CERN COMPUTER NEWSLETTER

Volume XL, issue 3 June–August 2005

Contents

Editors Nicole Crémel and Hannelore Hämmerle, CERN IT Department, 1211 Geneva 23, Switzerland. E-mail: cnl.editor@cern.ch. Fax: +41 (22) 7677155. Web: cerncourier.com (link CNL).

Advisory board Wolfgang von Rüden (head of IT Department), François Grey (IT Communication team leader), Christine Sutton (*CERN Courier* editor), Tim Smith (group leader, User and Document Services).

Produced for CERN by Institute of Physics Publishing
Dirac House, Temple Back, Bristol BS1 6BE, UK. Tel: +44 (0)117 929 7481. E-mail: chris.thomas@iop.org. Fax: +44 (0)117 920 0733. Web: iop.org.

Published by CERN IT Department

LCG takes on new challenges

This issue of CNL features a new section, LCG News. This will be accessible both as part of CNL and also from the LCG website as a standalone newsletter for the LCG community. Les Robertson, leader of the LCG project, talks to CNL about the challenges the project faces.

The success of the second service challenge in April – where a continuous data flow of on average 600 MB/s was sustained for 10 days from CERN to seven sites in Europe and the US – was widely reported. What is in store for the future?
The third service challenge, after a set-up phase before the summer, is due to start in September. This will involve not only the major Tier-1 centres, but also a number of Tier-2 centres – about 20 sites altogether. And the challenge will last until the end of the year. It will test not just data transfer but all aspects of a real Grid service, with a view to measuring the reliability and availability of the distributed resources. So this is a big step forward towards full Grid operations. A working group is already trying to define what basic services will be offered. The challenge here is not just technical, it is about getting people with diverse cultures and different agendas to work together. Last year during the data challenges we ran with the experiments, there were a dozen or so people actually active on the Grid for each experiment. This year, it could be in the hundreds. So we are counting on the experiments, with the help of the ARDA project, to get their analysis running.

How will these initial Grid operations impact on CERN? Will they clog up the Helpdesk?
That's a good question. Problems encountered by physicists in the experiments are best answered by people working closely with each experiment. There are people dedicated to helping each experiment in the Grid Deployment group, who will provide expert support. But ideally, the experiments should take over the responsibility for supporting their own users. One thing we have to realize is that the expert knowledge, like the resources, is distributed over the Grid. And local helpdesks will deal with local problems. The community still has a lot to learn about how to use such a distributed support effectively. But, no, the IT Helpdesk should not be noticeably affected by the service challenges.

How is work on the physical infrastructure for the CERN Tier-0 centre progressing?
Visitors to the Computer Centre can see the rewiring going on there, to prepare it to the same standards as in the basement. This should be completed soon. Also in the second half of the year, a new LAN (local-area network) structure will be deployed on the site. Planning for the WAN (wide-area network) is ongoing and is being coordinated with the Tier-1 centres, National Research Networks and GEANT. It's a big success just getting all these actors together and we are optimistic that this will result in a light path between CERN and most of the Tier-1 centres.

What is the status of the software for the LCG project?

The middleware stack gLite, developed in the EGEE project in close collaboration with the major US Grid middleware projects, will be available soon, and we are keen to test it. At the same time, we have to continue with the pragmatic approach we have had in the past, and use the best of whatever is out there. On the applications side, we are moving rapidly from development to deploying and consolidating what we have. A significant step is that the different stakeholders have agreed to merge the ROOT analysis software with SEAL, which supports the experiments' software frameworks. So this will be a challenging task for Pere Mato, the new manager of the LCG applications area, to put into practice (see p5).

The number of sites in LCG has grown much more rapidly than originally anticipated. How do you explain this?
Without doubt, this is a positive effect of the EGEE project, which has provided the seeds to encourage many sites to join. The leveraging effect of EGEE is very helpful in bringing on board some of the smaller sites. EGEE also makes an important contribution to funding the 120-odd people at CERN working on the LCG project, which includes about 90 people in the IT Department and 30 more in the PH Department.

And how do you see the future beyond 2005?
By September of next year a complete Grid service must be available, because this is when the experiments plan to start taking cosmic-ray data to calibrate their detectors. So we need a complete data-taking and analysis chain in place by then. That is a huge task. One point that is not widely realized is that once the Grid starts working in this way, it will never be able to stop. Unlike an individual computer centre, which can plan a Christmas shutdown, we are operating in a multi-institutional environment where there is practically no chance of defining a common shutdown period. This is surely going to have profound effects on operations at CERN and at the Tier-1 centres. It will be an exciting period!

Computing featured in this month's CERN Courier

The articles listed below have been published in the June 2005 issue of *CERN Courier*. Full-text articles can also be found on the *CERN Courier* website at www.cerncourier.com, together with the rest of the issue's contents.

Computing News

- **Inverted CERN School of Computing transforms students into teachers**
CERN shows that knowledge transfer can work both ways.

- **LHC Grid tackles multiple service challenges**
High-speed data transfer between sites in Europe and the US tests global computing infrastructure for the LHC.

- **LHC Grid accounting package clocks up 1 million job records**
In the three months since the release of the APEL package, more than 50% of LCG sites have published accounting data.

- **Software achieves breakthrough in data challenge**
IBM's storage virtualization software shatters records in reading and writing data to disk.

- **First industrial application runs on EGEE project infrastructure**
Geocluster enables researchers to process seismic data and to explore the composition of the Earth's layers.

- **Global Grid gets an Asian dimension**
GGF13 in South Korea shows the extent of Grid computing in Asia and reaches out to possible economic partners.

- **DO's data-processing record**
Six countries provide computing power to reprocess stored experimental data.

- **The world's first home PC is here!**
Popular Electronics magazine heralds the arrival of the world's first PC 30 years ago.

IT calendar

Feature article

- **Industrial solutions find a place at CERN**
David Myers and Wayne Salter describe the increasing use of commercial solutions in control systems at CERN.

PLUS accounts replace TAPES

Since March 2005 we have been phasing out the use of TAPES accounts in Computer Centre services. PLUS accounts are used instead.

Historically, a TAPES account was necessary to access data written to tapes. Over time, almost all users with TAPES accounts also obtained PLUS accounts. These have now also been created for the few users who only had a TAPES account, and the central tape services have been migrated from TAPES to PLUS. This means that it is no longer necessary to create TAPES accounts for users who need to access Castor data.

