On the Potential Use of Remote Computing Farms in the ATLAS TDAQ System

Chris Bee[1], Tomasz Bold[2], Bryan Caron[3], Robert W. Dobinson[†4], Gareth Fairey[5], Jorgen Beck Hansen[6], John Renner Hansen[5], Richard Hughes-Jones[5], Krzysztof Korcyl[2], Brian Martin[4], Catalin Meirosu[4,7] , Roger Moore[3], Jakob Langgard Nielsen[5], James Pinfold[3], Richard Soluk[3], Tadeusz Szymocha[2], Anders Waananen[5], Sarah Wheeler[8]

Abstract— **The ATLAS experiment at CERN will require a large amount of computing resources for the online analysis system. The software and communication protocols in the ATLAS Online analysis system are optimized for a cluster environment. We setup a geographically distributed testbed to evaluate the implications of integrating remote computing resources in this environment. This paper reports on the integration scenarios and analyzes the achieved performance. We highlight limitations in the communication protocols and suggest solutions for solving them. A proposal for employing Grid-enabled resources to allow for on-demand expansion of the computing capabilities is presented at the end of the paper.**

Index Terms—**data acquisition, real time systems, wide area networks**

I. Introduction

In 2007, CERN is scheduled to start operating the largest particle accelerator built to date, the Large Hadron Collider (LHC). ATLAS is one of the five experiments being implemented at the LHC. Within the ATLAS experiment, a worldwide collaboration of over 185 universities and research institutes will analyse the results of particle collisions at the centre of a large detector. The projected collision rate is 40 MHz. Each of the collisions generates data, referred as an "event" further in this paper, in the order of 1.5 MB. The amount of data generated by the detector requires the implementation of a multi-stage filtering system [1] to reduce the rate to a throughput that can be sustained by the storage system. The first two filtering levels base their decision on partial amounts of the event, while the last level analyses the event as a whole.

A first level of filtering [4] is implemented in hardware and reduces the event rate from 40 MHz to 75 kHz. An additional selection level (known as the High-Level Trigger - HLT) has to be employed before sending the events to permanent storage at a rate of O(300 MB/s). The HLT is partitioned in two systems: the Second Level Trigger (referred below as "Level 2") and the Event Filter (EF). The HLT is built using clusters of computers interconnected through high-performance Ethernet networks. The Level 2 was designed to perform event selection at a rate of 75 kHz, reducing the input to the EF down to 3.5 kHz of events. Only selected parts of an event, flagged by the first level of filtering as a Region of Interest, are made available to the Level 2 processors. The current architecture of the EF calls for running physics analysis algorithms on a massive computer farm of about 3200 processors, located at CERN. The selected data is sent to storage and later distributed to the collaborating institutes for off-line analysis [2].

Today's estimates for the EF computing power do not include the requirements of the online detector calibration and data monitoring tasks. Many institutes, members of the collaboration, have a strong interest in the analysis and performance monitoring of the detector components which they aided in constructing. They could easily leverage for this task computers already installed at their premises. We develop an argument for the use of remote resources to augment the computing power available at the experimental site, as envisaged by [3]. During experiments taking place simultaneously in several locations worldwide, we demonstrated that a standard version of the ATLAS Dataflow software may be operated at remote locations, under the control of an ATLAS Online Software console located at CERN. We present the measured event transfer rate using the real application, with no physics processing of the transferred dummy data. The limitations introduced by the use of software optimised for a LAN environment in a WAN scenario are presented, analysed and potential solutions are

[1] Centre de Physique des Particules de Marseille, IN2P3-CNRS-Université d'Aix-Marseille 2, France
[2] IFJ-PAN, Krakow, Poland
[3] University of Alberta, Edmonton, Canada
[4] CERN, the European Organization for Nuclear Research, Geneva, Switzerland
[5] University of Manchester, Manchester, UK
[6] Niels Bohr Institute, Copenhagen, Denmark
[7] "Politehnica" University, Bucharest, Romania
[8] University of California at Irvine, California, USA

Manuscript submitted June 9, 2005. This work was supported in part by the European Union under the IST-2001-33185 grant and also by the Polish Funding Agency under the grants 620/E-77/SPUB-M/CERN/P-03/DZ 110/2003-2005 and 620/E-77/SPB/5_PR UE/DZ 465/2002-2004.

