Trade-offs of Local SGD at Scale: An Empirical Study

Facebook AI Montreal

Jose Javier Gonzalez Ortiz

Nicolas Ballas

Mike Rabbat

Ari Morcos

Jonathan Frankle

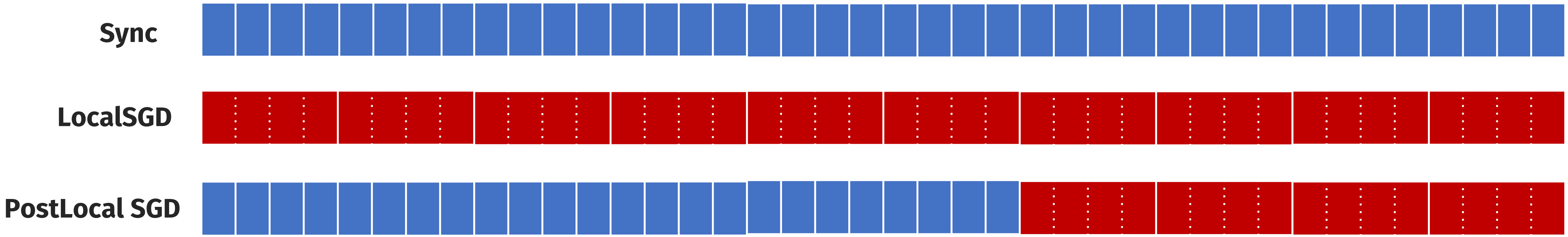
Overview

- First comprehensive empirical study of local SGD and post-local SGD on ImageNet.
- We find several trade-offs that impact the scalability of these methods, a departure from smaller-scale experiments in prior work
- We study the effect of learning rate and momentum, hinting at future directions to improve the trade-offs

1 - Preliminaries

Algorithms

- Distributed Data Paralellism (DDP) synchronizes gradients every step by averaging
- Local SGD instead averages the parameters every K steps
- PostLocal SGD does DDP for a while then switches to Local SGD



[2] Don't Use Large Mini-Batches, Use Local SGD, Lin et al (ICLR 2020)

Experimental Setup

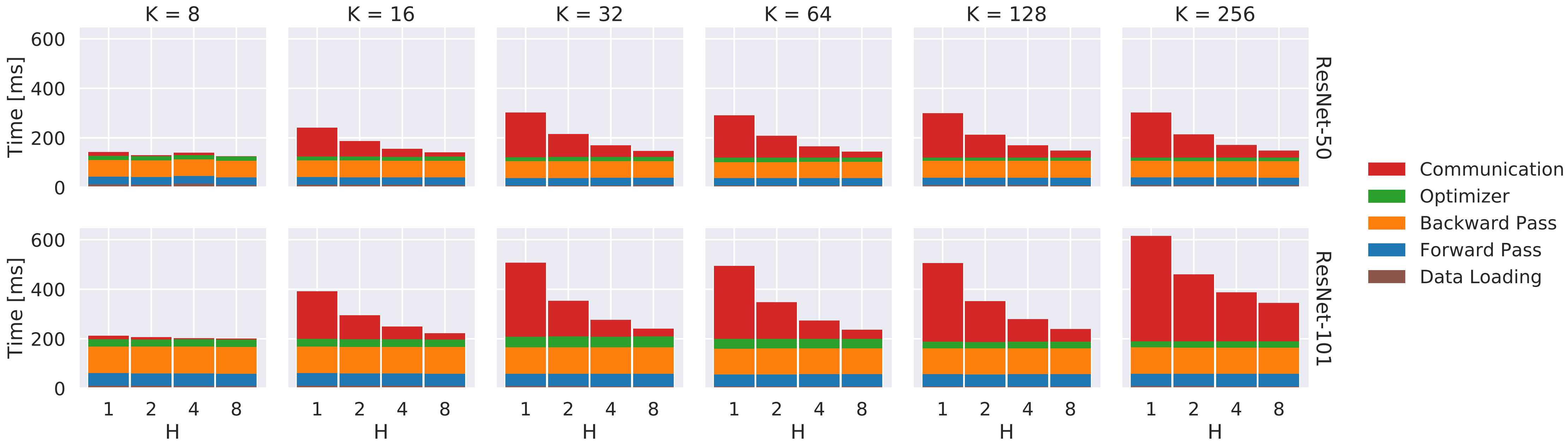
All Experiments are on:

- **Dataset:** ImageNet
- **Models**: ResNet50 & ResNet 101 (with Goyal init)
- **Optimizer**: SGD + Momentum(.9) + Nesterov
  - Momentum Correction
  - $\eta = N \cdot B \cdot 4 \times 10^{-4}$
- **Training:**
  - 90 Epochs
  - LR drop by 10 at epochs 30, 60, 80
  - Linear Warmup for 5 epochs

