A Virtual Data Grid Architecture for Medical Data using SRB

Lidia Leoni[1], Simone Manca[1], Andrea Giachetti[1,2] and Gianluigi Zanetti[1]
[1]CRS4, Polaris Scientific Park, Pula (CA), Italy
[2]Dip. Matematica e Informatica, Università di Cagliari, Italy
Corresponding author: Simone Manca,c/o CRS4 Parco Scientifico e Tecnologico Polaris edificio 1, Loc. Pixinamanna,09010 Pula (CA) Italy, simo@crs4.it

We present a method to include anonymized PACS data into data grids. It is based on a simple mechanism, SRB/DICOM driver, for the transparent integration of DICOM compatible PACS in SDSC Storage Resource Broker (SRB) based data grids. The SRB/DICOM driver provides an uniform interface between the SRB server component and medical images archives supporting DICOM Query/Retrieve functionalities. The driver thus allows SRB clients to access transparently, and in user specified data formats, anonymized medical images published by participating PACS. We expect this extension to be useful in the construction of general data exchange infrastructures for wide area biomedical data warehouse dedicated to research and clinical purposes.

INTRODUCTION

Data Grid is an emerging technological paradigm for the seamless access, via virtualized middleware, to heterogeneous and distributed ensembles of data storage resources (1). How this can be technically accomplished is the subject of an intense research activity, with several data grid frameworks and middleware packages being developed and tested by projects of various nature and geographical extension (2).
From the beginning, the data grid paradigm has been considered ideally suited to provide the advanced data exchange and collaboration infrastructure needed by multi-centric medical research efforts. Examples include the Biomedical Informatics Research Network (BIRN) (3), the National Digital Mammography Archive Grid (NDMA) (4), the eDiamond (5) and the Mammogrid (6) projects.
BIRN is a US National Institutes of Health initiative that has as a goal to foster distributed collaborations in biomedical science research, while NDMA, eDiamond and Mammogrid are instead centered on using data grid technology to realize wide area data infrastructures for the storage, retrieval and use of digital images (mainly mammograms) for clinical, epidemiological and training purposes.
One of the main features of the latter medical data grids is that they are designed to transparently support – via specialized gateways nodes – direct DICOM communications both from PACS to the data grid and from the data grid to the medics DICOM workstations. While this allows a potential better integration in the hospital workflow, it tends to focalize the data grid on the specificities of a particular image format. This could be asymptotically a problem, since medical data grid infrastructures are expected to be extended to contain other data sources – not necessarily DICOM – and to become the data exchange infrastructure of a more general wide area biomedical data warehouse (7) dedicated to research and clinical purposes.
From this point of view, PACS become simply one of the possible types of data sources that feed the data grid.
As an example of this strategy, in this paper we will report on how we extended the San Diego Supercomputing Center Storage Resource Broker (SRB) (8), one of the most popular middleware used to build a Data Grid, to directly support the use of DICOM3 storage servers as data grid data sources.

PURPOSE OF THE STUDY

Diagnostic imaging modalities are all rapidly becoming digital and connected to large on line archives called PACS (picture archiving and communicating system). PACS are conformant to the DICOM3 standard and accept queries for patient/study/series performed on secured networks.
The purpose of this study is to make accessible via a data grid infrastructure images that are physically stored in PACSs, while, at the same time, providing participating medical institutions with complete control on the selection of images they wish to publish.
The initial application of the system will be in a regional data grid for biomedical research, that will be extended, at a latter time, to support patient specific grid computing based analysis.
The data grid is expected to provide an abstraction layer between the medical image archives and end users typically interested in requiring semantically classified images for their research, while the image publication mechanism should:

- allow the participating institutions to easily select which of the cases resident on their PACS they want to share on the data grid;
- automatically anonymize the data so that no patient identifying detail is exported to the grid;
- automatically extract from the DICOM image all the information that can be used as meta data for image classification;
- guarantee that only the explicitly published data is accessible and that the PACS is effec-

tively shielded from unauthorized access via the data grid interface.

