Solution Brief

High-Performance Networking
Intel® Ethernet Fabric Suite

Deliver High-Performance Networking on Ethernet

Intel® high-performance networking supports demanding workloads on clusters using a high-bandwidth, scalable Ethernet fabric.

The pandemic has undoubtedly escalated the use of artificial intelligence (AI) in business enterprises, and it has contributed to the removal of adoption barriers in new sectors. This is evident in the number of AI and advanced-analytics projects in the production stage, which has increased 54 percent from 2020 to 2021.[1] To keep up with this growth, enterprise IT teams are building new high-performance clusters that can deliver the compute power they need.

But AI and other scalable workloads such as high-performance computing (HPC) require high-performance networking. IT teams face fresh challenges in trying to meet these dual demands. First, many must make numerous budget decisions when building a new cluster. This budget must cover not only new server nodes, but also the supporting infrastructure. These budget items might include cables, rack solutions for servers, software licensing, and fabric costs. Proprietary fabrics that can support AI and HPC application workloads, such as InfiniBand EDR, can consume much of the cluster budget. This fabric cost might reduce the number of nodes that can be incorporated. In addition, these proprietary fabrics are not Ethernet, and they require specialized knowledge to implement, integrate with other data center services, and maintain. This can drive up costs because these complicated fabrics require additional IT training and are time-consuming to learn and difficult to debug.

A solution would be to use what's readily available across networks that can scale for AI and HPC. Ethernet is a common interconnect, used in the cloud, across enterprise networks, in homes, and almost everywhere else. It's already used in AI and HPC small clusters, where cost is a key criterion when deciding what fabric to use. But while Ethernet has the advantage of ubiquity, standard Ethernet alone cannot scale to meet AI/HPC requirements.

Enabling Ethernet in a clustered and distributed compute environment

Intel high-performance networking with Ethernet tackles these challenges directly by delivering Ethernet anywhere a clustered, distributed compute environment is required: HPC, AI, HPC in the cloud, and across some enterprise applications. Because high-performance networking uses standard Ethernet components, it is less expensive than proprietary solutions, and it requires no additional training. But the solution helps achieve performance comparable to proprietary fabrics when building for the most common cluster sizes.

Intel high-performance networking with Ethernet includes a software stack similar to other high-performance fabrics. It extends the capabilities of Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) v2 to enable Ethernet to become a fabric for HPC/AI.

Intel Ethernet Fabric Suite

Intel Ethernet Fabric Suite is the next generation of mature fabric software, ported to Ethernet. Intel Ethernet Fabric Suite evolved from Intel Omni-Path Architecture (OPA), and it is tuned to optimal performance with the OpenFabrics Interfaces (OFI) to scale and provide the required performance.

