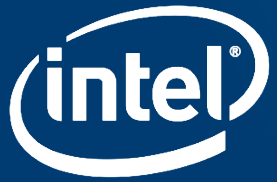


Intel® Omni-Path Architecture

Product Update

Connectivity Group

Data Center Group, Intel Corporation

September 2016

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.

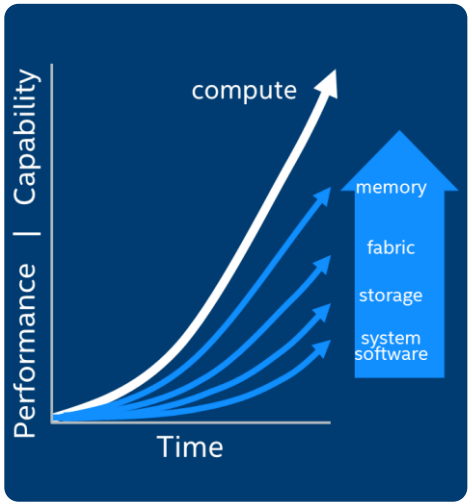
No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

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.

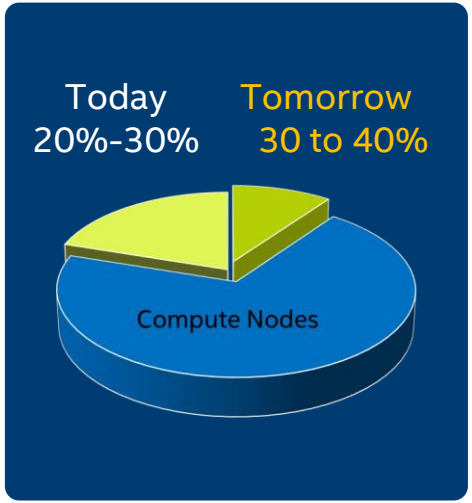
The Interconnect Landscape: **Why Intel® OPA?**

Performance



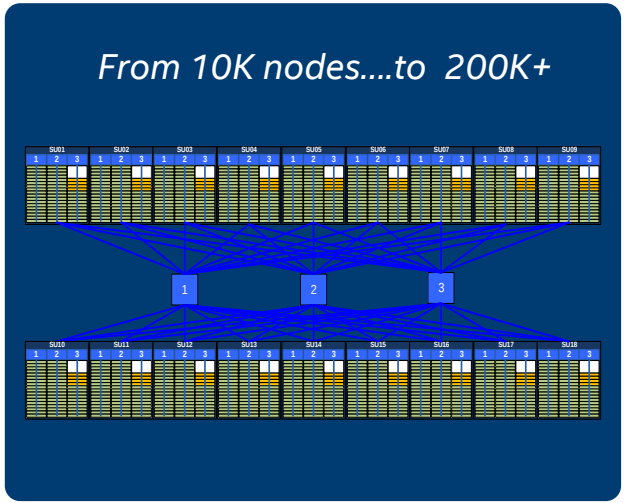
I/O struggling to keep up with CPU innovation

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.

Intel® Omni-Path Architecture

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

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
- Products from most HPC server and storage OEMs
- Strong ecosystem with 75+ Fabric Builders members

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

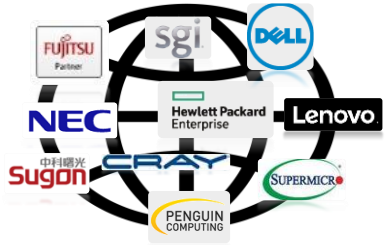
Passive Copper

Active Optical

Intel® Omni-Path Architecture is Quickly Gaining Industry Momentum

PENT-UP END USER DEMAND

METEORIC RAMP. WORLDWIDE COVERAGE



Major system deployments

US DoE CTS-1, Pittsburgh Supercomputing Center, Cineca, Alfred Wegener Institute (AWI), TACC, Rutgers University

8 clusters in the Top500!

2x InfiniBand EDR entries in Nov 2015 list*

Over 14K nodes shipped in Q1'16

2x InfiniBand EDR volume at the same point[1]*

Sold by *every* major HPC OEM. Delivered in *every* geography.

