WHITE PAPER Intel[®] True Scale Fabric Architecture HPC



Intel® True Scale Fabric Architecture: Enhanced HPC Architecture and Performance

Improved interconnect operability increases scalable performance for today's HPC clusters.

KEY FINDINGS

- InfiniBand* Architectures There are two types of InfiniBand architectures available today in the marketplace, the first being the traditional InfiniBand design, created as a channel interconnect for the data center. The latest InfiniBand architecture was built with HPC in mind. This enhanced HPC fabric offering is optimized for key interconnect performance factors, featuring MPI message rating, end-to-end latency and collective performance, resulting in increased HPC application performance.
- Enhanced Intel[®] True Scale Fabric Architecture Offers 3x to 17x the MPI (Message Passing Interface) message throughput of the other InfiniBand architecture. For many MPI applications, small message rate throughput is an important factor that contributes to overall performance and scalability.
- Improved End-to-End Latency End-to-end latency is another key determinant of an MPI application's performance and ability to scale. The Intel True Scale Fabric end-to-end latency is 50 percent to 90 percent lower at 16 nodes than the traditional InfiniBand offering available today.
- Increased Collective Performance Critical for an MPI application's performance and ability to scale. Intel True Scale architecture makes it possible to achieve significant collective performance at scale, without hardware based collective acceleration, resulting in 30 percent to 80 percent better collective performance for the three major collectives: Allreduce, Barrier, and Broadcast.
- Faster Application Performance Intel tested a number of MPI applications and found that they performed up to 11 percent better on the cluster based Intel True Scale Fabric QDR-40 than the traditional InfiniBand-based architecture running at FDR (56 Gbps).

TABLE OF CONTENTS

Key Findings 1
Executive Summary 2
Intel True Scale Fabric InfiniBand*- Based Architecture
MPI Message Rate Performance3
End-To-End Latency Performance4
Collective Performance5
Application Performance 5
Spec MPI20075
Conclusion 7
Appendix 1: Tested Configuration Information
Appendix 2: Disclaimers & Risk Factors 8
Legal Disclaimers8
Optimization Disclaimer8
Risk Factors8

EXECUTIVE SUMMARY

Today's high performance computing (HPC) clusters often take advantage of increased node counts, with each node utilizing faster, denser core count processors. With this advancement, scalable performance is a critical component of application optimization on larger, faster clusters. The interconnect is one of the key factors that influences overall performance of the HPC cluster at scale, accounting for up to 30 percent of the cost of an HPC cluster, making price an important consideration.

The Intel True Scale Fabric architecture was designed from the ground up for HPC, which means it offers improved HPC performance, at a competitive price point, especially for implementations requiring superior performance across large node counts. The following are key elements of the Intel True Scale Fabric Architecture.

- Improved On-Load Design Designed around a host on-load implementation for the host channel adapter (HCA). This implementation provides low endto-end latency, that stays low as a cluster is scaled. The reason for this is that the on-load design takes advantage of improvements in processor performance and maximizes the performance of clusters built using faster and higher core count processors. An on-load implementation takes full benefit of Moore's law by leveraging the increase in CPU performance resulting from high core count architectures.
- Increased Performance Scaled Messaging – Optimized interface library layer between the upper layer protocol, like MPI (Message Passing Interface), and the InfiniBand* driver. This library, called PSM (Performance Scaled Messaging), is lightweight in design and provides optimized performance capabilities for:
- -MPI Message Rate: Extremely high message rate throughput, especially with small message sizes

- -Latency: End-to-end latency that remains low, even at scale
- -Collective: Very low latency across all collective algorithms, even at scale
- Connectionless: This approach provides for low end-to-end latency, even at scale, thereby offering excellent scaling across large node/core count HPC clusters.

INTEL TRUE SCALE FABRIC INFINIBAND*- BASED ARCHITECTURE

There are two types of InfiniBand architectures available today in the marketplace: the traditional InfiniBandbased architecture, designed as a channel interconnect for the enterprise data center which features an offload host adapter and Verbs-based designs. The Intel True Scale Fabric is an HPC Enhanced version of InfiniBand, designed when it became clear that HPC was to be the major market for InfiniBand-based fabrics. Intel True Scale Fabric was purpose built to run HPC/MPI applications and take full advantage of today's latest processor technology, with its dense multi-core applications.

