INTEL® OMNI-PATH ARCHITECTURE

John Swinburne
HPC Fabric Technical Specialist
Intel Data Center Sales Group
john.swinburne@intel.com

Legal Notices and Disclaimers

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.

No computer system can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit http://www.intel.com/performance.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

Intel, the Intel logo and others are trademarks of Intel Corporation in the U.S. and/or other countries. *Other names and brands may be claimed as the property of others.

© 2016 Intel Corporation.

Optimization Notice

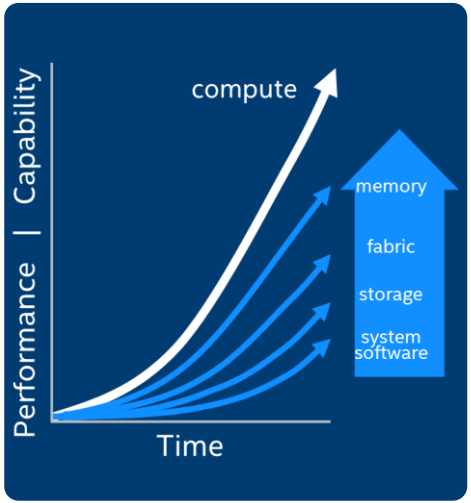
Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel.

Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice revision #20110804

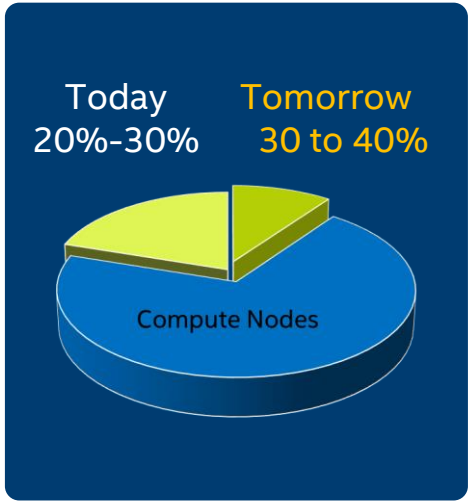
The Interconnect Landscape: **Why Intel® OPA?**

Performance



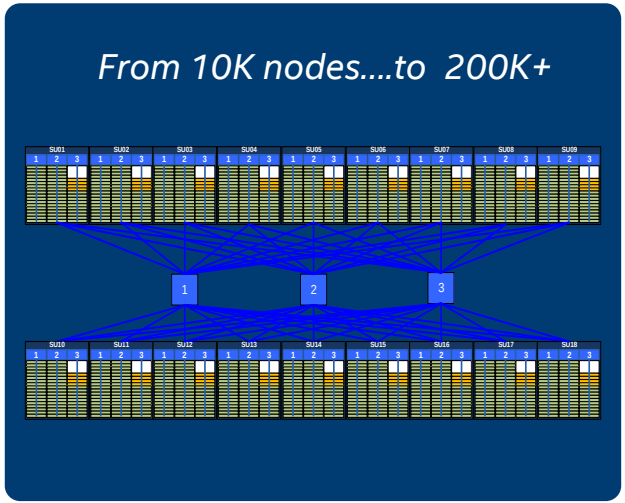
I/O struggling to keep up with CPU innovation

Fabric: Cluster Budget[1]



Fabric an increasing % of HPC hardware costs

Increasing Scale

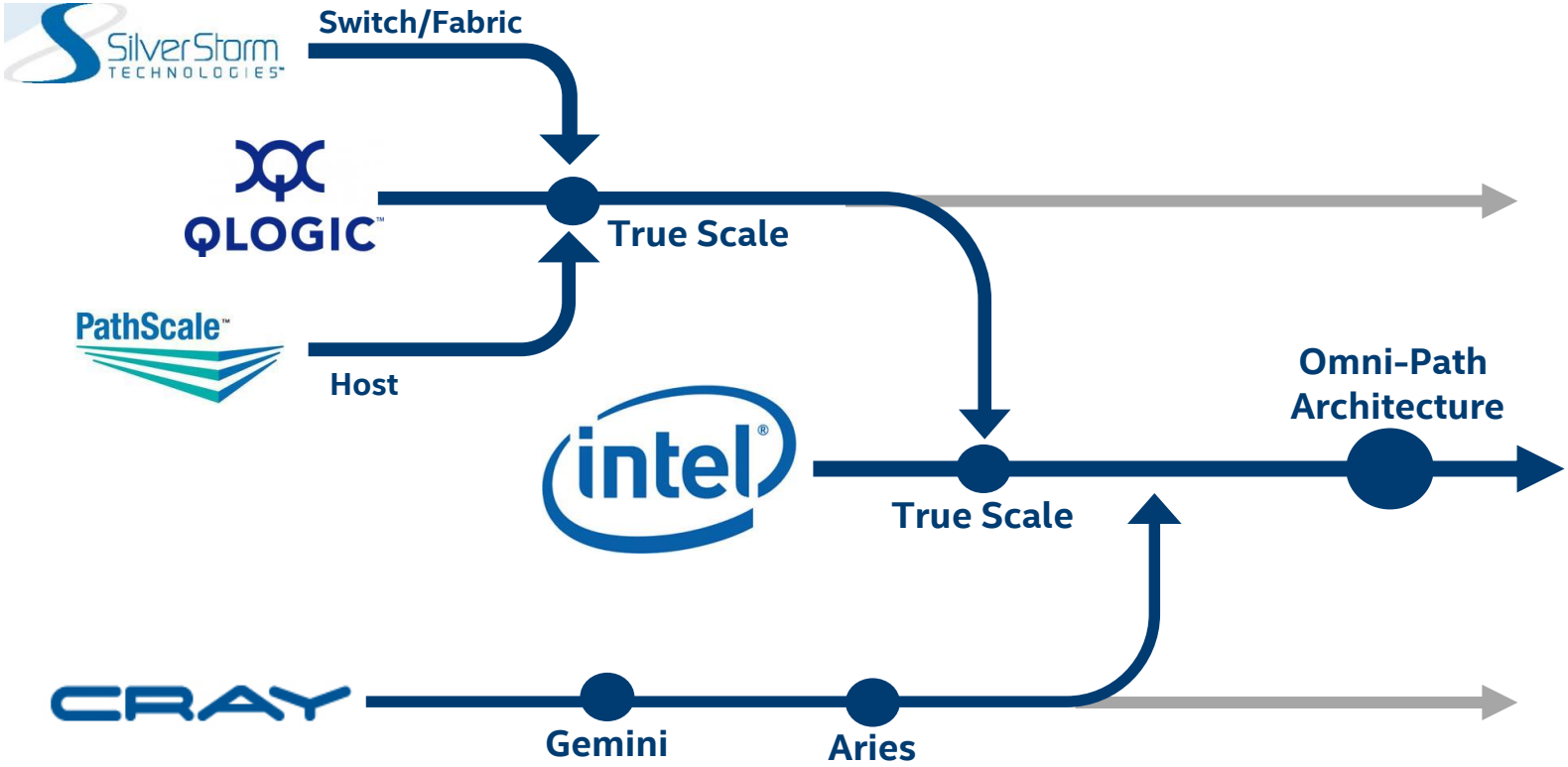


Existing solutions reaching limits

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.

A Brief History....



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

Intel® Omni-Path Architecture

Evolutionary Approach, Revolutionary Features, End-to-End Solution

HFI Adapters
Single port
x8 and x16

x16 Adapter (100 Gb/s)

x8 Adapter (58 Gb/s)

Edge Switches
1U Form Factor
24 and 48 port

48-port Edge Switch

24-port Edge Switch

Director Switches
QSFP-based
192 and 768 port

768-port Director Switch (20U chassis)

192-port Director Switch (7U chassis)

Silicon
OEM custom designs
HFI and Switch ASICs

HFI silicon Up to 2 ports (50 GB/s total b/w)

Switch silicon up to 48 ports (1200 GB/s total b/w

Software
Open Source
Host Software and Fabric Manager

Cables
Third Party Vendors
Passive Copper
Active Optical

Building on the industry's best technologies

- Highly leverage existing Aries and Intel® True Scale fabric
- Adds innovative new features and capabilities to improve performance, reliability, and QoS
- Re-use of existing OpenFabrics Alliance* software

Robust product offerings and ecosystem

- End-to-end Intel product line
- >100 OEM designs[1]
- Strong ecosystem with 70+ Fabric Builders members

[1] Source: Intel internal information. Design win count based on OEM and HPC storage vendors who are planning to offer either Intel-branded or custom switch products, along with the total number of OEM platforms that are currently planned to support custom and/or standard Intel® OPA adapters. Design win count as of November 1, 2015 and subject to change without notice based on vendor product plans. *Other names and brands may be claimed as property of others.

Intel® Omni-Path Host Fabric Interface

100 Series Single Port[1]

Low Profile PCIe Card

- 2.71”x 6.6“ max. Spec compliant.
- Standard and low profile brackets

Wolf River (WFR-B) HFI ASIC

PCIe Gen3

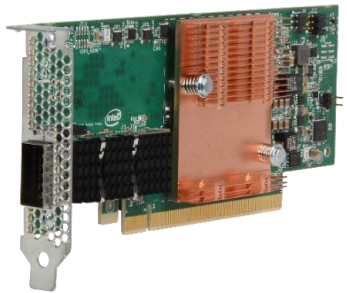
Single 100 Gb/s Intel® OPA port

- QSFP28 Form Factor
- Supports multiple optical transceivers
- Single Link status LED (Green)

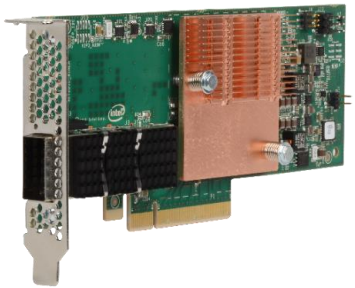
Thermal

- Passive thermal - QSFP Port Heatsink
- Standard 55C, 200lfm environment

Power	Copper		Optical (3W QSFP)	
Model	Typical	Maximum	Typical	Maximum
X16 HFI	7.4W	11.7W	10.6W	14.9W
X8 HFI	6.3W	8.3W	9.5W	11.5W



x16 HFI
(100Gb Throughput)



x8 HFI
(~58Gb Throughput)
PCIe Limited

[1]Specifications contained in public Product Briefs.

Intel® Omni-Path Edge Switch

100 Series 24/48 Port[1]

Compact Space (1U)
- 1.7”H x 17.3”W x 16.8“L

Switching Capacity
- 4.8/9.6 Tb/s switching capability

Line Speed
- 100Gb/s Link Rate

Standards-based Hardware Connections
- QSFP28

Redundancy
- N+N redundant Power Supplies (optional)
- N+1 Cooling Fans (speed control, customer changeable forward/reverse airflow)

Management Module (optional)

Power	Copper		Optical (3W QSFP)	
Model	Typical	Maximum	Typical	Maximum
24-Ports	146W	179W	231W	264W
48-Ports	186W	238W	356W	408W

24-port Edge Switch

48-port Edge Switch

[1]Specifications contained in public Product Briefs.

