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

Today 20%-30%

Tomorrow 30 to 40%

Increasing Scale

From 10K nodes... to 200K+

Fabric an increasing % of HPC hardware costs

Existing solutions reaching limits

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

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

<table>
<thead>
<tr>
<th>HFI Adapters</th>
<th>Edge Switches</th>
<th>Director Switches</th>
<th>Silicon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single port x8 and x16</td>
<td>1U Form Factor 24 and 48 port</td>
<td>QSF-based 192 and 768 port</td>
<td>OEM custom designs</td>
</tr>
<tr>
<td>x16 Adapter (100 Gb/s)</td>
<td>48-port Edge Switch</td>
<td>768-port Director Switch (20U chassis)</td>
<td>HFI and Switch ASICS</td>
</tr>
<tr>
<td>x8 Adapter (58 Gb/s)</td>
<td>24-port Edge Switch</td>
<td>192-port Director Switch (7U chassis)</td>
<td>Up to 2 ports (50 Gb/s total b/w)</td>
</tr>
</tbody>
</table>

- **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 designs¹
  - 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

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Copper Typical</th>
<th>Copper Maximum</th>
<th>Optical (3W QSFP) Typical</th>
<th>Optical (3W QSFP) Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>X16 HFI</td>
<td>7.4W</td>
<td>11.7W</td>
<td>10.6W</td>
<td>14.9W</td>
</tr>
<tr>
<td>X8 HFI</td>
<td>6.3W</td>
<td>8.3W</td>
<td>9.5W</td>
<td>11.5W</td>
</tr>
</tbody>
</table>

1Specifications contained in public Product Briefs.
Intel® Omni-Path Edge Switch

100 Series 24/48 Port

**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

<table>
<thead>
<tr>
<th>Model</th>
<th>Copper Typical</th>
<th>Copper Maximum</th>
<th>Optical (3W QSFP) Typical</th>
<th>Optical (3W QSFP) Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>24-Ports</td>
<td>146W</td>
<td>179W</td>
<td>231W</td>
<td>264W</td>
</tr>
<tr>
<td>48-Ports</td>
<td>186W</td>
<td>238W</td>
<td>356W</td>
<td>408W</td>
</tr>
</tbody>
</table>

1 Specifications contained in public Product Briefs.
## Intel® Omni-Path Director Class Systems

### 100 Series 6-Slot/24-Slot Systems

#### Highly Integrated
- 7U/20U plus 1U Shelf

#### Switching Capacity
- 38.4/153.6 Tbps 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

### Power

<table>
<thead>
<tr>
<th>Model</th>
<th>Copper</th>
<th>Optical (3W QSFP)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Typical</td>
<td>Maximum</td>
</tr>
<tr>
<td>6-Slot</td>
<td>1.6kW</td>
<td>2.3kW</td>
</tr>
<tr>
<td>24-Slot</td>
<td>6.8kW</td>
<td>8.9kW</td>
</tr>
</tbody>
</table>

*Specifications contained in public Product Briefs.*
CPU-Fabric Integration
with the Intel® Omni-Path Architecture

KEY VALUE VECTORS
- Performance
- Density
- Cost
- Power
- Reliability

Tighter Integration
Multi-chip Package Integration
Intel® OPA HFI Card
Twinax Cable

Future Generations
Additional integration, improvements, and features

Future Intel® Xeon® processor (14nm)
Intel® Xeon Phi™ processor 7200 Series
Intel® Xeon Phi™ coprocessor

NEXT GENERATION
Intel® Xeon® processor E5-2600 v4
Intel® Xeon® processor E5-2600 v3

PERFORMANCE
TIME
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
Generational Software Compatibility

**Intel® True Scale**

- Compiled Applications
  - I/O Focused Upper Layer Protocols (ULPs)
  - Verbs Provider/Driver
    - Intel® True Scale 
    - PSM Layer
    - (On-Load) Host Channel Adapter (HCA)
    - Wire Transport
    - Enhanced Switching Fabric

- Fast Data Path/ Low CPU Load/ High Performance

**Common base architecture** makes the transition smooth
Existing MPI programs and MPI libraries for True Scale that use PSM will work as-is with Omni-Path without recompiling providing **binary compatibility**

**~10% of Verbs based Code**

Programs can be recompiled for Intel® OPA to expose an **additional** set of features
PSM2 API is a **superset** of the PSM API used with True Scale