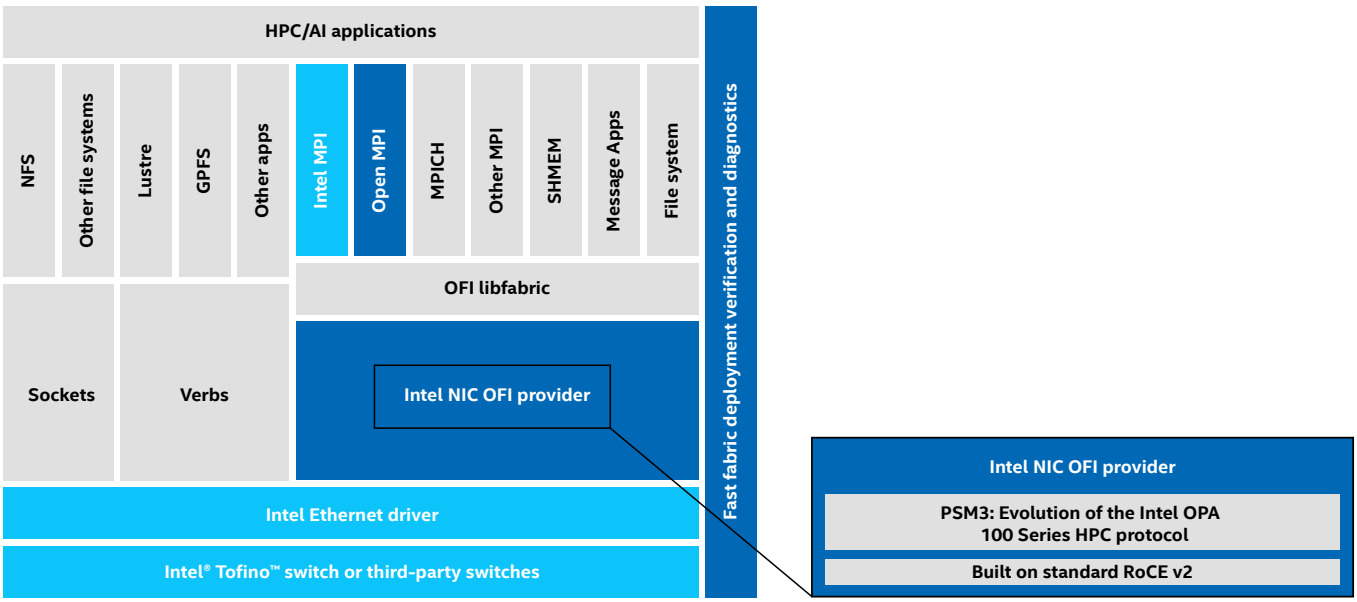

Figure 1. Intel Ethernet Fabric Suite is an evolution of Intel Omni-Path Architecture software, enhanced for Ethernet and RoCE v2

Intel Ethernet adapters

The Intel Ethernet 800 Series is a family of adapters and controllers capable of providing high bandwidth with RoCE v2 and support for PCIe Gen 3 and Gen 4. The latest-generation adapter can deliver up to 200 gigabit Ethernet (GbE) in a single PCIe Gen 4 x 16 slot to meet bandwidth-intensive application requirements.

Intel® Tofino™ and Intel Tofino 2 switches

To support high-bandwidth data transfers, Intel Tofino switches can deliver bandwidth up to 12.8 TB/s per switch application-specific integrated circuit (ASIC), with up to 400 gigabit Ethernet (GbE) per port. The latest generation of Intel Tofino switches scales performance for the most demanding use cases in distributed applications, virtual machines (VMs), AI, and serverless deployments. As the world's first P4-programmable Ethernet switch, Intel Tofino enables flexibility for many customer and Intel innovations, such as secure advanced telemetry to aid in debugging and tuning. This design opens the door to ecosystem and customer innovations.

Solution benefits

Intel high-performance networking with Ethernet is the result of Intel's leadership in providing continuous innovation for more than 35 years in Ethernet solutions. These efforts have resulted in a solution with numerous benefits.

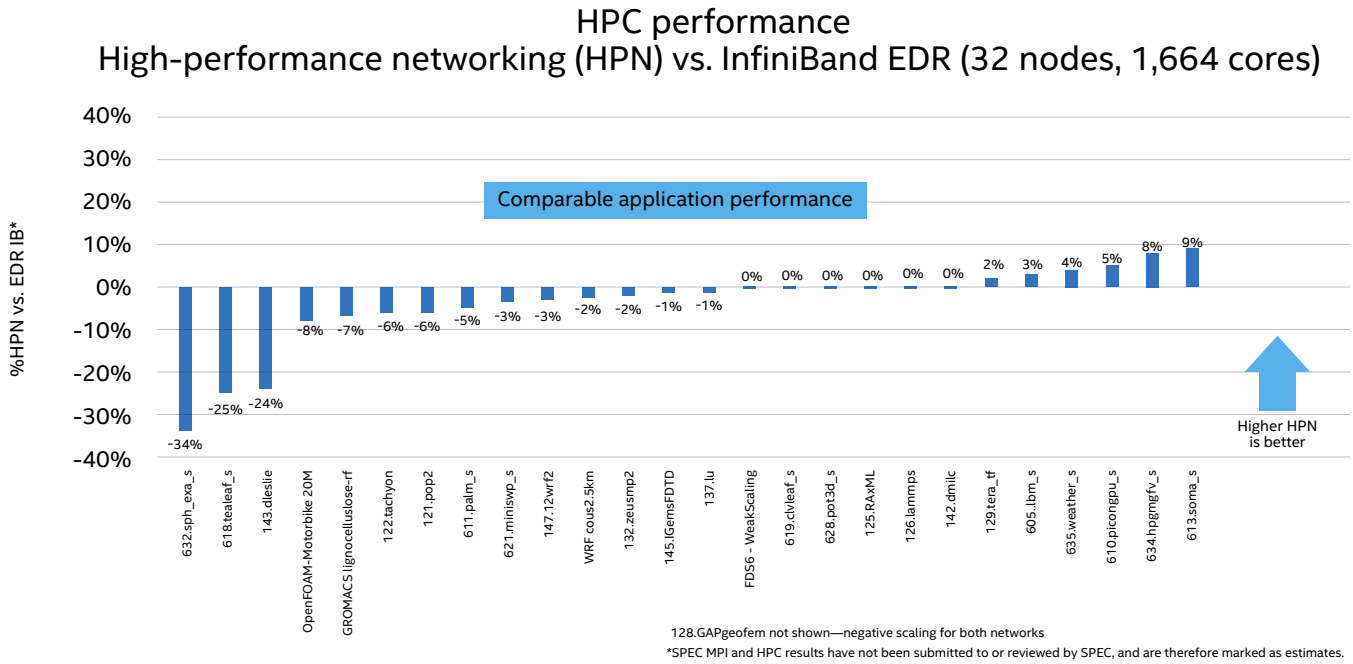

Figure 2. On average, application performance on Intel high-performance networking (HPN) with Ethernet is comparable to that of a proprietary fabric on similar systems[2]

