Contribution ID: **21** Type: **Oral presentation**

Developing Real-time Services for Large Volume Experiment-Data Analysis utilizing Supercomputing and Cloud technologies at CSCS (SELVEDAS)

Wednesday, 28 October 2020 16:50 (20 minutes)

Introduction

The ongoing developments in accelerators, detectors and experiment automation is leading to a rapid growth of data generated during experiments. A viable solution is utilizing suitable infrastructures that allow additional remote high performance capacity for processing and analysis of data from the experimental facilities with larger data volumes and higher processing needs. The SELVEDAS project proposes a hybrid cloud infrastructure, offering scalable and extensible services for data management and analysis to Swiss academic users by leveraging high performance computing (HPC), storage, networking as well as cloud technologies and orchestration. The on-demand services perform as a highly efficient remote data processing system providing fast feedback and analysis with the long time storage and archival of petabytes of data.

Approach

Hybrid cloud

The SELVEDAS project develops a hybrid cloud extension to the PSI infrastructure by giving access to the supercomputing resources at CSCS for experiment data analysis. The solutions rely on the easy transferability of workloads between systems comprising of HPC and cloud technologies. The architecture is data-driven (figure 1). The proposed services at CSCS are accessible through a RESTful API, FirecREST.

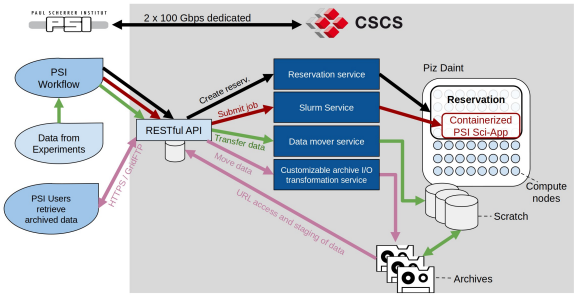

Figure 1. Hybrid cloud infrastructure with data driven workflow framework

Figure 1:

On demand service

Online analysis at PSI refers to processing of data while the scientist is using an instrument. Hence an on-demand service for advance resource reservation is implemented to realize the requirement of availability and to provide a tight feedback loop for the experiments 1. To end users, HPC resources can then be of no differences from the facility's on-site IT infrastructure resources.

Data catalog extension

Since 04/2018, CSCS and PSI jointly operate the PetaByte Archive located at CSCS. The PetaByte Archive pro-

vides user services for long term data storage and retrieval of experimental data from PSI large scale facilities. Archiving and retrieval of data is facilitated by the Data Catalogue (SciCat). The SELVEDAS project extends SciCat by integrating it with the FirecREST API and provides services for the analysis of experiment data stored in the PetaByte Archive at CSCS (figure 2).

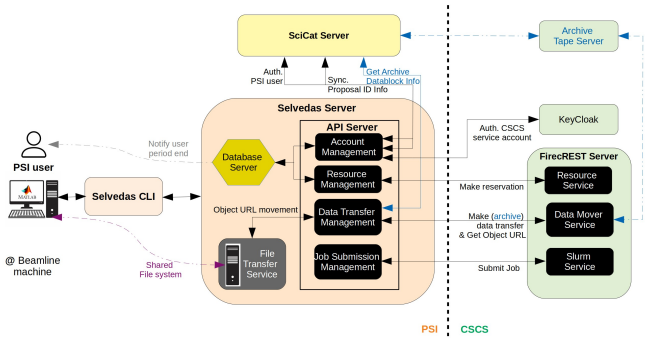

Figure 2. Overall data catalog architecture

Figure 2:

Cross-site authentication

Authentication and authorization are based on the PSI user account and access rights. PSI client uses service accounts from CSCS for accessing FirecREST API (figure 3). This separation of responsibilities frees PSI users from needing CSCS's personal accounts, improves user scalability, and allows decentralized authentication since it's done directly by PSI clients to an OIDC server (Keycloak) in CSCS.

