A SKIMMING PROCEDURE TO HANDLE LARGE DATASETS AT CDF

F. Delli Paoli*, D. Lucchesi, S. Da Ronco - INFN Padova, Italy
Armando Fella - INFN CNAF (Bologna), Italy
M. Casarsa, S. Belforte - INFN Trieste, Italy
G. Compostella - INFN Trento, Italy

Abstract

The CDF experiment has a new trigger which selects events depending on the significance of the tracks impact parameter. With this trigger, a sample of events enriched with Beauty and Charm mesons has been selected and is used for several important Physics analyses like the B_s mixing. The size of the dataset analysed for the Winter 2006 conferences is about 20 TB, which correspond to an integrated luminosity of 1 fb^{-1}.
In order to cope with the difficulties of handling large datasets CDF has developed a skimming procedure which reduces the dataset size by selecting events which contain only B mesons in specific decay modes. The rejected events are almost all background, and this guarantees that no signal is lost while the processing time is reduced by a factor of 10. This procedure is based on SAM (Sequential Access via Metadata), the CDF data handling system. The skimming procedure is described and the performances are summarized.

INTRODUCTION

CDF is an experiment running at the Tevatron, the $p\bar{p}$ collider at the Fermi National Accelerator Laboratory (FNAL) [1]. The CDF Collaboration comprises 624 physicists from 59 Institution around the world.
The experiment has collected up to today an integrated luminosity of 1.23 fb^{-1} and in the coming years the Tevatron is expected to deliver to CDF ~ 2 fb^{-1} per year. Figure 1 shows the design luminosity that the Tevatron is expected to deliver in the coming years. CDF has now nearly 1 PB of data under management and the amount of data is expected to grow in the future following the luminosity.
Physical events coming from the Collider are filtered by a three Level Trigger which can select events at a rate of 100 Hz (with the new coming upgrade we will reach 360 Hz). Events passing the trigger requirements are written on disk in Raw Data format and successively stored on tape. The Raw Data are processed to reconstruct physical objects on a dedicated Production Farm. After the production process data are split in different datasets on the basis of physical requirements and stored again on tape [4]. In order to do analysis the user needs to access these data.

* francesco.dellipaoli@pd.infn.it

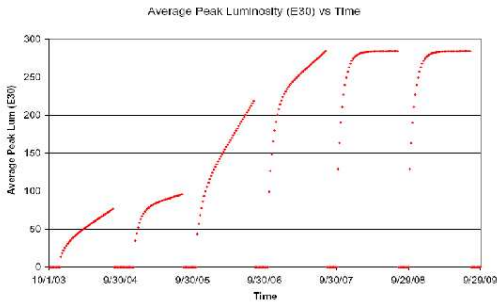

Figure 1: Design luminosity expected to be delivered by the Tevatron in the coming years.

WHAT IS AND WHY CDF NEEDS THE SKIM

For the Tevatron Run II CDF has a new trigger (based on the Online Silicon Vertex Tracker (SVT) [2] [3]) which selects a sample enriched with Beauty and Charm mesons by cutting on the significance of the track impact parameter with respect to the primary vertex. Figure 2 shows how is defined the impact parameter. This sample is very useful for several Physics measurements, one of the most important is the determination of the B_s mixing frequency.
The analysis framework allows to process events at

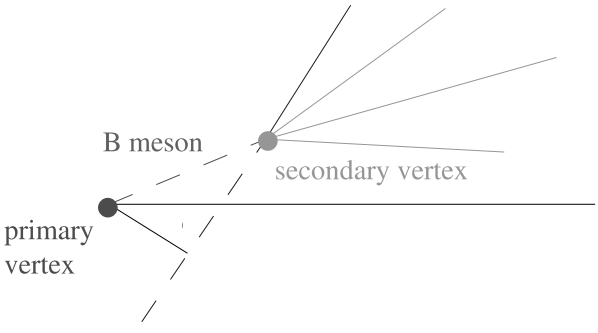

Figure 2: Track impact paramenter.

a rate of $0.1 - 0.5$ $s/event$, therefore to process the dataset for the Winter 2006 Conferences which corresponds to ~ 20 TB, on average months of user time is needed. Moreover, many physicists in CDF are interested in doing analysis with this sample and CDF needs to find a way to guarantee data avalaibility and enough resources to process them in reasonable time

for the entire Collaboration.
Two approches have been developed inside the collaboration:

- produce a common set of standard ntuples with all the information different analyses may require (processing ntuples takes tipically 0.002 $s/event$);
- reduce the dataset size by selecting only interesting events.

