

# Reprocessing DØ data with SAMGrid

Frédéric Villeneuve-Séguier

Imperial College, London, UK

*On behalf of the DØ collaboration and the SAM-Grid team.*

## Abstract

The DØ experiment studies proton-antiproton collisions at the Tevatron collider based at Fermilab. Reprocessing, managing and distributing the large amount of real data coming from the detector as well as generating sufficient Monte Carlo data are some of the challenges faced by the DØ collaboration. SAMGrid combines the SAM data handling system with the necessary job and information management allowing us to use the distributed computing resources in the various worldwide computing centers. This is one of the first large scale grid applications in High Energy Physics (in particular as we are using real data). After successful Monte Carlo production and a limited data reprocessing in the winter of 2003/04, the next milestone will be the reprocessing of the full current data set by this autumn/winter. It consists of ~500 TB of data, encompassing one billion events.

## 1. The DØ experiment

The DØ experiment is a worldwide collaboration of about 600 physicists from over 20 participating countries. The detector was designed to study high energy collisions between protons and antiprotons delivered by the Tevatron collider, based at Fermilab (Chicago, USA).

### 1.1 Tevatron at Fermilab

The Tevatron is the highest energy particle collider currently operational. A first successful data taking period from 1992 to 1996, Run I, led the two collider experiments CDF and DØ to the top quark discovery. A second run of collision, Run II, started in March 2001. The Tevatron has been significantly upgraded to deliver luminosity up to  $L = 2 \cdot 10^{32} \text{ cm}^{-2} \cdot \text{s}^{-1}$ , with an energy at the center of mass  $\sqrt{s} = 1.96 \text{ TeV}$ .

In order to reach such high luminosities, the whole accelerator chain had been upgraded as well. The “Main Ring” from Run I has been replaced by the “Main Injector”, housed in a separate tunnel to allow simultaneous operation of the Tevatron and other experiments (fixed target). An antiproton recycler has been installed in the same tunnel, and will allow a further increase in the total luminosity delivered by the accelerator.

The Fermilab accelerator system for Run II described in Figure 1.1, consists now of five stages of accelerators : a Cockcroft-Walton pre-accelerator, a linear accelerator (Linac), one

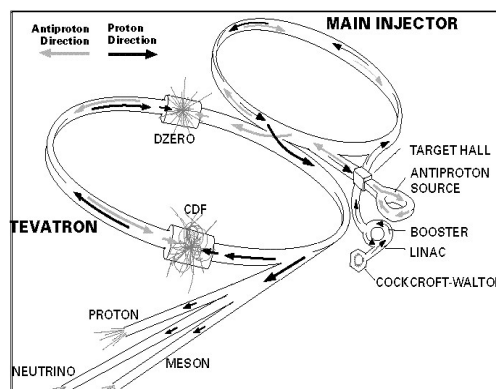


Figure 1.1: Tevatron and accelerator chain in the Run II configuration at Fermilab (Chicago).

synchrotron (Booster), the “Main Injector” and the Tevatron. Two storage rings device, the Debuncher and the Accumulator, are used to select and accelerate antiprotons after their production in the Target Station.

### 1.2 The DØ detector

The DØ detector, originally build for Run I [1], has been designed to study a large spectrum of high energy physics phenomena such as heavy flavour physics ( $b$  and  $t$  quarks), electroweak measurements, searches for Higgs boson(s), QCD studies, and searches for physics beyond the Standard Model.

Physics goals for Run II have led to several upgrades in order to improve performance. The detector setup for Run II is described in Figure 1.2. The tracking system now consists of a silicon vertex detector (SMT), a scintillating fibre detector (CFT), and a 2Tesla solenoid

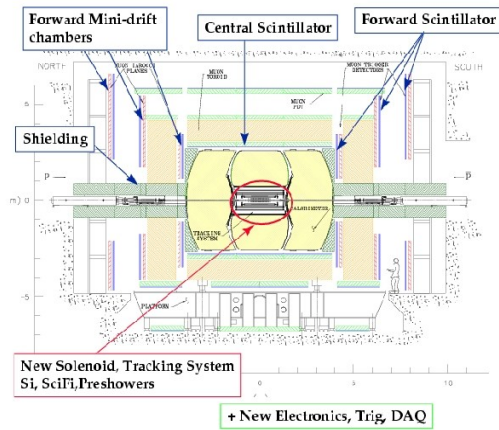


Figure 1.2 : The DØ detector and its major upgrades for Run II.

magnet. The trigger system and the readout electronics have been completely re-designed to better suit the Run II luminosity (more than 2 million interactions per second). The calorimeter has been enhanced with the addition of a central pre-shower system (CPS). The muon system upgrades extend the angular coverage, and new mini-drift chambers increase muon identification and resistance to radiation.

