Performance evaluation of a clustering approach based on thermophysical properties by using multiple platforms

Kou Murakami, Kazuhiko Komatsu, Masayuki Sato, Hiroaki Kobayashi
Tohoku University
Sendai, Miyagi, Japan
{kou.murakami.r1@dc.,masa@,komatsu@,koba@}tohoku.ac.jp

1 INTRODUCTION

Recently, the search and development of new materials have been a huge amount of time because they are developed through trials and errors using knowledge and empirical rules. However, the search and development of new materials are accelerated with the advent of material informatics (MI). In this field, the search and development of new substances have been energetically performed by exploiting machine learning (ML) algorithms.

At the same time, ML is recently expected to be accelerated by using accelerators instead of general-purpose CPUs. Therefore, there is a possibility to improve the performance of the search and development of new materials with ML by accelerators.

In this poster, we explore the possibility of enhancing MI using accelerators. For this purpose, we focus on a classification method of liquid fluids based on thermophysical properties [1]. The execution time of this method increases as the number of substances increases. We evaluate the performance of a part of this method when using large data sets. In the experiment, we measure the execution time of the liquid fluid classification code using CPU and two accelerators, GPU and vector processor, to examine the possibility of acceleration.

2 CLASSIFICATION METHOD OF LIQUID FLUIDS

A classification method of liquid fluids [1] visualizes thermophysical properties using clustering approaches. It is divided into two parts; the first part learns with SOM (Self-organizing map) and the second part visualizes the learning results by clustering with k-means. The original data set consists of 98 substances, each of which has 9 thermophysical properties. This is a 9-dimension data set with data size of 98. By SOM training, a 9-dimensional input vector is reduced to 2-dimensions. The visualization part classifies materials into several groups with similar properties by clustering the results of a 2-dimensional map.

3 EVALUATION

To investigate the performance of this method when changing data sets, we use large data sets from 1024 to 16384. For the evaluation by the multiple platforms, we prepare Intel Xeon Skylake (CPU), NVIDIA Tesla V100 (GPU), and NEC Vector Engine Type 10B (VE). CPU has 12 cores and its peak computing performance is 998.4 Gflop/s in the case of double-precision floating-point operations. GPU has 2560 cores and its peak computing performance is 7.8 Tflop/s in the case of double-precision floating-point. In order to exploit the high potential of GPU, a large number of cores need to be used. VE has 8 cores and its peak computing performance is 2.15 Tflop/s in the case of double-precision floating-point. VE has a high vector operation performance and a high vectorization ratio are required to achieve high performance. The ML frameworks used for these platforms are scikit-learn for CPU, NVIDIA RAPIDS for GPU, and NEC Frovedis for VE. These frameworks have the similar interfaces of scikit-learn. Note that we evaluate only the second part of the proposed method, which is the k-means clustering part for visualization. This is because only scikit-learn supports the key component of the first part, SOM, and the other two frameworks do not support it.

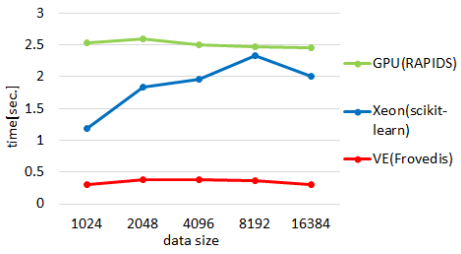

Figure 1: Performance evaluation of the visualization part on multiple platforms.

Figure 1 shows the execution time using the platforms. Figure 1 indicates that VE achieves the highest performance among these processors. The GPU did not perform as well as we expected. CPU has a trend to increase execution time as the data size increases, whereas those of VE and GPU do not change. To investigate this reason, the profile information in CPU and VE are examined. For all the data sizes, the vectorization ratio in Xeon is 10%, whereas that in VE is 50%. Since VE has a higher vectorization ratio, it is expected that VE can suppress the increase in the execution time even if the data size increases compared with CPU.

4 CONCLUSIONS

We evaluate the clustering approach based on thermophysical properties by using multiple platforms. The evaluation results show that the execution time almost unchanged even when the data size increased on GPU and VE. This indicates that accelerators such as GPU and VE are effective in clustering large data sets. For future work, we plan to evaluate the learning part in the classification method on multiple platforms.

REFERENCES

[1] Gota Kikugawa, Yuta Nishimura, Koji Shimoyama, Taku Ohara, Tomonaga Okabe, and Fumio S. Ohuchi. Data analysis of multi-dimensional thermophysical properties of liquid substances based on clustering approach of machine learning. *Chemical Physics Letters*, Vol. 728, pp. 109–114, August 2019.