DEPLOYMENT OF SAM FOR THE CDF EXPERIMENT

Stefan Stonjek[1,2], Andrew Baranovski[1], Valeria Bartsch[2], Stefano Belforte[3], Mòrag Burgon-Lyon[4],
Gabriele Garzoglio[1], Randolph Herber[1], Robert Illingworth[1], Rob Kennedy[1],
Ulrich Kerzel[5], Art Kreymer[1], Matt Leslie[2], Lauri Loebel-Carpenter[1], Lee Lueking[1], Adam Lyon[1],
Wyatt Merritt[1], Fedor Ratnikov[6], Alan Sill[7], Richard St. Denis[4],
Igor Terekhov[1], Julie Trumbo[1], Sinisa Veseli[1] Steve White[1]
[1]Fermi National Accelerator Laboratory, [2]University of Oxford, [3]INFN / Trieste
[4]University of Glasgow [5]Universität Karlsruhe [6]Rutgers University [7]Texas Tech University

Abstract

CDF is one of the experiments at the Tevatron at Fermilab. One dominating factor of the experiments' computing model is the high volume of raw, reconstructed and generated data. The distributed data handling services within SAM transport these data for physics analysis applications. The SAM system was already in use at the D-Zero experiment. Due to difference in the computing models of the two experiments some aspects of the SAM system had to be adapted. We will present experiences from the adaptation and the deployment phases. This includes the behavior of the SAM system on batch systems of very different sizes and types as well as the interaction between the datahandling and the storage systems, including disk pools and tape robot systems. In particular we will cover the problems faced using large scale compute farms. To accommodate the needs of Grid computing, CDF deployed installations consisting of SAM for datahandling and CAF for high throughput batch processing. The CDF experiment already had experiences with the CAF system. We will report on the deployment of the combined system.

INTRODUCTION

CDF [1] is an experiment at the proton–anti-proton collider Tevatron at Fermilab [2]. Due to the nature of proton–anti-proton physics, CDF needs to store and handle huge amounts of physics data. Since the begin of run-II in 2001 CDF accumulated more than a petabyte of data. A sophisticated datahandling system is necessary to store the data, to catalogue it and to make it available for physics analysis.

CDF decided to use the SAM[3] system which was already in use by the DØ experiment. At the core of the SAM system is an Oracle database, located at Fermilab, which contains, among other, file metadata, indexed by the file names. These meta data contain information about the files operating system attributes as well as their physics attributes. This database also stores all locations of all files in the world wide SAM system. These files can be either stored at a tape robot at Fermilab or at any of the participating machines around the world, called "SAM station".

SAM station

The SAM stations communicate between each other and with the central services via CORBA interfaces. One of the central services precesses the CORBA requests into SQL, queries the database and returns the results to the SAM station. This "db-server" contains most of the business logic of the SAM system, which are the rules which could not be efficiently implemented in the database itself. Another effect of this construction is, that the SAM system is just one client to the Oracle database and participating institutions are not required to purchase Oracle licences.

In addition to the file metadata, SAM also stores information about relationships between different files. The whole lineage from a raw file with data directly from the detector to the n-tuple used for a certain analysis is kept in the SAM system. This includes information about the application and its version used to create that file and the input files used by that application.

All the metadata for the files can be used to create sets of files or datasets. Datasets are named selection rules for files using metadata. The selected lists of files are called snapshots. Both, datasets and their associated snapshots, are stored in the database. Furthermore SAM also does bookkeeping for all physics analysis projects. This allows easy recovery in the case of incomplete analysis jobs.

SAM AND BATCH SYSTEMS

For it's batch processing needs, CDF introduced the CDF Analysis Farm (CAF) [4]. Currently the CAF system at Fermilab consists of 1200 nodes which represent 3200 GHz of compute power[1]. One part of the CAF is using "fbsng" [6] as its underlying batch system, the other part is using "Condor" [7]. The CAF system was originally designed to accommodate the needs of the physics analysis user, but it also performs well in processing raw data.

The user submits his job just once to the CAF system. The CAF runs it as many times as the user requested. But it is not guaranteed that all jobs run at the same time or in a certain order. It is even possible to have large time gaps between the different sections of a job.

