PROTOTYPE OF THE SWISS ATLAS COMPUTING INFRASTRUCTURE

A. Clark, M. Diaz Gomez, F. Orellana, I. Riu Dachs, DPNC, University of Geneva
I. Aracena, H.P. Beck, S. Gadomski[*], B.K. Gjelsten, C. Haeberli[†], S. Kabana[‡],
V. Perez-Reale[§], E. Thomas[¶], C. Topfel, LHEP, University of Bern
A. Reufer, Informatikdienste, University of Bern
G. L. Volpato, CSCS, Manno[‖]

Abstract

The Swiss ATLAS Computing prototype consists of clusters of PCs located at the universities of Bern and Geneva (Tier 3) and at the Swiss National Supercomputing Centre (CSCS) in Manno (Tier 2). In terms of software, the prototype includes ATLAS off-line releases as well as middleware for running the ATLAS off-line in a distributed way. Both batch and interactive use cases are supported. The batch use case is covered by a country wide batch system, while the interactive use case is covered by a parallel execution system running on single clusters. The prototype serves the dual purpose of providing resources to the ATLAS production system and providing Swiss researchers with resources for individual studies of both simulated data and data from the ATLAS test beam. The solutions used for achieving this are presented in the article. Initial experience with the system is also described.

MOTIVATION

The project described in this article is a part of the preparations for the analysis of ATLAS data [1] by physicists at the universities of Bern and Geneva. Large-scale data analysis is expected to start in 2007. A first prototype of the computing infrastructure has already been set up. It has provided input to resource estimates of the future Swiss ATLAS computing. The prototype has also enabled us to gain experience in the following areas:

- hardware of PC clusters,
- management of Linux farms, choice of Linux distributions,
- ATLAS off-line software and the ATHENA framework,
- distribution kit of ATLAS off-line software and issues related to working outside CERN,
- installation and use of Grid middleware,
- interactive analysis of ATLAS data,
- software tools for distributed interactive analysis.

The project was started two years ago. The group of people responsible for the computing infrastructure consists of four people, with some rotation in the course of the project. As we are all involved in other activities, such as ATLAS construction, teaching at the universities and other academic duties, the work on computing cannot become an important research activity in its own right. We have therefore put emphasis on finding practical and currently possible solutions to problems at hand.

The software involved must provide the needed functionality (or at least a well-defined subset of it), be documented, easy to install and to maintain. Development of Grid middleware and other “infrastructure” software is not in the scope of the project. Some development or customisation is in practice necessary, but must be very limited, always motivated by direct needs of the prototype.

HARDWARE INFRASTRUCTURE AND NETWORK

The prototype of Swiss ATLAS computing is using several clusters of PCs located at the University of Bern, the University of Geneva and the CSCS computer centre near Lugano. The clusters are:

The Bern ATLAS cluster consists of seven PCs owned by the Laboratory of High Energy Physics, University of Bern. The PCs have 16 CPUs in total and 13 TB of common storage space. The operating system is SUSE Linux. This small cluster is meant as a playground for testing of software. It contains several different architectures of PCs, including 32 bit and 64 CPUs.

The Bern Ubelix cluster is owned by the Informatikdienste of the University of Bern. The cluster serves several different research activities. The operating system is Gentoo Linux -this choice cannot be influenced by preferences of High Energy Physics. In exchange Ubelix offers significant resources, there are 288 CPUs at the moment.

The Geneva DPNC ATLAS cluster consists of one server and 5 worker nodes with 9 CPUs and a 300 GB RAID array for data storage at the Département de physique nucléaire et corpusculaire (DPNC) of the

[*] also with INP Cracow, Poland
[†] now with Inalp Networks AG, Bern
[‡] now with Univerity of Nantes, France
[§] now with CERN, Geneva
[¶] now with CERN, Geneva
[‖] now with RRZN, University of Hannover

University of Geneva. Three of the worker nodes serve simultaneously as shared desktop work stations. The hardware is highly heterogeneous, but all run CERN Scientific Linux 3, have a minimum of 1 GHz clock frequency, 1 GB of RAM and are interconnected with fast ethernet.

