Intel® Omni-Path Architecture

The Path to Exascale

John Swinburne

HPC Technical Specialist

Intel DCG Sales

Intel® Omni-Path Architecture

- OPA100 Update
- The Path to Exascale
- OPA100 Enhancements
  - NVMe over OPA
  - GPUDirect
  - Onload vs Offload (sorry…)
- ISV Performance Figures

Intel® OPA100: Continued Growth in 100Gb Fabrics

Top500 listings continue to grow

- 4 Top 15 systems, 13 Top 100 systems
- 36% more systems from Nov 2016 list
- First Skylake systems
- And almost 10% of Top500 Rmax performance

New deployments all over the globe

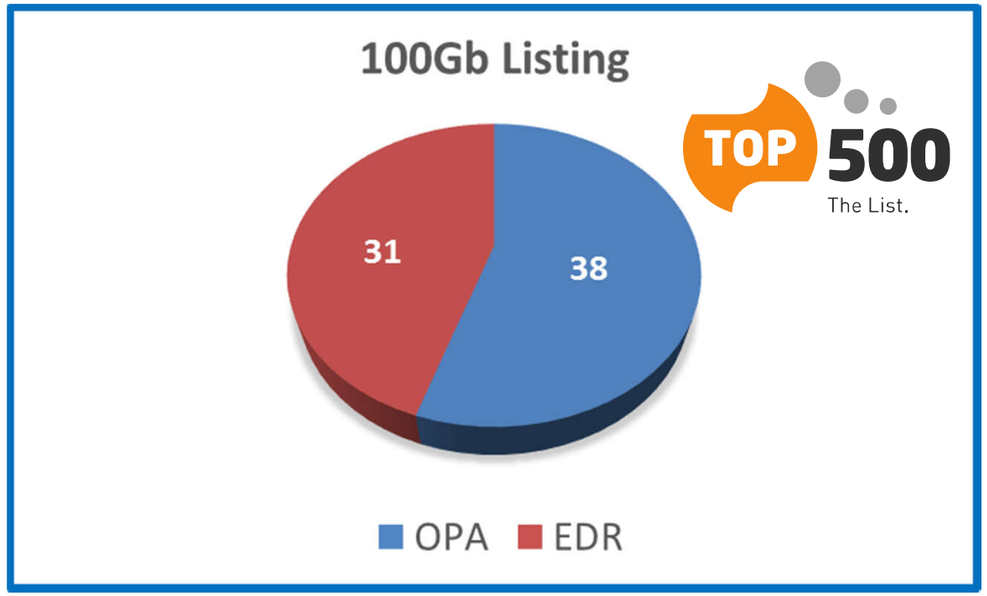
- All geos, and new deployments in HPC Cloud and AI

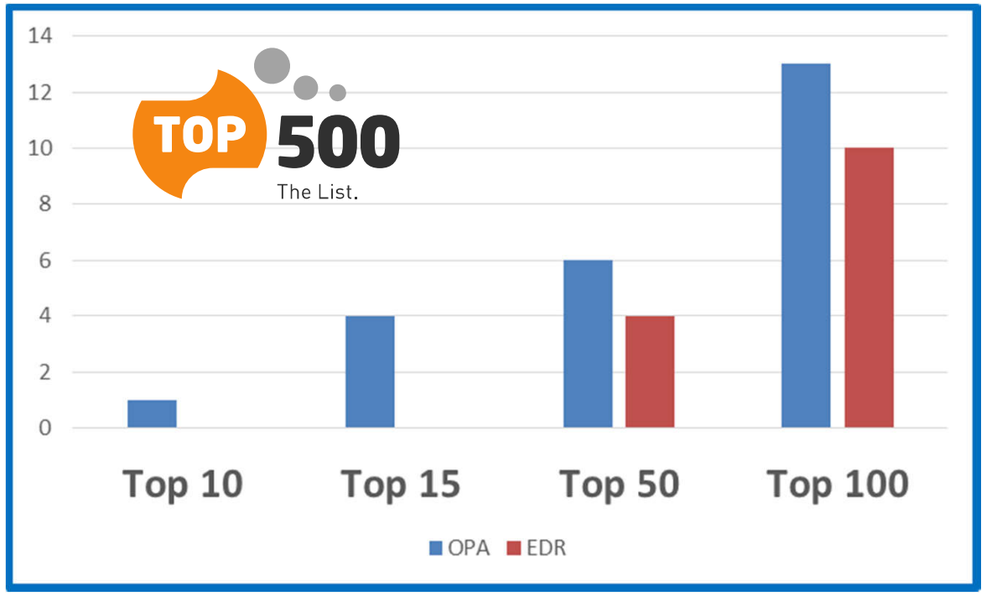
Expanding capabilities

- NVMe over Fabric, Multi-modal Data Acceleration (MDA), and robust support for heterogeneous clusters

