Last developments of the INFN CNAF Long Term Data Preservation (LTDP) project: the CDF data recover and safekeeping

P P Ricci[1,3], **A Cavalli**[1], **S Dal Pra**[1], **A Falabella**[1], **E Fattibene**[1], **M Pezzi**[1] **and S Amerio**[2]

[1] INFN CNAF, viale Berti Pichat 6/2 40127 Bologna, Italy.
[2] INFN Sezione di Padova, Via Marzolo 8, 35131 Padova, Italy.

E-mail: pierpaolo.ricci@cnaf.infn.it

Abstract. The INFN CNAF Tier-1 has become the Italian national data center for the INFN computing activities since 2005. As one of the reference sites for data storage and computing provider in the High Energy Physics (HEP) community it offers resources to all the four LHC experiments and many other HEP and non-HEP collaborations. The CDF experiment has used the INFN Tier-1 resources for many years and, after the end of data taking in 2011, it faced the challenge to both preserve the large amount of scientific data produced and give the possibility to access and reuse the whole information in the future using the specific computing model. For this reason starting from the end of 2012 the CDF Italian collaboration, together with the INFN CNAF and Fermilab (FNAL), introduced a Long Term Data Preservation (LTDP) project with the purpose of preserving and sharing all the CDF data and the related analysis framework and knowledge. This is particularly challenging since part of the software releases is no longer supported and the amount of data to be preserved is rather large. The first objective of the collaboration was the copy of all the CDF RUN-2 raw data and user level ntuples (about 4 PB) from FNAL to the INFN CNAF tape library backend using a dedicated network link. This task was successfully accomplished during the last years and, in addition, a system to implement regular integrity check of data has been developed. This system ensures that all the data are completely accessible and it can automatically retrieve an identical copy of problematic or corrupted file from the original dataset at FNAL. The setup of a dedicated software framework, which allows users to access and analyse the data with the complete CDF analysis chain, was also carried out with the addition of users and system administrators detailed documentation for the long-term future. Furthermore a second and more ambitious objective emerged during 2016 with a feasibility study for reading the first CDF RUN-1 dataset now stored as an unique copy in a huge amount (about 4000) of old Exabyte tape cartridges. With the installation of compatible refurbished tape drive autoloaders an initial test bed was completed and the first phase of the Exabyte tapes reading activity started. In the present article, we will illustrate the state of the art of the LTDP project with a particular attention to the technical solutions adopted in order to store and maintain the CDF data and the analysis framework, and to overcome the issues that have arisen during the recent activities. The CDF model could also prove useful for designing new data preservation projects for other experiments or use cases.

1. The INFN CNAF Tier-1 CDF use case

The INFN CNAF Tier-1 was involved in the CDF collaboration since the very beginning of its activity as computing center. At present the CNAF Tier-1 offers resources, in terms of computing, storage and general IT services to all the four LHC experiments and ~30 others non-LHC collaborations (including Astroparticle Physics). The CDF experiment (FNAL Tevatron) was one of the first CNAF Tier-1 users and after the end of the official CDF data taking activity (2011) we started the CDF long term data preservation LTDP collaboration with a close connection with FNAL, to preserve the amount of data

---

[3] Corresponding Author

produced during the last years of production of the detector. The technical term for defining this kind of activity is *"bit preservation"* since it usually only includes the maintenance of the data merely. The ability to access and reuse them in the future is generally defined as *"framework preservation"* and it includes all the software services that grant the accessibility and usability of the preserved datasets to the scientific communities.

