A Computing Platform Supporting AI

ISHIZAKA Kazuhisa, ARAKI Takuya, INOUE Hiroaki

Abstract

AI needs an extremely high computing performance due to an increase in data scale and complications in algorithms. Consequently it has become critical to use hardware accelerators arranged for specific purposes. Considering the difficulty for individuals to develop AI that covers a very wide range, the software for supporting the accelerators, called the framework software, has also become important. This paper is intended to introduce Flovedis, a framework for statistical machine learning using supporting accelerators developed by NEC as well as the Vector Engine, an accelerator supporting both statistical machine learning and deep learning.

Keywords

AI, platform, framework, accelerator, Frovedis, Vector Engine, SX-Aurora TSUBASA

1. Introduction

The AI uses various algorithms such as statistical machine learning, deep learning, reinforcement learning, combinational optimization and image analysis. As AI takes a very long time to learn the huge amount of data and also because of complications in the algorithms and the increase in the data scale, the required computations are tending to grow year on year[1)]. Meanwhile, the recent slowdown in the impact of the Moore's law claiming that computer chip performance would roughly double every 18 months, proficiency in the use of specific architectures is increasing in importance. Processing-specific architectures have already been released several times in the past but they were driven away by the CPUs, which are universal processors, as indicated under the Moore's law. Nevertheless, as the restrictions of the power consumption of chips caused the operating frequencies to reach a limit, the processing performance of the single CPU thread peaked at around 2005. Though this situation has advanced the multi-core design, an increase in the speed is restricted by the parts to be executed sequentially in the program, so improvement of performance is no longer expected, even if the number of cores is increased further (Amdahl's law). As a result, the recent trend is to use various accelerators (hardware) such as the GPU, vector processor, deep learning chip and FPGA in appropriate places, considering the characteristics of AI processing.

From the viewpoint of implementing social solutions that can deal with future issues of increased complexity, one of the issues for the AI platforms is the need for a huge input of precise man-hour deployment. The reason for this issue is the advancement of specific-type architectures as described above. It is necessary to understand completely the AI processing required for implementing this solution and to select/combine the right hardware for the processing in order to improve performance using specific-type architectures. This means that there will be increasingly precise man-hour deployment following increases in these types of accelerators.

What is important for dealing with the massive increase in man-hours is a software framework for AI (hereinafter "the framework"). The framework is the software that provides the components used to implement the various AI algorithms described above, which

are called the libraries. In many cases, the algorithms are implemented by combining libraries. It is because the framework provides the libraries matching the accelerators in advance that the AI developer can easily develop the AI optimized for the accelerators.

The following sections describe the accelerators and frameworks.

2. Accelerators

As described in the above, AI uses accelerators incorporating processors selected by emphasizing the performance rather than the universality. This is in addition to the CPU that is widespread and compatible with various kinds of software. Before introducing the accelerators, let us consider the techniques for improving processor performance. Recently, processors are improving performance by making use of the vector processing that refers to the processing for applying simultaneous and parallel computations to multiple items of different data. This is executed by incorporating and simultaneously running multiple computing units in a processor. It can improve performance by increasing the number of computing units and heightens their rate of utilization. Many of the libraries for AI are capable of vector processing.

On the other hand, processing that is incapable of running multiple computing units simultaneously is called the scalar processing. The performance of scalar processing cannot be improved by increasing the computing units. However, it is important to increase the operating speeds of computing units and to use high-function computing units. The required costs are higher than those for improving the vector processing performance. The frameworks for AI frequently use scalar processing outside of the libraries, for example in the context of library execution management.

It can be regarded that the processor putting importance on the performance of scalar processing is the CPU and the processors emphasizing the performance of vector processing are the accelerators. The CPU also introduces the vector processing but the accelerators achieve several times higher vector performance by incorporating larger numbers of simple computing units and by expanding the memory bandwidths to improve the computation and memory performances required for vector processing. With the design of AI, the library processing is executed by the accelerators and other processing by the CPU. The memory performance represents the capacity of the data supply to the computing units. Its insufficiency produces idle computing units, which means that it is an important performance indicator of processors.

The accelerators can be divided into the universal type and the deep learning-specific type. The universal type is not as universal as the CPU, but the incorporated programmable processor core makes it possible to execute various processing operations by rewriting the program. Examples of such accelerators is the CPU emphasizing the number of computing units and the Vector Engine emphasizing the memory performance. As it is difficult to achieve the highest levels in both the computing performance and the memory performance, the allocation of a balance between them defines the properties of accelerators.

