CDF long term data preservation

- November 19, 2012 -

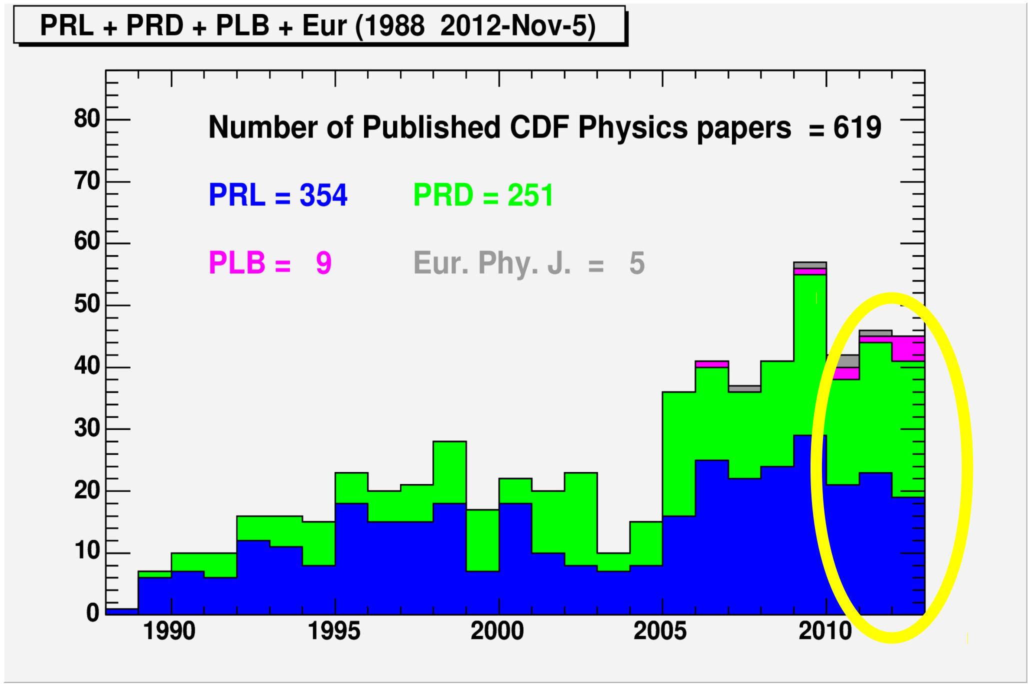
S. Amerio

(INFN Padova)

on behalf of CDF data preservation task force

Introduction

More than one year since the end of the operations.



CDF analysis plan:
- complete and publish current ongoing analysis with the full data sample
- Analysis with legacy potential:
  - QCD and EWK measurements @ 2 TeV
  - QCD analysis of the energy scan
  - t-ttbar and b-bbar asymmetry
  - W boson and top quark mass
  - ….

2011: 46 papers
2012: 40 papers

A lot of activity foreseen in the next few years.
But Tevatron data will remain unique and of great interest in the long term future.

CDF Long Term Data Preservation Project

Goal: preserve CDF data and analysis capability in the long term future (> 10 years from now)

We aim at Level 4 preservation: *full analysis chain* capability

- Data and technology to access it
- Analysis code
- Computing resources
- Knowledge!

A CDF long term data preservation task force is official since June 2012.

From our charge:
"... Production of physics results should not be more difficult in the future than it is at present despite the expected diminishing support and reduced availability of expert advice...."

Our target: physicists from CDF collaboration (and beyond !)

