AUSTRIAN GRID

STATUS UPDATE ON THE INTEGRATION OF SEE-GRID INTO G-SDAM AND FURTHER IMPLEMENTATION SPECIFIC TOPICS

Document Identifier:	**AG-DM-4aA-1c-1-2006_v1.doc**
Status:	**public**
Workpackage:	**A-1c, M-4a**
Partner(s):	**A-1c – RISC, Upper Austrian Research (UAR)**
Lead Partner:	
WP Leaders:	**Wolfram Wöß, FAW, Joh. Kepler University Linz** **Wolfgang Schreiner, RISC, Joh. Kepler University Linz** **Michael Buchberger, UAR**

Delivery Slip

	Name	Partner	Date	Signature
From	M-4a	A-1c	2006-07-31	A-1c M-4a
Verified by				
Approved by				

Document Log

Version	Date	Summary of changes	Author
1.0	2006-07-31	First stable version	A-1c, M-4a

1 Abstract

This document is an update to the topics discussed in the Austrian Grid deliverables AG-DM-4a-4-2005_v1 (Detailed requirement analysis concerning the cooperating WPs) and AG-DM-4aA-1c-1-2005_v1 (A report on a unified grid-aware access layer for SEE-Grid data sets). In the above mentioned deliverables the usage of G-SDAM (Grid Semantic Data Access Middleware) for the SEE-GRID application was discussed. Additionally the requirements of the SEE-GRID application for G-SDAM were composed. In this document the modifications to both the SEE-GRID application and G-SDAM since these deliverables are discussed and the consequences of these changes are documented and reviewed.

