Grid Computing at the University Hospitals of Geneva

Henning Müller, Arnaud Garcia, Jean-Paul Vallée and Antoine Geissbuhler
University Hospitals of Geneva, Division of Medical Informatics
Rue Micheli-du-Crest 24, 1211 Geneva 14, Switzerland
henning.mueller@dim.hcuge.ch

Running head: **Grid Computing at the University Hospitals of Geneva**

Contact author: **Henning Müller**
Tel ++41 22 372 61 75
Fax ++41 22 372 86 80

Abstract

Although no implementation with respect to grid technologies has taken place as of yet, the University Hospitals of Geneva plan to apply these technologies in several pilot studies. The implementations will first take place in research projects but grid networks are also seen as a strategic long-term technology. Grid-enabled applications can change the way the resources in a hospital are used. Computation on demand can soften performance peaks and allow new diagnostic tools. Storage grids can ease the exchange of information in a secure way, where all the partners have their view of the data that they are allowed to access and where distributed resources are seen as one single, large hard disk.

The fact that there are more than 5,400 computers in the University Hospitals of Geneva with most of them being used at most 5%-10% of the time shows how much potential there is to use the resources in a more efficient way. This article describes our vision for grids and the tools that need to be developed to use this technique in a hospital environment. The extension of this concept to the emerging community health information network is also considered.

Keywords: Grid computing, grid usage, image retrieval, text mining, hospital infrastructure

1. Introduction

In medical hospitals, the amount of information and especially digital information produced and used is rising strongly. In the Radiology Department of the University hospitals of Geneva, for example, the number of images produced daily (in 2002) is at more than 12,000 which results in more than 4 million images per year accounting to roughly 2 Terabytes of image data.

Other departments might not have quite as much data but it all still amounts to large quantities that have to be stored and also made accessible for example in the electronic patient records. Data comes from various sources in varying formats and is distributed using various information systems and protocols. The same is valid for the exchange of information with other institutions. Data have to be transmitted to other medical practitioners, to other hospitals and anonymized data also needs to be transferred to the public offices for statistics and to insurance companies for billing.

While this is mainly a data storage and communication problem, in several research fields as well as in imaging diagnostics, there is a need for powerful computational resources, regularly. In general, medical departments do not have the computing power that particle physics institutes such as the CERN[1] have as these can cost millions of Swiss Francs, which are not available for a hospital research computer infrastructure.

Grid networks [1,2] have the potential to help with these two problems. They are forecasted to be for computation what the Internet was for communication. A large set of computers can, in principle, be used as one large resource for the storage of data and for computation. With authentification of the users, access to computational and storage capacities can be granted or denied. Like this, virtual organisations can be created that share, depending on projects, the same resource for computing and storage. This can be the case on several levels, *i.e.* on a hospital level, a regional health care provider level and also on a project level, where resources from centres such as the CERN might be made available for certain computational tasks. Such a computation on demand can also open up new diagnostic methods and tools when speeding up procedures.

These virtual organisations to share data and computational power can also be imagined within research projects between several institutes.

Of course, the data also needs to be in a well-defined format such as XML to make the exchange possible and easy. Sharing of data alone does not necessarily help much. Proprietary applications need to be replaced by open ones.

2. Computer architectures in hospitals and the storage of data

Basically all hospitals have computerized information systems to store and access all sorts of data, from the administrative data such as goods purchase in the hospital information system to treatment data in clinical information systems and images in the PACS (Picture Archiving and Communication System). Various departmental information systems exist (RIS-Radiology, LIS- Laboratory).

The sources of the data are extremely varied and there exist information systems at various levels such as the systems for departments. Often, the various information systems for departments and on a global level are connected with communication interfaces using standard communication protocols such as HL7[2] for patient or diagnostic data or DICOM[3] [3] for image data. Although there exist standards describing these communication protocols, interfacing still remains a problem and many interfaces rest proprietary because the standards do not cover all the needed functionalities.