Corresponding author: Catalin Meirosu, CERN, PH Department, Bdlg. 513-R-017 G03610, Geneva, Switzerland (phone: +41-22-7673826; fax: +41-22-7673900; e-mail: catalin.meirosu@cern.ch).

suggested. Some considerations are made about challenges and additional studies that have to be made before taking a decision on whether to integrate remote computing facilities in the production Online system.

II. The ATLAS Event Filter system

This section will provide a brief introduction to the ATLAS EF system. A detailed view of the components, both software and hardware, is presented in [1].

Fig. 1 presents the hardware architecture of the EF. The functionality of the components is considered with respect to the standard event analysis data flow – the considerations for calibration and monitoring will be introduced later on.

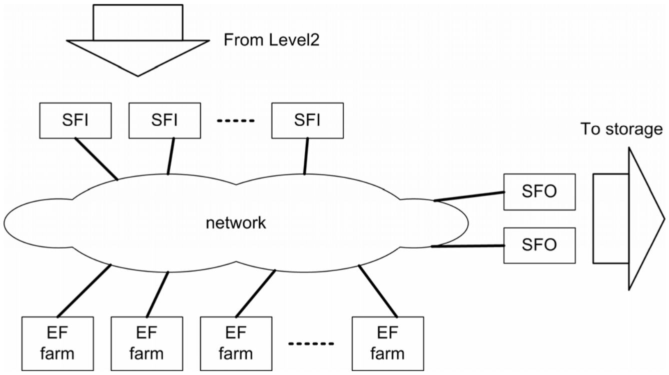

Fig. 1. - Generic architecture of the ATLAS EF system

The Sub Farm Input (SFI) computers receive the data selected by the Level 2 and act as a data source for the EF. The Sub Farm Outputs (SFOs) collect the events selected by the EF and act as a gateway to the permanent storage system. The Event Filter computer farms perform the physics analysis on the event data, deciding whether a particular event is deemed interesting from the physics point of view. The software packages running on the SFI, SFO and EF are part of the ATLAS DataFlow distribution [1], a software collection that implements the entire event transfer protocols and processing throughout the HLT.

The event transfer protocol between an SFI to a computer in a EF farm is based on a request-response mechanism, whereby the EF computer sends a request for data and the SFI responds by sending the entire event (about 1.5 MB of data). The transfer between the EF and the SFO is started by a request for temporary storage space sent by the EF computer, followed by the transfer of the accepted event when the positive response arrives from the SFO. Due to the very nature of the request-response mechanism, the throughput per host is dependent on the round trip time over the network. However, for the final system this would not count as a limitation due to the fact that an EF host would only need to process a number of events (per second) of the order of the number of installed processors, a transfer rate easily achievable in a cluster environment The data transfers are performed using a protocol based on TCP/IP.

The operation of the DataFlow is controlled by the ATLAS Online software [1]. The Online software provides a graphical user interface front-end to a human operator that can configure, control and monitor the data taking activities. The inter-computer communication infrastructure of the Online software is based on CORBA [5].

In addition to the main event filtering flow, two other important traffic classes will be passing through the Event Filter network: the online detector calibration and monitoring data. However, only rough estimates are available with respect to the computing and network bandwidth requirements for these tasks [6]. The components of the ATLAS detector are developed by a worldwide collaboration of universities and research institutes, therefore it is highly probable that a given sub-system expert is located at his university rather than at CERN. In this context, several institutes expressed their interest in having calibration and monitoring data flowing directly to and being processed by the institute's own computing infrastructure.

The use of remote processing power for real-time event handling in the ATLAS experiment was already suggested in [3]. We took a practical view of this approach and proceeded to determine how the DataFlow and Online software, designed and optimised for a cluster environment, would handle a system distributed over LAN and WAN.

The request-response nature of the traffic and the location of the sources and consumers in the online processing system mark a clear demarcation line between conventional high-energy physics data transfers, of GB size, for offline Grid processing and the experiments described in this paper. The traffic pattern generated by our application is similar to the one resulting from remote access to image databases or to iSCSI transactions over the Internet.

III. The Testbed

In the summer of 2004, the ATLAS experiment proceeded to testing components of the detector using real particle beams at an experimental facility located at CERN [7]. The latest (at the time) versions of the Dataflow and Online software were deployed in the testbed. The testbed was isolated from the standard CERN campus network infrastructure through a gateway computer. Only local IP addresses were allowed within the experimental site. Other restrictions were due to the use of a local shared file system (exported via NFS) to store the home directories of the users and the software distribution.

