Enhancing MPI Communication using Accelerated Verbs: The MVAPICH Approach

Talk at UCX BoF (SC ‘18)

by

Dhabaleswar K. (DK) Panda
The Ohio State University
E-mail: panda@cse.ohio-state.edu
http://www.cse.ohio-state.edu/~panda
Introduction, Motivation, and Challenge

- HPC applications require high-performance, low overhead data paths that provide
  - Low latency
  - High bandwidth
  - High message rate

- Hardware Offloaded Tag Matching

- Different families of accelerated verbs available
  - Burst family
    - Accumulates packets to be sent into bursts of single SGE packets
  - Poll family
    - Optimizes send completion counts
    - Receive completions for which only the length is of interest
    - Completions that contain the payload in the CQE

- Can we integrate accelerated verbs into existing HPC middleware to extract peak performance and overlap?
Verbs-level Performance: Message Rate

ConnectX-5 EDR (100 Gbps), Intel Broadwell E5-2680 @ 2.4 GHz
MOFED 4.2-1, RHEL-7 3.10.0-693.17.1.el7.x86_64
Verbs-level Performance: Bandwidth

ConnectX-5 EDR (100 Gbps), Intel Broadwell E5-2680 @ 2.4 GHz
MOFED 4.2-1, RHEL-7 3.10.0-693.17.1.el7.x86_64
The MVAPICH Approach

High Performance Parallel Programming Models

- Message Passing Interface (MPI)
- PGAS (UPC, OpenSHMEM, CAF, UPC++)
- Hybrid --- MPI + X (MPI + PGAS + OpenMP/Cilk)

High Performance and Scalable Communication Runtime

Diverse APIs and Mechanisms

- Point-to-point Primitives
- Collectives Algorithms
- Job Startup
- Energy-Awareness
- Remote Memory Access
- I/O and File Systems
- Fault Tolerance
- Virtualization
- Active Messages
- Introspection & Analysis

Support for Modern Networking Technology
(InfiniBand, iWARP, RoCE, Omni-Path)

- Transport Protocols
  - RC
  - XRC
  - UD
  - DC
- Modern Interconnect Features
  - UMR
  - ODP
  - SR-IOV
  - Multi Rail
- Accelerated Verbs Family*
  - Burst
  - Poll
  - Tag Match
- Modern Switch Features
  - Multicast
  - SHARP

* Upcoming