2 Description of the SEE-GRID Database Component

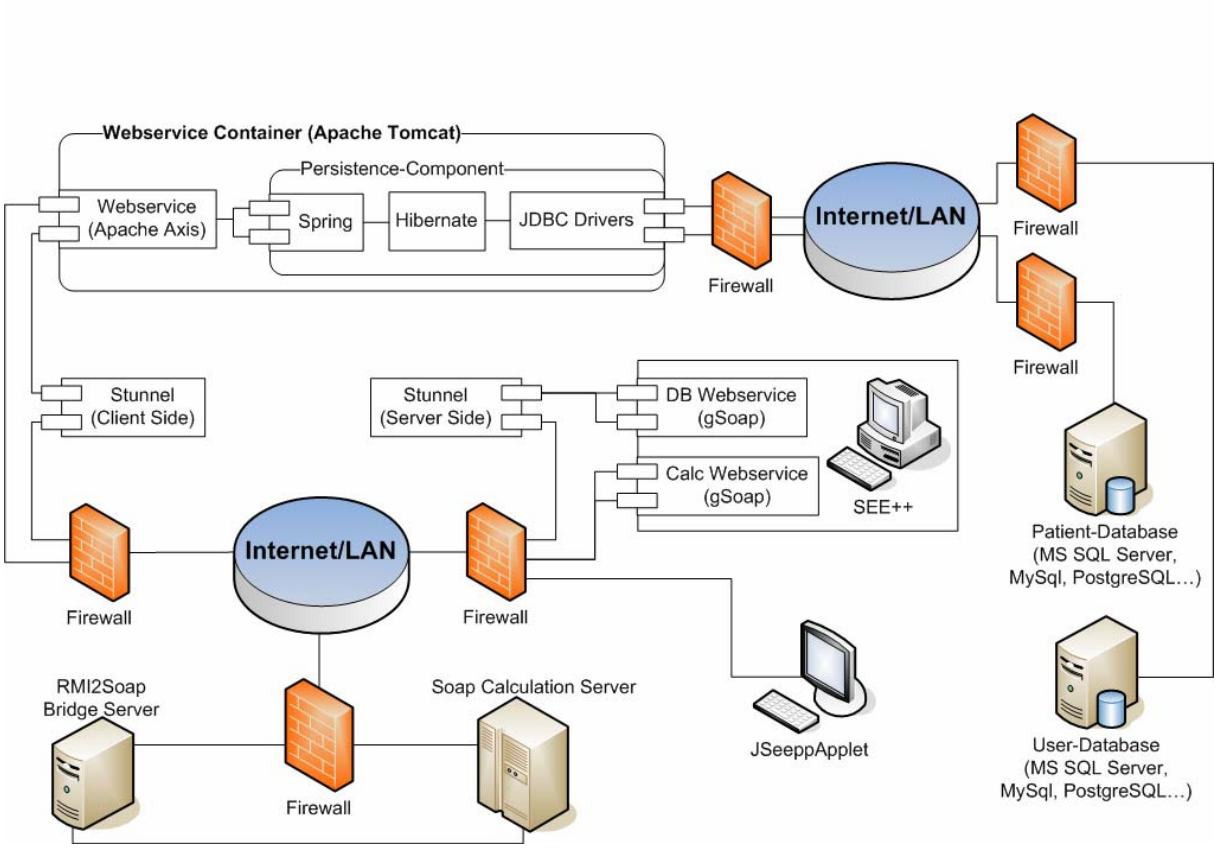

Figure 1: The Current Structure of SEE-GRID Database Component

As a starting point, the SEE-GRID database component was designed [Mitterdorfer, Bosa at al.] and prototyped as a web service application, see Figure 1. The SEE++ clients (JSeeppApplet as well) interact with the database via the SOAP protocol, the communication protocol of SEE++ was extended by additional SOAP messages used by the database application.

The medical data of SEE++ (patients' data, eye model parameters, measured gaze patterns, results of medical experiments, etc.) are stored in the patient database. The data model does not only support the needs of SEE++, it was designed for supporting general medical databases [Mitterdorfer]. Hence, the data model is based on a metamodel, which consists of generally applicable data structures based on various design patterns [Fow04, NHS] instead of specific data types used by SEE++.

For exploiting this general structure and keeping the maximum flexibility of the application, an *Object/Relational* (O/R) mapping solution (namely Hibernate together with Spring) [hibernate] is used. The O/R mapping technique ensures that the conversions between the object-oriented and the relational data structures are transparent to the source code.

Since the SEE-GRID database is designed for storing patient records, security is a very important aspect. The user database contains the user authentication and authorization information of the system. The security implementation ensures that every Web Service call is

secured appropriately by checking the caller's identity. Furthermore, the persistence component employs many techniques to maximize security like:

- intercepting every Web Service method call and checking the authorization for each method separately;
- Encrypting any network transfer via HTTPS or SSL (by Stunnel [stunnel]);
- Applying strong encryption of stored user passwords with a SHA-512 salted hash code.

The cryptographic algorithms used in the prototype are based on proven standards to maximize security [hl03, ssrb00, sch05]. Since the security component is not tied to the persistence component, it can be maintained separately and also used for other purposes.

Currently, the Web Service functionality on the database side is provided by the Apache Axis framework. The SEE-GRID team plans to supplement it by a grid-enabled database interface component in a later phase

3 G-SDAM, Globus and gLite as middleware for data integration

The next steps of the SEE-GRID team concentrate on developing a distributed grid-enabled database system that allows SEE-GRID to give efficient support to “Evidence Based Medicine”. This grid-based database also has to be able to perform various *data mining* algorithms on its data sets.

According to the scenario in [Bosa at al.], the doctors produce data for their local databases by the manual insertion of patient data. The data sets may be collected in a grid database by automatic transfer of medical data (like measured gaze patterns, eye model parameters, etc.) entered in local databases as well as by automatic insertion of the computed simulation data (the computed gaze patterns are never stored in databases, because these simulated patterns always depend on the biomechanical eye model used by the SEE-GRID software system; however they can be recalculated/recovered from the stored eye model parameters).