As an actual implementation of the abstraction layer between PACS and Data Grid users, we extended the San Diego Supercomputing Center Storage Resource Broker (SRB) (9), one of the most popular middleware used to build data grids, with a custom made driver enabling the integration in the grid of DICOM3 storage resources.
The SRB is a client-server middleware that provides a replication collection management across multiple storage systems. It allows the organization of data from multiple heterogeneous systems into easily accessible logical collections, and, in conjunction with the Meta data Catalog, it supports location transparency by accessing data and resources through queries on their attributes rather than their names or physical locations. SRB also provides capabilities to store replicas of data, for authenticating users, controlling access to documents and collections, and auditing accesses. It can also store user-defined metadata at the collection and object level and provides search capabilities based on these metadata.

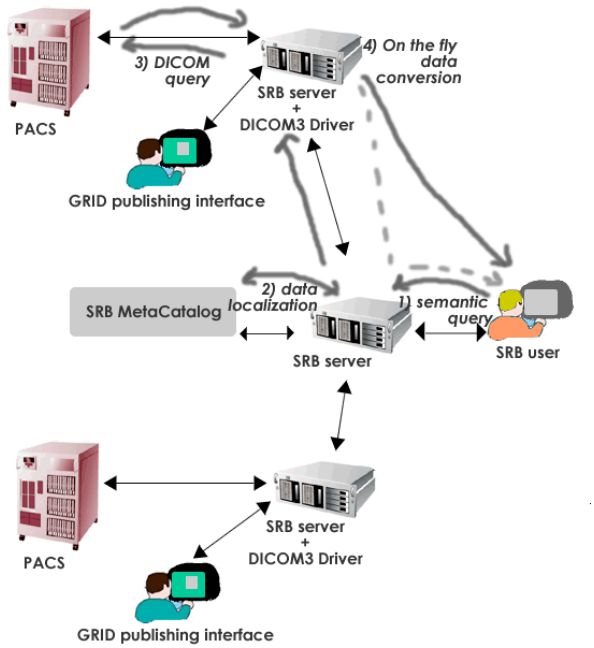

Figure 1. Example of SRB/PACS integration The SRB middleware is used in several successful applications, e.g., the examples described in (9).

The SRB/DICOM driver provides an uniform interface between the SRB server component and medical images archives supporting DICOM Query/Retrieve functionalities. The driver thus allows SRB clients to access transparently, and in user specified data formats, anonymized medical images published by participating PACS. Through the coupling of SRB servers and PACS systems, SRB clients can perform queries using meta information on the distributed "virtual" medical image database, and can access DICOM studies as hierarchically organized directories and files, in the same way they would have accessed them on a local or networked filesystem.
The selection and publication of local PACS data is managed through a simple web interface that allows radiologists to decide what data should be automatically anonymized and shared and what standardized meta-information should be associated to the latter.

MATERIAL AND METHODS

The transparent DICOM data access in our application works as described in Fig.1: when a SRB client ask for a resource, data are first localized through the Meta-Catalog database (step 1 in Fig.1), then retrieved from the correct SRB server (step 2 in Fig.1). If this server does not have data locally cached, it is able, through the DICOM3 driver, to retrieve them with a DICOM3 query to the PACS (step 3 in Fig.1). Finally, the data is sent back to the end user either directly or via the server that received the first SRB query depending on what the SRB middleware considers the most cost-effective network route (Step 4 in Fig.1).

Network connections between the SRB server and the external network, and between the DICOM server on the same machine and the DICOM network are kept logically separated so that there is no way to directly access the PACS server from outside. The only DICOM studies that can be queried from the SRB/Dicom server to the local PACS are those that had been previously published in the grid through a SRB/DICOM provided web interface.

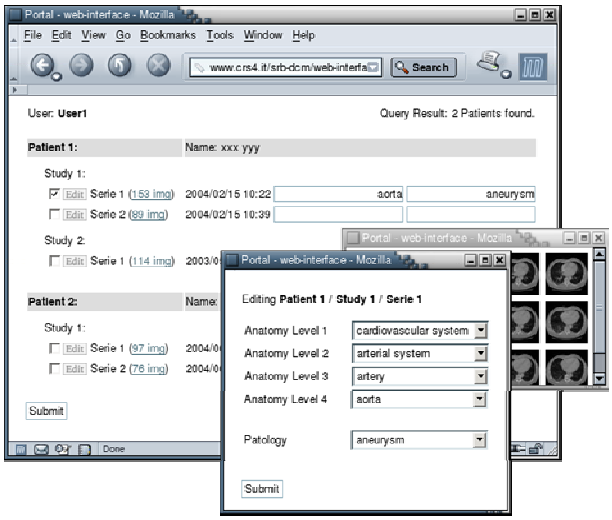

Figure 2. SRB/DICOM Grid publishing interface

The software driver is written in C++, exploiting the OFFIS DCMTK libraries (10) for DICOM data handling and communications. DICOM queries are formatted using the ID's stored in the MetaCatalog, image data are then retrieved as DICOM, and modified on the

fly. First the anonymization step is performed, removing a list of DICOM tags:

- DCM_PatientsName,
- DCM_PatientsBirthName,
- DCM_PatientsMothersBirthName,
- DCM_PatientsAddress,
- DCM_PatientsTelephoneNumbers,
- DCM_OtherPatientNames.

