NEW YORK UNIVERSITY

Final Project Presentation

Department of Electrical and Computer Engineering

Tandon School of Engineering, New York University

Present By: Haoze He

NYUID: hh2537

Agenda

01 / Executive Summary

02 / Approach

03 / Main Results

04 / Observation & Conclusion

05 / GitHub Link

Executive Summary

Asynchronized SGD

Advantages
High accuracy

Disadvantages
Waste time, Slow converge

Synchronous SGD

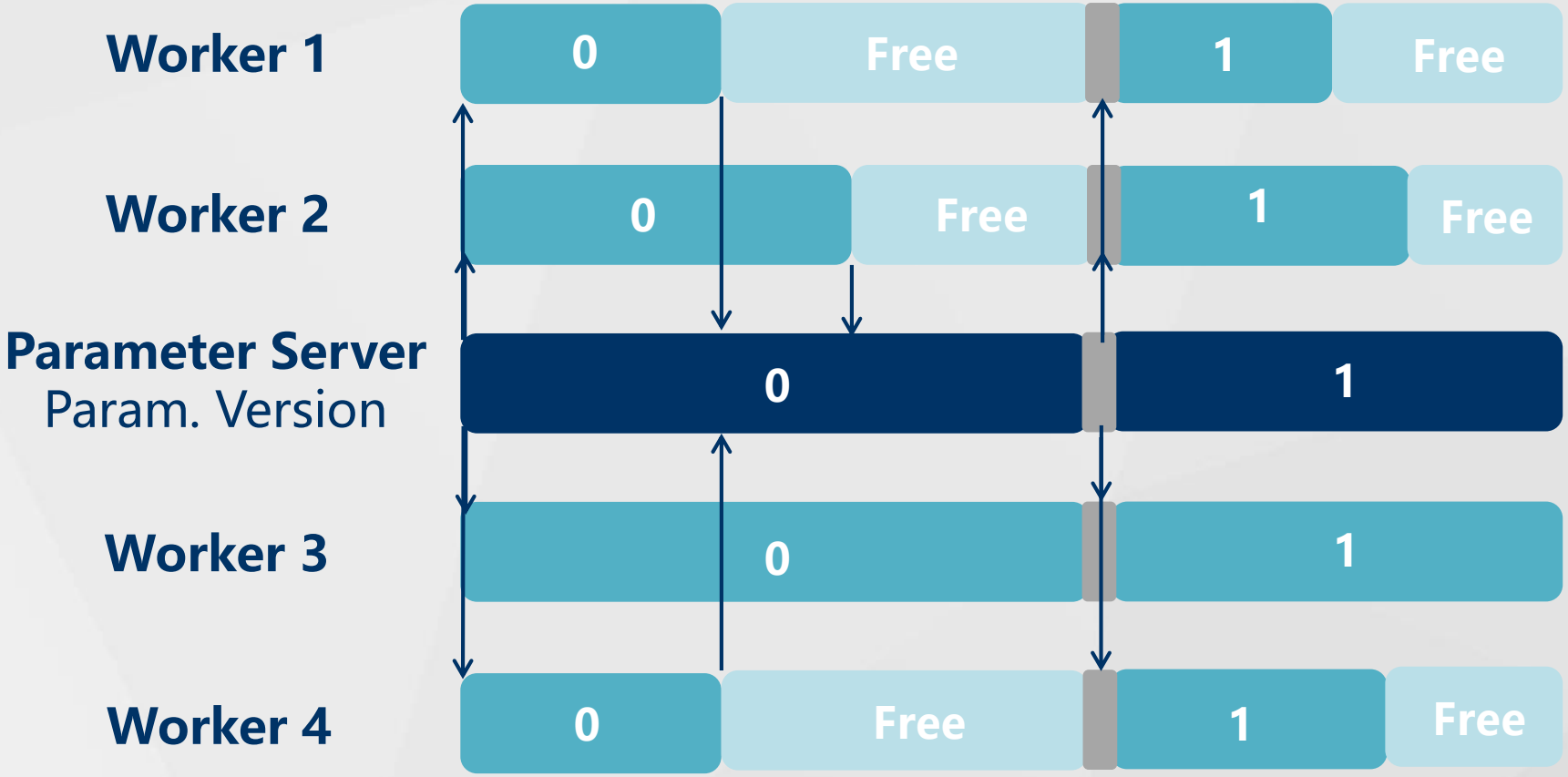
Advantages
Time saving

Disadvantages
Lower accuracy

Non-Block SGD

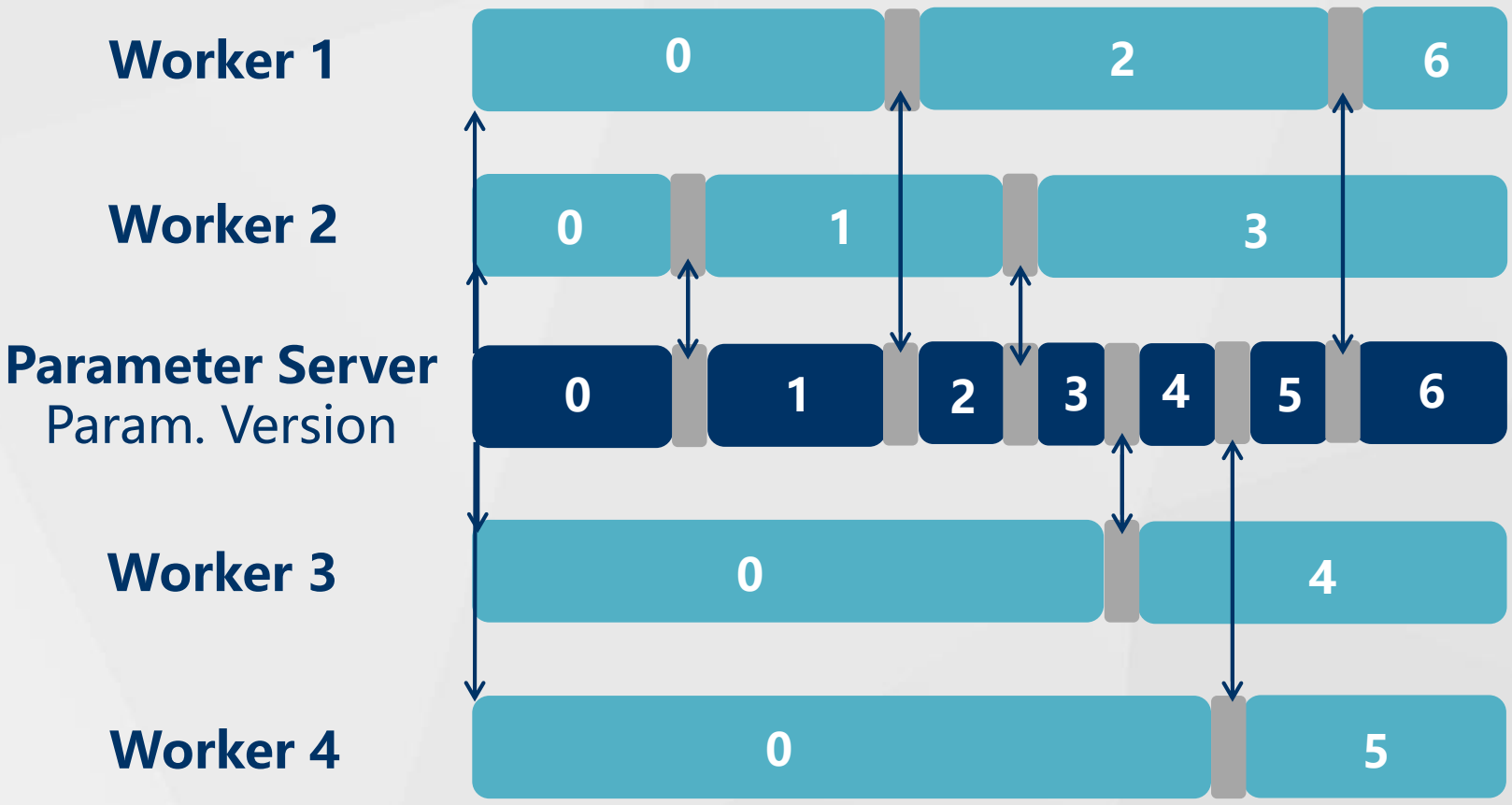
- **High accuracy as ASSGD**
- **Similar time usage as SSGD**

