HPC and AI Initiatives for Supercomputer Fugaku and Future Prospects

Atsushi Nukariya Kazutoshi Akao Jin Takahashi Naoto Fukumoto
Kentaro Kawakami Akiyoshi Kuroda Kazuo Minami Kento Sato
Satoshi Matsuoka

The development of AI technology is indispensable for the realization of Society 5.0, which has been announced as part of Japan's science and technology policy. As such development proceeds, the computational resources required for AI learning continue to increase. Through the development of the K computer and the supercomputer Fugaku (hereafter, Fugaku), Fujitsu has been providing high performance computing (HPC) systems with abundant computing resources. Now, in order to utilize the abundant computational resources of HPC systems for AI learning, we are working with RIKEN to develop an AI infrastructure on Fugaku. This article describes the current status of our joint project with RIKEN to test and evaluate the performance of AI-related software on Fugaku and our future work on HPC and AI.

1. Introduction

In 2016, Society 5.0 was proposed as part of Japan's science and technology policy [1]. In the society realized by Society 5.0, one aim is to add value to industry through the analysis of big data using AI.

The development of AI technology is essential for the utilization of AI in society and to lead the world in the realization of Society 5.0. As such development proceeds, the computational resources (e.g. processors, memory) required for AI learning and inference continue to increase. Against this background, the utilization of AI by high performance computing (HPC), which is rich in computational resources, is being promoted.

This article describes the relationship between HPC and AI, and the AI-related work being performed using the supercomputer Fugaku (hereafter, Fugaku). It goes on to describe the future outlook of this initiative.

2. HPC and AI

This section describes the relationship between HPC and AI from the perspective of the use of AI in HPC.

2.1 Fugaku as an AI platform

Three factors are considered to be key contributors to the development of AI technology, which is essential for the realization of Society 5.0 [2].

- Large amounts of data required for learning AI
- High-performance AI platform
- Talented AI personnel

We aim for Fugaku to become the core infrastructure of Society 5.0 by providing a high-performance AI platform that can process large amounts of data at high speed. Further, by providing a high-performance AI platform, we expect to attract talented AI personnel from around the world, which will promote development of AI technology.

2.2 Utilization of AI in HPC

This subsection describes the use of AI in HPC, including the acceleration of deep learning and examples of the application of AI to HPC applications.

1) Acceleration of deep learning

The main reason for the rapid development of AI in recent years is the improvement of deep learning technology. Deep learning can learn more features than conventional machine learning, but this requires a large number of computations. Moreover, since there is a lot of data to be used for learning, a mechanism to process the data at high speed is necessary.

On the other hand, in recent years, the abundance of computational resources has made learning in a realistic time possible. As a result, the use of deep

learning is expanding. The CPUs of the K computer and Fugaku have high computing performance, parallel performance, and wide memory bandwidth, which makes them suitable for deep learning. Going forward, AI applications that use deep learning are expected to be increasingly used on HPC systems rich in computational resources, with increasing acceleration of deep learning.

2) Application of AI technology to applications for HPC

Initiatives have also been launched to realize Society 5.0 through the application of AI technology to HPC applications. For example, deep learning is expected to be useful in coping with large-scale natural disasters by improving the resolution of satellite images and cloud removal to provide a more detailed picture of the ground surface [3]. AI is also being applied to system software tuning using deep learning for the enhancement of HPC [4]. On the other hand, AI is also expected to be used for the acceleration of data augmentation, which is a method for improving the accuracy of deep learning itself [5].

The development of an AI platform for HPC is also important for performing AI-related computations as part of applications for HPC.

3. Deep learning initiatives for Fugaku

This section describes the DL4Fugaku project and other deep learning initiatives for Fugaku.

3.1 Launch of DL4Fugaku project

The DL4Fugaku project was launched with the aim of operating AI software at high speed in order to provide Fugaku as an AI platform [6]. This project aims to develop AI software by tuning and operation verification, and RIKEN and Fujitsu have signed a memorandum of understanding under which they are conducting joint research and development [7]. Research and development of frameworks and applications that enable efficient distributed learning is also underway.

The software developed by the project will be released on an ongoing basis as open-source software (OSS) [8].

3.2 Development of a deep learning framework

This subsection gives an overview of the deep learning framework produced by the DL4Fugaku project and its performance evaluation.