Intel® Omni-Path Director Class Systems

100 Series 6-Slot/24-Slot Systems[1]

Power	Copper		Optical (3W QSFP)	
Model	Typical	Maximum	Typical	Maximum
6-Slot	1.6kW	2.3kW	2.4kW	3.0kW
24-Slot	6.8kW	8.9kW	9.5kW	11.6kW

Highly Integrated

- 7U/20U plus 1U Shelf

Switching Capacity

- 38.4/153.6 Tb/s switching capability

Common Features

- Intel® Omni-Path Fabric Switch Silicon 100 Series (100Gb/s)
- Standards-based Hardware Connections – QSFP28
- Up to Full bisectional bandwidth Fat Tree internal topology
- Common Management Card w/Edge Switches
- 32-Port QSFP28-based Leaf Modules
- Air-cooled, front to back (cable side) air cooling
- Hot-Swappable Modules
  - Leaf, Spine, Management, Fan , Power Supply
- Module Redundancy
  - Management (N+1), Fan (N+1, Speed Controlled), PSU (DC, AC/DC)
- System Power : 180-240AC

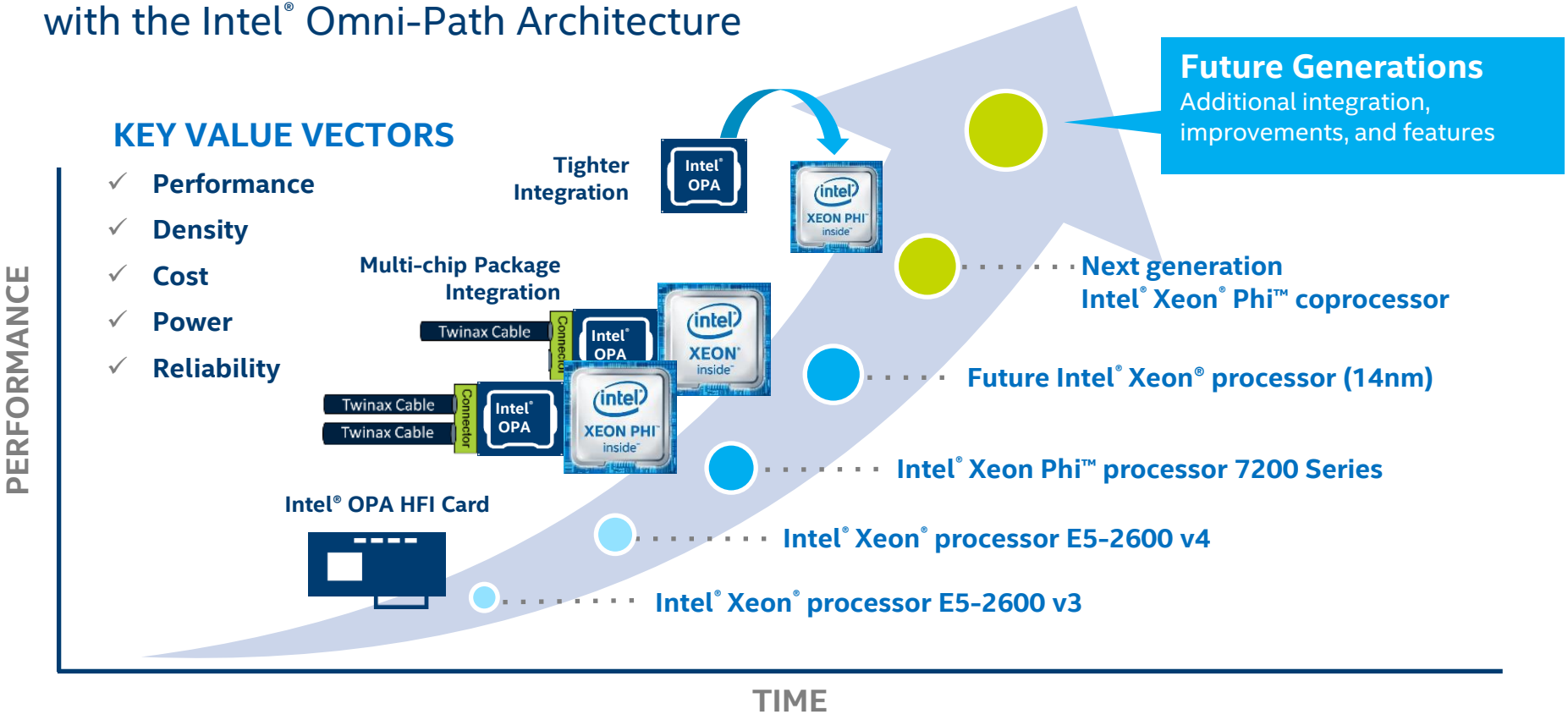
6-Slot Director Switch

24-Slot Director Switch

[1]Specifications contained in public Product Briefs.

CPU-Fabric Integration

with the Intel® Omni-Path Architecture



Intel® OPA Software Stack

Performance Scaled Messaging designed for HPC

Carefully selected division of responsibility

- MPI handles higher level capabilities (includes a wide array of MPI and user-facing functions)
- PSM focuses on HPC interconnect (optimized data movement, MPI performance, QoS, dispersive routing, resiliency)

Binary Compatible

- Common base architecture – PSM2 is a superset of PSM
- Applications built and tuned for PSM will just work

Connectionless with minimal on-adapter state

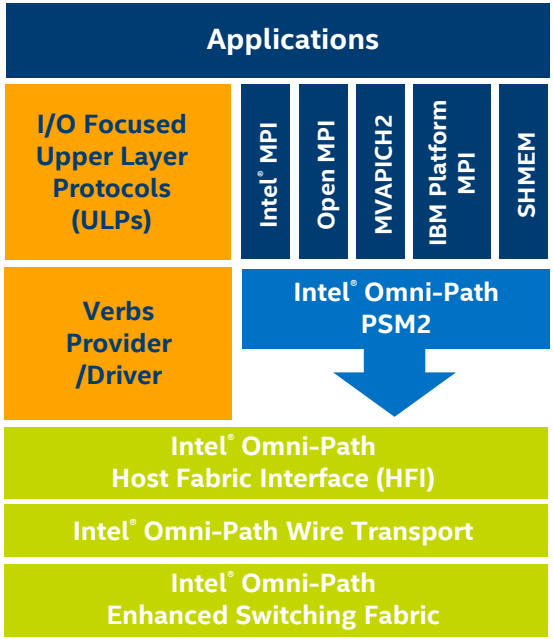
- No cache misses as the fabric scales

High MPI message rate – short message efficiency

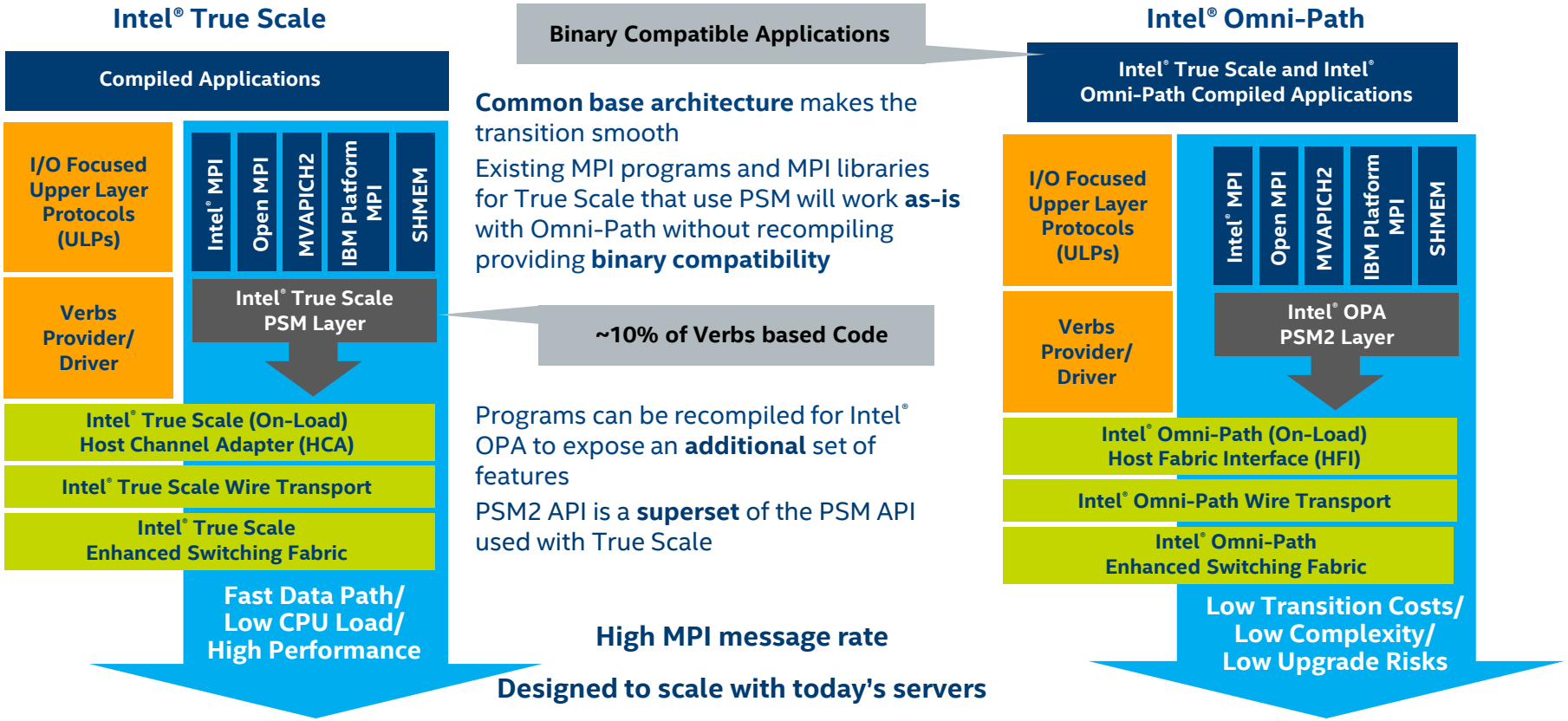
Designed to scale with today's servers

- Dense multi-core/multi-socket CPUs
- Fast processors, high memory bandwidth, more cores

Designed for Performance at Extreme Scale



Generational Software Compatibility



Intel® Omni-Path Software Strategy

- **Leverage OpenFabrics Alliance (OFA) interfaces so InfiniBand applications "just work"**
- **Open source all host components in a timely manner**
  - Changes pushed up stream in conjunction with Delta Package release
- **"Inbox" with future Linux OS releases**
  - RHEL, SLES and OFED (standalone distribution from OFA)
- **Deliver delta package that layers on top of the OS**
  - Updates before they are available inbox
  - Only change what's necessary. This isn't a complete distribution!
  - Delta packages will support N and N-1 versions of RHEL and SLES
  - Delta Packages available on Intel® Download Center
- Note: Intel-OFED only layers necessary changes **on top** of existing installations to reduce risk of compatibility issues with other interconnects.

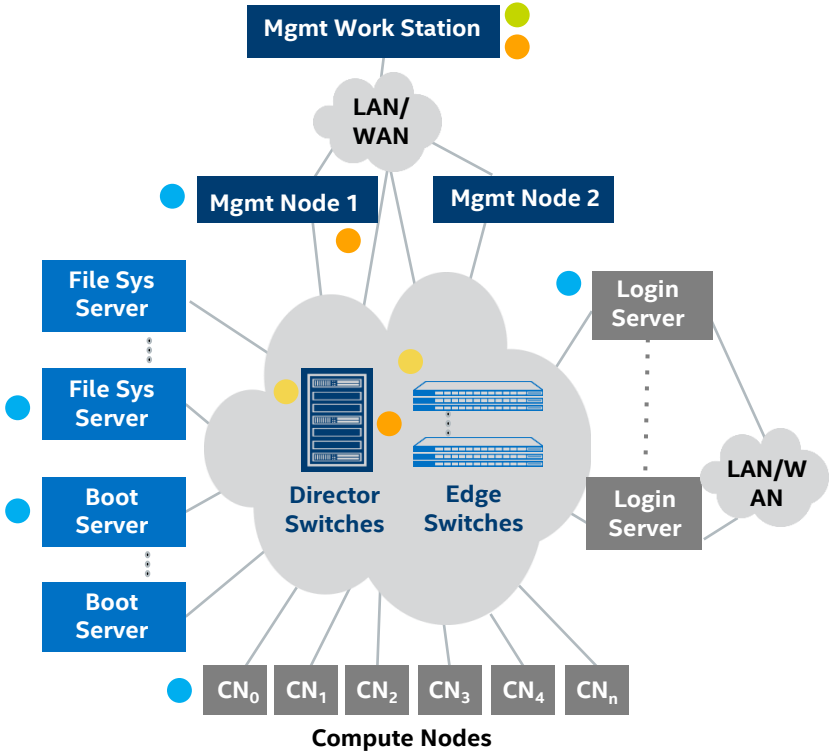
Comprehensive Intel® Omni-Path Software Solution

Element Management Stack

- "Traditional System Mgmt" included in all managed switches
- Functions: Signal integrity, Thermal, Errors

Host Software Stack

- Runs on all Intel® OP connected nodes
- High performance, highly scalable MPI implementation via PSM and extensive set of upper layer protocols
- Boot over Network



Fabric Management GUI

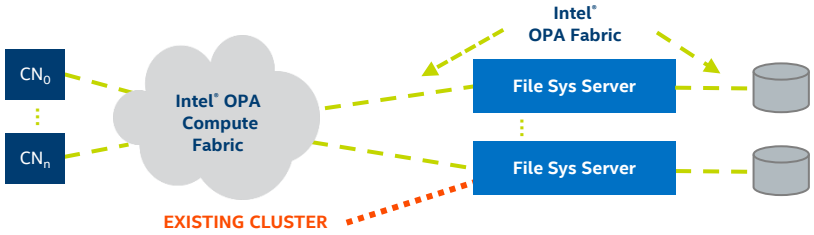
- Interactive GUI access to Fabric Management TCO features
- Configuration, monitoring, diags, element mgmt drill down

Fabric Management Stack

- Runs on Intel OP connected management nodes or switches
- Creates the "engineered topology" and controls flow of traffic in fabric
- Includes toolkit for TCO functions: Configuration, monitoring, diags, and repair

