Fsas Technologies

PRIMERGY CDI

Performance Evaluation of CPU-GPU Interconnect in Servers

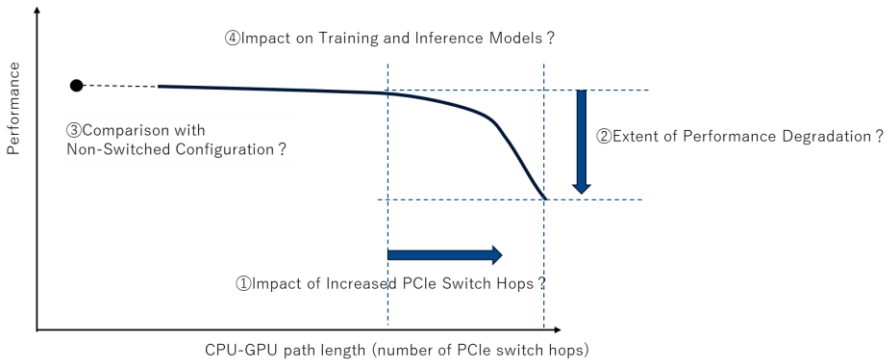
1. Introduction

Fsas Technologies Inc. offers PRIMERGY CDI, a novel server series that departs significantly from traditional designs. PRIMERGY CDI comprises compute servers, PCIe fabric switches, PCIe boxes, and a director. This design locates GPUs, SSDs, and NICs in external PCIe boxes, separate from the compute server chassis. A key feature of PRIMERGY CDI is its flexible allocation of devices within the PCIe boxes to multiple compute servers. This enables performance scaling, such as proactively adding GPUs to handle anticipated increases in inference workloads.

In previous white paper [1], we discussed the performance evaluation of inter-GPU communication, a critical factor for training models like ResNet. This white paper focuses on the performance of the CPU-GPU interconnect within PRIMERGY CDI, providing a detailed evaluation.
Unlike conventional servers, PRIMERGY CDI employs a configuration where the communication path between the CPU and GPU passes through multiple PCIe switches, utilizing PCIe fabric switches and PCIe boxes. Generally, each PCIe switch introduces latency, potentially affecting the performance of the CPU-GPU interconnect.
Therefore, this white paper verifies the following points based on measured data:

① Impact of Increased PCIe Switch Hops: How does increasing the number of PCIe switch hops in the CPU-GPU interconnect path affect performance?
② Extent of Performance Degradation: What is the rate of performance degradation due to an increased number of PCIe switch hops?
③ Comparison with Non-Switched Configuration: What is the performance difference compared to a configuration without PCIe switches between the CPU and GPU?
④ Impact on Training and Inference Models: How does interconnect performance affect training and inference models?

Through these verifications, we aim to provide a deeper understanding of the CPU-GPU interconnect performance of PRIMERGY CDI.



[1] https://www.fujitsu.com/jp/products/computing/servers/primergy/solution/cdi/cdi_whitepaper.pdf

---

---

2. System Configuration and Adjustment of CPU-GPU Interconnect Path

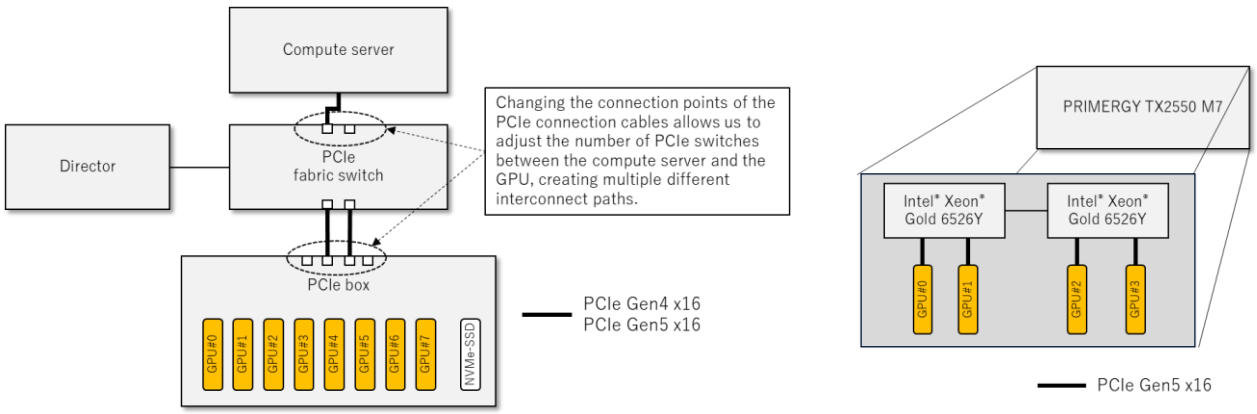
The block diagram of the server used in this evaluation and a comparison table of system specifications are shown below. PRIMERGY CDI allows us to adjust the number of PCIe switches between the compute server and the GPU by changing the connection points of the PCIe connection cables in the PCIe fabric switch and PCIe box, creating multiple interconnect paths. This enables a thorough verification of the CPU-GPU interconnect path's impact.

