Deep Learning Performance Benchmark on A100-PCIe with ResNet50 Model

Author: Rax Liang, Kevin Tsai

H3 Platform Inc. September 2020

Abstract

Falcon 4010 is a composable GPU solution ready for PCIe Gen 4 GPUs. This experiment will examine the deep learning training and inference performance on composable architecture using Nvidia PCIe-A100 GPU, ResNet-50 model and MXnet framework. The performance is measured in single node and scaled on up to two GPUs. The result is compared to Nvidia PCIe-P100 performance under the same platform. We found that Falcon 4010 does help GPUs to bypass host PCIe version limitations, and that increasing GPU units improves inference performance, but training performance is dependent to the least powerful GPU's processing capability.

Overview

Falcon 4010 GPU expansion chassis is a PCIe Gen 4 solution. Given A100-PCIe available, we are trying to examine whether it can help in improving deep learning performance. In this experiment, we specifically use a PCIe Gen 3 CPU for the test so the result can be easier compared to any existing PCIe Gen 3 platforms. Also, we will provide the deep learning benchmark result of P100 GPU for clearer head to head comparison. In this experiment, all GPU peer to peer communication are through PCIe.

Image 1. System structure

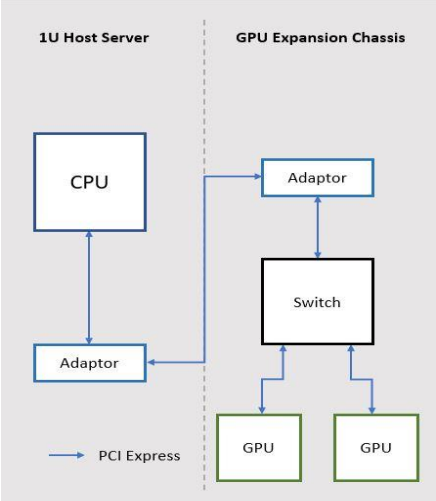
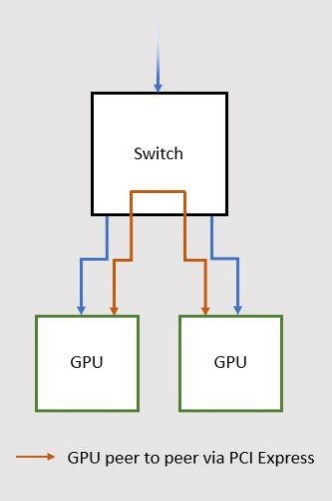


Image 2. GPU P2P structure



Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

Table 1. Hardware and software details

Hardware configuration	
Platform	Falcon 4010 + Supermicro SYS-5019GP-TT
CPU	Intel® Xeon® Gold 5120T @2.20GHz
Memory	64 GB RAM DDR4 2400MT/s
GPU	2x Nvidia P100-PCIe, 2x Nvidia A100-PCIe
Disk	Phison E12 PCIe SSD 1TB

Software and Firmware	
OS	Ubuntu 18.04
Kernel	4.15.0-101-generic
BIOS	American Megatrends Inc Version: 3.3
CUDA	CUDA 11.0.3_450.51.06
Docker	19.03.12 build 48a66213fe

Deep Learning Libraries and Frameworks	
MXNet	20.06
ResNet-50	V1_5

Data	
Dataset	ImageNet ILSVR2012
Batch size	192
Iterations	100
Precision	FP32

The performance benchmark uses ImageNet ILSVR2012 dataset. The number of iterations=100, number of classes=1000, number of groups=32, number of layers=50, with 1,281,167 test samples. The training is done with single-precision FP32. Performance for both training and inference are presented in this paper.

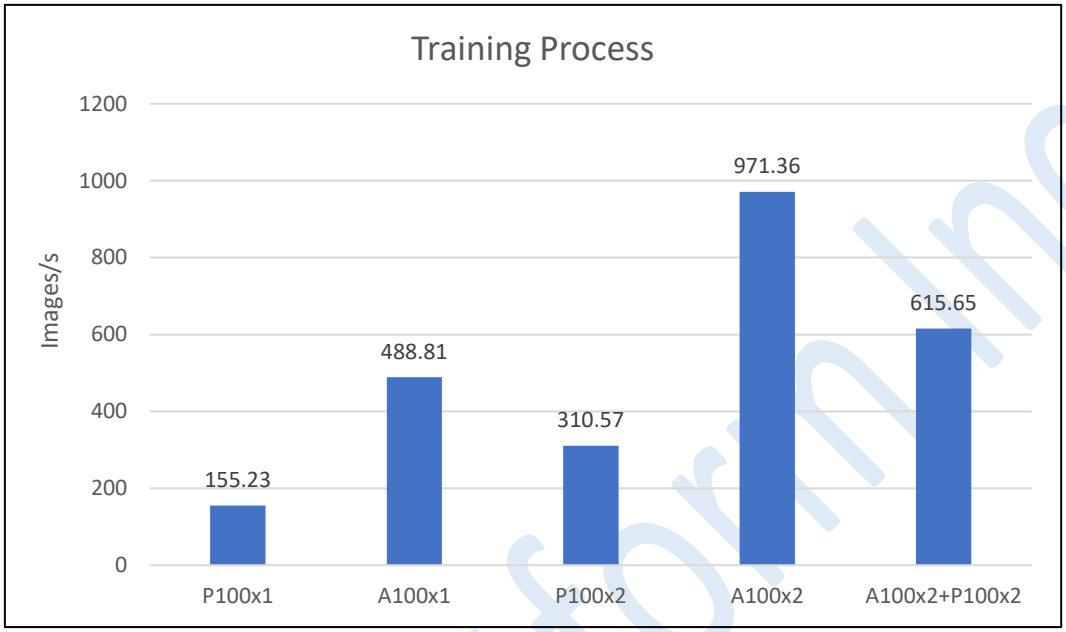
H3 Platform Inc.