Task force

Current members:

- Bo Jayatilaka *(coordinator), S.Amerio (co-coordinator)*
- *Ray Culbertson*
- *Rick Snider*
- *Steve Wolbers*
- *Tyler Parsons*
- *John Strologas*
- *Donatella Torretta*
- *Physics groups conveners are ex-officio members*
- *Physics groups representatives:*

  B: Satyajit Behari

  QCD+EWK*:* Niccolo Moggi

  Higgs: Craig Group

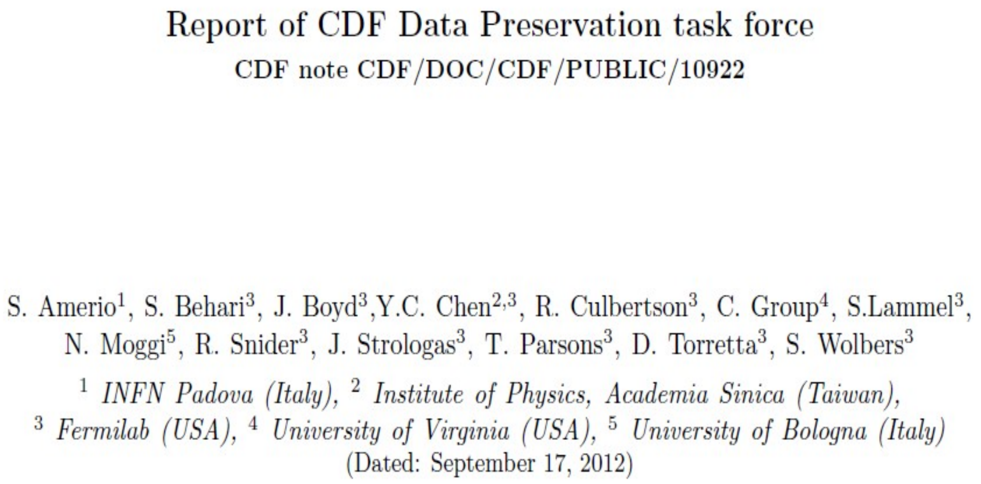
  Top: Yen-Chu Chen (all leptonic analysis), Jon Wilson (semi-leptonic analysis)

Task force activity and organization

An intense summer...

- Regular meetings every two weeks.
- Mid-August: Joint CDF/D0 meeting
- End of August: Report to Fermilab management

First report released on Sept.17th.



Report of CDF Data Preservation task force

CDF note CDF/DOC/CDF/PUBLIC/10922

S. Amerio[1], S. Behari[3], J. Boyd[3],Y.C. Chen[2,3], R. Culbertson[3], C. Group[4], S.Lammel[3], N. Moggi[5], R. Snider[3], J. Strologas[3], T. Parsons[3], D. Torretta[3], S. Wolbers[3]

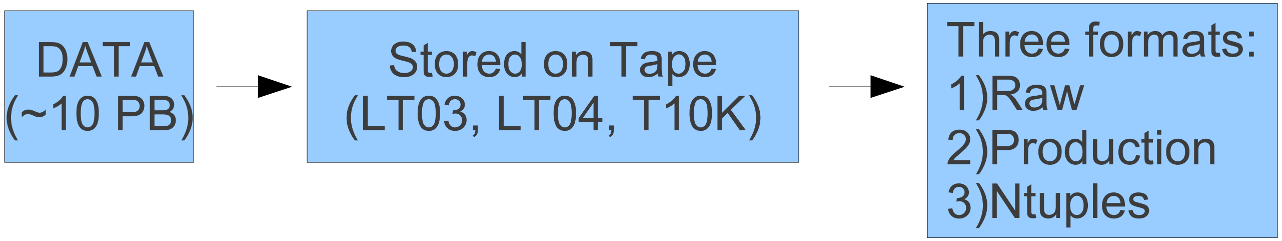
[1] *INFN Padova (Italy),* [2] *Institute of Physics, Academia Sinica (Taiwan),* [3] *Fermilab (USA),* [4] *University of Virginia (USA),* [5] *University of Bologna (Italy)*

(Dated: September 17, 2012)

In the report we identified requirements and possibile solutions for:
- *data access*
- *software preservation*
- *job submission*
- *documentation*