This evaluation collected data using two PRIMERGY CDI configurations:

- v1.0: Supports PCIe Gen4 and is equipped with NVIDIA L40S GPUs.
- v1.1: Supports PCIe Gen5 and is equipped with NVIDIA H100 NVL GPUs.

We performed performance evaluations with four interconnect paths for v1.0 and two for v1.1. Hereafter, PCIe Gen4 may be abbreviated as "Gen4" and PCIe Gen5 as "Gen5".

For comparison with a configuration without PCIe switches between the CPU and GPU, we used PRIMERGY TX2550 M7. This server supports PCIe Gen5, but operates as PCIe Gen4 when combined with the PCIe Gen4 L40S. To align with the evaluation target configurations of PRIMERGY CDI, we collected data with both Gen4 L40S and Gen5 H100 NVL configurations and compared their performance.



Server	PRIMERGY CDI v1.0	PRIMERGY CDI v1.1	PRIMERGY TX2550 M7
CPU	Intel® Xeon® Gold 6454S x2	Intel® Xeon® Gold 6530 x2	Intel® Xeon® Gold 6526Y x2
Frequency	2.2GHz	2.1GHz	2.8GHz
Core Count	32	32	16
Memory	16x 64GB	16x 64GB	16x 32GB
Storage	745.2GBx8 NVMe SSD	745.2GBx8 NVMe SSD	894.3GB (SATA SSD)
Interconnect	PCIe Gen4 x16	PCIe Gen5 x16	PCIe Gen5 x16
GPU	NVIDIA® L40S x8	NVIDIA® H100 NVL x8	NVIDIA® L40S x 4 NVIDIA® H100 NVL x 4
OS	Ubuntu 22.04.5	Ubuntu 22.04.5	Ubuntu 22.04.5
Software	CUDA 12.4	CUDA 12.4	CUDA 12.4
	cuda_driver_version: 550.90.07	cuda_driver_version: 550.90.07	cuda_driver_version: 550.90.07
HBA	PCIe HBA Card for CDI	PCIe Gen5 HBA Card for CDI	—
PCIe fabric switch	PCIe fabric switch (48port) for CDI	PCIe Gen5 fabric switch (48port) for CDI	—
PCIe box	PCIe box for CDI	PCIe Gen5 box (PCIe x 8) for CDI	—
Director	Controller Appliance for CDI	Controller Appliance for CDI	—

3. Evaluation Method Focusing on Data Transfer Rates for Training and Inference

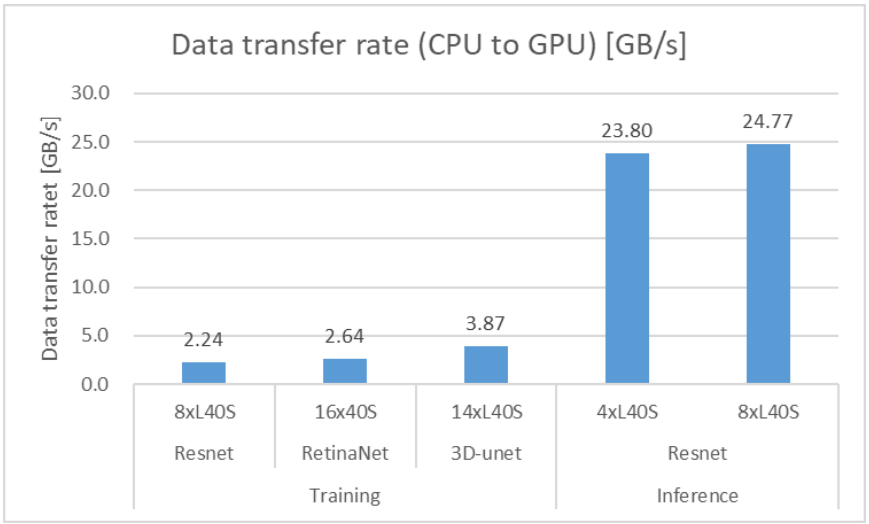
For training, current MLPerf™ [2] submission results [3] did not reveal any performance impact due to CPU-GPU latency.Two reasons are considered for this:

