INTERNATIONAL CONFERENCE ON PARALLEL PROCESSING
ICPP/2021/CHICAGO/USA
acm In-Cooperation
sighpc
AUGUST 9–12, 2021

CD-SGD: Distributed Stochastic Gradient Descent with Compression and Delay Compensation

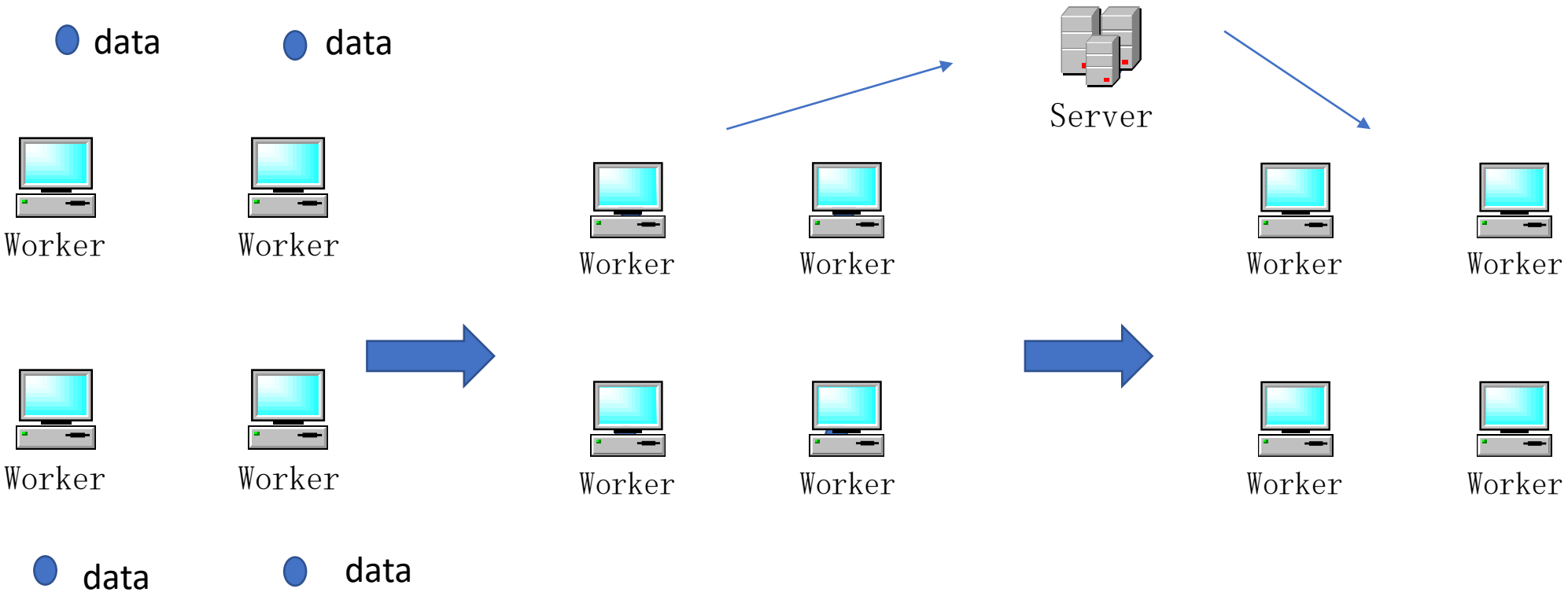
