Implementation of Service Level Management in PL-Grid Infrastructure

Tomasz Szepieniec, Małgorzata Tomanek, Marcin Radecki, Magda Szopa, and Marian Bubak

AGH University of Science and Technology, ACC Cyfronet AGH,
ul. Nawojki 11, 30-950 Kraków, Poland
t.szepieniec@cyfronet.pl

Abstract. One of the key concept of Grids – as stated in their definition – is providing nontrivial quality of service. However, there are many users whose specific needs for guaranteed resources (depending on the type of their applications and research schedule) are not satisfied. This fact is the main motivation for introducing Service Level Management in the Polish National Grid Infrastructure (PL-Grid). The paper presents a model of SLAs with quality properties required in HPC Grid. In order to properly manage these SLAs a negotiation process has been proposed. Finally, an implementation of the procedures and tools needed to support this processes is summarized.

Keywords: Service Level Management, Service Level Agreement, Grid.

1 Introduction

One of the key concepts of Grids – as stated in their definition [1] – is providing nontrivial quality of service. Several infrastructures have recently reached "production" quality, according to their claims. In parallel, there are several ongoing research projects to work out appropriate business model [7] and define related Service Level Management processes [4,5]. However, providing resources for scientific teams must take into account specific guarantees which are currently not available to users [8]. Such special needs might be related to the scientific teams' particular type of applications or, more commonly, their schedule of research actions. The main goal of introducing SLM in PL-Grid was to maximize scientific results achieved with the infrastructure. This is done by providing resource quality and availability guarantees for scientific groups, facilitating their research.

The primary challenge on the way to implementing this idea comes from the fact that PL-Grid provides services related to computations and data management on a large scale. In order to expose such services to users, vast amounts of elements related to computational and storage resources are needed. The crucial fact is that those elements are distributed and spread across different administrative domains. Each resource owner can be a provider in terms of SLM, and may introduce specific limitations and requirements related to SLA management. The resources themselves are of various types and thus their support

M. Bubak, T. Szepieniec, and K. Wiatr (Eds.): PL-Grid 2011, LNCS 7136, pp. 171–181, 2012.

levels may differ. Grid users, who are customers in terms of SLM, may have access to any grid resource on a technical level. The actual allocation of resources to customers is subject to Service Level Agreements (SLA). SLAs provide the users with certain guarantees related to the capacity and conditions of use of specified service elements. Detailed information is necessary to select suitable resources and then sign (conclude) an SLA. The complexity of the problem comes from the fact that, in typical scenarios, there are many providers for a user to contact and negotiate an SLA with, while at the same time each provider may provide resources to many (groups of) users. Therefore, SLM in a Grid forms a many-to-many network joining providers and clients. Two parties signing an agreement must be able to use the same language in terms of procedures, obligations and expectations. The above issues translate into requirements faced by the SLM model and any suitable support systems.

In this article we present an attempt at instituting service level management in the PL-Grid infrastructure. At the same time, we contribute to the discussion on how SLM should be structured for any national grid infrastructure or any federated e-infrastructure. Section 2 provides a short introduction to SLM which is subsequently mapped to PL-Grid actors. In Section 4 we list specific features of the SLA in terms of its lifecycle and quality metrics used in PL-Grid. Subsequently, we provide details on how the SLM-related processes, particularly the negotiation process, are implemented in PL-Grid operations. A summary of the tool environment is given in Section 7. The article ends with a description of related work and a summary.

2 Benefits of SLM

Service Level Management (SLM) is one of the processes defined within IT service management (ITSM), which is a discipline aiming at efficient delivery of quality IT services in a constantly changing environment. The standard process framework of ITSM is specified in ISO/IEC 20000 [15]. Best practices relevant to this field are listed in a set of handbooks collectively known as the IT Infrastructure Library (ITIL) [14]. The processes described there have been implemented by hundreds of IT providers worldwide. SLM includes the following actions:

- defining a catalog of IT service offerings,
- specifying services along with their service level options,
- negotiating and signing Service Level Agreements (SLAs) with customers,
- ensuring that each SLA is mirrored by Operation Level Agreements (OLAs) with suitable suppliers,
- monitoring and reporting on the fulfillment or violation of SLAs [13].

