New Fabric Interconnects

A comparison between Omni-Path and EDR Infiniband Architectures

Paolo Bianco

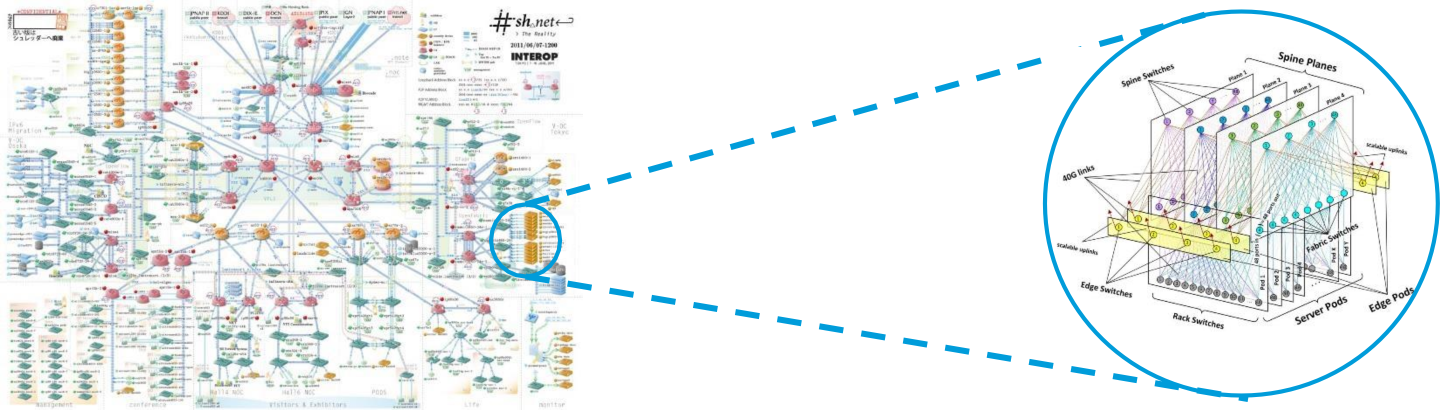
C&N Sales Engineer

paolo.bianco@dell.com

Networks and Fabrics

Network: Universal interconnect designed to allow any-and-all systems to communicate

HPC Fabric: Optimized interconnect to allow many nodes to perform as a single system



Intel® Omni-Path Architecture or Infiniband

Key NETWORK (Ethernet) Attributes:

- Flexibility for any application
- Designed for universal communication
- Extensible configuration
- Multi-vendor components

Key FABRIC Attributes:

- Targeted for specific applications
- Optimized for performance and efficiency
- Engineered topologies
- Single-vendor solutions

What is InfiniBand?

- A contraction of Infini(te) Band(width)
- Multi-lane, high-speed serial interconnect (Copper or Fiber)
- Standard Protocol, defined by IBTA (InfiniBand Trade Association)
- Multiplication of the link width and link speed
- Most common shipping today is 4x ports
- Mellanox is the only hardware vendor alive on the market
  - Intel announced EOL of TrueScale products on 31/12/2016
- Open Source software stack
- 100Gbps (EDR)/200Gbps(HDR) bi-directional speed
- 90ns port-to-port latency,149.5Mmsg/sec (EDR)
- 4KB MTU

What is Omni-Path?

- Multi-lane, high-speed serial interconnect (Copper or Fiber)
- Proprietary Architecture and Protocol
  - Evolutive path from Cray Aries and Qlogic TrueScale IB, with a flavour of Ethernet.
  - Replaces former Intel/Qlogic IB offering
- All OPA host software is Open-Source
- 100Gbps, 110ns port-to-port latency,160M msg/sec.
- CPU-Fabric integration available
- Larger MTU support (4KB, 8KB and 10KB)
- LNET routers to talk to existing IB storage
- Link Layer Traffic flow optimization

OmniPath Link Layer Innovations

- Congestion Avoidance through Virtual Lanes and Buffer Credits
  - VLs are separate logical communication links that share a single physical link.
    - OPA can have up to 32 VL (8 on adapters), IB up to 16 (8 on adapters)
  - Sender will transmit data only if sure that receiver has room to receive (buffer credits available) - Also IB and FC works in this way.

- Routing Enhancements
  - **Adaptive Routing**: monitors the routing paths of all the fabrics connected to the switch and selects the least congested path to balance the workload
  - **Dispersive Routing**: distributes the traffic across multiple paths as opposed to sending them to the destination via a single path

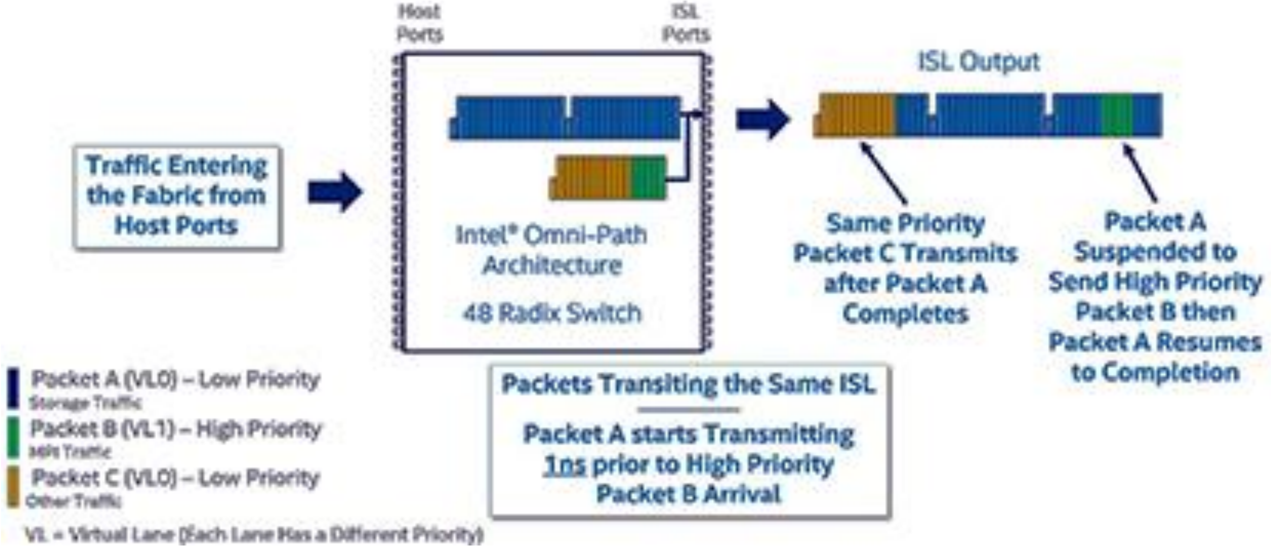
OmniPath Link Layer Innovations

- **Dynamic Lane Scaling**: physical link width (number of physical lanes used) can up/downscale based on BW needings or lane failures.
  - With IB the whole link downgrades to 1 lane if a physical lane fails, with OPA you just lose the lane.