Speaker：Enda Yu

Author: Enda Yu, Dezun Dong , Yemao Xu, Shuo Ouyang,

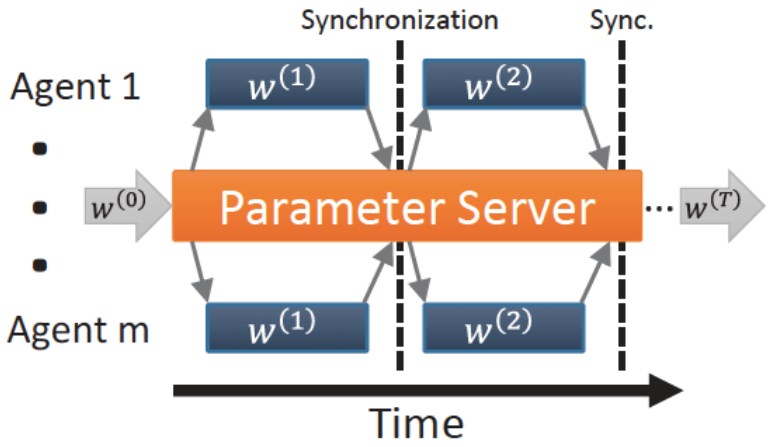
Xiangke Liao

From National University of Defense Technology

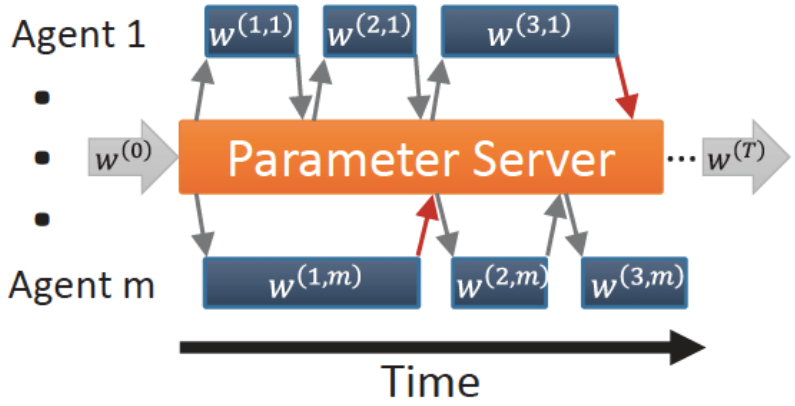