The Geneva DINF/DPNC cluster is a homogeneous system owned by the DPNC and the Division Informatique (DINF) of the University of Geneva, physically located at the DINF. It consists of 12 dual 2.4 MHz worker nodes with 4 GB of RAM each, a NorduGrid front-end machine and a files server with a 9.6 TB RAID array attached. All hosts run CERN Scientific Linux 3 and are interconnected with Gigabit-ethernet.

The CSCS Phoenix cluster is owned by CHIPP (Swiss Institute of Particle Physics) and is accessible for all Swiss researchers participating in the ATLAS, CMS and LHCb experiments. It is physically installed and managed at the CSCS computer centre in Manno. Phoenix is a homogeneous system consisting of 1 NorduGrid front-end machine, 1 machine serving as LCG front-end and local batch system master, 1 machine serving as file serve with a 9.6 TB RAID array attached and 10 dual 3.0 GHz worker nodes with 4 GB or RAM each. All hosts run CERN Scientific Linux 3 and are interconnected with Gigabit-ethernet.

The resources of the Swiss ATLAS Computing Prototype and CERN are connected via the SWITCH network [2], which provides a dark fibre infrastructure between CERN, the Swiss universities and the CSCS. The dark fibres are owned and exclusively used by SWITCH. Currently the dark fibre links are equipped with a bandwidth between 4 GBit/s and 10 GBit/s, depending on the importance of the connection.

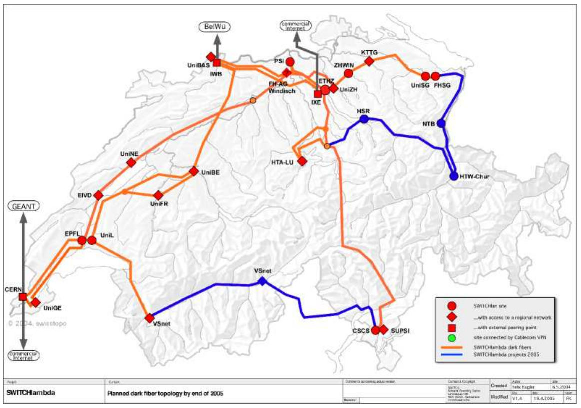

Figure 1: Current dark fibre connections between Swiss research institutions provided by SWITCH.

The advantage of the dark fibre infrastructure is that it allows SWITCH to ramp up the bandwidth continuously according to the growing needs by installing optical end devices using different colours on the same fibre. Maximally, SWITCH could provide a bandwidth of 16 × 10 GBit/s per dark fibre link. Technically, even 32 × 10 GBit/s would be feasible, but this would require a major investment.

The obtainable bandwidth is thus about two orders of magnitude above the foreseen output bandwidth of the ATLAS TDAQ system of 300 MB/s [3].

COUNTRY-WIDE BATCH SYSTEM BASED ON NORDUGRID ARC

By much the same reasoning as that underlying the original visions of Grid computing [4] it was judged that a sensible first goal was to establish a Swiss-wide batch system: a system for sharing computing resources, accessing data and submitting computing jobs, involving all Swiss institutes with ATLAS computing needs. Researchers should be able to join so-called virtual organisations and thereby gain authorisation to submit jobs in a transparent way to all Swiss clusters using the same simple user interface.

Given the non-interoperability of the available Grid middleware systems, this implied that all Swiss clusters had to be integrated into one Grid system. In practice this meant LCG [5] or NorduGrid [6]. After an informal evaluation, drawing on the experience of the people involved, NorduGrid was chosen. The reasons for this choice were:

- the DPNC test cluster and the Bern ATLAS cluster are too small to be integrated into LCG. An LCG cluster needed in the order of four PCs to run services. This would have been 50% of the Bern ATLAS cluster and 100% of the DPNC prototype cluster.

