

ACCELERATING GPU-STORAGE COMMUNICATION WITH NVIDIA MAGNUM IO GPUDIRECT STORAGE

ADDRESSING THE CHALLENGES OF GPU-ACCELERATED WORKFLOWS

The datasets being used in high-performance computing (HPC), artificial intelligence, and data analytics are placing increasingly high demands on the scale-out compute and storage infrastructures of today's enterprises. This, together with the computation shift from CPUs to faster GPUs, has rendered input and output (IO) operations between storage and the GPU to take on even greater significance. In some cases, application performance suffers as GPU compute nodes need to wait for IO to complete.

With IO bottlenecks in multi-GPU systems and supercomputers, the compute-bound problem translates into an IO-bound problem. Traditional reads and writes to GPU memory use POSIX APIs to read/write data from system memory as an intermediate bounce buffer. In addition some file systems need additional memory in kernel page cache. An extra copy of data through the system memory is the leading cause of the IO bandwidth bottleneck to the GPU, as well as overall higher IO latency and CPU utilization. This is primarily due to CPU cycles being used to transfer the buffer contents to the GPU.

NVIDIA MAGNUM IO GPUDIRECT STORAGE

NVIDIA Magnum IO™ GPUDirect® Storage (GDS) was specifically designed to accelerate data transfers between GPU memory and remote or local storage in a way that avoids CPU bottlenecks. GDS creates a direct data path between local NVMe or remote storage and GPU memory. This is enabled via a direct-memory access (DMA) engine near the network adapter or storage that transfers data into or out of GPU memory—avoiding the bounce buffer in the CPU.

With GDS, third-party file systems or modified kernel driver modules available in the NVIDIA Open Fabric Enterprise Distribution (MLNX_OFED) allow such transfers. GDS enables new capabilities that provide up to 2X peak bandwidth through the GPU while improving latency and overall system utilization of both the CPU and GPU.

By exposing GPUDirect Storage within CUDA® via the cuFile API, DMA engines near the network interface card (NIC) or storage device can create a direct path between GPU memory and storage devices. The cuFile API is integrated in the CUDA Toolkit (version 11.4 and later) or delivered via a separate package containing a user-level library (libcufile) and kernel module (**nvidia-fs**) to orchestrate IO directly from DMA and remote DMA (RDMA) capable storage. The user-level library is readily integrated into the CUDA Toolkit runtime and the kernel module is installed with the NVIDIA driver. MLNX_OFED is also required and must be installed prior to GDS installation.

NVIDIA Magnum IO GPUDirect Storage accelerates the data path to the GPU by eliminating IO bottlenecks.

- > Supports RDMA over InfiniBand and Ethernet RoCE
- > Supports distributed file systems: NFS, DDN EXAScaler, WekaIO, IBM Spectrum Scale
- > Supports storage protocol via NVMe and NVMe-oF
- > Provides a compatibility mode for non-GDS ready platforms
- > Enabled on NVIDIA DGX™ Base OS
- > Supports Ubuntu and RHEL operating systems
- > Can be used with multiple libraries, APIs, and frameworks: DALI, RAPIDS cuDF, PyTorch, and MXNet

BENEFITS

- > **Higher bandwidth:** Achieves up to 2X more bandwidth available to GPU than a standard CPU-to-CPU path.
- > **Lower latency:** Avoids extra copies in the host system memory and provides dynamic routing that optimizes path, buffers, and mechanisms.
- > **CPU utilization:** Use of DMA engines near storage is less invasive to CPU load and doesn't interfere with GPU load. At larger sizes, the ratio of bandwidth to fractional CPU utilization is much higher with GPUDirect Storage.

GPUDIRECT STORAGE DATA PATH

GPUDirect Storage enables a direct DMA data path between GPU memory and local or remote storage as shown in figure 1, thus avoiding a copy to system memory through the CPU. This direct path increases system bandwidth while decreasing latency and utilization load on the CPU and GPU.