- GPU Processing Time: In training, the GPU processing time is significant. Therefore, the latency caused by PCIe switches when transferring training data from the compute server to the GPU is expected to be negligibly small.
- Data Transfer Rate: Calculations show that even with large training datasets in image-based benchmarks like ResNet, RetinaNet, and 3D-Unet, the data transfer rate from the compute server to the GPU uses only up to 12% of the Gen4 (32GB/s) bandwidth. Therefore, any impact from PCIe switch latency is unlikely to be noticeable.

Next, for inference, previous PRIMERGY CDI results confirm that using four L40S GPUs in a Gen4 (32GB/s) environment for ResNet50 inference processing achieves an effective data transfer rate of approximately 25GB/s. This suggests that the interconnect is the bottleneck because the ResNet50 inference processing load is too light for the GPU performance, while other benchmarks are limited by GPU performance. Currently, this trend is observed only with ResNet50 in the MLPerf Inference benchmark tests.
Focusing on this effective data transfer rate in ResNet50, we will measure the impact of CPU-GPU path length (number of PCIe switch hops) on the data transfer volume. By reducing the number of GPUs from 4 (maximum 24GB/s) incrementally, we will adjust the interconnect bandwidth and then measure the impact of path length on the data transfer volume.

The following table and graph show the data transfer volume from the compute server to the GPU, calculated from the logs of each benchmark test. For Training, the learning time for one epoch is identified from the operation log, and the data transfer volume is calculated from the size of the training dataset. For ResNet50 in Inference, the score is output in QPS (queries/sec), so the data transfer volume is calculated from the data size of the query image. The table and graph show that bandwidth usage is low for Training and high for ResNet50 in Inference.

Category	Public ID	Benchmark Test	GPU Configuration	Score	GB/s	Bandwidth Utilization (% of Gen4 32GB/s)
Training(Submited)[2]	4.0-0026	Resnet	8xL40S	50.669[min]	2.24	6.99%
	4.0-0025	RetinaNet	16xL40S	58.647[min]	2.64	8.26%
	4.0-0024	3D-unet	14xL40S	16.860[min]	3.87	12.08%
Inference(Measured)	—	Resnet50	4xL40S	158,125[QPS]	23.80	74.38%
			8xL40S	164,531[QPS]	24.77	75.25%