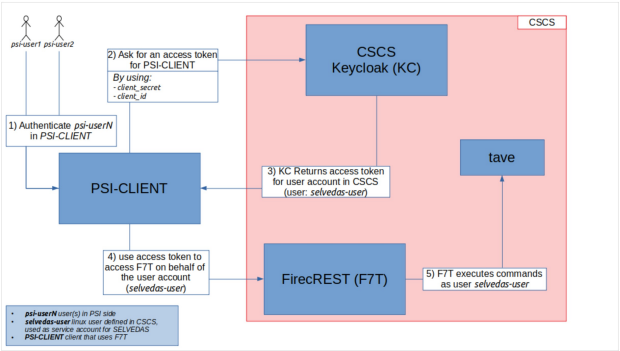

Figure 3. Cross-site authentication diagram

Figure 3:

Performance Result

The performance evaluation is tested on the typical experiment dataset. Figure 4&5 reports results of workflows and the job submission, one compared between two workflows with the large and small dataset, and another one compared between different GPUs for the job submission.

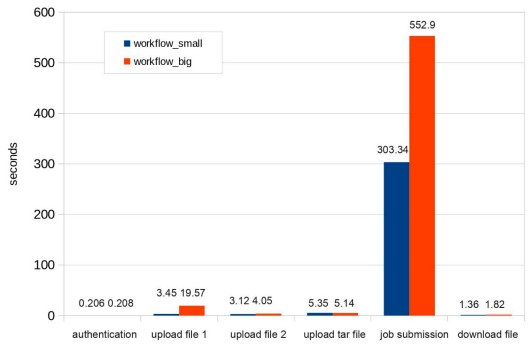

Figure 4. Two types of workflow with 2 GPU nodes

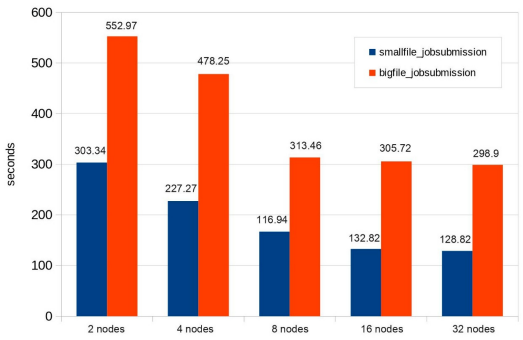

Figure 5. Performance for different nodes for job submission

Figure 4:

Conclusion

The SELVEDAS project demonstrates the feasibility of a hybrid cloud infrastructure supporting on-demand services with fast feedback analysis and analysis for archived data on the petabyte archive. The approach has been developed to provide the flexibility and extension to allow other institutions or domains to adopt similar approaches.

References

1. https://cug.org/proceedings/cug2019_proceedings/includes/files/pap138s2-file1.pdf.
2. Related Projects: EOSC, DAAS and PaNOSC

Primary authors: Dr CHANG, Mei-Chih (Paul Scherrer Institute (PSI)); Dr ASHTON, Alun W. (Paul Scherrer Institute (PSI)); Mr KLEEB, Hans-Christian S. (Paul Scherrer Institute (PSI)); Dr LEONG, Siew Hoon (Swiss National Supercomputing Centre (CSCS)); Mr DORSCH, Juan P. (Swiss National Supercomputing Centre (CSCS)); Mr ALIAGA, Tomas (Swiss National Supercomputing Centre (CSCS)); Dr MARTINASSO, Maxime (Swiss National Supercomputing Centre (CSCS)); Dr ALAM, Sadaf R. (Swiss National Supercomputing Centre (CSCS))

Presenter: Dr CHANG, Mei-Chih (Paul Scherrer Institute (PSI))

Session Classification: Scientific Computing, Machine Learning and large Data Management

Track Classification: Scientific computing, machine learning and large data management