Polish Grid Infrastructure for Science and Research

Ryszard Gokieli*, Krzysztof Nawrocki**, Adam Padee†, Dorota Stojda‡, Karol Wawrzyniak§ and Wojciech Wiślicki¶∥

* A. Sołtan Institute for Nuclear Studies, Laboratory for High Energy Physics, Warsaw, Poland, email: gokieli@fuw.edu.pl
** A. Sołtan Institute for Nuclear Studies, Laboratory for High Energy Physics, Warsaw, Poland, email: nawrocki@fuw.edu.pl
† Warsaw University of Technology, Institute of Radioelectronics, Warsaw, Poland, email: apadee@ire.pw.edu.pl
‡ Copernicus Science Centre, Warsaw, Poland, email: dorotas@icm.edu.pl
§ University of Warsaw, Interdisciplinary Centre for Mathematical and Computational Modelling, Warsaw, Poland, email: kwawrzyn@icm.edu.pl
¶ University of Warsaw, Interdisciplinary Centre for Mathematical and Computational Modelling, Warsaw, Poland, email: wislicki@fuw.edu.pl
∥ A. Sołtan Institute for Nuclear Studies, Laboratory for High Energy Physics, Warsaw, Poland

Abstract—Structure, functionality, parameters and organization of the computing Grid in Poland is described, mainly from the perspective of high-energy particle physics community, currently its largest consumer and developer. It represents distributed Tier-2 in the worldwide Grid infrastructure. It also provides services and resources for data-intensive applications in other sciences.

Keywords — Computer networks, distributed computing.

I. INTRODUCTION

Distributed computing was stimulated by human endeavours in the "big science" domain and, at the same time, by hopes of industry to commercialize developments in networking technologies. Some original ideas of organizing computing for science, as e.g. pioneering SETI@home [1], initiated by D. Gedye for search for signals of extraterrestrial civilizations, eventually evolved into coordinated networks supporting research projects of unprecedented scale, as e.g. the Large Hadron Collider (LHC) Computing Grid (LCG) in particle physics [2]. Nowadays, the Worlwide LHC Computing Grid (WLCG) is running on the infrastructure provided by the European initiative Enabling Grids for E-science (EGEE) [3], encompassing European national and regional Grids, coupled to the American Open Science Grid (OSG) and collaborating Grid centres in the Asia-Pacific region (cf. Fig. 1). This infrastructure is shared with a number of smaller projects.

A. The Large-scale and Local Grid Architecture and Middleware

Overall WLCG computing architecture is based on the hierarchical multi-tier model developed by MONARC Collaboration [4], as given in Fig. 2. The top Tier-0 is responsible for storage of raw data coming from the experiment Data Acquisition System, its first off-line processing and distribution of data over Tier-1’s. All data are copied to Tier-1 centres in order to speedup access during processing and ensure storage redundancy. Data reprocessing and higher-level reconstructions of real and simulated data are foreseen to be performed at Tier-1 and Tier-2 levels. Data and physics analyses are normally relegated to Tier-2 and -3 centres, closer to end-users. Tier-2s are powerful enough not only to support local needs but also to complement higher tiers with computing power and storage for more specialized purposes. Tier-2s are not required to provide with massive tape storage.
Fig. 2. Multitier tree-like LCG MONARC model. The DAQ stands for experiment’s Data Acquisition System and T0-T3 for Tier-0-3 levels of data processing. Required disc storage in petabytes (PB) and CPU in Mega SpecInt2000 (MSI2k), and connectivities between levels in bytes per second (bps) are indicated.

Fig. 3. Functional layers in Grid architecture. Physical resources and services in Local Area Networks (LANs) are available on the Wide Area Network (WAN), together with core and gateway services. End-users may access both the services and resources via dedicated portals ensuring easier workflow design.

Fig. 4. Records of numbers of CEs (upper) and jobs (lower) monitored over a year. Drops are seen during Easter in April, summer holidays in August and Christmas in December.

Based on the middleware gLite-3.0 [5], WLCG builds environment with the physical resource layer hidden behind core Grid services (cf. Fig. 3). Essential services, e.g. file and metadata catalogues, replica location, application resource catalogues or workflow engines, are either available directly for users or support other, complex services, e.g. Grid monitoring is used by resource brokers for process management. Intelligent scheduling and resource brokering are now combined in a complex Workload Management service. This system comprises a set of Grid middleware components responsible for the distribution and management of tasks across the Grid.