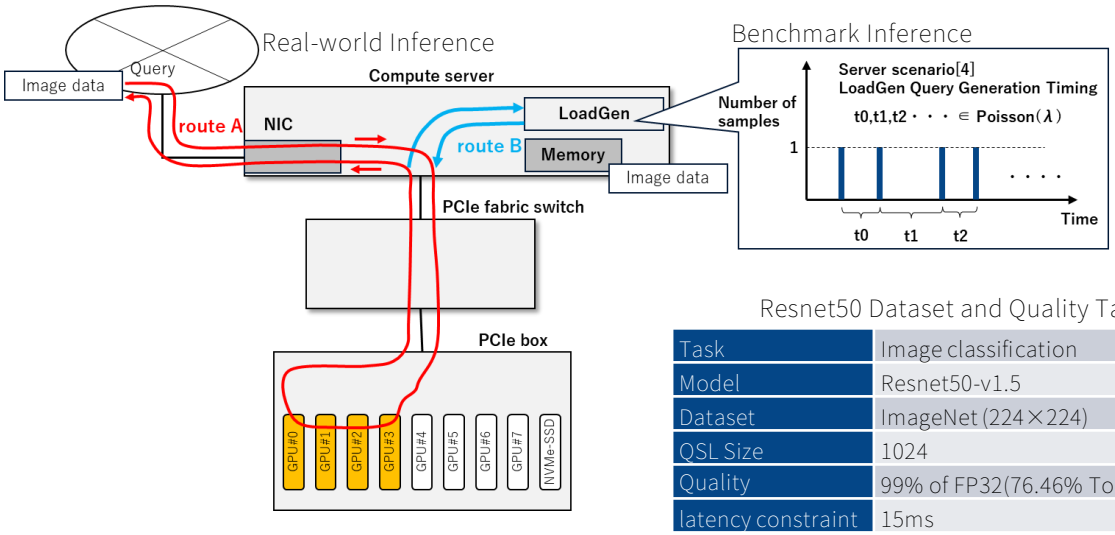


[2] MLPerf™ name and logo are trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.
[3] https://mlcommons.org/benchmarks/training/

4. Performance Evaluation Methodology Using a Benchmark Test

As shown in the figure below, the actual inference process occurs via route A. Each image data, serving as the inference query, is sent from the Internet, ingested into the compute server via the NIC, and then passed to the GPU. The inference result processed by the GPU is sent back to the Internet via route A. However, in this benchmark test, the LoadGen application, which generates the query, is located within the compute server. From here, according to the Server scenario [4] shown in the figure, the application generates a query using image data stored in the compute server's memory, and sends the data by first taking route B and then joining route A partway through, the inference result processed by the GPU is sent back to the LoadGen application from a point midway along route A, passing through route B.

The benchmark test's performance is measured by varying the target QPS (queries/sec) value output by the LoadGen application and searching for the maximum QPS value [5] that achieves a Top-1 accuracy of 76.46% while ensuring that 99% of queries are answered within a predefined time slot, which is 15ms for ResNet50. This QPS value search was previously performed manually, but we have automated it using a binary search algorithm [6], eliminating manual bias and improving data acquisition.



Resnet50 Dataset and Quality Target [5]

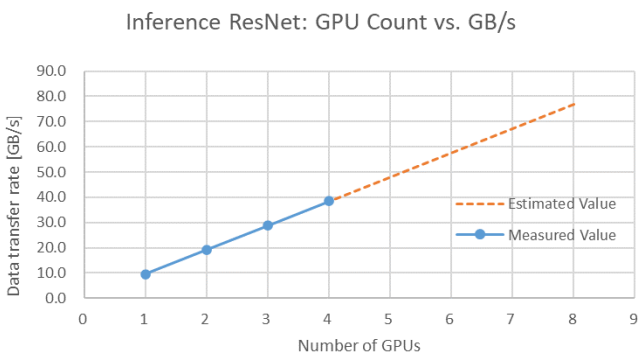
Task	Image classification
Model	Resnet50-v1.5
Dataset	ImageNet (224×224)
QSL Size	1024
Quality	99% of FP32(76.46% Top-1 accuracy)
latency constraint	15ms

5. Performance Evaluation of CPU-GPU Interconnect without PCIe Switch

We use the PRIMERGY TX2550 M7 server (max 4 GPUs) for comparison with a configuration without PCIe switches. For performance exceeding 4 GPUs, we use values estimated based on measured data (1-4 GPUs). This estimation is reasonable, as inference performance scales nearly linearly with the number of GPUs (based on our past MLPerf submissions and publicly available results from other companies).

The table and graph show measured values and values estimated based on measured data for Inference ResNet50 (4 H100-NVL GPUs on TX2550 M7). The graph confirms near-linear scaling, and the 8 GPU value estimated based on measured data aligns with published MLPerf results.