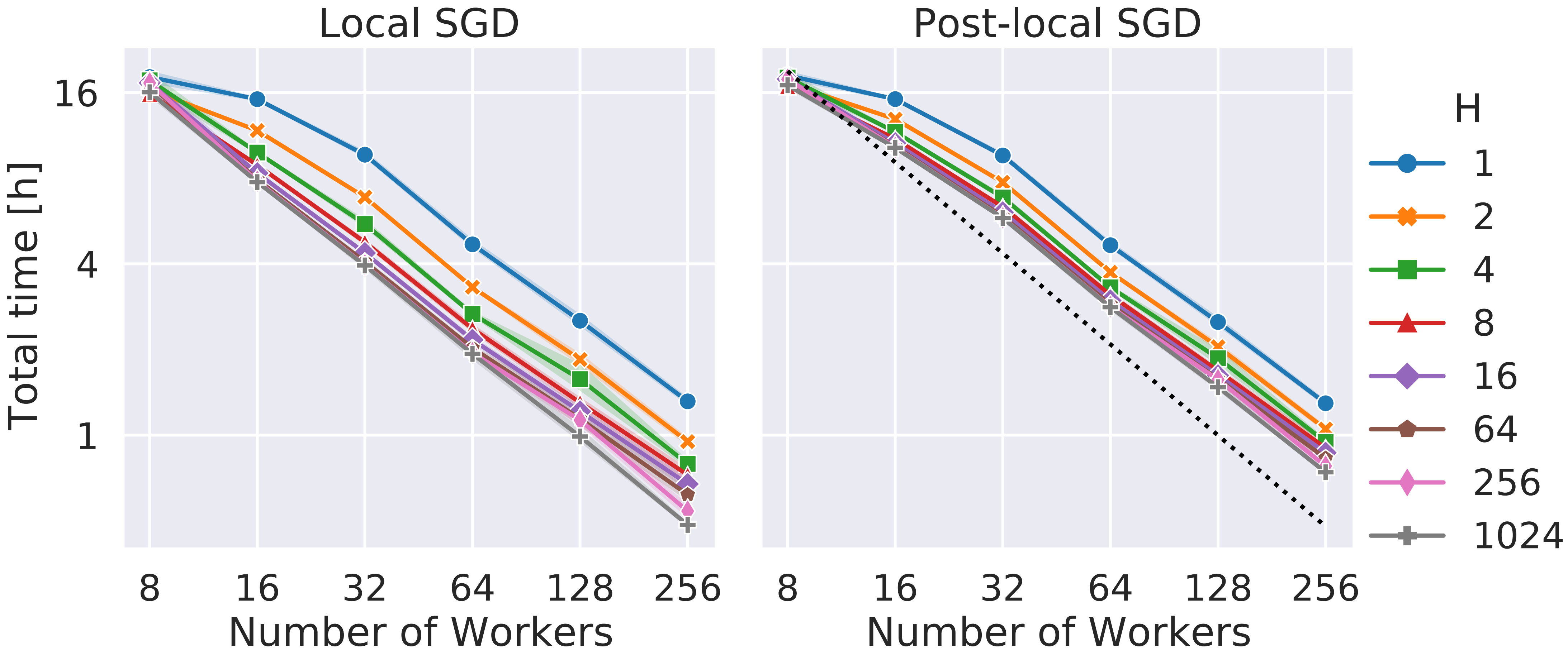
2 – Benefits of (Post-)Local SGD

Local SGD reduces communication overhead

8 Nvidia V100 per node over 10Gb/s Ethernet interconnect

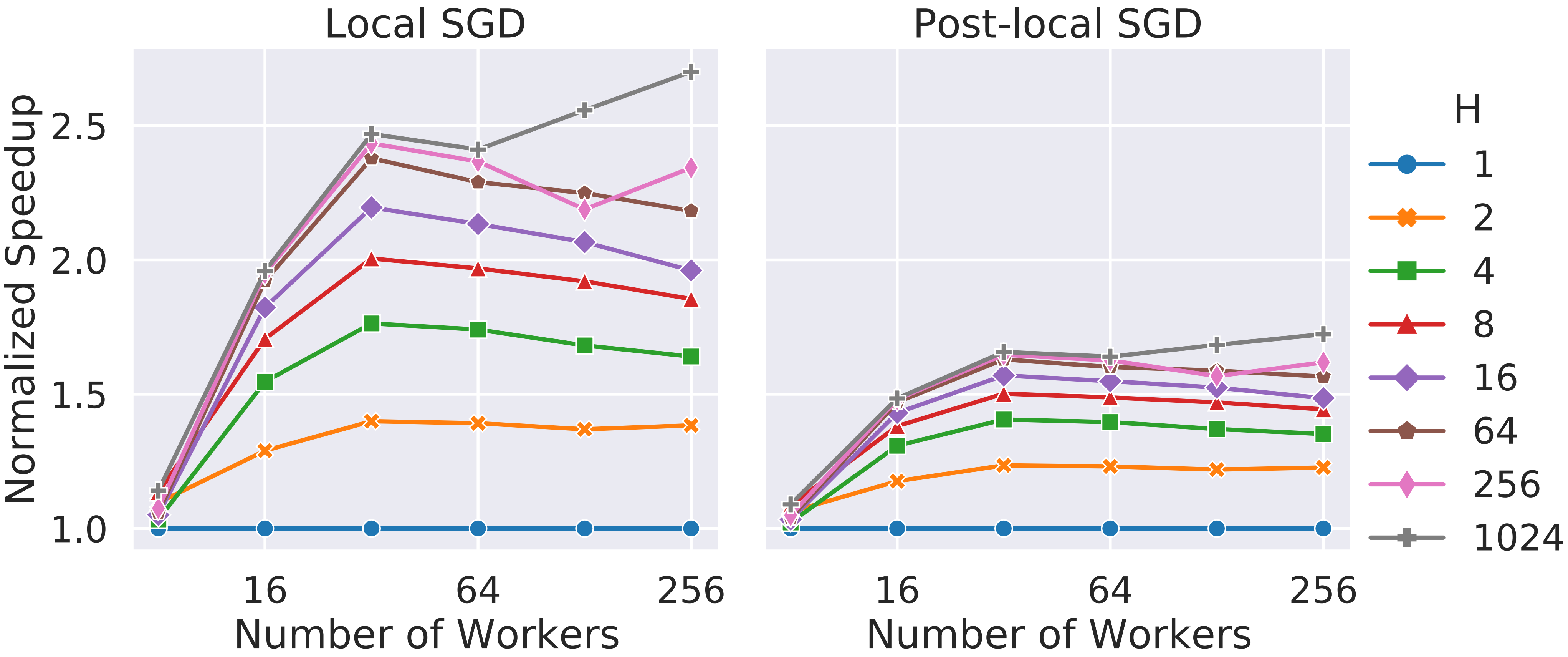


Local SGD scales to large distributed settings