[1] Mellanox node count based on reported EDR sales revenue reported in the Q2 2015 Mellanox 10Q. Intel estimates of $900 per node (6.6k nodes). [2] Configuration for performance testing: **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. Intel® Omni-Path Architecture (Intel® OPA)** Intel Fabric Suite 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. **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. IOU Non-posted prefetch enabled. **Applications: NAMD:,** NAMD V2.11, **GROMACS** version 5.0.4. **LS-DYNA** MPP R7.1.2 **LAMMPS** (Large-scale Atomic/Molecular Massively Parallel Simulator) Feb 16, 2016 stable version release. **Quantum Espresso** version 5.3.0, **WRF** version 3.5.1 **Spec MPI 2007:**. SPEC MPI2007, Large suite, https://www.spec.org/mpi/. *Intel Internal measurements marked estimates until published. [3] 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. For 16 Node configuration, Fabric hardware assumes one edge switch, 16 network adapters and 16 cables.

Intel® Omni-Path Architecture Product Family

Wolf River

HFI ASIC

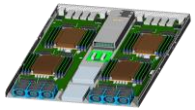
Storm Lake Gen1 **Host Fabric Interface (HFI) Silicon**
2 x 100 Gbps, 50 GB/sec Fabric Bandwidth

Prairie River

Switch ASIC

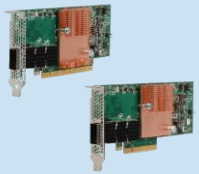
Storm Lake Gen1 **Switch Silicon**
48 ports, 1200 GB/sec Fabric Bandwidth

Intel® Product Line



Custom Mezz & PCIe Cards

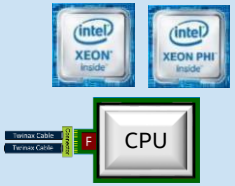
- OEM products based on Wolf River ASIC TBD



Standard PCIe Board[1]

(Chippewa Forest)

- Low Profile PCIe* v3.0 x16
- Low Profile PCIe* v3.0 x8
- Single Port QSFP28



Integrated Xeon® and Xeon Phi™

- Knights Landing: 2 x 100 Gbps ports
- Skylake (Xeon®): 1 x 100Gbps port

Edge Switch[1]

(Eldorado Forest)

- 24 / 48 port individual QSFP28 ports
  - Short reach – QSFP28 copper cables
  - Long reach – QSFP28 AOC, Transceivers
- Air cooling, N+1 redundant fans
- Optional N+N redundant power supply
- In-band management supported
- Optional management card

192-port (7U)

768-port (20U)

Director Class Switches (DCS)[1]

(Sawtooth Forest)

- Full Bisection Bandwidth
- QSFP-based leaf module
- N+1 (DC) and N+N (AC/DC) Redundant Power
- Air cooling, N+1 red. fans and chassis mgmt, hot plug

