BIOS IT

WHITE PAPER:

INTEL® OMNI-PATH: THE NEXT GENERATION FABRIC

SUMMARY AND PERFORMANCE OVERVIEW

POWERED BY SUPERMICRO intel

www.bios-it.com

INTRODUCTION

Omni-Path is the next generation high performance fabric from Intel®; the successor to the already very successful True Scale Fabric, Omni-Path is making a huge step in closing the gap to the market dominant Mellanox Technologies.

Intel® Omni-Path (OPA) implements a plethora of new technologies to the Technical Computing space with emphasis on High Performance Computing (HPC). Built on the foundations of Intel® True Scale Fabric and additional IP acquired from Cray, Intel® is looking to dominate the HPC arena with a low latency, high bandwidth cost efficient fabric.

OPA is not Infiniband; despite being a primary competitor to the Mellanox EDR technology Intel decided to move away from the Infiniband lock-in to a more functional fabric dedicated to HPC. Infiniband was never originally designed for HPC with adaptation in the early 2000's being slow after numerous setbacks; after which Infiniband was slowly adopted as a clustering interconnect. Intel® took a different approach using a technology called Performance Scaled Messaging (PSM) which optimised the Infiniband stack to work more efficiently at smaller message sizes, which is typically what you associate with HPC workloads; usually MPI traffic. For OPA Intel® have gone a step further, building on the original PSM architecture; Intel® acquired proprietary technology from the Cray Aries interconnect to enhance the capabilities and performance of OPA, these are at both the fabric and host level.

KEY FEATURES OF THE NEW INTEL® OMNI-PATH FABRIC

Some of the new technologies packaged in to the Omni-Path Fabric include:

ADAPTIVE ROUTING

Adaptive Routing monitors the performance of the possible paths between fabric end-points and selects the least congested path to rebalance the packet load. While other technologies also support routing, the implementation is vital. Intel's implementation is based on cooperation between the Fabric Manager and the switch ASICs. The Fabric Manager—with a global view of the topology—initializes the switch ASICs with several egress options per destination, updating these options as the fundamental fabric changes when links are added or removed. Once the switch egress options are set, the Fabric Manager monitors the fabric state, and the switch ASICs dynamically monitor and react to the congestion sensed on individual links. This approach enables Adaptive Routing to scale as fabrics grow larger and more complex.

DISPERSIVE ROUTING

One of the critical roles of fabric management is the initialization and configuration of routes through the fabric between pairs of nodes. Intel® Omni-Path Fabric supports a variety of routing methods, including defining alternate routes that disperse traffic flows for redundancy, performance, and load balancing. Instead of sending all packets from a source to a destination via a single path, Dispersive Routing distributes traffic across multiple paths. Once received, packets are reassembled in their proper order for rapid, efficient processing. By leveraging more of the fabric to deliver maximum communications performance for all jobs, Dispersive Routing promotes optimal fabric efficiency.

TRAFFIC FLOW OPTIMIZATION

Traffic Flow Optimization optimizes the quality of service beyond selecting the priority—based on virtual lane or service level—of messages to be sent on an egress port. At the Intel® Omni-Path Architecture link level, variable length packets are broken up into fixed-sized containers that are in turn packaged into fixed-sized Link Transfer Packets (LTPs) for transmitting over the link. Since packets are broken up into smaller containers, a higher priority container can request a pause and be inserted into the ISL data stream before completing the previous data.

The key benefit is that Traffic Flow Optimization reduces the variation in latency seen through the network by high priority traffic in the presence of lower priority traffic. It addresses a traditional weakness of both Ethernet and Infiniband in which a packet must be transmitted to completion once the link starts even if higher priority packets become available.

PACKET INTEGRITY PROTECTION

Packet Integrity Protection allows for rapid and transparent recovery of transmission errors between a sender and a receiver on an Intel® Omni-Path Architecture link. Given the very high Intel® OPA signalling rate (25.78125G per lane) and the goal of supporting large scale systems of a hundred thousand or more links, transient bit

errors must be tolerated while ensuring that the performance impact is insignificant. Packet Integrity Protection enables recovery of transient errors whether it is between a host and switch or between switches. This eliminates the need for transport level timeouts and end-to-end retries. This is done without the heavy latency penalty associated with alternate error recovery approaches.

DYNAMIC LANE SCALING

Dynamic Lane Scaling allows an operation to continue even if one or more lanes of a 4x link fail, saving the need to restart or go to a previous checkpoint to keep the application running. The job can then run to completion before taking action to resolve the issue. Currently, Infiniband typically drops the whole 4x link if any of its lanes drops, costing time and productivity.