Developers of the detector components under test were using the site computing and networking infrastructure at the same time as we were trying to integrate a remote computing infrastructure. Therefore, the preferred integration model called for zero disturbances of the local site activities, which only made our tests even more relevant with respect to the potential usage in the final system. This approach, however, had an impact on the overall architecture of the remote computing system, as explained later.

With support from the respective campus networks and national research network operators in Europe and Canada, we assembled a testbed connecting CERN to four collaborating institutes: the University of Alberta in Edmonton, Canada, the University of Manchester, in the United Kingdom, the Niels Bohr Institute in Copenhagen, Denmark and the IFJ PAN Krakow, Poland (Fig. 2). The functional diagram is over imposed on the connectivity diagram. The labels attached to the network connections denote the operator that provided the connection.

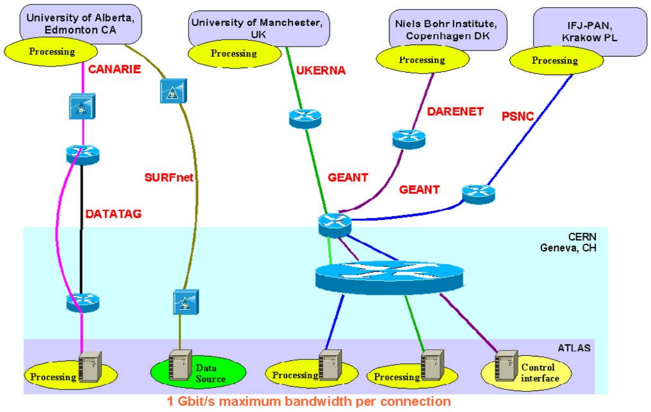

Fig. 2. - The configuration of the remote farms testbed

As mentioned above, the computers located at CERN were configured with local IP addresses that were non-routable over the Internet. We tried to assure a different type of network technology for the connection to each of the four sites. We succeeded in provisioning direct Ethernet connectivity to Edmonton (via an Ethernet over SONET circuit) and Krakow (via an Ethernet over MPLS tunnel). Therefore these two sites used non-routable IP addresses in the same subnet as the computers located at CERN.

It would be however unlikely that many sites could afford such circuits for production traffic. The sites of Manchester and Copenhagen were thus reached via standard IP connectivity, using Internet-visible IP addresses in order to explore the most likely scenario for production. Special servers that provided Network Address Translation (NAT) had to be installed at CERN in order to allow transparent access to these remote resources.

The NAT was implemented on dual-processor Xeon servers running the DevilLinux operating system [8], a free Linux distribution targeted at firewall and NAT functionality. The NAT solution was preferred to implementing a Virtual Private Network that included tunnelling via the Secure Socket Layer (SSL) technology due to the high bandwidth usage expected (in the hundreds of Mbit/s for the performance and protocol tests), that would have required much more computing power than a simple NAT operation. The NAT servers were capable of sustaining 800 Mbit/s of TCP/IP traffic passing through [9].

IV. Report on Experiments

The experiments were aimed at emulating a realistic operational scenario, whereby an operator located at CERN would control the system, including the computers located at the remote sites, from an Online software console.

The installation of DataFlow and Online software, officially supported only on some of the Linux versions installed at the remote sites, required the development of RPM packages for these systems in order to automate the procedure. The same user login and directory structure for the ATLAS-related software were defined at all sites. Eventually, scripts that allowed for the login and software installation directory structures to differ between the sites were developed.

The developers of the SFI, SFO and EF programs added parameters for increasing the socket buffers of their applications to values adapted to efficient transmission over long-distance networks. The events transferred contained dummy data, without any relevance for the physics. There was no physics analysis or processing performed on this data, nor dummy waiting intervals to simulate the analysis. The interest was to determine the maximum event transfer rate to each of the farms.

Event Filter farms were configured at the remote sites. An additional Event Filter farm, running in parallel with those located at the remote sites, was configured at CERN. The functionality of the SFI and SFO was configured on one computer located at CERN. The CORBA broker was forced to communicate over TCP/IP and to operate on a fixed network port, such that only a reduced number of ports would be opened on the firewall protecting the NAT servers.

The control function of the system performed as expected, once the right configuration settings were applied as described above. We were not expecting any problems, in view of the reduced scale of the test configuration and the fact that the communication was based on CORBA.

The results on the data path were, however, different from expectations. The event transfer rate was lower than the theoretical expectations. For example, the connection to IFJ PAN had a round-trip time to CERN of 54 ms, hence we were expecting an event transmission rate of about 18 Hz. Fig. 3 presents the results obtained during measurements performed between CERN and IFJ PAN.