- **Packet Integrity Protection**
  - Transmission errors are handled at the link level in addition to end-to-end level
  - If low-level CRC fails retransmission occurs per LTP (Link Transfer Packet, 128Bytes wide)

OmniPath Link Layer - QoS

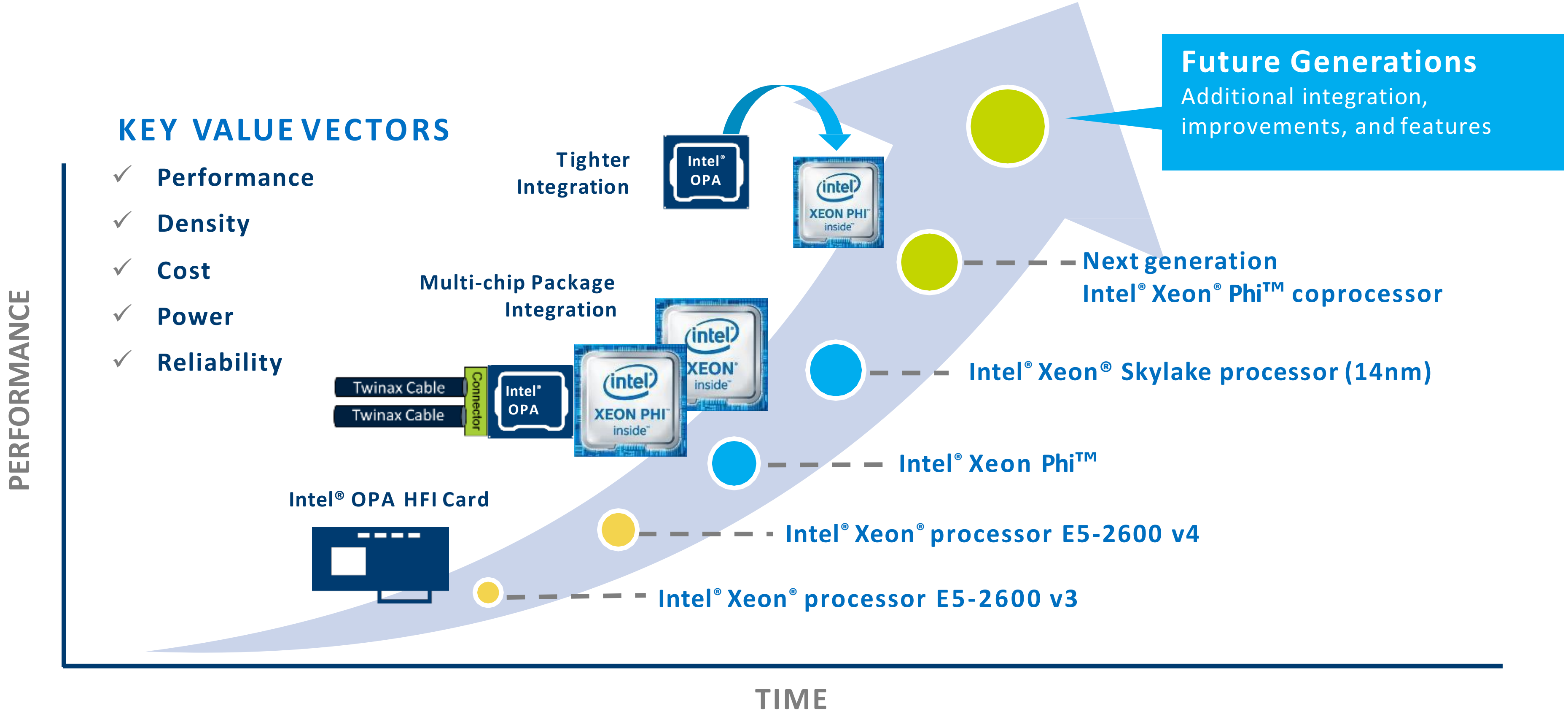
- **Traffic Flow Optimization**: QoS through Flow Priority (SL/TC) and FPs interleaving
  - FLITs (FLow Control digITs) pertaining to different Fabric Packets can be interleaved
  - High priority FPs can suspend transmission of other packets and be transmitted first (unique to OPA)


Host Ports
ISL Ports
ISL Output
Traffic Entering the Fabric from Host Ports
Intel® Omni-Path Architecture
48 Radix Switch
Same Priority Packet C Transmits after Packet A Completes
Packet A Suspended to Send High Priority Packet B then Packet A Resumes to Completion
Packets Transiting the Same ISL
Packet A starts Transmitting 1ns prior to High Priority Packet B Arrival
Packet A (VL0) – Low Priority
Storage Traffic
Packet B (VL1) – High Priority
MPI Traffic
Packet C (VL0) – Low Priority
Other Traffic
VL = Virtual Lane (Each Lane Has a Different Priority)

CPU-Fabric Integration

KEY VALUE VECTORS

- ✓ Performance
- ✓ Density
- ✓ Cost
- ✓ Power
- ✓ Reliability



Future Generations
Additional integration, improvements, and features

Tighter Integration

Multi-chip Package Integration

Intel® OPA HFI Card

Next generation Intel® Xeon® Phi™ coprocessor

Intel® Xeon® Skylake processor (14nm)

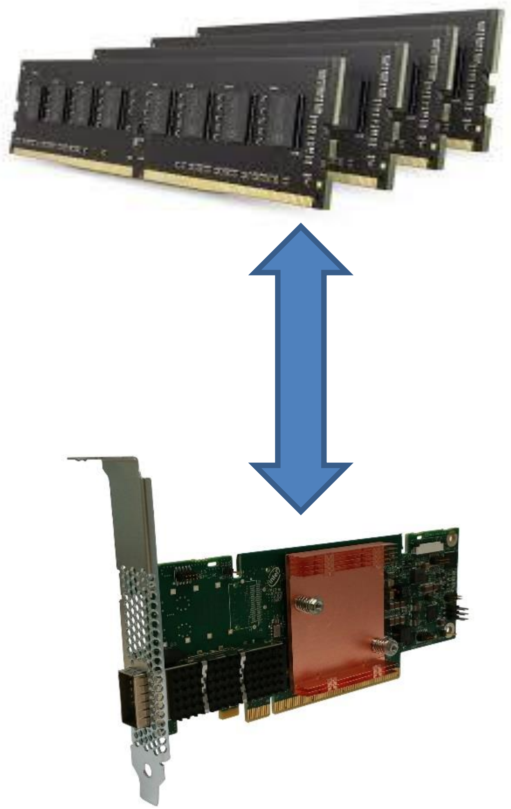
Intel® Xeon Phi™

Intel® Xeon® processor E5-2600 v4

Intel® Xeon® processor E5-2600 v3

Memory Performance considerations

- A 100Gbps adapter can give you 12.5GByte per second of throughput
- DDR3 memory at 2133MT/s runs at 17GByte per second.
- DDR4 memory at 2400 MT/s runs at 19.2GByte per sec.
- Some IB HCA adapters can give you dual 100Gbps connections.
- Memory read or write via the interface may potentially be able to saturate the system.



