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

Pier Paolo Ricci
on behalf of INFN CNAF LTDP Group
pierpaolo.ricci@cnaf.infn.it

ACAT 2017 Seattle

ACAT 2017
21-25 August 2017
University of Washington, Seattle

Summary

- The INFN CNAF Tier-1 LTDP CDF use case
- Bit preservation
  - CDF RUN-2 (2001-2011)
  - CDF RUN-1 (1992-1995)
- Framework preservation:
  - Software and job submission system
  - CDF database instances replica
- Conclusion and future plans

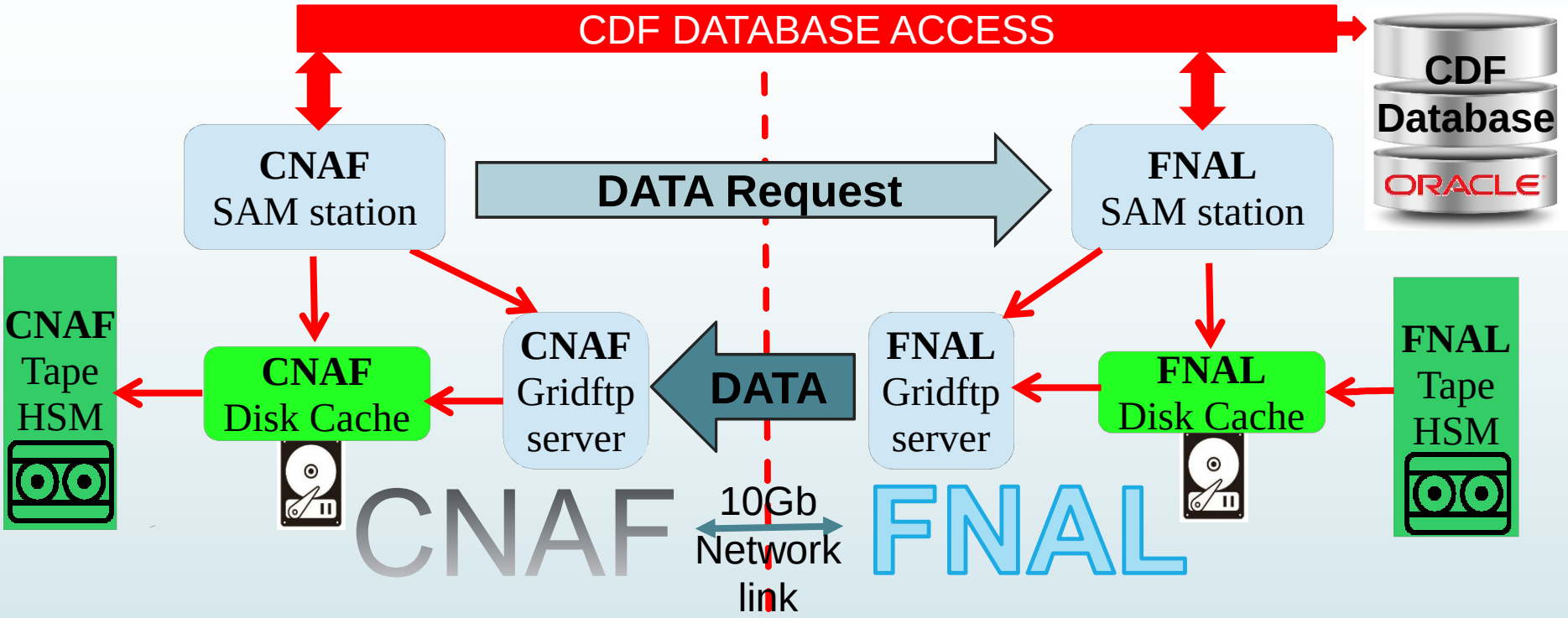
The Tier-1 LTDP CDF use case

- The INFN CNAF Tier-1 offers resources 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 INFN CNAF Tier-1 user
- After the end of CDF data taking (2011) we started a project (LTDP) in collaboration with FNAL for preserving:
  - the amount of data produced during the last years (**BIT PRESERVATION**)
  - the ability to access and reuse them in the future (**FRAMEWORK PRESERVATION**)
- **First Major task**: **4 Pbyte** (raw data and analysis-level ntuples) of RUN-2 data (2001-2011) to be copied from **FNAL => CNAF**