We are designing the project looking for
• The best compromise between available manpower from the collaboration and essential needs for analysis in the long term future
• Strict collaboration with D0 and FNAL Computing Sector.
• Feedback from DPHEP and other experiments

The starting point: CDF data



DATA (~10 PB) → Stored on Tape (LT03, LT04, T10K) → Three formats:
1)Raw
2)Production
3)Ntuples

Physical tape

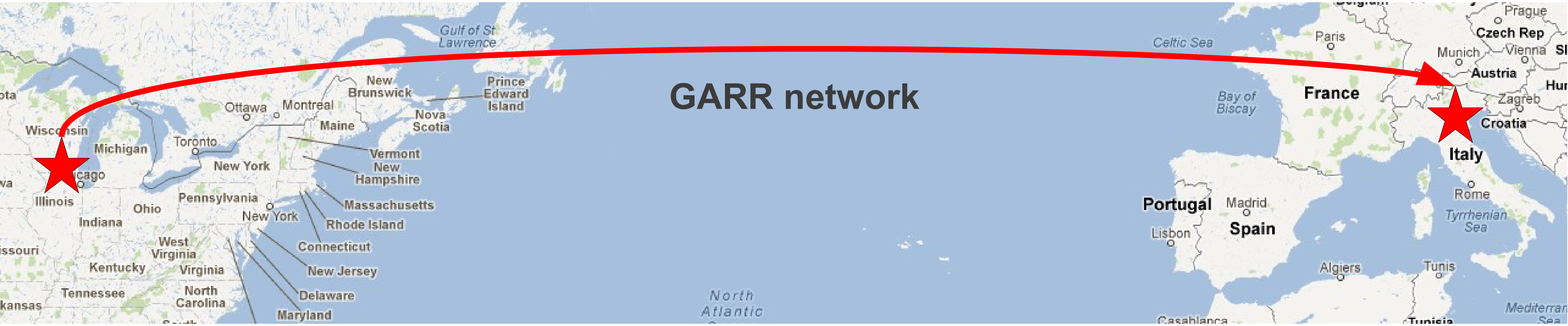
- Data will require to be regularly migrated to new tape technology
- Complete migration of LT03/LT04 data to T10K in the next two years.

Data redundancy

- Production data contains a copy of raw data
- A project to have a complete copy of CDF data offsite is being developed in collaboration with INFN

CDF data offsite: FNAL → CNAF

CDF Italian collaboration, together with CNAF computing center and FNAL, is developing a project to preserve at CNAF a copy of CDF data (raw and ntuples) and the analysis capabilities in the long term future.



- Use GARR network (Italian R&E network) for the copy
- Adapt the current copy mechanism between FNAL and CNAF
- First feasibility tests performed during the summer → successful copy retrieval of a small dataset to CNAF tape system and retrieval.

- The project got a first approval by INFN at the end of September; final approval is expected in December.

Data access

Data handling system based on

- **SAM (Sequential Access via Metadata)**
  - developed at Fermilab, used by D0 and CDF, it closely integrates with Fermilab Enstore tape manager and other data handling systems (e.g. dCache)
- **dCache**: it fetches files requested by the users and stores them on a distributed pool of disk servers (800 TB) for the user to access over the network.

Data access in the long term future:

We have to maintain SAM; two options under study:

1. Keep the current system as it is now
2. Update to the new SAM being developed for the Fermilab experiments

Option 2. requires more effort in the short time, but will ensure support in the long term future.

Other types of data: metadata in databases

Databases:

Oracle for

- run condition, configuration, trigger, luminosity, alignement, calibration → *necessary for new MC generation;*

- data metadata (dataset, fileset, file and run section information) → *essential part of SAM access method*

Long term future:

We prefer to keep Oracle; not enough manpower to migrate to a different DB and perform all the validation.

