Catch-All Virtual Organizations - Solution for Heterogeneous and Disperse Grid Users Communities

Jan Kmuníček[1], Daniel Kouřil[1], Eric Yen[3], Luděk Matyska[1], Zora Střelcová[2], Petr Kulhánek[2], Jaroslav Koča[2]

[1] CESNET, z. s. p. o., Zikova 4, 160 00 Praha 6, Czech Republic
[2] National Centre for Biomolecular Research, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic
[3] Academia Sinica Grid Computing Centre, Institute of Physics, No. 128, Sec. 2, Academia Road, Nankang 11529, Taipei, Taiwan
{kmunicek, kouril, ludek}@ics.muni.cz
{stre, kulhanek, jkoca}@chemi.muni.cz
Eric.Yen@twgrid.org

Abstract. The virtual organizations form a key concept for seamless utilization of all advanced features of available worldwide grid environments, the EGEE Grid in particular. However, apart from large, well organized user communities there is a substantial, non-trivial activation barrier in adopting grid concept and creating a new VO for small research groups. To minimize this barrier, a so-called "catch-all" virtual organization approach has been implemented. Here we present two successful use cases of catch-all virtual organizations (Virtual Organization for Central Europe and for EUAsia), demonstrating the strengths of this approach, discussing their setups, encountered challenges and suggesting corresponding solutions.

Keywords: virtual organization, grid, Enabling Grids in E-Science

1. Introduction

With increasing interest in worldwide grid environments that are becoming an ubiquitous media to perform up-to-date research tasks there is increasing importance in proper management and administration of involved end users communities and corresponding sets of computational and data resources. Nowadays, grids represent unique environments to fulfill many distinct requirements regardless whether the researchers need a solution for demanding computations requiring maximum of available computing power, data sharing across many institutions or cooperating institutes, interactive collaboration among geographically separated workplaces or they are looking for remote access to high-tech experimental devices.

The term *grid* [1] can be described as a large distributed system of organizationally separated elements that can be individual computers (memory, processor, hard discs), data storage (disk arrays, large HSM systems) and information systems jointly interconnected by a computer network. According to the concept of a grid

environment, the *Grid* is primarily an interconnection of resources of individual organizations with the main aim to ease collaboration and resource sharing for institutes and research groups involved in concrete cooperation. Subsequently, having in mind the main objective, this should lead to observable increase of effectiveness in utilization of incorporated resources. In words of I. Foster and C. Kesselman [2], "Grid computing is coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations."

Given the vast extent of grid environment, the users need a way how to effectively use the potential of it. Therefore, the so-called *virtual organization* (VO) has been introduced, which offers a way to organize work and appropriate utilization of the grid environment. In general, a virtual organization corresponds to a fellowship formed by individual persons, institutional computational/data resources with uniform willingness to work together on a specific, agreed topic or in a specific domain. Basically, a VO interconnects separated (even geographically, administratively, regionally) researchers and corresponding set of resources working together on the common field of interest. Each virtual organization is always fully described by its attached end users, dedicated resources available for them and application domains that the VO tries to profile in. A VO provides a platform through which the VO members can collaborate with each other. The VO management takes over many administrative tasks for their users, such as negotiation with resource providers and basic users' management. Users of the VO can focus solely on their research without being bothered with additional administrative overhead. In this paper we introduce the concept of catch-all VOs that provide an easy entry point to the grid for people not converged by any existing VO. We also describe a summary of our experiences gained during establishment and operation of two catch-all VOs.

