Contribution ID: **34** Type: **Presentation**

dCache in XDC and EOSC: Event-driven data placement and processing

Tuesday, 7 May 2019 13:30 (20 minutes)

Latest technologies for molecular imaging at state of the art photon and neutron facilities produce petabytes of data, challenging established data processing strategies. DESY develops innovative, flexible and scalable storage and compute services for collaborative scientific computing on fast growing cloud infrastructures like the European Open Science Cloud. Covering the entire data life cycle from experiment control to long term archival, the particular focus on re-usability of methods and results leads to an integrated approach that bundles data, functions, workflows and publications.

On the frontend, scientists increasingly build on the popular Jupyter ecosystem to compose, run and share analysis workflows. Usually, they have little control over the underlying resources and consequently discover boundaries when working with big data. On one hand, provisioning of Jupyter Servers with limited resources diminishes the user experience, while on the other hand allocating larger environments for exclusive access quickly becomes inefficient and unfeasible. Furthermore, it remains difficult to provide dedicated setups for all possible use-cases which often require special combinations of software components.

We demonstrate, that a Function-as-a-Service approach to this problem leverages efficient, auto-scaling provisioning of cloud resources for scientific codes from lambda functions to highly specialized applications. Scientists develop and deploy containerized micro-services as cloud functions, while at the same time they are preserving software environments, configurations and algorithm implementations. Codes that are well-adopted and are successfully delivering services to the scientific community automatically scale up, while less frequently used functions do not allocate idle resources, but still remain operable and accessible. Functions can be called from Jupyter Notebooks and in addition integrate as a backend service for distributed cloud computing applications.

In the eXtreme-DataCloud (XDC) project, DESY demonstrated that event-driven function execution as-a-service adds a flexible building block to data life-cycle management and smart data placement strategies. The peta-scale storage system dCache provides storage events which directly feed into automation on production systems. In response to incoming files, services are invoked to immediately create derived data sets, extract metadata, update data catalogues, monitoring and accounting systems. Enforcing machine actionable Data Management Plans (DMP), rule-based data management engines and file transfer systems consume storage events e.g. to create replicas of data sets with respect to data locality and Quality of Service (QoS) for storage.

With a focus on metadata and data interoperability, sequentially executed functions span pipelines from photon science to domain specific analysis and simulation tools e.g. in structural biology and material sciences. Well-defined interfaces allow users to combine functions from various frameworks and programming languages. Where data connectors or format converters are needed, scientists can deliver their solutions as additional micro-services and programmable interfaces.

This presentation addresses the perspective of both, users and providers, to a cloud based micro-service oriented architecture and illustrates how to share codes and continuously integrate them in automated data processing pipelines as well as interactive workflows.

Type of abstract

Presentation

Primary author: SCHUH, Michael (DESY)

Co-authors: STAREK, Juergen (DESY); FUHRMANN, Patrick (DESY); MILLAR, Paul (DESY)

Presenter: SCHUH, Michael (DESY)

Session Classification: Federated Data Management