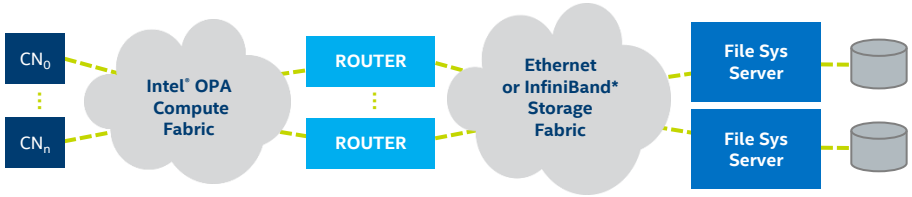
Intel® OPA Storage Solutions

Direct Attach to File System Server



- Engaging key HPC storage vendors to deliver Intel® OPA-based storage devices (new storage)
- Direct-attach Intel® OPA to existing file system servers -"dual-homed" approach

Routing through Ethernet/InfiniBand* Storage Fabrics



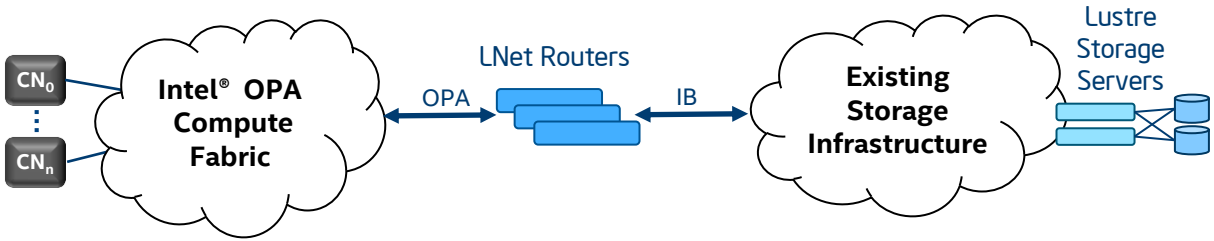
- **Lustre:** Supported via LNET Router
- **GPFS/NFS:** Supported via IP Router

Intel enabling plans:

- *Provide SW for LNET and IP router*
- *HW requirement specifications*
- *Documentation/user guides*

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

LNET Router with Intel® OPA



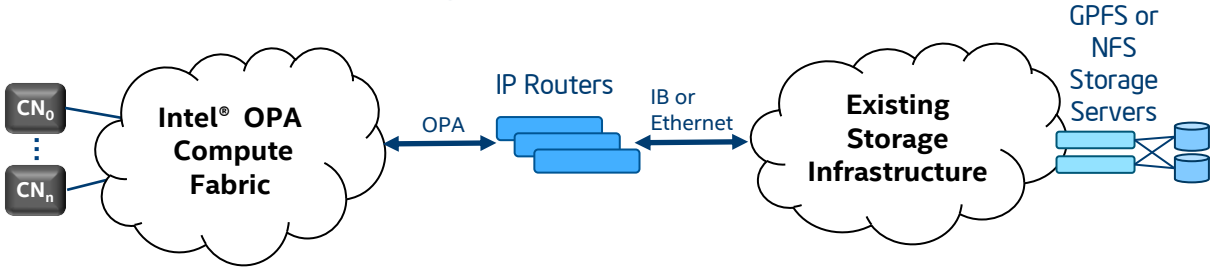
LNet Router Solution

- *Intel providing recipe for router solution*
- *HW requirement specifications & performance projections*
- *Documentation / design guide*

Software Stack	Configurations	Support
•Uses standard LNet router component in Lustre •IEEL version 2.4 or newer •RHEL and SLES solutions available	•Solutions for IB fabrics (QDR, FDR) •High availability solutions •Load balanced for performance	•Provided via Linux Distribution and Lustre

"Implementing Storage in Intel® Omni-Path Architecture Fabrics" white paper available now (*public link*)
"Intel® Omni-Path Storage Router Design Guide" available now (*public link*)

IP Router with Intel® OPA



IP Router Solution

- *Intel providing recipe for router solution*
- *HW requirement specifications & performance projections*
- *Documentation / design guide*

Software Stack	Configurations	Support
• Uses standard IP router available in Linux Distros • RHEL and SLES solutions available	• Solutions for IB and Ethernet fabrics • High availability solutions • Load balanced for performance	• Provided via Linux Distribution

"Implementing Storage in Intel® Omni-Path Architecture Fabrics" white paper available now (*public link*)
"Intel® Omni-Path Storage Router Design Guide" available now (*public link*)

Intel® OPA Industry **Momentum is Picking Up**

OEM Momentum	**Over 100** OEM and HPC storage vendor offerings expected for platforms, switches, and adapters[1]
Ecosystem Momentum	On track for robust hardware and software ecosystem at launch, **with 80+ members** in the Intel® Fabric Builders program[3]

1 Source: Intel internal information. Design win count based on OEM and HPC storage vendors who are planning to offer either Intel-branded or custom switch products, along with the total number of OEM platforms that are currently planned to support custom and/or standard Intel® OPA adapters. Design win count as of November 1, 2015 and subject to change without notice based on vendor product plans.
2 Expected membership in the Intel® Fabric Builders program at launch in Q4'15. Updated list of members can be found on our website (https://fabricbuilders.intel.com)
*Other names and brands may be claimed as property of others.

Intel® OPA's Industry Momentum is Picking Up:

Impact on the November Top500 List

Intel® OPA compared to InfiniBand* EDR (100Gb Fabrics)

- Share of clusters → 2x (OPA 28 vs EDR 14)
- Share of Flops → 2.5x (43.7PF vs 17.1PF)
- Top10 → 1 system
- Top15 → 2 system
- Top50 → 4 vs 2 systems
- Top100 → 10 vs 4 systems
- Xeon Efficiency → OPA 88.5% vs. EDR 83.7%

✓ *Momentum*

✓ *Performance*

✓ *Scalability*

✓ *Stability*

Top500 and Major Deployments

NNSA National Nuclear Security Administration · The University of Iowa · ANI · CINECA · JCAHPC · MIT Lincoln Laboratory Supercomputing Center · Pittsburgh Supercomputing Center · University of Washington · Uni Freiburg · University of Bristol · Dell EMC · 東京大学 The University of Tokyo · 筑波大学 University of Tsukuba · The Thomas Young Centre · MET Malaysia · RDI² Rutgers Discovery Informatics Institute · University of Pittsburgh

SCALABLE AND FLEXIBLE

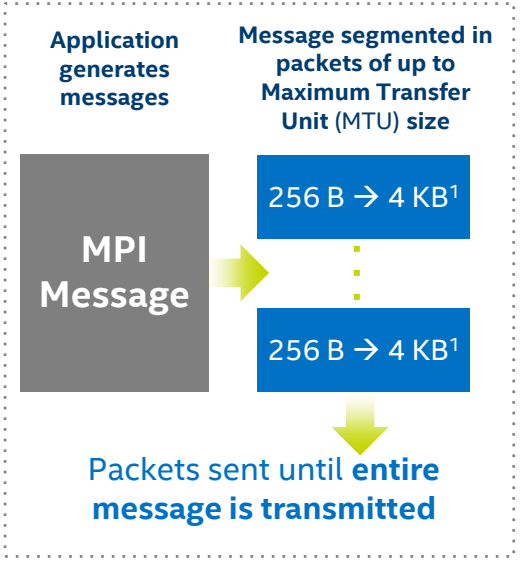
New Intel® OPA Fabric Features: Fine-grained Control Improves Resiliency and Optimizes Traffic Movement

	Description	Benefits
Traffic Flow Optimization	▪ Optimizes Quality of Service (QoS) in mixed traffic environments, such as storage and MPI ▪ Transmission of lower-priority packets can be paused so higher priority packets can be transmitted	▪ Ensures high priority traffic is not delayed →Faster time to solution ▪ Deterministic latency → Lowers run-to-run timing inconsistencies
Packet Integrity Protection	▪ Allows for rapid and transparent recovery of transmission errors on an Intel® OPA link without additional latency ▪ Resends 1056-bit bundle w/errors only instead of entire packet (based on MTU size)	▪ Fixes happen at the link level rather than end-to-end level ▪ Much lower latency than Forward Error Correction (FEC) defined in the InfiniBand* specification[1]
Dynamic Lane Scaling	▪ Maintain link continuity in the event of a failure of one of more physical lanes ▪ Operates with the remaining lanes until the failure can be corrected at a later time	▪ Enables a workload to continue to completion. **Note:** InfiniBand will shut down the entire link in the event of a physical lane failure

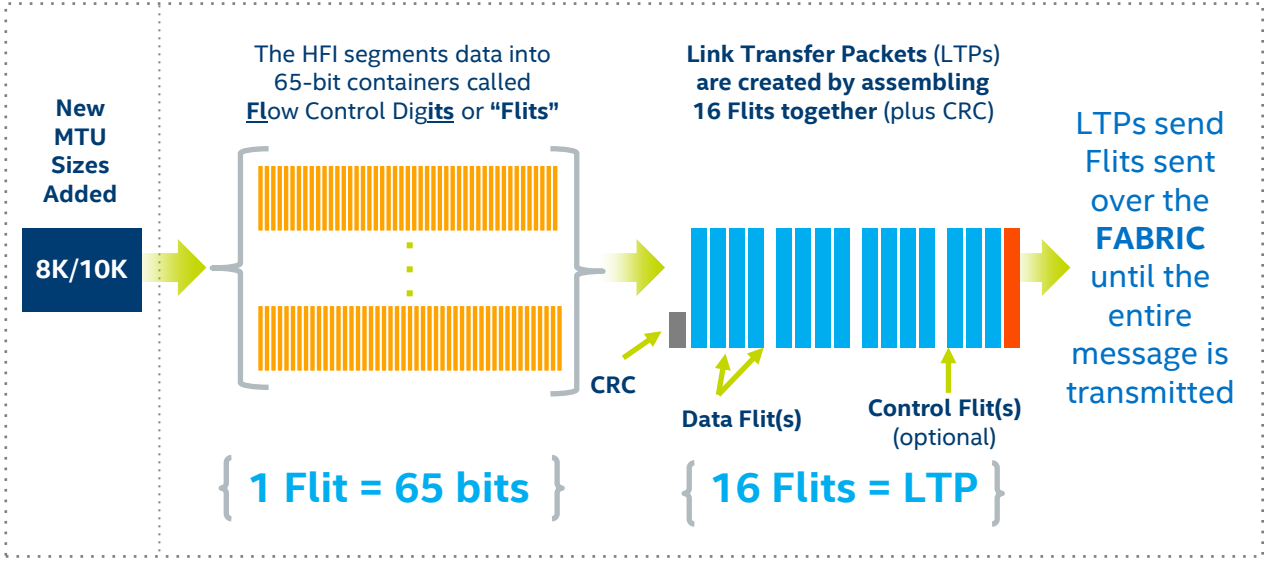
[1] Lower latency based on the use of InfiniBand with Forward Error Correction (FEC) Mode A or C in the public presentation titled "Option to Bypass Error Marking (supporting comment #205)," authored by Adee Ran (Intel) and Oran Sela (Mellanox), January 2013. Mode A modeled to add as much as 140ns latency above baseline, and Mode C can add up to 90ns latency above baseline. Link: www.ieee802.org/3/bj/public/jan13/ran_3bj_01a_0113.pdf

