Fujitsu Server PRIMERGY CDI 性能について

Fujitsu Server PRIMERGY CDI（以降、PRIMERGY CDI）は、GPUアクセラレータ、ストレージ、ネットワーク（NIC）などのPCIeデバイスリソースを、Server内部ではなくPCIe BOXに収納し、Serverの外に配置します。さらに、高速なPCIe Fabric Switchを用いてServerとPCIe BOXを接続し、効率的な構造を実現しています。 専用ソフトウェアによる管理機能を活用し、お客様のワークロードに応じてリソースを自由に配備・解放できるため、リソース使用率を最大化し、効率的な運用が可能です。

PRIMERGY CDIでは、 新たにHBA(Host Bus Adapter) 、PCIe Fabric Switch、PCIe BOXが新たに加わります。本資料では、これらが性能に与える影響について説明します。

説明は、PRIMERGY CDIとリファレンス機としてPRIMERGY RX2540 M7（以降、RX2540 M7）を用いて、比較しながら以下の順番で行います。

- GPU間通信の性能評価の内容と結果を説明。
- 実際にMLPerf™ [1] のベンチマークであるResNetを動かしてThroughput性能を測定し、NVIDIA Nsight™ Systems [2]を用いて分析を行った結果を説明。
- MLPerf™ Training v3.0 ResNetを、MLPerf™ のルールに従って取得したベンチマーク結果を説明。

なお、今回はResNet等のアプリケーションの学習で重要なGPU間通信を中心に評価を致しました。CPU-GPU間についての性能評価は、別途実施する予定です。

[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.

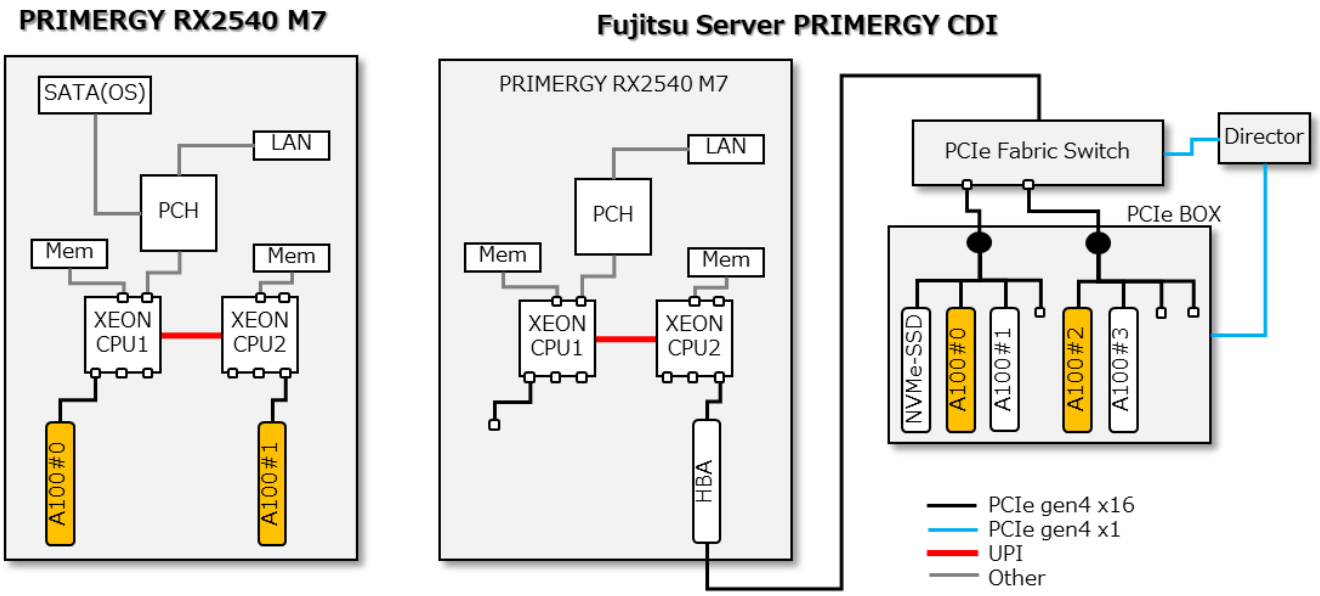
[2] https://developer.nvidia.com/nsight-systems