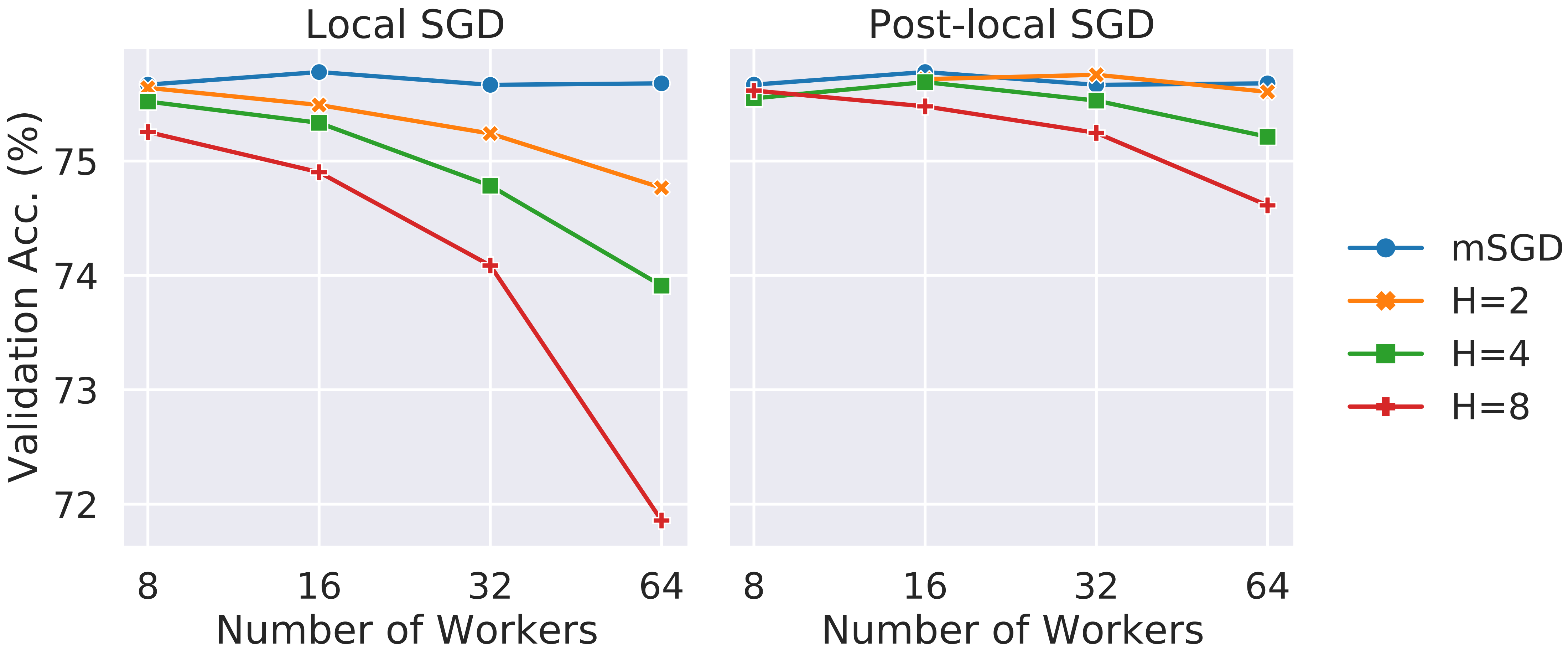
Post-local SGD is limited by the synchronous phase

For post-local SGD we get Ahmdal's law behaviour

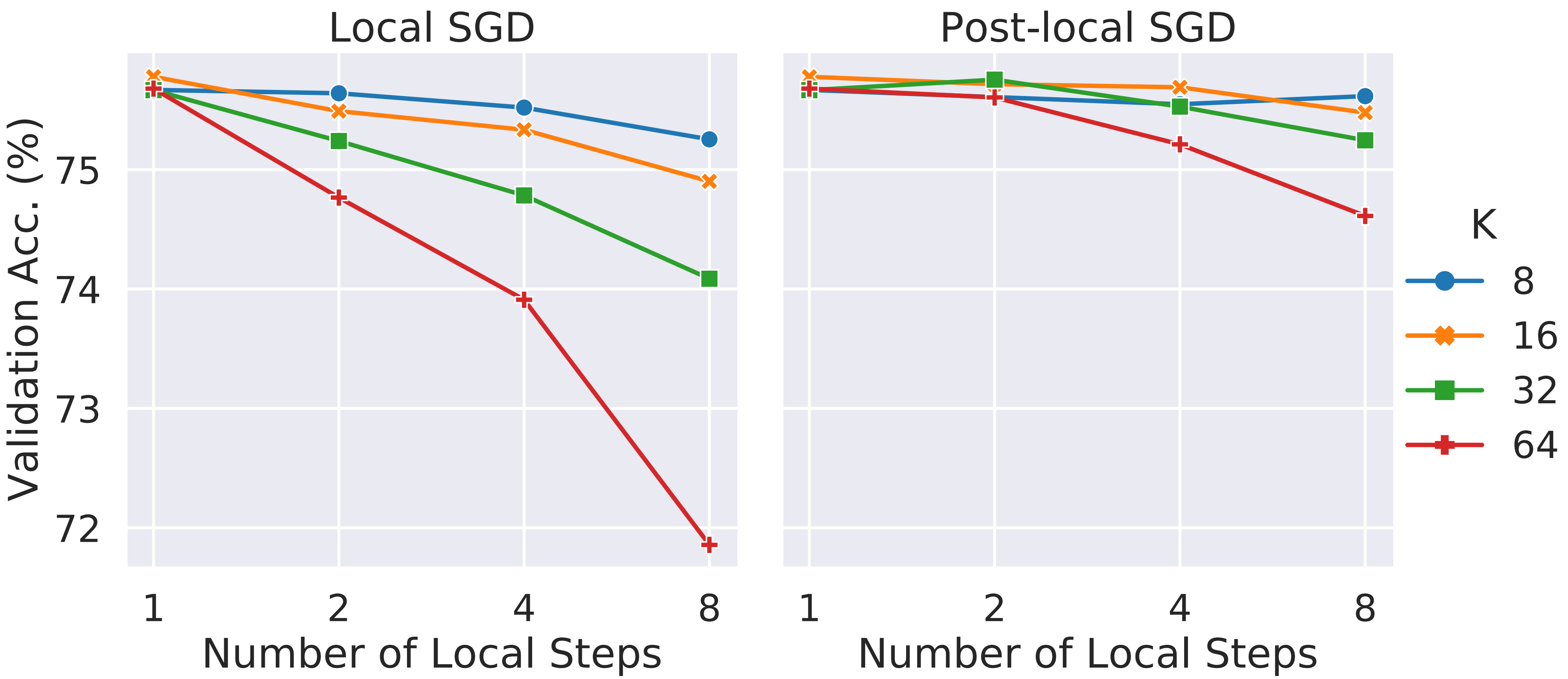


3 – Trade-offs of (Post-)Local SGD

Increasing the number of workers (*K*)

