Fsas Technologies

PRIMERGY CDI

Performance Comparison of Generative AI Inference Benchmarks

Introduction

Fsas Technologies Inc. offers PRIMERGY CDI, a new server series differing significantly from traditional designs. CDI, or Composable Disaggregated Infrastructure, consists of multiple compute servers, PCIe fabric switches, and PCIe boxes. This design locates GPUs, SSDs, and NICs in external PCIe boxes, separate from the compute server chassis.

One of the most outstanding features of PRIMERGY CDI is its flexible allocation of devices within the PCIe boxes to multiple compute servers. This allows for performance scaling; for example, adding GPUs proactively to handle anticipated increases in inference workloads.

Our previous white papers [1] showed a linear increase in performance with the number of GPUs (up to ten NVIDIA® L40S GPUs) in ResNet benchmarks.

We executed generative AI inference benchmarks on PRIMERGY CDI, using up to sixteen NVIDIA® L40S GPUs, and submitted the results to MLPerf™[2] Inference v4.0[3]. PRIMERGY CDI system achieved higher performance than the similarly submitted PRIMERGY GX2560 M7 system (four NVIDIA® H100-SXM GPUs). This white paper details these results.

This white paper is organized as follows:

- System configurations
- Generative AI Benchmark Results
- Summary and Analysis

PRIMERGY GX2560 M7

PRIMERGY CDI

[1] https://www.fujitsu.com/global/imagesgig5/cdi_whitepaper_scalalibity_en.pdf

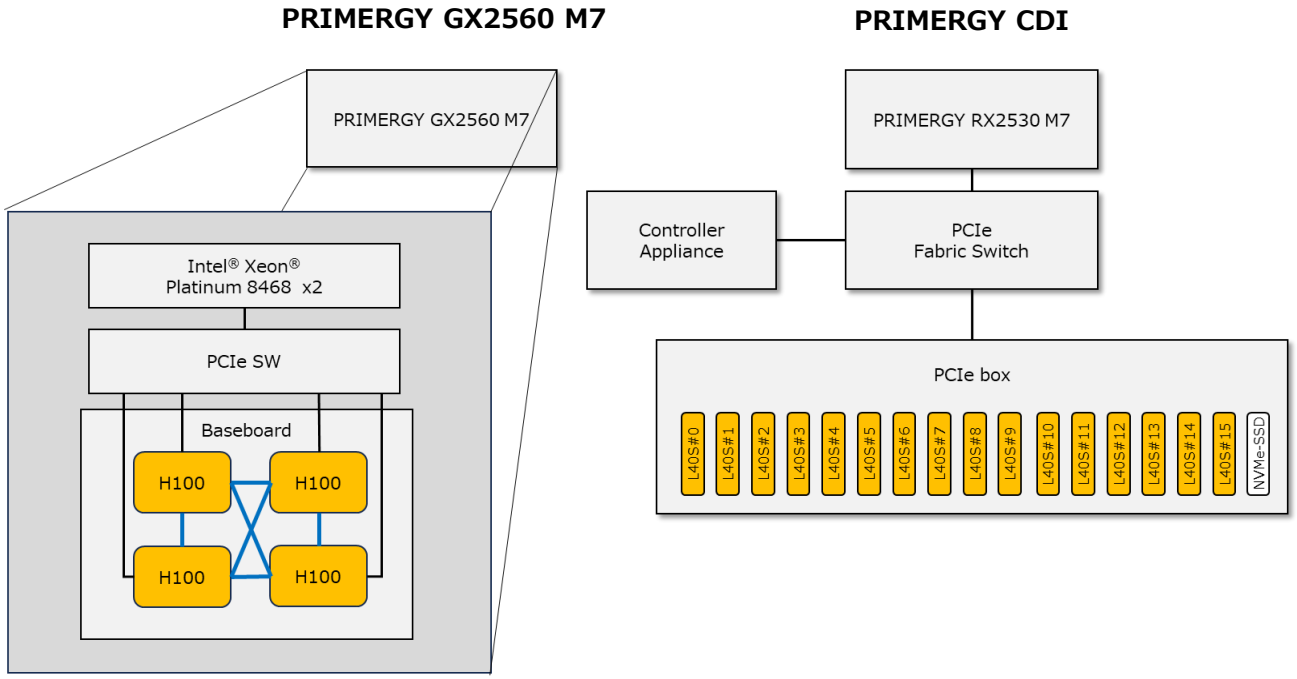
[2] MLPerf™ name and logo are trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.

