Introducing 100GBps with Intel® Omni-Path Fabric in HPC

By Munira Hussain, Deepthi Cherlopalle

This blog introduces the Omni-Path Fabric from Intel® as a cluster network fabric used for intra-node communication for application, management and storage communication in High Performance Computing (HPC). It is part of the new technology referring to Intel® Scalable System framework based on IP generated from the coalition of Qlogic, Truescale and Cray Aries. The goal of Omni-Path is to eventually be able to meet the demands of the exascale data centers in performance and scalability.

Dell provides complete validated and supported solution offering which includes the Networking H-series Fabric switches and Host Fabric Interface (HFI) adapters. The Omni-Path HFI is a PCI-E Gen3 x16 adapter capable of 100 Gbps unidirectional per port. The card supports 4 lanes supporting 25Gbps per lane.

HPC Program Overview with Omni-Path:

The current solution program is based on Red Hat Linux 7.2 (kernel version 3.10.0-327.el7.x86_64). The Intel Fabric Suite (IFS) drivers are integrated in the current software solution stack Bright Cluster Manager 7.2 which helps to deploy, provision, install and configure an Omni-Path cluster seamlessly.

The following Dell servers support Intel® Omni-Path Host Fabric Interface (HFI) cards

PowerEdge R430,PowerEdge R630, PowerEdge R730, PowerEdge R730XD, PowerEdge R930, PowerEdge C4130, PowerEdge C6320

The management and monitoring of the Fabric is done using the Fabric Manager (FM) GUI available from Intel®. The FM GUI provides in-depth analysis and graphical overview of the fabric health including detailed breakdown of status of the ports, mapping as well as investigative report on the errors.

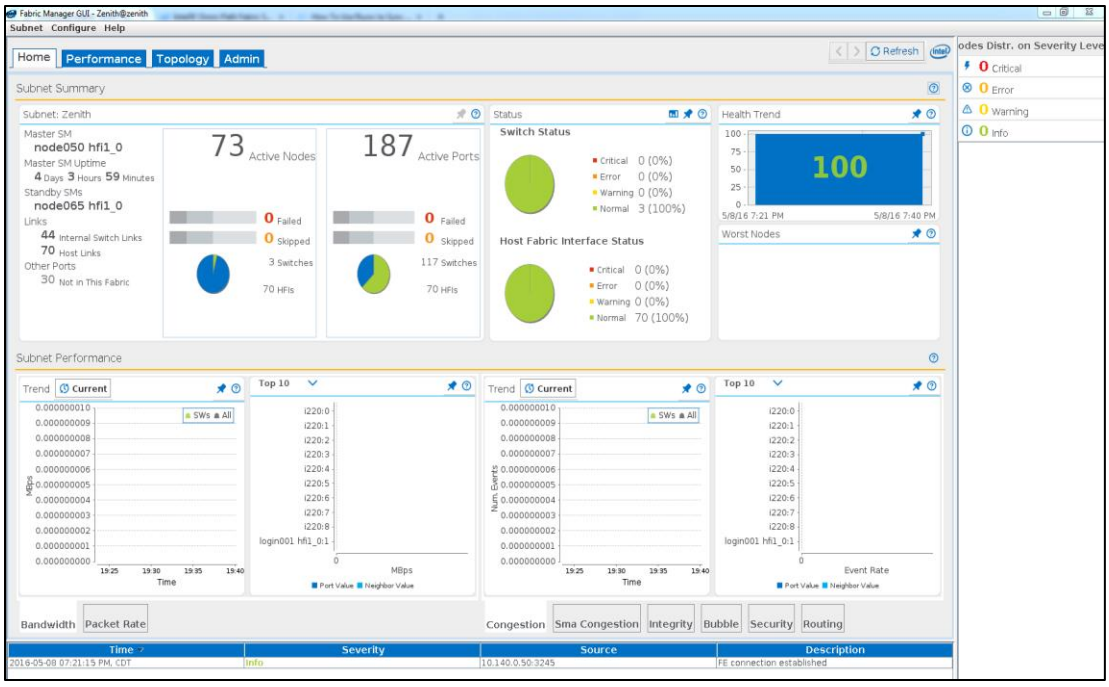


Figure 1: Fabric Manager GUI

The IFS tools include various debugging and management tools such as opareports, opainfo, opaconfig, opacaptureall, opafabricinfoall, opapingall, opafastfabric, etc. These help to capture a snapshot of the Fabric and to troubleshoot. The Host based subnet manager service known as opafm is also available with IFS and is able to scale up to 1000's of nodes.

The Fabric relies on the PSM2 libraries to provide optimal performance. The IFS package provides precompiled versions of the open source OpenMPI and MVAPICH2 MPI along with some of the micro-benchmarks such as OSU and IMB used to test Bandwidth and Latency measurements of the cluster.

Basic Performance Benchmarking Results:

The performance numbers below were taken on Dell PowerEdge Server R630. The server configuration consisted of the dual socket Intel® Xeon® CPU E5-2697 v4 @ 2.3GHz, 18 cores with 8*16 GB @ 2400MHz. The BIOS version was 2.0.2, and the system profile was set to Performance.

