Executive Summary

High-Performance computing starts with the “Scale-Up” paradigm to get the most power into a single node. Over time we have seen the gigahertz race (Clock speed), a new more efficient internal bus architecture (PCIe), faster and faster memory, the addition of processor sockets and additional processing cores to each processor and the proliferation of 64-bit applications. While improvements will continue in these areas, it is the advent of the coprocessor first brings the ability to have a terraFLOP of processing power in a single server utilizing Intel® Xeon® processors.

The Intel® Xeon Phi™ coprocessor is designed to help provide the highest level of power and flexibility in conjunction with Intel Xeon processor-based systems. Application(s) can be run on the Intel Xeon processors with highly-parallel code segments off-loaded to the Intel Xeon Phi coprocessor to accelerate performance. The Xeon Phi coprocessor has up to 60 processor cores running at gigahertz speeds so the application can perform more quickly and efficiently.

While the overall compute power in a single node is important, to further scale performance, these nodes must be interconnected by a highly efficient network infrastructure. This paper will lay out the Intel® Xeon Phi™ coprocessor HW and SW environment and compare performance data on the Verbs/Offload InfiniBand* architecture to the the Intel® True Scale Fabric with using the PSM/On-Load Architecture. To obtain best utilization from the a server utilizing the Intel® Xeon® processor and the Intel® Xeon Phi™ coprocessor, the Intel® Coprocessor Communications Link (Intel® CCL) SW technology is required. There are 2 versions, an early version was released that supports the Verbs/Offload architecture. A succeeding release has been optimized for the Intel® True Scale Architecture PSM/On-Load-based HCA. This Intel® True Scale Architecture HPC networking infrastructure enables a clustered application to approach its maximum possible efficiency.

Key Point - Architecture Matters:

- The True Scale architecture has been designed and optimized from the ground up for HPC. The latest Intel® CCL release, optimized for True Scale HPC, brings high-performance advantages to servers including the Intel® Xeon Phi™.
The Verbs/Offload architecture was originally designed for general purpose InfiniBand connectivity. It was designed to support “all” connectivity, not for any specific use. The initial version of Intel® CCL was based on Verbs/Offload.

Performance matters. The data highlights the difference.

**Intel® Xeon Phi™ Coprocessors**
(Complimentary Solutions for Parallel Workloads)

Today Xeon is the Right Solution for Most Workloads:
The vast majority of application workloads are and A class of applications exist that are referred to by many names including explicitly and highly parallel Intel® Xeon Phi™ products deliver leadership performance for these applications through greater compute density on a single processor die

The unique advantage of Intel® MIC architecture is the ability for customers to program it using the same tools, programming models, algorithms, and even source code that are used to build applications for Xeon maximizing the return on their programming investments. Algorithmic improvements continue to extract more parallelism in software applications which are built on Messaging, Threading, and Data Parallel constructs. As shown in the above figure, as the applications get more parallel, the more the Intel® Xeon Phi™ will aid in increasing the performance curve.

Intel continues to invest to expand the software tools supporting the Intel Xeon Phi architecture more transparently for the user.

**Code Execution Models:**
There is a set of execution models that support the Intel® Xeon® and Intel® Xeon PHI™ architecture. “Execution models” are largely determined by:

- The invocation location of the main program(s)
- Where serial and parallel portions of an app run

For Systems with the Intel® Xeon Phi™, Execution models are aided in efficiency by the Intel® Coprocessor Communications Link (Intel® CCL) SW technology. Intel CCL provides the optimized internal infrastructure that enables efficient intra-node communication and provides direct communications from the coprocessor to the HCA for all MPI processes. The early Intel CCL release was to support the general Verbs/Offload paradigm with the Intel® Xeon Phi™. Subsequently Intel® released a version that has for building optimized fabrics for the Intel® Xeon Phi™ around the Intel® True Scale Fabric using the
PSM/On-Load-based HCA. PSM has already been optimized for internode MPI communications. The Intel® True Scale Architecture with the Intel CCL has been specifically optimized to support high performance HPC-MPI applications on True Scale Fabrics.

**Intel® Xeon Phi™ Execution Model Overview**

**Many-Core Hosted Model**
It is best suited for highly-parallel codes. The ACN (autonomous compute node) machine model with Intel’s software stack and tools make this possible.

**Offload Models**
In offload models (forward and reverse), the main program executes on Xeon or MIC, respectively, but certain sections of code are run remotely. In the case of forward offload, highly-parallel work is sent from processor to the Co-processor. In reverse offload, highly-serial work is sent from MIC to main CPU.

**Symmetric Model**
In a symmetric model, main programs are run on both Intel® Xeon Phi™ and the Intel® Xeon® processor with the user explicitly balancing the work. MPI would be a typical means to use symmetric.

**The Intel® True Scale Architecture Overview**
The HPC interconnect that provides the high-performance network is InfiniBand-based. The Intel® True Scale Architecture provides high performance end to end connectivity.

The main item that greatly affects performance is the Host stack implementation. The Intel True Scale Architecture was engineered from the ground up to explicitly provide high performance in HPC fabrics even at scale.

Performance Scaled Messaging (PSM) is specifically designed and matched to support the MPI communications method. It is extremely lightweight comprising less than 10% of the lines of code that the “Verbs/Offload Driver” used in the other implementation thereby reducing the processor load on the host. The connectionless method of the True Scale Architecture further reduces network and processor load. Even with the On-Load model far less CPU is required for communications. This architecture leaves no chance of address cache misses as the fabric grows unlike the Verbs/Offload-based implementation. The On-Load HCA process allows a more streamlined HCA design. Connectionless communications has a significantly smaller User Space
footprint that the Verbs-based implementation. This architecture has been extended to support the Intel® Xeon Phi™.