Background of Distributed Training



S-SGD and A-SGD



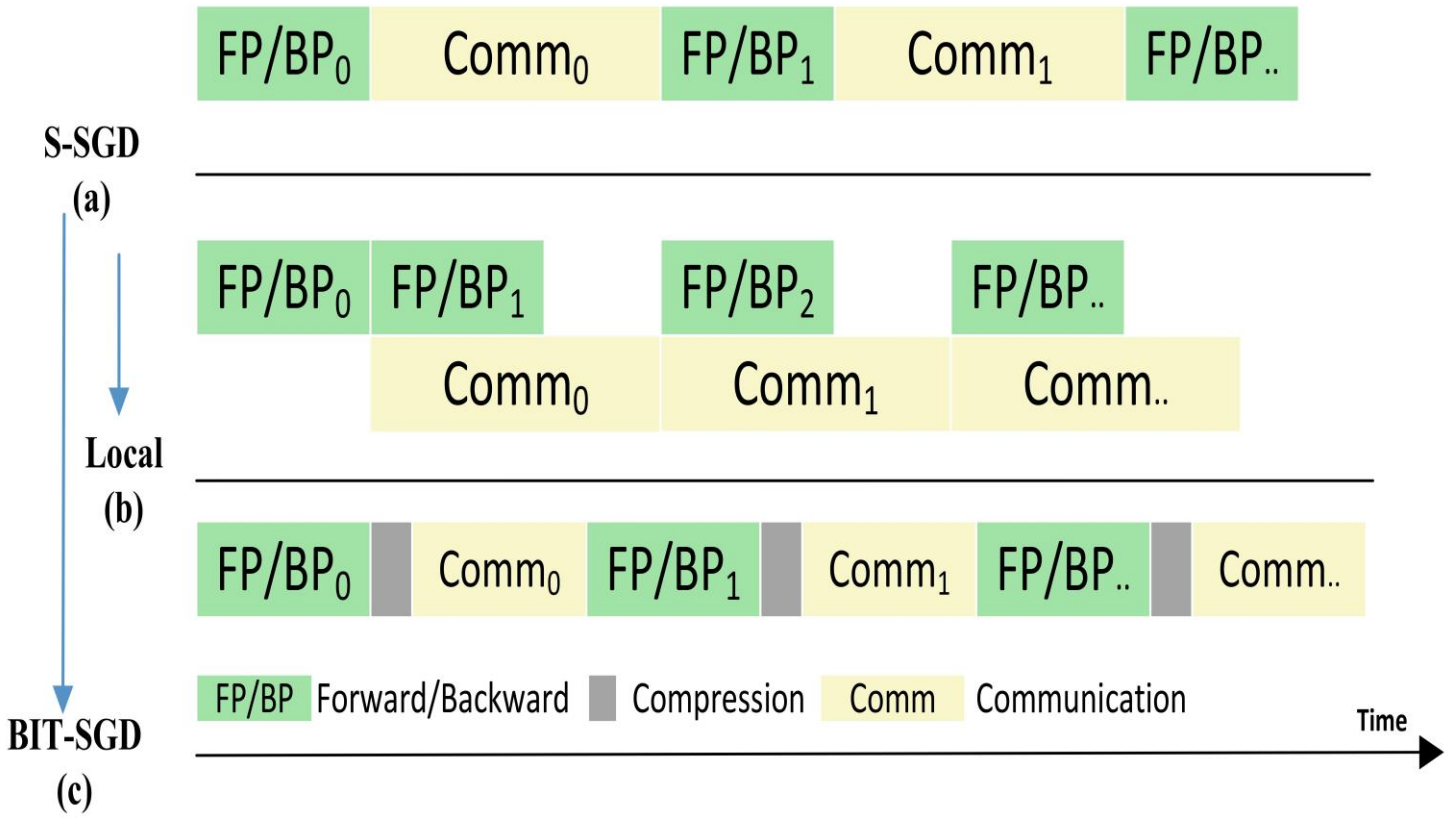
SSGD



ASGD

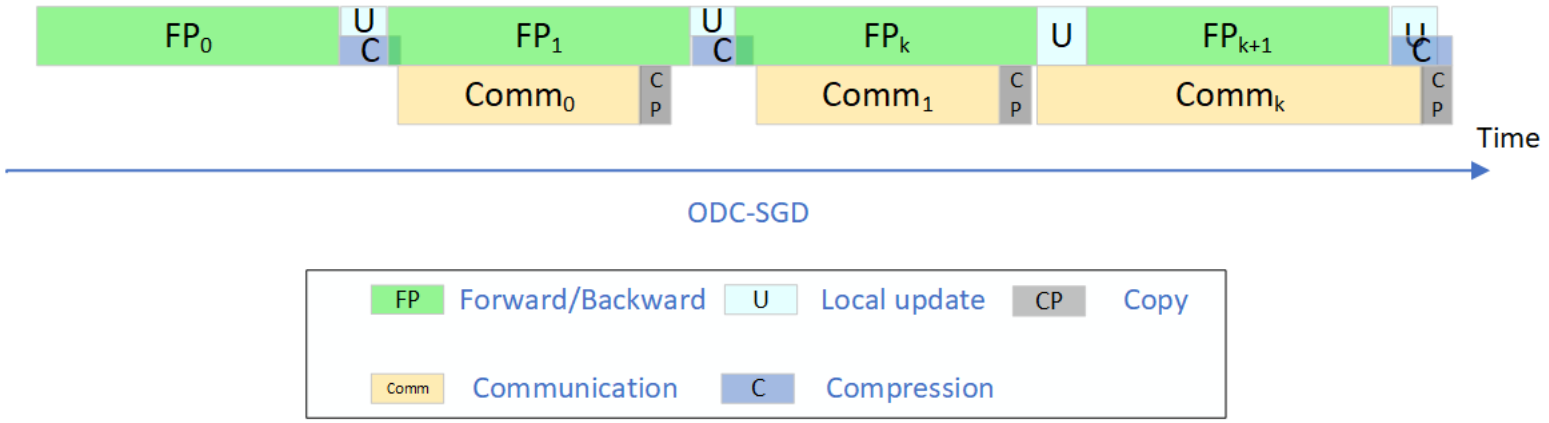
- ✓ **High accuracy&&Slow speed**
- ✓ **Low accuracy&&Fast speed**

