GLED within a HEP Collaboration: a Case Study

Matevž Tadel
Jožef Stefan Institute, Ljubljana, Slovenia
Email: Matevz.Tadel@ijs.si

Document in preparation

Started: December, 2002

Abstract: This document presents my views of future computing in *high-energy physics* (HEP). The arising problems, new to scientific computing, are quantified and discussed. The rest of paper is an attempt at defining possible roles of GLED framework within a HEP collaboration.

Disclaimer: this paper does not, in any way, represent the official view of any organization. All data, statements and implications originate from my understanding of affairs which may not be up to date or in accordance with official positions.

1 Context — the HEP challenge

The arising HEP experiments are facing a severe problem with respect to their computational infrastructure. As I am best acquainted with the state of affairs for experiments being built at the Large Hadron Collider (LHC, CERN, Switzerland), they will be used as a reference in the following article.

First a general outline of the data processing pipeline leading from raw data to physical results will be made. Then the most painful problems of the LHC computing will be presented together with solutions that are planned to overcome them.

1.1 An outline of data-flow in a HEP experiment

Not yet written

1.2 Problems of the next-generation computing at the LHC

1.2.1 Data production

Problem: Each of the new experiments will be producing ∼100 MB of raw data per second of operation. Further, a lifetime of 15 years is anticipated for the experiments, meaning that the lot of accumulated data will amount to 50 PB per collaboration. The data production rate as well as the final amount of generated data pose serious logistic problems.
Solution: Data will be distributed among several computing centers on behalf of organizations forming the HEP collaborations. Several copies of each recorded event will exist in parallel.

1.2.2 Low-level computing resources & intermediate data

Problem: A full cycle of a HEP analysis is very complex. Reprocessing of basic data is needed together with full simulation of physical processes and detector's response to fine-tune the reconstruction algorithms and to understand experimental and systematic errors. Different analyses employ different parameters for event selection and reconstruction. This leads to problems with respect to CPU power and storage needed. Sources of data and algorithms used to process it must be traceable.
Solution: Analysis will be segmented to allow for storage of intermediate results which will be efficiently tagged and databases will be used for indexing of the important tags. Concurrent software versions will be used. Regional centers will provide computing power for reprocessing of data and data-summaries will be broad-casted to other computing centers.

1.2.3 User access to data & low-level computing resources

Problem: User-analyses sometimes need access to low-level data to apply corrections or to perform additional event-selection. In the model of distributed data this means that users need access to foreign computing centers and to collaboration's resources. There will be ∼1000 users using the resources simultaneously.
Solution: `grid` technologies will provide users with the ability to access collaboration data and resources. Persons responsible for maintenance of some shared resource will have appropriate administrator-privileges to perform their task on collaboration-wide scale.
Comment: `grid` is a powerful weapon, but the complexity imposed by the nature of HEP data-flow can lead to ambiguities (poorly defined problems) and frustration of users. Intermediate information resources will be needed and experts for lower-level analysis will have to provide service to end-user-physicists.

1.2.4 Data analysis & visualization

Problem: Analysis on all levels of reconstruction is complex and has multi-dimensional dependencies. On the level of physical analysis a user has to consider physical process and its background channels, detector response, event selection and trigger mechanism as well as understand algorithms that produce objects of his analysis. A similar problem arises for optimization of low-level reconstruction algorithms which provide intermediate reconstructed objects and have to satisfy a set of (possibly contradicting) requirements.
Solution: Several work-groups have specific tasks assigned to them. They provide service in their area of expertise for the whole collaboration. This is not seen as a real problem.
Comment: Personally, I believe that the problem is underestimated, especially due to entangled levels. Mistakes in understanding and usage of the complex computational infrastructure will yield inconsistent (or simply wrong) results. Sources of errors will be hard to trace and correct. Experimental and systematic error estimation will also be problematic and will involve complex computations. Specialized data analysis and visualization tools will be needed to reduce technical complexity to a manageable level.

2 Gled as a monitor of collaboration-wide computing resources

In this section a GLED cluster spawning across all computing sites belonging, at least in part, to collaborating institutions is described. The main purpose of such cluster is to allow users to perform

any of the following tasks:

- Monitor the available computing and networking infrastructure and its performance.
- View (or query) available resources in general or on per computing center basis.
- View and modify parameters that control the behaviour of computing infrastructure. Make *reservations* for resources that will be needed in a near future.
- Instantiate user-defined monitors to collect data needed for optimization of cluster's performance.
- Instantiate control algorithms to perform maintenance of resources (e.g. mirroring of data, garbage collecting, back-up, update of databases).
- Communicate with other users by sending them a copy of a working environment (a view into the cluster structure) and messages.

