Multi-objective resource allocation in modern distributed infrastructures

Denis Trystram (DataMove)

1 Context and motivation

We are interested here in accompanying the huge development of the next generation of resource and compute distributed management systems of hybrid Big Data infrastructures, such as data servers, clouds, fog or edge computing. The technology moves fast and we are entering a new age of computing where the Internet of Things deal with petabytes of data while applications from virtual reality, smart cities to autonomous driving need extreme low latency responses. The efficient orchestration of compute workloads executed upon hybrid and complex infrastructures requires the latest innovations from different areas of computer science including stream and batch processing, distributed computing, large-scale system design, security and privacy, user interface design, machine learning, to cite only the most important ones.

Hybrid infrastructures have the characteristic of being composed of heterogeneous multiple computing units and networks, some being very close to the data sources – same building – (known as the Edge), some being geographically close (known as the Fog) and some geographically remote (known as the Cloud). In order to have efficient resource and compute managers on such hybrid distributed infrastructures, it is crucial to decide where computations will take place based on different types of objectives defined by the needs of each application.

The complexity and hybrid nature of big data infrastructures demand the development of multi-objective resource allocation techniques to take into account key parameters such as latency, cost, energy, data locality, security, privacy in order to schedule efficiently the workloads upon the right machines.

2 Description

The objective of this subject is to study different multi-objective scheduling algorithms and to model the hybrid infrastructure eco-system of a smart building. The focus is to study this distributed system and to develop a prototype that will allow several objectives to be taken into account simultaneously to determine a nearly optimal solution as fast as possible with a low overhead. The proposed study will be finalized with both theoretical analysis of algorithms and performance evaluation results in terms of overhead, efficiency and scalability of the new prototype in a simulated environment.