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

Test Results Summary

We present an overview of our test results in this section. More detailed results are available at the end of this report.

Training Process

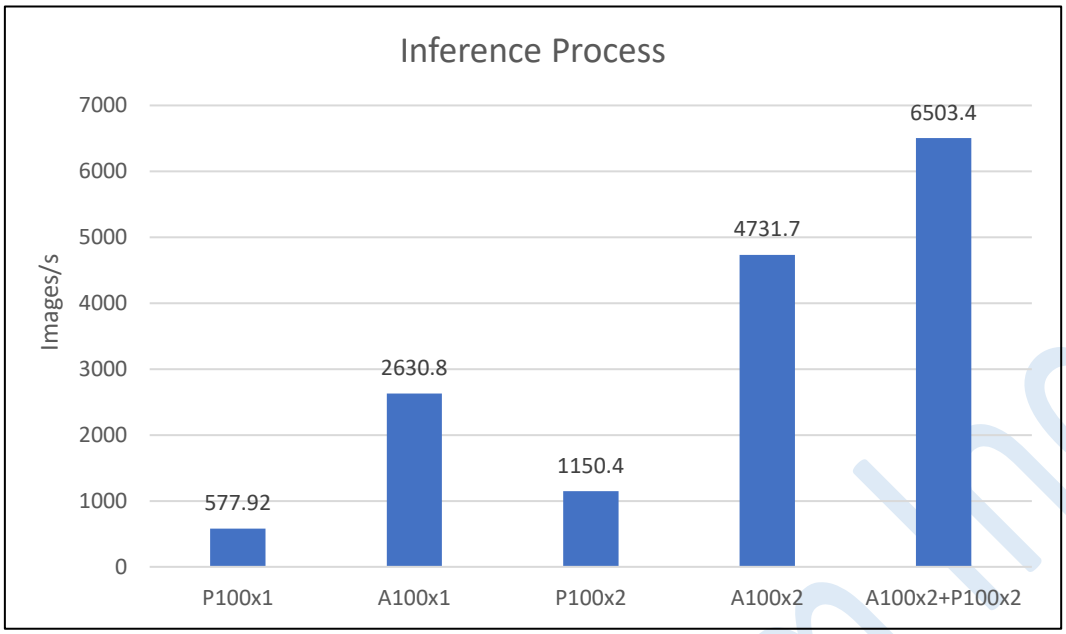


Inference speed measured in images/second					
GPU Model	P100x1	A100x1	P100x2	A100x2	A100x2+P100x2
Images/s	**155.23**	**488.81**	**310.57**	**971.36**	**615.65**
Performance %	**100**	**315**	**200**	**626**	**397**

From the training performance results, we can see that A100 trains approximately 3.15 times faster than P100 (488.81/155.23=3.1489), and the GPU scales well in our system setting (Two GPUs performance is doubled that of single GPU. A100- 971.36/488.81=1.99; P100-310.57/155.23=2.00). However, when we mixed P100 and A100 all together for training, the performance is worse than dual A100.

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

Inference Process



Inference speed measured in images/second					
GPU Model	P100x1	A100x1	P100x2	A100x2	A100x2+P100x2
Images/s	**577.92**	**2630.8**	**1150.4**	**4731.7**	**6503.4**
Performance %	**100**	**455**	**199**	**819**	**1125**

For the inference performance results, we get similar results as training performance when A100 and P100 are tested separately. But the performance for mixed GPU inference exceeds dual A100 inference.

H3 Platform Inc.

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

Findings

From our test results, we can see that **given Falcon 4010 GPU solution, the A100 deep learning performance is better than P100 when CPU supports only PCIe Gen 3** (Training- 315% performance. Inference- 455% performance), and the **GPUs scales nearly perfectly** for two GPUs (Two GPUs performance is doubled that of single GPU. A100- 971.36/488.81=1.99; P100- 310.57/155.23=2.00).

We also see different effects of mixing A100 and P100 to training and inference. We observed increased performance in inference, the inference performance of mixing two A100 and two P100 (6503.4) is approximately 10 percent better than the performance of two A100 plus the performance of two P100 (1150.4+4731.7=5882.1). We found that it is because the algorithm is set during inference process, as the GPUs are working independently to return inference result, the overall performance will be the sum of all GPUs. The results suggest that **increasing number of GPU does improve inference performance**. On the other hand, we saw that the training performance of mixing two A100s and two P100s is close to 4 times single P100 performance. We found that mixing A100 and P100 lowers training performance due to the fact that the workloads assigned to each GPU are equal in training, even if A100 complete its processing, it will have to wait for P100 to return the training results. The results indicate that **to maximize the performance of each GPU during training process**, one should **divide the workloads up according to GPU capabilities**.

H3 Platform Inc.