The two generations of InfiniBand architectures handle protocol processing very differently, with the Intel True Scale Fabric architecture based on a connectionless design. This approach does not establish connection address information between node/cores/process that is maintained in the cache of the adapter. The traditional InfiniBand implementation utilizes an offload implementation with a fairly heavyweight protocol control library called Verbs. Unlike the traditional InfiniBand-based architecture, with its offload/Verbs implementation where addressing/state information is kept in the cache of the host adapter, Intel True Scale Fabric's connectionless design does not have the potential for a cache miss on connection state as the HPC cluster is scaled. In offload/Verbs based implementations, when cache misses occur, address



Figure 1. Message Rate Profile for Small/Communication Dependent and Large/Processor Dependent Models

information must be obtained from main memory across the PCle* bus, significantly impacting performance as applications are scaled across a large cluster. The Intel True Scale architecture eliminates the potential for address cache misses, by utilizing a semantic tag matching approach for MPI messages. This implementation offers greater potential to scale performance across a large node/core count cluster, while maintaining low end-to-end latency as the application is scaled across the cluster.

The Intel True Scale Fabric innovative host design utilizes an HPC optimized library called PSM (Performance Scaled Messaging) for MPI communications. PSM is a "lightweight" library that is specifically built to optimize MPI performance requirements. PSM is built around semantic tag matching similar in concept to those used by high performance HPC interconnect pioneers Myricom* and Quadrics.* Intel True Scale Fabric's PSM divides the responsibilities between the host driver and the Host Channel Adapter differently than traditional Verbs-based implementations. In the PSM implementation, the host driver directly executes the InfiniBand transport layer, entirely eliminating both

the heavyweight Verbs interface on the host and any transport-layer bottlenecks in the Host Channel Adapter offload processor/micro-sequencer. This makes PSM, with its on-load approach, well-suited to take advantage of today's high-performance, dense multi-core processors.

The key measures of HPC performance are MPI message rate performance, end-to-end latency, collective performance and application performance. Tests in these areas show that Intel True Scale Fabric's on-load based InfiniBand architecture, with PSM implementation, is better at scaling, message processing, and latency when compared to the more traditional InfiniBand-based architecture.

MPI Message Rate Performance

For most HPC applications, MPI message throughput is the key factor that determines overall application performance and scaling. As an MPI application is scaled its message rate increases at a faster pace; this is especially true with small messages. The Intel True Scale Fabric architecture offers significantly higher message throughput vs. the traditional InfiniBand-based offerings. The graphs in Figure 1 are excellent examples of the increase in MPI message rate traffic as an application is scaled across a cluster. The HPC Zone for these two models is where 98 percent of messages occur. The interconnect performance within the HPC zone is key to overall application performance. The graph on the left has an HPC zone where 98 percent of the messages are 4K bytes or less. The HPC zone for the model on the right shows it takes up to 65K byte sized messages to reach the 98 percent mark.

It is important to note that for both models there is a significant increase in message rate for the 64 byte messages as a cluster when scaled from 8 to 16 to 32 nodes. The increase in MPI messages from 16 to 32 nodes for the Eddy_417K 64 byte messages is over 250 percent, which means the 64 byte messages now account for over 90 percent of all messages. For the Truck_111m model, the MPI message rate increase is 235 percent when going from 16 to 32 nodes and the 64 byte messages account for 64 percent of all traffic. The interconnect's ability to efficiently handle very small messages in volume is a key factor in determining application performance at scale.



Figure 2. Message Rate of Offload/Verbs vs. On-load/PSM using MVAPICH

The definitive test for measuring host rate message throughput is Ohio State University's (OSU's) MPI Message Rate test. The message rate test evaluates the aggregate unidirectional message rate between multiple pairs of processes. Each of the sending processes sends a fixed number of messages back-to-back to the paired receiving process before waiting for a reply from the receiver. This process is repeated for several iterations.

