Summary | Call 18 | Stipendium ID 6851

1. General

Latency-sensitive applications require strict adherence to quality-of-service (QoS) requirements, particularly response times, to avoid significant performance degradation or critical situations, such as motion sickness in augmented reality or accidents in autonomous vehicles. These applications have high resource demands, which local devices like AR headsets or autonomous vehicles often cannot meet. This limitation necessitates the offloading of tasks to edge servers, which offer greater computational power and can meet strict timing requirements.

However, offloading to edge environments presents several challenges, including unreliable infrastructure due to limited resources and unstable connections, resource heterogeneity that ranges from micro data centers to low-powered devices, dynamic failures and workload volatility, and mobility-related issues that expose devices to varying server behaviors. The primary objective of the research was to develop an offloading framework capable of guaranteeing reliable (near-)real-time performance in such volatile and unreliable environments. The focus was on ensuring guaranteed reliability rather than mere performance optimization, particularly for latency-sensitive applications.

2. Results

The research produced FRESCO, a framework that addresses critical challenges in edge offloading for latency-sensitive applications. It combines multiple components, including Satisfiability Modulo Theory (SMT) for formal guarantees of decision feasibility under strict timing constraints, a blockchain-based reputation system to track and ensure server reliability while preventing tampering, and hybrid smart contracts (HSC) to enable fast, off-chain decision-making while preserving secure on-chain reputation data.

FRESCO was evaluated using real-world datasets, such as Skype availability traces and OpenCellID locations, in simulated edge environments. Comparisons were made with state-of-the-art approaches, including Markov Decision Process (MDP), Mixed Integer Nonlinear Programming (MINLP), and Social Queueing (SQ). Experimental results demonstrated significant performance improvements, with response times reduced by up to 7.86x, energy efficiency gains of up to 5.4%, and minimized QoS violations to 0.4%. Decision-making efficiency was achieved with an average decision time of 5.05 milliseconds, ensuring suitability for latency-sensitive applications. Cost-effectiveness was also demonstrated through optimized server selection, balancing computational and monetary costs, and supporting practical commercial use.

3. Planned Future Activities

The completed research will be submitted to the IEEE Transactions on Service Computing by the end of 2024, with the review process expected to take up to a year. In parallel, further work will focus on runtime verification of edge offloading, aiming to formally verify non-functional requirements and enhance performance reliability during runtime operations. Detailed plans for this are outlined in an accompanying document.

4. Suggestions for Continuation by Third Parties

FRESCO offers significant potential for extension and application in various domains. It can be applied to cross-domain areas such as healthcare, autonomous systems, and industrial Internet of Things, where reliability and latency are critical. It also presents opportunities for interoperability studies to integrate FRESCO with existing edge computing frameworks and blockchain ecosystems. Furthermore, integrating machine learning solutions can help balance performance optimization and reliability guarantees, particularly for resource allocation and failure prediction in diverse application scenarios.