SLM is a vital part of customer-oriented provisioning of quality-aware IT services. The goal of SLM is to introduce a relationship between IT service providers and their customers. Typically, those two communities are separate and speak different languages. Thus, SLM promotes common understanding of customers' expectations, mutual responsibilities, communication channels as well as any constraints.

3 Actors and Relations in SLM Process

In classical SLM, relations are specified between a customer and a provider. The provider often relies on other (second-level) providers to deliver the agreed-upon services. In this section we specify how these roles are assigned in SLM for PL-Grid, in which case three types of actors are defined:

- *Customer: User Groups or Virtual Organizations* – sets of individual users who cooperate in order to achieve some scientific results; in PL-Grid user groups are registered in the PL-Grid Portal [16] and can apply for resources and access to specific domain services to support their research projects.
- *PL-Grid Operations Center (OC)* – integrates a robust set of technical services and provides a single point of contact for user groups and virtual organizations to apply for those services; it cooperates with 2^{nd} level providers to collect resource requests from customers, provide all the necessary tools and processes to monitor the use of resources and mediate in case of service delivery problems.
- *Individual Sites* – infrastructure providers who offer access to computing and storage resources through well defined protocols. Sites retain the right to manage their own resources, including allocating them to specific customers in accordance with site-specific procedures. Moreover, most sites do not provide the entire set of technical services required to support grid computing.

In order to introduce SLM, it is first necessary to define the relation between the provider and the customer, and reach an agreement upon services that are needed by customers. The crucial decision regarding our framework was to appropriately structure this agreement. The difficulty comes from the fact that services are delivered by sites, while the grid model assumes some level of unification under the umbrella of PL-Grid. It is also important that customers need

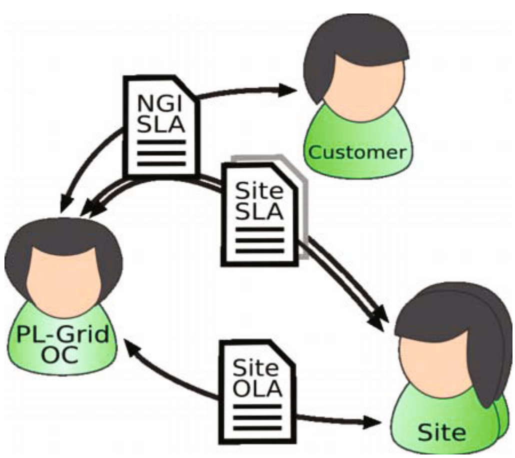

Fig. 1. SLAs and OLAs defining relations between actors in Grids

a single contact point for requesting resources, which, in our case, is the PL-Grid OC. The solution chosen for PL-Grid customizes the standard hierarchical structure of providers by adding some level of transparency in the SLA/OLA framework which enables convenient communication of customers with actual service providers. This resulted in three kinds of agreements, shown in Fig. 1, while the PL-Grid SLM framework is built on agreements defined below:

– *NGI Service Level Agreement* – the main SLA that is concluded between a customer and PL-Grid, i.e. the National Grid Initiative (NGI). It specifies the set of services that need to be delivered for a given customer, including service level descriptions and support information. The agreement must also include references to Site SLAs that provide the actual services.
– *Site Service Level Agreement* – concluded between a site (as a provider) and PL-Grid (representing the customer); it specifies the details of services, including instances of service elements. This agreement is linked to a specific NGI SLA.
– *Site Operations Level Agreement* – concluded between a site and PL-Grid to specify integration details and services that both parties need to deliver. This agreement is signed when a site joins PL-Grid; it also provides a base for integrated operations and efficient Site SLA settlement process.

