Analysis tools for the LHC experiments

Dietrich Liko

E-mail: Dietrich.Liko@cern.ch
CERN, CH-1211 Geneva 23, Switzerland

Abstract. In preparation for the startup of LHC the experiments have developed analysis models to address the computing needs given by the large amount of data and the required resources. As the computing needs exceed the capability of local resources, the experiments rely on large computing grids. It is clear that batch type analysis will play an important role, but there is also an increasing interest in more interactive applications. In the presentation the status of actual tools developed by the experiments for the analysis of data on the grid will be reviewed.

1. Introduction

As the startup of the LHC collider is coming closer, it is an interesting moment to compare the strategies the different experiments are taking to prepare themselves for the data analysis. The computing resources requirements for the operation of the LHC experiments exceed both in CPU and storage resources the capabilities of the computing center at CERN. It is therefore required to interconnect the worldwide resources available to the experiments using grid technology.

Grid technology is notoriously known to be difficult and up to now the grid has been mainly a tool for computing experts of the experiments. To enable the average user, the physicist of the experiments, the experiments have been developing analysis tools to simplify sending analysis jobs to the infrastructure. We will review the status of these tools and investigate their use. This will allow us to address the question if the grid is already a tool for everybody or only useful for experts.

1.1. The grid and the hierarchical tier model

Following the original ideas of the Monarc Working Group the LHC computing grid (LCG) is organized in a hierarchical fashion [1]. CERN as the place of data taking is the so-called Tier-0. Here computing resources are required to perform initial processing and calibrations. It is also required that data is written to permanent storage. As soon as possible data is being exported to a set of large computing centers, the so-called Tier-1's. To provide redundant data storage also these center will save data on permanent storage, which as of today is based on magnetic tapes. The CPU resources at these centers are then required to perform reprocessing of the data. A subset of data relevant for analysis will then be further exported to the so-called Tier-2 centers. These centers provide resources for analysis and for the production of simulated data. In addition small computing facilities at participating Universities and Laboratories are referred as Tier-3 centers. There main task is to provide computing capacity for local tasks, but opportunistic usage of this resources by the experiments is envisaged. Based on the common

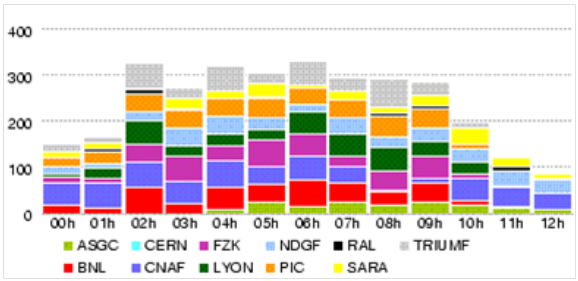

Figure 1. Data export rates in MB/second for the ATLAS experiment during the so-called M4 cosmic muon data taking.

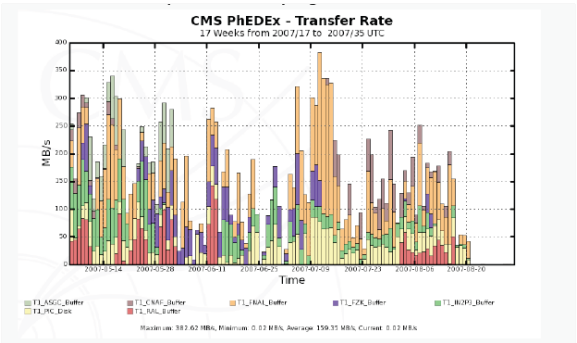

Figure 2. Data export rates for the CMS experiment.

grid infrastructure the experiments have arranged their computing models and prepare their operation (see for example [2]).

1.2. Data export

An essential feature to enable grid based analysis is efficient an efficient data distribution mechanism. The experiments have build data management layers on top of the middleware components that control and manage the data flow (DDM for ATLAS [3, 4] and PhEDEx for CMS [6]). These systems are responsible for data distribution according to the model. They have demonstrated their readiness by export of data from the Tier-0 to the associated Tier-1 centers.

