

PLGrid PLUS: Toward Domain-Specific Infrastructure for Supporting International Research Collaboration

Jacek Kitowski^{1,a,b}

a) AGH University of Science and Technology, Faculty of Electrical Engineering, Automatics, Computer Science and Electronics, Department of Computer Science, ul. Mickiewicza 30, 30-059 Krakow, Poland

b) AGH University of Science and Technology, ACC Cyfronet AGH, ul. Nawojki 11, 30-950 Krakow, Poland

E-mail: kito@agh.edu.pl

Lukasz Dutka

AGH University of Science and Technology, ACC Cyfronet AGH, ul. Nawojki 11, 30-950 Krakow, Poland

E-mail: l.dutka@cyfronet.pl

Zofia Mosurska

AGH University of Science and Technology, ACC Cyfronet AGH, ul. Nawojki 11, 30-950 Krakow, Poland

E-mail: z.mosurska@cyfronet.pl

Robert Pajak

AGH University of Science and Technology, ACC Cyfronet AGH, ul. Nawojki 11, 30-950 Krakow, Poland

E-mail: r.pajak@cyfronet.pl

Mariusz Sterzel

AGH University of Science and Technology, ACC Cyfronet AGH, ul. Nawojki 11, 30-950 Krakow, Poland

E-mail: m.sterzel@cyfronet.pl

Tomasz Szepieniec

AGH University of Science and Technology, ACC Cyfronet AGH, ul. Nawojki 11, 30-950 Krakow, Poland

E-mail: t.szepieniec@cyfronet.pl

¹ Speaker

The Polish Grid Initiative commenced in 2009 as a part of the PL-Grid project funded under the framework of the Innovative Economy Operational Programme. The main purpose of this Project was to provide the Polish scientific community with an IT basic platform making use of Grid computer clusters, enabling e-science research in various fields. The Project established a country-wide computing infrastructure, which supports scientific research through integration of experimental data and results of advanced computer simulations carried out by geographically-dispersed teams. The solutions applied in setting up this e-infrastructure ensured integration with other similar platforms around the world.

The Polish Grid Infrastructure created within the PL-Grid project will be continuously maintained and extended as a new, 3-year project (2012-2014) that has started recently: “PLGrid PLUS: Polish Roadmap toward Domain-Specific Infrastructure for Supporting Computational Science in European Research Area”. Its most important task is preparation of specific computing environments – so called domain grids – i.e., solutions, services and extended infrastructure (including software), tailored to the needs of different groups of scientists.

The paper will outline the current status of the basic IT infrastructure in Poland and plans for the future – federated, domain-specific solutions of the extended PL-Grid platform.

*EGI Community Forum 2012 / EMI Second Technical Conference,
Munich, Germany
26-30 March, 2012*

1. Introduction

Computing centers of Krakow (Academic Computer Center Cyfronet AGH), Poznan (Poznan Supercomputing and Networking Center, PSNC) and Warsaw (Interdisciplinary Centre for Mathematical and Computational Modeling, ICM Warsaw) participated in a number of Grid projects, coordinating several of them – this way experience and close links to the Grid communities have been established. In 2007 a PL-Grid Consortium has been formed, which – apart of the three institutions listed above – included computing centers of Gdansk (Academic Computer Centre, CI TASK) and Wroclaw (Wroclaw Centre for Networking and Supercomputing, WCNS), with a goal to provide the Polish scientific community with an IT platform based on Grid computer clusters, enabling e-science research in various fields. It was strongly felt that such infrastructure should be compatible with existing European and worldwide Grid frameworks.