Bit preservation RUN-2

- **First Major task extended goal:** preserve a complete copy of CDF RUN-2 data and Montecarlo samples at CNAF **(4 Pbyte)** + related transfer services *(access, data analysis)*
- Geographical connectivity on a dedicated 10 Gbps link from CNAF to FNAL
- The copy was splitted in roughly two phases:
  - **PHASE 1:** end 2013 - early 2014 → All data and MC user level n-tuples **(2.1 PB)**
  - **PHASE 2:** starting from mid 2014 → All raw data **(1.9 PB)**
- The Sequential Access via Metadata (**SAM**) data handling tool (developed at FNAL) was installed on dedicated servers at CNAF for orchestrating the data transfer (**GridFTP**)

RUN-2 Data transfer system layout



- CNAF requests data from FNAL that are **pre-staged in FNAL disk cache** from the tape backend (by CDF user request). The pre-staging of the data at FNAL is made in parallel with the data transfer
- **Data are copied via GridFTP protocol**, in a third party transfer (GridFTP options), the SAM station control the checksum (stored in a dedicated database)
- Once the data are in the CNAF disk cache, **they are automatically migrated to tape using custom integration of SAM and GridFTP command**
- A dedicated "copy" script was created in order to perform data transfer in a semi-automated way

Bit preservation RUN-2 status

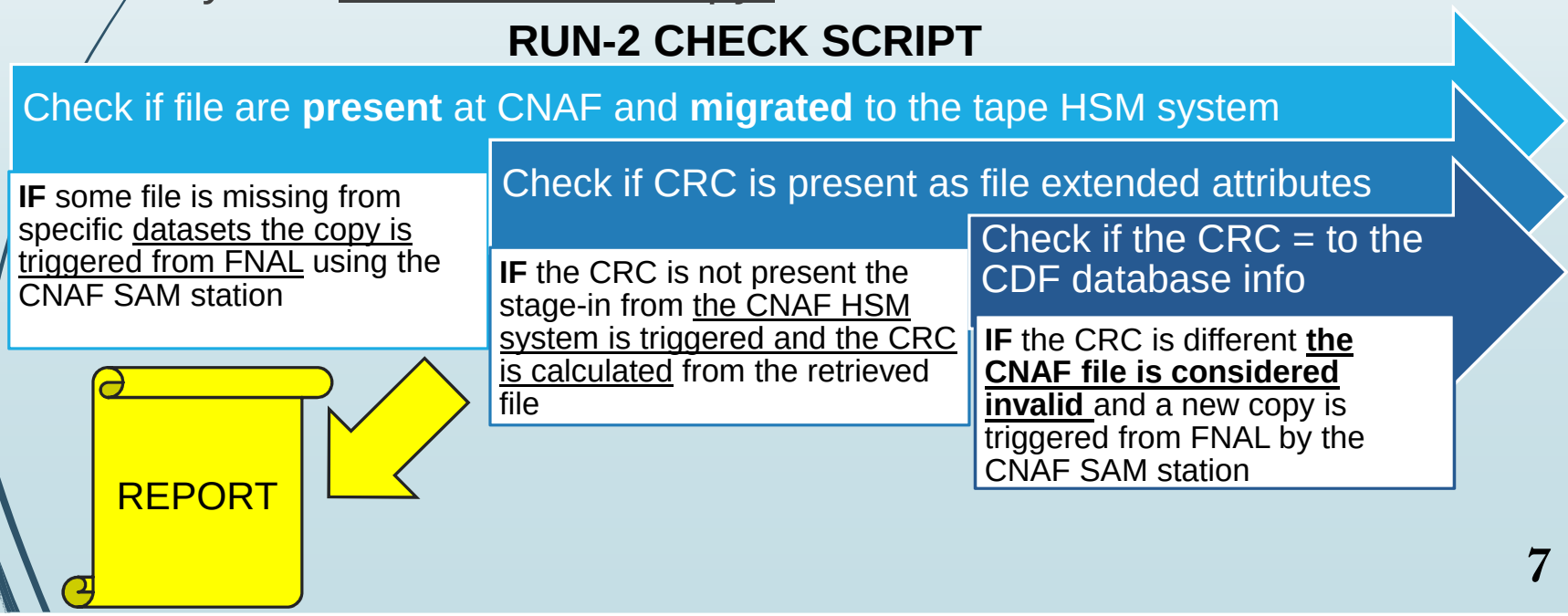
- Copy was completed during spring 2015☺
- All data are replicated at CNAF HSM facility (tape library)
- The "copy" script has evolved into a **check script** for verifying data and maintaining data alignment **FNAL => CNAF** (system for implementing regular integrity check of data).
- The new **check script** can be run periodically (cron) over specific CDF dataset.
- The **check script** also 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.
- Also the SAM station was upgraded with the **SAMWeb tool** (use http protocol for accessing the CDF database)

