Contribution ID: **57** Type: **Oral Presentation**

Processing of storage events as a service for federated science clouds

Friday, 28 August 2020 11:50 (30 minutes)

Extending large scale research infrastructures for the Photon and Neutron (PaN) science community is done in preparation for hundreds of Petabytes of data, that will be delivered over the next years with ground breaking molecular imaging technologies. These data volumes and rates as well as a growing number of data sources result in an increasing demand for innovative, flexible storage and compute services for low latency, high throughput data processing in distributed environments. Leveraging cloud computing and containerization, DESY has presented pilot systems for the Photon and Neutron Open Science Cloud (PaNOSC), and provides collaborative platforms like GitLab and JupyterHub as services to the European Open Science Cloud (EOSC). Building on cloud and container orchestration templates, elastic analysis services can be dynamically provisioned freeing users from burden with infrastructure management.

This allows to focus on enhancements for efficient resource provisioning and auto-scaling units from very granular functions to long-running complex HTC jobs e.g. automating data reduction services on classic batch systems. Scaling out both, modern Function-as-a-Service systems like Apache OpenWhisk and well known dynamic batch systems like HTCondor, to transient resources introduces a shift from local to federated resource management, from user based brokering to matching the current demand by a service.

With respect to the FAIR principles, we encourage the reuse of data and metadata as well as codes and configuration by providing computational microservices which typically run with user supplied software stacks that integrate analysis frameworks and share large portions of their codebase. Function-as-a-Service systems allow to run user provided containers as cloud functions, which we continuously integrate in shared container registries (GitLab) and deploy to OpenWhisk running on a linked Kubernetes cluster.

Adopting storage events, presented by the dCache project at ISCG 2019, as triggers for automated scientific data processing has led to the implementation of event brokers (Apache Kafka) and stream processing modules. This infrastructure is used for routing of generated events from the backend storage as well as from various other sources to streams which consumer services can subscribe to in order to invoke cloud functions. At the center of our discussion is the need for a comprehensive security model, where authenticated clients are authorised to filter dedicated subsets of events, which may contain tokenized access delegation for reading data and storing output as well as access tokens for compute services to run data processing pipelines.

In this context we show, how the function-as-a-service compute model itself provides the means to implement stream processing for storage events and to provide secure delivery to external systems. We compare this to a more centralized design with a multi-tenant event streaming platform. Finally, we discuss use cases spannig from data taking and online monitoring to offline batch processing and multi-cloud bursting.

Summary

Large scale research infrastructures for Photon and Neutron (PaN) science leveraging cloud computing and containerization for low latency, high throughput data processing. Building dynamically provisioned, elastic analysis services to provide enhancements for efficient resource provisioning and auto-scaling units from very granular functions to long-running complex HTC jobs e.g. automating data reduction services on classic batch systems. Scaling out both, modern Function-as-a-Service systems like Apache OpenWhisk and well known dynamic batch systems like HTCondor, to transient resources introduces a shift from local to federated resource management, from user based brokering to matching the current demand by a service.

Providing reuseable computational microservices by running user supplied containers as cloud funtions and adopting storage events as triggers for automated scientific data processing. Implementation of event brokers (Apache Kafka) and stream processing modules for routing of generated events from the backend storage as well as from various other sources to streams which consumer services can subscribe to in order to invoke cloud functions. At the center of our discussion is the need for a comprehensive security model, where authenticated clients are authorised to filter dedicated subsets of events, which may contain tokenized access delegation for reading data and storing output as well as access tokens for compute services to run data processing pipelines.

In this context we show, how the function-as-a-service compute model itself provides the means to implement

stream processing for storage events and to provide secure event and token delivery to external systems. We compare this to a more centralized design with a multi-tenant event streaming platform. Finally, we discuss use cases spannig from data taking and online monitoring to offline batch processing and multi-cloud bursting.

Primary authors: HANNAPPEL, Juergen (DESY); STAREK, Juergen (DESY); SAHAKYAN, Marina (DESY); Mr SCHUH, Michael (DESY); Dr FUHRMANN, Patrick (DESY/dCache.org); Dr MILLAR, Paul (DESY); Mr MKRTCHYAN, Tigran (DESY)

Presenter: Mr SCHUH, Michael (DESY)

Session Classification: Data Management Session

Track Classification: Data Management & Big Data