The next sections will describe how the second approach has been implemented, the skimming procedure working in the CDF Data Handling Framework. We will describe the procedure to skim data, concatenate the output and the method used to control that each input file is processed once and only once.

Since a specific tool to store files to disks and/or tape is needed we developed the so called "sam_upload" which consists of users authentication plus a transfer layer. Details on this tool, on its usage together with several definitions and performances will be also described.

DATA HANDLING MODEL AT CDF

The CDF Computing Architecture relies on the CDF Analysis Farm (CAF) model [5]; the CAF was originally a PC cluster localized at FNAL but successively the model was exported offsite and now many Decentralized CDF Analysis Farms exist in many sites worldwide. Currently two CAFs exist at Fermilab and nine dCAFs are hosted in offsite Institutions: $\sim 50\%$ of the Computing power of CDF resides outside Fermilab. Table 1 shows the current CDF CPU power for onsite and offsite resources. Onsite and offsite CAFs are basically identical but, since Data are stored in the Tape Robot at Fermilab, onsite CAFs are mainly used for User Data Analysis jobs or semi-coordinated activities like general Monte Carlo or common ntuples production. Offsite resources are basically devoted to Monte Carlo production. Recently, dataset replicas allow users to run analysis jobs in remote sites like CNAF at Bologna (Italy).

The Data Handling model relies on SAM (Sequential

Table 1: CDF computing resources

	CPU (MSpecInt2000)
CAF at Fermilab (onsite)	2.6
CAF offsite	2.5

Access via Metadata) [6]. SAM is organized as a set of servers which work together to store and retrieve files and associated metadata, including a complete record of the processing which has used the files. SAM is designed for the following tasks:

- track each file belonging to the system via a File Catalog declaration;
- track the location of each file in the system;
- provide storage utilities to copy files in a Permanent Mass Storage System implemented on a Robotic Tape Store;
- provide a way to cache files on local disk for the duration of the requesting job;
- deliver files on request from the closest location to local cache;
- track processing information of the consumption history of a specific dataset and of the processed and unprocessed files, permitting the construction of processing jobs which use information on which files have been previously successfully processed;
- track the lineage for all the processed files, identifying the input files used to produce certain files (parents files) or the output files produced by processing a certain files (children files).

SKIMMING MODEL

The skimming procedure basically consists of three steps:

- Skimming: select events based on useful decay modes from the large hadronic dataset and collect them in smaller datasets;
- Concatenation: concatenate the output files in order to reduce the number of files to be managed by the SAM Data Handling System;
- Storing: store the skimmed dataset on the Robotic Tape Store at Fermilab and on the storage disk located near Decentralized Analysis Farms like in CNAF, Bologna (Italy).

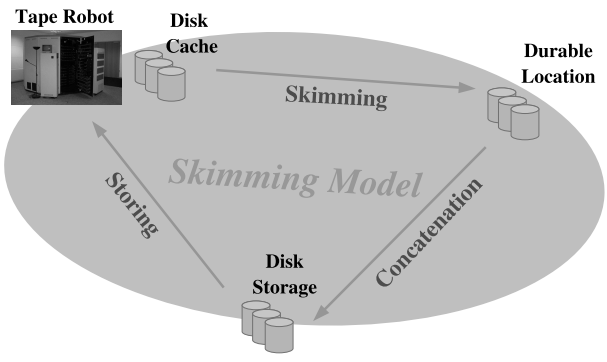

Figure 3: Skimming model.

Figure 3 shows the adopted model.

Skimming

The hadronic dataset has about 60000 files for a total of $\sim 20\ TB$ of disk space. Files are delivered by SAM and processed on the CAF farm in order to be filtered on the basis of eight physical decay modes; selection was performed simultaneously for each event to produce eight output datasets, Intermediate Datasets. Bookkeeping of all the skimming jobs and recovery of the failed jobs rely on the information provided by SAM about not delivered or not consumed files. The eight produced datasets are temporary stored on disks (Durable Location). Depending on the decay mode, the output sizes are reduced to less than 1% to 5% of the original input. The total space of all the output datasets is about 20% of the input dataset.