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

Detailed Test results

How to read the results

- `train.loss` - training loss
- `train.total_ips` - training speed measured in images/second
- `train.compute_ips` - training speed measured in images/second, not counting data loading
- `train.data_time` - time spent on waiting on data
- `train.compute_time` - time spent in forward/backward pass

- `val.loss` - validation loss
- `val.top1` - validation top1 accuracy
- `val.top5` - validation top5 accuracy
- `val.total_ips` - inference speed measured in images/second
- `val.compute_ips` - inference speed measured in images/second, not counting data loading
- `val.data_time` - time spent on waiting on data
- `val.compute_time` – time spent on inference

Find test results in next page.

H3 Platform

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

Test result of A100x2+P100x2

```
train.total_ips
           |    192    |
------------------------
     4     |   615.65  |

train.latency_avg
           |    192    |
------------------------
     4     |   1.2475  |

train.latency_50
           |    192    |
------------------------
     4     |   1.2426  |

train.latency_90
           |    192    |
------------------------
     4     |   1.2612  |

train.latency_95
           |    192    |
------------------------
     4     |   1.2855  |

train.latency_99
           |    192    |
------------------------
     4     |   1.3294  |

train.latency_100
           |    192    |
------------------------
     4     |   1.3365  |

val.total_ips
           |    192    |
------------------------
     4     |   6503.4  |
```

H3 Platform Inc.

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

```
val.latency_avg
           |    192    |
------------------------
     4     |  0.11809  |

val.latency_50
           |    192    |
------------------------
     4     |  0.11838  |

val.latency_90
           |    192    |
------------------------
     4     |   0.1387  |

val.latency_95
           |    192    |
------------------------
     4     |  0.14428  |

val.latency_99
           |    192    |
------------------------
     4     |  0.16103  |

val.latency_100
           |    192    |
------------------------
     4     |  0.16221  |
```

H3 Platform Inc.

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

Test result of A100x2

train.total_ips

	192
2	971.36

train.latency_avg

	192
2	0.39532

train.latency_50

	192
2	0.39401

train.latency_90

	192
2	0.40102

train.latency_95

	192
2	0.40447

train.latency_99

	192
2	0.42765

train.latency_100

	192
2	0.44195

val.total_ips

	192
2	4731.7

H3 Platform Inc.

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

```
val.latency_avg
           |    192    |
------------------------
     2     |  0.081155 |

val.latency_50
           |    192    |
------------------------
     2     |  0.078036 |

val.latency_90
           |    192    |
------------------------
     2     |  0.084778 |

val.latency_95
           |    192    |
------------------------
     2     |  0.087175 |

val.latency_99
           |    192    |
------------------------
     2     |  0.16651  |

val.latency_100
           |    192    |
------------------------
     2     |  0.23693  |
```

H3 Platform Inc.

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

Test result of P100x2

train.total_ips

	192
2	310.57

train.latency_avg

	192
2	1.2365

train.latency_50

	192
2	1.2358

train.latency_90

	192
2	1.2406

train.latency_95

	192
2	1.2419

train.latency_99

	192
2	1.2595

train.latency_100

	192
2	1.3085

val.total_ips

	192
2	1150.4

H3 Platform Inc.

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

```
val.latency_avg
           |    192    |
------------------------
     2     |   0.3338  |

val.latency_50
           |    192    |
------------------------
     2     |   0.3307  |

val.latency_90
           |    192    |
------------------------
     2     |  0.33292  |

val.latency_95
           |    192    |
------------------------
     2     |  0.33934  |

val.latency_99
           |    192    |
------------------------
     2     |  0.39329  |

val.latency_100
           |    192    |
------------------------
     2     |  0.39869  |
```

H3 Platform Inc.

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

Test result of A100x1

train.total_ips

	192
1	488.81

train.latency_avg

	192
1	0.39279

train.latency_50

	192
1	0.39278

train.latency_90

	192
1	0.39845

train.latency_95

	192
1	0.39938

train.latency_99

	192
1	0.40364

train.latency_100

	192
1	0.40411

val.total_ips

	192
1	2630.8

H3 Platform Inc.

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

```
val.latency_avg
           |    192    |
------------------------
     1     |  0.072981 |

val.latency_50
           |    192    |
------------------------
     1     |  0.072648 |

val.latency_90
           |    192    |
------------------------
     1     |  0.076749 |

val.latency_95
           |    192    |
------------------------
     1     |  0.078147 |

val.latency_99
           |    192    |
------------------------
     1     |  0.081338 |

val.latency_100
           |    192    |
------------------------
     1     |  0.081421 |
```

H3 Platform Inc.

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

Test result of P100x1

```
train.total_ips
           |    192    |
------------------------
     1     |   155.23  |

train.latency_avg
           |    192    |
------------------------
     1     |   1.2369  |

train.latency_50
           |    192    |
------------------------
     1     |   1.2354  |

train.latency_90
           |    192    |
------------------------
     1     |   1.2442  |

train.latency_95
           |    192    |
------------------------
     1     |   1.2472  |

train.latency_99
           |    192    |
------------------------
     1     |   1.2734  |

train.latency_100
           |    192    |
------------------------
     1     |   1.2824  |

val.total_ips
           |    192    |
------------------------
     1     |   577.92  |
```

H3 Platform Inc.

