GPU-centric Communication on NVIDIA GPU Clusters with InfiniBand

Achievement: Developed support for GPU-initiated intra- and inter-node communication on NVIDIA GPUs connected via PCIe, NVLink and InfiniBand.

Significance and Impact: This work enables applications to use OpenSHMEM programming model on extreme-scale systems with NVIDIA GPUs. Thus, enabling the applications to invoke one-sided remote communication and atomic operations on NVIDIA GPUs. While the target of the communication and atomic operations can either be host or device memory connected via PCIe, NVLink or InfiniBand.

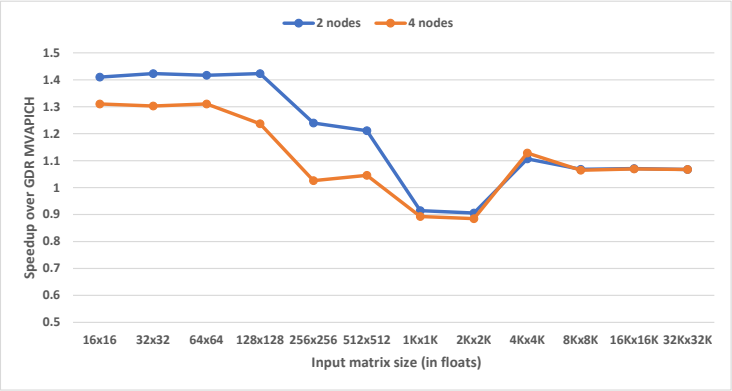
Research Details:

- Developed support for GPU-initiated intra- and inter-node communication on NVIDIA GPUs connected via PCIe, NVLink and InfiniBand.
- The capabilities are available as NVSHMEM, an implementation of OpenSHMEM for NVIDIA's GPUs
- Demonstrated NVSHMEM with support for Verbs kernels can achieve higher performance and throughput
- Demonstrated the productivity advantages of NVSHMEM with micro benchmarks and application kernels

Sponser/Facility: Work was performed with support from ORNL and DOD

PI and affiliation: Manjunath Gorentla Venkata from CSMD – Oak Ridge National Laboratory

Team: Manjunath Gorentla Venkata (ORNL), Neena Imam (ORNL), Sreeram Potluri (NVIDIA), Anshuman Goswami (NVIDIA), Davide Rosetti (NVDIA), CJ Newburn (NVIDIA)



Performance comparison of NVSHMEM-based 2D Stencil versus MPI-based (MVAPICH with GPUDirect) 2D Stencil

Publication: Sreeram Potluri (NVIDIA Corporation), Anshuman Goswami (NVIDIA Corporation), Davide Rossetti (NVIDIA Corporation), Manjunath Gorentla Venkata, Neena Inam (Oak Ridge National Laboratory), Chris J. Newburn (NVIDIA Corporation). GPU-centric Communication on NVIDIA GPU Clusters with InfiniBand: A Case Study with OpenSHMEM. HiPC 2017, Jaipur, India, December 2017.

Overview:

GPU-based extreme-scale systems are popular for scientific and data-intensive computing. As we move towards pre-exascale systems with multiple GPUs per node and tens of hundreds of GPUs in a system, efficiently moving data between the GPUs in the node and between the GPUs across node is critical to achieve performance. It is important the programming model supports such capabilities for both productivity and performance. NVSHMEM is an implementation of OpenSHMEM for NVIDIA GPUs which enables communication from the CUDA kernels over PCIe and NVLink. In this work, we focus on developing support for communication over Mellanox's InfiniBand in NVSHMEM suitable for Summit. We demonstrated the usage and productivity advantages of NVSHMEM with micro benchmarks and 2D stencil. To evaluate the efficiency of InfiniBand support, we showed that the network can be saturated with few streaming processors available on NVIDIA's Pascal P100 GPUs and achieve message throughput of 90 Million messages per sec on Mellanox's EDR InfiniBand HCA.