E2Clab: Reproducible Analysis of Complex Workflows on the Edge-to-Cloud Continuum

Daniel Rosendo, Alexandru Costan, Gabriel Antoniu, Patrick Valduriez

E2C*lab*: Reproducible Analysis of Complex Workflows on the Edge-to-Cloud Continuum

PhD Student: Daniel Rosendo* **Advisors:** Alexandru Costan*, Gabriel Antoniu*, Patrick Valduriez†
*University of Rennes, Inria, CNRS, IRISA - Rennes, France
†University of Montpellier, Inria, CNRS, LIRMM - Montpellier, France
{daniel.rosendo, alexandru.costan, gabriel.antoniu, patrick.valduriez}@inria.fr

***Abstract*—Distributed digital infrastructures for computation and analytics are now evolving towards an interconnected ecosystem allowing complex applications to be executed from IoT Edge devices to the HPC Cloud (aka the *Computing Continuum*, the *Digital Continuum*, or the *Transcontinuum*). Understanding end-to-end performance in such a complex continuum is challenging. This breaks down to reconciling many, typically contradicting application requirements and constraints with low-level infrastructure design choices. One important challenge is to accurately reproduce relevant behaviors of a given application workflow and representative settings of the physical infrastructure underlying this complex continuum. We introduce a rigorous methodology for such a process and validate it through E2C*lab*. It is the first platform to support the complete analysis cycle of an application on the Computing Continuum. Preliminary results with real-life use cases show that E2C*lab* allows one to understand and improve performance, by correlating it to the parameter settings, the resource usage and the specifics of the underlying infrastructure.**

***Index Terms*—Methodology, Computing Continuum, Reproducibility, Machine Learning, Optimization.**

I. Context

The explosion of data generated from the Internet of Things (IoT) and the need for real-time analytics has resulted in a shift of the data processing paradigms towards decentralized and multi-tier computing infrastructures and services. New challenging application scenarios are emerging from a variety of domains such as healthcare, asset lifetime monitoring in industry, precision agriculture, *etc.* This contributes to the emergence of what is called the *Computing Continuum* [1] (or the *Digital Continuum* or the *Transcontinuum*). It seamlessly combines resources and services at the center (e.g., in Cloud datacenters), at the Edge, and in-transit, along the data path. Typically data is first generated and preprocessed (e.g., filtering, basic inference) on Edge devices, while Fog nodes further process partially aggregated data. Then, if required, data is transferred to HPC-enabled Clouds for Big Data analytics, AI model training, and global simulations.

II. Problem Statement

However, despite an always increasing number of dedicated systems for data processing on each component of the continuum, this vision of ubiquitous computing remains largely unrealized. This is due to the complexity of deploying large-scale, real-life applications on such heterogeneous infrastructures, which breaks down to configuring a myriad of system-specific parameters and reconciling many requirements or constraints, e.g., in terms of communication latency, energy consumption, resource usage, data privacy. A first step towards reducing this complexity and enabling the Computing Continuum vision is to enable a **holistic understanding of performance** in such environments. That is, finding a rigurous approach to answering questions like: *(1) How to identify infrastructure bottlenecks? (2) Which system parameters and infrastructure configurations impact on performance and how? (3) Where should the workflow components be executed to minimize communication costs and end-to-end latency?*

III. State of the art

Approaches based on workflow modelling [2] and simulation [3] raise some important challenges in terms of specification, modelling, and validation in the context of the Computing Continuum. For example, it is increasingly difficult to assess the impact of the inherent complexity of hybrid Edge-Cloud deployments on performance. At this stage, experimental evaluation remains the main approach to gain accurate insights of performance metrics and to build precise approximations of the expected behavior of large-scale applications on the Computing Continuum, as a first step prior to modelling.

IV. Challenges

A key challenge in this context is to be able to reproduce in a representative way the application behavior in a controlled environment, for extensive experiments in a large-enough spectrum of potential configurations of the underlying Edge-Fog-Cloud infrastructure. In particular,

this means rigorously mapping the scenario characteristics to the *experimental environment*, identifying and controlling the relevant *configuration parameters* of applications and system components, defining the relevant *performance metrics*. The above process is non-trivial due to the multiple combination possibilities of heterogeneous hardware/software resources, system components for data processing, analytics or AI model training.