Approach A – Synchronize SGD



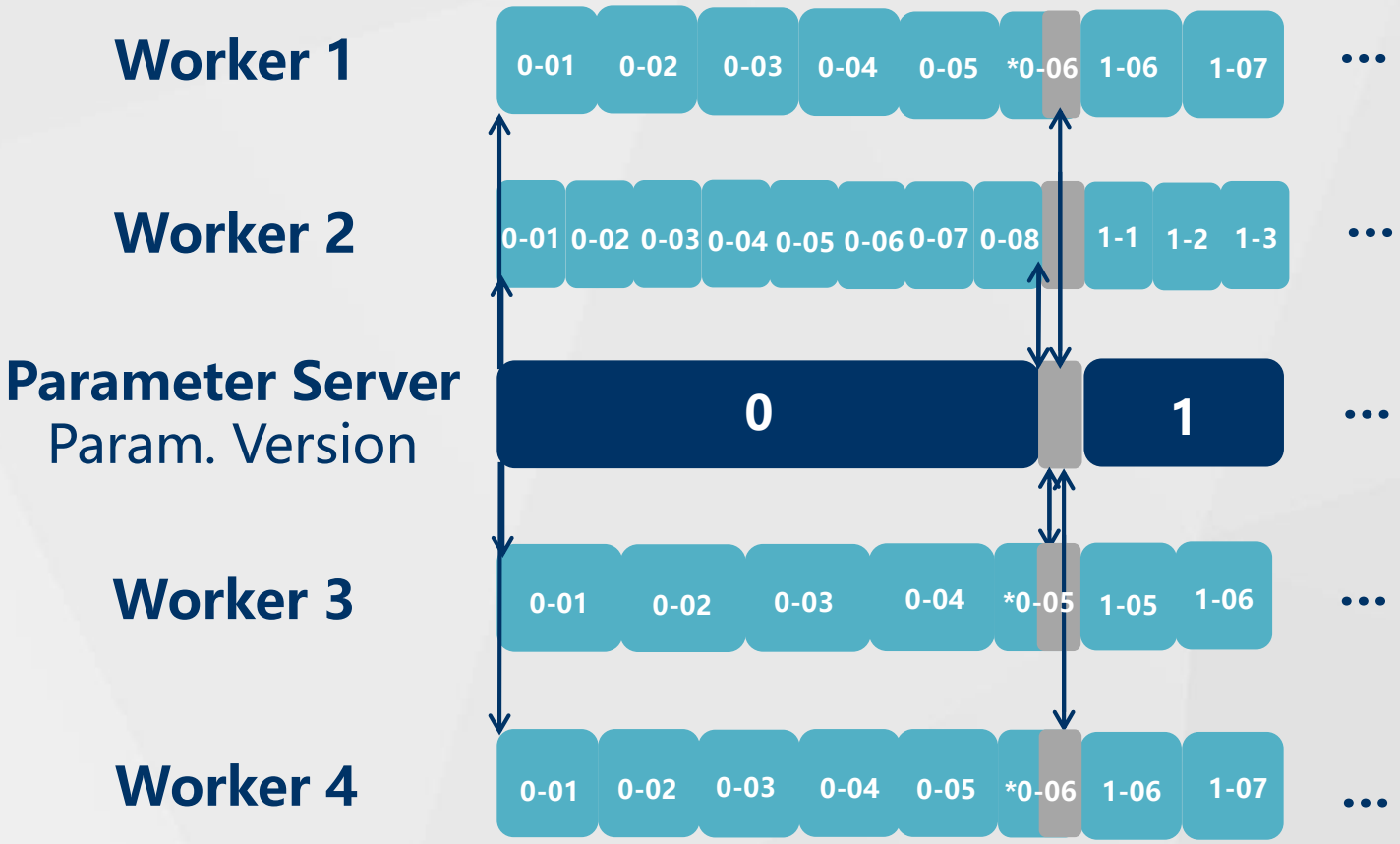
- Batch size =256
- Cons: Waste time
- : Communication time