In the near future (once we are certain that all the tools work successfully) all TAPES accounts will be deleted and will no longer be part of the account registration procedure.

Jan van Eldik and Harry Renshall, IT/FIO

Use of VPN access puts CERN's security at risk

Recently we have had incidents caused by people opening VPN connections on home computers and (unknown to them) spreading malicious software into CERN. VPN access to CERN should only be used for extreme and rare cases, and users are formally discouraged against using it as a general solution.

Users should also be aware that the availability of the VPN service may be discontinued in the future for security reasons. Some recommended methods for accessing CERN from the Internet are listed on the right.

VPN connections give access directly inside the CERN firewall. Your computer could therefore introduce viruses, worms or backdoor attacks against which the CERN site is normally protected. Similarly, a discovered password can give an attacker access inside the CERN firewall, putting the whole site at risk. In addition, personal software, such as P2P applications, can become visible on the Internet via CERN's infrastructure, requiring care to ensure compliance with CERN's Computing Rules. For these reasons we ask you to avoid using VPN.

The encryption capability of VPN, which allows you to connect to the CERN network without the data transmitted being visible to snoopers on the Internet, is also available in the recommended alternatives described on the right. Please use the methods we suggest. They are also described at http://cern.ch/security/vpn.

A detailed website about remote access to the CERN network is available at http://cern.ch/ras.

Computer Security Team, IT/DI

Recommended methods of connecting to CERN from the Internet

E-mail
To access your CERN mailbox use the Web-based client, Outlook Web Access (OWA), or configure your mail client to use IMAPS (IMAP over SSL). Windows users who have migrated to Exchange/Outlook 2003 can configure Outlook to use RPC over HTTPS. Configuration details for the CERN mail services are at http://mmm.cern.ch.

Internal Web servers
For access to internal CERN Web servers, use Windows Terminal Services (WTS) or open a browser on Lxplus.

Files
For access to the NICE DFS file system use WebDAV (Web Distributed Authoring and Versioning), which provides a Web interface to NICE files and folders (see https://dfs.cern.ch). We recommend that Linux users either use AFS or connect to Lxplus and use SFTP.

Interactive sessions
For an interactive session on Windows (NICE) use Windows Terminal Services (WTS) and for connections to other Windows systems use Remote Desktop from there. For an interactive Linux session use SSH to connect to Lxplus.

SUNDEV service set to come to an end on 1 July

As announced at the Desktop Forum in April 2005 (see the minutes at http://agenda.cern.ch/fullAgenda.php?ida=a052234) the SUNDEV service will be stopped on 1 July 2005.

This decision is associated with the reduction of scope of the Solaris support service because Solaris is no longer considered as an LHC physics platform.

Solaris 9 has just been made available with the Quattor automated system-management framework. Solaris 9 thus becomes the recommended system level at CERN. Please send mail to Solaris.Support@cern.ch to request Solaris 9 installations. Again, as it is no longer considered as an LHC physics platform, there is no formal certification process for Sun Sparc Solaris.

Solaris support will reduce the scope of its services, concentrating on the installation server and automated system administration with Quattor. All other questions concerning Solaris will be directed to the Desktop support contract or directly to Sun.

CS group has proposed to stop all non-routed protocols by the end of 2005. Solaris 7 and below will no longer be installable by the network. Half of the Solaris machines on the CERN site are running Solaris 7 and below – many of them cannot receive the update to Solaris 9 because they are not powerful enough, or because their hard drive is too small. Users are instead advised to consider moving to a PC or to a larger machine.

Stopping the SUNDEV service will also affect Unix users of FrameMaker, since FrameMaker is not available under Linux (Lxplus users currently launch FrameMaker on a SUNDEV server via ssh). Users are therefore strongly encouraged to move to Windows or Macintosh when using FrameMaker, and it will no longer be supported on Unix. If you have any questions about this matter, please contact the SDT (Software Development Tools) team at SDT@cern.ch.

Ignacio Reguero, IT/DES

Concurrent batch jobs get lower limit

ATLAS and LHCb have requested a lower limit on the concurrently running jobs of non-privileged users in the 1nw (one normalized week) batch queue. In the current system some users submit a series of 1nw jobs and when the shorter queues empty, such as at weekends, sets of 20 jobs will start and use up the groups' priority for many days following.

The mechanism is in place to make named accounts or user groups immune to this limit and give them their own tailored limits. The immune accounts can be seen by the command "bugroup u_special" and repeat command bugroup on the subgroups from the resulting output (e.g. "bugroup xu_prod"). It was ultimately decided to change this limit from 20 to 10.

Note that the service is typically running 2000–2500 jobs at a time and you can see the queue mixture at http://ccs003d.cern.ch/lsf.

Ordinary users (not members of the immune u_special group) cannot, of course, submit to the "prod" queues. This change will be a trade-off between potentially having idle cycles and having better throughput for shorter jobs. There will clearly be a period when users requiring a higher limit will have to be properly identified by those responsible for the experiment. Requests to add or subtract users from u_special must be sent to lsf.support@cern.ch.

Harry Renshall, IT/FIO

The maximum number of concurrent jobs for an an ordinary user

Per queue	Per user	Slots
8nm	all	120
1nh	all	120
8nh	all	100
cmsprs	all	75
1nd	(all ~u_special)	150
2nd	(all ~u_special)	75
1nw	(all ~u_special)	10
prod100	all	100
prod200	all	250
prod400	all	1000

LHC Grid accounting package clocks up 1 million job records

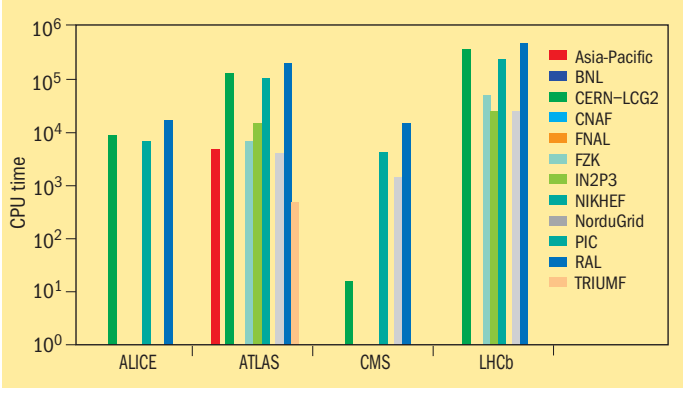


Fig. 1: the total CPU time provided by Tier-1 computing centres supporting LHC VOs during the latter part of 2004. Accounting data are gathered from sites running the LCG middleware suite, which currently excludes a few sites (such as Brookhaven, Fermilab and NorduGrid).