**High MPI message rate**
**Designed to scale with today’s servers**

**Intel® Omni-Path**

- Compiled Applications
  - I/O Focused Upper Layer Protocols (ULPs)
  - Verbs Provider/Driver
    - Intel® OPA
    - PSM2 Layer
    - (On-Load) Host Channel Adapter (HFI)

- Host Fabric Interface (HFI)

- Wire Transport

- Enhanced Switching Fabric

**Low Transition Costs/ Low Complexity/ Low Upgrade Risks**
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
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

**Intel® OPA Storage Solutions**

Direct Attach to File System Server

Routing through Ethernet/InfiniBand* Storage Fabrics

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

* 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**
- Uses standard LNet router component in Lustre
- IEEL version 2.4 or newer
- RHEL and SLES solutions available

**Configurations**
- Solutions for IB fabrics (QDR, FDR)
- High availability solutions
- Load balanced for performance

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

Software Stack
- Uses standard IP router available in Linux Distros
- RHEL and SLES solutions available

Configurations
- Solutions for IB and Ethernet fabrics
- High availability solutions
- Load balanced for performance

Support
- 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

<table>
<thead>
<tr>
<th>OEM Momentum</th>
<th>Over 100 OEM and HPC storage vendor offerings expected for platforms, switches, and adapters¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ecosystem Momentum</td>
<td>On track for robust hardware and software ecosystem at launch, with 80+ members in the Intel® Fabric Builders program³</td>
</tr>
</tbody>
</table>

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%

Top500 and Major Deployments

✓ Momentum
✓ Performance
✓ Scalability
✓ Stability
SCALABLE AND FLEXIBLE
## New Intel® OPA Fabric Features: Fine-grained Control Improves Resiliency and Optimizes Traffic Movement

<table>
<thead>
<tr>
<th>Traffic Flow Optimization</th>
<th>Description</th>
<th>Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Optimizes Quality of Service (QoS) in mixed traffic environments, such as storage and MPI</td>
<td>Ensures high priority traffic is not delayed → Faster time to solution</td>
</tr>
<tr>
<td></td>
<td>Transmission of lower-priority packets can be paused so higher priority packets can be transmitted</td>
<td>Deterministic latency → Lowers run-to-run timing inconsistencies</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Packet Integrity Protection</th>
<th>Description</th>
<th>Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Allows for rapid and transparent recovery of transmission errors on an Intel® OPA link without additional latency</td>
<td>Fixes happen at the link level rather than end-to-end level</td>
</tr>
<tr>
<td></td>
<td>Resends 1056-bit bundle w/errors only instead of entire packet (based on MTU size)</td>
<td>Much lower latency than Forward Error Correction (FEC) defined in the InfiniBand* specification¹</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dynamic Lane Scaling</th>
<th>Description</th>
<th>Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Maintain link continuity in the event of a failure of one of more physical lanes</td>
<td>Enables a workload to continue to completion. <strong>Note:</strong> InfiniBand will shut down the entire link in the event of a physical lane failure</td>
</tr>
<tr>
<td></td>
<td>Operates with the remaining lanes until the failure can be corrected at a later time</td>
<td></td>
</tr>
</tbody>
</table>

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](www.ieee802.org/3/bj/public/jan13/ran_3bj_01a_0113.pdf)
**InfiniBand**

- Application generates messages
- Message segmented in packets of up to Maximum Transfer Unit (MTU) size

  - 256 B → 4 KB

- MPI Message

- Packets sent until **entire** message is transmitted

**Goals:** Improved resiliency, performance, and consistent traffic movement

**Intel® Omni-Path Fabric**

- The HFI segments data into 65-bit containers called Flow Control Digits or "Flits"
- Link Transfer Packets (LTPs) are created by assembling 16 Flits together (plus CRC)

- LTPs send Flits sent over the FABRIC until the entire message is transmitted

- New MTU Sizes Added
  - 8K/10K

- 1 Flit = 65 bits
- 16 Flits = LTP

InfiniBand* supports up to 8KB for MPI Traffic and 10KB MTU for Storage

CRC: Cyclic Redundancy Check

1 Intel® OPA supports up to 8KB for MPI Traffic and 10KB MTU for Storage
Packet Integrity Protection (PIP)
Intel® Omni-Path Fabric Link Level Innovation

1. LTPs Enters the Fabric
   - Bad LTPs explicitly request correction
   - Good LTPs are implicitly acknowledged
   - No Overhead

2. Incoming LTPs Pass In-Line CRC Check
3. Fails CRC Check
4. Discards incoming LTPs after error detection until errored LTP is received

