SWALP: Stochastic Weight Averaging for Low-Precision Training

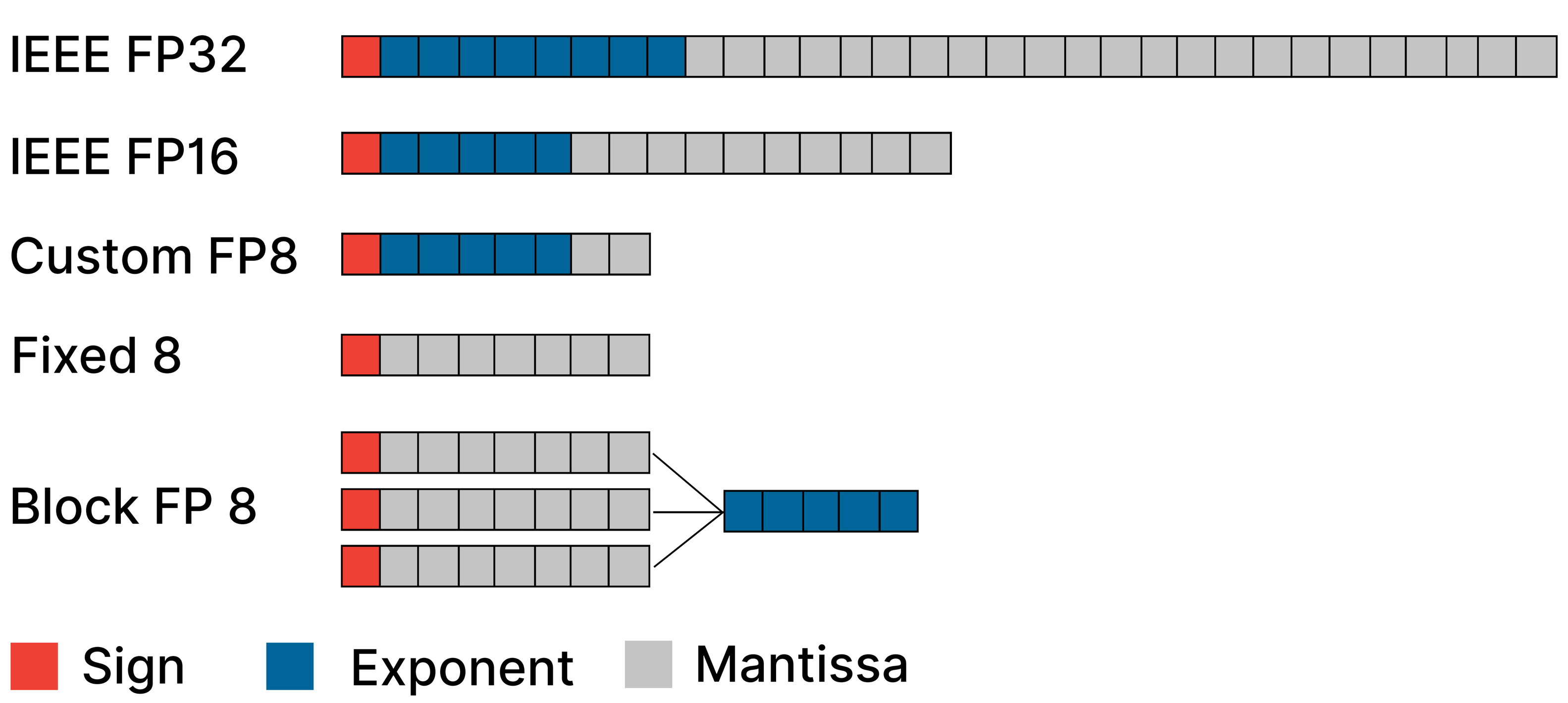
Guandao Yang, Tianyi Zhang, Polina Kirichenko, Junwen Bai, Andrew Gordon Wilson, Christopher De Sa

