LHCb Distributed Computing and the Grid

N. Brook[a], H. Bulten[b], J. Closier[c], D. Galli[d], C. Gaspar[c], F. Harris[e], K. Harrison[f], E. van Herwijnen[c], A. Khan[g], S. Klous[b], G. Kuznetsov[h], U. Marconi[d], P. Mato[c], I. McArthur[e], G. N. Patrick[h], A. Soroko[e], A. Tsaregorodtsev[i] and V. M. Vagnoni[d]

[a]H. H. Wills Physics Laboratory, Bristol, United Kingdom

[b]University of Vrije and NIKHEF, Amsterdam, The Netherlands

[c]European Laboratory for Particle Physics, Genève, Switzerland

[d]University of Bologna and INFN, Bologna, Italy

[e]University of Oxford, Oxford, United Kingdom

[f]University of Cambridge, Cambridge, United Kingdom

[g]University of Edinburgh, Edinburgh, United Kingdom

[h]CLRC Rutherford Appleton Laboratory, Chilton, Didcot, United Kingdom

[i]University of Aix, Marseille, France

The current architecture of the LHCb distributed system for Monte Carlo data production is described. An overview is given of the current and planned use of Grid technology from the European Datagrid Project (EDG), and of the development of an experiment-specific user interface to Grid services.

1. Introduction

LHCb has been developing a distributed system for Monte Carlo production since 1999. It currently operates in 15 sites distributed throughout the Collaboration, including both major national centres (CERN, RAL-UK, IN2P3-France, INFN-Italy, Nikhef-Netherlands), and regional and university centres (Bristol, Cambridge, Glasgow and Edinburgh, Liverpool, Oxford, ITEP/Moscow, UFRJ/Rio de Janeiro). Expansion to centres in Switzerland, Poland, Germany and Spain is planned for 2003.

This paper describes the current system architecture, based on Web Submission tools and Java servlet technology, the current use of Grid technology, and the future planning in terms of executing large-scale "Data Challenges" together with the phased integration of tools emerging from the European Datagrid (EDG) project. In addition we describe the developments for LHCb and LHC computing in the UK, and some of the work towards developing a user interface to Grid services.

2. The Monte Carlo Production System

SLICE (Simulation for LHCb and its Integrated Control Environment) [1] takes its input from a production manager who defines the required number of jobs, each job typically comprising 500 events, according to the physics request for events, physics channels, data type, software configuration and deadline for completion. The job creation system uses Java servlets to creates job scripts corresponding to the request inserted from a Web Server. The job configuration data includes program name and version number, input data type, output data type, and detector database version number.

Job scripts are written to an area that can be read by the batch worker nodes. Currently AFS is used for sharing data in the non-grid environment. The batch nodes must be able to execute Java programs, and hence the Java run time environment must be available. The Java programs must communicate with a centralized Java servlet that updates the bookkeeping database and is running at a machine at CERN, so the batch nodes need to be open to IP traffic.

The servlets create scripts that are different for each centre, depending on the facilities available at that centre. With Grid facilities this would be hidden from the production manager, with the detail of resource finding being accomplished by Grid software.

It should be noted that in using Datagrid middleware AFS cannot be used, and an "input sandbox" is used to package the LHCb runtime environment for the job.

A commercial control system (PVSS), adopted by the LHC experiments for control applications, is used to submit jobs, and to follow the status of the jobs through the various states prior to, during, and after execution.

With the current configuration all DST (Data Summary Tape) files are transferred to CERN. The scripts copy the output datasets to a staging area at the end of the job. A Java program uses the file transfer package BBFTP [2] to transfer the files to the correct path in the mass storage system CASTOR at CERN. This data transfer operation is very reliable. The Bologna group are currently achieving a throughput of up to 7 MB/s with a negligible transfer failure rate.

The Bookkeeeping service is accomplished using the central Oracle system at CERN.