Source: Top500.org

Top500 – 100Gb Fabric Only Listing



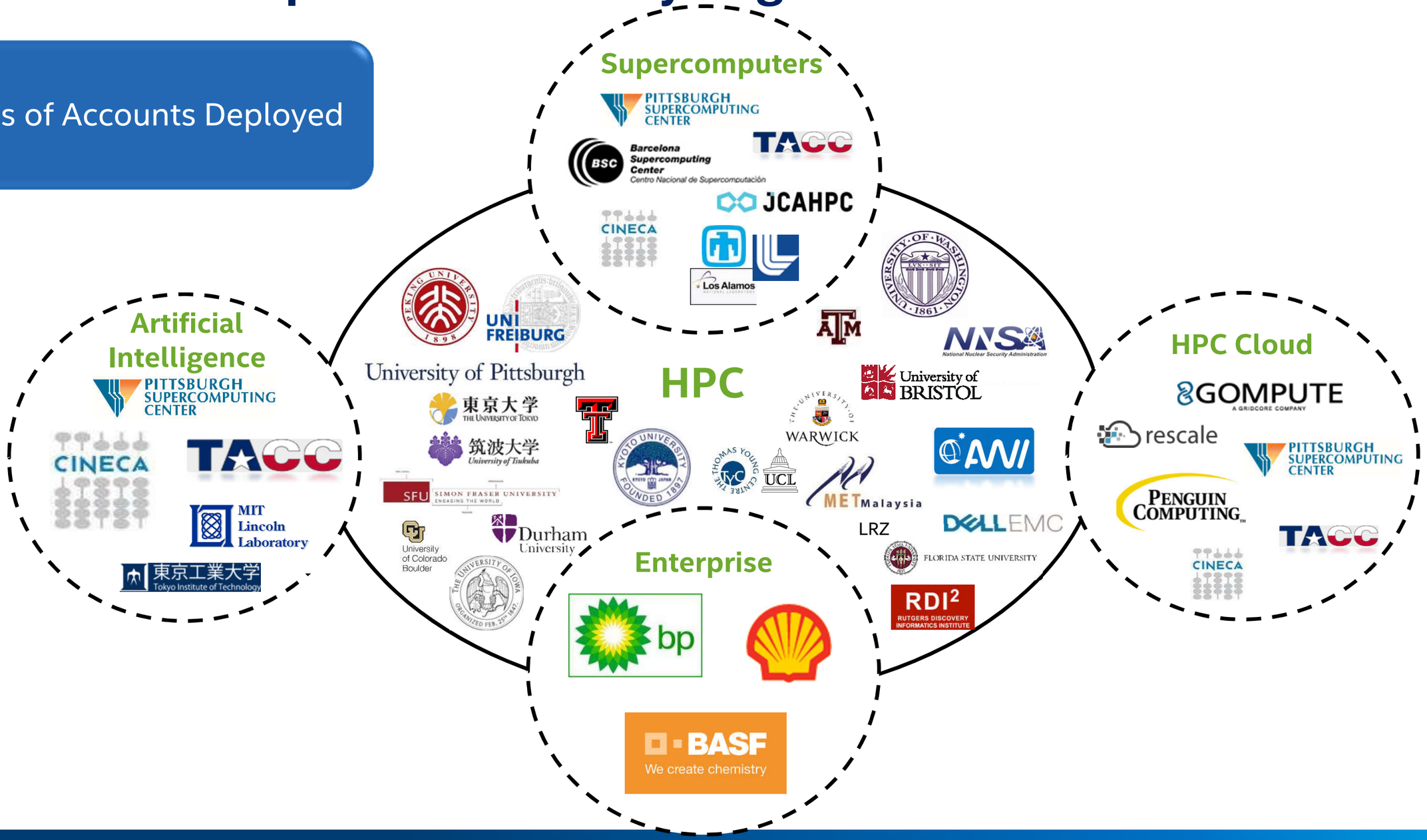


- Intel® OPA leads with 55% of the 100Gb listings
- Share of 100Gb Flops – 67.1PF OPA vs. 38.7PF EDR
- Intel Xeon® Processor HPL Efficiency: 74% OPA vs. 71% EDR

Source: Top500.org

Intel® OPA's Impact across Many Segments

100's of Accounts Deployed



Notable European Deployments

BASF – World's Leading Chemical Company

- **Goal:** Centralize and integrate number of smaller clusters to solve large problems, work more efficiently and effectively to meet new digitalization strategy in Ludwigshafen headquarters
- **Solution:** high performance cluster with Intel Omni-Path Architecture on HPE Apollo 6000 systems reduce modeling and simulations from months to days, or days to hours

Cineca – Largest Italian Super Computing Center

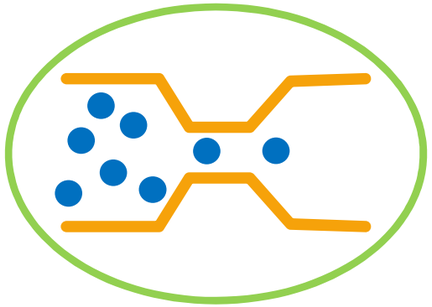
- **Goal:** Non-profit consortium of 70 Italian universities, 4 research institutions and Italian Ministry of Education desire to enable premier machine learning and AI system
- **Solution:** MARCONI – designed for advanced, scalable and energy-efficient high performance with >4000 Intel® Xeon® and Xeon Phi™ nodes implemented on Lenovo's NeXtscale platform

Barcelona Supercomputing Center – New MareNostrum 4

- **Goal:** Next generation platform and an expansion of Partnership for Advanced Computing in Europe (PRACE) high performance computing capability, servicing extensive engineering and scientific research
- **Solution:** Lenovo system with more than 3400 nodes of Intel® Xeon® processors networked with Intel® Omni-Path Architecture working alongside 3 smaller clusters

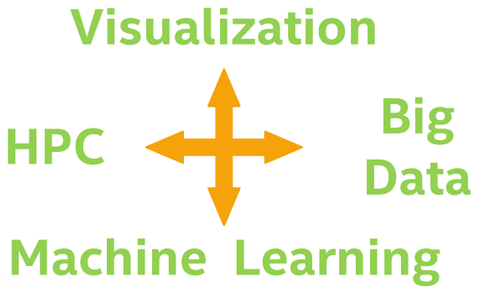
Growing Challenges in System Architecture

"The Walls" System Bottlenecks



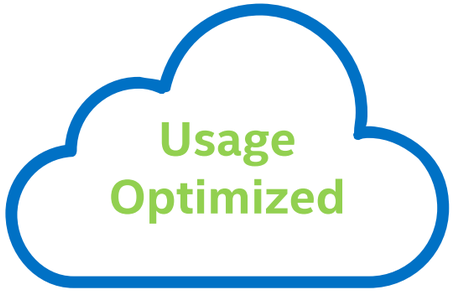
Memory | I/O | Storage Energy-Efficient Performance Space | Resiliency | Unoptimized Software

Divergent Infrastructure



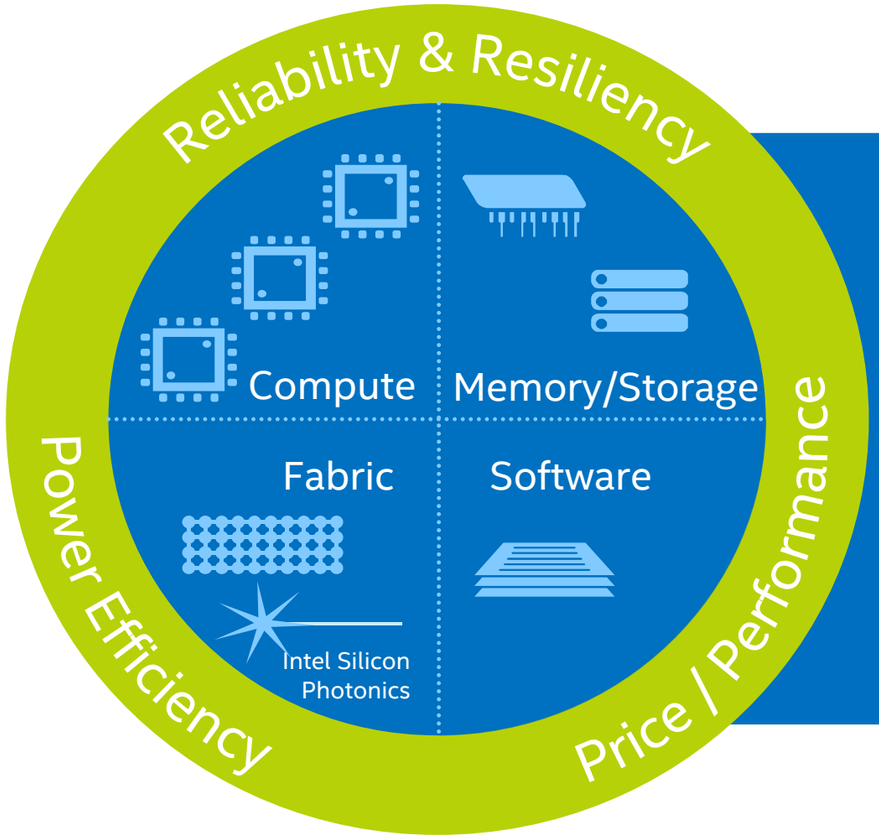
Resources Split Among Modeling and Simulation | Big Data Analytics | Machine Learning | Visualization

Barriers to Extending Usage



Democratization at Every Scale | Cloud Access | Exploration of New Parallel Programming Models

Intel® Scalable System Framework
Fuel Your Insight



Small Clusters Through Supercomputers

Compute and Data-Centric Computing

Standards-Based Programmability

On-Premise and Cloud-Based

Intel® Xeon® Processors
Intel® Xeon Phi™ Processors
Intel® Xeon Phi™ Coprocessors
Intel® Server Boards and Platforms

