Real-time dataflow and workflow with the CMS Tracker data

N De Filippis[1], G Bagliesi[2], R Bainbridge[3], T Boccali[4], V Ciulli[5], D Giordano[6], D. Hufnagel[7], D Mason[8], L Mirabito[9], C Noeding[10], F Palla[11], J Piedra[12] and S Sarkar[13]

[1] Dipartimento Interateneo di Fisica "M. Merlin" dell'Universitá e del Politecnico di Bari and INFN Sezione di Bari, Via Amendola 173, Bari, Italy
[2] INFN Sezione di Pisa, Polo Fibonacci Largo B. Pontecorvo 3, Pisa, Italy
[3] Blackett Laboratory, Imperial College, South Kensington, London
[4] INFN Sezione di Pisa, Polo Fibonacci Largo B. Pontecorvo 3, Pisa, Italy
[5] Universitá di Firenze e INFN Sezione di Firenze, Via G. Sansone 1, Sesto Fiorentino, Firenze, Italy
[6] Universitá di Bari and INFN Sezione di Bari, Via Amendola 173, Bari, Italy
[7] European Organization for Nuclear Research CERN CH-1211 23, Genve, Switzerland
[8] Fermi National Accelerator Laboratory, Batavia (IL), USA
[9] European Organization for Nuclear Research CERN CH-1211 23, Genve, Switzerland
[10] Fermi National Accelerator Laboratory, Batavia (IL), USA
[11] INFN Sezione di Pisa, Polo Fibonacci Largo B. Pontecorvo 3, Pisa, Italy
[12] Laboratory for Nuclear Science (LNS) Massachusetts Inst. of Technology (MIT), 77 massachusetts avenue, Cambridge, USA
[13] INFN Sezione di Pisa, Polo Fibonacci Largo B. Pontecorvo 3, Pisa, Italy

E-mail: Nicola.Defilippis@ba.infn.it, Giuseppe.Bagliesi@cern.ch, Robert.bainbridge@cern.ch, Tommaso.boccali@cern.ch, Vitaliano.ciulli@cern.ch, Domenico.giordano@ba.infn.it, Dirk.hufnagel@cern.ch, dmason@fnal.gov, Laurent.mirabito@cern.ch, noeding@fnal.gov, Fabrizio.Palla@cern.ch, Jonathan.piedra.gomez@cern.ch, Subir.sarkar@cern.ch

Abstract.

The Tracker detector took data with cosmics rays at the Tracker Integration Facility (TIF) at CERN. First on-line monitoring tasks were executed at the Tracker Analysis Centre (TAC) which is a dedicated Control Room at TIF with limited computing resources. A set of software agents were developed to perform the real-time data conversion in a standard format, to archive data on tape at CERN and to publish them in the official CMS data bookkeeping systems. According to the CMS computing and analysis model, most of the subsequent data processing has to be done in remote Tier-1 and Tier-2 sites, so data were automatically transferred from CERN to the sites interested to analyze them, currently Fermilab, Bari and Pisa. Official reconstruction in the distributed environment was triggered in real-time by using the the tool currently used for the processing of simulated events. Automatic end-user analysis of data was performed in a distributed environment, in order to derive the distributions of important physics variables. The tracker data processing is currently migrating to the Tier-0 CERN as a prototype for the global data taking chain. Tracker data were also registered into the most recent version of the data bookeeping system, DBS-2, by profiting from the new features to handle real data. A description of the dataflow/workflow and of the tools developed is given, together with the results about the performance of the real-time chain. Almost 7.2 million events were officially registered, moved, reconstructed and analyzed in remote sites by using the distributed environment.

1. Introduction

The CMS tracker detector was fully integrated at the Tracker Integration Facility (TIF) at CERN. The commissioning of the 25 % of the tracker was performed with cosmic muons since February 2007 until the middle of July.

A dedicated Control Room called Tracker Analysis Center (TAC) was setup as a common area for tracker operations and online and offline commissioning.

The computing resources at TAC were devoted to serve the needs of data quality monitoring and of the preliminary fast-response analysis. On the other side, a large community was expected to analyze data taken at the TAC, and this could not happen on TAC computers which were limited in number and with a limited local disk storage, just sufficient to cache the incoming data for about 10 days. The complete tracker data processing was then addressed by remote sites receiving data from the TAC, by using the CMS official tools.

The dataflow/workflow of the tracker data processing is reported in Figure 1. All the steps of the processing chain are described in the Section 2. In the Section 3 the migration to the Tier-0 of the tracker data processing and the new description of tracker data in the data bookeeping system is addressed.