Cornell University

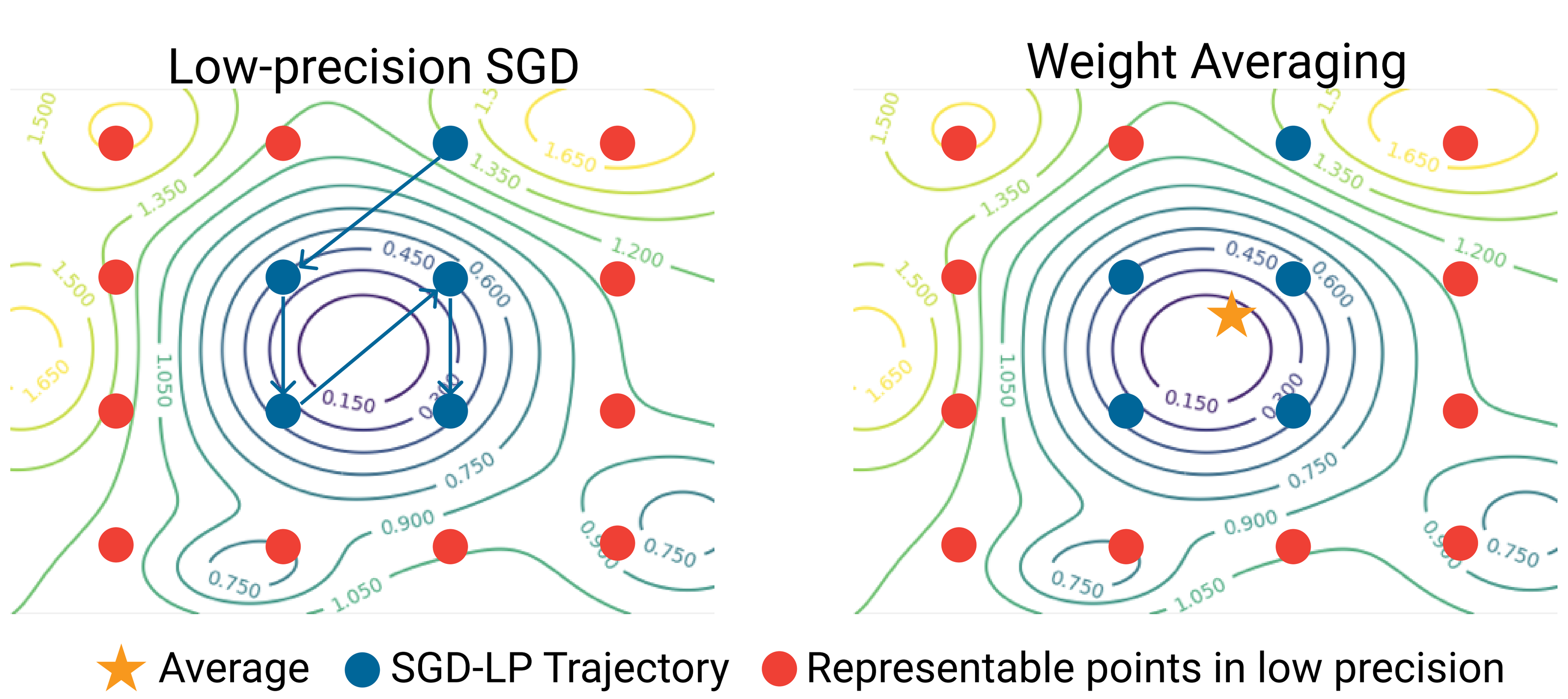
This work

- Studies how to leverage low-precision training to obtain a high-accuracy model, which may be higher-precision.
- Proposes a principled approach to using stochastic weight averaging in low-precision training (SWALP).
- Shows SWA sigificantly reduce the performance gap between low-precision and full-precision training.

