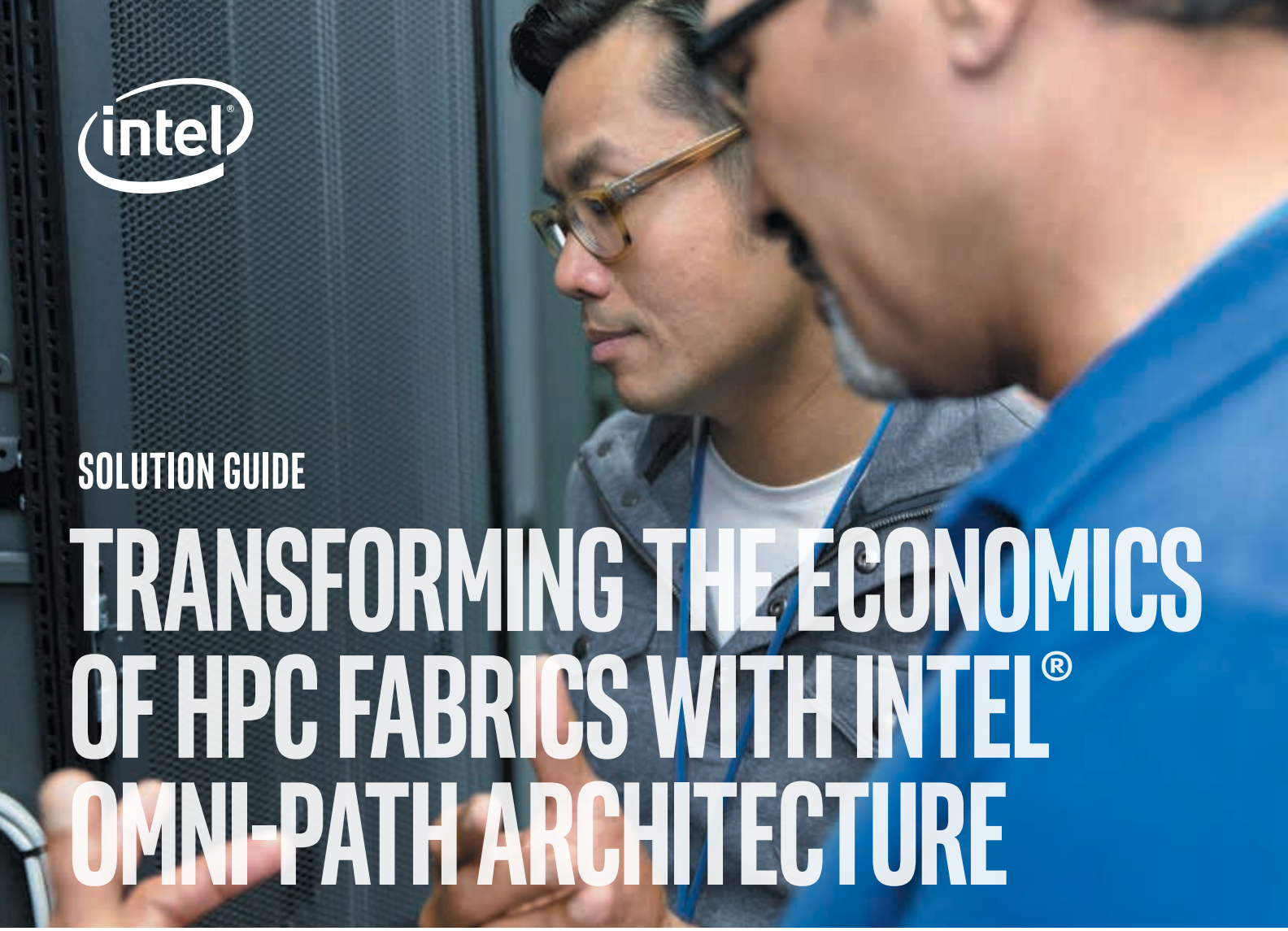

SOLUTION GUIDE

TRANSFORMING THE ECONOMICS OF HPC FABRICS WITH INTEL® OMNI-PATH ARCHITECTURE

Executive Summary

Today's rapid advances in processor and storage performance provide new opportunities for every organization that relies on high performance computing (HPC). Yet an HPC cluster is only as powerful as its weakest link. Compute, storage, and networking resources must be balanced to avoid bottlenecks that bring down the capability of the entire cluster. Intel® Omni-Path Architecture (Intel® OPA), an element of Intel® Scalable System Framework, answers this challenge by providing a major leap forward in fabric performance, scalability, and cost-efficiency.