Fujitsu Server PRIMERGY CDI性能について

1.　システム構成

Block Diagram比較

今回比較したシステム構成は、以下の図になります。PRIMERGY CDIは、RX2540 M7のNVIDIA® A100-PCIeのスロットにHBAを取り付け、そのHBA下にPCIe Fabric Switch、PCIe BOX(NVIDIA® A100-PCIe、NVMe-SSD)を接続しています。一方リファレンス機のRX2540 M7は、２つのCPUの各々に NVIDIA® A100-PCIeを接続する構成になっています。この２つの構成で性能測定を行い、HBA－PCIe Fabric Switch－PCIe BOXの有無による性能評価行います。



システム仕様比較

サーバ		PRIMERGY RX2540 M7	Fujitsu Server PRIMERGY CDI
CPU		Intel(R) Xeon(R) Gold 6430x2	
	周波数	2.1GHz	
	コア数	32	
メモリ		16GBx16	
ストレージ		SATA	8x 800GB NVMe SSD
インターコネクト		PCIe 4.0	
GPU		NVIDIA A100-PCIe-80GBx2	
OS		Red Hat Enterprise Linux release 8.6 (Ootpa)	
ソフトウェア		CUDA: 12.1.0.023	
		cuda_driver_version: 530.30.02	
HBA		–	PCIe HBAカード for CDI (帯域幅 64GB/s (双方向))
PCIe Fabric Switch		–	PCIe ファブリックスイッチ(48port) for CDI x1 (総合帯域幅 768 GB/s 双方向 48 ポート)
PCIe BOX		–	PCIe Box (PCIe×8) for CDI x1 (ポート帯域幅 最大 128GB/s (双方向))
Director		–	コントローラアプライアンス for CDI

2. GPU間通信の性能評価

性能評価ベンチマークテスト

学習時に行われるGPU間通信が、PCIe Fabric Switch、PCIe Boxを経由して行われることから、これらを含めて通信性能を評価するため、以下のベンチマークテストを使用します。

- p2pBandwidthLatencyTest [3]

  p2pBandwidthLatencyTestは、cuda-samplesに含まれるベンチマークテストです。GPU間通信の帯域幅とLatencyを測定します。

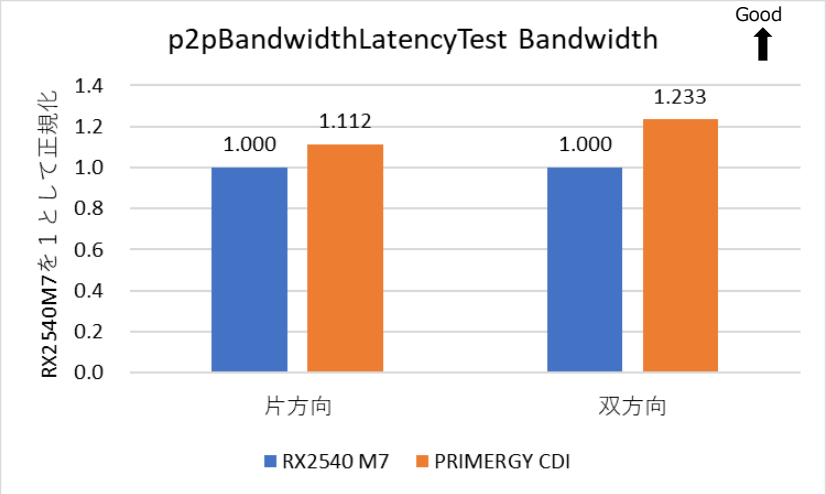
- nccl-tests [4]

