FUJITSU

About the Performance of Fujitsu Server PRIMERGY CDI

Fujitsu Server PRIMERGY CDI (hereinafter, PRIMERGY CDI) stores PCIe devices such as GPU accelerators, storages, and network interface cards (NICs) in external PCIe boxes, not in the server itself. Moreover, a high-speed PCIe fabric switch connects the server and the PCIe boxes, achieving an efficient structure. By utilizing the management functions provided by dedicated software, resources in the box can be freely deployed and released according to the customers' workload, maximizing resource utilization, and enabling efficient operation.

In PRIMERGY CDI, new components such as HBA (Host Bus Adapter), PCIe fabric switch, and PCIe box are introduced. This document will explain the impact these additions have on performance.

This document will provide a detailed explanation of the impact these additions have on performance in the following order. In doing so, we will use the PRIMERGY RX2540M7 (hereafter RGX2540M7) as a reference for comparison with PRIMERGY CDI.

- Performance evaluation of inter-GPU communication
- Performance evaluation and profiler analysis[1] of ResNet benchmark from MLPerf™[2]

This time, we focused on evaluating the inter-GPU communication, which is crucial for learning applications. Performance evaluation between CPUs and GPUs is planned to be conducted separately.

[1] https://developer.nvidia.com/nsight-systems

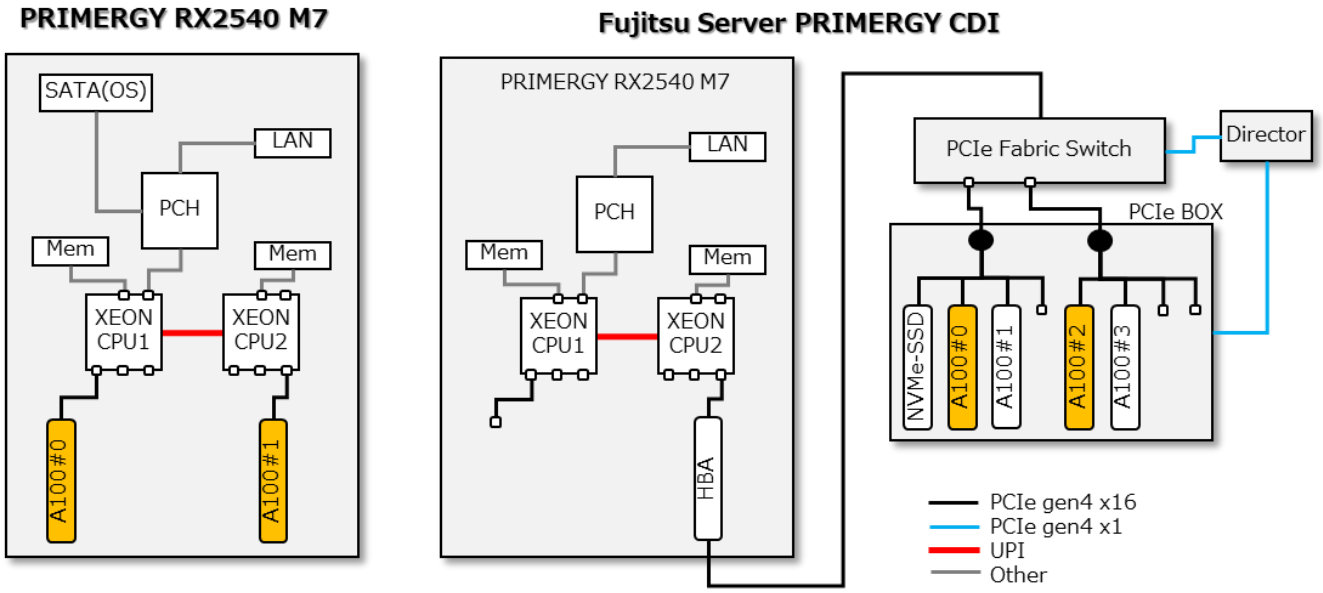
[2] MLPerf™ name and logo are trademarks of MLCommons™ Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.

About the Performance of Fujitsu Server PRIMERGY CDI

1. System Configuration

Block Diagram Comparison

The system configurations are shown in the following diagram. The PRIMERGY CDI system comprises of RX2540 M7 server, which includes an HBA attached to the PCIe slot of the server, a PCIe Fabric Switch and a PCIe box which includes four NVIDIA® A100-PCIe GPUs and an SSD. On the other hand, the reference machine, RX2540 M7, has a configuration where each of the two CPUs is connected to the NVIDIA® A100-PCIe GPUs. Performance measurements are carried out with these two configurations, and performance evaluations are conducted based on the presence or absence of the HBA-PCIe Fabric Switch-PCIe box.



