Emulating High-precision Matrix Operations on Low-precision Matrix Engines

Rio Yokota*[1]

[1]Institute of Science Tokyo – Japan

Abstract

Processor architecture design is now driven by the AI market, which can tolerate very low precision and does mostly dense matrix multiplications. For example, when comparing the latest generation of NVIDIA GPUs to the previous generation, the performance of Tensor Cores have increased significantly, while FP64 and FP32 performance has actually decreased. This trend is likely to continue, which makes it interesting to investigate the possibility of emulating high-precision matrix operations using multiple low-precision matrices. This talk will introduce several ways of doing this, and their pros and cons.

*Speaker