Available Offering – OPA

Dell Networking H-Series portfolio

HFI Adapter

x16 Adapter (100 Gb/s)

H1024-OPF Edge Switch

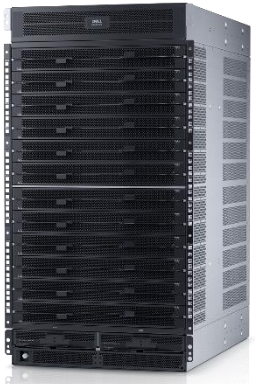
24 x 100 Gbps ports with up to 4.8 Tbps aggregate bandwidth for small to medium systems.

H1048-OPF Edge Switch

48 x 100 Gbps with up to 9.6 Tbps aggregate bandwidth for medium to large systems.

H9106-OPF Director-Class Switch

192 ports, 6 slots, 100 Gbps director-class switch supporting up to 38.4 Tbps switching capacity.



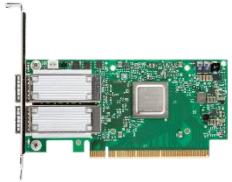
H9124-OPF Director-Class Switch

768 ports, 24 slots, 100 Gbps director-class switch supporting up to 153.6 Tbps switching capacity.

HFI	Edge switches	Director-class switches

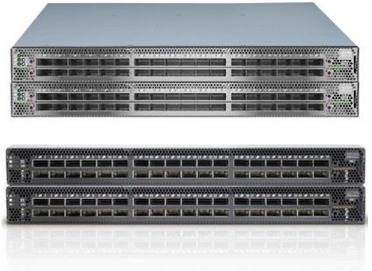
Available Offering – EDR IB

Entire Mellanox IB portfolio available trough Dell



HCA Adapter

x16 Adapter (100 Gb/s)
Single or Dual Port



CS7700/7800 Edge Switch

36 x 100 Gbps ports with up to 7 Tbps aggregate bandwidth for small to medium systems.

CS7520 Director-Class Switch

216 ports, 6 slots, 100 Gbps director-class switch supporting up to 43 Tbps switching capacity.

CS7510 Director-Class Switch

324 ports, 9 slots, 100 Gbps director-class switch supporting up to 64 Tbps switching capacity.

CS7500 Director-Class Switch

648 ports, 18 slots, 100 Gbps director-class switch supporting up to 130 Tbps switching capacity.

HFI	Edge switches	Director-class switches

Cabling

- Both OPA and EDR/HDR IB use passive (copper) or active (optical) cables with QSFP28 connectors, however:

- Intel Omni-Path
  - Max cable length: **3mt** (passive copper), **50mt** (active fiber)
  - Specs claim up to 5mt copper cable support
    - But no validated cables yet
  - Only validated cables are supported
  - Intel cables are rebranded from Finisar, Amphenol, Hitachi.
  - Turn to be standard 100GbE/IB cables from data sheet

- Mellanox EDR IB
  - Max cable length: **5mt** (passive copper), **200mt** (active fiber)
  - Mellanox allows the user to use any IBTA IB approved cable
  - Mellanox manufactures its own cables (claims an higher BER)
  - Mellanox IB cables turn to be different from Ethernet cables of the same speed

OPA vs IB - High Level Feature Comparison Matrix

Features	Intel® OPA	EDR	Notes
Link Speed	100Gb/s	100Gb/s	Same Link Speed
Switch Latency – Edge/DCS	100-110ns/300-330ns	90ns/~500ns	Intel® OPA includes "Load-Free" error detection • Application Latency Most important
MPI Latency (OSU pt2pt)	Less Than 1µs	~1µs	Similar 1 Hop Latency • Intel's OPA HFI improves with each CPU generation
Link Enhancements – Error Detection/Correction	Packet Integrity Protection (PIP)	FEC/Link Level Retry	Intel OPA is a HW detection solution that adds **no latency or BW penalty**
Link Enhancements – Data Prioritization across VLs	Traffic Flow Optimization (TFO)	No	Over and above VL prioritization. Allows High priority traffic to preempt in-flight low priority traffic (~15% performance improvement)
Link Enhancements – Graceful Degradation	Dynamic Lane Scaling (DLS)	No	Non-Disruptive Lane(s) failure. Supports asymmetrical traffic pattern. Avoids total shutdown,
RDMA Support	Yes	Yes	RDMA underpins verbs. Intel® OPA supports verbs. TID RDMA brings Send/Receive HW assists for RDMA for larger messages
Built for MPI Semantics	Yes – PSM (10% of code)	No - Verbs	Purpose designed for HPC
Switch Radix	48 Ports	36 Ports	Higher Radix means less switches, power, space etc.
Fabric Router	No	Future	Limited need to connect to older fabric technologies except for storage – Still not available

OPA vs EDR IB - Product Comparison

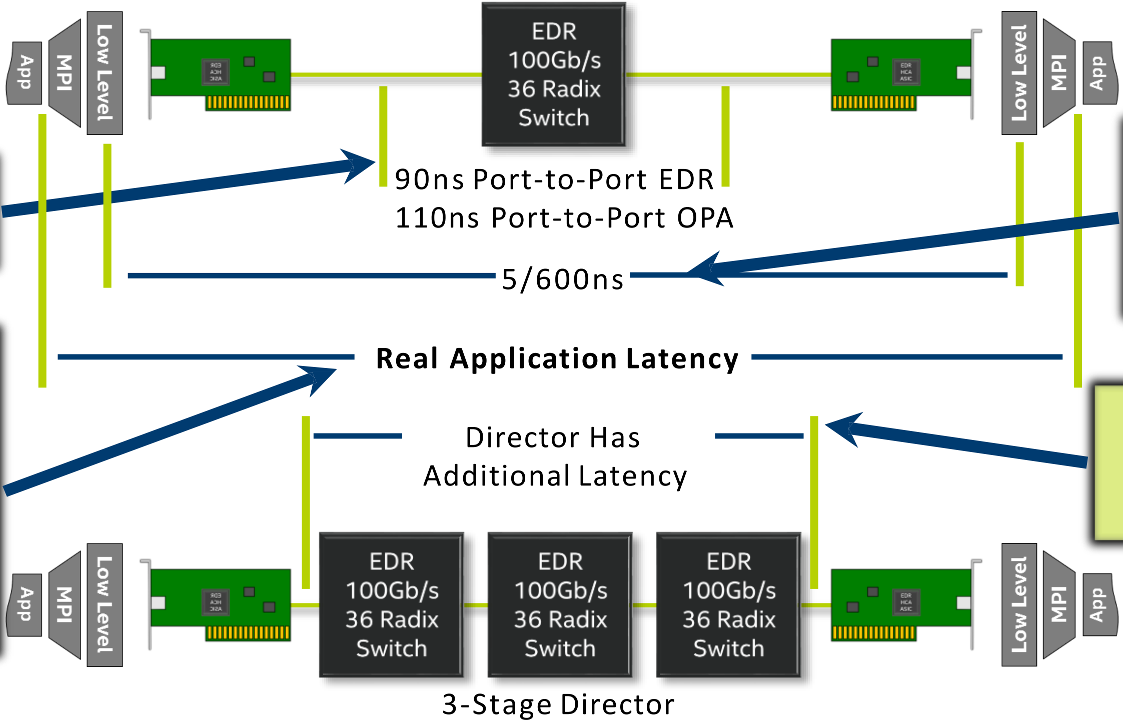
Feature	Intel® Omni-Path	EDR
Switch Specifications		
Link Speed (QSFP28)	100Gb/s	100Gb/s
Port Count: Director - Edge -	192, **768** **48**, 24	216, 324, **648** 36
Latency: Director - Edge -	**300-330ns** (Includes PIP) 100-110ns (Includes PIP)	<500ns[1] (Should be 3 x 90ns?) **90ns**[1] **(FEC Disabled)**
Packet Rate Per Port: Switch Host	195M msg/sec **160M** msg/sec (CPU Dependent)	150/195M msg/sec - Switch-IB/Switch-IB 2 150M msg/sec
Power Per Port ***(Typical Copper)***[2] : –24/18-Slot Director –48/36-Port Edge (M) –48/36-Port Edge (U)	 **~8.85 Watts** 3.87 W **3.48 W**	 14.1 Watts 3.78 W 3.78 W
Director Leaf Module: Size/Qty	32 / **(24-Slot),** (6-Slot)	**36** / (18-Slot), (6-Slot)
Largest 2 Tier Fabric (Edge/Director)	18,432	11,664
Host Adapter Specifications		
Host Adapter Model	Intel® OPA 100 Series (HFI)	HCA (ConnectX-4)
Protocol	Intel® OPA	InfiniBand
Speed Support (Host)	x16 = 100Gb/s – x8 = 58Gb/s	All Prior IB Speeds[1]
Power Per Port ***(Typical Copper)***[2] : –1-Port x16 HFI –1-Port x8 HFI	 **7.4 W Copper** **6.3 W Copper**	 13.9 W Copper

