PRIMERGY CDI
生成系AIベンチマーク（推論）の性能比較

はじめに

エフサステクノロジーズ株式会社は、従来のサーバー製品とは異なる革新的な新シリーズ「PRIMERGY CDI」を提供しています。 CDIは Composable Disaggregated Infrastructureの略称です。この製品は、計算サーバー、PCIe fabric switch、PCIe boxから構成されています。GPU、SSD、NICなどのデバイスは、計算サーバーの筐体内ではなく、外部のPCIe boxに収納されます。

PRIMERGY CDIの最大の特徴は、PCIe box内のデバイスを複数の計算サーバーへ自由に割り当てられることです。これにより、例えば推論処理で負荷の増加が予想される場合、事前にGPUを増設することで、性能を向上させることが可能になります。

これまで発表したホワイトペーパーでは、一般的な画像系のResNet学習ベンチマークテストプログラムにおいて、最大10枚のGPU（NVIDIA®L40S）を用いて、スループット性能がGPUの数に応じて向上していることを示しました。

今回、弊社はPRIMERGY CDIに最大16枚のGPU（NVIDIA®L40S）を搭載し、生成系のベンチマーク（推論）を動作させ、MLPerf™ [1] Inference v4.0に投稿[2]しました。その結果、PRIMERGY CDIは、同じく投稿したPRIMERGY GX2560M7（4×NVIDIA®H100-SXM）の処理性能を上回るスコアを達成しました。本ホワイトペーパーでは、これらの内容について説明します。

本書の構成は以下の通りです。

- システム構成（GX2560M7とPRIMERGY CDI）
- 生成系AIベンチマークの結果
- まとめ

PRIMERGY GX2560M7

PRIMERGY CDI

[1] 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] MLPerf™への投稿は「Fujitsu」として行っております。https://mlcommons.org/benchmarks/inference-datacenter/

