Contribution ID: **27** Type: **not specified**

Towards a cloud-based computing and analysis framework to process environmental science big data

Tuesday, 7 March 2017 14:00 (20 minutes)

Environmental sciences are quickly and increasingly adopting research methodologies based on data coming from satellites, large network of sensors installed on the ground or sea-floating stations, as well as from devices installed on balloons or aircraft. These networks produce a big amount of data that needs to be appropriately processed and analyzed to extract information useful for scientists to investigate natural phenomenas.
This require for example, the capacity to collect and store huge amount of data together with space and time information and the availability of large and powerful computing resources to run analysis and visualization codes. However, for environmental scientific community to develop or to have in house the infrastructures to accomplish and support efficiently all these steps can not be convenient, because they require a non negligible effort for maintenance.

On the other hand, in the last decades, scientific communities of other scientific fields, such as high-energy physics, have developed and acquired a large experience in computer grid infrastructures to store and analyze big amount of data coming from particle accelerator laboratories experiments. Computer grids have recently evolved into clouds which are more user-friendly allowing access to new and more heterogeneous communities with less computing expertise [1,2].

In this contribution we discuss how to bridge the experience acquired in using the grids for high energy physics experiments and the needs of environmental sciences. In particular, we apply this strategies in the context of the interdisciplinary EU-ERASMUS+ TORUS project which includes Europe's and south east Asia's partners with a strong expertise in distributed and cloud computing and earth and environmental sciences.
The TORUS project aims to make soon available to environmental scientists a cloud based computing and analysis framework to manage and process big-data.

This includes the ability to access clouds to virtualize the computing resources, and knowledges to use software tools to process and analyze data coming from the different data sources. We also describe how to store data together with meta-data information related to time and space, and how to present data at high-level that can be easily used and interpreted by user scientists.

Finally, we also discuss how to integrate into this framework high-performance computing to boost for example satellite image processing, that for the intrinsic computational complexity may require the use of recently developed accelerators like GP-GPUs or many-core processors [3,4].

References

[1] Fella, A., Luppi, E., Manzali, M., Tomassetti, L., A general purpose suite for Grid resources exploitation (2012) IEEE Nuclear Science Symposium Conference Record, art. no. 6154459, pp. 99-103.

[2] Roiser, S., et al., The LHCb Distributed computing model and operations during LHC runs 1, 2 and 3 (2015) Proceedings of Science, art. no. 005.

[3] Calore, E., et al., Massively parallel lattice Boltzmann codes on large GPU clusters, Parallel Computing 58 (2016), pp. 1–24.

[4] Adinetz, A. V., et al., Performance evaluation of scientific applications on POWER8, Lecture Notes in Computer Science 8966 (2015), pp. 24–45.

Primary authors: Prof. LUPPI, Eleonora (University of Ferrara and INFN); Dr TOMASSETTI, Luca (University of Ferrara and INFN); Dr SCHIFANO, Sebastiano Fabio (University of Ferrara and INFN)

Presenters: Prof. LUPPI, Eleonora (University of Ferrara and INFN); Dr TOMASSETTI, Luca (University of Ferrara and INFN); Dr SCHIFANO, Sebastiano Fabio (University of Ferrara and INFN)

Session Classification: Data Management & Big Data

Track Classification: Data Management & Big Data