Figure 1 shows the transfer of real data recorded recently by the ATLAS experiment during data taking using cosmic radiation. For actual data taking an export rate in the order of one GB/second is required. Figure 2 shows a similar measurement by the CMS experiment during a test of their system.

While these experiences are encouraging, the export to the larger number of Tier-2 centers is sometimes still problematic.

1.3. Batch based analysis and interactive models

As first consideration the experiments are concentrating on establishing analysis model based on the well known batch model. Computing tasks are formulated as jobs, which are then send to the infrastructure. The need for this strategy is simply given by the amount of data that has to be analysed for a typical year of data taking. It is well imaginable that the amount of data to be analysed is in the order of hundreds of terabytes (TB). To be able to process such a large amount of data it is required to split a computing task in many subjobs and perform processing in parallel. Nevertheless it is clear that obtaining the results of such a calculation will be rather in the order of hours or even days. A batch model is well suited for such a computing tasks and it offers also a well understood model of resource sharing.

The analysis chain being setup by the physicist can be understood as a continuous data reduction. While it is required to use worldwide resources to process the large datasets, the data used for the final analysis will be significantly more compact. This will allow to establish a much mode interactive mode of working where the effect of changes in the selection or the algorithms can be studied in a much more efficient way. Therefore also interactive analysis tools are also of considerable interest.

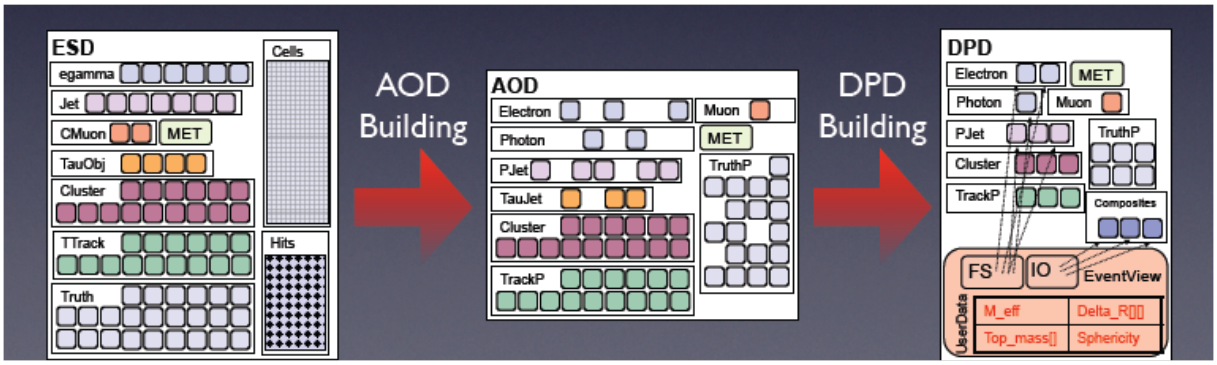

Figure 3. The ATLAS analysis model. The ESD data are the output of the detector reconstruction. By selecting quantities of interest for all physics analysis, the AOD format is derived. The physicist performing an analysis will further reduce the data size by defining a DPD format specific to his analysis.

2. Analysis Strategies

The data of the LHC experiment pass through an analysis chain. The data can be characterised by the format. This data is also send to the large Tier-1

- At the experiment so-called *RAW* data is recorded. It consist out of the direct measurements by the detector hardware. It has to be stored on permanent storage. Only a small fraction will be used for analysis, mainly to understand detailed characteristics of some detector component.
- The data passes a reconstruction step and so-called *event summary data (ESD)* are produced. In this stage the measurements are converted into physics measurements. While data produced in this stage contains all relevant information, the overall data size is still large and the full dataset is hard to analyse.
- By selecting quantities of particular relevance for the analysis by the physicists a more compact format is created, the so-called *Analysis Object Data (AOD)*. Data in this format will be replicated to many sites.
- As the final step in the analysis chain a final subset of analysis specific data is generated. They are sometimes referred as *Derived Physics Data (DPD)* or also N-tuples. This data are only relevant for a small set of persons and will be collected at a local site for the final interactive analysis.