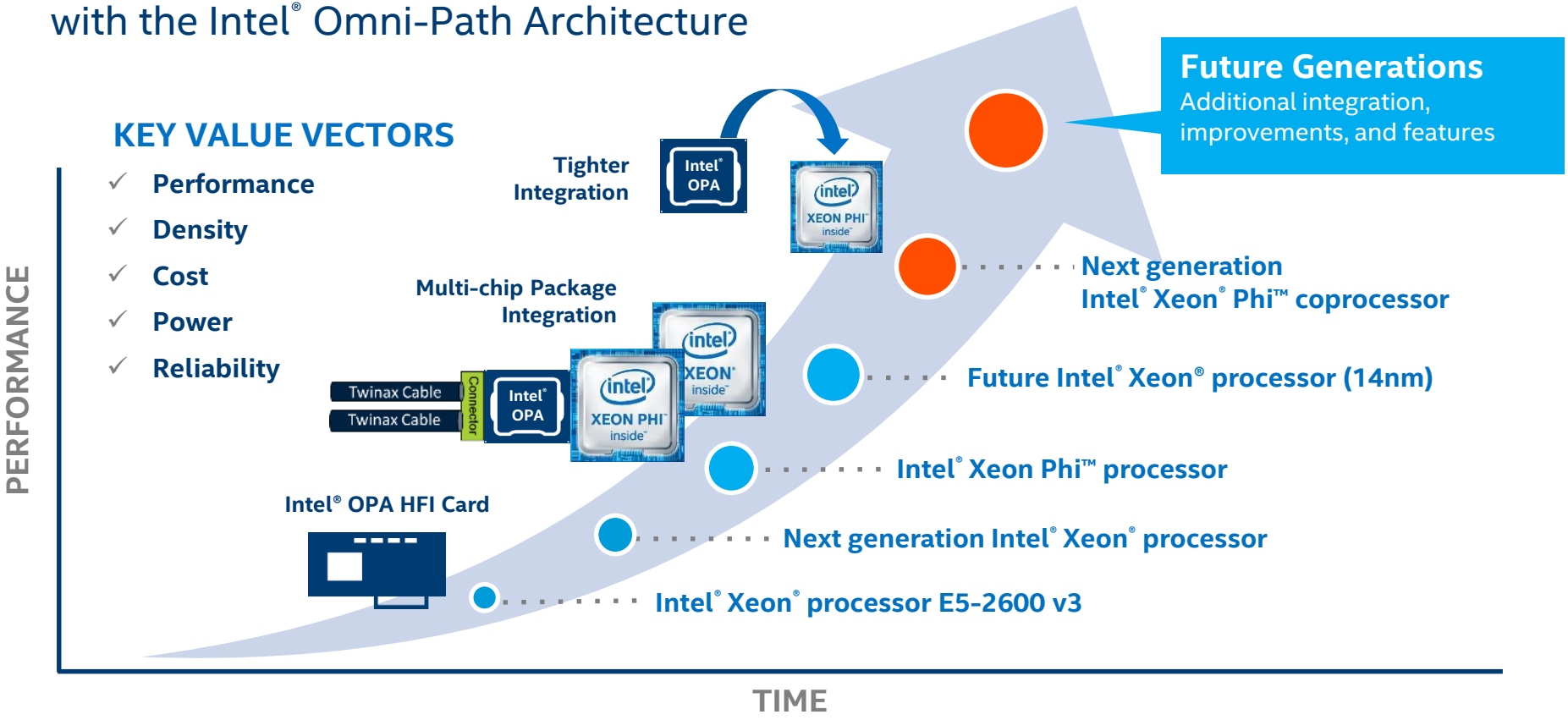
Custom Switches

- OEM products based on Prairie River ASIC TBD

[1] Available as a reference design and Intel product. Director class switch features and introduction in planning

CPU-Fabric Integration

with the Intel® Omni-Path Architecture



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

PERFORMANCE

Latency, Bandwidth, and Message Rate

Intel® Xeon® processor E5-2699 v3 & E5-2699 v4
Intel® Omni-Path Architecture (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

Intel® Omni-Path Scaling on ANSYS Fluent* 17
using Intel® Xeon® Processor E5-2600 v4 Product Family

ANSYS
www.ansys.com

Technical Computing

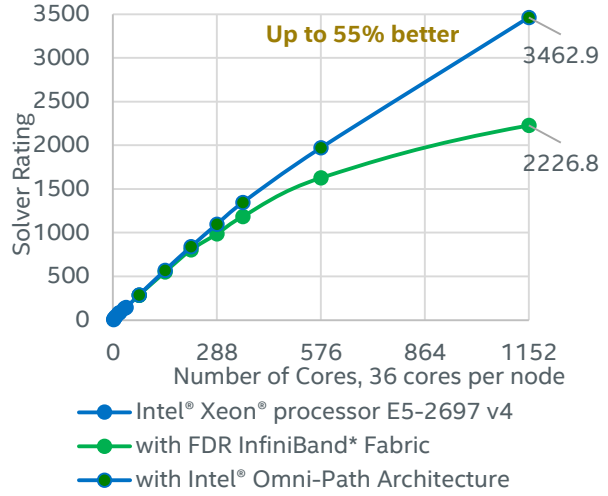
Fluent* 17 Computational Fluid Dynamics

*"Thanks to Intel® **OPA** and the latest Intel® **Xeon® E5-2600 v4** product family, ANSYS Fluent* is able to achieve performance levels **beyond our expectations**. Its unrivaled performance enables our customers to simulate higher-fidelity models without having to expand their cluster nodes ."*[1]

***Dr. Wim Slagter* – Director of HPC and cloud marketing, ANSYS**

- Intel® Omni-Path Architecture (Intel® OPA) is a powerful low latency communications interface specifically designed for High Performance Computing.
- Cluster users will get better utilization of cluster nodes through better scaling.
- Cluster performance means better time-to-solution on CFD simulations.
- Coupled with Intel® MPI, and utilizing standard Fluent runtime options to access TMI, Fluent is ready and proven for out-of-the-box performance on Intel OPA-ready clusters.

