Contribution ID: **14** Type: **Oral Presentation**

Smart Policy Driven Data Management and Data Federations, enabled by the H2020 eXtreme DataCloud project.

Tuesday, 20 March 2018 14:30 (30 minutes)

In November 2017, the H2020 "eXtreme DataCloud" project will be launched, developing scalable technologies for federating storage resources and managing data in highly distributed computing environments. The project will last for 27 months and combines the expertise of 8 large European organizations. The targeted platforms are the current and next generation e-Infrastructures deployed in Europe, such as the European Open Science Cloud (EOSC), the European Grid Infrastructure (EGI), the Worldwide LHC Computing Grid (WLCG) and the computing infrastructures that will be funded by H2020 EINFRA-12 calls.
One of the core activities within XDC is the policy-driven orchestration of federated and heterogeneous data management. The high-level objective of this work is the semi or fully automated placement of scientific data in the Exabyte region on the site (IaaS), as well as on the federated storage level. In the context of this Work Package, placement may either refer to the media the data is stored on, to guarantee a requested Quality of Service, or the geographical location, to move data as close to the compute facility as possible to overcome latency issues in geographically distributed infrastructures. The solutions will be based on already well established data management components as there are dCache, EOS, FTS and the INDIGO PaaS Orchestrator, Onedata and many more. The targeted scientific communities, represented within the project itself, are from a variety of domains, like astronomy (CTA), Photon Science (European X-FEL), High Energy Physics (LHC), Life Science (LifeWatch) and others.
The work will cover "Data Lifecycle Management" and smart data placement on meta-data, including storage availability, network bandwidth and data access patterns. The inevitable problem of network latency is planned to be tackled by smart caching mechanisms or, if time allows, using deep learning algorithms to avoid data transfers in the first place. Furthermore, data ingestion and data movement events, reported to the centralized INDIGO PaaS orchestration engine can trigger automated compute processes, starting from meta-data extraction tools to sophisticated workflows, e.g. to pre-analyze images from Photon Science or Astronomy detectors.
This presentation will present the overall architecture of this activity inside the eXtreme DataCloud project and will elaborate on the work plan and expected outcome for the involved communities.

Summary

The core activity within the newly created H2020 project "eXtreme DataCloud" will be the policy-driven orchestration of federated data management for data intensive sciences like High Energy Physics, Astronomy, Photon and Life Science. Well known experts in this field will work on combining already established data management and orchestration tools to provide a highly scalable solution supporting the computing models the entire European Scientific Landscape. The work will cover "Data Lifecycle Management" as well as smart data placement based on domain specific and technical meta-data, including storage availability, network bandwidth and data access patterns. Mechanisms will be but in place to trigger computational resources based on data ingestion and data movements. This presentation will present the first architecture of this endeavor.

Primary authors: Dr DONVITO, Giacinto (INFN/Bari); Dr DUTKA, Lukasz (CYFRONET); Dr KEEBLE, Oliver (CERN); Dr FUHRMANN, Patrick (DESY/dCache.org); Dr CESINI, daniele (CNAF/INFN)

Presenter: Dr FUHRMANN, Patrick (DESY/dCache.org)

Session Classification: Data Management & Big Data Session

Track Classification: Big Data & Data Management