SOFTWARE
PRODUCT BRIEF

Mellanox OFED GPUDirect RDMA

GPU-GPU Communication Acceleration Software

The rapid increase in the performance of graphics hardware, coupled with its programmability, has made graphic accelerators a compelling platform for computationally-demanding tasks in a wide variety of application domains. Because of the computational power demands of GPU computing, the GPU-to-GPU method has proven valuable in various areas of science and technology. GPU-based clusters are being used to perform compute-intensive tasks, like Finite Element Analysis, Computational Fluid Dynamics, Monte-Carlo simulations and Molecular Dynamics are just a few examples.

WORLD-LEADING SUPERCOMPUTERS LEVERAGE GPUs TO OPTIMIZE GPU-TO-GPU PERFORMANCE.

Given that GPUs provide high core count and floating point operations capabilities, high-speed InfiniBand networking is required to connect between the platforms in order to provide high throughput and the lowest latency for GPU-to-GPU communications. While GPUs have been shown to provide performance acceleration yielding price/ performance as well as power/performance benefits, several areas of GPU-based clusters could be improved in order to provide higher performance and efficiency. The main performance issue with deploying clusters consisting of multiple GPU-nodes involves the interaction between the GPUs, or the GPU-to-GPU communication model.

GPUDIRECT RDMA

While GPUs have been shown to provide performance acceleration yielding price/ performance as well as power/performance benefits, several areas of GPU-based clusters could be improved in order to provide higher performance and efficiency. The main performance issue with deploying clusters consisting of multiple GPU-nodes involved the interaction between the GPUs, or the GPU-to-GPU communication model.

A key technology advancement in GPU-GPU communications has been GPUDirect RDMA. Prior to GPUDirect RDMA, any communication between GPUs had to involve the host processor and required buffer copies of data via the system memory. GPUDirect RDMA is a technology introduced with Mellanox ConnectX-3® and Connect-IB® adapters and with NVIDIA® Kepler-class GPU's that enables a direct path for data exchange between the GPU and the Mellanox high-speed interconnect using standard features of PCI-Express®.

HIGHLIGHTS

BENEFITS

- Significantly boost message passing interface (MPI) applications with zero-copy
- Eliminate CPU bandwidth and latency bottlenecks for increased application performance
- No unnecessary system memory copies and CPU overhead
- Allow RDMA-based application to use GPU computing power, and RDMA interconnect simultaneously

KEY FEATURES

- Support for RDMA over InfiniBand and RoCE transports
- Support for peer-to-peer communications between Mellanox RDMA adapters and NVIDIA GPUs
- High-speed DMA transfers to copy data between peer devices
- Natively supported by Mellanox OFED 2.1 and later
- Supported by Open MPI, MVAPICH2 and other CUDA-aware MPI libraries

©2018 Mellanox Technologies. All rights reserved.

GPUDirect RDMA leverages PeerDirect RDMA and PeerDirect ASYNC™ capabilities of Mellanox network adapters to significantly reduce the communication latency between GPU devices of different cluster nodes. GPUDirect RDMA also completely offloads the CPU involvement, making network communication very efficient between GPUs.

The following figure shows the direct data path of the GPUDirect RDMA communication model:

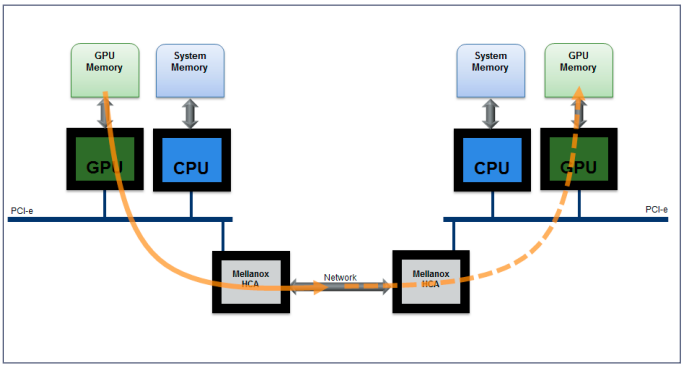

Figure 1. GPUDirect RDMA communication model

Effectiveness of GPUDirect RDMA on micro-benchmarks

Micro-benchmarks are useful in helping to understand the extent of the effectiveness of GPUDirect RDMA. Figures 2 and 3 show the benefits of using GPUDirect RDMA technology demonstrating the reduction in latency and improvements in bandwidth that is typical in achieving additional performance for GPU-to-GPU communications.

