The Swiss ATLAS Grid End 2008 Progress Report for the SwiNG EB

2009-02-06

E. Cogneras[a], S. Gadomski[b], S. Haug[a], Peter Kunszt[c]
Sergio Maffioletti[c], Riccardo Murri[c]

[a]Center for Research and Education in Fundamental Physics,
Laboratory of High Energy Physics
Sidlerstrasse 5, CH-3012 Bern, Switzerland

[b]Department of Nuclear and Particle Physics
24, Quai Ernest-Ansermet, CH-1211 Geneva, Switzerland

[c]Swiss National Super Computing Center (CSCS)
Galleria 2 - Via Cantonale, CH-6928 Manno, Switzerland

ABSTRACT

Since 2005 the Swiss ATLAS Grid (SAG) is in production. The SAG working group is based on a charter which was accepted by the SwiNG Executive Board in 2007. In summary it says that the SAG working group supports, monitors and manages the usage of the SAG. In 2009 the Swiss ATLAS Grid consists of four clusters with about 2000 shared computing cores and about 250 TB of disk space. It is based on middlewares provided by the NorduGrid Collaboration and the EGEE project. It supports multiple virtual organisations and uses additional middleware, developed by the ATLAS collaboration, for data management. The Swiss ATLAS grid is interconnected with both NorduGrid and the Worldwide LHC Grid. This infrastructure primarly serves Swiss research institutions working within the ATLAS experiment at LHC, but is open for about two thousand users on lower priority. The last three years about 80 000 wall clock time days have been processed by ATLAS jobs on the Swiss ATLAS Grid. In 2008 almost 50 thousand wall time days were processed. This is more than twice as much as in 2007.

1 Introduction

The Swiss ATLAS Grid (SAG) is a computing infrastructure serving Swiss ATLAS physicists. ATLAS is one of four large particle physics experiments at the Large Hadron Collider (LHC) in Geneva (CERN) [1][3]. The data from its detector is expected to answer fundamental questions about the universe, e.g. about the origin of mass and about the physical laws right after *Big Bang*. The LHC will deliver its first collisions this year (2009). ATLAS will record, replicate, simulate and analyze the data from these collisions. Several tens of petabytes per year will be produced in this process. A large effort has been invested into the world-wide distributed computing system, here called the ATLAS grid, with the SAG being the Swiss part of this system [2]. The SAG is realized as a collaboration between the Universities of Bern and Geneva, the Swiss National Super Computing Center (CSCS) and the Swiss Institute for Particle Physics (CHIPP).

This second progress report from the Swiss ATLAS Grid (SAG) as a working group of the Swiss National Grid Association (SwiNG) is an update of the first report on the activity in the first half of 2008. First we report on personnel changes and activity. Then we recapture the infrastructure and the usage. Finally we comment on the goals for 2008 and present those for the first half of 2009.

2 The Swiss ATLAS Grid Working Group 2008

The SAG working group is based on a charter which was accepted by the SwiNG Executive Board in 2007 [4]. In summary it says that the SAG working group supports, monitors and manages the usage of the SAG. Since the foundation it consists of Sigve Haug, who lead the group in 2008, Cyril Topfel and Szymon Gadomski. Further are Peter Kunszt and Sergio Maffioletti associated members from the Swiss National Super Computing Center (CSCS). For 2009 the group has reconstituted itself with S. Gadomski as lead. C. Topfel stepped down due to commitments at CERN and was replaced by Dr. Eric Cogneras from Bern (see Figure 1). In 2009 a CSCS replacement for P. Kunszt and S. Maffioletti is also expected.

Figure 1: The Swiss ATLAS Grid Working Group. From the left S. Haug (Bern, 2008 lead), E. Cogneras (Bern) who replaces C. Topfel, S. Gadomski (Geneva, 2009 lead), P. Kunszt and S. Maffioletti (CSCS).

