HackSA: Performance of VGG, Inception and ResNet with Horovod

Authors: Arias Rodrigo, Burca Horia

Date: 14/01/2019

Continuing the analysis with Horovod and the CIFAR10 dataset, we now focus on the performance of the models: VGG, Inception and ResNet

Performance metrics

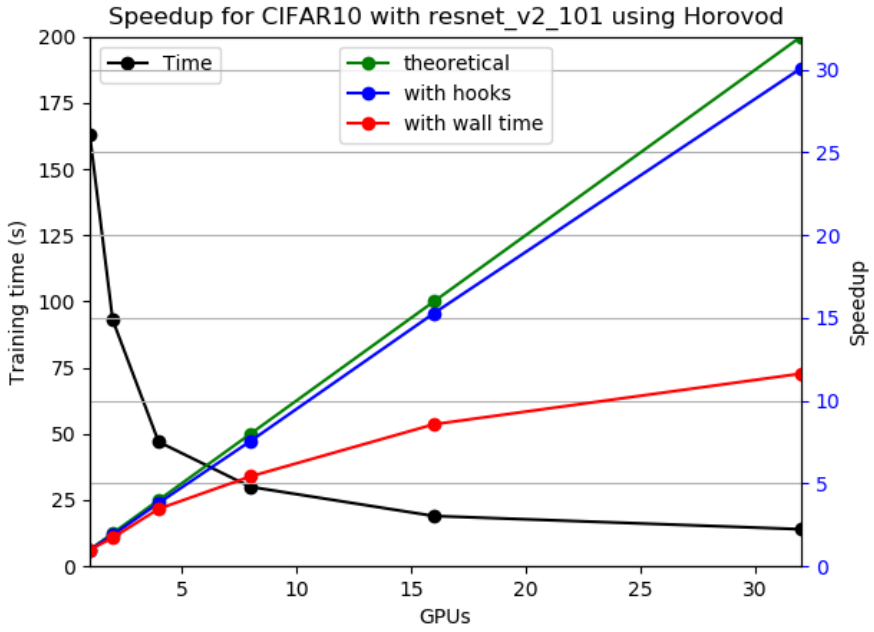
As previously stated in lab 11, we use 2 metrics to evaluate the overall performance. The speed of the training process, measured in the average number of samples per second. And the time it takes the complete training, skipping the initialization time.

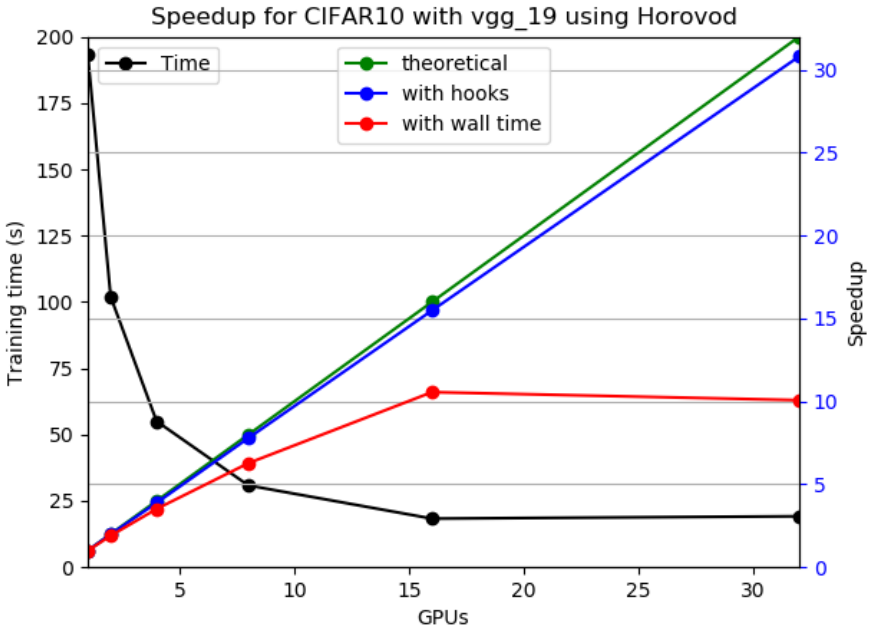
With that information, the speedup is computed as:

