A New High Performance Fabric for HPC

Michael Feldman **Addison Snell**

White paper

May 2016

EXECUTIVE SUMMARY

Across the cycles of new technology adoption, efficient performance at scale has endured as the driving purchase criteria for HPC. The newest generations of HPC systems are fueled by powerful multi-core and many-core processing elements, which have ushered in a new paradigm in computing parallelism. These powerful processors have created an even greater need for high-performance system interconnects, with low latency and high bandwidth being the most common metrics by which HPC cluster interconnects are evaluated.

Maintaining efficient performance at scale is a particularly relevant issue now, and with any component that must deliver performance at scale, cost often becomes the limiting factor. Larger fabrics, based on higher speed interconnects, can be expensive, not just in terms of the up-front capital, but also the operational costs of running them. In addition, computational price/performance has continued to advance at essentially a Moore's Law pace, while fabric cost, due to the complexities of implementing additional bandwidth and maintaining low latencies, is improving at a somewhat slower rate.

The introduction of the Intel Omni-Path Architecture product line marks the most significant new interconnect for HPC since the introduction of InfiniBand. Intel's rationale for its new architecture was to devise an interconnect that could scale more efficiently than other RDMA fabrics. In particular, the Omni-Path Architecture was designed with the technical and cost requirements of future exascale supercomputers in mind. At the same time, a number of features were added to make it more reliable and economical for modest-sized HPC systems.

The first generation Omni-Path products delivers 100 gigabits/sec of bandwidth per port, with port-to-port latencies on par with that of EDR InfiniBand. But the majority of Omni-Path's differentiation lies in its feature set, which aims to provide more robust error detection and traffic management than other RDMA fabrics:

- ***Packet Integrity Protection:*** a link-level error checking capability that is applied to all data traversing the wire. It allows for transparent detection and recovery of transmission errors as they occur.
- ***Dynamic Lane Scaling:*** maintains link continuity in the event of a lane failure. With the help of PIP, Omni-Path uses the remaining lanes in the link to continue operation.
- ***Traffic Flow Optimization:*** improves quality of service by allowing higher priority data packets to preempt lower priority packets, regardless of packet ordering.

Intel's Omni-Path Architecture will begin to capture a significant share of the HPC interconnect market, based in part on Intel's dominant position in the market, but also on significantly favorable end-user impressions. Intel will also be able to drive long-term cost, performance, and scalability improvements by integrating Omni-Path technology at the chip level. Omni-Path offers the HPC community a competitive and differentiated alternative InfiniBand, provided by a vendor that has demonstrated a long-term commitment to the HPC user community.

WHY HIGH PERFORMANCE FABRICS MATTER

Although high performance computing spans a wide list of domains, including manufacturing, finance, pharmaceuticals, energy, weather modeling, and scientific research, these fields are united under the HPC banner for one cause: performance. Across the cycles of new technology adoption, efficient performance at scale has endured as the driving purchase criteria for HPC.

The newest generations of HPC systems are fueled by powerful multi-core and many-core processing elements, which have ushered in a new paradigm in computing parallelism. These powerful processors have created an even greater need for high-performance system interconnects, with low latency and high bandwidth being the most common metrics by which HPC cluster interconnects are evaluated. Although interconnect performance is desirable anywhere, the nature of HPC applications is such that node-to-node communication must operate at high levels to support the kind of message passing that many science and engineering simulation codes require. Internode communication is the key to building efficient supercomputers out of powerful processing elements.

The standard MPI protocol, commonly employed in these applications, necessitates that latencies are minimized and bandwidth is maximized to support optimal application performance across these distributed memory systems. To the extent that can be achieved, these simulations can run at higher fidelity, with increased realism and greater degrees of freedom.

Maintaining efficient performance at scale is a particularly relevant issue now. In a recent study, HPC end users identified memory bandwidth as the most important criterion in evaluating processing architectures for HPC, and memory latency was about as important as the computational performance of the processors themselves.[1] As systems scale, more of the burden of bandwidth and latency falls to the system fabric, in order to maintain the high level of performance of the individual processing elements.

THE ECONOMICS OF PERFORMANCE

As with any component that must deliver performance at scale, cost often becomes the limiting factor. Larger fabrics, based on higher speed interconnects, can be expensive, not just in terms of the up-front capital, but also the operational costs of running them. This latter category includes both the administration of maintaining more complex networks and the cost of power consumption across the entire fabric infrastructure, including host adapters, edge switches, director switches, and either copper or optical cables.