Intel OPA builds on proven Aries fabric and Intel® True Scale Fabric, increasing link bandwidth to 100 Gbps and integrating new open source and Aries fabric technologies.[1] Optimized protocols provide low latency that stays low even at extreme scale.[2] The optimizations also provide fast and efficient file system throughput, high packet reliability, and low deterministic latency for high priority communications. Advanced traffic shaping features extend these advantages to deliver even higher levels of performance, scalability, and resiliency.

Perhaps most importantly, Intel OPA offers scalability and cost advantages that will increase with every new Intel platform generation. Today's family of 24- and 48-port edge switches and 192- and 768-port director class switches are based on new 48-port switch silicon. These high port densities can reduce infrastructure requirements by as much as 50 percent compared with leading InfiniBand* solutions.[3]

Next-generation Intel® Xeon Phi™ processors and select Intel® Xeon® processors will include integrated Intel OPA controllers to eliminate the need for separate host adapters. Future Intel platform generations will feature even tighter integration, providing an increasingly efficient foundation for continuing improvements in bandwidth, latency, and cost models.

This guide provides a high-level overview of Intel OPA, and identifies the key features and end-to-end product offerings that make it a superior HPC fabric at every scale.

Driving HPC Forward

The Intel® Scalable System Framework

Balanced resources are essential for optimizing HPC performance, yet achieving that balance can be a significant technical challenge. Intel® Scalable System Framework is a flexible blueprint for developing high performance, balanced, power-efficient, and reliable systems capable of supporting both compute- and data-intensive workloads.

The framework combines next generation Intel® Xeon® processors and Intel® Xeon Phi™ processors, Intel® Omni-Path fabric, silicon photonics, innovative memory technologies, and the Intel® Lustre* parallel file system, along with the ability to efficiently integrate them into a broad spectrum of system solutions. The framework also provides a ubiquitous and standards-based programming model, extending the ecosystem's current investments in existing code for future generations.

Learn more at: www.intel.com/SSF

Barriers to Growth in High Performance Computing

HPC requirements continue to grow. Organizations across almost every scientific and engineering discipline want to study more complex models with more variables and greater detail. They also want faster access to high quality results, so they can accelerate their research and their product development lifecycles. Meeting these needs requires new HPC solutions that are not only more powerful and scalable, but also more cost effective so organizations can harness more computing power without overextending their budgets.

Rising core counts in Intel Xeon processors and Intel Xeon Phi coprocessors help address these challenges by delivering great compute density at low cost and with low power consumption. However, high-density processing can strain the performance of other computing resources, such as memory and I/O in individual server platforms, and storage and interconnect fabrics in clustered architectures. Efficient HPC requires balanced performance across all these resources.

Breaking Down the Barriers through Synchronized Innovation

To address the need for balanced performance, Intel is driving synchronized innovation throughout the HPC solution stack. Improvements in core density are balanced by complementary improvements in memory, I/O, fabric, and storage performance. This helps to enable high performance throughout an HPC system, without bottlenecks and without customers having to overinvest in one cluster resource at the expense of others.

Intel OPA is a critical component of this synchronized innovation. It is a comprehensive fabric solution that includes host adapters, silicon, edge switches, and director switches. It also includes active and passive cables, and complete software and management tools.

End-to-end optimizations have been implemented to improve the speed, efficiency, and scalability of communications in clustered architectures. Intel OPA can support tens of thousands of cluster nodes today, and will scale in the future to support hundreds of thousands of nodes.