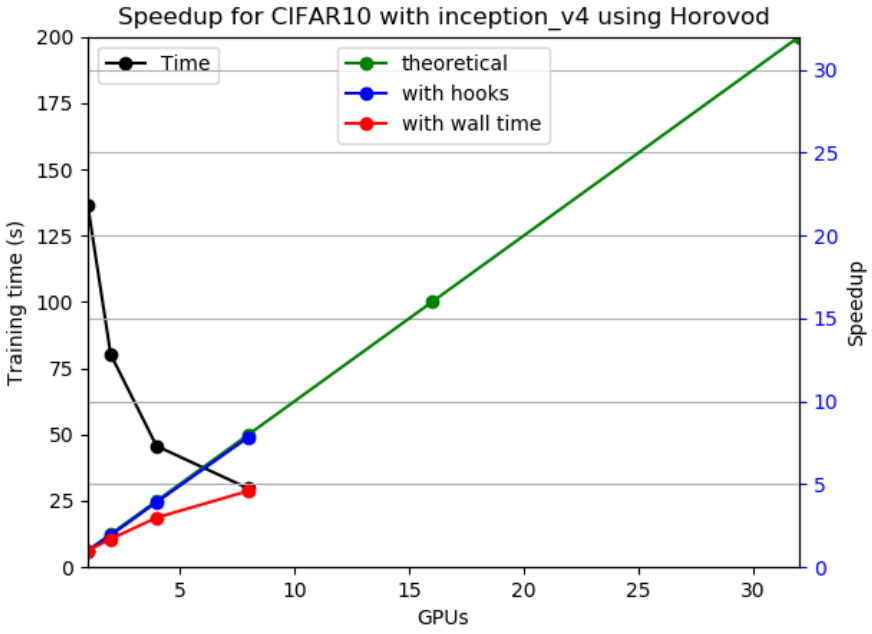
```
speedup_speed = gpus * speed_parallel / speed_serial
speedup_time = time_serial / time_parallel
```

Speedup

The speedup measured using both metrics (speed and wall time), and shown as the number of GPUS grow, in the three models:

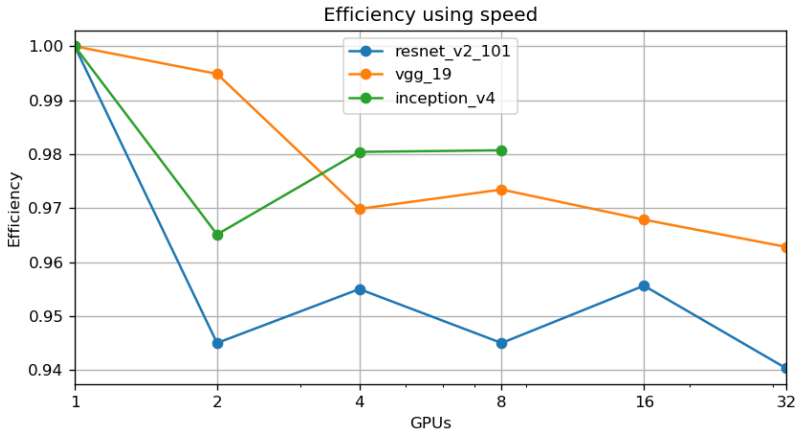


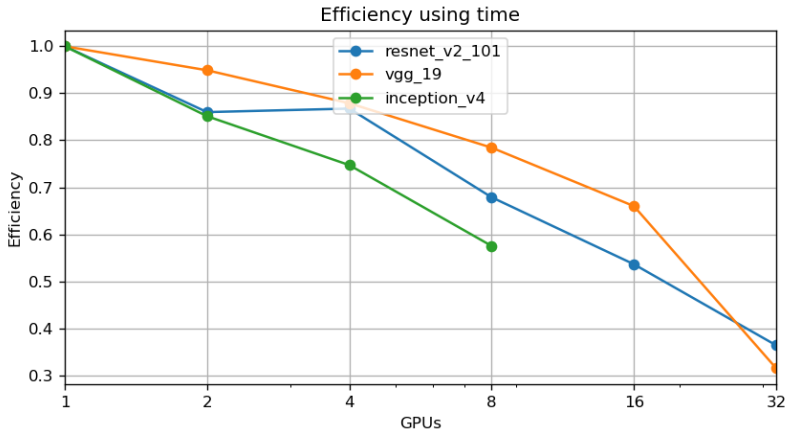

Speedup for CIFAR10 with vgg_19 using Horovod
Time
theoretical
with hooks
with wall time
Training time (s)
Speedup
GPUs
0
25
50
75
100
125
150
175
200
0
5
10
15
20
25
30
5
10
15
20
25
30


Speedup for CIFAR10 with inception_v4 using Horovod
Time
theoretical
with hooks
with wall time
Training time (s)
Speedup
GPUs
0
25
50
75
100
125
150
175
200
0
5
10
15
20
25
30
5
10
15
20
25
30

Efficiency

The efficiency is computed for each model, and plot using the two speedup metrics:





Results

The last model, `inception_v4` was unable to finish due to queueing constraints in the supercomputer. Some of the datapoints are missing, but the overall trend is similar to the other two models.

We see the first metric, using the speed, gives a very optimistic result, with an speedup very close to the theoretical maximum. However, when the wall time is used, the speedup is no longer linear but sublinear.

With respect to the efficiency given by the first metric, it seems to be kept constant and higher than 0.94, with the best efficient model being `vgg_19`. However, the second metric lead to a very different result, showing the efficiency

drop as the number of GPUs increase. It lowers below 0.4, which is quite a bad efficiency.

Conclusions

With these two contradictory metrics, we cannot establish a clear conclusion about what is the actual performance of the models, taking into account the speedup and the efficiency.

More research is required to understand why they don't bring similar results. However, due to the large waiting times of the queue, we couldn't advance any further our investigation.