It is important how the PL-Grid layer affects the guarantees offered to customers. In principle, any resource allocation and guarantee can be given by sites only based on their ability to manage their own resources. PL-Grid OC does not operate any resources directly. However, PL-Grid may upgrade customer guarantees by brokering Site SLAs to deliver the agreed-upon service level even in cases when a individual site violates its own Site SLA. The main benefit of maintaining an NGI SLA in addition to single-site SLAs lies in this increased level of guarantees. Other benefits are organizational in nature and related to the fact that customers can share a single point of contact for many sites.

This is a simplified model built upon the concepts presented in [13]. In the future, the model may be extended to support international VOs by adding an extra layer of hierarchy which can materialize as an international body, e.g. EGI.

4 SLA Life Cycle

Having identified the main types of SLAs, in this section we will focus on how the SLA state is described and how state transitions are performed during the SLA lifetime. To keep the description short we will focus on Site SLAs and then summarize how the NGI SLA differs from Site SLAs. Site OLAs are standard operation agreements – thus, we don't need to specify them in detail.

The state of a Site SLA is characterized by the pair $\alpha = (\alpha_1, \alpha_2)$. Individual elements are later referred to as sub-states. Some of the states make sense only if a previous state in the tuple has a specific value.

The first sub-state, α_1, is also called the *main sub-state*. Possible values of α_1 are as follows: `PROPOSAL`, `PREAGREED`, `AGREED` or `CANCELLED`. This sub-state

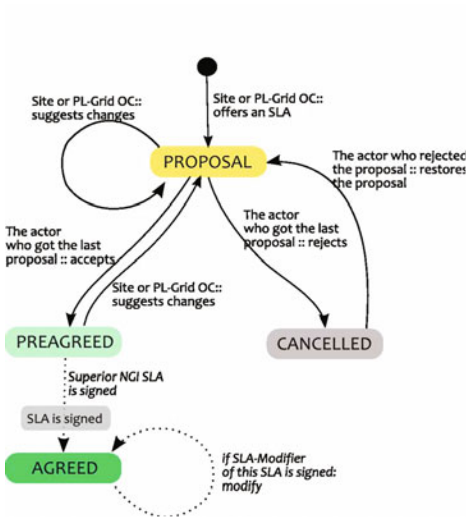

Fig. 2. Sub-states of SLA: main sub-states (α_1)

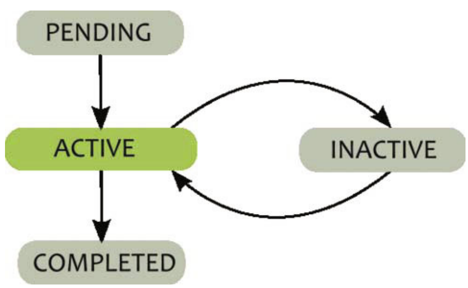

Fig. 3. Sub-states of SLA: activity sub-states (α_2)

determines whether the SLA is binding. If so, the sub-state is `AGREED`. To reach this sub-state both parties need to accede to each other's conditions and also the related NGI SLA needs to be in the `AGREED` state. Once the first condition is fulfilled, α_1 is switched to `PREAGREED`. The `PROPOSAL` value means that the SLA is in the process of negotiation. Finally, when an SLA is cancelled, its state switches to `CANCELLED`. The allowed transitions are presented in Fig. 2. Renegotiation is possible in the model by negotiating a new SLA (whose proposal can be based on the currently agreed version) and cancelling the current one.

The α_2 sub-state, also called the *activity sub-state*, is used to define the current status of SLA in relation to the time when the user utilizes the resources covered by the SLA. This sub-state is valid only if α_1 equals `AGREED`. When α_2 is `ACTIVE`, user is able to use the resources on demand (according to the SLA). Otherwise, this sub-state can be `PENDING` – which occurs before the first `ACTIVE` period, `INACTIVE` – between two `ACTIVE` periods or `COMPLETED` – following the last active period. These states are configured on the basis of the timeframe defined in the

SLA and additional administrative factors that may result in inactivating an otherwise ACTIVE SLA.