Performance Comparisons

System Configuration

	Omni-Path	EDR
Servers	32*PowerEdge R630	32*PowerEdge R630
Processors	Intel® Xeon® CPU E5-2697 v3 @ 2.60GHz No. Of Cores=14 Processor Base Freq: 2.6GHz AVX Base: 2.2GHz	
Memory	8*8GB DIMMS @2133Mhz	
HFI/HCA Card	Intel® Omni-Path Host Fabric Interface Adapter 100 Series 1 Port PCIe x16	ConnectX®-4 Single Port Adapter
BIOS	2.0.2	
System Profile	MaxPerformance • Turbomode : Enabled • Cstates : disabled • Nodeinterleave : disabled • Logical processor : disabled • Snoop mode : COD • IO-NonpostedPrefetch: Disabled	
Switch Firmware	Dell H1048 OPF Switch 10.0.1.0.21	Mellanox SB7790 11.0300.0354
IFS / MLNX-OFED version	10.0.1.1.5	Mellanox OFED 3.2-2.0.0.0

Understanding Latency


App
MPI
Low Level
EDR 100Gb/s 36 Radix Switch
Low Level
MPI
App
App
MPI
Low Level
EDR 100Gb/s 36 Radix Switch
EDR 100Gb/s 36 Radix Switch
EDR 100Gb/s 36 Radix Switch
Low Level
MPI
App

