Experiences with SAMGrid at the GermanGrid centre GridKa for CDF

M. Feindt, Th. Müller, U. Kerzel, Th. Kuhr, G. Quast (University of Karlsruhe)
A. Baranovski, L. Carpenter, S. Veseli (Fermi National Laboratory),
R. StDenis (University of Glasgow),
St. Stonjek (University of Oxford)

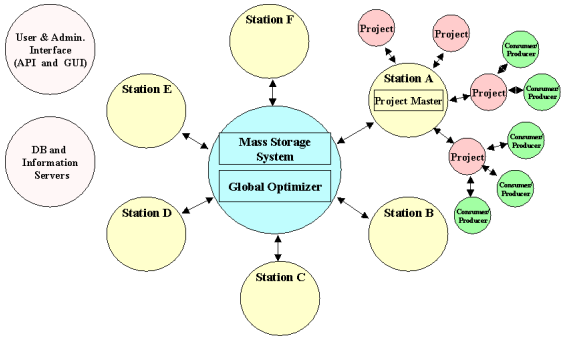

Figure 1: Illustration of the SAM architecture [6].

Abstract

The German Grid computing centre ”GridKa”[3] offers large computing and storing facilities to the Tevatron and LHC experiments, as well as BaBar and Compass. It has been the first large scale CDF cluster to adopt and use the FermiGrid software ”SAM” [1] to enable users to perform data-intensive analyses. The system has been operated on production level for about two years. We review the challenges and gains of a cluster shared by many experiments and the operation of the SAMGrid software in this context. Special focus will be given to the integration of the university based cluster (EKPPlus [5]) at the University of Karlsruhe, as well needs and use-cases of users who wish to integrate the software into their existing analyses.

THE FERMIGRID SYSTEM SAM

The competitive physics programme of the CDF collaboration requires an enormous amount of computing power. Besides the many analyses accessing the rapidly growing datasets, all raw data has to be reprocessed regularly to utilise recent developments and improvements in offline code such as tracking, etc.

The name SAM is an abbreviation for Sequential Access via Metadata which illustrates the basic principle: Data is being processed sequentially (i.e. one file after the other) and is associated with metadata. The metadata contains details about the content of the files such as file-type (e.g. processed file with the full event information, event simulation, analysis ntuples, etc.), run-numbers, trigger streams, etc. This information is stored in a central database and can hence be used to select specific sets of files for the respective analysis. Figure 1 illustrates the most important components of the SAM system. It consists of central systems which have to be deployed only once per experiment and components at each participating grid site. All data is stored on one or several mass storage systems (e.g. tape libraries). Access to tape is coordinated via a global optimiser in order to minimise tape mounts. Each participating grid site deploys a SAM station which manages locally available cache areas, imports requested files and serves user analysis jobs via project masters. One project master is started automatically per user analysis job. It coordinates which files are needed to match the user’s request and delivers them to the analysis program (“consumer”) running on the worker-nodes. All information such as metadata and user activity is logged into a central database. Command line and multiple interfaces to e.g. Python, ROOT and C++ exist to allow easy and flexible interaction with the SAM system and access to information stored in the database. It should be noted that SAM offers great flexibility in its configuration, e.g. the various sub-processes communicate via CORBA [2] which allows to run the services on distributed computers if necessary. Furthermore, SAM can be installed completely without `root` access to the involved computers and hence runs in user-space only. Aiming at minimal requirements, one dedicated computer per participating grid site is sufficient to serve a large computing facility and offer the full functionality to the user. Besides making the installation procedure rather easy and straight-forward, this has the additional advantage that the installation and maintenance of the SAM software is largely independent of the details concerning the setup of the computing facility.

THE GERMANGRID TIER1 CENTRE GRIDKA

GridKA is the German Grid computing centre located at the research-centre “Forschungszentrum Karlsruhe”. Eight high-energy physics experiments (Alice, Atlas, BaBar, CDF, CMS, Compass, DØ, LHCb) share a large cluster. In the respective terminology, GridKA is a Tier1-centre for the LHC experiments (Alice, Atlas, CMS, Compass, LHCb) and TierA for BaBar. For both CDF and DØ GridKA is used to process a significant amount of user jobs and to evaluate and test new Grid and data-handling technologies. Each experiment has access to a dedicated access computer which is used to both provide the experiment specific software and to allow users to submit jobs to the cluster and retrieve the output.