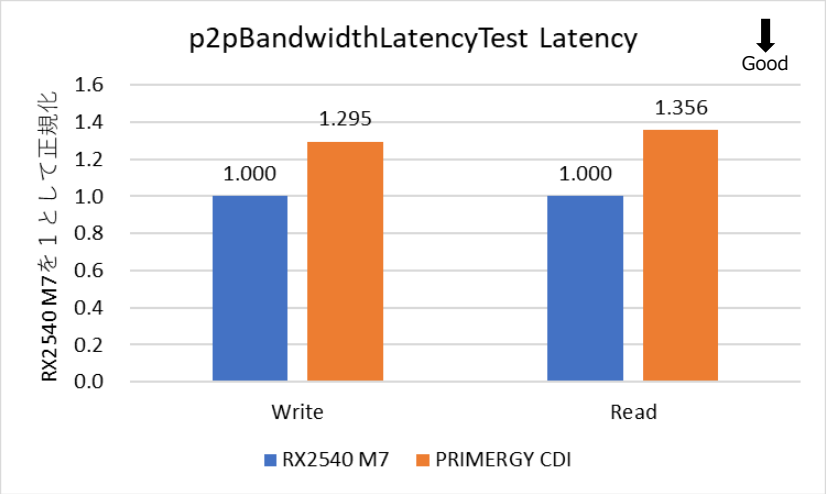
  nccl-testsは、学習で使用するAllReduceを実装しているNCCLの性能を測定するベンチマークテストです。

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

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

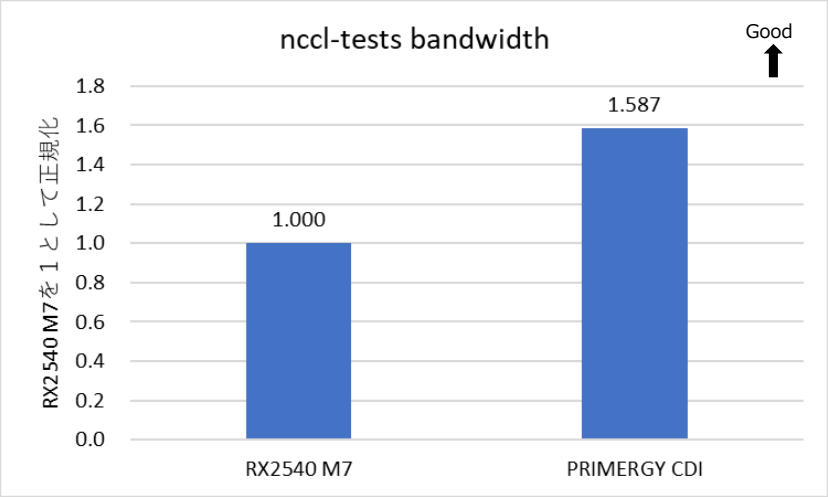
p2pBandwidthLatencyTest性能評価





Fujitsu Server PRIMERGY CDI性能について

nccl-tests性能評価



GPU間通信の性能評価のまとめ

- リファレンス機のRX2540 M7は2個のCPUを経由してGPU間通信するのに対し、PCIe接続だけでGPU間通信をする構成であるPRIMERGY CDI方が広い帯域幅となっています。
- 学習時に動作するNCCLの性能を測るnccl-testsを行うと、RX2540 M7とPRIMERGY CDIとの性能差が大きくなり、AllReduceを行う場合にPRIMERGY CDIの構成の方が有利に働くと予想されます。
- Latencyでは、RX2540 M7に比べPRIMERGY CDIのほうが長くなるため、Latencyが性能に影響することが懸念されます。

