Persistency framework manages LCG databases

Managing data in a distributed and heterogeneous Grid environment is one of the most challenging tasks of Large Hadron Collider (LHC) computing, in terms of both developing the required software and deploying the underlying services.

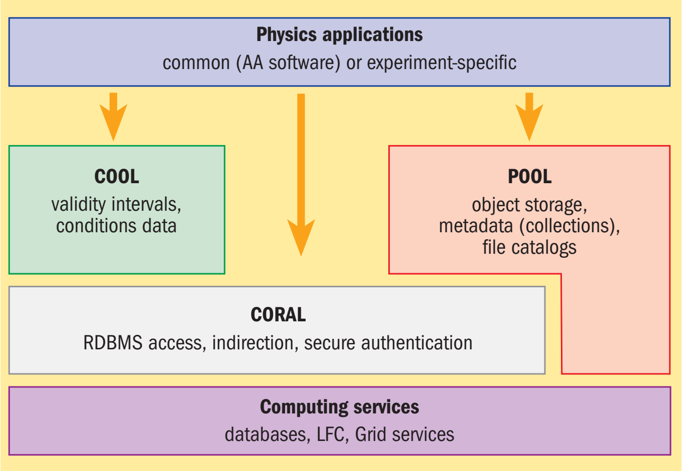
Relational database management systems (databases) play a central role, especially for conditions and event metadata, because they can provide consistent storage to many concurrent users.

Given the complexity of today's database systems, it is often difficult for users to exploit the underlying services in the most efficient way. However, many of the required optimizations can be delegated to an intermediate software layer if both the main physics use cases and service constraints are taken into account early on in its design.

LCG persistency framework

The persistency framework for the LHC Computing Grid (LCG), which is being developed jointly between the LHC experiments and the IT/PSS group, aims to provide such a software layer. Its purpose is to decouple the user code from the features of any particular database implementation.

The project started in 2002 in the LCG Applications Area, driven by the requirements of its users (ATLAS, CMS and LHCb experiments). Project priorities are set with the experiment representatives in the LCG Architects Forum, and experiment developers contribute actively to the software implementation. Development is also tightly coupled to service constraints, as the software is developed in close contact with the IT/PSS physics database service team and the LCG Distributed Deployment of Databases (3D) project (led by IT/PSS).



The LCG persistency framework: layering of software components.

The persistency framework project focused initially on the development of POOL, a hybrid store based on object streaming into ROOT files and metadata storage into databases. More recently the scope of the project was extended to provide a generic database access layer (CORAL) and a specialized component for storing and looking up conditions data (COOL).

Accessing databases with CORAL

Database access for all persistency framework components proceeds via the CORAL (COmmon Relational Abstraction Layer) package. CORAL is also being used in production by several LHC experiments directly from their offline and online applications.

CORAL provides a set of C++ interfaces that are independent of the database implementation and therefore enable the same code to be used against a variety of database systems. At the moment Oracle, MySQL, SQLite and FroNTier (a Web-based database caching package) are supported via plug-in libraries that can be loaded at application runtime.

Support for several database implementations is important, not only to minimize the risk of technology binding but also to cover the available database deployment infrastructure across LCG sites. The experiment deployment models foresee the use of Oracle at Tier-0 and Tier-1, and the use of MySQL, SQLite or FroNTier at other Tiers. More details will be available in a forthcoming *CNL* article about the LCG 3D project.

To exploit distributed database resources that are becoming available via the LCG 3D project, CORAL provides secure database authentication and indirection, including retrial/failover across multiple database replicas. CORAL resolves user-defined logical database names into physical connections to database servers that are now available. In the event of network or service problems, CORAL connections will failover to the next available database replica, if necessary at a different site.

CORAL implements several database access optimizations directly (such as row prefetching and the efficient use of server-side cursors) and significantly decreases the user effort to implement others (bind variables and bulk DML operations). These single-client optimizations are complemented by a connection pool that minimizes the number of concurrent server connections from larger applications with several database components and improves access to the database.