The worker-nodes are shared between all experiments and hence no experiment specific software is installed

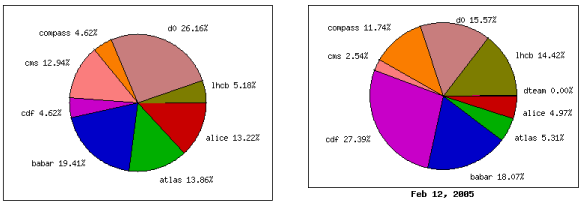

Figure 2: GridKa fair-share: nominal (left) and actual use at LHC idle time (right)

there. In total, approximately 310 TB of disk-cache, 500TB of tape storage and 1700 kSI2000 of computing power are available for all experiments. This large capacity in increased at least once per year until the LHC start up. Following this approach has the advantage to both benefit from falling prices for hardware and develop new techniques to build a large cluster in case currently available technologies do not scale to a computing cluster of this size. Sharing the resources does not only make the administration of the system considerably easier than maintaining eight dedicated computing clusters, but also results in other benefits. For example, LHC experiments require a large amount of computing power during the so-called data-challenges which test operational issues of the LHC grid and simulates data-taking, reprocessing, etc. However, once these data-challenges are completed the need for computing power drops significantly as the results need to be analysed. These otherwise idle computers can than be used e.g. by the CDF group as illustrated by figure 2.

GRIDKA FOR CDF

GridKA has been the first off-site computing centre where members of the CDF collaboration were able to analyse a significant amount of data, hence GridKA is considered as a prototype of a remote analysis farm. Many of the developments which lead to a successful operation of SAM at GridKA are of vital importance for both the operation of similar facilities in e.g. Italy, Asia and the USA, as well as for the further development of the on-line processing farms and user analysis facilities on-site Fermilab.

About 40 TB of disk cache and approximately 60TB of tape storage are currently available for the CDF group. The nominal share of CDF's access to the computing power is $\approx$ 5%. In addition, few TB of disk space are available for users to store executables to be submitted to the cluster, as well as intermediate files or output from analysis programs. The availability of this large capacity allows to store all datasets relevant for B and top physics analyses locally at GridKA and build up a replica of the files stored at FNAL. This is done by transferring all datasets needed for the various analyses also to the tape library. In addition, multiple wide-area file transfers are minimised and files can be automatically retrieved from tape in case they are replaced on the disk-cache SAM's dynamic cache management.

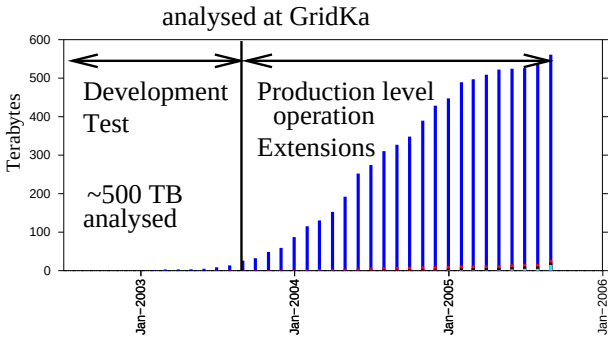

Figure 3: Amount of data analysed by CDF at GridKA

The German CDF group uses the large computing facilities at GridKA via SAM extensively. Figure 3 illustrates the success of the model described in this contribution. The figure shows the integrated amount of data analysed at GridKA using SAM in TB starting from the beginning of the project in the second half of 2002 to the end of 2005. Since the main development of the system and tests have been completed end of 2003, the system is used at production level in a wide range of B- and top-physics analyses. By the end of 2005, approximately 500 TB have been analysed at GridKA this way. About 5 TB per day have been processed at peak performance. This corresponds to roughly 20% of the data analysed by the whole CDF collaboration at the main computing centre on-site FNAL per day.