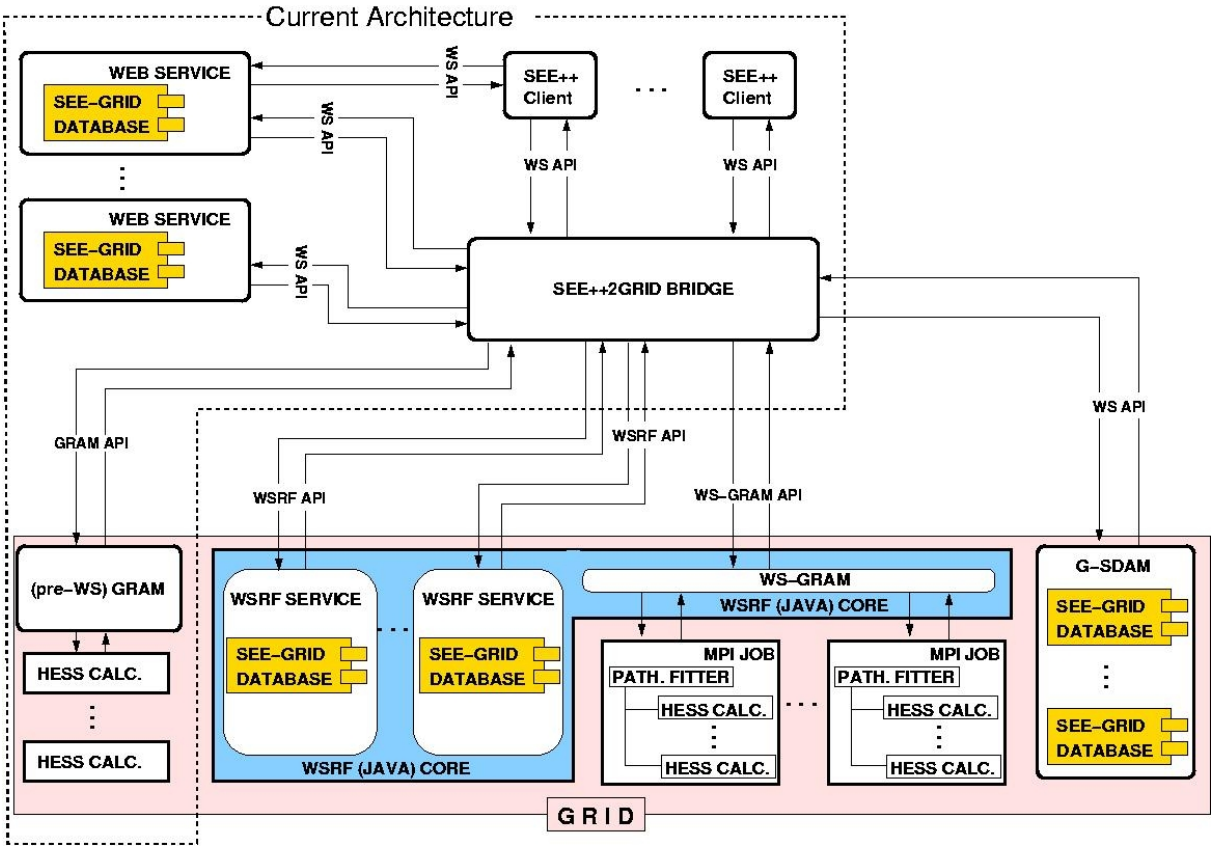

Figure 2: The Proposed Final Architecture of SEE-GRID (the Dashed Line Denotes the Existing Components)

To establish this proposed database without any major modification in the existing data access layer, an abstraction layer has to be introduced. Therefore, the proposed grid database will be based either on the *Grid Semantic Data Access Middleware* (G-SDAM) [G-SDAM] developed by the Institute for Applied Knowledge Processing (FAW) or on the *Web Service Resource Framework* (WSRF) integrated in Globus 4 [globus] (see Figure 2).