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

```
val.latency_avg
           |    192    |
------------------------
     1     |  0.33222  |

val.latency_50
           |    192    |
------------------------
     1     |  0.33017  |

val.latency_90
           |    192    |
------------------------
     1     |  0.33207  |

val.latency_95
           |    192    |
------------------------
     1     |  0.33362  |

val.latency_99
           |    192    |
------------------------
     1     |  0.38187  |

val.latency_100
           |    192    |
------------------------
     1     |  0.38333  |
```

H3 Platform Inc.

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

Test Records

Test records of A100x2+P100x2

```
Wed Sep  2 17:06:16 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 450.51.06    Driver Version: 450.51.06    CUDA Version: 11.0     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  A100-PCIE-40GB      Off  | 00000000:2B:00.0 Off |                    0 |
| N/A   74C    P0   141W / 250W |  12193MiB / 40537MiB |    100%      Default |
|                               |                      |             Disabled |
+-------------------------------+----------------------+----------------------+
|   1  A100-PCIE-40GB      Off  | 00000000:2E:00.0 Off |                    0 |
| N/A   63C    P0   125W / 250W |  12345MiB / 40537MiB |    100%      Default |
|                               |                      |             Disabled |
+-------------------------------+----------------------+----------------------+
|   2  Tesla P100-PCIE...  Off  | 00000000:35:00.0 Off |                    0 |
| N/A   63C    P0   131W / 250W |  11891MiB / 16280MiB |    100%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   3  Tesla P100-PCIE...  Off  | 00000000:38:00.0 Off |                    0 |
| N/A   67C    P0   142W / 250W |  12043MiB / 16280MiB |    100%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|    0   N/A  N/A     14432      C   python                          12189MiB |
|    1   N/A  N/A     14433      C   python                          12341MiB |
|    2   N/A  N/A     14434      C   python                          11887MiB |
|    3   N/A  N/A     14435      C   python                          12039MiB |
+-----------------------------------------------------------------------------+

[1,3]<stdout>:02aea22d2da4:148:574 [3] NCCL INFO Setting affinity for GPU 3 to 0fffffff
[1,2]<stdout>:02aea22d2da4:147:576 [2] NCCL INFO Setting affinity for GPU 2 to 0fffffff
```

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

```
[1,0]<stdout>:02aea22d2da4:145:575 [0] NCCL INFO Setting affinity for GPU 0 to 0fffffff
[1,1]<stdout>:02aea22d2da4:146:573 [1] NCCL INFO Setting affinity for GPU 1 to 0fffffff
[1,0]<stdout>:02aea22d2da4:145:575 [0] NCCL INFO Channel 00 :    0   1   2   3
[1,2]<stdout>:02aea22d2da4:147:576 [2] NCCL INFO Ring 00 : 2[2] -> 3[3] via P2P/IPC
[1,3]<stdout>:02aea22d2da4:148:574 [3] NCCL INFO Ring 00 : 3[3] -> 0[0] via direct shared memory
[1,3]<stdout>:02aea22d2da4:148:574 [3] NCCL INFO comm 0x7fc6782370e0 rank 3 nranks 4 cudaDev 3
nvmlDev 3 - Init COMPLETE
[1,1]<stdout>:02aea22d2da4:146:573 [1] NCCL INFO Ring 00 : 1[1] -> 2[2] via direct shared memory
[1,2]<stdout>:02aea22d2da4:147:576 [2] NCCL INFO comm 0x7fe470237020 rank 2 nranks 4 cudaDev 2
nvmlDev 2 - Init COMPLETE
[1,0]<stdout>:02aea22d2da4:145:575 [0] NCCL INFO Ring 00 : 0[0] -> 1[1] via P2P/IPC
[1,1]<stdout>:02aea22d2da4:146:573 [1] NCCL INFO comm 0x7f260c21abe0 rank 1 nranks 4 cudaDev 1
nvmlDev 1 - Init COMPLETE
[1,0]<stdout>:02aea22d2da4:145:575 [0] NCCL INFO Using 256 threads, Min Comp Cap 6, Trees disabled
[1,0]<stdout>:02aea22d2da4:145:575 [0] NCCL INFO comm 0x7fb87823c630 rank 0 nranks 4 cudaDev 0
nvmlDev 0 - Init COMPLETE
[1,0]<stdout>:02aea22d2da4:145:575 [0] NCCL INFO Launch mode Parallel
[1,1]<stderr>:2020-09-02 09:05:47,272:INFO: Epoch[0] Batch [0-19]      Speed: 494.86747232959243
samples/sec   LR: 0.0017485919712402636
[1,0]<stderr>:2020-09-02 09:05:47,273:INFO: Epoch[0] Batch [0-19]      Speed: 494.78042376152064
samples/sec   LR: 0.0017485919712402636
[1,3]<stderr>:2020-09-02 09:05:47,585:INFO: Epoch[0] Batch [0-19]      Speed: 489.8578871339673
samples/sec    LR: 0.0017485919712402636
[1,2]<stderr>:2020-09-02 09:05:47,591:INFO: Epoch[0] Batch [0-19]      Speed: 489.72420012929274
samples/sec   LR: 0.0017485919712402636
[1,1]<stderr>:2020-09-02 09:06:12,283:INFO: Epoch[0] Batch [20-39]     Speed: 614.1414760136486
samples/sec    LR: 0.0035892150988615936
[1,0]<stderr>:2020-09-02 09:06:12,285:INFO: Epoch[0] Batch [20-39]     Speed: 614.0996956816152
samples/sec    LR: 0.0035892150988615936
[1,3]<stderr>:2020-09-02 09:06:12,597:INFO: Epoch[0] Batch [20-39]     Speed: 614.090610953464
samples/sec     LR: 0.0035892150988615936
[1,2]<stderr>:2020-09-02 09:06:12,606:INFO: Epoch[0] Batch [20-39]     Speed: 614.054286434177
samples/sec     LR: 0.0035892150988615936
[1,0]<stderr>:2020-09-02 09:06:37,212:INFO: Epoch[0] Batch [40-59]     Speed: 616.2055718377686
samples/sec    LR: 0.005429838226482924
[1,1]<stderr>:2020-09-02 09:06:37,217:INFO: Epoch[0] Batch [40-59]     Speed: 616.0385673200454
samples/sec    LR: 0.005429838226482924
[1,2]<stderr>:2020-09-02 09:06:37,530:INFO: Epoch[0] Batch [40-59]     Speed: 616.2727986919555
samples/sec    LR: 0.005429838226482924
[1,3]<stderr>:2020-09-02 09:06:37,533:INFO: Epoch[0] Batch [40-59]     Speed: 615.9942137695475
samples/sec    LR: 0.005429838226482924
[1,1]<stderr>:2020-09-02 09:07:02,234:INFO: Epoch[0] Batch [60-79]     Speed: 613.9757170254576
samples/sec    LR: 0.007270461354104254
```