The high port densities of Intel OPA help to drive down the cost of deploying and scaling an HPC fabric. These cost savings are much needed, since fabric infrastructure can account for as much as 20 to 40 percent of total cluster costs.[4] By making HPC fabrics more powerful and affordable, Intel OPA can help organizations shift their investments to purchase more compute power and achieve higher overall cluster performance (Figure 1).

Why InfiniBand Can't Meet Growing Needs

The InfiniBand interconnect protocol dates back to the early 2000s. It was designed as a generic channel interconnect for the data center and was only later retrofitted for HPC. Although its overall bandwidth has increased to accommodate growing needs, it continues to rely on heavyweight, Verbs-based protocol libraries that add unnecessary latency to every packet.

InfiniBand also relies on a communications model that offloads much of the packet-processing workload to the host channel adapters (HCAs). Since the connection address information required for transmitting data is stored in the

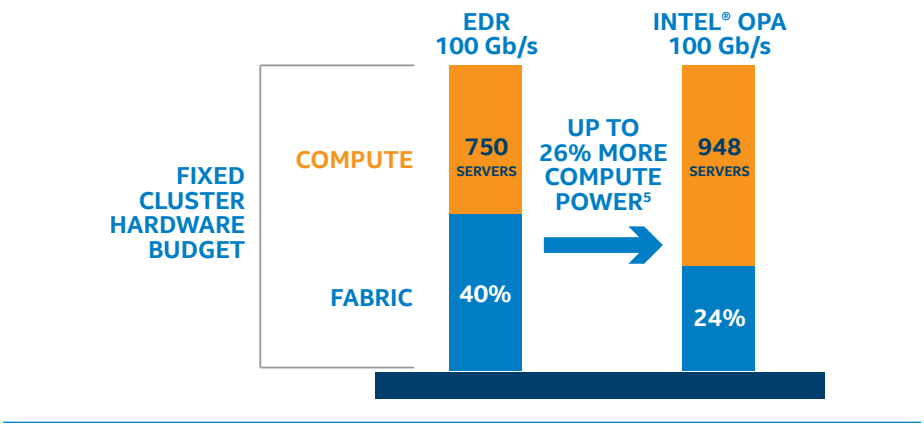

Figure 1. By delivering high performance, high port densities, and excellent scaling, Intel® Omni-Path Architecture helps to improve fabric cost models so organizations can purchase more compute capacity within the same HPC budget.[5]

HCAs, any change in routing patterns can lead to adapter cache misses. Address information must then be retrieved from host memory, a high-latency process that disrupts traffic flow. As a result, any attempt to optimize communication pathways during runtime are likely to be counterproductive, especially if the optimizations are performed frequently. A more efficient approach is required to meet today's growing needs.

Building on Proven Technologies to Evolve Fabric Capability

Intel's journey to transform fabric solutions began with the development of Intel True Scale Fabric, which addressed many of the inherent weaknesses of InfiniBand. Intel OPA builds on these advances, not only by increasing link bandwidth, but also by integrating innovations from the open source community and from Aries interconnect technology. These foundational technologies are described below.

On-Load Design Model

Instead of the offload design used by traditional InfiniBand, Intel True Scale introduced an on-load model that shifts some of the compute-intensive portions of the work from the HCAs to the more capable host system processors. This model has been adopted by Intel OPA and provides a number of benefits.

- **Low latency and high efficiency.** Connection address information is maintained in host memory so all inbound packets "hit" and can be processed with deterministic latency. Adapter cache misses are eliminated and routing pathways can be optimized during runtime to make better use of fabric resources.
- **Excellent performance scaling.** Packet throughput scales with the number of cores in the host system, allowing fabric performance to improve automatically as core densities increase in future Intel Xeon processors and Intel Xeon Phi coprocessors.
- **Simpler, more cost-effective HCAs.** Shifting compute-intensive elements of the workload to the host processors reduces HCA processing and power requirements. Designs are simpler, the potential for bottlenecks is reduced, and the need for packet processing firmware is eliminated.