Number of GPUs	Score [QPS]	GB/s	Remark
8	511,458	76.99	Estimated Value
7	447,526	67.37	
6	383,594	57.74	
5	319,661	48.12	
4	255,729	38.49	Measured Value
3	191,458	28.82	
2	127,839	19.24	
1	63,594	9.57	



[4] MLPerf Inference Benchmark : https://arxiv.org/pdf/1911.02549

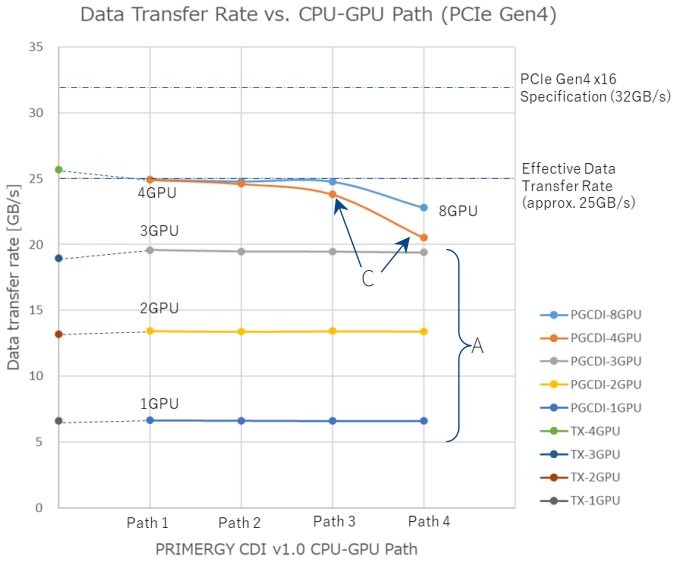
[5] Each benchmark is defined by a Dataset and Quality Target. : https://mlcommons.org/benchmarks/inference-datacenter/

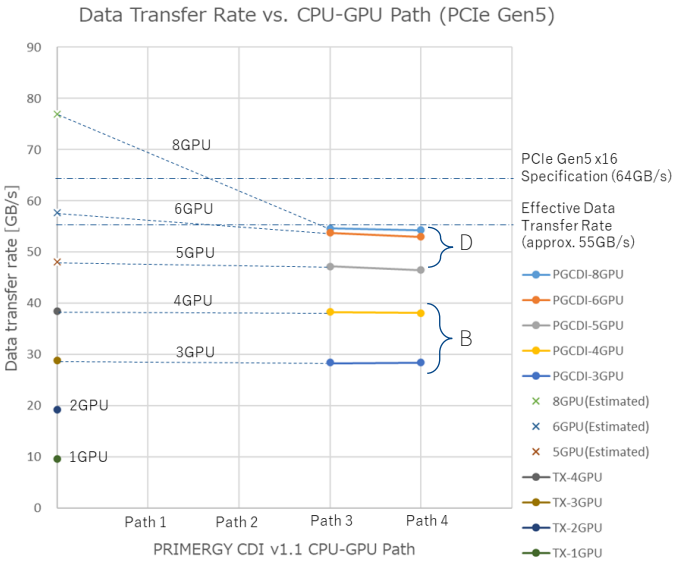
[6] Binary search : https://en.wikipedia.org/wiki/Binary_search

6. Benchmark Performance Measurement Results for CPU-GPU Interconnect

The following explains the graph below.

- Vertical axis: This indicates the data transfer rate per second, calculated as the Inference ResNet50 benchmark test score (queries/sec) multiplied by the image size per query (224x224x3 Bytes).
- Horizontal axis: This indicates the CPU-GPU interconnect path. In the PRIMERGY CDI used in this evaluation, the shortest path is labeled "Path 1," and the longest path is labeled "Path 4." Paths 3 and 4 of PRIMERGY CDI v1.0 have the same CPU-GPU path length (number of PCIe switch hops) as Paths 3 and 4 of v1.1.





PCIe Gen4	CPU-GPU Path	4GPU		3GPU		2GPU		1GPU	
		GB/s	Normalized	GB/s	Normalized	GB/s	Normalized	GB/s	Normalized
PRIMERGY TX2550 M7	—	25.67	1.00	18.95	1.00	13.21	1.00	6.60	1.00
PRIMERGY CDI **V1.0**	Path 1	24.92	0.97	19.57	1.03	13.44	1.02	6.66	1.01
	Path 2	24.60	0.96	19.47	1.03	13.39	1.01	6.64	1.01
	Path 3	23.80	0.93	19.46	1.03	13.42	1.02	6.62	1.00
	Path 4	20.50	C 0.80	19.40	1.02	13.41	1.01	6.62	1.00

