Efficient AI Applications in Edge-Cloud Environments

In-Young Ko[1], Michael Mrissa[2,3], Juan Manuel Murillo[4,5] and Abhishek Srivastava[6]

[1]*Korea Advanced Institute of Science and Technology, South Korea*
[2]*InnoRenew CoE, Slovenia*
[3]*University of Primorska, Slovenia*
[4]*University of Extremadura, Spain*
[5]*COMPUTAEX Foundation, Spain*
[6]*Indian Institute of Technology Indore, India*
E-mail: iko@kaist.ac.kr; michael.mrissa@innorenew.eu; juanmamu@unex.es; asrivastava@iiti.ac.in

The international workshop on Big Data-Driven Edge Cloud Services (BECS) provides a venue for scholars and practitioners to share their experiences and present their ongoing work in the development of data-driven AI applications and services in a distributed computing environment known as the edge cloud. The third edition of the workshop (BECS 2023)[1] was held in conjunction with the 23rd International Conference on Web Engineering (ICWE 2023),[2] which was held in Alicante, Spain on June 6–9, 2023.

This special issue of the Journal of Web Engineering focuses on addressing the challenges related to the development and provision of highly efficient AI applications and services in edge cloud environments. For this issue, we have selected papers from BECS 2023 that emphasize the creation of high-performance edge-cloud infrastructure and the development of AI applications optimized for such infrastructure.

Edge-cloud environments exhibit high dynamism with varying resource availability and user demands. Consequently, addressing load bursts becomes

[1]https://becs.kaist.ac.kr/iwbecs2023/
[2]https://icwe2023.webengineering.org/

crucial for maintaining service quality in such settings. In their article titled "Fully Decentralized Horizontal Autoscaling for Bursts of Load in Fog Computing," EunChan Park et al. propose a method where microservice instances autonomously make scaling decisions, encompassing self-monitoring, allowing for cloning or termination. This article extends their work from BECS 2023, enhancing the core algorithm, incorporating additional scale-up reaction time measurements, and conducting supplementary experiments and a case study.

Edge devices often incorporate multiple devices that harness asymmetric multi-cores to facilitate efficient data processing. To ensure top-notch performance, it becomes critical to judiciously allocate tasks to the appropriate cores within asymmetric multi-core processors. In their article titled "Priority-Based Scheduler for Asymmetric Multi-Core Edge Computing," Rupendra Pratap Singh Hada and Abhishek Srivastava present an algorithm that takes into account the capabilities of the cores and the priority of the tasks to address this challenge effectively. This article is an extension of their work presented at BECS 2023, with enhancements made to the discussion of related work in task scheduling, including a performance comparison. They also enhanced the evaluation by introducing a performance assessment of the proposed algorithm.

In the article titled "Exploring LLM-Based Automated Repair of Ansible Scripts in Edge-Cloud Infrastructures," Sunjae Kwon et al. delve into the feasibility of using Large Language Models (LLMs) for automated repair of Ansible scripts, a popular Infrastructure-as-Code (IaC) tool. The authors evaluate the performance of LLMs, including ChatGPT and Bard, across many Ansible script revision cases from Open Source Software (OSS). They also curate defective Ansible scripts from OSS repositories on GitHub to confirm the practicality of LLM-based Automated Repair of Programs (ARP). Furthermore, they conduct comparative assessments involving three LLMs, exploring various prompts and defect types to identify effective strategies for guiding LLMs in generating error-free Ansible scripts.

Urban Digital Twin (UDT) is a digital twin capable of modeling specific city aspects such as transportation, and environmental factors. Recently, the interest in the social aspects of smart cities has grown fast, and citizens are now considered as first-class entities of the UDT since they are the fundamental key to this ecosystem. In the article, titled, "Integrating Citizens' Avatars in Urban Digital Twins," Rafael García-Luque et al. propose to integrate citizens through their Digital Avatars (DAs) into UDTs. Particularly, the authors developed a framework that takes advantage of the Cloud-to-Thing

Continuum for optimizing the available processing resources. They extended their BECS 2023 paper by adding sections to evaluate the performance and accuracy of their UDT and to analyze how the citizen relates to the UDT through its own digital twin.

An Urban Digital Twin (UDT) is a digital model that focuses on specific city aspects like transportation and environmental factors. In recent times, there has been a rapidly growing interest in the social aspects of smart cities, with citizens now being recognized as first-class entities within the UDT ecosystem. In the article titled "Integrating Citizens' Avatars into Urban Digital Twins," authored by Rafael García-Luque et al., the proposal is to incorporate citizens into UDTs through their digital avatars. The authors have developed a framework that leverages the Cloud-to-Thing Continuum to optimize available processing resources. To enhance their BECS 2023 paper, they have expanded it to include sections for evaluating UDT performance and accuracy and for analyzing the interaction between citizens and the UDT via their digital avatars.

As mentioned earlier, these articles address a range of topics related to enhancing the efficiency of AI applications in edge-cloud environments. They also provide real-world examples of how these AI applications make use of computational resources and data in distributed edge-cloud environments.

Acknowledgment

The BECS 2023 workshop was supported by the Ministry of Science and ICT (MSIT), Korea, under the Information Technology Research Center (ITRC) support program (IITP-2023-2020-0-01795) supervised by the Institute of Information & Communications Technology Planning Evaluation (IITP).

Michael Mrissa gratefully acknowledges the European Commission for funding the InnoRenew CoE project (Grant Agreement #739574) under the Horizon2020 Widespread-Teaming program and the Republic of Slovenia (Investment funding of the Republic of Slovenia and the European Regional Development Fund).