POOL: object storage in databases

The POOL hybrid store functionality is now well integrated in the software frameworks of ATLAS, CMS and LHCb, and has been successfully tested in several large-scale data challenges using object storage in ROOT files.

Building on CORAL, POOL was recently generalized to store arbitrary C++ objects in any of the CORAL-supported database systems. This is particularly useful for calibration and configuration data, which cannot easily be managed in files. With this mechanism objects are decomposed according to their C++ type and stored as rows in relational tables. A set of customizable mapping rules enables the user to steer the automated table generation and to control the mapping of C++ data types to their relational counterpart.

Object-relational storage is now being integrated into the software frameworks of ATLAS and CMS.

COOL handles conditions databases

The COOL (LCG Conditions Database) package provides a software infrastructure for managing conditions data, focusing on the issue of their time variation and versioning.

The development of COOL began at the end of 2004 to replace several disjointed packages previously developed for MySQL and Oracle. COOL still shares their basic data model for conditions data but is now based on a single code implementation and the same relational schema for all supported back-ends, thanks to the use of CORAL.

In COOL, measured or calculated detector conditions (such as detector temperatures and alignment parameters) are associated with an interval of validity, the time range to which the stored conditions apply. Groups of similar condition items can be organized in a hierarchical structure similar to a file system. Multiple versions of condition data can be maintained (for example, originating from alternative alignment methods) and can be referred to by tag names (similar to release tags in the CVS code management system).

The COOL software provides a high-level C++ interface to store and retrieve the data according to the most important physics use cases. It takes over most of the physical management of database tables and the creation of indices for fast data access, enabling users to focus on the definition of the experiment conditions and their logical structure rather than on database access optimization. COOL enables users to store data either directly inside the database tables or to maintain references to data stored externally (such as XML or POOL files or other databases), depending on data volume and the experiment deployment model.

COOL is today the baseline conditions data implementation for the ATLAS and LHCb experiments. Although COOL is still being optimized, its performance already matches some of the experiment requirements. Sustained data rates over 20 MB/s and 20 k rows/s have been observed for retrieval from an Oracle RAC cluster database.

Summary

With the introduction of the CORAL and COOL packages alongside POOL, the LCG persistency framework is now providing storage functionality for all major physics data types as a consistent set of layered components. Designing these components in close contact with the experiments and the database service providers at CERN and other LCG tier sites will make it possible to successfully deploy both software and services for the start-up of the LHC.

Further information

- POOL (the LCG Persistency Framework): http://pool.cern.ch;
- CORAL (COmmon Relational Abstraction Layer): http://pool.cern.ch/coral;
- COOL (LCG Conditions Database): http://cern.ch/cool.

Radovan Chytracek, Dirk Düllmann, Giacomo Govi, Ioannis Papadopoulos and Andrea Valassi, IT/PSS, CERN

The Grid: revolution or evolution?

Why do the planets revolve around the Sun? Has genetic science shaken Darwin's theories to their foundations? Are viruses the champions of evolution? Is progress a form of tradition? On 8 and 9 July, Geneva's Science History Museum held science nights on the theme of "Evolution, Revolution".

CERN took part in the event, anticipating the (r)evolutions from the Large Hadron Collider (LHC). The future accelerator has already led to technological breakthroughs, it promises to deliver more scientific advances, and may even turn our understanding of the infinitesimally small on its head.

The CERN stand put a special emphasis on the computing revolution, from the Web to the Grid. Visitors were able to learn about the Grid in general, about Grid projects at CERN (EGEE, LCG and openlab) and about activities related to distributed computing such as Africa@home. Computer animations explained the technologies, which have spin-offs well beyond the field of particle physics that are of benefit to the whole of society.

Some 30 000 visitors attended the science nights, according to the organizers. Many people flocked to the stands and animations until late on Saturday night and all Sunday afternoon. The CERN stand received much interest and the 15 volunteers from the IT department and the Visits Service were kept busy answering questions. These ranged from the general "How does this work?" to more technical questions like "On what platforms does the Grid run?", and of course "How can I access this?" To make it more hands on, a computer game for children was available in which they could try to save the world by using the Grid to decrypt a message from outer space.