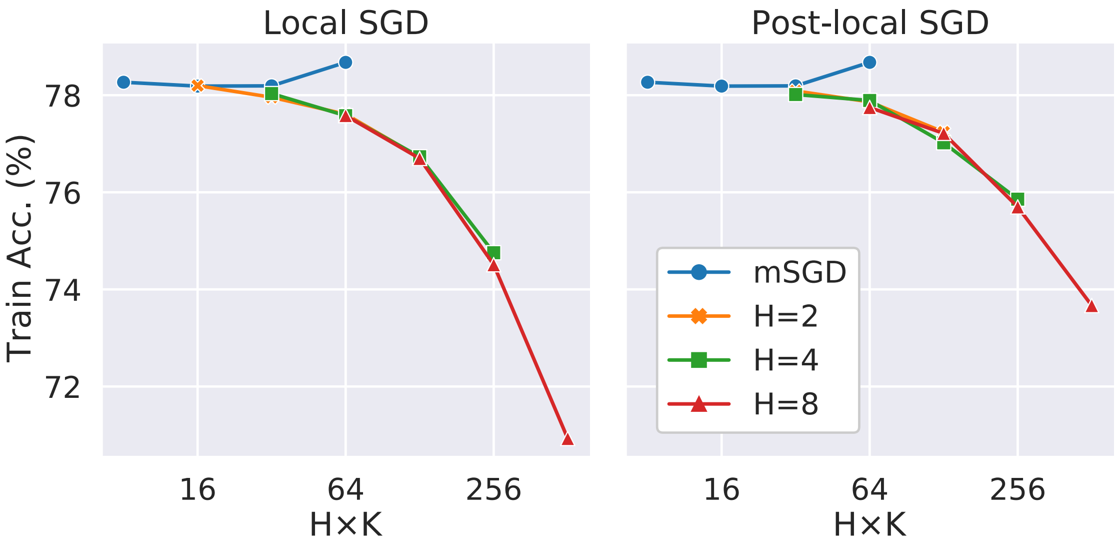


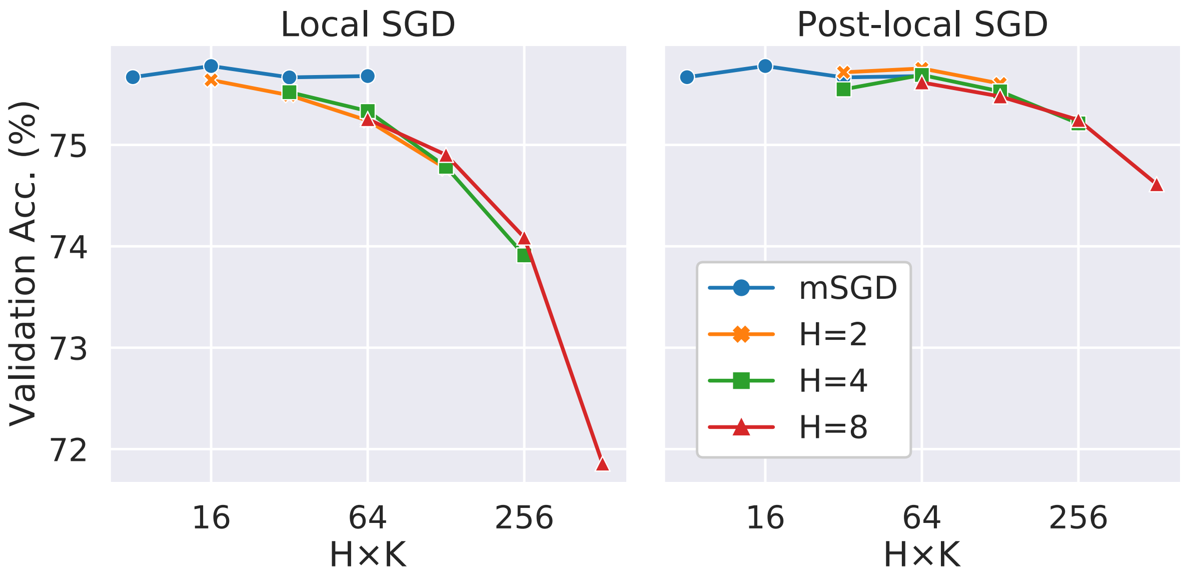
Reducing averaging frequency (*H*)