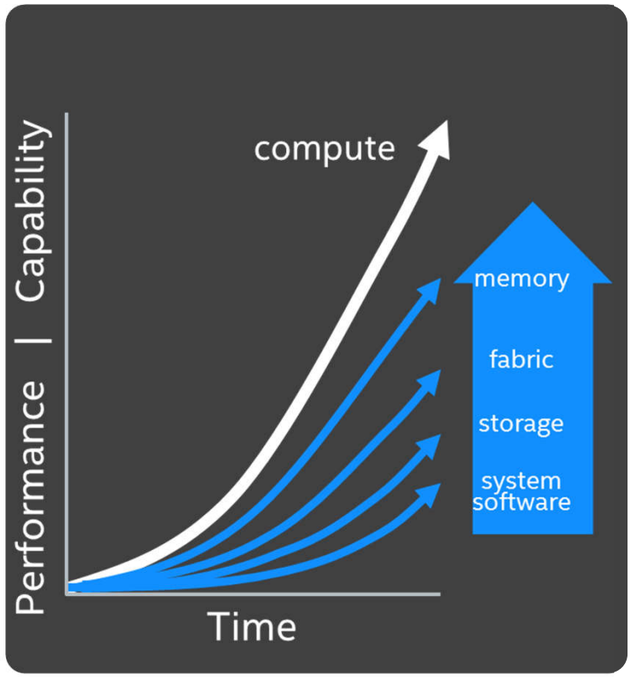
Intel® Solutions for Lustre*
Intel® Optane™ Technology
3D XPoint™ Technology
Intel® SSDs

Intel® Omni-Path Architecture
Intel® Ethernet
Intel® Silicon Photonics

Intel® HPC Orchestrator
Intel® Software Tools
Intel® Cluster Ready Program
Intel Supported SDVis

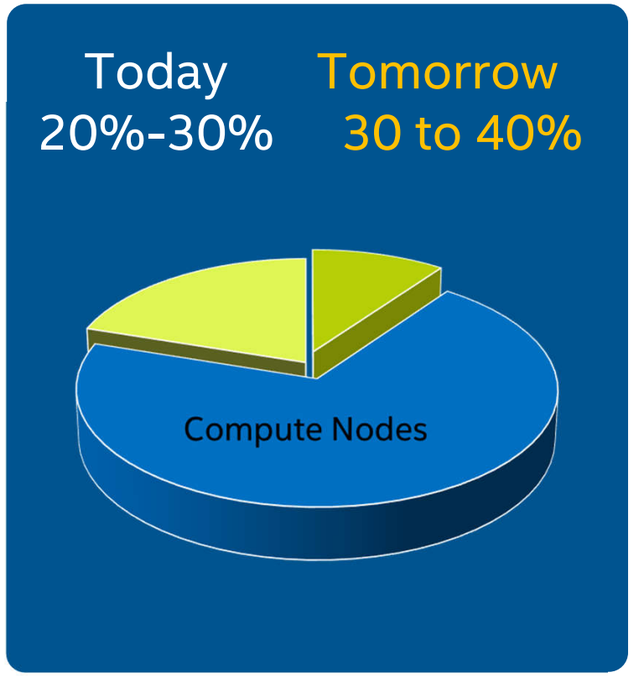
The Path to Exascale: Why Intel® OPA?

Performance



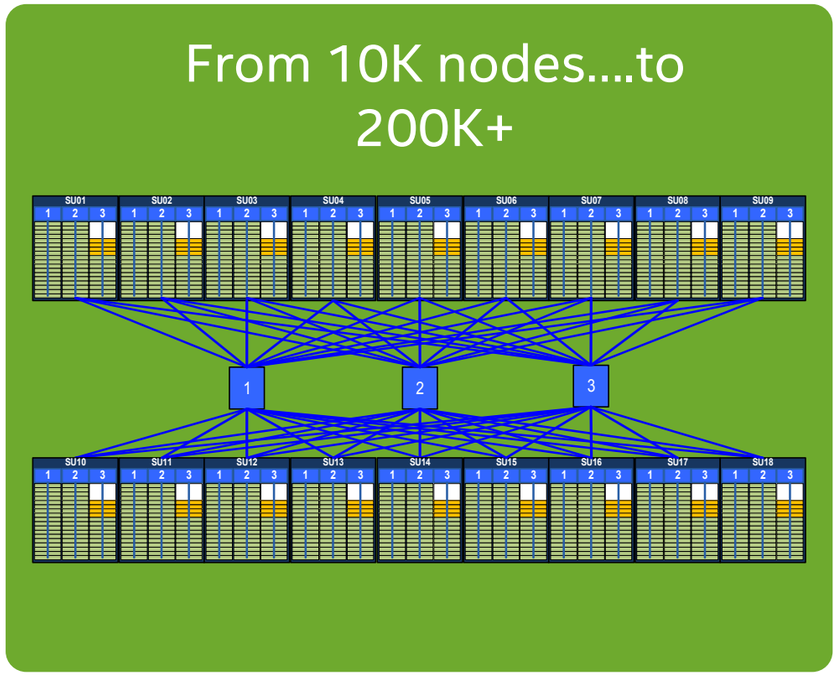
I/O struggling to keep up with CPU innovation

Fabric: Cluster Budget[1]



Fabric an increasing % of HPC hardware costs

Increasing Scale



Previous solutions reaching limits of scalability, manageability and reliability

Goal: Keep cluster costs in check → maximize COMPUTE power per dollar

1 Source: Internal analysis based on a 256-node to 2048-node clusters configured with Mellanox FDR and EDR InfiniBand products. Mellanox component pricing from www.kernelsoftware.com Prices as of November 3, 2015. Compute node pricing based on Dell PowerEdge R730 server from www.dell.com. Prices as of May 26, 2015. Intel® OPA (x8) utilizes a 2-1 over-subscribed Fabric. Intel® OPA pricing based on estimated reseller pricing using projected Intel MSRP pricing on day of launch.

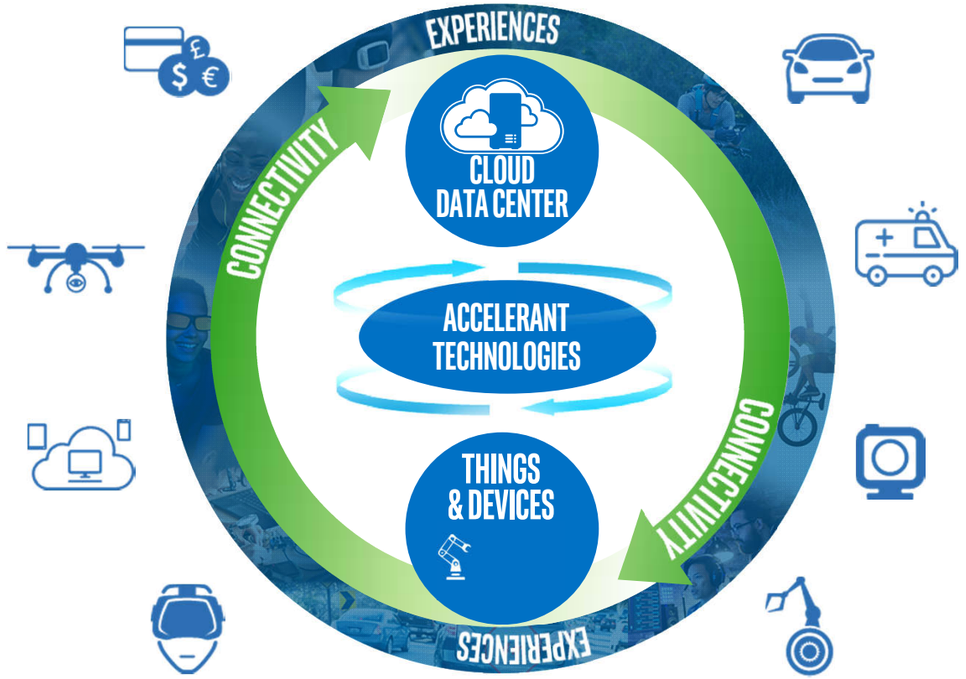
Next Up for Intel® OPA: Artificial Intelligence