Adults and children visited the CERN stand at Geneva's science nights.

So, is the Grid a revolution or an evolution? Historically the Grid can be seen as the latest step in the process of developing distributed computing solutions, an evolution that started many years ago. The switch from mainframe computers to clusters of PCs more than a decade ago is one step in this direction. The significance of the Grid today comes from its ambition to extend distributed computing to a global scale – now possible thanks to high-speed networks – and thus to change the way scientists work together.

In fact, every technological revolution conceals a great deal of evolution. The World Wide Web was itself based on other technologies, such as hypertext, and built on prior efforts to unify electronic information. The same is surely true for the Grid and distributed computing.

The IT communication team, CERN

The DIANE user-scheduler provides quality of service

Today the mainstream use of Grids resembles a large batch system: the goal is to maximize the computational throughput over long periods of time.

This fits many applications, in particular large data productions of the Large Hadron Collider (LHC) experiments, where the production manager puts thousands of jobs into the system and after several days they come out with the result. However, this model does not support other scenarios well. For example, in an interactive analysis the response of the system should be much faster and aligned with the activity of the user; and life-science applications often involve short-deadline jobs, that is many short jobs that must finish within a certain time limit. In general, such quality of service (QoS) characteristics are not present in today's Grid systems.

The scale and complexity of the Grid also has implications. The LHC Computing Grid/Enabling Grids for E-sciencE (LCG/EGEE) is the world's largest Grid system to date, comprising more than 20 000 worker nodes, some 200 computing sites and petabytes of storage. Such an impressive enterprise, which connects heterogeneous computing environments and organizations, comes at a cost: from the end-user's perspective, tracking problems can be time consuming and the system may sometimes be less efficient.

The DIANE project

User-level scheduling is a light software technique that enables new capabilities to be added and QoS characteristics and reliability to be improved, on top of the existing Grid middleware and infrastructure.

DIANE (DIstributed ANalysis Environment) is such a tool. It is an R&D project for parallel scientific applications in the master–worker model that was started at CERN in 2001. At the beginning the target was to investigate distributed ntuple analysis for particle physics. However, with time DIANE has become an application-independent user-scheduling tool on the Grid (see http://cern.ch/diane). It has been interfaced to a number of applications in high-energy physics, medical physics, life sciences and other fields.

DIANE is a python framework based on a master–worker processing model that is used on top of regular Grid middleware in a transparent way. Worker agents are sent to the Grid as regular Grid jobs. By opening a TCP/IP connection they register to the master agent that runs on the user's desktop computer and is the coordination point for the virtual worker pool. Workers may dynamically join and leave the pool, without disrupting the processing as a whole. The units of computation are many short tasks, which the master allocates to workers directly, bypassing the middleware scheduling layer.

This makes it possible to reduce the total job turn-around time and to react much faster to errors in task execution by reallocating them to other workers. Splitting the processing into many fine-grained tasks improves the load balancing and ensures that the workers are used efficiently. As a result the computing resources may be returned to the Grid faster, because worker agents are automatically terminated when the processing reaches its end.

DIANE's python framework enables existing applications to be integrated quickly, even those as complex as Athena, the analysis framework of the ATLAS experiment. Studies performed by members of the ATLAS collaboration showed that DIANE can be used to integrate local and Grid resources, and even resources from different Grid infrastructures at the same time. The DIANE-based parallel Athena prototype has been shown at EGEE conferences and has been included in the ATLAS Technical Design Report (TDR 2005).

DIANE has also been interfaced to Ganga, a user-friendly Grid interface created in the context of ATLAS and LHCb experiments at CERN. In future, physicists using Ganga will be able to choose the DIANE optimizer, which will be attached transparently to their jobs.

