Evaluation of the Pre-production Implementation of the Intel Omni-Path Interconnect Technology

Alexandr Zaytsev
alezayt@bnl.gov

BNL, USA
RHIC & ATLAS Computing Facility
Intel Omni-Path (OPA) Product Family Overview

Multi-generation high bandwidth/low latency fabric targeting leadership-class HPC facilities, large scale HTC facilities & data centers

100 Gbps links, Single port HFIs, 160M max. message rate, Switch latency < 110ns
Remote Test Setup

• 32-node “Diamond” cluster kindly provided to RACF by Intel for remote testing 3 days in Dec 2015:
  – Dual Intel Xeon E5-2697 v3 14-core CPUs (configured in a Turbo mode plus HT enabled) per node
  – 56 logical cores / 28 physical cores per node
  – 64 GB of 2133 MHz DDR4 RAM per node
  – Single Intel Omni-Path PCIe x16 HFI card (non-blocking 100 Gbps)

• 1792 HT CPU cores / 896 physical CPU cores + 2 TB of RAM in total

• Single 48-port Intel Omni-Path switch (9.6 Tbps fabric)
  – Maximum achievable switch fabric utilization by 32 nodes: 67%
  – All compute nodes are connected to the switch with 2m long passive copper QSFP28 cables

• OS and software stack configuration
  – RedHat Enterprise Linux 7 x86_64 (kernel version 3.10.0-123.el7.x86_64)
  – GCC v4.9.2 & ICC v15.0.2 compilers
  – OpenMPI v1.10.0, MVAPICH2 v2.1-3 from Intel OPA Fabric Suite v10.0.0.0-625
  – Kernel (hfi1) module version 0.9-418 with IPoIB layer configured on top (connected mode, MTU 65520)
Approach & Methodology

• Utilizing the low level fabric layer and MPI benchmarking tools available in the OFED, Intel Fabric Suite and MPI distributions built with both GCC and ICC provided with the setup

• The latency and bandwidth values are reported as given by the benchmarking tool, and unless otherwise stated the latency is given for the unidirectional message passing scenario, which corresponds to the RTT/2 in the interconnect system with symmetrical routing

• The results obtained were checked for reproducibility with three consequent benchmark iterations performed under the same background load conditions

• A subset of the tests were performed with benchmarking tools built with GCC and ICC, with and without explicit mapping of a test processes to the physical CPU cores on the compute nodes in order to determine wither these changes result in a statistically significant performance variation

• Only a subset of results is reported here – full report can be made available upon request (with Intel’s consent)
Aggregated bandwidth on the interface of the client node observed with the multithreaded iperf tests involving a single (1-to-1) and four (1-4) server nodes.
Average bandwidth per thread observed on the client node during the multithreaded \texttt{iperf} tests involving a single (1-to-1) and four (1-to-4) server nodes. The maximum value on this plot corresponds to the highest single threaded throughput observed with \texttt{iperf} between two compute nodes.
Latency observed with `osu_latency` test vs the message size between two selected compute nodes in the fabric with *(bind)* and without the MPI level binding of the processes to the first CPU socket on each compute node. The lower boundary for latency determined by minimum message transfer time *(trl)* is shown for comparison for a range of message sizes (4 kB to 4 MB)
Latency observed with osu_multi_lat test vs the message size between all 32 compute nodes in the fabric with MPI level binding of the processes to the first CPU socket on each compute node. The tests performed for 1, 2, 4, 8, 16 and 28 parallel processes running on each compute node. The lower boundary for latency determined by minimum message transfer time (trl) is shown for a range of message sizes (4 kB to 4 MB).
OpenMPI OSU Benchmarks: Bandwidth (1)

Bandwidth observed with osu_bw and osu_bibw tests performed with (bind) and without the MPI level binding of the processes to the first CPU socket on each compute node.

Alexandr Zaytsev (RACF, BNL) – CHEP2016 (San Francisco, USA, Oct 8-14, 2016)
OpenMPI OSU Benchmarks : Bandwidth (2)

Bandwidth observed with osu_mbw_mr tests performed with 1, 4, 16, and 28 parallel processes running on each compute node. The tests for 1 and 28 parallel processes were performed with (bind) and without MPI level binding of the processes to the first CPU socket on each compute node.
OpenMPI Intel Benchmarks : Bandwidth

Bandwidth utilization observed with the mpi_multibw tests for both ICC and GCC compiled versions of the OpenMPI
Performance impact of the selection between ICC and GCC compiler on the selected tests of the IMB test suite. The latency metrics are shown in blue and the bandwidth metrics are shown in red (positive value means that ICC version of the test is showing a better performance compared to GCC).