There are key factors that will determine high performance for Intel and other processors. This paper focuses on the Xeon® Phi Native & Hybrid modes. The criteria is shown below.

**Intel® Xeon® Processors/ Intel® Xeon Phi™ Coprocessor Performance Factors**

- High MPI Message Rate Performance, especially with small message sizes
- Low End-to-End Latency, even at scale
- Excellent MPI Collectives Performance
- Effective Bandwidth, capable of meeting the requirements of each HPC application
- Efficient Memory Utilization

---

**Execution Model Overview**

The execution models below are the focus of this paper.

**Native Mode:**

In the Native mode, there are separate applications in this Host. One is being run on the Xeon processor and the other on the Xeon Phi Coprocessor. These applications are independent. It is expected that Serial, Parallel or Vector programming functions will be located in the processor that best benefits the data parameters. Instead of all data results traversing the system to the main Xeon processor, each processor would independently communicate its own Function Step result data out to the cluster. The Xeon Phi Coprocessor would utilize the Intel® CCL if any intra-node communication is required.

**Native/Hybrid Mode:**

![Diagram of Native/Hybrid Mode](image)

---

4
OpenMP* is an API for intra-node multi-threaded, shared memory parallelism. Its key goals are to simplify parallel programming model based on:

- **Thread Based Parallelism**
  - Parallelism accomplished through threading
- **Explicit Parallelism**
  - Provide full control over parallelization in applications
  - Simply insert compiler directives
- **Multithreading – Fork-Join Parallelism**
  - Master thread that forks slave threads
  - Slave threads execute on different processors

MPI is still utilized for all inter-node process communications. There is not explicitly a requirement to have another MPI application on the main host processor.

---

**Intel® Xeon Phi™ Coprocessor with Intel® True Scale Fabric Performance**

When the Intel® Xeon Phi™ is used in concert with the Intel® True Scale Fabric, excellent MPI performance testing results are produced. All testing used tests used are from the HPC Challenge test suite except miniFE (Sandia Labs). All tests utilize the same test configurations. The tests are run with 2-nodes, 4- nodes and 8-nodes, each containing an Intel® Xeon Phi™ Coprocessor. The test results all showcase that the performance of the Intel® True Scale Architecture scales best as the cluster grows in size. Testing is performed to compare the performance difference between architectural implementations of InfiniBand. One is the “Verbs/Offload-based” Fabric and the “PSM/On-Load-based” Intel® True Scale Fabric. All tests were conducted in Intel’s lab using Industry accepted test criteria.
HPC Challenge – Latency

The following 3 tests are Communication latency tests. Each architecture is tested against standard benchmarks. Lower numbers are better. In each case, the Intel® True Scale architecture using the PSM/On-Load host stack out performs the Verbs/Offload architecture using the same host configuration and hardware.

Ping-Pong Latency
Measures non-simultaneous/no contention latency communication pattern. Shows a 27% advantage at scale.

Natural Order Ring Latency
Measurement is done using simultaneous communication, where it covers levels of contention caused by each process communicating with its closest neighbor. Shows a 21% advantage at scale.

Random Order Ring Latency
Measurement is done using simultaneous communication, where it covers levels of contention caused by each randomly chosen process in parallel that might occur in real application while communicating with its neighbors. Shows an 18% advantage at scale.
HPC Challenge – Bandwidth

Parallel Matrix Transpose Test (PTRANS)
This test exercises the communications where pairs of processors communicate with each other simultaneously. It tests the total communications capacity of the network. **At 8 nodes a 3% advantage is shown with a Scaling Efficiency of 94%**. What is also shown is the scale trend which shows performance increasing as the cluster grows. Performance must not suffer as the scale grows.

HPC Challenge – Performance

High Performance Linpack (HPL)
The Linpack TPP benchmark measures the floating point rate of execution for solving a linear system of equations and is used to calculate the FLOPs (Floating Point Operations per Second). **A performance advantage of 119% advantage at scale. The Scaling Efficiency is 98%**

Fast-Fourier Transform (FFT)
FFT measures the floating point rate of execution of double precision complex one-dimensional Discrete Fourier Transform (DFT). **A 79% performance advantage is shown. Also, the scaling efficiency is 77% for this test versus a 38% for the other architecture.**
**Application Performance Example**

**miniFE Time-to-Solution**
Sandia National Laboratories developed a simulation using an unstructured implicit finite element or finite volume application. The Native/Hybrid compute model and OpenMP were used in this test. Optimal Performance Result of a *17% advantage* scale. The optimal performance Host configuration was:

- Thread/Rank - 8
- Cores/Phi Utilized - 30
- Total Phi Cores/Cluster - 240

When used together the Intel® product suite including Intel® Xeon® Host processors, Intel® Xeon Phi™ coprocessors and the True Scale Fabric Architecture will bring Higher Performance to HPC clusters.

**Summary**

To achieve high performance, architecture matters. As shown in the performance results, the Intel® True Scale Architecture which was explicitly designed from the ground up for HPC performance using MPI does fare better overall in all of the HPC Challenge performance tests. The design criteria of high Message rate, low latency at scale and collectives performance along with the added nodal performance of the Intel® Xeon Phi™ Coprocessor and the Intel® CCL Host SW showcase the overall performance of the Intel® HPC portfolio. As shown, a general purpose host stack such as Verbs/Offload does not fare as well.
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. 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. Testing was done by Intel.

* Other names and brands may be claimed as the property of others.
1 As presented at the OPENFABRICS Alliance Workshop (Open MPI 1.4.3) – April 2013

Copyright © 2013, Intel Corporation. All rights reserved