The LHCb Computing Strategy

N. H. Brook[a] *

[a]H. H. Wills Physics Laboratory, University of Bristol
Bristol BS8 1TL, United Kingdom

The LHCb experiment, based at the Large Hadron Collider at CERN, Geneva, is preparing for the first data taking during 2008. The LHCb computing architecture for processing and analysing the data in a distributed computing environment is introduced. The readiness of the computing tools needed for physics analysis is addressed. The experience of transparently harnessing the distributed computing resources is reported.

1. COMPUTING MODEL

The dataflows of the LHCb computing model for all stages in the processing of the real and simulated LHCb events are described [1]. The roles of the various Tier centres are discussed and the distribution of the processing load and storage are outlined.

There are several phases in the processing of event data. The various stages normally follow each other in a sequential manner, but some stages may be repeated a number of times. The workflow reflects the present understanding of how to process the data. A schematic of the logical dataflow is show in Figure 1 and is described in more detail below.

The "real" raw data from the detector is produced via the Event Filter farm of the online system. The first step is to collect data, triggering on events of interest. The RAW data are transferred to the CERN Tier 0 centre for further processing and archiving. The RAW data, whether real or simulated, must then be reconstructed in order to provide physical quantities such as calorimeter clusters to provide the energy of electromagnetic and hadronic showers, trackers hits to be associated to tracks whose position and momentum are determined. Information about particle identification (electron, photon, π^0, hadron separation, muons) is also reconstructed from the appropriate sub-systems. The event reconstruction results in the generation of new data, the Data Summary

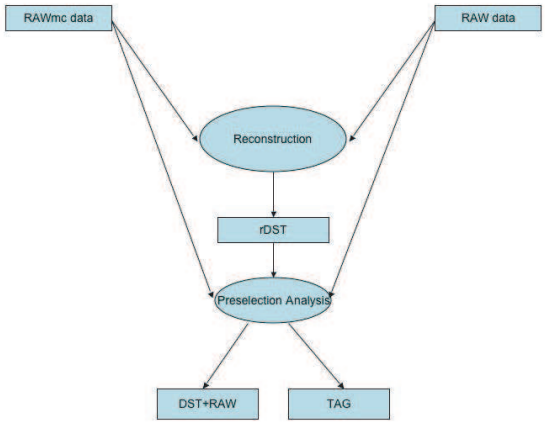

Figure 1. The LHCb computing logical dataflow model

"Tape" (DST). Only enough data will be stored in the DST, that is written out during reconstruction, to allow the physics preselection algorithms to be run at a later stage. This is known as a reduced DST (rDST.) The first pass of the reconstruction will happen in quasi-real time. It is planned to reprocess the data of a given year once, after the end of data taking for that year, and then periodically as required. This is to accommodate improvements in the algorithms and to make use of improved determinations of the calibration and alignment of the detector in order to regenerate new improved rDST information.

*on behalf of the LHCb collaboration

The rDST is analysed in a production-type mode in order to select event streams for individual further analysis. This activity is known as "stripping." The rDST information is used to determine the momentum four vectors corresponding to the measured particle tracks, to locate primary and secondary vertices and algorithms applied to identify candidates for composite particles whose four-momentum are reconstructed. Each particular channel of interest will have a preselection algorithm provided by the relevant physics working group. The events that pass a physics working group's selection criteria are written out for further analysis. Since these algorithms use tools that are common to many different physics analyses they are run in production-mode as a first step in the analysis process. The events that pass the selection criteria will be fully re-reconstructed, recreating the full information associated with an event. The output of the stripping stage will be referred to as the (full) DST and contains more information than the rDST. Before being stored, the events that pass the selection criteria will have their RAW data added in order to have as detailed event information as needed for the analysis. An event tag collection will also be created for faster reference to selected events. It contains a brief summary of each event's characteristics as well as the results of the pre-selection algorithms and a reference to the actual DST record. The event tags are stored in files independent of the actual DST files. It is planned to run this production-analysis phase 4 times per year: once with the original data reconstruction; once with the re-processing of the RAW data, and twice more, as the selection cuts and analysis algorithms evolve.

