A Cloud Infrastructure for Scalable Computing on Population Imaging Databanks

Ignacio Blanquer

Institute of Instrumentation for Molecular Imaging (I3M), Univ. Politèc. de València. E-mail: iblanque@dsic.upv.es
GIBI230, Polytec. Univ. Hospital La Fe (HUPLF). Valencia, Spain.

Miguel Caballer

Universitat Politècnica de València - I3M, Spain

Luis Martí-Bonmatí

Biomedical Imaging Research Group (GIBI230), Radiology Department and Biomedical (HUPLF). Spain

Ángel Alberich-Bayarri

Biomedical Imaging Research Group (GIBI230), La Fe Health Research Institute. Spain

María de la Iglesia

Centre of Excellence in Biomedical Image (CEIB), Regional Ministry of Health in the Valencia Region (CS).
Brain Connectivity Lab. Unidad Mixta FISABIO/CIPF, Spain

Jacobo Martínez

Valencian Biobank Network, Foundation for the Promotion of Health and Biomedical Research of Valencian Region (FISABIO), Valencia

Abstract: This article describes the software architecture designed to cope with the computing demand of research usage of complex data from the Imaging Biobank of the Regional Ministry of Health in the Valencia Region (CS). It proposes the use of self-configured virtual clusters on top of on-premise and public cloud infrastructures. It uses a model based on recipes and autoconfiguration to deploy virtual elastic clusters that adjust themselves to the actual workload of the study, therefore reducing operating costs and preventing the need of up-front investments both at the level of the Imaging Biobank or the final user. All the software used is released under open-source licenses.

Keywords: Biobanks; Imaging Population; Cloud Computing; Virtual Infrastructures.

Biographical notes: Dr Ignacio Blanquer, professor in the DSIC department of the Universitat Politècnica de València since 1999 and GRyCAP member since 1993. He is currently the head of this research group. He has authored and co-authored more than 70 papers in national and international journals. His research interests include Grid and Cloud systems, Medical Image Instrumentation and High Performance Computation.

Miguel Caballer obtained the B.Sc. M.Sc, and PhD degrees in Computer Science from the Universidad Politécnica de Valencia (UPV), Spain, in 2000, 2012 and 2014 repectively. He is part of the Grid and High Performance Computing group of UPV since 2001. He has participated in different research projects about the application of Grid and Cloud computing techniques to several areas of engineering. Other fields of interest include green computing.

Dr Luis Martí-Bonmatí, professor of radiology, is chairing the radiology department and the biomedical imaging research group (GIBI230) at La Fe University and Polithecnical Hospital since 2011. He is president of the European Society of Gastrointestinal and Abdominal Radiology and Secretary General of the International Society of Radiology. He has authored and co-authored more than 300 papers in national and international journals. His main research interests include Radiology, Imaging Biomarkers, Imaging Processing, Abdominal Imaging and Magnetic Resonance Imaging.

Dr Angel Alberich-Bayarri, is the Scientific-Technical director of the biomedical imaging research group of La Fe University and Polytechnics Hospital. He is also founder of the start-up QUIBIM. He has authored and co-authored more than 40 papers in national and international journals. His research interests include quantitative radiology, image processing techniques, imaging biomarkers and imaging biobanks.

Dra. Maria de la Iglesia-Vaya. She is civil servant in the Valencia Regional Ministry of Health; nowadays she is the head of the Brain Connectivity Lab. in the Department of Neurological Impairment at Príncipe Felipe Research Centre of Valencia (joint unit FISABIO-CIPF). Also she is part of the Interim Board in Euro-Bioimaging as scientific delegated by Spain in the field of Medical Imaging since 2014 in the expression of interest "Population imaging". She has authored and co-authored more than 20 papers in national and international journals. Her research interest includes methodologies with high performance computational image used for the analysis of functional and structural brain imaging and the development of statistical, data mining methods and bioinformatics tools to understand and analyze information of population imaging.

Jacobo Martínez. BSc in Chemistry by the Valencia University (1995). Post graduate Diploma in Biochemistry and Clinical Pathologies by the Valencia University (1996). Master Degree in Biotechnology by the International Business School ALITER (2003). He has worked for different biotech companies dealing with business development issues. Since 2008, he is Manager of a population-based biobank, the IBSP-CV Biobank, and Head of the Valencian Biobank Network. He has been principal investigator of several national funded projects. He is also member of the General Council of the Spanish National Biobank Network.