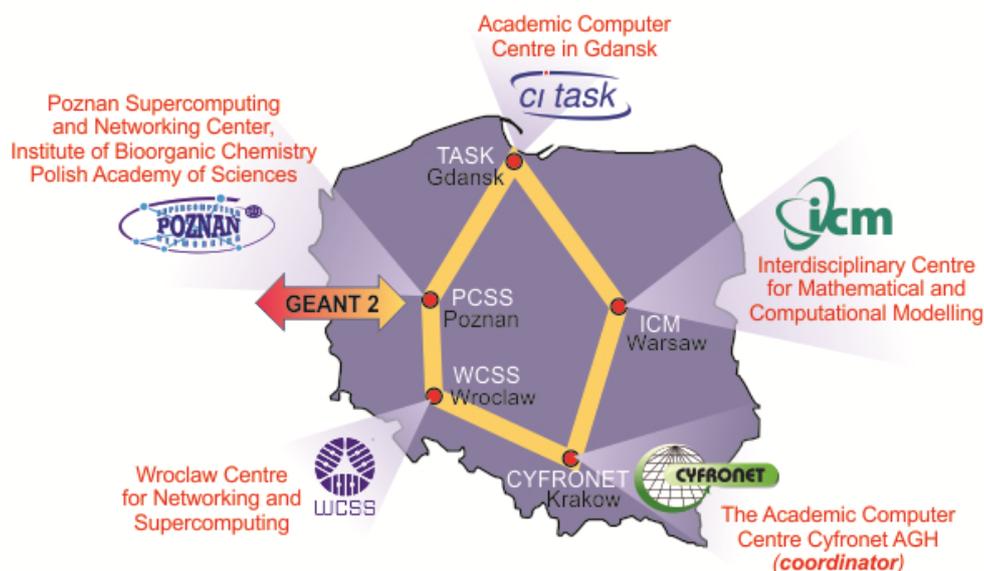


Fig. 1: PL-Grid Consortium members.

2. PL-Grid Project and Polish Grid Infrastructure

The PL-Grid Consortium prepared a project (called PL-Grid) [1] which obtained funding from the European Regional Development Fund as part of the Innovative Economy Program [2] and commenced in March 2009. The total Project budget was 21 M€ (including funding from EC: 17 M€) and Project completion was scheduled for March 2012, with the aggregated offer for the community consisting of 215 TFlops of computational power and 2.5 PBytes of storage resources. The overall goal of PL-Grid was to deploy a persistent national large-scale distributed computing infrastructure for Polish scientists, allowing them to conduct interdisciplinary research on a national scale, and giving them transparent access to international grid resources via affiliated international grid infrastructures.

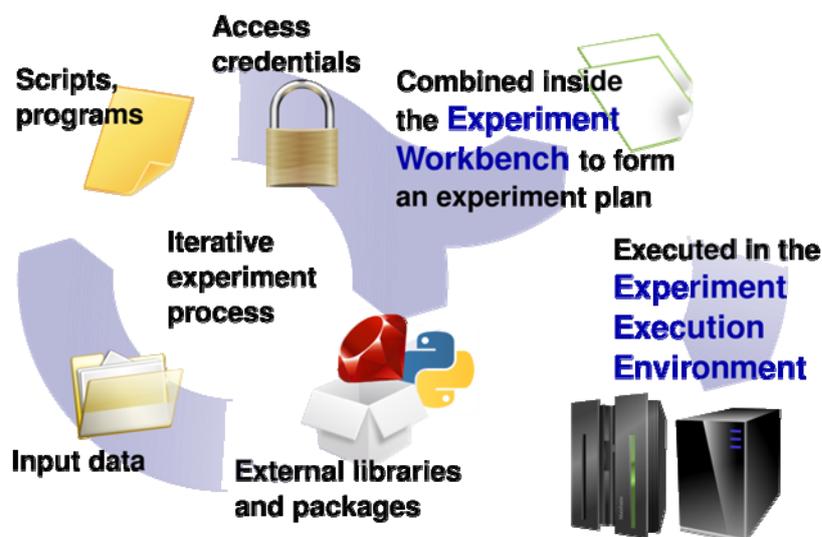
Within the PL-Grid project the Polish Grid Infrastructure (NGI) [3] has been built, which provides the Polish scientific community with an IT platform based on computer clusters, enabling research in various domains of e-Science. The infrastructure supports scientific investigations by integrating experimental data and results of advanced computer simulations carried out by geographically distributed research teams. PL-Grid infrastructure enables Polish scientists to carry out scientific research based on the simulations and large-scale calculations using the computing clusters as well as provides convenient access to distributed computing resources.

