Requirements of the German CMS groups concerning a National Analysis Facility ("NAF")

The German CMS Groups*

June 28, 2007

Abstract

The German groups participating in the CMS experiment at the future Large Hadron Collider (LHC) at CERN are in need of a national analysis facility for successful and internationally competitive analysis of the wealth of data expected from the LHC. This facility will complement the resources available at the German Tier-1 Centre GridKa and from the federated Tier-2 operated by DESY and RWTH Aachen by providing an efficient infrastructure for end-user data analysis. The foreseen structure will largely enhance the capability of German groups for collaborative analysis efforts. The detailed requirements and suggestions for their implementation are outlined in this document.

1 Introduction

According to the computing model of the CMS experiment, the data flow from the source, the CMS detector, to the desktops of the analysing physicists proceeds via several hierarchical Tiers. Tier-0 to 2 are well defined in the CMS Computing Technical Design Report [1], while the Tier-3, consisting of local computer cluster resources at universities and institutes, will only opportunistically be used for collaboration-wide tasks and still is to be shaped to optimise the analysis environment for the end-user.

In particular, such university and institute clusters and computing resources are important for interactive analysis and serve for development and testing of simulation, reconstruction, and analysis code. However, to allow shared analyses between the German groups[1] and to deal with the large data sets for physics analyses, a "National Analysis Facility" is planned at DESY. It is understood that members of all German CMS groups have equal and transparent access to the proposed facility.

In addition to the integration into the Grid, an effective analysis facility must provide extended functionality not foreseen by the standard Grid middleware and act as a link between the Grid-accessible Tier-2 resources and the user's desktop.

The additional requested functionality includes:

- computing, storage and analysis resources for n-tuples or equivalent data sets for end-user analysis selected from data sets stored on the Grid;
- support for both group-based and private usage of resources;

*This document is a result of the recent "CMS Analysis Infrastructure" workshop (DESY, June 4th 2007) and was approved by the CMS-FSP Meeting on June 26th.

[1]RWTH Aachen, DESY, Universität Hamburg, Universität Karlsruhe(TH)

- fast response for end-user analysis;
- user home directories;
- provision of a test-environment for code development and debugging.

In order to make most use of the resources in Germany, all computing resources provided by the German groups should be considered a potential addition to such an analysis facility. This approach will stimulate and considerably extend the ability of the groups to perform collaborative analyses. To this end, collaborative tools are still to be developed to facilitate user and data management across institute boundaries. This concept of a German "Virtual IT Centre (VITC)" is also subject to substantial funding by the Helmholtz Alliance. The VITC will provide the framework to bring together and bundle the resources from the participating institutes. For the rest of this document, we will use the name "National Analysis Facility (NAF)" for the prototype installation at DESY, which forms the starting point for the development of the VITC in such a way that additional sites can be embedded into the concept easily.

2 Analysis Steps and Required Components

As a realistic starting point, a prototype installation is proposed at DESY Hamburg with close connection to the CMS Tier-2 installation. This facility will form the nucleus for the envisaged distributed analysis environment and provide valuable operational experience as well as trigger the development of collaborative tools supporting analysis by working groups and individual users.

A typical analysis consists of several steps as outlined in Figure 1. The data source for analysis will be the reconstructed data stored at Tier-1 centres or the "Analysis Object Data (AOD)" stored at Tier-1 and Tier-2 centres. In a typical analysis, Grid tools will be used to submit a large number of jobs and collect the output, either in form of special AODs or in a n-tuple format, at the NAF. Periodically, these data sets will be replaced by improved ones considering the latest available calibration, improvements to the reconstruction code, or the selection procedure itself. Group data, *e.g.* common n-tuples or AOD, are stored on the NAF storage, either in the Grid storage area or on the workgroup storage.

These data usually form the basic data set for one or more physics working groups, and specialised n-tuples serving the need of individual analyses are then deduced from it. These are private data sets, which will undergo very frequent read access in the course of the analysis procedure. For end-user analysis, fast turn-around, i.e. fast response from job submission to the retrieval of the results or even interactive operation mode of the NAF is mandatory. The workgroup storage will allow to run parallel, I/O-intensive, or "burst", analysis as well as interactive data analysis.