With the deep learning used in image recognition, etc., the processing known as convolution becomes a bottleneck. Some of the recently launched accelerators incorporate the convolution-specific computing units in order to increase the number of computing units and to thereby improve the performance. These accelerators are categorized as deep learning-specific type accelerators and one of their examples is the TPU. Certain GPUs also incorporate convolution-specific computing units so that they can manifest the deep learning-type properties.

3. Frameworks

The framework performs the library execution management, memory management and accelerator management, in addition to the provision of the libraries for AI. It also performs optimization for increasing the processing speed, such as optimization by linking multiple libraries and optimization of inter-server communications. The visualization of the learning process and the profiling of performance are also among its roles. In this way, the framework has become indispensable for the development of AI by providing many functions for it. Further technological development of the framework is actively underway.

The framework has the characteristics that are useful from various viewpoints.

- Types of supported libraries
- Design of programming interfaces
- Compatible programming languages
- Types of supported accelerators
- Compatibility with distributed processing
- Presence of vendor support

The framework is selected according to the properties of the AI algorithms to be developed and the skills/experience of the development team. If the framework does not provide a library as necessary the user must implement one. The extendibility in such cases is therefore very important.

Among these characteristics, particularly important

are the "types of supported libraries" and the "types of supported accelerators". Whether or not the framework is equipped with sufficient libraries for the algorithms to be developed affects the development efficiency considerably. From the viewpoint of the types of libraries, the framework can be classified as the statistical machine learning emphasis type and the deep learning emphasis type. The recent attention to the deep learning in the field of image classification has led to the launch of several deep learning emphasis-type frameworks, including TensorFlow, Keras, PyTorch, Caffe, Theano MXNet and RAPID machine learning. On the other hand, the frameworks putting emphasis on the statistical machine learning include Frovedis, Spart (machine learning lb) and Scikit-learn.

From the viewpoint of processing performance, the accelerators supported by a framework are important. There are multiple accelerators for AI as described in section 2. Since the performance and programming characteristics of the accelerators differ between each other, it is difficult to be compatible with all of the accelerators. The technological development is therefore conducted to perform the execution management and profiling in an integrated manner and to achieve optimization for efficient use of the accelerators.

4. NEC's Approach

NEC believes that increasing the speed of AI for statistical machine learning as well as deep learning will be increasingly important in the future. Below we describe the accelerator under development by NEC, called the Vector Engine, and the statistical machine learning framework called Frovedia.

4.1 Accelerator Vector Engine

Vector Engine is an accelerator that inherits the technology of the SX series of vector-type supercomputers. Its main feature as hardware is the memory bandwidth at the world's top level at present (1.2 TB/s), and its main features as software are the programmability of general languages such as C/C++ and the compiler with a powerful auto vectorization capability. With these features that are important for library optimization, it can be regarded as the optimum accelerator for the developers of framework and the engineering for developing most advanced algorithms.

The SX-Aurora TSUBASA series are the servers incorporating Vector Engine. A wide range of products including the tower type, rack type and large type make the series applicable to various usage cases. The **Photo** shows an external view of Vector Engine.

Photo External view of Vector engine.

4.2 Statistical machine learning framework Frovedis

Frovedis is a high-performance framework for use in data analysis. It supports the CPU and Vector Engine and features compatibility with the distributed processing. It is consequently capable of fast execution of even the large-scale processing that needs multiple servers. It also features a rich range of libraries for statistical machine learning. The programming interfaces are compatible with programming languages Python and Scala as well as C/C++, and the APIs for Python's Scikit-learn machine learning library and Spark MLlib are provided. This means that any software based on them can be used with minor modifications of programs.

Although statistical machine learning libraries contain many memory bandwidth-dependent processing operations, Frovedis supports Vector Engine with a high memory bandwidth so it can execute such libraries at

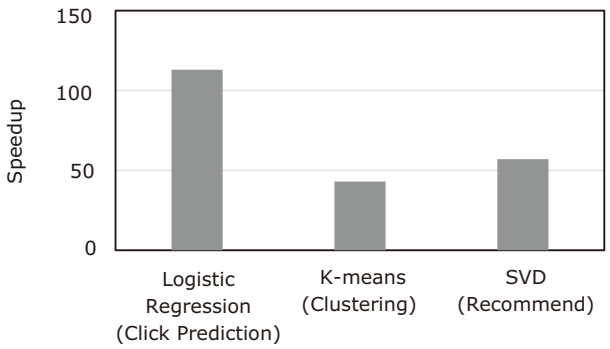

Fig. Performance of Frovedis on Vector Engine (compared to Spark on CPU).