3. GridPP and LHCb

In this section we discuss the developments that are occurring in the UK to establish computing facilites, for use by LHCb and other high-energy particle physics collaboration, under the auspices of GridPP. GridPP is a collaboration of Particle Physicists and Computing Scientists from the UK and CERN, who are building a Grid for Particle Physics. The project has now completed one year of operations and there is active engagement across 16 UK Particle Physics groups.

The MONARC model [3] of LHC computing involved a hierarchical arrangement of tiered regional centres, starting from Tier-0 at CERN, to major Tier-1 centres in several countries, to smaller Tier-2 and Tier-3 centres in institutes and departments. This model is now maturing; the hierarchical nature becoming less distinct with the concept of virtual Regional Centres in a so-called 'Cloud Model.' However the terms Tier-1, Tier-2 etc are still retained but now a service oriented view is developing within the LHC Grid project [4] that defines a categorization of regional centres [5]. Within the UK it is expected that there will be four Regional Tier-2 Centres in addition to the UK Tier-1 Regional Centre.

The prototype Tier-1 centre that exists for the LHC and the LHC Computing Grid also acts as a Tier-A centre for the BaBar collaboration is based at the Rutherford Appleton Laboratory (RAL). This centre consists of 156 dual CPU 1.4GHz rack mounted systems with 26 dual CPU disk servers with 2 RAID controllers/server each with 0.8TB of usable RAID5 disk i.e. 1.6TB/server. The resources purchased for the Tier-1/A centre were integrated into the existing linux farm at RAL. In addition, one extra STK 9940 tape driver was purchased for the tape robot, giving a total of 5 9940s and 5 IBM 3590B drives. The capacity of the tapes in the robot is 100TB. All the new equipment is attached directly or indirectly to a Summit 7i Gigabit ethernet switch. This is part of an ongoing hardware purchase programme to establish the prototype Tier-1 centre in the UK.

All UK institutes with major experimental particle physics groups have expressed an interest in being part of a Tier-2 centre. In fact several institutes were in a position to be standalone Tier-2 centres. In order to allow all institutes to participate fully it is envisaged that four Tier-2 Regional Centres will be developed. There are natural geographical groupings for these centres: London, Southern England, Northern England and Scotland (the ScotGrid project). All four groups have a strong representation of institutes involved in LHCb. The ScotGrid project is already well-established as a prototype Tier-2 centre. It was

proposed primarily for the analysis of data from the ATLAS and LHCb experiments. The centre consists of a 128CPU Beowulf Monte Carlo production facility maintained in Glasgow and a 5TB datastore and associated high-performance server run by the Edinburgh Parallel Computing Centre. The project is a partnership with IBM to study optimal configuration of a Monte Carlo production farm and a database server. The project also benefits from close association with the UK National e-science centre which is based in Edinburgh.

4. Current and planned use of Grid tools

LHCb has participated in the European Datagrid project since its birth in January 2001 and are co-authors of the project's application-requirements document [6], and the experiments' evaluation report for the first version of middleware [7]. Work in progress is focused on evaluating the project middleware, and on integrating this middleware into the production and analysis environment of LHCb.

Globus tools have been used for job submission to farms at RAL and Lyon since 2001 [8]. The Datagrid Testbed was integrated into the LHCb production system in March 2002, and so is one of the resources available for MC production. Consequent to this development, the system was used for the first EU project review in April 2002. The EDG Job Submission system using the Job Description Language (JDL) has succesfully been used to submit and execute jobs on all 5 major components of the testbed: CERN, INFN/CNAF, IN2P3, Nikhef and RAL. The "Sandbox" technology is used in place of AFS to transfer the job environment (executable, input and output files).

Tests of EDG tools are currently taking place for data replication and mass storage. In addition, the middleware for job monitoring will be evaluated in comparison with PVSS.

The advantage of Grid tools will be to provide a uniform interface to a heterogeneous environment, taking away the considerable effort that goes into producing servlets that are site dependent. Given the requirements of the job expressed in terms of software configuration, and of CPU and storage needs, the Grid will direct the job to the appropriate site.

