THE USE OF GRID STORAGE PROTOCOLS FOR HEALTHCARE APPLICATIONS

Flavia Donno
CERN, CH-1211, Geneva 23, Switzerland, flavia.donno@cern.ch

Elisabetta Ronchieri
INFN-CNAF, Viale Berti Pichat 6/2, I-40126 Bologna, Italy, elisabetta.ronchieri@cnaf.infn.it

Keywords: HealthGrid, Storage Grid Technology.

Abstract: Grid computing has attracted worldwide attention for a variety of domains. Healthcare projects focus on data mining and standardization techniques, the issue of data accessibility and transparency over the storage systems on the Grid has seldom been tackled. In this position paper, we identify the key issues and requirements imposed by Healthcare applications and point out how Grid Storage Technology can be used to satisfy those requirements. The main contribution of this work is the identification of the characteristics and protocols that make Grid Storage technology attractive for building a Healthcare data storage infrastructure.

1 INTRODUCTION

Grid computing aims at the definition of a global infrastructure able to share geographically distributed resources (such as data, storage, computers, software, tools, applications, instruments, and networks) in a secure context. The name "Grid" comes from the metaphor of "Electrical Grids" and the idea to get access to a resource (e.g., electricity) by using a plug (Foster, *et al.*, 2001). The integration of Grid and Healthcare has created a new area called HealthGrid (Breton, *et al.*, 2005), which has produced a great impact on almost every aspect of Healthcare.

Grid builds an infrastructure that provides resources and distributed information to the medical personnel. The available resources may include computational resources, storage, equipment (e.g., scanners, lab), or human resources (specialists). Healthcare specific policies (such as privacy or authorization) are enforced according to a given service agreement.

In this paper we focus on Grid Storage technologies and investigate how these could be exploited by healthcare applications. In particular, in Section 2 we describe Healthcare use cases and requirements. Section 3 reports on related work. In Section 4 we give an overview of Grid storage technology. Section 5 elaborates on existing Grid Storage protocols for data consistency, privacy, and preservation. A few examples are given for encoding medical metadata and preserve the information on a Grid Storage Service for later retrieval. Finally, Section 6 summarizes our conclusions and describes future work.

2 HEALTHCARE USE CASES AND REQUIREMENTS

A medical application can include a check-up, a specialist consultant exam, bio-signal or genomic measurements. The patient data are first collected. Old documentation concerning the patient is connected to the current patient file. New tests are performed and medical images or other exams are acquired. While the data concerning the patient are strictly protected and must not be exposed to insecure access, the images produced by the exams can be made available to several physicians for further comparative or statistics studies. By using the Grid, all the exams and measurements can be made available on geographically distributed

computing resources. Online medical information can be acquired by connecting to national or international medical information centres to help the diagnosis. Remote consultation can also be set up through the network. Cooperation and interoperability of clinical data is essential. Algorithms for searching and retrieving data from knowledge bases have to be properly defined. Furthermore, data mining is another essential process. Some of the distributed sources of data may have strong privacy rules. Therefore, the system must be proven highly reliable for privacy and security. Health Grid applications have specific data distribution and access requirements:

Security: By authentication we intend the process for which it is possible to know who a user is. Authorization implies the discovery of what a user is allowed to do in the system. Auditing is the process that keeps track of the actions performed by a user.

Privacy and confidentiality of data: It can be achieved by splitting sensitive patient data from the rest of the data content and using de-identification or encryption.

Legal and ethical enforcements: The legal and ethical requirements impose obtaining explicit consensus from the patients and clearance to use data from ethics' committees. Local policies must be enforced over global decisions for data distributed world-wide.

Interoperability of data: The interoperability between data stored in different centres can be achieved through standardization of data formats, structures and models.

Data access transparency: It is important to guarantee location independent transparent access to data. Therefore, a storage infrastructure should allow for the preservation of metadata (contextual data about the information being stored) and the negotiation of supported capabilities. The metadata information could concern for example the software that has generated the data, their encoding, their validity, security and privacy information and other data logically connected. A framework processing such a description could allow for the negotiation of given capabilities with the data access services.

Quality of service: The quality of service requirement is related to guarantee for example adequate access time and storage quality for the type of data being stored at a site. It is important to negotiate a Service Level Agreement (SLA) in terms of access latencies, integrity, etc. for specific important data stored in a Grid.

3 RELATED WORK

In 2002, the European Community has funded projects specifically on Healthcare. They are MammoGrid (Warren, *et al.*, 2007), and GEMSS (Grid-Enabled Medical Simulation Services Project, http://www.it.neclab.eu/gemss/). The UK e-Science program made funds available between 2001 and 2004 in Healthcare. These were used to sponsor projects such as eDiaMonND (UK e-Science eDiaMoND Project, http://www.ediamond.ox.ac.uk/), CLEF (CLEF – integrating information for the clinical e-Scientist, http://www.clinical-escience.org/start.html) and CareGrid (Liu, *et al.*, 2007).