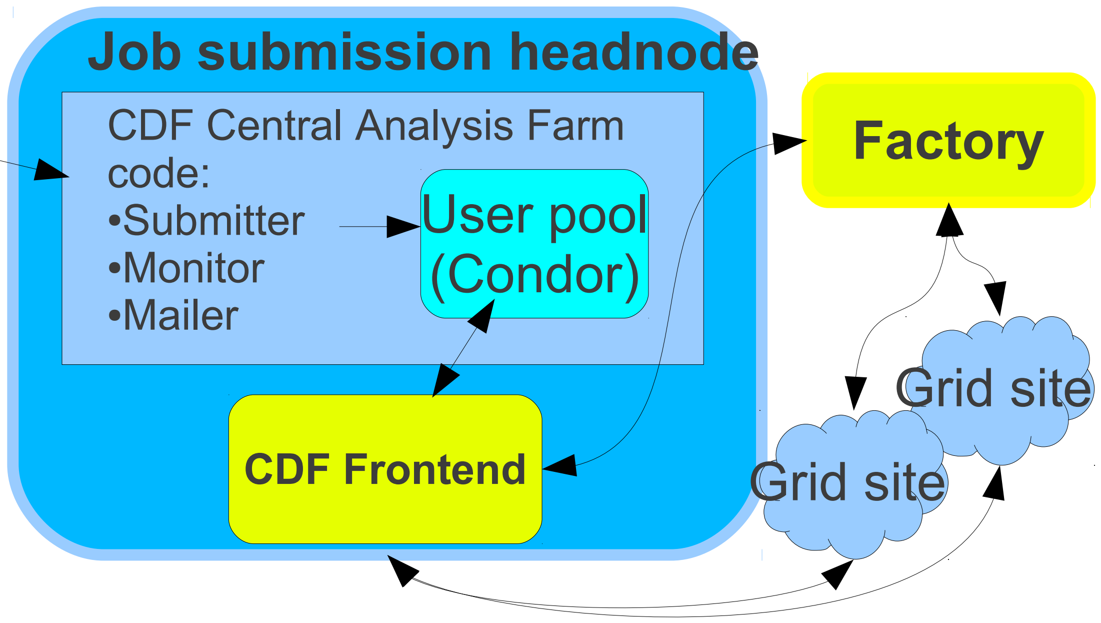
FNAL will keep using Oracle in the future; we may need to migrate to new Oracle versions in the future.

Experiment software

- CDF-supported software is archived as a set of packages in a **CVS repository:**
  - Fermilab is committed to preserve the content of thre repository indefinitely
  - Migration to SVN under investigation
  - ALL analysis code has to be archived (some still on user's desktops)
- Build configuration toolkit based upon **SoftRelTools** and **ups** products supported by Fermilab.
- **SL5; new SL6 version in 2013.**
  - Likely the port to SL6 will be the last
  - We are considering **virtualization** to preserve the existing build and run-time platforms
- A well documented and tested **validation procedure** is needed

Job submission

- CDF *Central Analysis Farm code (CAF)* provides the users with a uniform interface to resources on different Grid sites



- Three *portals* to access computing resources:
- CDFGrid → FNAL
- NamGrid → OSG
- Eurogrid → Tier1 @ CNAF and LCG

Based on *glideinWMS* workload management system (batch system = *Condor*)

Job submission in the future

In the near term future (< 5 years) the current system can be maintained with minimum effort.

Realistically, in the long term future access to CDF data will be greatly reduced; we need to move to a job submission system simple, flexible and easy to support.

Two options under study:

1) Development of a system which aligns with and integrates with the job submission system being developed for the new Fermilab experiments:

- We will take full advantage of developments and support.
- This solution requires consistent effort to adapt the current system.
- Security issues need to be addressed.

2) Build a virtualized system which can run in isolation

- No security issues.
- It requires dedicated support and maintenance.
- It may not be justified by the CDF analysis load.