Since March 2010 Polish Grid Infrastructure has been a part of the pan-European infrastructure built in the framework of the EGI (European Grid Initiative) project, which aims at integrating the national Grid infrastructures into a single, sustainable, production infrastructure. PL-Grid infrastructure is both compatible and interoperable with existing European and worldwide Grid frameworks.

Creation of the PL-Grid Infrastructure not only extended the amount of computational resources provided to the Polish scientific community by over 230 TFlops of computing power and more than 3600 TBytes of storage space, but – more importantly – facilitated effective use of these resources by providing innovative grid services and end-user tools, as well as continuous technical support.

Close to the end of the Project (in March 2012), the PL-Grid Infrastructure was characterized as maintaining a high level of availability and reliability. The total number of cores within the PL-Grid computing resources exceeded 26 000, the number of jobs per month oscillated between 750 000-1 500 000, and the number of users exceeded 900.

At present, PL-Grid users can choose of three major middleware suites to run their computations: gLite [4], UNICORE [5] and QosCosGrid [6]. To support grid users and administrators, a set of advanced tools and services has also been developed, which can be used e.g. in organization of computational experiments, visualization of applications' results and grid resources management. The end user, or the application developer, can use these modern tools, which will make the complexity of the e-infrastructure transparent.



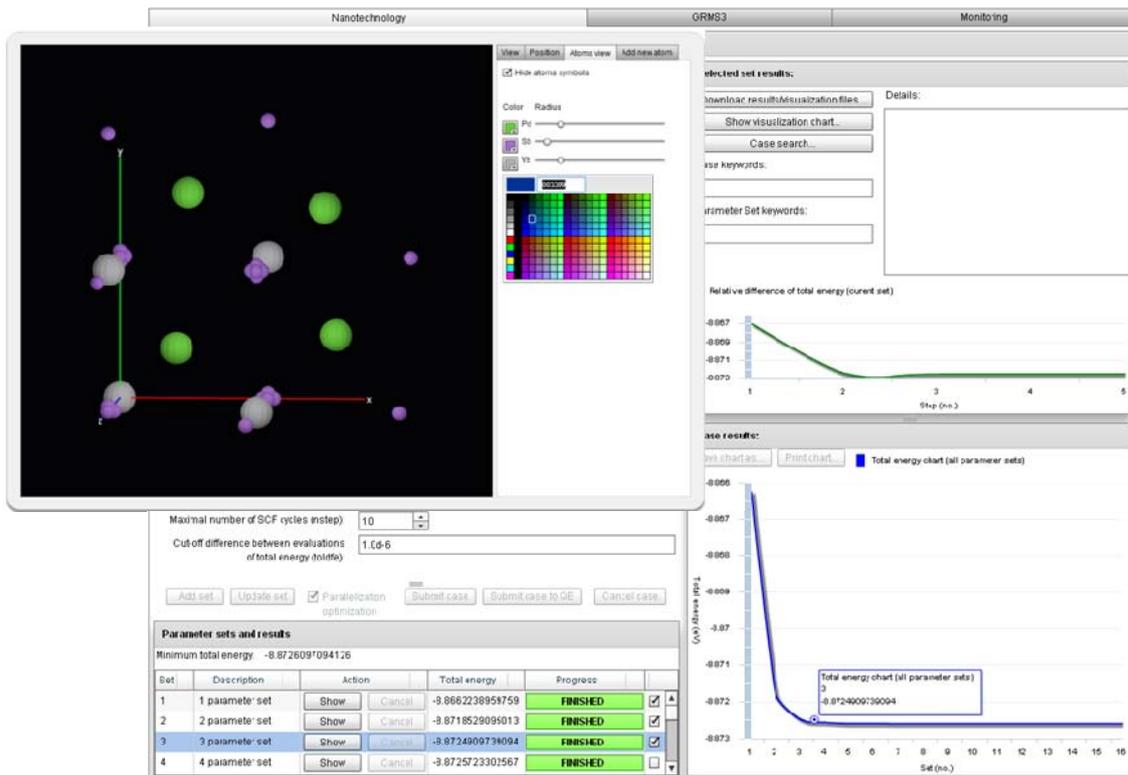


Fig. 2: Examples of tools developed within the PL-Grid project: GridSpace2 Experiment Workbench [7] and Vine Toolkit [8].

In order to fully support various scientific communities, each infrastructure has to provide a variety of scientific software packages for its users. For registered PL-Grid users a set of commercial and freely available scientific packages has been made available. It allows to perform scientific calculations in fields of biology, quantum chemistry, physics, numerical computations and simulations.