90ns Port-to-Port EDR
110ns Port-to-Port OPA

5/600ns

Real Application Latency

Director Has
Additional Latency

3-Stage Director

Minor Part of Overall Latency

Low Level Non-MPI Measurement Does not Include FEC Latency

Intel® OPA = 900ns
EDR = 1001ns without Added FEC Latency

OSU MPI Measurements

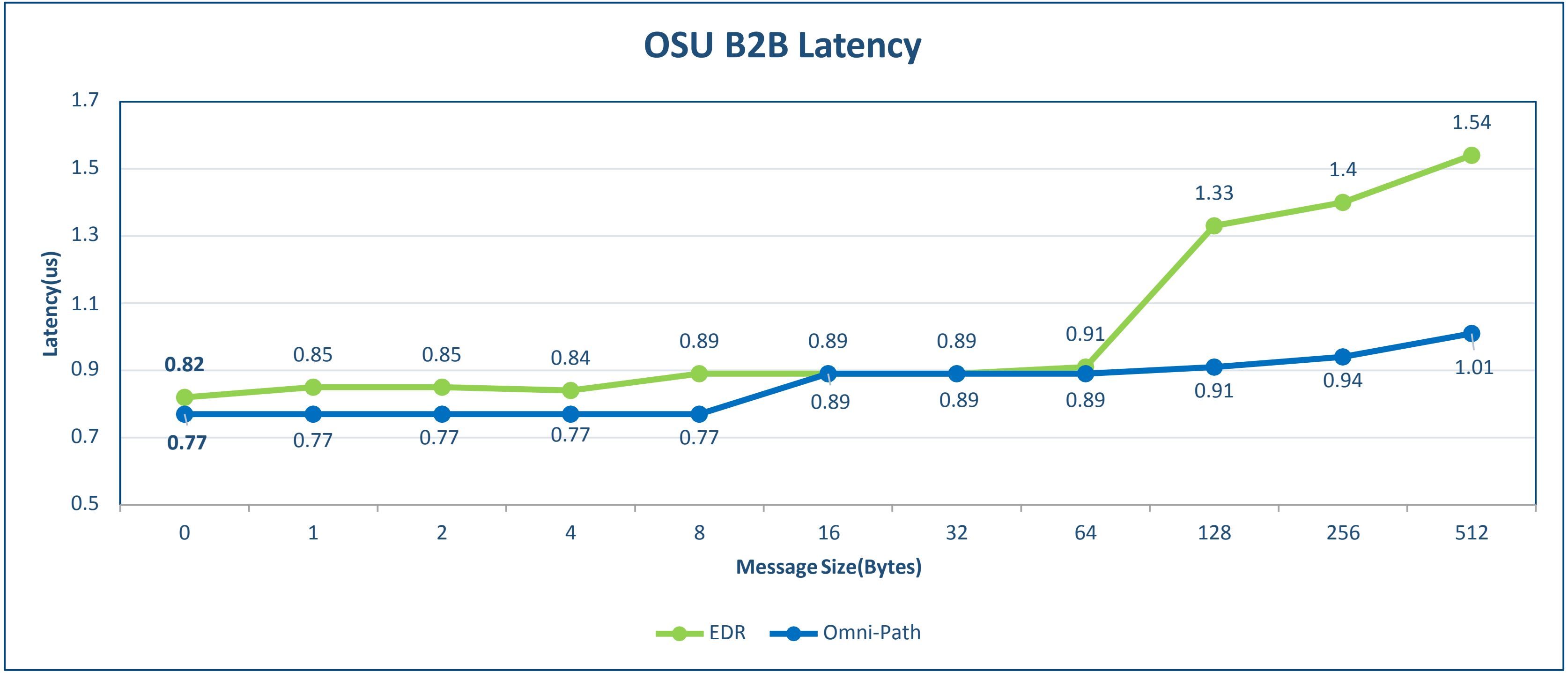
~500ns

Over 5x asingle Switch ASIC

Tests performed by Intel on Intel® Xeon® Processor E5-2697v3 dual-socket servers with 2133 MHz DDR4 memory. Turbo mode enabled and hyper-threading disabled. Ohio State Micro Benchmarks v. 4.4.1. Intel OPA: Open MPI 1.10.0 with PSM2. Intel Corporation Device 24f0 – Series 100 HFI ASIC. OPA Switch: Series 100 Edge Switch – 48 port. IOU Non-posted Prefetch disabled in BIOS. EDR: Open MPI 1.8-mellanox released with hpcx-v1.3.336-icc-MLNX_OFED_LINUX-3.0-1.0.1-redhat6.6-x86_64.tbz. MXM_TLS=self,rc tuning. Mellanox EDR ConnectX-4 Single Port Rev 3 MCX455A HCA. Mellanox SB7700 - 36 Port EDR InfiniBand switch 1. osu_latency 8 B message.

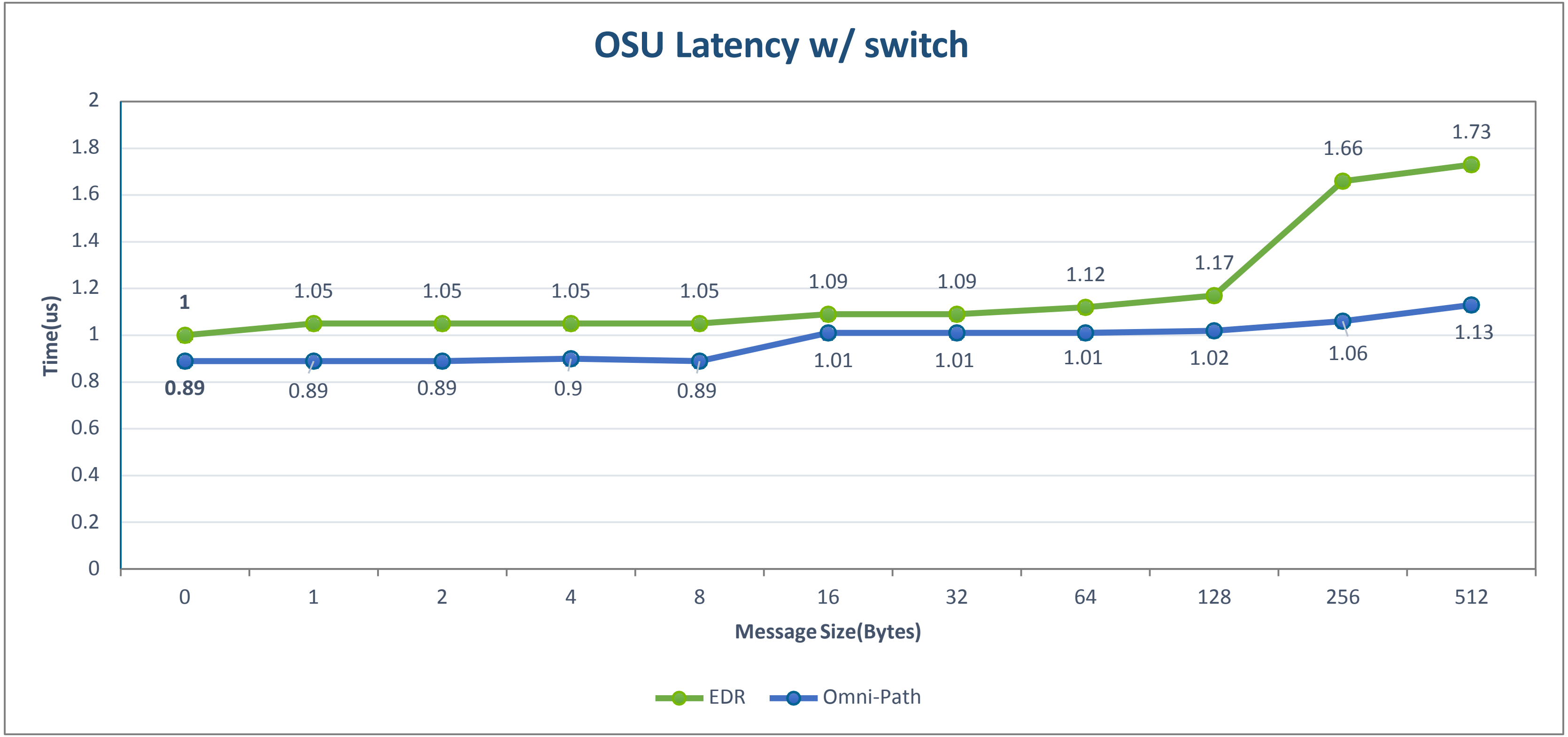
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may 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 complete information visit http://www.intel.com/performance.

OSU Latency (1/2)



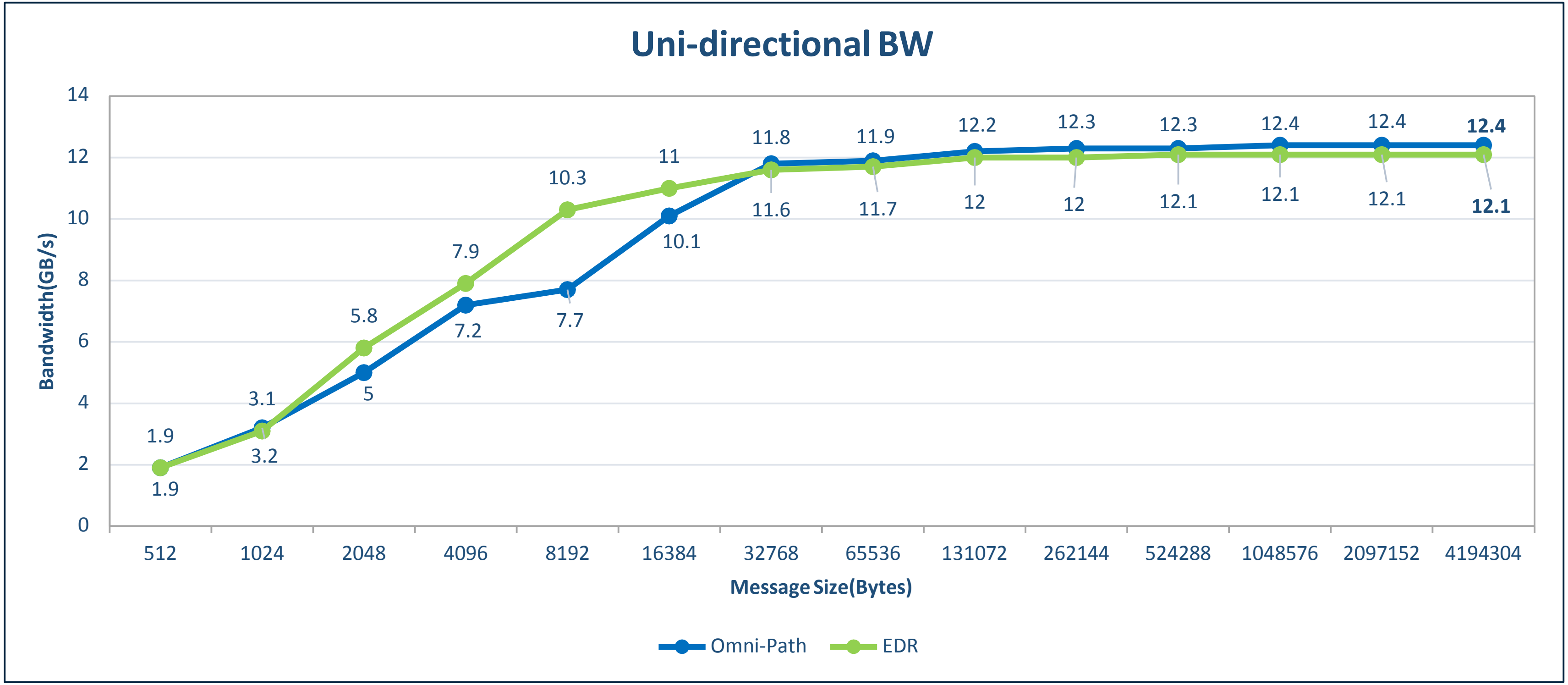
Back to Back Latency: EDR - 0.82µs, OPA - 0.77µs