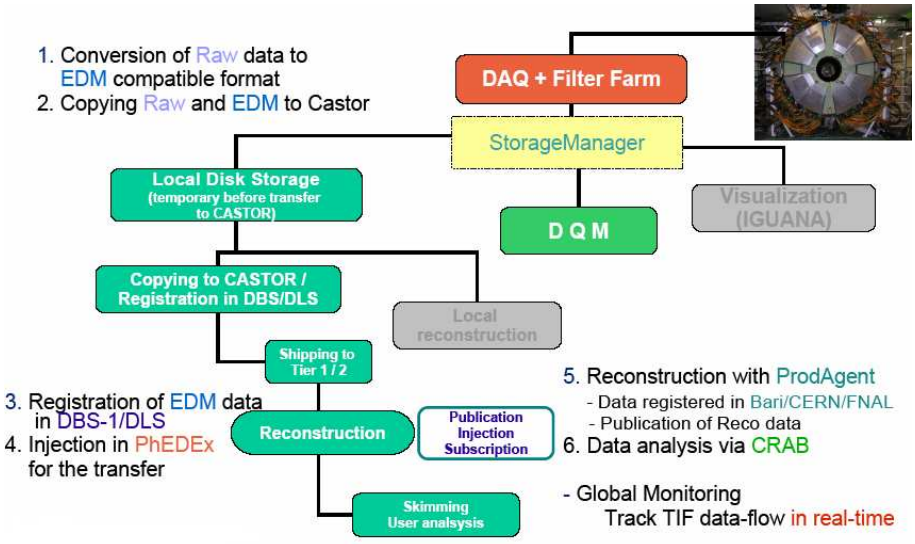

Figure 1. Dataflow/Workflow of the tracker data processing.

2. Tracker data processing

The tracker data processing consisted of the following steps: data archiving, the conversion of data from the raw format to the CMS Event data Model (EDM) format, the registration in the CMS official data bookeeping services, the transfer to remote tiers, the data reconstruction and the data analysis in a distributed environment. Some details are given in the following paragraphs.

2.1. Data archiving

All the data collected by the tracker detector were written into a main storage machine at TAC, which behaved like a temporary data buffer.

Data were backed-up to CASTOR storage [1] at CERN in real-time (every 5 minutes) with a job which scan the subdirectory structure creating the necessary directories on remote side, and

copying data files only when they did not exist on CASTOR side and had not been accessed in the last hour (to prevent the copy of still modified data). Once all the files belonging to a run were copied to CASTOR, a catalog was prepared for that run after a few checks.

2.2. Conversion to the EDM format

The strip tracker data acquisition system provided two mechanisms for writing data to disk, one using custom tracker-specific software and the other using "standard" CMS software (currently called the "StorageManager"). The former was in use since beam tests in 2001, when the definition of a standard mechanism for streaming data to disk was still far from complete. The latter mechanism has been the standard for CMS.

In both cases, the resulting data format was different to the"EDM compliant" format used within the CMS software framework (CMSSW). This approach was used to remove any overheads associated with the use of ROOT [2] and hence optimize the performances of the data acquisition. In this case, the resulting data files were known as *streamer* files and contain the raw FED buffers and a small number of header words that were used to define the structure of the data.

In order to perform reconstruction and analysis on the data, an intermediate step was implemented to convert data from the native formats to the EDM format; that procedure ran in real-time to allow the next steps of the processing to be triggered as soon as possible.

2.3. Data registration

As soon as converted data were made available on CASTOR at CERN they were registered and transferred in remote sites in order to make them officially available to the CMS community and ready to be analysed in a distributed environment, as expected in the CMS Computing model [5].

That was performed firstly by using a software agent responsible to look for new files archived on CASTOR and to trigger the registration in the CMS Data Bookeeping System, DBS, [3] which is the official database already used for Monte Carlo data samples. The concepts of "primary dataset", "data tier" and "processed dataset" were directly inherited from the DBS design project; the primary dataset identifies a class of data such as those taken from the TIB or those processed with a given CMSSW release; the data tier is related to the format of data, RAW, DIGI, RECO for example; the processed dataset gives an additional information related to a block of files of the same primary dataset. The processed dataset concept was used to distinguish runs of real data in this work so each run consisting of few files was registered as a different processed dataset.

The location of block of files, here data runs, were registered in the data location system, DLS [4], in terms of the storage elements hosting data blocks.

The time between subsequent checks of the availability of new files to be registered was set to 15 minutes and did not introduce any limitation to the processing.

2.4. Data transfer

Because of the limited resources at TAC and according to the CMS computing and analysis model [5], most of the processing of the real data was done in remote sites, Tier-1s and and Tier2s centres, as soon as data were officially published and transferred to them. The transfer was performed in remote sites using the CMS official data movement tool PhEDEx [6]. This operation required that data were injected in the PhEDEx transfer management database to be routed to destination sites.