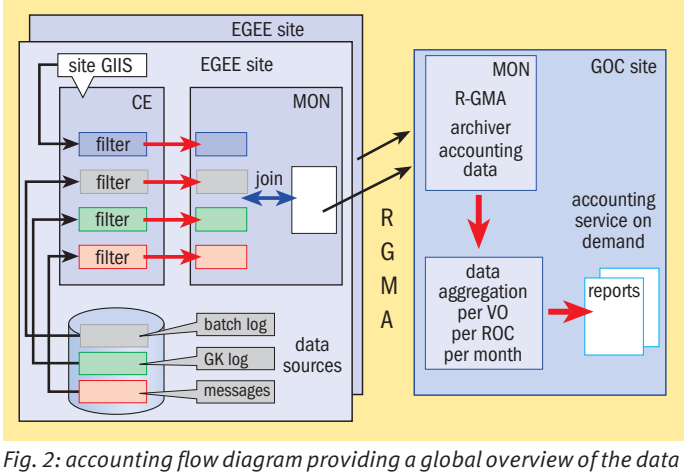


Fig. 2: accounting flow diagram providing a global overview of the data collection process (APEL) and the Web reporting service (GOC).

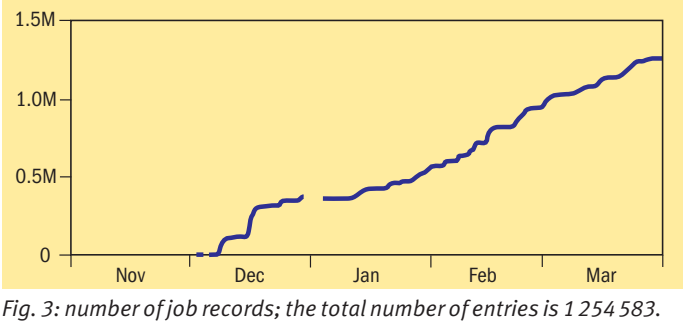


Fig. 3: number of job records; the total number of entries is 1 254 583.

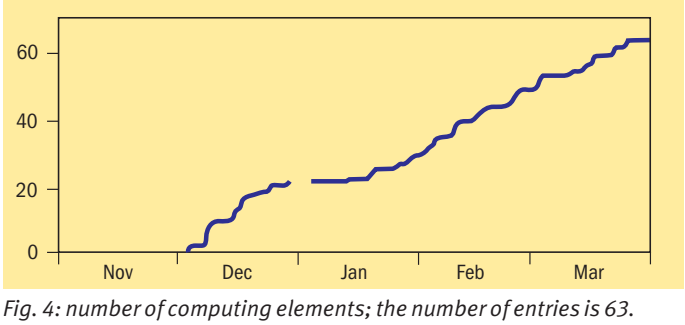


Fig. 4: number of computing elements; the number of entries is 63.

In the three months that followed the December release of the Accounting Processor for Event Logs (APEL) package in LCG-2 middleware, more than 50% of sites published accounting data, comprising over 1 million job records, to the Grid Operations Centre (GOC).

Accounting and the LCG Grid

In the LCG Grid environment, the computing resources, the application data and the Grid users belonging to Virtual Organizations (VOs) are distributed. Jobs submitted by these users may be sent to computing resources close to the data or may go to remote resources with available job slots, thus reducing queue times.

As a consequence, jobs that run on LCG resources must be properly accounted for. The consumption of resources (CPU time, wall-clock time and memory) by VOs, as well as the resources provided by sites as a function of time, can then be determined.

Grid Accounting has two main pricing models. In "Grid Job Accounting" a job-usage record provides the complete description of resource consumption. "Real-time Grid Accounting" is based on an incremental determination of resource value while the job is executed. In the former, the cost of computing is determined on the basis of what was done (after job execution), while in the latter, data feeds into a pricing model which determines the cost for consumption in real time.

An implementation of real-time accounting (DataGrid Accounting System) is under development within the gLite middleware framework of the EGEE Grid computing project, and will be the topic of a future article.

In LCG, the accounting records of each site are consolidated using Relational Grid Monitoring Architecture (R-GMA). They are aggregated in different ways to provide high-level views of the data for presentation on the Web. These data are grouped to provide information such as the CPU usage delivered by computing resources for each member country, or the CPU usage consumed by each of the VOs for each of the Tier-1 resource centres (figure 1).

Data collection mechanism

The collection of accounting usage records is done through R-GMA, an implementation of the Grid Monitoring Architecture (GMA) proposed by the Global Grid Forum (GGF). GMA models the information infrastructure of a Grid as a set of consumers (which request information), producers (which provide information) and a registry which mediates the communication between producers and consumers.

In R-GMA, the producers are the sites that contain a database of local accounting records for successful jobs. The consumer is a GOC that archives accounting records across all sites and provides a Web interface to view the data.

APEL action

APEL is a program that builds daily accounting records, based on information located in the log files of a computing element (CE). When installed on the CE, it will process log files provided by the batch farm (Portable Batch System and Load Share Facility)

and the gatekeeper, and then publish the data into a MySQL database local to the site. The data are assembled to form accounting usage records that identify the Grid user (through the unique Grid Distinguished Name), the VO and the resources used to execute the job (figure 2). Each record is unique as there is only one record per Grid job.

The successful implementation of APEL in the LCG project paves the way for the next major step in its development. It will be important in the future to also provide accounting for storage usage (disk and tape) on the Grid.