The tape storage is fully integrated into SAM via dCache [4] and customised access procedures. Thus the tape-access is fully transparent to the user who then neither notices that access to the tape library is required nor has to know any of the involved configuration details. Additionally, offering a permanent storage location on tape has the advantage that users can store new files into the SAM system. Example applications are the storage of newly created simulated events, skimmed files containing selected events with the full information or storing specialised analysis ntuples. This approach offers several advantages: All files are associated with metadata which describes the file content, i.e. all relevant information directly available and does not have to be summarised on e.g. individual user's private web-pages. This allows an automated bookkeeping procedure as all information is stored in a database. Furthermore, the metadata can be used to define a dataset of interest, e.g. asking for all files where a specific physics process has been simulated with a given generator, etc. Retrieving the requested files is then handled automatically by SAM without the need for further user intervention. In addition, all files stored this way are immediately available to the whole collaboration and do not have to be copied manually to various analysis centres. This has the additional benefit of removing the necessity to maintain dedicated file-servers which are used by various analysis groups not yet working in a grid-enabled way. Operated in this way, downtimes of

a computing facilities can be significantly reduced. For example, in case of hardware problems with the file-servers acting as disk-cache, they can simply be replaced by other disk-areas. After telling SAM to flush all files on the server in question, normal user operation can be resumed quickly as the requested files are automatically retrieved again.

Despite the fast up-link of the GridKA Tier1 centre, typical user analyses running more than 200 programs in parallel analysing the requested files will in general process the files faster than they can be retrieved from remote locations. Operational experience shows that this situation can be efficiently avoided by using SAM's ability to "pin" datasets on disk: Large datasets are imported before users start to analyse them. The imported files are transferred both to disk-cache and tape storage. By locking the files on disk-cache by a SAM administrator, they are never removed from the file-servers and hence always available for immediate and fast access by the user. Using different groups for scheduled file-import and user analysis, users are additionally able to import any files they need for e.g. small test, etc. Exploiting the feature of the SAM software to handle different groups with different quota of assigned disk space, accidental removal of data required by a large fraction of the user community by individual users can be easily avoided.

The close proximity of the GridKA Tier1 centre to the EKPPlus cluster at the University of Karlsruhe provides an ideal test-bed for future Tier1 to Tier2/3 connections in the LHC computing model. The EKPPlus cluster serves approximately 20 worker-nodes shared between users from the CDF, CMS and AMS experiment. A dedicated SAM station with approximately 1 TB of dynamic disk-cache is deployed at EKPPlus. Exploiting the fast connection to GridKA via a dedicated high-speed fibre connection, the SAM station at the EKPPlus cluster is configured to retrieve new files from the station at GridKA only: If a user requests a file not yet cached at the EKPPlus cluster, it is transferred from the GridKA station. In case a user requests a file which is not yet available at the Tier1 site, the SAM station at the University cluster uses the advanced routing features of SAM: It instructs the station at GridKA to import the required files to the large cache areas at GridKA first and then transferring it to the EKPPlus cluster in a second step. Following this approach has several advantages: Data is imported using the dedicated wide-area up-links of the Tier1 centre rather than the common University based Internet connection. Furthermore, due to the large storage capabilities at GridKA, large amounts of data can be provided to the user, thus minimising the need for multiple file transfers. This mode of operation is fully transparent to the user who only interacts with the SAM station at EKPPlus and does not have to know any detail about this specialised configuration. It should be noted that Karlsruhe is currently the only site where this setup is currently deployed for the CDF experiment. Being operational at production level for more than two years proves the feasibility of the intended tiered operation mode of the LHC computing model.