1.1. Enabling Grids in E-SciencE

The worldwide grid activities are currently greatly influenced by the series of the EU Enabling Grids in E-sciencE (EGEE) [3] project. The EU EGEE started in April 2004 and has been further developed in two following EU co-funded projects. The project has gone a long way already—starting with the establishment of underlying fabric infrastructure, through the enrichment of the application domains profiting from being gridified up to current transition towards sustainable grid environment forming the cornerstone of European Grid Infrastructure (EGI). The EGEE project brings together experts from more than 50 countries with the common aim of building on recent advances in Grid technology and developing a service Grid infrastructure which is available to scientists 24 hours-a-day. To be able to serve the enormous number of end users from numerous number of application fields (as the archeology, astronomy and astrophysics, biotechnology, bioinformatics and biomedical sciences, civil protection, computational chemistry, computer science, earth sciences, finances, fusion, geophysics, high-energy physics, life sciences, multimedia or material sciences) the establishment and corresponding management of more than several hundreds of virtual organizations has been required and performed. The EU EGEE

project has paid a big attention to the VO concept and provided tools, procedures and policies that are crucial to establish and run a VO.

2. Virtual Organizations

Three parallel tasks have to be fulfilled to provide end users with a properly managed virtual organization. The VO administration involves running of basic (gLite in case of EGEE Grid) services (such as WMS, BDII, VOMS, etc.), provision of the end users management, and access to user interfaces/high-level middleware frameworks to minimize the users' hassle with the underlying middleware to increase their real work efficacy. The experience obtained from the first phase of the EGEE project clearly indicated that these steps represents a high initial barrier that an individual researcher must be able to overcome before being able to even get familiar with the grid services. Therefore, it became apparent that there is a substantial need to operate a virtual organization that is immediately available to users from any scientific community. Clearly, small user communities were not going to profit from the delivered grid environment unless a suitable approach for involving a small research group (or even individual researchers) not tightly bound to a specific application domain or research field was available. Therefore a catch-all approach VO has been introduced and applied. The so-called "catch-all" virtual organizations represent an effective way for users to use grid environments. Catch-all VOs are provided usually as a service to users' communities as part of user support activities of several projects focused on provisioning of grid infrastructure. Linking providers and different end user communities, catch-all VOs form a crucial step towards a routine use of the worldwide grid platform, which is easily available to users. However, compared the majority of other VOs, they are not dedicated to any particular application domain, applications, or experiment. A catch-all VO represents a completely open grid environment willing to accept any potential user without restricting the user in the content of his/her research. Essentially, the openness of catch-all VOs can be recognized on three different levels:

- *open* from the **accepted users** point of view (limited by the VO politics and local resources access rules)
- *open* from the **application** point of view (no limitations in kind of application software being used until it breaches security and utilization rules)
- *open* from the **user tools** point of view (any advanced interface or high-level framework can be tested or utilized in fully production mode)

The daily routine VO management is then performed by dedicated administrators while the VO users focus themselves on purely research tasks. Decreasing the entrance barrier in this way is especially important for various regions with high heterogeneity and different grid knowledge of involved parties. The main advantages of the catch-all approach can be summarized as follows: dedicated administration of the catch-all VO by technically knowledgeable persons, initial outsourcing of critical

VO services for user communities new to the grid, freedom from establishing their own VO before trying a grid environment. In addition, the crucial advantage of catch-all virtual organizations lies in the fact that they provide three different functions in one simple setup. Each catch-all VO serves as an incubator for prospective grid users (primary aim). Simultaneously, it can function as a training environment (secondary aim) and a fully production grid environment especially for small groups (tertiary aim).

The catch-all VO can advantageously serve as the infrastructure for training purposes (t-infrastructure) as well as for standard production work. This is technically solved through dedication of a subset of VO resources as t-infrastructure thus forming an independent "sub-VO". The sharing of the infrastructure with the production services guarantees a high level of resources availability. As the training activity requires a reasonably secure environment, no anonymous users can be allowed to access the t-infrastructure and for the training itself only short-term certificates (days) are issued in advance. Using a production grid infrastructure (as is the case of the catch-all VO) as a t-infrastructure has the following advantages: the resources are allocated in advance, they are fully guaranteed, and the users are trained in the same environment to use in daily work later when the training activities are finished.

