# NVIDIA GeForce RTX 3000 Benchmark Report

2020/11/10

GDEP Advance Inc.

## Objectives

To measure the Deep Learning performance characteristics of NVIDIA GeForce RTX3000 series, by comparing the differences in GPU generations as well as platform generations.

## Benchmark Conditions

The combination of two systems and four GPU modules were used for the measurement. Two GPU units were equipped to enable multi-GPU environment.

Pytorch-1.6.0 was used as the framework, since it was only possible to build using CUDA-11.1 and cuDNN-11.1-v8.0.4.30. The benchmark suite chosen was Pytorch-benchmarks as it could measure many networks' performance numbers in one round, both in training in inference.

Each benchmark numbers are already calculated as an average of several test loops. However we executed at least two sets of those benchmark tests and took the average numbers.

We only used GPU performances, although Pytorch-benchmarks could measure CPU based performances.

The specifications are shown below.

**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<br>3300 MHz max | 2200 MHz<br>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) |

## Results

We show each performance numbers normalized by the results of GeForce RTX 2080 Ti as 100%. This means the higher the percentage, the better the performance. For example, 200% means the configuration was twice as fast as the base configuration.

The following two test cases were excluded from the results, since deviation in each test were too large.

- test_train[tacotron2-cuda-eager]
- test_eval[tacotron2-cuda-eager]

Some other test cases also showed relatively high deviation, however those were included to capture the tendency as a whole.

