Fsas Technologies

PRIMERGY CDI

GPU Scalability

1. Introduction

Fsas Technologies Inc. have introduced PRIMERGY CDI, a brand-new series of our server products. They consist of the following components: computational servers, PCIe fabric switches, and PCIe boxes. PCI devices such as GPUs, SSDs and NICs are installed not inside the servers but in the PCIe boxes outside the servers.
One of the most outstanding features of PRIMERGY CDI is that all devices in the PCIe boxes can be shared and allocated among the servers. Thus, users can modify the affiliation of the devices depending on workloads executed in the servers.

Our previous document has detailed the performance of inter-GPU communication, the analysis results using a profiler[1], and the results of benchmarks obtained in accordance with the rules of MLPerf™[2].
This white paper demonstrates that in the ResNet benchmark program, the throughput improves according to the number of GPUs used in the following order:

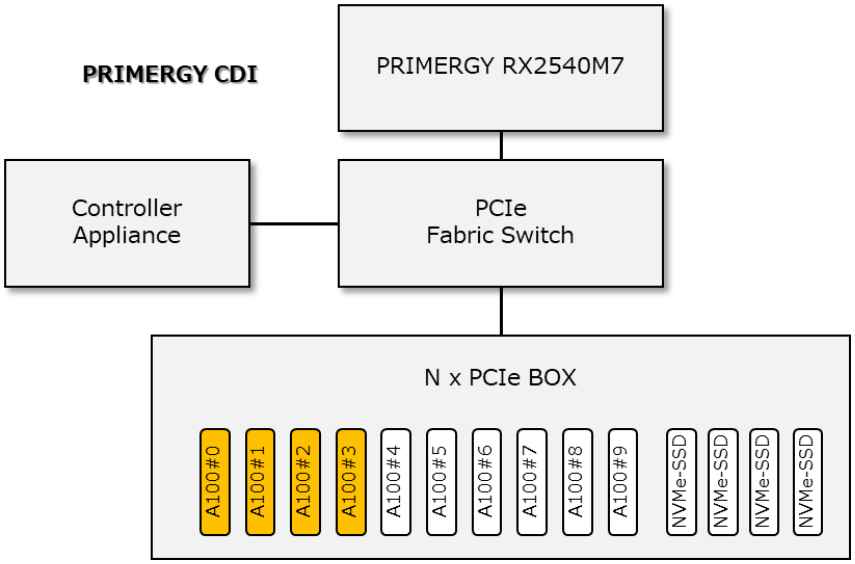
- Configuration of the PRIMERGY CDI used for the measurement,
- Performance at each number of GPUs, and
- Conclusion.

[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.

2. System configurations

2. 1 Block Diagram

The configuration of the PRIMERGY CDI system used for the measurements is as shown in the figure below. We use RX2540M7 server as a compute node, along with a PCIe fabric switch, two PCIe boxes, and a controller appliance. Each PCIe box is equipped with 10 GPUs (NVIDIA® A100 PCIe 80GB) and 4 SSDs. The figure shows the case of using 4 GPUs. In the measurements, we ran the benchmark program and obtained scores and logs when the number of GPUs used was 2, 4, 8, and 10.



2. 2 System configurations

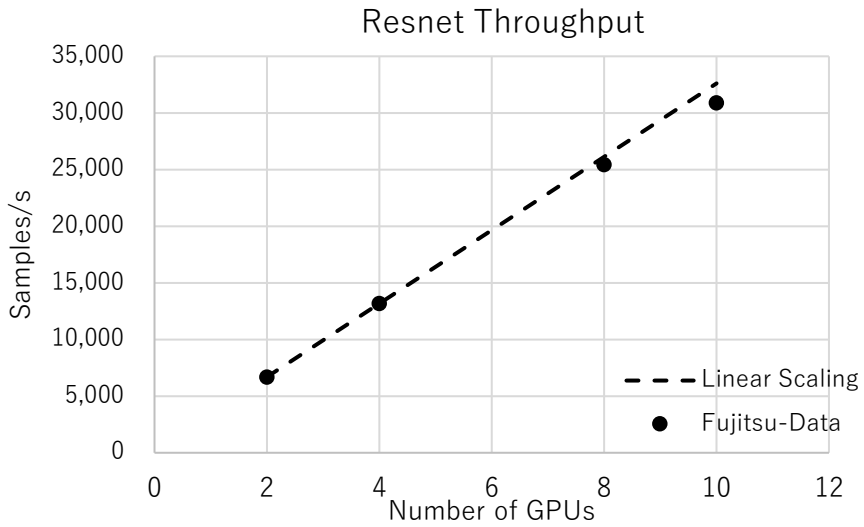
Server		PRIMERGY CDI
CPU		Intel(R) Xeon(R) Gold 6430x2
	Frequency	2.1GHz
	Core Count	32
Memory		16GBx16
Storage		8x 800GB NVMe SSD
Interconnect		PCIe 4.0
GPU		NVIDIA A100-PCIe-80GBx2~10
OS		Red Hat Enterprise Linux release 8.6 (Ootpa)
Software		CUDA: 12.1.0.023
		cuda_driver_version: 530.30.02
HBA		PCIe 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 for CDI xN (Maximum Port Bandwidth 128GB/s (Bidirectional))
Director		Controller Appliance for CDI

3. Performance at each number of GPUs

The table below shows the results of ResNet training by changing the number of GPUs used. Each column represents the actual training time (Training time), the number of images processed until the training is completed (Samples to train), and the throughput, respectively. The computation involved in training changes depending on the number of GPUs, which in turn alters the number of images processed. Furthermore, the throughput is the number of images processed per second, and this value is obtained by dividing the number of images processed by the time required for training.

GPU	Training time(sec)	Samples to train	Throughput (Samples/s)
2	6890.177	46,122,012	6,694
4	3501.386	46,122,012	13,173
8	1761.707	44,840,845	25,453
10	1451.635	44,840,845	30,890

The graph below shows the change in throughput in relation to the number of GPUs used. The black dots on the graph represent the actual measurements, and the dashed line is a straight line connecting the throughput of 2 GPUs and 4 GPUs, representing the ideal throughput if the performance was perfectly proportional to the number of GPUs. The difference between the ideal throughput and the actual measurement for 10 GPUs is calculated to be 94.73%. Therefore, it can be said that the performance during the training of ResNet is able to output performance relative to the number of GPUs.



4. Conclusion

- We demonstrated that the performance scales well according to the number of GPUs used by calculating the throughput from the training results submitted to MLPerf™ Training v3.1.
- PRIMERGY CDI can accommodate up to 20 GPUs. By equipping a large number of relatively inexpensive GPUs, it can also be used for training purposes.
- Given this scalability of performance, you can use PRIMERGY CDI systems starting with a smaller number of GPUs, and add GPUs as needed to improve performance. It also makes it easy to estimate performance improvements in response to adding GPUs.