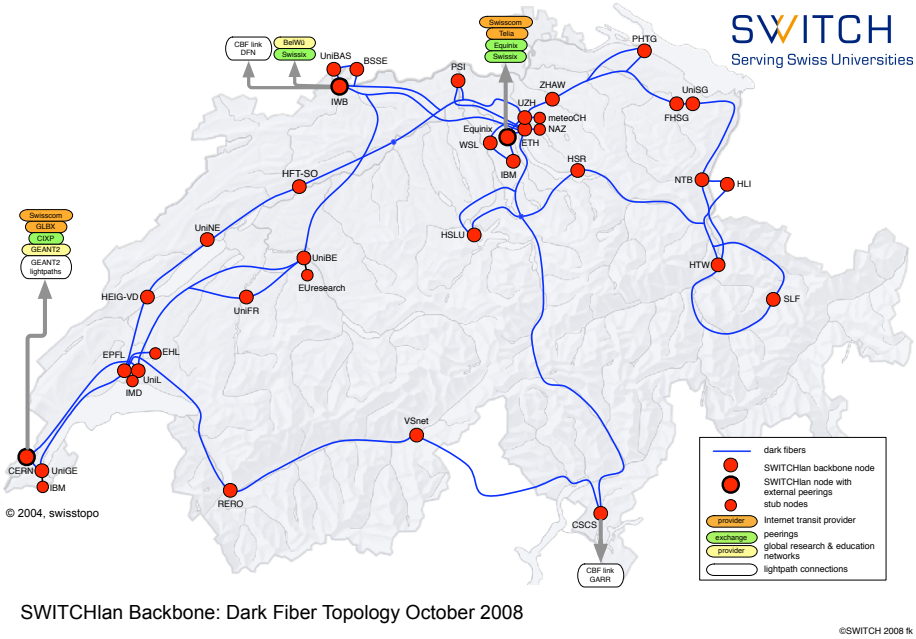

Figure 2: The SWITCHlan backbone. The SAG T2 is located at CSCS, while the T3s at UniBE and UniGE respectively.

The group communicates via wiki pages which are hosted by CSCS [4], a mailing list, which is provided by SwiNG (atlas-wg@swing-grid.ch) and managed by the group's lead, and monthly meetings which are normally virtual EVO meetings [5].

The group has five main outbound communication lines. It communicates with the Swiss T2 system administrators at CSCS via lcg@cscs.ch, with the German T1 via atlas-germany-computing@desy.de, with the SwiNG Executive Board via its mailing list, with the Swiss Institute for Particle Physics (CHIPP) and the T0 at CERN via personal contacts. For 2008 these lines have been sufficent.

In the second half of 2008 the human activity of the group has been the following. There were two virtual meetings whose minutes are collected on the group's wiki pages. Via S. Haug the group's engagement in the SWITCH/AAA project "Swiss Multi Science Computing Grid" was continued.

3 Infrastructure

The ATLAS grid is described in the ATLAS "Computing Technical Design Report" [2]. It is a hierarchical grid with a so called *four-tier* structure. Data from the tier zero (T0) at CERN in Geneva is pushed through dedicated network

hardware to about ten computing centers, so called "tier ones" (T1) around the world. They again duplicate and push data to their associated "tier twos" (T2), normally national or regional computing centers. The "tier twos" serve their "tier threes" (T3) which typically are clusters owned by single universities or research groups. The final tier four is the desk- or laptop of the ATLAS physicist. The SAG has one tier two at the *S*wiss Super Computing Center (CSCS) in Manno which is connected to the T1 at the *Forschungszentrum Karlsruhe*, Germany. In Switzerland two T3, in Bern and Geneva respectively, are being served by CSCS. This hierarchical structure is enforced in order to avoid the break down which a totally flat and chaotic structure can cause on the services.[1]

The SAG sites are connected by the SWITCHlan dark fibre network, i.e. the bandwith can be adjusted by illuminating the optical fibres with multiple frequencies [7]. This shared network is currently operated with one 10 Gb/s channel, but more bandwidth is possible. The network map is shown in Fig. 2. The foreseen output from the ATLAS detector is about 2.4 Gb/s, thus the Swiss capacity of the network meets the estimated ATLAS requirements for connectivity. The international BelWu 1 Gb/s connectivity to the T1 in Karlsruhe may have to be increased, in particular because this connection is also not dedicated to ATLAS. However, the redundant topology shown in Fig. 2 does ensure a stable connectivity at the low level.

A speciality is a direct and dedicated network link between the T0 at CERN and the Geneva T3. As only a small fraction of the data can be processed in Geneva, this option is not of interest for final physics analysis of the data, which will need to start with large datasets at T1 sites. However, the direct line will enable the users of the Geneva T3 to participate more effectively in the commissioning of the ATLAS experiment. During regular data taking the direct line can be used for data quality monitoring, which can be done by processing of the order of 1% of the data.

