Contribution ID: **11** Type: **Oral Presentation**

Function-as-a-Service and event-driven automation for the European Open Science Cloud

Wednesday, 3 April 2019 14:00 (30 minutes)

The Photon and Neutron science community (PaN) is pushing frontiers with ground breaking research and technologies in molecular imaging at the atomic level. State of the art Photon and Neutron sources, like the European XFEL and the European Spallation Source (ESS) will create hundreds of Petabytes of data per year, challenging established data processing strategies. Leveraging cloud computing methodologies, DESY develops innovative flexible and scalable storage and compute service, covering the entire data life cycle from experiment control to long term archival, with a particular focus on re-usability by the long tail of science.

Contributing to the European Open Science Cloud (EOSC) pilot, DESY, XFEL and ESS demonstrated cloud based solutions for FAIR access to large volumes of scientific data. The reproducibility of methods and results requires an integrated approach that bundles publications, data, workflows and functions. Fine grained access to functions-as-a-service (FaaS) infrastructures enables scientists to develop and deploy micro-services within minutes. Through shared container registries, those services will become available immediately as new cloud functions on the DESY compute cloud and be run by federated resources in the European Open Science Cloud.

The FaaS approach leads to evolving libraries and catalogues for standard functions, enhanced efficient resource provisioning and enables auto-scaling for compute intensive and repetitively used codes. For highly specialized applications, the platform preserves software environments, configurations and algorithm implementations, citable via DOIs.

We see the integration of our cloud functions into the European Open Science Cloud as an important incentive to further focus on metadata and data interoperability, feeding products from photon science into domain specific analysis and simulation tools e.g. in structural biology and material sciences. Well-defined interfaces enable users to route data through sequences of functions from various frameworks. Where data connectors or format converters are needed, they can be deployed as functions implementing additional micro-services. Providing scientists with the means to host and develop underlying codes, DESY runs collaborative platforms like GitLab and JupyterHub on auto-scaling clusters.

The interactive usage in scientific analysis is complemented by the directly achievable automation, which is closely integrated and designed to scale-up to high throughput use cases. Building on storage events, generated by the underlying dCache storage system, analysis pipelines can be triggered automatically for new, incoming data. The fully automated code execution in response to storage events directly extracts metadata, updates data catalogues, feeds into monitoring and accounting systems and creates derived data sets. In a decoupled design, storing derived data can trigger subsequent actions.

In the eXtreme-DataCloud (XDC) project, DESY demonstrates that event-driven code execution as a service adds a flexible building block to smart data placement strategies, enforcing machine actionable “Data Management Plans”. For this, the FaaS system interacts with rule-based data management engines and file transfer systems, e.g. to create replicas of data sets with respect to data locality and Quality of Service for storage. On data ingestion, files can be copied to cloud storage elements, which act as buffers next to strong clusters of compute elements, which carry out Function-as-a-Service pipelines and update data placement rules on success. This automatically deregisters files in the buffer next to the compute clusters, and triggers their distribution to offline storage and long term archival.

In our presentation, we will discuss both, the user and the provider perspective of a computing infrastructure, described above. We will elaborate on the benefits of such a decoupled cloud based service oriented architecture and we will provide a live demonstration, illustrating how to interactively install developed functions in an automated data processing pipeline.

Summary

The Photon and Neutron science community (PaN) will create hundreds of Petabytes of data per year and develops cloud based solutions for FAIR access and reproducibility of methods and results by bundling publications, data, workflows and functions-as-a-service (FaaS) infrastructures. This approach leads to evolving libraries and catalogues for standard functions, enhanced efficient resource provisioning and enables auto-scaling on cloud resources. Contributing to the European Open Science Cloud (EOSC) pilot, the platform preserves software environments, configurations and algorithm implementations.

Triggered in response to storage events from the underlying dCache storage system, analysis pipelines can be fully automated and event-driven code execution as a service adds a flexible building block to smart data placement strategies, enforcing machine actionable “Data Management Plans”.

This presentation, will discuss both, the user and the provider perspective of a decoupled cloud based service oriented architecture and will provide a live demonstration, illustrating how to interactively install developed functions in an automated data processing pipeline.

Primary author: Mr SCHUH, Michael (DESY)

Co-authors: Mr STAREK, Jürgen (DESY); Dr FUHRMANN, Patrick (DESY); Dr MILLAR, Paul (DESY); Dr FRANK, Schlünzen (DESY); Mr MKRTCHYAN, Tigran (DESY); Prof. GÜLZOW, Volker (DESY)

Presenter: Mr SCHUH, Michael (DESY)

Session Classification: Infrastructure Clouds and Virtualisation

Track Classification: Infrastructure Clouds and Virtualisation