In the following some relevant aspects of the analysis strategy of the different experiments will be discussed.

2.1. The CMS analysis environment

The experiment has developed a flexible approach, that allows the physicist to select their analysis strategy [7]. Their data model is based on the ROOT format. Their design choice was not to separate the persistent and the transient event model. While this reduces their flexibility for scheme evolution, they gain the possibility to analyse data in a simpler way. They forseen several analysis scenarios:

- An analysis can use the full framework and profit from all its features. This analysis can be performed locally or by sending jobs to the grid

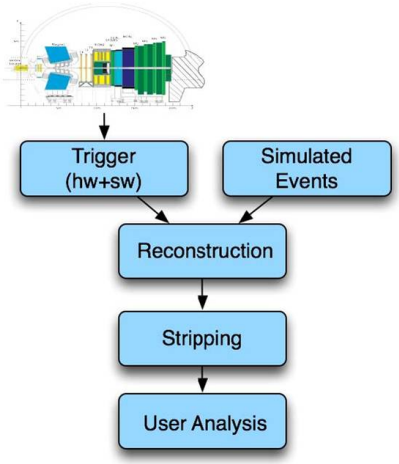

Figure 4. The LHCb analysis model. Simulated and real data follow the same analysis chain. Interesting events are collected by a central striping procedure.

- An analysis can be performed in a simplified version of the framework, called *FWLite*. This approach is well integrated in the ROOT framework. While not all features of the full framework are available, such a lighter approach can lead to a more interactive way of working
- An analysis based on *FWLite* can also be used in the context of the PROOF distributed analysis system. This system will be discussed the next section.
- Finally CMS data can be studied with "bare" ROOT giving direct access to the data members. This allows the user to study interesting quantities with a minimal overhead.

2.2. The ATLAS analysis model

The basic principle of the ATLAS analysis model is that a smaller dataset can be read faster and improve the turnaround in the analysis step [8]. Several approaches are used to achieve this goal

- During the *Skimming* procedure only events of interest for the particular analysis are kept. In function of the physics channel under study this can lead to a significant reduction of the overall data size.
- The data size of the selected events can be further reduced by *Thinning* . In this process objects within the events, that are of no interest for the analysis in question, are removed.
- An additional step is called *Slimming* . Here the size of objects of interest are further reduced by suppressing of data members not required for a specific analysis.
- A final step is called *Reduction* . A high level algorithm can be used on the data to determine a specific physics quantity. If the underlying quantities will not be used in the following, it is sufficient to keep the derived result.

For a specific analysis user will develop his own data format based on a common framework. This framework allows the user to preserve a specific object within the different format. For example an object describing a particle, as an electron, can be used in the ESD, the AOD and also the DPD format. This gives the possibility to develop programs and subroutines that can be applied in different stages of the analysis chain and encourage therefore the reuse of code. The DPD format adds the flexibity to add user derived data.

The analysis will be finalised in an interactive step, where DPD data can be studied within the ROOT framework. The small overall size of the DPD dataset improves the turnaround time.

2.3. The LHCb analysis model

For the LHCb analysis model stripping of interesting events takes a central role [9]. This reduces the size of the datasets to a managable quantity of 10^6 to 10^7 events per year. This data will be striped centrally and distributed to all Tier-1 sites. The grid is then used to send jobs to the data and perform the final analysis.

An interesting consequence of the LHCb analysis model is that user analysis is concentrated on the large Tier-1 centers. This is only possible as LHCb, a dedicated b-physics experimet, requires comparatively small computing resources. The advantage of the approach is that the larges Tier-1 center can provide their users with more reliable services.

2.4. The ALICE strategy

Alice has choosen ROOT as the basis of their analysis framework. Therefore there analysis strategy is well adjusted to that tool. Data is stored in ROOT format. For analysis a specific effort is made that algorithm can be used both interactively and in batch mode. That gives the Alice user the choice to perform analysis interctively in a local session, using more resources by using the PROOF model for disributed analysis or sending a batch job to the grid in the tranditional batch model.