System Specification Comparison

Server		PRIMERGY RX2540 M7	Fujitsu Server PRIMERGY CDI
CPU		Intel(R) Xeon(R) Gold 6430x2	
	Frequency	2.1GHz	
	Core Count	32	
Memory		16GBx16	
Storage		SATA	8x 800GB NVMe SSD
Interconnect		PCIe 4.0	
GPU		NVIDIA A100-PCIe-80GBx2	
OS		Red Hat Enterprise Linux release 8.6 (Ootpa)	
Software		CUDA: 12.1.0.023	
		cuda_driver_version: 530.30.02	
HBA		—	PCIe HBA Card for CDI (Bandwidth 64GB/s (Bidirectional))
PCIe Fabric Switch		—	PCIe Fabric Switch (48port) for CDI x1 (Total Bandwidth 768 GB/s Bidirectional 48 port)
PCIe BOX		—	PCIe Box (PCIe×8) for CDI x1 (Maximum Port Bandwidth 128GB/s (Bidirectional))
Director		—	Controller Appliance for CDI

About the Performance of Fujitsu Server PRIMERGY CDI

2. Performance Evaluation on Inter-GPU Communication

Performance Benchmark Test

During learning process, inter-GPU communication occurs via PCIe fabric switches and PCIe boxes . We conducted the following benchmark tests to measure communication performance including these components. Both tests are provided from NVIDIA® and available from the following sites.

- p2pBandwidthLatencyTest [3]

  The p2pBandwidthLatencyTest is a benchmark test that measures the bandwidth and latency of inter-GPU communication.

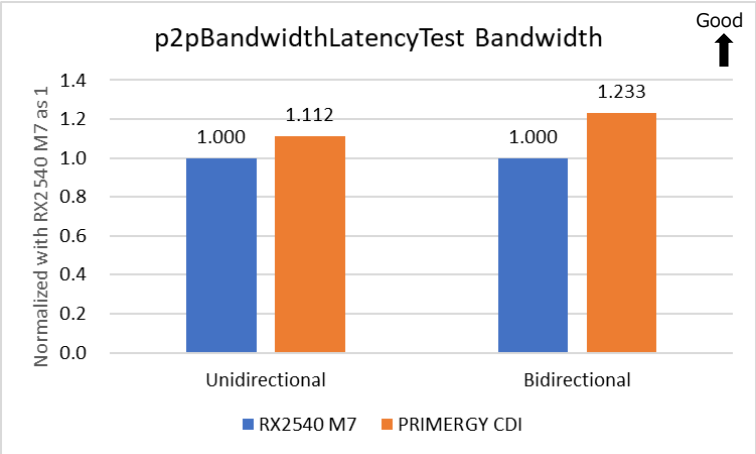
- nccl-tests [4]

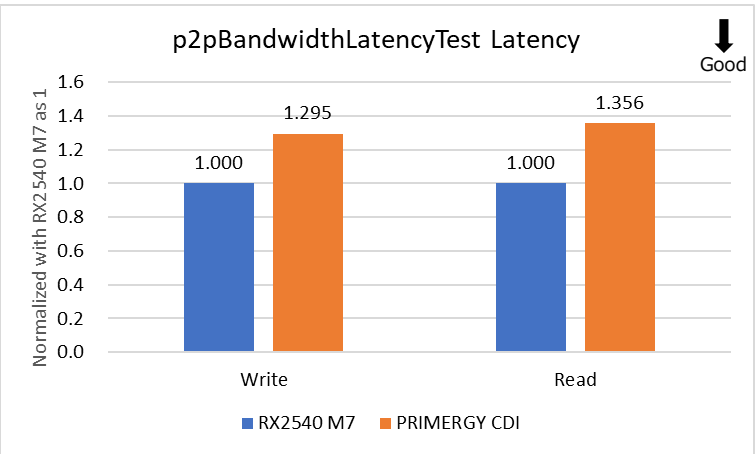
  The nccl-tests are benchmark tests that measure the performance of NCCL, which implements AllReduce used in learning.