The NGI SLA lifecycle follows quite similar rules, however transitions between states depend on the dependent Site SLAs. Moreover, in the NGI SLA α_1 cannot be set to PREAGREED. Setting α_1 to AGREED is possible only when all associated Site SLAs are AGREED or CANCELLED. α_2 is set according to the following rules, applied in the presented order:

1. ACTIVE if at least one associated site SLAs is also ACTIVE;
2. INACTIVE if at least one associated site SLAs is INACTIVE;
3. set to the state which is shared by all dependent site SLAs.

It can be easily shown that it is not possible for the first rules to not apply and for Site SLAs to differ in their sub-states.

5 SLA Negotiation Process

The conceptual model of the SLA negotiation process is shown in Fig. 4. In the case of hierachical SLA models which have been introduced in PL-Grid, we can distinguish two types of related negotiation processes:

– NGI SLA Negotiation – initiated upon customer request. The SLA is negotiated with PL-Grid represented by the Operator. Each actor may change the SLA proposal but any modification must be confirmed by the other side.

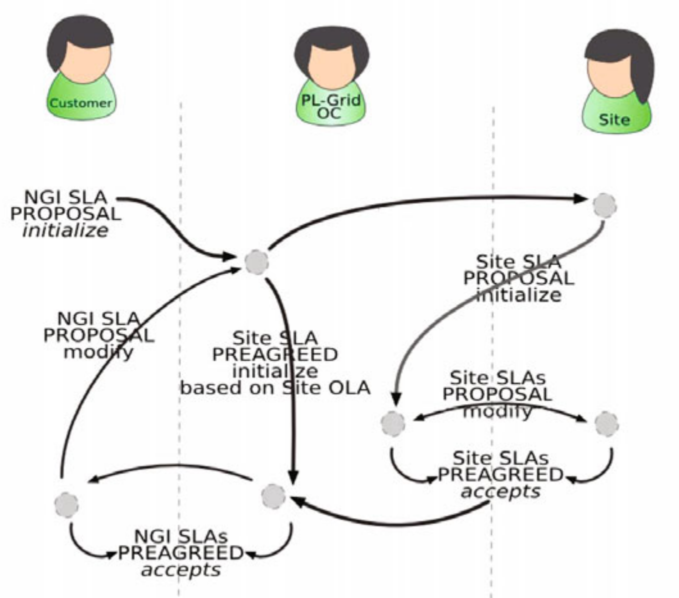

Fig. 4. SLAs negotiation process in PL-Grid

– Site SLA Negotiations – initialized as the site's answer to a customer request. The Site SLA is negotiated between a site and NGI. Agreement is reflected by setting the Site SLA status to `PREAGREED`. Several Site SLAs may be involved in negotiations at the same time, each covering a part of the customer's request and each realized by different providers.

The key task in this process falls to the PL-Grid Operator who brokers the site offers and aggregates them into a single SLA. The SLA is understood as a set of dependent Site SLAs; moreover, the SLA itself can contain an additional set of quality-related properties. The process is shaped in such a way that the act of accepting the NGI SLA is associated with final acceptance of the underlying Site SLAs. Without an accepted SLA those Site SLAs are not binding. Moreover, customers can review details of Site SLAs and propose changes to them in the process of SLA negotiations.

6 SLA Quality Properties

The usability of the proposed SLM model strongly depends on how SLA details are formulated and, in particular, on what resource quality properties are expressed in such agreements. The main requirement in defining them is that they should be both measurable and easy to verify in post-mortem analysis. Additionally, they should be configurable, which means that there should exist a method of translating these properties into service configuration. The set of properties included in SLA should not be too large, but should remain meaningful for the parties. Thus, they should have practical meaning for users as well as for providers. As different users' needs may vary, the properties included in a particular SLA can be selected from a larger set of available properties. If no properties are specified, a minimum service level should be guaranteed by some general documents. It is important to define the responsibilities for quality properties, each of which may refer to:

– provider – to define details of the offer and guarantees which the provider agrees to deliver,
– customer – to define limits within which the offer is valid or to restrict malpractice.