Information, monitoring and logging are available through the Relational Grid Monitoring Architecture (R-GMA) service, being an implementation of the GMA standard. Fig. 4 presents counting of numbers of CEs and numbers of jobs monitored during one year, as provided by the R-GMA Monitoring Service. Users interact with R-GMA through Application Programming Interfaces available for high-level programming languages.

Crucial for Grid management is the Distributed Grid Accounting System (DGAS). Its purpose is to implement resource usage metering, accounting and account balancing in a fully distributed Grid environment. The last function, i.e. account balancing, is still not in use, because the Grid has not yet come into commercial phase.

Large amounts of data are handled with DCache system [8]. Terabytes of data are distributed over many disc storage nodes but the name space is uniquely represented within a single file system tree. The system has shown to significantly improve the efficiency of connected tape storage systems, through caching, optimizing buffers and scheduled staging techniques. Furthermore, it optimizes the throughput to and from data clients as well as smoothing the load of the connected disc storage nodes by dynamically replicating datasets.

According to recent trend towards Service Oriented Architecture (SOA), Grid components are reengineered as Web-services and published on the net. Dedicated portals are used for designing complex workflows within SOA.

Users are organized in Virtual Organizations (VOs) and managed within VOs through the Virtual Organization Membership Service (VOMS).

Each site is designed in a way typical for GLOBUS-operated [6] Grids (cf. Fig. 5). It contains the Computing Element (CE), normally consisting of the front-end machine playing gatekeeping role and a set of Worker Nodes (WNs). The Storage Element (SE) is a main data container in a site and usually consists of disc matrices managed
Fig. 5. Minimal site structure in computing Grid based on GLOBUS Toolkit. The CE and SE are on public IP numbers and can be accessed directly from WAN.

by dedicated machine. The User Interfaces (UIs), enabling user access to the infrastructure, may be either located on the spot or installed remotely.

B. Connectivity

The Grids of EGEE and WLCG are built on top of the GÉANT network [7] – a collaboration of 26 national and research networks in Europe, led by the DANTE Company. GÉANT Project aims to deliver a quality-of-service gigabit speed backbone network for research in Europe. The GÉANT connectivity scheme is given in Fig. 6. GÉANT has 12 Gbps connectivity to North America and 2.5 Gbps to Japan and to Trans-Eurasia Information Network (TREIN2), thus ensuring collaboration of EGEE with OSG, Japanese National Research Grid Initiative (NAREGI) and Asian-Pacific Grids. The entry point to Polish National Regional Network has bandwidth of 10 Gbps with 4470 B maximum transition unit on the switch.

C. Computing Models of Principal End-users

Four LHC experiments: ALICE [9], ATLAS [10], CMS [11] and LHCB [12], represent the largest consumers of resources on the Grid. Raw data (RAW) coming from real experiment’s DAQ or Monte Carlo simulation are recorded and processed off-line by reconstruction programs giving Event Summary Data (ESD), programs extracting physical variables and providing with Analysis Object Data (AOD), and to further data reduction, selection and filtering, leading to Event Tags (TAG). In addition, derived streams of filtered data at the ESD and AOD levels and specialized data for detector alignment and calibration are recorded and analysed. Grid is also used for the quasi on-line processing of data used for on-line calibration and filtering of data in the framework of the Interactive European Grid (IEG) project [13]. Summary of data-flow parameters and requirements for off-line resources is given in Tab. 1. Such features of computing models as e.g. data flows, number of computing passes at each level, data redundancy, interactions of streams etc., differ between experiments according to specifics of physics processes and detectors to observe them. Important differences in implementations of those models result in many experiment-specific tools and services used by each experimental group.

| TABLE I |
|-----------------|---------|---------|---------|---------|
|                | ALICE  | ATLAS   | CMS     | LHCB    |
| event rate (Hz) | 50     | 100     | 100     | 200     |
| Byte flow (MB/s)| 1250   | 100     | 100     | 25      |
| CPU (kSI2k)     | 30     | 23      | 40      | 13      |
| Storage (PB/y)  | 25     | 17      | 30      | 7       |

models as e.g. data flows, number of computing passes at each level, data redundancy, interactions of streams etc., differ between experiments according to specifics of physics processes and detectors to observe them. Important differences in implementations of those models result in many experiment-specific tools and services used by each experimental group.