The MammoGrid (European Federated Mammogram Database Implemented on a GRID Structure) project aimed to develop a pan-European distributed database of mammography images using Grid technologies. The GEMSS (Grid-Enabled Medical Simulation Services) project aimed at providing an innovative Grid middleware to support several medical service applications including maxillofacial surgery simulation, neuro-surgery support, radio-surgery simulation, inhaled drug delivery simulation, cardio-vascular system simulation and advanced medical image reconstruction. The eDiaMoND (on digital mammography) project aimed at delivering a prototype which could support breast screening in the UK through screening tests, computer-based training, epidemiology studies and computer aided detection of breast cancer. It gave patients, physicians and hospitals fast access to a vast database of digital mammogram images. The database was also used for image analysis algorithms research, for data mining and imaging standardization techniques. The CLEF (CLinical E-science Framework) project aimed at creating a generic scalable architecture based on Grid technology for capturing, integrating, interpreting and using clinical data with genomic data and images within practical clinical systems. The CareGrid project focuses on developing software for supporting decisions based on trust, privacy, security and context models in Healthcare application domain.

One topic not yet covered is how to provide support to Healthcare applications through the Grid Storage Technology.

4 GRID STORAGE TECHNOLOGY

A **Grid Storage Element (SE)** *is the set of hardware and software solutions adopted in order to realize a storage Grid service. It hides the difference among specific solutions and allows users for consistently and securely storing files at a Grid site*.

An SE provides a set of useful functionalities. Grid users and applications migrate across multiple administrative domains. Both a **transparent interface** for specific, very frequent operations and a **set of different communication protocols** are supported. S**ecurity protocols** used across domains might differ and are equally and transparently supported. In particular, authentication and authorization mechanisms valid across domains are honored. Reliable and high performing **data transfers protocols** across domains are available. **Local policies and priorities** have priority over global ones. A set of **storage classes** is advertised so that application and services can take advantage of them. A storage class is a type of storage space with specific attributes such as reliable, persistent storage or unreliable scratch space for temporary usage only with very low access latency. Grid users also need to have available a set of operations for **managing and reserving storage space** and for **filesystem-like operations** (such as ls, mkdir, ln, and file locks). Grid Storage services differentiate between valuable and expendable data (**volatile** vs. **permanent data**). Operations such as **transparent**, **automatic** or **forced migration to tertiary storage** (tapes) are available. Mechanisms for transparently **locating data** on any storage device are provided for debugging, disaster recovery, and security reasons. Storage systems also provide a mechanism to **advertise capacity, status, availability and content** to an information system. **Management** and **monitoring functions** for Grid global control of service behavior and functionality are available. **Data integrity mechanisms** allow through the usage of checksum and location independent information to ensure correctness of data even after replication or reprocessing. **Data privacy mechanisms** is ensured with the support of side services that allow for instance the storage of encryption keys on distributed servers so to protect also from malicious attempts at a specific site.

The **Storage Resource Manager (SRM)** specification (Shoshani, *et al.*, 2004) is used today by many Grid storage services. Other emerging protocols are available and could be used in the context of healthcare. In the next section, we provide an overview of the most common protocols used today in the Grid environment for storage services together with some Healthcare application examples.

5 GRID STORAGE PROTOCOLS

A **Storage Resource Manager** (Sim, *et al.*, 2008) *is a middleware component whose function is to provide dynamic space allocation and file management on shared storage components on the Grid*. It satisfies most of the requirements explained in the previous section. Data sets and metadata information can be packaged in a file that SRM can manage.

Let us illustrate the functionality of an SRM through a typical healthcare use case. A physician would like to make available a set of files through the Grid. The application used checks that the amount of space needed for storing the data is. Therefore, after authentication and authorization are performed, the application can issue an SRM request to reserve the space to store the data on an SE with a certain quality (magnetic device/online disks/fast access) to allow for security and preservation policies to be applied. Additionally, the application is enabled to only use a certain set of file protocols, as in the case of DICOM (http://medical.nema.org) images. Therefore, it can negotiate the provision of an area of space where the requested protocol can be used. As a result, the SRM provides a protocol dependent file handle that the application can use to store the data. The application initiates the store session that includes data encryption and storage of encryption keys on specific servers, like Hydra Keyserver (Gilliam, 2002). Metadata information can also be stored through interfaces such as the SNIA XAM (Storage Networking Industry Association, *Information Management – Extensible Access Method (XAM) – Part 1: Architecture, Version 1.0*, 2 April 2008, http://www.snia.org/forums/xam/technology/specs/XAM_Arch_v1.0_Apr6.pdf.) described later. Access Control Lists can be defined on both the data namespace and the storage space to control access. The original set of data together with its metadata bundle can be automatically replicated at different sites for better availability and preservation or to optimize data access. Other physicians at a different location on the Grid can check for the availability of the data in order to start further processing. Data integrity methods guarantee the correctness and authenticity of the data and the associated metadata. Once done, the space reserved by the can be