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

```
[1,0]<stderr>:2020-09-02 09:07:02,238:INFO: Epoch[0] Batch [60-79]      Speed: 613.7718031172357
samples/sec    LR: 0.007270461354104254
[1,2]<stderr>:2020-09-02 09:07:02,537:INFO: Epoch[0] Batch [60-79]      Speed: 614.2241753306702
samples/sec    LR: 0.007270461354104254
[1,3]<stderr>:2020-09-02 09:07:02,552:INFO: Epoch[0] Batch [60-79]      Speed: 613.9421675892387
samples/sec    LR: 0.007270461354104254
[1,1]<stderr>:2020-09-02 09:07:27,068:INFO: Epoch[0] Batch [80-99]      Speed: 618.5221558003361
samples/sec    LR: 0.009111084481725585
[1,1]<stderr>:2020-09-02 09:07:27,069:INFO: Validating epoch 0
[1,1]<stderr>:2020-09-02 09:07:27,069:WARNING: DALI iterator does not support resetting while
epoch is not finished. Ignoring...
[1,0]<stderr>:2020-09-02 09:07:27,070:INFO: Epoch[0] Batch [80-99]      Speed: 618.5514981479366
samples/sec    LR: 0.009111084481725585
[1,0]<stderr>:2020-09-02 09:07:27,072:INFO: Validating epoch 0
[1,0]<stderr>:2020-09-02 09:07:27,072:WARNING: DALI iterator does not support resetting while
epoch is not finished. Ignoring...
[1,2]<stderr>:2020-09-02 09:07:27,386:INFO: Epoch[0] Batch [80-99]      Speed: 618.1389893870124
samples/sec    LR: 0.009111084481725585
[1,2]<stderr>:2020-09-02 09:07:27,387:INFO: Validating epoch 0
[1,2]<stderr>:2020-09-02 09:07:27,388:WARNING: DALI iterator does not support resetting while
epoch is not finished. Ignoring...
[1,3]<stderr>:2020-09-02 09:07:27,388:INFO: Epoch[0] Batch [80-99]      Speed: 618.4563609117052
samples/sec    LR: 0.009111084481725585
[1,3]<stderr>:2020-09-02 09:07:27,389:INFO: Validating epoch 0
[1,3]<stderr>:2020-09-02 09:07:27,389:WARNING: DALI iterator does not support resetting while
epoch is not finished. Ignoring...
[1,1]<stderr>:2020-09-02 09:07:40,743:INFO: Experiment took: 173.26722264289856 sec
[1,0]<stderr>:2020-09-02 09:07:40,912:INFO: Experiment took: 173.43508219718933 sec
[1,3]<stderr>:2020-09-02 09:08:03,050:INFO: Experiment took: 195.57389616966248 sec
[1,2]<stderr>:2020-09-02 09:08:03,595:INFO: Experiment took: 196.11434698104858 sec
```

H3 Platform Inc.

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

Test records of A100x2

```
Wed Sep  2 17:21:12 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 450.51.06    Driver Version: 450.51.06    CUDA Version: 11.0     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  A100-PCIE-40GB      Off  | 00000000:2B:00.0 Off |                    0 |
| N/A   71C    P0   220W / 250W |  12231MiB / 40537MiB |    100%      Default |
|                               |                      |             Disabled |
+-------------------------------+----------------------+----------------------+
|   1  A100-PCIE-40GB      Off  | 00000000:2E:00.0 Off |                    0 |
| N/A   65C    P0   220W / 250W |  12343MiB / 40537MiB |    100%      Default |
|                               |                      |             Disabled |
+-------------------------------+----------------------+----------------------+
|   2  Tesla P100-PCIE...  Off  | 00000000:35:00.0 Off |                    0 |
| N/A   62C    P0    32W / 250W |      2MiB / 16280MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   3  Tesla P100-PCIE...  Off  | 00000000:38:00.0 Off |                    0 |
| N/A   61C    P0    36W / 250W |      2MiB / 16280MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|    0   N/A  N/A     17399      C   python                          12227MiB |
|    1   N/A  N/A     17400      C   python                          12339MiB |
+-----------------------------------------------------------------------------+

[1,1]<stdout>:02aea22d2da4:2522:2736 [1] NCCL INFO Setting affinity for GPU 1 to 0fffffff
[1,0]<stdout>:02aea22d2da4:2521:2735 [0] NCCL INFO Setting affinity for GPU 0 to 0fffffff
[1,0]<stdout>:02aea22d2da4:2521:2735 [0] NCCL INFO Channel 00 :    0   1
[1,0]<stdout>:02aea22d2da4:2521:2735 [0] NCCL INFO Ring 00 : 0[0] -> 1[1] via P2P/IPC
[1,1]<stdout>:02aea22d2da4:2522:2736 [1] NCCL INFO Ring 00 : 1[1] -> 0[0] via P2P/IPC
[1,0]<stdout>:02aea22d2da4:2521:2735 [0] NCCL INFO Using 256 threads, Min Comp Cap 8, Trees
disabled
```