Performance Scaled Messaging (PSM) Software

Intel True Scale Fabric replaced the Verbs protocols of traditional InfiniBand with the lightweight, higher-performing protocols of open source PSM software. Since PSM is semantically matched with message passing interface (MPI) software, it decreases the amount of required code by as much as 90 percent.[6] As a result, processing overhead and packet latency are both reduced.

PSM has proven its capability in production environments for more than 10 years. It provides extremely high message rates, especially with small message sizes. It also delivers very low fabric latency that remains low at scale.[2]

Run-Time Traffic Shaping

Intel True Scale fabric introduced run-time traffic shaping techniques that make better use of all available fabric resources. Congestion control technologies identify bottlenecks and reroute traffic to less congested pathways. Adaptive Dispersive Routing takes advantage of multiple pathways to improve utilization, latency, and redundancy for every communication.

Innovations from the Aries Interconnect

Intel OPA also builds on a number of innovative, forward-looking fabric technologies that were acquired from Cray and are based on Cray's next-generation Aries interconnect solutions.

- **Flow Control Digits (FLITS)** allow data transport to be controlled at a more granular level.
- **No-load error detection** allows corrupted packets to be detected without increasing latency, and then corrected with relatively little impact on performance [this technology is implemented as Packet Integrity Protection (PIP) in Intel OPA].

Delivering a Major Leap Forward with Intel Omni-Path Architecture

Intel OPA builds on the innovations described in the previous section to deliver major gains in fabric capability. The most important advances are described below and summarized in Table 1.

2.5X Higher Bandwidth

Intel OPA operates at 100 Gbps, providing 2.5X the bandwidth of 40 Gbps Intel True Scale Fabric solutions. Intel OPA also delivers extremely low latency that stays low at scale (100 to 110 ns per port[7]). With up to two fabric ports per two-socket server, Intel OPA supports up to 50 GB/s of bidirectional bandwidth per cluster node.

Traffic Flow Optimization (TFO) for Advanced Quality of Service (QoS)

Unlike current InfiniBand solutions, Intel OPA can interrupt the transmission of large data packets in order to transmit higher priority packets. All communications are broken into 65-bit FLITS, which are bundled into 1056-bit Link Transfer Packets (LTPs), and then into larger message transmission units (MTUs). InfiniBand solutions have to wait for an MTU to finish

TABLE 1. INTEL® OMNI-PATH ARCHITECTURE: BUILDING ON THE FOUNDATION OF INTEL® TRUE SCALE FABRIC[8]

FEATURE	INTEL® TRUE SCALE FABRIC	INTEL® OMNI-PATH ARCHITECTURE	BENEFIT
Speed and Throughput			
Link Bandwidth	40 Gb/s (10 GB/s)	100 Gb/s (25 GB/s)	2.5X-plus higher bandwidth
Port Latency	140 ns	100-110 ns[7]	Up to 25 percent lower latency
Message Rate (per port)	40M pp/s	195M pp/s	Up to 4.8X more message rate of Intel® True Scale Fabric
Architectural Enhancements			
Edge Switch Size	18/36	24/48	Fewer switches, fewer switch hops for large clusters
Director Port Counts	72/216/384/648	192/768	Higher port density, flexible fabric designs
Switch Leaf Size	18	32	Higher port density
Director Port Counts	72/216/384/648	192/768	Flexible fabric designs
Max Compute Nodes 5 hop fabric	11,664	27,628	Higher scalability with lower latency, power, and cost
MTU Size	Up to 4K	Up to 10K	Improved file transfer efficiency

processing before other packets can be sent. Intel OPA can halt transmission at the end of any FLIT to send higher priority packets.

The ability to provide low, deterministic latency for high priority traffic makes Intel OPA ideal for consolidating MPI, storage, and other communications on a single fabric. Intel OPA also supports MTUs of up to 10 KB and provides additional optimizations to improve file system performance.[9] Throughput is optimized for both small packet and large packet traffic, and critical MPI traffic receives top priority.

