### New concept for processing medical data in distributed environments

Viezens F, Sax U
*Abteilung Medizinische Informatik, CIOffice Forschungsnetze, Bereich Humanmedizin, Universität Göttingen, Deutschland*
*fred.viezens@med.uni-goettingen.de*

**Introduction and Question** Grid Computing offers new possibilities to interdisciplinary research. The advantage of this cooperative research does not only consist in the more effective use of resources, e.g. arithmetic performance (CPU) and storage capacity, but also in the faster processing and use of the newest algorithms in medical research. The biomedical community recognized this development early. By standardizing data formats, the open source applications can easily be parallelized for grid computing relying on the Internet and own Web servers. Such an application is DIALIGN [1], a software tool for the alignment of multiple DNA and protein sequences. This was taken up in the BMBF funded D-Grid-initiative by MediGRID [2]. Applications in the medical community in the MediGRID projects currently include the areas biomedical informatics, medical image processing and clinical research. As a special requirement of the medical community the security of person related data has to be ensured at any time. This is not sufficiently solved in current applications and is not transferable to the grid environment. A new concept has to be developed, in order to introduce the potentials of the grid computing to life science and the medical field.

**Materials and Methods** Biomedical informatics with their applications are already in the position to run a grid application. This creates synergies for the other communities, who could benefit from this experience in the grid later. Synergies arise from solutions concerning the requirement to homogenize the raw data. Data sources in biomedical grids consist of heterogeneous data sources with different of no standards. Person related data in a medical grid additionally requires specific security methods in the grid. Traditional security mechanisms do not meet today's requirements at distributed procedures or jobs, where resources in a virtual environment handle the processing of complex data for research purposes. The new concept defined not only the requirements by security aspects for such a system to be processed medical data, but also the pseudonymization of identifying data (IDAT) in the grid before processing. Today's development in the area of such distributed systems goes toward grid and grid services, ajar the web service paradigm [3], based on such standards, like simple object access protocol (SOAP), web services description language (WSDL) and universal description discovery and integration (UDDI). The substantial difference between both insists in the status less condition of web services on transportation protocol layer level. A grid service is session-based. This difference reaches from the protocol layer up to the application level. On transportation level mechanisms already exist, which ensure secure and confidential communication of users and grid resources. Virtualization technologies like Xen have to be evaluated for these purposes. On application level a grid service goes through different transaction-conditioned sections: from authentication and authorization, to audit, tracking and accounting. This communication will be based on the eXtensible markup language (XML). With XML based communication in the grid it is possible to describe and supervise accesses to documents and resources exactly and secure. Under use of certificates and attributes in XML message header the users are in the position to assess resources in order to classify and to release the document or parts of it for grid applications or grid user for use. With fine-granular access rights [4] and a certificate-based infrastructure such multilevel security architecture on application layer with XML documents is possible. Among them is a secured transport layer with well-known protocols secure socket more layer (SSL), secure hypertext transmission protocol (https) and so on, as well as were based Firewalls, which extended and complete this architecture for the lowest layers of the OSI model.

**Results** The following exemplary scenario shows the current – preliminary - state of medical grid applications. Before processing personal data in grid applications, a separation from identifying data and the raw data is to be made. The identifying data (IDAT) and personal identifier (PID) of a record are pseudonymized to an object identifier (OID). That happens in a PID generator. The PID generator is already in use in medical research networks like congenital heart disease [5, 6]. Following the paradigm of web services we use the possibility of the integration of the PID generator like a legacy system. The communication with the PID generator, the transfer of the metadata and the composition of the raw data and IDAT, and the transmission of the OID to the PID dispatcher, are realized in SOAP. The SOAP messages are secured accordingly in the header and provided with attributes, so that secure communication without can take place. Thus security-relevant requirements are fulfilled, because raw data with appropriate OID and life science identifier (LSID) are pseudonymized data in the sense of the processing of personal medical data for maintaining the patient rights. This data are currently fed via protocol services (Gridftp) into the grid. Then they get assigned on resources and processes, which are explicitly mentioned in the document header or in the respective description of job. The processing can be examined and verified by dedicated log files for the respective session at any time. Besides certificates, the header also includes attributes, so that one can exactly identify the user and specify the machine. The organizational model of a patient consent derived, the patient can determine to which machine, hospital area, health insurance or commercial range, her data should be processed. By consistent use of XML techniques on application level, like security assertion markup language (SAML) and eXtensible access control markup language (XACML), the representation of person-relevant information and the processing of authorize policies are possible.

**Discussion** Secure grid computing requires an appropriate Public key infrastructure as a basic condition. Users and services are authenticated by X.509-certificates. In a virtual organization, like the grid, an authentication and an authorizing infrastructure (AAI) play a paramount role. Security flaws can be fixed under adherence of attributes strictly assigned on document and resources level and roll-based access control. The illustration of organizational procedures on these new concepts helps during the conversion of the demands specified above. As some life science communities already use Shibboleth within their AAI, compatibility to grid shib should be considered. Existing and future documents standards should support the protection of person related data. They are based on a continuing application of the web service paradigm and therefore seem to be promising for such privacy architecture, which concerns the substantial demand for the use of the grid technology in life science and the medical field.

This work was supported by the D-Grid Project MediGRID, funded by the Federal Ministry of Education and Research (BMBF), FKZ 01AK803H.

### Literature

[1] Morgenstern B. DIALIGN: multiple DNA and protein sequence alignment at BiBiServ. Nucleic Acids Research 2004; 32: W33-W36.

[2] MediGRID. Medical Grid Computing. 2005 [cited 07.02.2006]; Available from: http://www.medigrid.de.

[3] Burwitz V, Roth M, Schäfer Th. Sichere Web Services. 2003; Retrieved February 16, 2006.

[4] Bourka, A., et al., Enriching healthcare applications with cryptographic mechanisms and XML- based security services. Technol Health Care, 2003. 11(1): p. 61-76.

[5] Pommerening, K., et al., Pseudonymization in medical research - the generic data protection concept of the TMF. GMS Medizinische Informatik, Biometrie und Epidemiologie, 2005. 1(3).

[6] AHF. Kompetenznetz angeborene Herzfehler. 2005 [cited 18.07.2005]; Available from: http://www.kompetenznetz-ahf.de.