The DIANE scheduler will also be used to operate the statistical regression testing part of the Geant4 release validation procedure. It enables turnaround time to be reduced and provides a more stable and predictable job output rate. This is because the worker agents acquired at the beginning of processing are held inside the pool and are shielded from the instabilities in the Grid brokering. Stable job output rate is an important QoS feature because it enables testing operations on the Grid to be planned with more reliability.

Practical applications

Earlier this year DIANE was used to perform a sizeable fraction of an *in silico* drug discovery application using the EGEE and other Grid infrastructures. The challenge was to analyse possible drug components against the avian flu virus H5N1.

This activity showed that a user-level scheduler like DIANE can improve the distribution efficiency on the Grid from below 40% to above 80% by optimizing the allocation of the fine-grained computing tasks. Automatic error-recovery mechanisms proved to be efficient in extended periods of continuous work: the part performed with DIANE lasted around 30 days.

In May and June, CERN successfully supported a series of large-scale data-processing activities carried out by the International Telecommunications Union (ITU) as part of the ITU's Regional Radiocommunication Conference. Several sites of the EGEE infrastructure provided a computing Grid of more than 400 PCs to work on each analysis in parallel, and the processing was conducted using the DIANE scheduling layer.

The system completed more than 200 000 very short frequency analysis jobs (clustered in around 40 000 processing tasks) in around one hour, proving that on-demand computing with a short deadline is possible on the Grid. The frequency allocation plan that was optimized with the help of the Grid enabled more than 1000 delegates from 104 countries to adopt the treaty agreement that will replace the analogue broadcasting plans that have existed since 1961 for Europe and since 1989 for Africa.

In the future, closer integration with Ganga will enable access to all of DIANE's capabilities. Ongoing PhD research is aimed at supporting hard QoS requirements with novel techniques such as a floating worker pool, extending scalability above 500 worker agents, and supporting inter-dependent tasks for workflow applications.

Jakub Moscicki, IT/PSS, CERN

LCG launches online bulletin

The *LCG Bulletin* was launched at the beginning of the summer. It is available at https://cern.ch/twiki/bin/view/LCG/LcgBulletins.

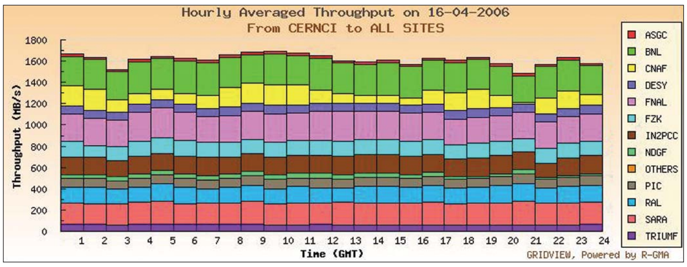
The aim of the bulletin is to streamline the distribution of practical information for the LCG community, in the form of short summaries. If you wish to publish information in the bulletin, such as news or future events involving your area of the LCG project, mail this to me with the relevant links.

Alberto Aimar, IT/LCG, CERN

BARC collaborates with LCG

Situated on the shores of the Arabian Sea, on the outskirts of Mumbai in India, lies the Bhabha Atomic Research Centre (BARC). The research centre is an Indian Government facility that employs some 10 000 people, and that has been contributing human resources to the LCG project for the past 21 months. I was one of several members of a CERN contingent from the LCG project who were welcomed to BARC during the CHEP'06 conference in Mumbai.

A remarkable range of projects was presented during the visit. In the area of security, a range of solutions for intrusion detection and incident analysis were discussed, along with a hand-scanning biometric system. This is especially important for BARC, which maintains several isolated networks. Another interesting project was a fast, low-cost 2 million pixel display with 16 separate panels driven by a cluster of machines. This display is used for applications such as satellite imagery, CAD/CAM, medical analysis and tsunami simulation.



GridView monitor of data transfer rates from CERN to Tier-1s in April, showing peaks above 1.6 GB/s. The monitor was developed by BARC.

