2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS) | 978-1-6654-7177-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICDCS54860.2022.00127

Performance Modeling of Weather Forecast Machine Learning for Efficient HPC

Karthick Panner Selvam
SEDAN, SnT
University of Luxembourg
Luxembourg
karthick.pannerselvam@uni.lu

Mats Brorsson
SEDAN, SnT
University of Luxembourg
Luxembourg
mats.brorsson@uni.lu

***Abstract*—High-performance computing is a prime area for many applications. Majorly, weather and climate forecast applications use the HPC system because it needs to give a good result with low latency. In recent years machine learning and deep learning models have been widely used to forecast the weather. However, to the best of the author's knowledge, many applications do not effectively utilise the HPC system for training, testing, validation, and inference of weather data. Our experiment is to conduct performance modeling and benchmark analysis of weather and climate forecast machine learning models and determine the characteristics between the application, model and the underlying HPC system. Our results will help the researchers improvise and optimise the weather forecast system and use the HPC system efficiently.**

***Index Terms*—High-Performance Computing, Deep Learning, Weather Forecast, Distributed Computing, Performance modeling, Benchmark Analysis.**

I. Introduction

Weather and climate science play a vital role in all living forms, mainly human health, protection and economy. Sudden weather changes are causing various problems to human beings. Predicting the weather and climate is one of the ways to minimise hazardous impacts. Unfortunately, even fine weather and climate prediction models are time-consuming and unreliable, and cannot forecast for more than a week. Recently machine learning and deep learning models are being used to augment existing numerical simulation models. MAELSTROM[1] is a large-scale EuroHPC[2] project with an aim to improve on the use of machine learning in weather and climate modelling in terms of: i) applications amenable for ML augmentation, ii) workflow and iii) machine architectures suitable for ML-augmented Workload Characterization modelling [1].

MAELSTROM has six different machine learning and deep learning applications, as shown in table I. Most of the applications are for faster weather prediction using neural networks, e.g. downscaling temperature, weather forecasts to support energy production, and forecast post-processing for better local weather predictions [1]. Application two is still in work, so it is not mentioned in Table I. Most of the applications have a large amount of data for training, and testing.

The estimated average of 10 TB data will be gathered in future of each application, so the complexity of training, testing and running this application will increase. The core problem is making all six applications fully utilise the computing hardware very effectively and provide results with low latency. Our objectives within the MAELSTROM project, are to develop a thorough understanding of the MAELSTROM applications' characteristics on modern hardware and to develop performance prediction models that allow us to search the design space of suitable future architectures with the need to build them. The approach we take is to perform a workload characterization of all applications in different HPC architectures, and with different machine learning software frameworks. This should provide us with a deep understanding of the interplay between application and hardware architecture. This result will help choose a better hardware configuration and software framework to effectively utilise the HPC system for weather and climate prediction.

II. Methods

For this workload characterization, our target is to use a combination of multiple HPC architectures using a combination of CPU, GPU and (Artificial Intelligence) AI accelerators like GraphCore[3] IPU[4] [2] [3]. IPU is the Intelligence Processing Unit. It contains 1,216 processors called tiles. Each tile has a local memory of 256 KiB, a computing core, and six threads. IPU contains a total of 7,296 threads to perform multithread operations. IPU is a true MIMD (Multiple Instructions, Multiple Data) architectures. For this advantageous, we included IPU in our benchmark analysis [4]. We have access to various HPC systems like Intel, AMD, and NVIDIA to run this experiment. For the initial stage, we plan to use the below HPC system.

- Aion[5] is a supercomputer constructed by AMD that has 318 computing nodes, 40704 compute cores, and 81408 GB RAM, with a prime speed of around 1,70 PFLOP/s.

This work has been funded by EuroHPC JU under contract number 955513 and by the Luxembourg National Research Fund (FNR) under contract number 15092355.

[1] https://www.maelstrom-eurohpc.eu/

[2] https://eurohpc-ju.europa.eu/

[3] https://www.graphcore.ai/

[4] https://www.graphcore.ai/products/ipu

[5] https://hpc-docs.uni.lu/systems/aion/

2575-8411/22/$31.00 ©2022 IEEE
DOI 10.1109/ICDCS54860.2022.00127

TABLE I
MAELSTROM Application Lists

Application	Resolution	Grid Size	Data Size	Data format	Arithmetic Precision
App 1: Postprocessing	1 km	1796×2321	$\sim$ 5 TB	NetCDF	FP32
App 3: Radiation	40 km	137 vertical levels	$\sim$ 2 TB	NetCDF/TFRecords	FP32, FP16
App 4: ENS10	0.5°	720x361x11x11	$\sim$ 2.6 TB	GRIB/NetCDF	FP32
App 5: Downscaling	0.1°	96x128	$\sim$ 300 MB	NetCDF	FP32, FP16 mixed
App 6: Power production	0.1°	351x55110 vertical levels	$\sim$ 1 TB	NetCDF	FP64