Several typical optimization methods



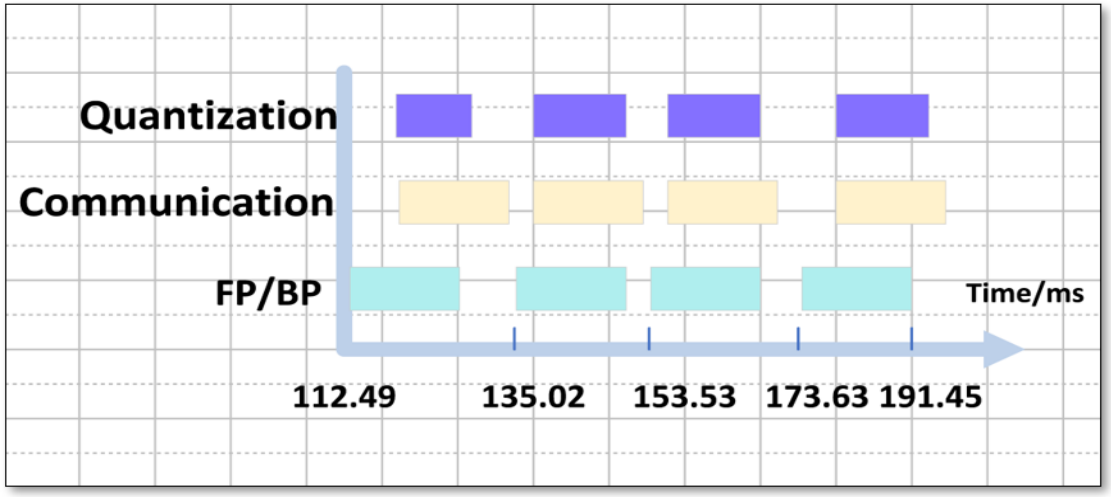
- ✓ **SSGD(a):Synchronous Stochastic Gradient Descent**
- ✓ **Local(b):A local update method.**
- ✓ **BIT-SGD(c):2-bit quantization method provided by MXNet**

ODC-SGD

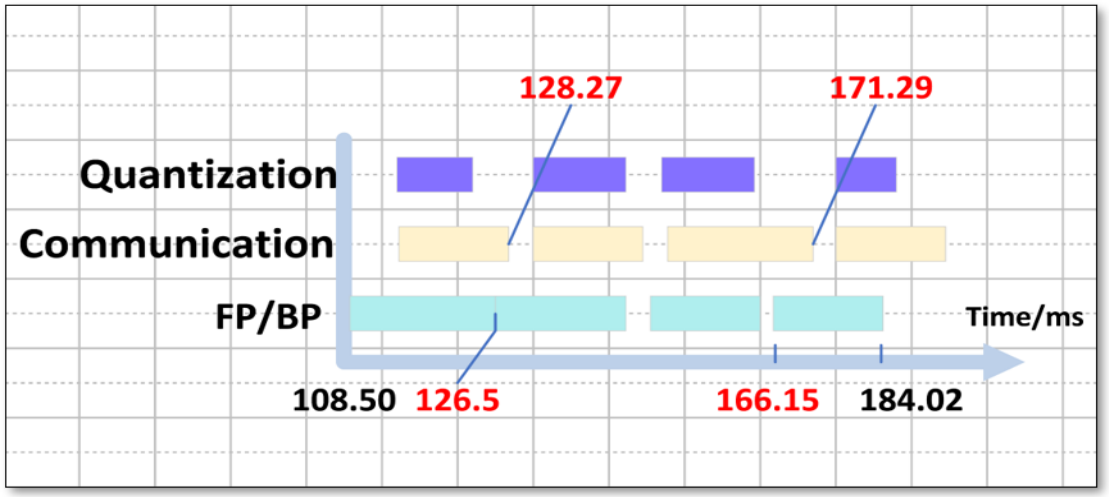


- ✓ **The local update mechanism is introduced into quantization method.**
- ✓ **K-step correction method is used to periodically repair the loss of accuracy caused by quantization.**

