Contribution ID: **154** Type: **Presentation**

Using Onedata for data caching in hybrid-cloud environments

Tuesday, 9 October 2018 11:45 (15 minutes)

Onedata [1] is a global high-performance data management system, that provides easy and unified access to globally distributed storage resources and supports a wide range of use cases from personal data management to data-intensive scientific computations. Onedata enables the creation of complex hybrid-cloud deployments, using private and commercial cloud resources. It allows users to share, collaborate and publish data as well as perform high-performance computations on distributed data. Onedata system consists of zones (Onezone) which enable the creation of federations of data centres and users, storage providers (Oneprovider) which expose storage resources, and clients (Oneclient), who can access their data via a virtual POSIX file system. Onedata introduces the concept of space, a virtual volume, owned by one or more users, where the data is stored. Each space can be supported by a dedicated amount of storage supplied by one or multiple storage providers. Storage providers deploy Oneprovider instance near the storage resources, register it in selected Onezone service to become part of a federation and expose those resources to users. By supporting multiple types of storage backends, such as POSIX, S3, Ceph and OpenStack Swift, and GlusterFS.

In large-scale hybrid cloud deployments, it is often the case that data maintained in the private cloud has to be processed on-demand in the public cloud. While deploying remote jobs is today fairly straightforward, and can be automated using several orchestration platforms, making the data available for processing in the remote cloud is a significant challenge. Onedata makes this easy, by enabling automatic, on-demand, block-based data prefetching based on the POSIX requests from user applications and automatically caching the files based on analysis of file popularity. In most cases, prestaging is not necessary at all, as the data blocks are fetched on the fly when requested for reading, however it provides REST API for controlling data replication manually or integrating it with 3rd party services.

Currently, Onedata is used in Helix Nebula Science Cloud [2], eXtreme DataCloud [3], PLGrid [4], European Open Science Cloud Hub, and European Open Science Cloud Pilot [6], where it provides data transparency layer for computation deployed on hybrid-clouds. In EOSC-hub [5] it serves as the basis of EGI Open Data Platform, supporting open science use cases such as open data curation (metadata editing), publishing (DOI registration) and discovery (OAI-PMH protocol).

1. Onedata project website. http://onedata.org.
2. Helix Nebula Science Cloud (Europe's Leading Public-Private Partnership for Cloud). http://www.helix-nebula.eu.
3. eXtreme DataCloud (Developing scalable technologies for federating storage resources). http://www.extreme-datacloud.eu.
4. PL-Grid (Polish Infrastructure for Supporting Computational Science in the European Research Space). http://projekt.plgrid.pl/en.
5. European Open Science Cloud Hub (Bringing together service providers to create a contact point for European researchers and innovators.). https://www.eosc-hub.eu.
6. European Open Science Cloud Pilot (Development of the EOSC-hub). https://eoscpilot.eu.

Type of abstract

Presentation

Summary

A brief description of Onedata solution and demonstration of caching and file popularity mechanisms for the hybrid cloud use cases.

Primary authors: KRYZA, Bartosz (CYFRONET); Dr DUTKA, Lukasz (CYFRONET); ORZECHOWSKI, Michal (CYFRONET)

Presenters: Dr DUTKA, Lukasz (CYFRONET); ORZECHOWSKI, Michal (CYFRONET)

Session Classification: Thematic Services

Track Classification: Area 5. Digital Infrastructures for EOSC and/or EDI