[3] https://github.com/NVIDIA/cuda-samples/tree/master/Samples/5_Domain_Specific/p2pBandwidthLatencyTest

[4] https://github.com/NVIDIA/nccl-tests

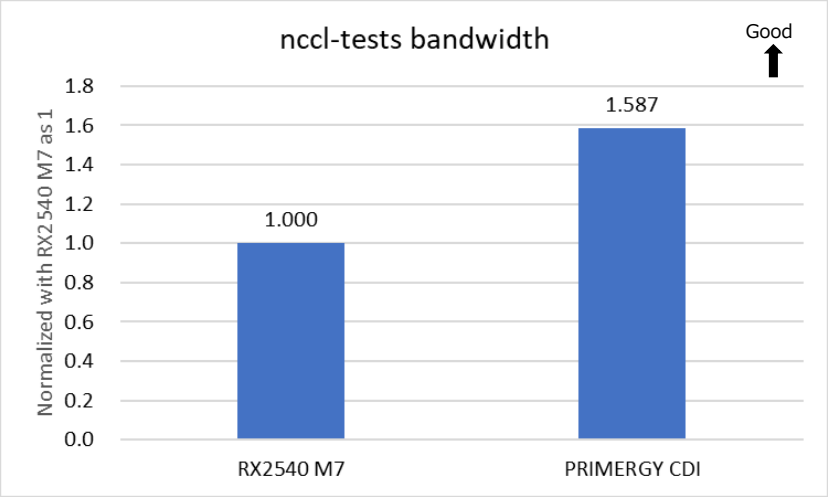
Performance Evaluation of p2pBandwidthLatencyTest





About the Performance of Fujitsu Server PRIMERGY CDI

Performance Evaluation of nccl-testsccl-tests



Summary of Inter-GPU Communication Performance Evaluation

- In the reference machine RX2540 M7, inter-GPU communication needs to go through two CPUs. On the other hand, in PRIMERGY CDI, communication can be done just through PCIe connections, resulting in a wider communication bandwidth compared to the reference machine.
- According to the results of the nccl-tests, it is clear that the performance of AllReduce on PRIMERGY CDI significantly surpasses that of RX2540 M7. Therefore, it is expected that the configuration of PRIMERGY CDI will be advantageous in regular learning as well.
- From the perspective of inter-GPU communication latency, the latency in the PRIMERGY CDI configuration is longer than that in the RX2540 M7 environment. Therefore, there are concerns about the impact on learning performance.

About the Performance of Fujitsu Server PRIMERGY CDI

3. Throughput Performance Evaluation of ResNet

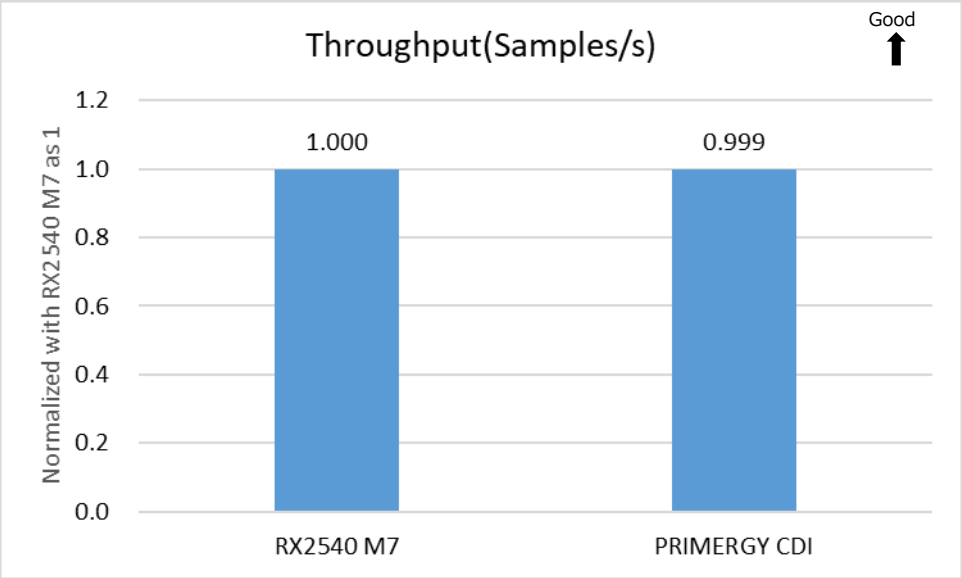
In this section, we will compare the throughput during training and conduct performance analysis using a profiler between the PRIMERGY CDI and the RX2540M7 machine. The training program used will be the ResNet benchmark program adopted in MLPerf™ Training.

