# Trade-offs of Local SGD at Scale

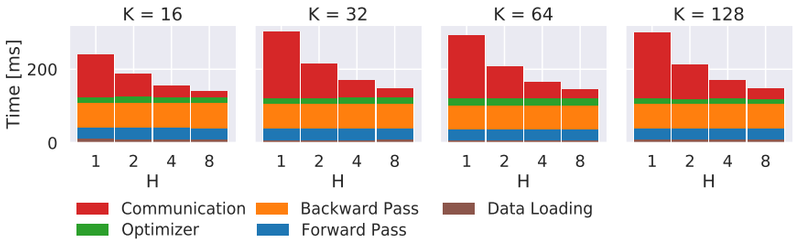
Jose Javier Gonzalez Ortiz  Jonathan Frankle  Mike Rabbat  Ari Morcos  Nicolas Ballas

FACEBOOK AI

- In distributed training, *local SGD* reduces communication compared to *minibatch SGD* (mSGD). Local SGD averages the weights across the ***K*** workers every ***H*** iterations.
- Local SGD speeds up training but often at the cost of accuracy.
- *Post-local SGD* is a variant that performs mSGD training for ***T*** epochs and then switches to Local SGD.

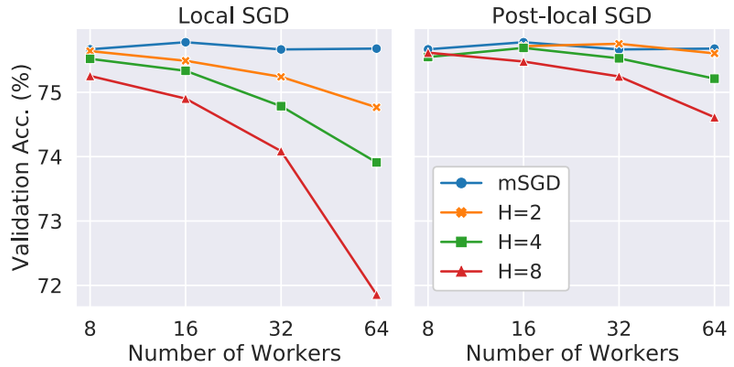
We study training ResNet-50 and ResNet-101 on ImageNet using local and post-local SGD as we vary ***H***, ***K*** and ***T***. We identify several scalability limitations of local and post-local SGD.

Minibatch SGD (H=1) spends the majority of the time communicating, whereas local SGD amortizes communication across several iterations.



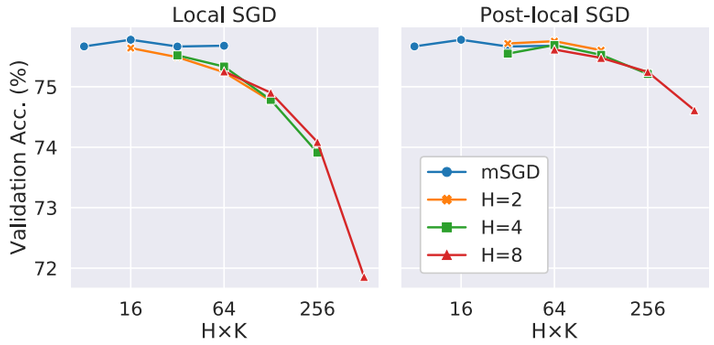

8 V100 GPUs per Node – 10GBps Ethernet interconnect

As either the number of workers ***K*** or the number of local steps ***H*** increases, final model accuracy decreases.

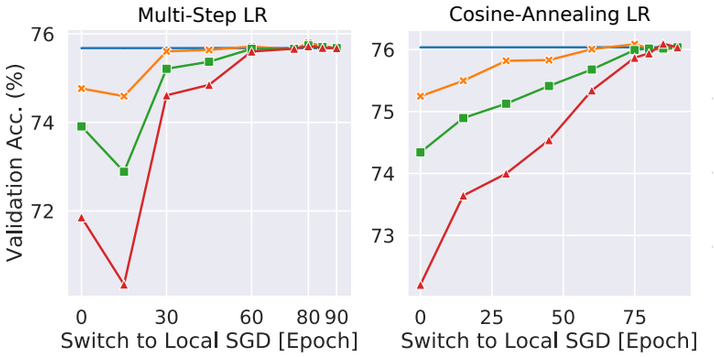



## Our empirical study on ImageNet identifies previously unknown scalability issues of [Post]Local SGD.

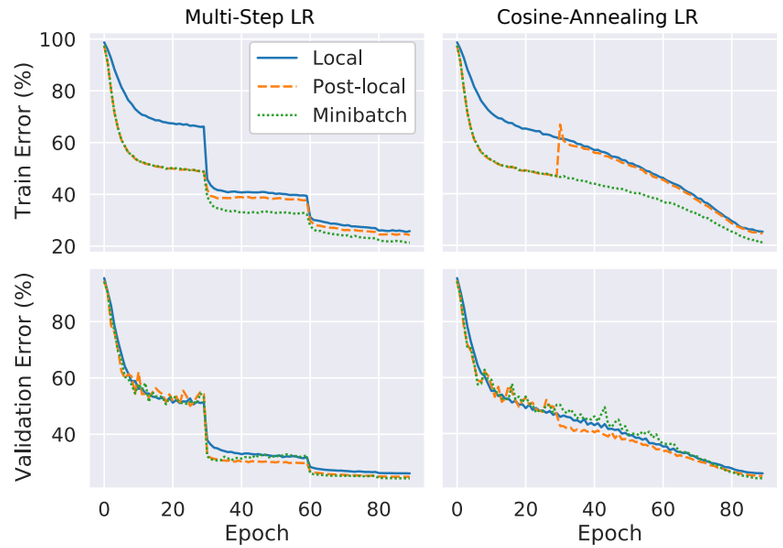
The accuracy-communication trade-off is closely related to the product of ***H*** and ***K***, i.e. the total number of local model updates between synchronizations.




- The switching point in post-local SGD presents a trade-off between training time and final accuracy.
- The choice of learning rate schedule has a large impact for the final accuracy of post-local SGD.




Switching to local SGD has a regularization effect on optimization that is beneficial in the short term, but ends up being detrimental to final model accuracy




Regardless of when the switch to local SGD is performed, the loss and error curves follow the same general trajectory.