The 38th International Symposium on Lattice Field Theory

LATTICE 21

JULY 26-30 2021, ZOOM/GATHER@MIT

Contribution ID: **610** Type: **Poster**

D9: Use tensor cores to accelerate math intensive kernels in QUDA

Wednesday, 28 July 2021 15:40 (20 minutes)

We will present our recent efforts on using tensor cores, which are available on NVIDIA GPUs starting from the Volta architecture, to speed up the math intensive kernels in QUDA. A light-weighted abstraction of the CUDA PTX matrix multiply-add (MMA) instruction is added in order to efficiently stage data through the different layers of GPU memory. Specifically the efforts include:

- Use tensor cores to accelerate the 5th dimension DWF operators in the multi-splitting preconditioned conjugate gradient algorithm, utilizing the HMMA tensor core instruction;
- Use tensor cores to accelerate the dense matrix multiplications in the set up steps in multi-grid;
- Use tensor cores to accelerate the math intensive multi-BLAS kernels;
- Use double precision DMMA instruction to accelerate the contraction workflow.

Primary authors: TU, Jiqun (NVIDIA Corporation); WEINBERG, Evan; CLARK, Kate (NVIDIA); WAGNER, Mathias (NVIDIA)

Presenter: TU, Jiqun (NVIDIA Corporation)

Session Classification: Poster

Track Classification: Software development and Machines