The responsibility of both actors related to a single measurable value may create conditions under which utilization of resources is expected and guaranteed.

While dealing with multi-level SLAs we need to distinguish between the following types of quality properties:

– *qualitative* – defining services, service elements or available service features;
– *quantitative summable*, called *capacity metrics* – measurable properties that specify the parameters of a service or service level; for metrics of this kind the NGI SLA value should not exceed the sum of values in the Site SLAs; usually these properties relate to the capacity of some resource;

– *quantitative threshold*, called *quality metrics* – measurable properties that specify parameters of services or service levels; for such properties the NGI SLA value should not exceed the minimum value listed in the Site SLAs unless service elements are configured to increase the overall threshold; usually these properties are related to service level limits, deadlines, etc.

Table 1 defines both capacity and quality metrics describing the main types of resources applied in PL-Grid. In the case of computational resources, the capacity is expressed by "total normalized wall clock time", which limits the aggregated use of computational power defined in the SLA. Normalization is performed on the basis of benchmarks run on each type of available machines. Wall clock time values are used because concurrency is limited by the number of so-called machine slots, thus minimal computing efficiency for each job should be guaranteed and reflected in the benchmark. Other properties listed in the table define conditions for using computational power, including minimum machine configurations and time-dependent distribution of computations.

Similarly, a list of metrics for storage resources is provided. Not surprisingly, the capacity metric here is defined as "soft quota". Quality metrics specify the types of resources, assessment parameters, additional restrictions and extra features (e.g. backup processes). Apart from specific properties, there are several metrics that are common to all types of resources. Those metrics, presented in

Table 1. SLA quality properties for key types of resources

Name [unit]	Description/Notes	Resp.
Computational resources		
Total normalized wall time [normalized hours]	Sum of allocated time for user jobs in the queue system. The sum is normalized according to slot efficiency	provider
Max wall clock time for a single job [hours]	Limits the wall clock time for a single job. When this value is exceeded the RP is allowed to abort the job.	customer
Max single job parallelism [no. of slots]	Peak number of slots used by a single parallel job.	customer
Max slots used concurrently [no. of slots]	Number of slots that are used at the same time	customer
No. of slots reserved	Number of slots available for computation with no delay	provider
Storage Resources		
Soft Quota [GB]	Maximum volume of stored data that should not be exceeded	provider
Type of storage	Disks or tapes	provider
Hard Quota [GB]	Maximum volume of stored data that may not be exceeded	customer
Grace Period [days]	Maximum time allowed for soft quota breaches	customer