Intel offers a complete AI Portfolio

- From CPUs to software to computer vision to libraries and tools

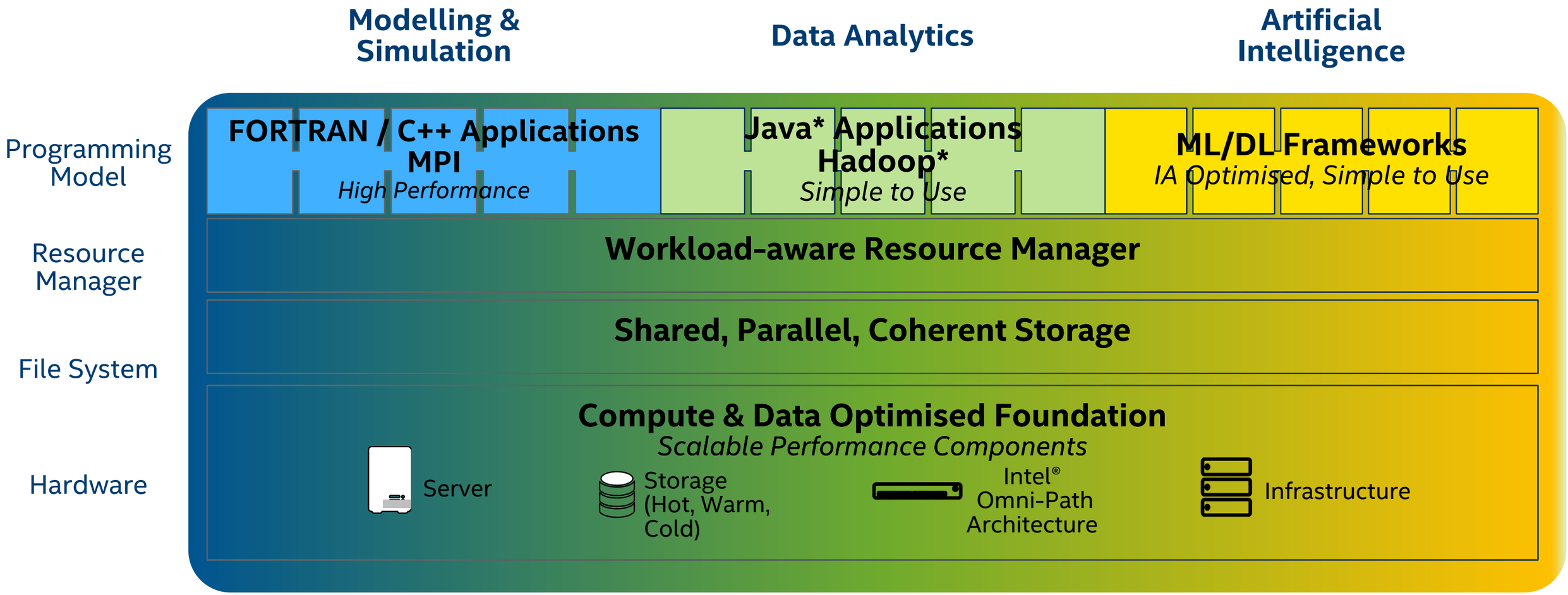
Intel® OPA offers breakthrough performance on scale-out apps

- Low latency
- High bandwidth
- High message rate
- GPU Direct RDMA support
- Xeon Phi Integration



World-class interconnect solution for shorter time to train

Converged Architecture for HPC, Analytics and AI



*Other names and brands may be claimed as the property of others

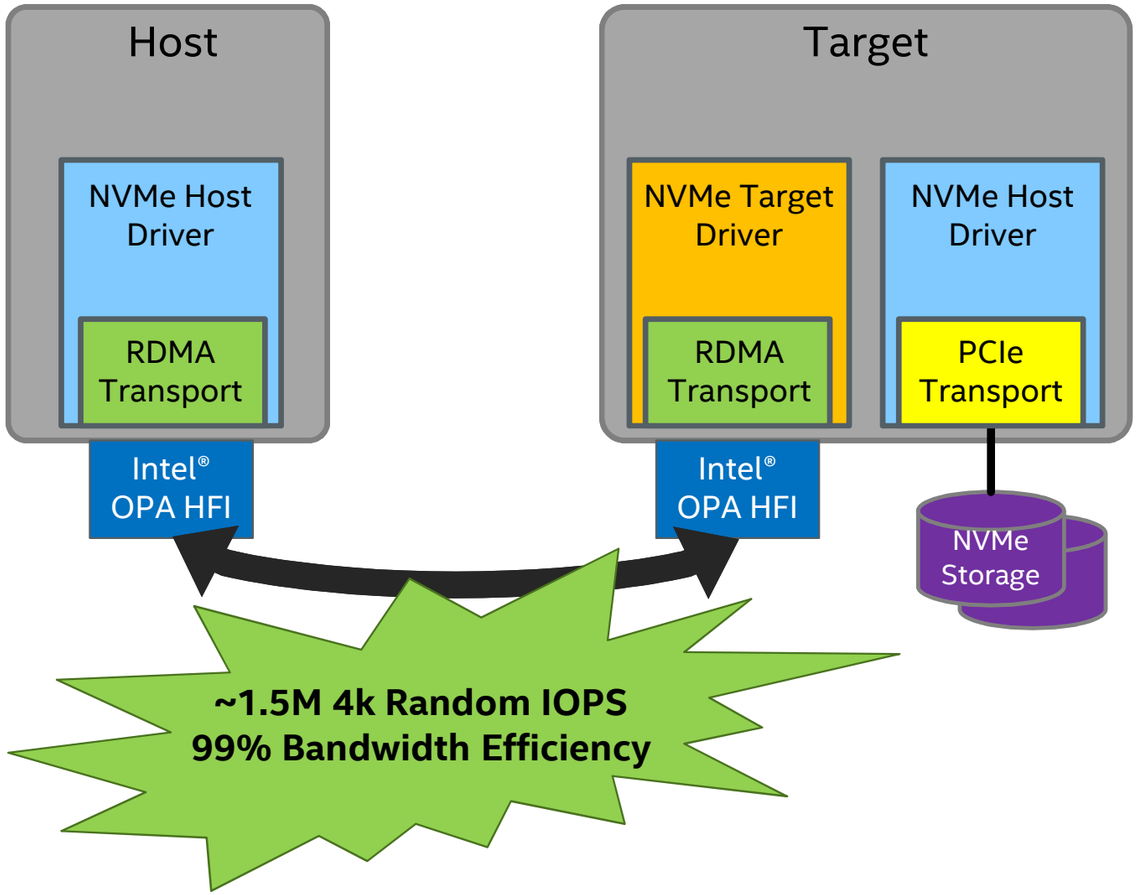
NVMe* over OPA

Intel® OPA + Intel® SSD and Optane™ Technology

- High Endurance
- Low latency
- High Efficiency
- Complete NVMe over Fabric Solution