Intel® OPA Link Level Innovation Starts Here

InfiniBand*



Intel® Omni-Path Fabric



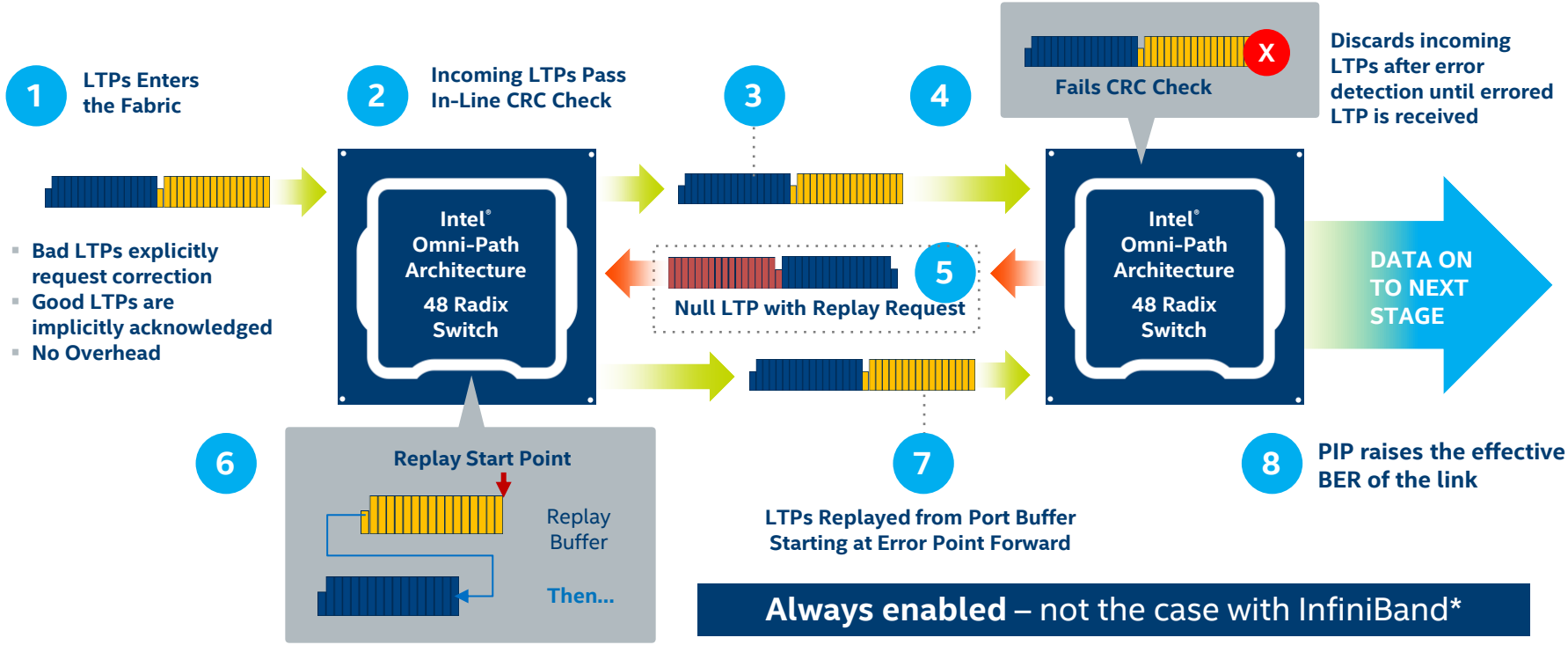
[1] Intel® OPA supports up to 8KB for MPI Traffic and10KB MTU for Storage

CRC: Cyclic Redundancy Check

Goals: Improved resiliency, performance, and consistent traffic movement

Packet Integrity Protection (PIP)

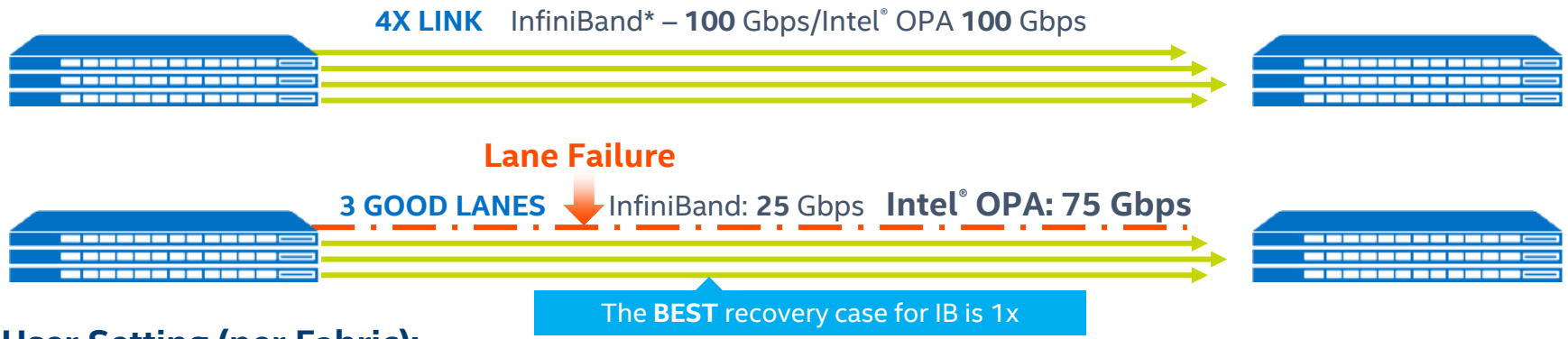
Intel® Omni-Path Fabric Link Level Innovation



Always enabled – not the case with InfiniBand*

Dynamic Lane Scaling (DLS) Traffic Protection

Intel® Omni-Path Fabric Link Level Innovation



User Setting (per Fabric):

- Set maximum degrade option allowable
  - 4x – Any lane failure would cause link reset or take down
  - 3x – Still operates at degraded bandwidth (75 Gbps)
  - 2x – Still operates at degraded bandwidth (50 Gbps)
  - 1x – Still operates at degraded bandwidth (25 Gbps)

Link Recovery:

- PIP is used to recover link without reset – An Intel® OPA innovation

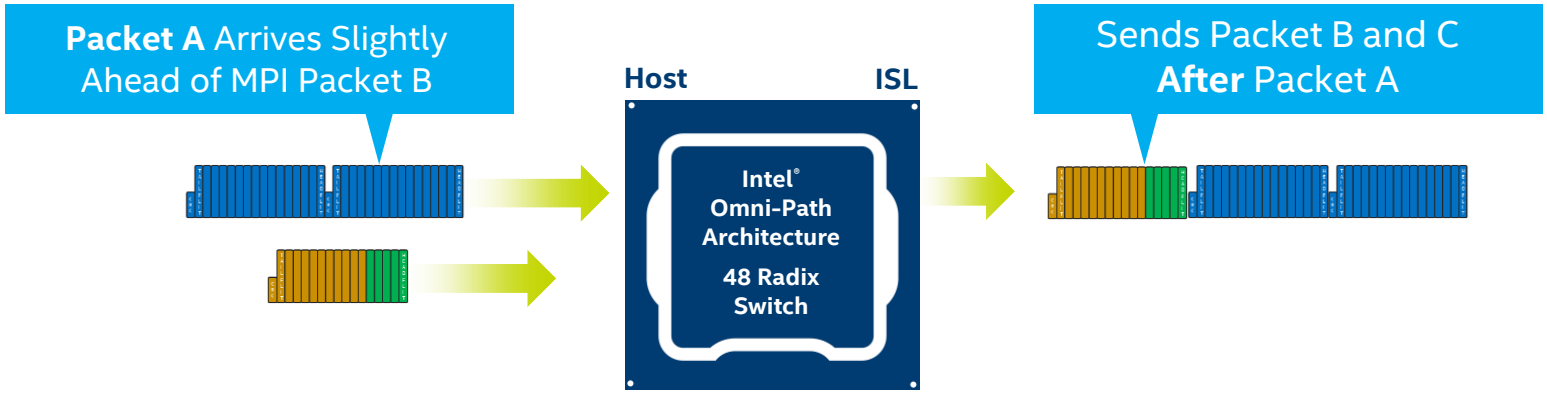
Intel® OPA still passing data at reduced bandwidth with link recovery via PIP

InfiniBand* may close entire link or reinitialize @1x introducing fabric delays or routing issues

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. *Other names and brands may be claimed as property of others.

Traffic Flow Optimization (TFO) – Disabled

Intel® Omni-Path Fabric Link Level Innovation



Packet A (VL0)
Storage Traffic

Packet B (VL0)
MPI Traffic

Packet C (VL0)
MPI Traffic

Configured for Same Priority

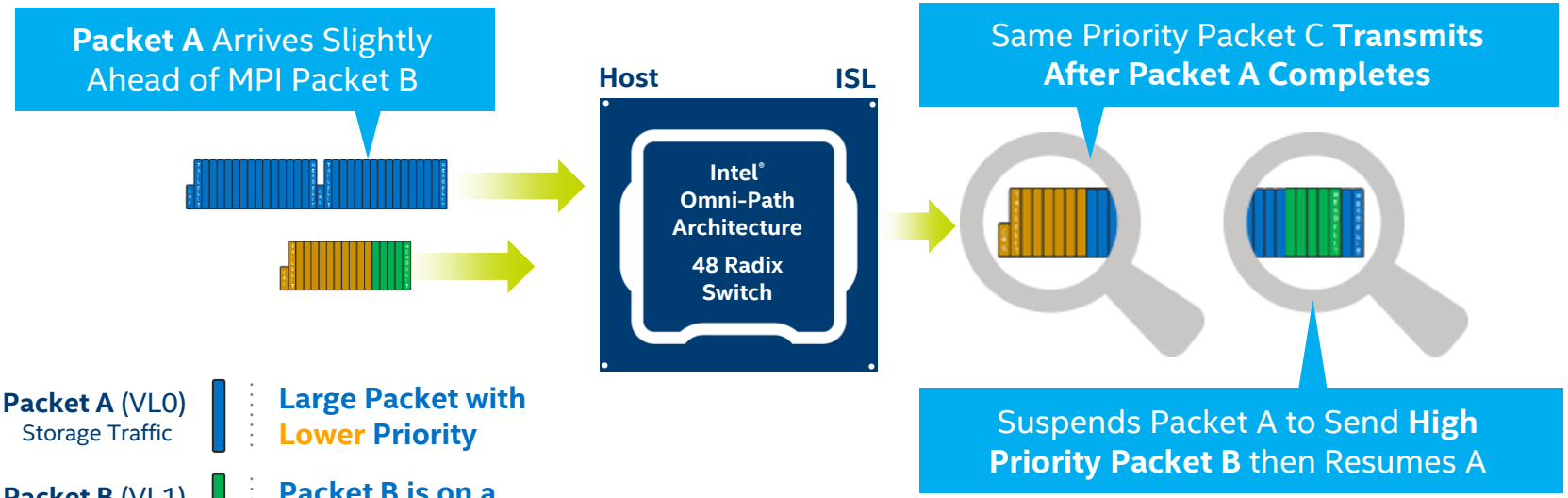
Standard InfiniBand* operation

VL = Virtual Lane (Each Lane Has a Different Priority)

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. *Other names and brands may be claimed as properties of others.

Traffic Flow Optimization (TFO) – Enabled

Intel® Omni-Path Fabric Link Level Innovation

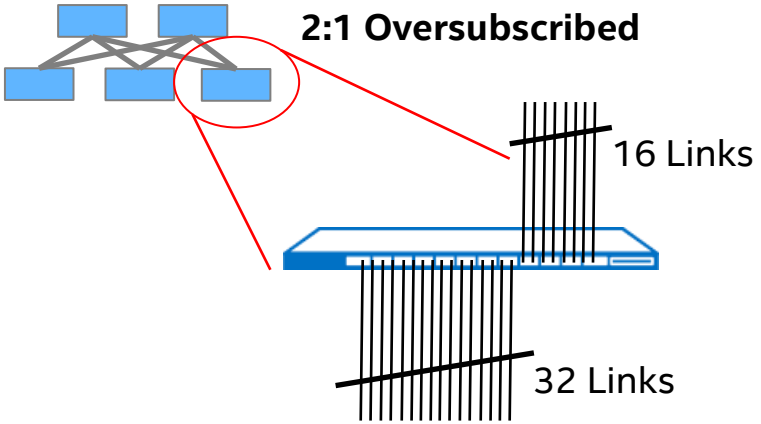


VL = Virtual Lane (Each Lane Has a Different Priority)

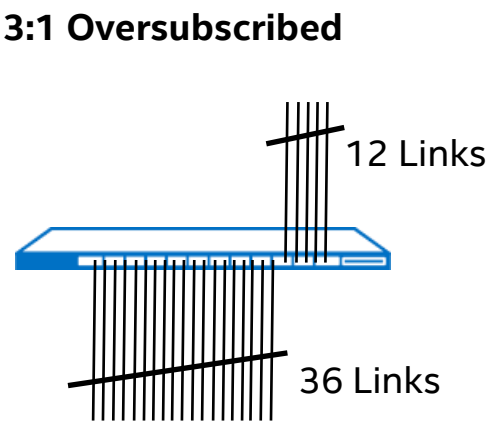
Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration.