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

```
[1,1]<stdout>:02aea22d2da4:2522:2736 [1] NCCL INFO comm 0x7f7e9c216650 rank 1 nranks 2 cudaDev 1 nvmlDev 1 - Init COMPLETE
[1,0]<stdout>:02aea22d2da4:2521:2735 [0] NCCL INFO comm 0x7f9c40216610 rank 0 nranks 2 cudaDev 0 nvmlDev 0 - Init COMPLETE
[1,0]<stdout>:02aea22d2da4:2521:2735 [0] NCCL INFO Launch mode Parallel
[1,0]<stderr>:2020-09-02 09:20:46,272:INFO: Epoch[0] Batch [0-19]      Speed: 778.085024513156 samples/sec     LR: 0.0004372789931075817
[1,1]<stderr>:2020-09-02 09:20:46,276:INFO: Epoch[0] Batch [0-19]      Speed: 777.7247298742562 samples/sec    LR: 0.0004372789931075817
[1,0]<stderr>:2020-09-02 09:20:54,145:INFO: Epoch[0] Batch [20-39]     Speed: 975.5710225352337 samples/sec    LR: 0.0008975726700629308
[1,1]<stderr>:2020-09-02 09:20:54,150:INFO: Epoch[0] Batch [20-39]     Speed: 975.3899695567004 samples/sec    LR: 0.0008975726700629308
[1,0]<stderr>:2020-09-02 09:21:02,080:INFO: Epoch[0] Batch [40-59]     Speed: 967.9061283910776 samples/sec    LR: 0.0013578663470182799
[1,1]<stderr>:2020-09-02 09:21:02,090:INFO: Epoch[0] Batch [40-59]     Speed: 967.246559203417 samples/sec     LR: 0.0013578663470182799
[1,0]<stderr>:2020-09-02 09:21:09,951:INFO: Epoch[0] Batch [60-79]     Speed: 975.6910821443795 samples/sec    LR: 0.0018181600239736292
[1,1]<stderr>:2020-09-02 09:21:09,958:INFO: Epoch[0] Batch [60-79]     Speed: 976.1346693977534 samples/sec    LR: 0.0018181600239736292
[1,1]<stderr>:2020-09-02 09:21:17,890:INFO: Epoch[0] Batch [80-99]     Speed: 968.2780444990161 samples/sec    LR: 0.002278453700928978
[1,1]<stderr>:2020-09-02 09:21:17,891:INFO: Validating epoch 0
[1,1]<stderr>:2020-09-02 09:21:17,891:WARNING: DALI iterator does not support resetting while epoch is not finished. Ignoring...
[1,0]<stderr>:2020-09-02 09:21:17,898:INFO: Epoch[0] Batch [80-99]     Speed: 966.4313487545099 samples/sec    LR: 0.002278453700928978
[1,0]<stderr>:2020-09-02 09:21:17,900:INFO: Validating epoch 0
[1,0]<stderr>:2020-09-02 09:21:17,900:WARNING: DALI iterator does not support resetting while epoch is not finished. Ignoring...
[1,1]<stderr>:2020-09-02 09:21:26,469:INFO: Experiment took: 57.95062565803528 sec
[1,0]<stderr>:2020-09-02 09:21:26,748:INFO: Experiment took: 58.23367977142334 sec
```

H3 Platform Inc.

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

Test records of P100x2

```
Wed Sep  2 17:17:31 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 450.51.06    Driver Version: 450.51.06    CUDA Version: 11.0     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  A100-PCIE-40GB      Off  | 00000000:2B:00.0 Off |                    0 |
| N/A   56C    P0    40W / 250W |      2MiB / 40537MiB |      0%      Default |
|                               |                      |             Disabled |
+-------------------------------+----------------------+----------------------+
|   1  A100-PCIE-40GB      Off  | 00000000:2E:00.0 Off |                    0 |
| N/A   49C    P0    36W / 250W |      2MiB / 40537MiB |      0%      Default |
|                               |                      |             Disabled |
+-------------------------------+----------------------+----------------------+
|   2  Tesla P100-PCIE...  Off  | 00000000:35:00.0 Off |                    0 |
| N/A   74C    P0   149W / 250W |  11821MiB / 16280MiB |    100%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   3  Tesla P100-PCIE...  Off  | 00000000:38:00.0 Off |                    0 |
| N/A   62C    P0   145W / 250W |  11921MiB / 16280MiB |    100%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|    2   N/A  N/A     16936      C   python                          11817MiB |
|    3   N/A  N/A     16937      C   python                          11917MiB |
+-----------------------------------------------------------------------------+


[1,0]<stdout>:02aea22d2da4:2151:2365 [2] NCCL INFO Setting affinity for GPU 2 to 0fffffff
[1,1]<stdout>:02aea22d2da4:2152:2366 [3] NCCL INFO Setting affinity for GPU 3 to 0fffffff
[1,0]<stdout>:02aea22d2da4:2151:2365 [2] NCCL INFO Channel 00 :    0   1
[1,1]<stdout>:02aea22d2da4:2152:2366 [3] NCCL INFO Ring 00 : 1[3] -> 0[2] via P2P/IPC
[1,0]<stdout>:02aea22d2da4:2151:2365 [2] NCCL INFO Ring 00 : 0[2] -> 1[3] via P2P/IPC
```