Fujitsu Server PRIMERGY CDI性能について

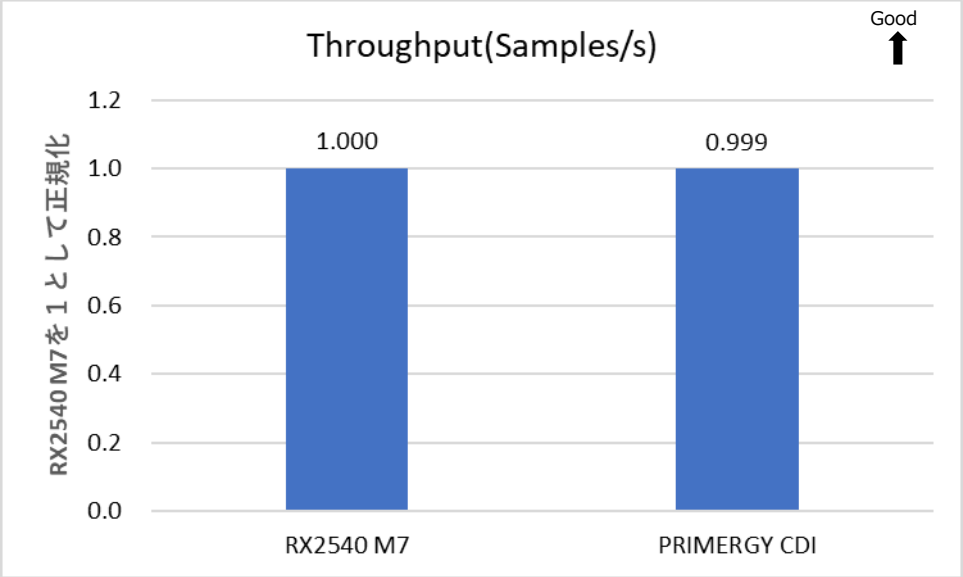
３．　ResNetのThroughput性能評価

３．１　Throughput性能評価

評価方法

評価に用いるResNetは、MLPerf™ Training v2.1 ResNet 、MXNet NVIDIA Release 22.04を使用します。ベンチマークを動作ささせて実行ログを取得し、そのログのタイムスタンプから算出した1epochの処理時間と、Datasetにある学習用画像の枚数の２つから、Throughput(Samples/s)を算出しています。

Throughput性能評価



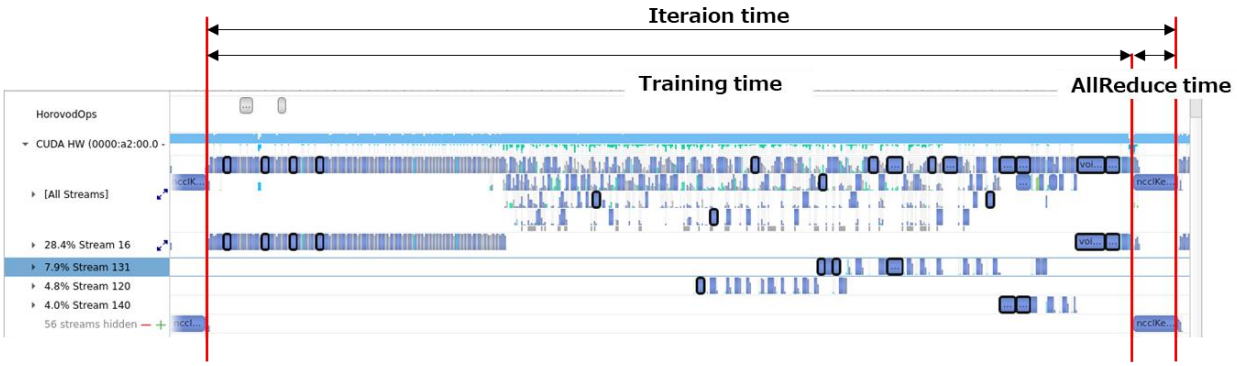
Throughput性能評価のまとめ

- GPU間通信の性能評価から、PRIMERGY CDIはNCCL部分の性能が良いこととLatencyが遅いことが判りました。一方Throughput性能評価では、0.1% 差でRX2540 M7とPRIMERGY CDIが同じ結果になりました。良いところと悪いところが打ち消しあって、このような結果になったものと考えられます。