OSU Micro-benchmarks were used to determine latency. These latency tests were done in Ping-Pong fashion. HPC applications need low latency and high throughput. As shown in Figure *2*, the back to back latency is 0.77µs, and switch latency is 0.9µs which is on par with industry standards.

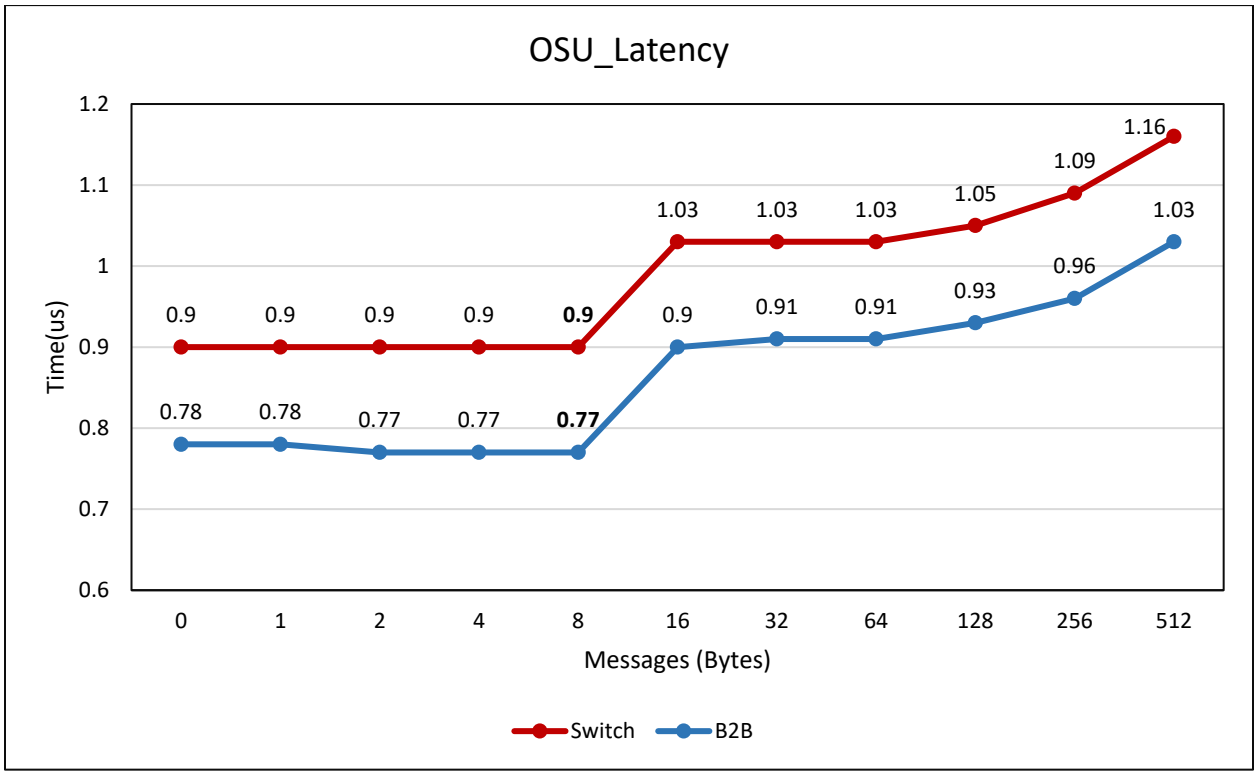


Figure 2: OSU Latency - E5-2697 v4

Figure *3* below shows the OSU Uni-directional and bi-directional bandwidth results with OpenMPI-1.10-hfi version. At 4MB Uni-directional bandwidth is around 12.3 GB/s, and bi-directional bandwidth is around 24.3GB/s which is on par with the theoretical peak values.