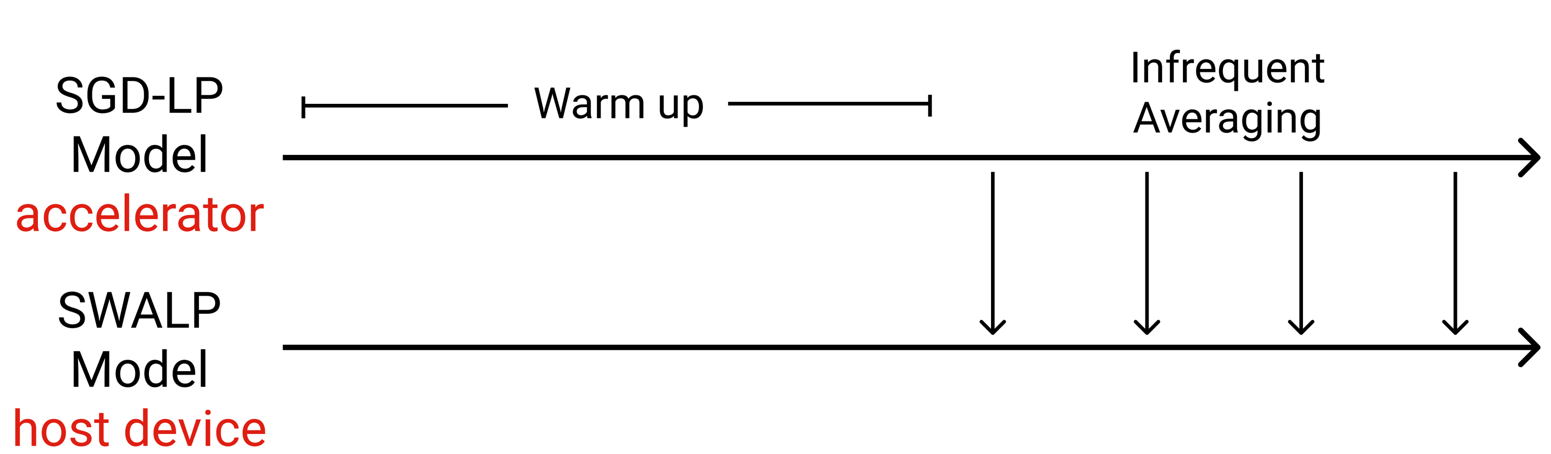
Low-Precision Computation



SWALP



- Low-precision representation inherently limits the accuracy.
- By averaging, we hope to recover a better solution.



Convergence Analysis

Let T be the number of iterations and δ be the quantization gap (the difference between two successive representable numbers). With standard assumptions and fixed point quantization, we can prove the following statements.