Approach B – Asynchronized SGD



- Batch size =256
- Cons: Lower Accuracy
- : Communication time

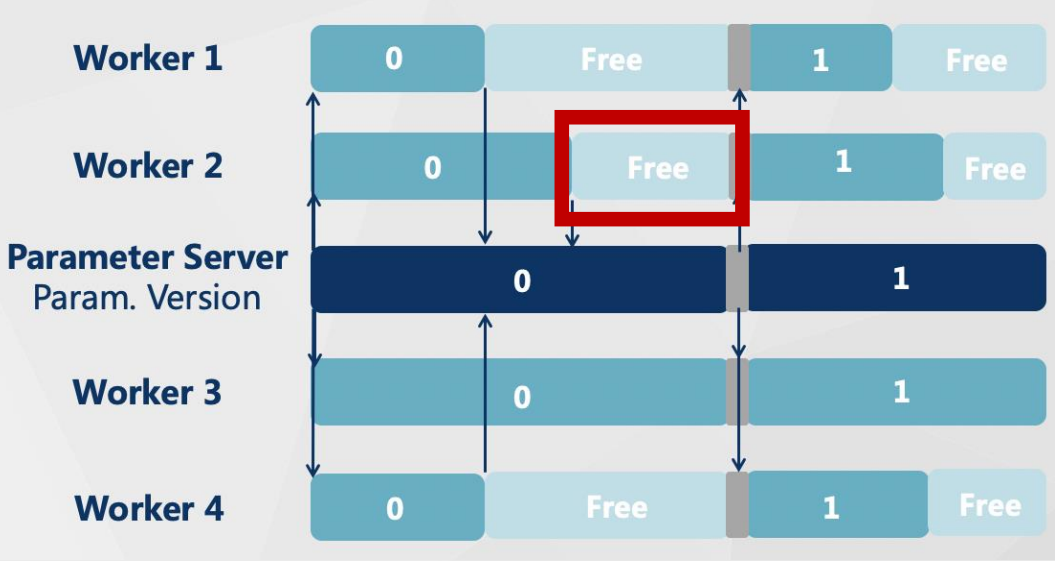
Approach C – Non-Block SGD



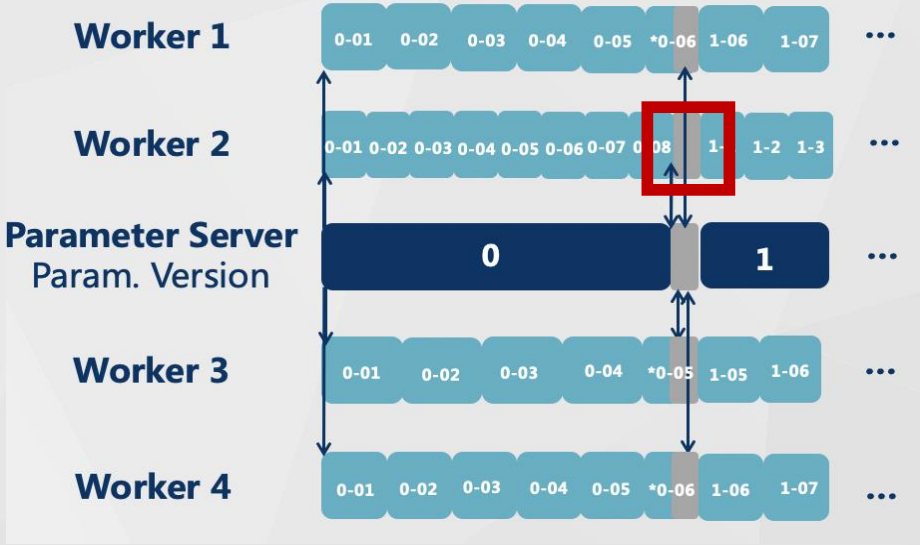
- Batch size = 256
- Mini-batch size = 32
- : Unfinished part and communication time

Differences – Synchronized SGD and Non-Block SGD

Synchronized SGD



Non-Block SGD



- The proportion of free time remains large even though we set the batch size relatively small, especially when the large performance gap is huge in different workers

- By using mini-batch size, we can decrease the proportion of free time dramatically

Approach - Implementation

Hardware Approach

Results are test on the NYU Greene HPC clusters with mi50 nodes. To get computing resources shortly, I use multiple CPU nodes as workers and the parameter server instead of multiple GPUs (will be hard to get resources).