PRIMERGY CDI 生成系AIベンチマーク（推論）の性能比較

１．システム構成

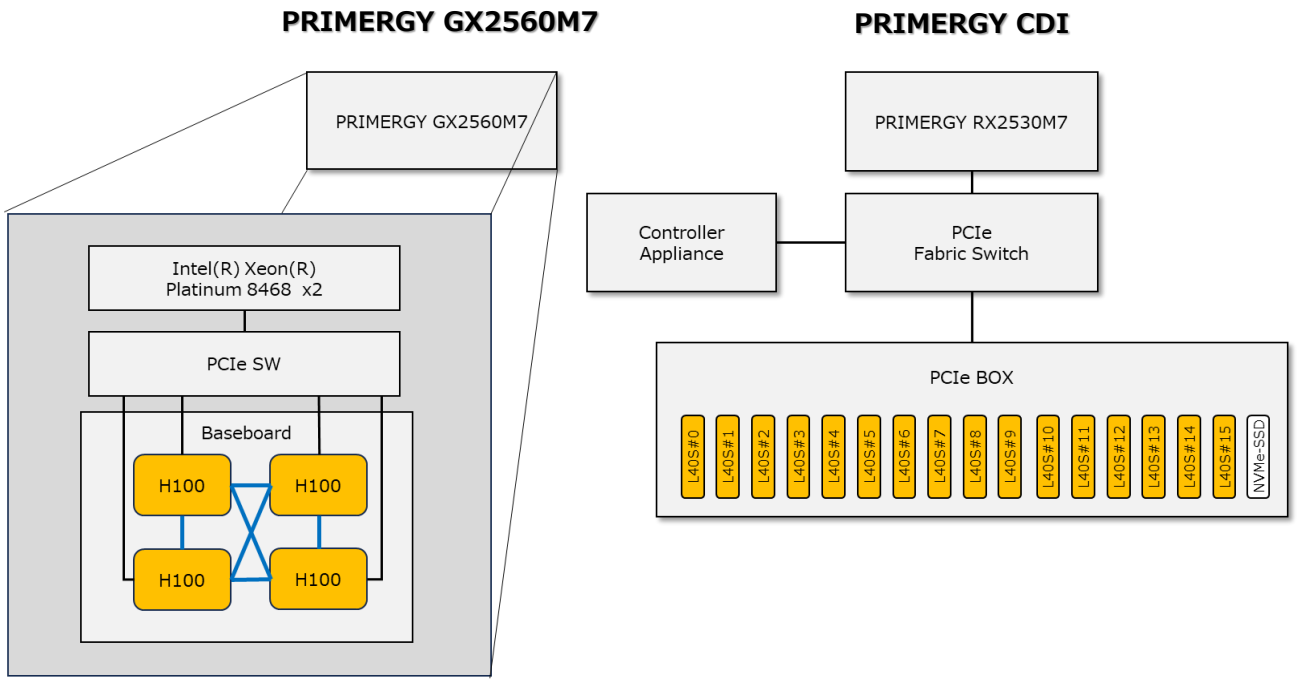
Block Diagram比較

今回比較した2つのシステム構成は、以下の図のようになります。

PRIMERGY GX2560M7は、1つの筐体内に4枚のNVIDIA®H100-SXM-80GBを搭載しており、同じ筐体内のPCIe switch経由でCPUに接続されています。

PRIMERGY CDIは、計算サーバーとしてPRIMERGY RX2530M7を使用し、独立した筐体のPCIe fabric switch、PCIe box、controller applianceから構成されています。PCIe boxには、16枚のGPU（NVIDIA®L40S）とNVMe-SSDが搭載されています。

それぞれのシステムの詳細な仕様については、以下の表をご覧ください。



システム仕様比較

Server		PRIMERGY GX2560M7	PRIMERGY CDI
CPU		Intel(R) Xeon(R) Platinum 8468 x2	Intel(R) Xeon(R) 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-SXM5-80GB 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．生成系AIベンチマークの結果

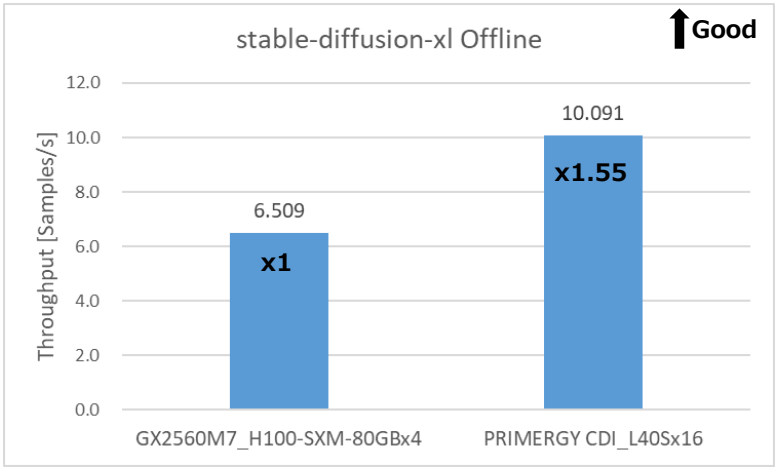
以下の表は、公開されたMLPerf™ Inference v4.0 Resultsから、画像生成系AIベンチマークのstable-diffusion xl [3]（以降SDXLと表記）と言語生成系AIのgptj-99[4]（以降gptjと表記）の結果を抜粋してまとめたものです。