Concerning the computation and storage resources in the SAG, about 2000 worker node cores and 250 TB disk space are comprised by four clusters. The cluster hardware is summarized in Tab. 1. In 2004 the sites *gridified* with some 32 bit one core desktop boxes and then evolved to the current, in the Swiss context, considerable resources made up by 64 bit four and eight core servers. Both Intel and AMD processors are represented. The storage systems are all disk based. All clusters use Gigabit ethernet for interconnections, and at least 2 GB RAM is available per core. The size of the resources will be growing with the needs in ATLAS, along a timeline exceeding a decade.[2]

The choice of operating system and middleware has been a compromise between maintenance minimization and feasibility. The ATLAS software has a size

[1] Admittedly this hierarchical structure is not fully respected. A considerable amount of horizontal and vertical data pulling between tiers is ongoing and crucial for the functioning of the system.

[2] This long timeline gives rise to many challenges, i.e. the transition from the 32 bit to the 64 bit infrastructure is not trivial. The applications still have to run in a 32 bit compatibility setup. Another example is the bankrupt of hardware suppliers and the related loss of warranties.

Table 1: The Swiss ATLAS Grid sites in 2008. The second column shows the worker node cores, the third the disk storage in terabyte (TB), the fourth the operating system (OS), the fifth the middleware (MW), the sixth the local resource management systems (LRMS), and the seventh the storage elements (SE). The ^{s} indicates that the resource is shared and not use by ATLAS only. The numbers typically undergo a 10% fluctuation following the actual status of the resources.

Cluster	CPU Cores	DISK TB
Bern T3 LHEP	30	30
Bern T3 UBELIX	1000^s	0
CSCS T2	1000^s	150
Geneva T3	188	70

of several hundred GB of which a so called "distribution kit" is extracted, validated and pulled to the tiers. Kits are typically released several times a month and occupy about 10 GB each. They are developed, compiled and validated on Scientific Linux Cern (SLC) [8]. Experience has shown that the deployment of this software on other operating systems may imply a significant additional workload. Thus, the SLC has become the preferred operating system on the SAG clusters. However, on a shared cluster like Bern T3 B, which is running Gentoo, it is not possible for one project to determine the operating system. On such clusters the additional workload is accepted. Similar is the situation for the choice of middleware. EGEE's gLite practically does not support other platforms than SLC or related operating systems [9]. Further, gLite has historically been quite worker node intrusive and manpower demanding. These are the reasons for the extended use of the NorduGrid Collaboration's *Advanced Resource Connector* and for the rapid start up in 2005 as a light weight grid [10]. The only cluster also deploying gLite is the T2 at CSCS which is serving additional LHC experiments.

The distributed data management (DDM) system on the ATLAS grid requires some specific storage element features [11]. The system is a set of central databases at the T0 which organize files on the grid into data containers. Containers have locations which are file catalogs, gLite's LFC, deployed at the T1. These again contain the physical file names of the files stored on storage elements in their respective T2. The actual movements are issued by Python services at the T1 which in turn issue gLite's File Transfer Service (FTS) service with SRM endpoints. Virtual splitting of the storage into so called *space tokens* is extensively used. Until now this has effectively excluded the ARC SE as an option and the most used solutions are dCache and EGGE's DPM [12] [13]. Since the ATLAS DDM system normally does not contain T3 sites, only the Swiss T2 is connected. However, a DDM based storage element is now being installed in Geneva.

