Report on the outcomes of a Short-Term Scientific Mission[1]

Action number: E-COST-GRANT-CA20111-95815235

Grantee name: Stefania DUMBRAVA

Details of the STSM

Title: **Automated Verification of a Conflict-Free Replicated Property Graph Data Structure**

Start and end date: **12/07/2022 – 21/07/2022**

Description of the work carried out during the STSM

Description of the activities carried out during the STSM. Any deviations from the initial working plan shall also be described in this section.

(500 words)

The initial order of the following STSM research and dissemination actions was switched, to accommodate the availability of the researchers and students interested in participating.

1. **July 12th - 15th:** The host, Mario Pereira, and I worked on implementing a verified library of conflict-free replicated data structures (CRDTs) for distributed graph processing. We realized the following:

- **State-of-the-art review on CRDTs**. We have analyzed surveys and tutorials of convergent and commutative replicated data types and articles on invariant-preserving applications for weakly consistent replicated databases.

- **Specifying a custom CRDTs for property graphs (PG-CRDT)**. Based on the gathered insights, we have identified a novel CRDT for property graphs. This **PG-CRDT** structure supports sharding and is composed of a node CRDT and an edge CRDT. Each CRDT combines a Two-Phase Set data type — comprising Grow-Only Sets for insertions and deletions — with a Map CRDT, for storing key/value pairs, attached to nodes and edges.

[1] This report is submitted by the grantee to the Action MC for approval and for claiming payment of the awarded grant. The Grant Awarding Coordinator coordinates the evaluation of this report on behalf of the Action MC and instructs the GH for payment of the Grant.

- **Building a formal library of verified components for modular PG-CRDTs in Why3**. CRDT are state- or operation-based and follow specific conflict resolution strategies. Hence, PG-CRDT can have variations, each requiring formal strong eventual consistency (SEC) guarantees. We have completed an initial Why3 library, which supports declarative specification, verification automation, and extraction of provably correct executable code. This contains the formal specification of key components, building up to state-based PG-CRDTs, with a last-writer wins resolution strategy.

- **State-of-the-craft review on implementing persistent mergeable data structures.** To comprehend the limitations, challenges, and open issues concerning CRDTs, Mario and I undertook research meetings with:

  - Carla Ferreira (NOVA), Kevin Porre, and Elisa Boix (VUB), the developers of the VeriFx system, on implementing and verifying CRDTs in Scala.
  - Nuno Preguiça (NOVA) on the implementation of the AntidoteDB and IPA systems and the types of properties maintainable across replicas. This exchange helped pinpoint feasible applications for our verified PG-CRDT.

- **State-of-the-craft review on using Why3 to formally verify program properties.** To understand the challenges of fully-automatizing the formal verification of PG-CRDT SEC properties, we have:

  - Experimented with different ways of specifying CRDTs in Why3, thus gaining knowledge of the good deductive verification practices to follow.
  - Discussed with Jorge Sousa Pinto (Universidade do Minho) about his Why3-do system for distributed system verification in Why3, without using proof assistants for inductive proofs. As most interesting graph properties are inductive, the meeting opened the perspective of leveraging Why3-do to facilitate their maintenance within PG-CRDTs.

2. **July 18th - July 21st: Research Planning Activities and Dissemination.**

- **July 18th - 19th:** Mario and I planned joint future research collaborations.
- **July 20th:** I have held a hybrid research seminar on "Graph Data Processing Techniques", at the NOVA laboratory, in which I outlined my recent research on graph shaped data processing and graph databases. This enriching experience allowed me to interact with laboratory members, from various areas, all of whom had an interest in graph-centric technologies.

- **July 21st:** I gave a hybrid introductory graduate class on graph databases (data model, storage, query languages, and applications), followed by an afternoon of exercises using the Neo4j graph database.

Description of the STSM main achievements and planned follow-up activities

Description and assessment of whether the STSM achieved its planned goals and expected outcomes, including specific contribution to Action objective and deliverables, or publications resulting from the STSM. Agreed plans for future follow-up collaborations shall also be described in this section.

(500 words)

The STSM achieved its planned goals of fostering collaboration between Mario and I on formally verifying property graph CRDTs.

As deliverables, we have drafted the specification of PG-CRDT structures and a formal library containing base building blocks for verifying distributed graph databases in Why3.

We have disseminated findings through a joint research seminar, a course and lab session, discussions with potential partners, as well as the two month supervision of a joint intern.

We plan to continue the formalization and complete the verification of SEC properties for PG-CRDTs. Short-term, we aim to disseminate findings in an article on formalizing PG-CRDTs in Why3, detailing the capabilities and limitations of SMT solvers for automatic verification. Mid-term, we plan on extending the work and illustrating how the extracted code can be used as an oracle to verify a real-world commercial graph database engine.

- **Completed Research Coordination Objectives.**

- Point 2 - “Promote the output of detailed, checkable proofs from automated theorem provers”. The topic of verifying PG-CRDT structures using Why3 and the integrated automated theorem provers was promoted through: discussions with researchers from NOVA LINCS, VUB, and Minho University, the *jointly supervised ENSIIE-NOVA internship of Aurelien Coureau (ENSIIE student) between the May 30th and July 29th of July,* and the submission of a joint second year master thesis topic for 2023.

- Point 3 - “Make techniques for program verification more effective and more accessible to all stakeholders”. This objective is a consequence of the fact that the verified PG-CRDT library is a use-case for the interoperability of Cameleer and Why3.

- **Completed Capacity Building Objectives.**

- Point 1 - “Bring together members of the different communities working on proofs in Europe”. The seminar facilitated discussions with NOVA researchers working in different formal methods approaches, e.g., behavioral type systems and information control flow.

- Point 2 - ”Act as a stakeholder platform in the field of formal proofs from its theoretical grounds to its industrial applications'. The topic of our seminar was a useful opportunity to interact and gain feedback from the local developers of AntidoteDB.

- Point 3 - “Create an excellent and inclusive network of researchers in Europe with lasting collaboration beyond the lifetime of the Action". The visit provided an occasion to plan further bi-lateral collaborations:, a French national project submission, another internship supervision, and further mobility and dissemination actions.

- Point 4 - “Ease access to formal verification techniques in education and other areas of science”: The use-case on distributed graph processing data structures and algorithms will be used for pedagogical purposes in courses taught by the partners.

- Point 5 - “Actively support young researchers, the under-represented gender, and teams from regions with less capacity”. This mission provided a useful opportunity to advance on an interdisciplinary topic and plan common projects. A joint internship on the topic was completed, another one was submitted, the submission of a national level research grant and of a joint publication was planned, as well as the upcoming visit of Mario at the SAMOVAR laboratory in Evry, hosted by the grantee.