Application performance

Application performance is the most important HPC/AI performance metric. A recent performance test demonstrated that Intel high-performance networking results were comparable to those of a proprietary fabric on a range of applications (see Figure 2).[2]

Scalability

The verbs interface, used traditionally with Ethernet, does not normally scale well for HPC and AI applications. At the heart of Intel high-performance networking with Ethernet is the Performance Scaled Messaging (PSM) layer. PSM addresses scalability challenges such as communications-stack memory footprint, latency at scale, and latency jitter and resiliency at scale. The solution, which incorporates the third generation of PSM (PSM3), enables Ethernet to scale with increasing node counts.

Lower total cost of ownership (TCO)

With comparable application performance, Intel high-performance networking with Ethernet provides the required performance for less in upfront costs than proprietary fabrics (see Figure 3).[3] This allows the purchase of additional servers for even higher cluster performance. And in addition to lower overall costs than proprietary fabrics, Ethernet switches are available with 64 ports, compared to only 36 ports on InfiniBand EDR. As a result, when the cluster size increases, the number of required switches is smaller. Fewer switches mean less data center space, fewer cables, and lower power consumption.

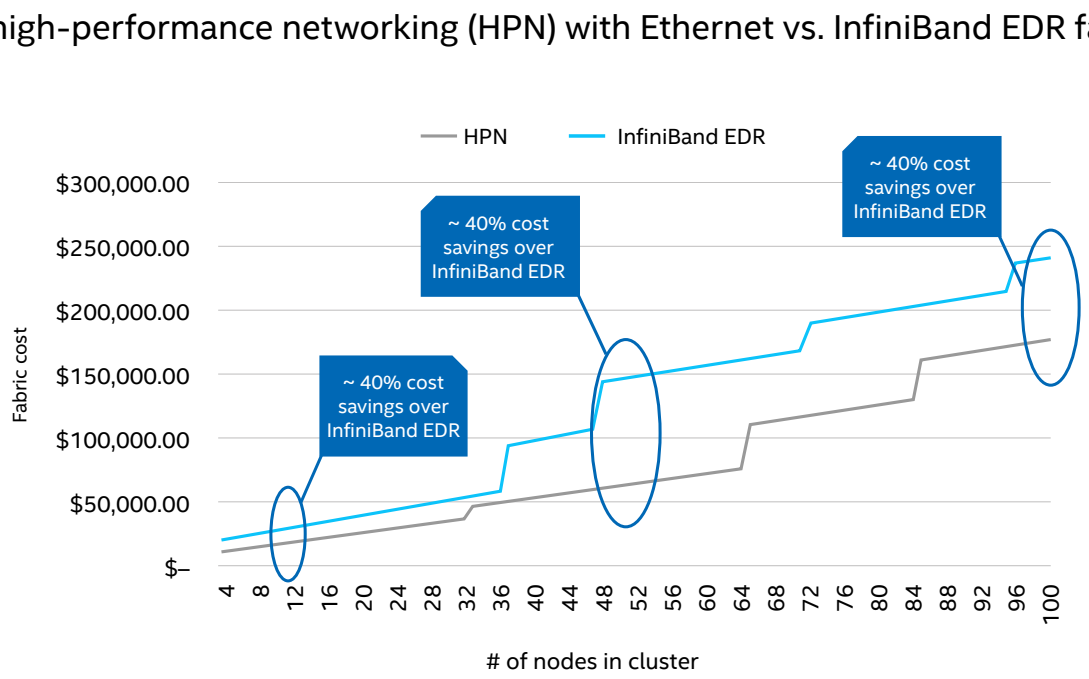

Figure 3. Customers can achieve the required performance for less in upfront costs than with InfiniBand EDR, or they can purchase additional servers for more performance[3]

No code changes

Because Intel high-performance networking is compatible with existing message passing interface (MPI) applications and supports OpenFabrics Alliance (OFA) interfaces, applications work right away. The Intel solution also offers strong verbs applications performance for file systems, reliable sockets performance, and compatibility with existing applications using OFA verbs or sockets.

