Eur Phys J C **33**, s01, s1008–s1010 (2004)
Digital Object Identifier (DOI) 10.1140/epjcd/s2004-03-1797-6

**EPJ C direct**
**electronic only**

# Distributed computing concepts in DØ

Daniel Wicke

Bergische Universität Wuppertal, Gaußstr. 20, D-42097 Wuppertal, e-mail: `Daniel.Wicke@physik.uni-wuppertal.de`

Received: 15 October 2003 / Accepted: 10 March 2004 /
Published Online: 31 March 2004 – 

**Abstract.** The DØ experiment faces many challenges enabling access to large datasets for physicists on four continents. The new concepts for distributed large scale computing implemented in DØ aim for an optimal use of the available computing resources while minimising the person-power needed for operation. The real live test of these concepts is of special interest for the LHC Computing GRID, LCG, which follows a similar strategy.

**PACS.** 07.05.Kf

## 1 Introduction

Some of the most interesting events in $p\bar{p}$ collisions at the Tevatron are very difficult to distinguish from the overwhelming QCD background. Top and Higgs particles are moreover very rare. The Tevatron experiments DØ and CDF therefore record large amount of data for later analysis and detailed studies. While taking data each experiment records roughly 500 GB of raw data per day. Reconstructing these events adds another 1.1 TB/day in DØ. Within the last two years around 350 TB have been accumulated. Providing this data volume for physics analysis performed by more than 100 people is one of the challenges DØ faces.

With this amount of data providing the necessary overall IO rate is a difficult task. As only a fraction of the data can be stored on disks, tape mounting leads to major dead-times. DØ follows a combined concept of locally optimising the resource usage and distributing the data globally. For an international collaboration like DØ with around 50% of the collaborators working at non-US institutes the second step is of special importance to provide easy data access not only for those resident near Fermilab but also to those working remotely often on another continent.

In addition, to analyse these data, a sufficient number of simulated events need to be produced for selection studies and the estimation of detector effects. To fulfil the requests the production is distributed to many sites. Production chains with ever changing versions and parameters are however complicated to handle. To ease the production an automatic handling of large batches of jobs was developed.

The mentioned concepts to meet the outlined requirements are discussed in the following.

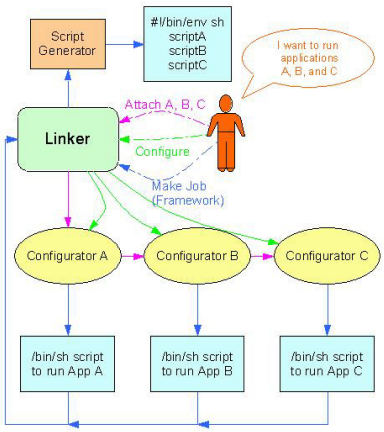


**Fig. 1.** `Runjob` separates the details of how to call individual programs ($A$, $B$ and $C$) from the operational details of which programs (or program versions) to combine

## 2 Management of large batches of jobs

To reduce the person-power needed for Monte Carlo production with its ever changing versions and parameter settings, D0 developed the work-flow management system `Runjob` [1,2]. A work which is continued in collaboration with CMS [3].

It automates the linking of several processing tasks into a single job based on a job description and allows to separate the configuration of the individual programs from the definition of the work-flow (Fig. 1). Thus the individual steps of Monte Carlo production, like event generation, detector simulation and reconstruction, can be configured by the corresponding experts. Those performing the production can in turn concentrate on the work-flow, i.e. which program versions to combine.

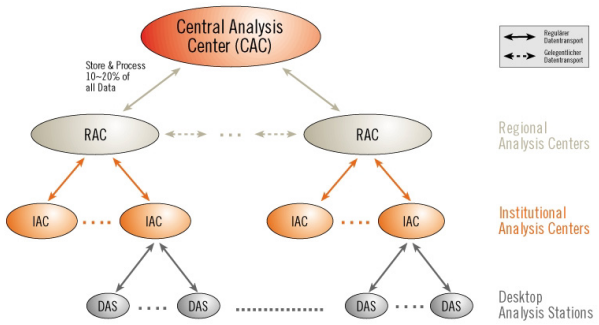


**Fig. 2.** Data distribution scheme within DØ. The distribution of data is done in a tree structure from the central repository (CAC) to regional analysis centres (RACs), further to the institutional centres (IACs) and finally to the physicists desktop (DAS)

Furthermore, the coherent description of the production work-flow allows the preservation of the meta-data describing what actually was run during the production. DØ stores these meta-data along with the produced simulated data in the SAM-system described below.

Recently the work-flow description was extended to allow for job parallelisation. This way a single production can be run in many parallel jobs at a given site. `Runjob` takes care about the necessary initialisation and termination tasks, e.g. the combination of results of those many parallel jobs.

## 3 Optimised use of local resources

Besides person-power the optimised use of computing resources is important to fully exploit the physics potentials of DØ. With the given amount of simulated and real data only a fraction of the available information can be kept on disks. All other data are stored within tape robots.

To avoid inefficiencies due to tape mounting the access to data needs to be optimised. DØ has developed a data management system (named Sequential Access through Meta-data, SAM [4]) which exploits that the order of the stored physics events is of no interest for the analyses: Instead of looping over a given list of files the user requests a dataset. The order in which the files belonging to this dataset are processed is optimised to minimise the overall number of tape mounts. The definition of a dataset can be performed by the individual users based on the meta-data which describe the content of the files.

SAM also includes user-level bookkeeping and is a central component of the DØ software environment. It has recently been adopted by CDF.

## 4 Worldwide distribution of data

Even after optimising local resource utilisation and tape mounting, the access to the data is a bottleneck. To further improve the accessibility of data especially for physics analysis in summer 2002 a scheme for data distribution was outlined [5] (Fig. 2). Besides the central data repository, which holds all DØ-data, several regional centres should hold copies of the data used for analysis and provide them to users of their region.