The first major task that we granted to the CDF collaboration was the maintenance (bit-preservation) of the whole 2 PByte (raw data and analysis-level ntuples) that made up the RUN-2 first dataset collected during 2001-2011. The idea was copying them from the FNAL storage facilities "master copy" to the CNAF Tier-1 as a backup copy. We have decided, as an extended goal, to preserve a complete copy of the CDF RUN-2 whole amount of data and Montecarlo samples at CNAF in addition to the software transfer services (access, data analysis), for a total of roughly 4 PByte of data and associated software. To accomplish these two tasks, a geographical connection on a dedicated 10 Gbps link from CNAF to FNAL was made available from the GARR network association [1]. Consequently, the copy process was split in two separated phases:

- PHASE 1: end 2013 - early 2014 → All data and MC user level ntuples (2.1 PB)
- PHASE 2: starting from mid 2014 → All remaining raw data (1.9 PB)

The PHASE 1 also includes the installation and configuration of the necessary software services for the data transfer as outlined in the following section.

2. The CDF RUN-2 bit preservation activity

2.1. The PHASE 1 and PHASE 2 data transfer activity

The Sequential Access via Metadata (SAM) data handling tool (developed at FNAL) [2] was installed on dedicated servers at CNAF for orchestrating the data transfer that is actually made using the GridFTP protocol. In Figure 1 an overall schema of the FNAL/CNAF data transfer system is reported. The CDF user data request triggers a pre-staging process in the FNAL disk cache from the tape backend. This operation is made in parallel with the data transfer and permits a direct disk-to-disk transfer of only the pre-staged data at FNAL. Data are copied via GridFTP protocol, using the "third party transfer" option.

The SAM station performs a real time validation of the checksum (stored in a dedicated Oracle database) on the transferred data, in order to grant the reliability, as soon as it is saved to the CNAF Tier-1 hierarchical storage management (HSM) system [3]. Once the data are in the CNAF disk cache, they are automatically migrated to the HSM tape backend using custom integration of SAM and GridFTP commands. The Oracle CDF database also stores information about the specific dataset locations and metadata and it is the authoritative source of all the query requests about the CDF stored data. A specific dedicated copy script was created in order to automatize the data transfer process. The whole copy process (PHASE 1 and PHASE 2) was completed during spring 2015. This means that all CDF RUN-2 data are replicated at CNAF HSM facility (tape library) and fully checked. Furthermore, the SAM station was upgraded with the SAMWeb tool which now uses HTTP protocol for accessing the CDF database [2].

2.2. The CDF RUN-2 dataset check script

During the last phase of the copy process the idea of an automated system for regular checking of the data integrity has arisen. For this reason the copy script has evolved into a check script for verifying data and maintaining alignment from the FNAL "master copy" to the CNAF secondary one. The idea was to obtain a system for implementing regular integrity check on data and provide automatic correction of corrupted or unavailable datasets. This check script can be executed periodically (using automated scheduling tools or user commands) over specific CDF datasets, ensuring that all data are completely accessible. It can also automatically retrieve an identical copy of problematic or corrupted file from the original dataset at FNAL.

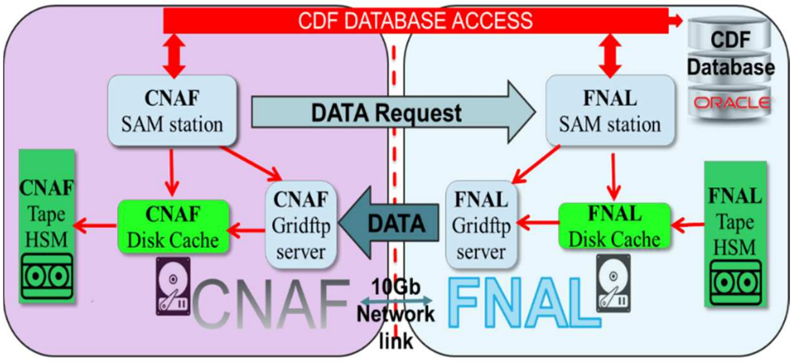

Figure 1. The FNAL/CNAF data transfer system.

The script performs three main actions as detailed below:

- Get information for the specific dataset.
- Check the file information checksum (CRC).
- Eventually heal the dataset and create a report of all the fulfilled operations.