Since most of the code development for the Grid-based selection jobs will also be performed on the NAF, it is essential to provide an interactive environment for code development, debugging, and testing on small data samples. Output of small amounts of data, *e.g.* histograms and text files, is directed to the home directory space, which is made visible to the users' desktops in the home institutes.

3 Requested Resources and Services

To support the above activities, the Tier-2 infrastructure at DESY should be extended by additional hardware and services. Production of private or group-specific Monte Carlo data sets, CPU power for Fast Monte Carlo generation or Toy Monte Carlo studies can easily be provided by extending the Grid computing cluster at the Tier-2 with worker nodes that are primarily dedicated to German CMS

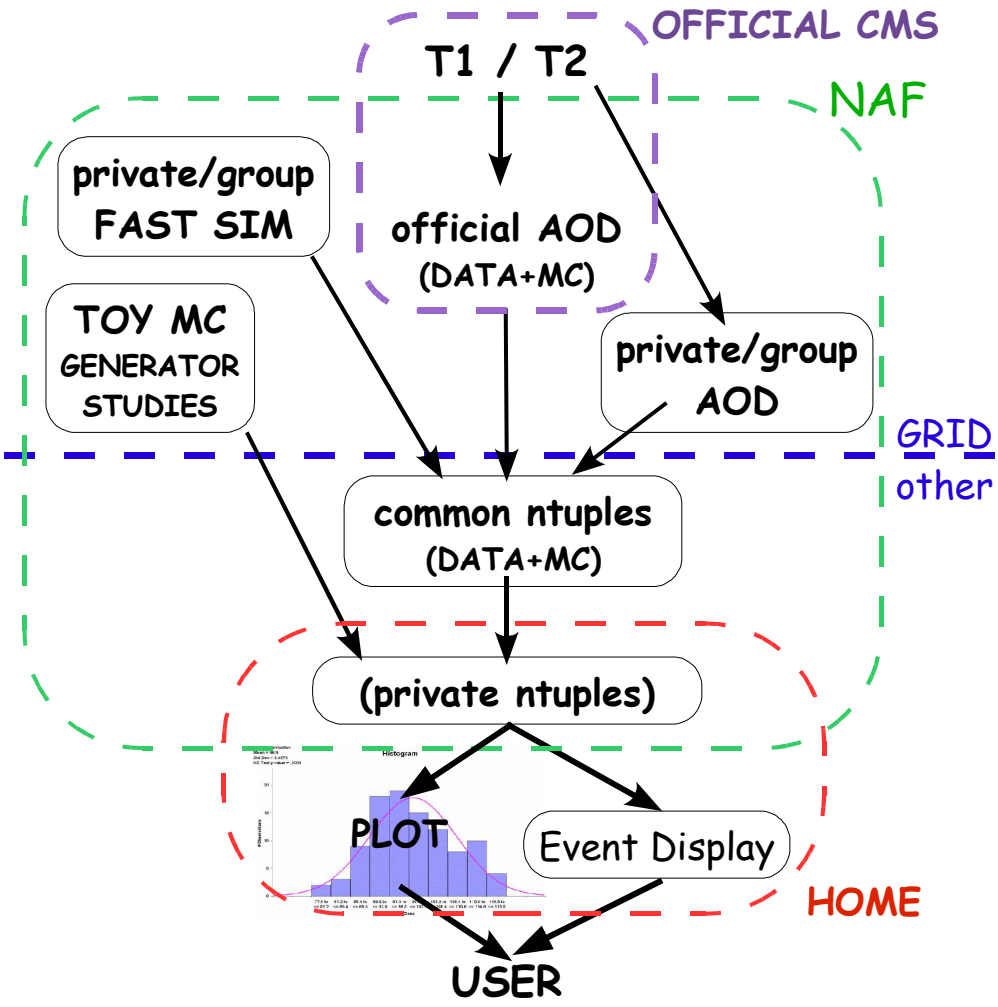

Figure 1: Diagrammatic steps of a user analysis. In the proposed NAF structure, direct access from the German Tier-2/Tier-3 to workgroup servers and Grid storage complement the standard Tier-1/Tier-2 Grid setup.

users. A special tag for German users, a so-called *VOMS* , has already been defined within CMS, thus allowing local users to be distinguished from general Grid users.