Packet Integrity Protection (PIP) for Latency-Efficient Error Detection

Intel OPA implements error checking and correction for high packet reliability with no increase in latency. A cyclic redundancy check (CRC) is performed at line speed for every LTP at each fabric node. The CRC is checked when the LTP is received at the next node, also at line speed. If an error is detected, the LTP (not the entire MTU) is resent from the previous node's retry buffer. There is no added latency unless a retry is triggered. Even then, the only additional latency is the time required to communicate the need for a retry and to transmit and receive the retry across the link.

This strategy eliminates the long delays associated with end-to-end error recovery techniques that require error notices and retries to traverse the full length of the fabric. Some recent InfiniBand implementations support link-level error correction through a technology called Forward Error Correction (FEC). However, FEC introduces additional latency into the normal packet processing pipeline. Intel OPA provides similar levels of integrity assurance without the added latency.

Dynamic Lane Scaling

Each 100 Gbps Intel OPA link is composed of four 25 Gbps lanes. In traditional InfiniBand implementations, if one lane fails, the entire link goes down, which often causes a running HPC application to fail. In contrast, if an Intel OPA lane fails, the rest of the link remains up and continues to provide 75 percent of the original bandwidth.[10] If packets have been lost or damaged, PIP automatically fixes most errors. Bandwidth remains high, communications remain error-free, applications remain up and running, and the cable can be serviced at a more convenient time (such as when running jobs have completed). Once the cable is replaced, full service is automatically restored.

CPU-Fabric Integration

Intel OPA Host Fabric Interface (HFI) silicon will be integrated into next-generation Intel Xeon Phi processors and select Intel Xeon processors, so separate fabric add-in cards will no longer be required, further reducing fabric cost and complexity. This integration will also help to deliver improvements in performance, density, power efficiency, and reliability. Tighter integration in future product generations will provide a foundation for delivering increasing performance and scalability with improved cost models.

End-to-End Product Support

Intel OPA offers complete product support for building HPC fabrics at every scale, including host fabric adapters, edge switches, director switches, and silicon for custom adapter and switch designs. Passive copper and active fiber optic cables are also available.

Host Fabric Adapters (and Silicon)

The Intel® Omni-Path Host Fabric Adapter 100 Series provides up to 100 Gbps of bandwidth per port, for up to 25 GBps of bidirectional bandwidth. This low profile PCIe x16 card supports both passive copper and active optical fiber connections. It consumes a maximum of 12 watts (8 watts is typical) and supports multi-core scaling to help increase performance in combination with today's dense, multi-core processors.

HOST ADAPTERS	
Intel® Omni-Path Host Fabric Adapter 100 Series	Single port x16 Host Fabric Interface (HFI), 100 Gbps
	Single port x8 HFI, 58 Gbps
Intel® Omni-Path Host Fabric Interface Silicon 100 Series	For customized OEM designs (supports up to 2 ports)

The Intel Omni-Path Host Fabric Adapter 100 Series is also available in a PCIe x8 configuration that supports up to 58 Gbps of bandwidth for less demanding environments. Intel OPA Host Fabric Adapters are based on an Intel application specific integrated circuit (ASIC), which is available to third-party fabric vendors for customized adapter designs.

Edge and Director Switches (and Silicon)

One of the most effective ways to improve scalability and cost models in HPC fabrics is to increase switch port densities. To deliver on this need, Intel developed a high-density, 48-port ASIC for Intel OPA switches. This silicon reduces the number of required switch chips by up to 50 percent in a typical fat tree fabric configuration[3] (see the sidebar, Quantifying the Cost Benefits of Intel Omni-Path Fabric).