Firstly, the script checks if the files which make up the dataset are correctly migrated to the CNAF HSM tape system. If some files are missing from a specific dataset, the copy is automatically triggered from FNAL to CNAF using the CNAF SAM station. After that, the script checks whether the file information checksum (CRC) is present as file extended attribute. If the CRC is absent or incorrect, the stage-in from the CNAF HSM tape system is triggered and the CRC is calculated on the retrieved file once available on the disk cache. Finally the script checks if the CRC is exactly the same stored in the CDF Oracle database info. If the CRC is different, the CNAF file (as secondary copy) is considered invalid and a new copy is triggered from the FNAL master copy by the use of the CNAF SAM station. It is clear that at the moment the FNAL master copy of the data is supposed to be proof of corruption and thereby represents the CDF data integrity model. In the future it will be probably possible to implement a “two way” system where in case of FNAL data corruption or unavailability, the CNAF “backup copy” could be used for recovering the original dataset.

3. The CDF RUN-1 bit preservation activity

3.1. The data recovery layout

The RUN-1 of CDF (1992-1995) contains events that demonstrate the "top quark" existence and it represents unique dataset in physics since the D0 collaboration did not keep the data at that time. The energy and conditions were different from RUN-2 but still proton-antiproton collisions [4]. This Tevatron conditions uniqueness brings both scientific and educational value for the RUN-1 datasets. After an initial review it was found that all valuable data are stored in ~4000 Exabyte 8 mm (Data8) tape cartridges with a maximum capacity of 5GB each. The old text database and Fortran software code (for analysis, simulation and visualization) are still available. Some preliminary tests using the RUN-2 CDF software were carried out at FNAL and indicated that probably the RUN-1 dataset can still be used with modern software. For this reason a bunch of 20 test tapes was delivered to CNAF in early 2016 in order to check the feasibility of a large scale data retrieve operation. Preparatory attempts showed that the tape data on the Data8 tape cartridges could be accessed using old Exabyte compatible tape drives connected via SCSI to modern O.S. servers (Linux S.L.6). Since a single tape contains variable amount of data (from 10Mbyte to the nominal capacity of 5Gbyte) the reading speed rates from 400Kbyte/s to 1.2MB/s and this translates in a reading time of 3-4 hours for a full tape. This means that the usage of single tape drives should clearly be excluded giving preference to tape drive autoloaders with 7-10 slots capacity each that could speed up the whole process. Furthermore, Exabyte latest generation Mammoth drives are backward compatible (read-only) with our tapes and this means that last generation autoloaders should be preferred since they are easier to find on the refurbished market. Technically speaking, the

RUN-1 data retrieval was considered possible and therefore the full amount of ~4000 tapes were packed and sent to CNAF at end of 2016. We acquired in the refurbished market one EXB-210 10-slots Exabyte Autoloader and two EZ17 7-slots Mammoth Autoloaders (all Linux compatible) with one year "swap warranty". Using old Dell 1950 with PCI-X servers, dismissed SCSI cards and cabling recycled from INFN Tier-1, we set up a hardware system whose layout is depicted in Figure 2.

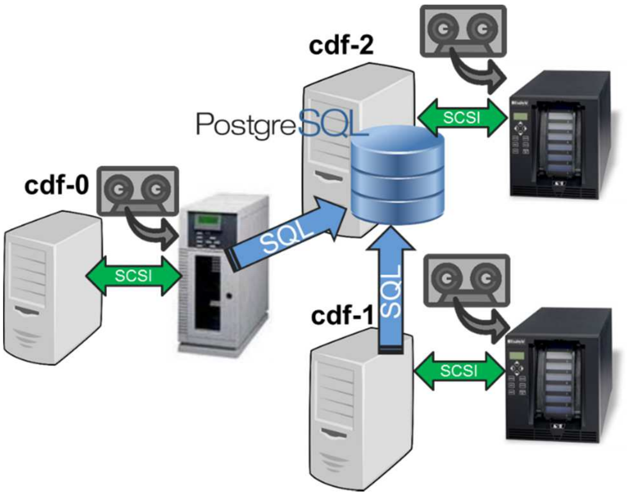

Figure 2. The CDF RUN-1 data recovery system.