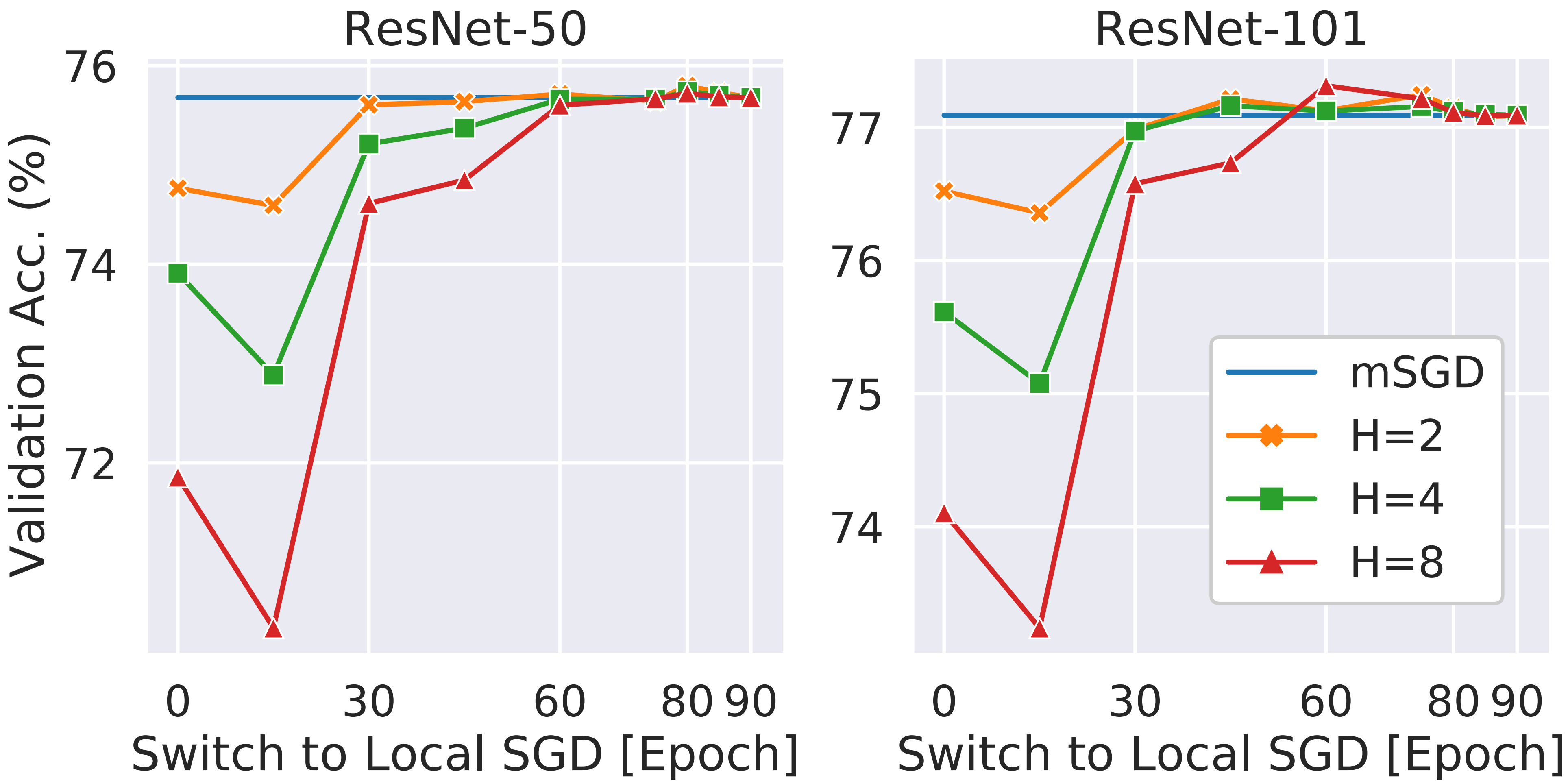
Towards a unified trade-off

Trade-off is better expressed as *H·K*. I.e. number of local model updates between synchronizations