OPTIMIZED COSTS

Increased capacity vs 36 port switch



CPU	# Cores	% vs 36p
E5-2650 v4	768	+33%
E5-2697 v4	1152	

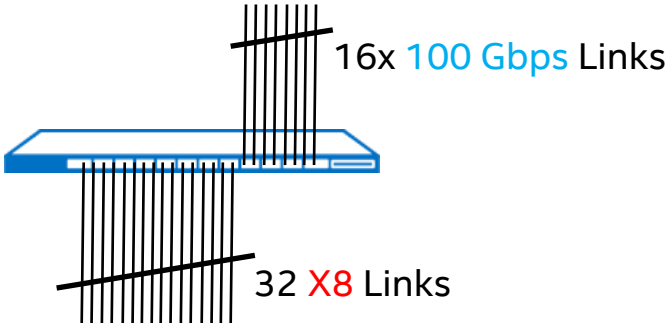


CPU	# Cores	% vs 36p
E5-2650 v4	864	+33%
E5-2697 v4	1296	

Increased flexibility with X8 HFA

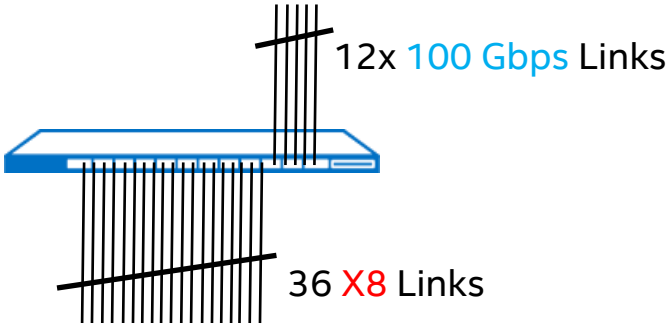
- Intel OPA Links negotiate at 100Gbps
- X8 HFA links limited by PCIe Gen3 X8 slot (56-58 Gbps depending on protocol/encoding)

2:1 Port Oversubscribed



Aggregate Host BW	Aggregate ISL BW	Effective BW Oversub
Up to 1792 Gbps	1600 Gbps	1.12:1

3:1 Port Oversubscribed



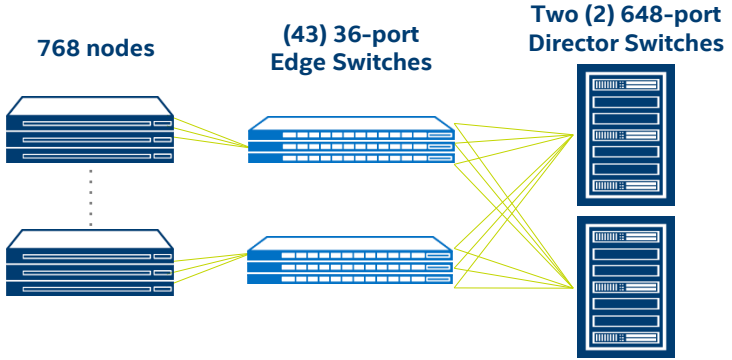
Aggregate Host BW	Aggregate ISL BW	Effective BW Oversub
Up to 2016 Gbps	1200 Gbps	1.68:1

Intel® Omni-Path Fabric's 48 Radix Chip

It's more than just a 33% increase in port count over a 36 Radix chip

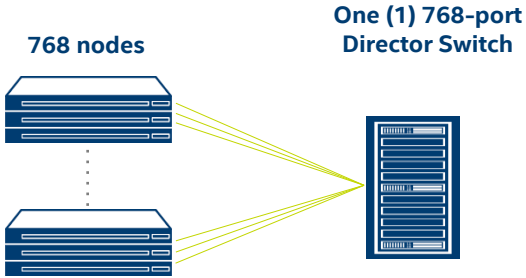
InfiniBand* (36-port Switch Chip)

FIVE-hop Fat Tree



Intel® Omni-Path Architecture (48-port)

THREE-hop Fat Tree



(43) 36-port	**Edge Switches**	Not required	**100%**
1,542	**Cables**	768	**50%**
99u (2+ racks)	**Rack Space**	20u (<½ rack)	**79%**
~680ns (5 hops)	**Switch Latency[1]**	300-330ns[2] (3 hops)	**51-55%**

% Reduction

[1] Latency numbers based on Mellanox CS7500 Director Switch and Mellanox SB7700/SB7790 Edge switches. See www.Mellanox.com for more product information.

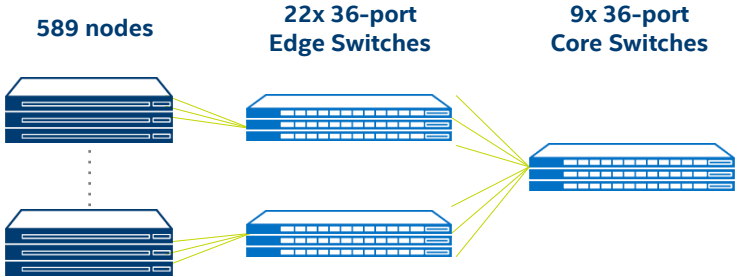
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may 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. *Other names and brands may be claimed as the property of others.

Intel® Omni-Path Fabric's 48 Radix Chip

Real World Example – sub-648 node design still provides advantages

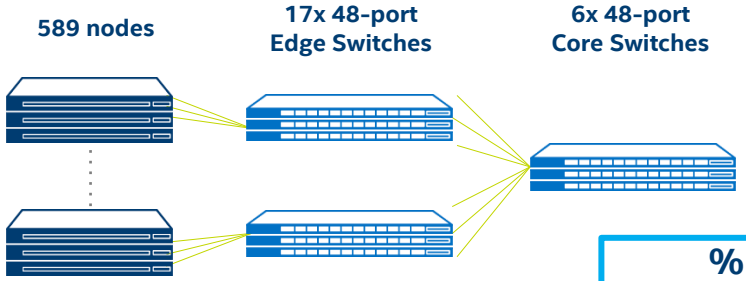
InfiniBand* (36-port Switch Chip)

THREE-hop 3:1 Tree



Intel® Omni-Path Architecture (48-port Switch Chip)

THREE-hop 3:1 Tree

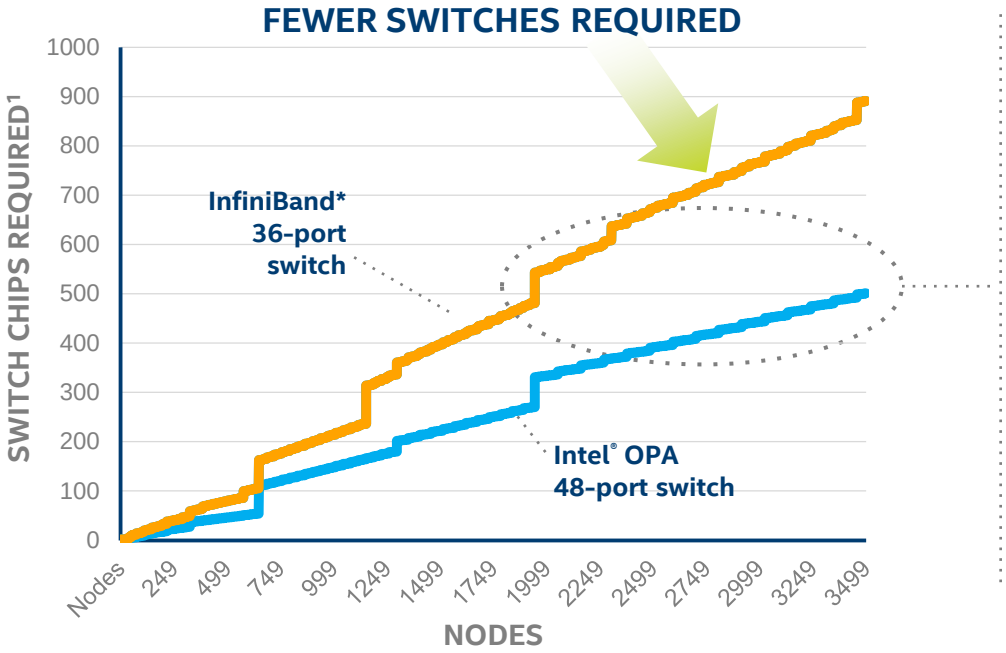


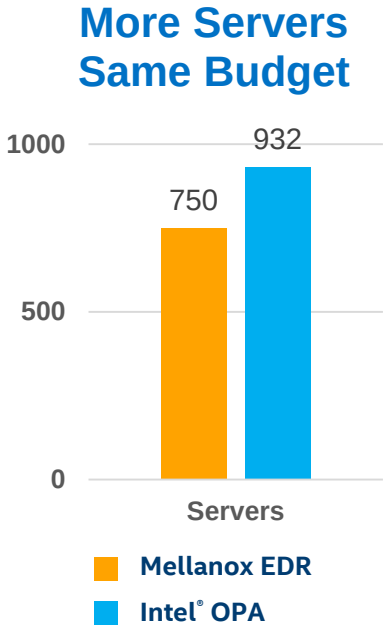
31x 36-port	**Switches**	23x 48-port	**% Change**
787	**Cables**	793	**<1%**
31u	**Rack Space**	23u	**-26%**
27 nodes (756 cores)	**Largest Non-Blocking Set**	36 nodes (1008 cores)	**+33%**

[1] Latency numbers based on Mellanox CS7500 Director Switch and Mellanox SB7700/SB7790 Edge switches. See www.Mellanox.com for more product information.

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

Are You Leaving **Performance** on the Table?





Up to **24%** more Servers[1]

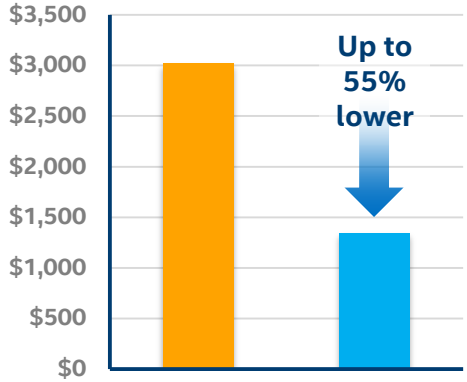
Or
More storage
More software licenses
Higher support SLA
Additional consultancy

[1] Configuration assumes a 750-node cluster, and number of switch chips required is based on a full bisectional bandwidth (FBB) Fat-Tree configuration. Intel® OPA uses one fully-populated 768-port director switch, and Mellanox EDR solution uses a combination of 648-port director switches and 36-port edge switches. Intel and Mellanox component pricing from www.kernelsoftware.com, with prices as of October 20, 2016. Assumes $6,200 for a 2-socket Intel® Xeon® processor based compute node. * Other names and brands may be claimed as property of others.

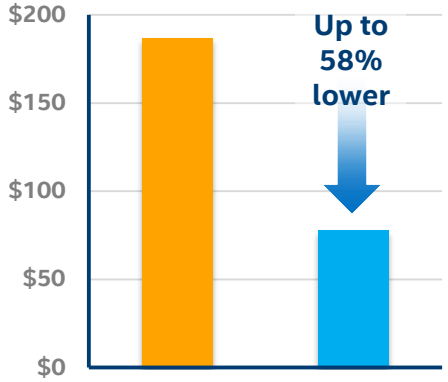
3-Year TCO Advantage

Based on HW acquisition costs (server and fabric) and 3-year power and cooling costs

Fabric Cost Comparison[1]



Fabric Power and Cooling Costs[1]



Intel® OPA can deliver **up to 64% lower Fabric TCO** over 3 years[1]