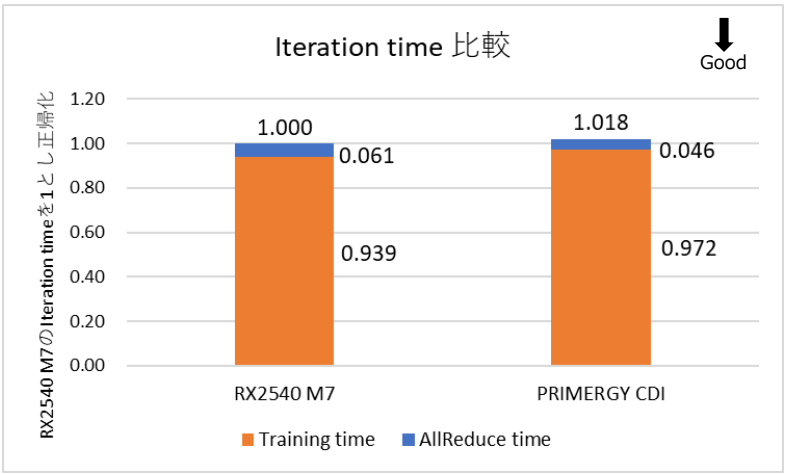
3．2　NVIDIA Nsight™ Systemsによる分析

Nsight™ Systemsによる分析方法

NVIDIA Nsight™ Systemsは、ResNet等のアプリケーションを動作させて詳細なデータを収集し、取得データからアプリケーションの動作を可視化して分析できるツールです。これを利用してThroughput性能の分析で重要な1 Iteration内の各処理の時間を取得します。測定結果の表示画面の一部を、以下に示します。図のようにItereation time 、AllReduce time、そして Itereation timeからAllReduce timeを引いた残りの時間をTraining timeとして、各処理時間を取得データから算出して比較します。



性能評価



Nsight™ Systemsによる分析方法のまとめ

- AllReduce timeでは、PRIMERGY CDIの方が処理時間が短くなっています。nccl-tests評価から、RX2540 M7よりPRIMERGY CDIの方が帯域幅が広い結果が出ており、それに合う結果となっています。
- Training timeでは、RX2540 M7が処理時間が短く、PRIMERGY CDIが長い結果となりました。おそらくLatencyが長くなっていたことがTraining timeに影響しているものと考えられます。
- Iteration timeでは、RX2540 M7に対しPRIMERGY CDIが1.8%増になっています。しかし前ページのThroughput性能が同じ結果であることから、 PRIMERGY CDIの Iteration time 1.8%増は、問題にならない差であるようです。

４．　MLPerf™ベンチマークテスト評価

評価方法

実際にMLPerf™ Training Rules[5]に従ってベンチマークテストを実施し、Submitと同じ手順でスコアを算出します。ResNetの場合、連続5回動作させてログを取得し、動作時間の最大最小を除いた３つのログの動作時間の平均値を今回のスコアとしています。

評価はMLPerf™の最新データと比較するため、現時点で最新のMLPerf™ Training v3.0 ResNet 、MXNet NVIDIA Release 23.04を使用しています。またTraining v3.0において複数のSubmit実績のある４GPUのシステムでPRIMERGY CDIのデータを取得しています。

測定結果

ResNetの測定結果、**60.444分[6]を達成**。MLPerf™ Training v3.0 Results [7]の形式で記載すると、以下のようになります。

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

MLPerf™ベンチマークテスト評価のまとめ

- 最新のベンチマークテストによる測定の結果、PRIMERGY CDIは、MLPerf™ Training v3.0 Results[7]に掲載された同じような仕様の他社と、同等なスコアを得ることができました。PRIMERGY CDIは、従来Serverに対し新たにHBA、PCIe Fabric Switch、PCIe BOXがあるものの、その影響は無いと考えられます。

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

[6] 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.

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

Fujitsu Server PRIMERGY CDI性能について

5．　まとめ

- リファレンス機RX2540 M7との比較であるGPU間通信性能評価、ResNetのThroughput性能評価をまとめると以下になります。個々の項目を見ると新たに加わった構成の影響は確認できますが、ResNetを動かしたトータルのThroughput性能を見ると影響は無いと言えます。

テスト種類	詳細	RX2540 M7		PRIMERGY CDI	
		結果 *3	値 *2	値 *2	結果 *3
p2pBandwidthLatencyTest	片方向通信		1.000	1.112	○
	両方向通信		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：値が小さい方が良いテスト、*2：各テストで測定結果を正規化した値、*3：○は良い結果、または同等を示す。

- MLPerf™の他社スコアと比較するため、ResNetのベンチマークテストを行った結果、PRIMERGY CDIは、同様な仕様の他社スコアと同等なスコアを得ることができました。
- 以上の評価結果から、各テストの詳細を見ると新たに加わった構成の影響を確認できますが、ResNetのThroughput性能と実際のMLPerf™ベンチマークスコアでは、影響は無いと考えられます。結果としてPRIMERGY CDIは、従来と同等の性能を確保しながら、PRIMERGY CDIの特徴である「お客様のワークロードに応じてリソースを自由に配備・解放して、リソース使用率を最大化し、効率的な運用をする」ことを可能にします。