Of specific interest are also their activities to established an organised analysis model using so-called analysis trains. This is a well established analysis procedutre, where the analysis programs of the physicist are run together in a an organised way. For Alice this is essential as the sheer size of the data will make it difficult for individual physicists to analyze the full dataset.

3. GRID Analysis Tools

To ease the acces to the grid, the experiments have developed grid analysis tools. They aim to shield the user from the complexity of the grid and give users a transparent access to remore resources. These tools are reviewed in turn.

3.1. CMS

The CMS experiment has developed a command line tool, the *CMS Remote Analysis Builder (CRAB)* to send jobs to the grid infrastructure [10].

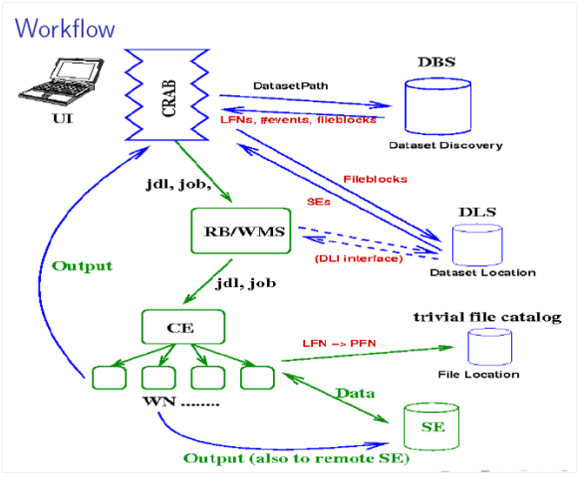

Figure 5. Workflow of the CMS Remote Analysis Builder (CRAB).

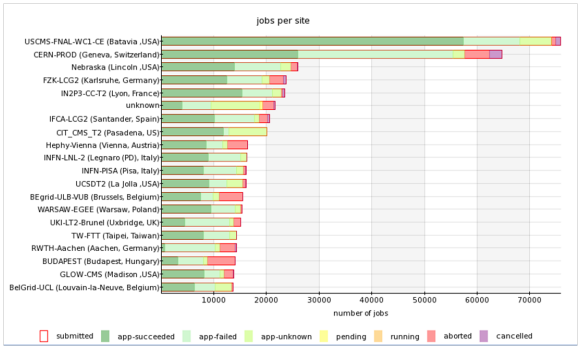

Figure 6. CRAB usage observed with the ARDA dashboard. The diagram shows the number of jobs for several sites in a one month periode of summer 2007.

The tools is simplyfing the submission by wrapping up the user environment as sandbox for an analysis job on the grid (see figure 5. The tools is also contacting the relevant databases to get information on the required datasets and their location. Using this information jobs are send to the sites and their Computing Elements (CE). The results of the analysis jobs can be stored on remote Storage Elements (SE) or can be returned to the user.

For job submisison the tools is supporting both submission using the gLite Workload Management system and direct submisison based on CondorG. For the near future it is planned to provide an additional server, that would complement the current implementation. Itis expected that a sever implementation can reduce the load on the human operator, increase the reliability of job exectution and improve the scalability of teh overall system.

The ARDA dashbord [11] is used to monitor usage of the tool. In Figure 6 the number of jobs in a one month periode of summer 2007 are shown. In that periode 645K jobs were observed, corresponding to about 20K jobs per day. On overall a job success rate of 89% has been achieved.

3.2. LHCb

The LHCb experiment is using the GANGA job management system for preparation and submission of analysis jobs. This system is being jointly developed with the ATLAS experiment and will be discussed in the next section. For job execution they rely on their own submission system.

