FEATURED CONTRIBUTIONS AT EGEE USER FORUM

VERSION 3.0

TABLE OF CONTENTS

1. INTRODUCTION

The Strategic Discipline Clusters of the EGEE-III NA4 activity have played a key role in the definition of the scientific programs of the User Forum events. More importantly they have worked to increase the use of grid technologies within their communities, directly and indirectly aiding many of the contributions to the User Forum events. Prior to each User Forum, the NA4 Steering Committee, including of the leaders of the Strategic Discipline Clusters, identifies featured contributions, from the consistently excellent User Forum contributions, that showcase the work and/or impact of the Strategic Discipline Clusters within their scientific domain. The descriptions explain the importance of the technical and/or scientific of these featured contributions for User Forum 5 as well as how the EGEE-III NA4 activities contributed to it.

2. ASTRONOMY & ASTROPHYSICS

The Grid for Astronomy and Astrophysics in Italy and Europe

Dr. TAFFONI, Giuliano (INAF - IASFBO)

http://indico.cern.ch/contributionDisplay.py?contribId=159&sessionId=26&confId=69338

The chosen featured contribution will provide a report of all relevant activities carried out within the A&A cluster during the two years of the EGEE-III project.

The primary author (G. Taffoni) has written the abstract of the contribution having in mind the objective of providing an exhaustive overview of what has been produced during the two years of EGEE-III and the most important results achieved in this period, some sort of final evaluation of the participation of the A&A activity in EGEE-III now that the end of the project is approaching.

To produce a report that really covers all activities carried out in the A&A cluster, the primary author will collect inputs from by all Institutes and research group that actively contributed in terms of gridified applications tools and services.

The oral contribution will report not only on applications but also on relevant activities such as training and dissemination events; in practice all those initiatives aimed at extending and making more substantial the participation of the A&A community in the project.

This featured contribution, therefore, is exactly focused on the role played by the A&A cluster to foster the penetration and adoption of grid computing by the European astrophysical community.

This contribution indeed is ideal to provide a summary of what happened in these two years, much more effective than a contribution focused on a single specific application. The contribution should have the advantage of stimulating discussion, especially in this phase where there are not clear perspectives for a continuation of the activities started through the A&A cluster in the EGI era.

The contribution will report about the most relevant technical and scientific advances in grid computing achieved thanks to the A&A cluster. This snapshot of the current state of the art for what concerns grid and the A&A research is fundamental as it can stimulate the development of new ideas and techniques on the basis of what has been achieved so far.

The scientific and technological impact of the contribution is, therefore, evident. The A&A cluster has certainly been dominated by some pilot applications: Planck and MAGIC being the most important of them. But, besides these, other projects and applications have had the chance to approach and use the grid.

In this period that comes before the User Forum 5 in Uppsala, a discussion is already in progress within the A&A cluster for what concerns the work done so far, the lessons learnt, and how the grid related activity should progress in EGI to improve the results achieved. The featured contribution should act as the main reference for this discussion twenty days before the end of EGEE-III although future perspectives are now negatively influenced by the lack of funds to support the communities of users.

3. COMPUTATIONAL CHEMISTRY

Computational chemistry experiments over a web service-oriented grid

Prof. BARONE, Vincenzo (Scuola Normale Superiore)

http://indico.cern.ch/contributionDisplay.py?contribId=76&sessionId=13&confId=69338

The constantly increasing complexity of chemical computations caused by the demand to study the properties of bigger and bigger molecules is the main reason for the search for new numerical algorithms and techniques in the field of computational chemistry. Among many types of computations performed for molecules, those of electronic energy derivatives are of particular importance for chemists. Depending on the derivation order, different molecular properties can be predicted, ranging from molecular geometry or dipole moment (first order energy derivatives) through molecular vibrations or nuclear magnetic resonance chemical shifts and coupling constants (second order properties), up to non-linear optical phenomena or anharmonic corrections for vibrational spectra (third and higher orders). Because of the high complexity, as well as the huge demands for computational power and large memory requirements, such computations were previously possible mainly on HPC architectures. Fortunately due to the parallelisation of existing methods such computations are also possible now on the grid infrastructure. As a result of the Computational Chemistry Cluster efforts, many computational chemistry packages have been ported to the grid and enabled for parallel execution.

Further steps in applying the grid architecture to chemical computations have been introduced by a group of researchers led by Vincenzo Barone. The computational model proposed by the group consists of a three-layer set-up, with each layer performing a different task (see the detailed description in the abstract). The tasks are responsible for (1) interaction with the user, input and jobs coordination, (2) creation and execution of all separate inputs and (3) post processing including extraction of important information from resulting the output files and the computation of final results. The most significant achievement of the model proposed by Barone's group is related to the computations of individual components required to obtain the desired order of energy derivative. To enable this step legacy software codes have been updated. Several test computations have been performed and the results are very promising. A set of tests showed the computational benefits of this kind of solution for the computation of molecular properties. Also it turned out that grid computations scale much better then locally executed parallel computations for medium to large size molecules.