1) Trends in and structure of deep learning frameworks

PyTorch, TensorFlow, Chainer, and MXNet, among other deep learning frameworks, have been released as OSS. PyTorch and TensorFlow adopt a design in which the application programming interface (API) part, which is provided for various languages such as Python, and the operation part, which is implemented in C++, are separate (**Figure 1**). The reason why the operation part is implemented in C++ is that most of the deep learning and inference processing that require high-speed operation is performed in the operation part.

The main processing content of the operation part in deep learning is matrix operations. Thus learning and inference are sped up by calling on arithmetic libraries optimized for Intel CPUs [9], such as the oneAPI Deep Neural Network Library (oneDNN) and the Basic Linear Algebra Subprograms (BLAS), from the operation part.

2) Demonstration of deep learning performance of Fugaku with Chainer

To verify the feasibility of improving the performance of AI software by Fugaku, we performed performance tuning and evaluation during massively parallel execution using Chainer, which is widely used in Japan.

Although there are several algorithms for convolution operations, which are frequently used in deep learning, in Chainer, they are often implemented using GEMM (General Matrix Multiply) operations. Therefore, with a view to combining performance tuning and evaluation with the deep learning library dnnl_aarch64 and the mathematical library of the system software Technical Computing Suite (TCS) of the K computer, we created a prototype of a general-purpose high-speed kernel using GEMM operations so that it could be used on the CPUs of the K computer and Fugaku, and we tuned it on the K computer. The tuning performed was thread parallelization of convolution operations with the mathematical library of TCS and acceleration of the Adam optimization algorithm. This made the learning operation on the K computer 36.4 times faster.

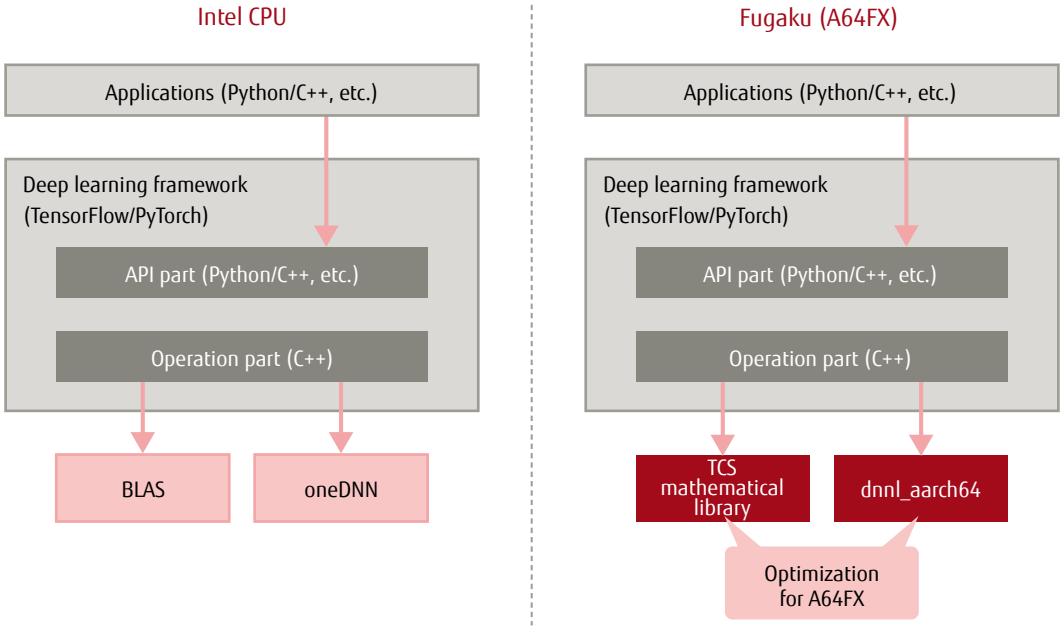


Figure 1
Internal structure of deep learning frameworks.

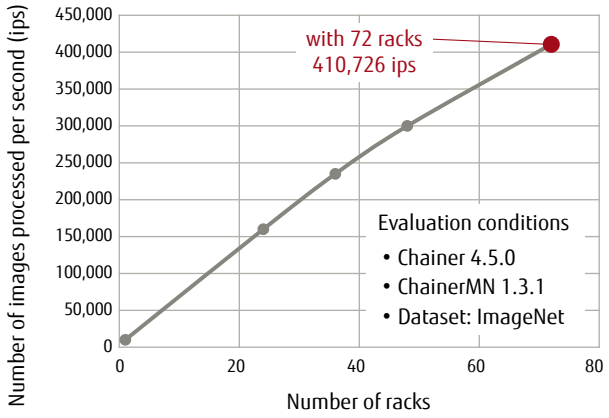


Figure 2
Performance of distributed learning using Chainer (ResNet-50).