II. POLISH TIER-2 INFRASTRUCTURE

Polish physics groups are involved in four LHC experiments and in other high-rate experiments using the Grid, e.g. COMPASS [14] at CERN and ZEUS [15] at Deutsches Elektronen-Synchrotron (DESY). These groups are mostly affiliated at Cracow and Warsaw high-energy physics laboratories.

A. Tier-2 – Tier-1 Connectivity

Large part of computing resources is located in these two cities and, in addition, in Poznań, with no physics groups but where the operator of Polish backbone computing network PIONIER resides. These computing centres constitute Polish distributed Tier-2 connected to the Tier-1 centre at Forschungszentrum Karlsruhe (FZK) in Germany. The PIONIER network interconnects Polish Tier-2 computing centres with a dedicated bandwidth of 1 Gbps.
and provides a bandwidth-splitting DWDM interface to the 10 Gbps backbone Deutsche Forschungsnetz (DFN) (cf. Fig. 7). Polish Tier-2 centres and FZK constitute a

Virtual Local Area Network (VLAN) with the address pool 212.191.227.xxx.

B. Computing Infrastructure at Polish Tier-2 Centres

Computing resources provided by Polish centres to Tier-2 are summarized in Tab. II. The clusters are not homogenous and different computing platforms and server hardware solutions are used. As for the CPU, the 64-bit processor architecture prevails. For example, at ICM the CE is based on AMD Opteron 250 processors assembled in Sun Fire v20 and v40 servers. Rack-mounted WNs and StorEdge SE are interconnected with routable Nortel Baystack 5510-48T and Nortel Baystack 425-24T switches. An automated IPMI tool was developed for efficient cluster management [17].

Just before start-up of the LHC Collider, Polish Tier-2 resources amount to almost 900 CPU and 37 TB of disc space on storage elements. This represents 2.5% and 0.3% of the total resources being 36,000 CPUs and 13.5 PB storage. Similar figures for many Tier-2s are still below requirements mandatory at the LHC running time. Depending on the experiment and its computing model, the CPU and disc storage are expected to be higher a couple of times and an order of magnitude, respectively. Resource doubling every year, i.e. faster than the Moore’s Law, is planned during LHC operation time (cf. Fig. 8). Suitable investment for Polish Tier-2 is underway.

Each site on the Grid is permanently monitored by Grid Operational Centre and its status, both physical and functional, is made available on the Academia Sinica Web Host [16]. Example plots showing actual numbers of CPU usage, numbers of jobs and disc storage usage are given in Fig. 9. More detailed insight into disc storage distribution over VOs is displayed in Fig. 10.

C. VO Support and Resource Sharing

Tab. III shows resource sharing over virtual organizations in Polish Tier-2 clusters. Besides already mentioned VOs, one finds VOs related to the Baltic Grid project (BALTGRID), biology (BIOMED), chemistry (COMPCHEM), internal EGEE development VO (DTEAM), the EU-China Grid initiative (EUCHINA), Central European Federation VO (VOCE) and a couple of minor VOs. In this report we do not distinguish between EGEE VOs, official global VOs, official local VOs and others, although these distinctions are important from managerial viewpoint.

Inspection of the table reveals differences in local policies of resource allocation to VOs. There is no yet official regulation for these policies and resource allocations are usually negotiated between user communities and site managements. In order to ensure optimal CPU usage, fair share system is normally implemented in queues, unless

| site name             | CPU available | Storage on SE (TB) |
|-----------------------|---------------|--------------------|
| AMD64.PSNC.PL         | 222           | 4.3                |
| CYFRONET-IA64         | 34            | 0.3                |
| CYFRONET-LCG2         | 274           | 21.3               |
| egee.man.poznan.pl    | 132           | 5.2                |
| WARSAW-EGEE           | 224           | 5.9                |
| Total                 | 886           | 37                 |
Fig. 9. Monitoring plots provided by Grid Index Information Service tool. In the upper panel, total number of Working Nodes’ CPUs available for users is shown in red and free CPUs in blue. The middle panel presents numbers of running (green) and waiting (blue) jobs. Disk storage on Storage Elements available (green) and used (blue). Temporary connectivity drop is seen as a dip in two upper figures.