Compression overhead hiding



Timeline tracing of BIT-SGD

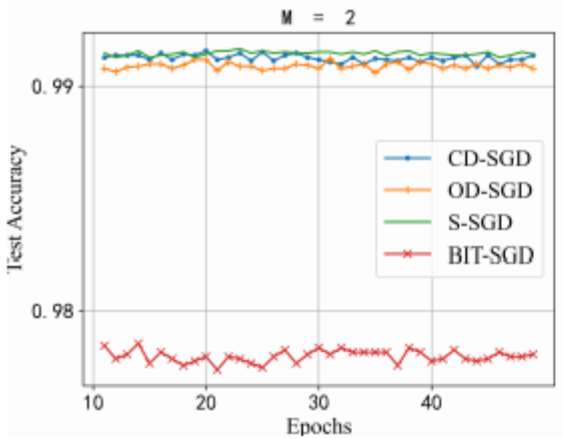


Timeline tracing of CD-SGD

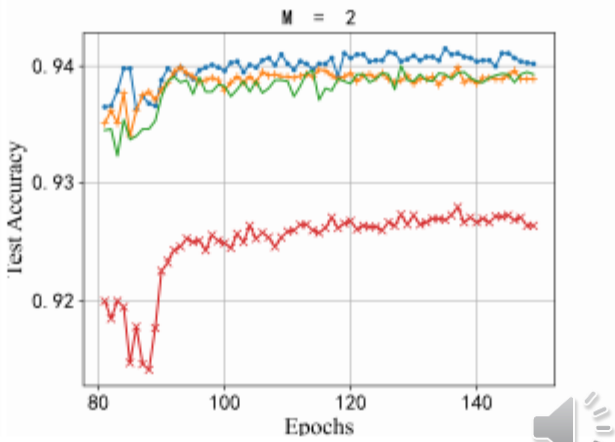
- Quantization always has an impact on the timing of the end of communication.

- The next FP/BP(calculation) in CD-SGD does not need to wait for the current communication, solving the speed impact caused by quantization.

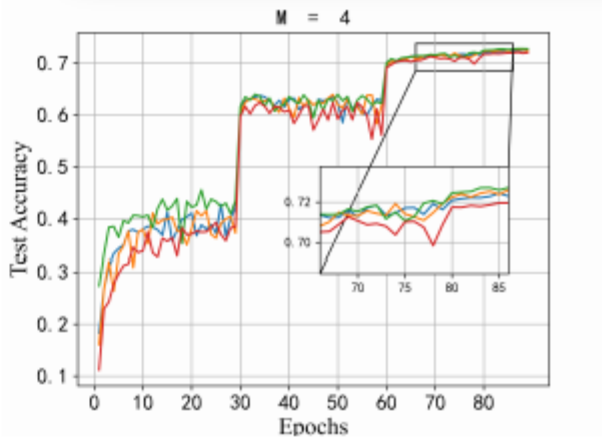
Convergence Accuracy



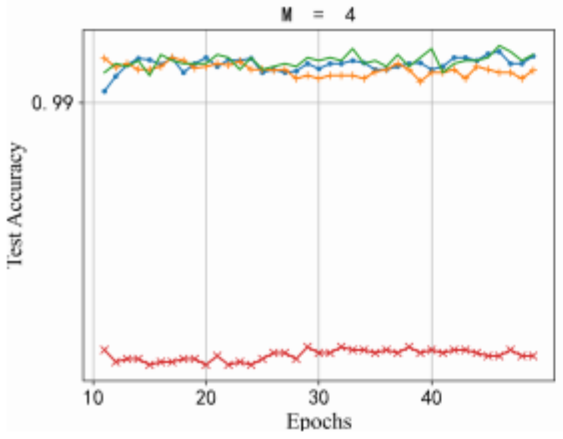
(a) Lenet-5 on MNIST (M=2)



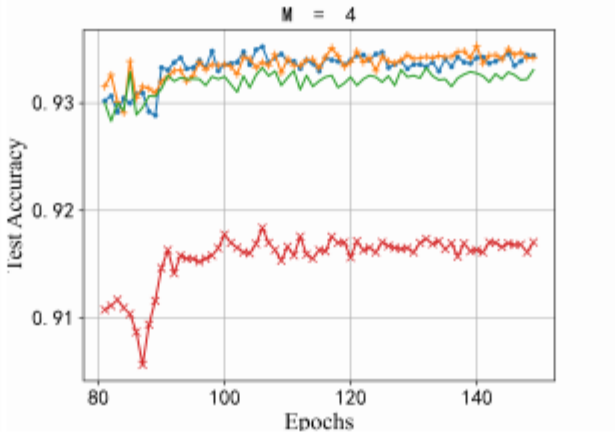
(b) Inception-bn on CIFAR-10 (M=2)