recollected by the system and reused to satisfy other requests. Files can be removed automatically by the system for further security, allowing only a specific time window for given operations. SRM offers a **namespace** similar to the one of a UNIX filesystem. The Grid Security Infrastructure (Campana, *et al.*, 2004) is also used.

Another emerging protocol that is acquiring popularity is NFS v4 (http://www.nfsv4.org/). Among the features of NFS v4 we can list: improved access and good performance on the Internet; strong security, with security negotiation built into the protocol; enhanced cross-platform interoperability; extensibility of the protocol. Furthermore, NFS v4.1 provides support for metadata encapsulation.

SRM and NFS v4 address the problem of providing control and access uniform interfaces to world-wide distributed Grid Storage Services. The Storage Networking Industry Association (SNIA) is promoting the definition of the eXtensible Access Method (XAM). This protocol proposes a way to handle reference information (metadata) at the level of a storage device. Such metadata provide a way to relocate data across diverse local hardware platforms, without compromising data integrity. Based on the XAM-retrieved metadata content, the storage service can trigger automatic data operation before serving the data to a user for access or manipulation. XAM provides: **a Global location independent unique name associated with reference information,** that allows for implementing data integrity practices; metadata strictly coupled with its data, facilitating the data management, access and manipulation together with data interoperability; **storage systems accessed via a standard, pluggable architecture**. XAM also provides a standardized set of management disciplines and semantics for fixed content. Retention and expiration policies can be enforced. An example of application of the XAM technology to HealthCare application is for instance the encoding of the Clinical Document Architecture (CDA) in XAM objects. CDA is a schema for recording clinical events in documents. The schema is composed of two entities: a header that contains information like global-unique identifiers, document type code, timestamp, confidentiality code, patient, author and custodian; a body that contains tables, lists, sounds, and video clips. CDA schema can be translated in XAM, simply abstracting data and metadata into XSet contained into a logical XSystem. Therefore it is possible to represent for example a mock up patient data by using an XSystem characterized by several XSets one for each CDA entity.

6 CONCLUSIONS

In this work we have identified the Grid storage services and protocols that could facilitate the integration and interoperation of Healthcare data and frameworks world-wide. While many of the current Healthcare Grid projects address issues such as simulation, data location and description on the Grid and security, the problems connected to data storage, integrity, preservation and distribution have been neglected. We are planning to make available a prototype infrastructure using EGEE gLite that integrates all protocols mentioned as a proof of concept for distribution and sharing of medical data.

REFERENCES

Breton, V., *et al.*, 2005, "The HealthGrid White paper," In *Proceedings of HealthGrid 2005* (From Grid to HealthGrid), IOS Press, vol. 112, pp. 249-321.

Campana, S., *et al.*, 2004, "Toward a Grid Technology Independent Programming Interface for HEP Applications," In *Proceedings of CHEP2004*, Sept. 27 - Oct. 1, 2004 Interlaken, Switzerland.

Foster, I., *et al.*, 2001, "The Anatomy of the Grid: Enabling Scalable Virtual Organizations," In *International Journal of Supercomputer Applications*, vol. 15, no. 3, pp. 200-222, http://www.globus.org/research/papers/anatomy.pdf.

Gilliam, D., 2002, "Summary report on enterprise security workshop," In *Proceedings of the Eleventh IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises*, pp. 43-46.

Liu, Y., *et al.*, 2007, "On smart-care services: Studies of visually impaired users in living context," In *Proceedings of the First International Conference on the Digital Society (ICDS'07)*, pp. 32-40, IEEE Press, January 2007.

Shoshani, A., *et al.*, 2004, "Storage Resource Management: Concepts, Functionality, and Interface Specification," GGF 10, The Future of Grid Data Environment, 9-13 March 2004, Humboldt University, Berlin Germany.

Sim, A., *et al.*, 2008, "The Storage Resource Manager Interface Specification Version 2.2," OGF-GSM, 24 May 2008.

Warren, R., *et. al.*, 2007, "MammoGrid - A Prototype Distributed Mammographic Database for Europe," In *Clinical Radiology*, vol. 62, no. 11, pp. 1044-1051, June 2007.