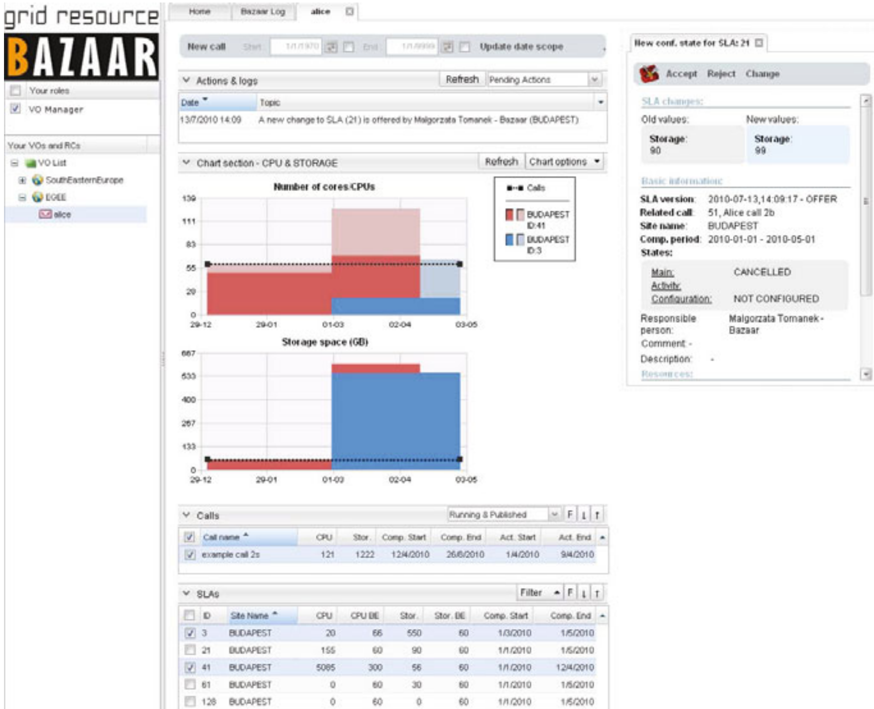

Fig. 5. Grid Resource Bazaar graphical user interface

the last section of the table, are mainly inherited from EGEE SLA and address service availability and related staff support.

7 Tool Support

The main requirement regarding SLM was to simplify the communication related to the process of SLA negotiation. Additionally, managing many SLAs requires support related to managing site capacity over time. This motivation inspired the development of the Grid Resource Bazaar [10] – a web platform which, once integrated with other PL-Grid operational tools, provides a convenient SLA and OLA negotiation framework. This collaborative tool simplifies communication by implementing communication patterns crafted according to the SLM model. The Bazaar GUI was designed for easy SLA handling, using complexity management techniques. As can be seen in Fig. 5, the portal is organized in the form of a dashboard with resource allocation views for customers and providers. The resources are visualized on a chart and SLAs are listed beneath. The user can manage SLA proposals and examine the influence of new SLAs on the resources.

Bazaar is integrated with the PL-Grid Portal authorization service, therefore in Bazaar users can negotiate resources only on behalf of user teams, which can be defined in the PL-Grid Portal. Information regarding SLA status is also made available via APIs which enable integration with other services, including automatic site configuration tools. Additionally, SLA metric monitoring is being integrated with the PL-Grid monitoring infrastructure and an accounting module is under development, but these features are out of scope of this article.

8 Related Work

There is currently no coordinated effort to introduce Service Level Managements in the main European Grid Initiatives, besides providing an initial business model [7]. There have been several attempts to implement some of its aspects in the infrastructures, mainly at the national level. Moreover, the EU-funded gSLM Project intends to stimulate the development of SLMs for Grids. The arguments supporting the need of such actions are collected in [13]. Leff et al. in [6] draw the conclusion that Grids without an SLM framework can only have very limited use.

Examples of such projects are SLA@SOI[1] and SLA4D-Grid[2]. The former project is concerned mainly with Service Oriented Infrastructures and aimed at industrial use cases. Its main concern is ensuring predictability and dependability for business partners. These goals are achieved by introducing an SLA framework for automatic SLA negotiation and management, which may not be possible in such large infrastructures as Grids. The aim of the second project is to design and implement a Service Level Agreement layer in the middleware stack of the German national Grid initiative D-Grid. The target of introducing SLAs into the project was to guarantee the quality of service and fulfillment of prenegotiated business conditions. The SLA4D-Grid project focuses on tools for automatic SLA creation and for negotiations, also offering support for monitoring and accounting. It does not, however, provide a model of an integrated SLA framework which would enable interaction with grid infrastructures other than D-Grid. An important aspects of its activities involves standardizing SLA negotiation protocols on the basis of WS-Agreement [17].

In general, SLA awareness in grids and other computing infrastructures remains a challenge [12].

9 Summary

This article presented several important parts of the PL-Grid SLM concept, including a model SLA/OLA framework, as well as details regarding internal SLA states and resource usage metrics. We described the negotiation process and how it is supported by specialized tools. We believe that the presented solution remains valid for other federated infrastructures such as grids and clouds.

