# Panasas®, Penguin Solutions™, and Cornelis Networks™ Team Up to Transform HPC

***Solution Brief***

# *Introduction*

Today's high-performance computing (HPC) continues to transform from the modeling and simulation of traditional research labs to an ever-evolving convergence of workloads and technologies. HPC workloads are more complex and mixed with applications integrating simulations and high-performance data analytics. HPC infrastructure is also expanding with compute cluster fabrics like Omni-Path Express™ (OPX) delivering higher performance and lower latency to help brace the foundation for application performance and scalability.

To that end, Penguin Solutions™, Panasas®, and Cornelis Networks™ have come together to define and test a reference design of Penguin Solutions Altus® servers and Panasas ActiveStor® storage appliances connected with the Cornelis OPX fabric. Industry standard HPC application benchmarks spanning climate/weather modeling, computational fluid dynamics for manufacturing, and molecular dynamics simulations in life sciences were conducted and documented.

Tested for three different verticals, the results prove an HPC solution ideal not only for traditional HPC workloads, but also able to meet the dynamics and challenges of today's modern HPC. This solution meets the challenges by being scalable from small to large computing cluster configurations up to multi-petabytes of high-performance storage, and by providing flexible bandwidth and connectivity as needed. Designing HPC solutions can be difficult, and we hope this turnkey solution makes that task a little less challenging.

Data centers with existing file system storage using Ethernet in the back end do not have to invest in new storage when deploying OPX networking. Cornelis Networks offers high performance gateway solutions that allow Omni-Path HPC clusters to connect to these existing storage systems, saving time and money.

# System Under Test

A Panasas PanFS ActiveStor parallel file system on ActiveStor storage appliances is used for the storage. Four Penguin Computing Altus XE nodes are each installed with a 100GbE NIC and a 100-series Cornelis Omni-Path Host Fabric Interface (HFI). The 100GbE NIC directly connects through the Ethernet fabric to the Panasas ActiveStor director nodes. The Omni-Path HFI is connected to the Cornelis Omni-Path fabric. Also, on the Omni-Path fabric is the Cornelis Omni-Path Gateway, which contains a 100GbE NIC connecting to the Ethernet network. The test setup is illustrated in Figure 1.

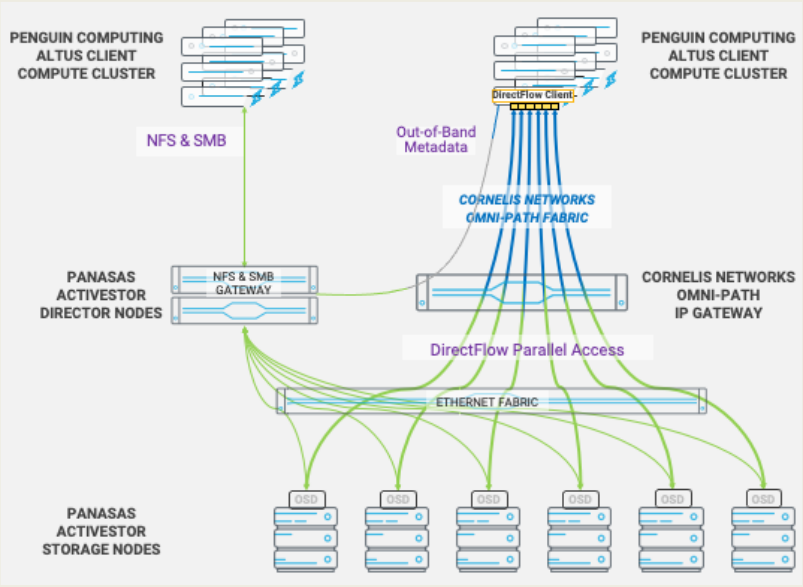


**Figure 1. Panasas PanFS ActiveStor.**

When the client nodes are operating in direct-connected Ethernet mode, the file system is mounted over the installed 100GbE NICs and no storage traffic goes over the Omni-Path network. When the client nodes are operating in the gateway mode, the file system is mounted over the Omni-Path network and routes through the gateway with a final destination to the Ethernet-based ActiveStor storage nodes. Standard kernel IP routing is handled by the Cornelis Omni-Path gateway. Out-of-band metadata between Penguin Computing Altus XE nodes and Panasas ActiveStor director nodes similarly runs over the Cornelis Omni-Path fabric and Cornelis Omni-Path Gateway when operating in the gateway mode. Complete configuration for all hardware is at the end of the paper.