ENHANCED PERFORMANCE SCALED MESSAGING (PSM).

The application view of the fabric is derived heavily from, and has application-level software compatible with, the demonstrated scalability of Intel® True Scale Fabric architecture by leveraging an enhanced next generation version of the Performance Scaled Messaging (PSM) library. Major deployments by the US Department of Energy and others have proven this scalability advantage. PSM is specifically designed for the Message Passing Interface (MPI) and is very lightweight—one-tenth of the user space code—compared to using verbs. This leads to extremely high MPI and Partitioned Global Address Space (PGAS) message rates (short message efficiency) compared to using Infiniband verbs.

UPGRADE PATH TO INTEL® OMNI-PATH

Despite not being truly Infiniband, Intel® have managed to maintain compatibility with their previous generation True Scale Fabric meaning that applications that work well on True Scale can be easily migrated to OPA. OPA integrates support for both True Scale and Infiniband API's ensuring backwards compatibility with previous generation technologies to support any standard HPC application.

INTEL® OMNI-PATH HARDWARE

HOST FABRIC INTERFACE ADAPTERS (HFI'S)

Intel® currently has two offerings on the host fabric interface (HFI) adapter side these include a PCIe x8 58Gbps adapter and a PCIe x16 100Gbps adapter, both of these are single port adapters. Both HFI's use the same silicon so offer the same latency capabilities and features of the high end 100Gbps card.

Along with the physical adapter cards, Supermicro will also be releasing a range of Super Servers with the Omni-Path fabric laid down on the motherboard, this will offer a tighter layer of integration and enable a more compact server design. To take this design even further Intel® have announced that they will be intergrading OPA on to future Intel® Xeon® processors, this will reduce latency further and overall increase performance of all applications.

FIGURE 1: HOST FABRIC INTERFACE ADAPTORS

SOME KEY FEATURES:

- Multi-core scaling – support for up to 160 contexts
- 16 Send DMA engines (M2IO usage)
- Efficiency – large MTU support (4 KB, 8 KB, and 10KB) for reduced per-packet processing overheads. Improved packet-level interfaces to improve utilization of on-chip resources
- Receive DMA engine arrival notification
- Each HFI can map ~128 GB window at 64 byte granularity
- Up to 8 virtual lanes for differentiated QoS
- ASIC designed to scale up to 160M messages/second and 300M bidirectional messages/second

INTEL® OMNI-PATH EDGE AND DIRECTOR CLASS SWITCH 100 SERIES

The all new Edge and Director switches for Omni-Path from Intel® offer a totally different design from traditional Infiniband switches. Incorporating a new ASIC and custom front panel layout, Intel® have been able to offer up to 48 Ports at 100Gbps from a single 1U switch, this is 12 ports higher than its nearest competitor. The higher switching density allows for some significant improvements within the data centre, some include:

	Intel® Omni-Path Host Fabric Adapter 100 Series 1 Port PCIe x16	Intel® Omni-Path Host Fabric Adapter 100 Series 1 Port PCIe x8
Adapter Type	Low Profile PCIe Card (PCIe x16)	Low Profile PCIe Card (PCIe x8)
Ports	Single	Single
Connector	QSFP28	QSFP28
Link Speed	100Gb/s	~58Gb/s on 100Gb/s Link
Power (Typ./Max) – – Copper – Optical	7.4/11.7W (Copper) 10.6/14.9W (Optical)	6.3/8.3W (Copper) 9.5/11.5W (Optical)
Thermal/Temp	Passive (55° C @ 200 LFM)	Passive (55° C @ 200 LFM)

TABLE 1: INTEL® OMNI-PATH HOST FABRIC ADAPTER 100 SERIES 1 PORT PCIE X16 AND PCIE X 8

- Reduced switching cost due to needing less physical switching (over 30% reduction in switches for most configurations)
- Lower amount of fabric hops for reduced latency
- 100–110ns switch latency
- Support for fabric partitioning
- Support for both active and passive cabling
- Higher node count fabric: support for up to 27,648 nodes in a single fabric that is up by nearly 2.3x of traditional Infiniband.

Intel's Director switch range offers a very similar feature set to the Edge switches with various chassis options as you may expect. Currently there are 20U and 7U variants available supporting various Spine and Leaf modules.

NTEL® OMNI-PATH SOFTWARE COMPONENTS

Intel® Omni-Path Architecture software comprises the Intel® OPA Host Software Stack and the Intel® Fabric Suite.