Concatenation

The Intermediate Datasets produced by the skimming procedure constitutes of about 30000 files of $5 - 10\ MB$ each: a huge number of relatively small files would end up wasting time and resources. It has been necessary to develop and then perform a Concatenation procedure to produce datasets with a lower number of files but larger sizes. The Concatenation produces Output Datasets which consist of $100 - 1000$ files with sizes of $\sim 1\ GB$.

Output Datasets storing

In order to allow CDF users to access these data, the Output Datasets are stored on the FNAL Tape Robot; fast access for offsite Institutions is granted by caching them on disks near remote sites via SAM. A copy is also keept at CNAF on Disk Storage.
Transfers from local disk to Durable Location, to Disk Storage and to tape is done using the sam_upload tool.

THE SAM_UPLOAD TOOLS

The sam_upload tool has been developed on top of the SAM Data Handling System; it can transfer files from local disk to the FNAL Tape Robot, to a Durable Location and to Disk Storage [7] [8]. Figure 4 shows the model which implements this tranfer.

First of all the file which has to be transferred is declared in the SAM Catalog; in this way it can be properly tracked by the SAM System. Then the transfer starts; the authentication is performed by Globus Security Infrastructure which relies on x509 certificates. Files are copied to a Temporary Location by the gridftp protocol. The transfer is wrapped around an enqueuing system in order to avoid overload of the gridftp server in the destination machine. If the transfer fails for some reasons, the file is unregistered so the procedure can be executed a second time in a clean way.
As soon as the file is in the Temporary Location

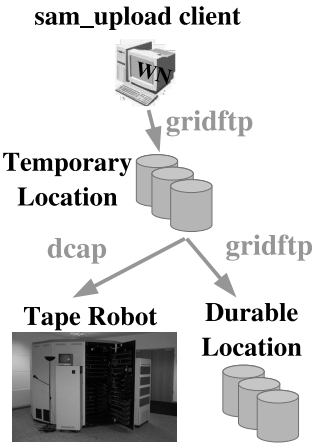

Figure 4: The sam_upload model.

sam_upload delegates the transfer to the Tape Robot or to a specific Durable Location to the SAM System. Tranfers to tape are performed by the File Storage System (fss), a SAM component which wraps the dcap protocol used to talk to the FNAL Tape Robot around an enqueuing system. Transfers to Durable Locations are simply performed by the gridftp protocol. When the transfer is executed, the file is removed from the Temporary Location to save disk space on the temporary machine.
The sam_upload tool relies on many external services such as SAM Catalog, gridftp servers and the File Storage System; in order to limitate damage due to temporary lack of these services sam_upload implements a retry procedure around these sensitive parts of the procedure.

RECOVERY AND CHECKING

Recovery and checking of the skimmig procedure rely on the SAM System and are performed in two distinct but cooperating ways.
Failures of the jobs producing Intermediate Datasets are recovered using the consumption history information of the dataset provided by the SAM System. SAM jobs can keep track of the status of each file in the input dataset: file status can be "delivered", "not delivered yet", "consumed" and "unconsumed". At the end of the job some files in the input dataset can be "not delivered" or "unconsumed". SAM allows creation of a recovery dataset, which is basically a sub-sample of the input dataset containing only files not correctly processed by the previous job. In this way it is possible to run iteratively on the input dataset in order to have the 100% of the inputs processed [9].
Checking of the Concatenation procedure is done using SAM file lineage information. Figure 5 shows how this information is organized for the skimming. Each file in the Hadronic Dataset has eight children in the Intermediate Datasets and each file of the Intermediate

Dataset has one child in the Concatenated Dataset. In

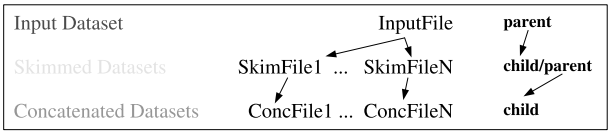

Figure 5: The SAM file lineage capability.

this way SAM can keep track of the parent/children of every input file in the skimming chain, in this way it is possible to know if every file has all the children or parents it is expected to have. The whole procedure has been checked in order to:

- avoid or find file input duplication;
- avoid file input loss.