Job submission and data analysis at CNAF: WNoDeS

WNoDeS → Worker Nodes on Demand Service

Virtualization architecture which provides transparent user interfaces to Grid, Cloud and local access to resources

In **production at several Italian centers, including the INFN Tier-1 since November 2009** (Currently managing about 2000 on demand Virtual machines there)

New feature under development: Dynamic virtual networks:
dynamic instantiation of **private** VLANs and address assignement for VM isolation.

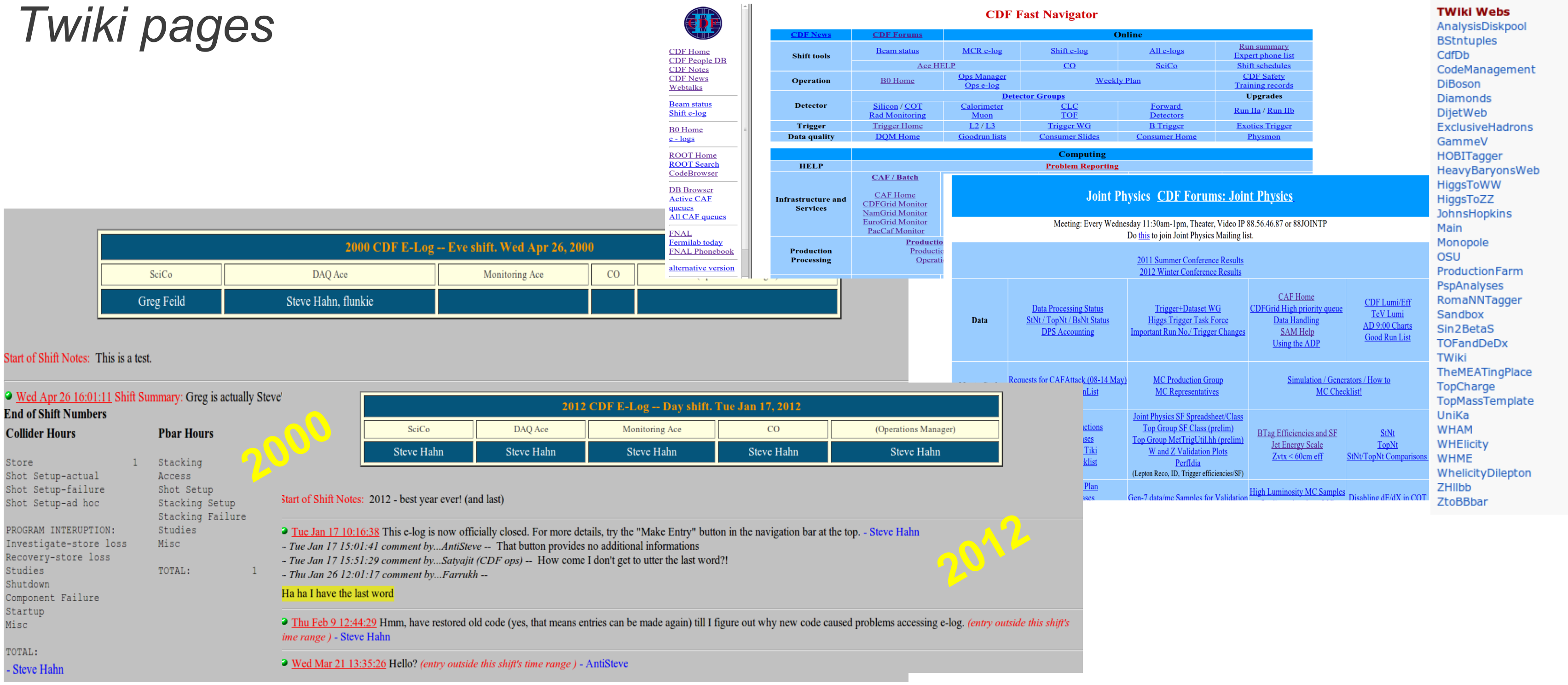
↓

In the long term future at CNAF: CDF services and analysis computing resources can be instantiated on demand on pre-packaged VMs in a controlled environment.