The figure shows that three servers are directly connected via SCSI to the three autoloaders. The original text database (with number of files per tape and other relevant info about stored data) was imported into a PostgreSQL database running on one of the servers and this database also keeps track of the reading progress (success/failure statistics, size, CRC, etc…) into a table populated by the script operating the tape drives.

3.2. First results of the RUN-1 data recovery process

We decided to divide the RUN-1 data recovery operation in phases similarly to the RUN-2 data copy. PHASE 1 was defined with a target of successfully reading a subset of 10% (400) data tapes, followed by an evaluation of the success and failure ratio in order to estimate the sustainability of the operation. Specific scripts were also designed for operating the autoloaders, managing read operation and dealing with error conditions (i.e. partially read tapes). After the end of PHASE 1 we realized that two different tape media were used at the time: the Sony Data type QG-112M (~1000 tapes) and the Fuji "consumer" type P6-120 (the remaining ~3000 tapes). The Fuji brand tapes tend to slow down and block the tape drives as well as dirty the drive head. They also showed a considerably high *completely unreadable* status (15-20%) from all drives. Focusing on the Sony tapes we got better results, as we read 440 tapes with only 4 *completely* and 8 *partially unreadable* tapes. Anyway we noticed that, after a few months of continuous usage, all drives tended to malfunction and they stopped working permanently. Since the on-site assistance is not available due to lack of expertise, the only solution was to send the autoloaders to the refurbish expert sites worldwide, with a significant loss of time. In Figure 3, a graph describing the read operation trend during the first two months of activity is reported. As shown, a total of ~1350Gbyte over ~580 tapes were successfully read which means roughly 10 tapes/day with an average of 2.3Gbyte per tape. The occasional interruptions in the increasing trend confirm that the autoloaders drives need constant human intervention to retry read of failed tapes or to deal with struck drives. For this reason, the sustainability of the whole read operation should be seriously reconsidered. Therefore, we defined and started PHASE 2, which consists of reading the subset of all the Sony tapes (~1000 cartridges roughly 25% of the total) with the addition of the rebuilding of the RUN-1 data framework software. Since the remaining 3000 Fuji "consumer" tapes showed a higher failure rate (almost unsustainable) compared to the Sony ones, it will be hardly possible to complete the whole operation given the limits of the current hardware.

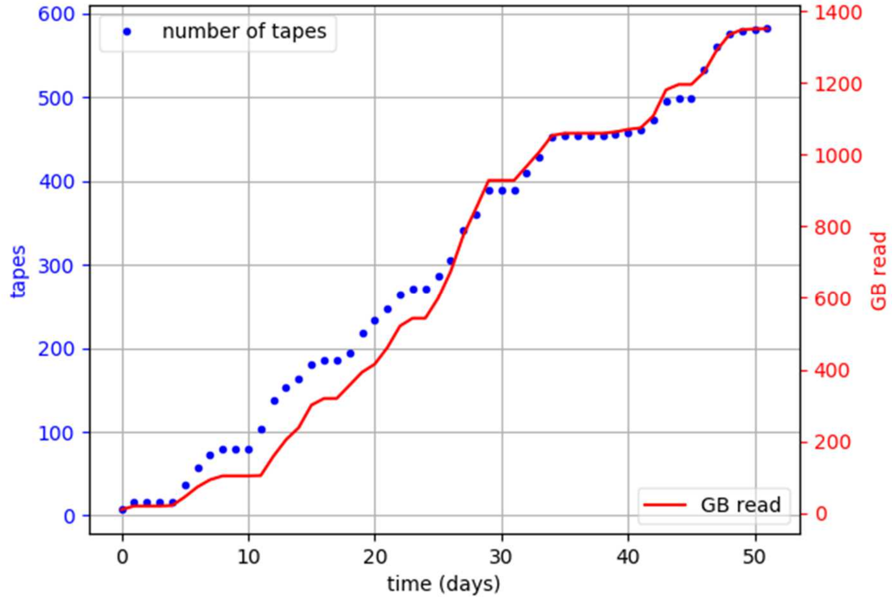

Figure 3. The results of two months of CDF RUN-1 read activity.