EDGE SWITCHES	
Intel® Omni-Path Fabric Edge Switch 100 Series	48 ports, 1U chassis
	24 ports, 1U chassis
DIRECTOR SWITCHES	
Intel® Omni-Path Fabric Director Switch 100 Series	786 ports, 20U chassis
	192 ports, 7U chassis
SILICON	
Intel® Omni-Path Fabric Switch Silicon 100 Series	For customized OEM designs (supports up to 48 ports)

Intel offers two families of Intel OPA switches based on this high-density ASIC. Both switch families provide 100 Gbps per port with port-to-port latency of 100-110 ns.[7] They also support both passive copper and active optical fiber connections, and include redundant fans. Redundant power is optional.

- **The Intel Omni-Path Fabric Edge Switch 100 Series** comes in a 1U chassis design with either 24 or 48 ports. An optional management board provides out-of-band chassis management. It also provides tools for managing subnets and performance.
- **The Intel Omni-Path Fabric Director Switch 100 Series** comes in a 20U chassis design that supports up to 768 ports for up to 153.6 terabits per second aggregate bandwidth. It is also available in a 7U chassis design that supports up to 192 ports for a maximum aggregate bandwidth of up to 38.4 terabits per second. In each case, ports can be added in 32-port increments up to the full capacity of the switch. Both director switches include integrated chassis management.

Quantifying the Cost Benefits of Intel® Omni-Path Architecture

With their high port densities, Intel Omni-Path Architecture switches can help organizations reduce the number of required switches, cables, and racks compared with traditional InfiniBand solutions. This not only reduces infrastructure costs, but also helps to decrease both power consumption and overall fabric latency.

For a quantitative example of the potential benefits, see the Intel Solution Brief, "Higher Performance at Lower Cost for HPC Fabrics."

All Intel OPA switches can be used either as stand-alone switches or combined to build much larger fabrics. A single 5-hop fabric built with these edge switches can support up to 27,648 nodes, which is up to 2.3 times more nodes than can be supported with Intel True Scale switches or other current InfiniBand switch designs.[11] The scalability of Intel OPA allows large fabrics to be implemented with few network hops for low end-to-end latency. For example, an 11,664 node fabric based on traditional InfiniBand would require a 7-hop fabric versus a 5-hop fabric (using Intel OPA).

Passive and Active Cables

Intel offers passive copper cables, as well as active optical fiber cables. The active cables are based on Intel® Silicon Photonics Technology, and can be used to implement high-performance links over longer distances. Both types of cables are also available from third-party vendors.

CABLES	
Passive copper cables	• See intel.com/omnipath for specific offerings • Cables are also available form third-party vendors
Active fiber optic cables (based on Intel® Silicon Photonics Technology)	

A Complete Software Stack and Management Suite

Intel OPA uses the same Open Fabrics Alliance (OFA) interfaces used by InfiniBand,* thus ensuring that the vast majority of commercial and community HPC applications should "just work" with no code changes. This ensures a robust software ecosystem. It also reduces the cost, complexity, and risk of an upgrade to Intel OPA.

Intel OPA Host Software includes an enhanced version of PSM software called PSM2 that provides high bandwidth and more efficient packet processing. It builds on the Intel True Scale software stack, which has been proven in production environments for more than 10 years.

Intel has contributed all of its host fabric software code to the open-source community, including PSM2. Intel OPA support will be incorporated into future "in box" Linux operating system (OS) distributions. Please contact your Linux distribution vendor for specific support plans and timing.

This suite of tools provides the advanced routing algorithms, diagnostic tools, and failover capabilities needed to optimize end-to-end fabric performance and uptime. It provides fabric-wide visibility and control and includes Fast Fabric tools for automated fabric deployment, verification, and debugging.

Conclusion

Faster and more cost-effective fabrics are needed to support continuing growth in HPC performance. Intel OPA provides an end-to-end 100 Gbps solution that delivers high bandwidth and low latency that stays low at scale—all while keeping fabric costs in line with previous-generation solutions. It also includes advanced traffic optimization technologies that make more efficient use of all availabe resources.

Intel OPA is designed to scale over time to balance the growing performance of Intel processors, coprocessors, and storage solutions. Tighter integration, higher port densities, lower power consumption, and improved efficiency will help take performance versus cost to new levels, so organizations can continue to get higher total performance from every dollar they invest in their HPC clusters.