[3] We submit to MLPerf™ under the Fujitsu name. https://mlcommons.org/benchmarks/inference-datacenter/

1. System configurations

Block Diagram Comparison

The diagrams below show the configurations of the two systems compared. PRIMERGY GX2560 M7 system has four NVIDIA® H100-SXM-80GB GPUs in a single chassis, connected to the CPU via an internal PCIe switch.
PRIMERGY CDI consists of a compute server (PRIMERGY RX2530 M7), a PCIe fabric switch, PCIe boxes (containing sixteen NVIDIA® L40S GPUs and NVMe SSDs), and a controller appliance.
A detailed specification table is provided below.



System Specification Comparison

Server		PRIMERGY GX2560 M7	PRIMERGY CDI
CPU		Intel® Xeon® Platinum 8468 x2	Intel® Xeon® Platinum 8452Y x2
	Frequency	2.1GHz	2.1GHz
	Core Count	48	32
Memory		32x 32GB DDR5	16x 16GB
Storage		877GB (NVMe SSD) + 14TB (SATA SSD)	745.2GBx8 NVMe SSD
Interconnect		PCIe Gen5 x16	PCIe Gen4 x16
GPU		NVIDIA® H100-SXM-80GB x4	NVIDIA® L40S x8,16
OS		Ubuntu 20.04.4	Ubuntu 20.04.4
Software		CUDA 12.2	CUDA 12.2
		cuda_driver_version: 535.129.03	cuda_driver_version: 535.129.03
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

2. Generative AI Benchmark Results

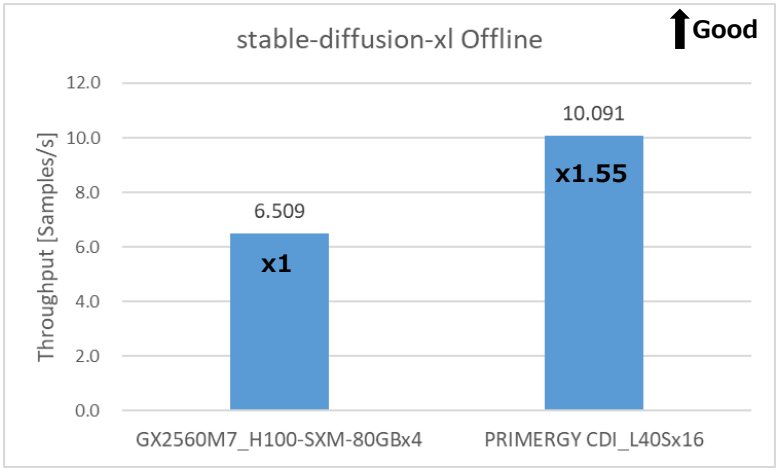
The following table presents a summary of results from the publicly available MLPerf™ [2] Inference v4.0 [3] results. Specifically, we have extracted the benchmark scores for the image generation benchmark, stable-diffusion xl [4] (referred to as SDXL), and the language generation benchmark, gptj-99 [5] (referred to as gptj).

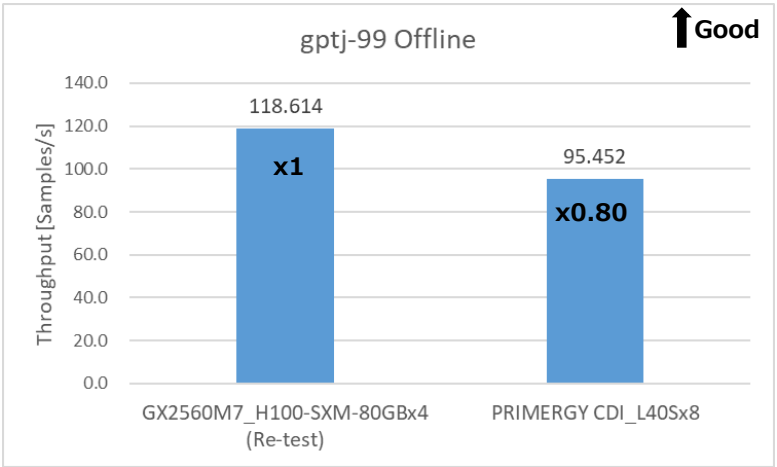
- Public ID: This column corresponds to the unique identifier for each result in the official MLPerf™ [2] results.
- gptj 4.0-0041: This submission used 8 GPUs because the benchmark test implementation did not fully support a larger number of GPUs.
- Re-test [6]: This row presents results from a post-submission re-measurement of gptj 4.0-0043 using the latest software version to ensure consistency with other submissions.

