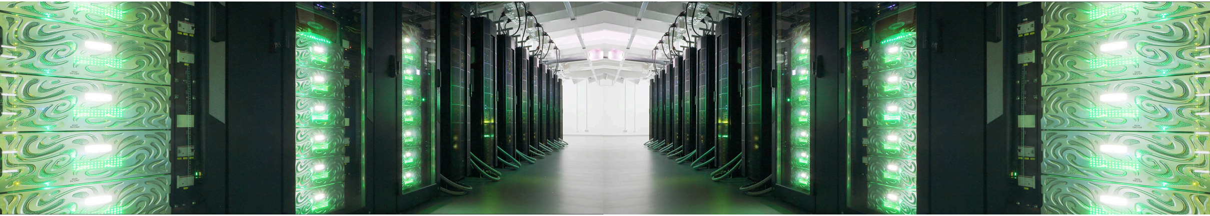
GPU Direct Storage: performance comparison

Sebastian Krey



28.09.2023

ZKI Supercomputing Herbsttagung 2023

Table of contents

What is GPUDirect Storage

Benchmark scenario

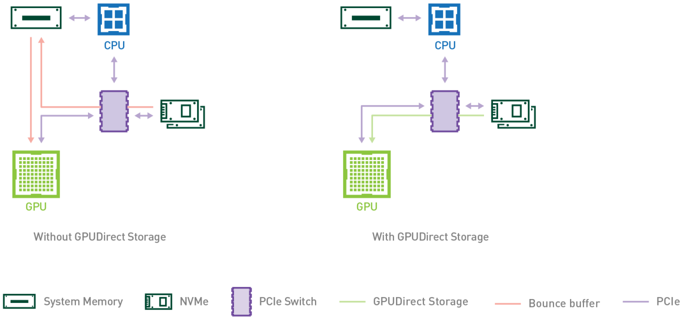
Results

Further benchmarks

Summary

What is GPUDirect Storage

Nvidia GPUDirect Storage (GDS) or Nvidia Magnum IO provides a direct DMA path between GPU and PCIe attached storage via the `cuFile` API in a Nvidia ConnectX-4+ based fabric.



Without GPUDirect Storage

With GPUDirect Storage

Source: https://developer.nvidia.com/blog/gpudirect-storage/

Examples for useable storage: local NVME drives, Lustre, BeeGFS, GPFS, WekaFS, VASTData, NetApp ONTAP, RDMA enabled NFS

Hardware: NHR System Grete

- 34 nodes
- 2 Epyc Milan 7513 32 Core CPUs per node
- 4 Nvidia A100 40GB GPUs per node
- 2 Intel P4510 1TB PCIe 3 NVME SSDs per node
- Dual rail Infiniband HDR interconnect
- Cluster local GPU Direct enabled SSD storage (2 DDN ES400NVX with ExaScaler 6.1)

Benchmarktool: elbencho

- Combined benchmarking tool for multiple purposes
- I/O and metadata benchmarking
- Metrics similar to fio available (bandwidth, iops)
- POSIX IO, CUDA, GDS via cuFile API and S3
- Client/server model for multi node testing
- Tools for parameter tuning and visualization
- Used 4k, 512k and 4M blocksizes (very nice I/O patterns)
- CPU based POSIX I/O, CUDA I/O with all GPUs and cuFile API via option (-gds)

Results 1

local NVME	CPU POSIX	CUDA	cuFile (GDS)
4k rand read IOPS	934.276	349.722	828.364
4M seq read MiB/s	5.336	5.325	5.311
4k rand write IOPS	348.002	138.581	171.072
4M seq write MiB/s	2.129		

Results 2

Lustre 1 node	CPU POSIX	CUDA	cuFile (GDS)
4k rand read IOPS	1.124.942	263.014	1.054.308
4M seq read MiB/s	37.521	22.139	38.389
4k rand write IOPS	88.652	84.524	88.794
4M seq write MiB/s	21.879	10.260	20.575

Results 3

Lustre 2 nodes	CPU POSIX	CUDA	cuFile (GDS)
4k rand read IOPS	2.190.572	497.150	2.215.756
4M seq read MiB/s	45.988	26.347	70.726
4k rand write IOPS	152.127	146.593	147.537
4M seq write MiB/s	30.287	24.153	29.728

Strange observation

Suspicous error message in `cufile.log`:

```
27-09-2023 18:05:21:334 [pid=21356 tid=21356] ERROR  cufio-fs:152 EXT4
 journal options not found in mount table for device,can’t verify
  data=ordered mode journalling
27-09-2023 18:05:21:334 [pid=21356 tid=21356] NOTICE  cufio:1538
 cuFileHandleRegister GDS not supported or disabled by config,
  using cuFile posix read/write with compat mode enabled
```

Results 4

Bypassing GDS for blocksizes <1MB on Lustre with 2 nodes:

Lustre 2 nodes	cuFile (POSIX)	cuFile (GDS)
4k rand read IOPS	1.933.847	2.215.756
512k rand read MiB/s	57.396	55.650

MLPerf Storage

MLPerf Storage is a benchmark suite to characterize the performance of storage systems that support machine learning workloads.

Available at `https://github.com/mlcommons/storage`

Based on work of the *Deep Learning I/O (DLIO) Benchmark*, available at `https://github.com/argonne-lcf/dlio_benchmark`

Aims to simulate I/O parttern of different deep learning workloads to drive a certain amount of processors (e.g. simulated Nvidia V100).

Rather low peak bandwidths, but consistent low latencies are important.

IO500

- Benchmark collection with different configurations based on ior, mdtest, mdworkbench and pfind
- MPI enabled for multinode operation to benchmark large scale systems
- POSIX IO and for some benchmarks MPIIO and/or S3 object IO.
- Results show a frame for expectable performance for a user
- Easy benchmarks are a upper limit, hard benchmarks lower limit for sensible IO
- Individual benchmarks for metadata and objectdata, streaming and randiom IO, small files and large file, memory aligned large IO sizes and unaligned small IO sizes, individual files and shared files
- Measurement of write, read and (for metadata) delete performance
- Recently updated to include GPU based storage benchmarks
- Traditional focus on peak bandwidth and IOPS

Summary

- I/O based on `cuFile API` provides large performance gains for reading compared to traditional CUDA I/O
- GPUDirect storage limited to systems with Nvidia interconnect and installed MOFED
- Several GDS enabled storage plattforms available, even DIY based on RDMA enabled NFS possible
- Additional performance gains of GDS vs cuFile API in POSIX compatibility mode seems small
- Additional benchmarks with other tools are needed for full overview
- MLPerf Storage interesting approach to generate workloads that are more realistic for actual user workloads
- Nothing prevents a storage system from choking on stupid I/O decisions of its users