5. Null LTP with Replay Request
   - LTPs Replayed from Port Buffer Starting at Error Point Forward

6. Replay Start Point
   - Replay Buffer
   - Then...

7. PIP raises the effective BER of the link

8. Always enabled – not the case with InfiniBand*

DATA ON TO NEXT STAGE
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
Traffic Flow Optimization (TFO) – Disabled
Intel® Omni-Path Fabric Link Level Innovation

Packet A Arrives Slightly Ahead of MPI Packet B

Sent Packet B and C After Packet A

Host ISL

Packet A (VL0) Storage Traffic
Packet B (VL0) MPI Traffic
Packet C (VL0) MPI Traffic

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

Packet A Arrives Slightly Ahead of MPI Packet B

Packet A (VL0) Storage Traffic
Packet B (VL1) MPI Traffic
Packet C (VL0) MPI Traffic

Large Packet with Lower Priority
Packet B is on a Higher Priority VL
Other Traffic on a Lower Priority VL

Suspends Packet A to Send High Priority Packet B then Resumes A

Same Priority Packet C Transmits After Packet A Completes

Host ISL

Intel® Omni-Path Architecture
48 Radix Switch

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.
### Increased capacity vs 36 port switch

#### 2:1 Oversubscribed

<table>
<thead>
<tr>
<th>CPU</th>
<th># Cores</th>
<th>% vs 36p</th>
</tr>
</thead>
<tbody>
<tr>
<td>E5-2650 v4</td>
<td>768</td>
<td>+33%</td>
</tr>
<tr>
<td>E5-2697 v4</td>
<td>1152</td>
<td></td>
</tr>
</tbody>
</table>

#### 3:1 Oversubscribed

<table>
<thead>
<tr>
<th>CPU</th>
<th># Cores</th>
<th>% vs 36p</th>
</tr>
</thead>
<tbody>
<tr>
<td>E5-2650 v4</td>
<td>864</td>
<td>+33%</td>
</tr>
<tr>
<td>E5-2697 v4</td>
<td>1296</td>
<td></td>
</tr>
</tbody>
</table>
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)

<table>
<thead>
<tr>
<th>2:1 Port Oversubscribed</th>
<th>3:1 Port Oversubscribed</th>
</tr>
</thead>
<tbody>
<tr>
<td>16x 100 Gbps Links</td>
<td>12x 100 Gbps Links</td>
</tr>
<tr>
<td>32 X8 Links</td>
<td>36 X8 Links</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Aggregate Host BW</th>
<th>Aggregate ISL BW</th>
<th>Effective BW Oversub</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up to 1792 Gbps</td>
<td>1600 Gbps</td>
<td>1.12:1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Aggregate Host BW</th>
<th>Aggregate ISL BW</th>
<th>Effective BW Oversub</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up to 2016 Gbps</td>
<td>1200 Gbps</td>
<td>1.68:1</td>
</tr>
</tbody>
</table>
Intel® Omni-Path Fabric's 48 Radix Chip
It’s more than just a 33% increase in port count over a 36 Radix chip

Intel® Omni-Path Architecture (48-port)

THREE-hop Fat Tree

One (1) 768-port Director Switch

Two (2) 648-port Director Switches

768 nodes

InfiniBand® (36-port Switch Chip)

FIVE-hop Fat Tree

(43) 36-port Edge Switches

768 nodes

(43) 36-port Switch Chip

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

<table>
<thead>
<tr>
<th>589 nodes</th>
<th>22x 36-port Edge Switches</th>
<th>9x 36-port Core Switches</th>
</tr>
</thead>
</table>

Intel® Omni-Path Architecture (48-port Switch Chip)
THREE-hop 3:1 Tree

<table>
<thead>
<tr>
<th>589 nodes</th>
<th>17x 48-port Edge Switches</th>
<th>6x 48-port Core Switches</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>31x 36-port Switches</th>
<th>23x 48-port Core Switches</th>
</tr>
</thead>
<tbody>
<tr>
<td>787 Cables</td>
<td>793</td>
</tr>
<tr>
<td>31u Rack Space</td>
<td>23u</td>
</tr>
<tr>
<td>27 nodes (756 cores)</td>
<td>Largest Non-Blocking Set</td>
</tr>
<tr>
<td>(1008 cores)</td>
<td></td>
</tr>
</tbody>
</table>

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?