In addition, the relative costs of compute and fabrics have diverged. Computational price/performance has continued to advance at essentially a Moore's Law pace, while fabric cost, due to the complexities of implementing additional bandwidth and maintaining low latencies, is improving at a somewhat slower rate. Essentially, it is easier to improve computational performance than communication performance. That's not to say fabric price/performance isn't improving – it is – but the relatively slower rate of improvement means users must spend proportionally more on interconnects than computing as time passes.

This trend is revealed in the Intersect360 Research report on HPC users' budget allocations, which tracks spending at HPC sites.[2] According to these surveys, over the last five years, the share of network costs rose from 4.8 percent in 2011 to 5.8 percent in 2015, representing a 20 percent increase during this period. That means less money, relatively speaking, was available to spend on other categories, like storage, software, staffing, or processors.

[1] Intersect360 Research special study, "Processing Architectures in HPC," 2016.
[2] Intersect360 Research, "HPC Budget Allocation Map: HPC Budget Distribution," January 2016.

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

A NEW INTERCONNECT FOR HPC

The introduction of the Intel Omni-Path Architecture product line marks the most significant new interconnect for HPC since the introduction of InfiniBand. Although its roots are in two acquisitions – the InfiniBand technology Intel acquired from QLogic and the Aries interconnect purchased from Cray – Omni-Path represents a unique fabric, incorporating the QLogic True Scale architecture and its associated software stack with high-performance features from Cray. As a result, Omni-Path is software compliant with the Open Fabrics Alliance (OFA) stack for RDMA fabrics, while at the same time, it introduces a number of features that differentiate it from InfiniBand technology.

Intel's rationale for its new architecture was to devise an interconnect that could scale more efficiently than other RDMA fabrics. In particular, the Omni-Path Architecture was designed with the technical and cost requirements of future exascale supercomputers in mind. At the same time, a number of features were added to make it more reliable and economical for modest-sized HPC systems.

Advanced Features

The first generation Omni-Path products delivers 100 gigabits/sec of bandwidth per port, with port-to-port latencies on par with that of EDR InfiniBand. In fact, internal testing conducted by Intel has yielded better fabric latencies (17 percent lower) and messaging rates (27 percent higher) than EDR. For each host interface, Intel has stated that their host architecture supports message rates of up to 160 million messages per second (unidirectional), depending on the CPU. The switch ASIC itself is capable of 195 million messages per second.

But the majority of Omni-Path's differentiation lies in its feature set, which aims to provide more robust error detection and traffic management than other RDMA fabrics. These enhancements are largely hidden from the application or end user, since they operate at the level of the Omni-Path wire protocols.

One of those features is Packet Integrity Protection (PIP), which is a link-level error checking capability that is applied to all data traversing the wire. It allows for transparent detection and recovery of transmission errors as they occur. Since it is built into the wire protocol, it doesn't add latency, as has been the case for the traditional forward error detection scheme employed with InfiniBand. Therefore, latency and overall network behavior becomes more predictable, even as the fabric scales.

Note that with InfiniBand, error detection and correction can be turned on or off. The decision to build in error detection into Omni-Path was driven by the fact that at faster data rates, more errors are going to occur, even at the relatively small distances as would be encountered in a modest-sized cluster. At slower data rates, with relatively short cables, turning off error detection was a reasonable option. End-to-end retries were few and far between. At 100 GB/sec, even moderately large clusters (i.e., more than 1,000 nodes) are going to necessitate error detection in most situations. As data rates continue to rise, it will be even more critical.

A feature related to Packet Integrity Protection, known as Dynamic Lane Scaling (DLS), maintains link continuity in the event of a lane failure. With the help of PIP, Omni-Path uses the remaining lanes in the link to continue operation. The faulty lane can be addressed after the application completes. In another network, a lane failure could result in the shutdown or re-initialization of the link, which may prevent the application from completing.

Another important Omni-Path feature is Traffic Flow Optimization (TFO), which is a capability designed to improve quality of service (QoS). It works by allowing higher priority data packets to preempt lower priority packets, regardless of packet ordering. In practice, this means MPI packets can be given preference to bulk storage or checkpoint traffic, even when these packets arrive ahead of MPI packets. By doing so, it reduces latency variation (jitter) when MPI and storage data are being transferred on congested links. This, in turn,

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

allows an application to run in a much more deterministic manner from one run to the next. Deterministic behavior for HPC applications is highly desirable, since it helps ensure consistent results and makes it easier to find software bugs.