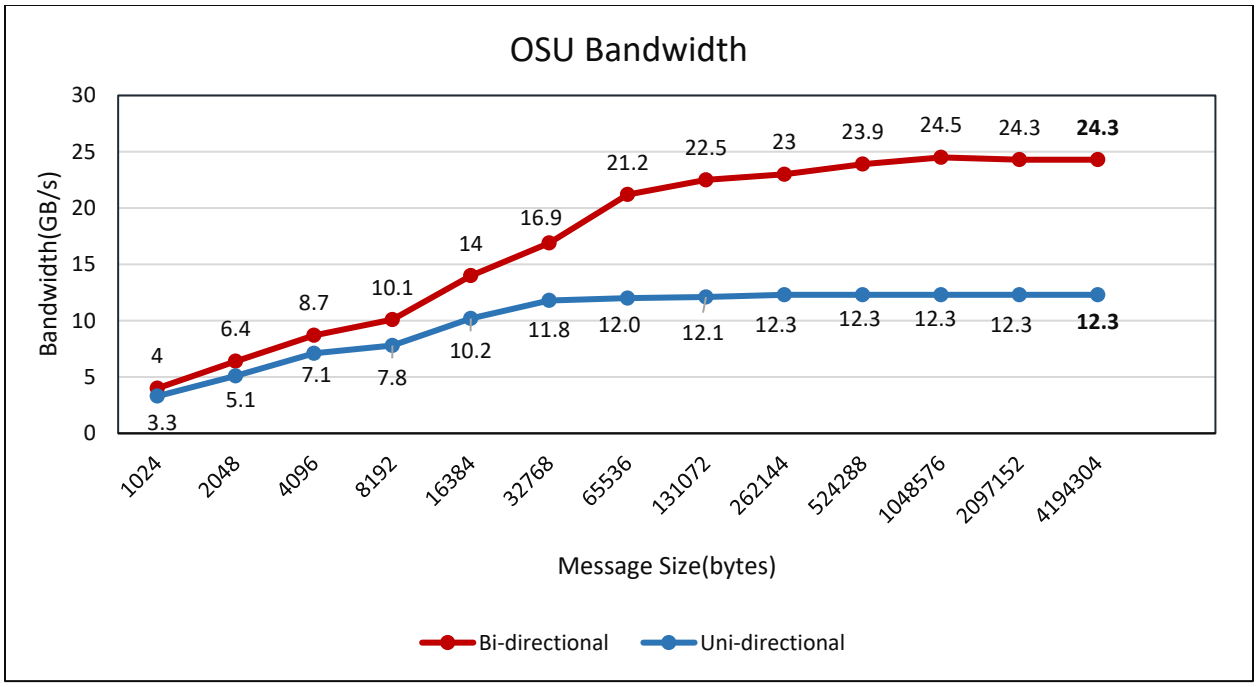
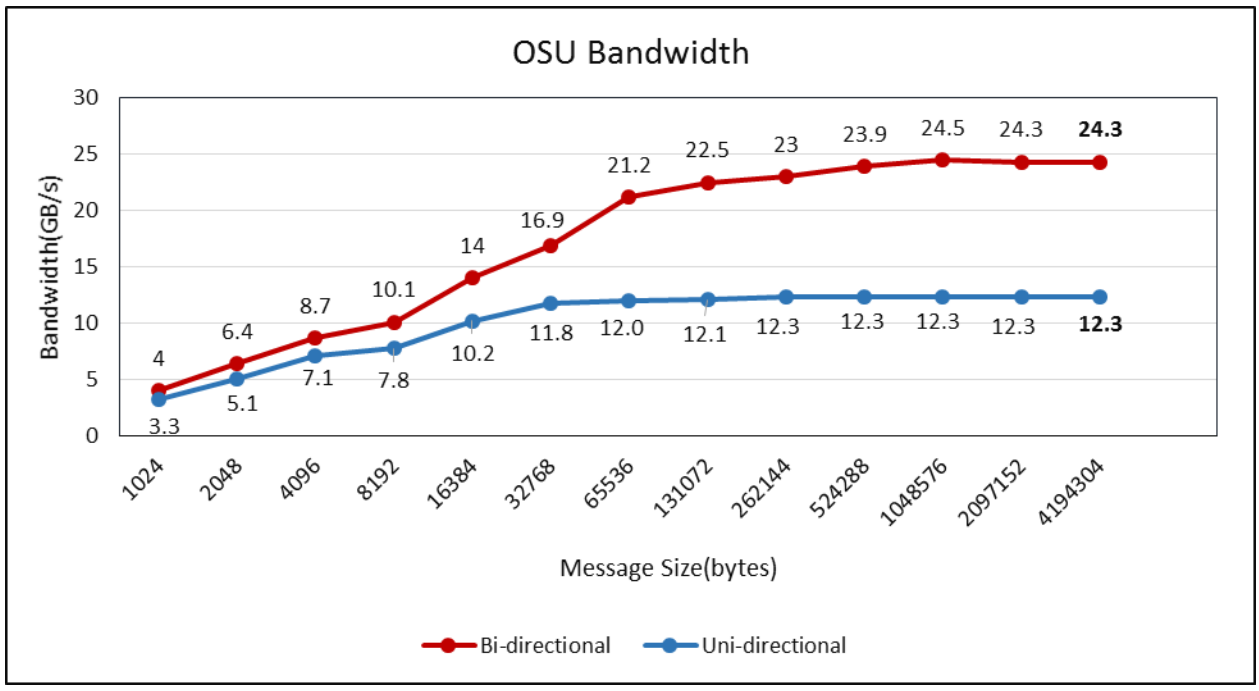


Figure 3: OSU Bandwidth – E5-2697 v4

Conclusion:



Omni-Path Fabric provides a value add to the HPC solution. It is a technology that integrates well as a high speed fabric needed for designing flexible reference architectures with the growing need for computation. Users can benefit from the open source fabric tools like FMGUI, Chassis Viewer and also FastFabric that is packaged with the IFS. The solution is automated and validated with Bright cluster Manager 7.2 on Dell Servers.

More details on how Omni-Path perform in the other domains is available here <link to the benchmarking white paper>. This document provides Intel® Omni-Path Fabric technology key features and provides a reference to performance data conducted on various commercial and open source applications.