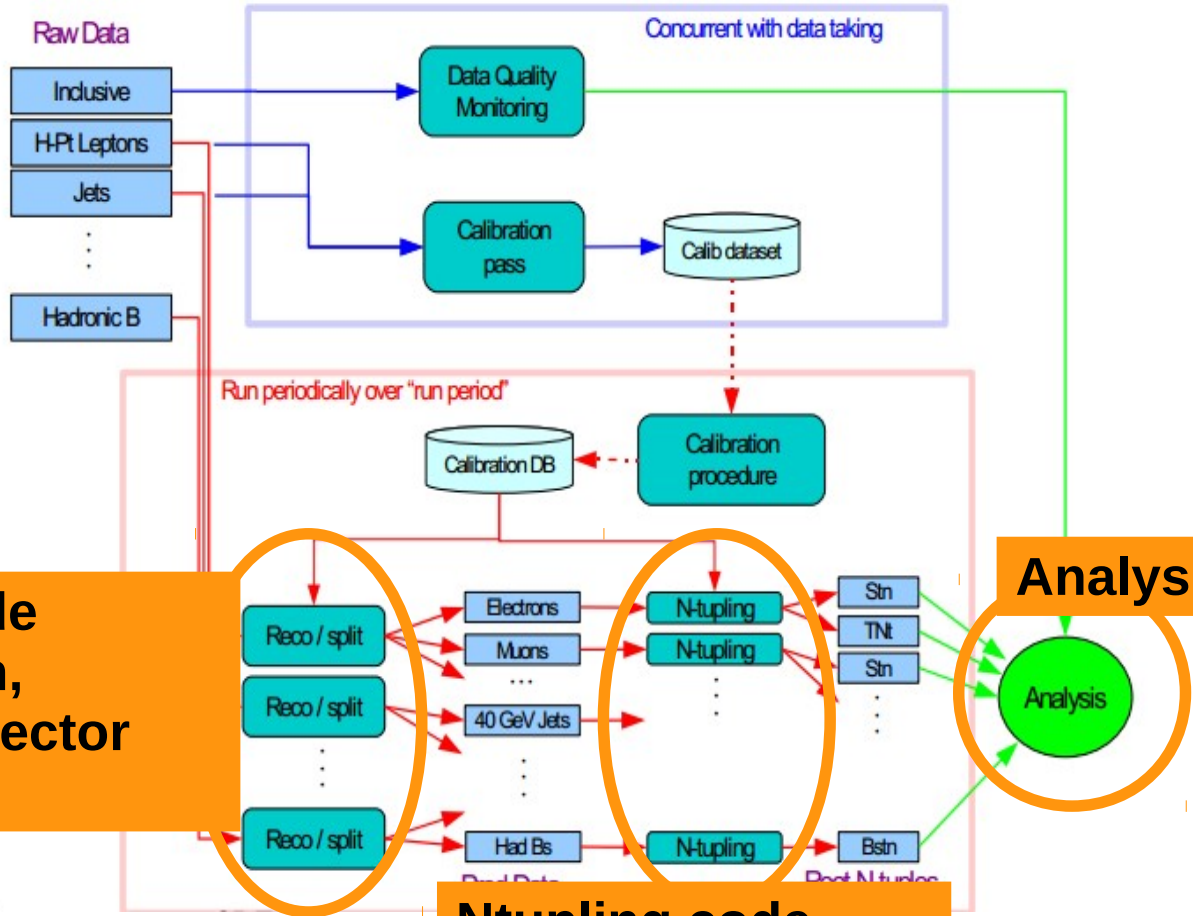
Both CDF and D0 use Oracle → licence cost is a long term future challenge.

Migration to open source db would require a considerable human effort (need to rewrite the analysis software)

Current strategy:
- upgrade to the most recent version of Oracle at the time of shutdown (2011)
- if future upgrades break the access, freeze the current db version and run in network isolation if necessary.