1 Introduction

Population imaging and large data sets of medical images are a clear need. To further develop personalized medicine, quantitative medical imaging needs to clinically assess value, reproducibility and universality. By ensuring the development and implementation of signal modeling and image features extraction strategies, quantitative imaging will help to prevent or effectively treat diseases. Virtual biobanks, as repositories of digital information, have increased the opportunities for sharing, federating and exploiting large data sets of patients' information. Medical Imaging Biobanks can be considered as virtual biobanks. They recently emerged for advancing on the study of diseases, identification of early disease biomarkers and surrogates, and development of population studies. These imaging biobanks will evaluate the impact of new quantitative biomarkers, both resolved in space and time on parametric and nosologic images, on early disease diagnosis, disease phenotyping, disease grading, targeting therapies and evaluation of disease response to treatment.

Imaging and engineering technologies support retrospective and prospective epidemiological studies on the population level. Specific markers obtained by large scale image processing of pre-symptomatic diseases might be used to evaluate different biological and pathological abnormalities, new disease phenotyping and early identification of people at risk.

The European Strategy Forum on Research Infrastructures (ESFRI) mission leads to the better use and development of research infrastructures, both at the European and international levels. Within this initiative, Euro-BioImaging is an European project establishing a distributed set of open access nodes. Although many new imaging technologies and quantitative initiatives are arising, such as multiparametric approaches, their application to clinical practice is limited, most remaining on the research field. One important reason for this limitation is the need for standardized platforms in which different procedures, protocols and techniques can be evaluated on suitable images. Such platforms should be able to solve the strong requirements on data computing and storage. Our Euro-BioImaging node main objective is to set up an innovative platform for the execution of multicentre clinical trials and research projects based on medical imaging and imaging biomarkers. This advanced imaging platform will enable to test and validate state-of-the-art image analysis techniques, increase quality and accuracy of clinical trials through the use of advanced quantitative imaging techniques, promote new visualization tools and, finally, reduce overall costs.

Simlarly, in the last years several collaborative environments have been developed together with powerful tools for the exchange of knowledge among medical institutions, improving diagnosis and research. Most of them primary rely on medical images, and they use Grid and Cloud computing infrastructures to organize the access and the use of the various services offered by them. Such collaborative environments include the Biomedical Informatics Research Network (BIRN) from *Grethe (2005)*, the e-Science platform for researchers of Alzheimer's disease neuGRID (*Redolfi et. al. (2009)*), its recent successor N4U or the integrated healthcare platform for European paediatrics Health-e-Child from *Marcus et al. (2007)*. Representative examples of medical imaging applications exclusively using Cloud computing include the work of *Silva et. al. (2012)* for the storage of medical studies using public Cloud providers, and a Windows Azure-based DICOM server from *Teng et al. (2010)*. Those platforms offer fixed environments which are infrastructure-dependent.

2   Requirements

According to the European Society of Radiology (ESR) working group on Imaging Biobanks definition, Imaging Biobanks are defined as organized databases of medical images and associated imaging biomarkers (radiology and beyond) shared among multiple researcher, that should be linked to a biorepository as stated in *ESR WG on Imaging Biobanks (2014)*. The definition already considers the interaction of specialists from different disciplines with the platform.

2.1   Use cases as defined by applications of the platform

Therefore, the needs to be covered by this advanced imaging platform are directly related to the most relevant use cases for medical imaging and that are considered in its design, consisting on: clinical research, development of new imaging biomarkers, imaging-based clinical trials and biobanks.

The platform will be an environment for clinical research, where the users will have the capability of choosing the data they want to statistically process for their study, including not only images but also associated imaging biomarkers and meta-data in a way of customized research based on imaging. The data in the platform will be properly anonymised if not having written informed consent and de-identified in the case of having a written informed consent (e.g. those coming from clinical trials). The data will cover the entire spectrum of human disease and provide a platform where state-of-the-art image analysis algorithms can be evaluated on own data.

This population-based imaging biobank will be used also to fill the gap between research and clinical application in quantitative imaging and imaging biomarkers. Nowadays, there is a need to validate either existing or new quantitative imaging methods. As an example, several imaging biomarkers exist as candidates to early evaluate treatment response in oncology, but large cohorts are required in order to extract robust conclusions and relevant evidence for clinical practice and clinical trials. Image processing scientists will have the capability of uploading and sharing their own algorithms and methods with other researchers, to be tested with the studies in our platform. The platform will also ease the development and validation of novel image analysis techniques and their comparison with implementations of other research groups applied to the same cohorts.

The number of clinical trials based on imaging is progressively increasing, due to the better technological capabilities of the different modalities that allow for more precise and accurate diagnosis. However, the use of quantitative imaging in clinical trials is still an important challenge, since there is a lack in the standardization of image analysis procedures. Our platform will include image analysis algorithms following criteria of the scientific community and novel standardization initiatives like the QIBA (Quantitative Imaging Biomarkers Alliance).