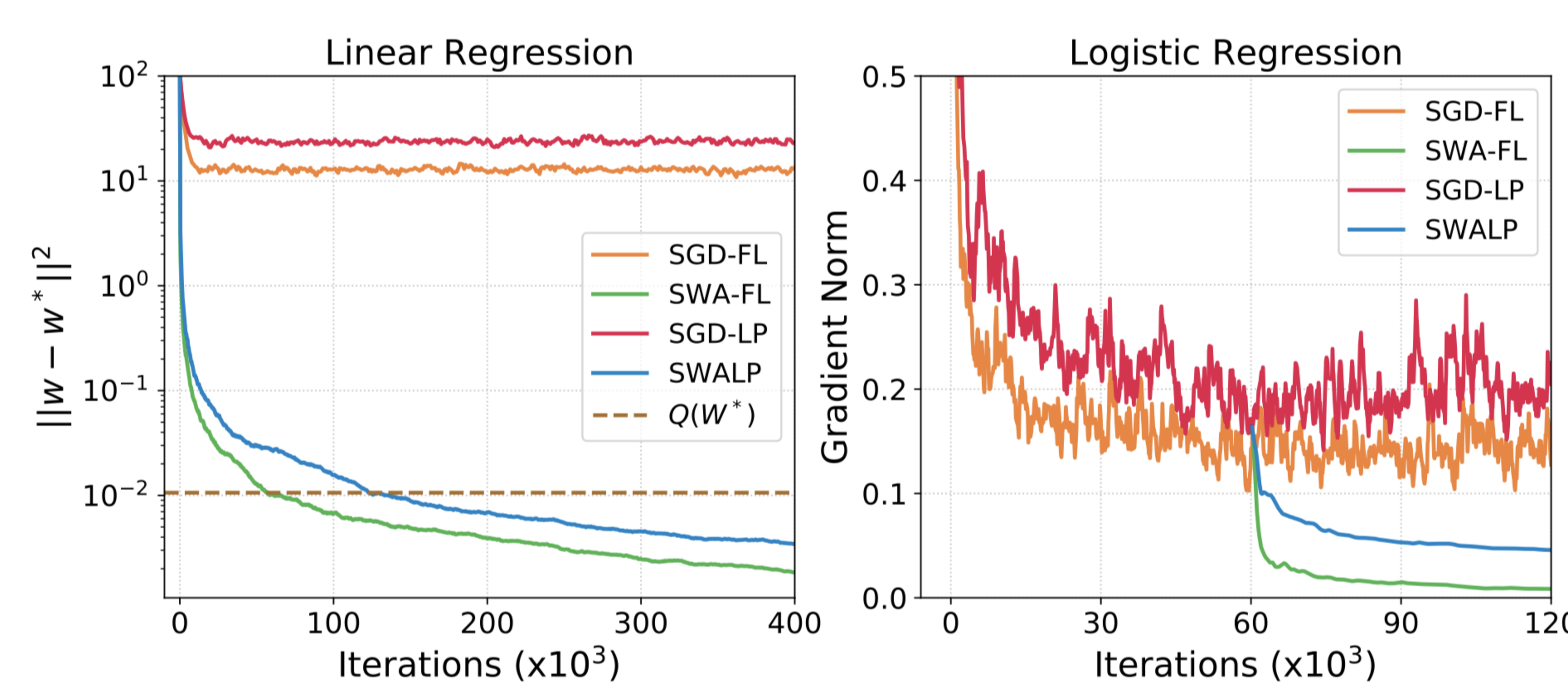
Theorem 1 (Quadratic)
The expected squared distance between the SWALP solution and the optimal one converges to 0 at a O(1/T) rate.

- SWALP has the same O(1/T) convergence rate with full-precision SGD.
- SWALP converges to the optimal solution regardless of the numerical precision.

Theorem 2 (Strongly Convex)
The expected squared distance between the SWALP solution and the optimal one has a $O(\delta^2)$.