Further reading

- **LHC Computing Grid** http://lcg.web.cern.ch/LCG/
- **DGAS** http://egee-jra1-wm.mi.infn.it/egee-jra1-wm/org.glite.dgas.shtml
- **gLite** http://glite.web.cern.ch/glite/
- **EGEE** http://public.eu-egee.org/
- **R-GMA** www.r-gma.org/index.html
- **GOC Accounting website** http://goc.grid-support.ac.uk/gridsite/accounting/

Rob Byrom and Dave Kant, Rutherford Appleton Laboratory, UK

Dave Kant at Rutherford Appleton Laboratory's computing centre.

Mato becomes new manager of LCG Applications

The LCG project is often described in terms of petabytes and processors. But the numbers are just the foreground of the hard work done by groups of dedicated people. This is the first in a series of profiles that introduce some people in this project to the wider community.

To ensure a smooth transition from the software development stage to its maintenance, on 1 March this year Pere Mato took on the job of leading the LCG Applications area involving about 45 people. The Applications area develops and maintains that part of the physics applications software and associated infrastructure shared among the LHC experiments.

Mato will be responsible mainly for four projects: software process and infrastructure (SPI), persistency framework (POOL and conditions database), core libraries and services (SEAL), and the simulation project. He plans to reduce duplication in the current code, thus making the software easier to maintain. "We have to take this occasion to re-engineer some parts of the code for longevity while keeping the functionality," he said.

Mato: ready for the challenge.

The choice of the 45-year-old Catalan was based not only on his expertise in the applications field, but also on the management experience he acquired while working on the LHCb project, according to Les Robertson, LCG project leader. Mato was the key person in convincing the LHCb experiment to migrate from the then much trusted Fortran software to C++, a process which has lasted almost six years.

In his new job, Mato will follow the merging of the ROOT software framework with the SEAL software components he developed while working in LHCb. He hopes to develop a structured process to facilitate the maintenance of the LCG software when its development is largely finished.

Mato came to CERN in 1986 as a student from Barcelona University, where he majored in theoretical physics. He then started a classic CERN career, joining the LEP experiment ALEPH, where he stayed for 13 years before moving to the LHCb collaboration in 1998.

The nature of the LCG applications area is not new to Mato. It resembles large experiments he has participated in before, where compromises have to be continually brokered in order to move forward. "Nothing is black or white, the world is made up of greys," he said. "The art is to choose the right grey."

Leticia Martignon, IT/DI, CERN

Tier-2: focus on Czech computing centre 'Golias'

The Czech LCG Tier-2 computing centre "Golias" in Prague is one of the LCG's mid-sized sites with just over 200 processors. However, its contributions to the data challenges last year were significant. Golias delivered 6% of the jobs submitted by the ATLAS experiment and 7% of the ALICE jobs. Only much larger computing centres delivered a greater percentage of the hundreds of thousands of jobs submitted during the data challenges.

The Golias site, located at the Institute of Physics of the Academy of Sciences of the Czech Republic, is entirely dedicated to the HEP community's needs. It provides roughly half of its computing power for ATLAS and half for the D0 experiment at Fermilab. A small fraction of the capacity is used for the ALICE experiment.

The site will stock up its computing centre and increase 10 times the computing power to prepare for the LHC challenge in 2007. The centre's storage capacity currently amounts to 40 TB. Partially, the storage space is also used for the second Czech EGEE computing centre "Skurut", run by CESNET, the Czech network provider.

The collaboration between LCG and Golias dates back to 2002, when the centre first established contact with the new Grid project for LHC computing. The site therefore got involved with the computing project even before the Memorandum of Understanding, which officially integrated the site into the Grid, was signed in 2003.

The centre has been funded through the Academy of Sciences of the Czech Republic. "Unfortunately, there is no direct governmental funding for Grid computing as in other countries," explained Milos Lokajicek, who is responsible for the five dedicated Grid staff at the centre.

Lokajicek leads the team at Golias, which is made up of two physicists, who make sure that the software for ATLAS and the other experiments runs smoothly on the nodes, and of three computing specialists, who mainly work on giving support for the operating system, the Grid middleware and the networking.

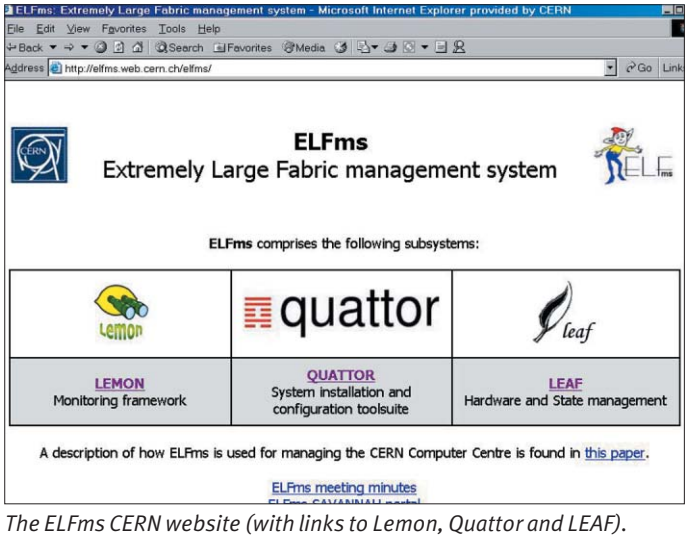
Leticia Martignon, IT/DI, CERN

CERN uses ELFms to manage enormous range of systems

ELFms: Quattor, Lemon and LEAF

With the scale-up needed to serve LHC's Tier-0/1 centre, CERN's Computer Centre will increase from its current 3000 nodes to approximately 8000 in less than three years. At the end of the European DataGrid (EDG) project, remaining developments as well as the maintenance and deployment of Quattor (www.quattor.org) and Lemon (http://cern.ch/lemon) were taken over by the ELFms (Extremely Large Fabric management system) project, a project coordinated by CERN's IT Department with contributions from many HEP institutes. ELFms also accommodates the LEAF (http://cern.ch/leaf) system, developed with support from the UK's GridPP as part of LCG.

The three toolsets have been integrated and form a fully modular, interoperating framework. With ELFms, we manage a large heterogeneous environment. This includes large general-purpose batch and interactive farms, test and certification clusters, disk and tape servers, and database and Web servers. Various generations of hardware co-exist, leading to a multiplicity of set-ups, for example in terms of CPU types, memory and disk sizes. Supported operating systems include RedHat (7.3, RHES2, RHES3), Scientific Linux 3 and Solaris 9.



