NVIDIA RTX 3000 ベンチマーク報告書

2020/11/10

株式会社ジーデップ・アドバンス

目的

NVIDIA RTX3000 シリーズの Deep Learning に関する性能をプラットフォームシステム毎に評価し、各 GPU の世代による性能の変化に関する知見を得ることを目的とする。

ベンチマーク条件

システムは 2 種類、GPU は Turing 世代 2 種類と Ampere 世代 2 種類、計 4 種類を用いた。各システムには GPU を 2 枚実装し、マルチ GPU 環境でのテストを行った。

フレームワークは GeForce RTX 3000 世代に対応した CUDA-11.1 / cuDNN-11.1- v8.0.4.30 でビルドが可能であった Pytorch-1.6.0 をベースとし、ベンチマークプログラムにはトレーニングとインファレンスを網羅的にテスト可能な Pytorch-benchmark を使用した。

各ベンチマーク指標は、それ自体それぞれのテスト内部で繰り返した平均値で示されているが、その一連のベンチマークを最低 2 セット行い、その平均値を使用した。

Pytorch-benchmark には CPU の性能評価も含まれるが、今回は GPU をターゲットとするもののみ使用した。

諸元を以下に示す。

System :

- Deep Learning Box II (DLBox II)
- TR3/G2

System	DLBox II	TR3/G2
Chipset	X299	TRX40

CPU Version	Intel(R) Core(TM) i9-7900X	AMD Ryzen Threadripper 3970X
Clock Speed	1200 MHz 3300 MHz max	2200 MHz 3700 MHz max
L3 Cache	14,080 kB	131,072 kB
Cores / Threads	10 / 20	32 / 64
Memory Size / Speed	4x 16384 MB / 2400 MT/s	8x 16384 MB / 2933 MT/s

GPU:

- GeForce RTX 2080 Ti
- Titan RTX
- GeForce RTX 3080
- GeForce RTX 3090

GPU	**GeForce RTX 2080 Ti**	**TITAN RTX**	**GeForce RTX 3080**	**GeForce RTX 3090**
PCIe	16x Gen 3	16x Gen 3	16x Gen 4	16x Gen 4
CUDA Capability	7.5	7.5	8.6	8.6
CUDA Cores	4352 (68sm x64)	4608 ‘(72sm x64)	8704 (68sm x128)	10496 (82sm x128)
Memory	11019 MiB	24220 MiB	10014 MiB	24268 MiB
Max Power	250.00 W	280.00 W	320.00 W	350.00 W
Graphics Clock	2100 MHz	2100 MHz	2100 MHz	2100 MHz
Memory Clock	7000 MHz	7001 MHz	9501 MHz	9751 MHz

Operating System:

OS	**Kernel**
Ubuntu 18.04.5 LTS (Bionic Beaver)	5.4.0-48-generic

Software Libraries:

Module
cmake/3.17.1
compiler/gcc-7.5.0

cuda/11.1
cudnn/11.1-v8.0.4.30
ffmpeg/4.3.1
opencv-3.4.11-gcc-7.5.0
python-3.7.7
pytorch-1.6.0
intel-perflib/2020.2
lmdb/0.9.24
nccl/2.7.8/cuda-11.1
openmpi/4.0.5/gcc-7.5.0.lp
protobuf/3.13.0
pytorch-benchmark (git 86a5e8f80d249dbc47a6a2ed9911ecc9df808fb5　2020-10-03)

結果

結果は、各システム上のそれぞれのパフォーマンス指標を RTX 2080 Ti の値を 100%とした相対値で示した。数字が大きいほど性能が高いことを意味する。例えば相対性能 200%は単純に基準に比べて 2 倍高速という意味である。

ベンチマーク指標のうち以下のテストは毎回の値のばらつきが大きすぎ、信用できないため結果からは除外した。

test_train[tacotron2-cuda-eager]

test_eval[tacotron2-cuda-eager]