NVMe-over-OPA status

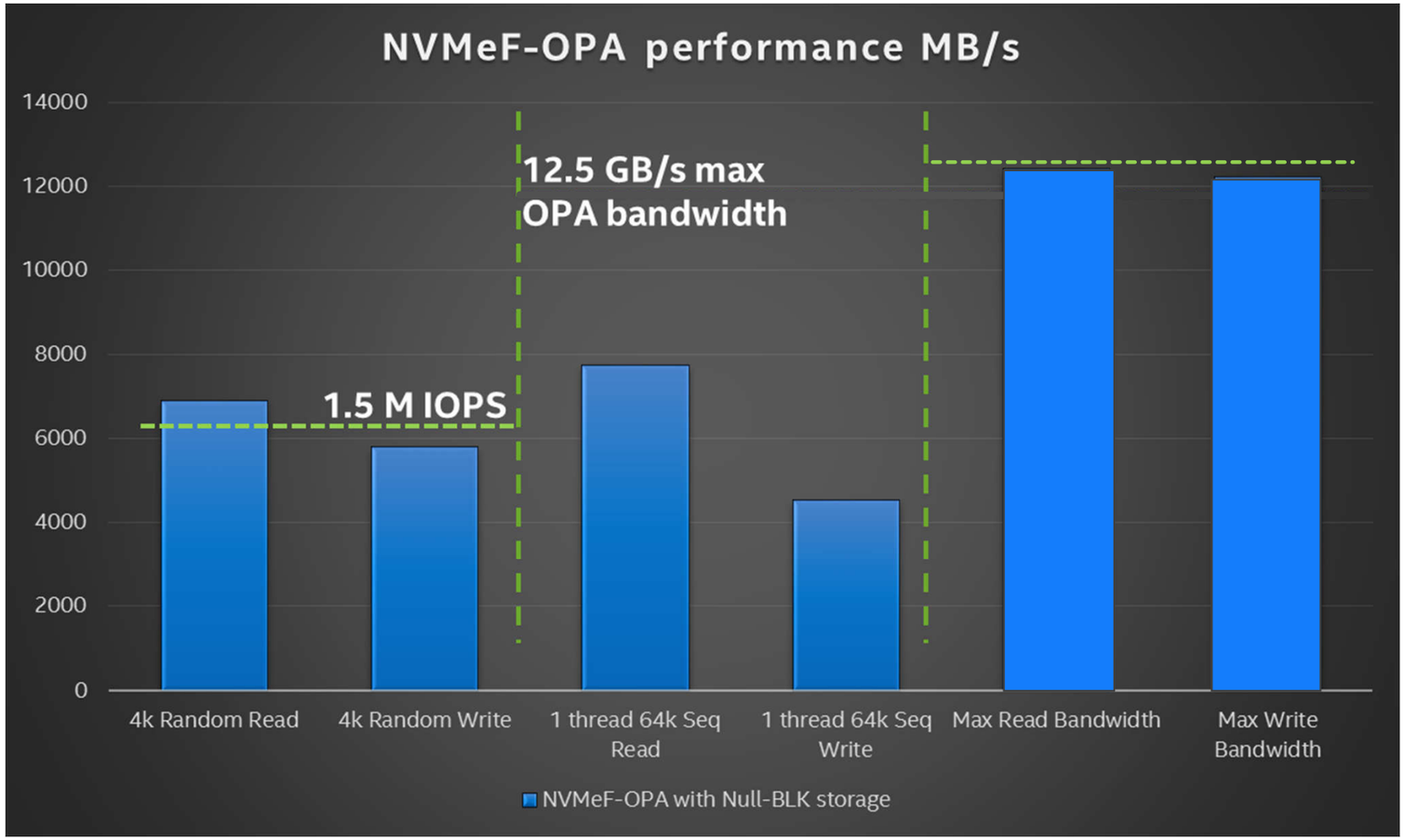
- Supported in 10.4.3 IFS release
- Compliant with NVMeF spec 1.0



Only Intel is delivering a total NVMe over Fabric solution!

Target and Host system configuration: 2 x Intel® Xeon® CPU E5-2699 v3 @ 2.30Ghz, Intel® Server Board S2600WT, 128GB DDR4, CentOS 7.3.1611, kernel 4.10.12, IFS 10.4.1, NULL-BLK, FIO 2.19 options hfi1 krcvqs=8 sge_copy_mode=2 wss_threshold=70

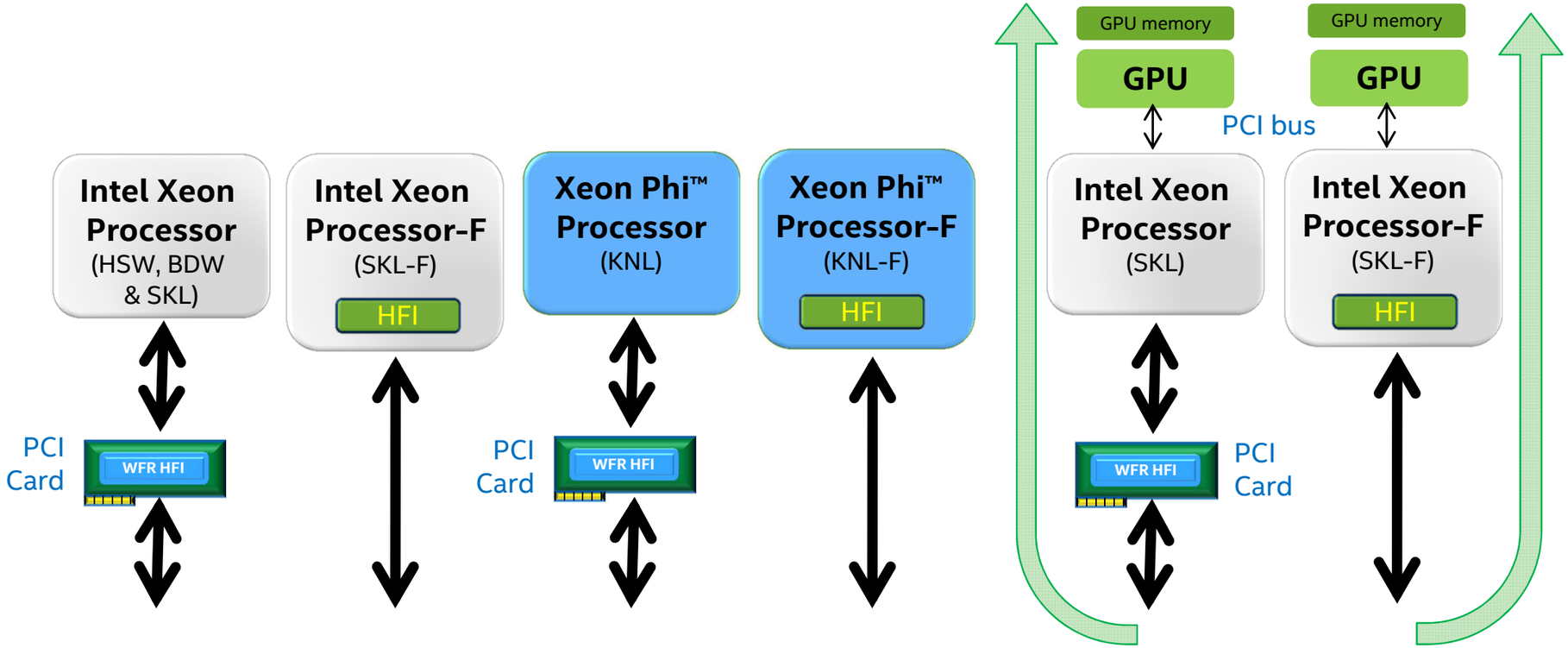
NVMe* over OPA: Initial Performance Figures