To help researchers in exploiting the PL-Grid Infrastructure, a helpdesk system has been introduced, providing users with expert technical support and consultations. In addition, many events have been organized: conferences, seminars and training sessions where various promotional materials were distributed among participants.

2.1 PL-Grid Computing Resources at TOP500

Extension of the hardware resources enabled the computer centres (being PL-Grid partners) to be highly located at the November 2011 TOP500 list (with ranks: ACC Cyfronet AGH: 88, TASK: 279, ICM: 296, PSNC: 298 and WCNS: 360), thus offering substantial amount of resources to the users.

The Zeus cluster from ACC Cyfronet AGH (co-funded from the PL-Grid project funds) took the highest place among the computers from Poland. At present, Zeus operates in the following configuration: 132 TFlops of computing power, 22 TBytes of RAM and 1.8 PBytes of memory storage capacity.

Zeus uses the Scientific Linux (SL) operating system and is assigned to the scientific computations. Zeus has been lately extended with a modern vSMP (versatile SMP)

environment, which offers the ability to dynamically create virtual SMP machines (Symmetric Multiprocessing, i.e. systems with shared memory), adapted to the different needs of users. In addition, within the Zeus cluster, computing nodes with GPGPU cards have been made available for GPU-accelerated computations².

2.2 PL-Grid Book

In mid-2011 the idea was born within the Project management to prepare a set of scientific articles presenting work being done in the Project and publish them in form of a book.

After several months of intensive work, the common effort of 112 authors, 54 reviewers and the editors' team was rewarded – the book titled “Building a National Distributed e-Infrastructure – PL-Grid” was published in March 2012 by the Springer Publisher in the series: Lecture Notes in Computer Science, Vol. 7136. The volume has been made available at the Springer website [9].

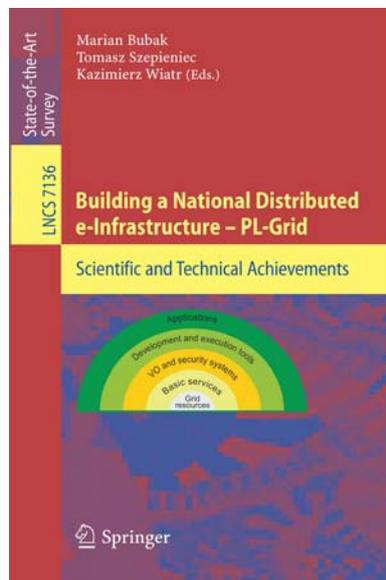


Fig. 3: PL-Grid Book cover.