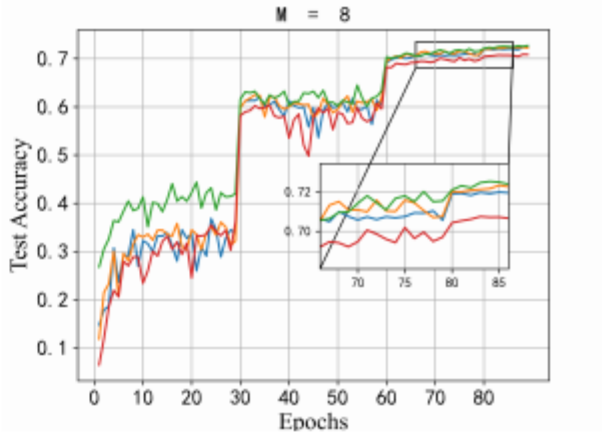
(c) ResNet-50 on ImageNet (M=4)



(d) Lenet-5 on MNIST (M=4)



(e) Inception-bn on CIFAR-10 (M=4)

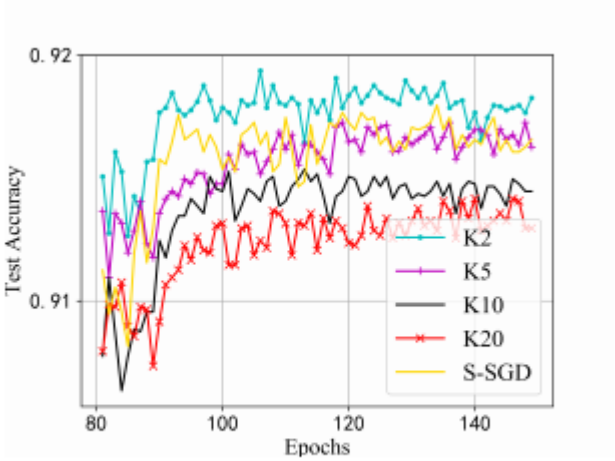


(f) ResNet-50 on ImageNet (M=8)

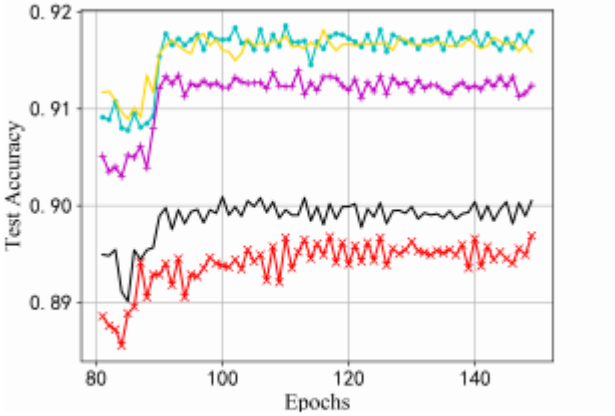
CD-SGD can help BIT-SGD achieve accuracy close to or even better than S-SGD

The upper limit of CD-SGD accuracy is affected by the quantification method used and the warm-up time

Analysis about k-step correction



(a) K-step Sensitivity Analysis (M=2)



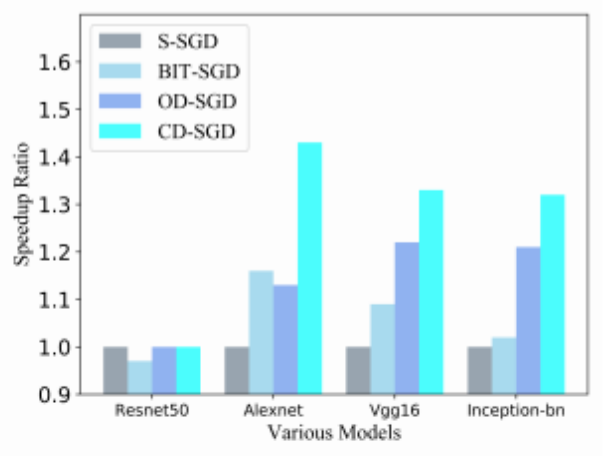
(b) K-step Sensitivity Analysis (M=4)

- Setting k equal to 5 is suitable for most situations.