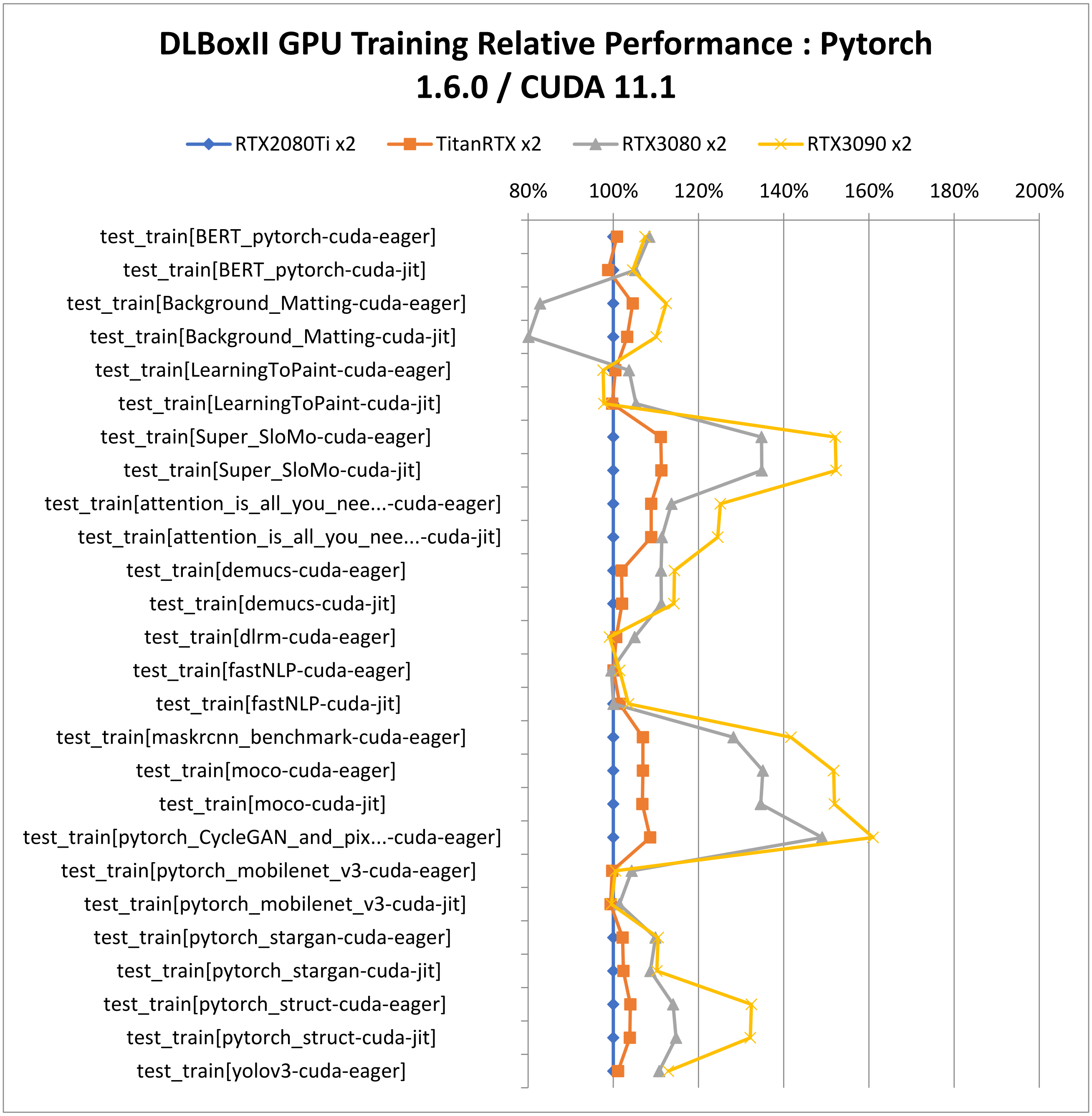

**Figure 1 DLBox II Training Relative Performance**

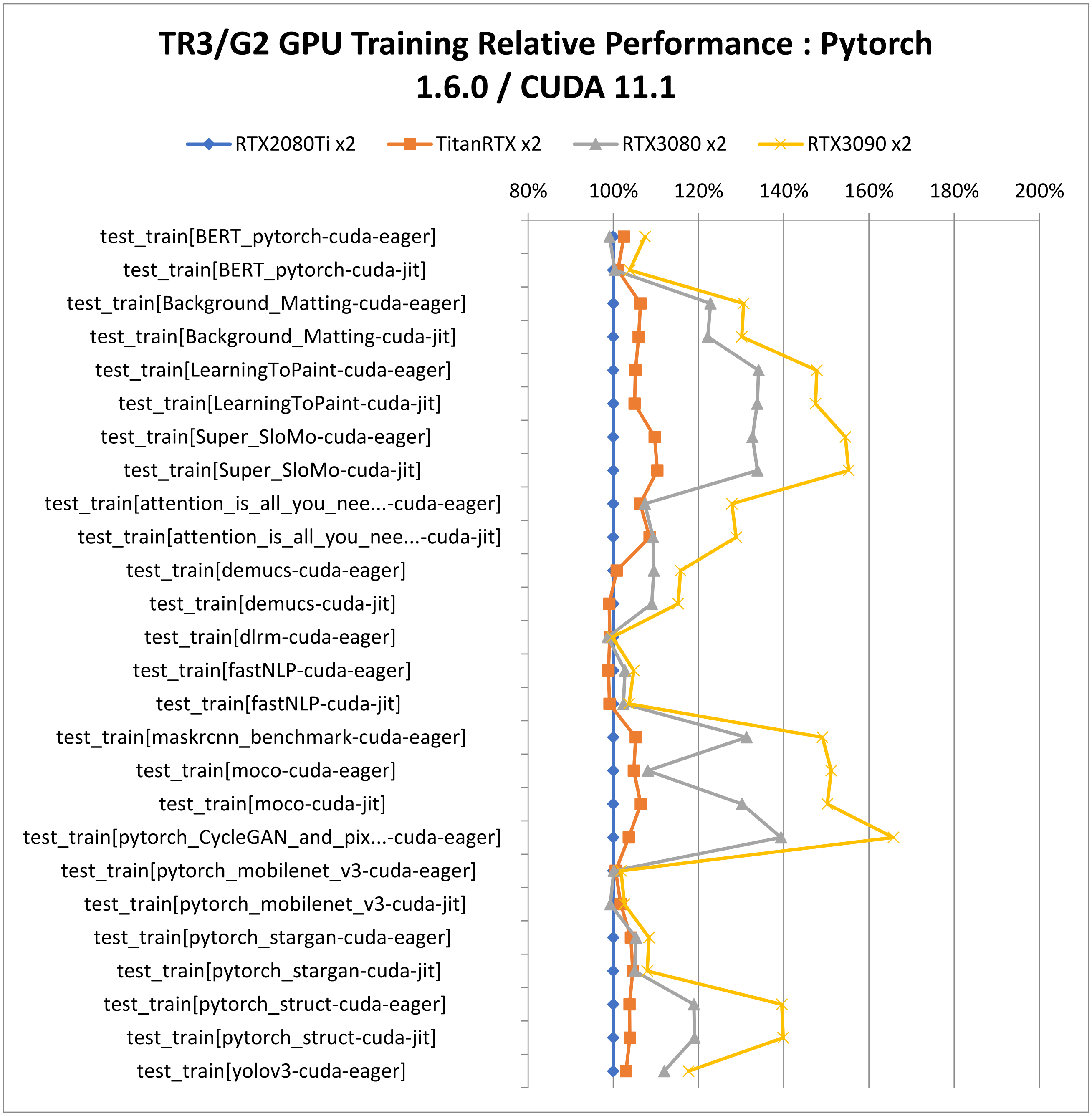

**Figure 2 TR3/G2 Training Relative Performance**

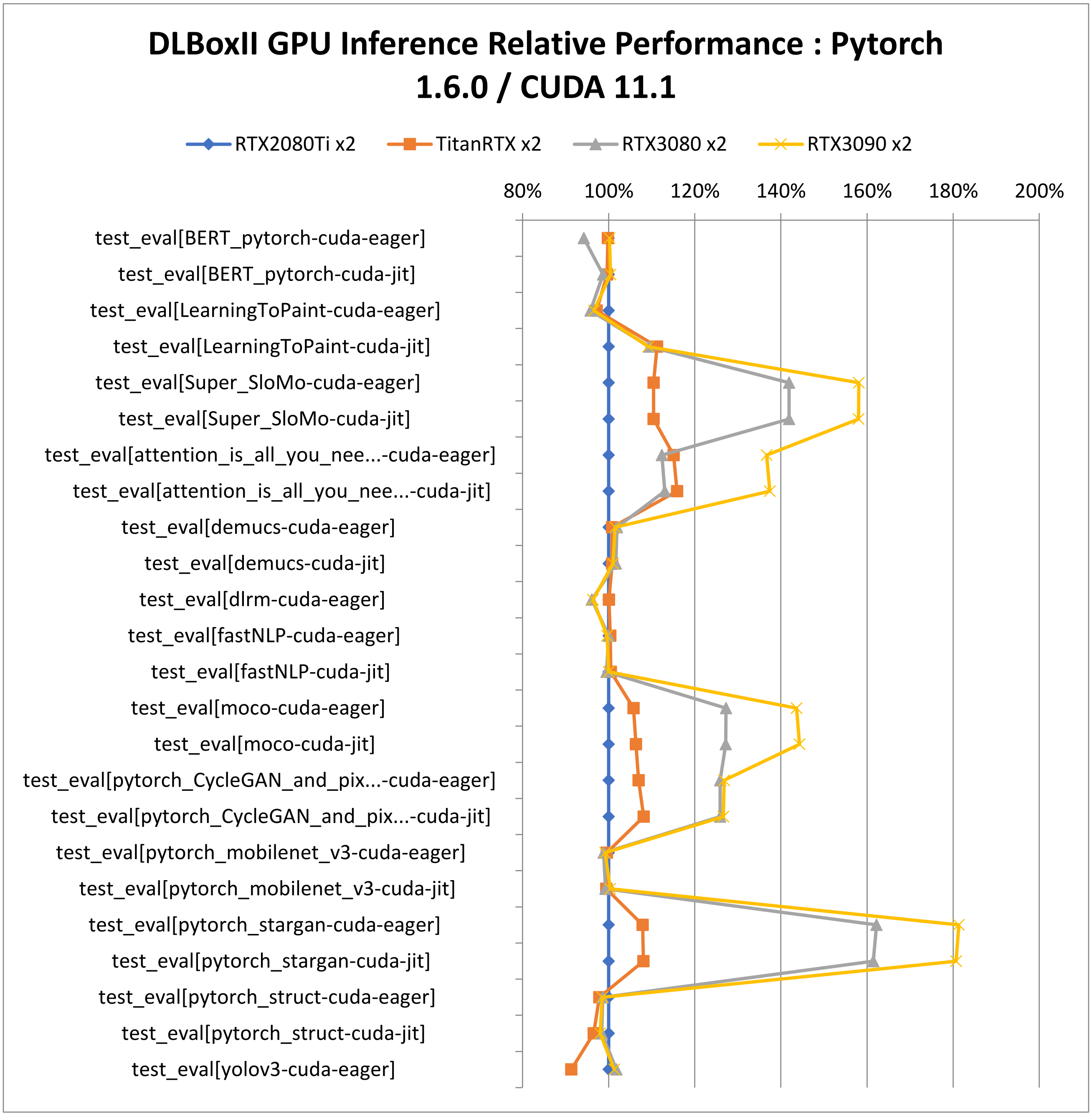

**Figure 3 DLBox II Inference Relative Performance**

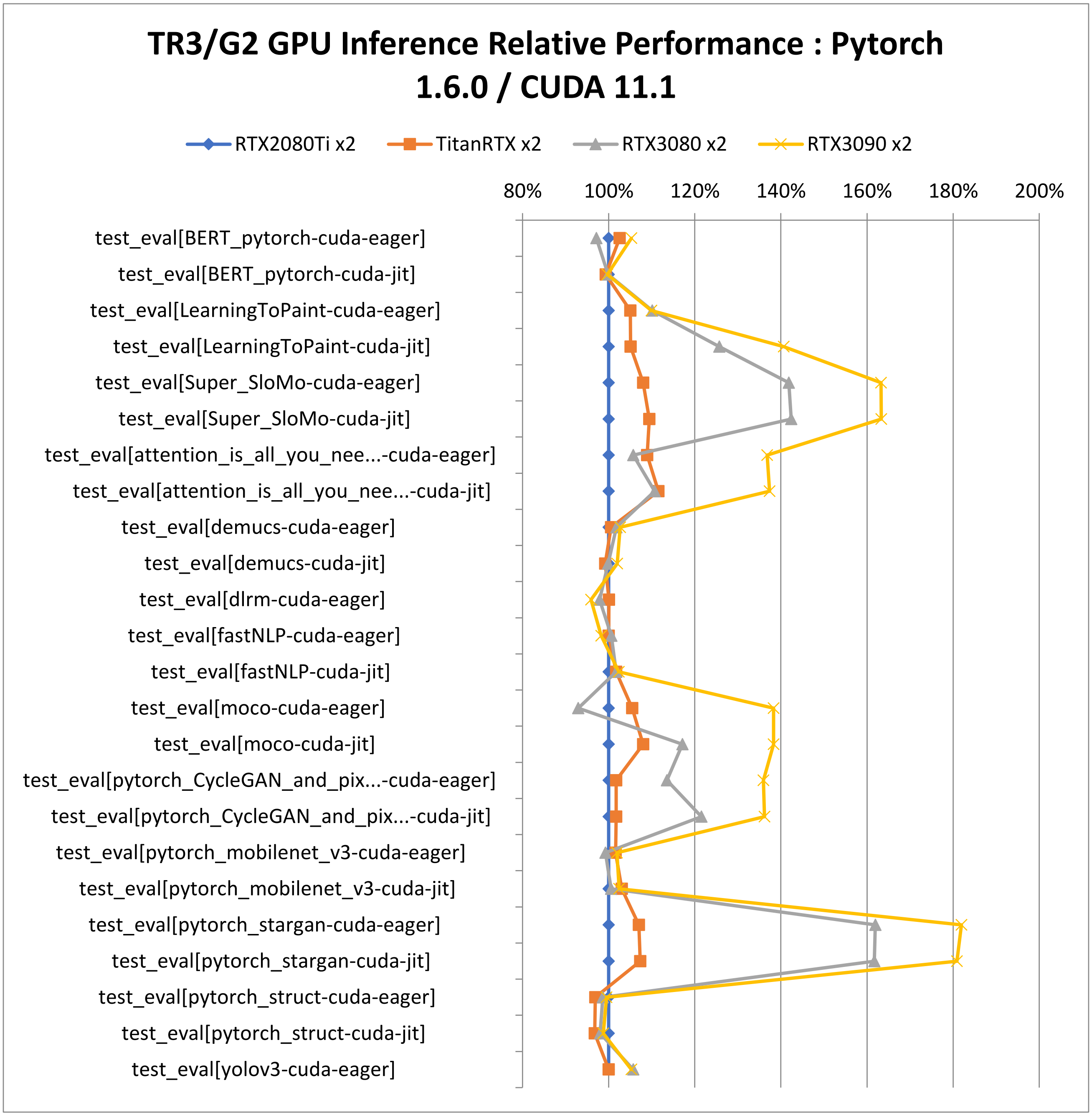

**Figure 4 TR3/G2 Inference Relative Performance**

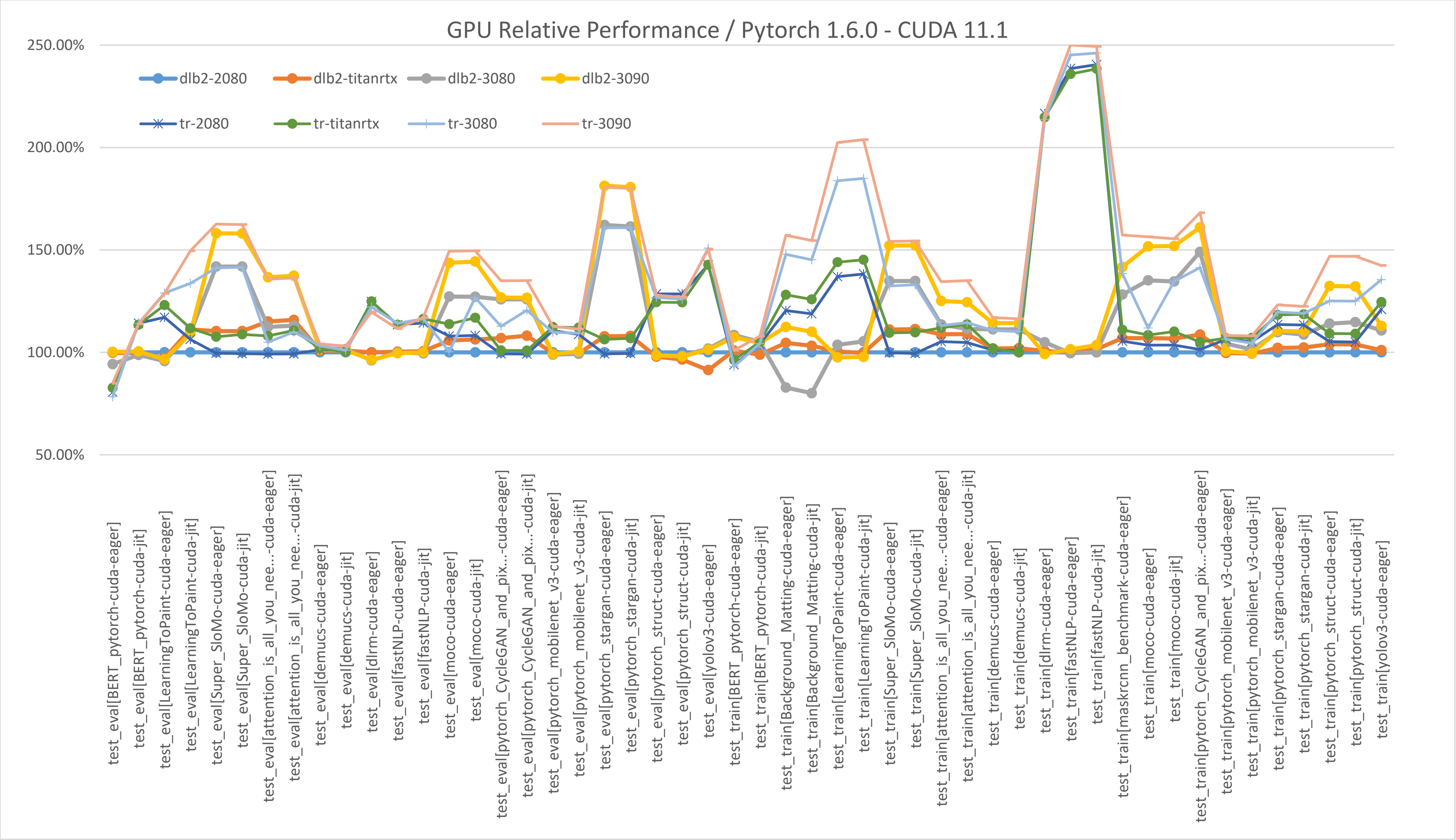