# Benchmarks

To prove the functionality and performance of the solution, a distributed storage benchmark called elbencho (https://github.com/breuner/elbencho) as well as three HPC-application benchmarks spanning the climate, life sciences, and manufacturing verticals were executed. Performance is compared when writing to the file system using legacy infrastructure versus through the Cornelis Omni-Path Gateway.

# Results

The elbencho benchmarks compared the sequential read performance with increasing thread counts (concurrency) from a single client node between the Omni-Path gateway and direct-connect Ethernet modes. As shown in Figure 2, PanFS ActiveStor bandwidth saturation is achieved using 16 threads at a block size of 512kB in the Omni-Path gateway mode. Figure 3 shows the same elbencho performance benchmark results run in direct-connect Ethernet mode where scaling and maximum bandwidth achieved are very comparable to the Omni-Path gateway mode.

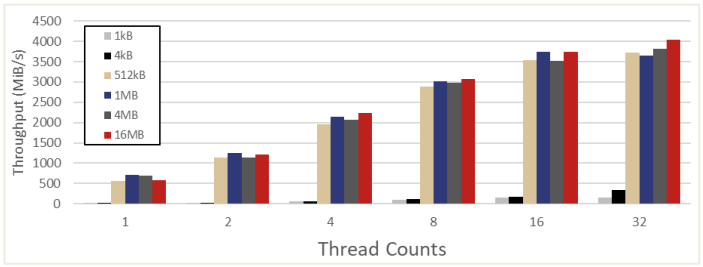


**Figure 2. elbencho sequential read performance using a single AMD client node with Cornelis Gateway.**

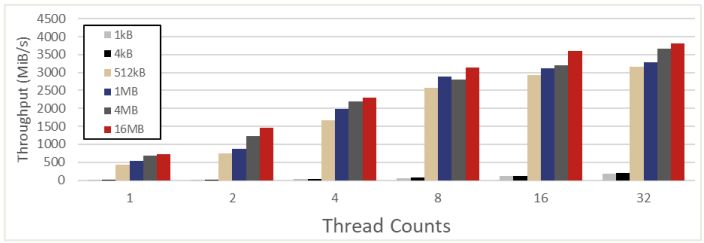


**Figure 3. elbencho sequential read performance using a single AMD client node with direct connected Ethernet.**

To further test the gateway solution, three HPC applications were selected, each representing a unique scientific vertical. For climate/weather, the Weather Research and Forecasting[1] (WRF) application running a continental U.S. 2.5-kilometer benchmark was chosen. For life sciences, the Nanoscale Molecular Dynamics[2] (NAMD) application running the Satellite Tobacco Mosaic Virus (STMV) benchmark was chosen. Lastly, for the manufacturing vertical, the OpenFOAM[3] computational fluid dynamics application was chosen using a 20M cell motorbike benchmark. For all three of these applications, the inputs were modified such that the outputs to the ActiveStor file system are artificially heavy, stressing the file system more than what would be expected under typical production runs. Exact testing details are in the configuration section of this document.

---

[1] Weather/Climate: WRF 3.9.1.1. The conus 2.5km benchmark was run with 1 MPI rank per core. The input file to WRF was modified to write output files every 5 timesteps instead of the default of 180 timesteps, with the entry in namelist.input: history interval = 5. The performance metric is average timesteps per second. Each file written by the master process is approximately 3.4GB in size.

[2] Life Sciences: NAMD 2.15a2. The NAMD molecular dynamics application was compiled using Charm 6.10.2. The STMV virus benchmark was run with increased IO by modifying `stmv.namd` as follows: `restartfreq 10; restartsave yes`. An example run command using the Charm communication library is: `mpirun -genv FI_PROVIDER=psm2 -n 16 -ppn 4 -f hostfile ./namd2 +ppn 11 +pemap 2-23,26-47 +commap 0-1,24-25, +isomalloc_sync stmv.namd`. An empirically determined optimized value of `OMP_NUM_THREADS=12` was chosen. During the IO-heavy case, each file written is approximately 50MB per output step. https://www.ks.uiuc.edu/. NAMD was developed by the Theoretical and Computational Biophysics Group in the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign.

