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

TOHOKU UNIVERSITY

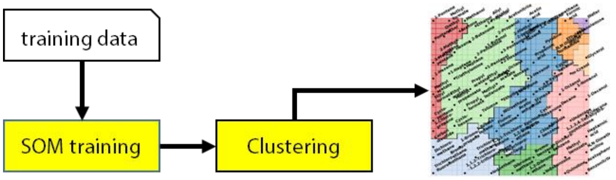
Kou Murakami, Kazuhiko Komatsu, Masayuki Sato, Hiroaki Kobayashi

Motivation

- **Conventional liquid fluid material development**
  - It is necessary to consume a lot of time in developing new materials by experimental trials and errors.
  - Researchers rely on intuition to develop new materials.
- **New Approach: Material Informatics (MI)**
  - New materials are discovered by informatics approaches.
    - Candidates by simulation and experiment are analyzed with machine learning (ML).
- **Problem on MI: Computing cost**
  - A huge amount of computing resource is requested to analyze the properties of a large number of materials.
  - The computation time increases by the growth of the number of material combinations.

Evaluate clustering for MI using various accelerators
- GPU, vector processor, and so on

A workflow of clustering liquid fluid materials[1]



Visualized result

- This workflow is implemented by SOMPY.
  - SOMPY is implemented by scikit-learn.
- The materials are classified based on their thermophysical properties.
  - The visualized result makes understanding material properties easy.

- This workflow consists of the SOM training, and the k-means clustering and visualization parts.

1. **SOM training**
   - A 2-D map is created (input vector is 9-D).
     1. Codebook is initialized by principal component analysis (PCA).
     2. The best match unit (BMU) is searched by the k-nearest neighbor algorithm.
     3. BMU and its neighbors are updated.
     4. A 2-D map is output.
2. **K-means clustering and visualization parts**
   - Materials are classified based on similar properties.
     1. A SOM training result is clustered by the k-means algorithm.
     2. The clustering result is visualized.

[1] G. Kikugawa; et al. 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.

Performance evaluation

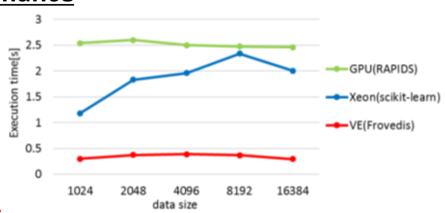
Platforms

Processor	Xeon Gold 6126	Tesla V100	Vector Engine 10B
FLOPS (Double)	998.4 GFLOPS	7.8 TFLOPS	2.15 TFLOPS
Memory bandwidth	128 GB/s	900 GB/s	1.2 TB/s
# of cores	12	2560	8
ML library	scikit-learn	RAPIDS	Frovedis

Experimental environment

- The k-means clustering algorithm on multiple platforms is evaluated.
  - The execution time does not include that of the visualization part because this part is negligible.
- Large data sets are created in the experiment.
  - The number of material: 1024~16384
  - The number of dimension: 9

Performance



- VE achieves the shortest execution time among these processors in large codebook sizes.
  - The demand B/F is high.
  - The execution time of GPU has a large overhead except for the k-means calculation.
- The execution times of GPU and VE do not change as the data size increases, whereas that of Xeon increases.
  - GPU and VE do not make the most of their processing and vector processing performance within the range of this data sizes.

Conclusions and future work

- **Conclusions**
  - The k-means clustering and visualization parts are evaluated by multiple platforms.
  - VE achieves the highest performance among these processors.
- **Future work**
  - The SOM training part in the classification method is implemented and evaluated on multiple platforms.
  - The workflow is accelerated by offloading k-NN considering data set size to the suitable processor.

Acknowledgements

The authors would like to thank Prof. Gota Kikugawa for providing the code of clustering liquid fluid materials. This work was supported in part by MEXT as "Next Generation High-Performance Computing Infrastructures and Applications R&D Program," entitled "R&D of A Quantum-Annealing-Assisted Next Generation HPC Infrastructure and its Applications" and Grants-in-Aid for Scientific Research(A) #19H01095.

(URL) https://www.cal.is.tohoku.ac.jp
(E-mail) kou.murakami.r1@tohoku.ac.jp