INTEL® OPA HOST SOFTWARE

Intel's host software strategy is to utilize the existing OpenFabrics Alliance interfaces, thus ensuring that today's application software written to those interfaces run with Intel® OPA with no code changes required. This immediately enables an ecosystem of applications to "just work."

All of the Intel® Omni-Path host software is being open sourced.

	Intel® Omni-Path Edge Switch 100 Series: 48 Port	Intel® Omni-Path Edge Switch 100 Series: 24 Port
Ports	48 up to 100Gbps	24 up to 100Gbps
Rack Space	1U (1.75")	1U (1.75")
Capacity	9.6Tb/s	4.8Tb/s
Port Speed	100Gb/s	~58Gb/s
Power (Typ./Max) – – Input 100–240 VAC 50–60Hz – Optical	189/238 W (Copper) 356/408 W (Optical)	146/179 W (Copper) 231/264 W (Optical)
Interface	QSFP28	QSFP28
Fans & Airflow	N+1 (Speed Control) Forward/Reverse	N+1 (Speed Control) Forward/Reverse)

TABLE 2: INTEL® OMNI-PATH EDGE SWITCH 100 SERIES: 48 PORT AND 24 PORT

	Intel® Omni-Path Director Class Switch 100 Series: 24 Slot	Intel® Omni-Path Director Class Switch 100 Series: 6 Slot
Ports	48 up to 100Gbps	24 up to 100Gbps
Rack Space	1U (1.75")	1U (1.75")
Capacity	9.6Tb/s	4.8Tb/s
Management Modules	1/2	1/2
Leaf Modules (32 Ports)	Up to 24	Up to 6
Spine Modules	Up to 8	Up to 3
Power (Typ./Max) – – Input 100-240 VAC 50-60Hz – Optical	6.8/8.9 KW (Copper) 9.4/11.6 KW (Optical)	1.8/2.3 KW (Copper) 2.4/3.0 KW (Optical)
Interface	QSFP28	QSFP28
Fans & Airflow	N+1 (Speed Control) Forward/Reverse	N+1 (Speed Control) Forward/Reverse)

TABLE 3:INTEL® OMNI-PATH DIRECTOR CLASS SWITCH 100 SERIES: 24 SLOT AND 6 SLOT

As with previous PSM generations, PSM provides a fast data path with an HPC-optimized lightweight software (SW) driver layer. In addition, standard I/O-focused protocols are supported via the standard verbs layer.

INTEL® FABRIC SUITE

Provides comprehensive control of administrative functions using a mature Subnet Manager. With advanced routing algorithms, powerful diagnostic tools and full subnet manager failover, the Fabric Manager simplifies subnet, fabric, and individual component management, easing the deployment and optimization of large fabrics.

INTEL® FABRIC MANAGER GUI

Provides an intuitive, scalable dashboard and analysis tools for viewing and monitoring fabric status and configuration. The GUI may be run on a Linux or Windows desktop/laptop system with TCP/IP connectivity to the Fabric Manager.

PERFORMANCE SUMMARY:

INTEL® OMNI-PATH FABRIC VS. MELLANOX EDR INFINIBAND

EXECUTIVE SUMMARY

SYNTHETIC BENCHMARKS: OSU MICRO BENCHMARKS

For the synthetic testing we are going to compare two currently very popular Infiniband derivatives, Mellanox FDR and Mellanox EDR against the new Intel® Omni-Path fabric. For the testing we are going to use Ohio State University (OSU) Micro Benchmarks version 5.1, to determine bandwidth, message rate and latency of each of the interconnects – ultimately this should provide us with a good understanding of which one is best performing. http://mvapich.cse.ohio-state.edu/benchmarks/

"osu_latency" – in figure 3 you can see a graph outlining the results from the test:

It is clear that the initial tests of the interconnect show Intel® have significantly reduced the latency of its new Omni-Path Interconnect for the smaller message sizes; excluding one anomaly in the data, Omni-Path also provides significantly lower latency with the larger message sizes.

FIGURE 2: INTEL FABRIC MANAGER GUI

Moving to the other end of the spectrum, the same story as before is clear, Omni-Path is still providing significantly lower latency with the larger messages sizes also, despite using the PSM technology which focuses on the smaller MPI type traffic.

It is clear that at least for latency bound applications, or applications which have a large number of Collective MPI calls will significantly benefit from suing Omni-Path interconnect.