The ELFms CERN website (with links to Lemon, Quattor and LEAF).

Quattor

In 2003 we decided to migrate the dispersed Computer Centre configuration information into the Quattor Configuration Database (CDB). Configuration data from more than 20 different places (databases, shared files and installation scripts) were identified and appropriate CDB data models and schemata defined. The resulting configuration templates were arranged into hierarchical template-based structures, which matched service and cluster descriptions. As of today CDB holds configuration information (such as networking parameters and physical location) for 95% of all systems in the Computer Centre. For systems managed by Quattor, all configuration details including the precise hardware set-up, list of software packages and running services are stored in CDB. An SQL-based back-end is used for general-purpose access to the database contents. A KickStart generation tool interfacing with CERN's legacy RedHat Linux installation system (the Automated Installation Management System – AIMS) has been written, but might in the future be replaced by Quattor's AII solution (which is itself based on KickStart).

Software deployment and node configuration

In addition to centrally configurable and reproducible installations via CDB-generated KickStart files, nodes should be actively managed during operation in order to maximize availability without needing to reinstall owing to functional or security updates, or service reconfigurations. Quattor addresses this requirement by providing two subsystems, SPMA (Software Package Manager Agent) for software deployment and NCM (Node Configuration Manager) for system configuration.

NCM is a framework where service-specific plug-ins, called "Components", are responsible for making the necessary system changes to bring the node to the desired state as defined in its CDB configuration profile, regenerating or updating local service configuration files and restarting services if needed. Configurations of NCM components are stored in CDB, making full use of its hierarchical inheritance and overwriting mechanisms in order to define service definitions and software environments.

SPMA manages the software packages installed on a node and can handle multiple package formats, including RPM (RedHat Package Manager) and PKG (package system) for Solaris. SPMA's main advantage over other software distribution systems (like apt-get or yum) is the clear separation of node configuration and repository contents: adding a new version of a package to a repository does not imply its deployment on client nodes without an explicit configuration change in CDB. This way, external drifts are avoided and reinstallations of systems result in exactly the same set-up as defined in the SPMA configuration.

It is possible to configure nodes with different package version selections, which is sometimes essential – e.g. because of backward incompatibilities on production services, which would otherwise block a package upgrade required elsewhere. SPMA also allows rollback to older, well known software configurations.

SPMA replaced ASIS (the Application Software Installation Server), the legacy application software distribution system developed and used previously at CERN, in spring 2003. CERN's legacy Unix configuration system, SUE (Standard Unix Environment), has been completely replaced by NCM on newly certified OS versions. Around 90 NCM configuration components have been produced, ranging from basic system configuration up to the complete set-up of LCG Grid services.

Scalability via proxy caching

Uniform deployment of large software updates onto clusters with thousands of nodes requires scalable solutions. By default, Quattor uses HTTP as the transport protocol for RPM packages and XML configuration profiles. A replicated two-level "reverse proxy" server architecture is being deployed. A server cluster, consisting of two back-end servers and a variable number of front-end servers, handles requests for configuration profiles and software packages. A second level of proxy nodes mediates and caches requests between racks of client nodes and the front-end servers, and also provides serial-line-based console access to the rack nodes. This second level avoids wasted network bandwidth caused by multiplication of

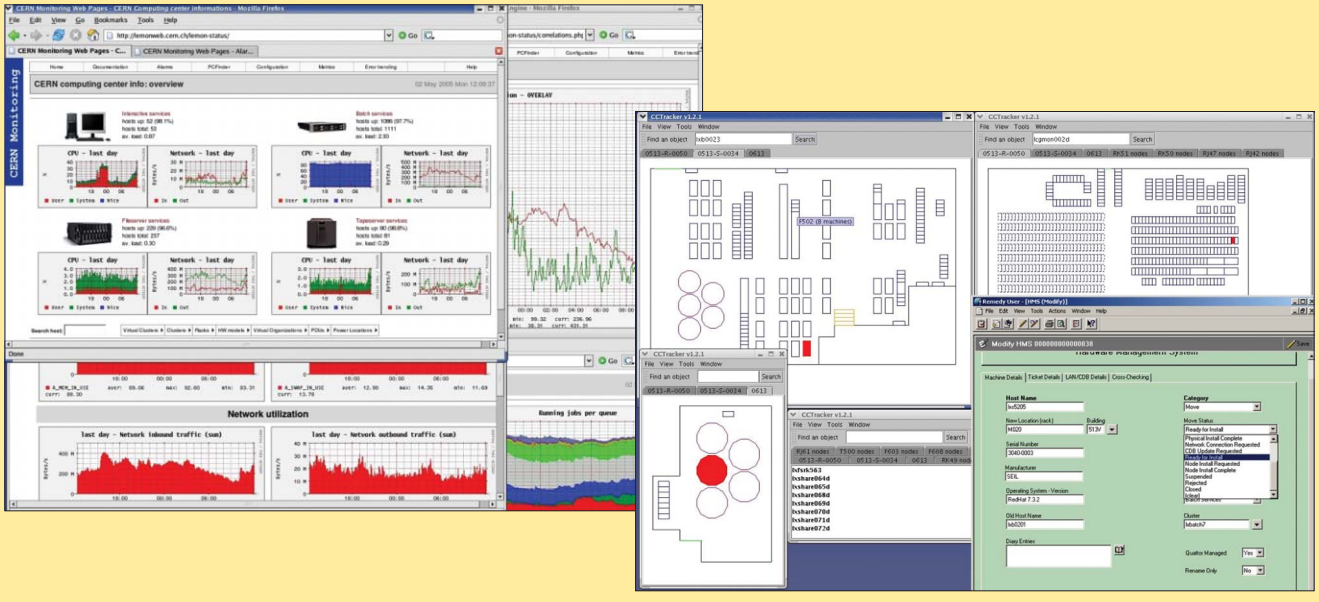


Fig. 1, left: the visualization interface for the monitoring system Lemon. Fig. 2, right: user interface for the LEAF toolset.

identical transfers and it increases reliability by reducing the ratio of clients per server from hundreds or thousands to the number of nodes in a rack.