In a large hospital there are thousands of computers to access the stored data when and where they are needed. To have all the data accessible from all these computers, large integrated systems are needed to administer the integrated, electronic patient record.

2.1 University Hospitals of Geneva

The University Hospitals of Geneva[4] (HUG) have been pioneering the development of hospital information systems since the 1970s. The PACS system for image storage was one of the first of its kind in the early 1990s. By now, the entire image producing, diagnostics and storing process is digital.

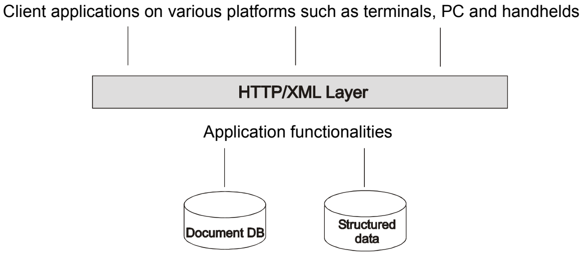
Client applications on various platforms such as terminals, PC and handhelds
HTTP/XML Layer
Application functionalities
Document DB
Structured data

Figure 1: The model of the hospital information system at the University Hospitals of Geneva.

The information system has recently been redeveloped using a component architecture, with many technologies being based on web standards such as XML for describing data and HTTP for communication [4]. Figure 1 demonstrates this model when clients access functionalities via HTTP/XML

[1] http://www.cern.ch/

[2] http://www.hl7.org/

[3] http://medical.nema.org/

[4] http://www.hcuge.ch/

that have access to the various databases containing documents and structured patient data.

Such an architecture allows having a set of standard interfaces and eases the reuse of functionalities and the data interchange. Functionalities can be used by several subsystems in exactly the same way.

A tool that is used for many research projects internally is the image case database system Casimage[5] [5]. This system is routinely used by Medical Doctors (MDs) in the hospital to store interesting or typical cases with the corresponding images. Its main use is for research and for teaching but it can as well be used for diagnostics in domains such as evidence-based medicine or case-based reasoning. For this, new search facilities might be useful or even necessary.

3. Architecture for research in the medical environment

Departments of Medical Informatics in large University Hospitals often have a significant number of employees, in the case of Geneva, more than 50 people. Still, most activities are not for research but for service with respect to the hospital infrastructure such as the electronic patient record or the PACS system to store images. Often, the two are combined to transfer results from research into routine use.

Research groups in medical informatics work on a large variety of different projects from image processing to database technologies, text mining and security aspects or mobile computing. Many of these activities demand the storage and analysis of very large quantities of data and the computational need can as well be very high. Still, rarely, very powerful computer are accessible by these research departments as the focus of the departments work is often on the service of the important hospital infrastructure for diagnostics, data storage and data access rather than research.

For new research areas such as the integration of grids into the hospital information structure the means are often even more restricted. At the University Hospitals of Geneva, currently two persons are working to obtain knowledge on the implementation and use of grid technologies while developing pilot applications. All this is an additional workload to the usual charges and no dedicated person for applying these technologies exists. Other hospitals will most likely not have more resources for similar technologies and projects.

4. A vision of grids

Some of the ideas on grid networks in this section might yet be a bit naïve but they are not from a network expert's point of view but from a user standpoint and might omit some of the domain problems.

The big idea of coupling several single resources to create a larger one sounds very appealing. Especially the integration of strong authentification routines into this concept is very important for the medical field. Such an infrastructure can allow having a physical structure almost as it is at the moment, and at the same time the logical structure can be fairly different. Virtual organisations can be created in research projects or in other cooperations that share storage and/or computational resources. Figure 3 demonstrates this idea, where different institutions like the CERN, the University Hospitals of Geneva (HUG), statistics institutes and also other MDs can create a single grid network where depending on the access rights of a certain person or institution, different resources can be used and different data can be accessed. Normally, the logical structure is larger than the physical one as resources are shared between partners.