[3] This offering is not approved or endorsed by OpenCFD Limited, producer and distributor of the OpenFOAM software via www.openfoam.com, and owner of the OPENFOAM and OpenCFD trademarks. OPENFOAM is a registered trademark of OpenCFD Limited, producer and distributor of the OpenFOAM software. Manufacturing: OpenFOAM v 22.06. OpenFOAM was run using a 20M cell motorbike benchmark with a mesh size of 100x40x40 and using scotch decomposition, also with 1 MPI rank per core. To increase the rate of IO, 'writeInterval' in system/controlDict was set to 2 for the write case, and 50000 for the no-write case, meaning no IO was performed during the simulation. Each MPI rank/process reads from and writes to its own directory structure. Per write interval, approximately 2.1GB of files are written in aggregate.

Results of the testing are summarized below showing performance versus node count. For each comparison, the performance of the baseline "no write" runs are compared to the same simulations but when the IO is enabled, either utilizing the Cornelis Networks gateway (blue) or the direct-connect 100GbE NIC. In this latter configuration, the MPI traffic is over Cornelis Omni-Path while the storage traffic goes over 100Gb Ethernet.

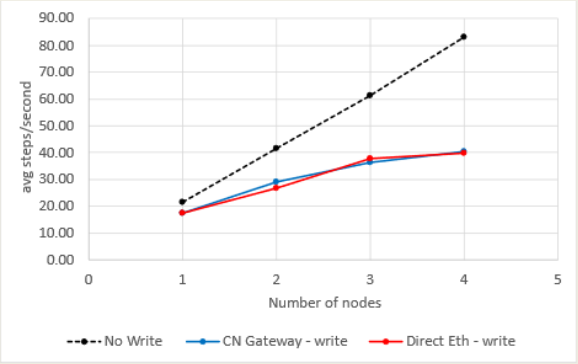


**Figure 4. WRF Performance.**

Performance in terms of average steps per second are shown in Figure 4 for WRF. As expected, nearly linear scaling is achieved for each run without file IO, as seen by the black dashed line. The solid blue line shows each run with the IO turned on and writing to Panasas ActiveStor storage nodes over the Cornelis Omni-Path Gateway. The solid red line shows each run when connected to ActiveStor directly across the Ethernet fabric. As expected, there is a performance drop when the IO is enabled, and this is for two reasons. Firstly, there are costly reduction operations to the master MPI process every time the file IO is performed. This increases the MPI communication in the application. Secondly, the IO operation itself consumes time on the master MPI process and other processes must wait to continue until the IO is finished. It can be seen in Figure 4 that the performance of the Cornelis Omni-Path Gateway is the same as when the file system is directly connected over Ethernet.

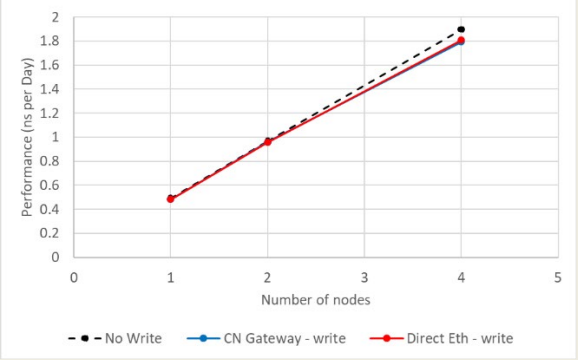


**Figure 5. NAMD Performance.**

Results for NAMD running the STMV benchmark are shown in Figure 5. Nearly linear scaling is achieved without file IO. Almost no impact to performance is seen on both the gateway and direct-connected Ethernet modes. No strong conclusions can be made from the performance data except that the Cornelis Omni-Path Gateway is a viable alternative to direct-connect Ethernet with Panasas ActiveStor storage.

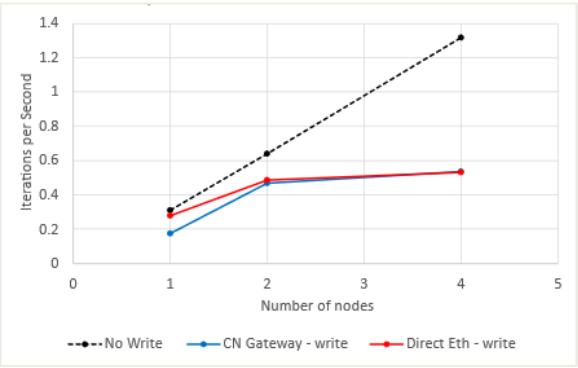


**Figure 6. OpenFOAM Performance[3].**

Results for OpenFOAM running a 20M cell motorbike benchmark are shown in Figure 6. Note that near-linear scaling is achieved as expected for the "no write" case. At two and four nodes, the performance of the Cornelis Omni-Path Gateway and the direct-connected Ethernet is almost indistinguishable and should be considered equivalent. The impact of the file IO (every two timesteps) is quite large such that there is almost no performance scaling from 2 to 4 nodes. Though this is an artificially high rate of IO for this application, it proves the viability of the Cornelis Omni-Path Gateway.