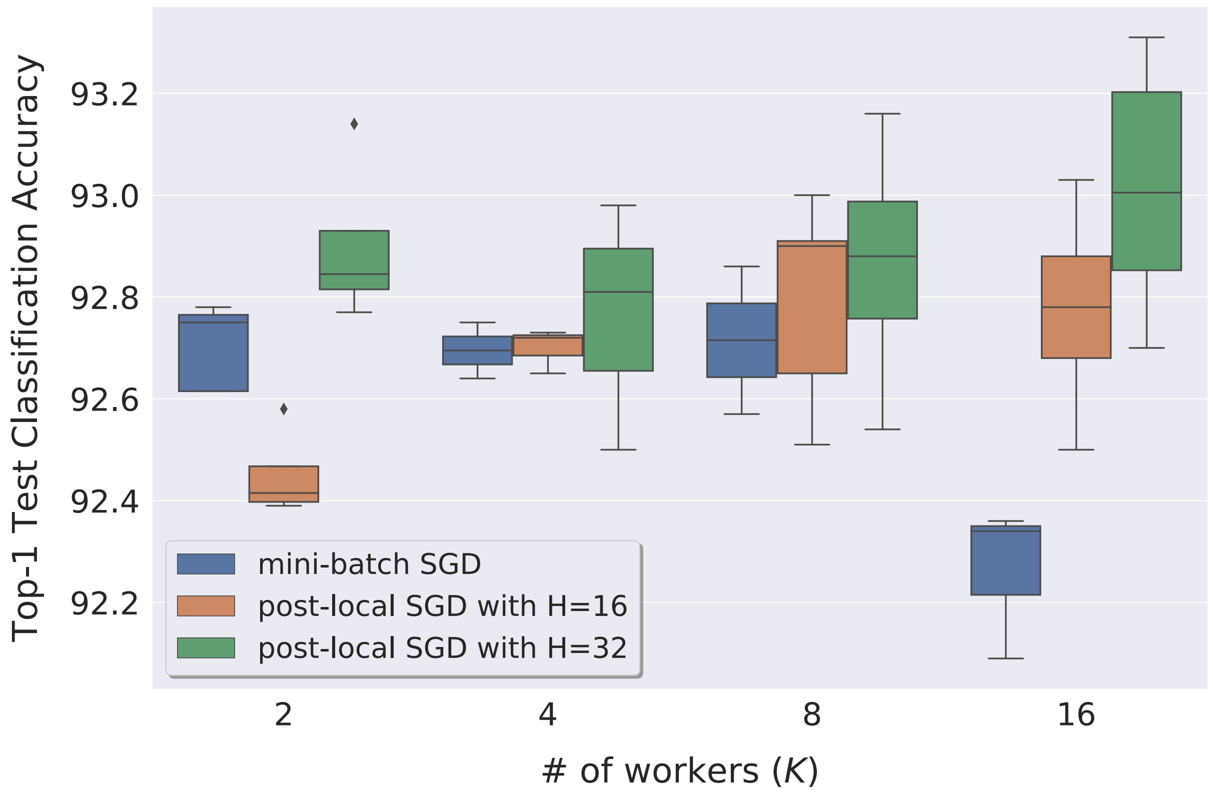


The switching point poses another trade-off



Post-local SGD behavior depends on the task scale

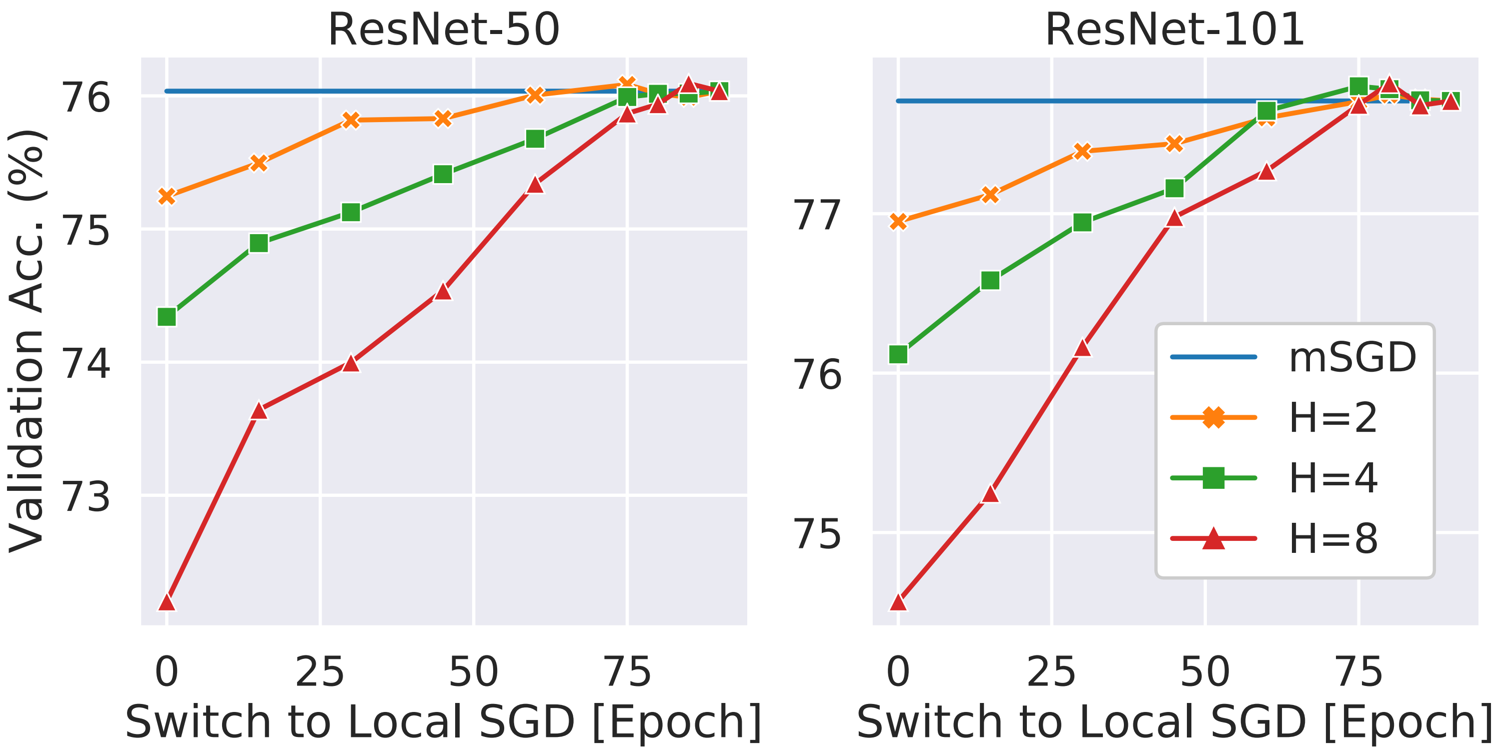
Departure from Lin et al (2018) results for ResNet-20 on CIFAR-10 (Improved Acc as H or K increases)



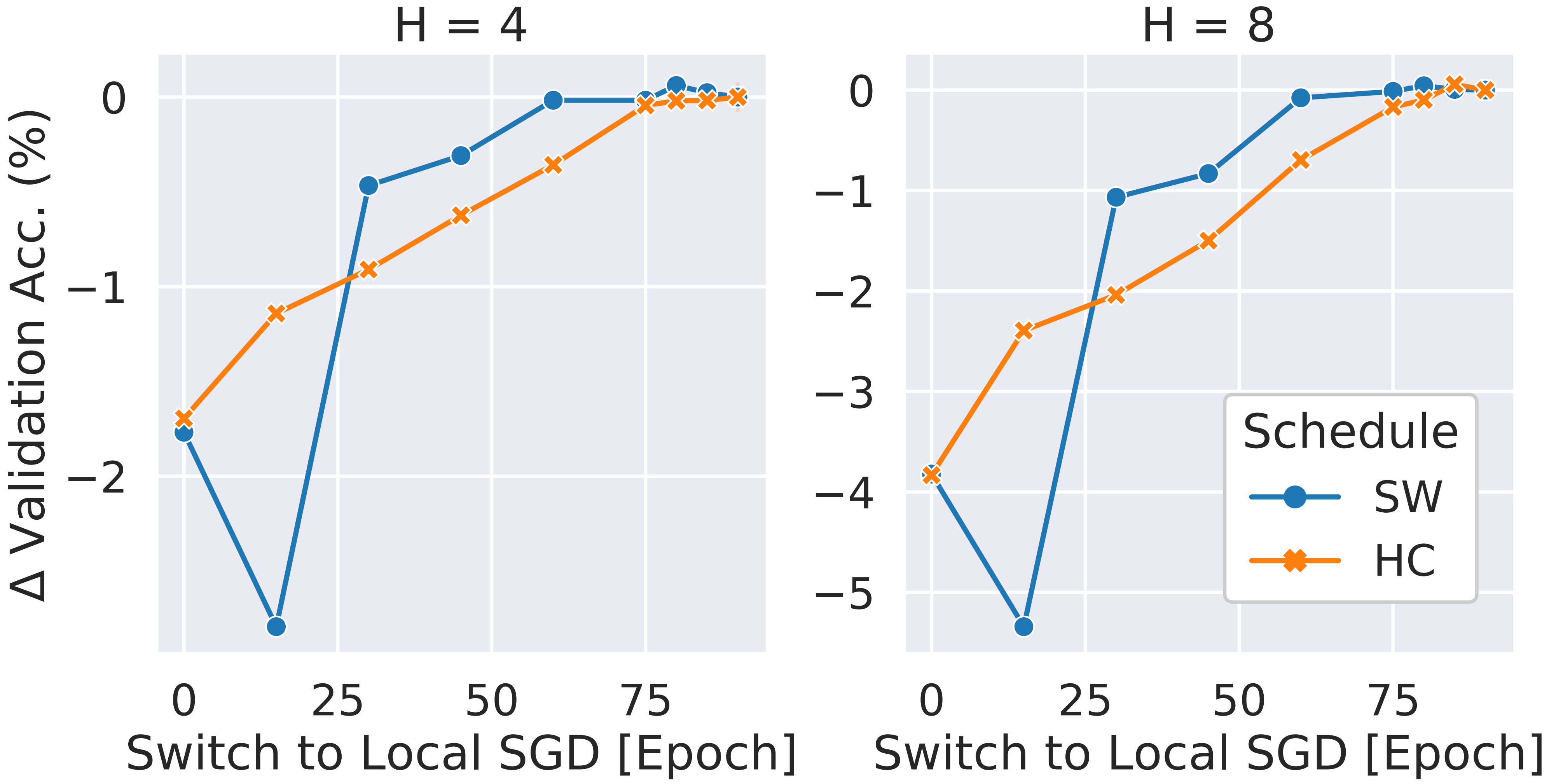
4 – Expanding the design space: Learning Rate & Momentum