The LHCb VO (Virtual Organisation) is already in place, and work is proceeding on the mapping of individuals and their responsibilities in the experiment to the grid-map-file defining user privileges at each site. It is expected that the EDG tool VOMS (VO Management System) will be used for this task.

5. Data Challenges

LHCb is performing Physics Data Challenges (PDC) at the end of 2002 and beginning of 2003. These will generate of the order 10 million events, including signal, specific backgrounds and minimum bias, to be used for detector, physics and trigger studies.

In parallel, Computing Data Challenges (CDC) will be performed to check software developments. Components, including Grid tools, will be incorporated into the PDC activity once proven in the CDCs. It is hoped that a reasonable percentage of essential physics data can thus be generated using Grid tools in 2003.

6. LHCb-GRID R&D: the Ganga interface

A user-Grid interface allowing configuration of applications implemented in the Gaudi/Athena OO software framework of ATLAS and LHCb [9] is being developed as a joint ATLAS/LHCb project: the Gaudi/Athena and Grid Alliance (Ganga) [10]. The design of Ganga implies that it will be a GUI-based application providing user support for the complete job life-time. In particular, Ganga services should be developed for job preparation and configuration, resource booking, job submission, job monitoring and control. The requirements for the interface include the following:

- The user will interact with a single application covering all stages of the job life-time.
- The user will be able to restore his or her workspace (list of files, state of tools, jobs

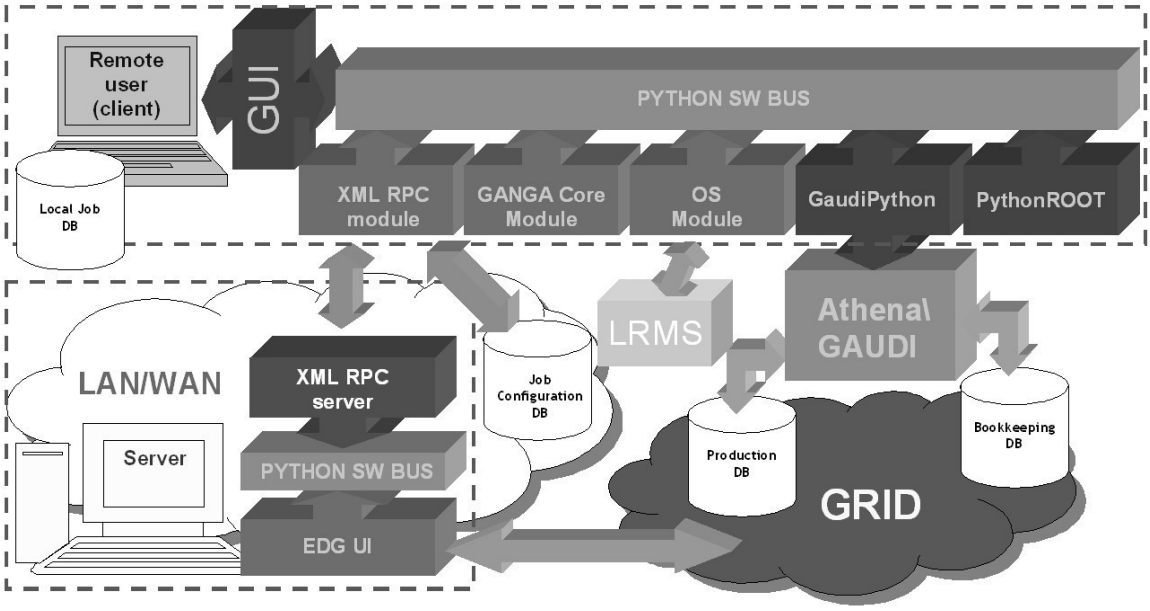

Figure 1. A schematic of the Ganga design.

in preparation) at the beginning of each session.

- The GUI will allow the user to submit jobs locally as well as to the Grid.
- The user will be able to access the interface not only from the computer where the Grid UI program runs, but also from a remote "thin" client.