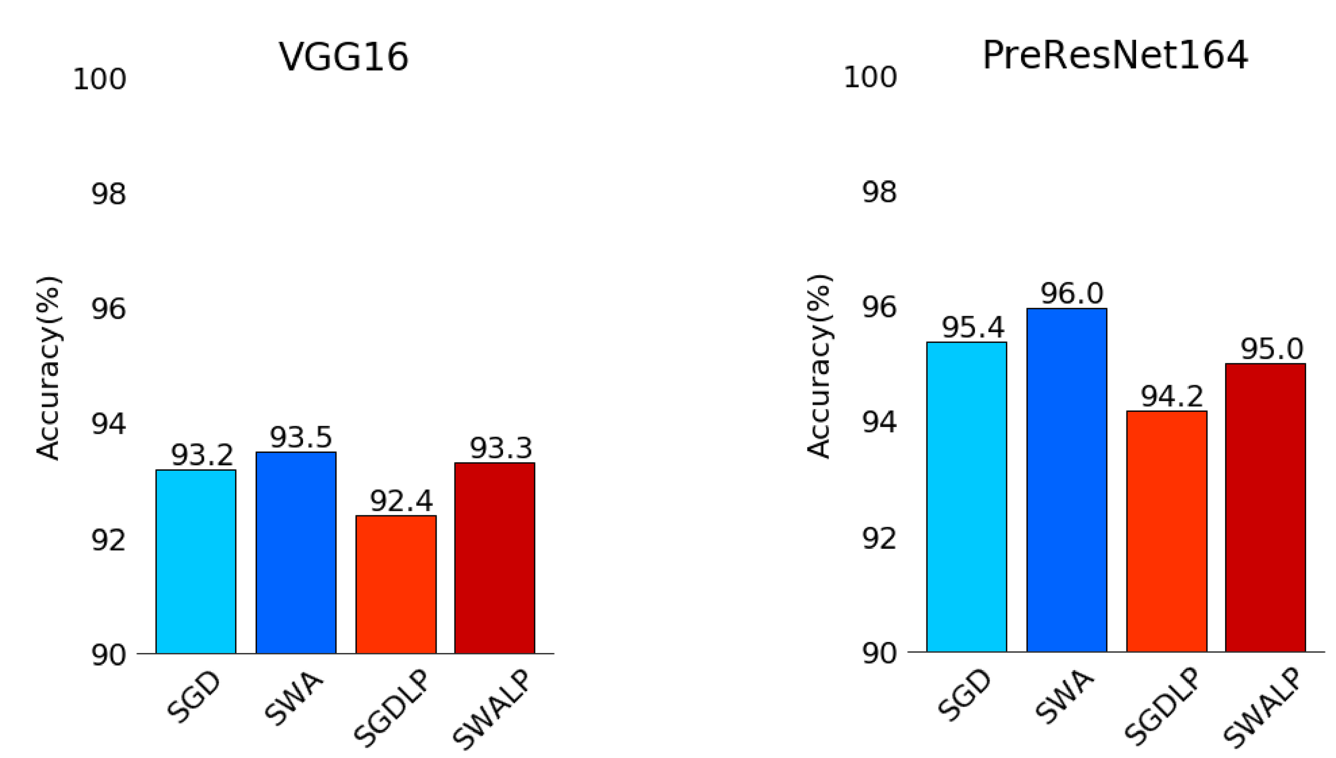
- The best bound for low-precision SGD is O(δ) (Li et al, 2017).
- SWALP requires half of the number of bits to reduce the noise ball by the same factor.

