Dear Reviewers,

This cover letter briefly introduces the composition of the supplementary materials. It consists of three parts:

1. appendix.pdf

This document includes the theoretical proof of DaSGD, mainly the formula derivation of convergence rate.

2. fastai-dasgd code

The algorithm simulation is under the fastai-dasgd folder. It is developed from the fast.ai packages and examples from https://github.com/fastai/fastai. Due to the limitation of the computing resources, we use 8 GPUs to simulate the algorithmic behavior of DaSGD on a large cluster (up to 256 workers). The experiment creates a model instance to simulate the behavior of a worker. A simulation of n workers means n model instances. A model instance takes a set of data batches and iterates τ local SGD updates. At the time of the global synchronization, all of the models are averaged together and every model instance is updated after d more iterations.

We have modified the train.py file in the Fast.Ai and added a function named "fit_asynchronous_average_models_one_cycle" for supporting DaSGD. The added code is compatible with the original code and the one cycle policy in Fast.Ai can be used. In addition, we add many popular neural network models based on CIFAR-10 in the /vision/models file, which can support various experimental results. Finally, we give an example of CIFAR-10 training based on the DaSGD algorithm in train_cifar-10_dasgd.py. A number of smaller modifications were made in the fastai library.

Dataset: Two open datasets, CIFAR-10 (https://www.cs.toronto.edu/~kriz/cifar.html) and ImageNet (http://www.image-net.org), were used in the experiment.

How to run the code? Please see README in the code to reproduce the experimental results.

3. paleo-dasgd code

The performance simulation is under paleo-dasgd folder. It is developed from https://github.com/TalwalkarLab/paleo, including the following changes.

- A system configuration with a server connecting 8 Nvidia TITAN X with PCIE3.0, servers connecting with each other on EtherNet 20Gbps, and 32 such a server.
- Bufferfly All-Reduce scheme.
- The simulation support for such a configuration.
- A new GPU configuration simulating Nvidia Tesla V100.
- A system configuration with a server including 8 Nvidia Tesla V100 interconnecting NVLink2.0, servers connecting with each other on InfiniBand 70Gbps, and 32 such a server.
- Two-level hierarchical All-Reduce scheme, including Tree All-Reduce between servers and Ring All-Reduce between GPUs on the same server.

- The simulation support for such a configuration. This document includes the theoretical proof of DaSGD, mainly the formula derivation of convergence rate.

How to run the code? Please see README in the code to reproduce the experimental results.

Thank you very much! If you have any question, please do not hesitate to contact us.

With best regards!