この他にも多少ばらつきが大きめの指標も存在するが、ここでは全体の傾向としてとらえるため、結果に含めてある。

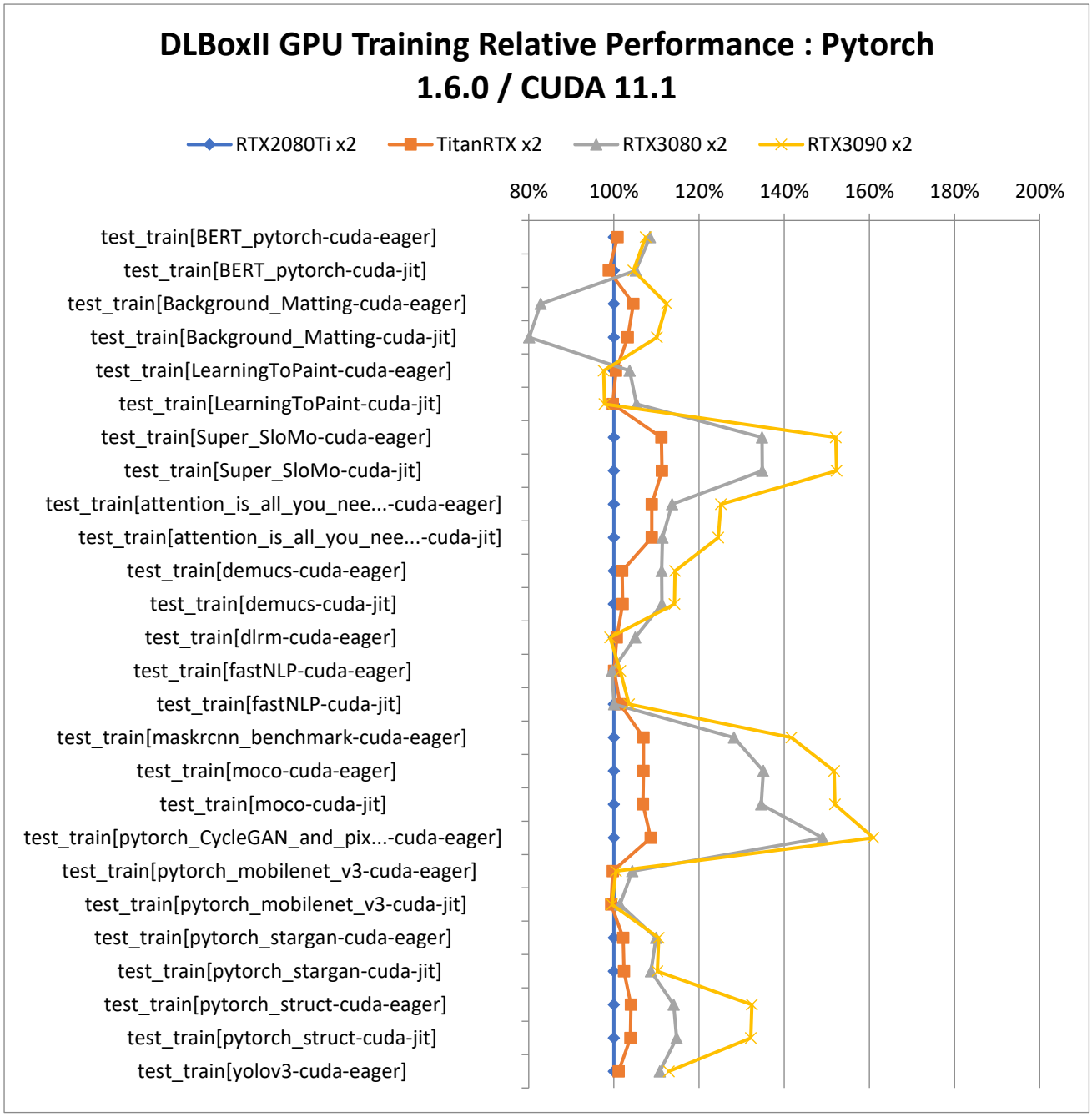

Figure 1 DLBox II Training Relative Performance

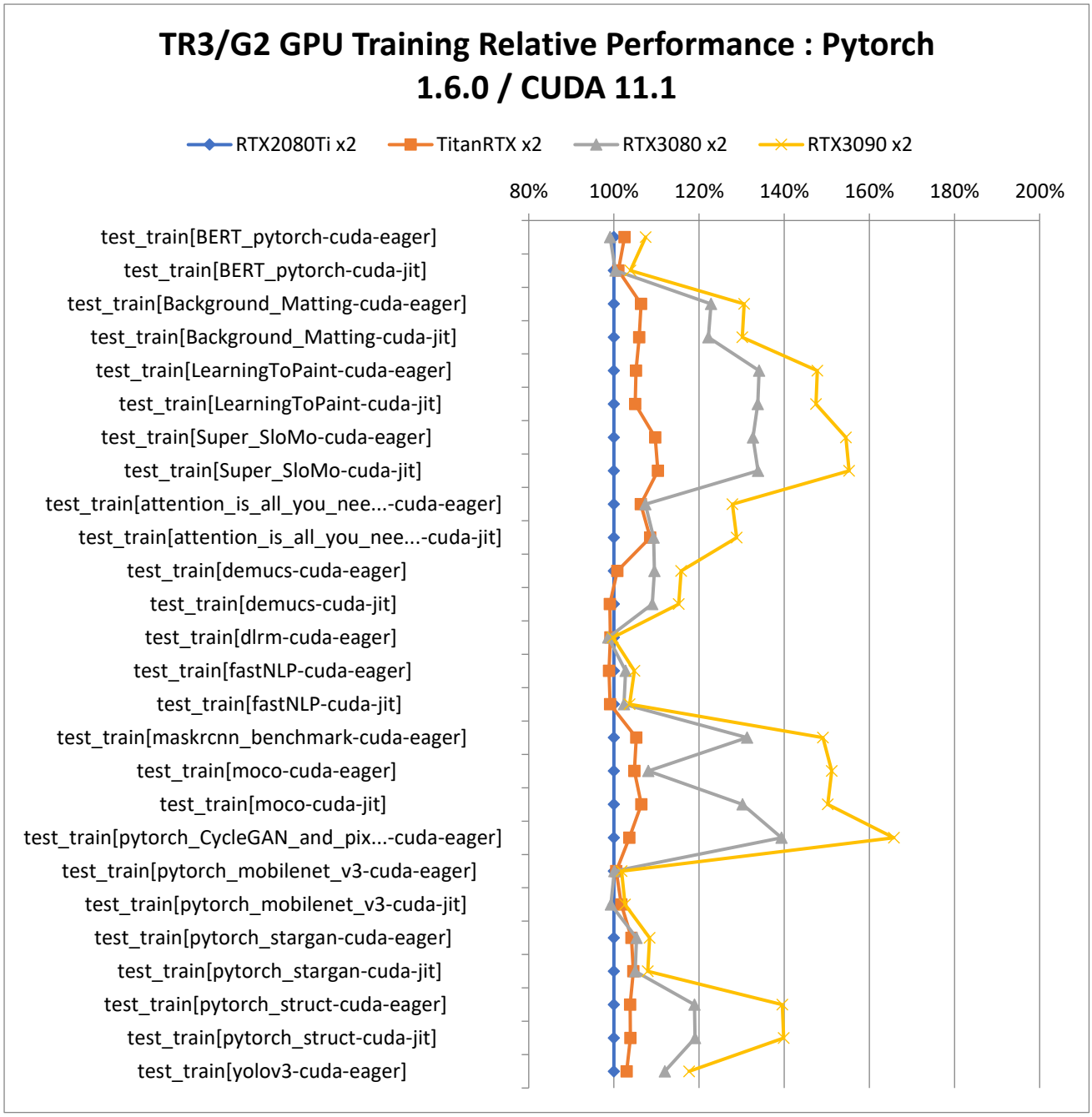

Figure 2 TR3/G2 Training Relative Performance