Users incubation is basically fulfilled through the smooth transition from a trained user into an advanced user with the production use of the infrastructure with the ultimate goal—to support preparation of a new, application/domain/purpose-specific VO (as soon as users will get more organized and experienced and willing to take responsibility of their own VO).

Despite indisputable advantages of catch-all VOs, there are also few issues that should be taken properly into account. First, the utilization rules for a catch-all VO have to be set up properly and strictly to prevent unwanted misuse or security breaches. The VO usage rules should help to avoid any unnecessary overloading of the infrastructure, which may be difficult to detect for the VO management. Enforcing the rules may be difficult, however, mainly since the catch-all VO manager lacks a tight relationship with their users. For example, acting as the CE VO managers we were notified once about a user running a set of seti@home tasks, which seemed as exemplary example of misusing of the infrastructure. However, having approached the user we found out he was doing research aimed at the utilization of the BOINC framework in grids and using seti@home to evaluate the infrastructure.

The application management is also different compared to the application specific VOs, where the Software Group Manager (SGM) is responsible for application software. The SGM is a privileged role allowed to maintain software used by common VO members. The software repository is located on each cluster supporting the VO and the SGM users are provided with write access to the repository. This concept cannot work in the catch-all VOs correctly, since the users community does not use a single set of software that could be maintained by a single user or a small group of users. The catch-all VOs can provide SGM accounts to anyone who asks for them, causing the repository not being maintained in a uniform way. Namely, no one can really ensure the integrity of the software placed in the repository since virtually uncontrolled set of users can change the applications stored. If users require a strict control over their applications, they have to ensure their fetching to the local cluster as a part of the job and/or its preparation.

3. VOCE Use Case

The Virtual Organization for Central Europe (VOCE) provides a complete grid infrastructure under the EGEE project umbrella. It was established to serve the requirements from all researchers belonging to the EGEE Central Europe region (Fig.1). VOCE is officially registered as the standard EGEE VO; the detailed VO card is available from the CIC portal [4]. VOCE was setup as the very first catch-all VO with regional scope spanning the whole Central Europe federation. The core VOCE services are operated by CESNET [5], the Czech NREN/NGI provider, while the computational and storage resources are contributed by many institutions across the whole Central Europe region. Currently, the portfolio of participating countries involves Austria, Croatia, Czech Republic, Hungary, Poland, Slovakia, and Slovenia that indicates well the heterogeneity in partners as well as the main common regional characteristic.

Fig. 1. Virtual Organization of Central Europe logo.

The current VOCE status is summarized in Table 1. As the annual renewal of the VOCE account is required to continue the user membership in VOCE there was registered so far 642 users from which 232 user account are currently active (in the year 2009).

Table 1. VOCE statistics (data taken from *lcg-infosites* command output).

VOCE statistics	
CPU cores	8474
Storage	196,7 TB
Users	232

Based on the report during account extension for year 2009 the VOCE is mainly utilized for evaluation of testing jobs (collections, parametric or DAG jobs), free energy calculations, docking, molecular dynamics/quantum chemistry studies, development of various portlets, numerical astrophysics simulations, phylogeny studies, plasma analysis, and generic porting applications to the EGEE infrastructure. The utilization report clearly indicates successful achievement of the primary catch-all VO goal—to serve distinct user groups and support large set of scientific application domains. The steady and continuous increase in number of jobs within VOCE is indicated by accounting plot in Fig. 2. Moreover, a few computational

challenges (as the one dealing with large scale free energy calculations testing the multiple walkers approach accelerating such kind of calculations) have been successfully completed within the VOCE environment, demonstrating thus the applicability and usability of the worldwide grid environments represented by the EGEE Grid.