These regional centres in addition serve the institutions in that region such that collaborators in these institutions can develop and test their analysis at their local computer cluster or even their desktops.

With the tree structure DØ hopes to minimise the necessary amount of data copies over long distances. Besides storage such regional centres should provide a reasonable amount of CPU power to analyse the stored data.

### 4.1 Prototype

To test the concept of regional centres which serve the associated institutes, a prototype was set up at the German Grid Computing Centre in Karlsruhe (GridKa) [6] with the five (now six) German institutes associated.

With this prototype the main parameters of the system should be tested: The required network bandwidth and the manpower to run the regional centre. Moreover shortcomings of the DØ software should be discovered.

GridKa was established in 2002 and is rapidly growing. The centre is currently shared by 8 HEP experiments. Its (at the time) roughly 180 compute nodes are set up with NFS-shared user home directories and NFS-shared experiment specific disk areas. No experiment specific code can be set up on the compute nodes. To allow for pre-Grid usage of the system, each experiment has a so called software server, on which experiment specific software can be installed and which serves for user login. These specifications are quite different from the setup used on the systems at FNAL, which are exclusively used by DØ.

Unfortunately, the required changes to the DØ-software couldn't, in all cases, be implemented in a site independent way, such that an adaption of each version and to each computing cluster is still necessary. To avoid these tasks in the future a standard DØ computing environment needs to be defined which is flexible enough to deal with all possible cluster configurations.

The network bandwidth to Fermilab is sufficient to continuously download the most condensed data format (Thumbnails). When requesting large datasets in one go transport rates of around 3.3MB/s (8TB/day) are observed.

All changes required for doing DØ analysis at GridKa were available by end of 2002, with the exception of luminosity access. During January several German collaborators used the additional resources to finish their analysis for Moriond in time. With this successful end-user test the value of regional centres was finally proved.

The experiences with this prototype show, that besides the technical achievements, it is of great advantage for the users to work on a centre in their own time-zone, where user problems can be solved by the operators during the usual working hours.

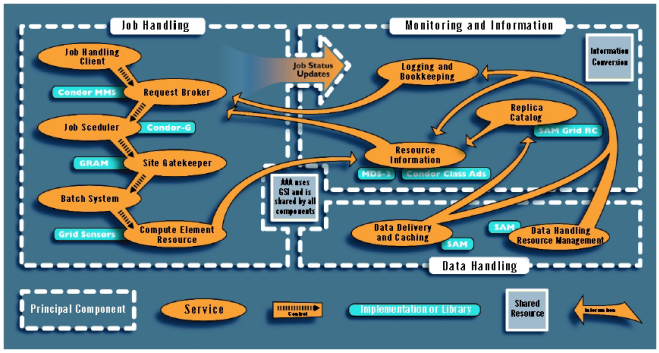


**Fig. 3.** Components diagram of the SAM Grid project

## 5 A GRID for DØ

While the regional analysis centre in Germany has proven its value, the effort that needs to be taken by the user is still large. The analysis code needs to be copied manually to GridKa and needs to be recompiled on the head node before it can be run. This procedure currently needs to be repeated for each site that will be used. Moreover the user needs to track the availability and performance at different sites in order to make even a reasonable choice about where to run his/her project.

To allow for an automatic dynamic adaption to the actual situation GRID tools are required. The JIM project [7,8,9,10] is an integration of existing Globus [11] tools with the SAM system [4] used by both DØ and CDF. JIM uses Condor [12] as its resource broker (Fig. 3).

An initial version of JIM has been installed at several sites within DØ including the GridKa cluster. First experiences with physics analyses in a GRID like environment are expected in the near future.

## 6 Summary

To maximise the physics output DØ aims to optimise the use of its resources while reducing the person-power needed to maintain operation.

The data management system SAM provides user-level bookkeeping and optimises the use of storage resources. World wide distribution of data through regional centres to individual institutions adds additional resources for standard tasks like physics analyses, Monte Carlo production or data (re)processing. JIM integrates Globus and Condor based GRID tools to reach a coherent access to the globally distributed resources and to allow for global optimisation of their usage. `Runjob` eases the handling of large batches of jobs.

With these concepts DØ profits from additional, better exploited resources and reduced operation tasks. At the same time many of the distributed computing concepts foreseen for LCG [13] are tested in a real live environment, which is of strong interest for the preparation of the LHC experiments.

## References

1. G. Graham and D. Evans: CHEP Proceedings (2001)
2. G. Graham, D. Evans, and I. Bertram: CHEP Proceedings (2003), ArXiv:cs.dc/0305063
3. `http://www.uscms.org/s&c/testbed/Tiger/SHAHKAR/Shahkar.htm`
4. Mike Diesburg et al.: DØ-Note 3464, (1998)
5. I. Bertram et al.: DØ-Note 3984, (2002)
6. Grid Computing Centre Karlsruhe (GridKa): `http://www.gridka.de/`
7. Rod Walker: Beauty 02, (2002)
8. I. Terekhov: Nucl. Instrum. Meth. A **502**, 402–406 (2003)
9. A. Baranovski et al.: Nucl. Instrum. Meth. A **502**, 423–425 (2003)
10. A. Baranovski et al.: (2003), ArXiv:cs.dc/0307007
11. I. Foster and C. Kesselman: Intl J. Supercomputer Applications **11(2)**, 115–128 (1997)
12. D.H.J. Epema, M. Livny, R. van Dantzig, X. Evers, and J. Pruyne: Future Generation Computer Systems **12**, 53–65 (1996)
13. LHC Computing Grid Project: `http://cern.ch/lcg/`