Contribution ID: **33** Type: **Sessions contributions**

Long Term Data Preservation for CDF at CNAF

Wednesday, 21 May 2014 16:40 (20 minutes)

After the end of data taking in 2011, CDF is now facing the challenge to both preserve the large amount of data produced during several years and to retain the ability to access and reuse it in the future. The CDF italian collaboration, together with INFN-CNAF computing center, has developed and is now implementing a long term future data
preservation project. The project comprises the copy of all CDF raw data and user level ntuples (about 4 PB) at CNAF and the setup of a
framework which will allow to access and analyse the data in the long term future. In this talk we first illustrate the difficulties and the
technical solutions adopted to copy, store and maintain CDF data at CNAF. We then describe how we are exploiting virtualization techniques to build the long term future analysis framework, and the validation tests under development to check data integrity and software over the time.

Wider impact and conclusions

During the implementation of this project we are facing several issues typical of data preservation, and sharing experiences with other
experiments and laboratories, as Babar at SLAC, Desy and of course Fermilab. The project is being developed within the DPHEP collaboration. Data maintenance, validation systems, virtualization
techniques to run CDF legacy software are key areas for our project where we aim at solutions highly sustainable in the long term future, as much flexible as possible and easy to be adapted to other experiments. The project described in this contribution is the first INFN supported
project on data preservation. It will serve as a prototype for other experiments –inside and outside HEP - which are currently storing
their data at CNAF computing center.

Description of work

Total CDF collision and simulation data amount to about 10 PB. These samples are currently stored on T10K technology tapes at Fermilab and we plan to copy to CNAF all raw and ntuple data (about 4 PB) before the start of LHC data taking in 2015. To meet this tight constraint we setup a dedicated system able to transfer data at 5 Gb/s rate, and
copy it to CNAF tape system automatically updating the FNAL database. The copy is driven by CNAF using the CDF SAM data handling system: upon a request from CNAF, data are retrieved from the FNAL tape system and copied via gridftp to CNAF, where they are automatically uploaded to tape. The data storage layout consists of a pool of disks managed
by GPFS, a tape library infrastructure for the archive back-end and an integration system to transfer data from disk to tape and vice versa.
The CNAF storage solution is GEMSS, an integration of GPFS, TSM and StoRM, which is completely transparent to CDF data handling system. As
far as data analysis is concerned, CNAF already offers a set of services to analyse CDF data. Data can be accessed via SAM and stored on a dedicated cache. Users can submit their analysis jobs to LCG via a dedicated portal. CDF analysis code is accessible via AFS. All these services are replicas of CDF services at Fermilab, installed as
virtual machines on SL5 and SL6 operating systems. For the long term future, running CDF legacy code requires addressing several issues, like availability of suitable hardware resources, software maintenance and handling of computer and network security. Services used to access
CDF data be eventually migrated to a dynamic virtual infrastructure. We are implementing this infrastructure so that CDF services can be
instantiated on-demand on pre-packaged virtual machines (VMs) in a controlled environment, where in- and

out-bound access to these services and connection to storage data is administratively controlled.

Primary author: DELL'AGNELLO, Luca (INFN)

Co-authors: Dr PROSPERINI, Andrea (INFN-CNAF); Dr GREGORI, Daniele (INFN-CNAF); SALOMONI, Davide (INFN); Dr DE GIROLAMO, Donato (INFN-CNAF); Dr ROSSO, Felice (INFN-CNAF); Dr CHIARELLI, Lorenzo (GARR); Dr PEZZI, Michele (INFN-CNAF); Dr RICCI, Pierpaolo (INFN-CNAF); AMERIO, Silvia (INFN); Dr ZANI, Stefano (INFN-CNAF)

Presenter: Dr PEZZI, Michele (INFN-CNAF)

Session Classification: New data management solutions for EGI

Track Classification: Requirements and solutions for data management and computing (Track Leaders: B. Konya, H. Heller, S. Tarkoma)