Contribution ID: **33**

Type: **Presentation**

Onedata Virtual Filesystem for Hybrid Clouds

Tuesday, 30 January 2018 12:10 (20 minutes)

Onedata is a complete high-performance storage solution that unifies data access across globally distributed environments and multiple types of underlying storages, such as NFS, Lustre, GPFS, Amazon S3, CEPH, as well as other POSIX-compliant file systems. It allows users to share, collaborate and perform computations on their data.

Globally Onedata comprises of: Onezones, distributed metadata management and authorisation components that provide entry points for users to access Onedata; and Oneproviders, that expose storage systems to Onedata and provide actual storage to the users. Oneprovider instances can be deployed, as a single node or a HPC cluster, on top of high-performance parallel storage solutions with ability to serve petabytes of data with GB/s throughput.

Onedata introduces the concept of Space, a virtual directory, owned by one or more users. The Spaces are accessible to users via an intuitive web interface, which allows for Dropbox-like file management and file sharing, Fuse-based client that can be mounted as a virtual POSIX file system, or REST and CDMI standardized APIs. Onedata does not provide users with any physical storage, each Space has to be supported with a dedicated amount of storage by one or more providers, who are running Oneprovider component.

Fine-grained management of access rights, including POSIX-like access permissions and access control lists (ACLs), allow users to share entire Spaces, directories or files with individual users or user groups. Onedata user groups are particularly suitable for managing communities of users that wish to share common resources. Access to Spaces is managed by flexible authentication and authorisation methods such as: access tokens, OpenID, X.509 certificates and Macaroons.

Furthermore, Onedata features local storage awareness that allows users to perform computations on the data located virtually in their Space. When data is available locally, it is accessed directly from a physical storage system where the data is located. If the needed data is not available locally, the remote data is fetched in real-time from remote facility, using a dedicated highly parallelized dedicated protocol with block-level data transfer, that also provides common features such as pre-staging, data migration and data replication.

Currently Onedata is used in Indigo-DataCloud and PLGrid projects as a federated data access solution, aggregating either computing centres or whole national computing infrastructures; and in EGI-Engage, where it is a basis for Open Data Platform prototype for dissemination and exploitation of open data sets.

Open Data Platform prototype is capable of publishing Spaces as data containers with assigned globally unique identifier, such as DOI (Digital Object Identifier) or PID (Persistent Identifier), to selected communities or public open data portals. It features metadata management editor with ability to add custom metadata-schemes to open data containers, files, and folders. Detailed data management plan, can be defined for individual containers characterising the overall data lifetime: from generation, generated data formats, curation, licensing and long term preservation policies.

Primary author: DUTKA, Lukasz

Co-authors: KRYZA, Bartosz (ACC Cyfronet-AGH); ORZECHOWSKI, Michał (AGH University of Science and Technology, Academic Computer Centre Cyfronet AGH, Krakow, Poland); Mr OPIOLA, Lukasz (ACK Cyfronet AGH); Mr WRZESZCZ, Michal (ACK Cyfronet AGH)

Presenter: DUTKA, Lukasz

Session Classification: Scalable Storage Backends for Cloud and HPC: Foundations