The DIRAC system provides a layer on top of the grid middleware. The system is using the gLite Workload Management system to send so-called pilot jobs to the sites, which in turn retrieve jobs from a central taskqueue (see figure 7). This allows the Virtual Organisation (VO) not only to provide its own accounting and its own job prioritisation, but it also improves the user experience by hiding problems of the infrastructure. The late bindig of the pilot to the actual job avoids problems as unexpected downtimes or misconfigurations of Computing Elements. In addition the system can interact with the Storage Element (SE) on the site and provides a mean to prestage data from tape, if required.

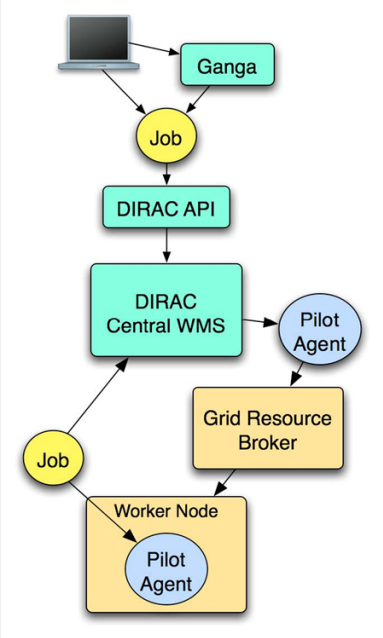

Figure 7. The workflow of the DIRAC system for Distributed MC production and analysis.

3.3. ATLAS

The ATLAS experiment uses several grid infratsructures, EGEE/LCG, OSG and Nordugrid. For user analysis on EGEE and Nordugrid the GANGA job management tool is used to prepare and submit analysis jobs. For the OSG infrastructure a combined production and analysis system PANDA provides a layer on top of middleware. While both system are being actively developed, for the future various interoperabiliy options are under investigation.

The GANGA job management system, which is jointly developed with LHCb, is used to prepare and submit jobs to submit directly to the EGEE and Nordugrid infrastructure. In case of EGEE jobs are sent using the gLite Workload management system, in case of Nordugrid ARC middleware is being used.

A GANGA job is described by an object model (see figure 9). The actual job is described by several componets:

- The *Application* object describes the actual application and its specific parameters. In case of ATLAS the main application is *Athena*, in case of LHCb *Gaudi*. There are a number of other applications for different use cases, as *AthenaMC* to support user production in ATLAS or *ROOT*, which allows to execute ROOT macros on the grid.
- The *Inputdata* object describes the input dataset for the application in question. Notions of a dataset are specific to the experiments and typically require a connection the the Bookkeeping database and the data management tools of the experiment in question.
- The *Outputdata* object describes the way the application should store results. Typically the user can request tresults to be returned directly to the user using the sandbox mechanism or to register them on storage elements (SE) on the grid.
- The *Splitter* object contains the logic to generate subjobs during submission. The plugable architecture allows for different strategies.
- The *Merger* object provides functinallity to merge results from the execution of jobs on the grid.

In total 920 persons have used tried the tool, within ATLAS more then 540 users. On average about 280 persons are using GANGA, again within ATLAS about 150.

Analysis on OSG is supported by PANDA. PANDA is a pilot based system and uses concepts similar to DIRAC (see figure 10). For analysis a command line based tool is used to wrap up the user environment and send it together with the job description to the PANDA server. A Web based interface is used to monitor job progress and get information on the job status and the output. A particular strength of the system is its support for fast submission of large number of jobs.

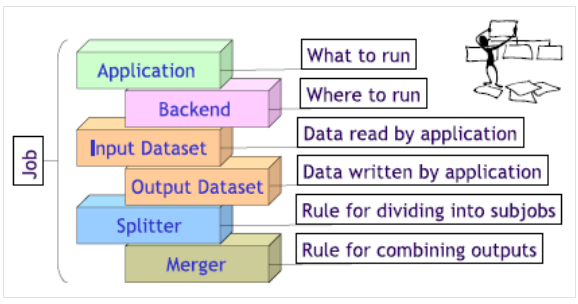

Figure 8. The components of a job object in GANGA.