The dependency between the transfer rate and the size of the event would be normal only for the first few events sent after establishing the connection, until the TCP window increases to the value of the bandwidth-delay product and allows for one event to be transmitted in a single burst of packets. However, such dependency was unexpected in the case of long intervals (such as an 18 hours-test between CERN and IFJ PAN). A detailed description and analysis of these results can be found in [10].

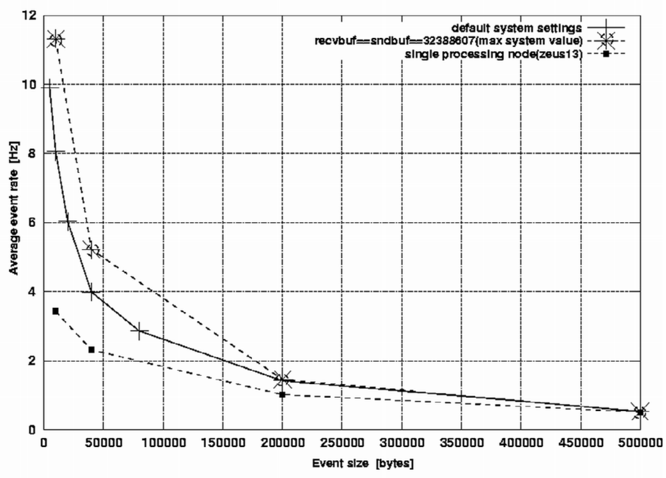

Fig. 3. - Rate versus event size

The transfer rate of the local processing farm remained constant throughout the duration of the tests (at a value of about 700 Hz), with no slowdown from the integration of the remote farms in the testbed setup.

The processing rate required in a production environment is only expected to reach 1-2 Hz per processing node, depending on the number of processors installed in the computer [1]. Hence the use of remote resources, even using the current mechanism for transferring the event data, would provide an almost linear increase in the computing power, comparable to locating those resources at CERN. A way of improving the remote transfer and processing model would be through integrating Grid-enabled resources.

Grids are expected to become widely deployed, at least in the academic world, by the time of the projected start of the ATLAS experiment in 2007. Efforts are under way, for example [4], to adapt the traditional batch processing-oriented implementation of Grids to interactively service demands. The major advantages brought by integrating interactive Grid services would be:

- de-couple the long-distance event transfer protocol from the transfers performed on the local farm, with the possibility of using a transfer protocol optimised for operation over long-distance networks
- the authentication, authorisation and accounting would be performed by existing Grid middleware
- the allocation of resources would be transparent, computers may be replaced in the remote site's configuration without the need to inform or modify the running configuration of the ATLAS Online system
- advanced load-balancing and scheduling schemes may be implemented to determine the most efficient resource allocation (in terms of bandwidth and computing power) for a given run

Work is carried on within the ATLAS Remote Farms community in order to determine and evaluate the best models for integrating Grid-enabled resources [11] [12].

V. Applications and Challenges

In the following, we will only address the technical implications of integrating remote computing power in the ATLAS Online system. Three potential types of processing may be applied to the event data:

- online detector calibration: determining incremental adjustments to the detector calibration settings that have to be available at the start of the next run
- online event monitoring: detailed monitoring of the physics content of events from the current run
- real-time event filtering: running physics analysis code to determine whether an event is considered interesting enough to be stored for detailed offline analysis

The calibration and monitoring tasks are similar with respect to the fact that they can be considered "interactive" but not necessarily "real-time". Variation of the processing and transmission times would not affect the outcome of the tasks. The event filtering, however, would require the provision of certain guarantees with respect to the largest network transit time due to the integration onto the data path of the HLT. These guarantees would have to include delays incurred by partial retransmissions due to packet losses over long-distance networks.

There are two important factors, relevant to a potential deployment, that we were unable to evaluate during our experiments: the cost of the long-distance bandwidth and the access to the ATLAS conditions database.

The long-distance connectivity that we used in our experiments was offered free of charge and on a best effort basis by the respective national research and education networks. It might be possible to benefit from similar arrangements when considering a production environment, but only on a case-by-case basis and as long as the bandwidth usage falls below a certain threshold. For the sake of the argument, we may consider this threshold as being equal to 10% of the bandwidth required for the Tier0-Tier1 communication, which roughly estimates the need to about 80 MB/s. Many national research networks have the infrastructure that would allow permanent access to a fraction of this bandwidth today.