On Fugaku, we performed further parallel tuning, such as accelerating I/O processing by pseudo-staging of the newly designed file system Lightweight Layered IO Accelerator (LLIO) and acceleration of the import processing for Python packages. Through this tuning, we achieved learning processing performance of 410,726 ips (images per second) on 72 racks (27,648 nodes) using the ResNet-50 v1.5 image recognition neural network. As shown in **Figure 2**, the performance improved in proportion to the number of racks, and one can expect considerable performance improvement in larger configurations.

Chainer moved into a maintenance phase in 2019 and migration to PyTorch is now being encouraged [10]. Therefore, future performance evaluation of massively parallel execution will be performed using PyTorch.

3) Acceleration initiatives for Fugaku

Fujitsu is developing the dnnl_aarch64 and TCS math libraries optimized for the Fugaku CPU (A64FX), which uses the ARM architecture. Replacing oneDNN and BLAS optimized for Intel CPUs with the dnnl_aarch64 and TCS math libraries, respectively, is expected to accelerate deep learning frameworks on Fugaku.

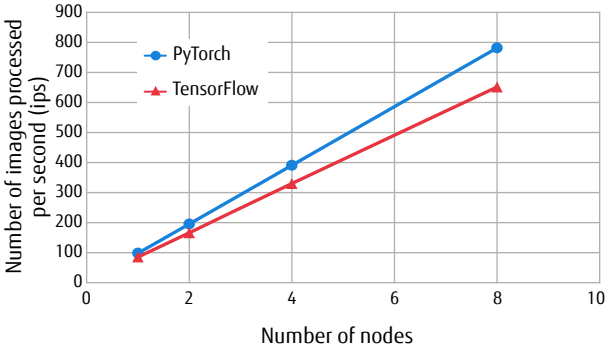
4) ResNet-50 evaluation with PyTorch/TensorFlow

The dnnl_aarch64 and TCS mathematical libraries were incorporated into PyTorch and TensorFlow, and ResNet-50 v1.5 performance evaluation was performed using A64FX.

With PyTorch, we achieved 98.8 ips for learning and 383.7 ips for inferencing on a single node (red figures in **Table 1**). We also evaluated the performance of Horovod, a distributed learning library for deep

Table 1
Single-node performance with PyTorch/TensorFlow.

Framework	Dataset	Number of processes	Learning performance (ips)	Inferencing performance (ips)
PyTorch 1.5.0	Dummy	1	82.3	337.6
	Dummy	4	98.8	383.7
TensorFlow 2.1.0	Dummy	1	60.5	176.8
	Dummy	4	86.9	295.6



Evaluation conditions

[PyTorch]
- PyTorch 1.5.0
- Horovod v0.19.0
- Dataset: Dummy data

[TensorFlow]
- TensorFlow 2.1.0
- Horovod v0.19.2
- Dataset: Dummy data

Figure 3
Distributed learning performance.

learning frameworks, during distributed learning, and achieved 781.5 ips with 8 nodes (**Figure 3**).

We also verified the operation of TensorFlow, which along with PyTorch has come into wide use, and we achieved 86.9 ips for learning and 295.6 ips for inferencing on a single node (red figures in Table 1). In distributed learning using Horovod, we achieved 651.5 ips (Figure 3).

Going forward, we will evaluate performance in larger environments with more than 8 nodes based on the results of massively parallel execution in Chainer.

4. Future initiatives

1) Collaboration with the ARM community

Developed for A64FX, dnnl_aarch64 has been released as OSS [11]. Further, information on how to use PyTorch and TensorFlow with CPUs that use the ARM architecture is provided by the Arm HPC Users Group [12]. Looking ahead, we will contribute to the development of AI platforms for ARM CPUs through feedback of the knowledge gained from the performance evaluation and operation verification on Fugaku.

2) Application of new technologies

Fujitsu is developing technologies to accelerate AI processing, such as content-aware computing (CAC) [13], which automatically adjusts the accuracy of calculations according to the content of calculations during execution. We aim to apply the technologies we develop to AI software and provide higher-performance AI platforms. We will also continue to investigate applications that A64FX excels in.