Documentation

We need to carefully **preserve and re-organize** as much documentation as possible

- *Internal webpages*
- *Webtalks pages*
- *CDF notes archive*
- *Logbooks*
- *Twiki pages*



CDF Home
CDF People DB
CDF Notes
CDF News
Webtalks

Beam status
Shift e-log

B0 Home
e - logs

ROOT Home
ROOT Search
CodeBrowser

DB Browser
Active CAF queues
All CAF queues

FNAL
Fermilab today
FNAL Phonebook

alternative version

CDF Fast Navigator

CDF News	CDF Forums	Online				
Shift tools	Beam status	MCR e-log	Shift e-log	All e-logs	Run summary Expert phone list
	Ace HELP		CO	SciCo	Shift schedules
Operation	B0 Home	Ops Manager Ops e-log	Weekly Plan		CDF Safety Training records
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

	Computing
HELP	Problem Reporting
Infrastructure and Services	CAF / Batch CAF Home CDFGrid Monitor NamGrid Monitor EuroGrid Monitor PacCaf Monitor
Production Processing	Productio Productio Operati

Joint Physics CDF Forums: Joint Physics

Meeting: Every Wednesday 11:30am-1pm, Theater, Video IP 88.56.46.87 or 88JOINTP
Do this to join Joint Physics Mailing list.

2011 Summer Conference Results
2012 Winter Conference Results

Data	Data Processing Status StNt / TopNt / BsNt Status DPS Accounting	Trigger+Dataset WG Higgs Trigger Task Force Important Run No./ Trigger Changes	CAF Home CDFGrid High priority queue Data Handling SAM Help Using the ADP	CDF Lumi/Eff TeV Lumi AD 9:00 Charts Good Run List
	Requests for CAFAttack (08-14 May) nList	MC Production Group MC Representatives	Simulation / Generators / How to MC Checklist!	
	ictions ises Tiki klist	Joint Physics SF Spreadsheet/Class Top Group SF Class (prelim) Top Group MetTrigUtil.hh (prelim) W and Z Validation Plots PerfIdia (Lepton Reco, ID, Trigger efficiencies/SF)	BTag Efficiencies and SF Jet Energy Scale Zvtx < 60cm eff	StNt TopNt StNt/TopNt Comparisons
	Plan ises	Gen-7 data/mc Samples for Validation	High Luminosity MC Samples	Disabling dE/dX in COT

TWiki Webs
AnalysisDiskpool
BStntuples
CdfDb
CodeManagement
DiBoson
Diamonds
DijetWeb
ExclusiveHadrons
GammeV
HOBITagger
HeavyBaryonsWeb
HiggsToWW
HiggsToZZ
JohnsHopkins
Main
Monopole
OSU
ProductionFarm
PspAnalyses
RomaNNTagger
Sandbox
Sin2BetaS
TOFandDeDx
TWiki
TheMEATingPlace
TopCharge
TopMassTemplate
UniKa
WHAM
WHElicity
WHME
WhelicityDilepton
ZHllbb
ZtoBBbar

2000 CDF E-Log -- Eve shift. Wed Apr 26, 2000

SciCo	DAQ Ace	Monitoring Ace	CO	
Greg Feild	Steve Hahn, flunkie			

Start of Shift Notes: This is a test.

