Reviewer #1:

"limited tweak relative to previous work" The contributions of CSER put it beyond merely a "tweak". Our two key contributions are: (1) A novel mechanism "error reset" that uses arbitrary compressors in a way different from error feedback. Our theoretical analysis (Theorem 1, Lemma 2, Remark 1,2) shows that error reset achieves better error bounds than error feedback. Empirical results show better convergence of error reset, especially for high compression ratios. (2) New combination of partial gradient and model synchronization–in particular, by carefully distributing the communication budgets between two synchronizations, we can further improve the convergence. As shown in Table 4 in Appendix P30, CSER combining 2 synchronizations works better than only using one of them (CSEA) in most cases.

"results not very surprising" Our work is the first to push the compression ratio to 1024 for both worker-to-server and server-to-worker communication, where CSER shows much better convergence than previous work. Particularly, at points when EF-SGD and QSparse diverge, CSER converges well. So far, the experiments with largest compression ratio were proposed by Deep Gradient Compression [Lin et al., 2018], which only considers worker-to-server communication with the compression ratio ≈ 600, and it had no theoretical analysis. Our GRBS compressor (Section 3.3) reduces bidirectional communication with the desired high compression ratio (≥ 256), which is not explored in previous work.

"exact accuracy results for the Imagenet" See the table below.

R_C/ Optimizer	1	16	32	256	1024
SGD	**76.41**$\pm$0.03				
EF-SGD		76.34$\pm$0.06	76.19$\pm$0.07	69.73$\pm$0.66	diverge
QSparse		76.40$\pm$0.05	73.89$\pm$0.09	diverge	diverge
CSER		**76.53**$\pm$0.05	**76.33**$\pm$0.06	**75.94**$\pm$0.09	**74.93**$\pm$0.11

"No accuracy loss" is conditional. For both CIFAR and ImageNet, the accuracy is slightly better than full-precision SGD when $R_C \leq 16$. With larger compression ratios, CSER has some accuracy loss compared to full-precision SGD, but performs much better than EF-SGD and QSparse. We will revise the claim by adding context.

"is it using NCCL" All the algorithms use the same communication library: Horovod with NCCL.

"How is the experimental setup chosen" Multiple nodes connected by 10Gb/s Ethernet is a typical setup used in previous works such as [Lin et al., 2018], signSGD [Bernstein et al., 2019] and EF-SGD [32]. When using single node with multiple GPUs connected by NVLink, the communication will be extremely fast and compression is less necessary. In this work, we aim to show that we can significantly reduce the heavy inter-node communication. Indeed, one may increase number of GPUs per node to do large batch training, but this becomes prohibitively expensive due to GPU cost.

"WMT/Transformer" typically uses SGD variants with adaptive learning rates such as ADAM. In this paper we focus on SGD with momentum without adaptive learning rates. Applying error reset to ADAM is future work.

"speedups relative to QSparse" CSER, EF-SGD and QSparse use exactly the same amount of communication, thus theoretically having the same training time with the same overall R_C. The advantage over QSparse is that CSER converges much better with the same low amount of communication. Figure 1(e), 2(d) show slightly shorter training time of CSER because of less memory copy in computation, which is irrelevant to communication overhead.

Reviewer #2:

"High compression ratios" are useful when network bandwidth is very low and model sizes are very big.

"other compressors" Yes. Our theoretical analysis works for arbitrary compressors.

"non-i.i.d." Yes. Our theoretical analysis already applies to non-iid case. In Assumption 2, we do not assume identical workers. In our proof (line 404 in Appendix), we only need independence to obtain V_1/n variance.

how to choose two efficient compressors Theorem 1 shows how the configurations of compressors affect convergence. With fixed overall R_C, we can enumerate possible configurations (as shown in Table 3 in Appendix P29) to get relatively smaller error bounds. To find the best configuration in practice, we do grid search.

"choose a good β" is possible, but irrelevant to communication compression. So we just use the common value 0.9.

"how to choose the learning rate" We use grid search to tune the learning rate. Details can be found in Section 5.1.

"detail advantage of the error reset" We will highlight the advantages of error reset in the revision.

Reviewer #3:

only reported on ResNet Prior work such as Deep Gradient Compression [Lin et al., 2018], EF-SGD [32] and QSparse [3], all used CIFAR-10/100 and ImageNet + ResNet in the experiments. We choose the same to directly contrast results.

"relationship between training loss and the configuration of H, RC1, and RC2" Theorem 1 shows the relationship between squared gradient norm and configurations ($R_{C_1} = 1/\delta_1, R_{C_2} = 1/\delta_2$). Though we cannot translate the convergence rate of gradient norm into the one of training loss due to non-convexity, in practice better convergence on the gradient norm implies faster convergence on the training loss. With fixed overall $R_C = 1/[1/(R_{C_1} \times H) + 1/R_{C_2}]$, we can enumerate possible configurations to get relatively smaller error bounds, but the optimal configuration is unknown. To find the best configuration in practice, we do grid search.

"influence from the number of machines" This is identical behavior as full-precision SGD, in that changing the number of machines affects the global batch sizes, thus affects the testing accuracy and requires different learning rates.

Reviewer #4:

"H=1 as a baseline" CSER with $H = 1$ is a special case called "CSEA", which is also novel. The results are in Appendix P30, Table 4 and subsequent figures. CSEA uses only **one compressor**. In Table 4, we can see that CSER with 2 compressors outperforms CSEA in most cases.

By **"state-of-the-art"** we mean the best latest work combining both local SGD and compression. We will clarify it, and add decentralized SGD and PowerSGD to the related work.