The performance impact of choosing the compiler is bound to the range of −6.8% to +3.1%, yielding on average over all test only −0.4% of residual performance difference (in favor of GCC).
OpenMPI Intel Benchmarks: NxN Stress Tests

Latency observed by mpi_nxnlatbw (ICC) tests

Overall latency spectrum observed with the mpi_nxnlatbw test for the cases of 1, 2, 4 and 8 test processes running on every compute node in the cluster.
OpenMPI Intel Benchmarks: mpi_nxnlatbw

Latency vs number of active communication sessions

Latency profile for both local and fabric communication peaks in the spectrum (ICC compiled versions only)

Latency profile for the fabric communication peak in the spectrum (ICC and GCC compiled versions)
OpenMPI Intel Benchmarks : Latency Variations

48-port OPA switches are designed with two block of ports for latency equalization between the ports and ASIC

Ports are merged into 12 “megaport” blocks with 4 ports in each; latency between ports within “megaport” is lower than the latency between two different “megaports”

The overall latency budget in a single switch system is expected to be:
+ ≈ 2x 400 ns (Endpoints)
+ ≈ 110 ns (Switch fabric)
+ ≈ 2x 10 ns (2 m passive copper cables)

The latency asymmetry experimentally observed (intentionally shifted by +1%) with the test setup stays within the range from −0.68% to +0.96% (< 10 ns deviation from the mean value)

Oct 11, 2016
Alexandr Zaytsev (RACF, BNL) – CHEP2016 (San Francisco, USA, Oct 8-14, 2016)
Slice of the latency ($\tau$) spectrum obtained with the `mpi_nxnlatbw` tests with 8 processes per compute node for the range of values $0.70 \mu s \leq \tau < 0.95 \mu s$

Slice of the latency ($\tau$) spectrum obtained with the `mpi_nxnlatbw` tests with 8 processes per compute node for the range of values $\tau \geq 0.95 \mu s$
Pairwise bandwidth observed with mpi_nxnlatbw (ICC) tests

Bandwidth, Gbps

Number of communication sessions

Bandwidth per OPA interface, Gbps
Summary (1)

- Remote evaluation of the pre-production Intel Omni-Path hardware and software was performed by the RACF in Dec 2015 using the Intel owned “Diamond” cluster provided with a non-blocking Intel Omni-Path fabric based on a single 48-port pre-production Omni-Path switch
  - Even though the scale of this remote installation was limited, the test environment provided was sufficient to carry out a large variety of RDMA, native OpenMPI, Intel OpenMPI, and IP over Omni-Path performance measurements and functionality tests
- Performance measurements have demonstrated that it is feasible to extract close to the specification limits performance from the pre-production Intel Omni-Path hardware for the latency-bound applications utilizing the range of MPI message sizes from 1 byte to 4 MB:
  - Carrying up to **96.9 Gbps** of aggregated unidirectional MPI message traffic though a single interface (with 100 Gbps as the hardware limit)
  - Observe point-to-point unidirectional latency between two compute nodes in a single switch based fabric that is as low as **0.93 µs**
  - Bring a single IP over Omni-Path interface up to the level of **99.2 Gbps** of aggregated unidirectional TCP/IP traffic with multi-threaded iperf tests
Summary (2)

- The following observations about the fabric operational features were made while conducting performance tests:
  - The fabric (of a given scale of 32 compute nodes) remains stable even under usage scenarios implying severe network congestion under which every port participating in the fabric operates close to bandwidth saturation
  - The control over the fabric with the Intel Omni-Path diagnostic tools is not deteriorating under such usage scenarios

- Special attention was given in our tests to the latency jitter and latency systematics observed between compute nodes in the fabric
  - The observed latency jitter averaged across multiple MPI sessions was staying within ±1% of the mean value across the entire fabric
  - Systematic differences in node-to-node latency observed is staying within ±5% of the mean value for individual MPI sessions equally distributed among multiple compute nodes, even under a significant (yet uniformly distributed) workload

- The hardware and software features of pre-production Intel Omni-Path technology make it a viable alternative to the Mellanox 4X EDR Infiniband for both latency-bound and bandwidth-bound HPC and HTC workloads
Q & A