The Long Term Data Preservation (LTDP) project at INFN CNAF: CDF use case

S. Amerio[1], L. Chiarelli[2],L. Dell'Agnello[3], D. Gregori[3], M. Pezzi[3], P. Ricci[3], F. Rosso[3], S. Zani[3]

[1] University of Padova, Italy
[2] GARR
[3] INFN-CNAF, Bologna, Italy

E-mail: michele.pezzi@cnaf.infn.it

Abstract. In the last years the problem of preservation of scientific data has become one of the most important topics inside international scientific communities. In particular the long term preservation of experimental data, raw and all related derived formats including calibration information, is one of the emerging requirements within the High Energy Physics (HEP) community for experiments that have already concluded the data taking phase. The DPHEP group (Data Preservation in HEP) coordinates the local teams within the whole collaboration and the different Tiers (computing centers). The INFN-CNAF Tier-1 is one of the reference sites for data storage and computing in the LHC community but it also offers resources to many other HEP and non-HEP collaborations. In particular the CDF experiment has used the INFN-CNAF Tier-1 resources for many years and after the end of data taking in 2011, it 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 the whole amount of it in the future. According to this task the CDF Italian collaboration, together with the INFN-CNAF computing center, has developed and is now implementing a long term future data preservation project in collaboration with Fermilab (FNAL) computing sector. The project comprises the copy of all CDF raw data and user level ntuples (about 4 PB) at the INFN-CNAF site and the setup of a framework which will allow to access and analyze the data in the long term future. A portion of the 4 PB of data (raw data and analysis-level ntuples) are currently being copied from FNAL to the INFN-CNAF tape library backend and a system to allow data access is being setup. In addition to this data access system, a data analysis framework is being developed in order to run the complete CDF analysis chain in the long term future, from raw data reprocessing to analysis-level ntuples production and analysis. In this contribution we first illustrate the difficulties and the technical solutions adopted to copy, store and maintain CDF data at the INFN-CNAF Tier-1 computing center. In addition we describe how we are exploiting virtualization techniques for the purpose of building the long term future analysis framework.

Introduction

Long-term data preservation can be defined as the ability to provide continued access to digital materials and the capability to analyze them. Interest in the long term preservation of scientific data and their availability to general public is growing.

Data collected in High Energy Physics (HEP) experiments are the result of a significant human and financial effort. The preservation of HEP data beyond the lifetime of the experiment is of crucial importance to ensure the long term completion and extension of scientific programs, to allow cross collaboration analysis, analysing data from several experiment at once, to perform new analysis with new theoretical models and techniques and for education, training and outreach.

HEP data preservation poses many technical and organizational challenges: data preservation implies migration to new storage media when available, adjusting data access methods if needed; moreover data analysis capabilities must be preserved ensuring the experiment legacy software runs on new platforms, or on old ones with no security issues; validation systems have to be set up to regularly check data access and analysis framework; all the information needed to access and analyze data has to be properly organized and archived. For these reasons a data preservation project can be divided into two main areas: the first consists in "bit preservation", how we preserve the data, while the second part consists in the preservation of the analysis capabilities.

In this paper we describe a project to preserve at INFN-CNAF computing center in Italy a complete copy of the data of CDF experiment which run at the Fermilab Tevatron until September 2011. The project aims at having at CNAF a copy of all raw data and user level ntuples (4 PB) and providing users with data access and data analysis capabilities in the long term future.

2. Bit preservation and storage configuration

For the bit preservation, there are two important aspects to consider: first how the data is transferred and second where the data is stored.

As shown in Figure 1, we are using dedicated network resources for the CDF long term data preservation project.

At CNAF two servers with 10GE network interfaces are directly connected to the core switch/router of the Computing Center in order to guarantee the maximum network performance and the minimum round trip time (RTT).

In Local Area Network, servers involved in this project are isolated in a spacific VLAN and are using a dedicated IP network.

Geographical connectivity is provided by GARR and ESNET and consists on a dedicated 10 Gbps link from CNAF to FNAL.

