Scaling Higher with OpenRadioss™ and Cornelis Networks™ Omni-Path Express™

Article

The open-source software model has proven that software developed and tested by the community delivers more reliable, functional, and performant solutions to users. In a pioneering strategy to accelerate innovation around the industry-proven Altair® Radioss® finite element analysis solver, Altair has released an open-source version named OpenRadioss.[1] Cornelis Networks has likewise embraced the open-source community with the release of Cornelis Omni-Path Express™ (OPX), an enhanced version of the high-performance Cornelis Omni-Path interconnect based on a new, open-source software stack.

The first bake-off between Cornelis OPX and NVIDIA InfiniBand HDR showed that with Cornelis OPX, users will experience better OpenRadioss performance per fabric cost than NVIDIA InfiniBand HDR. The comparisons used up to 8 AMD EPYC™ 7713 dual-socket nodes, for a total of 1024 cores.[2]

In this paper, performance is updated using the latest software recipes for both Cornelis OPX and NVIDIA HDR InfiniBand, increasing the scale further to 16 AMD EPYC 7713 dual-socket nodes.

OpenRadioss software simulates how materials interact and deform based on outside influences, such as a car crash, bridge deformations under a heavy load, or even a cell phone dropping on a kitchen floor. These simulations model how millions of elements react to external forces through each millisecond of an event. For small workloads that can be performed in a single compute node, CPU and memory bandwidth are key to performance. However, as the simulation size grows beyond the capability of a single node, the fabric becomes a critical consideration.

Cornelis OPX is designed specifically for high-performance, parallel computing environments. It is built utilizing a unique link-layer architecture and a highly optimized OFI libfabrics provider[3] that delivers higher message rates and lower latencies than competing interconnects with a leadership price/performance value proposition.

In this paper, the 10-million cell taurus model (TAURUS_A05_FFB50[4]) shown in Figure 1 is used to demonstrate how the network fabric affects application run time. The time of the physical simulation was increased from 2ms in the original study[2] to 10ms in this paper, which is shortened from 120ms in the full model. Altair recommends 10ms for fast performance and scalability testing.[1]

Figure 1. Taurus 10M cell model.

Figure 2 compares the performance of the benchmark using up to 16 AMD EPYC 7713 dual-socket nodes, for a

[1] Industry-Proven Altair Radioss Finite Element Analysis Solver Now Available as Open-Source Solution: https://www.altair.com/newsroom/news-releases/industry-proven-altair-radioss-finite-element-analysis-solver-now-available-as-open-source-solution. www.openradioss.org

[2] https://www.cornelisnetworks.com/wp-content/uploads/Delivering-Leadership-Performance-with-OpenRadioss%E2%84%A2-and-Cornelis%E2%84%A2-Omni-Path-Express%E2%84%A2-1.pdf

[3] https://ofiwg.github.io/libfabric/

[4] HPC Benchmark Models, https://openradioss.atlassian.net/wiki/spaces/OPENRADIOSS/pages/47546369/HPC+Benchmark+Models

total of 2048 cores, connected with a 100Gbps Cornelis OPX fabric and the same nodes connected with a 200Gbps NVIDIA InfiniBand HDR fabric. OpenMPI is used and OpenRadioss was compiled with gcc10.2 using the default build flags. The performance shown is job throughput (the number of cases able to run in one day if they were executed back-to-back without down time). Each data point is the result of five runs, eliminating the minimum and maximum performance and averaging the middle three runs. Every data point has a relative standard deviation of less than 1%. Since the simulation time was shortened by a factor of 12 from the original model, the job throughput is reduced by the same factor of 12 to represent the original model. The upgrade of UCX from version 1.14.0 to 1.15.0 has improved the HDR result at 8 nodes to be on-par with Cornelis OPX but scaling further to 16 nodes reveals a 5.9% advantage with Cornelis OPX. The best performance for both fabrics is achieved at 16-nodes using 32 MPI ranks per node with 4 OpenMP threads per rank, consistent with the previous study, leveraging the hybrid parallelization mode offered with OpenRadioss.

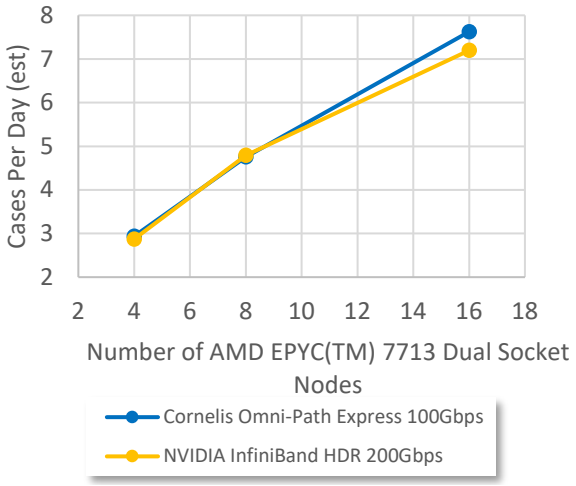

Figure 3. Performance of the OpenRadioss Taurus T10m benchmark.