This can help in medical research projects, where the computational power of the CERN can be used for short periods of time and also the CERN that might need test users for their technologies, can profit from these tests. Statistics institutes can directly access the data in the hospital information system that they are allowed to use and the information exchange can get much easier as data do not need to be transmitted and be stored in different versions and with changes being made regularly. MDs can also access actual data which can help to avoid errors due to a bad information flow which are frequent in medicine [6,7] and according to several studies cause thousands of deaths per year in the US alone.

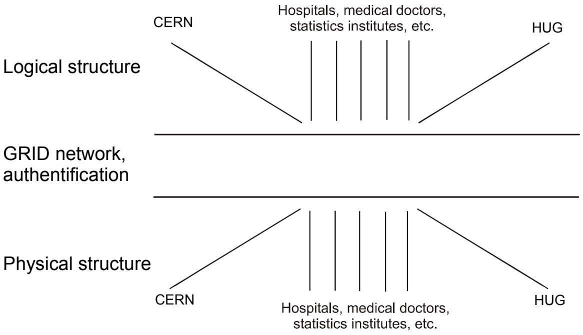
CERN
Hospitals, medical doctors, statistics institutes, etc.
HUG
Logical structure
GRID network, authentification
Physical structure
CERN
Hospitals, medical doctors, statistics institutes, etc.
HUG

Figure 2: A grid network where the same players have a physical infrastructure and a larger logical structure including the access to all resources that they are permitted to use.

Such a common infrastructure can also ease other development projects such as a common Geneva health information system [8]. This can not only improve the quality of the data but also reduce development and data interchange costs and accelerate the development. Health professionals could efficiently share patient-related information, and decision makers would have easy access to up-to-date anonymized data.

[5] http://www.casimage.com/

Still, not all problems will be solved that easily. The various different data formats and many proprietary applications will stay to be used for a while and format converters will continue to be useful. But it is at least a first step to create an infrastructure for easy communication and sharing of computational capacities.

5. Useful applications of grid technologies in the University Hospitals of Geneva

This section compiles a few test applications that are planned to be "griddified" in a first step. These applications are chosen because of their relevance to current practice and the expected benefit that the grid can have for these specific applications. The applications already exist and they are relatively easy to parallelize. They can as well be tested with anonymized data so it is not necessary to have a grid cluster within the hospital but resources can be used elsewhere for some first tests.

5.1 Content-Based Image Retrieval

Content-based image retrieval is an extremely active research domain that is trying to develop tools and methods to manage the large amount of visual and multimedia data that are produced and made accessible via the Internet. In a hospital environment, the generation of multimedia data for diagnostics, teaching and therapy planning is also on a rise. These methods normally use automatically extracted visual features such as colours, contours and textures to describe and search images as contrast to other search methods, which are usually text-based. A good overview of the field with many references can be found in [9].

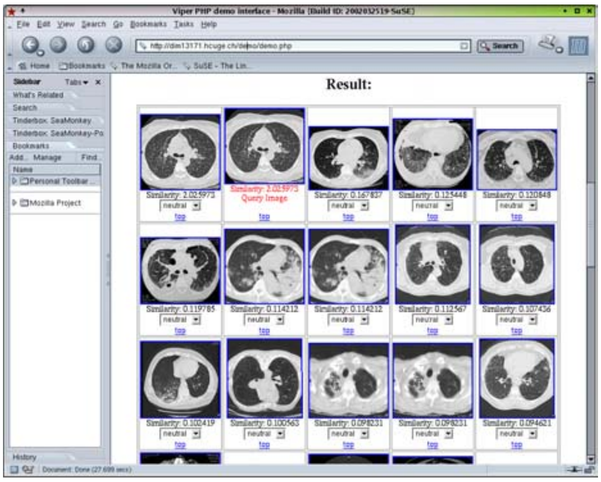

Figure 3: The interface of the medGIFT.

With the availability of the open source image retrieval system GIFT[6] (GNU Image Finding Tool) a tool is available to easily build content-based retrieval applications for several domains. This project was founded from the University of Geneva's Viper[7] project [10,11]. An overview of the architecture of the system can be found in [12]. The system is component-based and to access the retrieval engine, a communication protocol has been developed to separate interface and search engine and also to allow access for other applications such as meta search engines. The language for the communication between the interface/programs and the search engine is called MRML[8] (Multimedia Retrieval Markup Language). Figure 3 shows the interface of the program with the result of an example query.