Grid integration

Integrating fabric resources into Grid environments requires deploying additional software and configurations to cluster and server nodes. Since configuration errors for Grid software may affect not just a site, but also the response and stability of the Grid as a whole, the use of automated tools to avoid error-prone human intervention is particularly desirable. For achieving integration of Quattor with the LCG Grid, NCM configuration components needed to be developed for multiple Grid services and server types. At the same time some 10 Tier-1/Tier-2 centres have started to use Quattor for managing their Grid and/or production clusters.

In order to minimize work duplication and advance towards a uniform and end-to-end Grid fabric management solution, a community taskforce (with the participation of LCG sites including IN2P3/LAL, NIKHEF and INFN/CNAF) has been set up and configuration components and templates have been produced for most of the LCG-2 services. In the CERN Computer Centre, batch nodes have been converted with Quattor into LCG-2 Worker Nodes. With the deployment on other sites, site assumptions and dependencies in the core Quattor software and configuration components have been identified and replaced by more general solutions, and the quality and robustness of the software have consequently been increased.

Quattor status and next steps

Quattor is currently managing around 2200 Linux nodes in the CERN Computer Centre and providing a uniform management solution. SPMA and NCM are used on a daily basis to deploy security, functional and configuration updates. One example of a large deployment (which used to take days) was the upgrade of the CERN batch system from LSF 4 to LSF 5, which took place in less than 15 min on more than 1000 nodes, and without service interruptions.

The inclusion of the EGEE and LCG testbeds into the existing infrastructure, as well as new deployments outside CERN, will imply significant functional growth. For example, a "light" version of Quattor will be released in the near future, containing NCM and configuration templates for configuring Grid services on top of existing installations.

Also, the ramp-up for LHC computing will require optimizing potential bottlenecks. For example, parallelism will be used for XML profile generation inside CDB to reduce recompilation times.

Security is another critical aspect: a prototype CDB fine-grained authorization module will be tested and deployed this summer, as well as an improved mechanism for handling host certificates that will allow encrypted HTTPS to be used for configuration profile transport.

Lemon

Lemon (LHC EDG monitoring) is a client/server-based monitoring system. On every monitored node, a monitoring agent launches, and communicates using a push–pull protocol with sensors that are responsible for retrieving monitoring information. The extracted samples are stored in a local cache and forwarded to a central Measurement Repository. Sensors can collect information on behalf of remote entities like switches or elements of the electrical power distribution systems. The Measurement Repository can interface with a relational database or a flat-file back-end for storing the received samples, and can be queried via an API (Application Program Interface) based on SOAP (Simple Object Access Protocol).

Thanks to the modular agent design, it has been possible to add to the sensors provided by the EDG project (mostly performance-related) new sensors for providing a broader fabric view and accurate exception and inventory information. Some 120 new metrics have been provided, ranging from general information (for hardware information and the status of daemons) over batch system specifics (for LSF) to service specific metrics (e.g. for disk and tape servers, and Oracle databases). In addition to the existing C++ API, a Perl sensor API has been provided to ease the task of writing sensors. Some of the metrics and sensors have been external contributions, e.g. coming from IBM (for StorageTank) or BARC India (LSF quality of service and Grid VO usage). In total, we collect around 150 metrics organized in 70 metric classes and implemented in 12 sensors.

For reliable archiving of monitoring information we adopted and enhanced the "OraMon" Oracle back-end-based instance of the Measurement Repository. OraMon (developed with support from Weizmann Institute and LCG) has been in operation at the CERN Computer Centre since November 2003, collecting around 1.2 GB of data daily from approximately 2000 monitored nodes. ▷

Technical briefs

Fault recovery and derived metrics
In fabric environments with thousands of nodes, the ability to perform automated fault recognition and self-healing recovery becomes paramount to release operators and technicians from repetitive operations (like cleaning up temporary space or restarting daemons). We re-engineered the initial EDG fault-tolerance prototype and now provide two independent modules for global and local scope action. A lightweight self-healing module running on every monitored node can compare cached metric samples to absolute or relative reference values and trigger recovery actions or activate exceptions, which trigger an operator alarm.

For correlations involving several nodes, a Perl plug-in-based framework allows execution of advanced user-defined correlations, runs recovery actions and injects derived metrics back to the Monitoring Repository.

Server redundancy
For enhancing server robustness, a high-availability solution is being developed, providing fail-over recovery at the database and OraMon level. At the database level, an Oracle Streams-based replication between two database servers has been enabled using a master–master scenario.

Visualization
In order to present the monitoring information in an easily understandable way, we have created a graphical Web interface using PHP and RRDTool. Data from the Monitoring Repository (and other data sources such as the Quattor CDB or LSF) are retrieved via a Python application and stored into RRD on a per-node, per-cluster or per-service basis (e.g. power-supply consumption).

Lemon status and next steps
The Lemon agent and the OraMon repository have been in production for three years on Linux nodes, and also on Solaris nodes for the past year. Lemon is also used outside CERN's Computer Centre: it has been interfaced with CalTech's MonALISA system, and the GridICE project uses it for monitoring Grid services. Recently the CMS experiment decided to use Lemon for monitoring their DAQ prototype nodes. Lemon has been integrated into the current operator alarm system at CERN and SURE, and we are working on a gateway into the future LHC controls alarm system, LASER.

LEAF
The successful deployment of Quattor at CERN has provided a platform for some advanced components to be added to the fabric management stack. These components, known collectively as the LHC-Era Automated Fabric (LEAF) toolset, consist of a State Management System (SMS), which enables high-level commands to be issued to sets of Quattor-managed nodes, and a Hardware Management System (HMS), which manages and tracks hardware workflows in the Computer Centre and allows visualization of equipment location.

State management
SMS enables sets of nodes to be automatically reconfigured to be in production or on standby during operational and service management Use Cases. By leveraging the Quattor framework, a set of machines may be removed from production (for instance, during a kernel upgrade or a physical move), undergo the intervention and be put seamlessly back into production once the activity is complete. Concurrent events, such as a simultaneous kernel upgrade and physical move, are also correctly handled, with machines not going into production until both interventions are complete. All parties can see who is doing what, when and why. SMS ensures auditing, authorization, authentication and validation.