CDF LTDP RUN-2 check script

3 **Main actions**:

1. Get information for the dataset
2. Checking file information checksum (CRC) and create a report
3. Eventually "heal" the dataset

Can be periodically used on specific dataset in order to check consistency with FNAL "master copy"



Bit preservation RUN-1

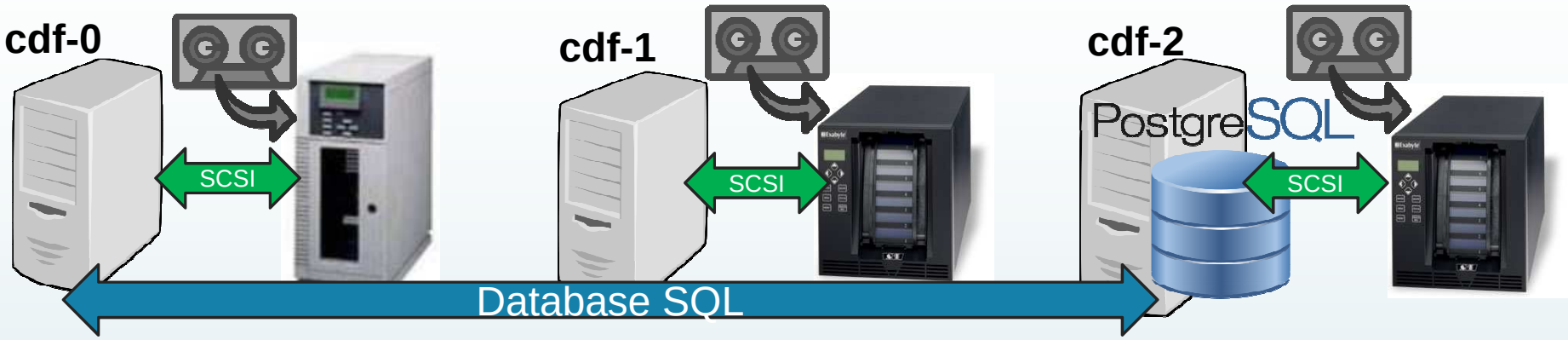
- RUN-1 of CDF (1992-1995) contains events that prove the "top quark" existence.
- Unique dataset in Physics (the D0 collaboration did not keep the data…)
- Different energy and conditions from RUN-2 but still proton-antiproton collisions => Tevatron conditions "uniqueness" means both scientific and educational value.
- All data are stored in **~4000** Exabyte 8 mm (Data8) tape cartridges with a maximum capacity of 5GB.
- Text database and Fortran software code (for analysis, simulation and visualization…) is still available.
- Test with RUN-2 software was carried out at FNAL **=> RUN-1 data can still be used!**

Bit preservation RUN-1

- A "bunch of test tapes" was sent to CNAF in early 2016.
- Preliminary tests show that the tape data could be accessed using old Exabyte compatible tape drives connected via SCSI to "modern" O.S. server (Linux S.L.6).
- A single tape contains greatly variable amount of data (from 10Mbyte to the nominal capacity of 5Gbyte).
  Reading speed rates from 400Kbyte/s to 1.2MB/s => 3-4h for a "almost standard" full tape.
- Single tapes drives should clearly be excluded => autoloaders (7-10 slots capacity each) speed up the whole process.
- Since Exabyte latest generation drives (Mammoth) are backward compatible (read-only) last generation autoloader should be preferred (easier to find on refurbished market).

Technically speaking retrieving of data is possible => full ~4000 tapes are packed and sent to CNAF at end of 2016!