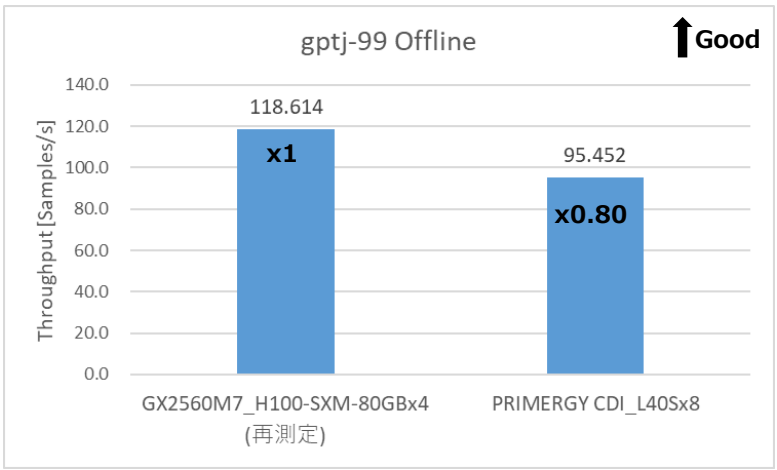
- 表にあるPublic ID列は、公開されているResultsの識別IDです。
- 4.0-0041のgptjは、ベンチマークテストの都合で、8GPUで投稿しています。
- 一番下の再測定[5]の行は、投稿締切後に4.0-0043のgptjに対して他社と同じ最新版を用いて再測定した結果になります。

※単位、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(R) Xeon(R) Platinum 8452Y	2	NVIDIA L40S	16	10.116	10.091	–	–
4.0-0041	PRIMERGY CDI (8x L40S, TensorRT)	Intel(R) Xeon(R) Platinum 8452Y	2	NVIDIA L40S	8	–	–	95.797	95.452
4.0-0043	GX2560M7_H100_SXM_80GBx4 (4x H100-SXM-80GB, TensorRT)	Intel(R) Xeon(R) Platinum 8468	2	NVIDIA H100-SXM	4	6.118	6.509	85.908	112.015
再測定[5]						–	–	115.298	118.614

以下の図は、上記表の結果をグラフ化したものです。画像生成系AIのSDXLでは、16枚のNVIDIA®L40Sを搭載することで、4枚のNVIDIA®H100-SXMを搭載したシステムと比べて1.55倍の性能が得られました。言語生成系AIベンチマークのgptjでも、GPU数を増やすことでSDXLと同様に、性能向上が期待できます。





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

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

[5] MLPerf™によるレビューは受けていません。Unverified MLPerf™ v4.0 Inference Datacenter gptj. Result not verified by MLCommons Association.

３．まとめ

- 投稿内容
  - PRIMERGY CDI に最大16枚のNVIDIA®L40S を搭載し、MLPerf™ Inference v4.0に投稿しました。性能比較の対象として、PRIMERGY GX2560M7+4×NVIDIA®H100-SXMも投稿しました。
  - NVIDIA®L40Sを用いたMLPerf™ Inferenceへの投稿は、他社含めて初めてとなります。NVIDIA®L40Sに対応するベンチマークテスト側実装が十分でない状況下で、生成系AIベンチマークのSDXLと gptj に投稿しました。SDXLは16GPU、gptj は8GPUでの投稿となりました。

- 性能比較の結果
  - 本ホワイトペーパーで用いたベンチマーク結果を、PRIMERGY GX2560M7の性能スコアを 1 としてPRIMERGY CDIのスコアを正規化して以下の表にまとめました。
  - 表からわかるように、NVIDIA®L40S を複数枚動作させることで、NVIDIA®H100-SXMを上回る性能スコアを実現することができました。

System	GPU	#	SDXL	gptj	gptj 再測定
PRIMERGY GX2560M7	NVIDIA H100-SXM-80GB	4	1.00	1.00	1.00
PRIMERGY CDI	NVIDIA L40S	16	1.55	–	–
		8	–	0.85	0.80

- 今回の結果から
  - PRIMERGY CDIは、生成系AIの推論において、比較的安価なNVIDIA®L40S を複数枚利用することで、高性能なNVIDIA®H100-SXMに匹敵する性能を出すことが可能です。
  - PRIMERGY CDIは、PCIe box内のデバイスを複数の計算サーバーに自由に割り当てることができるため、推論の負荷に合わせてGPU数の増減を柔軟に行えます。余ったGPUは他の計算サーバーに割り当てることも可能です。
  - PRIMERGY CDIは、導入当初は少ないGPUで始め、必要に応じてGPUを追加することで性能を向上させることができます。これにより、初期投資を抑えながら、将来的な性能向上を見据えた導入が可能になります。