Orchestrating FaaS Applications across the Cloud-Edge Continuum

***Keywords*:** cloud-edge continuum, FaaS, serverless, edge, fog

Contact:

Nikos Parlavantzas, IRISA, Rennes (Nikos.Parlavantzas@irisa.fr**)**

François Lemercier, IRISA, Rennes (Francois.Lemercier@irisa.fr)

Team: MAGELLAN (https://team.inria.fr/magellan), IRISA

Location: Rennes

Description

Modern data-intensive applications, such as AI-driven Internet of Things (IoT) applications [1], require efficient processing of large-scale data generated at the edge of the network. The current trend leans towards deploying such applications closer to data sources along the cloud-edge continuum, ensuring reduced latency, improved reliability, and enhanced privacy. A promising model for developing and deploying applications on the cloud-edge continuum is serverless computing, and in particular the Function-as-a-Service (FaaS) model [1]. In the FaaS model, applications are built as dynamic compositions of functions that can be flexibly scaled and migrated along the cloud-edge continuum to optimise application performance and cost. Although applying the FaaS model to edge applications is gaining increasing research attention [3,4], much further research is needed before practical solutions begin to emerge.

Indeed, managing complex FaaS applications across the cloud-edge continuum poses significant challenges. First, these applications typically involve large-scale, dynamic workflows comprising tasks with continuously fluctuating computational requirements. These workflows must execute on geographically dispersed and diverse resources, including sensors, mobile devices, and micro data centers. As a result, meeting application Quality of Service (QoS) requirements, such as performance and energy consumption requirements, becomes complex [5]. Second, the performance of data-intensive FaaS applications strongly depends on the allocation of network resources [6]. Despite this dependency, current research has not yet addressed network resource allocation for FaaS deployments across the continuum. Third, FaaS applications require cost-effective and efficient mechanisms for exchanging data among functions [7,8]. Cloud-based storage services, such as S3 from AWS, impose delays and incur operational costs, while edge-based services suffer from scarce bandwidth and limited resources.

This thesis will explore methods and tools to enable effectively orchestrating FaaS applications across the cloud-edge continuum. Specifically, the thesis will propose intelligent management policies to dynamically orchestrate workflows composed of FaaS functions with the goal of satisfying application requirements. The work will include developing network management mechanisms for guaranteeing performance as well as data management mechanisms for optimizing data exchange. The work will build upon and extend a framework for placing FaaS functions in fog environments, currently under development in the Magellan team [9]. Additionally, the work will leverage recent research within the team on FaaS scheduling using Deep Reinforcement Learning (DRL) to jointly optimise latency and energy reduction objectives. The results of this thesis will be evaluated through the deployment of various AI-

based applications in both the Grid'5000 testbed and a fog infrastructure dedicated to environmental monitoring located at the Beaulieu campus.

Qualifications

- Excellent communication and writing skills
- Strong programming and scripting skills in Linux environments
- Knowledge and experience in one or more of the following areas: distributed systems, adaptive systems, FaaS, IoT, machine learning, cloud, edge

References

[1] R. Singh and S. S. Gill, "Edge AI: a survey", Internet of Things and Cyber-Physical Systems 2023, Vol. 3, Page 71-92, doi:10.1016/j.iotcps.2023.02.004

[2] J. Schleier-Smith, V. Sreekanti, A. Khandelwal, J. Carreira, N. J. Yadwadkar, R. A. Popa, J. E. Gonzalez, I. Stoica, and D. A. Patterson, "What serverless computing is and should become: The next phase of cloud computing", Commun. ACM, vol. 64, no. 5, p. 76–84, Apr. 2021. Available: https://doi.org/10.1145/3406011

[3] M. S. Aslanpour, A. N. Toosi, C. Cicconetti, B. Javadi, P. Sbarski, D. Taibi, M. Assuncao, S. S. Gill, R. Gaire, and S. Dustdar, "Serverless Edge Computing: Vision and Challenges", in 2021 Australasian Computer Science Week Multiconference (ACSW '21), ACM, New York, NY, USA, Article 10, 1–10, doi:/10.1145/3437378.3444367

[4] P. Raith, S. Nastic and S. Dustdar, "Serverless Edge Computing—Where We Are and What Lies Ahead", in IEEE Internet Computing, vol. 27, no. 3, pp. 50-64, May-June 2023, doi: 10.1109/MIC.2023.3260939

[5] A. Tariq, A. Pahl, S. Nimmagadda, E. Rozner, and S. Lanka. "Sequoia: enabling quality-of-service in serverless computing". In Proceedings of the 11th ACM Symposium on Cloud Computing (SoCC '20). Association for Computing Machinery, New York, NY, USA, 2020, 311–327, https://doi.org/10.1145/3419111.3421306

[6] J. Kim and K. Lee, 2020. "I/O resource isolation of public cloud serverless function runtimes for data-intensive applications", Cluster Computing 23, 3 (Sep 2020), 2249–2259. https://doi.org/10.1007/s10586-020-03103-4

[7] C. Cicconetti, M. Conti, and A. Passarella. "In-network computing with function as a service at the edge", Computer, 55(09):65–73, sep 2022

[8] Y. Tang and J. Yang, "Lambdata: Optimizing serverless computing by making data intents explicit," in 2020 IEEE 13th International Conference on Cloud Computing (CLOUD), 2020, pp. 294–303

[9] V. Parol-Guarino, N. Parlavantzas, "GIRAFF: Reverse Auction-based Placement for Fog Functions", 9th International Workshop on Serverless Computing (WoSC'23), Bologna, Italy, Dec 2023

[10] C. Latreche, N. Parlavantzas, H. Duran-Limon, "FoRLess: A Deep Reinforcement Learning-based approach for FaaS Placement in Fog", 17th IEEE/ACM International Conference on Utility and Cloud Computing (UCC 2024), Sharjah, United Arab Emirates, Dec 2024