The space of the CMS data storage at the Tier-2 should be enlarged by additional storage resources. This will allow to import the data sets the German CMS groups are interested in by using the CMS data transfer infrastructure. Analysis jobs running on a NAF node should have access to the data via protocols supported by the CMS software. The direct access to a large number of CMS data sets will accelerate physics analyses and can only be achieved at the Tier-2.

In addition to the Grid resources, dedicated compute nodes, named “workgroup servers” in the following, are needed, which provide access to workgroup storage and compute power for n-tuples and other data under group and private ownership. These workgroup servers should provide local login for each user as well as access to home directories based on *AFS* to be usable across institute boundaries. This requires a central registry for all German CMS users at the NAF. The workgroup servers should have access to the Grid storage space of the Tier-2, thus allowing data import and export via Grid mechanisms. A Grid user interface and the CMS software should also be provided.

The computing power should be made available by a batch system and a special system for parallel interactive analysis, *e.g.* a “Parallel ROOT Facility (PROOF)”. The workgroup storage will house

user data, like n-tuples or dedicated data sets used for calibration and detector commissioning, under group and private ownership. The connection to this storage must allow for the highest possible bandwidth, since typical end-user analyses tend to be I/O-limited. Grid access to this storage would allow to write such data files directly from the data processing job run on the Grid to the workgroup storage. *POSIX* compliant access protocols would guarantee compatibility with most tools and programs used in high energy physics.

Figure 1 indicates the proposed locations of the NAF components, either as an extension of Tier-2 Grid resources and services, or dedicated hardware and services and the "home space" connected to the users desktop. Access to the extended Tier-2 resources is exclusively via the Grid, while the workgroup servers are accessible in a direct way. Authentication for user login may be generated automatically from Grid certificates, or the gsissh protocol may be used directly for users‘ login. Certainly, development of suitable tools is still needed in this area.

Storage space for local users should be provided both within the Tier-2 framework as Grid-enabled storage space accessible via *SRM*, and as private space on the workgroup storage. We propose to let the working groups manage this disk space under the responsibility of the group leader. Backup services should be made available for identifiable parts of the storage space, *e.g.* the home directories, software areas and dedicated parts of the data storage.

The long-term planning of NAF resources depends on further progress with grid tools, and on operational experience gained with the prototype installation. The required size of the National Analysis Facility suitable to support the German CMS user community is comparable to an average CMS Tier-2. However, the NAF should offer relatively more disk space to house complete analysis data sets.

It is considered important to start the prototpe installation of each proposed component as soon as possible. A reasonably estimated prototype of a NAF to be installed in the year 2007 should consist of 80 TB Grid storage, 15 TB workgroup storage and 30 TB tape. As the Grid computing cluster exists already at DESY, the additional NAF computing resources within the Grid can be small for the first year. Thus, we propose to start with computing power equivalent to 30 kSI2k for the Grid part and 60 kSI2K for the workgroup part which should allow batch and interactive analysis with I/O-intensive jobs. While all CMS members at DESY and Universität Hamburg already have AFS home directories at DESY, additional $\sim$100 active CMS users from Aachen and Karlsruhe can be expected.

Although some aspects of the NAF setup are independent of the needs of a particular experiment, support for CMS specific software and services, e.g. for data management and ROOT-based analysis within the CMS framework, is an indispensable requirement and can best be guaranteed by the local proximity of an active CMS group with a substantial number of CMS software and computing experts. Furthermore, coordination and optimisation of the resource usage and the development of a standardised analysis framework can be done more easily in collaboration with a strong local CMS group.

References

[1] CMS Collaboration, Computing Project Technical design report, CERN/LHCC 2005-023;

[2] LHC computing Grid Technical design report, CERN/LHCC 2005-024;

[3] WLCG Memorandum of understanding, actual version on the internet, see http://lcg.web.cern.ch/LCG/C-RRB/MoU/WLCGMoU.pdf