At a lower level, each node is configured to execute a specific script when asked to perform a particular state transition. For example, when the desired state of an interactive node is changed from "standby" to "production", logins are enabled and monitoring alarms switched on. Given this encapsulation, it becomes a trivial matter to issue high-level configuration commands to a heterogeneous set of nodes, such as "go into production", because callers do not need to know how this is achieved.

Hardware management
By LHC start-up mass installations, moves, renames and retirements are to be expected, along with daily hardware failures. A product of extensive workflow analysis and process re-engineering, HMS facilitates predictable, consistent, traceable and automatic workflows that are designed to scale up to the future needs of CERN's Tier-0/1 facility. The system:

- automates the update of all databases and repositories participating in its Use Cases;
- issues formal work orders wherever people are required to perform an action;
- provides statistics for management reporting;
- removes the dependency on specific individuals or informal communication for fulfilment.

LEAF status and next steps
Since its first production release in late 2002, where it was used to manage the installation of 400 new machines, HMS has evolved rapidly, with 16 new releases last year. With nearly 1500 machines relocated in just four months during 2004, HMS has been well tested and has proven successful. The first full production release of SMS was in January 2004 to coincide with a stable configuration database schema, and is currently deployed for all Quattor-managed nodes. HMS will continue to evolve smoother, more automatic processes and to handle more secondary scenarios on demand. It may be necessary in the future to track other types of hardware and to integrate new or modified components.

SMS is currently deployed for all farm PCs but needs to be extended to other node types, such as disk or tape servers. In addition, more service-specific SMS clients need to be written to allow service managers and system administrators to perform maintenance tasks and upgrades more easily. One such client will be a more advanced GUI, currently being developed to allow easy invocation of service management and operational interventions across sets of selected nodes, automatically initiating HMS workflows and invoking SMS state changes as necessary.

Conclusions
The ELFms tool suite is in full production at CERN's Computer Centre. We are benefiting from the results of three years of development, in which the components have been stabilized and hardened from prototype to production-quality. Thanks to their modular architecture, Quattor and Lemon have been integrated into the Computer Centre's site specific set-up and procedures, and have allowed us to progressively phase out legacy solutions. In their quality as framework systems, it has been possible to enhance the original set of plug-ins for system configuration and monitoring, thereby increasing the automation level of our complex and dynamically evolving computing fabric. With LEAF as a "glue" layer on top of Quattor and Lemon we are now able to define complex high-level workflows for full lifecycle management.

Although a number of non-core developments and integration steps remain to be done, we have reached a high level of management automation. This is a requisite for providing quality fabric services in Grid environments, where the correctness of configuration is critical and may affect the stability of the complete Grid workload system.

Other projects and sites are starting to deploy ELFms modules. This is leading to improved software portability and quality, and also increased functionality, in particular for configuring and managing LCG/EGEE Grid services. In this shared collaboration context, the coordination and standardization of developments and configuration schemas will become very important for streamlining new developments aiming to instrument Grid services, and avoiding incompatibilities and work duplication.

ELFms project team (based on the CHEP2004 paper by G Cancio, T Kleinwort, W Tomlin *et al.*)

Meeting brings PVSS users together to share knowledge

PVSS (Prozeßvisualisierungs- und Steuerungssystem) is a highly sophisticated Supervisory Control and Data Acquisition (SCADA) package, developed by the Austrian company ETM Professional Control. In 2000, after a detailed evaluation, PVSS was selected to be used to build the supervisory layer of the control systems for the LHC experiments, and a contract was established with ETM for unlimited use of PVSS for the LHC collaborations.

In 2002, after a review of SCADA products used at CERN, the CERN Controls Board recommended that PVSS should be used for all CERN projects requiring SCADA functionality. As a result, the contract with ETM was extended to cover all projects in the research and accelerator domains at CERN. Since then members of the four LHC experiment collaborations in approximately 100 institutes in 26 countries have started working with PVSS to build their part of the overall control systems for these experiments.

In addition, PVSS has been and is still being used for many other systems at CERN, including several fixed-target experiments (COMPASS, NA60 and HARP), the Gas and Magnet Control Systems for the LHC experiments, the LHC Cryogenics and Vacuum Control Systems, and the supervision of several LHC machine and experiment safety systems.

As PVSS offers extensive functionality, its correct use needs detailed knowledge and CERN users regularly hold discussions to exchange their experience. Furthermore, the features of PVSS lend themselves well to building reusable components to incorporate into application development. Within CERN this approach is being used extensively and the various projects have been able to benefit from one another's developments. Based on this experience it is clear that the exchange of ideas, and possibly of developments, with other PVSS users is of great interest. So in collaboration with ETM, CERN formed a PVSS Users' Group. The first meeting took place at CERN on 5–6 April 2005.

Audience at the PVSS Users' Meeting in the Council Chamber.

Vincent Lambercy from Skyguide, discussing the use of PVSS to supervise the technical system for air-traffic control in Switzerland.

At this meeting there were almost 150 participants from a wide range of different application domains, including HEP, radio astronomy, air-traffic control, traffic monitoring, gas production and distribution, water distribution and purification, maritime navigation systems and many others. Approximately three-quarters of the participants were from outside of CERN and the majority of these from industry. The meeting programme included 14 interesting and diverse presentations on experience in PVSS and special developments using it. Among these were presentations on the use of PVSS for:

- monitoring and control of the world's longest pipeline;
- monitoring and control of the world's largest radio telescope;
- monitoring and control of the H1 experiment at DESY;
- monitoring and control of the LHC cryogenics systems;
- monitoring and control of underground railway stations in Vienna;
- remote monitoring and control of the Main–Danube canal locks;
- supervision of parts of the technical infrastructure of Skyguide, the company that handles Swiss air-traffic control;
- development of the controls framework for the LHC experiments and the integration of a Finite State Machine toolkit.

These presentations provided a good insight into many applications being developed with PVSS, which although being developed for very different domains, often exhibit many similarities. The presentations highlighted two of the major advantages of PVSS: the device structuring and the possibility of building configuration tools, template applications and/or application generators to ease the development of final applications. Many of the presentations described one or another of these approaches.

It is important that the users of a system know how the product will evolve. To address this ETM gave a presentation on the features planned for the next 18 months. It should be noted that many of the new features were requested by CERN users.

As well as the presentations, there were three lively discussion sessions covering:

- design aspects in large and complex PVSS applications;
- how to develop company standards with PVSS;
- user evolution wishes for PVSS.

