Task Orchestrator for *edge* platforms

D. Trystram
Univ. Grenoble Alpes, Grenoble INP

October 7, 2020

1 Context

1.1 Emerging Edge Architectures

The main idea behind the recent paradigm of *edge computing* is to be an alternative to the highly energy consuming cloud computations. Instead of transferring all data across the network through the Cloud to be executed and stored in large remote centralized data centers, most processing should be transferred closer to the source of data (where they are produced and exploited).

This promising paradigm, based on distributed computing, is expected to improve cloud service deployments by using opportunistic local computing resources. It is studied for several years in the DataMove team at Inria-LIG.

1.2 A new type of services

Serverless Computing is a very recent paradigm that could be integrated into edge computing. It has been introduced in 2019 for simplifying the usage of distributed computing resources by removing the burden of resource management while enabling programmers to develop applications as group of individual *functions* that can run independently. Furthermore, while the Cloud provider manages completely the resources provisioning, the application is charged only for the execution time of each function thus, minimizing the cost of the whole infrastructure usage.

The challenge of managing such a complex framework is further intensified by the heterogeneity of the available computational resources, which may include general-purpose CPUs, low powered CPUs, accelerators, dedicated embedded systems, etc. and the critical isolation demands to support the expected related levels of security.

1.3 Objectives

In this context we need to develop mechanisms to support coherent, low bring-up times, fast execution, resource savings, fault-tolerance and flexible resource management in a secure manner.

The combination of edge computing with the serverless approach pose new challenges to compute virtualization with the need for short-lived functions with low latency demands to be executed on resource constrained devices.

The proposed subject deals with the efficient management of resources in this context and the orchestration of the distributed (local) jobs.

2 Characteristics of edge platforms

The execution of workflows in edge platforms using a serverless runtime arise interesting challenges that are described below.

As the team DataMove has a strong expertise on IA and data analytics applications (that constitute most of the computations running on existing edge platforms), we will concentrate to this class of applications.

- **Heterogeneity.**

  The hyper-heterogeneity of the hardware composing the edge computing environment, the mobility of some edge compute resources, the continuous change of network links along with connection intermittence impose dynamicity/elasticity in the way the resource management and orchestration should be done.

  Based on the characteristics of edge resources and the type of workflows (again, we target IA applications), different configurations of computing units can be proposed, which can evolve dynamically depending on their geo-localization, and the needs of the edge nodes along with the system's needs and its constraints. The runtime provides mechanisms to seamlessly deploy workflows on a possibly continuously evolving distributed computing environment. This aspect is studied by other researchers in DataMove and in the Edge Intelligence MIAI research project.

  In particular, the resource management and services orchestration should be transparent to the developer who may only dictate high-level constraints/criteria/objectives (such as latency, locality, cost, energy saving, data privacy, etc.). Several optimization objectives and constraints will be investigated.

- **Mobile and distributed management.**

  In such hybrid and heterogeneous computing environments with possibly disconnecting parts, a fully distributed control of compute resources where each separate compute node can take part in the decision process, could seem ideal at the first glance. Nevertheless, these solutions suffer of too heavy schedule times (and also implementation complexity). On another hand, a purely centralized approach would not allow independence of compute intelligence when a part is disconnected. Hence, a decentralized approach based on the federation of computing units would bring the best of both worlds.

  We need to enable intelligent scheduling of different compute clusters that have to be managed separately but collaborate in the execution of cross-domain workflows.

- **Dealing with uncertain data**

  The last characteristics of edge platforms is to maintain the most precise environment as possible.

  Computing local AI analytics will be provided by light-way algorithms. This makes the whole federated picture more difficult to construct.

  The propsoed subject will investigate one of the three previous directions.

The collaboration with the start-up Ryax Technologies (in Lyon) will benefit from their software suite.