CDF Legacy Data Preservation Project

Joint FNAL-SLAC-DESY Data Preservation Meeting
28 Mar 2014

Bo Jayatilaka,
Willis Sakumoto, and
CDF Data Preservation Group

- Project goal
  - Hand-off of our legacy CDF analysis/documentation infrastructure to FNAL Scientific Computing Division (FSCD) operations
- Current integration into FSCD
  - Maintenance of Run II data, processed data, and simulated data
    - Copy to T-10K archival tape: 95% complete
    - SAM infrastructure for access/storage: local and FermiGrid
  - Generic operations
    - Interactive login VMs, data, simulated data servers
    - Access to GRID computing and other such computing services
    - Calibration database and associated servers
      Database is essentially frozen
      Migration to latest Oracle DB release: complete and validated
    - CDF code file system and its CVMFS subset
- Archive of CDF Run II data outside of FNAL: INFN/Padova, S. Amerio
  - Hardware and software infrastructure: up and running since Feb 2014
  - Transfer rate 5 GB/sec, 700 TB on tape, 240 TB in disk queue

- CDF legacy software and documentation
  - Current effort: software preservation and validation, finish this by May
  - Documentation has been languishing, but will pick up in a few months
- CDF software preservation and validation general goals
  - Validate current SL5/ROOT4 software on SL6
  - ROOT4 → ROOT5 for the archival legacy code
  - Prepare/validate archival legacy code for long term support
    CDF code with ROOT5 on SL5 – run on SL5/6
    CDF code with ROOT5 on SL6
    Integration into FNAL Intensity Frontier Computing infrastructure
  - Code repository
    Full legacy on FNAL file system, ie, everything we used
    CVMFS subset for running CDF code on a generic GRID
- ROOT4 → ROOT5: we settled on ROOT 5.34-12
  - Major upgrade: completed acceptance test a week ago
    ROOT4 and 5 are structurally quite different
    CDF archival files (data/simulation) are ROOT4 files
  - Previous ROOT5 versions have file streamer issues with ROOT4 files
    Checksum/ClassVersion handling – could not read CDF ROOT4 files
    Worked with ROOT developers to fix this (5.34-12)

- CDF specific software preservation
  - Current system obsolete: SL5 with gcc/++ 3.4, g77, and ROOT 4.0
  - Modernize our software
    Near term: SL5 with gcc/++/gfortran 4.1 and ROOT 5.34-12 (32bit)
    Long term: SL6 with gcc/++/gfortran 4.4 and ROOT 5.34-12 (32bit)
  - Acceptance criteria
    Run with an acceptable level of crashes with optimized binaries (-O2)
    Validation of output relative to SL5/ROOT4 output
    For now, we allow the use and access of the CDF code file system
- Short term bridge: run and validate SL5/ROOT4 code on SL6
  - Will run as is on SL6/CDFGrid but needs our legacy code file system
  - Validation of SL5/SL6 runs: Completed Mar 2014
- Near and long term status: mostly done
  - Modernization has been difficult because of two requirements
    ROOT4 → ROOT5
    gcc/++ optimization, different for SL5/3.4 → SL5/4.1 → SL6/4.4
  - To do:
    Large simulation jobs on CDFGrid for final validation
    Full integration into FNAL Intensity Frontier infrastructure

- CDF legacy file system: two separate file systems
  - Full Legacy: all Run II software, exists but too big for FNAL CVMFS
  - CVMFS version: currently working to implement this – HIGH priority
    - Needed for runs of CDF legacy code on future computing GRIDs
    - We have an empty CVMFS image for CDF
      Needed packages of the Full Legacy system have been identified
      These packages need to be put into the image
      Test image functionality on GRID nodes without CDF software
- Additional CDF work over the next few months
  - Run II documentation: within scope of current project
    - Legacy release: update and add to current CDF html pages
    - CDF internal notes: transfer to Inspire
  - CDF Run I data tapes: this is a new item – any comments?
    - Investigate condition of tapes – 8mm, non-controlled storage
    - If feasible, investigate possibility of climate controlled storage