CDF Software preservation



Software preservation, distribution and validation

At the time of shutdown
- all code in frozen releases or in CVS repositories
- based on 32-bit frameworks built on Scientific Linux 5 (but with compatibility libraries to older OSs)

Long term future solution: build l*egacy release* that contains no pre-SL6 libraries
- Build and test completely on SL6, drop support for all previous releases
- Release now available for general use

CVMFS for code distribution:
- Used by other FNAL experiments →support in the long term future
- Can be distributed to computing centers outside FNAL (e.g. MIT, CNAF)
- User setup scripts modified → transparent for users

Validation suite implemented in a simple web accessible package that tests functionality of supported components.

CDF MC production

MC production and upload made easier → MCMaster tool.

It consists of a set of scripts and control file which combine MC production, ntupling and bookkeeping into a single job.
The job can be run locally or on the grid.

In the control file user specifies
- generator
- analysis ntuple format (three “flavours” at CDF, TopNtuple, BstnNtuple, StnTuple)
- run mode (test or full job)
- output location

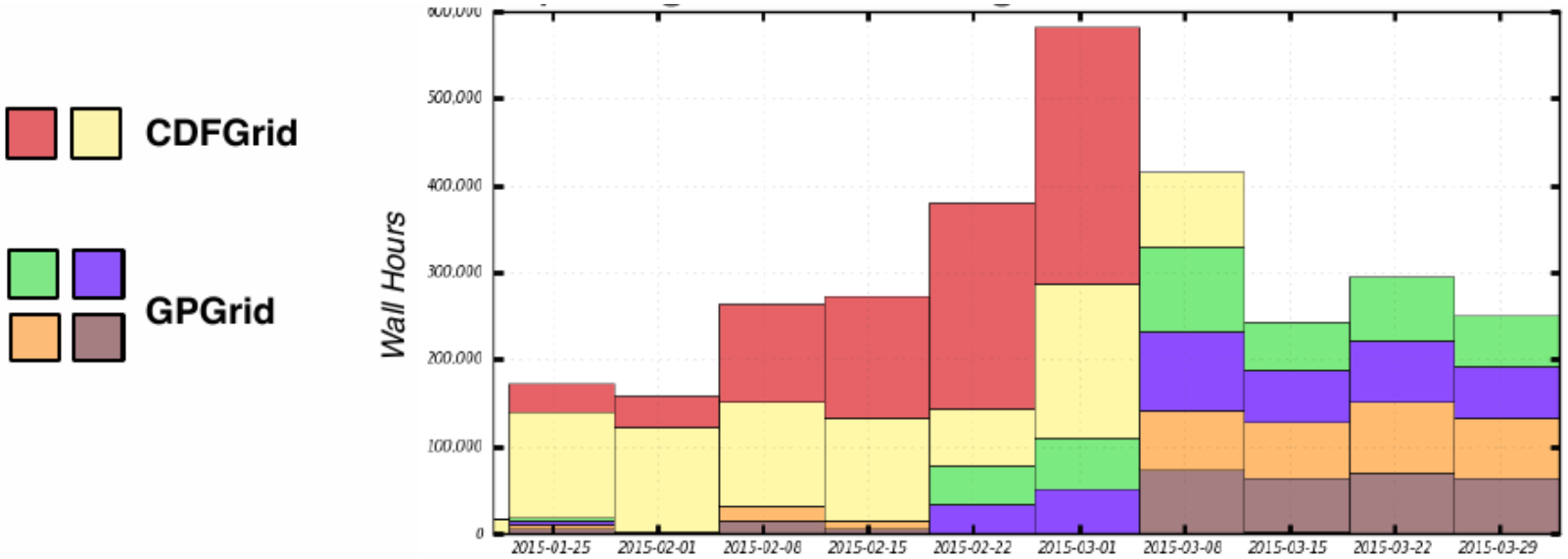
The bookkeeping is handled by Mcmaster.

