Tutorial and Live Demo

Accelerating HPC Applications with MVAPICH2-DPU

Donglai Dai, Richmond Liew, and Nick Sarkauskas

X-ScaleSolutions

http://x-scalesolutions.com

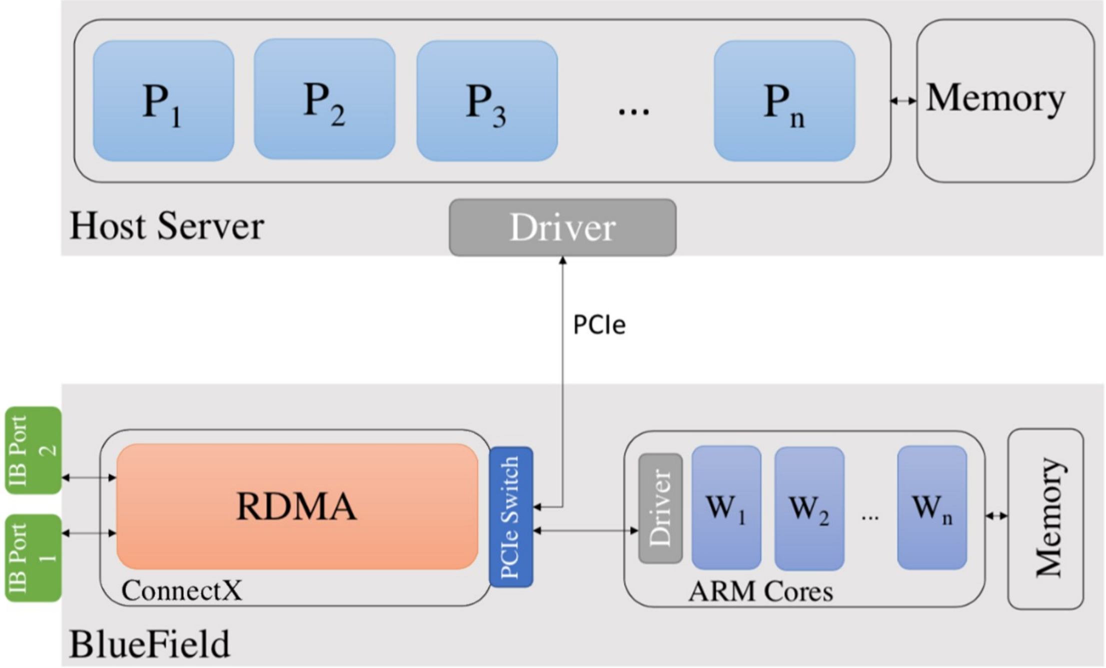
Requirements for Next-Generation MPI Libraries

- Message Passing Interface (MPI) libraries are used for HPC and AI applications
- Requirements for a high-performance and scalable MPI library:
  - Low latency communication
  - High bandwidth communication
  - Minimum contention for host CPU resources to progress non-blocking collectives
  - High overlap of computation with communication
- CPU based non-blocking communication progress can lead to sub-par performance as the main application has less CPU resources for useful application-level computation

Network offload mechanisms are gaining attraction as they have the potential to completely offload the communication of MPI primitives into the network

Overview of BlueField-2 DPU

- ConnectX-6 network adapter with 200Gbps InfiniBand
- System-on-chip containing eight 64-bit ARMv8 A72 cores with 2.75 GHz each
- 16 GB of memory for the ARM cores



How to Re-design an MPI library to take advantage of DPUs and accelerate scientific applications?

MVAPICH2-DPU Library 2021.08 Release

- Based on MVAPICH2 2.3.6
- Released on 08/22/2021
- Supports all features available with the MVAPICH2 2.3.6 release (http://mvapich.cse.ohio-state.edu)
- Novel framework to offload non-blocking collectives to DPU
- Supports offloads of the following non-blocking collectives
  - Alltoall (MPI_Ialltoall)
  - Allgather (MPI_Iallgather)
  - Broadcast (MPI_Ibcast)

MVAPICH2-DPU Library 2021.08 Release (Cont'd)

- Significantly increases (up to 100%) overlap of computation with any mix of MPI_Ialltoall, MPI_Iallgather, or MPI_Ibcast non-blocking collectives
- Accelerates scientific applications using any mix of MPI_Ialltoall , MPI_Iallgather, or MPI_Ibcast non-blocking collectives

Available from X-ScaleSolutions, please send a note to contactus@x-scalesolutions.com to get a trial license.

Today’s Live Demo

- Being run on the HPC-AI Advisory Council cluster
  - 32 Xeon nodes connected with 32 DPUs over 200Gbps InfiniBand
  - 1,024 CPU cores (Xeons) and 256 ARM cores (DPUs)
- Configuration
  - Server HW:
    - CPU: Dual Socket Intel® Xeon® 16-core CPUs E5-2697A V4 @ 2.60 GHz
    - Adapter: Nvidia BlueField-2 DPU, 8 ARM cores 2.75 Ghz, 16GB DDR4
  - Software/Firmware:
    - OS version: CentOS 8.3
    - Driver version: 5.2-1
    - Firmware version : 24.30.1004
  - MPI:
    - MVAPICH2-DPU 2021.08
  - OSU Micro-Benchmarks (OMB) 5.7.1
  - P3DFFT application v2.3

Today's Live Demo (Cont'd)

- Four parts on performance benefits
  - OSU MPI Micro-Benchmarks (OMB 5.7.1) with Ialltoall
  - P3DFFT application (using non-blocking Alltoall)
  - OMB with Ibcast
  - OMB with Iallgather

Upcoming Support to Accelerate DL Training Using DPU

- Support for distributed CPU-based DL training using NVIDIA Bluefiled-2 DPUs
- Intelligent designs to accelerate DL training
- Up to 15% performance improvement in DL training time compared to without DPU offloading
- Support for PyTorch/Torchvision and user defined DNN models and datasets
- To be available with the next release of MVAPICH2-DPU

The design is based on a recent research paper “Accelerating CPU-based Distributed DNN Training on Modern HPC Clusters using BlueField-2 DPUs” by A. Jain, N. Alnaasan, A. Shafi, H. Subramoni, D. Panda, 28th IEEE Hot Interconnects, Aug 2021

Future Releases and Engagement Plan

- Offloading designs for other non-blocking collectives
  - All-reduce, Reduce, etc.
- Offloading designs for other MPI functions
- Application-level and scalability studies
- Co-designing MPI and AI applications with DPU support

X-ScaleSolutions will be happy to get engaged, please send a note to contactus@x-scalesolutions.com.

Thank You!

contactus@x-scalesolutions.com

X-ScaleSolutions

http://x-scalesolutions.com/