The baseline LHCb computing model is based on distributed multi-tier regional computing centres. It attempts to build in flexibility that will allow effective analysis of the data whether the Grid middleware meets expectations or not. A schematic of the LHCb computing model is given in Figure 2.

CERN is the central production centre and will be responsible for distributing the RAW data in quasi-real time to the Tier-1 centres. CERN will also take on a role of a Tier-1 centre. An additional six Tier-1 centres have been identified: CNAF(Italy), FZK(Germany), IN2P3(France), NIKHEF(The Netherlands), PIC(Spain) and RAL(United Kingdom.) There are also a series of Tier-2 computing centres. CERN and the Tier-1 centres will be responsible for all the production-processing phases associated with the real data. The RAW data will be stored in its entirety at CERN, with another copy distributed across the other 6 Tier-1 centres. The re-processing of the RAW data, during the LHC shutdown, will also use the resources of the LHCb online farm. As the production of the stripped DSTs will occur at these computing centres, it is envisaged that the majority of the distributed analysis of the physicists will be performed at CERN and at the Tier-1 centres. The current year's stripped DST will be distributed to all centres to ensure load balancing. The Tier-2 centres will be primarily Monte Carlo production centres, with both CERN and the Tier-1 centres acting as the central repositories for the simulated data. It should be noted that although we do not envisage any user analysis at the Tier-2's in the baseline model presented, it should not be proscribed, particularly for the larger Tier-2 centres.

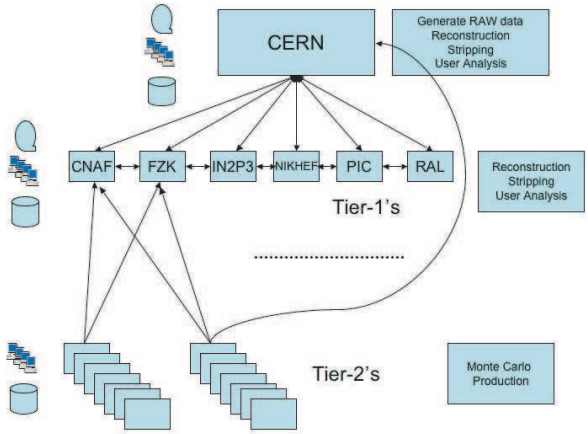

Figure 2. Schematic of the LHCb Computing Model.

It is expected that the reconstruction and the

first stripping of the data at CERN and at the Tier-1 centres will follow the production in quasi real-time, with a maximum delay of a few days. The DST output of the stripping will remain on disk for analysis and be distributed to all other Tier-1 centres and CERN, whilst the RAW and rDST will be migrated to the mass storage system, MSS.

The re-processing of the data will occur over a 2-month period. During this process the RAW data will need to be accessed from the MSS both at CERN and the Tier-1 centres. The CPU resources available at the pit allow a significant fraction of the total re-processing. Hence at CERN there is an additional complication that the RAW data will also have to be transferred to the pit; similarly the produced rDST will have to be transferred back to the CERN computing centre. To enable later stripping it is necessary to distribute a fraction of the rDST produced at CERN during this re-processing to the Tier-1's; this is a consequence of the large contribution from the online farm.

The (two) stripping productions outside of the reconstruction of the RAW data will be performed over a one-month period. Both the RAW and the rDST will need to be accessed from the MSS to perform this production. The produced stripped DSTs will be distributed to all production centres.

The Monte Carlo production is expected to be an ongoing activity throughout the year The whole of the current year's Monte Carlo production DST will be available on disk at CERN and another 3 copies, on disk, distributed amongst the other 6 Tier-1 centres.

2. HARNESSING THE GRID

2.1. DIRAC

LHCb will have to integrate a coherent system of resources and Grid services to carry out its computing tasks in the distributed environment. DIRAC [2] is designed to be highly adaptable to the use of heterogeneous computing resources available to the LHCb Collaboration. These are mainly resource provided by the LHC computing Grid, WLCG [3]. However, other resources provided by sites not participating in the WLCG, as well as a large number of desktop workstations can be easily incorporated. One of the main design goals is the simplicity of installation, configuring and operation of various services. Once installed and configured, the system automates most of the management tasks, which allows all the DIRAC resources to be easily managed by a single Production Manager.