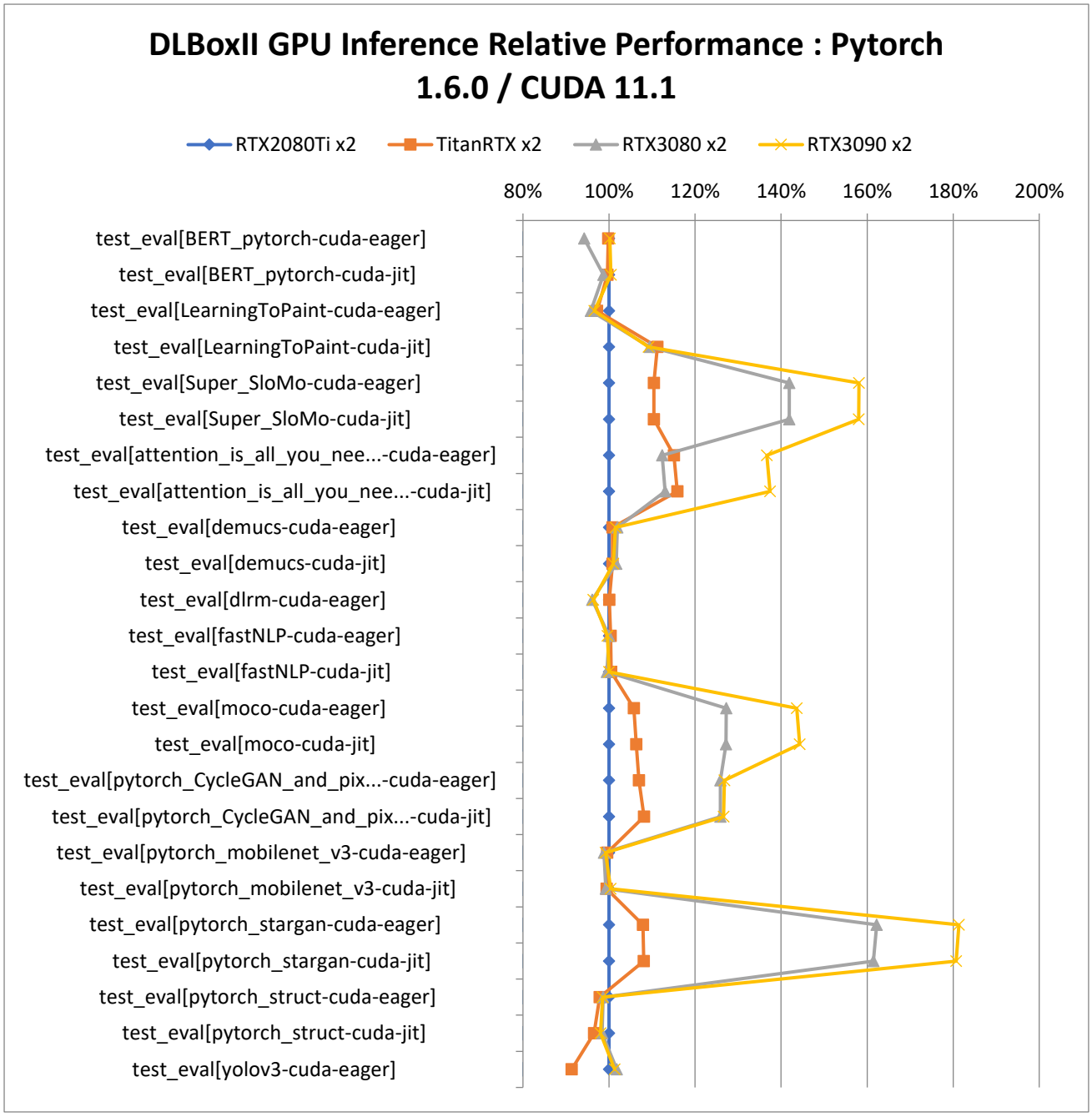

Figure 3 DLBox Inference Relative Performance

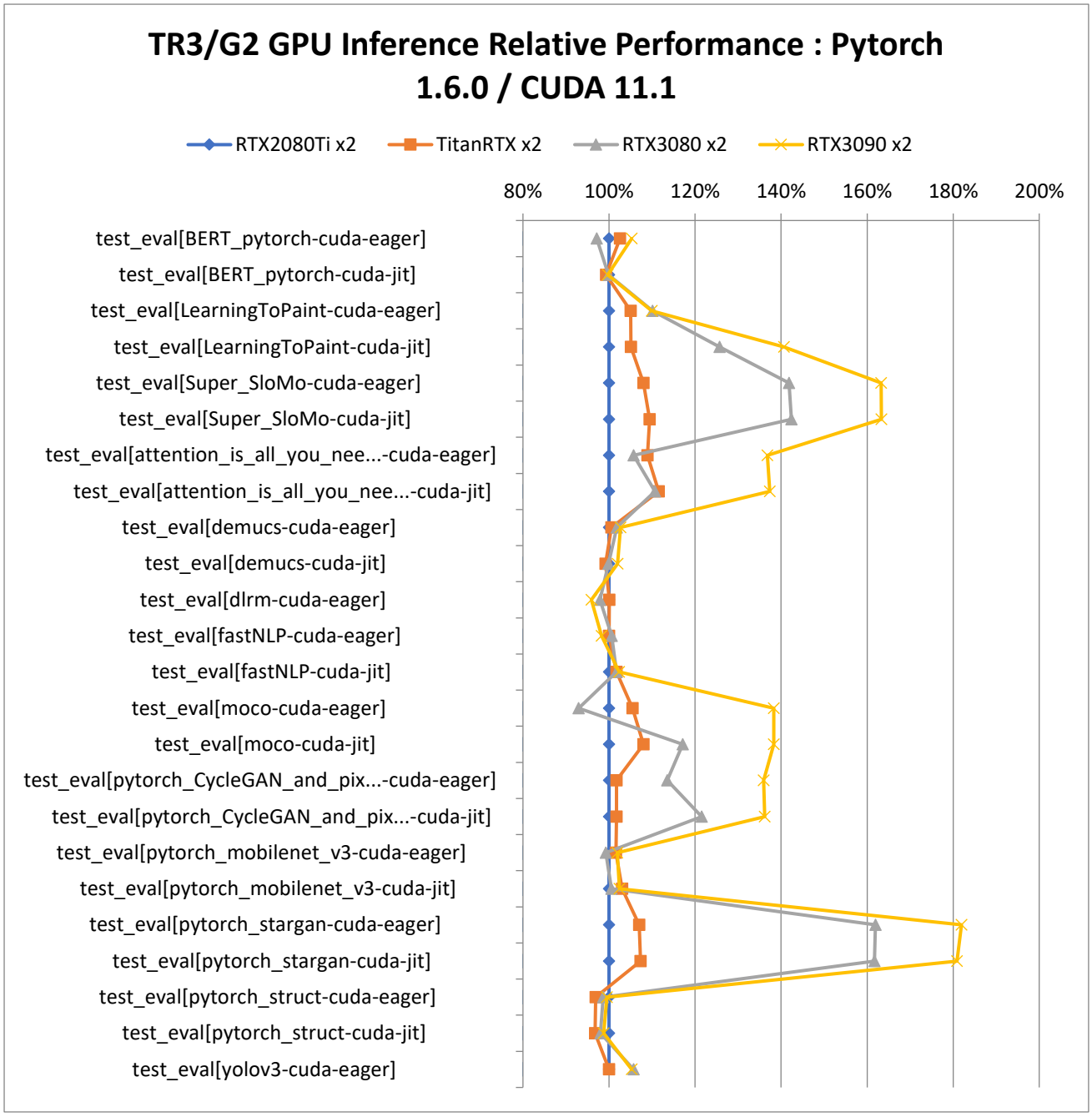

Figure 4 TR3/G2 II Inference Relative Performance

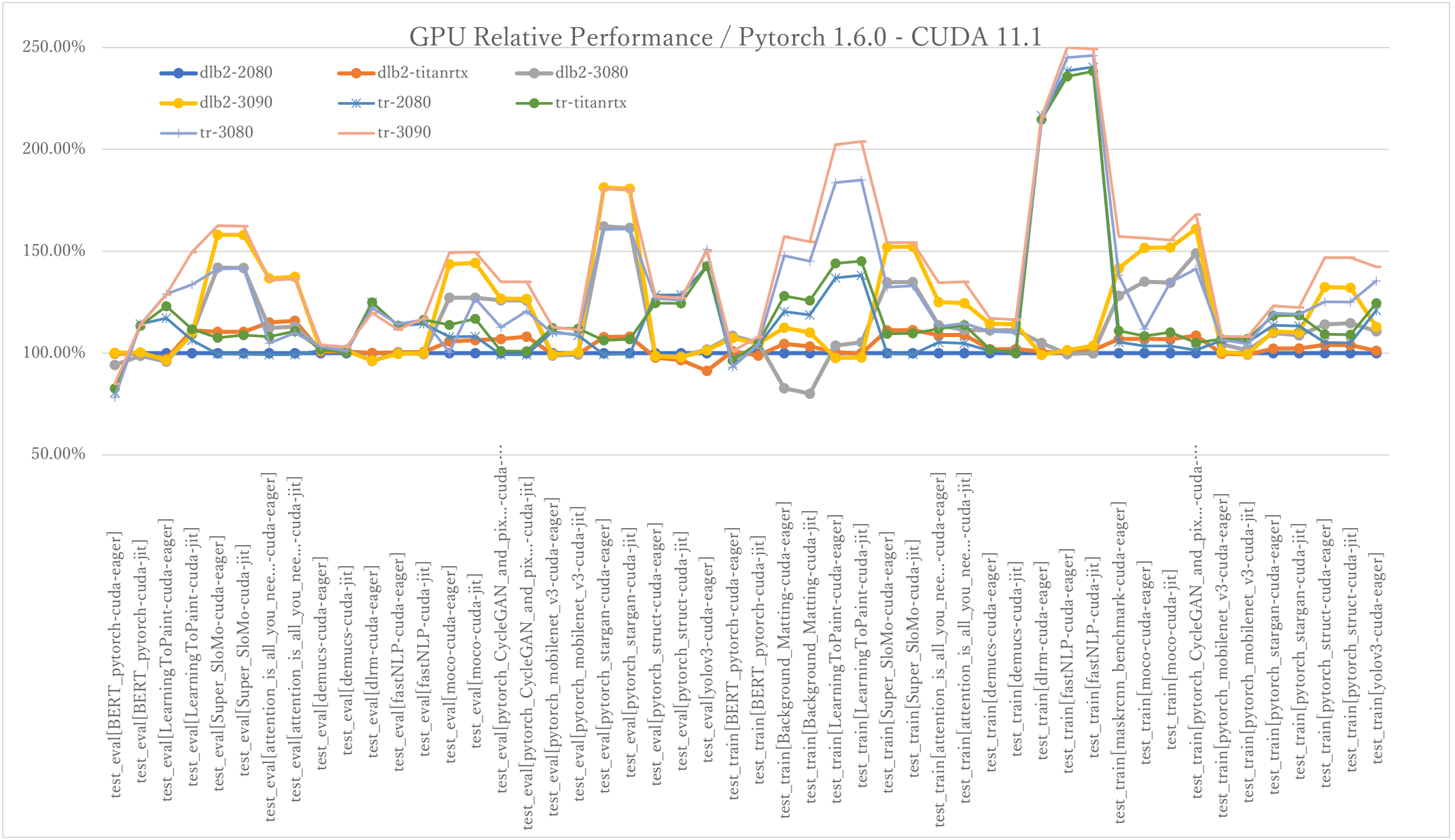