Wed Apr 26 16:01:11 Shift Summary: Greg is actually Steve'

End of Shift Numbers

```
Collider Hours                 Pbar Hours

Store                   1      Stacking
Shot Setup-actual              Access
Shot Setup-failure             Shot Setup
Shot Setup-ad hoc              Stacking Setup
                               Stacking Failure
PROGRAM INTERUPTION:           Studies
Investigate-store loss         Misc
Recovery-store loss
Studies                        TOTAL:          1
Shutdown
Component Failure
Startup
Misc

TOTAL:
```

- Steve Hahn

2000

2012 CDF E-Log -- Day shift. Tue Jan 17, 2012

SciCo	DAQ Ace	Monitoring Ace	CO	(Operations Manager)
Steve Hahn	Steve Hahn	Steve Hahn	Steve Hahn	Steve Hahn

Start of Shift Notes: 2012 - best year ever! (and last)

Tue Jan 17 10:16:38 This e-log is now officially closed. For more details, try the "Make Entry" button in the navigation bar at the top. - Steve Hahn
- *Tue Jan 17 15:01:41 comment by...AntiSteve --* That button provides no additional informations
- *Tue Jan 17 15:51:29 comment by...Satyajit (CDF ops) --* How come I don't get to utter the last word?!
- *Thu Jan 26 12:01:17 comment by...Farrukh --*

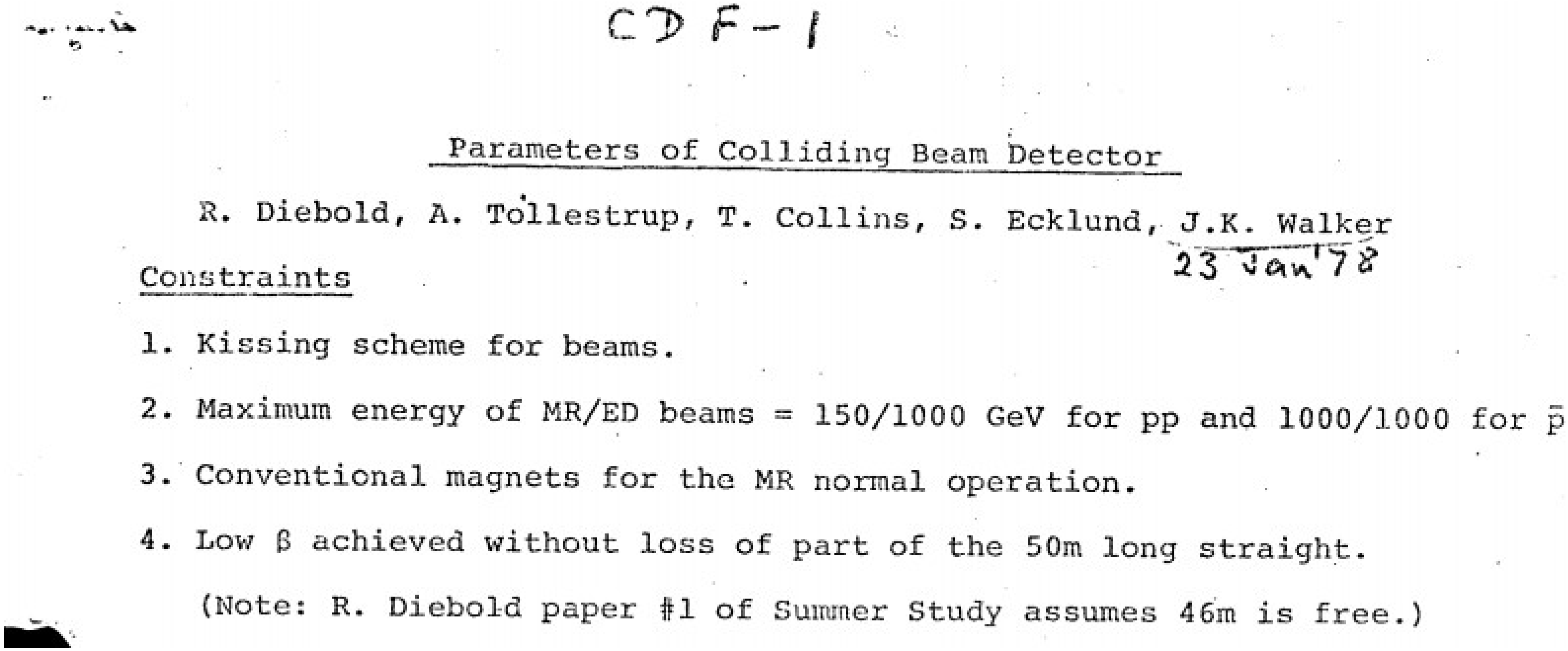
Ha ha I have the last word

Thu Feb 9 12:44:29 Hmm, have restored old code (yes, that means entries can be made again) till I figure out why new code caused problems accessing e-log. *(entry outside this shift's time range)* - Steve Hahn