EDR IB

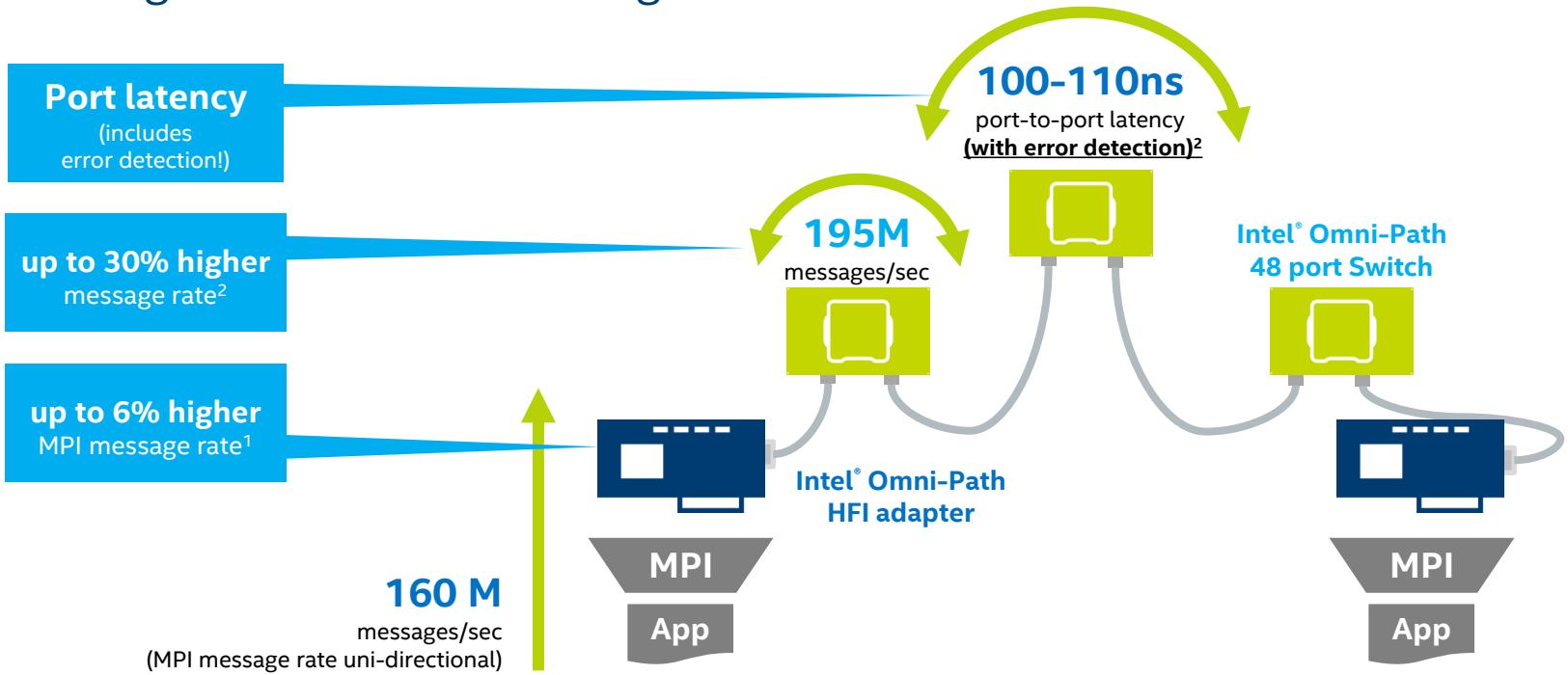
Intel® OPA

[1] Configuration assumes a 750-node cluster, and number of switch chips required is based on a full bisectional bandwidth (FBB) Fat-Tree configuration. Intel® OPA uses one fully-populated 768-port director switch, and Mellanox EDR solution uses a combination of director switches and edge switches. Includes hardware acquisition costs (server and fabric), 24x7 3-year support (Mellanox Gold support), and 3-year power and cooling costs. Mellanox and Intel component pricing from www.kernelsoftware.com, with prices as of October 20, 2016. Mellanox power data based on Mellanox CS7500 Director Switch, Mellanox SB7700/SB7790 Edge switch, and Mellanox ConnectX-4 VPI adapter card product briefs posted on www.mellanox.com as of November 1, 2015. Intel OPA power data based on product briefs posted on www.intel.com as of November 16, 2015. Power and cooling costs based on $0.10 per kWh, and assumes server power costs and server cooling cost are equal and additive. * Other names and brands may be claimed as property of others.

HIGHER PERFORMANCE

Intel® Omni-Path Architecture

Accelerating data movement through the fabric



[1] Based on Intel projections for Wolf River and Prairie River maximum messaging rates, compared to Mellanox CS7500 Director Switch and Mellanox ConnectX-4 adapter and Mellanox SB7700/SB7790 Edge switch product briefs posted on www.mellanox.com as of November 3, 2015.

[2] Latency reductions based on Mellanox CS7500 Director Switch and Mellanox SB7700/SB7790 Edge switch product briefs posted on www.mellanox.com as of July 1, 2015, compared to Intel measured data that was calculated from difference between back to back osu_latency test and osu_latency test through one switch hop. 10ns variation due to "near" and "far" ports on an Intel® OPA edge switch. All tests performed using Intel® Xeon® E5-2697v3 with Turbo Mode enabled.

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

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.

Latency, Bandwidth, and Message Rate

Intel® Xeon® processor E5-2699 v3 & E5-2699 v4 with Intel® OPA

Metric	E5-2699 v3[1]	E5-2699 v4[2]
Latency (one-way, 1 switch, 8B) [ns]	910	910
Bandwidth (1 rank per node, 1 port, uni-dir, 1MB) [GB/s]	12.3	12.3
Bandwidth (1 rank per node, 1 port, bi-dir, 1MB) [GB/s]	24.5	24.5
Message Rate (max ranks per node, uni-dir, 8B) [Mmps]	112.0	141.1
Message Rate (max ranks per node, bi-dir, 8B) [Mmps]	137.8	172.5

Near linear scaling of message rate with added cores on successive Intel® Xeon® processors

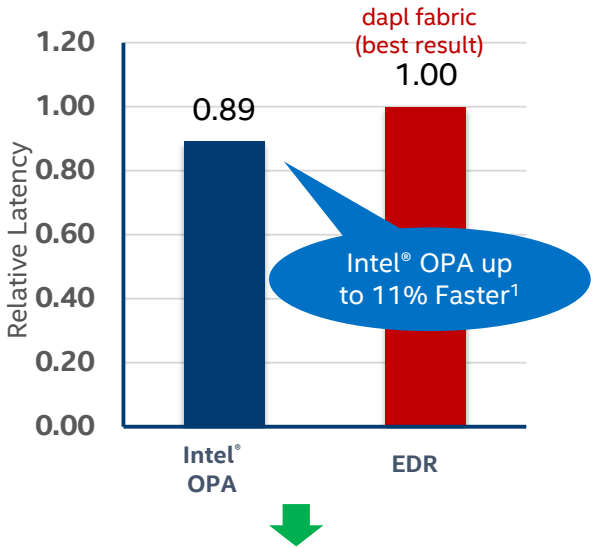
Dual socket servers. Intel® Turbo Boost Technology enabled, Intel® Hyper-Threading Technology disabled. OSU OMB 5.1. Intel® OPA: Open MPI 1.10.0-hfi as packaged with IFS 10.0.0.0.697. Benchmark processes pinned to the cores on the socket that is local to the Intel® OP Host Fabric Interface (HFI) before using the remote socket. RHEL 7.2.Bi-directional message rate measured with osu_mbw_mr, modified for bi-directional measurement. We can provide a description of the code modification if requested. BIOS settings: IOU non-posted prefetch disabled. Snoop timer for posted prefetch=9. Early snoop disabled. Cluster on Die disabled.
1. Intel® Xeon® processor E5-2699 v3 2.30 GHz 18 cores, 36 ranks per node for message rate test
2. Intel® Xeon® processor E5-2699 v4 2.20 GHz 22 cores, 44 ranks per node for message rate test

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.

MPI Performance - Ohio State Microbenchmarks

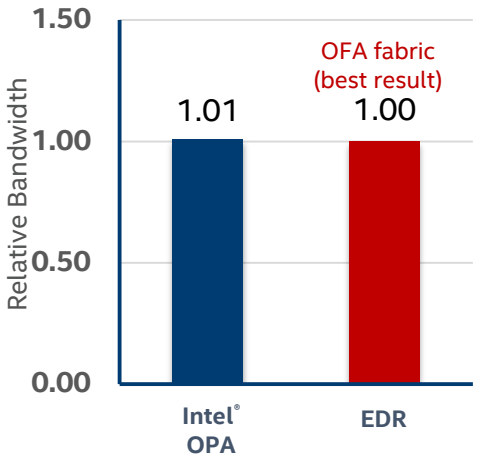
Intel® Omni-Path Architecture (Intel® OPA) vs. InfiniBand* EDR - Intel® MPI

MPI Latency[1]



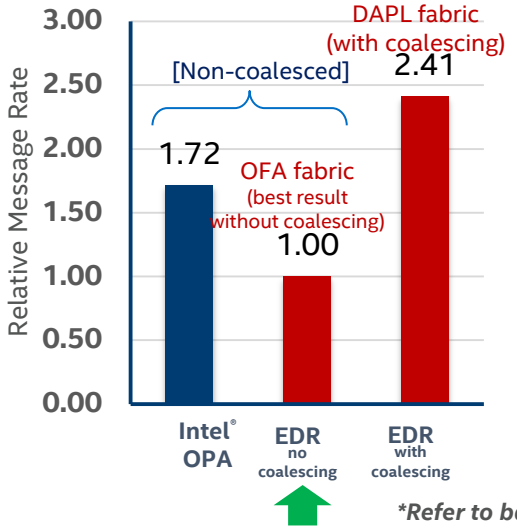
LOWER is Better

MPI Bandwidth[2]



HIGHER is Better

MPI Message Rate[3]



HIGHER is Better

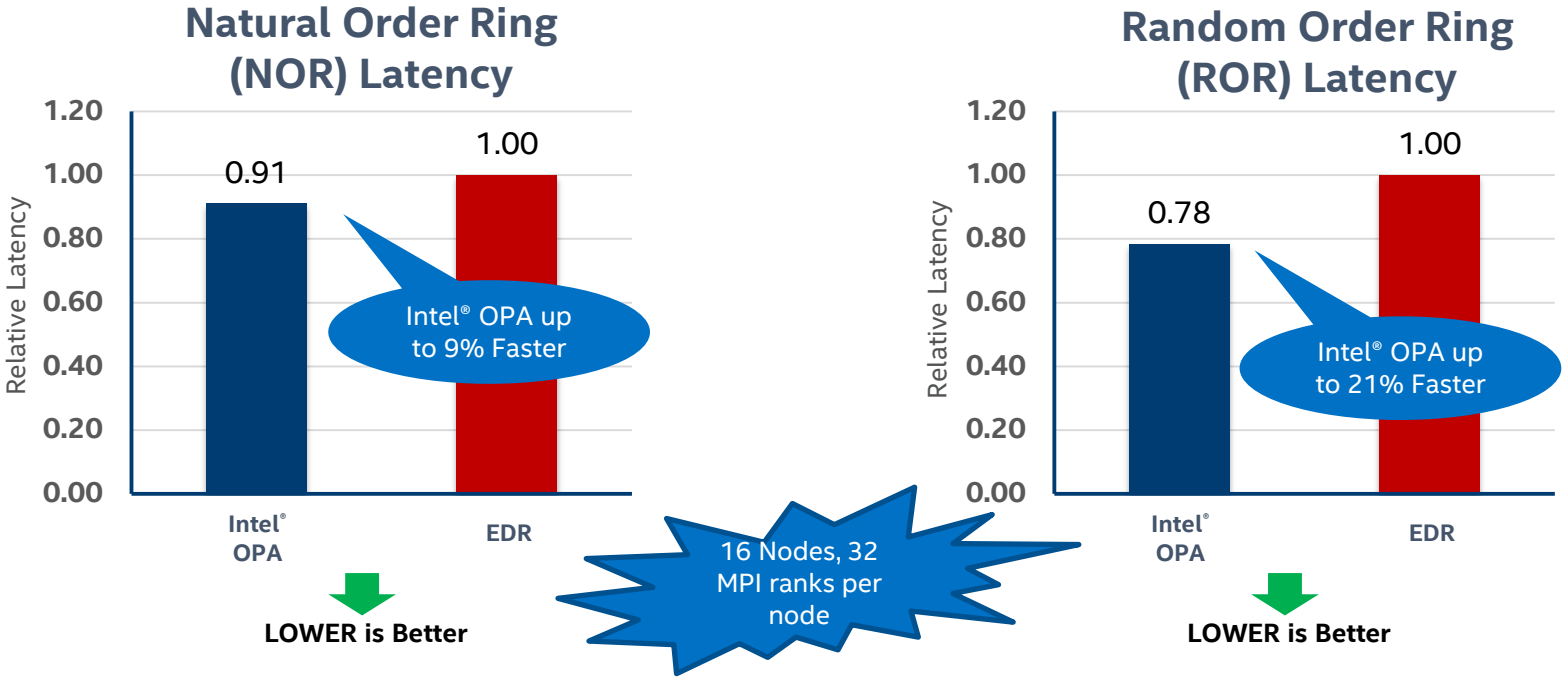
Refer to backup slide for more detail

Tests performed on Intel® Xeon® Processor E5-2697A v4 dual-socket servers with 2133 MHz DDR4 memory. Intel® Turbo Boost Technology and Intel® Hyper-Thread Technology enabled. Ohio State Micro Benchmarks v. 5.0. Intel MPI 5.1.3, RHEL7.2. Intel® OPA: tmi fabric, I_MPI_TMI_DRECV=1. Intel Corporation Device 24f0 – Series 100 HFI ASIC (B0 silicon). OPA Switch: Series 100 Edge Switch – 48 port (B0 silicon). IOU Non-posted Prefetch disabled in BIOS. Snoop hold-off timer = 9. EDR based on internal testing: shm:dapl fabric. **-genv I_MPI_DAPL_EAGER_MESSAGE_AGGREGATION off.** Mellanox EDR ConnectX-4 Single Port Rev 3 MCX455A HCA. Mellanox SB7700 - 36 Port EDR Infiniband switch. MLNX_OFED_LINUX-3.2-2.0.0.0 (OFED-3.2-2.0.0). 1. osu_latency 8 B message. 2. osu_bw 1 MB message. 3. osu_mbw_mr, 8 B message (uni-directional), 32 MPI rank pairs. Maximum rank pair communication time used instead of average time, average time introduced into Ohio State Micro Benchmarks as of v3.9 (2/28/13).