Bit preservation RUN-1



- **1 EXB-210** 10-slots Exabyte Autoloader and **2 EZ17** 7-slots Mammoth Autoloader acquired in the refurbished market (Linux compatible) with 1 year "swap warranty".
- Servers (old Dell 1950 with PCI-X), scsi cards and cables "recycled" from INFN Tier-1 dismissed hardware.
- Very simple layout: **3 servers** directly connected via SCSI to the **3 autoloaders.**
- Original "text" database (with #files/tape and other info) imported into **PostgreSQL** running on the 3rd server. **PostgreSQL** database also keeps track of reading progress (success/failure statistics, size, CRC…) from all servers read activity.

PHASE 1: A target of 10% (400) data tapes must be successfully read.

…after that, a deep analysis of success and failures could be carried out!

Bit preservation RUN-1

- Specific scripts designed for operating autoloader, reading operation and error management (i.e. partially read tapes)
- 2 different media used for tapes: **Sony Data type** QG-112M (~1000) and **Fuji "consumer" type** P6-120 (~3000)

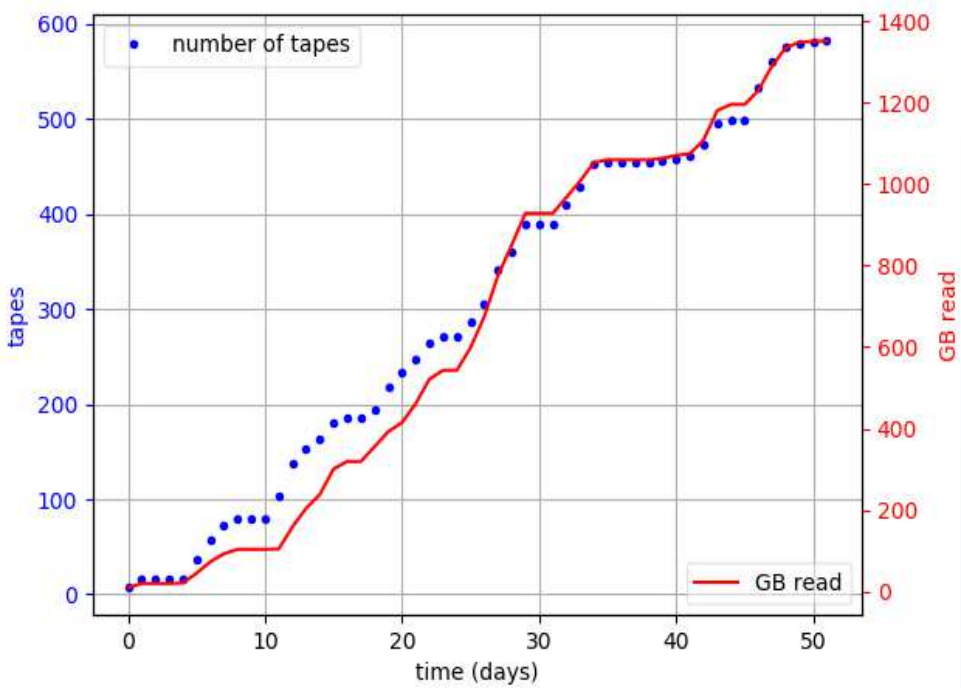
After end of **PHASE 1**…

- The Fuji "consumer" type tapes tends to slow down and block the tape drives and to "dirty" the drive head. Have a great "completely unreadable" status ratio (15-20%) from all the autoloaders. Not a good start ☹
- Focusing on the Sony tapes we get better results reading 440 tapes with only 4 "completely" and 8 "partially" unreadable tapes☺
- Anyway after some use all drives tend to malfunction and stop working. On-site assistance is not possible due to lack of expertise => The only solution is send the autoloader to the "refurbish" tech expert worldwide ☹

=> PHASE 2: **Target of all Sony tapes read (~1000 tapes)**

Bit preservation RUN-1

- After 2 months of activity **~1350Gbyte of data over ~580 tapes** were successfully read (10 tapes/day with an average 2.3Gbyte/tape)
- The autoloaders & drives need constant "babysitting" for retrying failed tapes and drive stuck situation => ***VERY TIME CONSUMING!***
- 2 autoloader (EZ17) are now under assistance (weeks of delay); the remaining 1 has just been swapped



The sustainability of the activity for the **PHASE 3** (likely the whole amount of tapes) should be seriously considered…

With the current hardware we are now focusing on **PHASE 2** (25% of total tape read) and on the rebuilding of the RUN-1 data framework software.