Table Frameworks and compatible algorithms and processors

Algorithm	Framework	Supported Acceralator		
		General Purpose		DL
		GPU	Vector Engine [2]	TPU
Deep Learning	TensorFlow	✓	✓	✓
	PyTorch	✓	×	×
Statistical Machine Learning	Frovedis [1]	×	✓	×
	scikit-learn	×	×	×

Points to be noted:
1. Among statistical machine learning-oriented frameworks, only Frovedis supports accelerators.
2. Only Vector Engine is supported both by the deep learning- and statistical machine learning-oriented frameworks.

very high speeds. The **Figure** shows the performance of Frovedis by comparing the performances of machine learning libraries for the logistic regression, k-means clustering and singular-value decomposition between Frovedis on Vector Engine and Spark on the CPU (Intel Xeon). These results indicate that Frovedis on Vector Engine is by tens or hundreds of times faster and has a potential for finishing processing that used to take tens of hours in a few minutes.

4.3 Framework compatibility situation

The **Table** shows the algorithms and accelerators supported by representative frameworks. TensorFlow is deep learning oriented but is actually the framework to be most widely adopted worldwide. Its special feature is the large number of compatible accelerators. Among the statistical machine learning-oriented frameworks, only Frovedis is compatible with accelerators. On the other hand, from the standpoint of accelerators, Vector Engine is the sole accelerator supported by both the statistical machine learning-oriented framework (Frovedis) and the deep learning-oriented framework (TensorFlow). This fact supports the important positioning in AI of Frovedis among frameworks and of Vector Engines among the accelerators.

5. Conclusion

In the above, this paper describes Vector Engine and Frovedis, which are respectively the particularly important accelerator and framework for the computing platform designed to support AI. NEC will continue preparation of the frameworks capable of utilizing various accelerators in the appropriate positions for use in various AI processing operations.

Reference

1) AI and Compute, May 2018
https://blog.openai.com/ai-and-compute

Authors' Profiles

ISHIZAKA Kazuhisa
Manager
Data Science Research Laboratories

ARAKI Takuya
Senior Principal Researcher
Data Science Research Laboratories

INOUE Hiroaki
Senior Manager
Data Science Research Laboratories

The details about this paper can be seen at the following.

Related URL:

Frovedis
https://github.com/frovedis/frovedis

Information about the NEC Technical Journal

Thank you for reading the paper.
If you are interested in the NEC Technical Journal, you can also read other papers on our website.

Link to NEC Technical Journal website

Japanese | English

Vol.14 No.1 AI and Social Value Creation

Remarks for Special Issue on AI and Social Value Creation
Data — Powering Digitalization and AI

Vol.14 No.1
January 2020

Special Issue TOP

Papers for Special Issue

NEC's Efforts Toward Social Applications of AI
NEC's Commitment to Its New "NEC Group AI and Human Rights Principles" Policy
Human Resource Development in the Age of AI

AI-Enhanced Services/Solutions to Accelerate Digital Transformation
NEC Advanced Analytics Platform (AAPF) Promoting "AI Co-Creation"
Use of Individual Identification Based on the Fingerprint of Things Recognition Technology
Visual Inspection Solutions Based on the Application of Deep Learning to Image Processing Controllers
Remote Vehicle Surveillance Solution Based on Communication Prediction/Control Technology
NEC's Emotion Analysis Solution Supports Work Style Reform and Health Management
Facial Recognition Solution for Offices — Improved Security, Increased Convenience
Outline of an Auto Response Solution (AI Chatbot) for Assisting Business Automation and Labor Saving
AI for Work Shift Support — Accelerating the Transition to Human-Centered Business Value Creation
NEC Cloud Service for Energy Resource Aggregation Leveraging AI Technology
Patient Condition Change Signs Detection Technology for Early Hospital Discharge Support
Effective Data-Based Approaches to Disease Prevention/Healthcare Domains
Co-creation of AI-Based Consumer Insight Marketing Services
"Anokorowa CHOCOLATE" Lets People Savor Delicious Chocolates that Reflect the Mood of Special Moments in History

Cutting-Edge AI Technologies to Create the Future Together With Us
Heterogeneous Object Recognition to Identify Retail Products
Optical Fiber Sensing Technology Visualizing the Real World via Network Infrastructures
Intention Learning Technology Imitates the Expert Decision-Making Process
Graph-based Relational Learning
Retrieval-based Time-Series Data Analysis Technology
New Logical Thinking AI Can Help Optimize Social Infrastructure Management
Deep Learning Technology for Small Data
A Computing Platform Supporting AI