PCIe Gen5	CPU-GPU Path	6GPU		5GPU		4GPU		3GPU	
		GB/s	Normalized	GB/s	Normalized	GB/s	Normalized	GB/s	Normalized
PRIMERGY TX2550 M7	—	57.74	1.00	48.12	1.00	38.49	1.00	28.82	1.00
PRIMERGY CDI **V1.1**	Path 3	53.75	D 0.93	47.22	0.98	38.28	0.99	28.43	0.99
	Path 4	53.01	0.92	46.50	0.97	38.14	0.99	28.46	0.99

※Normalized: Data transfer rate relative to TX2550 M7 (1.00)

The preceding graphs and tables reveal the following:

- PRIMERGY CDI connects the compute server and PCIe fabric switch via PCIe Gen4 x16 (32GB/s) or PCIe Gen5 x16 (64GB/s), enabling flexible configuration changes. It was confirmed that this architecture has a limitation in the CPU-GPU interconnect speed.
- Below the effective data transfer rate, performance equivalent to a configuration without PCIe switches between the CPU and GPU is achievable in the Gen4 environment with 3 or fewer GPUs (A in the graph) and in the Gen5 environment with 4 or fewer GPUs (B in the graph).
- In the Gen4 environment, significant performance degradation is observed from Path 3 to Path 4 at the 4-GPU data transfer rate. Path 4 exhibits an approximately 20% performance decrease (C in the graph and table).
- In the Gen5 environment, at data transfer rates involving 5 or more GPUs, Path 4 exhibits a performance decrease of 3 to 8% near the effective data transfer rate (D in the graph and table).

However, this performance degradation does not halt responses (interrupting communication); it only increases response times. Even a minor drop in the 99% latency compliance rate (e.g., from 99% to 98%) renders the benchmark score "INVALID" according to MLPerf rules.

7. Final Remarks

This white paper investigated the impact of CPU-GPU path length (number of PCIe switch hops) on performance in PRIMERGY CDI, which utilizes multiple PCIe switches, and summarizes the results.

- Overview of Investigation
  - Evaluation Method Focusing on Data Transfer Rate for Training/Inference
    - ✓ Training: MLPerf results show no performance degradation due to CPU-GPU path length. This is because GPU learning times are significantly longer than CPU-GPU communication times, rendering PCIe switch latency negligible. Additionally, CPU-GPU communication is at most 4GB/s (12% of Gen4 bandwidth), suggesting path length has minimal impact on typical training.
    - ✓ Inference: Past PRIMERGY CDI results confirm that ResNet50 inference processing, using 4 L40S GPUs in a Gen4 environment, achieves a bandwidth close to the effective data transfer rate of 25GB/s. This suggests that the interconnect is the bottleneck because the ResNet50 inference processing load is too light for the GPU performance, while other benchmarks are limited by GPU performance. We leveraged this characteristic to adjust the data transfer volume between the CPU and GPU by varying the number of GPUs and investigated the effect of CPU-GPU path length.
  - Performance Evaluation Method
    - ✓ Using the Server scenario method of Inference ResNet50, we calculated the data transfer volume per second by multiplying the "QPS value that achieves a Top-1 accuracy of 76.46% and allows 99% of queries to be answered within 15ms," as determined by LoadGen on the compute server, by the image size per query. Binary search was used to automate the QPS value calculation.
- Conclusion
  - In this CPU-GPU interconnect performance evaluation, we investigated the effect of CPU-GPU path length (number of PCIe switch hops) using Inference ResNet50. As a result, we confirmed that performance decreases with increasing path length and data transfer volume. The impact was generally greater in the Gen4 environment and smaller in the Gen5 environment.
  - Training: CPU-GPU path length is unlikely to affect typical AI training, as data transfer volume remains ≤4GB/s even in demanding MLPerf benchmarks. However, future models exceeding 25GB/s (Gen4) or 55GB/s (Gen5) may require further consideration.
  - Inference: Normal AI models are expected to operate without issues and achieve sufficient response performance. Even with particularly high data transfer volumes, such as in the ResNet50 inference benchmark test, the flexible configuration change function of PRIMERGY CDI allows for reducing the data transfer load per virtual server. As a result, CPU-GPU path length is unlikely to affect performance.