Then the required metacatalog resource is generated: in the metacatalog there are not only the original DICOM data set, but also other "virtual" images not corresponding to the physical data, like jpeg converted images, thumbnails, volumetric raw data, etc.
In the current implementation the data conversions are performed on the fly by the SRB/DICOM driver.

DATA Publication/metadata addition

In our data publication model only data selected by the local PACS administrator can be accessed through the GRID.
The publishing mechanism is implemented as a web service, the SRB/DICOM publishing interface, that allows PACS administrators and registered operators to select between the studies present on the PACS using a set of standard forms. It is then possible to assign to selected studies further pre-defined attributes from given attribute lists (two for the time being: anatomical regions and pathologies).
Once the extended attributes have been selected is then possible to proceed with the actual publication of the studies, that is performed by SRB/DICOM by sending all the relevant information, (data location, ID's, selected DICOM tags and added attributes) to the SRB MetaCatalog.

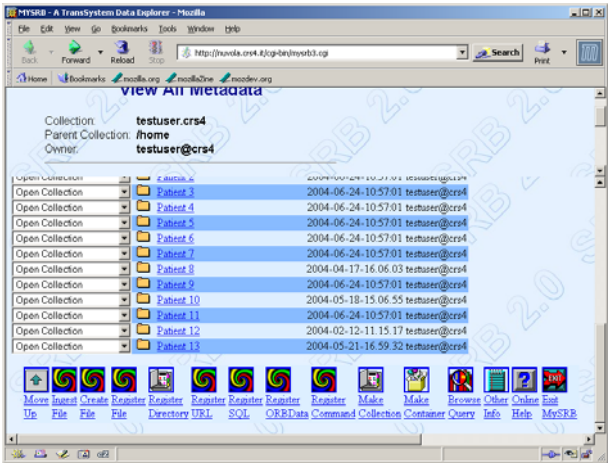

Figure 3. Access to the data via the standard mySRB interface

In the figure above we illustrate an access to the data grid using a typical web interface, mySRB (11). Using mySRB the user can select the relevant data and then download it for local processing.

Global system architecture

The global system architecture is, obviously, directly related to the architecture of standard SRB based grids, with a common metacatalog server and SRB servers located at each participating institutions.

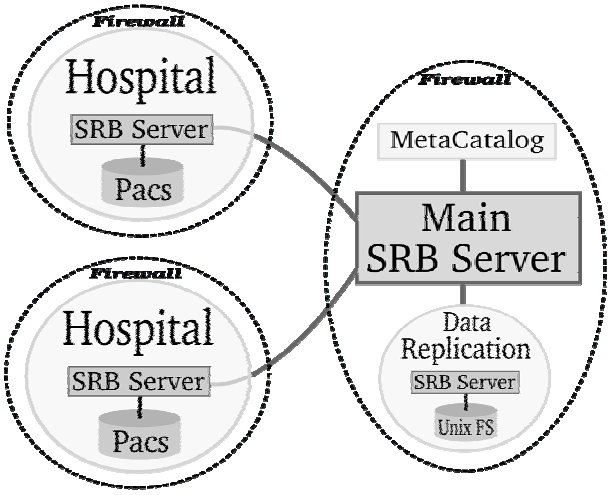

Figure 4. Datagrid architecture. SRB data replication functionality will be used to provide back-up capabities to the system.

The actual deployment of the system will be based on the installation in the participating clinical institution of a specialized, self contained, SRB/DICOM box – similar to the one described in (12) – that will contain both the SRB/DICOM server and the support for SRB/DICOM publishing. The boxes will be connected by DICOM and https to the internal hospital network, and by SRB to the data grid. The metacatalog server, as well as the SRB replicated data server are hosted at CRS4.

