SOLUTION OVERVIEW

NVIDIA GPUDIRECT RDMA

Accelerating GPU-GPU Communication with NVIDIA GPUDirect RDMA

Whether you're exploring mountains of data, researching scientific problems, training neural networks, or modeling financial markets, you need a computing platform with the highest data throughput available. GPUs consume data much faster than CPUs and as the GPU computing horsepower increases, so does the demand for IO bandwidth.

World-leading Supercomputers Leverage NVIDIA Technology

Today, NVIDIA powers many of the fastest supercomputers in the world, fusing high performance computing (HPC) and artificial intelligence (AI) to accelerate scientific discovery. In addition, supercomputer centers around the world are adopting the NVIDIA® Ampere architecture to bring science into the Exascale Era and simulate larger models, train and deploy deeper networks, and pioneer an emerging hybrid field of AI-assisted simulations.

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. Given that GPUs provide a much higher core count and floating point operations capabilities, NVIDIA Quantum InfiniBand networking is required to connect between the nodes in order to provide high throughput and the lowest latency for GPU-to-GPU communications.

NVIDIA GPUDirect RDMA

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. Introduced with NVIDIA ConnectX® InfiniBand and Ethernet (RoCE) smart adapters, GPUDirect RDMA enables a direct path for data exchange between the GPU and the NVIDIA high-speed interconnect using standard features of PCI-Express®. It serves as an API between peer memory clients, providing NVIDIA network adapters access to peer memory data buffers. As a result, RDMA-based applications can leverage the peer device computing power via the RDMA network, without the need to copy data to host memory. This capability is supported with any NVIDIA ConnectX InfiniBand and Ethernet RoCE adapters (starting with ConnectX-4). Figure 1 shows the direct data path of the GPUDirect RDMA communication model:

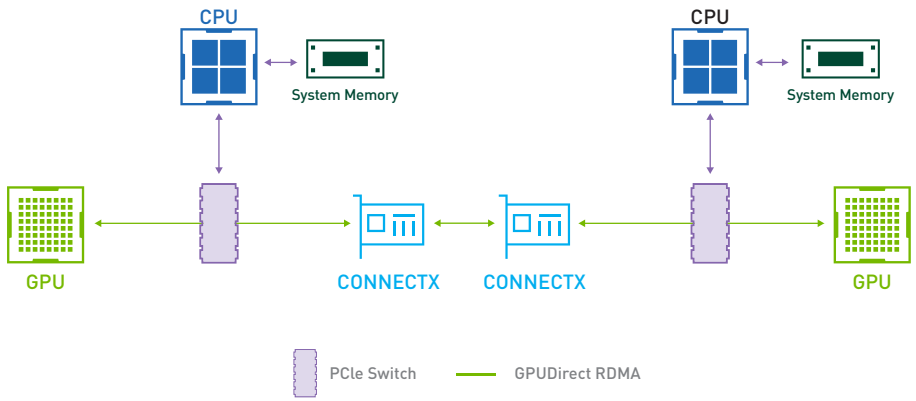

Figure 1. GPUDirect RDMA communication model.

Guarantee fast and reliable data transfers and access for NVIDIA data center GPUs, while decreasing CPU overhead and reducing latency, resulting in significant performance improvement.

- Supports NVIDIA Quantum InfiniBand and RDMA over Converged Ethernet (RoCE)
- Enables direct communications between NVIDIA RDMA adapters and NVIDIA GPUs
- Supported by NVIDIA HPC-X, Open MPI, MVAPICH2. and other CUDA-aware MPI libraries

BENEFITS

- **Enhanced Performance:** Designed specifically for low-latency and high data-throughput to provide scalable performance for GPU-based applications.
- **Decreased CPU Overhead:** Using GPUDirect, network adapters and storage can directly read and write to/from GPU memory, eliminating unnecessary memory copies, decreasing CPU overhead, and reducing latency, resulting in significant performance improvements.

Enhancements with GDRCopy—A Fast Copy Library

GDRCopy is a low-latency fast copy library that is based on NVIDIA GPUDirect RDMA technology. While GPUDirect RDMA is meant for direct access to GPU memory from the network, it is possible to use these same APIs to create perfectly valid CPU mappings of the GPU memory. The advantage of leveraging GDRCopy, is the very small overhead involved and enhanced performance.