Finally, the mechanisms for the connection of this imaging platform with existing biorepositories will be also considered. A biorepository is s biological material repository that collects, processes, stores, and distributes biospecimens to support future scientific investigation. Although a biorepository can contain or manage specimens from animals, including humans, and many other living organisms, in some discussions, such as in *ISBER (2001)*, the term is reserved for human specimens.

Since the late 1990s, biorepositories have became an important resource in medical research, supporting many types of contemporary research like genomics and

personalized medicine. For example, many diseases are associated with single-nucleotide polymorphisms, and performing genome-wide association studies using large collections of samples which represent tens or hundreds of thousands of individuals can help to identify disease biomarkers as stated by *Greely (2007)*.

Hewitt (2011) and many other researchers have identified biobanking as a key area for infrastructure development in order to promote drug discovery and drug development. These data are currently available in local databases and the mechanisms of connectivity and access to them will be addressed.

The connection of Biorepositories and medical image databanks enable the guidance to the development of advanced imaging surrogates that could lead to early diagnosis and to anticipate the most effective therapy.

2.2   Use cases as defined by the types of interaction

According to the way the resources will be used, three use cases are considered:

- *Interactive usage*. A project requires to analyse and process manually a set of images from the case studies selected. A single Virtual Machine - with a specific virtual hardware configuration- is needed. The Virtual Machine is properly configured with a set of tools and Interactive access is required.

- *Batch*. A project requires applying a straight but computationally intensive processing on a wide range of images. Unattended execution is possible and desirable. Once the parameters are fixed, a simple combination of programs is executed over the whole bunch of images. It will require multiple Virtual Machines and a local resource management system (such as Torque, PBS, LSF, OGE, etc.)

- *Pipeline*. A project requires applying a complex processing on a set of images requiring multiple steps and synchronizations with different requirements (for example, it requires merging the results from previous steps). It will require multiple Virtual Machines and a pipeline manager able to work with the infrastructure.

2.3   Image processing requirements

The platform will integrate data with a high degree of heterogeneity, like medical images with all the interpretations of DICOM (http://dicom.nema.org) standards from the different manufacturers, parametric maps of imaging biomarkers, 3D reconstructions of anatomy, source codes of image processing algorithms, and associated clinical data and variables. The main data types managed by the platform will be plain DICOM files (1 DICOM file per image) and also NIFTii and Analyze formats (*.nii and *.hdr/*.img file extensions, respectively), that can handle entire volumes in a single file and are widely extended among medical image processing scientists community.

For data conversion, the mricron (McCausland Center for Brain Imaging, Columbia SC, USA) package will be included in the platform. For data analysis the platform will include a broad range of tools, such as the FSL (FMRIB Software Library, Analysis Group, Oxford, UK) and SPM (Wellcome Trust Center for Neuroimaging, University College London, London, UK) packages that are specially suited for neuroimaging analysis, and a connector for Python scripts, in order to allow the evaluation of algorithms developed by the users of the platform.

2.4   Pipelining and processing

A pipeline-based architecture will be used for the development of quantitative imaging procedures. Image analysis methods will be classified in those used for quality control data (i.e. signal noise ratio plot), data pre-processing (i.e. segmentation, filtering, interpolation), data analysis (i.e. brain volumes quantification, T2 mapping, perfusion analysis, lung emphysema quantification) and data measurement (i.e. histogram based analysis, multi-variate statistical analysis). Rules will be defined for the interconnection of the different types of processing modules for pipeline creation.

3   Materials

According to the above requirements, the management of individual subprojects has to deal with the following points:

- Each subproject will have specific requirements (in terms of resources and applications) that have to be fulfilled individually.

- Multi-tenancy of the computing infrastructure does not need that data can be accessible by different projects. Data, even anonymised, and produced results must be protected from the access of unauthorized users.

- Projects may require resources or have external resources available, and the applications should work seamlessly.

In this scenario, we define a software architecture that instantiates Virtual Appliances customized for each subproject. A Virtual Appliance is a set of fully configured virtual machines that run coordinately.

In all cases, the infrastructure management service will provide fully-configured Virtual Appliances. The software required will be described as auto configuration recipes. In this part we envisage two types of software components:

- *Commodity components*. The Valencia EuBio node (BIMCV) offers a range of widely used open-source processing programs and toolkits such as FSL, SPM, mricron. Those components are available in a market place as a set of customised recipes that are used by the Infrastructure Manager for setting up the resources. Those recipes include the minimum virtual hardware requirements, which can be overwritten in each specific project if needed.

- *Custom components*. Each project can include their own components to the Virtual Appliances. Users are encouraged to prepare configuration recipes to facilitate auto-scaling automatic configuration. Nevertheless, users can manually administer their own virtual resources adding more components. However, this last feature is limited to the configuration running on end-users' premises or public clouds, to reduce the risk of running malicious software on BIMCV's premises.