The book includes 26 articles accompanied by the subject index and a glossary. It describes the experience and the scientific results obtained by the Project partners as well as the outcome of research and development activities carried out within the Project. A very important contribution to the book comes in the form of five articles prepared by end-user teams. They describe how the PL-Grid middleware and infrastructure are applied to solving scientific problems such as grain morphology analysis, seeking common structural motifs in protein families, chemistry computations, molecular dynamics models, LHC computing, and Monte Carlo simulations for the Cherenkov Telescope Array.

The book editors meant for it to serve as an important intellectual resource for researchers, developers and system administrators who work on their own grid infrastructures, and to

² Both these extensions were financed by the POWIEW project funds (<http://wielkiewyzwania.pl/>), but have been made available to users of the PL-Grid Infrastructure.

promote collaboration and exchange of ideas in the process of constructing a common European e-Infrastructure.

3. PLGrid PLUS Project

The PL-Grid project has finished in March 2012; however, the Polish Grid Infrastructure is continuously maintained and extended. The PL-Grid Consortium carries on its activities in the area of grids and clouds for scientific community and has prepared a follow-up project, called “PLGrid PLUS: Polish Roadmap toward Domain-Specific Infrastructure for Supporting Computational Science in European Research Area” [10]. It was funded in October 2011 by the European Regional Development Fund as a part of the Innovative Economy Program, with the total budget of 18 M€ (including funding from EC: 15 M€). The Project is being carried out by the PL-Grid (Polish NGI) Consortium, coordinated by the ACC Cyfronet AGH and will last until September 2014.

3.1 Project Aims

The main objective of the PLGrid PLUS project is to increase the potential of the Polish Science by providing necessary IT services for research teams in Poland, in line with European solutions.

The e-infrastructure requirements of the scientific community working on Big Science problems are highly diversified and depend on the scientific field. To harmoniously support, in terms of IT, both the development of scientific research in the various problem areas and researchers, it is necessary to fit the characteristics of the IT infrastructure to problems being the subject of research. Preparation of specific computing environments – so called domain grids – i.e., solutions, services and extended infrastructure (including software), tailored to the needs of different groups of scientists, will be the most important task implemented within the 3-year-long PLGrid PLUS project. These domain-specific solutions will be created for 13 groups of users, representing strategic areas and important topics for the Polish and international science: AstroGrid, HEPGrid, Nanotechnologies, Acoustics, Life Science, Chemistry and Physics, Ecology, SynchroGrid, Energetics, Bioinformatics, Health, Materials, and Metallurgy. Subsequently, launching support and adaptation services, as well as training for other user groups is foreseen within the Project.

In the framework of the PLGrid Plus project we also plan to:

- extend the existing resources available in the PL-Grid infrastructure by ca. 500 TFlops of computing power and ca. 4.4 PBytes of storage capacity, together with keeping the diversity of these resources: clusters (thin and thick nodes), clusters with GPGPU, SMP machines and vSMP,
- design and start-up of support for possible new domain grids,
- deploy new infrastructure services,
- enable System-level Research,
- deploy Quality of Service system for users by introducing SLA agreement,
- implement Service Level Management procedures,

- deploy Cloud infrastructure for users,
- provide broad consultancy, training and dissemination offer.

3.2 Domain Grids

Domain grids will be enabled in the infrastructure by simple physical or logical separation, depending on the user communities' requirements; however more focus will be put on virtualization and Cloud approach.

Specific tasks of PLGrid PLUS that will be developed in domain grids are highly diverse and include:

- provision of appropriate specialized software and platforms enabling its effective use (e.g. in chemical science, nanotechnology, material science, or energetics),
- mechanisms to ensure computing facilities and data storage for important scientific experiments (e.g. Synchrotron, experiments related to the Large Hadron Collider and Astrophysics),
- development of specific new platforms and integration of new types of tools so that they can be useful in studies on important aspects of social life (including support for medical diagnostics, environmental studies, research related to the spread of noise or computational chemistry).