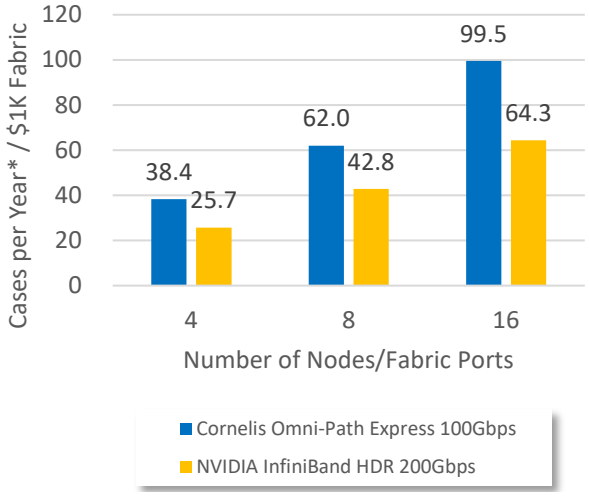

Figure 2. Performance of the OpenRadioss Taurus T10m benchmark, per fabric cost.

In addition to performance, another important consideration in fabric selection is price. For this second comparison, MSRP pricing was used[5] to build an 8-node cluster consisting of a single edge switch, 8 cables, and 8 host adapters. Performance is shown in terms of job throughput per year normalized by the cost of the 8-node fabric. As seen in Figure 3, the Cornelis OPX cluster at 16 nodes delivers up to 54% better job throughput per fabric cost compared to the NVIDIA InfiniBand HDR cluster. This means users can obtain peak OpenRadioss performance with a lower budget, or they can deploy more nodes with the same budget to shorten the time to results.

In conclusion, the OpenRadioss open-source software combined with Cornelis Networks Omni-Path Express interconnect fabric delivers leadership performance and more than a 50% higher return on investment. Cornelis Networks Omni-Path (100-series) hardware is available now, contact sales@cornelisnetworks.com to get started!

[5] MSRP Pricing obtained on 7/11/2023 from https://store.nvidia.com/en-us/networking/store. Mellanox MCX653105A-HDAT $1628 per adapter. Mellanox MQM8700-HS2F managed HDR switch, $25555. MCP1650-H002E26 2M copper cable - $281. Cornelis Omni-Path Express MSRP pricing as of 7/11/2023. Cornelis 100HFA016LSN 100Gb HFI $880 per adapter. Cornelis Omni-Path Edge Switch 100 Series 48 port Managed switch 100SWE48QF2 - $19750. Cornelis Networks Omni-Path QSFP 2M copper cable 100CQQF3020 - $147. Exact pricing may vary depending on vendor and relative performance per cost is subject to change.

System Configuration

Tests performed on 2 socket AMD EPYC 7713 64-Core Processors. Rocky Linux 8.4 (Green Obsidian). 4.18.0-305.19.1.el8_4.x86_64 kernel. 32x16GB, 256 GB total, 3200 MT/s. BIOS: Logical processor: Disabled. Virtualization Technology: disabled. NUMA nodes per socket: 4. CCXAsNumaDomain: Enabled. ProcTurboMode: Enabled. ProcPwrPerf: Max Perf. ProcCStates: Disabled.

OpenRadioss-latest-20230209 compiled with gcc 10.2. Example run command: `mpirun -np 256 --map-by numa:PE=4 -x OMP_PLACES=cores --bind-to core -x OMP_NUM_THREADS=4 -mca btl self,vader -x OMP_STACKSIZE=400m -hostfile 16nodes ./engine_linux64_gf_ompi -i TAURUS_A05_FFB50_0001.rad -nt 4`.

Cornelis Omni-Path: Open MPI 4.1.4 compiled with gcc 10.2. libfabric 1.18.0 compiled with gcc 10.2 Additional run flags: `-x FI_PROVIDER=opx -mca mtl ofi -x FI_OPX_HFI_SELECT=0`

NVIDIA HDR: OpenMPI 4.1.5a1 as provided by hpcx-v2.15-gcc-MLNX_OFED_LINUX-5-redhat8-cuda12-gdrcopy2-nccl2.17-x86_64, UCX version 1.15.0. Additional run flags: `-x UCX_NET_DEVICES=mlx5_0:1 -mca coll_hcoll_enable 0` (performance was lower with hcoll enabled).