The system can profit from both main aspects of grids, the storing of a large amount of visual data and the need for computing power to quickly return calculated retrieval results to the user. Most of the algorithms are fairly easy to parallelize which has already been implemented using PVM (Parallel Virtual Machine) in [13].

Main applications for the use in grids are the computationally intense extraction of visual features and as well the execution of queries that access the large amounts of visual features that are stored in an inverted file index for efficient searching. Our current image database contains more than 25,000 images which results in a feature index of more than 1 Gbyte. The GIFT system is in the process of being integrated into the CasImage system, which is a medical case database system that includes image data and corresponding annotations. This will insure the routine use of this search technology and will create a need for fast indexing and quick responses to visual queries. CasImage is in routine, daily use in the hospital since more than two years.

5.2 Text Mining

Text mining in the medical environment poses different problems, as it is more a question of the large amount of data that needs to be analysed. Text mining techniques can be used on various medical data, mostly on the electronic patient record that contains large amounts of more or less structured data. The quality of the data is not always extremely high as much of the content is intended for internal use only, anyways. Thus algorithms to correct these data are one first step before further analysis and retrieval can take place [14].

These technologies can profit from the use of storage grids where large amounts of data can be accessed quickly and in a distributed way.

Text mining techniques can be used for text categorisation where concepts are automatically assigned to pieces of text. This often implies the use

[6] http://www.gnu.org/software/gift/

[7] http://viper.unige.ch/

[8] http://www.mrml.net/

of extremely large, controlled vocabularies that could also be used efficiently in storage grids.

Text categorisation is often used in medical informatics: for billing or decision-making purposes or for cross lingual information retrieval (UMLS, [15]). Most advanced categorisation approaches to handle large sets of categories (between 20,000 (MeSH) to 2,000,000 (UMLS)) cannot use machine-learning strategies (k-nearest neighbours, SVM, Linear Least Square Fit) for computational reasons. Therefore, the use of a GRID-like architecture could help in order to scale up such data-intensive approaches.

5.3 3D Image Processing

For 3D image processing or other image processing algorithms such as fusion of images from various modalities, large computational power is needed. Especially the idea to make these algorithms interactive needs access to fast computing resources.

To integrate these techniques even better into the diagnostic process, it is necessary to make them fast. Having much shorter reaction time for image fusion or functional MRI (fMRI) can also open up new diagnostic possibilities as images can be created and processed during an operation and not necessarily a day in advance. This can help to recognize the shift of organs in the body after opening and allow easier recognition and navigation, for example.

Many of these algorithms are relatively easy to parallelize. Segmentation algorithms can be executed on several machines slice-by-slice and only little communication is needed for optimisation. The same holds for 3D reconstruction where nodes can work on parts of the data and communicate the results for display.

6. Needs for grid technologies and steps towards an implementation

Although several projects on grid technologies exist and a large amount of software has been developed in these projects, there is not yet a free “out-of-the-box-solution”. Most projects such as the Globus toolkit or the DIET software [16] are fairly difficult to install and work only in a very strictly defined environment on one operating system. Creating local miniGrids is thus fairly difficult and demands very experienced system administrators.

Current technologies are also not really compatible among each other. They are developed in projects and it is important for choosing one of the packages that they are continued to be supported in the long term. Otherwise, a project can be hindered strongly by a wrong choice of platform or middle ware.

The usage of running grids is not extremely easy, either. Normally, no easy-to-use application programming interface (API) exists for the current middle ware applications.

6.1 Needs for grid software to make it usable in a medical environment

There are a number of projects that produce software that can help to install and use grids. This includes the Globus[9] toolkit that is part of the European dataGrid[10] project.

The Unicore[11] forum is another option that allows downloading software to ease the development of grids.

