Performance Characteristics of Modern Distributed Stream Processing Engines

Keywords: stream processing, workload characterization, micro-architectural analysis, benchmarking

Problem: Nowadays, we observe an ever-increasing amount of data-in-motion that appear in the form of streams. A stream is a sequence of data that is not persisted anywhere but must be processed on-the-fly. Thus, modern stream processing engines (SPEs) need to process huge data volumes under stringent latency constraints. For this purpose, many systems have been developed and all follow the same principle: they execute processing pipelines on a distributed architecture and apply a scale-out strategy in order to deal with the volume and velocity of the incoming streams. However, it is still unclear how existing systems differ to each other in terms of hardware resources utilization. Which SPEs better exploit the characteristics of modern hardware? What is the main bottleneck for each SPE? What are the main defining characteristics for the performance of a stream processing system?

Project: In this project, we will try to answer the above questions by investigating the micro-architectural behavior of four widely used stream processing frameworks, namely of Spark [1], Flink [2], Storm [3] and Kafka Streams [4]. To achieve that, we will follow a similar methodology to the one used in [5]. First, we are going to benchmark some typical streaming applications (e.g., Yahoo! Streaming Benchmark [6]) by measuring high-level performance metrics such as throughput and end-to-end latency. Then, in order to understand how the available hardware affects performance, we will try to associate the high-level metrics with the hardware utilization in terms of: i) CPU cycles, ii) memory bandwidth and iii) network bandwidth.

Plan:
1. Setup Spark in a cluster of machines.
2. Profile performance and hardware utilization for various configurations of a single node deployment.
3. Investigate how performance characteristics vary when we scale out to multiple nodes.
4. Repeat 1 – 3 for Flink, Storm and Kafka Streams.

Supervisor: Prof. Anastasia Ailamaki, anastasia.ailamaki@epfl.ch

Responsible collaborator(s): Dr. Ioannis Mytilinis, ioannis.mytilinis@epfl.ch,
Utku Sirin, utku.sirin@epfl.ch

Duration: 1 semester
References