CDF distributed computing

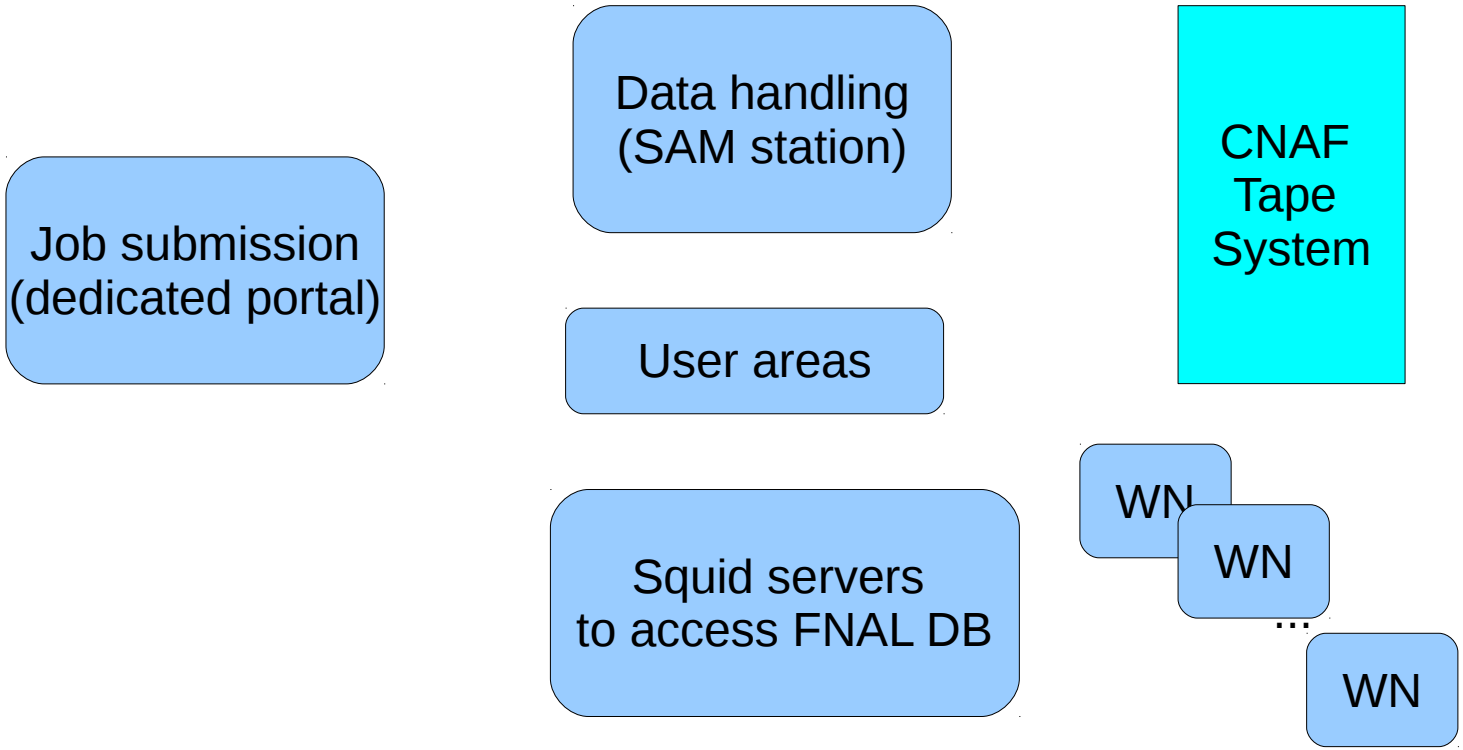
During Run II CDF maintained
- a dedicated cluster “CDFGrid” at FNAL
- OSG and LCG sites outside FNAL

CDFGrid shut down on 1-Mar-15 and all CDF jobs now sent to “GPGrid” cluster shared with other FNAL experiments

Adopted the job submission tool used by other FNAL experiments, jobsub → long term support.
Implemented a wrapper to emulate previous job submission commands → almost same commands for users, easy migration to the new system