Areas where BARC has been contributing expertise to the LCG project include the Extremely Large Fabric management system (ELFms), as well as Grid-monitoring technology. As Les Robertson, the LCG project leader, commented during the visit, "the collaboration enables LCG to benefit from BARC's long experience with high-performance computing".

BARC has been instrumental in the development of GridView to support the Grid-level monitoring for the service challenges (http://gridview.cern.ch/GRIDVIEW). BARC has also contributed to LEMON, the lower-level LHC-Era MONitoring component of ELFms, which is being used to monitor 80 metrics across approximately 2500 hosts in 100 clusters. Some 80 automatic recovery actions have now been defined, resulting in a reduction in problem tickets and the associated reduction in human intervention.

CCTracker is a tool used to visualize and plan the physical layout of computer centres such as the LCG Tier-0, and to search for equipment. BARC plans to develop it further to perform high-level service management use cases across sets of nodes and clusters.

Speaking at the CHEP'06 conference, the president of India, Dr A P J Abdul Kalam, referred to the collaboration with CERN in areas such as Grid monitoring and fabric management as a model for how India can contribute to the LHC project.

In addition to the advances in the quality and scalability of fabric-management components, the exchange of ideas, experiences and solutions between BARC and the LCG project has proved beneficial to both parties. BARC is planning to adopt several ELFms components for their operations, including CCTracker and LEMON, and the LCG project is benefiting from the additional skilled resources at this critical time.

Bill Tomlin, IT/FIO, CERN

IN2P3 Computing Centre prepares for the LHC

The IN2P3 Computing Centre (CC-IN2P3), located in Lyon, is the national facility for data storage and processing of the French National Institute of Nuclear Physics and Particle Physics (IN2P3).

Funded by the National Centre for Scientific Research (CNRS) and the Atomic Energy Commission (CEA/DSM/Dapnia), it has provided computing services for more than two decades to several experiments in the fields of nuclear physics, particle physics and astroparticle physics.

Experiments of the Large Electron Positron (LEP) collider, and more recently DZero and Babar, are examples of major users of the centre. Since the early 2000s it has also provided computing services to research institutions in the field of biomedical applications.

CC-IN2P3 has set up and operates a Tier-1 centre to process data from the four Large Hadron Collider (LHC) experiments. It is planned to eventually contribute about 9% of the total worldwide Tier-1 computing capacity for ALICE, 13% for ATLAS, 10% for CMS and 27% for LHCb.

Becoming an LHC Computing Grid (LCG) Tier-1 centre means many changes to the site. A major upgrade of the cooling and power infrastructure of the centre's computer room is scheduled for the second half of 2006. This will enable it to host the data-processing equipment that is essential for the LHC experiments and for other scientific experiments that the centre will continue to support in the coming years. A second machine room in a new building is planned to extend the site's capacity by 2009.

A dedicated optical network circuit that links CC-IN2P3 and CERN at 10 Gbit/s has been in operation since early 2006. It is being used to validate the data-exchange infrastructure for LHC experiments in the context of the LCG project. Data transfer rates of 250 MB/s have been demonstrated between CERN and CC-IN2P3. This circuit, along with the site's links to the national and international networks, are operated by RENATER, the French telecoms network for research and education.

The IN2P3 Computing Centre operates a Tier-1 site for the LCG.

Work has also begun to upgrade the disk- and tape-based data storage and the computing infrastructure of the centre to the levels required for the LCG project. The new Grid components at the site are being integrated progressively into the production-level procedures, with the aim of reaching a high quality of continuous service by the time LHC starts.

CC-IN2P3 has been actively involved in Grid activities since Datagrid, the first European-level Grid project. It also contributes to Enabling Grids for E-sciencE (EGEE), both as a regional operations centre for France and as the developer and operator of the EGEE central daily operations portal.

The LCG, like other large-scale projects, is both a major technological push and a significant human enterprise. In the year of the 20th anniversary since CC-IN2P3 moved from Paris to Lyons, its staff are working hard to meet the challenges presented by the LHC over the coming years.

Fabio Hernandez, CC-IN2P3