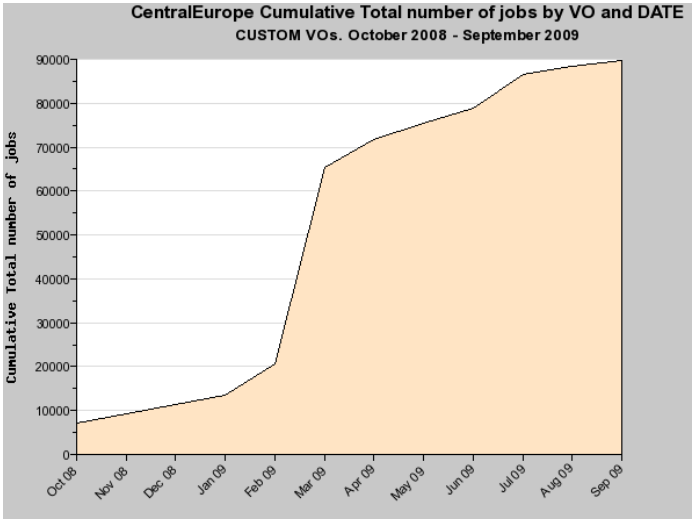

Fig. 2. Cumulative total number of jobs in VOCE (data from EGEE accounting portal – http://www3.egee.cesga.es/gridsite/accounting/CESGA/egee_view.html).

VOCE has proven useful for many users seeking experience with a grid. For example, the VOCE infrastructure has been utilized by the group of researchers from National Centre for Biomolecular Research (NCBR, http://ncbr.chemi.muni.cz/) that performed chemical computations (free energy calculations and novel methods of their accelerating in particular) at their local resources and were looking for ways how to test their applications in an environment of larger scale. The VOCE infrastructure and its user support managed to help the users to get familiar with the grid environment and submit first jobs. Following that, the users were able to run their regular jobs and even perform a quite large scale challenge utilizing a great deal of the infrastructure. Utilizing the VOCE support and core services, users can fully focus on preparation of their applications to the grid environment, which enabled to finish all the steps quite quickly. After this experience we worked together with the users to adapt their tools for job submission and management so that they are able to seamlessly support both the local and grid facilities. Using these tools users could easily use both environments, in a transparent way. The tools are also available to other user communities in VOCE and Czech national grid infrastructure.

During the VOCE existence we have learnt that the fundamental benefit of the catch-all approach is the fact the VO administration is disburdened from the users. When we accompany this with automatic obtaining an account on preinstalled User

Interface (UI) machine, the maximal simplification of the registration procedure followed by semi-automatic user membership acceptance and dedicated helpdesk (running specific instance of the request tracking system) we obtain a unique set of added value attractive enough for the researchers to test and use.

4. EUAsia Virtual Organization

Built on the top of the expertise gained with the catch-all concept and its implementation in VOCE, a know-how transfer through partnership within the EUAsiaGRID consortium (participating in the EUAsiaGrid project [6]) was recently introduced, the applicability of the catch-all approach in Asia-Pacific (AP) region was tested and resulted into a newly established, regional virtual organization named EUAsia.