Work towards the utilization of the grid infrastructure for fundamental algorithms of computational chemistry methods has already been performed by other researchers. One of the examples is an execution of the SCF cycle on the EGEE grid for the Wien2k package. Nevertheless the talk by Barone presents a novel and a unique technique. The method has been proven not only to be useful for the calculation of molecular properties, but also has shown the advantages of grid computing over locally executed parallel computations. This is a significant step in advanced computations of the properties for large molecules on the EGEE grid architecture. We look forward to the existence of such method capabilities in broadly available chemical software packages.

4. EARTH SCIENCE

Bridging the gap between applications geospatial data and the Grid

PINA, António (Universidade do Minho)

http://indico.cern.ch/contributionDisplay.py?contribId=91&sessionId=14&confId=69338

Geospatial and Grid infrastructures interoperability in enviroGRIDS

Prof. GORGAN, Dorian (Technical University of Cluj-Napoca)

http://indico.cern.ch/contributionDisplay.py?contribId=161&sessionId=14&confId=69338

The earth science data always contain at least three coordinates, geographical location (latitude and longitude) and time, and in many cases four coordinates as the altitude is added. Then Geographical Information Systems (GIS) are used in all kind of activities related with spatial data, such as societal applications (cadastre, topography, census, traffic, all public information), civil protection (flood, fire, earthquake), industry, research (earth observation, global and regional climatology) and so forth. New technologies, such as Web and grid services, and easier access to powerful HPC resources permit earth science users to solve complex issues. New requirements arise, such as an easy and standard way to access geographically distributed data, the need for more intensive computing related to the increase of the data volume and the complexity of the simulations, and the need for quick results, especially for weather and risk predictions. Those requirements and solutions, addressed by the ES cluster and by the ES grid community for several years, are expressed in the two selected abstracts from the Cross-fire and EnviroGrids projects. Cross-fire is a Portuguese NGI funded project aiming to develop a grid-based risk management decision support system, especially for civil protection. The EU FP7 project, EnviroGrids, concerns the observation of the Black Sea Catchment and Geospatial functionalities are needed for decision makers as well as all public. The ES cluster is in contact with both project teams.

In previous years many European and international projects and initiatives (e.g. INSPIRE, GEOSS, GMES) aimed to define an architectural framework for the realization of the so-called Spatial Data Infrastructure (SDI). The Open Geospatial Consortium (OGC) Inc.®, a non-profit, international, voluntary consensus standards organization, is leading the development of standards for geospatial and location based services. OGC has defined specifications for many different geospatial Web-based services: the Open Geospatial Web Services (OWS). The main services already used by different ES applications in hydrology and fire (cf. the EU CYCLOPS project) are the Web Coverage Service (WCS), Web Map Service (WMS), Web Processing Service (WPS) and Sensor Web Services (SWS). As an example, WPS is a computer program to publish and perform geospatial processes (e.g. a simple geometric calculation or a complex simulation model) over the Web with a standardized and open interface. Extending grid services with OGC web services (OWS) is well suited for providing high processing performance and storage capacity along with improved service availability. The OGC and OGF are working together to specify standard interfaces and best practices. Instances of the OGC service are developed by the G-OWS (the grid-enabled Open-Geospatial Web Services Working Group) on gLite for example, the Catalog service for Earth Observation data (CSW), Web coverage service (WCS) for land and Meteo datasets, WPS for risk management tools and the sensor observation service (SOS). Representatives of the ES cluster have participated in the G-OWS initiatives.

Several members of the wider ES grid community have actively participated in the OGC activities. The ES cluster informed the European Services Director of OGC about the extension of the ES cluster to a wider ES grid Community, and a letter of OGC support for the SAFE project. OGC is interested in the sharing of results and experiences about the exploration of synergy effects between distributed computing and geospatial data as well as processing infrastructures. Additionally OGC has welcomed and encouraged the Earth Science VRC to actively participate in the OGC process.

In both abstracts and also in G-OWS, the key point addressed is the OGC and grid interoperability, and in particular with gLite. The advances in this domain will attract new ES domains of applications that need intensive computing, to be used by decision makers, industry, research and the general public. These different initiatives and collaborations are crucial for Earth Science.

