Using Thrill to Process Scientific Data on HPC

Mariia Karabin[1], Xinyu Chen[2],Supreeth Madapur Suresh[3], Ivo Jimenez[4], Li-Ta Lo[5], Pascal Grosset[5]

1.Clemson University, 2.University of New Mexico, 3.University of Wyoming, 4.University of California, Santa Cruz, 5.Los Alamos National Laboratory

Abstract

With ongoing improvement of computational power and memory capacity, the volume of scientific data keeps growing. To gain insights from vast amounts of data, scientists are starting to look at Big Data processing and analytics tools such as Apache Spark.

In this poster, we explore Thrill, a framework for big data distributed computation to help scientists at the Los Alamos National Laboratory (LANL) post-process and analyze data from plasma physics and molecular dynamics simulations.

Using Thrill, we implemented analytics operations with less programming efforts than handcrafted data processing programs and obtained results which were verified by scientists at LANL.

Data Flow Example for Time Series

To introduce the data flow we present here an example for time series, which was part of our data analysis for the Accelerated Molecular Dynamics (AMD). We show the basic Thrill operations that we used as well as final result.

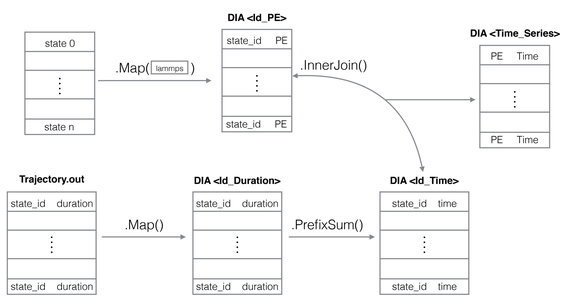

Figure: Data flow example for time series.

Data Analysis for Accelerated Molecular Dynamics

Data analysis for the accelerated molecular dynamics (AMD) included clustering (using k-means algorithm), time series and other data analysis such as sorting and finding minimum/maximum of the total potential energy per state.

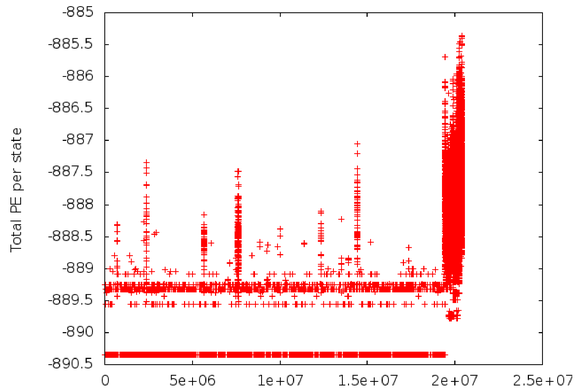

Figure: Time Series for the Accelerated Molecular Dynamics, 62265 states.

Clustering the states with k-means algorithm based on the CNA(Common Neighbor Analysis) structures calculated from the LAMMPS function call.

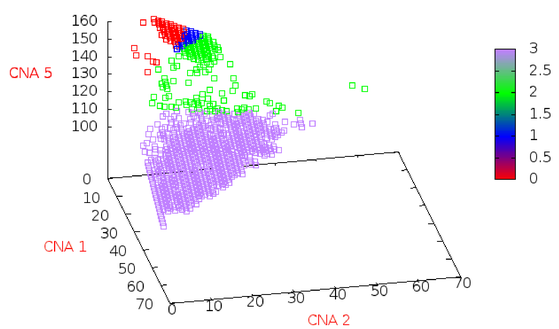

Figure: Clustering analysis for the Accelerated Molecular Dynamics using k-means algorithm.

Data Analysis for Vector Particle in Cell (VPIC)

Data Analysis for the Physicists from the VPIC group included trajectory analysis for which we present here graphs to represent some of the results.

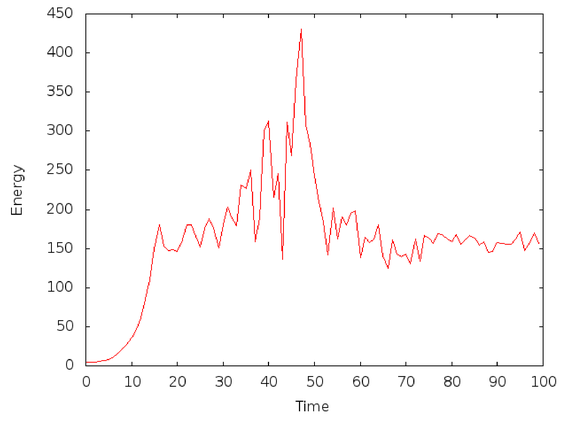

Figure: Energy versus time for the VPIC data set.

Based on the trajectory analysis scientists are interested in getting the coordinates for the particle with the highest energy to understand what causes the acceleration and use this information for further applications.

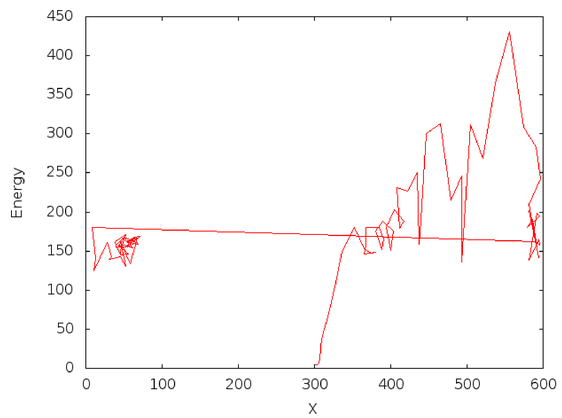

Figure: Trajectory for the particle with the highest energy from the VPIC data set.

Scalability tests

- The scalability tests show that applications have strong scalability to some extent.

Scalability Tests For AMD and VPIC

The timing measurements showed strong scalability to 128 MPI ranks for querying AMD potential energy time series. We did not observe the strong scalability for VPIC due to our implementation.

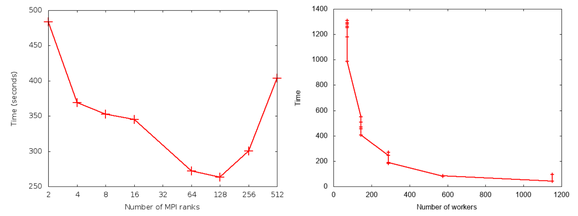

Figure: Thrill implementations show strong scalability to some extent for both (a) AMD and (b) VPIC data analysis.

Conclusion

We explored the Thrill library to process and analyze simulation data and showed the usefulness of the declarative language paradigm in solving big data scientific problems. This helped scientists to deal with many files in parallel and get the data analysis done in a more efficient way. The results were verified and displayed strong scalability to some extent.

Acknowledgements

We want to thank our collaborators from Los Alamos National Laboratory: Dr. Danny Perez, Dr. Fan Guo and Dr. Xiaocan Li (T division); as well as Data Science at Scale School and ISTI for funding.

[1] Bingmann, Timo, et al. "Thrill: High-performance algorithmic distributed batch data processing with C++." arXiv preprint arXiv:1608.05634 (2016).

LA-UR-18-27362