Therefore, our platform defines the following services:

- A Catalogue of images of preconfigured Virtual Machines including plain images for further contextualization and complete stacks.

- A recipe registry with automatic installation instructions for a wide range of processing components.

- An Infrastructure manager for the deployment and reconfiguration of virtual appliances.

- A service for the automatic management of elastic clusters based on the workload.

- A workflow manager able to deal with elastic hybrid pipelines.

- A computing infrastructure based on a cloud on-premise.

- A final user application that enables requesting the virtual appliances and deploying them on the desired infrastructure.

The previous services are described in the following sub-sections.

3.1 Virtual Resource Management Catalogues

The widespread usage of virtualization requires creating Virtual Machine Images (VMIs), which encapsulate the required hardware and software configuration required for the applications. Different applications impose different software and hardware requirements, which leads to a massive creation of VMIs which must be properly indexed in order to leverage both their reuse and sharing. For that, *Carrión et. al (2010)* proposes a catalog and repository of Virtual Machine Images (VMIs) that enables the users (and/or Cloud administrators) to index and store the VMIs together with the appropriate metadata that describes their hardware and software features. The VMRC system can be deployed within a private Cloud to store the different VMIs required to provision the services. It can also index VMIs stored within other Cloud providers. In addition, it can be employed to federate different private Clouds so that a single VMRC entity can provide the images for different Clouds. The catalog itself is hypervisor-agnostic and it uses the Open Virtual Format (OVF) specification. The catalogue is extended with a repository of Ansible-based recipes (*Hochstein (2014)*), which enables setting-up not only the proper basic VMI, but additionally the software to be configured in the Virtual Appliance. In this way, software configurations are described as a set of modules to be included in the final set-up.

3.2 Automatic Configuration of Virtual Appliances

Cloud infrastructures can be used to address the computational needs in multiple domains, including scientific applications. However, the use of public or on-premises Infrastructure as a Service (IaaS) clouds require non-trivial system administration skills. If a single Virtual Machine (VM) is needed, custom configurations available in public clouds may be sufficient to meet users' requirements. When multiple VMs are needed, users have to configure shared directories, batch queues, administer users and customise software. For that, the Infrastructure Manager (IM), firstly presented in *Caballer et. al (2014); de Alfonso et. al. (2011)*, is a tool that ease the access of IaaS clouds by automating the selection of the proper Virtual Machine Image, and the further deployment, configuration, software installation, monitoring and update of Virtual Appliances.

IM provides an abstraction layer to enable the interoperability with different IaaS clouds. This layer has been designed using a plug-in scheme that currently provides plug-ins for:

OpenNebula, OCCI, Amazon EC2, Google Cloud, Docker, OpenStack and libvirt. This set of plug-ins enables the access to a large number of cloud deployments and virtualization platforms, thus enabling the user to start using a simple virtualization system and then transparently migrating to the cloud. It integrates Ansible as the contextualization system to enable the installation and configuration of all the user required applications providing the user with a fully functional infrastructure. Despite that multiple cloud providers offer services for the deployment of software configurations (e.g. HEAT from OpenStack or the Software Management and Catalog service of OpenNebula), IM recipes are compatible to all the platforms supported via plug-ins.

Figure 1 shows two examples of configuration recipes. The one in the left describes the structure of a configuration of a medical imaging storage node, which includes four parts:

- The description of the virtual hardware.

- The deployment of the node.

- The common configuration recipes.

- The instantiation of variables specific of this node.