5. FUSION

Complex Scientific Workflows exploiting Grid and HPC

Dr. CASTEJóN, Francisco (CIEMAT)

Mr. GóMEZ-IGLESIAS, Antonio (CIEMAT)

http://indico.cern.ch/contributionDisplay.py?contribId=17&sessionId=27&confId=69338

The importance of this milestone relies on the fact that simulating a full fusion Tokamak requires a range of codes and applications that address different aspects of plasma physics and work at specific ranges of space and time scales. The necessity of such disparate codes is based on the different physical models and techniques that are used for solving such specific problems. The computational complexity of all these tools is so high that only paradigms like grid computing or HPC allow carrying out all the simulations.

The full simulation of a magnetic confinement fusion device needs the implementation of workflows between such applications that can run on the grid or on an HPC. In this work we show a complex workflow that uses two applications that run on such different computer architectures, grid and HPC, taking advantage of the specificity of those computing platforms.

As previously shown by the fusion community in the EGEE project, the usage of grid infrastructures for fusion research has provided interesting and relevant results that have created a large number of new possibilities to explore in the future. The fusion community has been traditionally focused on HPC and it is still reluctant to use grid computing, despite the fact that some very good results have been obtained using grid techniques. We go beyond previous work here, showing the feasibility to join both technologies and take the maximum advantages from each of them. Our experience shows that creating and executing complex workflows to simulate plasma physics is a critical point nowadays and the key point for the future of the simulation in our field.

The EUFORIA project supports fusion modellers in this simulation work by providing programming expertise for porting and optimizing codes on a range of computing platforms, providing grid and HPC resources, and providing the tools that are required to orchestrate the range of simulation codes necessary to simulate the full fusion reactor. We use some developments achieved in the EUFORIA and EGEE projects to implement such a workflow. An HPC from the EUFORIA project and the EGEE grid fusion VO are the infrastructures used for this workflow.

Kepler, a workflow orchestration tool, has been modified by the EUFORIA team to enable fusion modellers to submit simulations to both grid and HPC resources from their desktops and to visualise the results they obtain. The project partners collaborate with DEISA and EGEE in order to ensure a wide adoption of the tools developed and deployed by EUFORIA in the infrastructure of DEISA and EGEE, by using the Fusion VO resources. Applications previously ported to EGEE gird are linked by this workflow with others that run either on the grid or on HPCs.

We demonstrate with this milestone that the feasibility of creating complex workflows, with some components running on HPC, others on the grid, and with the use of the results obtained from these components by another component that can also run on a grid or HPC, becomes a reality with these developments. The technical feasibility of launching jobs from this workflow orchestration tool to a mixed DEISA and EGEE environment provides a basic tool for building a new fusion computing paradigm in which the users can run their applications using the most suitable infrastructure for that.

This technique allows the fusion researchers to exploit the applications in a new way that will produce new relevant scientific results, showing the importance of this development in future researches.

6. GRID OBSERVATORY

Addressing Complexity in Emerging Cyber-Ecosystems - Exploring the Role of Autonomics in E-Science

Prof. PARASHAR, Manish (Rutgers University)

http://indico.cern.ch/contributionDisplay.py?sessionId=9&contribId=175&confId=69338

This keynote speech is a new contribution to building bridges between grid researchers and practitioners on one hand and those in the autonomic computing community on the other. Towards the same goal, the Grid Observatory has initiated the Grids Meet Autonomic Computing (GMAC) series of workshops (first one in June 2009, and continued in June 2010, associated with the IEEE International Autonomic Computing and Communication Conference (ICAC). As acknowledged in the GMAC'09 panel called "Grids/Clouds/Autonomics Convergence" (with Bob Jones and Manish Parashar among the panellists), the establishment of a data repository and the exploration of behavioural models provide an important basis for the subsequent development of autonomic software and systems.

Infrastructures and applications for e-science are complex systems, that is systems that exhibit collective properties going well beyond those of their individual components. The complexity of the EGEE grid stems from both sources: its sharing paradigm, and its unprecedented scale in resources, data, and expected timeline. As such, EGEE epitomizes all the unprecedented challenges in development, configuration and management associated with computing systems' complexity in the Internet age. The promise of autonomics is to transform the painful and not always successful process by which administrators and users design their configurations and applications into a hierarchy of self-governing systems driven by high-level goals, through self-configuration, self-optimization, self-healing and self-protection.

Existing cloud computing infrastructures already support some autonomic concepts, such as dynamic resource provisioning, and migrating workloads across computational platforms. A key differentiator between grids or clouds and enterprise data centres, from the viewpoint of the role of autonomics, is the nature of the applications and usage modes. The grid modality of e-science is thus a specific area for autonomics that can greatly benefit from the cross-fertilization of autonomic research and grid production. Autonomics offer principled methods to manage the systems' scale, dynamism, and heterogeneity at multiple levels (systems, application, and data). Autonomics address a fundamental requirement of large-scale distributed systems and computational science applications: decision making under uncertainty because of incomplete or inaccurate monitoring data.