~1.5M 4k Random IOPS
99% Bandwidth Efficiency over OPA

Target and Host system configuration: 2 x Intel® Xeon® CPU E5-2699 v3 @ 2.30Ghz, Intel® Server Board S2600WT, 128GB DDR4, CentOS 7.3.1611, kernel 4.10.12, IFS 10.4.214, NULL-BLK, FIO 2.19 options hfi1 krcvqs=8 sge_copy_mode=2 wss_threshold=70
http://www.intel.com/content/www/us/en/solid-state-drives/ssd-dc-p3700-spec.html
http://www.intel.com/content/dam/www/public/us/en/documents/product-briefs/optane-ssd-dc-p4800x-brief.pdf

Maximizing Support for Heterogeneous Clusters

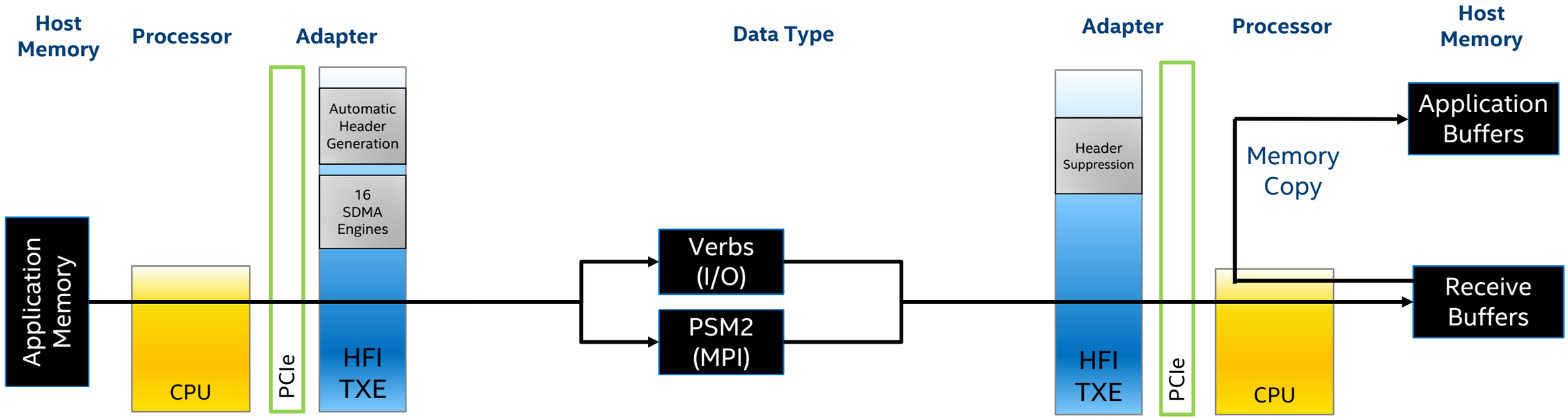


GPU Direct v3 provided in Intel® OPA 10.3 release

Greater flexibility for creating compute islands depending on user requirements

Multimodal Data Acceleration

Highest performance small message transfer: Programmed I/O



Host Driven Send

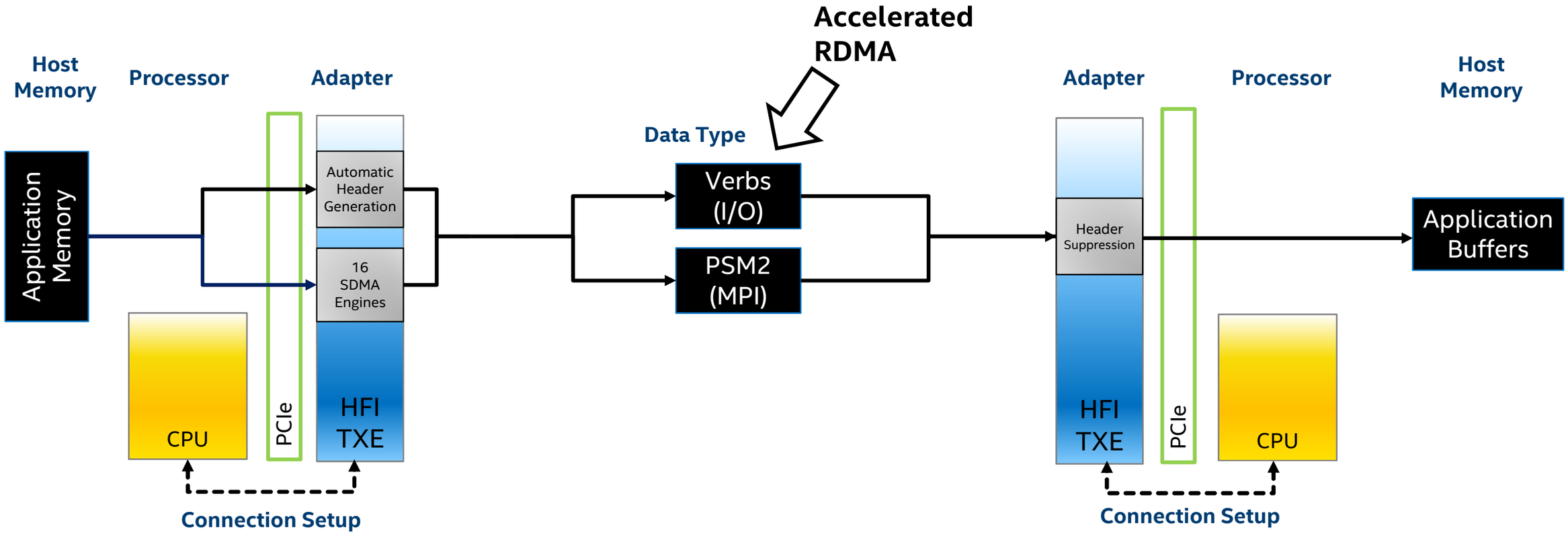
- Optimizes latency and message rate for high priority messages
- Transfer time lower than memory handle exchange, memory registration