To accomplish these tasks, the Project relies on broad cooperation with representatives of various disciplines, often grouped in domain consortia. Direct cooperation with them guarantees matching of computing, software, databases and storage services to actual needs.

The activities planned to be realized in domain grids can be grouped into several categories:

- **Integration Services:** national and international levels, dedicated portals and environments, unification of distributed databases, virtual laboratories, remote visualization,
- **Computing Intensive Solutions:** specific computing environments (platforms), adoption of suitable algorithms and solutions, workflows, cloud computing, porting scientific packages,
- **Data Intensive Computing:** access to distributed scientific databases, organization of scientific databases, data discovery, process, visualization, validation, 4th paradigm of scientific research (data intensive scientific discovery),
- **Instruments in Grid:** remote transparent access to instruments, sensor networks,
- **Organizational:** organizational backbone, professional support for specific disciplines and topics.

The new services, foreseen to be developed within PLGrid PLUS, will provide a significant extension of the Polish computing infrastructure, which has been built since 2008 within the PL-Grid project.

3.3 New Infrastructure Services

The new services planned to be developed within the PLGrid PLUS project involve:

- **Cloud Computing for Polish Science** – new computing paradigm foreseen as a natural extension of the current PL-Grid Infrastructure offer,

- **Platform for supporting e-Science**, resulting from the need for an international cooperation between various disciplines of scientific domains,
- **Production infrastructure** oriented towards domain specific services, tools, environments and software packages,
- **Professional support** for specific disciplines and topics important for Polish e-Science,
- **Visualisation** of scientific results via shared infrastructure servers equipped with possibility of binding domain specific visualisation tools.

3.4 Innovative Infrastructure Environment – PL-Grid Extensions

The set of advanced tools and services developed within the PL-Grid project will be used and extended within PLGrid PLUS. The tools can be grouped into several categories:

- **Efficient Resource Allocation:**
 - Grid Resource Bazaar, mobile access to the infrastructure, new security modules and other tools for users and systems administrators, management of users requests,
- **Experimental Workbenches:**
 - GridSpace2 platform extension for supporting new domains and integration with new grid/cloud services,
 - InSilicoLab – integrated environment for chemists and biologists,
- **Tools and Middleware:**
 - Migrating Desktop, VineToolkit and gEclipse tools integration with various PL-Grid domain services,
 - QStorMan Toolkit – extension for domain requirements on optimization of data access,
 - QosCosGrid development continuation,
 - Liferay Portal framework(s) – adoption to specific needs,
 - Helpdesk Portal for the users (specialized versions).

3.5 First Results – Use Cases

It is impossible to mention all the work planned by Project communities in this short report, but below we provide a few use case scenarios, which will be implemented in the PL-Grid Infrastructure. We have chosen three topics from the fields of chemistry, bioinformatics and astrophysics. In all three cases proposed scientific services will be provided via domain gateway(s)/portal(s).

For chemistry, one of the important topics is the analysis of series of molecular geometries (called trajectories and usually coming out of molecular dynamics simulations, MD). Every snapshot of such trajectory usually contains a molecule whose properties are studied, and a set of accompanying solvent molecules. The key part is not only to decide which snapshots from the trajectory should be analyzed in detail, but also, for each snapshot, to minimize the number of solvent molecules taken into account in quantum chemistry computations. The service we aim at developing will perform both functionalities based on user requirements. After the analysis of the trajectory, the service will perform desired quantum chemical computations on

the set of modified snapshots with the computational method chosen by the user. Required functionality is currently under development within the InSilicoLab work environment.

The service planned to be built for bioinformatics aims at providing an automated tool for protein folding. Within the service two main blocks will be integrated. First one will be responsible for modification of the ϕ and ψ angles between amino acids in protein. This part will be based on “fuzzy oil drop” model [11]. Second part, based on one of the molecular dynamics codes (Gromacs, NAMD), will include protein structure optimization (modified previously by “fuzzy oil drop” model) and/or MD simulation. Both parts will be executed sequentially until user’s convergence criteria will be reached.