OSU Latency (2/2)



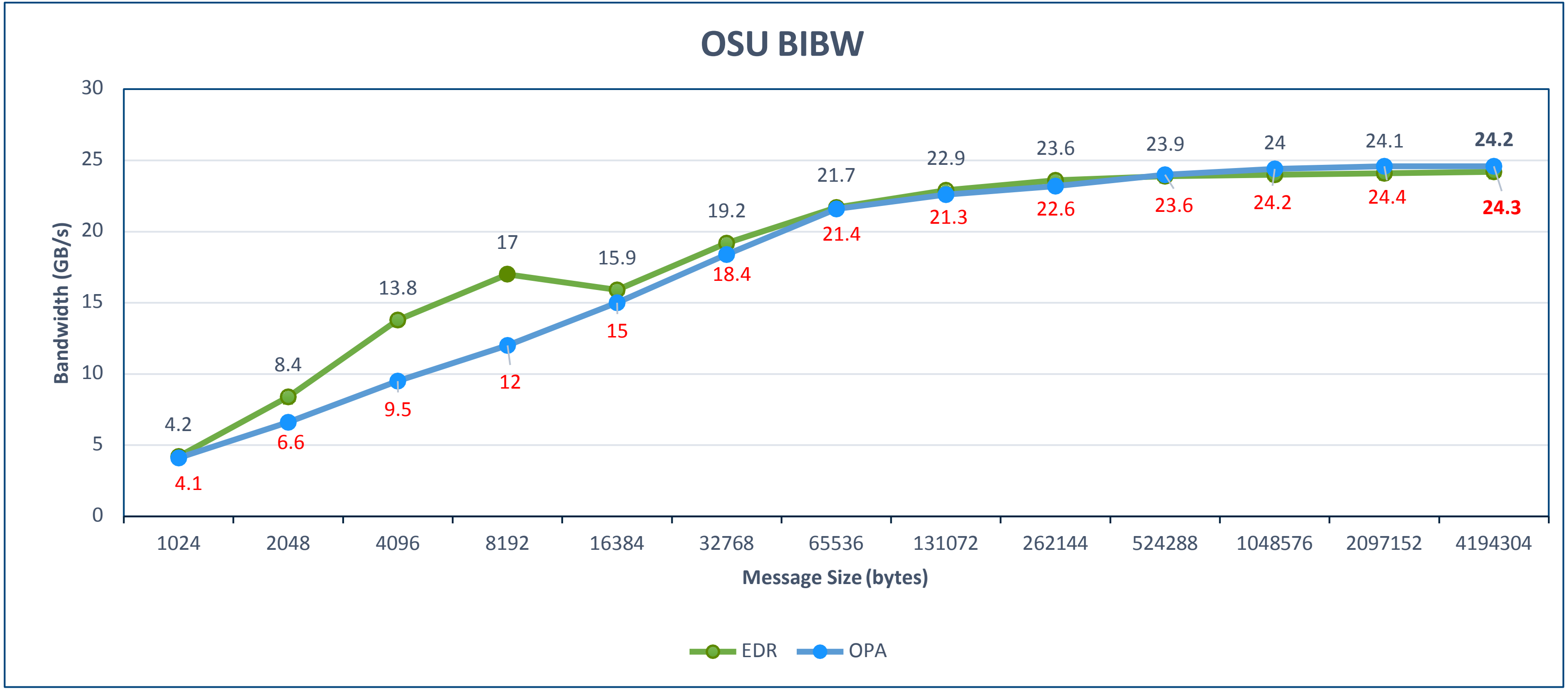
OSU Latency w/ switch: EDR-1µs, OPA-0.89µs

OSU uni-directional BW (1/2)



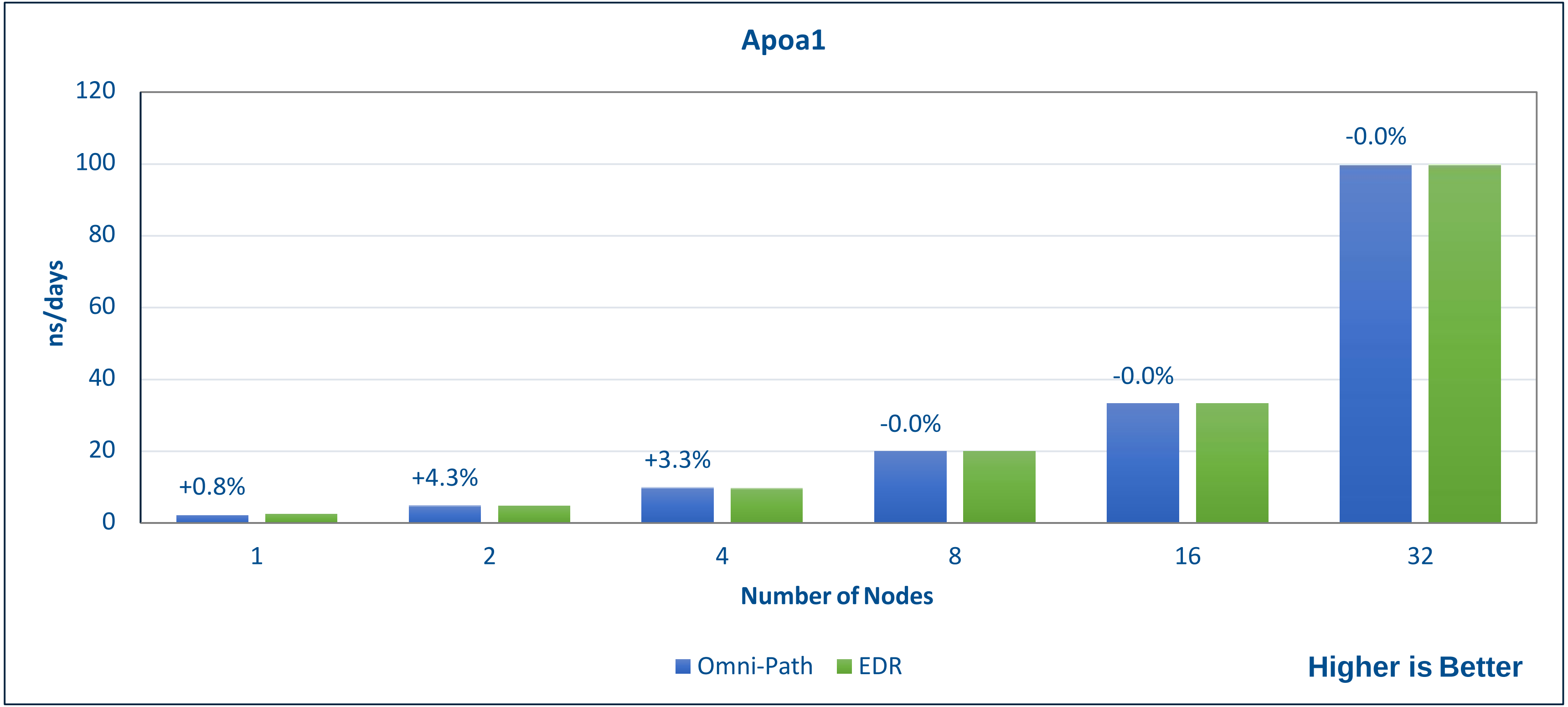
Uni-directional BW: EDR-12.1 GB/s, OPA-12.4GB/s