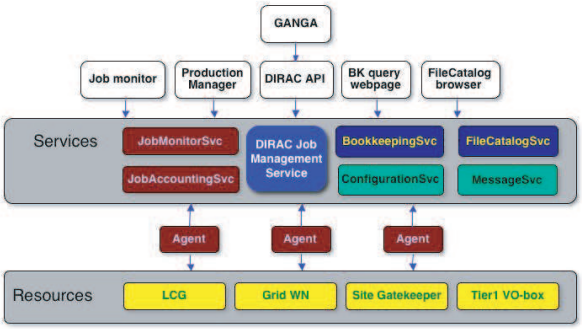

Figure 3. A schematic representation of the DIRAC architecture of resources, services and agents. Examples of services, resources and clients are illustrated.

DIRAC uses the paradigm of a Services Oriented Architecture (SOA). The main DIRAC components are Resources, Services and Agents as illustrated in Figure 3. Resources represent the Grid computing and storage elements and provide access to their associated capacity and status. Services are providing access to the full functionality of the DIRAC system in a well controlled way. Jobs are interacting with the system via services to accomplish their work. Agents are lightweight software components usually running close to the computing resources. The DIRAC main subsystems, Workload Management (WMS) and Data Management, are combinations of central Services and distributed Agents. This achieves an efficient operation of the distributed system with an easy and non-intrusive deployment of its distributed part. Since the Grid environment is intrinsically very dynamic, the efficient deployment is one of the most important characteristics

of the system.

The WMS allows reservation of computing resources. This takes advantage of having a light easily deployable agent, which is part of the DIRAC native WMS. The jobs that are sent to the Grid Resource Broker (RB) are just executing a simple script, which downloads and installs a standard DIRAC agent on the worker node, WN. Once this is done, the WN is reserved for the DIRAC WMS and is effectively turned into a virtual DIRAC production site for the time of reservation. The reservation jobs are sent whenever there are waiting jobs in the DIRAC task queue eligible to run on the site.

2.2. GANGA

A physicist analysing data from LHCb will have to deal with data and computing resources that are distributed across multiple locations. GANGA [4] has been developed, in cooperation with ATLAS, to help with this task by providing a uniform high-level interface to the different low-level implementations for the required tasks, ranging from the specification of input data to the retrieval and post-processing of the output. GANGA presents the user with a single interface rather than a set of different applications. It uses pluggable modules to interact with external tools for operations such as querying metadata catalogues, job configuration and job submission. At start-up, the user is presented with a list of templates for common analysis tasks, and information about ongoing tasks is persisted from one invocation to the next. GANGA can be used either through a command line interface or through a Graphical Graphical User Interface. Their behaviour are completely linked allowing easy transition from one to the other.

A job in Ganga is constructed from a set of building blocks. All jobs must specify the software to be run (application) and the processing system (backend) to be used. Many jobs will specify an input dataset to be read and/or an output dataset to be produced. Optionally, a job may also define functions (splitters and mergers) for dividing a job into subjobs that can be processed in parallel, and for combining the resultant outputs. In this case after splitting the job becomes a master job and provides a single point of access for all its subjobs. The user can define the operations to be performed within a job, and to store information returned by the processing system, allowing tracking of job progress.

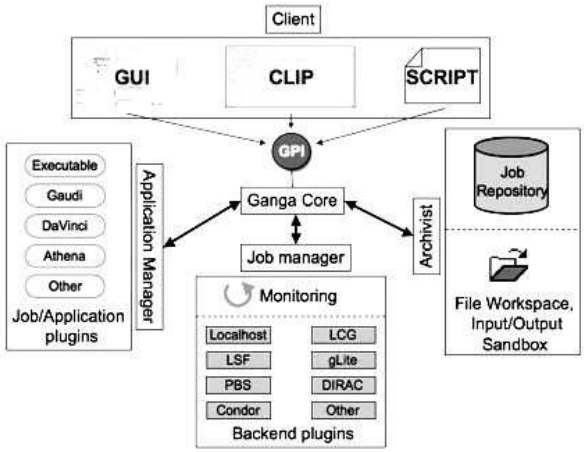

Figure 4. Schematic representation of the GANGA architecture. The main functionality is divided between Application Manager, Job Manager and Archivist. A client can run the GUI, the Command-Line Interface (CLIP) or scripts

The architecture of GANGA is such that the functionality of GANGA is divided between a number of components illustrated in Figure 4. A number of common tasks are provided by the core and in addition it links the components together. The components are categorised as Application and Job Managers, Job Repository, and File Workspace. All the components communicate via a well-defined interface. There are three ways a client can communicate with GANGA: through a shell (CLIP); using scripts, or through the GUI.