Despite that on-the-fly configuration introduces an overhead, this overhead is only compulsory for creating the virtual images for the first time. VMs can be preserved without requiring being active along the life-time of the project. Snapshots of the VMs can be performed and registered in a catalogue for longer-term hybernation. An interesting alternative, which is also supported in IM, is the use of docker containers. This will enable a binary-compatibility level with cloud providers (including public clouds like the container service of Amazon). IM can instantiate those containers and they can be registered in a catalogue for further reuse.
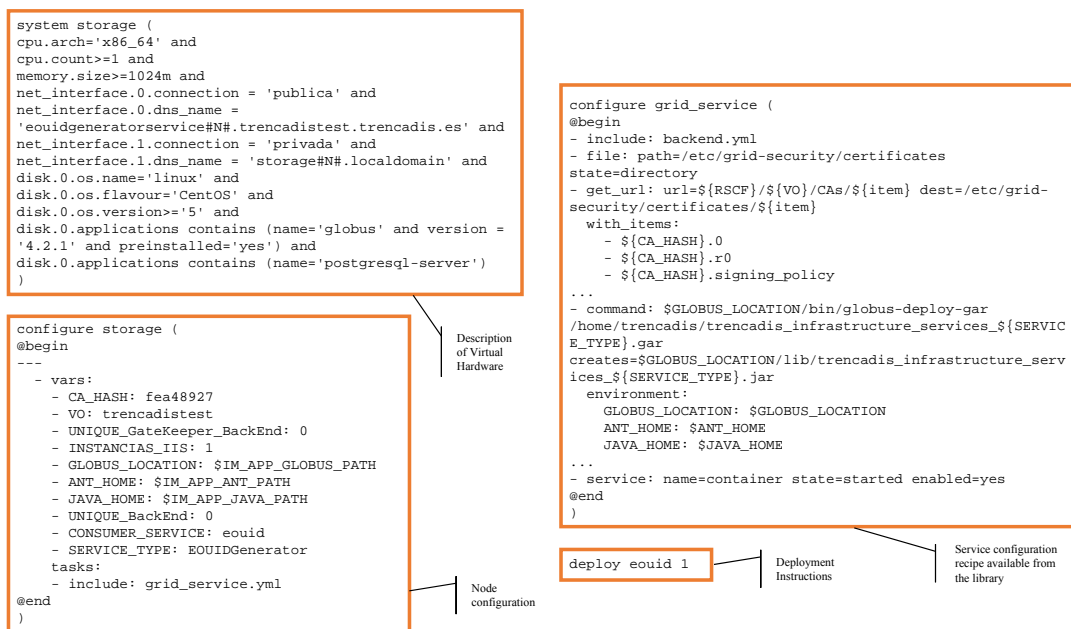
3.3   Elastic Virtual Applicances

Elastic Cloud Computing Cluster (EC3) (*Caballer et. al (2013)*) is a tool to create elastic virtual clusters on top of Infrastructure as a Service (IaaS) providers, either public (such as Amazon Web Services) or on-premise (such as OpenNebula and OpenStack). We offer recipes to deploy TORQUE (optionally with MAUI), SLURM and SGE clusters that can be self-managed with CLUES: it starts with a front-end node and working nodes will be dynamically deployed and provisioned to fit increasing load (number of jobs at the LRMS). Working nodes will be undeployed when they are idle. This introduces a cost-efficient approach for Cluster-based computing. Clusters can be complemented in a hybrid model by the creation of VPNs that extend a virtual cluster (*Amanda et. al. (2014)*) deployed in on-premise infrastructures and complemented with additional resources from a public cloud provider. This model is feasible when the inter-node communication and the access to a shared file system is small in comparison with the computational cost.

3.4   Infrastructure

The platform is designed to be 'cloud-agnostic', being able to work with multiple infrastructures. In the present time, three kinds of infrastructures are considered: Local infrastructure of BIM-CV. This infrastructure is provided by the UPV and the CIPF, with a total capacity of 100 virtual instances. It is currently provisioned using OpenNebula.

Figure 1 Description of the configuration of a Virtual Appliance. Example of a storage resources

```
system storage (
cpu.arch='x86_64' and
cpu.count>=1 and
memory.size>=1024m and
net_interface.0.connection = 'publica' and
net_interface.0.dns_name =
'eouidgeneratorservice#N#.trencadistest.trencadis.es' and
net_interface.1.connection = 'privada' and
net_interface.1.dns_name = 'storage#N#.localdomain' and
disk.0.os.name='linux' and
disk.0.os.flavour='CentOS' and
disk.0.os.version>='5' and
disk.0.applications contains (name='globus' and version =
'4.2.1' and preinstalled='yes') and
disk.0.applications contains (name='postgresql-server')
)
```

Description of Virtual Hardware

```
configure storage (
@begin
---
  - vars:
    - CA_HASH: fea48927
    - VO: trencadistest
    - UNIQUE_GateKeeper_BackEnd: 0
    - INSTANCIAS_IIS: 1
    - GLOBUS_LOCATION: $IM_APP_GLOBUS_PATH
    - ANT_HOME: $IM_APP_ANT_PATH
    - JAVA_HOME: $IM_APP_JAVA_PATH
    - UNIQUE_BackEnd: 0
    - CONSUMER_SERVICE: eouid
    - SERVICE_TYPE: EOUIDGenerator
    tasks:
    - include: grid_service.yml
@end
)
```

