*Highlights of Spanish Astrophysics VIII, Proceedings of the XI Scientific Meeting of the Spanish Astronomical Society held on September 8 – 12, 2014, in Teruel, Spain. A. J. Cenarro, F. Figueras, C. Hernández-Monteagudo, J. Trujillo, and L. Valdivielso (eds.)*

# Calibration of radio-astronomical data on the cloud. LOFAR, the pathway to SKA.

**J. Sabater[1,2], S. Sánchez-Expósito[2], J. Garrido[2], J. E. Ruiz[2], P. N. Best[1], and L. Verdes-Montenegro[2]**

[1] Institute for Astronomy (IfA), University of Edinburgh, Royal Observatory, Blackford Hill, EH9 3HJ Edinburgh, U.K.
[2] Instituto de Astrofísica de Andalucía, CSIC, Apdo. 3004, 18080 Granada, Spain

## Abstract

The radio interferometer LOFAR (LOw Frequency ARray) is fully operational now. This Square Kilometre Array (SKA) pathfinder allows the observation of the sky at frequencies between 10 and 240 MHz, a relatively unexplored region of the spectrum. LOFAR is a software defined telescope: the data is mainly processed using specialized software running in common computing facilities. That means that the capabilities of the telescope are virtually defined by software and mainly limited by the available computing power. However, the quantity of data produced can quickly reach huge volumes (several Petabytes per day). After the correlation and pre-processing of the data in a dedicated cluster, the final dataset is handled to the user (typically several Terabytes). The calibration of these data requires a powerful computing facility in which the specific state of the art software under heavy continuous development can be easily installed and updated. That makes this case a perfect candidate for a cloud infrastructure which adds the advantages of an on demand, flexible solution. We present our approach to the calibration of LOFAR data using Ibercloud, the cloud infrastructure provided by Ibergrid. With the calibration work-flow adapted to the cloud, we can explore calibration strategies for the SKA and show how private or commercial cloud infrastructures (Ibercloud, Amazon EC2, Google Compute Engine, etc.) can help to solve the problems with big datasets that will be prevalent in the future of astronomy.

## 1 Introduction

With the advent of new generation astronomical facilities that can deliver data volumes of several PB, a new computing model must be developed. The calibration of the Low Frequency Array (LOFAR) data requires powerful computing and storage resources and can be considered as an example of data-driven astronomical computing [1]. Using this early example we can test the use of distributed computing infrastructures like the grid or the cloud to to deal with large amounts of data.

We present a preliminary study of the use of cloud technologies for the calibration of radio-interferometric LOFAR data from the point of view of the final user.

## 2 LOFAR

LOFAR is a new generation phased-array interferometer that operates at frequencies between 10 and 240 MHz [5]. Its core is located in the north of the Netherlands where 38 receiver stations are located and it also contains several additional stations spread across Europe. It is considered to be one of the pathfinders of the Square Kilometre Array (SKA) [2]. LOFAR has no moving parts with observations being entirely controlled in software. It is a software defined telescope: the data are correlated in a Graphics Processing Unit cluster and then processed and analysed using specialized software running in common computing facilities. That means that the capabilities of the telescope are virtually defined by software and mainly limited by the available computing power.

One of the science cases of LOFAR consists of deep and wide surveys of the sky, with a very wide range of science goals ranging from detailed studies of nearby galaxies and active galactic nuclei through to searches for the most distant radio sources [4]. In order to do that LOFAR uses its interferometric imaging capabilities. The field of view of a single observation beam at the higher frequencies reach a diameter of 6 to 7 degrees. We will focus our study in this type of data.

Following correlation and pre-processing in the computing facilities located in Groningen, the imaging data are stored in a Long Term Archive based on grid technologies. At this point the final user can download and calibrate the data, but this calibration process still presents some challenges:

- The size of the data can be considerable. A single observation of 10 hours can amount up to 20 TB if the data are not pre-averaged. Even when the data are averaged to 1 channel per sub-band and a time resolution of 10 s, the size of the each dataset is about 3 TB.

- The calibration of the data requires intensive processing. A single calibration round of an observation of 10 hours requires (with current software) at least 2 CPU-years of computing. The calibration is only possible if some kind of parallelization technique is used.

- The software required for the calibration is very specific and is under continuous development. The installation of the software is not trivial for some operating systems and versions, although this is improving quickly.

We explored the use of a cloud infrastructure to deal with these remaining challenges.