To meet all these requirements several software components must work in harmony. Even though most of them could be written as GLED classes, this is not at all necessary and in some cases it is even undesirable. Instead wrappers querying the respective information servers can be used to export the needed information into a GLED system for viewing and user interaction.

2.1 Cluster organization

Sun Absolute of the collaboration is located at the central site (at CERN for LHC experiments). *Regional centers* (RCs) as well as CERN's off-line computing centre connect to it and are visible as its first-level Saturns.

2.1.1 Sun Absolute

Sun-space of the Sun Absolute is partitioned into several queen-spaces which group together different services.

1. *Sun queen* holds information about all Saturns connected into the cluster and therefore provides the structural view of the GLED cluster. All users connected into the cluster via Eyes are also visible in the sun-queen and various messages can be sent to them.
2. *Global configuration* queen holds objects controlling the collaboration wide behaviour of RCs. Policies for access to data and information services can also be specified here. This queen is mandatory.
3. *General statistics* queen encapsulates data representing usage and availability of resources for each RC. The availability and planned usage is particularly interesting for RCs which share their computational infrastructure with other *virtual organizations* (VOs). This is also the natural place for gathering statistics of inter-site network traffic.
4. *On-line* queen shows statistics of newly collected data and allows for basic monitoring of experiment operation. Information on routing algorithms and RCs that receive copies of new events are also exported from this queen.

5. *RC queens.* Each RC has its representative queen on the Sun Absolute where a more detailed information on its status can be found. These queens also provide a more detailed structure of each RC and can be used as generators of MIRs that will perform more detailed queries targeted at the RC in question. Information stored in them is produced by the lower-level Saturns themselves, either as a result of a query or as a periodical update from lower-level monitors.

This partitioning allows basic status to be exported to all collaboration members while allowing them to retrieve a more detailed summary of each centre (by requesting mirroring of its representing queen) or to access its information servers directly to perform the desired query.

2.1.2 Dependent Saturns

Regional centers are represented by first level Saturns. These are computing nodes with access to local information servers (some of them might be part of GLED itself). They observe local infrastructure and report (by pull or push method) its status to queens on the Sun Absolute.

Furthermore, they can spawn lower level Saturns and delegate them to perform sub-tasks such as reprocessing of data, low-level analysis, Monte Carlo event generation and detector response simulation. Each of these needs special configuration and has common I/O targets. Therefore it is, especially for large computing centers, profitable to fragment their operation.

Layout of their sun-spaces is dictated by two main features: first they must provide information for the Sun Absolute and second they control (at least in part) their lower-level resources. Again, several queens provide the required services:

1. *Local configuration* queen holds configuration information for the RC in question (this Saturn and its moons).

2. *Reporter* queen collects the monitoring information and stores it the format of GLED objects. It also creates summaries of these information and makes them available to Sun Absolute or any other interested party.

3. *Monitoring* queens perform specialized tasks of actively monitoring the resources. They can be wrappers for other services or implemented in GLED itself. Information is pushed into reporter queen for further processing.

4. *Active* queens are meant as the entry point for initiating collaboration-wide data manipulation and transfer services. They are meant for users with administrative privileges.

2.2 User's perspective

Outline only

Connection via local master. On request can gain all information from the (non mandatory) queens at the Sun Absolute and thus gain insight into status of RCs.

For further information requests can be sent directly to RCs.

In case of need a user can connect directly to the Saturn of the RC and perform the desired tasks on the spot. This, of course, implies the creation of second level Saturn on the user's desktop machine.

Or any other mirroring option.

3 Gled in a local computing environment

Outline only

Set-up of small computing environment (e.g. HEP department of JSI): desktop machines as computational nodes, a few servers and some pure computational nodes

Local master: non-aggressive cluster monitor ... just provides availability of CPU power

per-collaboration information servers: *queens* ... software status, file catalogues, job options, available n-tuples

job submission & monitoring: Delegation of Saturns for processing clusters of jobs with common I/O. Prepares data sources and sinks, then claims available nodes. Post processing, packing of results.

analysis: a user can spawn his own GLED cluster and attach computational nodes. Also ... other users can connect to his Saturn and participate in the analysis or its preparation.

4 Conclusion & prospects

Not yet written