III. OPERATIONS OF POLISH TIER-2

A. Daily Operations and User Support

European Grid is supposed to provide a permanent and reliable infrastructure for research and science. The hardware is run by staff of participating institutions and is under local responsibility. Both central and local services are run by dedicated groups of people and are shared between partners, depending on their competence, size and needs of regional scientific groups. The case of large LHC collaborations, consisting even of thousands of researchers, somewhat violates this scheme. While the lowest-order Tier-3 nodes are traditionally situated in scientific institutes, Tier-2’s and Tier-1’s are often run by large regional or national computing centres, capable of fulfilling operational requirements.

Daily operations are monitored by Grid Operational Centre (GOC) located at UK [18]. The GOC is responsible for coordinating the overall operation of the Grid. It acts as a central point of operational information such as configuration information and contact details. The GOC has responsibility for monitoring the operation of the Grid Infrastructure as a whole, devising and managing mechanisms and procedures which encourage optimal operation of the Grid, and working with Local Support Groups to assist them in providing the best possible service while their equipment is connected to the Grid.

Basic functionality of services is regularly tested and monitored using Site Functional Tests (SFTs). The SFT uses a small test job that runs at each site and determines the availability of the main Grid functions. Similarly, the Grid Status Monitor retrieves information published by each site about its status. Their use and subsequent triggering of follow-up action is supervised by the Core Infrastructure Centre (CIC) on Duty staff raising operational tickets against sites to resolve observed problems.

Two-level ticketing system is incorporated. Ticket flow diagram is displayed in Fig. 11.

Global Grid User Support (GGUS) portal run by FZK is a principal entry point for all sorts of tickets [19]. Daily ticket operations, including initial recognition of the type of the problem, opening and assigning ticket to supporters, directing it to Local Support units, care about timely solving and contact with users, is a duty of Ticket Processing Management Group (TPM). The TPM works in a shift system, 5 days a week, 8 hours a day. SFT failure tickets are send to GGUS and redirected automatically to Local Support units and from there to the front-line supporters at sites. Another type tickets are issued by end-users who may encounter many kinds of problems with a system or with applications. User tickets, depending on the type of the problem, are either solved by dedicated group of supporters asked by TPM from GGUS, or redirected to Local Support units and solved there.

There is at least one Local Support unit in Federation. For Polish Tier-2, ticketing tool [20] is running using One-or-zero portal software [21].
TABLE III
VIRTUAL ORGANIZATIONS SUPPORTED BY POLISH TIER-2 CENTRES

| Site                | VO       | Normalized CPU share (%) |
|---------------------|----------|--------------------------|
| AMD64:PSNC.PL       | ATLAS    | 73.9                     |
|                     | BALTICGRID | 6.9                     |
|                     | CMS      | 4.1                      |
|                     | COMPCHEM | 11.5                     |
|                     | DTEAM    | 0.2                      |
|                     | OPS      | 0.1                      |
|                     | VOCE     | 3.3                      |
| CYFRONET-LCG2       | ALICE    | 14.3                     |
|                     | ATLAS    | 26.8                     |
|                     | BALTICGRID | 6.6                     |
|                     | BIOMED   | 2.0                      |
|                     | COMPCHEM | 25.2                     |
|                     | DTEAM    | 0.3                      |
|                     | EUCHINA  | 18.9                     |
|                     | LHCB     | 0.4                      |
|                     | OPS      | 0.1                      |
|                     | VOCE     | 3.5                      |
| CYFRONET-IAG64      | ALICE    | 14.3                     |
|                     | ATLAS    | 26.8                     |
|                     | BALTICGRID | 6.6                     |
|                     | BIOMED   | 2.0                      |
|                     | COMPCHEM | 25.2                     |
|                     | DTEAM    | 0.3                      |
|                     | EUCHINA  | 18.9                     |
|                     | LHCB     | 0.4                      |
|                     | OPS      | 0.1                      |
|                     | VOCE     | 3.5                      |
| egee.man.poznan.pl  | ALICE    | 0.0                      |
|                     | ATLAS    | 83.3                     |
|                     | BALTICGRID | 1.6                     |
|                     | CMS      | 0.1                      |
|                     | COMPCHEM | 12.30                    |
|                     | DTEAM    | 0.1                      |
|                     | EUCHINA  | 0.1                      |
|                     | LHCB     | 0.0                      |
|                     | OPS      | 2.6                      |
|                     | VOCE     | 2.1                      |
| WARSAW-EGEE         | ALICE    | 11.9                     |
|                     | CMS      | 54.1                     |
|                     | COMPASS  | 0.1                      |
|                     | DTEAM    | 0.1                      |
|                     | LHCB     | 32.0                     |
|                     | OPS      | 0.0                      |
|                     | VOCE     | 2.1                      |