The objective of this benchmark is to determine the achieved message rate from one node to another with a configurable number of processes running on each node.

Note: This is a test and the results in Figure 2 are based on non-coalesced message rate performance. As message rate has become more recognized as an important indication of HPC performance, technology providers want to portray their products in the best possible way. Coalescing artificially increases the overall message rate, but it requires sending one stream of messages to only one other process, which is not typical of MPI interprocess messaging patterns. In addition, coalescing adds latency to the transaction because the sending process must wait and decide whether to send the packet of messages to as is or wait for other messages to add to the packet.

Figure 2 illustrates that the traditional InfiniBand-based offload/Verbs architecture "tops out" at roughly 10 million messages per second. More significantly, the performance of the offload/Verbs solution actually declines as the number of processor cores moves beyond four. In contrast, the HPC enhanced Intel True Scale Fabric with its on-load/PSM architecture offers up to 17x more message throughput at 16 cores than the offload/Verbs architecture.

Key Findings:

- Host-based adapters achieve significantly more messages per second at scale
- Offload/Verbs implementation performance peaked at four cores
- Intel True Scale Fabric QDR-80 provides near-linear scaling ~60M messages per second

End-To-End Latency Performance

Latency, especially end-to-end latency, is another key factor of an HPC application's performance and ability to scale. The Intel True Scale Fabric's enhanced HPC architecture provides for low end-toend latency that remains low as an application is scaled across an HPC cluster. There are several ways to measure latency, the easiest being a two node test. Figure 3 (on next page) shows the latency for the two different InfiniBand implementations using this simple node-tonode test with the OSU Latency test.

As Figure 3 shows, the two different InfiniBand-based architectures have similar latency to one another in this simple test. The question is what do latencies look like with a set of more realistic tests at scale?

HPCC (HPC Challenge) has a set of latency tests that are more representative of HPC/ MPI latency at scale. The latency tests used in this study determine end-to-end latency, which is a function of the Infini-Band adapter and the host InfiniBand stack and switch. The following tests were used to determine and analyze the performance of the InfiniBand architectures:

- Maximum Ping-Pong Latency reports the maximum latency for a number of non-simultaneous ping-pong tests. The ping-pongs are performed between as many distinct pairs of processors as possible.
- Naturally Ordered Ring Latency reports latency achieved in the ring communication pattern.
- Randomly Ordered Ring Latency reports latency in the ring communication pattern. The communication processes are ordered randomly in the ring.

Figure 4 (on next page) summarizes the results of HPCC latency Ping-Pong, NOR and ROR tests at 16 nodes. The fourth set of bars is an average of the three tests. In each of the tests, the Intel True Scale InfiniBand architecture achieved significantly lower latency than its counterpart. The Randomly Ordered Ring Latency test showed the most performance difference; Intel True Scale Fabric had a 5x latency advantage. The Intel True Scale Fabric average latency is over 70 percent lower than traditional InfiniBand architecture running at FDR speed.



Figure 3. Two Node Latency Test with OpenMPI



Figure 4. HPCC Latency Tests using OpenMPI

Key Findings:

- Latency is a key factor impacting the performance of most MPI applications
- The Intel True Scale Fabric design provides lower latency at QDR versus the traditional InfiniBand designed offering at FDR speed
- Intel True Scale Fabric has a 20 percent to 82 percent latency advantage depending on the test
- Average latency advantage for Intel True Scale Fabric is 72 percent

Collective Performance

A collective operation is a concept in parallel computing in which data is simultaneously sent to, or received from, many nodes. Collective functions in the MPI API involve communication between all processes in a particular group (which can mean the entire process pool or a program-defined subset). These types of calls are often useful at the beginning or end of a large distributed calculation, where each processor operates on a part of the data and then combines it into a result. The performance of collective communication operations is known to have a significant impact on the scalability of most MPI applications. The nature of collectives means that they can become a bottleneck when scaling to thousands of ranks (where a rank is an MPI process, typically running on a single core).

Collective performance is critical for the ability to scale the performance of an MPI application, especially on an HPC cluster. It is possible to achieve significantly improved collective performance at scale without hardware based collective acceleration. The Intel True Scale Fabric InfiniBand architecture is highly optimized for the HPC marketplace. Because of this focused design Intel True Scale Fabric does not require special or retrofitted collective acceleration hardware or software to achieve collective performance at scale.

