

Using SAM datahandling in processing large data volumes

Stefan Stojek

Fermi National Accelerator Laboratory / University of Oxford

Mòrag Burgon-Lyon, Richard St. Denis
University of Glasgow

Valeria Bartsch, Todd Huffman
University of Oxford

Elliot Lipeles, Frank Würthwein
University of California, San Diego

25 June 2004

Abstract

SAM is the new, grid enabled datahandling system of the CDF experiment at Fermilab. Its core capabilities are storing metadata and locations for data files and transferring the files to the nodes to process them. The system is known to work for physics analysis jobs. To accommodate the compute needs for physics analysis jobs, CDF introduced the CAF batch farm system. Now SAM and CAF are used together to allow the reprocessing of large dataset on big compute farms. The problems faced during the transition and their solutions will be discussed.

1 Introduction

The CDF [1] detector is an elementary particle detector at the Fermi National Accelerator Laboratory (FNAL) [2]. This detector produces large amounts of raw physics data during normal operations. These first level data are also called raw data. To be useful for physics analysis, the raw data has to be processed on large compute farms to produce the data of the second level, the so called reprocessed data. These data will then be read by physics analysis programs which produce the results used in scientific publications.

So far all data of the CDF experiment were simply stored on tape and nearly all metadata about the content of the file where kept in the file name. CDF intended to have, by June 2004, 25% of it's computing

CPU from sources not located at Fermilab. Therefore CDF is introducing a new datahandling infrastructure. Its main component SAM [3], a system to store the file's metadata and to transfer the files to the locations where they are needed.

At the CDF experiment the SAM system is accompanied by the farm batch system CAF [4] to analyze the large amount of data from the CDF detector. The CAF system is based on FBSNG [5], a Fermilab farm batch system. In addition CAF includes sandbox handling, Kerberos authorization and Ganglia monitoring. In the future CAF will use Condor [6] instead of FBSNG as its underlying batch system.

In addition the CDF compute model foresees Grid tools to interconnect different sites, whereby the sites have installed the above mentioned tools. The grid exten-

tion to SAM is called Job and Information Management (JIM). The project we present has so far being tested locally at Fermilab but is designed to run at any participating site.

2 SAM metadata storage

The SAM system stores metadata about all files which are stored to the system. File name, size and cyclic redundancy checksum (CRC) are mandatory information for every file with the file name as a unique identifier. If available SAM also stores the information about the physics content of a file. These includes when and with which setup those data were taken and whether the data are real or simulated data. For every file which is the result of processing a file which already exists in the SAM database, this relationship is stored. All these information are important to select the files for a given physics analysis. This information can be used to group files into named entities which form the input dataset for physics analysis.

In addition the SAM system is able to move files to participating sites and clusters if needed. To use this system the user just has to specify the criteria, for the files he wants to analyze, by their physics content. The SAM system will automatically deliver all files to the site /cluster which hosts the users jobs. After the successful transfer SAM will notify the users process about the location of the files at the execution site.

All the file information is stored in a central Oracle database at Fermilab. This includes the static file information as well as the information about the current locations of every file. In addition the database contains information about all processes which ever requested a file and whether they successfully processed this file.

This central database constitutes a single point of failure. We are aware of this problem. Therefore we currently try to move as much information as possible from the central database to the sites.

So far the main usage of the SAM system at CDF has been to deliver files to indi-

vidual physics analyzes. The SAM system knows about all the raw files of the CDF detector this system. Therefore it should be possible for the SAM system to deliver data to the compute farms which process the “raw data” into “reprocessed data”.

3 Reprocessing

Analyzing reprocessed data is typically done with small amounts of data (~ 1 TB) on a small amount of processors (~ 10), or in a short time (\sim days). But reprocessing raw data means processing large amounts of data (> 10 TB) on many CPUs (several hundreds) for a long time (\sim month).

The aim of our project is to enable the system composed of a farm batch system (CAF [4]) and a data handling system (SAM) to reliably reprocess the largest raw data stream of the CDF experiment. This exceeds the requirements for an individual physics analysis by far. The amount of data taken at CDF is larger than those taken at D0, the other high energy physics experiment which uses SAM. Therefore this project is a good test for the scalability of SAM.

Before one can actually process raw data into reprocessed data one has to define the amount of data one wants to process without interruption. Due to the design of the SAM tool we limited the number of parallel processes for a given set of files to 100. This was to ensure that the bookkeeping process would not cause any delays. We also limited the uninterruptable time span to not more than four days.

These datasets were then processed at a CAF system at Fermilab. Since the input files came from a tape system they were each ~ 1 GB in size. Each input file produced several output files, each of them smaller than 1 GB. The information in these files have to go back to a tape system. Therefore it is necessary to concatenate files into 1 GB files. At this stage it became apparent that exact bookkeeping is essential and the SAM system could deliver this.

Typically the files produced by the concatenation will be accessed by individual

physics analyses. In a typical analysis, those files whose raw files were produced at the same time, will be accessed together. Therefore the concatenation process has to ensure that files which will probably be accessed together end up at the same tape.

Another important issue surfaced in connection with the intermediate files. Due to the large amount of data the statistical probability of bit errors became so big that we computed a CRC checksum for each file directly after production and stored it into the SAM database.

Summary

We found that the SAM metadata and file handling system was able to scale well into the region required by this project. This success already triggered the decision of the CDF collaboration to use this system for the entire reprocessing step. Since the stream which was reprocessed in the presented project is by far the biggest at the CDF experiment, this extension should

not present any scalability problems.

In the course of this project we learned that good bookkeeping is essential. And we found that we could not use the tape system as an abstract storage entity, but we needed to adjust to the physical parameters of the media.

References

- [1] CDF home page
<http://www-cdf.fnal.gov/>
- [2] FNAL home page
<http://www.fnal.gov/>
- [3] SAM home page at CDF
<http://cdfdb.fnal.gov/sam/>
- [4] CAF home page
<http://cdfcaf.fnal.gov/>
- [5] FBSNG home page
<http://www-isd.fnal.gov/fbsng/>
- [6] Condor Project
<http://www.cs.wisc.edu/condor/>