IV. INTERACTIVE GRID INFRASTRUCTURE

Bulk of applications in experimental particle physics and other sciences needs a high-throughput batch processing of large amounts of data. For a number of applications, however, often interaction with intermediate results or fast response to a well defined computational problem is desired. This sort of applications, called (quasi-) interactive, draws an attention of Grid community since almost beginning. In Poland, involvement in deploying such applications on the Grid and building appropriate tools and infrastructure for them, dates back to the CrossGrid Project [22] and is nowadays continued in the framework of IEG project [13].

The IEG resources are split into two separated infrastructures:
- the production infrastructure aimed at providing computing and storage resources for the end-users running scientific applications,
- the development infrastructure being fully independent of the production and aimed at supporting the Project software development, the test of new middleware and its rollout process. As such this infrastructure does not provide a service as stable and reliable as the production. Development sites may also be occasionally reconfigured with specific setups to evaluate or validate software components. Currently, the production infrastructure provides with 300 CPU cores and 8 TB disc storage, located in 8 computing centres in Europe, with a considerable contribution of three Polish computing centres.

The IEG supports the following interactive applications:
- Ultra Sound Computer Tomography.
- Medical Applications on Brain Images.
- Flood Forecasting application. This application was first deployed on the CrossGrid testbed.
- Visualisation of Baltic Wave Model.
- Evolution of pollution clouds in the atmosphere. This application was first deployed on the CrossGrid testbed.
- ATLAS online monitoring and calibration system. This application is related to ATLAS experiment at LHC but it does not run on the WLCG infrastructure.
- Analysis of Maps of Cosmic Microwave Background.
- Visualization of Plasma in Fusion Reactors. This application runs also in less interactive mode on the EGEE infrastructure.

V. TRAINING, DEMONSTRATION AND DIFFUSION ACTIVITIES

Being a global-scale initiative with large investment and social impact, computing Grid needs associated actions attracting and training users, and explaining the newest Grid technology to wider public.

Training is provided by organizing courses, normally given by staff members of academic partner institutes of large Grid projects (cf. e.g. refs [13], [23]), and using

Fig. 11. Tickets flow scheme. Tickets issued by SFT failures are put onto GGUS and automatically directed to Local Support units and solved there. End-users’ tickets can be either put to GGUS or to Local Support units. These tickets are solved either by local staff or by supporters from dedicated support groups, depending on the type of the problem. Daily ticket handling is performed by TPM group. Solved tickets are stored in a database.
dedicated training infrastructure. Courses are attended by students of universities at the engineering, M.Sc. and Ph.D. levels, practicing scientists of informatics, natural sciences and engineering, developers and managerial staff from commercial companies.

Dissemination of Grid technology is assured by its active promotion in communities of the actual and potential users by using press news and specialized publications, participating in conferences of possibly wide spectrum of subjects, and through media [23], [24]. The on-demand TV, being nowadays a distributor of knowledge, may shortly become Grid’s customer exploiting its computing power for own programme casting and production.

VI. OUTLOOK AND PERSPECTIVES

The WLCG is an example of well developed Grid infrastructure for science. To large extent, however, it was designed for specific needs of experimental particle physics where high-throughput, massive, asynchronous data-intensive processing of segmented data is needed. Parallel processing and using Message Passing Interface software is rather rare. Occasional usage of such gateway services like application resource catalogs or workflow engines is not a common practice. Workflow management is quite often done semi-manually and resource brokerage is still far from being adaptive and autonomous. Prospective line of development guides toward SOA where applications are distributed over the network and are accessible from everywhere as services. Data are going to be virtualized (dCache) and workflows are dynamically designed and redesigned according to needs (cf. e.g. PGrade portal [25]).