4. The CDF Framework preservation use case

4.1. The software and job submission system replica

The primary goal of the framework preservation is to safeguard the full infrastructure for using the experiment software and data in the long term future. We based the implementation on the instantiation of virtual machines (VMs) that runs the necessary services and can be put to "sleep" and resumed as needed. There are 3 principal service nodes in the CDF production framework:

- **cdfsam:** the SAM Station (for accessing the CDF datasets)
- **cdfdata:** the CDF users area (for user-level data e.g. *home* directory or custom software)
- **cdfheadsrv:** the CDF job submission frontends (head node)

Currently the SAM station software code is based on SL6 and therefore all cdfsam stations will remain SL6 virtual machines. The CDF analysis software has been using SL5 but a new software legacy release based on SL6 has been officially released (Feb. 2015) and therefore must be preserved [5]. In addition the CDF software is distributed using the CernVM File System [6] (CVMFS) which uses a server currently located only at FNAL for distributing the software to the end users.

The CDF job submission system is now based on JobSub [7] a system already in use at FNAL from other communities. In Figure 4 a scheme of the CDF software framework is represented. As reported in the figure for a "FNAL-independent" long term future job submission and software analysis system we need to replicate at CNAF the job submission head node (using JobSub), the CVMFS server and the CDF Oracle database. In particular a preliminary study for the CDF database replication was carried out and it is described in the following sub-section.

4.2. *The CDF Oracle database replica*

The CDF Oracle database is composed by two main instances: the *offline* instance (information for offline data processing and bookkeeping) and the *online* instance (data taking condition). We decided to use for the CNAF dedicated installation the Oracle version 11gR2 running on Oracle O.S. Linux 7.3. This is different from the O.S. used at FNAL (SUN Solaris) but aligned to the other database in production at our site. A single VM machine was installed with two Oracle database instances: the instances were named ***cdfofdp*** for the *offline* database and ***cdfondp*** for the *online* one. The VM is hosted on a four nodes KVM virtualization cluster, with the virtual machine images residing on a GPFS filesystem. The main reason was that the Oracle VM can be managed with the usage of GPFS *file clones* capability, a sort of snapshots where we can easily go back to a clean state in case of serious problems during the database import test. Using the Oracle Data Pump technology two export files were produced on February 2017 and copied from FNAL to CNAF containing a replica of both the *offline* and *online* complete databases with a size of 230GB each. The files contains a huge number of entries to be imported into our installation in particular:

- the number of table rows in the *offline* DB is ~3.3×10^9 rows on 958 tables
- the number of table rows in the *online* DB is ~4.2×10^9 rows on 766 tables

For this reason we decided to limit the import test to only two tables with ~9.1×10^6 rows with the addition of the relative indexes. The imported tables contain data that seems to be reasonable, however a cross-check with the original data is mandatory to demonstrate the real data alignment between the two databases. As a conclusion, the import from the Oracle database dump on a new installation with a different O.S. requires some special investigation and adjustment of the data definition statements included in the exports, in order to fit the data into the destination environment.