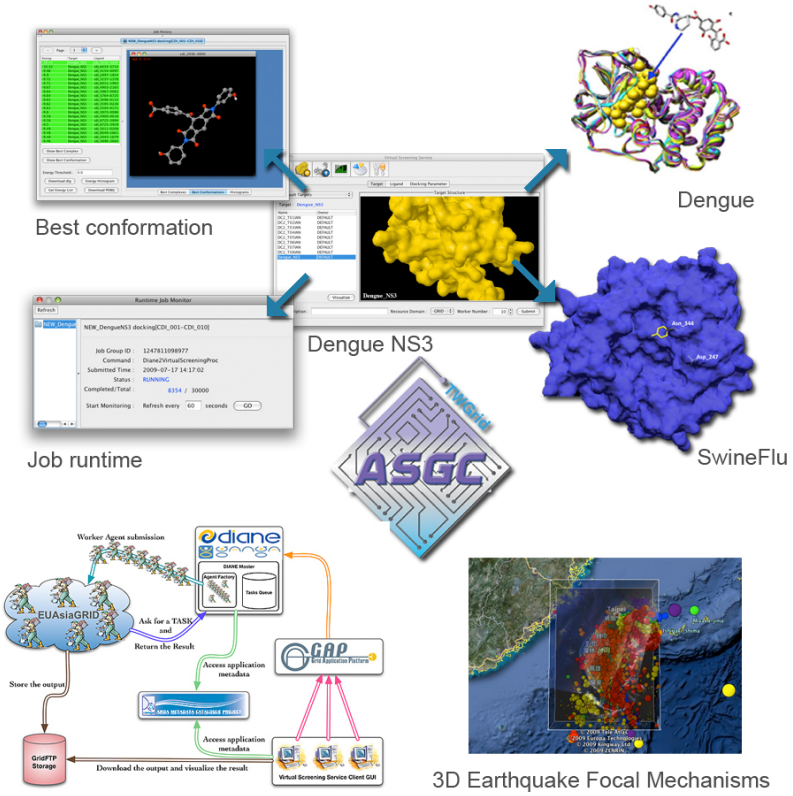

Fig. 3. Key application-related activities within the EUAsia catch-all VO.

Apart from helping project partners to obtain experience with setting up and running a VO, the EUAsia VO serves also as a template for the planned new

EUAsiaGrid initiated VOs,. The EUAsia VO is operated by ASGC, with help of the European partners (namely CESNET).

Compared to VOCE the EUAsia VO has during its startup phase a slightly different registration procedure (two steps) to meet the very specific regional requirements. However, based on the first user feedback the registration process was simplified, the automatic creation of an account at a predefined UI has been added and further small modifications to the whole process have been introduced. The current EUAsia status is summarized in Table 2. The application profile of EUAsia VO is rather specific and strongly related to the AP region itself (Fig. 3).

Table 2. EUAsia VO statistics (data taken from *lcg-infosites* command output).

EUAsia VO statistics	
CPU cores	2225
Storage	4,8 TB
Users	133

The application portfolio within the EUAsia virtual organization profits from the heterogeneity of the Asia-Pacific region. Apart from the obvious, well-know application areas covering research domains like high energy physics, computational chemistry, bioinformatics and biomedicine, there is also a set of application areas interested in exploiting the available EUAsia VO resources through solving tasks in social science, digital culture and heritage and mitigation of natural disasters—crucial domain for the region naturally dealing with earthquakes, typhoons, floods and landslides. Moreover, the continuous search for new research domains within the EUAsiaGrid project brought further areas that could benefit from grid-enabling of their respective codes as weather forecast and climatology, mathematical modeling and biodiversity studies.

5. Conclusions and Future Prospective

In this paper we have showed that the introduced concept of catch-all virtual organizations has proven justified and viable. The catch-all VO approach was easily adopted by several distinct communities. The flexibility in the VO policy and registration procedure set up towards different user communities with the same underlying core services and fabric infrastructure allows provision of customized virtual environments according to different user communities needs. In addition, the availability to serve as training and production environments in parallel makes catch-all virtual organizations a crucial, non-separable part of any advanced e-infrastructure.

Acknowledgements. The presented work is supported by the following grants: The Ministry of Education of the Czech Republic (contracts LC06030, MSM0021622413 and MSM6383917201), by EGEE III project funded by European Commission (contract number INFSO-RI-222667), and by the EUAsiaGrid project (contract number FP7-INFRA-223791).

References

1. Foster, I., Kesselman, C.: Computational grids—Invited talk (Reprinted from The Grid: Blueprint for a new computing infrastructure, 1998). Vector and Parallel Processing—Vecpar. 1981, 3–37 (2001)
2. Foster, I.: The anatomy of the grid: Enabling scalable virtual organizations. First IEEE/ACM International Symposium on Cluster Computing and The Grid. 6–7 (2001)
3. Enabling Grids in E-Science (EGEE) project (http://www.eu-egee.org/)
4. CIC portal (http://cic.in2p3.fr/)
5. CESNET, Czech NREN and NGI (http://www.cesnet.cz/)
6. EUAsiaGrid project (http://www.euasiagrid.eu/)