CHALLENGES IN A RUNNING EXPERIMENT

Building a working computing Grid in a running experiment leads to unique challenges: While the LHC experiments still have a few years of preparations ahead, Run II of the Tevatron experiments has already started in 2002. Since then both CDF and DØ have recorded a significant amount of data – and more data is taken each day. It is therefore imperative that the use of new technologies must not interrupt neither data-taking nor physics analyses. The software SAM is jointly being developed by CDF and DØ. A practical approach is taken in its design and development: Starting from working technologies, these are replaced by more "grid-like" tools as the software evolves. For example, earlier versions of the SAM software used the FTP implementation developed by the BaBar experiment [7] which was replaced by the GridFTP implementation from the Globus toolkit [8] once it reached production level stability. This Ansatz has the advantage that emerging technologies can be immediately tested in "real life" situations, minimising the risk of wrong developments and missing essential features. As users utilising the system for their own analyses are directly involved in this process, weakness in design and implementation of the system are quickly observed by a large group of non-expert users. Furthermore, all resources are available to the users for their analysis which also means that a reliably working system has to be maintained at all times.

User interaction

A key ingredient of a working grid system is its user interface: Generally speaking, users are neither interested in getting involved in future development of the grid nor in being involved in testing its functionality or features. This point is of particular importance for running experiments, where users expect the system to "just work". Furthermore, the grid-based infrastructure has to be integrated in already developed analysis code which uses conventional means such as static file lists, etc. to access data. It is therefore imperative to provide a wide range of flexible user-interfaces for the grid-based system. SAM supports a broad range of applications making it accessible from all types of applications encountered in typical physics analyses. In detail, interfaces exist to:

- CDF analysis framework `AC++` [9]
- C++ API
- Python API
- command line interface
- ROOT [11] class TSam [10]

This allows each user to benefit from the advantages of the powerful SAMGrid system at all stages of the analysis, e.g. when accessing the full event information via the `AC++`

framework, storing newly created analysis ntuples into the system and analysing them later within the user-specific analysis programs and ROOT macros. Since SAM takes automatically care of all aspects of file-handling, the user does not have to worry about insufficient disk-space, failing file-servers, etc. Using the provided metadata functionality does not only allow a flexible and automated bookkeeping but also offers versatile means of selecting (sub-) sets of the data.

CONCLUSION

The FermiGrid software "SAM" has been successfully deployed at the GermanGrid Tier1 centre "GridKA". The system has been successfully used by members of the CDF collaboration at production level for more than two years at the time of writing, analysing more than 500 TB of data until the end of 2005. GridKA has been the first site outside Fermilab to adopt the SAM system for CDF and offer a significant amount of storage and computing capacity to its users and is hence considered a prototype of a remote analysis farm. Many of the features developed in this context are of vital importance for both the deployment of similar clusters elsewhere and remodelling of the central computing farms on-site Fermilab. A distinct feature of the SAM deployment at GridKA is its fully integrated tape storage. This allows both to build a replica of files used in B and top physics analyses and offer users a permanent storage location for simulated events and analysis ntuples. Exploiting the fast connection of the GridKA Tier1 centre to the EKPPlus cluster of the University of Karlsruhe provides an ideal test-bed environment for Tier1 - Tier2/3 interaction within the context of the LHC computing model. Using the advanced routing features of SAM, it has been demonstrated that this approach works reliably at production level without requiring that the user is aware of this setup: SAM provides a uniform interface to the user and hence hides all details of the complexity of such a setup from the user.

ACKNOWLEDGEMENTS

This work was supported by the German Federal Ministry for Science and Education (BMBF).

REFERENCES

[1] *The SAMGrid project* `http://projects.fnal.gov/samgrid`

[2] *The common object request broker: Architecture and specification* `www.corba.org`

[3] `www.gridka.de`

[4] *dCache* `www.dcache.org`

[5] *P. Schemitz et al.: "An Offsite Local Analysis Facility for CDF II"* `CDF Note 6385`

[6] *The SAM architecture* `http://cdfdb.fnal.gov/sam/doc/architecture/sam_architecture.html`

[7] *The* bbftp *file transfer protocol* `http://doc.in2p3.fr/bbftp`

[8] *The Globus Alliance* `http://www.globus.org`

[9] *E. Sexton-Kennedy: A user's guide to the AC++ framework* `CDF Note 4178, 16th May 1997`

[10] *Th. Kuhr: TSam: Access to SAM files in ROOT* `http://www-ekp.physik.uni-karlsruhe.de/~tkuhr/TSam/`

[11] *R. Brun: ROOT: An object oriented data analysis framework* Nucl. Instrum. Meth, A389, 1997