Post-local SGD depends on learning rate schedule

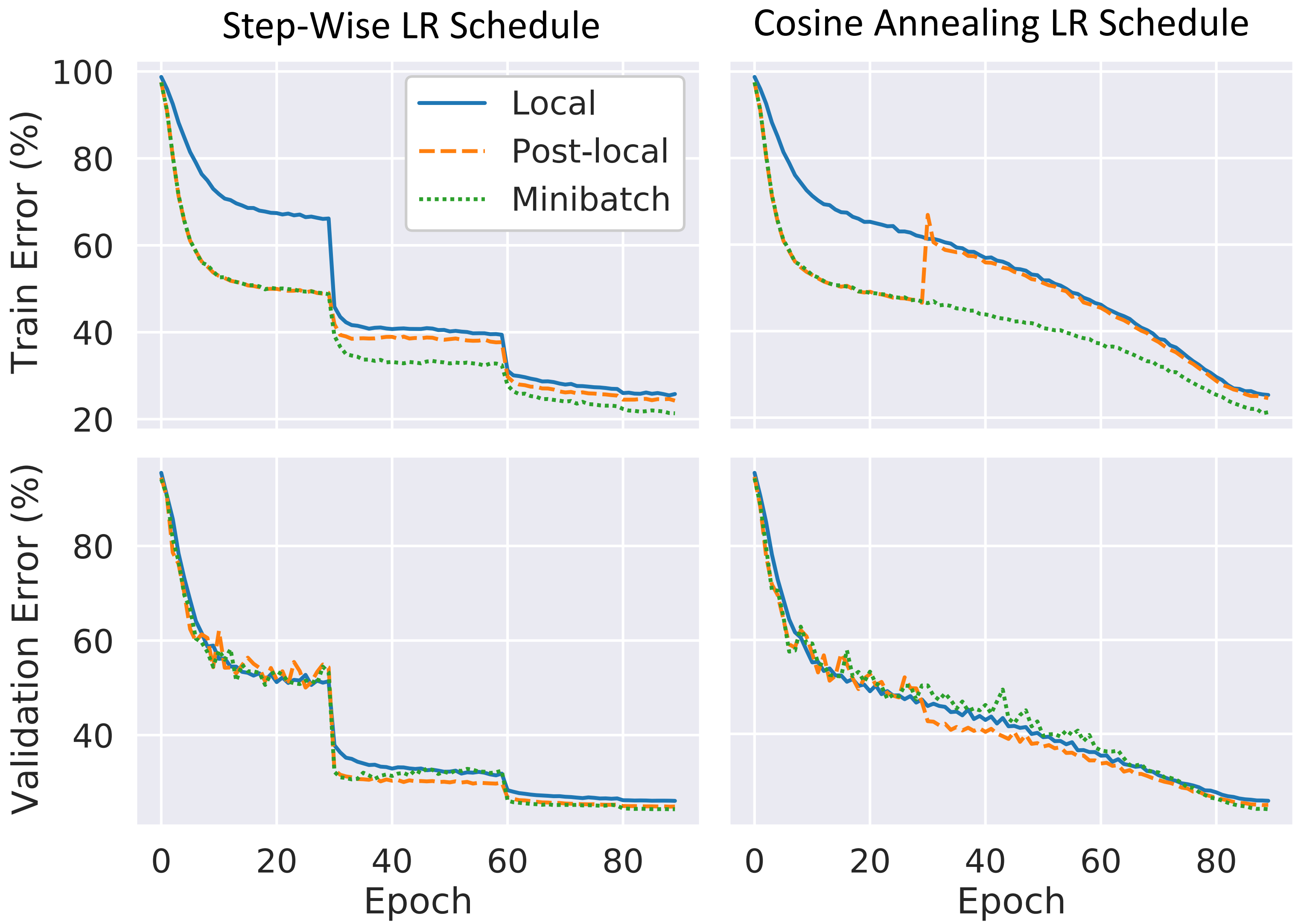
- Half Cosine Schedule



Post-local SGD depends on learning rate schedule



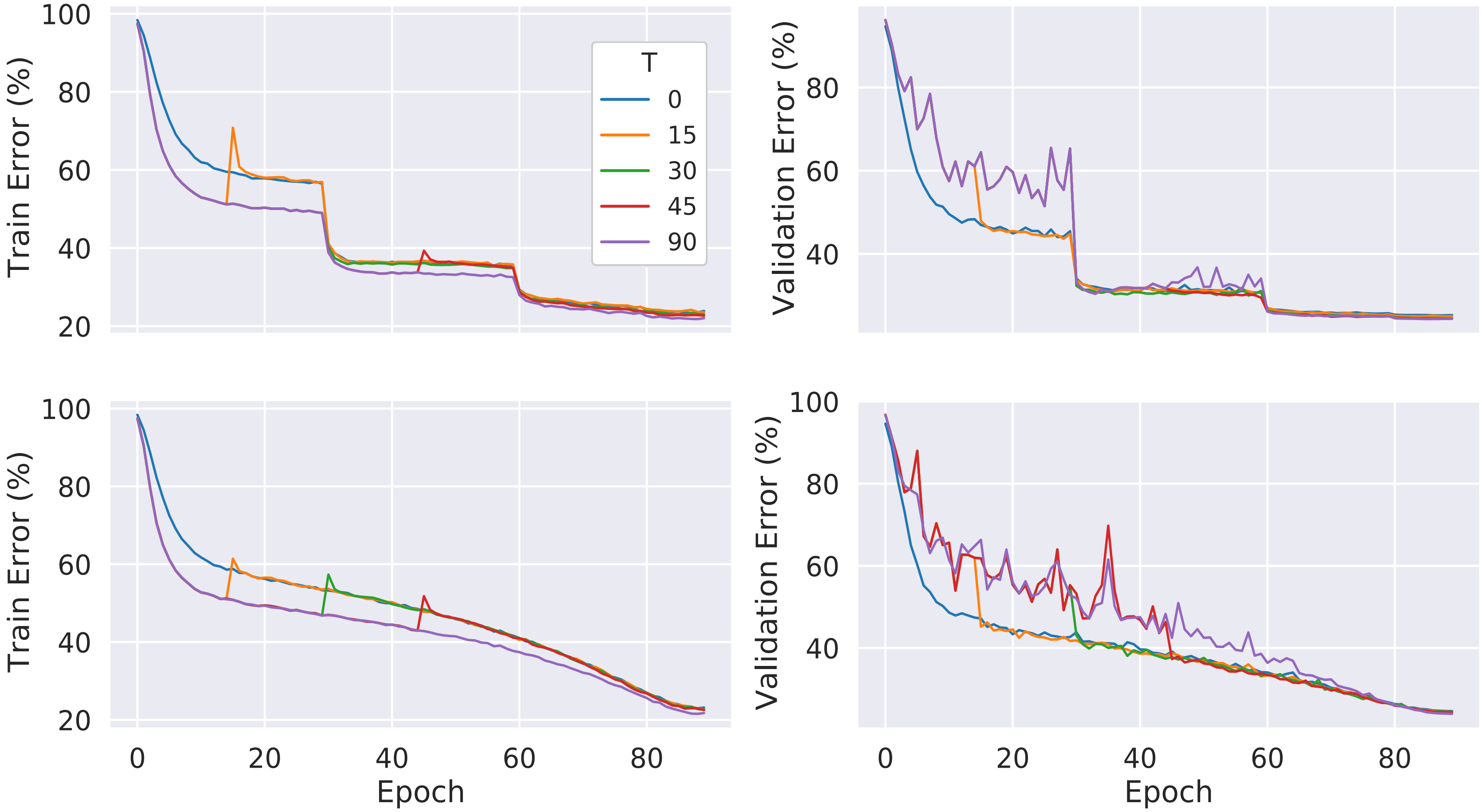
Local SGD as a regularizer



Switching to Local SGD is beneficial in the short term but it's detrimental to final model accuracy

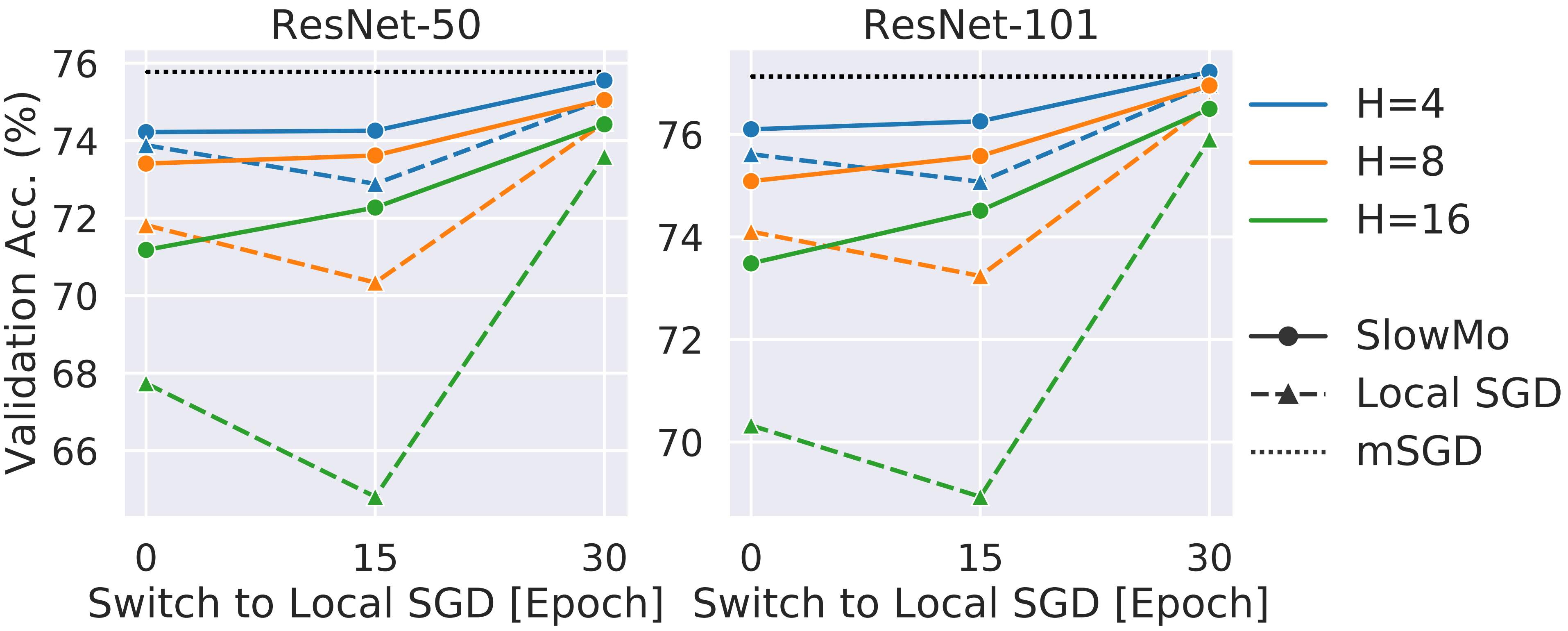
Local SGD as a regularizer

The amount of regularization effect is related to Local SGD



Improving accuracy with slow momentum

SlowMo consistently improves accuracy, specially for early switch points



Conclusion

- (Post-)Local SGD has several **trade-offs** that impact its **scalability** (# workers, averaging frequency, switching point)
- Switching to **Local-SGD** has a **regularization** effect, beneficial in the short term but detrimental to final model accuracy
- Post-local SGD viability heavily relies on the **LR schedule**
- **Slow Momentum** improves accuracy for Post-local SGD, achieving a better trade-off

Questions?