Figure 5 All GPU / All System Relative Performance

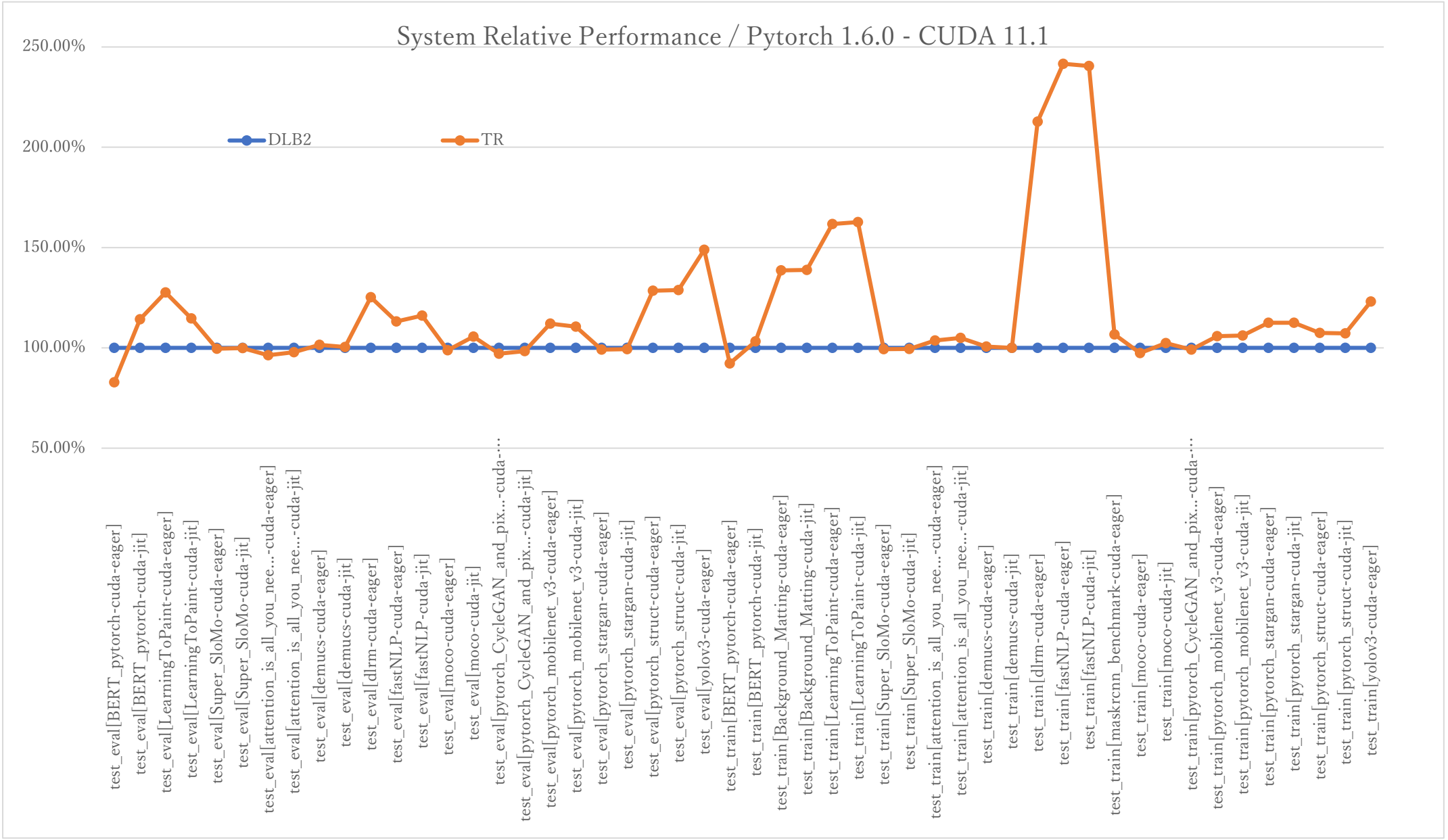

Figure 6 System Relative Performance

考察

トレーニング、インファレンスとも Super_SloMo や stragan、moco などは GeForce RTX3080 / RTX3090 の性能が大きく寄与しています。