### 1.3 Data and computing distribution

Collisions that produce potentially interesting events are selected through the trigger system. If any trigger condition is fired, digitized output from each sub-detector is stored in a format called RAW data. These data are then processed through the DØ reconstruction software in order to create higher level objects (such as tracks, vertices, electromagnetic objects, muons, etc ...) , which are then used for physics analysis. Calibration constants from sub-detectors needed for object reconstruction are stored in a calibration database.

Reconstructed events are then condensed in another data format called TMB (Thumbnails). In addition, metadata for both RAW and TMB files are created, with information on software version, input file, and data format. This procedure is described in Figure 1.3.

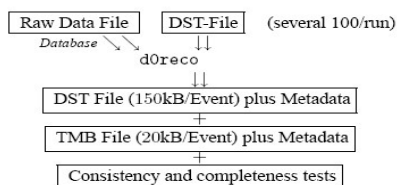


Figure 1.3 : Application flow of a DØ reconstruction job. Initially, RAW data were stored in an intermediate format (DST), and the TMB file size had to be increased up to 50-70 kB/Event.

The DØ experiment has recorded more than 1000 million collisions since the beginning of the Tevatron Run II. This corresponds to more than 500 TB of data. In addition, huge samples of Monte-Carlo simulated events have been produced for physics analysis, which contribute consistently to the whole storage volume.

Distributed computing at “remote” sites is the solution adopted to enable access to such large datasets. All DØ Run II Monte-Carlo data are produced outside of Fermilab at remote sites. A system of regional analysis centers (RAC) was established to supply the associated institutes with the data for analysis needs. Remote MC production and automatic on-demand data distribution for remote analysis have been part of the DØ model since Run II was started.

The reconstruction software is regularly improved as our understanding of the detector improves. To benefit from these developments, both data and Monte-Carlo files have to be reprocessed with the new software. Fermilab on-site computing resources are not sufficient to perform such tasks. Over than 10 sites in 6 different countries have joined this effort. A limited data reprocessing was carried out during winter 2003 [3], and Monte-Carlo production using SAM-Grid in 2004 [4].

## 2. SAM-Grid

The DØ computing model, originally based on distributing computing through a data grid (SAM), has progressively evolved to the use of a computational grid (SAM-Grid) based on standard grid common tools.

### 2.1 Introduction

SAM-Grid is an integrated grid infrastructure for job, data and information management with SAM (Sequential Access via Metadata [5, 6]), the data handling system used by both the DØ and CDF experiments. Standard grid tools and protocols, such as the Globus Toolkit and Condor-G, are used for job and information management (JIM) [7, 8].

The full SAM-Grid infrastructure, job and information management system as well as data handling, has been deployed since January 2004 (SAM has been operational since 1999).

Over the last year, the system has been tailored for both Monte-Carlo production and data reprocessing within DØ. These experiences lead to a better understanding of the SAM-Grid service requirements, the standard middleware, resources and their management. During the 2004 MC production phase, 50 million events were generated by DØ.