5. Conclusion

This article has described the relationship between HPC and AI, and AI initiatives for Fugaku.

These efforts targeting deep learning frameworks on Fugaku are expected to contribute to further development of AI platforms for HPC systems. In the future, we will evaluate the performance of massively distributed learning and investigate the applications that A64FX excels in to promote the use of AI in HPC. We aim also to contribute to the realization of Society 5.0 by developing AI platforms in ARM through collaboration with ARM Ltd. and other OSS communities.

References and Notes

[1] Cabinet Office: Society 5.0. *https://www8.cao.go.jp/cstp/english/society5_0/index.html*

[2] S. Matsuoka: Supercomputer 'Fugaku'. p. 17 (May 2020). (in Japanese) *https://www.r-ccs.riken.jp/wp-content/uploads/2020/05/20200515_matsuoka.pdf*

[3] B. Adriano et al.: Cross-domain-classification of tsunami damage via data simulation and residual-network-derived features from multi-source images. Proceedings

of 2019 IEEE International Geoscience and Remote Sensing Symposium, pp. 4947-4950.
[4] T. Dey et al.: Optimizing Asynchronous Multi-Level Checkpoint/Restart Configurations with Machine Learning. 2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), New Orleans, USA, 2020.
[5] R. Hataya et al.: "Faster AutoAugment: Learning Augmentation Strategies using Backpropagation." arXiv:1911.06987, 2019.
[6] S. Matsuoka: Fugaku as the Centerpiece of Society5.0 Revolution. p. 42 (Feb 2020). *https://www.r-ccs.riken.jp/R-CCS-Symposium/2020/shared/images/under/program/1-01_Matsuoka.pdf*
[7] RIKEN Center for Computational Sciences: Signing of Memorandum of Understanding with Fujitsu to Build AI (Artificial Intelligence) Framework for Fugaku. (in Japanese) *https://www.r-ccs.riken.jp/library/topics/191126.html*
[8] GitHub: DL for Fugaku. *https://github.com/dl4fugaku*
[9] Intel Open Source Technology Center: oneAPI Deep Neural Network Library (oneDNN). *https://01.org/onednn*
[10] Preferred Networks: Preferred Networks Migrates its Deep Learning Research Platform to PyTorch. *https://preferred.jp/en/news/pr20191205/*
[11] GitHub: Fujitsu. *https://github.com/fujitsu*
[12] GitLab: Arm HPC Users Group. *https://gitlab.com/arm-hpc*
[13] Fujitsu Laboratories: Content-Aware Computing: Automatically Adjusting Computational Accuracy to Accelerate AI Processing Tenfold (researcher interview). *https://www.fujitsu.com/jp/group/labs/en/about/resources/article/202004-cac.html*

Atsushi Nukariya
Fujitsu Limited, Platform Software Business Unit
Mr. Nukariya is currently engaged in R&D of AI software for HPC.

Kazutoshi Akao
Fujitsu Limited, Platform Software Business Unit
Mr. Akao is currently engaged in R&D of AI software for HPC.

Jin Takahashi
Fujitsu Limited, Platform Software Business Unit
Mr. Takahashi is currently engaged in R&D of AI software for HPC.

Naoto Fukumoto
Fujitsu Laboratories Ltd, ICT Systems Laboratory
Dr. Fukumoto is currently engaged in R&D of AI software for HPC.

Kentaro Kawakami
Fujitsu Laboratories Ltd, Platform Innovation Project
Dr. Kawakami is currently engaged in R&D of AI software for HPC.

Akiyoshi Kuroda
RIKEN, Center for Computational Sciences, Operations and Computer Technologies Division
Dr. Kuroda is currently engaged in development of application development platforms for Fugaku.

Kazuo Minami
RIKEN, Center for Computational Sciences, Operations and Computer Technologies Division
Dr. Minami is currently engaged in development of application development platforms for Fugaku.

Kento Sato
RIKEN, Center for Computational Sciences, High Performance Big Data Research Team
Dr. Sato is currently engaged in R&D of system software for AI/big data processing for HPC.

Satoshi Matsuoka
RIKEN, Head of the Center for Computational Sciences, High Performance Artificial Intelligence Systems Research Team
Dr. Matsuoka leads the DL4Fugaku Project.

This article first appeared in Fujitsu Technical Review, one of Fujitsu's technical information media. Please check out the other articles.

https://www.fujitsu.com/global/technicalreview/