Cornelis Networks™ Omni-Path™ High-Performance Networking for NVIDIA® Ampere GPUs

White Paper

Cornelis Networks^{TM} Omni-Path 100 products support NVIDIA GPUDirect® technology and perform well on state-of-the-art datacenter GPUs. Data analytics, scientific computing, artificial intelligence, and financial trading markets can reap the benefits of multi-GPU scale-out environments when using Cornelis Omni-Path fabrics. In this article, the network latency and bandwidth of Cornelis Omni-Path is shown when connecting two NVIDIA A40 GPUs through an Omni-Path switch.

High performance fabrics must have low latency and high bandwidth to keep up with increasing demands of the latest datacenter CPUs and GPUs. A wheels-off-the-ground analysis of how fast the network can perform when transferring GPU buffers with GPUDirect RDMA technology is desired to ensure that the communication performance is not a bottleneck when scaling out clusters in a datacenter. These benchmarks show the theoretical limit of the platform and communication fabric, though not all parallel frameworks and applications will fully utilize the capability. In this article, we are not stressing the compute ability of the GPU but rather the ability to communicate efficiently over Cornelis Omni-Path. Subsequent papers will expand testing to cover other important communication frameworks in the GPU-enabled high-performance computing (HPC) and artificial intelligence (AI) community.

Cornelis Omni-Path currently supports NVIDIA GPUDirect technology, which includes running NCCL with Performance Scaled Messaging (PSM2). PSM2 has been optimized and tuned for low latency and high bandwidth applications, including when using NVIDIA GPUs.[1] The next generation 400Gb Cornelis Networks CN5000 fabric will have full support for the RDMA buffer transfer technologies of AMD, Intel, and NVIDIA GPUs with the Omni-Path Express^{TM} (OPX) libfabric provider. An upcoming release of OPX will also extend support for RDMA GPU buffer transfers on Omni-Path 100 to Intel® Data Center GPU Max Series and AMD Instinct™ MI200-MI300 Series accelerators.

MPI Latency

Ohio State University Micro-Benchmarks v7.3[2] are used to measure the point-to-point latency between two NVIDIA A40 GPUs using osu_latency with the device to device (D D) option, shown in Figure 1. Using standard CUDA cuMemcpy functions in PSM2, the latency is measured to be roughly 27 microseconds – this is reported at the MPI (user) level starting in one GPU, through the host fabric interface (HFI) adapter, through an Omni-Path switch, and to the receiving GPU on the other node. Enabling GPUDirect RDMA with Cornelis Networks' implementation of GDRCopy improves the latency down to only 3.8 microseconds, an 80% improvement. The small-message latency of Cornelis Omni-Path with GPUDirect technology is as good or better than HDR published numbers.[3]

[1] Cornelis™ PSM2 Programmer's Guide, Doc No. H76473 R18.0, accessed via www.cornelisnetworks.com customer portal.

[2] https://mvapich.cse.ohio-state.edu/benchmarks

[3] GPUDirect example over 200Gb/s HDR InfiniBand: https://hpcadvisorycouncil.atlassian.net/wiki/spaces/HPCWORKS/pages/2791440385/GPUDirect+Benchmarking#Appendix:-GPUDirect-example-over-200Gb/s-HDR-InfiniBand, accessed December 7, 2023.

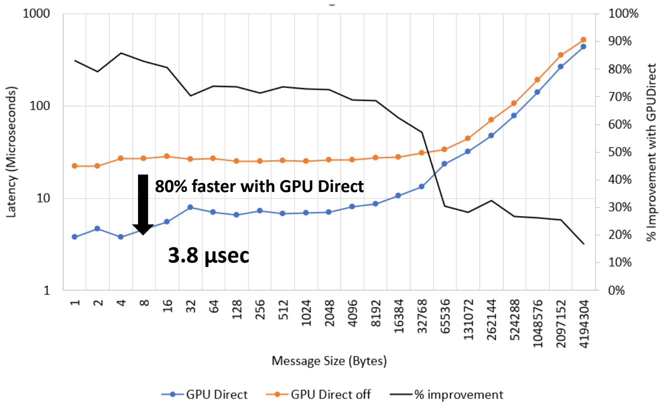

Figure 1. GPU buffer latency with Cornelis Omni-Path.

While latency is typically reported for smaller message sizes, the latency benefits of GPUDirect are seen even for larger message sizes up to the benchmark default maximum of 4MB, where a 17% improvement is seen with GPUDirect.

MPI Bandwidth

The point-to-point bandwidth is shown in Figure 2 using the osu_bw benchmark with the D D option. GPUDirect is shown to improve streaming bandwidth between GPUs for all message sizes up to 32KB, with up to a 4x throughput/ message rate improvement at 4 bytes. Since the non-GPUDirect performance is able to achieve line rate, the percent improvement with GPUDirect diminishes at larger message sizes. Ninety percent of line rate performance is reached at 256KB message sizes.