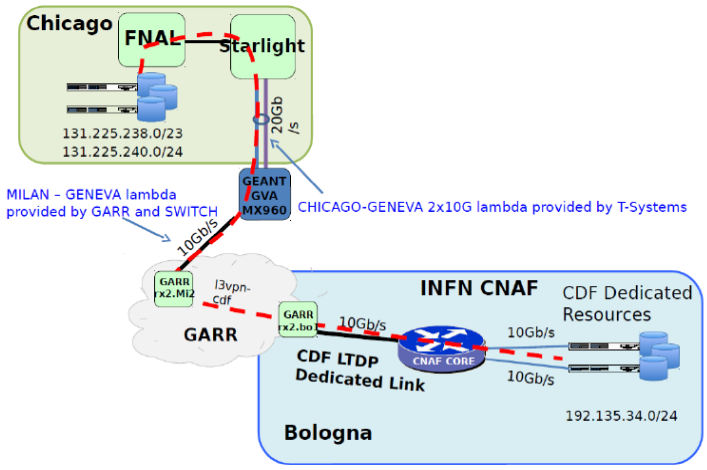

Figure 1 : Layout of the FNAL-CNAF copy network.

For the data transfer Sequential Metadata Access (SAM)[1], a data handling tool developed at Fermilab, has been installed on a dedicated machine at CNAF and it is used to pilot the transfer. Figure 2 shows how the data transfer works.

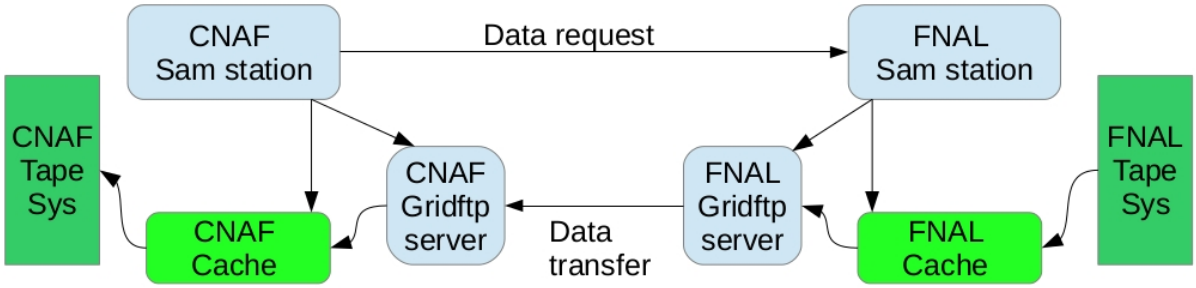

Figure 2 : Data transfer layout.

Integrating the SAM and GridFTP transfer commands within the same bash script, the data transfer takes place in a semi-automatic way:

1) CNAF requests from FNAL data that are staged at FNAL cache.
2) Data are copied using GridFTP protocol, using a third party transfer.
3) To check data integrity the file checksum is computed and compared with the value stored in SAM database.
4) Once the data files are in the CNAF cache and if the data integrity check is successfull, they are automatically migrated to tape and the SAM bookeeping database is updated.

The adopted solution at CNAF for the "bit preservation" of CDF data is GEMSS[2], a well established system, in production since several years at CNAF. GEMSS consists of a pool of disks managed by GPFS (IBM General Parallel File System), a tape library infrastructure for the archive back-end managed by Tivoli Storage Manager[3] (TSM) and an integration system to transfer data from disk to tape and vice versa.

The storage layout system is composed by several elements as shown in Figure 3.