3.1. Throughput Performance Evaluation

Evaluation Method

Throughput is defined as the number of images processed per second. It is calculated by dividing the number of whole images in the training dataset by the processing time for one epoch.

Throughput Performance Evaluation



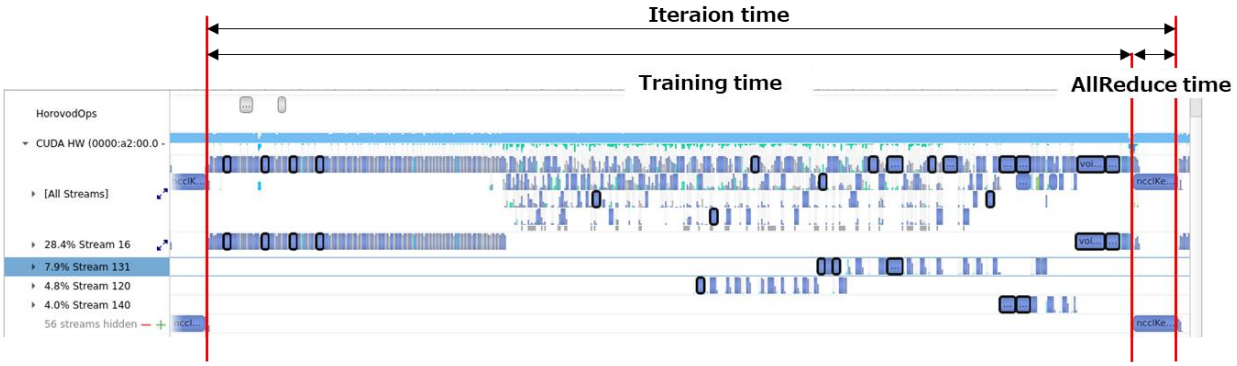
Summary of Throughput Performance Evaluation

- The difference in throughput between PRIMERGY CDI and RX2540 M7 was 0.1%, with the RX2540M7 machine performing better. From the results of the previous section, although the bandwidth of inter-GPU communication on PRIMERGY CDI is wider than that of RX2540M7, the latency is longer. Therefore, it is believed that the advantages and disadvantages cancelled each other out, resulting in this outcome.

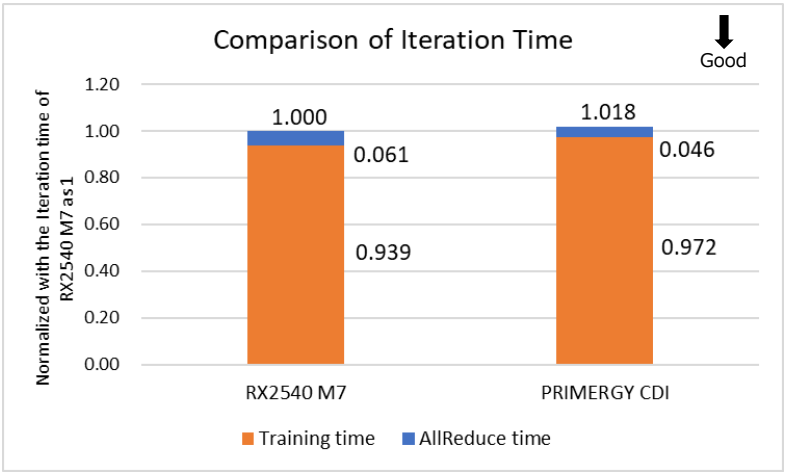
3.2. Analysis by NVIDIA Nsight™ Systems

Analysis Method Using Nsight™ Systems

In the performance analysis we used Nsight™ Systems, a profiler made by NVIDIA. The profiler records and analyzes operations in the ResNet benchmark program. We analyzed the processing content of one loop of learning. Then, we measured the time required for learning, the time required for inter-GPU communication, and the time required for one loop. Finally, we compared these measurements between the PRIMERGY CDI and the RX2540M7 machine.



Performance Evaluation



Summary of Analysis Method Using Nsight™ Systems

- The time required for AllReduce is shorter on the PRIMERGY CDI. This is consistent with the evaluation results from the nccl-tests.
- The time required for training is shorter on the RX2540M7. This is likely due to the longer latency of inter-GPU communication.
- The time required for one loop is shorter on the RX2540M7, and 1.8% longer on the PRIMERGY CDI. This result is a larger difference than in the throughput evaluation in the previous section. It is believed that the difference occurred because the measurement range is limited in the evaluation by the profiler.