# Conclusions

The results of the testing clearly demonstrate that the Cornelis Omni-Path fabric with the Omni-Path Gateway solution provided by Cornelis Networks is a highly efficient and cost-effective alternative to Ethernet only fabrics for data centers with high-performance computing applications accessing Panasas ActiveStor storage using Ethernet in the back end. The performance benchmarks, including the elbencho distributed storage benchmark and the HPC-application benchmarks representing different scientific verticals, consistently showed comparable performance between the Omni-Path Gateway mode and direct-connected Ethernet mode.

The elbencho benchmarks revealed that the Omni-Path gateway mode achieved expected bandwidth saturation, matching the performance of direct-connect Ethernet mode. Furthermore, the HPC-application benchmarks, including the WRF, NAMD, and OpenFOAM applications, demonstrated nearly linear scaling without file input/output (IO) and only marginal performance impact when IO was enabled through the Cornelis Omni-Path Gateway. These results indicate that the gateway solution is capable of handling heavy IO operations without sacrificing performance.

Overall, the Cornelis Omni-Path Gateway solution presents a compelling proposition for data centers looking to leverage the high-performance OPX networking while utilizing their existing storage infrastructure. By eliminating the need for investing in new storage systems, organizations can scale while saving both time and money. The successful collaboration between Cornelis Networks, Penguin Solutions, and Panasas in validating this solution adds further credibility to its effectiveness and reliability.

# About the Partners

Cornelis Networks is a technology leader delivering purpose-built, high-performance fabrics accelerating HPC, High Performance Data Analytics (HPDA), and Artificial Intelligence (AI) workloads. The company's current and future product lines enable customers to efficiently focus the computational power of many processing devices on a single problem, simultaneously improving both result accuracy and time-to-solution for highly challenging application workloads. Cornelis Networks delivers its end-to-end interconnect solutions worldwide through an established set of server OEM and channel partners. For more information, visit: cornelisnetworks.com

Penguin Solutions accelerates digital transformation by deploying the power of emerging HPC, AI, and IoT technologies with solutions and services that span the continuum of Edge, Core, and Cloud. From artificial intelligence and deep learning to video analytics and digital signage, Penguin Solutions delivers the real-world deployment of high-performance applications for your science or business needs. Penguin Solutions have engineered the Altus family of compute and accelerated compute servers to maximize the benefits of AMD EPYC processors that allow for unmatched performance and compute density for the most demanding HPC workloads. For more information, visit: www.penguinsolutions.com

Panasas builds a portfolio of data solutions that deliver exceptional performance, unlimited scalability, and unparalleled reliability—all at the best total cost of ownership and lowest administrative overhead. The Panasas data engine accelerates AI and high-performance applications in manufacturing, life sciences, energy, media, financial services, and government. The company's flagship PanFS data engine and ActiveStor storage solutions uniquely combine extreme performance, scalability, and security with the reliability and simplicity of a self-managed, self-healing architecture. The Panasas data engine solves the world's most challenging problems: curing diseases, designing the next jetliner, creating mind-blowing visual effects, and using AI to predict new possibilities. For more information, visit: www.panasas.com

# Configuration

Client nodes: AXE2214GT Dual socket AMD EPYC 7413 24-core CPU @ 2.60GHz, 1024GB DDR4-3200MHz ECC RDIMMs, One dual-port 100Gb Mellanox NIC, One single-port Cornelis OPA Host Fabric adapter 100 Series, Panasas DirectFlow for Linux client software, Single 960GB SATA boot drive. RHEL 8.4.

Storage: Panasas ActiveStor Ultra ASU-100 storage enclosures x4 (16 ASU-100 storage nodes with dual 25GbE, 1.213PB raw capacity), ActiveStor Director ASD-200 enclosure (3 ASD-200 director nodes with dual 25GbE), PanFS 9.3.3.

Cornelis Networks Omni-Path 200 Gb/s 2x100 GbE Ethernet Gateway 100GWYE2G02: The data path through the gateway was limited to 100 Gbps to match the out of band Ethernet link speed. Only one Omni-Path adapter and one port of the Ethernet NIC were active for testing.