Network performance analysis in real time

Other benefits include support for a wide array of telemetry and insights capabilities. Because Intel high-performance networking uses Ethernet, it offers capabilities not available with InfiniBand or other proprietary fabrics, such as end-to-end network visibility with in-band network telemetry (INT). A novel framework originated by the P4 Language Consortium, INT provides unprecedented, real-time telemetry data for every packet traversing the network. For an in-depth analysis of INT information, Intel Deep Insight Network Analytics Software collects, compiles, and reports on all telemetry data for a holistic view of the network.

Network owners can now answer critical questions about a particular packet. These questions might include:

- Which path did my packet take?
- Which rules did my packet follow?
- How long did it queue at each switch?
- Who did it share the queue with?

These technologies empower end users to remediate throughput issues with nanosecond accuracy and optimize packet flows for challenging workloads such as HPC and AI.

Support for intelligent networking

Intel high-performance networking with Ethernet supports HPC/AI clusters built on high-performing Intel® Xeon® Scalable processors. It helps ease deployment and deliver optimal performance on Intel-based clusters by supporting numerous tools and specifications such as:

- Intel Cluster Checker: Enhances system reliability and verifies that cluster components work seamlessly together for improved uptime and productivity.
- Intel MPI Library: Delivers flexible, efficient, and scalable cluster messaging on the leading MPI implementation for Intel architecture–based systems.
- Intel HPC Platform Specification: Defines foundational software and hardware requirements for broad application compatibility and interoperability across a range of HPC workloads.
- Intel Select Solutions for HPC: Eliminates guesswork with rigorously benchmarked and quick-to-deploy infrastructure optimized for analytics clusters and HPC applications. Intel Select Solutions for HPC help accelerate time to breakthrough, actionable insight, and new product design.

Learn more

Intel high-performance networking with Ethernet delivers a reliable out-of-the-box experience and proven interoperability for current and future networking infrastructure.

To learn more, contact your Intel representative or visit **intel.com/ethernet**.

Solution provided by:

intel

[1] Avasant. "Avasant Applied AI and Advanced Analytics Services 2021 RadarView." April 2021. https://avasant.com/report/applied-ai-and-advanced-analytics-services-2021-radarview/.

[2] Performance results are based on testing by Intel as of February 2021. See configuration disclosure for details. Performance varies by use, configuration, and other factors. Learn more at www.intel.com/PerformanceIndex. Configuration: Tests performed on 2-socket Intel Xeon Platinum 8170 processor at 2.10 GHz. Intel Hyper-Threading Technology (Intel HT Technology) enabled. Intel Turbo Boost Technology enabled with Intel P-State driver. Red Hat Enterprise Linux 8.1 (Ootpa) 4.18.0-147.el8.x86_64 kernel. 12 x DDR4, 196,608 MB, 2,666 megatransfers per second (MT/s). irdma version 1.3.19. ice version 1.3.2. CVL firmware-version: 2.15 0x800049c3 1.2789.0. Ethernet switch: Arista DCS-7170-32CD-F, 4.22.1FX-CLI. PFC enabled on priority 0. CX-5 InfiniBand Mellanox SB7800 Switch-IB2. CX-5: MLNX_OFED_LINUX-5.1-2.3.7.1. Intel MPI 2019.10, FI_PROVIDER=psm3 and mlx (UCX) for Mellanox. HPN: Intel Ethernet Fabric Suite 11.0.0.162. EDR InfiniBand: Mellanox OFED with UCX. Contact HPN org for application testing details.

[3] Tests performed on 2-socket Intel Xeon processor E5-2680 v4 at 2.40 GHz with Intel HT Technology enabled, Intel Turbo Boost Technology enabled with Intel P-State driver, CentOS Linux 8 (Core), kernel 4.18.0-147.8.1.el8_1.x86_64, 8 x DDR4 for 256 GB, 2,400 MT/s, irdma version 1.2.22, ice version 1.2.0_rc36, CVL firmware-version: 2.15 0x800049c3 1.2789.0, Intel MPI 2019.9, FI_PROVIDER=psm3 and VERBS, 2-tier switch fabric, Arista DCS-7060CX-32S TOR/edge, Arista DCS-7260CX-64 spines, PFC enabled on priority 0, 145.lGemsFDTD, 142.dmilc, 126.lammps stand-alone runs from SPEC MPI2007—results have not been submitted to or reviewed by SPEC and are therefore marked as estimates—WRF v3.9.1.1, and conus2.5km. Tested by Intel 11/10/2020.

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Printed in USA 0621/SB/PRW/PDC Please Recycle 346833-001US