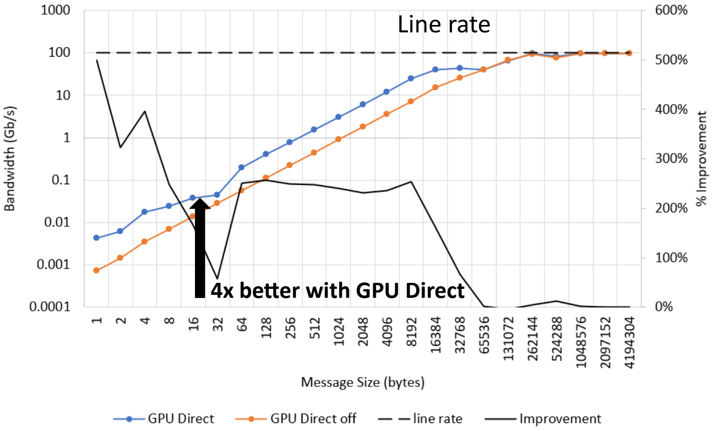

Figure 2. GPU buffer bandwidth with Cornelis Omni-Path.

Cornelis Omni-Path supports multiple 100 PCIe HFIs per server. Typically, customers may choose to install one HFI per GPU and each GPU process communicates through a dedicated HFI on the fabric. Cornelis Networks recommends testing your applications to be sure about your network bandwidth requirements before concluding you need more than 100Gb. A single HFI often provides sufficient bandwidth and has extreme performance/price advantages over the competition.[4&5]

Conclusions

This paper shows the fundamental capability of Cornelis Omni-Path to transfer NVIDIA GPU buffers with GPUDirect. Over many years, the performance of Omni-Path is shown to match or exceed the

[4] OpenFOAM® Performance and Scalability with Cornelis Networks™ OPX on 3rd Generation Intel® Xeon® Processors: https://www.cornelisnetworks.com/openfoam-performance-and-scalability-with-cornelis-networks-opx-on-3rd-generation-intel-xeon-processors/

[5] Scaling Higher with OpenRadioss™ and Cornelis Networks™ Omni-Path Express™: https://www.cornelisnetworks.com/scaling-higher-with-openradioss-and-cornelis-networks-omni-path-express/

competition for many HPC workloads on Intel and AMD-powered servers, and the same technologies are being applied to GPU enabled applications in HPC and AI. While the latency of Omni-Path 100 Cornelis Omni-Path is proven to be as good or better than the competition, ever-evolving parallel frameworks will continue to be tuned and optimized to take advantage of increased fabric bandwidth.

Cornelis Networks' next generation CN5000 fabrics will deliver 400Gb/s of network bandwidth per adapter, as well as maintaining the low latencies and high message rates that are seen today. For now, the Omni-Path 100 HFIs deliver 100Gb/s line-rate performance to and from GPU memory, and multiple Omni-Path 100 HFIs can be used to increase total bandwidth to the server, if required. The future of Cornelis Networks is bright including support for not only NVIDIA GPUs, but Intel Data Center GPU Max Series as well as AMD Instinct series GPUs. Contact your Cornelis sales representative today and get started with Omni-Path!

Configurations & Recipe

Tests performed on 2 socket AMD EPYC 7252 8-Core Processor. CcxAsNumaDomain: Enabled, ProcTurboMode: Enabled, ProcCStates: Enabled. Red Hat Enterprise Linux 8.6 (Ootpa). 4.18.0-372.9.1.el8.x86_64 kernel. 16x16GB, 128 GB total, 3200 MT/s. GPU config: slot:e2:00.0 3D controller: NVIDIA Corporation GA102GL [A40] (rev a1) NUMA node: 1. Driver Version: 545.23.08 CUDA Version: 12.3. Cornelis Omni-Path Express Fabric Suite 10.13.0.0.10 installed with -G, PSM2 as provided by openmpi-4.1.4-cuda-hfi. OSU Microbenchmarks v7.3.

Example run command: mpirun -np 2 --map-by ppr:1:node -host host1,host2 -mca btl self,vader -mca mtl psm2 -x PSM2_CUDA=1 -x PSM2_GPUDIRECT=1 ./osu_latency D D. Set PSM2_GPUDIRECT=0 to measure without GPUDirect.

Legal Disclaimer

You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Cornelis Networks products described herein. You agree to grant Cornelis Networks a non-exclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

All product plans and roadmaps are subject to change without notice.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Cornelis Networks technologies may require enabled hardware, software, or service activation.

No product or component can be absolutely secure.

Your costs and results may vary.

Cornelis, Cornelis Networks, Omni-Path, Omni-Path Express, and the Cornelis Networks logo belong to Cornelis Networks, Inc. Other names and brands may be claimed as the property of others.