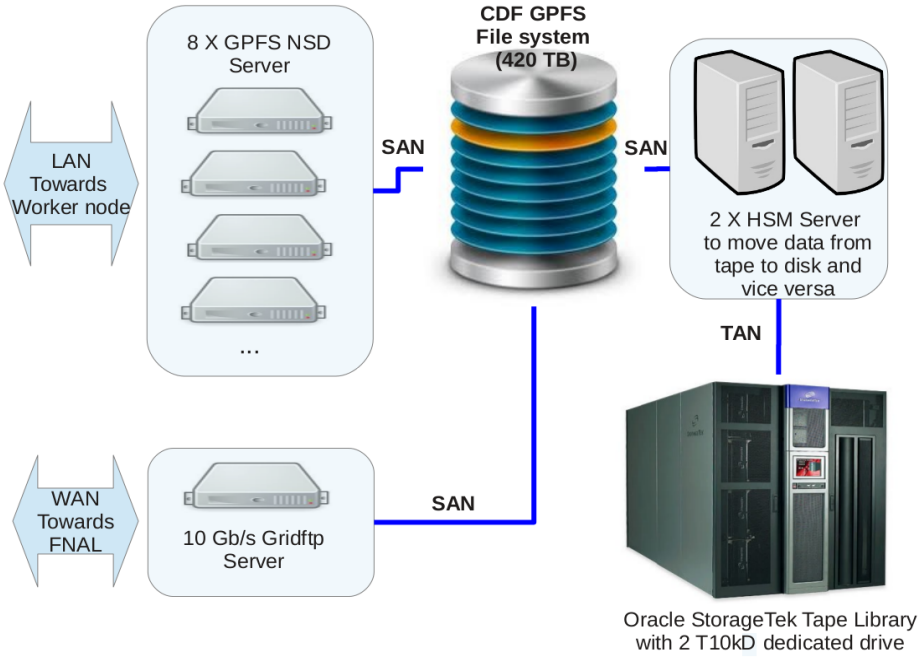

Figure 3 : Storage system layout.

The transferred data are stored on a dedicated GPFS cluster file system. On this file system two different directories were created. The first directory (cache) is the actual SAM disk cache that is used during network transfer. After a data file is copied to the disk cache and its data integrity is checked, it is moved to a second directory (durable) from which files are regularly and automatically migrated to

tape. Two dedicated HSM servers are used for moving data from tape to disk and vice versa using 2 tape drives through the Tape Area Network (TAN). During the copy phase, once a file is migrated to tape it is automatically removed from disk, in order to maintain a significant (~ 300 TB) disk buffer for the copy. In the short term future, once the copy is completed, we will keep a small (~ 100 TB) CDF-dedicated disk area to cache datasets requested by users for the analysis. In the long term future instead, when intermittent usage is foreseen, the disk cache will be allocated on demand. In any case a small disk cache (~ few tens of TB) is needed for periodic data read for integrity checks. In case of damage or loss, the file can be recovered by Fermilab, where two additional copies of the data are archived.

3. Long term future analysis framework

In order to access and process the data in the long term future it is essential to have an infrastructure that allows to use the experiment software. The goal is to make the software and data available and functional even when the experiment is no longer active. In the following we will briefly describe the main ingredients of the framework under development at CNAF.

In Figure 4 the main elements of CDF analysis frameworks are shown. The user submits a job from his/her user area. The job submission system contacts the SAM station to access the data. Data is sent to worker nodes together with a tarball containing the analysis code. If needed, e.g. for MC production, detector and run conditions are retrieved from a dedicated database. All these services are already installed at CNAF and for the long-term future we intend to replicate this system as much as possible.

We expect in the long-term future data will be accessed and processed very rarely. For this it is necessary to study a solution which is robust but at the same time allows to minimize the needed resources (number of physical machines within the framework). As shown in the Figure 4 the only real machine, except the storage system, is the database, which is currently located at FNAL. All other services are instantiated on virtual machines. The virtual machines will be handled by WNoDeS[4], an INFN-developed framework that makes possible to dynamically allocate virtual resources out of a common resource pool, in order to instantiate the virtual machines when a job is started.

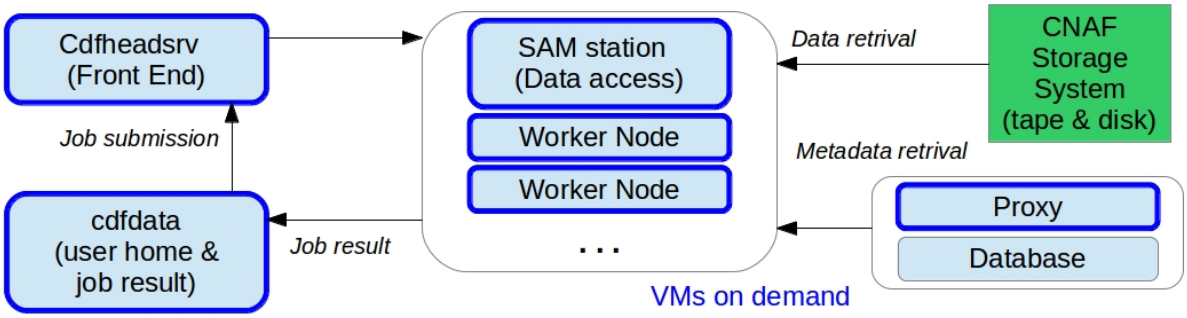

Figure 4 : Long term analisys framework.