Wed Mar 21 13:35:26 Hello? *(entry outside this shift's time range)* - AntiSteve

2012

Archived the first 1000 CDF internal notes!



CDF-1

Parameters of Colliding Beam Detector

R. Diebold, A. Tollestrup, T. Collins, S. Ecklund, J.K. Walker

23 Jan '78

Constraints

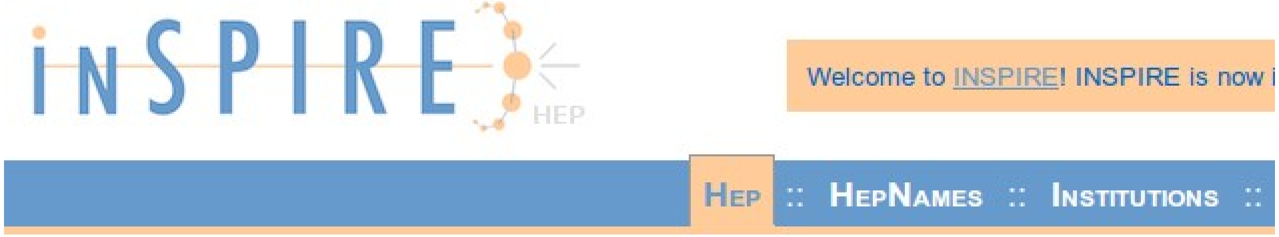
1. Kissing scheme for beams.
2. Maximum energy of MR/ED beams = 150/1000 GeV for pp and 1000/1000 for $\bar{p}$
3. Conventional magnets for the MR normal operation.
4. Low β achieved without loss of part of the 50m long straight.

(Note: R. Diebold paper #1 of Summer Study assumes 46m is free.)

1000 notes from the Berkeley collection *scanned and uploaded* to our archive.

Thanks to Lina Galtieri, Stephanie Schuler and Barb Hehner!

CDF internal notes on Inspire



Welcome to INSPIRE! INSPIRE is now i

- 11000 internal notes, from 1978!
- We will archive them in Inspire
- Still in the testing phase: a first set of notes will be uploaded in the next month
- If no major issues, all notes arvchived by next Spring.

Many thanks to Zaven Akopov, Heath O'Connell and Alan Jonckheere.

New CDF website

CDF website **www-cdf.fnal.gov**

- Public section
- Private section
  - Online
  - Computing
  - Physics (with sub-pages for each physics groups)
  - Organization (CDF organization, meetings, webtalks, internal notes, ...)

CDF online webserver **www-cdfonline.fnal.gov**

- Data acquisition and sub-detector e-logs, details about detector systems, daq and trigger operations

Re-organize all the essential information in a new web-site

Target: a physicist, non necessarily from CDF
Goal: he/she has to be able to perform an analysis from the very beginning
Timeline: 2 years
How: 1) identifiy representatives for the different sections 2) identify well defined tasks that should be considered as service tasks

Conclusions and next steps

The CDF data preservation task force is official since June 2012.

First report released In September 2012.

We have identified necessary requirements to preserve in the long term future CDF analysis capabilities.

Preservation of the documentation has already started (high priority, given the natural reduction of the collaboration).

CDF/D0 requirements are being discussed with Fermilab Computing Sector experts; both task forces will be part of the Fermilab data preservation project team (details in Steve's talk).

- Backup -