ANSYS Fluent* 17 solver rating increased by up to 1.55X with Intel® Omni-Path Architecture scaling on a 32-node cluster



Workload: 12 million cell combustor model, part of the Fluent benchmarking suite. Fluent 17.0

Up to 55% performance advantage with Intel® OPA compared to FDR fabric on a 32 node cluster

1 - Testing conducted on ISV* software on 2S Intel® Xeon® Processor E5-2697 v4 comparing Intel® OPA to FDR InfiniBand* fabric. Testing done by Intel.

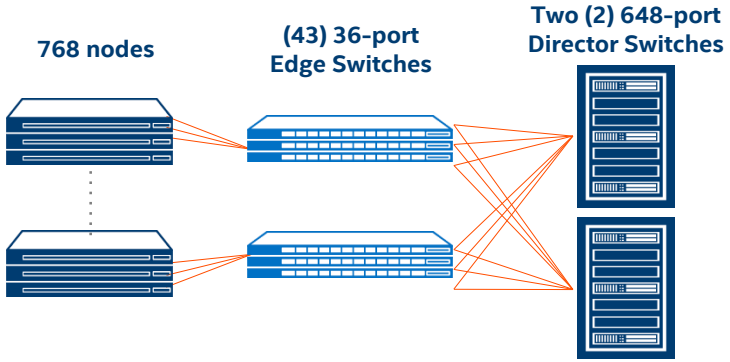
COST BENEFITS

Intel® Omni-Path Fabric's **48 Radix Chip**

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

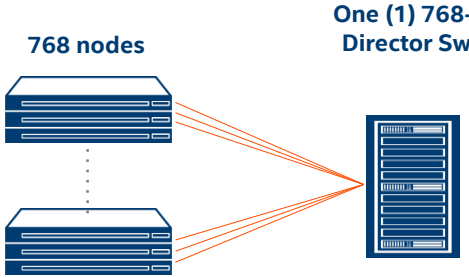
InfiniBand* EDR (36-port Switch Chip)

FIVE-hop Fat Tree



Intel® Omni-Path Architecture (48-port)