H3 Platform Inc.

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

```
[1,0]<stdout>:02aea22d2da4:2151:2365 [2] NCCL INFO Using 256 threads, Min Comp Cap 6, Trees disabled
[1,0]<stdout>:02aea22d2da4:2151:2365 [2] NCCL INFO comm 0x7fab802341b0 rank 0 nranks 2 cudaDev 2 nvmlDev 2 - Init COMPLETE
[1,1]<stdout>:02aea22d2da4:2152:2366 [3] NCCL INFO comm 0x7f7ff4234750 rank 1 nranks 2 cudaDev 3 nvmlDev 3 - Init COMPLETE
[1,0]<stdout>:02aea22d2da4:2151:2365 [2] NCCL INFO Launch mode Parallel
[1,1]<stderr>:2020-09-02 09:17:07,983:INFO: Epoch[0] Batch [0-19]      Speed: 305.9280616972206 samples/sec    LR: 0.0004372789931075817
[1,0]<stderr>:2020-09-02 09:17:08,066:INFO: Epoch[0] Batch [0-19]      Speed: 304.9505557427164 samples/sec    LR: 0.0004372789931075817
[1,0]<stderr>:2020-09-02 09:17:32,853:INFO: Epoch[0] Batch [20-39]     Speed: 309.8398063341463 samples/sec    LR: 0.0008975726700629308
[1,1]<stderr>:2020-09-02 09:17:32,860:INFO: Epoch[0] Batch [20-39]     Speed: 308.7185082201739 samples/sec    LR: 0.0008975726700629308
[1,0]<stderr>:2020-09-02 09:17:57,568:INFO: Epoch[0] Batch [40-59]     Speed: 310.7400430921222 samples/sec    LR: 0.0013578663470182799
[1,1]<stderr>:2020-09-02 09:17:57,569:INFO: Epoch[0] Batch [40-59]     Speed: 310.82600498857704 samples/sec   LR: 0.0013578663470182799
[1,1]<stderr>:2020-09-02 09:18:22,279:INFO: Epoch[0] Batch [60-79]     Speed: 310.8066939775967 samples/sec    LR: 0.0018181600239736292
[1,0]<stderr>:2020-09-02 09:18:22,279:INFO: Epoch[0] Batch [60-79]     Speed: 310.79328654760053 samples/sec   LR: 0.0018181600239736292
[1,0]<stderr>:2020-09-02 09:18:46,982:INFO: Epoch[0] Batch [80-99]     Speed: 310.89850492029285 samples/sec   LR: 0.002278453700928978
[1,0]<stderr>:2020-09-02 09:18:46,983:INFO: Validating epoch 0
[1,0]<stderr>:2020-09-02 09:18:46,983:WARNING: DALI iterator does not support resetting while epoch is not finished. Ignoring...
[1,1]<stderr>:2020-09-02 09:18:46,985:INFO: Epoch[0] Batch [80-99]     Speed: 310.85797134931624 samples/sec   LR: 0.002278453700928978
[1,1]<stderr>:2020-09-02 09:18:46,986:INFO: Validating epoch 0
[1,1]<stderr>:2020-09-02 09:18:46,986:WARNING: DALI iterator does not support resetting while epoch is not finished. Ignoring...
[1,1]<stderr>:2020-09-02 09:19:22,396:INFO: Experiment took: 163.78442239761353 sec
[1,0]<stderr>:2020-09-02 09:19:22,482:INFO: Experiment took: 163.87777996063232 sec
```

H3 Platform Inc.