Node configuration

```
configure grid_service (
@begin
- include: backend.yml
- file: path=/etc/grid-security/certificates
state=directory
- get_url: url=${RSCF}/${VO}/CAs/${item} dest=/etc/grid-
security/certificates/${item}
  with_items:
    - ${CA_HASH}.0
    - ${CA_HASH}.r0
    - ${CA_HASH}.signing_policy
...
- command: $GLOBUS_LOCATION/bin/globus-deploy-gar
/home/trencadis/trencadis_infrastructure_services_${SERVIC
E_TYPE}.gar
creates=$GLOBUS_LOCATION/lib/trencadis_infrastructure_serv
ices_${SERVICE_TYPE}.jar
  environment:
    GLOBUS_LOCATION: $GLOBUS_LOCATION
    ANT_HOME: $ANT_HOME
    JAVA_HOME: $JAVA_HOME
...
- service: name=container state=started enabled=yes
@end
)
```

```
deploy eouid 1
```

Deployment Instructions

Service configuration recipe available from the library

External infrastructures. In this point we consider both public infrastructures such as Amazon AWS and research infrastructures such as EGI Federated Clouds. The framework supports already both AWS EC2 (Amazon proprietary API) and r-OCCI (the standard interface for managing cloud infrastructures, supported by EGI Federated Clouds). We use end-users' own credentials for the former and robot certificates for the latter. End-users' own infrastructure. The infrastructure management service is able to deploy the VMs directly on an end-user's on-premise cloud (supporting both ONE and OpenStack, or even libvirt-based resources).

The virtual infrastructures will expose a XNAT (eXtensible Neuroimaging Archive Toolkit, http://xnat.org/) instance. *Marcus et al. (2007)* is an open source platform designed to facilitate the management of image sets and associated data (assessments, reconstructions and any other information). Initially it is designed to work with neuroimaging but the open data model and customizable XML-based technologies allow to adapt the platform for any type of bioimaging. XNAT follows a three-tier architecture that includes a data file, an user interface and a middleware engine. The data file can be incorporated into the platform through different ways, such as XML files, web forms, DICOM transfers from image capture devices or image viewers like Oxiris and so on. Among its most important features are the personalized safe access to information, quality control processes of data and image information, classification and storage of data, ability to run custom searches, communication with bioimaging generating systems, programmability of process flows

using scripts (pipelines), the incorporation of intermediate results and conclusions to the study, etc. All these features make XNAT an ideal platform for the management of clinical trials.

4   Methods

The general procedure BIMCV addresses goes through the following steps:

- A proposal of a project requesting data is prepared using a set of standardized forms. These forms include the needs for data, software and computing resources. The proposals are presented in a two-stage model. In the first stage, the proposal is evaluated by a technical committee, that checks the viability of the request, the availability of data and the ethical compliance. After a positive answer, proposers must present a more detailed version of the project which is evaluated by a scientific committee.

- The scientific committee revise the project and proposes a final assignment of resources (if needed). Those resources are defined as a credit in hours that can be spent on BIMCV's computing infrastructures.

- Data is extracted and anonymised from the regional PACS and manually verified by an expert, which complements the data with the necessary clinical information. The result is a dataset that is copied into a virtual disk volume. This virtual disk volume will be attached to the virtual infrastructure.

- A technical expert defines the proper configuration of the virtual appliance. A recipe is created with the components selected and the features of the virtual hardware. This recipe will be used to instantiate and reconfigure the virtual appliance.

- Deploy and monitor the virtual appliance in the infrastructure proposed. In case of using BIMCV's infrastructure, a real-time accounting is provided to check the availability of resources.

Next section describes the deployment architecture.

Figure 2   Deployment of a project-specific Virtual Appliance
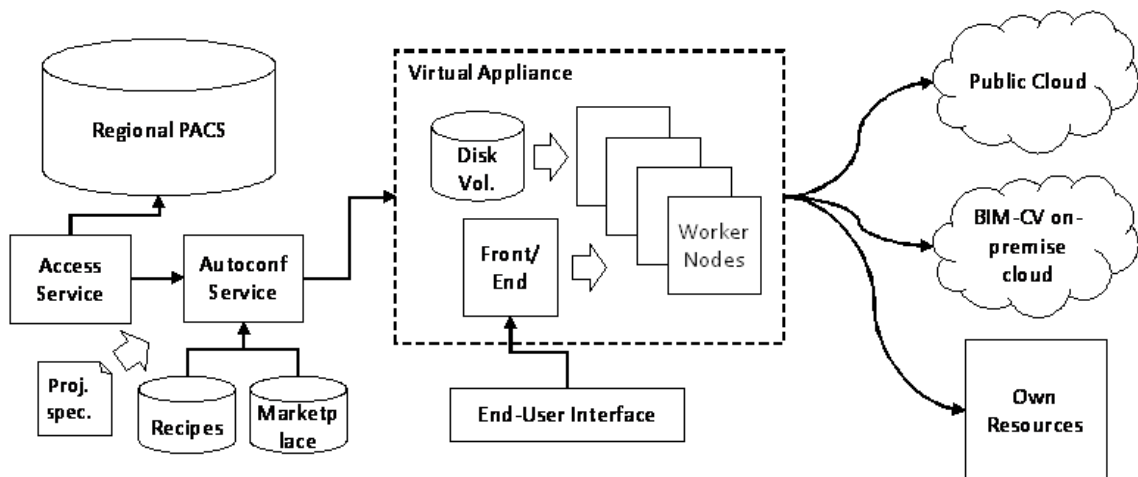
4.1 Architecture Deployment