- UBELIX is a resource shared within the University of Bern. But LCG requires a dedicated cluster, because it imposes cluster management and strict operating system requirements. The Laboratory for High Energy Physics cannot impose LCG cluster management and a CERN Linux distribution on a common university resource.

- NorduGrid ARC is far simpler to install and configure than LCG. It delivers the functionality we need to build a country-wide batch system.

- Users should be able to submit jobs from their own desktop or laptop. It did not appear feasible demanding of normal users to install an LCG user interface.

Currently the country wide batch system unifies about 370 CPUs into one big batch cluster. The five clusters (see section) making up this Swiss ATLAS Grid not only have different hardware configurations, they also have rather different usage patterns and have different levels of maintenance and support.

The Phoenix cluster runs both LCG and NorduGrid, supporting two interfaces for job submission.

USE OF THE BATCH SYSTEM

Reconstruction of simulated $Z\to e^+e^-$ data.

Physicists at the University of Geneva analysed $Z\to e^+e^-$ events [7], simulated during the so-called 'Data Challenge 2' and 'Rome' productions of ATLAS. 100'000 events spread over 2000 files, taking up 200 GB of disk space were located, replicated to national resources and reconstructed on clusters in Bern, Manno and Geneva using release 10.0.1 of the ATLAS software kit plus an extra shared library. The actual dispatching of the jobs and the subsequent bookkeeping of which jobs had finished, which had failed, resubmission and retrieving of results was taken care of by a customised script we wrote for that purpose. The jobs were submitted both via our NorduGrid system and on the worldwide LCG infrastructure.

Production of simulated SUSY data.

At the University of Bern, physicists were interested in producing their own simulated data on Super-symmetry events by running the full chain of simulation, digitisation and reconstruction, using releases of the ATLAS software plus their own compiled code. The productions amounted to 100 SUSY event generation jobs (5500 events per job) and 2000 SUSY simulation, digitisation and reconstruction jobs (50 events per job). The latter step required substantial CPU power 1 event needed about 650 s on 2.8 GHz CPU.

Production of simulated test beam data.

Part of the simulation activity of the ATLAS test beam which was previously run at the CERN batch facility (lxbatch) was migrated onto our system [8]. This production was done using a GUI for preparing, submitting and monitoring jobs and retrieving and cataloguing their output. Each submission took ∼1.4 seconds.

Reconstruction of real test beam data

This task was carried out by physicists of the universities of Geneva and Bern. Analysis of beam test data introduced new requirements for the Swiss ATLAS computing system and provided additional user feedback. Around 100000 beam test events were reconstructed in about 25 jobs.

DISTRIBUTED-INTERACTIVE ANALYSIS

It is expected that the volumes of data with which physicists will be working regularly during ATLAS operation will be much larger than the test-beam or simulation data currently used. At the same time there is an advantage in working with the data in an interactive way. Obtaining a result rapidly speeds up development, enabling physicists to correct their algorithms and to refine their selection criteria in a shorter time. Efficiency of working with the data is an important factor in producing good quality physics results.

These premises motivated us to look for ways of working with large data samples in an interactive way. Use of a local cluster of PCs, instead of a single PC, seems to be a natural step towards making interactive analysis more powerful. As is well known, ROOT software [9] already offers a way to work with n-tuples in a parallel way, using PROOF. However, for data formats specific to ATLAS, another solution is necessary. We have investigated ways of working interactively with the ATLAS software framework ATHENA [3].

In order to simplify the problem, we are making the following assumptions:

- All the software that a user wants to execute is already installed on a shared file system of the cluster. The software can be partly in an ATLAS release (installed in one place for all users) and partly in a directory owned by the user, but both are visible to all the nodes of the cluster.
- All the data that the user wants to process is also already available to all the nodes of the cluster.

With the above assumptions the task of the software infrastructure becomes relatively simple. It needs to be able to:

- split the input data,
- distribute the processing,
- collect and merge the results.

