Contribution ID: **30** Type: **Practical Demonstration (15')**

Parallel FaaS on the Edge-Cloud continuum

Thursday 13 October 2022 13:45 (15 minutes)

The design and execution of AI applications on Edge-Cloud infrastructures require a set of programming tools to seamlessly design and partition the models and a runtime that properly assigns the different components on the available nodes, also distributing the data.

The design abstractions allow to provide high-level annotations to specify QoS constraints and code dependencies and to introduce performance parameters for the allocation of tasks to computing continuum resources and security and privacy annotations for data allocation and processing.

The runtime leverages the embedded computing resources of each node to host the execution of functions in a service manner and generates hybrid workflows, composed of atomic and continuous processing tasks, achieving distribution, parallelism and heterogeneity across edge/cloud resources transparently to the application developer.
The adoption of such a distributed model for executing the applications allows the user to concentrate on the application development and rely on the infrastructure management by the serverless platform.

The programming framework runtime parallelizes the execution of the different parts of the applications that can be invoked in a FaaS way according to the QoS constraints. The runtime is able to schedule the tasks on both edge and cloud devices, orchestrating the execution and leveraging on fault tolerance mechanisms to react to the dynamicity of the edge.

The aim of the AI-SPRINT "Artificial intelligence in Secure PRIvacy-preserving computing coNTinuum"project is to develop a platform composed of design and runtime management tools to seamlessly design, partition and operate Artificial
Intelligence (AI) applications on the Edge-Cloud continuum, providing resource efficiency, performance, data privacy, and security guarantees.
In this presentation we will demonstrate the execution of a healthcare application that is built using the design time tools and FaaS components of AI-SPRINT, namely the PyCOMPSs programming framework and the OSCAR event-driven serverless applications manager.

Primary authors: LEZZI, Daniele (Barcelona Supercomputing Center); Dr LORDAN GOMIS, Francesc (Barcelona Supercomputing Center)

Presenter: LEZZI, Daniele (Barcelona Supercomputing Center)

Session Classification: IBERGRID Contributions

Track Classification: Cooperation between Iberian research communities