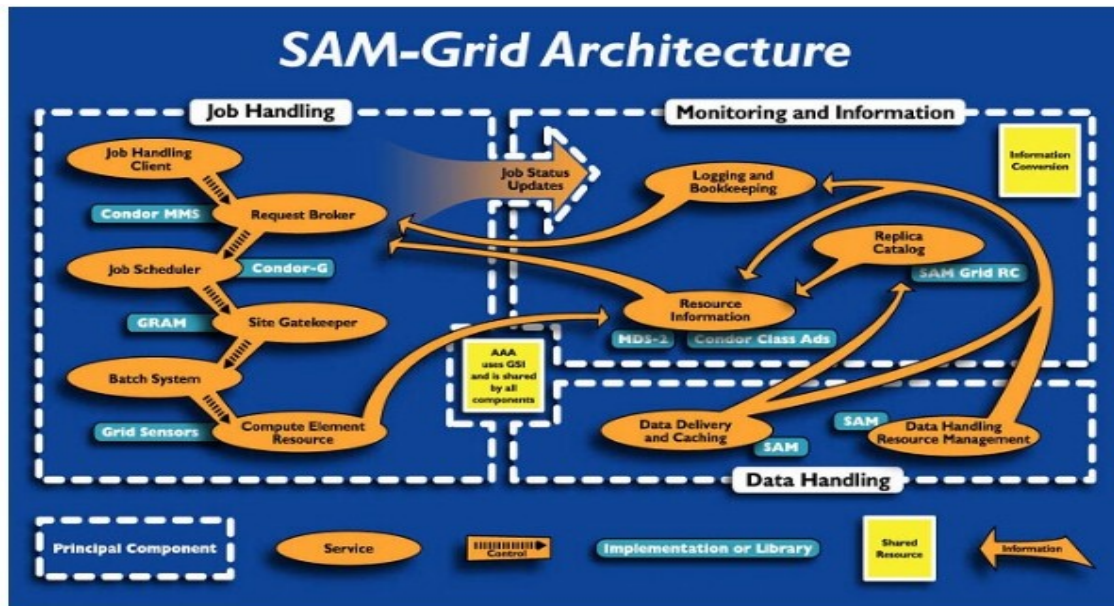


Figure 2.1 : The SAM-Grid architecture. Jobs handling and management integrate standard grid tools, such as the Globus Toolkit and Condor-G. Data handling is performed with SAM (Sequential Access via Metadata system).

The inefficiency in event production due to the grid infrastructure has been reduced from 40% to 1-5%.

## 2.2 Data handling management

SAM is a data handling system 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.

The SAM system has been mainly designed to track locations and informations about a file (or a set of files) requested by a user, provide storage and delivery utilities on request, and provide methods of job submission for local or grid-aware systems. SAM-Grid uses the “data grid” represented by the SAM stations system, as shown in Figure 2.1.

## 2.3 Architecture and Deployment

The three components of the SAM-Grid architecture, “Job handling”, “Monitoring and information”, and “Data handling”, are presented in Figure 2.1. This figure describes the main services performed for job submission and execution.

A grid architecture can generally be organized in two layers : the grid layer, which encompasses global services, and the fabric layer, which includes services whose scope is restricted to a local individual site. Both layer interact via an interface, which adapts the directives of the grid services to the specific configuration of the fabric at the site.

The “grid” and “fabric” services of the SAM-Grid architecture are shown in Figure 2.2.

The SAM-Grid “grid” layer includes the resource selection service, the global data handling service, such as metadata and replica catalogue, and the submission services, which are responsible for queue maintenance and for interacting with resources at the remote site. The “fabric” layer includes the local data handling and storage services, the local monitoring, and the local job scheduler.

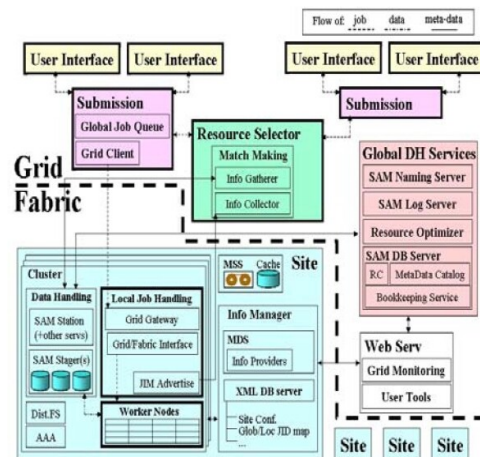


Figure 2.2 : Diagram of the SAM-Grid architecture organized in grid and fabric services. The grid services are global in nature, while the fabric services are limited to the scope of a single site.

The SAM-Grid had to develop its own interface between both layers of services, as standard grid/fabric interfaces, provided by the Globus Toolkit in the form of job-managers,

