Contribution ID: **8**   Type: **not specified**

Flexible and Cost-effective Petabyte-Scale Architecture with HTCondor Processing and EOS Storage Backend for Earth Observation Applications

Wednesday 7 June 2017 10:15 (25 minutes)

Flexible and Cost-effective Petabyte-Scale Architecture with HTCondor Processing and EOS Storage Backend for Earth Observation Applications

Veselin Vasilev, Dario Rodriguez, Armin Burger and Pierre Soille

European Commission, Joint Research Centre (JRC)
Directorate I. Competences. Unit I.3 Text and Data Mining
Via E. Fermi 2749, I-21027 Ispra (Va), Italy

The Copernicus programme[1] of the European Union is delivering massive amounts of satellite image data of interest to a range of European policies supported by activities of the JRC. In this context, the JRC faces the challenges of storage and processing of Earth Observation data at Petabyte-scale. This led to the design and development of the JRC Earth Observation Data and Processing Platform (JEODPP) [1]. In order to address the needs for high data throughput and scalability, multi-purpose usage, as well as budgetary and data accessibility constraints, an implementation based on commodity hardware and open source solutions was developed. The infrastructure consists of a processing cluster, built upon a scalable set of processing nodes, and a storage cluster which sits on top of a Just a Bunch Of Disks (JBODs) attached to dedicated storage nodes. HTCondor was chosen as a processing workload manager for its maturity to work in Docker universe. As storage layer, CERN's in-house developed storage solution EOS[2] was chosen. It fits the requirements for scientific data processing in a cluster environment, scales well and integrates into an existing Kerberos realm for data access management. The current set-up, of gross 1.8 PB, has the processing nodes (37 nodes for a total of 952 CPUs) mounting the storage using EOS's own FUSE client as wrapper around XRootD software framework. This allows a unified POSIX-like data access that grants clients to run applications in a performant cluster environment without any special modification. The usage of Docker universe for HTCondor allows the flexible handling of very diverse processing environments developed over time by various JRC projects. Using HTCondor as a job scheduler for interactive web processing requests is among the future challenges for the setup.

[1] P. Soille, A. Burger, D. Rodriguez, V. Syrris, and V. Vasilev. Towards a JRC Earth observation data and processing platform. In P. Soille and P.G. Marchetti, editors, Proc. of the 2016 Conference on Big Data from Space (BiDS'16), pages 65{68. Publications O_ce of the European Union, 2016. doi: http://dx.doi.org/10.2788/854791.

————————-

1. http://copernicus.eu/.
2. https://eos.web.cern.ch/.

Authors: VASILEV, Veselin (European Commission, Joint Research Centre (JRC)); RODRIGUEZ, Dario (European Commission - DG JRC); SOILLE, Pierre (European Commission); BURGER, Armin (European Commission - Joint Research Centre)

Presenter: VASILEV, Veselin (European Commission, Joint Research Centre (JRC))

Session Classification: Workshop

Track Classification: HTCondor presentations and tutorials