Besides these developments what is really needed is an easy to use *application programming interface* (API) that allows the development of applications that use the resources a grid provides.

This would perfectly be usable on *different platforms* such as PCs under Windows and Linux, Apple computers and Unix systems from various vendors, so that existing, heterogeneous computer infrastructures with a large number of PCs (as they exist in the hospital) can be reused. Tools for *load balancing* are also indispensable to be able to distribute the charge on the machines depending on their computational power and the needed bandwidth to transmit data.

Extremely *fast communication links* are needed if a lot of data needs to be transmitted by the used algorithms

The *parallelisation* of existing algorithms also needs to be made as easy as possible so existing software can easily be transferred into efficient grid-applications. Partly this might be possible within compilers but for compilers there will definitely be limits and some manual intervention and program planning is for sure necessary.

The software to install and *manage grid nodes* also needs to be as easy as possible to install. A good option would also be to not only use specific grid computers for inclusion into a grid node but to be able to demand further resources among the PCs that are, for example, not used at a certain point. This can be somewhat similar to the SETI@home[12] project but would most likely require a much more complex structure to protect the data, for example.

Maybe the most important points to make grids usable in a medical environment are *security issues*! This goes beyond simple access right to data and encryption of the communication. In first steps, it is clear that only anonymized data can be used outside of the hospital environment to test applications on grids, even if all security mechanisms to protect the data do exist. Only after some confidence has been build up, it might be possible to use clinically relevant and personalized files for griddified

[9] http://www.globus.org/

[10] http://eu-datagrid.web.cern.ch/eu-datagrid/

[11] http://www.unicore.org/

[12] http://setiathome.ssl.berkeley.edu/

applications. Tests will need to be performed regularly to insure the security of grid systems.

For running large-scale grid applications these security mechanisms need to be well in place and tested.

6.2 Our further steps for implementation

Although our first goal was originally to build a miniGrid at the interior of the hospital to test technologies and couple a number of PCs for larger calculations, this idea has been omitted for now due to the large effort that would be necessary to maintain such a network.

Our first steps will thus be the *use of resources* from the CERN, which offered us this possibility. This will include the comparison of execution times of programs on normal PCs and in a grid environment. Later on, a *miniGrid* in the hospital seems reasonable that can be used for data storage and exchange and sharing of computational load in a couple of projects for research or for data exchange with statistical offices.

A first application to test in a grid environment is the pilot application for *content-based image retrieval*. This has several reasons. (1) The application is available in open source and thus relatively easy to change and adapt to grid specifications. (2) The application has two different parts that can both be parallelized separately, fairly easily. (3) The feature extraction and index generation phase can be extremely parallelized as the features are extracted from every single image in the database. The Casimage database currently contains 25,000 images that could be sent to 25,000 computers, each to extract the features of one image, in an extreme scenario for feature extraction. This could reduce the visual feature calculation phase from 14 hours (2 seconds per image) to a moderate 2 seconds plus a few seconds for all the data transmission.

The query phase is not as easy to parallelize but several computer can do part of the evaluation, and then, the results can be merged in the end. The gain for this will not be as high as the cost for merging is significantly high. Evaluations have to verify the gain or loss.

After this first test phase, the other pilot applications need to be implemented and also evaluated. Also the integration of the University Hospitals into a Geneva-wide medical information system to gather epidemiological data about the health system is an option, as proper data interchange between various partners can be eased by such a Geneva-wide grid network.

7. Conclusions

Grid technologies have a large potential in the medical domain as well as in others with respect to better use of existing resources but also for obtaining computation and storage on demand to make applications faster and data access easier.

This can not only be the case for medical research departments as it might be in a first phase but also in clinical routines where the availability of diagnostic tools in real time can help the decision making process and open up new possibilities. The storage of data within the hospital can also profit from the tools for secure storage and access to data and the exchange of information can be eased.

The idea to use all the 4,500 computers available for computationally intensive tasks when the computers are not used otherwise sounds very logical and can remove the need to spend much money on specialized high performance computers to realize quick response times. By simply using the existing resources more efficiently this money can be used elsewhere.