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.

MPI Latency at Scale

Intel® Omni-Path Architecture (Intel® OPA) vs. InfiniBand* EDR - Open MPI

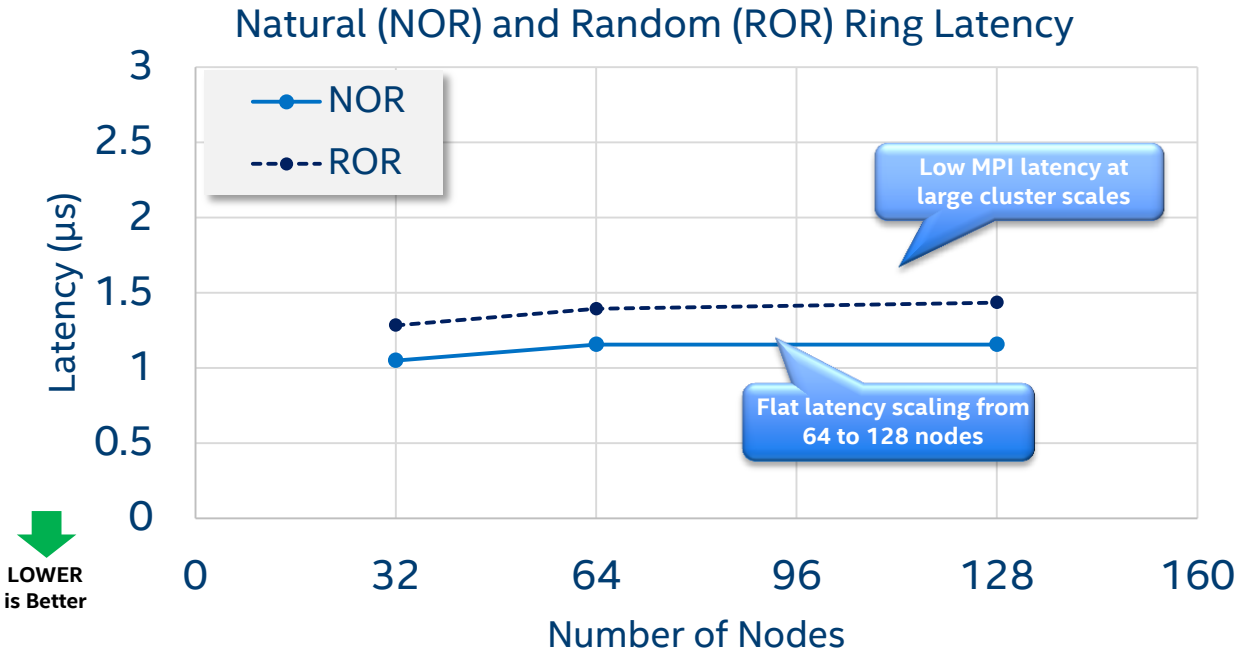


Tests performed on Intel® Xeon® Processor E5-2697A v4 dual-socket servers with 2133 MHz DDR4 memory. Intel® Turbo Boost Technology and Intel® Hyper-Thread Technology enabled. HPCC 1.4.3. RHEL7.2. Intel OPA: Open MPI 1.10.0 with PSM2 as packaged with IFS 10.0.1.0.50. Intel Corporation Device 24f0 – Series 100 HFI ASIC (B0 silicon). OPA Switch: Series 100 Edge Switch – 48 port (B0 silicon). IOU Non-posted Prefetch disabled in BIOS. EDR: Open MPI 1.10-mellanox released with hpcx-v1.5.370-gcc-MLNX_OFED_LINUX-3.2-1.0.1.1-redhat7.2-x86_64. MLNX_OFED_LINUX-3.2-2.0.0.0 (OFED-3.2-2.0.0). Mellanox EDR ConnectX-4 Single Port Rev 3 MCX455A HCA. Mellanox SB7700 - 36 Port EDR Infiniband switch.

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.

MPI Latency at Scale

Intel® Omni-Path Architecture (Intel® OPA)

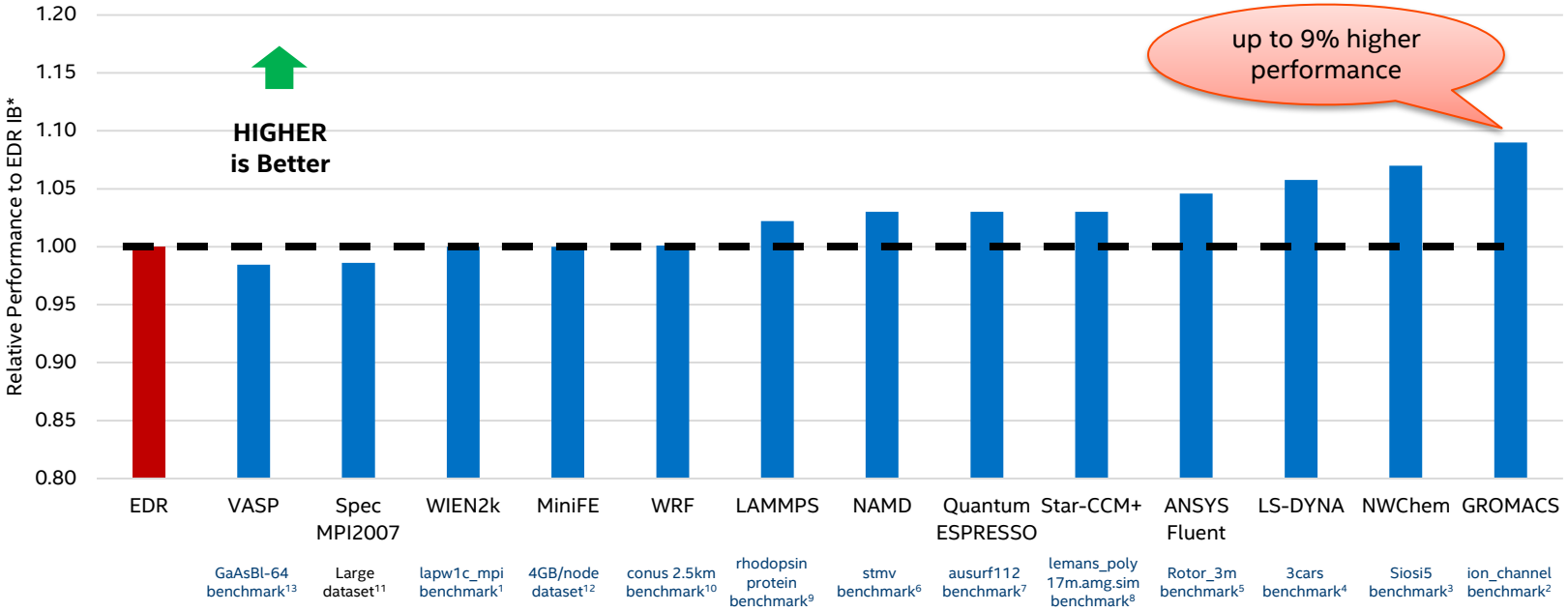


Nodes	MPI Ranks
32	1152
64	2304
128	4608

Intel® Xeon® processor E5-2699 v3, Intel® Turbo Boost and Intel® Hyper-Threading Technology disabled. 36 MPI ranks per node. HPCC 1.4.3. Intel MPI 5.1.1. /opt/intel/composer_xe_2015.1.133/mkl. -O2 -xCORE2-AVX -ip -ansi-alias -fno-alias -DLONG_IS_64BITS -DRA_SANDIA_OPT2 -DUSING_FFTW -DHPCC_FFT_235. Pre-production Intel® Omni-Path hardware and software. IFS 10.0.0.0.625. 2133 MHz DDR4 memory per node. RHEL 7.0.

Intel® Omni-Path Architecture (Intel® OPA)

Application Performance - Intel® MPI - 16 Nodes



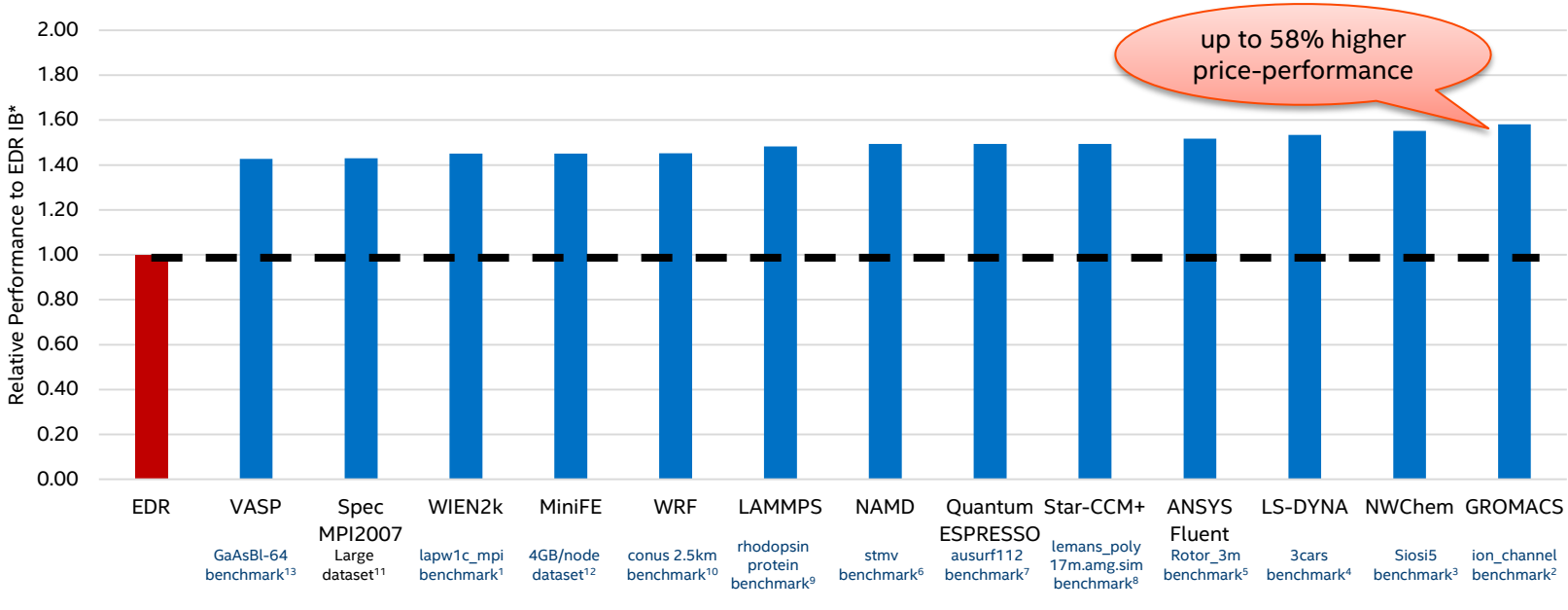
*Spec MPI2007 results estimates until published
**see following slide for system configurations

No Intel® OPA or EDR specific optimizations applied to any workloads except LS-DYNA and ANSYS Fluent: Intel® OPA HFI driver parameter: eager_buffer_size=8388608
WIEN2k comparison is for 8 nodes because EDR IB measurements did not scale above 8 nodes*

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.

Intel® Omni-Path Architecture (Intel® OPA)

Application Performance Per Fabric Dollar* - Intel® MPI - 16 Nodes



*Spec MPI2007 results estimates until published

**see following slide for system configurations

No Intel® OPA or EDR specific optimizations applied to any workloads except LS-DYNA and ANSYS Fluent: Intel® OPA HFI driver parameter: eager_buffer_size=8388608

WIEN2k comparison is for 8 nodes because EDR IB measurements did not scale above 8 nodes*

*All pricing data obtained from www.kernelsoftware.com May 4, 2016. All cluster configurations estimated via internal Intel configuration tool. Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction. **Fabric hardware assumes one edge switch, 16 network adapters and 16 cables.**

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.