When the user submits jobs to the CAF he also has to

[1] measured in Intel Pentium IV equivalents

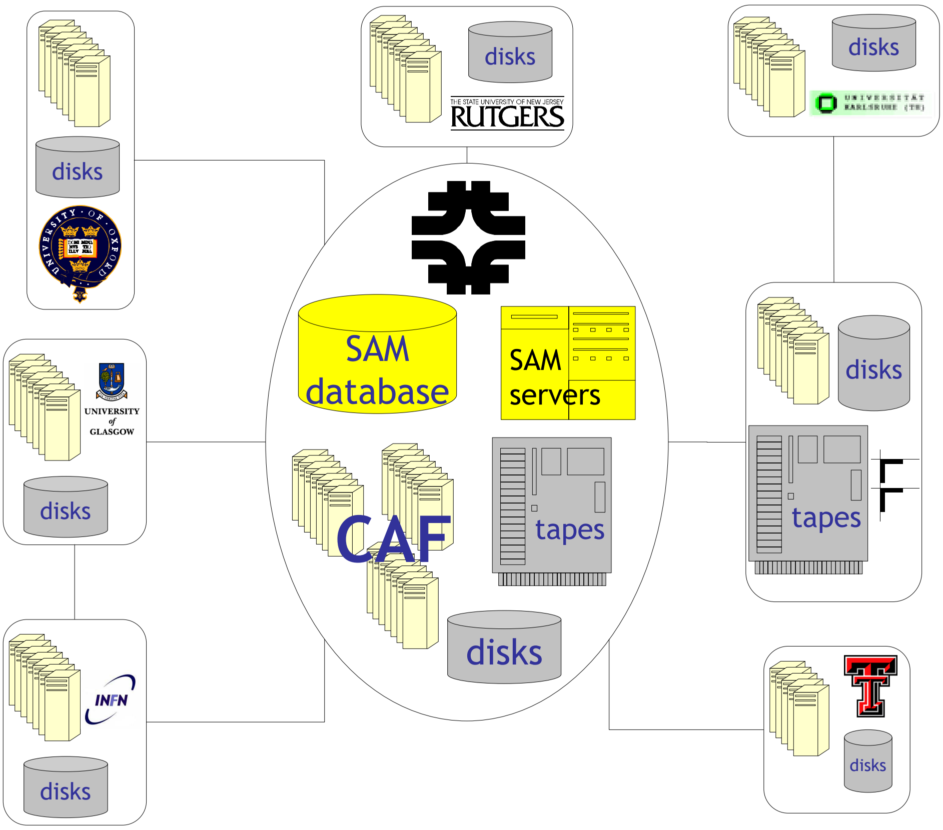

Figure 1: CDF currently has SAM installations in eight countries on three continents. Most stations transfer files directly to and from Fermilab. Some stations use other stations to relay files. Some files get transfered directly from station to station

specify the SAM dataset or snapshot the job should process. This enables the CAF system to notify the SAM system before the files are needed by the processes. At Fermilab all regularly used datasets are kept on disk. This big disk pool is part of a dCache [5] system. dCache is optimized for high performance access from a large number of clients. This allows easy and transparent access from all worker nodes. In this environment SAM is able to handle more than 60 terabytes per day.

Issues with SAM and CAF

One problem arises from a particular feature of the Condor batch system. Enabling "fair share" requires that every job can be preempted within a certain time interval at the beginning. This generates a problem for any system outside of this particular worker node, which records actions from this worker node. The assumption that the job on that worker node is stateless, with respect to the outside world, is not true anymore. To ensure integration of SAM and Condor-CAF, we had to make the preemption interval as short as possible and had to prevent any interaction between the job on the worker node and the SAM system during the interval in which preemption could occur.

Since the SAM system keeps track of all the files delivered to the different processes of a job, the data in the SAM system are important to recover jobs in which some processes have failed. To facilitate recovery better communications between the SAM datahandling and the batch system would be helpful.

WORLDWIDE DEPLOYMENT OF SAM FOR CDF

Since the compute power of the computer center at Fermilab does not satisfy the compute needs of the CDF experiment, CDF decided to use compute power located at the participating institutes. To facilitate the installation of SAM at the remote locations, the software necessary for the operation of a SAM station was packaged into several software products which can be easily deployed by Fermilabs ups/upd [8] system. We were able to bring the installation to a state where a person at a remote institute has to type just one command and send one e-mail to install a functional SAM station.

Due to local differences such as preexisting batch systems and different hardware configurations, all installations are different. Only a few installations use local dCache systems. Several installations use different batch systems.

So far CDF has functional installations of SAM software in eight countries on three continents. They represent a total compute power of 1800GHz with 35 terabyte disk capacity and they are mainly used for physics analysis and Monte-Carlo generation.

Issues with the worldwide deployment

During the deployment phase we faced many minor and some more difficult problems.

All SAM stations communicate which each other and with the central services via CORBA calls. Since the execution of the respective CORBA functions might take time, CORBA-callbacks were selected as the appropriate solution. But those callbacks were often stopped by local firewalls. Another possible implementation would have been to keep the TCP/IP connection open for the duration of the execution of the CORBA function. But this would have caused firewall problems too. So far, we depend on the cooperation of the local firewall administrators.

Another problem is related to the ongoing development of the system. Generally it is no problem to change software versions whenever the responsible administrator has enough time to perform that task. But for some software updates, in particular those which included a CORBA interface change, it was critical to change software versions at all participating institutions during a small period of time. Sometimes we were able to relax that requirement by providing central services with the old and the new interface.

We are able to run CDF software at all participating institutions because they ensure the installation of the correct release of the entire CDF software package, consisting of all the products a CDF physicist requires to run analysis. This requirement may cause problems with Grid installations which will not be experiment specific. Some sites solved this problem by dedicating specific machines in their cluster to each participating experiment. The experiments were allowed to export their software from that machine to the cluster. Another option would be to include all necessary parts of the CDF software into the submitted job. Since this causes problems due to the total size of the CDF software distribution we are examining the possibility to store, during the job submission process, the executables and the CDF software packages into SAM. From there they can be retrieved by the worker node when they are needed.

CONCLUSIONS

CDF was able to deploy the SAM datahandling system to many of its collaborating institutions. Thereby CDF gained 35% of its total compute capacity. Due to the success of this first step, CDF plans to increase this number to 50% in the future.

Despite the problems faced during the deployment phase, the CDF compute model is a first big step to a compute model in the Grid-age.

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] dCache home page
http://www.dcache.org/

[6] FBSNG home page
http://www-isd.fnal.gov/fbsng/

[7] Condor Project
http://www.cs.wisc.edu/condor/

[8] ups/upd home page
http://www.fnal.gov/docs/products/ups/