However, this amount of bandwidth would only cover for calibration and monitoring application requirements. In the case of event filtering, exporting 10% of the processing needs to remote institutes translate into an aggregate traffic of about 500 MB/s. The fastest long-distance network connections available in production today can transfer about 9.5 Gbit/s of user data payload, therefore the remote event filtering application would mean provisioning a significant number of fast connections in order to make an impact.

The availability of the latest version of the ATLAS conditions database at a remote site is another problem that has to be tackled when considering remote event processing applications. The conditions database stores a set of constants related to the Detector Control System (DCS), online and offline calibration and detector alignment data and monitoring data characterising the performance of the detector hardware and software components during any particular time interval [13]. This means that the database might be updated after any given run and certainly updated when the calculation of the calibration constants becomes available.

Therefore, it is important that all the computers in the Online system access the same version of the database during a particular run. This would not be an issue for the computers that are part of the farms located at CERN. Different solutions are being considered in order to ensure a fast distribution of the about 100 MB that have to be accessed by every computer at the beginning of a run [13]. The transfer of 100 MB of data to remote locations has to be carefully implemented, when considering real-time or quasi-real time constraints. For example, it is unlikely that computers located remotely will be allowed to delay the start of a data taking session just because a certain network path is experiencing unusually high packet loss rate, requiring partial retransmission of the data.

VI. Conclusion and Future Work

Through experimentation on a widely distributed testbed we demonstrated that remote computing facilities could be used to analyse, in real time, for the purposes of the ATLAS Trigger and Data Acquisition System, data produced at the future LHC at CERN. We found that the ATLAS DataFlow and Online Software can be configured in a way to seamlessly integrate remote processing capabilities, with virtually no impact on the local event processing rate.

We outline directions of further studies on how to integrate such resources in our data processing with real-time constraints scenario. These directions are currently pursued within the ATLAS Remote Farms collaboration.

Acknowledgments

The authors would like to express their gratitude to the members of the TDAQ testbeam support teams and to the software developers of the different HLT system components for their help during the experiments.

The connectivity for the testbed was kindly offered by the following research network operators: CANARIE, Darenet, Geant, PSNC, SURFnet, UKerna and also by the DataTAG project (IST-2001-32459). We are thankful to the operators of the respective research networks and to the CERN external networking team for quickly troubleshooting various network problems that we encountered throughout the testing interval.

References

[1] ATLAS High-Level Trigger, Data Acquisition and Controls, Technical Design Report, CERN/LHCC/2003-022, July 2003.

[2] Adams, D; Barberis, D; Bee, C P; Hawkings, R; Jarp, S; Jones, R; Malon, D; Poggioli, L; Poulard, G; Quarrie, D; Wenaus, T; The ATLAS Computing Model, ATL-COM-SOFT-2004-009, December 2004.

[3] R.W. Dobinson, K. Korcyl, J. Hansen and M. Turala; Prospects for the use of remote real time computing over long distances in the ATLAS on-line system, Contribution to the Int. Europhysics Conference on High Energy Physics, July 17th-23rd 2003, Aachen, Germany

[4] The ATLAS Level 1 Trigger Technical Design Report, CERN/LHCC/98-14, June 1998

[5] S. Kolos et al. Experience with CORBA communication middleware in the ATLAS DAQ, CHEP 2004, Interlaken, Switzerland

[6] R. Hawkings, F. Gianotti, ATLAS detector calibration model – preliminary subdetector requirements, 28 February 2005

[7] S. Gadomski et al., Integration of ATLAS Trigger and Data Acquisition system in the Combined Beam Test, the 14th IEEE Real Time Conference 2005, Stockholm, Sweden

[8] Devil Linux, http://www.devil-linux.org

[9] N. Guillod, NAT & Firewall @ 1 Gbit/s, Ecole de Metiers de Sainte-Croix, June 2004

[10] R. Hughes-Jones et al., Investigating the Network Performace of Remote Real0Time Computing Farms For ATLAS Trigger DAQ, the 14th IEEE Real Time Conference 2005, Stockholm, Sweden

[11] K. Korcyl, T. Bołd, T. Szymocha, Crossgrid Resources as Remote Farms for the ATLAS Online Calibration, Monitoring and Filtering, ATLAS TDAQ week workshop, April 2005

[12] B. Caron, High Bandwidth Real-Time Remote Processing Systems and Grids for the ATLAS High Level Trigger, CANARIE's 10th annual Advanced Networks Workshop, November 2004

[13] The ATLAS Computing Technical Design Report, draft 2.0, 28 May 2005