Recovering single precision accuracy from Tensor Cores while surpassing the FP32 theoretical peak performance

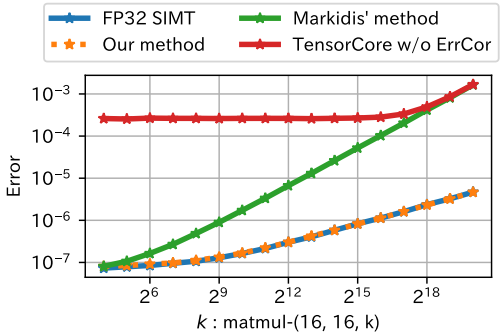
Hiroyuki Ootomo and Rio Yokota

IHPCSS 2022

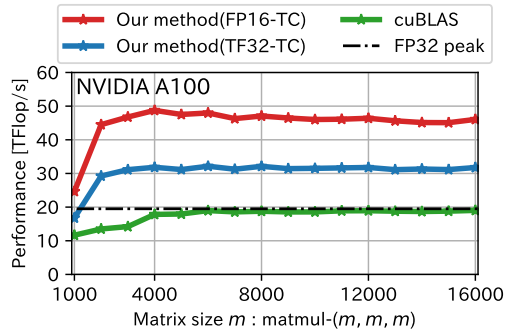
Achievements of this work

Our SGEMM emulation on Tensor Cores outperforms the theoretical peak performance of FP32 SIMT Core while achieving the same level of accuracy.

Accuracy



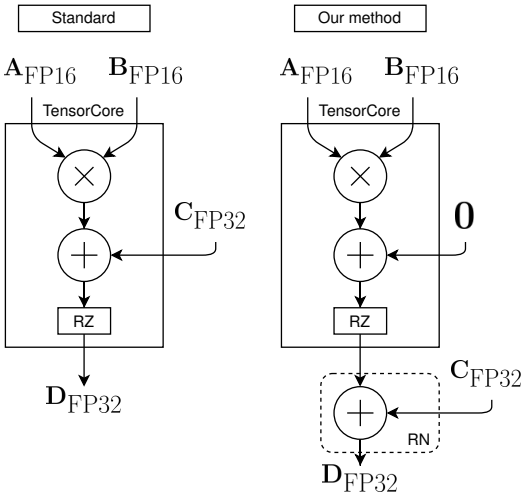
Throughput



Paper : `https://arxiv.org/abs/2203.03341`

Contribution (1/2)

- We have found that the rounding for accumulator inside Tensor Cores – RZ – causes the low accuracy of Markidis' method.
- To avoid this rounding, we use FP32 SIMT Core for the accumulation outside of Tensor Cores.

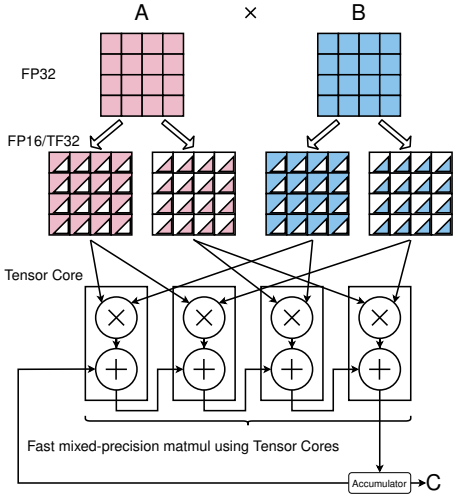


Contributions (2/2)

- Improve the accuracy of Markidis' method
  1. Calculat expectation mantissa length
  2. Found the causes the low accuracy: rounding inside Tensor Core
  3. Develop a method to avoid this rounding.
4. Reduce the underflow probability during the error correction by scaling error correction terms
5. Reduce computational complexity by omitting negligible error correction step
6. Demonstrate that our method outperforms the FP32 SIMT Core peak performance and consumes lower consumption while the the same level accuracy.

Summary: Comparison of our method and Markidis' method

Markidis' method



Our method