As for the first possibility, G-SDAM is an open and easy extensible grid-based software system focusing on seamless data access. It is developed as a standalone grid middleware, which does not require any underlying software layer (like Globus), however it takes over and uses the applications Grid CA and GridFTP developed by the Globus project.

G-SDAM is capable to add or remove data sources/resources at runtime. Its other core features are the possibility to distribute and launch a query on all participating distributed data source and then to collect and compose the query results received from various data sources into one query result. To a client the query seems to be processed in a central database. The implementation of the G-SDAM architecture is not yet complete, such that SEE-GRID may only use a simplified version of it. This will yield the previously mentioned advantages without the support of nodes containing heterogeneous data structures.

An additional advantage of the usage of G-SDAM is that in its final version it will support the security concepts of Virtual Organizations (VOs). The authentication and authorization of the

data sources on the grid are critical issues for SEE-GRID, because it has to be sure that its published medical data will be hosted only by some certain trusted grid nodes.

The other possibility is to deploy data (re)sources as grid services in Globus 4. In this case, SEE-GRID team must develop a parallel/distributed search algorithm so that they will be able to access and collect the desired information from the distributed database. For achieving this, they plan to utilize some of the features of WSRF, like

- *Resource Properties* for discovering the available data source nodes,
- *Stateful Services* for performing queries and for submitting some computational jobs,
- *Notifications* either for reporting the changes in the database triggered by different operations (e.g.: insert, update, delete) or for informing the client about the status of the ongoing/finished computations and
- *BaseFault* for handling the reported faults during the Web Service invocations.

Since both the G-SDAM and the Web Service Architecture in Globus 4 are able to communicate via the SOAP protocol with other grid-based applications, the SEE-GRID database implementation is flexible enough for later adaptation.

3.1 A gLite Port of SEE-GRID Database

In the frame of the EGEE2 project [EGEE], the SEE-GRID team is developing a SEE-GRID version that is compatible with the new gLite 3.0.0 grid middleware [gLite].

Under gLite, they may exchange the software architecture and authentication methods applied earlier for the SEE-GRID medical databases to an AMGA-based solution. AMGA [AMGA] is a database access service for grid applications, which is a part of the latest release of gLite. It is able to hide the differences of the user interfaces of the supported underlying database systems and provides a unified access to them with the grid style certificate-based authentication. Since AMGA supports among others MySQL as well, it would be possible to use the same medical databases in the Globus 4 and the gLite environments.

4 Applying G-SDAM to the SEE-GRID application

Currently a prototype for G-SDAM is implemented that includes a DAS (Data Access Service) for relational databases. For further information on this prototype please refer to the Austrian Grid deliverable AG-DM-4a-2-2006_v1 (Testbed Report – Evaluation of the adapted and proposed algorithms and framework applicability).

Though the SEE-GRID application does not need the ability of G-SDAM to dynamically add heterogeneous data sources with the aid of ontologies, because at the moment SEE-GRID only supports data sources that all have identical data structures, SEE-GRID can utilize the advantage of the facility to add various data sources at a later date, when heterogeneous data

sources are added. Therefore the support and utilization of ontologies in G-SDAM is not set aside for the SEE-GRID application. The usage of ontologies as means to transport semantic knowledge between data sources is at the moment not as important for the SEE-GRID application as the ability to address many data sources through a single query and gain one unified result set from all participating data sources.

The data interchange of sensitive data, e.g. medical data, has its own requirements especially in the area of security and privacy. Especially the various laws concerning privacy demand flexible and extensible security measures. Beside the security measures it is of great importance to monitor data during the data interchange process, as well as prevent unauthorized access of data. Another security measure has to be the usage of trusted sites (e.g. Grid nodes) during the data interchange process and later processing of the acquired data. Therefore it has to be guaranteed that during the data interchange process only known and trusted users and Grid nodes had access to the accumulated data. These security features have to be implemented and introduced into G-SDAM.