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

Test records of A100x1

```
Wed Sep  2 17:22:53 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 450.51.06    Driver Version: 450.51.06    CUDA Version: 11.0     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  A100-PCIE-40GB      Off  | 00000000:2B:00.0 Off |                    0 |
| N/A   77C    P0   242W / 250W |  12183MiB / 40537MiB |    100%      Default |
|                               |                      |             Disabled |
+-------------------------------+----------------------+----------------------+
|   1  A100-PCIE-40GB      Off  | 00000000:2E:00.0 Off |                    0 |
| N/A   57C    P0    39W / 250W |      2MiB / 40537MiB |      0%      Default |
|                               |                      |             Disabled |
+-------------------------------+----------------------+----------------------+
|   2  Tesla P100-PCIE...  Off  | 00000000:35:00.0 Off |                    0 |
| N/A   58C    P0    31W / 250W |      2MiB / 16280MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   3  Tesla P100-PCIE...  Off  | 00000000:38:00.0 Off |                    0 |
| N/A   59C    P0    35W / 250W |      2MiB / 16280MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|    0   N/A  N/A     17860      C   python                          12179MiB |
+-----------------------------------------------------------------------------+
[1,0]<stdout>:02aea22d2da4:2891:2998 [0] NCCL INFO Setting affinity for GPU 0 to 0fffffff
[1,0]<stdout>:02aea22d2da4:2891:2998 [0] NCCL INFO Using 256 threads, Min Comp Cap 8, Trees
enabled up to size -2
[1,0]<stdout>:02aea22d2da4:2891:2998 [0] NCCL INFO comm 0x7fa638209610 rank 0 nranks 1 cudaDev 0
nvmlDev 0 - Init COMPLETE
[1,0]<stderr>:2020-09-02 09:22:55,941:INFO: Epoch[0] Batch [0-19]       Speed: 416.87420121959457
samples/sec   LR: 0.00010933613067585794
[1,0]<stderr>:2020-09-02 09:23:03,787:INFO: Epoch[0] Batch [20-39]      Speed: 489.4114890094625
samples/sec    LR: 0.0002244267945451821
```

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

```
[1,0]<stderr>:2020-09-02 09:23:11,637:INFO: Epoch[0] Batch [40-59]      Speed: 489.16785024548096
samples/sec   LR: 0.00033951745841450623
[1,0]<stderr>:2020-09-02 09:23:19,505:INFO: Epoch[0] Batch [60-79]      Speed: 488.0815335178691
samples/sec    LR: 0.00045460812228383037
[1,0]<stderr>:2020-09-02 09:23:27,364:INFO: Epoch[0] Batch [80-99]      Speed: 488.63320489090484
samples/sec   LR: 0.0005696987861531545
[1,0]<stderr>:2020-09-02 09:23:27,366:INFO: Validating epoch 0
[1,0]<stderr>:2020-09-02 09:23:27,366:WARNING: DALI iterator does not support resetting while
epoch is not finished. Ignoring...
[1,0]<stderr>:2020-09-02 09:23:35,182:INFO: Experiment took: 56.11194610595703 sec
```

H3 Platform Inc.

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

Test records of P100x1

```
Wed Sep  2 17:26:49 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 450.51.06    Driver Version: 450.51.06    CUDA Version: 11.0     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  A100-PCIE-40GB      Off  | 00000000:2B:00.0 Off |                    0 |
| N/A   44C    P0    37W / 250W |      2MiB / 40537MiB |      0%      Default |
|                               |                      |             Disabled |
+-------------------------------+----------------------+----------------------+
|   1  A100-PCIE-40GB      Off  | 00000000:2E:00.0 Off |                    0 |
| N/A   37C    P0    33W / 250W |      2MiB / 40537MiB |      0%      Default |
|                               |                      |             Disabled |
+-------------------------------+----------------------+----------------------+
|   2  Tesla P100-PCIE...  Off  | 00000000:35:00.0 Off |                    0 |
| N/A   78C    P0   222W / 250W |  12039MiB / 16280MiB |    100%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+
|   3  Tesla P100-PCIE...  Off  | 00000000:38:00.0 Off |                    0 |
| N/A   43C    P0    29W / 250W |      2MiB / 16280MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|    2   N/A  N/A     18138      C   python                          12035MiB |
+-----------------------------------------------------------------------------+

[1,0]<stdout>:02aea22d2da4:3098:3205 [2] NCCL INFO Setting affinity for GPU 2 to 0fffffff
[1,0]<stdout>:02aea22d2da4:3098:3205 [2] NCCL INFO Using 256 threads, Min Comp Cap 6, Trees
enabled up to size -2
[1,0]<stdout>:02aea22d2da4:3098:3205 [2] NCCL INFO comm 0x7fcfc022dcb0 rank 0 nranks 1 cudaDev 2
nvmlDev 2 - Init COMPLETE
[1,0]<stderr>:2020-09-02 09:24:54,452:INFO: Epoch[0] Batch [0-19]       Speed: 150.76769797378185
samples/sec   LR: 0.00010933613067585794
```

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.

```
[1,0]<stderr>:2020-09-02 09:25:19,257:INFO: Epoch[0] Batch [20-39]      Speed: 154.8106913113547
samples/sec    LR: 0.0002244267945451821
[1,0]<stderr>:2020-09-02 09:25:43,972:INFO: Epoch[0] Batch [40-59]      Speed: 155.37567072395325
samples/sec   LR: 0.00033951745841450623
[1,0]<stderr>:2020-09-02 09:26:08,698:INFO: Epoch[0] Batch [60-79]      Speed: 155.3030121883166
samples/sec    LR: 0.00045460812228383037
[1,0]<stderr>:2020-09-02 09:26:33,403:INFO: Epoch[0] Batch [80-99]      Speed: 155.43513115780172
samples/sec   LR: 0.0005696987861531545
[1,0]<stderr>:2020-09-02 09:26:33,404:INFO: Validating epoch 0
[1,0]<stderr>:2020-09-02 09:26:33,404:WARNING: DALI iterator does not support resetting while
epoch is not finished. Ignoring...
[1,0]<stderr>:2020-09-02 09:27:08,750:INFO: Experiment took: 164.37539100646973 sec
```

H3 Platform Inc.

Copyright©2020 H3 Platform Inc. All rights reserved. Information in this paper is subject to change without notice.