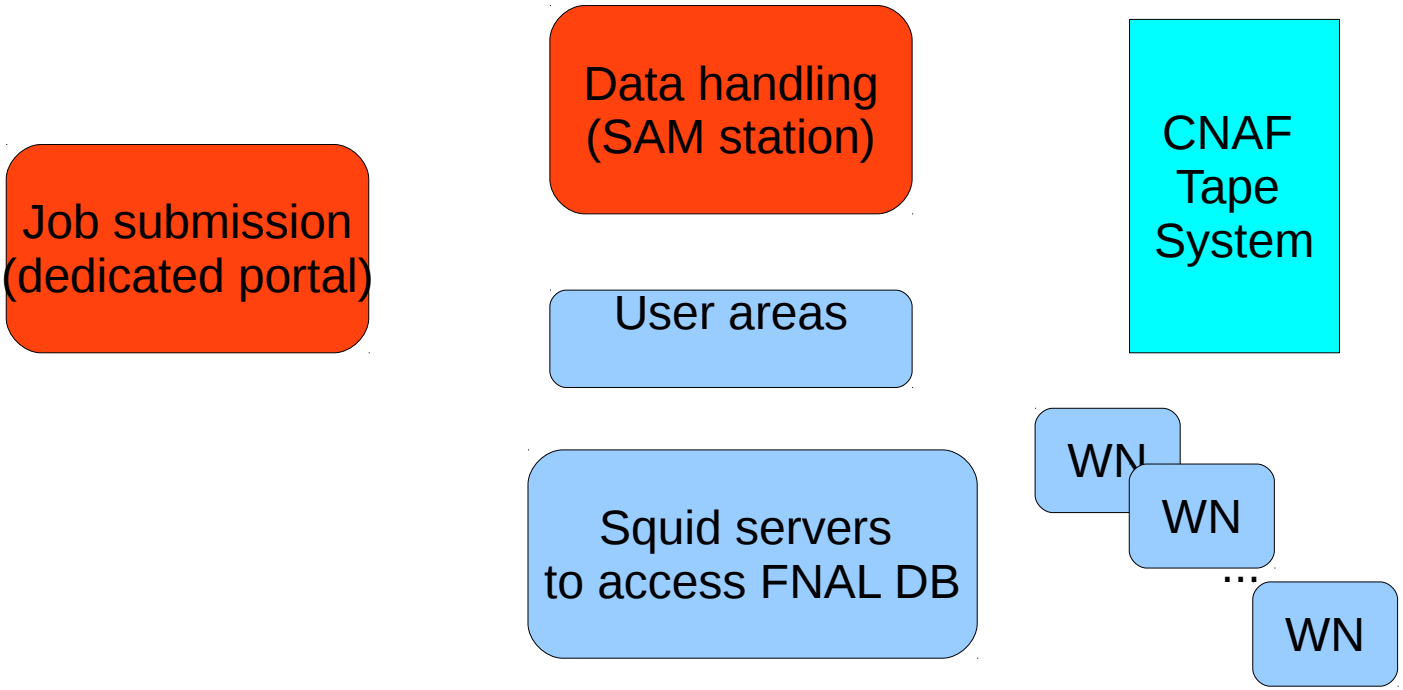


Current CDF computing services @ CNAF
All virtual machines.



CDF software now accessible at CNAF via CVMFS (previously was AFS)

Ongoing work:
- Upgrade current SAM station to SAMWeb
- Install jobsub
- CDF users will select the computing resources for their job (FNAL, MIT, CNAF) via a jobsub parameter.



Database: curently accessing FNAL DB. In the long term future CNAF aims at having a replica. Costs and implementation under discussion.

CDF Documentation

CDF notes (~ 11k) archived to INSPIRE

Two collections, for public and internal notes respectively.

Experiment-specific account for note upload and internal notes access.

CDF webpages updated with detailed instructions to use the legacy computing model.

MySQL Dbs (notes, talks, CDF members) virtualized.

Twiki/wiki pages moved to static web pages.

Web page at CNAF with
- List of all datasets on tape/disk at CNAF
- Instructions to access data and run CDF code

Support

At FNAL:

Users can request support through the Computing Service Desk interface.

Infrastructure issues → Computing Division personnel.

Physics-analysis software issues → CDF

At CNAF:

Currently 0.8 FTE dedicated to data preservation

Infrastructure issues → CNAF personnel (via GGUS ticket)

Physics-analysis software issues → CDF (via FNAL Computing Service Desk)

Summary

CDF Run II data preservation project at FNAL completed → *it ensures full analysis capability up to 2020 (and beyond)*

Mirror archival at CNAF under completion.

• Long term support thanks to the migration to tools and software used by other experiments
• Weak points: Oracle; possible loss of knowledge as collaboration naturally decreases

The project target collaboration members, no plan for open data yet.

- Backup -

CDF and D0 tape migration