Technically the injection is handled by the PhEDEx tool with a specific software agent originally developed to inject Monte Carlo data samples. For the purpose of the tracker data registration an agent was setup on the official PhEDEx node at Bari and used together with another external agent based on one component of the official tool for Monte Carlo production,

ProdAgent [7]. The time between subsequent checks of the availability of new runs was configurable and was set to about half an hour. The registration in DBS/DLS and the injection in PhEDEx were handled indepedently by the agents, asyncronously.

The subscription of data to the destination sites was performed by the PhEDEX administrators. At the end of the transfer once all the files of a given block were really on disk at remote sites, the location of that block was registered in the DLS in terms of the remote storage element hosting the block.

Tracker data were trasferred and registered successfully at Bari, Pisa and FNAL.

2.5. Data reconstruction

As soon as data were registered and transferred to Tier sites they were reconstructed or re-reconstructed, in case of some reconstruction parameters were changed from time to time.

The ProdAgent tool was used to perform the reconstruction at the same way that is performed with simulated samples. The use of ProdAgent ensured that reconstructed data were automatically registered in DBS and DLS, ready to be shipped to Tier sites for subsequent analysis. Input data were processed run by run; a software agent was developed to trigger the real-time processing. The offline database was accessed in remote via FroNTier software [8] at Tier-1/2.

Two different strategy for reconstruction were used:

- Standard Reconstruction:
  The standard reconstruction was based on the code included in official releases of the CMS software. The procedure was triggered by a Bari machine running some agents. Jobs ran at Bari and CERN where raw data were published.
- Fermilab (FNAL) reconstruction:
  The latest code under development was used for the processing of cosmic runs at FNAL. The focus was to allow immediate feedback to the tracking developers by using corrected geometry and latest algorithm changes.

4.7 million events were reconstructed between the 7 million registered; the other ones were not collected in physics configuration while were used to test the performances of the silicon modules and of the read.out chain. The event display of a cosmic muon is given in Figure 2.

2.6. End-user analysis with CRAB

The official CMS tool for running end-user analysis in a distributed environment is CRAB [9]. CRAB runs in a User Interface environment and requires a Grid personal user certificate to start the processing.

In order to process events as soon as they were reconstructed an automatic procedure based on CRAB was setup, too. Runs were processed mostly at Bari, CERN and FNAL. ROOT macros and scripts were used to extract the interesting quantities.

The monitoring of the progress of the analysis was executed via web interfaces. The output retrieval was triggered automatically. Log files of CRAB and macro steps results (ROOT files with histograms) were also linked from web pages. Summary result tables were also compiled.

Physics results were extracted by using the previous analysis chain; for example noisy strips, modules with low signal cluster occupancy, tracks were looked for and plots and distributions were automatically made available via web interfaces.

The prioritization of analysis jobs was setup at Tier centers by using a dedicated VOMS group called "Tracker commissioning"; this allowed site administrators to identify tracker analysis jobs, to raise their priority if needed, to give permissions to access files and to write files to dedicated disk areas.

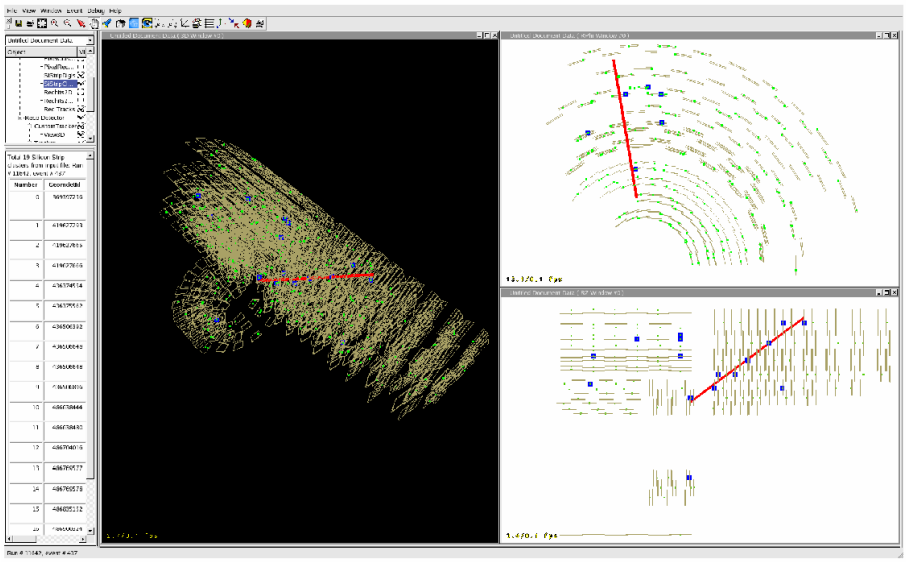

Figure 2. Event display of a cosmic reconstructed event.