Intel® Omni-Path Architecture

Disruptive innovations to knock down the "I/O Wall"

HIGHER PERFORMANCE

Accelerates discovery and innovation

Up to **21% lower latency at scale,** up to **17% higher messaging rate,** and up to **9% higher application performance** than InfiniBand EDR[1]

BETTER ECONOMICS

Reduces size of fabric budgets.
Use savings to purchase more compute

up to **24% more compute nodes**
Better price-performance than InfiniBand* EDR reduces fabric spends for a given cluster size. Use savings to get more compute nodes with same total budget[2]

MORE POWER EFFICIENT

more efficient switches and cards and a reduction in switch count and cables due to the 48-port chip architecture

Up to **60% lower power** than InfiniBand* EDR[3]

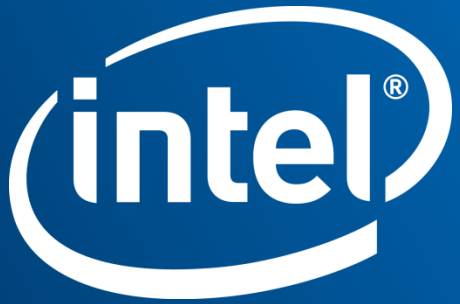
GREATER RESILIENCY

"no compromise" error detection and maintains link continuity with lane failures

No additional latency
penalty for error detection with Packet Integrity Protection[4]

[1] Intel® Xeon® Processor E5-2697A v4 dual-socket servers with 2133 MHz DDR4 memory. Intel® Turbo Boost Technology and Intel® Hyper Threading Technology enabled. BIOS: Early snoop disabled, Cluster on Die disabled, IOU non-posted prefetch disabled, Snoop hold-off timer=9. Red Hat Enterprise Linux Server release 7.2 (Maipo). Intel® OPA testing performed with Intel Corporation Device 24f0 – Series 100 HFI ASIC (B0 silicon). OPA Switch: Series 100 Edge Switch – 48 port (B0 silicon). Intel® OPA host software 10.1 or newer using Open MPI 1.10.x contained within host software package. EDR IB* testing performed with Mellanox EDR ConnectX-4 Single Port Rev 3 MCX455A HCA. Mellanox SB7700 - 36 Port EDR Infiniband switch. EDR tested with MLNX_OFED_Linux-3.2.x. OpenMPI 1.10.x contained within MLNX HPC-X. Message rate claim: Ohio State Micro Benchmarks v. 5.0. osu_mbw_mr, 8 B message (uni-directional), 32 MPI rank pairs. Maximum rank pair communication time used instead of average time, average timing introduced into Ohio State Micro Benchmarks as of v3.9 (2/28/13). Best of default, MXM_TLS=self,rc, and -mca pml yalla tunings. All measurements include one switch hop. Latency claim: HPCC 1.4.3 Random order ring latency using 16 nodes, 32 MPI ranks per node, 512 total MPI ranks. Application claim: GROMACS version 5.0.4 ion_channel benchmark. 16 nodes, 32 MPI ranks per node, 512 total MPI ranks. Intel® MPI Library 2017.0.064. Additional configuration details available upon request. [2] Configuration assumes a 750-node cluster, and number of switch chips required is based on a full bisectional bandwidth (FBB) Fat-Tree configuration. Intel® OPA uses one fully-populated 768-port director switch, and Mellanox EDR solution uses a combination of 648-port director switches and 36-port edge switches. Intel and Mellanox component pricing from www.kernelsoftware.com, with prices as of October 20, 2016. Assumes $6,200 for a 2-socket Intel® Xeon® processor based compute node. [3] Configuration assumes a 750-node cluster, and number of switch chips required is based on a full bisectional bandwidth (FBB) Fat-Tree configuration. Intel® OPA uses one fully-populated 768-port director switch, and Mellanox EDR solution uses a combination of director switches and edge switches. Mellanox power data based on Mellanox CS7500 Director Switch, Mellanox SB7700/SB7790 Edge switch, and Mellanox ConnectX-4 VPI adapter card installation documentation posted on www.mellanox.com as of November 1, 2015. Intel OPA power data based on product briefs posted on www.intel.com as of November 16, 2015. [4] A CRC check is performed on every Link Transfer Packet (LTP, 1056-bits) transmitted through a switch hop as defined by the Intel® OPA wire protocol, so stated switch latencies always include error detection by definition.

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.


intel®

PERFORMANCE BACKUP

System & Software Configuration

Application Performance and Application Performance per Fabric Dollar slides

Common configuration for bullets 1-11 unless otherwise specified: Intel® Xeon® Processor E5-2697A v4 dual socket servers. 64 GB DDR4 memory per node, 2133 MHz. RHEL 7.2. BIOS settings: Snoop hold-off timer = 9, Early snoop disabled, Cluster on die disabled. IOU Non-posted prefetch disabled. Intel® Omni-Path Architecture (Intel® OPA):Intel Fabric Suite 10.0.1.0.50. Intel Corporation Device 24f0 – Series 100 HFI ASIC (Production silicon). OPA Switch: Series 100 Edge Switch – 48 port (Production silicon). EDR Infiniband: MLNX_OFED_LINUX-3.2-2.0.0.0 (OFED-3.2-2.0.0). Mellanox EDR ConnectX-4 Single Port Rev 3 MCX455A HCA. Mellanox SB7700 - 36 Port EDR Infiniband switch.

- WIEN2k version 14.2. http://www.wien2k.at/. http://www.wien2k.at/reg_user/benchmark/. Run command: "mpirun ... lapw1c_mpi lapw1.def". Intel Fortran Compiler 17.0.0 20160517. Compile flags: -FR -mp1 -w -prec_div -pc80 -pad -ip -DINTEL_VML -traceback -assume buffered_io -DFTW3 -I/opt/intel/compilers_and_libraries_2017.0.064/linux/mkl/include/fftw/ -DParallel. shm:tmi fabric used for Intel® OPA and shm:dapl fabric used for EDR IB*.
- GROMACS version 5.0.4. Intel Composer XE 2015.1.133. Intel MPI 5.1.3. FFTW-3.3.4. ~/bin/cmake .. -DGMX_BUILD_OWN_FFTW=OFF -DREGRESSIONTEST_DOWNLOAD=OFF -DCMAKE_C_COMPILER=icc -DCMAKE_CXX_COMPILER=icpc -DCMAKE_INSTALL_PREFIX=~/gromacs-5.0.4-installed. Intel® OPA MPI parameters: I_MPI_FABRICS=shm:tmi, EDR MPI parameters: I_MPI_FABRICS=shm:dapl
- NWChem release 6.6. Binary: nwchem_comex-mpi-pr_mkl with MPI-PR run over MPI-1. Workload: siosi3 and siosi5. Intel® MPI Library 2017.0.064. 2 ranks per node, 1 rank for computation and 1 rank for communication. shm:tmi fabric for Intel® OPA and shm:dapl fabric for EDR, all default settings. Intel Fabric Suite 10.2.0.0.153. http://www.nwchem-sw.org/index.php/Main_Page
- LS-DYNA MPP R8.1.0 dynamic link. Intel Fortran Compiler 13.1 AVX2. Intel® OPA - Intel MPI 2017 Library Beta Release Candidate 1. mpi.2017.0.0.BETA.U1.RC1.x86_64.ww20.20160512.143008. MPI parameters: I_MPI_FABRICS=shm:tmi. HFI driver parameter: eager_buffer_size=8388608. EDR MPI parameters: I_MPI_FABRICS=shm:ofa.
- ANSYS Fluent v17.0, Rotor_3m benchmark. Intel® MPI Library 5.0.3 as included with Fluent 17.0 distribution, and libpsm_infinipath.so.1 added to the Fluent syslib library path for PSM/PSM2 compatibility. Intel® OPA MPI parameters: -pib.infinipath, EDR MPI parameters: -pib.dapl
- NAMD: Intel Composer XE 2015.1.133. NAMD V2.11, Charm 6.7.0, FFTW 3.3.4. Intel MPI 5.1.3. Intel® OPA MPI parameters: I_MPI_FABRICS=shm:tmi, EDR MPI parameters: I_MPI_FABRICS=shm:dapl
- Quantum Espresso version 5.3.0. Intel Compiler 2016 Update 2. ELPA 2015.11.001 (http://elpa.mpcdf.mpg.de/elpa-tar-archive). Minor patch set for QE to accommodate latest ELPA. Most optimal NPOOL, NDIAG, and NTG settings reported for both OPA and EDR. Intel® OPA MPI parameters: I_MPI_FABRICS=shm:tmi, EDR MPI parameters: I_MPI_FABRICS=shm:dapl
- CD-adapco STAR-CCM+® version 11.04.010. Workload: lemans_poly_17m.amg.sim benchmark. Intel MPI version 5.0.3.048. 32 ranks per node. OPA command: $ /starccm+ -ldlibpath /STAR-CCM+11.04.010/mpi/intel/5.0.3.048/linux-x86_64/lib64 -ldpreload /usr/lib64/psm2-compat/libpsm_infinipath.so.1 -mpi intel -mppflags "-env I_MPI_DEBUG 5 -env I_MPI_FABRICS shm:tmi -env I_MPI_TMI_PROVIDER psm" -power -rsh ssh -np 512 -machinefile hosts -benchmark:"-nps 512,256,128,64,32 -nits 20 -preits 40 -tag lemans_opa_n16" lemans_poly_17m.amg.sim. EDR command: $ /starccm+ -mpi intel -mppflags "-env I_MPI_DEBUG 5" -power -rsh ssh -np 512 -machinefile hosts -benchmark:"-nps 512,256,128,64,32 -nits 20 -preits 40 -tag lemans_edr_n16" lemans_poly_17m.amg.sim
- LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator) Feb 16, 2016 stable version release. MPI: Intel® MPI Library 5.1 Update 3 for Linux. Workload: Rhodopsin protein benchmark. Number of time steps=100, warm up time steps=10 (not timed) Number of copies of the simulation box in each dimension: 8x8x4 and problem size: 8x8x4x32k = 8,192k atoms Intel® OPA: MPI parameters: I_MPI_FABRICS=shm:tmi, I_MPI_PIN_DOMAIN=core EDR: MPI parameters: I_MPI_FABRICS=shm:dapl,, I_MPI_PIN_DOMAIN=core
- WRF version 3.5.1, Intel Composer XE 2015.1.133. Intel MPI 5.1.3. NetCDF version 4.4.2. FCBASEOPTS=-w -ftz -align all -fno-alias -fp-model precise. CFLAGS_LOCAL = -w -O3 -ip. Intel® OPA MPI parameters: I_MPI_FABRICS=shm:tmi, EDR MPI parameters: I_MPI_FABRICS=shm:dapl
- Spec MPI 2007: 16 nodes, 32 MPI ranks/node. SPEC MPI2007, Large suite, https://www.spec.org/mpi/. *Intel Internal measurements marked estimates until published. Intel MPI 5.1.3. Intel® OPA MPI parameters: I_MPI_FABRICS=shm:tmi, EDR MPI parameters: I_MPI_FABRICS=shm:dapl

Common configuration for bullets 12-13: Intel® Xeon® Processor E5-2697 v4 dual socket servers. 128 GB DDR4 memory per node, 2400 MHz. RHEL 6.5. BIOS settings: Snoop hold-off timer = 9. Intel® OPA: Intel Fabric Suite 10.0.1.0.50. Intel Corporation Device 24f0 – Series 100 HFI ASIC (Production silicon). OPA Switch: Series 100 Edge Switch – 48 port (Production silicon). IOU Non-posted prefetch disabled. 2). Mellanox EDR based on internal measurements: Mellanox EDR ConnectX-4 Single Port Rev 3 MCX455A HCA. Mellanox SB7700 - 36 Port EDR Infiniband switch. IOU Non-posted prefetch enabled.

- MiniFE 2.0, Intel compiler 16.0.2. Intel® MPI Library version 5.1.3. Build settings: -O3 –xCORE-AVX2 -DMINIFE_CSR_MATRIX -DMINIFE_GLOBAL_ORDINAL="long long int", mpirun -bootstrap ssh -env OMP_NUM_THREADS 1 - perhost 36 miniFE.x nx=200 ny=200 nz=200, 200x200x200 grid using 36 MPI ranks pinned to 36 cores per node. Intel® OPA MPI parameters: I_MPI_FABRICS=shm:tmi, EDR MPI parameters: I_MPI_FABRICS=shm:dapl . Intel® Turbo Mode technology and Intel® Hyper threading technology disabled.
- VASP (developer branch). MKL: 11.3 Update 3 Product build 20160413. Compiler: 2016u3. Intel MPI-2017 Build 20160718 . elpa-2016.05.002. Intel® OPA MPI parameters: I_MPI_FABRICS=shm:tmi, EDR MPI parameters: I_MPI_FABRICS=shm:dapl, I_MPI_PLATFORM=BDW, I_MPI_DAPL_PROVIDER=ofa-v2-mlx5_0-1u, I_MPI_DAPL_DIRECT_COPY_THRESHOLD=331072. Intel® Turbo Mode technology disabled. Intel Hyper Threading technology enabled.

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.