A virtual cluster is assigned to each project. This virtual cluster uses a base Virtual Image from the Virtual Machine Resource Catalogue (VMRC) and contextualizes it with two types of instances:

- A master node that manages the virtual cluster through a batch queue (using torque). This master node is also instrumented with the automatic elasticity component [10]. This master node will interact with the IM to deploy or undeploy a number of working nodes.

- A working node image that will be used for processing the individual batch jobs that a pipeline generates. This stateless working nodes will be instantiated by the master node.

Figure 2 shows the deployment of a project-specific Virtual Appliance. The data resulting from a query to the regional PACS is prepared in a virtual disk volume, which will be part of the Virtual Appliance. The autoconfiguration accesses the recipes database and using the project specification configures two VM images with the front-end and the working nodes. Those images are instantiated in the desired infrastructure. The service is infrastructure-agnostic, supporting OCCI, EC2, ONE, OSTACK and libvirt interfaces currently, so it can work with public clouds, the on-premise cloud of BIM-CV or end-users' own resources. End-users can connect through the front/end node, which provides the desired interface according to the components and recipes available in the marketplace.

5 Conclusions and Further Work

The need of computing platforms in population imaging analysis is clearly justified by the huge demand of resources that requires processing large cohorts of complex images. In this sense, cloud computing has emerged as a solution to feed this request. However, the direct usage of public cloud infrastructrues undercover several challenges that are addressed by the work presented.

Each public cloud provider has its own (incompatible) way to represent VMIs. Migration from one provider to another is not automatic since it will require converting VMIs and reimplement the elasticity management. Moreover, the configuration of virtual clusters require system administration skills. Vendor lock-in, lack of control, IaaS management and security refrain Medical Imaging researchers to adopt cloud computing in a wider scale. In this sense, the use of recipes and on-the both eases infrastructure deployment and isolates from vendor's VMI particularities. Moreover, infrastructures can be even deployed on local on-premise infrastructures. Finally, recipes can be exposed in a library and re-used by non-expert users to deploy complex configurations.

The deployment on the fly increases the delay on the availability of resources. However, this delay is only suffered in the first instatiation of the virtual infrastructure and *Caballer et. al (2014)* shows reasonable re-configuration delays when new resources are added or removed from the infrastructure. The scenario of BIM-CV focuses on long-lasting infrastructures that can be elasticaly increased or decreased.

Elasticity is a requirement for budget contention (either real or virtual credits) in any cloud infrastructure. Elasticity should be automatic and depending on the workload. The

use of pipelines facilitates the implementation of IaaS elasticity rules based on the length of the processing queue. Results in *Caballer et. al (2013)* show very good performance, especially in coordination with workflow engines.

Finally, the deployment of individual virtual infrastructures is a better approach than sharing a physical infrastructure even in the case that computing and software requirements of the subprojects concurr. Multitenancy at the level of the infrastructure enables isolating users of different subprojects even more, and reduces the risks of crashes, infrastructure overloading and maintenance failures with respect to the traditional mechanisms based on quotas.

Therefore, the model of virtual elastic infrastructure fits well the requirements both of the users and the infrastructure providers.

EuroBioImaging is currently in the intermim phase, and planning to start prototyping services during 2015. The BIMCV has identified preliminary sub-projects to be used as prototypes for the infrastructure proposed in the article. Currently, the platform is implemented at the level of the provisioning services. Current work is being focused on the access service to data. In this piloting phase, BIMCV will run standard medical-imaging benchmarks and populate the software repository.

In a second stage, BIMCV is planning to upgrade the software architecture with the support of vertical elasticity and the use of containers. Vertical elasticity will enable running memory-demanding applications in a workflow without requiring the pre-reservation of large-scale instances.

References

Myneni, S. and Patel, V.L. (2010) 'Organization of biomedical data for collaborative scientific research: a research information management system', *International Journal of Information Management*, Vol. 30, No. 3, pp.256–264

Caballer, M. , de Alfonso, C., Moltó, G., Romero, E., Blanquer, I. and García, A. (2014) 'CodeCloud: A platform to enable execution of programming models on the Clouds', *Journal of Systems and Software*, Vol 93, pp. 187-198.

de Alfonso, C., Caballer, M., Alvarruiz, F. and Moltó, G. (2013) 'An Economic and Energy-Aware Analysis of the Viability of Outsourcing Cluster Computing to the Cloud', *Futur. Gener. Comput. Syst. (International J. Grid Comput. eScience)*, vol. 29, pp. 704–712.