Ganga will also include a GUI for browsing Grid resources, (e.g., Virtual Organisation - VO - active services, the list of Computing Elements - CEs - and Storage Elements - SEs -), and a GUI for data management tools (e.g., registration of datasets to the Grid). During the job preparation stage the user will be able to access the Job Configurations Database and retrieve the required configuration for his or her Gaudi application using high-level commands. The user will have the possibility of modifying settings and storing personalized configurations in his or her own area. Ganga will assist in the installation of missing components, if necessary. Ganga will then translate the user configuration and data requests into one or more files of JDL. Most likely it will contact the Gaudi Bookkeeping Database and the Grid Replica Catalogue to obtain the list of Logical File Names (LFNs) from high-level physics selection criteria. If requested by the user, Ganga will pre-book Grid resources before actual job submission. During the execution stage it will be possible to monitor the internal job state using special Gaudi services publishing information to the Grid. These are currently under development. Some functionality from the ROOT package (e.g., histogram drawing) may be added to complete this task. After job execution Ganga will provide tools for the user (Production manager) for updating the Bookkeeping Database used by Gaudi applications.

The architecture of the Ganga program should be modular, to permit flexibility in the implementation. At present, a design that uses Python as

a software bus is being developed. A schematic of the Ganga design is shown in Figure 1. Python is used to glue together the different modules that provide the functionality of the interface, allowing interaction and communication between them.

The first steps to create a Ganga prototype have been accomplished with the development of Python classes representing jobs, a local job registry, and some GUI elements. Remote access to the LHCb Job Configurations Database has been implemented using the xmlrpclib Python extension module. Serialization of objects (user jobs) is done using the pickle module. Ways in which to exploit EDG Grid tools are currently under study.

7. Conclusions

LHCb is proceeding in parallel with Data Challenges and Grid developments. It is expected that simple Grid middleware functionality for job submission, data replication and storage will be integrated into the production system by the end of 2002. As a necessary corollary to this, LHCb will develop its VO in terms of "groups" and "group authorisations" mapped onto the sites available to the experiment.

A longer term development is taking place in the area of GANGA, which will provide an interface to Grid facilities for both production managers and physicists performing analysis. A first prototype of Ganga should be available early in 2003.

REFERENCES

1. *SLICE - the LHCb distributed Monte Carlo production system* `http://lhcb-comp.web.cern.ch/lhcb-comp/ ComputingModel/datachallenges/slice.doc`
2. IN2P3 Computing Centre, *bbftp home page*, `http://doc.in2p3.fr/bbftp`
3. MONARC - Models of Networked Analysis at Regional Centres for LHC Experiments, `http://monarc.web.cern.ch/MONARC/docs/phase2report/Phase2Report.pdf`
4. LHC Computing Grid Project - `http://lcg.web.cern.ch/lcg`
5. L. Bauerdick et al., *Regional Centre Category and Service Definition* CERN-LCG-2002-16 `http://lcg.web.cern.ch/LCG/SC2/RTAG6/finalreport.doc`
6. *DataGrid User Requirements and Specifications for the DataGrid Project*, `http://datagrid-wp8.web.cern.ch/DataGrid-WP8/Documents/Workspace/D8.1/ D8.1.a_v2.1.pdf`
7. *Datagrid Report on results of run #0 for HEP applications*, `http://datagrid- wp8.web.cern.ch/DataGrid-WP8/Documents/Approved/D8.2_3.0.pdf`
8. F. Harris, E. van. Herwijnen *et al*, *Moving the LHCb Monte Carlo Production System to the Grid*, Proc. CHEP 2001, Science Press, pp. 676-680
9. P. Mato *et al*, *Status of the GAUDI Event-processing Framework*, Proc.CHEP 2001, Science Press, pp. 209-213
10. *GANGA project development*, `http://lhcb-comp.web.cern.ch/lhcb-comp/Frameworks/Ganga/default.htm`