Contribution ID: **10** Type: **Hands-On Tutorial**

Thrill: High-Performance Algorithmic Distributed Batch Data Processing with C++

Thursday, August 29, 2019 1:15 PM (4h 45m)

In this tutorial we first present our new distributed Big Data processing framework called Thrill [1,2]. It is a C++ framework consisting of a set of basic scalable algorithmic primitives like mapping, reducing, sorting, merging, joining, and additional MPI-like collectives. This set of primitives goes beyond traditional Map/Reduce and can be combined into larger more complex algorithms, such as WordCount, PageRank, and suffix sorting. These complex algorithms can then be run on very large inputs using a distributed computing cluster with external memory. Among the main design goals of Thrill is to lose very little performance when composing primitives such that small data types are well supported. Thrill thus raises the questions of a) how to design algorithms using the scalable primitives, b) whether additional primitives should be added, and c) if one can improve the existing ones using new ideas to reduce communication volume and latency. Our performance evaluations show that Thrill is always faster than Apache
Spark and Apache Flink on a set of five microkernel benchmarks.

After introducing the audience to Thrill we continue by guiding participants through the initial steps of download and compiling the software package. After a short sight-seeing tour of the framework's internal source code structure, the tutorial group will together go through the steps to develop and run a simple K-means clustering implementation. As an intermezzo, we will present more details on the implementations of Sort and Reduce in Thrill. In the last part, participant are then given a set of free exercises to choose from and to work on together.

Pre-requisites

- Participants need a computer to follow the hands-on parts of the tutorial.
- Knowledge of C++ is required for enjoyment of this tutorial.
- Thrill's primary platform is Linux or MacOS, but Windows may also work.

References

[1] Timo Bingmann, Michael Axtmann, Emanuel Jöbstl, Sebastian Lamm, Huyen Chau Nguyen, Alexander Noe, Sebastian Schlag, Matthias Stumpp, Tobias Sturm, and Peter Sanders. "Thrill: High-Performance Algorithmic Distributed Batch Data Processing with C++". In: IEEE International Conference on Big Data. preprint arXiv:1608.05634, pages 172–183. IEEE.
Dec. 2016.

[2] http://project-thrill.org

Presenter: BINGMANN, Timo (KIT)

Session Classification: Tutorials