Special Issue: Serverless Computing

Shadi Ibrahim, *Inria centre at Rennes University, Rennes, 35042, France*

Omer Rana, *Cardiff University, Cardiff CF24 4AG, UK*

Olivier Beaumont, *Inria centre at the University of Bordeaux, Talence, 33405, France*

Xiaowen Chu, *The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, 511453, China*

Abstract—***This special issue covers emerging research and development challenges in Serverless Computing, ranging from the development of an allocation and scheduling simulator to evaluate various orchestration policies, to support for scalable and cost-effective deployment of serverless machine learning models in heterogeneous edge environments. The articles also demonstrate how the Serverless Computing paradigm has been deployed across different systems architectures, from edge-only deployments to those that make use of a combination of edge and cloud environments. There is clear enthusiasm in the systems community in utilizing this paradigm for a wide range of applications.***

Serverless Computing extends cloud computing by allowing programmers to develop, run and scale their applications without having to worry about infrastructure provisioning and management. By breaking an application into small standalone functions, Serverless Computing can offer several advantages, such as high elasticity, cost efficiency, and ease of deployment. Serverless Computing has received significant attention in the past few years and has been adopted for a wide range of applications and infrastructures. This paradigm has primarily focused on supporting short running functions, especially where startup time (referred to as *cold start time*) can be low. Conversely, *warm start time* refers to functions that have already been initialized and are already waiting for trigger events to execute. Serverless Computing, however, brings challenges and new opportunities in many aspects such as system architectures, scheduling, data sharing, energy, privacy and adoptions for Machine Learning (ML) applications – especially where ML functions take a long time to execute. This special issue aims to bring together the state-of-the-art research results of Severless Computing.

XXXX-XXX © 2021 IEEE
Digital Object Identifier 10.1109/XXX.0000.0000000

IN THIS ISSUE

The four articles in this special issue covers emerging research and development challenges in Serverless Computing. The first article deals with providing an allocation and scheduling simulator to evaluate serverless orchestration policies. The second article covers the issue of supporting scalable and cost-effective deployment of serverless machine learning models in heterogeneous edge settings. The third article looks at providing energy-efficient execution of serverless applications by exploring CPU share and core frequency. The fourth article discusses and evaluates a serverless platform designed specifically for edge-only deployments.

The first article is "HeROsim: An Allocation and Scheduling Simulator for Evaluating Serverless Orchestration Policies" by V. Lannurien et al.[A1] The authors present the design and implementation of the heterogeneous resources orchestration simulator (HeROsim), a discrete-event simulator that helps to evaluate resource allocation and task scheduling policies in serverless platforms. HeROsim supports a fine-grained description of serverless applications on heterogeneous resources, taking into account the characteristics of functions (e.g., execution time, cold start time, energy consumption) and the temporal dependencies between functions. It also supports different performance metrics ranging from latency to energy efficiency. The authors present two case studies where

HeROsim is used to evaluate orchestration policies. The simulator is publicly available to the research community.

The second article "ARASEC: Adaptive Resource Allocation and Model Training for Serverless Edge Computing" by D. Katare et al.,[A2] focuses on the challenge of developing and deploying resource-aware AI/ML models in serverless edge computing. The authors develop a serverless-oriented framework (ARASEC) that facilitates model deployment in edge-cloud environments. ARASEC makes use of an adaptive partitioning methodology to adapt AI/ML models to heterogeneous computing environments, thereby improving resource allocation; and incorporates advanced machine learning algorithms including distributed nesterov accelerated gradient (D-NAG) and asynchronous parallel stochastic gradient descent (APSGD) to reduce operational costs and improve efficiency.

The third article is "Power-Aware CPU Cap Mechanism in Serverless Computing Environments" by M. R. HoseinyFarahabady et al.[A3]. The authors show the benefits of CPU frequency scaling along with resource allocation for power savings in serverless platforms. They present a CPU cap controller that dynamically adjusts CPU share and core frequency according to the Quality of Service (QoS) of serverless applications, thereby achieving lower power consumption while consolidating functions. The authors integrate their controller into Apache OpenWhisk and demonstrate its effectiveness on different cloud workloads.

The fourth article is titled "WebAssembly at the Edge: Benchmarking a Serverless Platform for Private Edge Cloud Systems" by G. De Palma et al.[A4] The authors present *FunLess* – a Function-as-a-Service (FaaS) platform specifically designed for private edge cloud systems. FunLess uses WebAssembly as its runtime environment to provide performance and lightweight isolation for serverless functions, and to offer support for heterogeneous devices. The authors implement and evaluate FunLess and show that FunLess achieves performance comparable to other platforms using binary native code on constrained-resource devices, but with significantly less memory and bandwidth. The sources and documentation of FunLess are publicly available to the research community.

Overall, this special issue demonstrates key novel themes in Serverless Computing, focusing on the deployment of machine learning algorithms to support resource constrained execution (including a focus on power efficiency). This special issue received twelve submissions, and after a thorough peer review process (including second round reviews), four papers were selected for publication. We are grateful to all the colleagues involved in the production of this special issue, from authors who submitted their papers, to the reviewers who undertook timely reviews of these submissions. We are especially grateful to the IEEE Internet Computing editorial team and the editor in chief for their support in making this special issue a reality.

APPENDIX: RELATED ARTICLES

A1. V. Lannurien, L. d'Orazio, O. Barais, S. Paquelet and J. Boukhobza, "HeROsim: An Allocation and Scheduling Simulator for Evaluating Serverless Orchestration Policies," in IEEE Internet Computing, doi: 10.1109/MIC.2024.3511332.

A2. D. Katare, E. Marin, N. Kourtellis, M. Janssen and A. Y. Ding, "ARASEC: Adaptive Resource Allocation and Model Training for Serverless Edge Computing," in IEEE Internet Computing, doi: 10.1109/MIC.2024.3514670.

A3. M. R. HoseinyFarahabady and A. Y. Zomaya, "Power-Aware CPU Cap Mechanism in Serverless Computing Environments," in IEEE Internet Computing, doi: 10.1109/MIC.2024.3513446.

A4. G. De Palma, S. Giallorenzo, J. Mauro, M. Trentin and G. Zavattaro, "WebAssembly at the Edge: Benchmarking a Serverless Platform for Private Edge Cloud Systems," in IEEE Internet Computing, doi: 10.1109/MIC.2024.3513035.

Shadi Ibrahim is a Research Scientist at Inria centre at Rennes University, Rennes, 35042, France. Contact him at shadi.ibrahim@inria.fr.

Omer Rana is Professor of Performance Engineering at the School of Computer Science and Informatics at Cardiff University, UK. Contact him at ranaof@cardiff.ac.uk.

Olivier Beaumont is a Senior Research Scientist at Inria centre at the University of Bordeaux, Talence, 33405, France. Contact him at Olivier.Beaumont@inria.fr.

Xiaowen Chu is a Professor and Head at the Data Science and Analytics Thrust of Hong Kong University of Science and Technology (Guangzhou), Guangzhou, 511453, China. Contact him at xwchu@hkust-gz.edu.cn.