Three of the most widely used collectives are AllReduce, Barrier, and Broadcast.

As shown in Figure 5, Intel True Scale Fabric shows excellent performance across the above set of collectives, especially in the HPC Zone where most of the HPC/MPI traffic occurs.

Key Findings:

- Performance of collective operations has an impact on the overall performance and scalability of applications
- Intel True Scale Fabric architecture shows excellent collective performance across the key collective operations—AllReduce, Barrier, and Broadcast—particularly in message sizes that would be within the HPC Zone

APPLICATION PERFORMANCE Spec MPI2007

Spec MPI2007 is a benchmark suite for evaluating MPI-parallel, floating point, and compute intensive performance across a wide range of cluster implementations. MPI2007 is designed to measure and comparing high-performance computer



Figure 5. Collective Performance using OpenMPI

systems and clusters. The benchmark programs shown in Figure 6 are developed from native MPI-parallel end-user applications, as opposed to synthetic benchmarks or even parallelized versions of sequential benchmarks. (http://www.spec.org/mpi).

The Intel True Scale Fabric, with its enhanced HPC architecture, shows excellent performance across the Spec MPI2007 suite of applications when compared to the more traditional InfiniBand implementation. The percentages (above each of the application tests in Figure 7) illustrate the performance differential of Intel True Scale Fabric on-load/PSM architecture to the traditional offload/Verbs. The first percentage is Intel QDR-40 and the second percentage is QDR-80; where blue represents better performance for Intel True Scale Fabric on-load/PSM architecture. In summary, Intel QDR-40 shows better performance in 7 out of 12 tests and QDR-80 has better performance in 10 out of 12 tests.

Key Findings:

- Intel True Scale Fabric QDR-40 shows on average an 11 percent performance advantage
- The QDR-80 average performance advantage is 18 percent

BENCHMARK	APPLICATION DOMAIN	SUITE	LANGUAGE
milc	Physics: Quantum Chromodynamics (QCD)	medium	С
leslie3d	Computational Fluid Dynamics (CFD)	medium	Fortran
GemsFDTD	Computational Electromagnetics (CEM)	medium	Fortran
fds4	Computational Fluid Dynamics (CFD)	medium	C/Fortran
рор2	Ocean Modeling	medium, large	C/Fortran
tachyon	Graphics: Parallel Ray Tracing	medium, large	С
lammps	Molecular Dynamics Simulation	medium, large	C++
wrf2	Weather Prediction	medium	C/Fortran
GAPgeofem	Heat Transfer using Finite Element Methods (FEM)	medium, large	C/Fortran
tera_tf	3D Eulerian Hydrodynamics	medium, large	Fortran
zeusmp2	Physics: Computational Fluid Dynamics (CFD)	medium, large	C/Fortran
lu	Computational Fluid Dynamics (CFD)	medium, large	Fortran

Figure 6: Spec MPI2007 Benchmark Test List

The Spec MPI2007 test results shown in Figure 7 (on next page) compared the performance of an HPC cluster environment where all the components were kept the same with the exception that the interconnect was varied between the two major InfiniBand implementations.



Figure 7: Spec MPI2007 Benchmark Test Results using Open/MPI

CONCLUSION

The interconnect architecture has a significant impact on the performance of a cluster and the applications running on the cluster. Intel True Scale Fabric host and switch technologies provide an interconnect infrastructure that maximizes an HPC cluster's overall performance. The Intel True Scale Fabric Architecture, with its onload protocol processing engine, connectionless implementation, and lightweight semantic-based PSM interface, provides an optimized environment that maximizes MPI application performance. With the use and size of HPC clusters expanding at a rapid pace, Intel True Scale Fabric InfiniBand architecture and technology extracts the most out of your investment in compute resources by eliminating adapter and switch bottlenecks.

APPENDIX 1: TESTED CONFIGURATION INFORMATION On-Load/PSM Configuration

Location: HPC Lab, Intel, Swindon, UK 16 nodes.

