Contribution ID: **29** Type: **not specified**

A versatile environment for large-scale geospatial data processing with HTCondor

Thursday, 6 September 2018 16:30 (25 minutes)

Geospatial data are one of the core data sources for scientific and technical support to the European Commission (EC) policies. For instance, the Copernicus programme of the European Union provides a vast amount of Earth Observation (EO) data for monitoring the environment through the Sentinel satellites operated by the European Space Agency. In terms of data management and processing, big geospatial data streams and other data sources have motivated the development of a petabyte-scale computational platform at the EC Joint Research Centre (JRC). This platform is called the JRC Earth Observation Data and Processing Platform (JEODPP) [1]. Thematic applications at the JRC rely on a variety of data sources each with their own data formats and protocols. In addition, experts from different domains build on different software, tools and libraries, making difficult knowledge sharing and the reproducibility of the experimental results. Taking into consideration all these challenges, the JEODPP has been designed by following the principles of modularity, parallelization and virtualization/containerization. In this way, it provides a flexible working environment where the users are able to deploy and optimize software and algorithmic workflows specialized for their tasks while fostering knowledge and data sharing.

Although there is no constraint on the type of data that can be processed, the main focus of the platform is currently on geospatial analysis and on the processing of satellite images. The Sentinel satellites are following a series of fixed orbits with image data delivered on a continuous basis and with a revisit time depending on the Sentinel mission type. The image data are stored in the form of flat files with each file mapping a given portion of the Earth surface. This drove both the architectural decisions and the physical/logical implementations regarding the JEODPP set up. In particular, the platform supports batch processing via mainly high-throughput computing where large collections of files are processed in parallel. Besides the batch farm, JEODPP offers other services such as interactive data analysis and visualization, data sharing, data storage, remote desktop access and experimental results dissemination. The operation of all these services is based on Docker containerisation.

HTCondor was chosen as workload manager, a versatile and robust job scheduler. Taking advantage of the Docker universe that HTCondor inherently supports, massive batch processing runs successfully on JEODPP since 2016. Besides, HTCondor functionalities allow a flexible combination of both types of nodes, workers, and managers. For example, it is possible for the user to submit jobs from different nodes, containers, or IPython notebooks using varying methods for authentication. Since it requires no external services for storage, HTCondor can use both the local and the network file system such as the EOS open source storage solution developed by CERN and deployed on the JEODPP. In practice, HTCondor shares features of a resource manager combined with those of a job scheduler. By integrating these features into a single system, it allows complex policy configurations and sophisticated optimizations. In this presentation, we show two applications that fully rely on HTCondor as workload manager and provide suggestions and lessons learnt based on our experience.
- Mosaicking Copernicus Sentinel-1 Data at Global level [2,3]: An algorithmic workflow for producing mosaics based on the dual polarisation capability of Sentinel-1 SAR imagery;
- Optimizing Sentinel-2 image selection in a Big Data Context [4]: An optimization scheme that selects a subset of the Sentinel-2 archive in order to reduce the amount of processing, while retaining the quality of the resulting output. As a case study, the focus is on the creation of a cloud-free composite, covering the global land mass and based on all the images acquired from January 2016 until September 2017.
- Marine ecosystem modelling in the SEACOAST project comprises types of modelling codes that are relevant

to Marine Framework Strategy Directive [5], implemented on different spatial and temporal scales, complemented by essential data (bathymetry, initial, boundary forcing, in and output) that are inherently coupled to each other. These models are implemented as an MPI application based on FORTRAN and it is running by using the parallel universe of HTCondor. We add a network file system NetApp beside EOS, which improves the performance of the MPI jobs over 80%.

In the near future, the possibility to combine HTCondor with Apache Mesos will be investigated. The aim is to provide a flexible, reconfigurable and extendable infrastructure to cover a wide range of different scientific computing use cases like HTC, HPC, Big Data analytics, GPU acceleration and Cloud technologies.

References

[1] P. Soille, A. Burger, D. De Marchi, D. Rodriguez, V. Syrris, and V. Vasilev.; *A versatile data-intensive computing platform for information retrieval from big geospatial data*; Future Generation of Computer System, pages 30-40, 2018. Available from: https://doi.org/10.1016/j.future.2017.11.007

[2] V. Syrris, C. Corbane, and P. Soille; *A global mosaic from Copernicus Sentinel-1 data* in Proc. Big Data Space, 2017, pp. 267–270. Available from: http://dx.doi.org/10.2760/383579

[3] V. Syrris, C. Corbane, M. Pesaresi, and P. Soille; *A global mosaic from Copernicus Sentinel-1 data* IEEE Tr. on Big Data. Available from: http://dx.doi.org/10.1109/TBDATA.2018.2846265

[4] P. Kempeneers and P. Soille.; *Optimizing Sentinel-2 image selection in a Big Data context*; Big Earth Data, pages 145-148, 2017. Available from: https://doi.org/10.1080/20964471.2017.1407489

[5] D. Macias and E. Garcia-Gorriz and A. Stips.; *Productivity changes in the Mediterranean Sea for the twenty-first century in response to changes in the regional atmospheric forcing* Frontiers in Marine Science, pages 70, 2015. Available from: https://doi.org/10.3389/fmars.2015.00079

Primary authors: Dr RODRIGUEZ ASERETTO, Dario (European Commission); Dr SOILLE, Pierre (European Commission)

Presenter: Dr RODRIGUEZ ASERETTO, Dario (European Commission)

Session Classification: Workshop presentations

Track Classification: HTCondor presentations and tutorials