※ Units: Server – Queries/s, Offline – Samples/s

Public ID	System	Processor		GPU		stable-diffusion-xl		gptj-99	
		name	#	name	#	Server	Offline	Server	Offline
4.0-0040	PRIMERGY CDI (16x L40S, TensorRT)	Intel® Xeon®Platinum 8452Y	2	NVIDIA L40S	16	10.116	10.091	–	–
4.0-0041	PRIMERGY CDI (8x L40S, TensorRT)	Intel® Xeon®Platinum 8452Y	2	NVIDIA L40S	8	–	–	95.797	95.452
4.0-0043	GX2560 M7_H100_SXM_80GBx4 (4x H100-SXM-80GB, TensorRT)	Intel® Xeon®Platinum 8468	2	NVIDIA H100-SXM	4	6.118	6.509	85.908	112.015
Re-test[6]						–	–	115.298	118.614

The following figure graphs the results shown in the table above. The stable-diffusion-xl image generation benchmark shows a 1.55x performance improvement with sixteen NVIDIA® L40S GPUs compared to four NVIDIA® H100-SXM GPUs. Similarly, the gptj-99 language generation benchmark shows potential for improved performance with increased GPU count.





[4] stable-diffusion xl, https://github.com/mlcommons/inference/tree/master/text_to_image

[5] gptj-99, https://github.com/mlcommons/inference/tree/master/language/gpt-j

[6] Unverified MLPerf™ v4.0 Inference Datacenter gptj. Result not verified by MLCommons Association.

3. Summary and Analysis

- Submission Details
  - We submitted results to MLPerf™[2] Inference v4.0[3] using PRIMERGY CDI with up to sixteen NVIDIA® L40S GPUs. For comparison, we also submitted results for PRIMERGY GX2560 M7 with four NVIDIA® H100-SXM GPUs.
  - This is the first submission to MLPerf™ Inference using NVIDIA® L40S GPUs (including submissions from other vendors). Due to limitations in benchmark test implementations supporting NVIDIA® L40S, our submissions for the generative AI benchmarks (SDXL and gptj) used 16 GPUs for SDXL and 8 GPUs for gptj.

- Performance Comparison Results

  The table below presents the benchmark results from this white paper, normalized to the PRIMERGY GX2560 M7 score (set to 1). The results show that deploying multiple NVIDIA® L40S GPUs achieved higher performance than the NVIDIA® H100-SXM GPUs.

System	GPU	#	SDXL	gptj	gptj (Re-test)
PRIMERGY GX2560 M7	NVIDIA H100-SXM-80GB	4	1.00	1.00	1.00
PRIMERGY CDI	NVIDIA L40S	16	1.55	—	—
		8	—	0.85	0.80

- Conclusion
  - PRIMERGY CDI achieves performance comparable to systems using high-performance NVIDIA® H100-SXM GPUs by leveraging multiple, more cost-effective NVIDIA® L40S GPUs.
  - PRIMERGY CDI offers flexible GPU scaling for inference workloads. It allows you to dynamically assign PCIe devices within the box to multiple compute servers, readily increasing or decreasing GPU resources as needed. Unused GPUs can be allocated to other servers.
  - PRIMERGY CDI allows for a phased deployment strategy. Users can start with a minimal GPU configuration and incrementally add more GPUs as needed, balancing initial investment with future performance requirements.

◆ About Trademarks
- 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 (® , ™).

◆ 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 Fsas Technologies Inc., 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.