Cost Advantages

One of the critical design goals of Omni-Path was to reduce fabric cost. The technical rationale for doing so is provided in the "The Economics of Performance" section above. As a vendor, Intel also has a natural incentive to develop a cost-optimized fabric, given its desire to compete effectively against vendors offering InfiniBand and Ethernet alternatives.

One of the main paths of attack to lower cost involves denser switching. The first generation Omni-Path product line uses a 48-port switch chip, a 33 percent increase from the traditional 36-port switch ASIC used for InfiniBand. That density can translate to less switches and cables in a system, saving not just in upfront purchase expenses, but also in power and maintenance costs, as well as data center floor space.

For example, a relatively modest-sized cluster of 768 nodes, would need just a single Omni-Path director switch to support a non-blocking fat-tree topology. For InfiniBand, that would entail two director switches plus forty-three additional edge switches, along with the extra cabling required to link the switches together. In addition, since the Omni-Path configuration in this scenario eliminates the edge switches, there are two less hops (three hops versus five hops) compared to InfiniBand. That will reduce end-to-end latency by about half. As cluster size grows into the thousands of nodes, cost savings and latency reductions grow proportionally.

Given a fixed budget, the cost savings could be applied to purchasing more performance, in the form of additional servers, faster processors, or faster storage. Intel is projecting customers will be able to buy up to 26 percent more servers due to Omni-Path cost savings, relative to EDR InfiniBand. More to the point, if users are able to spend proportionally less on fabric costs relative to overall infrastructure spending, customers will gain additional financial flexibility for other data center spending; it need not be for more performance.

Software Compatibility

The introduction of a new architecture is often difficult. The most pressing challenge is that of software, which often must be re-targeted to the new hardware. As mentioned previously, Omni-Path avoids this pain since it uses the OpenFabrics Alliance interfaces that are in common use across software designed to run on InfiniBand. To the application, Omni-Path looks like any other RDMA fabric.

For MPI, Intel is relying on the True Scale software technology it inherited from QLogic. Existing MPI programs and libraries for True Scale that use the Performance Scaled Messaging (PSM) library work as is without recompilation. The additional Omni-Path features supported by an enhanced version of the library (PSM2), programs do not need to be updated to utilize these features. Since Omni-Path supports larger fabrics, applications will need to be recompiled to take advantage of the available larger scale. With PSM2 being a superset of PSM, backward compatibility is maintained.

Intel supplies Omni-Path Linux drivers and other software components required for Red Hat Enterprise Linux (RHEL), SUSE Linux Enterprise Server (SLES), and other Linux operating systems. In addition, Intel provides a complete, open-source software solution for server nodes and switches, including a system management suite, a host software stack, a fabric management GUI, and a fabric management stack.

