Contribution ID: **188** Type: **Parallel**

Breaking the latency barrier: Strong scaling LQCD on GPUs

Tuesday 18 June 2019 15:40 (20 minutes)

The ability to strong scale is crucial for Lattice QCD simulations. Since the creation of the QUDA library for Lattice QCD on NVIDIA GPUs, this has always been a key development goal. Techniques like GPUDirect RDMA and NVLink allow for fast intra-node and inter-node data transfer and QUDA makes extensive use of them. However, API overheads and necessary synchronizations between GPU and CPU are increasingly limiting the ability to strong scale with MPI communication. Fine-grained GPU-centric communication provides a way out as it completely removes these bottlenecks by moving the communication to the GPU kernels. We will discuss the techniques that QUDA implements to achieve the best scaling with MPI and novel improvements using NVSHMEM for GPU-centric communication. Finally, we will show scaling results on x86 and POWER systems.

Authors: CLARK, Kate (NVIDIA); WAGNER, Mathias (NVIDIA); WEINBERG, Evan (NVIDIA Corporation)

Presenter: WAGNER, Mathias (NVIDIA)

Session Classification: Algorithms and Machines

Track Classification: Algorithms and Machines