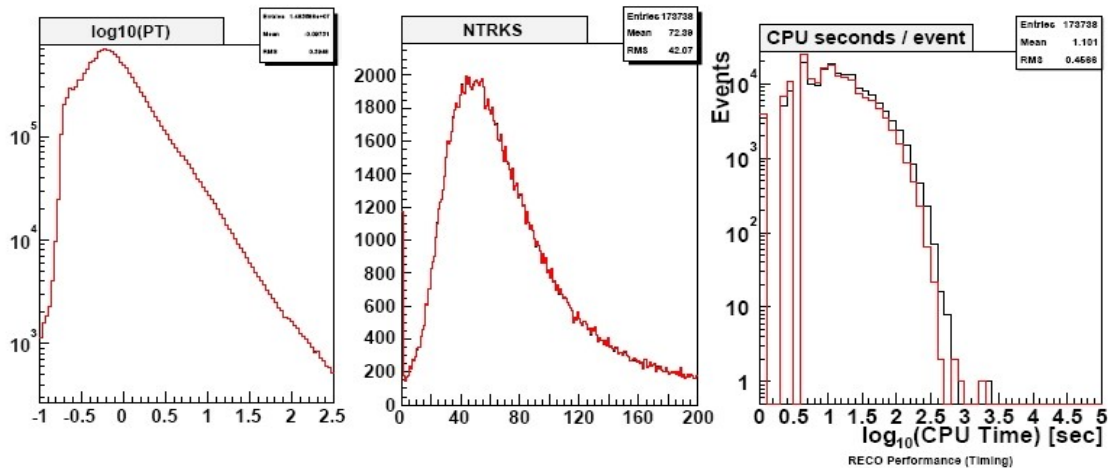


Figure 3.1 : Remote sites have to be certified before joining the official production. Here, comparison of results produced in Lyon and Manchester show good agreement, except for the variables which depends on the speed of the CPU which is related to site configuration.

presented several problems with jobs on the SAM-Grid (Flexibility, Scalability [4]).

The job-managers interact with the local batch system through a “batch adapter”, which is configured for each specific remote site installation. In order to reduce the scalability problems encountered with standard job-managers, the SAM-Grid job-managers aggregate multiple batch jobs from single grid jobs : each grid job is split at the gateway node into multiple local jobs according to the local or experiment policies.

The deployment of SAM-Grid consists of installing and configuring at remote sites the software required to accept jobs from the grid services. Sites generally provide a gateway machine and support to install standard middleware from the Virtual Data Toolkit (VDT) distribution, the SAM-Grid grid/fabric interface, and the client software for the fabric services. The worker nodes of the cluster do not require any pre-installed software or running daemons. SAM-Grid is now deployed at about a dozen sites in the Americas, Europe, and India.

There is an ongoing programme of work on SAM-Grid interoperability with other grid, in particular LCG and OSG. Some prototype mechanisms to forward jobs automatically from SAM-Grid to LCG have been developed, and established in Germany and in France..

### 3. DØ P17 data reprocessing

During winter 2003/2004, about 100 million data events were reprocessed with the P14 version of the DØ software. Since then, the understanding of the detector and the reconstruction software have made major improvements (objects identification, detectors calibration).

The reprocessing of the whole DØ data (~ 1000 million events, ~ 500 TB) with the P17 version of the reconstruction software started in April 2005. While the P14 reprocessing was achieved with distributed computing rather than a grid structure, this new campaign will fully benefit from the improved SAM-Grid expertise with the DØ infrastructure.

#### 3.1 Contributing sites

The P17 reprocessing involves computing resources in six countries : Germany, France, Czech Republic, the United Kingdom, Canada, and the USA. The computing power available including all participating institutes is equivalent to more than 3000 CPUs. This CPU size represents the computer power estimated to be available in number of CPUs equivalent to a 1 GHz Pentium-III processor.

#### 3.2 Sites certification

Each participating site has to be certified to be allowed to join the official production. It is necessary to check for any numerical deviation due to varying operating system versions or configuration problems.

The certification procedure consists of two steps : TMB production from RAW data files, and TMB file merging. At each step, many physics variables are plotted and compared with a central reference dataset, but not processed using SAM-Grid. Any discrepancy in the distributions could indicate a problem in the configuration of the site. Figure 3.1 shows comparison plots between 2 production sites (Lyon and Manchester). Plots show good agreement, except for the variables which depend on the speed of the CPU.