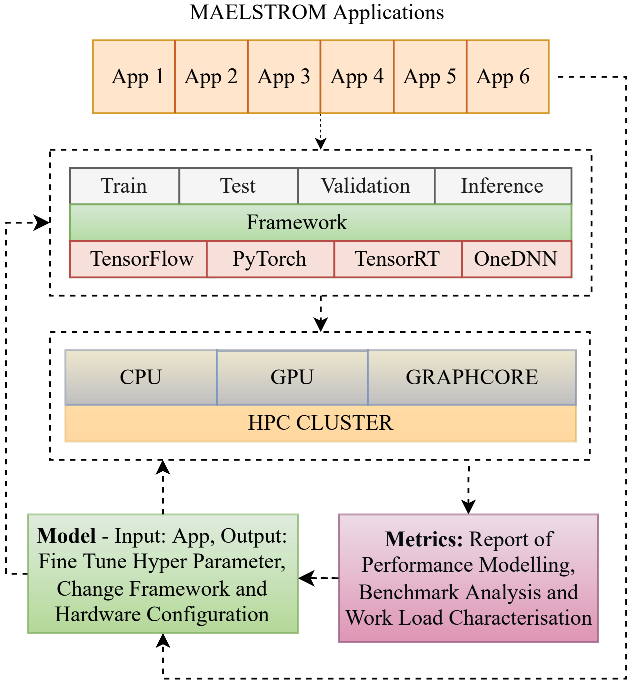
MAELSTROM Applications
App 1
App 2
App 3
App 4
App 5
App 6
Train
Test
Validation
Inference
Framework
TensorFlow
PyTorch
TensorRT
OneDNN
CPU
GPU
GRAPHCORE
HPC CLUSTER
Model - Input: App, Output: Fine Tune Hyper Parameter, Change Framework and Hardware Configuration
Metrics: Report of Performance Modelling, Benchmark Analysis and Work Load Characterisation

Fig. 1. Maelstrom Performance modeling Architecture

- Iris[6] is an Intel supercomputer with 196 computing nodes, 5824 compute cores, and 52224 GB RAM, with a peak speed of around 1,072 PFLOP/s. In addition, Iris has 96 NVIDIA V100 GPU-AI accelerators, allowing GPU-enabled applications and Deep Learning workflows to run faster. Each node has a memory capacity of 3 TB RAM.

In future, we will do performance modeling of the applications with more HPC clusters and GraphCore systems. All applications will run in the HPC system with their default deep learning framework in the first stage. Then we calculate the training epoch time and inference, application and individual component execution time like loss functions and matrix multiplications. And then collect threading and microarchitecture utilisation. After that, we continue the process for different hardware architectures like CPU, GPU and IPU, and then calculate the number of CPU, GPU and nodes used, performance per watt, hardware heat, overall memory consumption, and nodes usage and VRAM consumption. Finally, we store all the metrics in the database.

In the second stage, we will continue the process with different deep learning frameworks like TensorFlow[7] PyTorch[8], TensorRT[9], and OneDNN[10] as shown in Fig. 1. For example, App 1 default framework is TensorFlow, and we have already calculated all metrics in different architectures in the HPC system. We will change the App 1 framework to PyTorch and calculate all the metrics. Likewise, we calculate metrics on the combination of all deep learning frameworks with different HPC systems. We will continue this process for all applications. Then we will create a model using these metrics. Our model will help researchers fine-tune the application's hyperparameters and change the frameworks and hardware configurations to effectively utilise the HPC system and reduce latency.

III. Conclusion

Our performance modeling and benchmark analysis of weather and climate forecast ML and DL model results will provide complete characteristics between weather forecast application and HPC system. To the best of the author's knowledge, our result will help the researchers to identify better hardware configurations and understand application characteristics for weather and climate prediction application to effectively utilise the HPC system. Furthermore, our results will help the researchers to understand the characteristics between underlying hardware system and ML models from different domains for effective High-performance computing.

References

[1] G. Bing, L. Michael, D. Peter, C. Matthew, N. Thomas, A. Markus, and B.-N. Tal, "Report on a survey of MAELSTROM applications and ML tools and architectures. Deliverable 2.1." MEAELSTROM EuroHPC project., Tech. Rep., 2021. [Online]. Available: https://www.maelstrom-eurohpc.eu/deliverables

[2] M. Brorsson, "Roadmap analysis of technologies relevant for ML solutions in W&C. Deliverable 3.2." MEAELSTROM EuroHPC project., Tech. Rep., 2021. [Online]. Available: https://www.maelstrom-eurohpc.eu/deliverables

[3] C. Zhang, F. Zhang, X. Guo, B. He, X. Zhang, and X. Du, "iMLBench: A Machine Learning Benchmark Suite for CPU-GPU Integrated Architectures," *IEEE Transactions on Parallel and Distributed Systems*, pp. 1–1, 2020.

[4] Z. Jia, B. Tillman, M. Maggioni, and D. P. Scarpazza, "Dissecting the graphcore ipu architecture via microbenchmarking," 2019. [Online]. Available: https://arxiv.org/abs/1912.03413

[6] https://hpc-docs.uni.lu/systems/iris/

[7] https://www.tensorflow.org/

[8] https://pytorch.org/

[9] https://developer.nvidia.com/tensorrt

[10] https://github.com/oneapi-src/oneDNN