OSU BiBW (2/2)

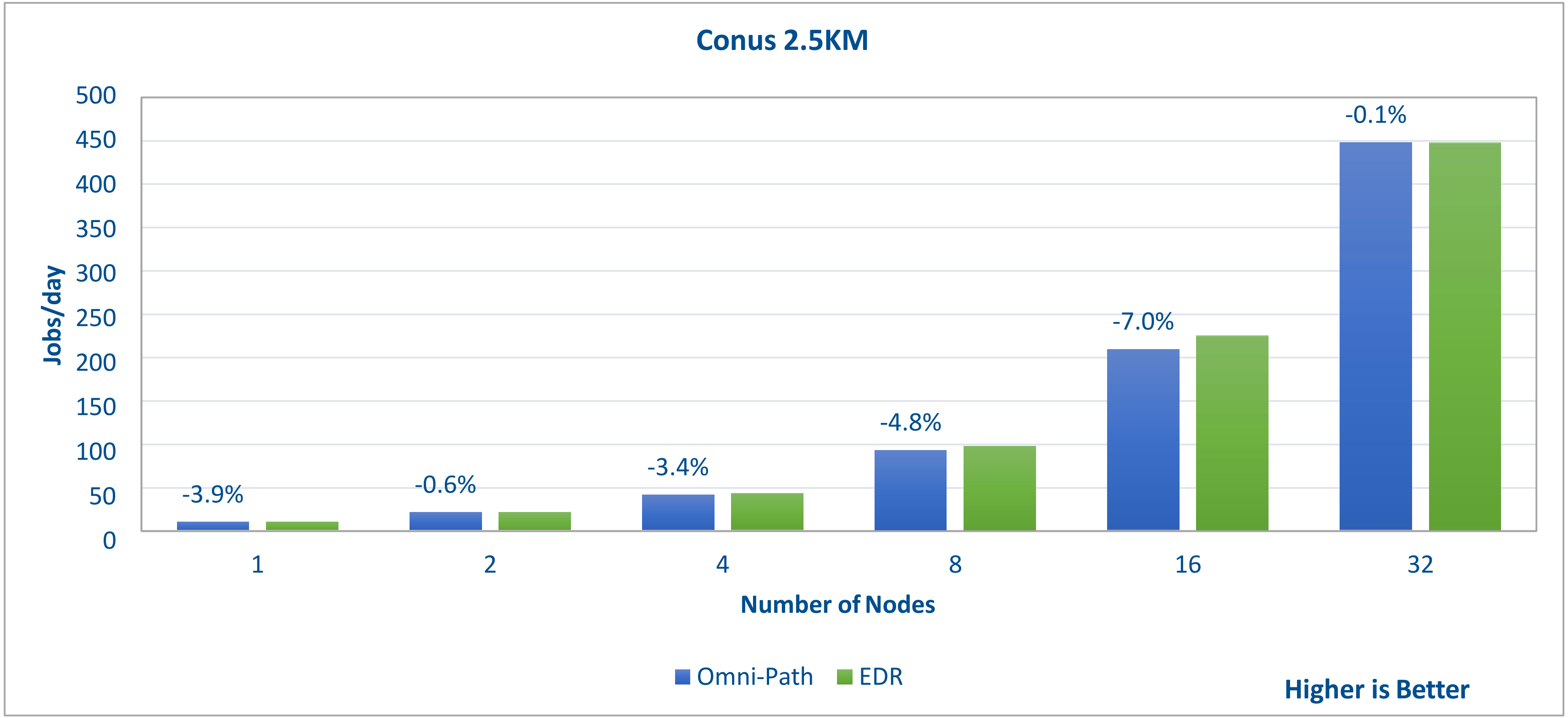


Bi-directional BW: EDR - 24.2 GB/s, OPA - 24.3GB/s

NAMD (Molecular Dynamics)