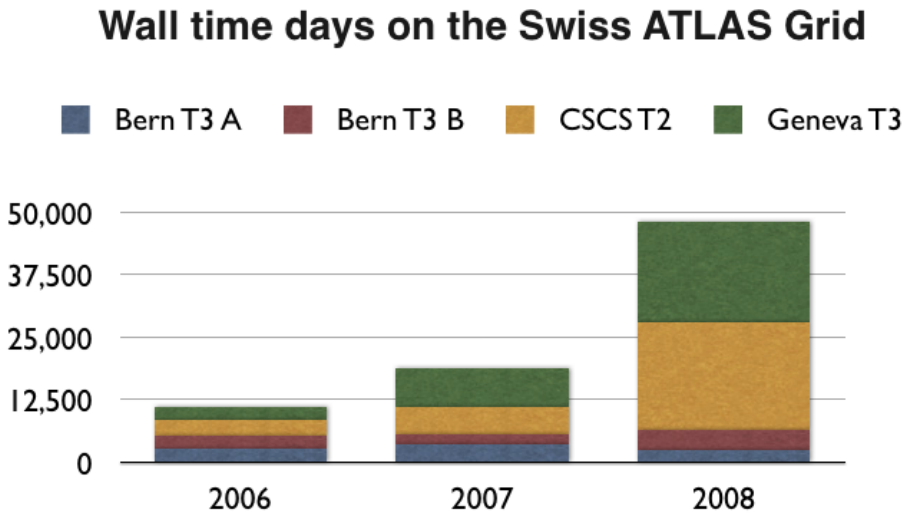

Figure 3: Wall clock time days on the Swiss ATLAS Grid. The total wall clock time in 2008 corresponds to 1% of the world wide ATLAS wall clock time as accounted by the ATLAS dashboard.

4 Monitoring and Accounting

The SAG monitoring is locally done with custom made scripts and Ganglia [14]. On the grid level the ARC monitor and the ATLAS dashboard are used. The ARC and gLite information systems, on which these web applications are based, provide sufficient data for identifying problems on an hourly basis. Concerning computational issues these solutions are sufficient. Monitoring of the data transfers still requires a lot of attention due to the immature state of the storage elements and the middleware services.

The usage of the Swiss ATLAS Grid has increased repidly in 2008 due to an increase in our resources. The usage is measured in three ways. For the T3 clusters the Torque and SGE accounting files are analyzed. For the T2 cluster the EGEE accounting portal and the ATLAS dashboard are consulted [15] [16]. Both sources rely on EGEE's APEL database. In 2007 the T2 wall clock time from these sources were cross checked several times with the analysis of the local accounting. The numbers were consistent within 5%, which is the estimated uncertainty for the SAG accounting.

Figure 3 shows the ATLAS usage of the clusters in wall clock time days. Compared to 2006 the usage increased by a factor of five to approximately 48000 wall clock time days in year 2008, i.e. 130 wall clock time years. This corresponds to more than 1% of all the accounted ATLAS computing [16].[3] Of

[3]Neither on the Swiss ATLAS Grid nor on the world wide ATLAS grid is all usage accounted, e.g. interactive work and usage by other projects are not contained in the numbers.

the 220 sites which contributed to the 11 000 wall clock time years on the ATLAS Grid in 2008, only 10 sites (the number of T1) contributed with more than 2% and no site contributed with more than 10%. Considering the increase in the available worker nodes on the SAG the usage in 2009 is expected to double. However, it is difficult to predict the usage for 2009 since the experiment is expected to start collecting collider data for the first time. On one hand it is likely that this will increase the usage further. On the other many more unexperienced users will submit jobs which eventually will cause new unforeseen challenges for the sites.[4]

On the SAG there is no automated disk space accounting. The storage is mostly inspected in an *ad hoc* and in a manual manner. An integrated storage element solution with detailed information of all disk operations down to the level of distinguished names is very much desired. However, such solutions are not yet provided by the middleware.

5 Evaluation of the 2008 Objectives

The 2008 goals, which were staked out in the charter of the SAG group, are repeated and commented below.

- *"Monitor, account and manage the ATLAS data transfer in Switzerland, i.e. the T1-T2 and the T2-T3s transfers. Ensure dataset completeness, sufficient rates, LFC-dCache consistency. It includes work with ATLAS DQ2 system, FTS, LFC, SRM, GridFTP and dCache. Communication with T1, T2 and among T3s is required."* Some of the mentioned tasks are being taken care of by the T1 operators. The SAG working group did not achieve the aimed transfer overview in 2008. The reason was low priority due to no real need of meeting this aim.

- *"Monitor and account the ATLAS jobs on the Swiss ATLAS Grid. It includes work with the batch systems, ARC front-end, LCG/gLite CE and deployment of ATLAS releases. Communication as for goal 1."* As was seen from Figure 3 this goal was met.

- *Support local users, i.e. Geneva and Bern physicists, in issues related to 1 and 2.* This goal has been met by personal contact between the users in Geneva and S. Gadomski and between the users in Bern and S. Haug. Further wiki pages have been created for the SAG at CSCS and for the Geneva T3 at CERN [4][19]. The SAG also has some old wikipages at CERN [20]. The goal was met.

- *Enable 5% of the Tier 3 CPU resources to at least one "SwiNG application". Access will be given via the Advanced Resource Connector of the NorduGrid collaboration.* This goal has been met in the sense that the

[4] Already now jobs with enormous inputs and outputs, i.e. several tens of GBs, output files with 40 GB size etc have been observed. Such usage may rapidly bring down the services.