After fulfilling LHC commitments, Polish Tier-2 should evolve towards new computational paradigm where complex reserach scenarios are executed in response to external events (e.g. rapid weather change) in closed loops with instruments and humans (interactivity). On-demand allocation of computing resources should ensure solving identified important problems.

From commercial perspective, other aspects of Grid should be underlined. For pure research, robustness and security for applications is not really critical unless facilities are being built. Scientific groups gladly relegate security for applications is not really critical unless facilities are being built. Scientific groups gradedly relegate operations to commercial entities. But this practice often results undesirably for science because business itself is only interested in research in case of visible income. Interesting game between two aspects of distributed, large-scale computing: the economic and the research, is in front of us.

REFERENCES

[1] D.P. Anderson, J. Cobb, E. Korpela, M. Lebofski and D. Werthimer, "SETI@home - An experiment in public-resource computing", Communications of the ACM, vol. 45, pp. 56-61, November 2002, (http://setiathome.ssl.berkeley.edu/)

[2] M. Lamanna, "The LHC computing grid project at CERN", Nuclear Instruments and Methods in Physics Research A, vol. 534, pp. 1-6, November 2004, (http://lcg.web.cern.ch/LCG/)

[3] EGEE is the project funded by the European Commission in the 6th Framework Programme under contract INFSO-RI-031688 (http://www.eu-egge.org/)

[4] http://cern.ch/MONARC/

[5] E. Laure et al., "Programming the Grid using gLite", EGEE-PUB-2006-029, submitted to Computational Methods in Science and Technology

[6] B. Sotomayor and L. Childers, Globus Toolkit 4: Programming Java Services, Morgan Kaufmann (December 16, 2005), (http://www.globus.org/)

[7] http://www.geant.net

[8] M. de Ruise et al., "dCache, the Book", (http://www.dcache.org/ manuals/Book/)

[9] ALICE: B. Alessandro et al., "ALICE: Physics Performance Report, volume II", Journal of Physics G, vol. 32, pp. 1295-2040, September 2006, (http://aliceinfo.cern.ch/)

[10] ATLAS: A. Di Ciacco et al., "The ATLAS experiment at CERN: Two years before the start of the data taking", 9th ICATPP Conference on Astroparticle, Particle, Space Physics, Detectors and Medical Physics Applications, Villa Erba, Como, Italy, 17-21 Oct 2005, (http://atlas.web.cern.ch/Atlas/index.html)

[11] CMS: K. Hoepfner et al., "Physics with CMS - potential and challenges", Particles and Nuclear International Conference (PANIC 05), Santa Fe, New Mexico, 24-28 October 2005, American Institute of Physics Conference Proceedings, vol 842, pp. 1094-1096, July 2006, (http://cms.cern.ch/)

[12] LHCb: T. Nakada et al., "Status of the LHCb experiment", Conference on Physics at LHC, Vienna, Austria, 13-17 July 2004, Czech Journal of Physics B vol. 55, pp. 405-418, July 2005, (http://lhcb.web.cern.ch/lhcb/)

[13] IEG is the project funded by the European Commission in the 6th Framework Programme under contract FP6-2005-Infrastructures-7, (http://dissemination.interactive-grd.eu/)

[14] COMPASS: E. S. Ageev et al., Nuclear Physics B, vol. 765, pp. 31-70, January 2007, (http://wwwcompass.cern.ch/)

[15] H1 and ZEUS: G. Brandt et al., "Exotic physics at HERA", 33rd International Conference on High Energy Physics (ICHEP 06), Moscow, Russia, July/August 2006, (http://www-zeus.desy.de/)

[16] http://goc-grd.smaica.edu.tw/gstat/index.html

[17] http://www.polgrid.pl/modules.php?name=Documentation&content=contact

[18] http://goc-grd-support.ac.uk/gridsite/gocmain/?page=2

[19] https://ger.itk.de/pages/home.php

[20] https://helpdesk.polgrid.pl/

[21] http://www.onecorzer.com/

[22] http://www.crossgrid.org/

[23] http://www.polgrid.pl/

[24] Academic Scientific TV distributed on the Internet, (http://atv.scri-fzk.de)

[25] http://www.ipds.sztaki.hu/pgrade/