Receive Buffer Placement

- Data placed in receive buffers
- Buffers copied to application buffer

Multimodal Data Acceleration

Lowest overhead RDMA-based large message transfer: Accelerated RDMA



Send DMA (SDMA) Engine

- Stateless offloads on send side
- DMA setup required

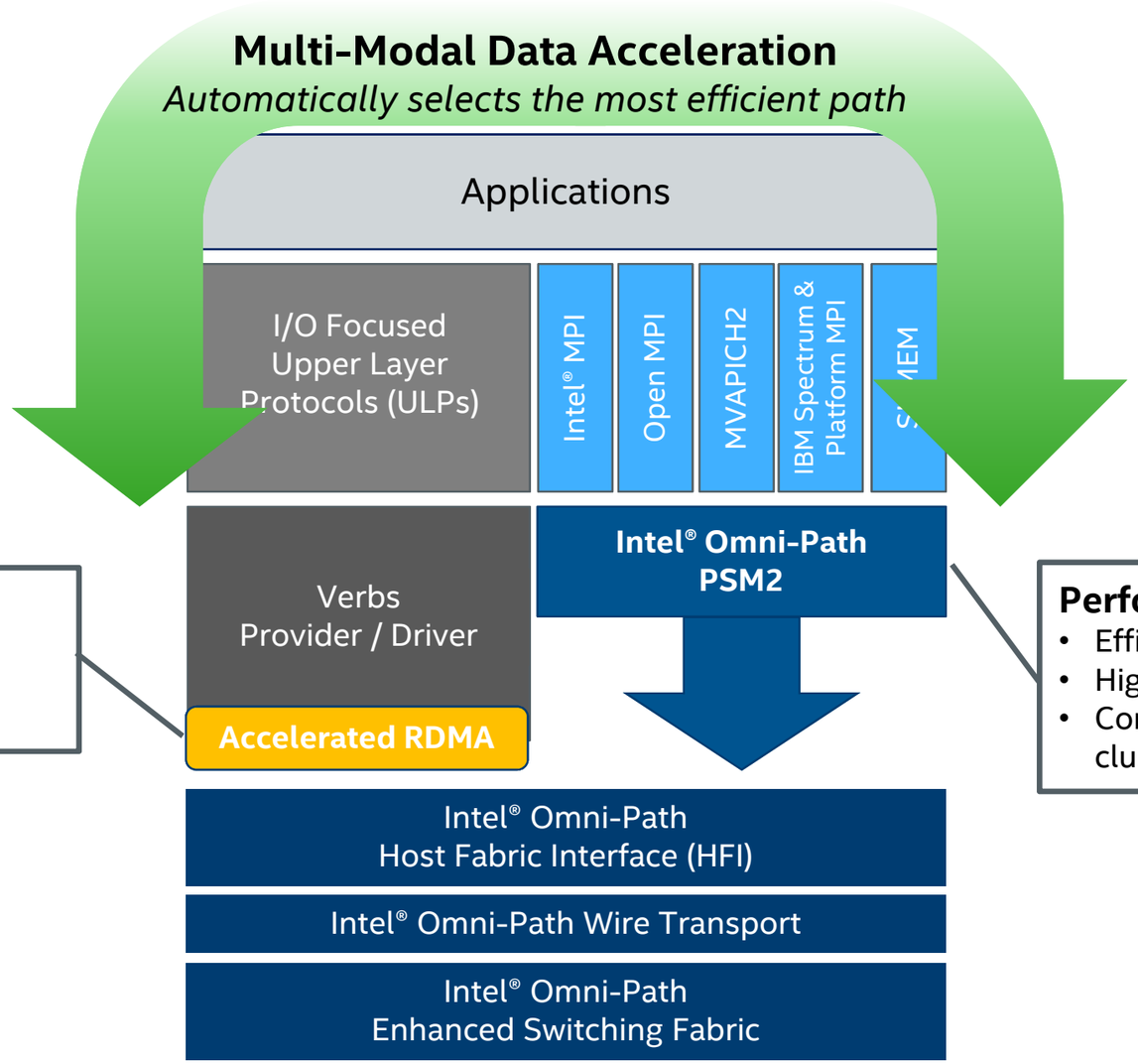
Direct Data Placement

- Direct data placement on receive side
- Eliminates memory copy

Multi-Modal Data Acceleration (MDA):
Optimizing Data Movement through the Fabric

VERBS Traffic

Large data packets
Bandwidth sensitive



MPI Traffic

Small - Med data packets
Latency & message rate sensitive

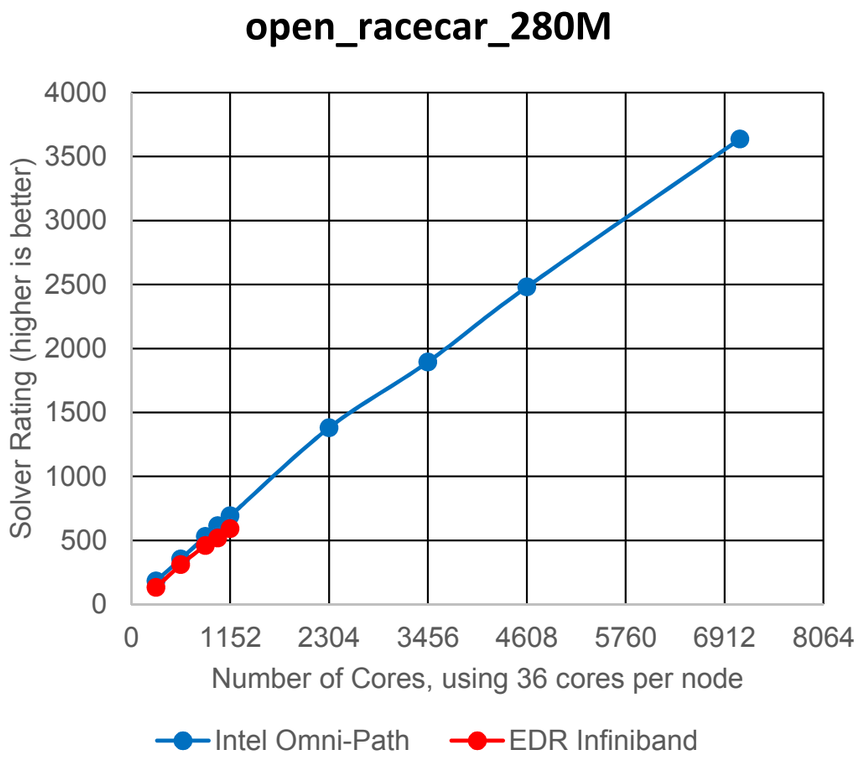
Accelerated RDMA:
Performance enhancements for large message read or writes

Performance Scaled Messaging 2 (PSM2):
- Efficient support for MPI (1/10 the code path)
- High message rate and bandwidth
- Consistent, predictable latency independent of cluster scale