2.7. Problematic issues

Some problems were experimented through the processing of tracker data. All of them were indetified and solved quickly.

Data conversion suffered from CASTOR problems to stage-in and stage-out files. The mismatch of file sizes between the complete files and those archived on CASTOR was also a problematic issues due to the fact that sometimes happened that files were archived before being closed by ROOT. NFS service on the storage machine slowed down the processing when a large number of clients accessed the data volume.

Data registration had no hiccups at all, it was fast and robust.

Data transfer via PhEDEx suffered from the fact that if CASTOR failed to deliver files, PhEDEx could wait indefinitely without work-around the problem. File size mismatch between CASTOR and PhEDEx database were also found because some files were overwritten after injection to PhEDEx.

CERN to FNAL data transfer was affected by other transfers with higher priority like the Monte Carlo Production; the eventually importance of tracker data was recognised and transfer was streamlined.

Problems also happened with the standard reconstruction processing, mostly related to the scarse resources in terms of dedicated CPUs and disk. The disk became full at Bari after some days of reconstruction while there was no disk space available at Pisa to host the data because of Monte Carlo samples. CERN resources were shared with the official production teams so the queues became full of jobs and the tracker processing was stopped and restarted more times in order to drain the CERN queues. Standard reconstruction went slowly with respect to FNAL one and with a lower efficiency becuase of problems mentioned previously.

3. Tracker data processing at Tier-0

Most of tracker processing chain hosted at TAC machines, run at Bari and in remote sites is supposed to be run at Tier-0, where the data registration and the prompt reconstruction will be run with real CMS data. The migrstion of some processing tasks to the Tier-0 is in progress and the main goal is to integrate the tracker processing workflow with the global data taking

effort of the data operation team at CERN.

The Tier-0 data operation team joined the tracker effort and helped to solve the technical aspects of the migration.

The most critical stuff was the registration of tracker data in version 2 of the data bookkeeping service database, as detailed below. The DBS team was effective in the support of the new tracker data description.

3.1. Data description in DBS-2

Migration to DBS-2 profits from the new DBS-2 features about real data handling that means a hopefully better organization of data. Just one primary dataset is defined. Run and luminosity info can be stored in the database to describe real data. One processed dataset for all the runs belonging to a datatier (RAW, RECO) was defined.

An important concept was introduced in DBS-2: the analysis dataset. Homogeneous runs (such as pedestal runs, physics runs or any set of run taken in a well defined configuration) can be grouped in analysis dataset according to a given set of algorithms/rules.

Scripts were developed with the purpose of extracting informations related to streamer and converted files from many sources (Storage Manager and RunSummary database, etc.). The registration of streamer and converted files was performed according to the name conventions used for global data taking. Streamer and converted files were also copied from the current locations on CASTOR to new ones according to the new conventions; this new copy of data caused a delay in that registration because of the problems in stage-in and stage-out files from either tape or disk.

7.2 million events, 1574 runs, 2348 files were registered in DBS-2 for both streamer and converted files; the former were 13 TB disk, the latter almost 7 TB.

The reconstruction of tracker data starting from DBS-2 information was tried successfully but the complete processing is not still started. The access both to single run and range of runs in processed dataset and analysis dataset via CRAB is supported and tested.

4. Conclusions

Tracker data processing was the first experience with data which used official CMS tools in a distributed environment. It was a successfull and complete exercise and an example of the integration between the detector and the computing communities.

7 million events were registered, 4.7 million events from physics runs were reconstructed and analyzed successfully in the distributed environment.

The tracker data processing is actually considered the prototype for the Tier-0 global data taking chain. Tracker data description in DBS-2 was also the first example of real data handling for analysis. A lot of feedback was given and received by the Tier-0, the PhEDEx, the DBS and the CRAB teams.

Acknowledgments

A special thank to the CMS Tracker community for the large support and collaboration. Thank also to the Tier-0, the Phedex, the DBS and the CRAB teams for their effective contribution in the implementation choices and the effective interaction.

References

[1] http://castor.web.cern.ch/castor
[2] httsp://root.cern.ch
[3] https://uimon.cern.ch/twiki/bin/view/CMS/DBS-TDR
[4] https://uimon.cern.ch/twiki/bin/view/CMS/DLS
[5] CMS Computing TDR, CERN-LHCC-2005-023, 20 June 2005

[6] http://cmsdoc.cern.ch/cms/aprom/phedex
[7] https://uimon.cern.ch/twiki/bin/view/CMS/ProdAgent
[8] http://lynx.fnal.gov/ntier-wiki
[9] https://twiki.cern.ch/twiki/bin/view/CMS/WorkBookRunningGrid