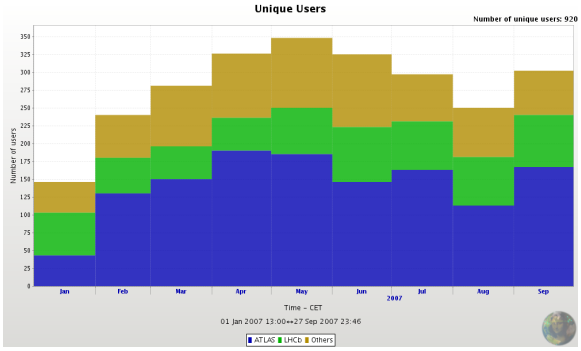

Figure 9. Usage of the GANGA job management tool for the year 2007.

For the ATLAS users interoperability is of great importance. At this point several strategies are under evaluation. On the one hand PANDA is used to send analysis jobs to several EGEE sites, on the other GANGA is used to define analysis jobs in the PANDA system. The flexibility is considered as an advantage and allows for different evolution path in the future.

3.4. Alice

As already discussed the Alice experiment is putting a strong emphasis the possibility to perform an analysis in batch and interactive mode.

For batch analysis they have updated their Alien system. It is designed as grid middleware and offers the user a central point to enter in the system. The emphasis of tool is to support the Alice experiment, but it is used also by other communities.

The system contains several components necessary to build up a grid infrastructure and to communicate with other structures. Alien 2 provides

- a filecatalog with associated metadata
- mechanisms for authorisation, authentication, job optmisation and execution and storage management
- a system for auditing, quota management and monitoring
- interfaces to other grid systems

The user interface of Alien is integrated with the Unix shell and provides a seamless integration with the user desktop. The filecatalog is presented as a virtual file system and the taskqueue as a virtual batch system. The system is in use since 2002 for production and since 2006 all Alice users have access.

The interactive system is based on PROOF. Thye system present a computer cluster to the user as an extension of the local PC. The same ROOT macro syntx can be used loacly and in the distributed environment. The system gives access to a more dynamic use of resources and provides real-time feedback to the user. Automatic splitting of the computing task and merging of the results hides all complications from the user.

A PROOF cluster has been resently studied as a prototype for the Alice version of the CERN analysis facility (CAF). The systems aims to provide support for prompt and for pilot analysis, for calibration and alignment and for fast simulation and reconstruction. A test setup has been put in place build out of 40 machines, with 2 CPUs and 250 GB disk space. The system is setup as a xrootd pool and provides for local caching of data. Local access to data has significant adavantage for the analysis of the data and can avoid bottlenecks when accessing the central storage element.

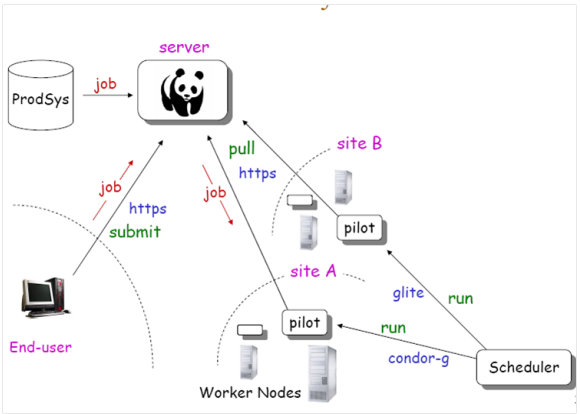

Figure 10. The workflow of the Production and Distributed Analysis system (PANDA).

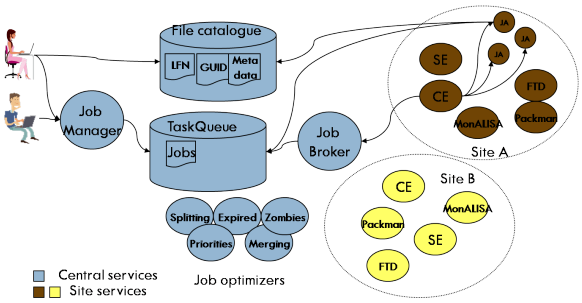

Figure 11. The workflow of the Alien2 system.