Servers: Each:

- 2x Intel[®] Xeon[®] Processors E5-2670
- Processor speed 2.60 GHz
- Memory 32 GB 1666 MHz DDR3

CPU Setting: TURBO

Interconnect: QDR-40 & QDR-80 Intel True Scale Fabric (2xQLE7340), 1x12300 Intel True Scale Fabric 36 port switch.

IB Switch F/W: 7.0.1.0.43

OS: RHEL6.2 Kernel - 2.6.32-220.el6. x86_64

IB Stack: IFS 7.1.0.0.55 with ib_qib from PR 120677 build qib-qofed-1.5.4.1_120677

Compiler: gcc + Intel CC Version 12.1.3.293 Build 20120212

Math Library: Intel® MKL

MPI: Various as noted in each test

Testing Methodology: Out-of-Box Testing

Offload/Verbs Configuration

Location: HPC Lab, Intel, Swindon, UK 16 nodes.

Servers: Each:

- 2x Xeon Processors E5 2680
- Processor speed 2.70 GHz
- Memory 32 GB 1666 MHz DDR3

CPU Setting: TURBO

Interconnect: Single Rail Mellanox FDR MT4099 Dual Port (MCX354A-FCBT), 1 x SX6036 Mellanox FDR 36 port switch

IB Switch F/W: 2.10.600

OS: RHEL6.2 Kernel - 2.6.32-220.el6. x86_64

IB Stack: mlnx-ofa_kernel-1.5.3-OFED.1.5.3.3.0.0 (options mlx4_core log_ num_mtt=21 og_mtts_per_seg=7)

Compiler: gcc + Intel CC Version 12.1.3.293 Build 20120212

Math Library: Intel® MKL

MPI: Various as noted in each test

Testing Methodology: Out-of-Box Testing

APPENDIX 2: DISCLAIMERS & RISK FACTORS

Legal Disclaimers

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

A "Mission Critical Application" is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL'S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS' FEES ARISING OUT OF, DIRECTLY, OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENT IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Intel product plans in this presentation do not constitute Intel plan of record product roadmaps. Please contact your Intel representative to obtain Intel's current plan of record product roadmaps.

Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families. Go to: http://www.intel.com/products/processor_number.

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or go to: http://www.intel.com/design/literature.htm

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark* and MobileMark,* are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to: http://www.intel.com/performance

Sandy Bridge and other code names featured are used internally within Intel to identify products that are in development and not yet publicly announced for release. Customers, licensees and other third parties are not authorized by Intel to use code names in advertising, promotion or marketing of any product or services and any such use of Intel's internal code names is at the sole risk of the user

Optimization Disclaimer

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice revision #20110804

Risk Factors

The above statements and any others in this document that refer to plans and expectations for the second quarter, the year and the future are forward-looking statements that refer to or are based on projections, uncertain events or assumptions also identify forward-looking statements. Many factors could affect Intel's actual results and dirariances from Intel's current expectations regarding such factors could actual exalt Intel's actual results, and viranices from Intel's current expectations regarding such factors could actual exalt Intel's actual results and dirariances from Intel's current expectations. Demand could be different from Intel's expectations due to factors including changes in the level of inventory at customers. Uncertainty in global economic and filtons poses a risk that consumers and businesses may defer purchases in response to negative financial econditions poses are instepsic to exist that are characterized by a high percentage of costs that are fixed or difficult to reduce in the short term and product demand that is highly variable and difficult to forecast. Revenue and the gross marging percentage are affected by the timing of Intel's ability for sequention of products; actions taken by Intel's competitors, including product offerings and introductions, marketing and introductions poses are interestive. Since activity in global excession and timing issues associated with these changes, including products fers ale concess, franking and introductions, including variations related to the timing of qualifying product financial conditions poses are appresented to excess or obsolete inventory valuation, including variations related and eventage are resources; product manufacturing aleids to the sponse on equitive valuation is the product introductors, market acceptance of Intel's products; actions taken by Intel's competitors, including product offerings and introductors, market acceptance of Intel's products; actions taken by Intel's competitors, including product fires and comperise in the