1 http://www.sla-at-soi.eu/
2 http://www.sla4d-grid.de/

References

1. Foster, I.: What is the Grid: A Three Point Checklist. Grid Today (2002)
2. Gerndt, H.M., Rana, O.F., Von Laszewski, G., Ziegler, W.: Service level agreements in grids, Schloss Dagstuhl, Germany, March 22-27 (2009), abstracts collection
3. Battré, D., Hovestadt, M., Keller, A., Kao, O., Voss, K.: Virtual Execution Environments and the Negotiation of Service Level Agreements in Grid Systems. In: Boursas, L., Carlson, M., Hommel, W., Sibilla, M., Wold, K. (eds.) SVM 2008. CCIS, vol. 18, pp. 1–12. Springer, Heidelberg (2008)
4. Koller, J., Oliveros, E., Sanchez-Macian, A.: Service Level Agreements (SLAs) in the Grid Environment. In: Buyya, R., Bubendorfer, K. (eds.) Market-Oriented Grid and Utility Computing. John Wiley (2010)
5. Guitart, J., Macias, M., Rana, O., Wieder, P., Yahyapour, R., Ziegler, W.: SLA-Based Resource Management and Allocation. In: Buyya, R., Bubendorfer, K. (eds.) Market-Oriented Grid and Utility Computing. John Wiley (2010)
6. Leff, A., Rayfield, J.T., Dias, D.M.: Service-level agreements and commercial grids. IEEE Internet Computing 7(4), 44–50 (2003)
7. Candiello, A., Cresti, D., Ferrari, T., et al.: A Business Model for the Establishment of the European Grid Infrastructure. In: Proceedings of the 17th Int. Conference on Computing in High Energy and Nuclear Physics (CHEP 2009), Prague (March 2009)
8. Schwiegelshohn, U., Badia, R.M., Bubak, M., Danelutto, M., Dustdar, S., Gagliardi, F., Geiger, A., Hluchy, L., Kranzlmüller, D., Laure, E., Priol, T., Reinefeld, A., Resch, M.M., Reuter, A., Rienhoff, O., Rüter, T., Sloot, P.M.A., Talia, D., Ullmann, K., Yahyapour, R.: Perspectives on grid computing. Future Generation Comp. Syst., 1104–1115 (2010)
9. EGEE-III: Service Level Agreement Between ROC and Sites (template) (2008), https://edms.cern.ch/file/860386/0.7/EGEE-ROC-Site-SLA-v1.6.pdf
10. Szepieniec, T., Tomanek, M., Twarog, T.: Grid Resource Bazaar. In: Proc. of Cracow Grid Workshop 2009, Kraków (2010)
11. von Laszewski, G., Alunkal, B.K., Veljkovic, I.: Toward Reputable Grids. Scalable Computing: Practice and Experience, Scientific International Journal for Parallel and Distributed Computing 6(3), 95–106 (2005)
12. Sakellariou, R., Yarmolenko, V.: Job Scheduling on the Grid: Towards SLA-Based Scheduling. In: Grandinetti, L. (ed.) High Performance Computing and Grids in Action. The Advances in Parallel Computing series, vol. 16, pp. 207–222. IOS Press (2008)
13. Szepieniec, T., Kocot, J., Schaaf, T., Appleton, O., Heikkurinen, M., Belloum, A., Serrat-Fernandez, J., Metzker, M.: On Importance of Service Level Management in Grids. Presented in Euro-PAR 2011. LNCS. Springer, Heidelberg (2011)
14. ITIL Lifecycle Publications Suite, ISBN 13: 9780113313235
15. ISO/IEC 20000-1:2011 IT Service Management System (SMS) Standard
16. Radecki, M., Szymocha, T., Harężlak, D., Pawlik, M., Andrzejewski, J., Ziajka, W., Szelc, M.: Integrating Various Grid Middleware Components and User Services into a Single Platform. In: Bubak, M., Szepieniec, T., Wiatr, K. (eds.) PL-Grid 2011. LNCS, vol. 7136, pp. 15–26. Springer, Heidelberg (2012)
17. Andrieux, A., Czajkowski, K., Dan, A., Keahey, K., Ludwig, H., Pruyne, J., Rofrano, J., Tuecke, S., Xu, M.: Web Services Agreement Specification (WS-Agreement). Global Grid Forum (May 2004)