These sessions allowed users to discuss detailed technical issues with representatives of ETM and each other.

In addition to the technical aspects of the Users' Meeting, the participants were given the opportunity to visit the ATLAS cavern and the CMS installation hall to see for themselves the size and complexity of the experiments for which PVSS will be used. Furthermore, during the two days there were discussions between various ETM Partners (typically system integrators) regarding the possibility of exchanging software components.

The consensus is that the meeting was an excellent forum for users to meet and exchange ideas, and brought together an interesting mix of users from research and industry. Although the industrial approach was sometimes quite different from that in the research domain, there were many similarities in the problems encountered and the solutions chosen. The seeds of many new collaborations were sown at the meeting and it will be interesting to watch these develop in the weeks and months to come. We are all looking forward to the next PVSS Users' Meeting in 2006!

- See also the feature on control systems in *CERN Courier* June 2005 p20.

Wayne Salter IT/CO

Information corner

Questions and answers from the Helpdesk

The User Assistance Team in IT/UDS maintains a database for Questions and Answers that have been dealt with by the Computing Helpdesk. This provides many tips on daily computing issues. You can search the database at http://consult.cern.ch/qa/.

Below is an example of a Question and Answer (Q&A) related to Windows PCs (http://consult.cern.ch/qa/3831).

Most desktop icons lost on PC

Question

I seem to have lost all but a few of my desktop icons. Is there a quick way of getting them back?

Answer

The desktop icon shortcuts reside in the user's home directory in the form \\cernhome0X\Desktop_InitialLetter\Userid. Replace the "X" by the number of your home server (if unknown, the phone book and "more info" will tell you), "InitialLetter" by the first letter of your login ID, and "Userid" by your login ID.

Use the shadow copy client described in http://weba5/winservices/docs/ShadowCopyClient/ to recover the latest version. If none is visible, please e-mail helpdesk@cern.ch, asking for a file reload.

Other general-interest Q&As and their corresponding websites

Windows (NICE – Office) related

- http://consult.cern.ch/qa/3839 — Errors from .Net Framework
- http://consult.cern.ch/qa/3691 — NICE departmental or workspace share (large NICE disk quota needs)

Unix (AFS-Lxplus/Lxbatch) related

- http://consult.cern.ch/qa/0360 — AFS space (users quota and project space)
- http://consult.cern.ch/qa/3825 — rcp – connection refused
- http://consult.cern.ch/qa/3807 — Add font to PC Linux SLC3

Find out more about CERN IT Department activities

Every Friday morning, the managers of the main services offered from within the IT Department meet to coordinate their activities; this is the so-called C5 meeting.

It is called the "C5" meeting because it was known as "CCCCC" or "CERN Computer Centre Coordination Committee" in its original form, but today it includes representation from IT groups offering services not directly connected to the Computer Centre.

The minutes of this meeting are published every week on the World Wide Web and are therefore accessible to a wider audience at CERN. No reminder will be issued about the availability of the latest copy, but the minutes of a Friday meeting normally appear the following Thursday morning. You will find the C5 website at http://cern.ch/it-dep-c5/.

How to install the latest release of LaTeX on Solaris

TeXLive 2004

In January 2005 the latest release of TeXLive (TL2004, compiled in November 2004) was installed at CERN on AFS for general use on Linux. (See the article "New release of LaTeX at CERN" in CNL April–May 2005 p10, which you can browse online at www.cerncourier.com/articles/cnl/2/4/15/1.)

Recently, the binaries for Solaris (2.7) were added.

Running LaTeX on Solaris systems

On a Solaris (2.7) machine with AFS installed you get access to TL2004 by adding the directory containing the binaries to your PATH variable, as detailed in the box below.

Michel Goossens, IT/UDS

- If you are running a Bournelike shell (such as sh, bash and ksh):
  PATH=/afs/cern.ch/sw/ XML/TL2004/bin/ sparc-solaris:$PATH
  export PATH
- If you are running a C-type shell (such as csh and tcsh):
  setenv PATH /afs/cern.ch/sw/XML/TL2004/bin/sparc-solaris:$PATH

Recent changes to IT services

Changes to services in the IT Department are published on the Service Status Board (SSB) which is located at http://cern.ch/it-servicestatus. The most recent changes and their dates of posting are shown below. The SSB also includes service incidents, scheduled interventions, power cuts and the status of most services.

29 April	Xterminal server move (May)
25 April	Rundown of SUNDEV facility (1 July)
22 April	Restriction on remaining RedHat 7 capacity to shorter jobs (25 April)
20 April	Adobe Acrobat and Adobe Reader 7.0 (21 April)
18 April	Proposal to reduce the number of concurrent 1 nw LSF jobs per user (19 April)
14 April	Phase out of TAPES accounts (April–May)
30 March	New version of the Oracle Tools on Windows (4 April)

Calendar

June

26–29 **GGF14** Chicago, Illinois, US, www.ggf.org

July

24–27 **The 14th IEEE International Symposium on High Performance Distributed Computing (HPDC-14)** Research Triangle Park, North Carolina, US, www.caip.rutgers.edu/hpdc2005/

24–27 **The 3rd International Conference on Computing, Communication and Control Technologies (CCCT '05)** Austin, Texas, US, www.iiisconfer.org/ccct05/website/default.asp

August

17–19 **1st WSEAS International Symposium on Grid Computing** Corfu Island, Greece, www.worldses.org/conferences/2005/corfu/smo/grid/index.html

September

5–9 **Parallel Computing Technologies PaCT-2005** Krasnoyarsk, Russia, http://ssd.sscc.ru/conference/pact2005/

12–18 **XX International Symposium on Nuclear Electronics & Computing (NEC'2005)** Varna, Bulgaria, http://sunct2.jinr.ru/NEC-2005/first_an.html

17–22 **Oracle OpenWorld 2005** San Francisco, California, US, www.oracle.com/openworld/sanfrancisco/conference/index.html

The deadline for submissions to the next issue of CNL is

25 July 2005

E-mail contributions to cnl.editor@cern.ch

If you would like to be informed by e-mail when a new issue of CNL is available, subscribe to the mailing list cern-cnl-info. You can do this from the CERN CNL website located at http://cern.ch/cnl