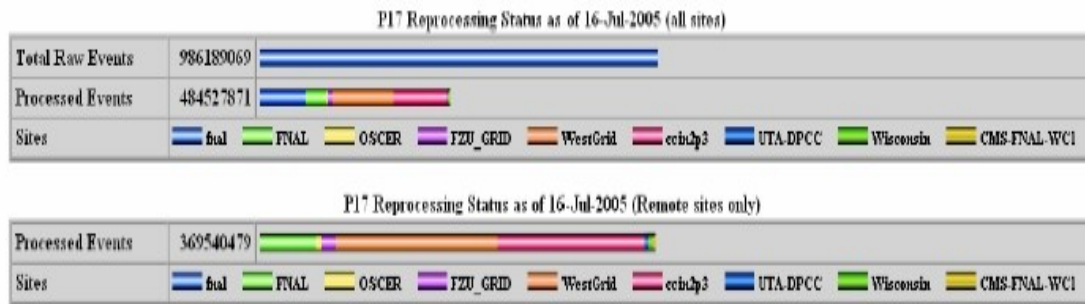


Figure 3.1 : Current P17 reprocessing statistics, as of 16/07 (not all participating sites shown).

### 3.3 Operation and status

All the data to be reprocessed are split into hundreds of dataset, created manually, assigned and distributed to the contributing sites. The number of dataset was adapted to each site's expected capacity and available disk storage. In order to speed up the reprocessing task, most sites already have pre-staged the files of the assigned dataset into the SAM cache before the processing jobs were started.

The reprocessing of one dataset proceeds in 2 steps (as for the certification). In a first grid job, each RAW data file from the dataset is sent to the batch system and is reprocessed with the P17 DØ reconstruction software. The second grid job consists of merging the TMB files produced in the previous job. The TMB files created are too small (~50 kB) to allow an efficient access to them via the tape system. Around 8-10 of them are merged into a single file. Merged TMBs and associated metadata are then stored in SAM.

The P17 reprocessing has been ongoing since 3 months. As of July 2005, 2 sites are fully operational (IN2P3, Westgrid), and 7 other sites have been certified or are currently under certification. However, nearly 500 million events (50% of the P17 dataset) have been produced up to now.

### 4. Conclusion and outlook

Since the beginning of the Tevatron Run II in March 2001, the DØ experiment has recorded about 1000 millions events, which corresponds to more than 500 TB. Managing such amounts of data requires distributed computing, which lead the DØ collaboration to adopt SAM-Grid, a grid infrastructure for job and information management as well as data handling.

Reprocessing the entire DØ dataset using SAM-Grid, deployed at 9 remote sites from 6 countries, is the most recent and ambitious reprocessing challenge faced by the experiment. The P17 reprocessing is ongoing since April 2005, and 25% of the participating sites are fully operational while the other 75% were

recently certified or are being under certification. However, already 500 million events have been reprocessed so far (~50% of the P17 dataset). The initial plan was to run the reprocessing for 6 months, and this goal seems reasonably achievable.

### 5. References

- [1] DØ Collaboration, S.Abachi et. al., *The detector*, **Nucl. Instrum. Meth.** A338 (1994) 185-253.
- [2] DØ Collaboration, S. Abachi, et al., *The upgrade : the detector and its physics*, FERMILAB-PUB-96-357-E.
- [3] M. Diesburg, D. Wicke, *Data reprocessing n worldwide distributed systems*, FERMILAB-CONF-04-512-CD, Dec 2004, To appear in the proceedings of Computing in High-Energy Physics (CHEP '04), Interlaken, Switzerland, 27 Sep - 1 Oct 2004.
- [4] G. Garzoglio, I Terekhov, J. Snow, S. Jain, A. Nishandar, *Experience producing simulated events for the Dzero experiment on the SAM-GRID*, FERMILAB-CONF-04-480-CD, Dec. 2004, Prepared for Computing in High-Energy Physics (CHEP '04), Interlaken, Switzerland, 27 Sep - 1 Oct 2004.
- [5] I. Terekhov, et al., *Meta-Computing at DØ*, in Nuclear Instruments and Methods in Physics Research, Section A, NIMA14225, vol 502/2-3 pp 402-206.
- [6] I Terekhov, et al., *Distributed Data Access and Resource Management in the DØ SAM System*, in Proceedings of the 10<sup>th</sup> IEEE International Symposium on High Performance Distributed Computing (HPDC-10), San Francisco, California, Aug 2001.
- [7] I. Terekhov, et al., *Grid Job and Information Management for the FNAL Run II Experiments*, in Proceedings of Computing in High Energy and Nuclear Physics (CHEP03), La Jolla, Ca., USA, March 2003.
- [8] G. Garzoglio, et al. *The SAM-Grid project : architecture and plan*, in Nuclear Instruments and Methods in Physics Research, Section A, NIMA14225, vol 502/2-3 pp 423-425.