Analysis of databases of spectral data is one of the key parts of astrophysicists work. Therefore one of the services planned by astrophysics community is development of Virtual Observatory. In the first step this task aims at providing access to already available several databases of spectral data. In the second phase Virtual Observatory will be enriched by data gathered by Polish scientists over last several decades. In addition to access to the databases, Virtual Observatory will provide a set of tools allowing detailed analysis of the spectral data based on user requirements. The part of the spectral data library coming out of observations recorded by Polish scientists will also be accessible by scientists from other countries.

4. Project Impact

Current technologies enable integration of computing resources, data storage systems, unique scientific instruments, applications and databases. An important factor is the virtualization of resources and adoption of methodology of Service Oriented Architecture (SOA), whereby it becomes possible to create a system of services for e-Science. The current level of development of the computing infrastructure in Poland will result in creation of a universal, well tuned, production platform at the national level.

At the international level there are many ongoing projects supporting computing infrastructures for e-Science. One example is EGI-InSPIRE [12], therefore the PLGrid PLUS project has been elaborated in such a way that in the end it shall be consistent with the assumptions of the pan-European structure, whose development is coordinated by the EGI.eu [13] organization.

At the creation of the concept and during subsequent stages of development of the Polish computing infrastructure, one can take advantage of many experiences on both the organization of projects and technical solutions adopted. The primary objective of such infrastructure should be to provide the scientific community with unified computing services enabling implementation of the e-Science model in various fields of science.

This will be achieved within the PLGrid PLUS project through actions such as:

- extending and maintaining the computing infrastructure fully compatible and interoperable with worldwide solutions,
- provision of a common infrastructure for the development of domain grids,
- the use and promotion of standards for generic and domain services,
- integration, testing, assembly and installation of software from leading Polish and international projects in the PLGrid PLUS infrastructure.

5. Conclusions

The presented PLGrid PLUS project fits well with the need of development of an advanced IT infrastructure designed for the implementation of modern scientific research, and providing Polish academic units with capability for collaboration with international research organizations. Consortium members and cooperating research units have the skills necessary both in the operation of infrastructure and its use in research, while international cooperation provides contemporary solutions of an innovative character. Organizational experience of all Project partners allows for high scores of realization of the Project aims.

Pilot introduction of 13 domain grids opens the scope of use of the Project results to various research communities. However, the scope is not limited to the selected domains. By using the developed general services and experience in building the domain ones, the integration of new groups will proceed smoothly and at lower cost.

References

- [1] PL-Grid project, <http://projekt.plgrid.pl>
- [2] Innovative Economy Operational Programme, <http://www.poig.gov.pl/english/>
- [3] PL-Grid Infrastructure, <http://www.plgrid.pl>
- [4] gLite – Lightweight Middleware for Grid Computing, <http://glite.cern.ch/>
- [5] UNICORE – Uniform Interface to Computing Resources, <http://www.unicore.eu>
- [6] QosCosGrid Middleware, <http://www.qoscogrid.org/>
- [7] GridSpace2 Experiment Workbench, <https://gs2.cyfronet.pl/>
- [8] Vine Toolkit, <http://vinetoolkit.org/>
- [9] PL-Grid Book: “Building a National Distributed e-Infrastructure – PL-Grid”, <http://www.springer.com/computer/communication+networks/book/978-3-642-28266-9>
- [10] PLGrid PLUS project, <http://www.plgrid.pl/plus>
- [11] I. Roterman, L. Konieczny, M. Bryliński, *Late stage intermediate in silico model*, in Structure-Function Relation in Proteins, pp. 79-103, 2009
- [12] EGI-InSPIRE project, <http://www.egi.eu/about/egi-inspire/>
- [13] EGI, <http://www.egi.eu/>