3.1 Data access

We will keep using SAM to access data in the long term future. CNAF tape system is transparent for SAM: upon a request to access a file, the file is recalled from tape on a disk cache before being sent to the final destination (worker node or user area). As already stated in the previous section, in the long term future we foresee intermittent access to CDF data, so the necessary disk cache will be allocated on-demand using CNAF resources in opportunistic mode. We estimate that a user area of 50 TB will be sufficient for each analysis, as the majority of CDF datasets amounts to few TB and the biggest datasets do not exceed 20 TB. As described in section 2, CNAF used GPFS filesystem, and the tape library which is developed and maintained by IBM. SAM code is based on SL6 and will not be ported to new operating systems thus compatibility problems may arise if IBM will not support GPFS SL6 version in the long term future. A possible solution is accessing the data via NFS protocol, which is

expected to be compatible with older operating system. The first tests of this access method were successful: CDF GPFS filesystem was exported by a server via NFS protocol and data accessed via standard SAM commands. Scalability tests will follow on extended datasets (hundreds of files).

3.2 Data analysis

Future analysis on CDF data will use a software legacy release based on Scientific Linux 6 (SL6) distributed through CVMFS. At CNAF as a first step squid proxy servers have been setup to access the CVMFS server located at FNAL. In a second phase the FNAL CVMFS server will be replicated at CNAF.

The legacy release has not yet been officially released by FNAL; in the meantime the development version is being tested on CNAF SL6 nodes. As far as job submission is concerned, the system currently installed for CDF at CNAF is based on glideinWMS [5] to exploit computing resources at CNAF and additional LCG resources at different Tier-2 sites in Italy and other European countries. In the short term future the current system will be maintained, upgrading it to use the legacy release. For the long term future instead a simpler job submission system is under study which will exploit only CNAF CPU resources and LSF workload management system, already used at CNAF by other experiments.

3.3 Long term database issues

Oracle databases for data bookkeeping and run conditions are currently located at FNAL. In order for the CDF analysis framework at CNAF to be independent from FNAL, they would need to be locally replicated. There are basically two options currently under review. As first straightforward approach, we could plan to setup the needed Oracle database instances at CNAF, and keep them synchronized with one of the high availability or disaster recovery features available with Oracle. Choosing this solution we could keep compatible with the CDF analysis framework, but it would be quite expensive in terms of Oracle licenses, especially in the long term. Therefore, as an alternative approach, we are investigating the idea to rewrite the database interface inside the CDF software, opening the path to use open-source RDBMSs such as PostgreSQL or MySQL. This second approach is under discussion, because it needs some assessment of the needed developing effort. However it could be worthwhile to reduce the fixed costs, especially considering that this is a long term data preservation project.

4. Conclusion

A project to preserve CDF data and analysis capabilities is being implemented at CNAF. The project is be divided into two areas: the "bit preservation"and the preservation of the analysis framework. For the bit preservation a system to copy the data from FNAL has been setup; the data is being copied and about 2 PB of data out of 4 PB are already on tape. For the analysis framework, a system is under development: the goal is to exploit as much as possible the CDF analysis framework already existing at CNAF, investigating solutions to keep it functional for many years to come, e.g. using NFS to access data and virtualization techniques to run CDF legacy code.

References

[1] Stonjek S, Baranovski A, Kreymer A, Lueking L, Ratnikov F et al., *Deployment of SAM for the CDF Experiment*, 2005 1052–1054 2, Proceedings of 2004 CHEP conference

[2] D. Bonacorsi et al., *The Grid Enabled Mass Storage System (GEMSS): the Storage and Data management system used at the INFN Tier1 at CNAF*, 2012 J. Phys. Conf. Ser. 396 042051 Proceedings of 2012 CHEP conference.

[3] IBM website references for Tivoli Storage Manager info and documentation https://www.ibm.com/developerworks/wikis/display/tivolidoccentral/Tivoli+Storage+Manager

and http://www-03.ibm.com/software/tivoli/products/storage-mgr/
[4] D.Salomoni et all, WNoDeS, a tool for integrated Grid and Cloud access and computing farm virtualization 2010 J. Phys. Conf. Ser. 331 052017 Proceedings of 2010 CHEP conference
[5] S.Amerio et al, *Eurogrid: a new glideinWMS based portal for CDF data analysis*, Journal of Physics: Conference Series 396 (2012) 032001
[6] S. Amerio et al., *Long Term Data Preservation for CDF at INFN-CNAF* 2013 J. Phys. Conf. Ser. 513 (2014) 042011 Proceedings of 2013 CHEP conference