V. PhD Objectives

In order to allow other researchers to leverage the experimental results and advance knowledge in different domains, the testbed needs to enable three R's of research quality: **Repeatability, Replicability,** and **Reproducibility (3R's)**. This translates to establishing a *well-defined experimentation methodology* and providing *transparent access to the experiment artifacts* and *experiment results*.

The Computing Continuum vision calls for a rigorous and systematic methodology to map real-world application components and dependencies to infrastructure resources, a complex process that can be error prone. Key research goals are: *1)* to identify relevant characteristics of the application workloads and of the underlying infrastructure as a means to enable accurate experimentation and benchmarking in relevant infrastructure settings in order to understand their performance; and *2)* to ensure research quality aspects such as the 3R's.

VI. Our Contribution: **E2C***lab*

E2C*lab* [4], implements a methodology that supports the complete experimental cycle across the edge-to-cloud continuum, including deployment, configuration, optimization, and experiment execution in a reproducible way. It may be used by researches to deploy real-life applications on large-scale testbeds and perform meaningful experiments in a systematic manner. The **main contributions** of this work are:

A **rigorous methodology for designing experiments with real-world workloads on the Computing Continuum** spanning from the Edge to the Cloud; this methodology provides guidelines to move from real-world use cases to the design of relevant testbed setups for experiments enabling researchers to understand performance and to ensure the 3R's properties.

A novel **framework named E2C***lab* that implements this methodology and allows researchers to deploy their use cases on real-world large-scale testbeds, e.g., Grid'5000 [5]. To the best of our knowledge, **E2C***lab* is the first platform to support the complete analysis cycle of an application on the Computing Continuum: *(i)* the configuration of the experimental environment; *(ii)* the mapping between the application parts and machines on the Edge, Fog and Cloud; *(iii)* the deployment and monitoring of the application on the infrastructure; and *(iv)* the automated execution and gathering of experiment results.

A **large scale experimental validation** on the Grid'5000 testbed [5] with Pl@ntNet [6], a real-life use case. Our framework allows optimizing the Pl@ntNet's performance based on the analysis of the parameter settings and correlation to processing time and resource usage.

VII. Preliminary Results

We illustrate [4] **E2C***lab* usage with a **real-life Smart Surveillance System** deployed on the Grid'5000 testbed, showing that our framework allows one to understand how the Cloud-centric and the hybrid Edge-Cloud processing approaches impact performance metrics such as latency and throughput. Besides, we are also validating **E2C***lab* with **Pl@ntNet**, another **real-life use case**. We demonstrate that **E2C***lab* guides on the optimization of the Pl@ntNet performance based on the analysis of the parameter settings and correlation to processing time and resource usage. Preliminary results show that Pl@ntNet's deployment configurations found by **E2C***lab* perform better than the current ones used in the production servers.

VIII. Next Research Steps

We will explore **scalable optimization** techniques that supports surrogate modeling optimization for large-scale multi-objective optimization problems. In this direction, we have an ongoing collaboration with Argonne National Laboratory, where we are using *DeepHyper* as a support to the optimization of application workflows. Furthermore, since **E2C***lab* supports **reproducible experiments**, we will explore and propose techniques for **runtime provenance** collection in large-scale and distributed experimental environments. The goal is to provide additional context that more accurately explains the experiment execution and results. This research direction is a collaboration with the Federal University of Rio de Janeiro, Brazil.

References

[1] ETP4HPC Strategic Research Agenda. [Online]. Available: https://www.etp4hpc.eu/sra.html

[2] S. Sadiq *et al.*, "Data Flow and Validation in Workflow Modelling," in *Proceedings of the 15th Australasian database conference-Volume 27*, 2004, pp. 207–214.

[3] S. Svorobej *et al.*, "Simulating Fog and Edge Computing Scenarios: An Overview and Research Challenges," *Future Internet*, vol. 11, no. 3, p. 55, 2019.

[4] D. Rosendo *et al.*, "E2clab: Exploring the computing continuum through repeatable, replicable and reproducible edge-to-cloud experiments," in *IEEE CLUSTER*. IEEE, 2020, pp. 176–186.

[5] R. Bolze *et al.*, "Grid'5000: A Large Scale And Highly Reconfigurable Experimental Grid Testbed," *IJHPCA*, vol. 20, no. 4, pp. 481–494, 2006.

[6] A. Joly *et al.*, "A look inside the pl@ntnet experience," *Multimedia Systems*, vol. 22, no. 6, pp. 751–766, 2016.