Real-time Critical Path Identification for Optimizing System Performance

Master Thesis Proposal

Problem Description

Parallel data processing can significantly improve the latency and throughput of a computation but requires careful decisions on the implementation, including the synchronisation of individual workers (e.g., threads in a multi-core environment, nodes in a distributed system, etc.). Poor design of parallel data processing systems results in performance bottlenecks that are surprisingly hard to spot. In such cases, the optimization of a computation requires knowledge about its 'internals'; for instance, the blocking dependencies amongst the various computation steps. Critical path [1] is a valuable tool in this direction: the critical path of a computation is defined as the set of computation steps for which a differential increase in the step execution time results in the same differential increase in the end-to-end computation time. Although the identification of the critical path of a sequential computation is trivial, the task gains in complexity with parallel jobs and complex (e.g., cyclic) dataflows. In addition, the critical path of a computation is not a static feature; it depends on the particular workload and may also change dynamically as the computation proceeds.

The proposed master thesis will focus on real-time critical path identification for parallel processing jobs with complex dataflows. The analysis will be based on Timely Dataflow [2], a novel data-parallel streaming engine that supports incremental computations and arbitrary (even iterative) dataflows. The goal of the thesis is fourfold. First, we will investigate the usefulness of critical path in optimizing parallel jobs with non-trivial inter-dependencies, possibly by rethinking the current definition of critical path in the literature. Second, we will study lightweight instrumentation techniques to collect and process data from Timely Dataflow in real time. At a next step, we will specify a general schema of the data one needs to obtain in order to apply our critical path approach to any data processing system (e.g., progress messages, notifications, task scheduling information, etc.). Finally, we will compare our approach with existing profiling- and instrumentation-based approaches for performance optimization, e.g., [3] and [4]. The outcome of the project will also include an interactive visualization front-end where the user can see the critical path of the computation in real time.

If you are interested in this project and want to discuss further, please contact John Liagouris (liagos@inf.ethz.ch) and Desislava Dimitrova (desislava.dimitrova@inf.ethz.ch). The proposed Master Thesis will be supervised by Prof. Timothy Roscoe.