**Figure 5 All GPU / All System Relative Performance**

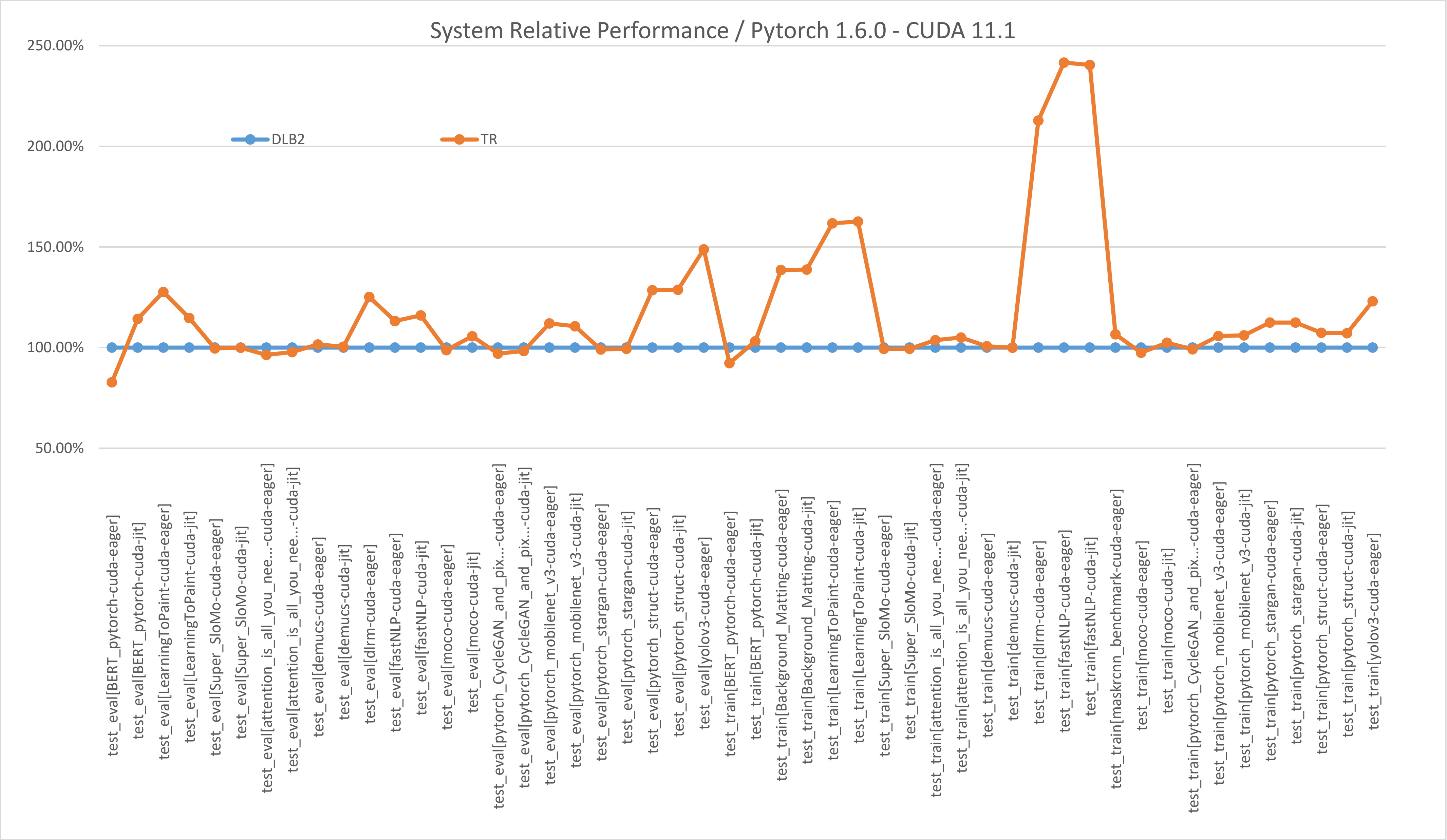

**Figure 6 System Relative Performance**

## Discussion

GeForce RTX3080 / RTX3090 contributed largely in the performance of Super_SloMo, stragan and moco tests. Other types of networks including BERT might require some more time before they could get much benefits from the Ampere architecture.

On the other hand, the performance largely depended on the base system. The Figure 6 shows the average of all GPUs used on each base system. This figure indicates that the GWS-TR3/G2 is a very good choice as far as price per performance was considered.

At the same time, the Deep Learning Box II is still a good choice for stability and expandability point of view, even though the architecture is based on very much matured Intel Core series of CPUs.

## Conclusion

It might require little more time before all kinds of AI applications could enjoy the performance of Ampere architecture. However, some are already getting benefits and it could be safe to say that it is worth investing into the new GeForce RTX3080 / RTX3090, than to older GeForce 2080 and Titan RTX when factors like price and memory capacity are considered as a total.

It was made clear that the base system had large impact on the performance numbers. From these results, here is our recommended configurations.

For cost per performance enthusiast
GWS-TR3/2G

For stability and expandability professional

DeepLearningBOXII