A DISTRIBUTED PROCEDURE

The skimming is a totally distributed procedure which has been implemented in two different ways.
Data collected up to December 2004 have been processed by the Fermilab CAF resources. Input files have been delivered from the FNAL Tape Robot and Intermediate Datasets have been written on FNAL Durable Locations; Concatenated Datasets have been stored on CNAF at Bologna (Italy) and then copied back to the FNAL Tape Robot.
Data collected from December 2004 up today have been processed on the CNAF CAF resources in Italy via the GlideCaf system [10]: the whole hadronic dataset has been mirrored to CNAF disk cache. Intermediate Datasets have been written on CNAF Durable Locations and the Concatenated Datasets stored to the FNAL Tape Robot.

PERFORMANCES

The Skimming Procedure is now in production with a total failure rate of 1%, basically due to SAM file delivery problems. Problems essentially are of two kinds:

- not all the files of the Input Dataset are correctly delivered. This is probably due to SAM Calatog problems and is under investigation;
- sometimes ($2-3$ times over the whole procedure) files are delivered and considerated well processed even if they are actually not processed. This is due to worker nodes failures and it is very difficult to track it because there is no way to make SAM know about farm problems. However, the incidence of this is very low and can be kept under control by checking the worker node status.

This skimming part has been performed using both CDF onsite and offsite resources. Processing the whole sample has taken two months of user time running with four standard CDF users over an average of 600 KSpecint2000. It has been extimated that running on 2.6 MSpecInt2000, corresponding to all the CDF onsite resources, would have taken 2 weeks. The investment of this time to produce Skimmed Datasets allows CDF users to save time for their analysis jobs: a standard analysis job running over the large hadronic dataset takes at least two months but running on a skimmed dataset takes a maximum of 24 hours, depending on the decay mode.

CONCLUSIONS

The new CDF hadronic trigger produced a large dataset of $\sim$ 60000 files for a total of 20 TB of disk space corresponding to $\sim$ 1 fb^{-1} of data. In the future CDF will collect $\sim$ 2 fb^{-1} per year. Processing this sample directly takes months and wastes resources useful for the Collaboration, therefore it is necessary to find another way to guarantee the availability of this kind of data.
The solution proposed in this article consists in skimming the dataset in order to produce smaller datasets on the basis of the physical decay modes, useful for the Beauty and Charm mesons measurements. Processing these datasets saves CPU resources and allows event reconstruction jobs to be performed in 24 hours maximum. Samples produced are now available both in the FNAL Tape Robot and at offsite Institutions such as CNAF in Bologna (Italy). It is fully integrated with CDF Data Handling and very rubust as demonstrated by the performances.

REFERENCES

[1] http://www.fnal.gov

[2] S. Belforte et. al., IEEE Trans. Nucl. Sci. 46, 933 (1999)

[3] 2. W. Ashmanskas et. al., Nucl. Instr. Meth. A 447, 218 (2000)

[4] CDF Collaboration, "The CDF II Technical Design Report", FERMILAB-Pub-96/390-E (1996)

[5] M. Casarsa, S.-C. Hsu, E. Lipeles, M. Neubauer, S. Sarkar, I. Sfiligoi and F. Wurtwein, "The CDF Analysis Ferm", AIP Conference Proceedings (2005) 794.

[6] http://projects.fnal.govsamgrid/WhatisSAM.html

[7] http://www-cdf.fnal.gov/tiki/tiki-index.php?page=CdfSamUserDocumentation, "How To Store Files In SAM"

[8] A. Fella, S. Belforte, G. Garzoglio, "The sam_upload", CDF Note 7748, (2005).

[9] A. Baranovki, V. Bartsch, E. Lipeles, A. Lyon, I. Sfiligoi, D. Benjamin, K. Genser, "Automated recovery of data-intensive jobs in D0 and CDF using SAM", CHEP 2006 Proceedings (Mumbai 2006)

[10] S. Sarkar et. al, "GlideCAF - A Late Binding Approach to the Grid", CHEP 2006 Proceedings (Mumbai 2006)