Experimental Validation

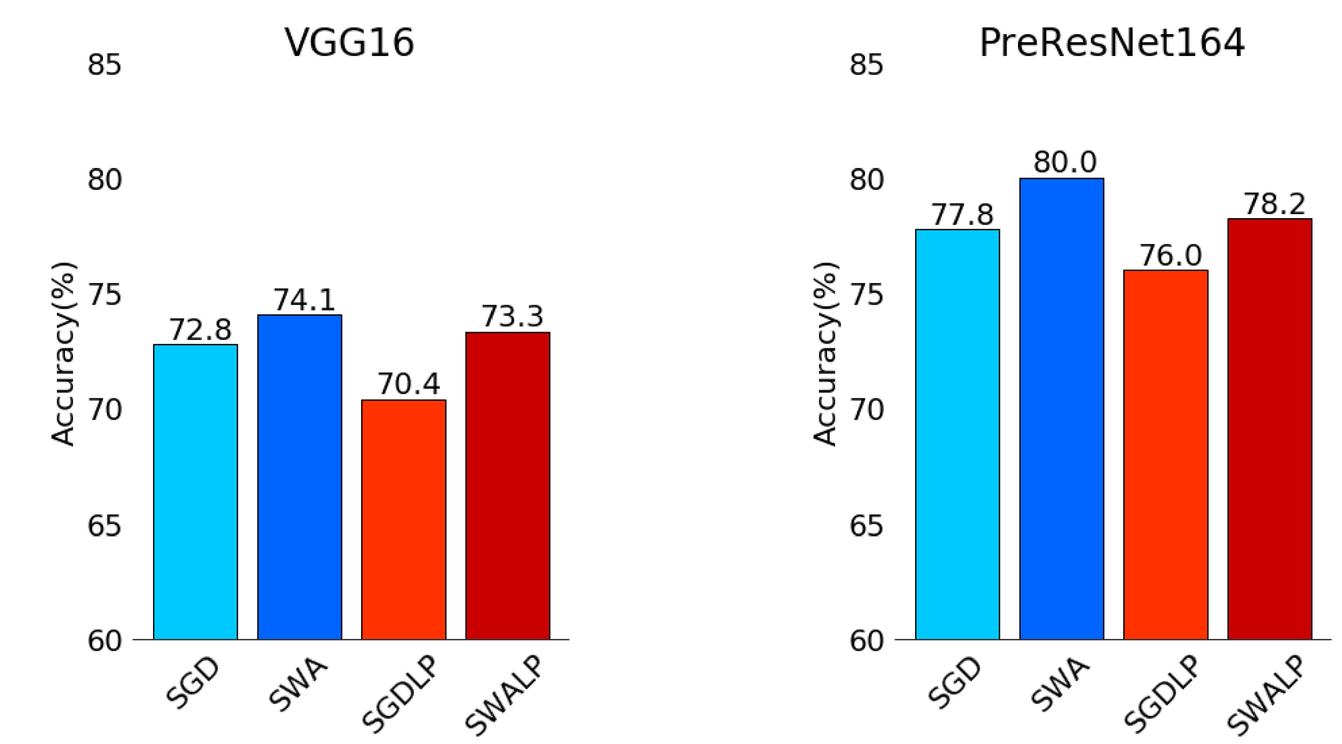


Experiments

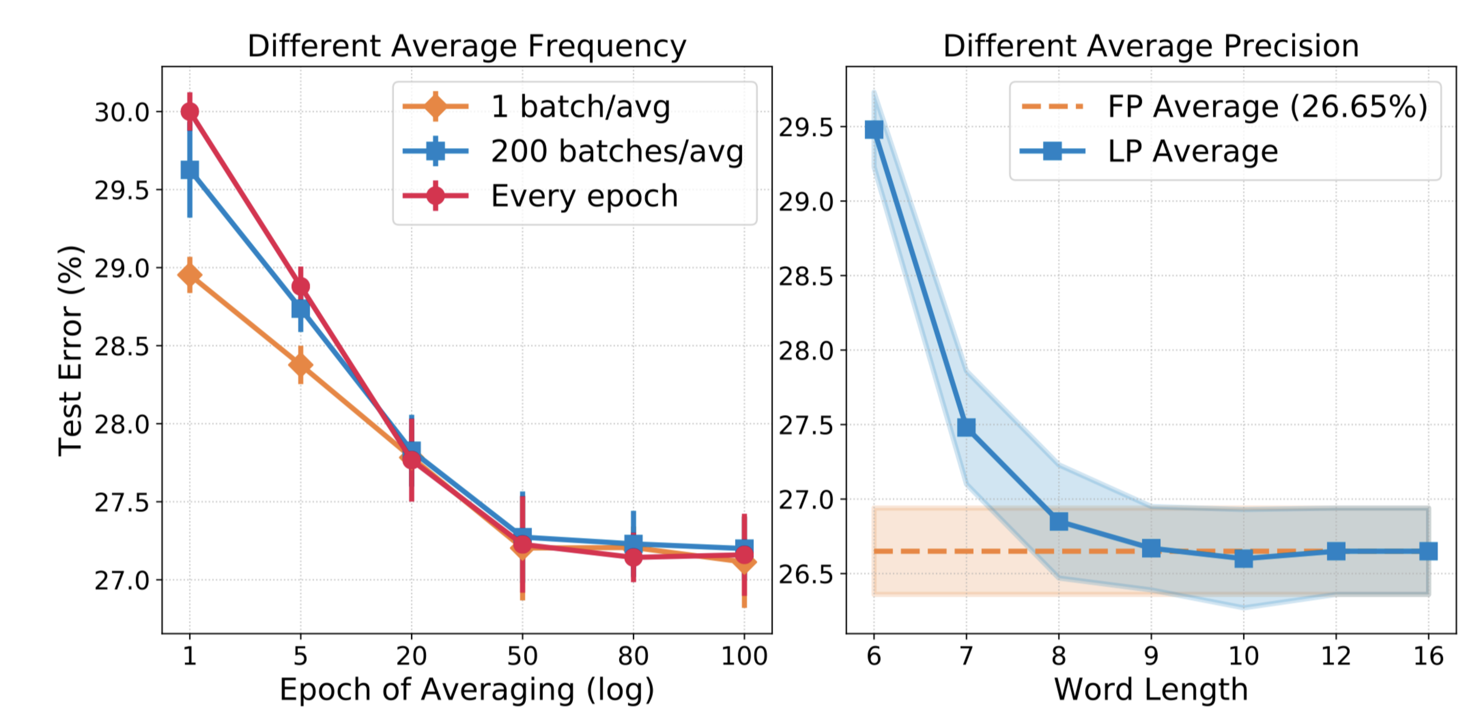
Results: CIFAR10



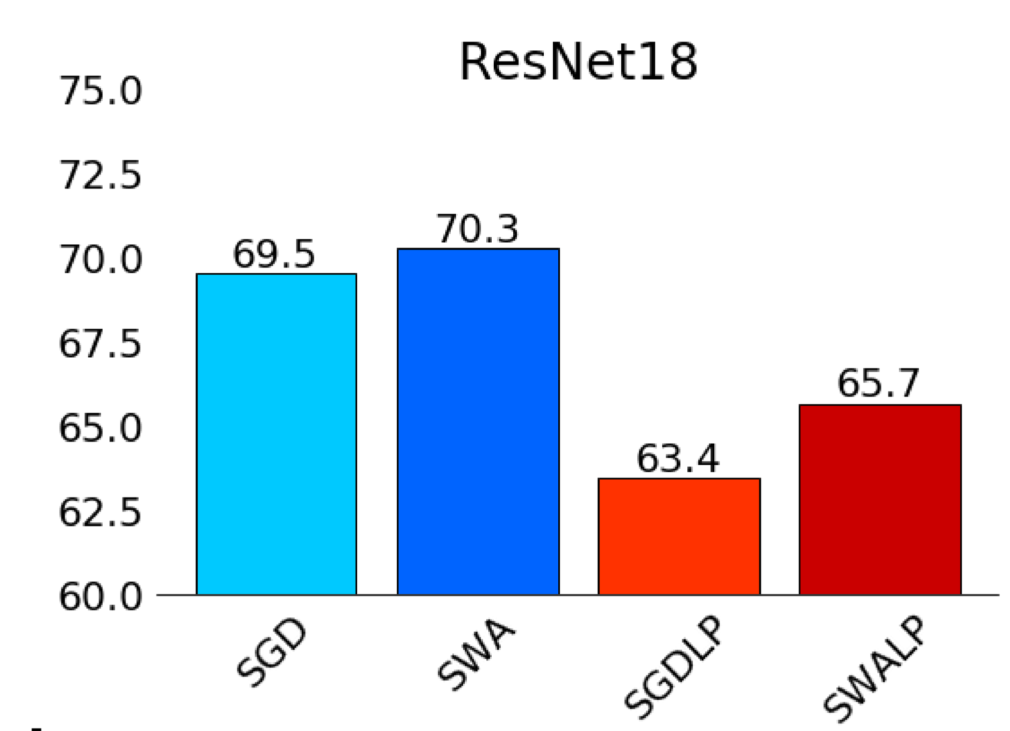
Results: CIFAR100



Averaging in Different Precision and Frequency



Results: ImageNet



QPyTorch

We release QPyTorch, a low-precision arithmetic simulation package in PyTorch. A diverse range of quantization methods is supported with GPU acceleration.