Model Used

I utilize RESNET50 model and CIFAR10 datasets to test the final accuracy of communication scheme.

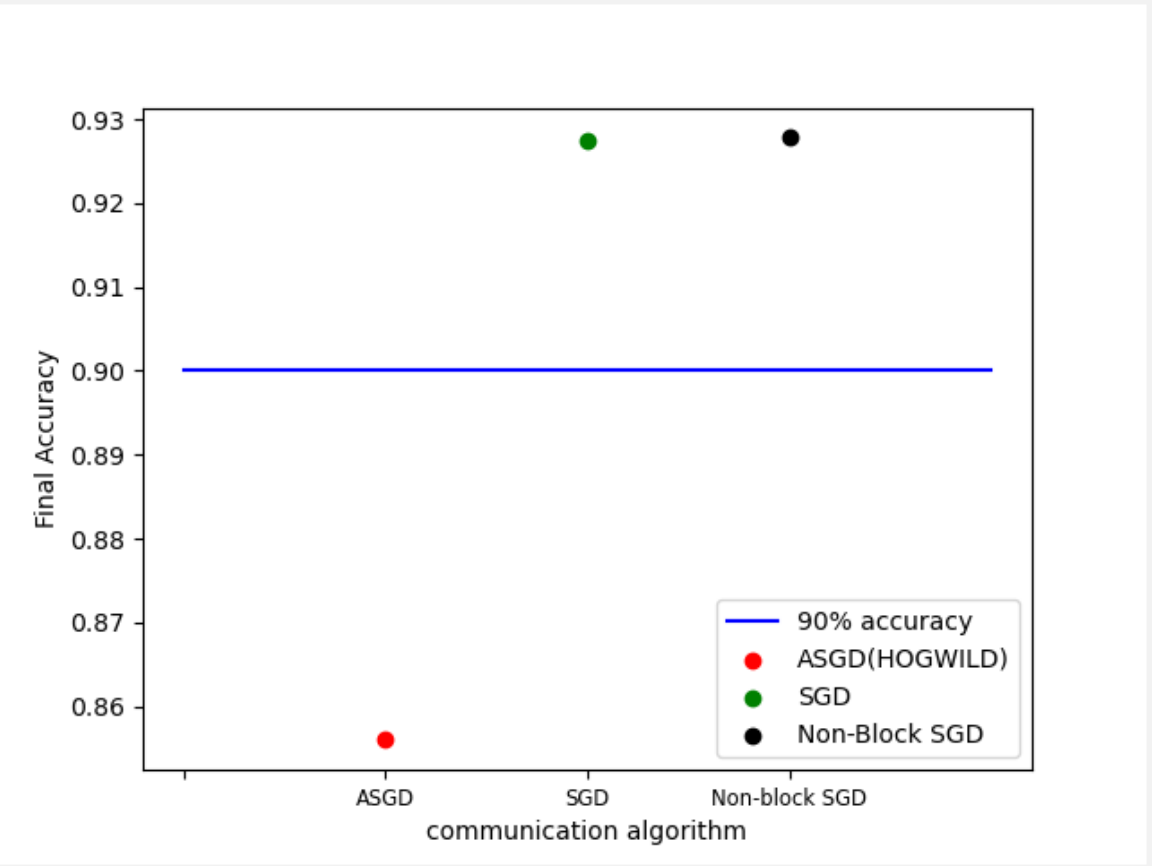
Performance Test

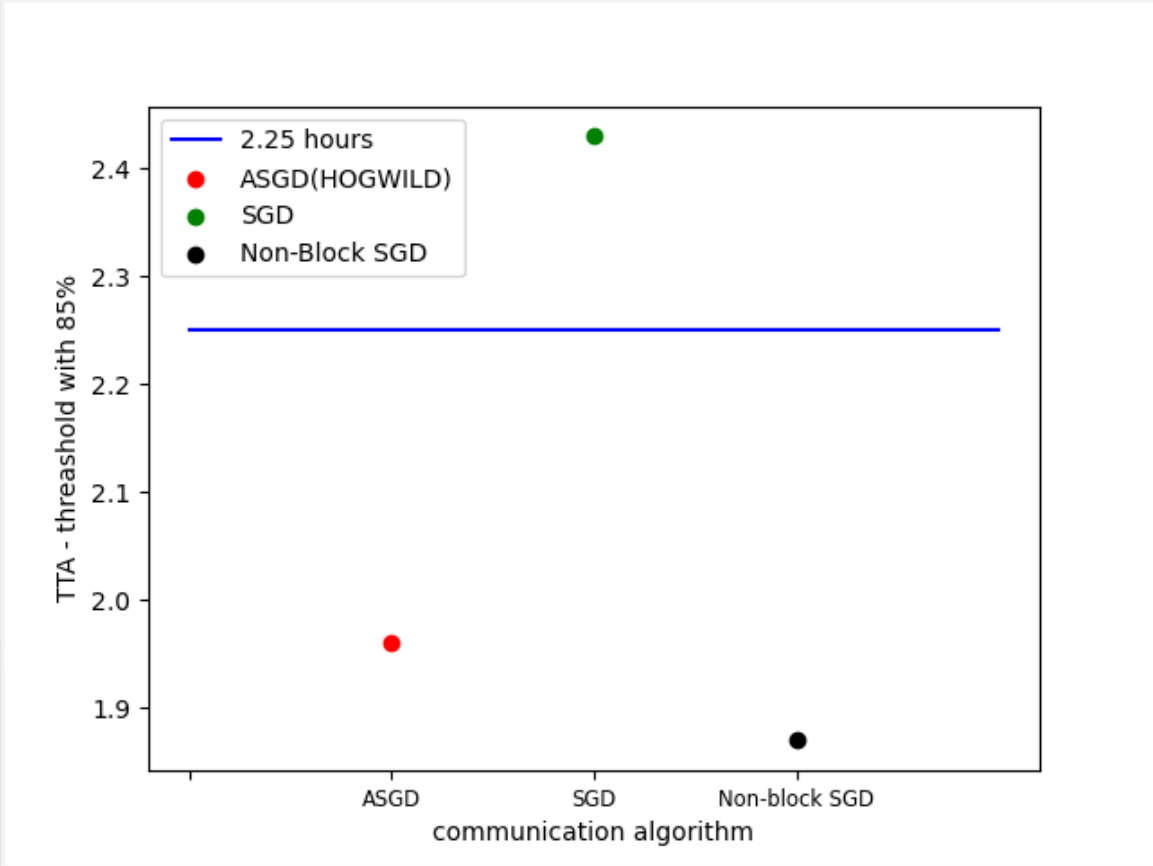
We use PyTorch programs to observe the effect and test our method and measure the TTA as well as the final accuracy compared to other methods such as ASGD, HOGWILD(SGD), Downpour. Timing(TTA) and accuracy are tested.

Code Approach

Using torch.Distributed and MPI as a backend to build the communication scheme. Run the code on NYU HPC with multiple nodes. Single CPU/task per node.

Main Results





- Results are got from average among 5 experiments with the same configurations
- Non-block SGD has around 23% more time saving than SGD

Observations & Conclusions

- Non-blocking SGD has similar accuracy to SGD and is much higher than ASGD
- Non-blocking SGD has a similar TTA to ASGD and is much quicker than SGD
- Non-blocking SGD has even better TTA than ASGD when a threshold is high enough(85%)

GitHub Link

GitHub Link: https://github.com/HectorHHZ/HPML

NEW YORK UNIVERSITY

THANK YOU!