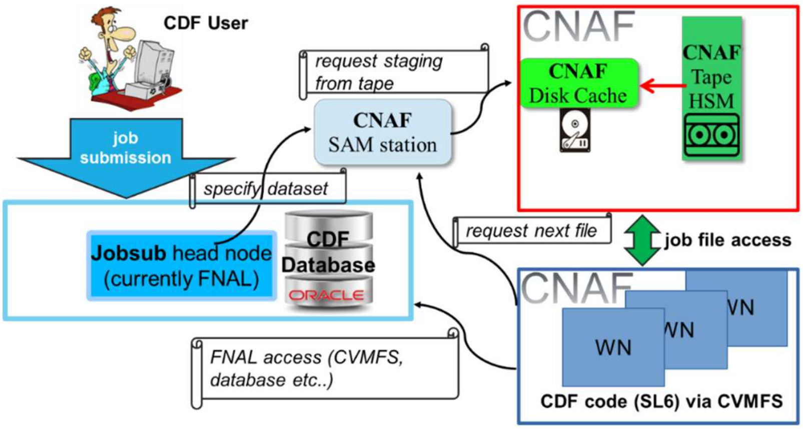

Figure 4. CDF software analysis framework.

5. Conclusion and future plans

This paper describes the latest developments in the INFN CNAF Tier-1 LTDP project. During the last years the complete CDF RUN-2 dataset was successfully copied to the INFN CNAF Tier-1. Nevertheless the integrity of the stored data should be regularly verified. For this reason, we are going to improve the check script described in section 2 in order to validate files on the CNAF HSM tape backend with periodically scheduled stage-in operation from tape to disk. About the CDF RUN-1 datasets, the PHASE 2 data retrieval should be completed soon, but some uncertainties still remains about the remaining consumer Fuji tapes that contains the residual 75% amount of data. Furthermore, the RUN-1 Fortran software framework could be revived using VMs and perhaps the RUN-2 software could also be adapted to work with the RUN-1 dataset. Regarding framework preservation, the job submission system could be fully replicated at CNAF ensuring full autonomy for these activities from FNAL. This includes replicating the FNAL CDF Oracle database, which can be carried out using FNAL database export files imported at CNAF. However, the problem of maintaining over time the synchronization between the two databases over two different sites (FNAL/CNAF) still remains. A dedicated website is under completion for collecting all relevant documents and could contain dedicated pages with specific information about CNAF archived datasets. Finally, and in conclusion, the CDF use case model will certainly be useful for designing future data preservation projects for other experiments.

Acknowledgments

We would like to thank the whole FNAL CDF Data Preservation team and all the CDF experiment people in particular Ray Culbertson, Eugene Schmidt Jr, Jonathan Lewis, Giorgio Bellettini and Donatella Lucchesi for the valuable support in the development of the present work.

References

[1] Amerio S et al., The Long Term Data Preservation (LTDP) project at INFN CNAF:CDF use case *J. Phys. Conf. Ser.* ***608*** *(2015) 012012, proceedings of the 2014 ACAT Conference.*

[2] Illingworth R A, A data handling system for modern and future Fermilab experiments *J.Phys. Conf. Ser.* ***513*** *(2014) 032045, proceedings of the 2013 CHEP Conference.*

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

[4] Abe F et al., Evidence for top quark production in $\bar{p}p$ collisions at $\sqrt{s}$ =1.8 TeV *Phys. Rev.* ***Vol.50****, Iss. 5, September 1994 D 50, 2966.*

[5] Amerio S et al., Data preservation at the Fermilab Tevatron *Elsevier Nuclear Instruments and Methods in Physics Research Section A* **Vol. 851***, 11 April 2017,* pp 1–4.

[6] Buncic P et al., CernVM - a virtual appliance for LHC applications *PoS(ACAT08)****012****, proceedings of the 2008 ACAT Conference.*
Info about CVMFS also avaliable online: *https://cernvm.cern.ch/portal/filesystem*

[7] Box D et al., Progress on the Fabric for Frontier Experiments Project at Fermilab *J. Phys. Conf. Ser.* ***664*** *(2015) 062040, proceedings of the 2015 CHEP Conference.*
Info about JobSub also avaliable online: *https://cdcvs.fnal.gov/redmine/projects/jobsub/wiki*