それ以外のテストについては現在自然言語処理などで多用されている BERT も含め、新しい Ampere アーキテクチャ GPU への最適化についてもう少し時間が要する結果となっています。

一方、システムによる性能差もかなり大きく、Figure-6 は使用した全ての GPU の性能値を平均した場合のシステムによる性能差を示していますが、ここから見られるように、総じて GWS-TR3/G2 はコストパフォーマンスが高いモデルであることがお判りいただけます。

また、DeepLearningBOX II もインテル製 CPU と Gen3 の PCI-Express バスの組み合わせではありますが、実ベンチとしてはトレーニング、インファレンス双方においておいても、また拡張性という部分でも充分に検討に値するスコアを出しています。

結論

全てのトレーニングやインファレンスが新しい GPU アーキテクチャ（Ampere）に対応するにはもう少し時間が掛かりそうですが、一部のトレーニング、インファレンスにおいては GeForce RTX3080 / RTX3090 の性能を充分に引き出しており、価格、メモリ容量など総合的に考慮しても従来の GeForce RTX2080Ti や TITAN RTX と比較し GeForce RTX3080/RTX3090 を導入する効果は小さくないと考えられます。

また、システム側の影響も大きく、コストパフォーマンスという観点からも考慮し、以下の製品を当社のリコメンドとします。

■コストパフォーマンス指向

GWS-TR3/2G

■拡張性・安定性指向

DeepLearningBOXII

以上