Fluent R18.0 Performance on Intel® Xeon Processor and OPA

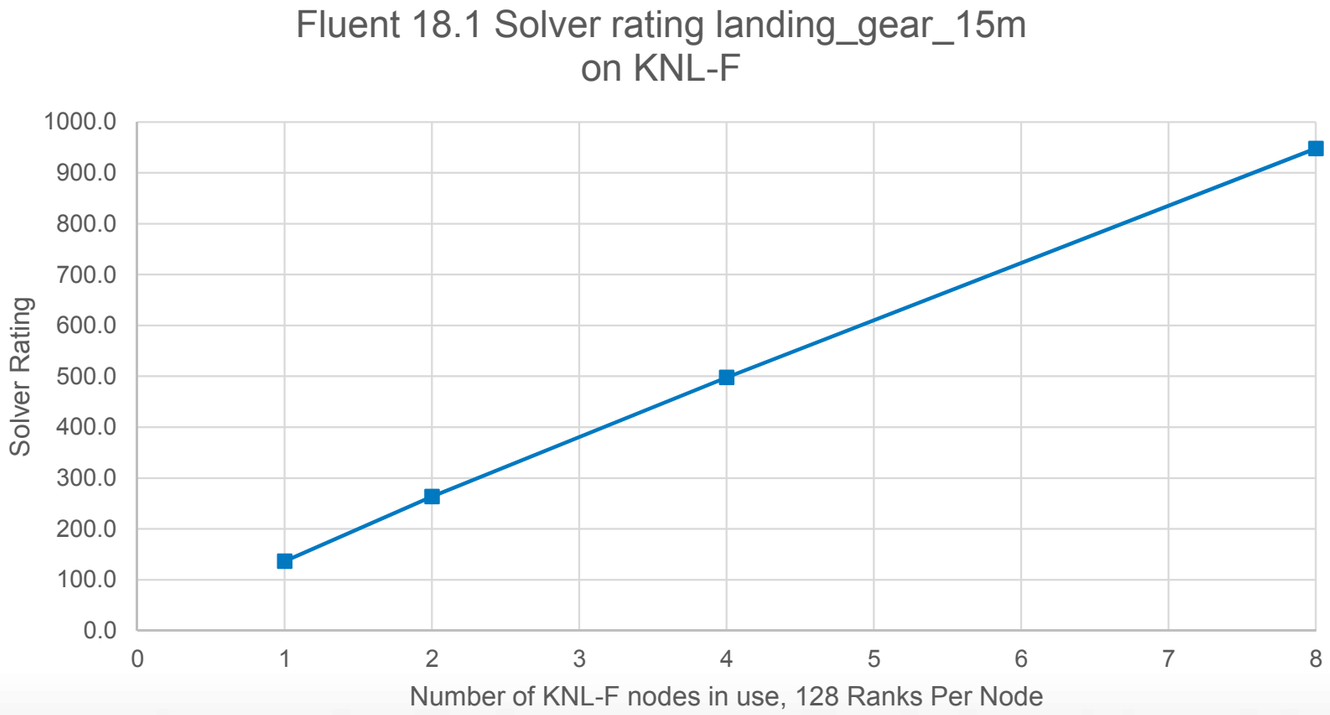
- **Fluent R18.0 performance measured using benchmark sets ranging from 2 to 14 Million cells.**
- **Intel Xeon E5 v4 processor family – up to 96 nodes (3456 cores)**
- **At lower core counts (~576 cores) the performance between Intel Omni-Path vs EDR Infiniband is comparable and at higher core counts Omni-Path outperforms by ~25-47%**

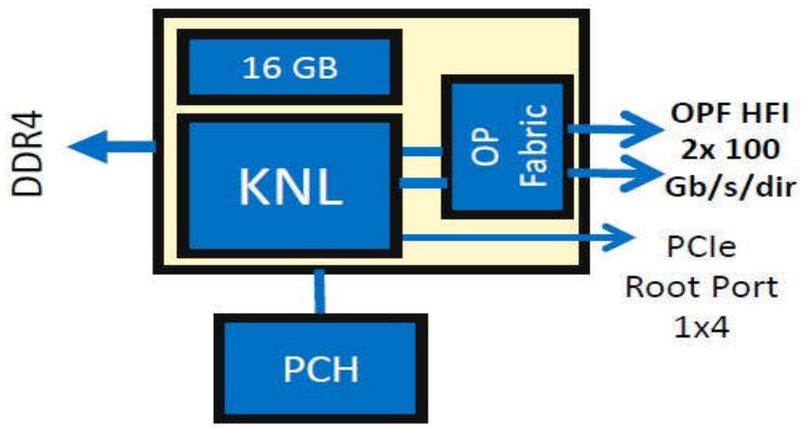
Digital Model	Relative Performance (Intel OPA/EDR Infiniband)								Performance Gain (using Intel OPA on the largest tested cluster size)
	Number of Clustered Servers (cores)								
	1 (36)	2 (72)	4 (144)	8 (288)	16 (576)	32 (1,152)	64 [a] (2,304)	96[a] (3,456)	
Pump_2m	1.00	1.0	0.98	1.02	1.04	1.16	1.30	—	30%
Rotor_3m	1.00	0.99	1.00	1.03	1.05	1.17	1.47	—	47%
Fluidized_bed_2m	1.00	1.00	0.98	1.04	1.06	1.25	—	—	25%
Sedan_4m	1.00	1.00	0.99	1.00	0.98	1.14	1.39	—	39%
Combustor_12m	1.00	1.00	0.98	1.00	1.00	1.02	1.19	1.33	33%
Aircraft_wing_14m	1.00	0.99	0.99	0.99	1.00	0.94	1.11	1.25	25%



Fluent R18.1 Performance on KNL Fabric Integrated (KNL-F)

- **Measured parallel performance up to 8 nodes of KNL with Fabric.**
- **Near linear speed up is observed for relatively larger cases (>10M cells) like combustor_12m and landing_gear_15m cases**





KNL with Omni-Path

DDR Channels: 6

MCDRAM: up to 16 GB

Gen3 PCIe (Root port): 4 lanes

Omni-Path Fabric: 200 Gb/s/dir

*Image courtesy of Intel®

Intel® Omni-Path Benchmarking Resources

Intel Internal

- Intel “Endeavour” Cluster (US) – large scale RFP support
- Intel Swindon HPC Labs (UK) – Direct end user access
- Intel “Diamond” Cluster (US) – Direct end user access

Intel Partners

- Several OEM and Integrator Partner benchmarking and solution centres;
  - HPE, Lenovo, Dell EMC, Fujitsu…
  - See Intel Fabric Builders members at https://fabricbuilders.intel.com

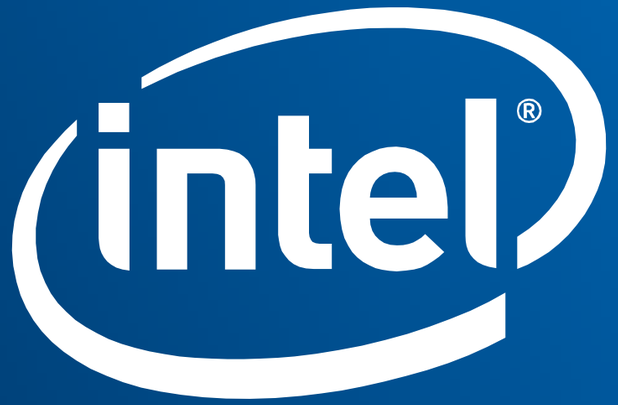
Summary

Intel® OPA continues its 100Gb HPC fabric leadership in the Top500 list

As we move to Exascale; ***Fabric Cost***, ***Error Detection/Correction*** and ***Quality of Service*** become increasingly important alongside ***Performance***.

Enhanced capabilities opening up new opportunities for greater ***Scale***, ***Performance*** and ***Efficiency***

Intel® Omni-Path Architecture is a core ingredient of Intel's Exascale strategy.



Thank You

John.Swinburne@Intel.com