## 3 LOFAR on Ibercloud

Ibercloud is the cloud infrastructure provided by Ibergrid, the Portuguese-Spanish grid initiative. The infrastructure is based on OpenStack. We tested the implementation of a data calibration pipeline on this infrastructure.

First, we created the disk images where the operating system (OS) and the LOFAR software are installed. They were based on OS images available on the image repository of Ibergrid that were provided by other users. This step is isolated from the rest of the work-flow and allow us to focus only once on the installation of the software. The possibility to choose the underlying OS facilitates the process. We can choose the same OS used by the LOFAR software developers as reference. This avoids compatibility issues between the OS and the installed software. It is also very useful to be able to isolate the installation and software support process from the rest of the scientific work-flow.

We deployed a testing virtual cluster composed of one head node and two worker nodes. The memory capacity and the number of processors of the worker nodes can be adapted to the requirements of the calibration pipeline. For example, if the software of one of the calibration steps requires more memory the working nodes can be created with more memory. This allows the user to adapt the computing resources available to the pipeline, which can be very useful for testing purposes.

At the moment, the test data are transferred using Internet from the archive to the processing facility. The time to transfer the data is relatively small in comparison with the time needed to run the calibration (less than 1/3 of the time). However, in other facilities (and in the future) the transfer of data could suppose a problem that should be considered.

We used IPython Parallel [3] to orchestrate the calibration tasks between the different nodes. The preliminary tests run successfully. We could not find any strong performance penalization in comparison with typical clusters. However, more testing and profiling are needed to draw any conclusion about performance.

Finally, we would like to note that the Ibercloud infrastructure is being integrated into the European GRID Infrastructure (EGI) Federated Cloud. This will provide a pan-European homogeneous way to access the cloud resources that are integrated into the infrastructure. The use of commercial clouds like Amazon Web Services, Google Compute Engine, etc. could be possible but has not been tested yet.

## 4 Summary and conclusions

LOFAR, one of the SKA pathfinders, is completely operational and providing high quality data. From the point of view of the user there are some challenges remaining for the data calibration, in particular the high computational requirements. We explored the use of a cloud infrastructure (Ibercloud) to deal with the problems we found.

We found the cloud very flexible to adapt the computing, storage and networking resources to our needs. This allows the easy exploration of different calibration strategies that require different hardware architectures. The ability to isolate the installation of software from the rest of the calibration work-flow has proven to be very useful in our case.

The cloud solution provided enough power to deal efficiently with the data. It was possible to process in parallel the data like in a traditional cluster. Additionally, the cloud allows an on-demand consumption of resources that can be shared by different users, including users working on different areas.

The possible problems with the storage and data transfer were not considered in this study. They must be taken into account when the size of the data makes undesirable or impossible their transport. Probably a new model in which processing is co-located with existing data should be considered.

A cloud infrastructure could provide a powerful high throughput computing (HTC) resource that can deal with the big amount of data produced by new and future astronomical facilities.

## Acknowledgments

The authors would like to acknowledge the support given by the Ibercloud team as well as the use of this infrastructure. JS would like to acknowledge the travel grant of the Society of Spanish Researchers in the U.K. and the Office for Cultural and Scientific Affairs of the Embassy of Spain in London. JS and PNB are grateful for financial support from STFC. Part of this work was funded by the grant AYA2011-30491-C02-01, co-financed by MICINN and FEDER funds, and the Junta de Andalucía (Spain) Grants P08-FQM-4205 and TIC-114. LOFAR, the Low Frequency Array designed and constructed by ASTRON, has facilities in several countries, that are owned by various parties (each with their own funding sources), and that are collectively operated by the International LOFAR Telescope (ILT) foundation under a joint scientific policy.

## References

[1] Berriman, G.B. et al. 2012, Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 371 doi:10.1098/rsta.2012.0066.

[2] Norris, R. P. et al. 2013, PASA, 30, 20 doi:10.1017/pas.2012.020, arXiv:1210.7521.

[3] Pérez F., Granger B. E., 2007, Computing in Science and Engineering, 9, 21 URL:http://ipython.org, doi:10.1109/MCSE.2007.53.

[4] Röttgering, H. et al. 2011, Journal of Astrophysics and Astronomy, 32, 557. doi:10.1007/s12036-011-9129-x , arXiv:1107.1606

[5] van Haarlem, M.P. et al. 2013, A&A, 556, A2. doi:10.1051/0004-6361/201220873, arXiv:1305.3550.