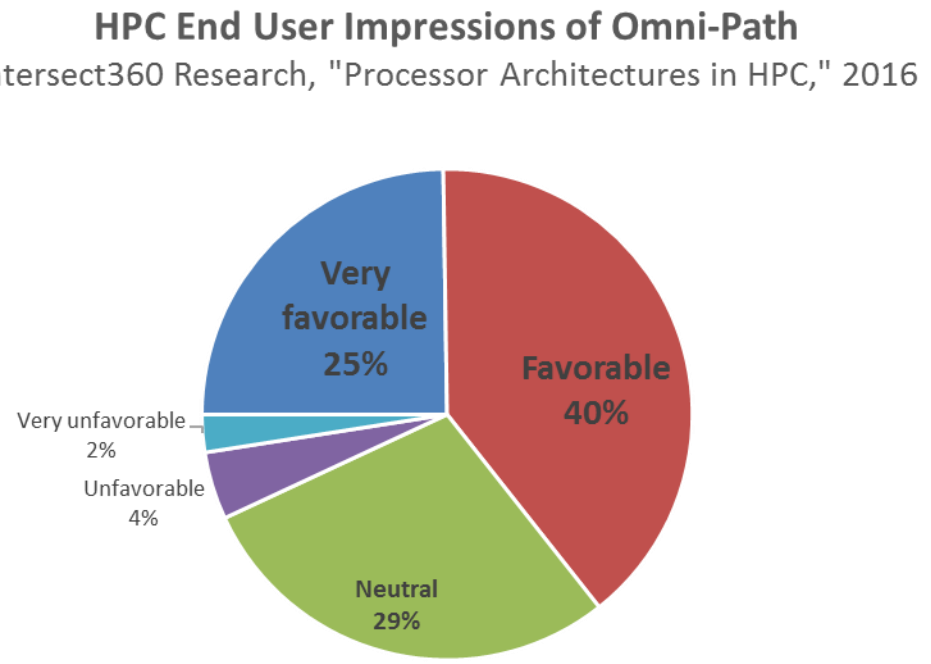
Host Integration Roadmap

The initial release of the Omni-Path utilizes traditional host adapter cards for server-side connectivity. However, subsequently Intel is planning to offer an in-package host adapter configuration, where the fabric ASIC is integrated into the processor socket. This capability will first appear with Xeon Phi processors, and later with Xeon processors. Further

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

down the road, the Omni-Path host interface will be integrated directly into the processor.

When the host adapter circuitry is moved closer to the processing element the advantages will be three-fold: performance, power efficiency, and cost. All are improved by removing the extra hardware required when a host adapter device is connected to a PCIe bus, which is used to transfer data to and from the processor. Latency, in particular, can be optimized by avoiding the bus transfer. The performance enhancement will be especially applicable to processors destined for HPC work, where an in-package or on-chip host interface can be used to deliver both lower latency and greater effective bandwidth.



INTERSECT360 RESEARCH ANALYSIS

Omni-Path presents a viable alternative to current interconnect fabrics in current use by HPC. We expect a large number of end users to trial Omni-Path over the next few years, and as a result, Omni-Path will begin to capture a significant share of the HPC interconnect market. That assessment is based on four principle contentions, which are further elaborated below:

- The favorable impression Omni-Path by the HPC end user community;
- The ability of Intel to differentiate Omni-Path hardware based on its core competencies in integrated circuit design and manufacturing;
- Intel's position as a market leader in HPC;
- The Omni-Path technology's performance and differentiated feature set outlined in this paper.

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

Favorable expectations: When HPC users were asked about their forward-looking impressions of Omni-Path, 65 percent reported they had a favorable opinion of the technology (ranked 4 or 5 on a scale of 1 to 5), based on what they knew about Omni-Path prior to its commercial debut (survey conducted in Q1 2016; see chart above). For all users, the average favorability rating was 3.8.[3] That represents a remarkably positive outlook for a product that had not yet been released at the time of the survey.

Design and Manufacturing Expertise: As a manufacturer of semiconductor devices, Intel possesses some of the best design expertise and technology in the computer industry. As it has successfully done with its line of processors and SSDs, the company should be able to leverage this expertise to its host adapter and switch devices. Moreover, Intel's leadership in process technology manufacturing confers performance and energy efficiency advantages across to its semiconductor products.

Market Leadership: Intel processors are currently used in 90 percent of all new HPC systems deployed.[4] That level of dominance gives the company an enviable base on which to build a business based on products that connect to those processors. Although Power and ARM may erode some of that market over the next few years, they are not an existential threat to Intel's position. Leveraging this market position, Omni-Path products are expected to be offered with more than 100 OEM server and storage platforms at introduction.

Performance and Feature Set: With Omni-Path, Intel has made a strategic decision to build an interconnect to compete at the performance end of the market. While the fabric will almost assuredly move into non-HPC use cases in the future, Intel chose to establish its products at the high end of the market first. To that end, they included the performance, scalability and QoS features that will be necessary to succeed in HPC against well-established InfiniBand and Ethernet product lines

That's not to say Omni-Path will become a dominant fabric in the near-term. InfiniBand enjoys an established HPC customer base, a mature and growing software stack, and an aggressive performance roadmap. Ethernet solutions also claim a large part of the HPC market, especially where fabric performance is not paramount and where the attraction of employing the most widely used network standard is an important consideration. Omni-Path offers the HPC community a competitive and differentiated alternative to these legacy fabrics, provided by a vendor that has demonstrated a long-term commitment to the HPC user community.

.

[3] Intersect360 Research special study, "Processor Architectures in HPC," 2016.
[4] Intersect360 Research, "HPC User Site Census: Processors," October 2015.

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com