4. Evaluation of MLPerf™ Benchmark Test

Evaluation Method

We ran ResNet benchmark program on the PRIMERGY CDI system equipped with four GPUs (NVIDIA A100 PCIe 80GB), and measured a training time that can be compared with the results submitted by other companies to MLPerf™. Therefore, measurements must be carried out in accordance with the rules of MLPerf™[5].

According to the rules, the benchmark program must be run a specified number of times, and the time it takes to reach a certain level of accuracy is measured each run. The average of the measured times, excluding the maximum and minimum, is taken as the score. The software used for evaluation is the one published in MLPerf™ Training. The source code can be obtained from the following website [6].

Measurement Results

The measurement result for ResNet achieved 60.444 minutes[7]. This result has also been submitted to MLPerf Training v.3.0[8] and is published as follows.

ID	Submitter	System	Processor	#	Accelerator	#	Software	ResNet
Unverified								
	Fujitsu	PRIMERGY CDI mxnet	Intel(R) Xeon(R) Gold 6430	2	NVIDIA A100-PCIe-80GB	4	MXNet NVIDIA Release 23.04	60.444

Summary of MLPerf™ Benchmark Test Evaluation

- Based on the comparison of the results submitted to MLPerf, it was found that PRIMERGY CDI achieves a similar score to systems of similar specifications from other companies. From this, it can be inferred that while the PRIMERGY CDI adds a new HBA, PCIe Fabric Switch, and PCIe Box to the conventional server, the impact on their performance is negligible.

[5] https://github.com/mlcommons/training_policies/blob/master/training_rules.adoc

[6] The source code of ResNet benchmark program: https://github.com/mlcommons/training_results_v3.0/tree/main/NVIDIA/benchmarks/resnet/implementations/mxnet

[7] Unverified MLPerf™ v3.0 Training Closed ResNet. Result not verified by MLCommons Association. The MLPerf™ name and logo are trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.

[8] MLPerf™ Training v3.0 Results: https://mlcommons.org/en/training-normal-30/

FUJITSU

About the Performance of Fujitsu Server PRIMERGY CDI

5. Summary

- The table below compares the PRIMERGY CDI and the reference machine, the RX2540M7. The comparison items are the inter-GPU communication measured so far, the throughput of the benchmark program, and the time obtained from the profiler analysis. It can be said that it does not affect the actual throughput obtained by running the benchmark program although slight impact of the new configuration added in PRIMERGY CDI can be confirmed in some item.

Evaluation test	Details	RX2540 M7		PRIMERGY CDI	
		result *3	value *2	value *2	result *3
p2pBandwidthLatencyTest	Unidirectional		1.000	1.112	○
	Bidirectional		1.000	1.228	○
	Latency *1	○	1.000	1.356	
nccl-tests	nccl-tests		1.000	1.590	○
MLPerf™ Training v2.1 ResNet	Throughput	○	1.000	0.999	○
NVIDIA Nsight™ Systems	ALLReduce time *1		0.061	0.046	○
	Training time *1	○	0.939	0.972	
	Iteration time *1	○	1.000	1.018	

*1: A test where a smaller value is better, *2: A value normalized from the measurement results in each test,
*3: A circle indicates a good result or equivalent.

- As a result of conducting the benchmark test of ResNet to compare with the scores of other companies in MLPerf™, PRIMERGY CDI was able to obtain a score equivalent to that of other companies with similar specifications.
- Based on the throughput measurement results of ResNet and the benchmark score measured according to the rules of MLPerf, the new configuration added in PRIMERGY CDI does not affect performance and can maintain performance equivalent to the conventional one. As a result, it enables "freely deploying and releasing resources according to customer's workload, maximizing resource utilization, and operating efficiently."

[Precautions]

- We prohibit the redistribution of information, such as forwarding this document to a third party or uploading the contents of this document to a website.
- The copyright belongs to Fujitsu Ltd., or its information providers, and unauthorized reproduction of the contents is prohibited.
- The performance information contained in this document does not guarantee performance improvement in customer systems.

[About Trademarks]

- NVIDIA® is a registered trademark or trademark of NVIDIA Corporation in the United States.
- Intel, Xeon are registered trademarks or trademarks of Intel Corporation in the United States.
- Other company names, product names, etc. mentioned are registered trademarks or trademarks of their respective companies.
- In addition, not all company names, system names, product names, etc. described in this document are marked with trademark symbols (®, ™).