"osu_mbw_mr" – in figure 5 you can see a graph outlining the results from the test:

Message rates for the smaller message sizes tell a different story to latency, it is clear here that from all of the data points for Omni-Path, excluding a few early points in the data; that Infiniband wins this battle. However looking at further data and some of the other tests, it looks like the designers of Omni-Path have purposely take a hit with small message bandwidth and message rates to get the latency down, and within typical HPC latency is usually what matters.

The larger message sizes show a differing set of results, you can see by the end of the small message size graph that by 32,768 Byte messages Omni-Path was on its way back, this has then continued for the

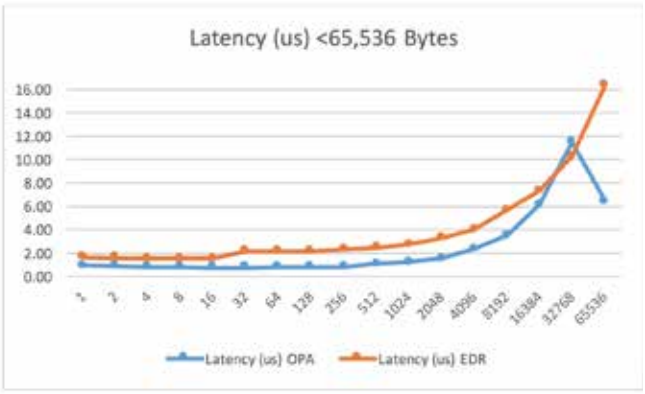

FIGURE 3

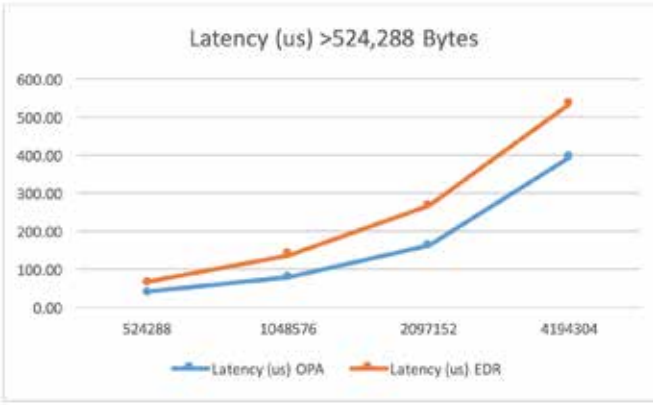

FIGURE 4

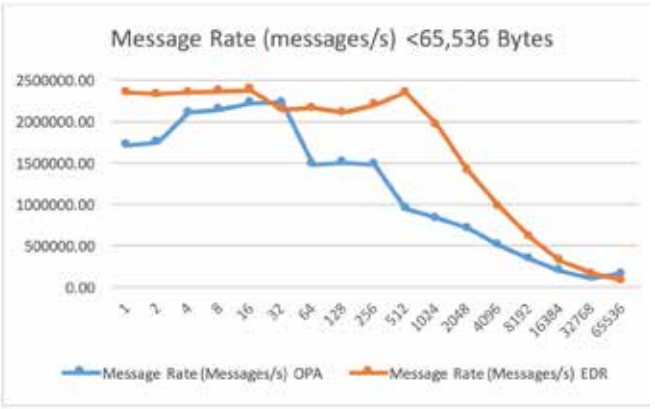

FIGURE 5

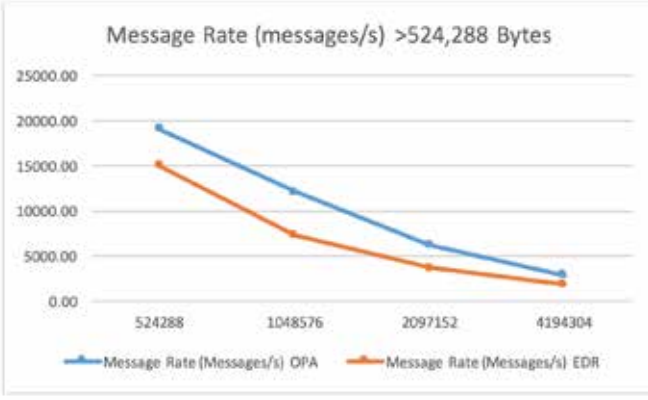

FIGURE 6

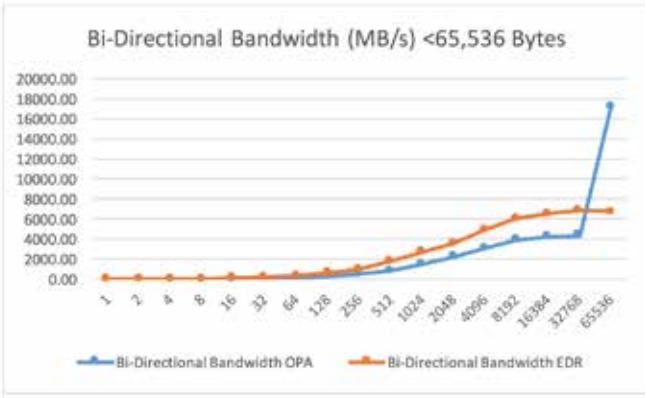

FIGURE 7

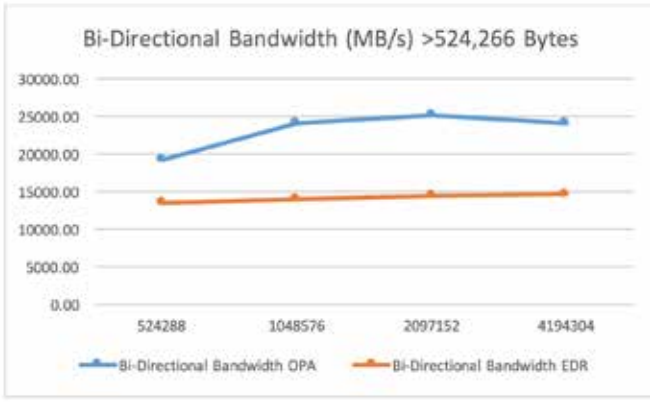

FIGURE 8

larger message sizes, clearly showing Omni-Path with a higher overall message rate at the top end of nearly 4MB messages.

"osu_bibw" & "osu_bw" – in figure 7 you can see a graph outlining the results from the test:

As discussed previously it looks like Intel® have purposely taken a hit on the small message bandwidth and message size, this is made clear by the graph above. At 32,768 Byte message size there is a huge spike in bandwidth, if you check the latency graph, you can see at the same 32,768 Byte message size the latency goes up by 100% compared to the previous 16,384 Byte message. It is clear that this is where the 'PSM sweet spot' ends as the bandwidth and message rate both become significantly more competitive by this point.

As per the final two points on the small message graph the large message graph tells pretty much the same story, with consistently higher bi-directional bandwidth than the comparable EDR solution, despite both using technology with the same 100Gbps bandwidth.

SUMMARY AND SOLUTIONS OVERVIEW