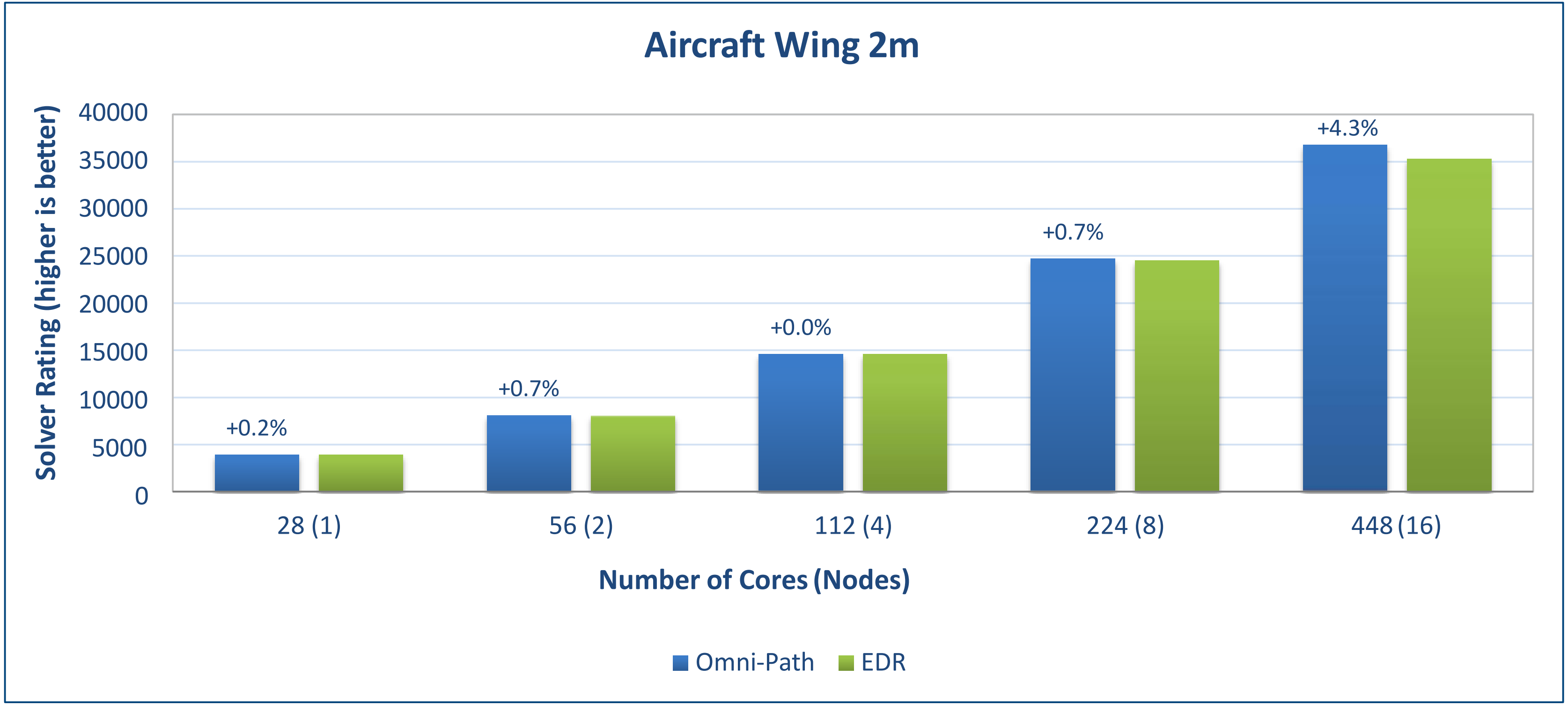


WRF (Weather Research and Forecast)

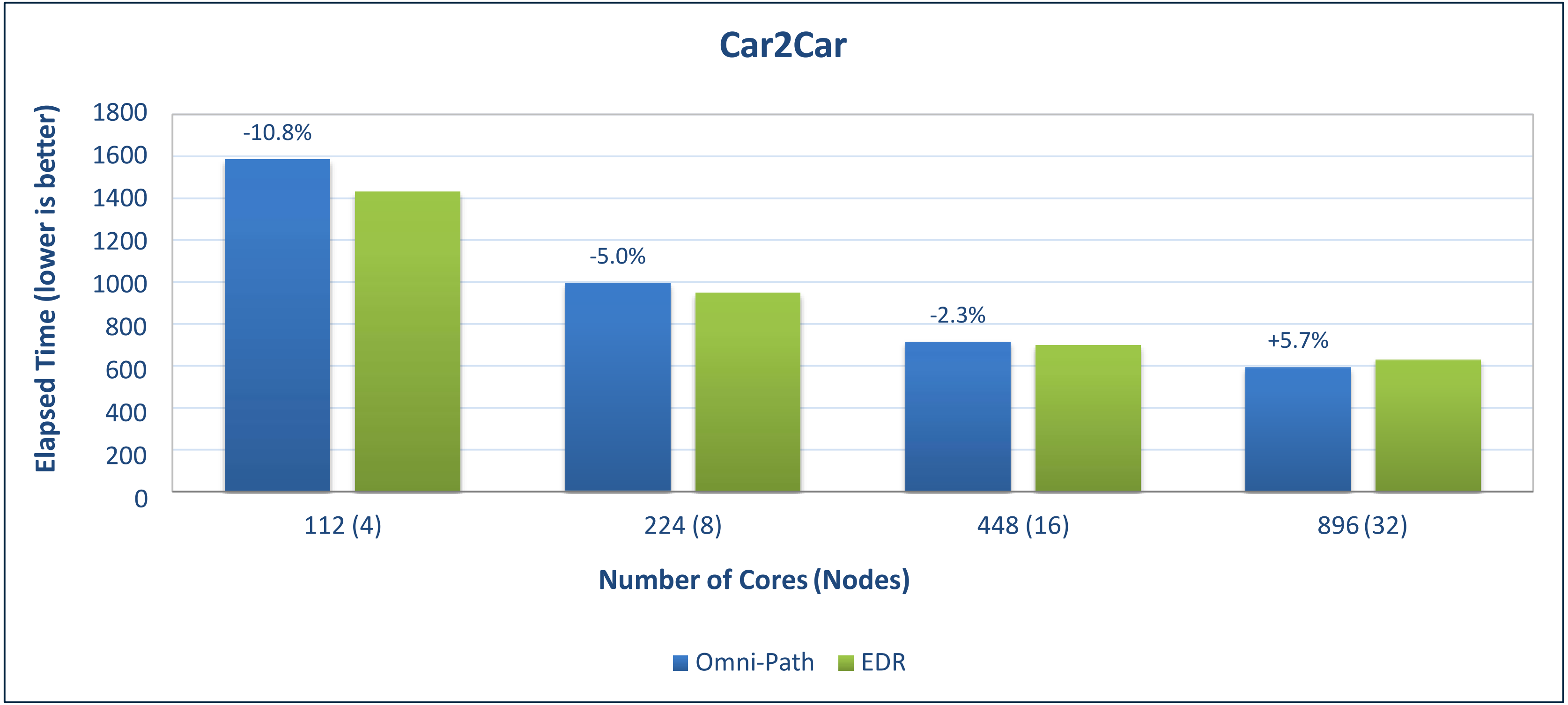


DELLEMC

ANSYS Fluent (Computational Fluid Dynamics)



LS-DYNA (Finite Element Analysis)



In turns: EDR InfiniBand Advantages

- Support for GPU Direct in current implementation
- Supports virtualisation (Xen, VMware, Hyper-V)
- True in-hardware RDMA
- True CPU Offload
- Multiple topologies supported
- Supported on Ceph, CephFS, Gluster, GPFS, NFS, Lustre, etc…
- Backwards compatible with existing InfiniBand fabrics
- Dual Port InfiniBand NICs for active/active or high-availability fabric designs

In turns: Intel Omni-Path Advantages

- Dynamic Lane Scaling (when one or more lanes fails, the fabric continues to function)
- Adaptive Routing (monitors the routing paths of all the fabrics connected to the switch and selects the least congested path to balance the workload)
- Dispersive Routing (distributes the traffic across multiple paths as opposed to sending them to the destination via a single path)
- Traffic Flow Optimization (prioritizes packets in mixed traffic environments like storage and MPI)
- GPU Direct (on firmware version > 10.4)
- Per-port switch list price cost about 50% less than EDR IB, at roughly same performances.

DELL EMC