A software tool that was able to provide the needed functionality, in a way that could easily be combined with the ATHENA framework, was DIANE [10].

The installation of DIANE on the Bern cluster was done in March 2005 and turned out to be relatively straightforward. DIANE is a set of Python scripts and can easily be customised. Once the software was installed and customised, a user needed to provide the following files in order to run ATHENA with DIANE:

1. A script to run ATHENA including environment setup. The environment variables define also the user's code to be run in ATHENA. A "job options" file, which describes the configuration of ATHENA, is also defined in the script.
2. A "job description" file that defines some parameters of the process, such as the name of the output file that should be fetched from each node after the processing is finished.
3. A list of input data files.
4. A list of cluster nodes to be used.

Using the mentioned files, the DIANE software can distribute processing to the nodes of the cluster. Splitting of input data is done on a per-file level. The result files are collected and copied to one directory in the user's home directory. Files containing histograms in ROOT format are merged together and the histograms are added. The user finds the merged result file as well as the individual histogram files produced by each process.

The processes run interactively (under the user's account) and the processing is started within a few seconds on all the worker nodes. Up to eight processors were used in parallel. An analysis of simulated ATLAS data is in our experience usually CPU limited. In the limit of large data samples, where the initialisation time of the job is short compared to the total CPU time needed, one would expect the time to process a given data sample to be inversely proportional to the number of processors that available. In our tests, the data samples were relatively small. Also the data splitting was not perfect (i.e. some processors had less data to process then others), but a gain in processing time of up to a factor of four could nevertheless be achieved.

In summary, the DIANE tool offered an interesting possibility for distributing the data processing in an interactive session, using the ATLAS off-line software framework.

PLANS

The Swiss ATLAS computing prototype offers functional infrastructure to Swiss physicists working on ATLAS simulation and beam test data analysis. In order to face the challenges of real data analysis, the system will need to grow in size and will need to be more systematically supported.

More functionality is also needed. In particular a practical way to process large data samples needs to be implemented. This requires working on the infrastructure related to data sets, in particular on the database infrastructure (file and dataset catalogues). Integration with the Distributed Data Management (DDM) system of ATLAS will also be crucial, as it will allow Swiss physicists access to ATLAS data samples processed in other countries. An automation of job submission (dataset splitting, monitoring of jobs, resubmission of failed jobs) will also be needed. We therefore intend to continue installing and testing available software tools. We may also opt for limited development efforts, shared between Bern and Geneva, to provide practical solutions to our needs.

ACKNOWLEDGEMENTS

We thank the support teams of our computing centres for keeping the clusters running smoothly, SWITCH for hosting our country-level GIIS server and Jakub Moscicki for DIANE support.

REFERENCES

[1] The ATLAS experiment at CERN,
http://cern.ch/atlas/

[2] SWITCH - Swiss Academic and Research Network,
http://www.switch.ch/

[3] G. Duckeck et al. [ATLAS Collaboration] (2005), "ATLAS computing: Technical design report",
CERN-LHCC-2005-022,
http://weblib.cern.ch/abstract?CERN-LHCC-2005-022

[4] I. Foster and C. Kesselman: *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann; 1st edition (November 1998),
ISBN: 1558604758

[5] The LHC Grid Computing Project,
http://cern.ch/lcg/

[6] NorduGrid Collaboration,
http://www.nordugrid.org/

[7] Manuel Diaz-Gomez, Doctoral Thesis (2006),
DPNC, University of Geneva

[8] Mireia Dosil and Frederik Orellana, "Massive data processing for the ATLAS Combined Test Beam", Proceedings of CHEP06, Computing in High Energy and Nuclear Physics, 13-17 February 2006, T.I.F.R. Mumbai, India,
http://www.tifr.res.in/chep06/

[9] ROOT, An Object-Oriented Data Analysis Framework,
http://root.cern.ch/

[10] The DIANE (Distributed Analysis) Project,
http://cern.ch/diane/