In summary, the Omni-Path Architecture (OPA) fabric solutions provide a competitive alternative to the existing 100GB Infiniband solutions in the market today with performance on par with and in some cases, better. BIOS-IT has integrated OPA products in to its HPC range with full turnkey solutions which are available today. From the HCA (58Gb / 100Gb) to switches (24/48 Port) and all the software and integration required in between, BIOS IT's solutions ensure the best price/performance on the market for your HPC applications . All our OPA solutions are also certified as 'Intel Cluster Ready'. Remote access for testing and evaluation is also available through BIOS-IT labs.

BIOS IT'S SOLUTIONS ENSURE THE BEST PRICE & PERFORMANCE ON THE MARKET FOR YOUR HPC APPLICATIONS

ABOUT BIOS IT

BIOS IT delivers global first-to-market technology together with High Performance Computing products and techniques, previously exclusive to academia and scientific research, into the real world. With a number of key hardware and software partners, BIOS IT are able to design and develop unique and manageable compute and storage clusters with industry leading value/performance ratios.

Privately held since inception, we have grown from humble beginnings to become a global leader in enterprise information technology with over 20 years' experience. Although during this period technology architectures have evolved, our mantra for delivering high quality, first to market products and services has always been the same. This set of core values has allowed us to grow organically to a turnover of over $60million with a foot hold in the world's leading economies.

As a dedicated division, BIOS IT has then taken this innovation a step further to enable constant investment in new technologies and subsequently allowed us to design and manufacture our own HPC systems, the first of which was the micro-server Viridis platform, the world's first ARM server for the enterprise. This revolutionary architecture has enabled us to deliver supercomputing performance from as little as 5W per server, paving the way to exascale computing.

CONTACT US

www.bios-it.com | sales@bios-it.com
AMERICAS: 1-800-654-BIOS
EMEA: +44 (0) 203 178 6467
APAC:: +61(0) 2 9959 1010

BIOS IT