Framework preservation

- Goal: preserve the "full infrastructure" for using the experiment software and data in the *long term* future.
- Based on instantiation of virtual machines (VMs) that runs the necessary services (VMs can be put to "sleep" and resumed as needed).
- 3 principal nodes in the CDF framework
  - **cdfsam:** SAM Station (for accessing the CDF datasets)
  - **cdfdata:** user area (for user-level data e.g. */home/user1*)
  - **cdfheadsrv:** job submission frontends (head node)
- The SAM code is based on SL6 => cdfsam stations are and will be SL6 VMs
- The CDF analysis has been using SL5 but a new software legacy release based on SL6 has been officially released (Feb. 2015) and must be preserved!
- The CDF software is distributed using the **CernVM File System (CVMFS[1]).** The CVMFS system uses a server currently located at FNAL for distributing the software
- The CDF job submission system is now based on **JobSub[2]** a system already in use at FNAL from other communities

1) https://cernvm.cern.ch/portal/filesystem

2) https://cdcvs.fnal.gov/redmine/projects/jobsub/wiki

CDF job submission system@CNAF

job submission

FNAL

Jobsub head node (currently FNAL)

CDF Database

ORACLE

specify dataset

CNAF SAM station

request staging from tape

CNAF

CNAF Disk Cache

CNAF Tape HSM

All CDF Data and MC samples

request next file

job file access

CNAF

WN

WN

WN

FNAL access (CVMFS, database etc..)

CDF code (SL6) via CVMFS

For a "FNAL-independent" **long term future job submission** system we need to replicate at CNAF:

- job submission head node (with Jobsub)
- CVMFS replica server
- CDF Database

N.B. In the old system (SL5) **virtual worker node (WNoDeS)** were used =>**CAN BE USEFUL IN FUTURE**

Framework preservation: CDF DB

2 main DB instances: **OFFLINE** (info for offline data processing & data bookkeeping) and **ONLINE** (data taking condition) => **WE'LL TRY TO REPLICATE @CNAF**

- **CNAF INSTALLATION**: Oracle 11gR2 on Oracle Linux 7.3 *(Solaris@FNAL...)*
  - Single machine with 2 Oracle DB instances: **cdfofdp, cdfondp**
  - OS and Oracle patched (June 2017)
  - VM hosted on a 4-node KVM cluster, with VM data on GPFS filesystem
  - Oracle VM managed with GPFS “file clones” => we can easily go back to a clean state in case of serious problems/mistakes testing the database import
- **ORACLE IMPORT TEST**
  - 2 Oracle "Data Pump exports files" copied from FNAL: OFFLINE and ONLINE DBs, size 230GB each, produced on February 2017
    - **Number of table rows in offline DB: ~ 3.3 x 10^9 rows on 958 tables**
    - **Number of table rows in online DB: ~ 4.2 x 10^9 rows on 766 tables**
  - Tested the import of 2 tables: ~ 9.1 x 10^6 rows + indexes
    - The imported tables contain data that seems to be reasonable… ☺
    - …but a cross-check with the original data is mandatory

The import of a Oracle DB dump on a new installation requires some inspection and adjustment of the data definition statements included in the exports, in order to fit the data into the destination environment.

Conclusion and future plans

- **RUN-2** => the **check script** could be improved for checking file integrity on CNAF tape backend (scheduled stage-in tape => disk)
- **RUN-1** => **PHASE 2** should be completed (25%) and some uncertainties remains about the remaining "consumer" tapes
- **RUN-1** => Run-1 Fortran software framework could be "revived" on VMs and also Run-2 software could be adapted for working with the Run-1 (partial) dataset
- **Framework preservation** => The job submission system could be fully replicated at CNAF (full autonomy from FNAL)
- **Framework preservation** => The FNAL CDF Oracle database could be "freezed" and "imported" at CNAF (*but how keep the 2 databases synchronized?*)
- **Documentation** => A dedicated website is under completion for collecting all relevant document and could contain dedicated pages with info about CNAF archived dataset

The CDF use case model is certainly useful for designing new data preservation projects for other experiments!

Thank you! Questions?