Still, much work needs to be done before grids can be programmed as easily as single computer systems at the moment. The middleware needs to be much easier to install grids and use the grid services within programs. A simple-to-use API needs to be created for such an easy access.

For the use in storage grids, a security infrastructure needs to be created that can subsequently be applied to the hospital so only anonymized data can get to the outside and only authorized personnel can access personal patient data.

Such a security and computing infrastructure can make the exchange of data within research projects, to official statistics offices and to other partners much easier than it is at the moment as this can create virtual organisations that share part of the data and/or part of the computing resources depending on projects or institutional links.

8. References

[1] I. Foster, C. Kesselman, Steven Tuecke, The Anatomy of the Grid – Enabling scalable virtual organisations, *International Journal on Supercomputer Applications*, 2001.

[2] F. Gagliardi, B. Jones, M. Reale, S. Burke, European Datagrid project: Experiences of deploying a large scale testbed for e-science applications, *Performance 2002*, pp. 480-500, 2002.

[3] B. Revet, DICOM Cookbook for implementations in modalities, *technical report Philips medical systems*, 1997.

[4] A. Geissbuhler, C. Lovis, A. Lamb, S. Spahni, Experience with an XML/http-based federative approach to develop a hospital-wide clinical information system, *Medinfo 2001*, IOS Press, Amsterdam, 2001.

[5] A. Rosset, O. Ratib, A. Geissbuhler and J-P. Vallée, Integration of a Multimedia Teaching and Reference Database in a PACS Environment, *Radiographics*, 22(6):1567-1577, 2002

[6] L. L. Leape, Error in medicine, *JAMA, The Journal of the American Medical Association* 272(23), pp. 1851-1861, 1994.

[7] L. T. Kohn, J. M. Corrigan, M. S. Donaldson, editors, To err is human – Building a Safer Health System, *National Academic Press*, Washington D.C., 1999.

[8] D. Hausser, Analyse stratégique de l'implantation du système de santé à Genève, *technical report*, IDHEAP, June 2002.

[9] AWM. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain, Content-Based Image Retrieval at the End of the Early Years, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(12):1349-1380, 2000.

[10] DMcG. Squire, W. Müller, H. Müller and J. Raki, Content-Based Query of Image Databases, Inspirations from Text Retrieval: Inverted Files, Frequency-Based Weights and Relevance Feedback, *Scandinavian Conference on Image Analysis (SCIA)*, pp. 143-149, Kangerlussuaq, Greenland, 1999

[11] D. McG. Squire, H. Müller, W. Müller, S. Marchand-Maillet and T. Pun, Design and evaluation of a content-based image retrieval system, *Chapter 7, Design and Management of Multimedia Information Systems: Opportunities and Challenges, Idea Group Publishing*, S. M. Rahman, editor, pp. 125-151, 2001.

[12] H. Müller, W. Müller, DMcG. Squire, Z. Pecenovic, S. Marchand-Maillet and T. Pun, An Open Framework for Distributed Multimedia Retrieval, *Computer Assisted Information Retrieval (RIAO)*, volume 1, pp. 701-712, Paris, 2000.

[13] H. Müller, D. McG. Squire, W. Müller, T. Pun, Efficient Access methods for content-based image retrieval with inverted files, *Multimedia Storage and Archival Systems IV, SPIE proceedings volume 3846,* Boston, Massachussetts, 1999.

[14] P. Ruch, R. Baud, A. Geissbuhler, Evaluating and reducing the effect of data corruption when applying bag of words approaches to medical records, *International Journal of Medical Informatics* 67, pp. 75-83, 2002.

[15] D. Eichmann, M. Ruiz, P. Srinivasan, Cross-Language Information Retrieval with the UMLS Metathesaurus, *Research and Development in Information Retrieval*, pp. 72-80, 1998.

[16] F. Desprez, DIET: A distributed interactive engineering toolbox for client-server applications in a grid environment, *ERCIM News*, No. 45, April 2001.