RESULTS

We have developed a preliminary version of the SRB/DICOM software and it is currently under test in our labs. The current version of the software supports the following capabilities:

- DICOM3 retrieve functionality via C-MOVE messages;
- Automatic anonymization of DICOM dataset;
- Automatic format conversion to various image and raw data formats.

The SRB/DICOM publishing module provides a web interface that it is currently supporting the following operations:

- DICOM3 query/retrieve functionality via SCU class;
- List of studies, and related hierarchies, by patient, date and other attributes;
- Association of attributes, currently only pa-

thology and anatomical region tags, to selected studies;
- Publication to the SRB metacatalog.

As mentioned above, the system will be deployed as self contained SRB/DICOM boxes. Our current reference implementation of the box is a dual AMD Opteron Linux machine with 2GB of memory, three 140GB SATA disks and two Gigabit Ethernet cards. The boxes will be initially deployed in the Medical center of the University of Cagliari and at AOB, Cagliari main public hospital.

DISCUSSION

We presented a simple mechanism for the transparent integration in SRB-based data grids of DICOM compatible PACS. The SRB/DICOM driver provides an uniform interface between the SRB server component and medical images archives supporting DICOM Query/Retrieve functionalities. The driver thus allows SRB clients to access transparently, and in user specified data formats, anonymized medical images published by participating PACS. We expect this extension to be useful in the construction of general data exchange infrastructure for wide area biomedical data warehouse dedicated to research and clinical purposes.

ACKNOWLEDGEMENTS

We would like to thank George Kremenek for enlightening discussions on SRB internals and Matteo Vocale for support on network security issues.

References

1 Data and Metadata Collections for Scientific Applications, Arcot K. Rajasekar and Reagan W. Moore, European High Performance Computing conference, Amsterdam, Holland, June 26, 2001..

2 Grid Computing: Making the Global Infrastructure a Reality, F.Berman, G. C. Fox, and A. J. G. Hey (eds.), Wiley, 2003.

3 The Biomedical Informatics Research Network, The Grid, Blueprint for a New Computing Infrastructure Peltier ST, Ellisman MH (2003). 2nd edition: Elsevier (in press).

4 http://nscp.upenn.edu/ndma.

5 eDiaMoND: A grid-enabled federated database of annotated mammograms, J. M. Brady, D. J. Gavaghan, A. C. Simpson, M. M. Parada, and R. P. Highnam.. In F. Berman, G. C. Fox, and A. J. G. Hey, editors, Grid Computing: Making the Global Infrastructure a Reality, pages 923--943. Wiley Series, 2003.

6 MammoGrid: Large-Scale Distributed Mammogram Analysis, S. R. Amendolia, M. Brady, R. McClatchey, M. Mulet-Parada, M. Odeh and T. Solomonides. Proceedings of the XV111th Medical Informatics Europe conference (MIE'2003). St Malo, France May 2003. Volume 95 of Studies in Health Technology and Informatics, pp 194-199. ISBN 1 58603 347 6. IOS Press, Amsterdam.

7 Virtualization services for Data Grids, Reagan W. Moore and Chaitan Baru in "Grid Computing: Making the Global Infrastructure a Reality", F.Berman, G. C. Fox, and A. J. G. Hey (eds.), Wiley, 2003.

8 Storage Resource Broker, Version 3.1, SDSC (http://www.npaci.edu/dice/srb).

9 Storage Resource Broker - Managing Distributed Data in a Grid, Arcot Rajasekar, Michael Wan, Reagan Moore, Wayne Schroeder, George Kremenek, Arun Jagatheesan, Charles Cowart, Bing Zhu, Sheau-Yen Chen, Roman Olschanowsky, Computer Society of India Journal, Special Issue on SAN, Vol. 33, No. 4, pp. 42-54 Oct 2003.

10 DICOM@OFFIS project Web site: http://dicom.offis.de/

11 MySRB & SRB - Components of a Data Grid, Arcot Rajasekar, Michael Wan and Reagan Moore, The 11th International Symposium on High Performance Distributed Computing (HPDC-11) Edinburgh, Scotland, July 24-26, 2002

12 Data Grids, Collections and Grid Bricks, Arcot Rajasekar, Michael Wan, Reagan Moore, George Kremenek, and Tom Guptill, 20th IEEE/ 11th NASA Goddard Conference on Mass Storage Systems & Technologies (MSST2003) San Diego, California, April 7-10, 2003.