The transition from conventional systems and application to autonomic ones is a long process. Besides the fundamental scientific issues, a global question is how to build trust between autonomic-based systems and its potential users and administrators that have to deliver a 24/7, production-quality service. The keynote speech is an important step on this direction, by promoting community wide discussion of, and collaboration on potentially high-impact ideas that will influence and foster continued research in improving the scalability, manageability and reliability of grids.

7. HIGH-ENERGY PHYSICS

LHCb operations: organizations, procedures, tools and critical services.

Dr. SANTINELLI, Roberto (CERN/IT/GD)

http://indico.cern.ch/contributionDisplay.py?contribId=19&sessionId=19&confId=69338

This talk gives an insight into the LHCb grid computing operations.

It will focus on aspects related to the monitoring of the services and resources describing the pros and cons of the tools in place used for checking the health of the system. Limitations found, actions performed and plans to address these issues are described. From the level of the fabric infrastructure level to the VO specific services, passing through the general grid service layer, we will provide an audit of the instruments available to the operations crew to address problems with the system. Emphasis will be given to the grid services monitored by SAM, SLS and Dashboard and to the Tier 0 critical services (VOBOX, DIRAC), whose monitoring has been integrated into the CERN IT infrastructure in order to exploit the 24x7 support offered. Furthermore, alarms and procedures defined for the IT piquet service are described. These sensors and alarms (after years of tuning) can now spot problems before users.

The talk will also give a breakdown of the structure, organization and procedures within the LHCb grid production team describing how the various support layers interact, the way information is conveyed between stakeholders, and solutions are shared amongst them.

Details of a new DIRAC component for service and resource monitoring and management–the Resource Status Service–are given. This service addresses the existing lack of procedures inside the LHCb production team about site management and policies, as well as enforcing procedures that today are based on the common sense of the shifters and experts.

An overview on the operations of the other LHC experiments will be provided: their view of service criticality, tools & procedures and operations. The main differences and the common aspects are presented. This will help to define the best practices and share them among all communities.

In summary, this talk presents key aspects of the WLCG service from the users' perspective, that is, from the LHC community. The effort of the Experiment Support team at CERN, and in particular of the NA4-HEP community, in providing operations monitoring tools, support infrastructures, and user analysis tools with special emphasis to the data access solutions, is reflected in this presentation. This is the culmination of a common effort performed for eight years between the experiments, the WLCG and the EGEE projects.

8. LIFE SCIENCE

Grid-based International Network for Flu Observation

Ms. DA COSTA, Ana Lucia (HealthGrid)

http://indico.cern.ch/contributionDisplay.py?contribId=158&sessionId=11&confId=69338

This contribution introduced a grid-enabled flu information collection network (g-INFO: grid-based International Network for Flu Observation) that aims at (i) integrating existing data sources to assemble a global surveillance network and (ii) providing a reactive online computing facility able to process in near real-time the information collected daily by national health surveillance institutions. The output of g-INFO is a grid database populated with viral sequences information and a portal delivering daily results on the analysis of these sequences to scientists. Currently, sequence alignment and phylogeny tree extraction bioinformatics tools are used to monitor the evolutions of virus gene sequences and to identify mutations leading to the emergence of new diseases.

Viruses know no state borders and the International Network for Flu Information is intrinsically a Virtual Organization-scale effort (Europe-Asia) to track the emergence and evolutions of diseases. Beyond the multi-national collaborative work fostered by the Life Science cluster, the establishment of this health surveillance network also provides the operational procedures and the tooling (e.g. pilot job framework, grid databases, workflow engine) designed within the cluster or in collaboration with other grid partners to develop innovative in-silico approaches to the disease monitoring challenge.

The technical platform relies on data-intensive and reliable grid-computing services designed, tested or improved in the context of the Life Science cluster. In particular, the WISDOM pilot jobs framework, originally designed for the drug discovery Data Challenges, was extended to provide a multi-purpose, reliable, and efficient job submission environment. The AMGA grid database front-end also benefitted from earlier performance and scale improvements. The MOTEUR workflow engine is currently tested to implement the logic of the data phylogeny pipeline. In the future it will be used to describe and enact new data analysis procedures.

In recent years, several rapidly expanding pandemic diseases have spread worldwide, including SARS (Severe Acute Respiratory Syndrome), the avian flu, Influenza A, and the swine flu. Among the biggest challenges from emerging infectious diseases is the relation to early detection and surveillance of the diseases, as new cases can appear anywhere. Existing data sources are integrated towards a global surveillance network for molecular epidemiology. The g-INFO portal is expected to have an impact by adding a new weapon to the health researchers' arsenal: the grid.