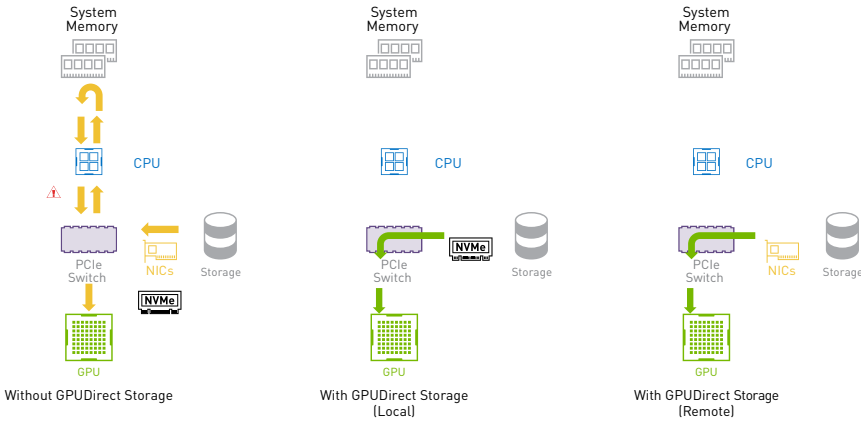


Figure 1. GPUDirect Storage data path

EFFECTIVENESS OF GPUDIRECT STORAGE ON MICROBENCHMARKS

> GDSIO Benchmark

Figure 2 shows the benefits of using GDS with the gdsio benchmarking tool that’s available as part of the installation. The figure demonstrates up to a 1.5X improvement in the bandwidth available to the GPU and up to a 2.8X improvement in CPU utilization compared to the traditional data paths via the CPU bounce buffer.

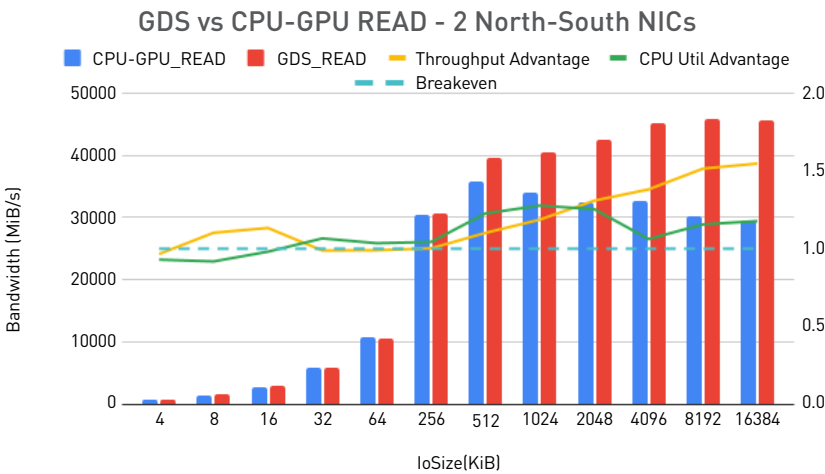
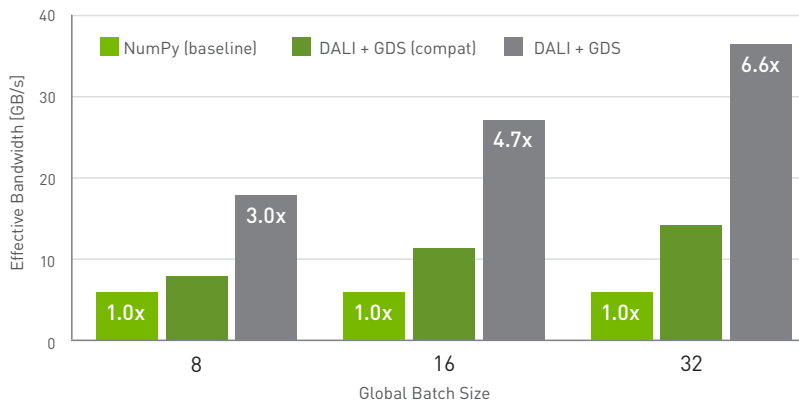


Figure 2. The benefits of using GDS with the gdsio benchmarking tool

> DeepCAM Benchmark

Figure 3 demonstrates another benefit of GDS. When optimized with GDS and the NVIDIA Data Loading Library (**DALI**®), DeepCAM, a deep learning model running segmentation on high-resolution climate simulations to identify extreme weather patterns, can achieve up to a 6.6X speedup compared to out-of-the-box NumPy, a Python library used for working with arrays.

Accelerating DeepCam Inference



Performance benchmarking done for DeepCAM Inference using standard GDS configuration in DGX A100, UB 20.04 MLNX_OFED 5.3 GDS 1.0 DALI 1.3. Application for batch size ≥ 32 limited by GPU compute throughput

Figure 3. Performance of DeepCAM with DALI 1.3.0 and GDS

EVOLUTIONARY TECHNOLOGY, REVOLUTIONARY PERFORMANCE BENEFITS

As workflows shift away from the CPU in GPU-centric systems, the data path from storage to GPUs increasingly becomes a bottleneck. NVIDIA Magnum IO GPUDirect Storage enables DMA directly to and from GPU memory. With this evolutionary technology, the performance benefits can easily be seen with a variety of benchmarks and real-world applications resulting in reduced runtimes and faster time to data-insight.

Learn more about GPUDirect Storage Acceleration at: developer.nvidia.com/gpudirect