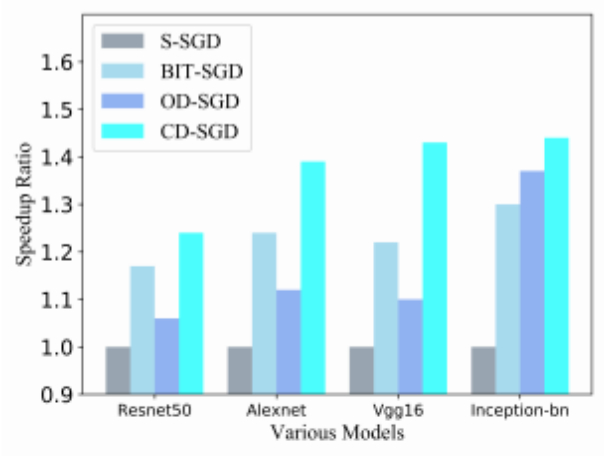
The average epoch wall-clock time and communication time of ResNet-20 on CIFAR-10 (/sec).

Time	SSGD	BIT-SGD	$k2$	$k5$	$k10$	$k20$
Te (M=4)	2.24	2.22	1.88	1.86	1.86	1.85
Te (M=2)	4.32	3.65	3.61	3.51	3.46	3.44
Tc (M=4)	2.19	2.26	2.14	2.08	2.05	2.11
Tc (M=2)	4.11	4.10	3.67	3.62	3.81	3.90

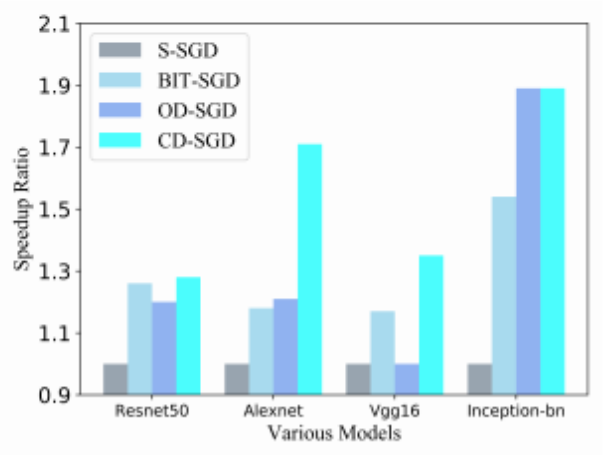
Training Speed



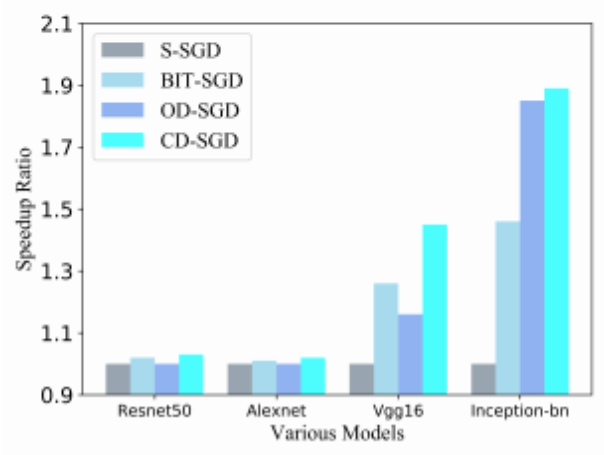
(a) batch size 32 per GPU on K80



(b) batch size 32 per GPU on V100



(c) batch size 64 per GPU on V100



(d) batch size 128 per GPU on V100

- The performance improvement range of CD-SGD to S-SGD and BIT-SGD is 0 to 89% and 3% to 45%, respectively.

Example: the average epoch time of ResNet50 training with 4 workers batch size 32 per GPU on V100 (/sec).

	S-SGD	BIT-SGD	OD-SGD	CD-SGD
Epoch-time	824.8	601.2	712.2	567.9

Conclusion

- **Distributed training communication optimization is necessary**
- **Three challenges in combining compression methods with system-level methods: Additional compression overhead, Accuracy, Sufficient benefits**
- **The local update mechanism can cover the additional computational overhead of compression.**
- **K-step correction method can solve the accuracy reduction problem of the compression method.**

CD-SGD: Distributed Stochastic Gradient Descent with Compression and Delay Compensation

THANKS