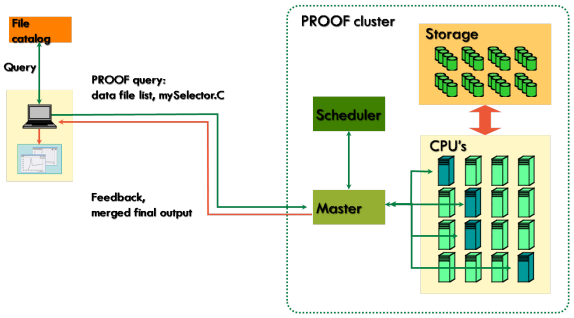

Figure 12. The workflow of the PROOF system.

4. User Experiences

We are observing thar larger number of users are profiting from the grid for their daily work. An imporatnt aspect was the development of user friendly access tools that hide the complexity from the user. Nevertheless experts are still important to solve underlying problems with the infrastructure.

While experiences are in generally good, users still have to accept some complications. Some users have reported that they feel that they are not only doing their work, but also debugging the grid.

It has been found that the sites play a very important role, not only as administartor of the middleware, but also in the support they have to provide. Their responsability extends not only to data availibility, but also to software configuration and other experiment specific issues. It has been found that site that have a high investment in that area ca provide a significantely better service to the users.

At the same time also user support has found to be very important, especially for new users. Here the first line support by the experiment plays an important role. Experts from the experiment have to understand if problems are experiment specific or can be assigned to the grid. They have then to direct the issues either to the developers in the experiment or, using systems like GGUS, to the grid support.

5. Conclusion

The LHC experiments rely on the grid to deliver the resources required both in computing and in storage capacity. The grid has been organized in a hierarchical way following a tiered model to organize the data handling. The experiments are preparing themselves to use the grid not only for simulation production, but also to analyze the real data. They have developed tools to enable the physicists to use remote resources for their work. These tools aim to simplify the access to the infrastructure and aim to address the current shortcomings.

We have seen that the use of the grid is on a steady rise during the last years. Hundreds of users are using it as a tool for their work and send an increasing number of jobs to the infrstructure. Nevertheless it has to be stated that the grid remains a complex tool. Data distribution and data storge is a central issue. A strong user support is required to address the complications.

We have also observed a rising interest in more interactive applications. This is leading to an evolation of the data formats used by the experiments.

References

[1] M. Campanella and L. Perini, *The analysis model and the optimization of geographical distribution of computing resources: a strong connection*, CERN, Monarc Note 98/1, 1998.

[2] R. Jones et al., *The ATLAS computing Model*, this proceedings.

[3] M. Lassnig et al., *Managing ATLAS data on the petabyte-scale with DQ2*, this proceedings.

[4] A. Klimentov et al., *ATKAS Distributed Datamanagement Operations. Experience and projection.*, this proceedings.

[5] R. Rocha et al., *Monitoring the ATLAS Distributed Datamanagement System* , this proceedings.

[6] L. Tuura et al, *Scaling CMS data transfer system for LHC startup*, this proceedings.

[7] C. Jones et al., *Analysis Environments for CMS* , this proceedings.

[8] A. Farbin et al., *The ATLAS Analysis Model*, this proceedings.

[9] P. Stuart et al., *Distributed Data Analysis in LHCb*, this proceedings.

[10] D. Spriga et al., *CRAB (CMS Remote Analysis Builder)*, this proceedings.

[11] J. Andeerva et al.,*Grid Monitoring from the VO/User perspective. Dashboard for the LHC experiments.*, this proceedings.

[12] J. Elmsheuser et al., *Distributed Analysis using GANGA on the EGEE/LCG infrastucture* , this proceedings.

[13] A. Maier et al., *Ganga - a job management and optimising tool*, this proceedings.

[14] J.F. Grosse, *The CERN Analysis Facility - A PROOF cluster for Day-One Physics analysis*, this proceedings.

[15] F. Rademakers, *latest Developments in the PROOF System*, this proceedings.