Compas'2023 : Parallélisme / Architecture/ Système
LISTIC/USMB - Annecy France, 4-7 juillet 2023

Towards an Energy-Aware Multi-objective Scheduling Policy for Server-based Edge-Cloud Continuum

Anderson Andrei Da Silva[✦✧], Yiannis Georgiou[✧], Michael Mercier[✧],
Gregory Mounié[✦], Denis Trystram[✦]

[✦]Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, Grenoble, France,
[✧]Ryax Technologies, Lyon, France
anderson-andrei.da-silva@inria.fr, yiannis.georgiou@ryax.tech,
michael.mercie@ryax.tech, gregory.mounie@imag.fr, denis.trystram@inria.fr

---

Résumé
The cloud is extended towards the edge to form a computing continuum while managing resources' heterogeneity. The serverless technology simplified how to build cloud applications and use resources, becoming a driving force in consolidating the edge-cloud continuum with the deployment of small functions with short execution. Function as a service (FaaS) is an approach encompassed by serverless, extending the FaaS concept by avoiding server infrastructure management. The edge-cloud continuum can be defined by an architecture with different levels composed of several cloud clusters, edge clusters, and possibly edge resources. There, the high-level global continuum has the view of all possibly heterogeneous clusters and resources of the continuum ; and the local cluster level is composed of, in general, groups of homogeneous resources. However, the adaptation of serverless to the continuum brings new challenges mainly related to resource management and scheduling. Kubernetes-based platforms are used for serverless proposes, and Kubernetes provides a few scheduling mechanisms to address some of these challenges, but they are not ideal nor designed for serverless. Standard cloud scheduling policies are based on greedy algorithms that do not efficiently handle platforms' heterogeneity, do not deal with problems such as cold start delays, nor present efficient energy-aware approaches. Therefore, this work introduces a new scheduling policy that tries to address these issues. It is based on a two-level scheduling policy. The first level relies on multi-objective optimization for energy consumption and makespan while considering the platform heterogeneity. The second is a ContainerLayer-Aware policy that optimizes function's placements by selecting machines with the biggest rate of container layers that can be shared. Such a multi-level scheduling policy targets the edge-cloud continuum. The first scheduling level makes the decisions regarding the clusters on the global-level of the continuum, and the second scheduling policy level makes the decisions regarding the machines on the local level of the continuum. Using simulations that vary workloads, platforms, and heterogeneity levels, we study the system utilization and the trade-offs between the targets. We perform comparisons with a baseline inspired by a Kubernetes-based policy, and our experiments show representative gains, in order of magnitude, for all metrics and targets.

Mots-clés : Scheduling Policies, Serverless Computing, Edge-Cloud Continuum, Heterogeneous Platforms, Energy Consumption.