3. EXPERIENCES

3.1. Production

LHCb have been very successful in running production on the Grid since 2004. The overlay paradigm of DIRAC, described in section 2.1, has been critical to this success. An agent checking the run time environment prior to downloading a workload has been crucial for efficient running with a rather modest manpower effort. This was particularly relevant when the Grid was in its infancy.

A LHCb simulation job runs for approximately 24 hours. Over the last year typically 5000 simultaneous Monte Carlo jobs have been running, with a peak of 10000 jobs. The only limiting factor on the number of jobs encountered was the available WLGC capacity for LHCb. LHCb have run at 80 distinct sites during the production, including many sites where there are no LHCb collaborators. The production, across all these sites, is run by a single person entering jobs into the central DIRAC WMS.

A data challenge testing the re-processing of the data through the reconstruction software is still ongoing. This processing of the data is only performed at the LHCb Tier-1 sites. The data to be processed have to be re-staged from the sites' mass storage system, MSS. This has revealed many issues. The MSS systems have often been optimised to deal with large transfer rates into the site from external sources. This led to inefficiencies and instabilities in many MSS. A pre-stage command issued from the WN prior to running the application proved not to be the optimal approach to access the data due to these instabilities. LHCb developed, within the framework of the DIRAC data management system, a remote tape stager agent. This ensured all files required by the application were accessible and were staged prior to the submission of the job to the site. In the development of this stager other problems associated with the implementations of the generic grid storage interface, SRM [5], to some of the MSS backends were revealed. One limitation of this approach is the inability to pin files on disk until there is free CPU for the reconstruction job to run at the site. This limitation will be addressed in version 2.2 of SRM. Despite the many issues surrounding data access LHCb achieved 450 simultaneous reconstruction jobs running. This corresponds to simultaneously processing the order of 10000 files.

The LHCb computing model envisages the data needed for analysis will be stored on disk to avoid the need to re-stage data from the MSS. The wisdom of the decision has only be reinforced by the experiences gained during the re-processing challenge. LHCb analysis is organised through GANGA, using DIRAC as a backend, to submit user jobs to the Grid. Using DIRAC allows LHCb to keep control of the resources allocated to them and to match the priorities of jobs to reflect those of the collaboration. Since the start of 2007 they have been 99 unique LHCb users of GANGA with 10000 LHCb GANGA sessions. These users have submitted 393k analysis jobs through the DIRAC system, of which 85% were executed at the LHCb Tier-1 centres. This reflects the fact that data for analysis is held at these centres. Those jobs executed outside of the Tier-1 centres have no need to access the simulated data, for example Toy Monte Carlo jobs.

4. SUMMARY

The LHCb computing model is currently being finalised and is under stress test, in particular the issues associated with access to data. LHCb have developed the DIRAC system in order to allow efficient use of Grid resources. Monte Carlo production is now routine with major effort now invested in understanding how best to reprocess the data. There has been a major increase in the number of LHCb physicists using the Grid. LHCb, in collaboration with ATLAS, have developed GANGA in order to assist in the preparation of analysis jobs for the Grid. GANGA presents a simple interface between Grid resources, accessed via DIRAC, and the LHCb software framework.

REFERENCES

1. R. Atunes Nobrega et al., LHCb Computing TDR, CERN/LHCC 2005-019.

2. A. Tsaregorodtsev et al., DIRAC, the LHCb Data Production and Distributed Analysis System, Proc. of CHEP 2006, Mumbai.
3. I. Bird et al., LHC Computing Grid TDR, CERN/LHCC 2005-024.
4. U. Egede et al., GANGA – A Grid User Interface, Proc. of CHEP i 2006, Mumbai.Mumbai
5. SRM working group, http://sdm.lbl.gov/srm-wg/