More Information

Literature:

Higher Performance at Lower Cost for HPC Fabrics

Intel® Omni-Path Director Class Switch 100 Series Product Brief

Intel® Omni-Path Edge Switches 100 Series Product Brief

Intel® Omni-Path Host Fabric Interface Adapter Product Brief

Intel Web Site:

Intel High Performance Computing Fabrics www.intel.com/hpcfabrics

[1] Intel acquired certain assets from Cray Inc. related to Cray's high performance computing (HPC) interconnect program, including access to the company's world-class interconnect experts and next-generation Aries interconnect technologies.

[2] Low latency at scale based on preliminary simulations for Intel® Omni-Path Host Fabric Interface (HFI) and switches, which utilize the same connectionless messaging implementation as Intel® True Scale. See the following Intel® True Scale white paper for more details: http://www.intel.com/content/dam/www/public/us/en/documents/white-papers/true-scale-architecture-three-labs-white-paper.pdf. Messaging rate based on Intel simulations using Intel® Xeon® processors populated with Intel® Omni-Path Architecture Adapters, connected with Intel® Omni-Path Architecture Switch products, compared to a comparable configuration utilizing shipping Intel® True Scale Fabric adapters and switches.

[3] Reduction in up to ½ fewer switches claim based on a 1024-node full bisectional bandwidth (FBB) Fat-Tree configuration, using a 48-port switch for Intel Omni-Path cluster and 36-port switch ASIC for either Mellanox or Intel® True Scale clusters.

[4] Internal analysis based on a 256-node to 2048-node clusters configured with Mellanox FDR and EDR InfiniBand products. Mellanox component pricing from www.kernelsoftware.com Prices as of November 3, 2015. Compute node pricing based on Dell PowerEdge R730 server from www.dell.com. Prices as of May 26, 2015. Intel® OPA (x8) utilizes a 2-1 over-subscribed Fabric. Intel® OPA pricing based on estimated reseller pricing using projected Intel MSRP pricing on day of launch. All amounts in US dollars.

[5] Assumes a 750-node cluster, and number of switch chips required is based on a full bisectional bandwidth (FBB) Fat-Tree configuration. Intel® OPA uses one fully-populated 768-port director switch, and Mellanox EDR solution uses a combination of 648-port director switches and 36-port edge switches. Mellanox component pricing from www.kernelsoftware.com, with prices as of November 3, 2015. Compute node pricing based on Dell PowerEdge R730 server from www.dell.com, with prices as of May 26, 2015. Intel® OPA pricing based on estimated reseller pricing based on projected Intel MSRP pricing at time of launch. All amounts in US dollars.

[6] Source: Intel internal estimates based on file size comparison for the standard host Infiniband* software stack based on Intel® Performance Scale Messaging library.

[7] Intel measured data that was calculated from difference between back to back osu_latency test and osu_latency test through one switch hop. All tests performed using Intel ® Xeon® E5-2697v3. Pre-production Intel Corporation Device 24f0 – Series 100 HFI ASIC, Series 100 Edge Switch – 48 port.

[8] All comparisons are based on internal Intel design documentation and product briefs for each product line unless otherwise indicated. Product briefs are available at www.intel.com.

[9] Intel® Omni-Path optimizations for accelerating file system traffic include larger MTU support, 16 SDMA engines for better large-packet parallelization (versus one SDMA in Intel® True Scale Fabric solutions), 160 send/receive contexts for improved mapping across large numbers of CPU cores, improved receive side scaling and interrupt coalescing, automatic header generation.

[10] Each link consists of 4 Lanes. Per Intel® OPA design specifications, if enabled this feature continues to pass data on the remaining 3 links should one fail.

[11] Actual number is 27,628 nodes based on a cluster configured with Intel® Omni-Path Architecture using 48-port switch ASICs, as compared with a 36-port switch chip that can support 11,664 nodes.

Printed in USA 1115/EG,CO/HBD/PDF ♻ Please Recycle 332939-001US