M. Caballer, C. de Alfonso, F. Alvarruiz, and G. Moltó, 'EC3: Elastic Cloud Computing Cluster,' *J. Comput. Syst. Sci.*, vol. 79, pp. 1341–1351.

Carrión, J. V., Moltó, G., De Alfonso, C., Caballer, M., and Hernandez, V. (2010). 'A generic catalog and repository service for virtual machine images'. *2nd International ICST Conference on Cloud Computing (CloudComp 2010)*.

Hochstein, L. 'Ansible: Up and Running, Automating Configuration Management and Deployment the Easy Way', (2014) ISBN: 978-1-4919-1532-5, *Ed. O'Reilly Media*.

Calatrava, A., Moltó, g., Caballer, M. and De Alfonso, C. 'Virtual Hybrid Elastic Clusters in the Cloud' (2014). *8th Iberian Grid Infrastructure Conference (IberGrid 2014)*.

ESR Working Group on Imaging Biobanks. (2014). Mission Statemens ESR Working Group on Imaging Biobanks. http://www.myesr.org/cms/website.php?id=/en/research/esr_wg_on_imaging_biobanks/esr.htm. Last visit: 02/01/2015.

Caballer, M., Blanquer, I., Moltó, G., and de Alfonso, C. 'Dynamic management of virtual infrastructures' (2014). *Journal of Grid Computing*, doi: 10.1007/s10723-014-9296-5.

de Alfonso, C., Caballer, M., Alvarruiz, F., Moltó, G., and Hernández, V., 'Infrastructure Deployment over the Cloud' (2011). *3rd IEEE International Conference on Cloud Computing Technology and Science (CloudCom 2011)*, pp. 517 - 521, 2011, 10.1109/CloudCom.2011.77.

de Alfonso, C., Caballer, M., Alvarruiz, F.,and Hernández, V., 'An Energy Management System for Cluster Infrastructures', (2013) *Computers and Electrical Engineering*, Volume 39, Issue 8, pp. 2579-2590, ISSN 0045-7906, 10.1016/j.compeleceng.2013.05.004.

Greely, H. T. (2007). 'The Uneasy Ethical and Legal Underpinnings of Large-Scale Genomic Biobanks'. Annual Review of Genomics and Human Genetics 8: 343-364.

Hewitt, R. E. (2011) 'Biobanking: The foundation of personalized medicine'. Current Opinion in Oncology 23 (1): 112-119.

International Society for Biological and Environmental Repositories (ISBER) Newsletter Volume No.1, Issue No. 1, 2001.

Marcus, D.S., Olsen T., Ramaratnam M., and Buckner, R.L. (2007). The Extensible Neuroimaging Archive Toolkit (XNAT): An informatics platform for managing, exploring, and sharing neuroimaging data. Neuroinformatics 5(1): 1134.

Freund, J., Comaniciu, D., Ioannis, Y., Liu, P., McClatchey, R., Morley-Fletcher, E., Pennec, X., Pongiglione, G., Xiang, Zhou Health-e-child: an integrated biomedical platform for Grid-based paediatric applications. Stud. Health Technol. Inform. 120, 259–270 (2006)

Grethe, J.S., Baru, C., Gupta, A., James, M., Ludaescher, B., Martone, M.E., Papadopoulos, P.M., Peltier, S.T., Rajasekar, A., Santini, S., Zaslavsky, I.N., Ellisman, M.H. Biomedical informatics research network: building a national collaboratory to hasten the derivation of new understanding and treatment of disease. Stud. Health Technol. Inform. 112, 100–109 (2005)

Redolfi, A., McClatchey, R., Anjum, A., Zijdenbos, A., Manset, D., Barkhof, F., Spenger, C., Legré, Y., Wahlund, L.O., Pietro, C.B.d.S., B Frisoni, G. Grid infrastructures for computational neuroscience: the neuGRID example. Future Neurol. 4(6), 703–722 (2009). doi:10.2217/fnl.09.53

Silva, L., Costa, C., Oliveira, J. A PACS archive architecture supported on cloud services. Int. J. Comput. Assist. Radiol. Surg. 7(3), 349–358 (2012). doi:10.1007/s11548-011-0625-x

Teng, C.C., Mitchell, J., Walker, C., Swan, A., Davila, C., Howard, D., Needham, T. A medical image archive solution in the cloud. In: 2010 IEEE International Conference on Software Engineering and Service Sciences (ICSESS), pp. 431–434 (2010). doi:10.1109/ICSESS.2010.5552343