4.1 Benefits of G-SDAM

The first prototype of G-SDAM is used to study various scenarios for seamless data interchange in Grid environments. The SEE-GRID application is used as a use case to examine the applicability of G-SDAM as data interchange service in a highly secure environment. This scenario is used to determine whether if the preliminary Relational DAS is usable and in which way the security measures have to be integrated at a later date. The ability of G-SDAM to operate only with known and trusted grid sites is of imminent importance for the data interchange of sensitive data. The introduction of a Virtual Organization (VO) responsible for the registration of new data sources as well as monitoring institution during the data interchange is very important. Additionally the possibility to determine the origins of the accumulated data (data provenance) is another key feature that is of great importance for the users of G-SDAM as well as for the monitoring VO to backtrack the data interchange process.

G-SDAM will implement most of the above mentioned features, though the first prototype is implemented to determine the applicability of G-SDAM and at the moment does not implement particular security measures. Later versions of G-SDAM however will implement these security features to ensure the appliance of G-SDAM as data interchange service for sensitive data. Another future core feature of G-SDAM is the ability the generate data interchange workflows (DIW). These DIWs can be used to predetermine Grid sites for particular data manipulations (e.g. manipulation of personal information regarding patients, to ensure their privacy) during the data interchange process.

5 Outlook

During the next few months the prototype of G-SDAM will be further implemented and a testbed for the SEE-GRID application will be established. This testbed will be used to observe the applicability of G-SDAM to the SEE-GRID application and will allow us to find and review additional requirements as well as necessary security measures for the implementation of a comprehensive security manager that will be added to G-SDAM at a later date.

References

[AMGA] AMGA User's and Administrator's Manual
http://project-arda-dev.web.cern.ch/project-arda-dev/metadata/downloads/amga-manual_1_2_3.pdf

[Bosa at al.] Karoly Bosa, Wolfgang Schreiner, Michael Buchberger, Thomas Kaltofen, *SEE-GRID, A Grid-Based Medical Decision Support System for Eye Muscle Surgery*, 1st Austrian Grid Symposium, December 1-2, 2005, Hagenberg, Austria. OCG Verlag, 14 pages.

[EGEE] EGEE-II homepage, 2006. http://www.eu-egee.org

[Fow04] Martin Fowler, *Analysis Patterns: Reusable Object Models*, Addison-Wesley, 2004.

[gLite] gLite 3.0.0 homepage, 2006. http://www.glite.org

[globus] The Globus Toolkit. http://www.globus.org/toolkit/

[G-SDAM] *A Report on a Unified Grid-aware Access Layer for SEE-GRID Data Sets*, Austrian Grid Deliverable M-4aA-1c, FAW Institute and RISC Institute, Johannes Kepler University, Linz, August 2005. http://www.faw.uni-linz.ac.at

[hibernate] Hibernate homepage, 2005. http://www.hibernate.org/

[hl03] Michael Howard and David LeBlanc, *Writing Secure Code*, Microsoft Press, 2nd edition, 2003.

[Mitterdorfer] Daniel Mitterdorfer, *Grid-Capable Persistance Based on a Metamodel for Medical Decision Support*, Diploma thesis, Upper Austria
University of Applied Sciences, Hagenberg, July 2005.

[NHS] *NHS Healthcare Modelling Programme*, 1995,
http://www.standards.nhsia.nhs.uk/hcm/index.htm

[ssrb00] Douglas Schmidt, Michael Stal, Hans Rohnert, Frank Buschmann, *Pattern-Oriented Architecture*, Vol. 2: Patterns for Concurrent and Networked Objects, John Wiley and Sons Ltd, 2000.

[sch05] Bruce Schneider, Cryptanalysis of SHA-1, 2005.
http://www.schneier.com/blog/archieves2005/02/

[stunnel] Stunnel – Universal SSL Wrapper. http://www.stunnel.org/