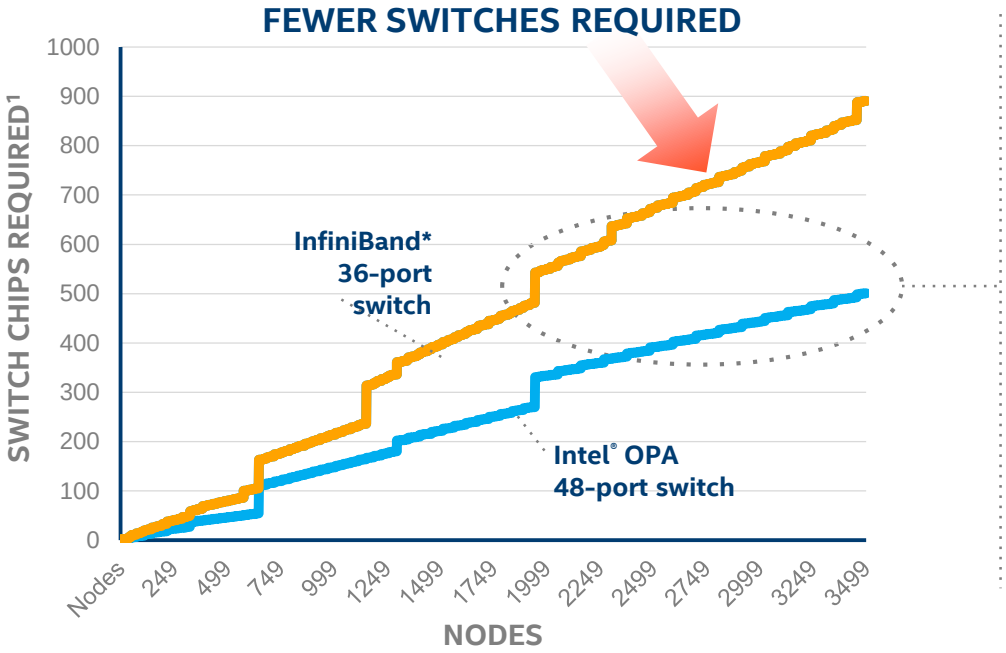
THREE-hop Fat Tree

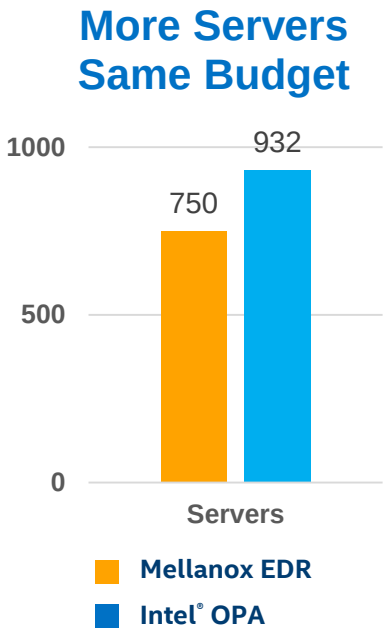


			% Reduction
(43) 36-port	**Edge Switches**	Not required	**100%**
1,542	**Cables**	768	**50%**
99u (2+ racks)	**Rack Space**	20u (<½ rack)	**79%**
~680ns (5 hops)	**Switch Latency[1]**	300-330ns[2] (3 hops)	**51-55%**

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

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





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

[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 May 5, 2016. Compute node pricing based on Dell PowerEdge R730 server from www.dell.com, with prices as of November 3, 2015. Intel® OPA pricing based on estimated reseller pricing based on projected Intel MSRP pricing at time of launch. * Other names and brands may be claimed as property of others.

STORAGE, SOFTWARE, AND SUPPORT

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: Mellanox’s OFED (aka “MOFED”) is a complete overwrite that may impact compatibility with other interconnects. We only layer the necessary changes on top of what’s inbox.

Proven Technology Required for Today's Bids:

Intel® OPA is **the Future** of High Performance Fabrics

QLOGIC intel Aries

Highly Leverages
existing Aries and Intel®
True Scale technologies

Open Source software and supports standards like the **OpenFabrics Alliance***

Innovative Features
for high fabric performance, resiliency, and QoS

Leading Edge Integration
with Intel® Xeon® processor and Intel® Xeon Phi™ processor

Robust Ecosystem
of trusted computing partners and providers

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

BACKUP:
PERFORMANCE TEST CONDITIONS

System & Software Configuration for Application Performance and Price Performance Slides

System configuration: 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. Intel® Omni-Path Architecture (Intel® OPA): Intel Fabric Suite 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. 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. IOU Non-posted prefetch enabled.

Workloads:

- 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
- 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
- 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.
- 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
- 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
- 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
- Spec MPI 2007: To be completed. 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

Configuration Details for Ansys Fluent* 17

ANSYS Fluent 17.0: Combustor_12m workload; Intel OPA vs. FDR. Testing by Intel, 3/10/2016.

BASELINE: Intel® Xeon® processor E5-2697 v4, 2.3 GHz, 36 cores, Grantley-EP (Wellsburg),128GB DDR4/2400 DIMM, Mellanox FDR HCA, Lustre cluster file system used, 36 cores per cluster node used (fully subscribed), Intel® MPI 5.0.3 as distributed with ANSYS Fluent, home snoop, Intel® Hyper-Thrading Technology and Intel® Turbo Boost on, Red Hat Enterprise Linux* 6.4kernel 2.6.32-358, Request Number: 1907

NEW: Intel® Xeon® processor E5-2697 v4, 2.3 GHz, 36 cores, Grantley-EP (Wellsburg),128GB DDR4/2400 DIMM, Intel® Omni-Path Architecture (Intel® OPA) interconnect, Lustre cluster file system used, 36 cores per cluster node used (fully subscribed), Intel® MPI 5.0.3 as distributed with ANSYS Fluent, home snoop, HT and turbo on, Red Hat Enterprise Linux* 6.4kernel 2.6.32-358, Request Number: 1907

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.

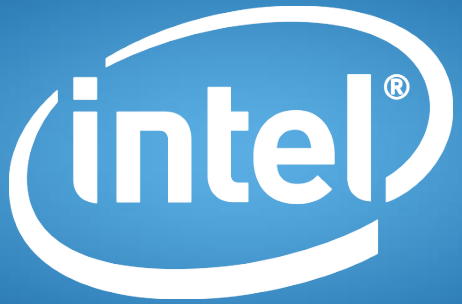
Intel, the Intel logo and others are trademarks of Intel Corporation in the U.S. and/or other countries. *Other names and brands may be claimed as the property of others.

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


intel®