RSA768 application has been given access to the Geneva and one of the Bern clusters.

- *Apart from uncoordinated group communitcation, the group meets virtually monthly, preferably the same day as the ATLAS FZK T1 meeting.* This goal has been met.
- *The group will reconsider its existence in the latest 2008 report to the SwiNG executive board.* This goal has been met (see next section).

6 Objectives for Q1 and Q2 2009

- Continue production at the CSCS T2 via LCG and ARC. S. Haug is the experimental contact. E. Cogneras takes the responsibility for the ATLAS releases for the ARC part.
- Continue production at both clusters in Bern. S. Haug is responsible for bringing the production back to UBELIX. E. Cogneras then takes the production responsibility for both clusters.
- Continue both user analysis activity and production via the existing NorduGrid interface at the T3 in Geneva.
- Setup a Storage Element in Geneva and have it integrated with the ATLAS Distributed Data Management sysem. Get data transfers from CERN to the new SE. Try some use cases for Trigger Data Quality work.
- Try user data transfers and user job submissions to CSCS from Berne and from Geneva.
- Continue with monthly EVO meetings organized and chaired by S. Gadomski.
- Follows the T1 activity (S. Haug).
- Organize an in-person meeting at the CSCS in Manno.
- Present and publish proceedings at GPC 2009 (S. Haug) and CHEP 2009 (S. Gadomski).

7 Summary

The Swiss ATLAS working group manages a Grid system in which about 2000 shared cores are available. During 2008 the Swiss ATLAS Grid has provided almost 50 thousand wall time days to the computing of the ATLAS experiment. Compared to 2007 this number has more than doubled.

The activity has been reported in four talks at the CHIPP Winter School, the annual meeting of the Swiss Physics Society, at the SwiNG Scientic Council

meeting, at University of Bern. Further the group has engaged in other SwiNG projects.

For the next six months the group foresees two publications and increased processing of ATLAS production jobs. The lead of the group went from S. Haug to S. Gadomski.

References

[1] The ATLAS Collaboration, G Aad , et al.: The ATLAS Experiment at the CERN Large Hadron Collider, 2008 JINST 3 S08003

[2] ATLAS Collaboration, *Computing Technical Design Report*, CERN-LHCC-2005-022, ATLAS-TDR-017.

[3] S. Gadomski *et al.*, *The Swiss ATLAS Computing Prototype*, ATL-SOFT-PUB-2005-03, CERN-ATL-COM-SOFT-2005-07.

[4] https://twiki.cscs.ch/twiki/bin/view/LCGTier2/SwissATLASGridWorkingGroup.

[5] Enabling Virtual Organizations, http://evo.caltech.edu/evoGate/about.jsp.

[6] ARC meets SwiNG, www.lhep.unibe.ch/gridbern08.

[7] SWITCH: `http://www.switch.ch`

[8] Scientific Linux homepage: `https://www.scientificlinux.org`

[9] Generic Installation and Configuration Guide for gLite 3.1: `https://twiki.cern.ch/twiki/bin/view/LCG/GenericInstallGuide310`

[10] M. Ellert, et al.: Advanced Resource Connector middleware for lightweight computational Grids, Future Generation Computer Systems 23 (2007) 219-240

[11] S.Haug, et al.: Data Management for the Worlds Largest Machine, PARA 2006, LNCS 4699, pp. 480-488, Springer-Verlag Berlin Heidelberg 2007

[12] dCache homepage: `http://www.dcache.org`

[13] Disk Pool Manager: `https://twiki.cern.ch/twiki/bin/view/LCG/DpmGeneralDescription`

[14] B. N. Chun, et al.: The Ganglia Distributed Monitoring System: Design, Implementation, and Experience, Parallel Computing, Vol. 30, Issue 7, July 2004

[15] EGEE Accounting Portal: `http://www3.egee.cesga.es`

[16] ATLAS Collaboration's Dashboard: `http://dashboard.cern.ch/atlas`

[17] The Swiss National Grid Association: `http://www.swing-grid.ch`

[18] The ATLAS Dashboard, http://dashboard.cern.ch/atlas/.

[19] https://twiki.cern.ch/twiki/bin/view/Atlas/GenevaATLASClusterDescription

[20] https://twiki.cern.ch/twiki/bin/view/Atlas/SwissAtlasComputing