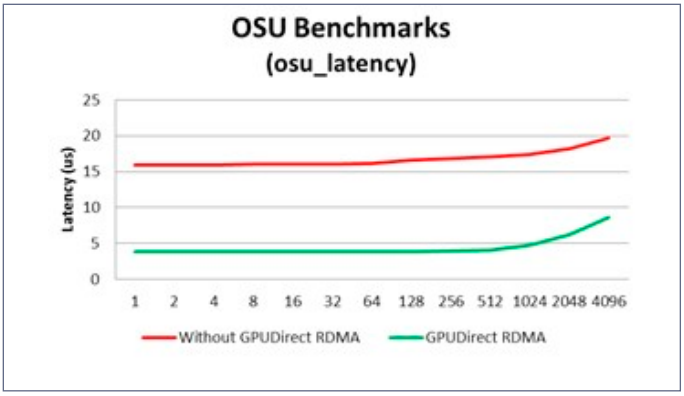

Figure 2. GPU-to-GPU internode of MPI latency is lowered by 76% on a single FDR InfiniBand link

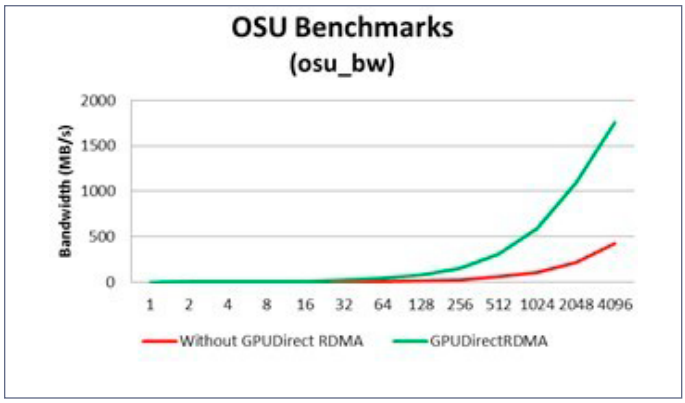

Figure 3. GPU-to-GPU internode of MPI bandwidth is improved by more than 4X on a single FDR InfiniBand link

Superior High Performance Computing using GPUDirect RDMA with real applications

HOOMD-blue (Highly Optimized Object-oriented Many-particle Dynamics -- Blue Edition) performs general purpose particle dynamics simulations. HOOMD-blue provides an optimized code

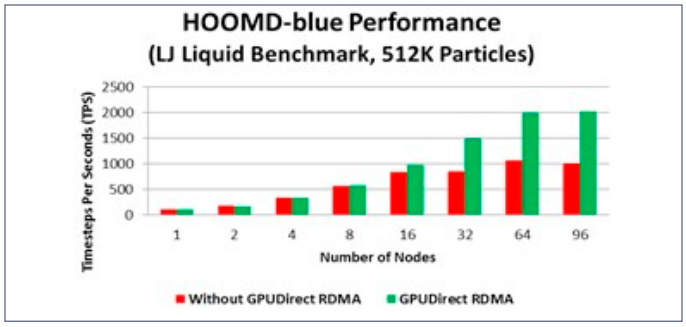

Figure 4. GPU-to-GPU internode of MPI bandwidth is improved by more than 4X on a single FDR InfiniBand link

path that utilizes standard host-based MPI, it can also take advantage of cuda-aware MPI implementations using GPUDirect RDMA over Mellanox InfiniBand.

350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com

© Copyright 2018. Mellanox Technologies. All rights reserved.
Mellanox, Mellanox logo and ConnectX are registered trademarks of Mellanox Technologies, Ltd. PeerDirect ASYNC is a trademark of Mellanox Technologies, Ltd.
All other trademarks are property of their respective owners.

60278PB
Rev 1.1