Today's modern communication libraries, such as NVIDIA HPC-X, OpenMPI, and MVAPICH2 can easily take advantage of GPUDirect RDMA and GDRCopy to exploit the lowest latency and highest bandwidth when moving data to utilize today's unprecedented acceleration capabilities of the NVIDIA A100 GPUs.

Effectiveness of GPUDirect Technology on Micro Benchmark

Micro benchmarks are useful for understanding the effectiveness of GPUDirect RDMA. Figures 2-4 show the benefits of using GPUDirect RDMA technology with the GDRCopy library, demonstrating the reduction in latency and improvements in bandwidth that are typical when it comes to achieving additional performance for GPU-to-GPU communications.

From significantly boosting message passing interface (MPI) applications, to eliminating CPU bandwidth and latency bottlenecks, GPUDirect RDMA increases application performance.

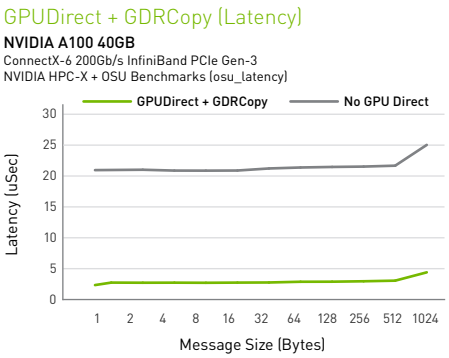

Figure 2. GPUDirect + GDRCopy consistently deliver low latency performance across all message sizes.[1]

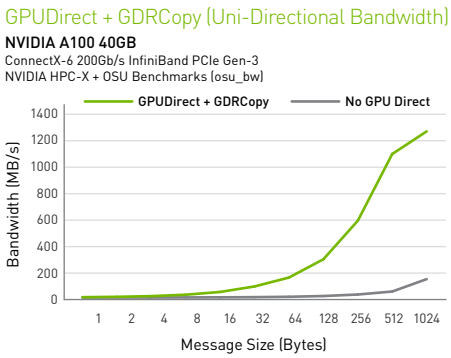

Figure 3. GPUDirect + GDRCopy delivers superior uni-directional bandwidth as message sizes increase.[1]

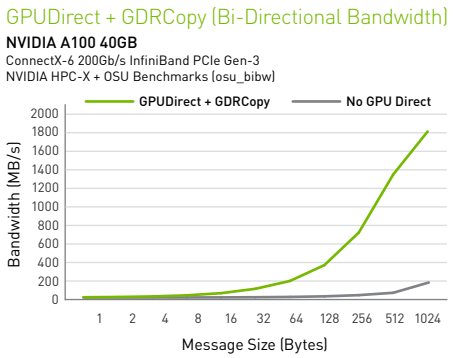

Figure 4. GPUDirect + GDRCopy delivers superior bi-directional bandwidth as message sizes increase.[1]

NVIDIA NCCL, GPUDirect RDMA, and Deep Learning Frameworks

The NVIDIA Collective Communication Library (NCCL) natively supports GPUDirect RDMA and implements multi-GPU and multi-node communication primitives optimized for NVIDIA GPUs and networking. NCCL is topology-aware and provides routines, such as all-gather, all-reduce, broadcast, reduce, and reduce-scatter, as well as point-to-point send and receive that are optimized to achieve high bandwidth and low latency over PCIe and NVLink high-speed interconnects, within a node and over NVIDIA Quantum InfiniBand across nodes.

Leading deep learning frameworks, such as TensorFlow, PyTorch, MxNet, Caffe2, and Chainer, have integrated NCCL to accelerate deep learning training on multi-GPU multi-node systems.

[1] Test Configuration: Colfax CX41060t-XK7, Dual Socket Intel(R) Xeon(R) Gold 6138 CPU @ 2GHz, NVIDIA ConnectX-6 200Gb/s InfiniBand (PCI-e Gen 3), NVIDIA A100 GPU 40GB, https://github.com/NVIDIA/gdrcopy, NVIDIA HPC-X v2.9, OpenMPI 4.1.1, OSU Micro-Benchmarks 5.8

Learn more

Learn more about GPUDirect RDMA: **developer.nvidia.com/gpudirect**