DIAS: Data-Intensive Applications and Systems Laboratory
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Building BC, Station 14
CH-1015 Lausanne
URL: http://dias.epfl.ch/

Accelerating Complex Imperative Workflows Using Speculations

Keywords: speculative execution, approximate query processing, imperative programs

Problem: An increasing number of companies and organizations use complex workflows, either custom or based on data mining algorithms, to extract insights from large amounts of data. Developers often implement such workflows as imperative programs over distributed computing frameworks, such as Spark. As a result, multiple analytical tasks are interconnected through control flow and loop constructs, as well as producer-consumer dependencies. The dependencies, however, limit parallelization opportunities. While each individual task can be run in a data-parallel way, the algorithm cannot predict which task will run next or what its input will be. Hence, algorithms cannot exploit task-parallelism in complex workflows despite the multi-task execution environment.

Project: The goal of this project is to parallelize complex workflows using speculations. Recent work [1] parallelizes inter-dependent sub-queries in complex analytical queries using speculations. However, analytical query languages lack the expressive power of workflows, which include programming language constructs, such as loops and control flow. Hence, the speculative framework needs to be extended to cover complex imperative programs. The student is expected to produce a generalized formulation that can be applied on a wide class of workflows.

Plan:

1. Perform a case study on well-known data mining and data analysis algorithms, such as k-Means.
2. Formulate speculation in imperative programs and propose a unified framework for complex workflows and complex queries.
3. Implement optimization rules in a prototype API that uses an established distributed DBMS as a back-end and demonstrate the effect of speculative execution.

References

1. P. Sioulas, V. Sanca, I. Mytilinis, A. Ailamaki. “Accelerating complex analytics using speculation”. Conference on Innovative Data Systems Research (CIDR) 2021

Supervisor: Prof. Anastasia Ailamaki, anastasia.ailamaki@epfl.ch
Responsible collaborator(s): Viktor Sanca, viktor.sanca@epfl.ch

Duration: 1 semester