**FEWER SWITCHES REQUIRED**

<table>
<thead>
<tr>
<th>NODES</th>
<th>InfiniBand® 36-port switch</th>
<th>Intel® OPA 48-port switch</th>
</tr>
</thead>
<tbody>
<tr>
<td>249</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>498</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>749</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>999</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>1249</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>1499</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>1749</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>2049</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>2399</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>2749</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>3099</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>3449</td>
<td>11</td>
<td>11</td>
</tr>
</tbody>
</table>

**More Servers Same Budget**

- **Intel® OPA**
- **Mellanox EDR**

**Up to 24% more Servers**

- 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

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

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

Port latency (includes error detection!)

100-110ns port-to-port latency (with error detection)²

195M messages/sec

160 M messages/sec (MPI message rate uni-directional)

Intel® Omni-Path 48 port Switch

160 M messages/sec

MPI HFI adapter

MPI App

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.
## Latency, Bandwidth, and Message Rate

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

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

**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.
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.

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 (80 silicon). OPA Switch: Series 100 Edge Switch – 48 port (80 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.

MP Latency at Scale
Intel® Omni-Path Architecture (Intel® OPA) vs. InfiniBand® EDR - Open MPI

Natural Order Ring (NOR) Latency

- Intel® OPA: 0.91
- EDR: 1.00

Relative Latency

Intel® OPA up to 9% Faster

Random Order Ring (ROR) Latency

- Intel® OPA: 0.78
- EDR: 1.00

Relative Latency

Intel® OPA up to 21% Faster

LOWER is Better

16 Nodes, 32 MPI ranks per node
MPI Latency at Scale
Intel® Omni-Path Architecture (Intel® OPA)

Natural (NOR) and Random (ROR) Ring Latency

- NOR
- ROR

Lower is Better

Latency (µs)

Number of Nodes

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/composerxe_2015.1.133/mkl -O2 -xCORE2-AVX -ip -ansi-alias -fno-alias -DLONG_IS_64BITSDRA_SANDBIA_OPT2 -DUSING_FFTW -DHPCC_FFT_235. Pre-production Intel® Omni-Path hardware and software. IFS 10.0.0.625. 2133 MHz DDR4 memory per node. RHEL 7.0.
Intel® Omni-Path Architecture (Intel® OPA)

Application Performance - Intel® MPI - 16 Nodes

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

*Spec MPI2007 results estimates until published

**see following slide for system configurations

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.
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.

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

*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.

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
Intel® Omni-Path Architecture
Disruptive innovations to knock down the “I/O Wall”

21% 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 EDR1

24% 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 budget2

60% 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* EDR3

GREATER RESILIENCY
“no compromise” error detection and maintains link continuity with lane failures

No additional latency penalty for error detection with Packet Integrity Protection4

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 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 v3.0 (2/25/13). Best of default, MXM_TLS=sel,rc, and -mca pml yaalla 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. 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 15, 2015. 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.

- LST-DYNA MPF R8.1.0 dynamic link. Intel Fortran Compiler 17.0.0.20160512. Intel® MPI Library 2017.0.2.0.0 (OFED 2.0.0.0 (OFED)). Mellanox EDR ConnectX-4 Single Port Rev 3 MCX455A HCA. Mellanox SB7700 - 36 Port EDR Infiniband switch.

System & Software Configuration


- LST-DYNA MPF R8.1.0 dynamic link. Intel Fortran Compiler 17.0.0.20160512. Intel® MPI Library 2017.0.2.0.0 (OFED 2.0.0.0 (OFED)). Mellanox EDR ConnectX-4 Single Port Rev 3 MCX455A HCA. Mellanox SB7700 - 36 Port EDR Infiniband switch.

- LST-DYNA MPF R8.1.0 dynamic link. Intel Fortran Compiler 17.0.0.20160512. Intel® MPI Library 2017.0.2.0.0 (OFED 2.0.0.0 (OFED)). Mellanox EDR ConnectX-4 Single Port Rev 3 MCX455A HCA. Mellanox SB7700 - 36 Port EDR Infiniband switch.

- LST-DYNA MPF R8.1.0 dynamic link. Intel Fortran Compiler 17.0.0.20160512. Intel® MPI Library 2017.0.2.0.0 (OFED 2.0.0.0 (OFED)). Mellanox EDR ConnectX-4 Single Port Rev 3 MCX455A HCA. Mellanox SB7700 - 36 Port EDR Infiniband switch.


- LST-DYNA MPF R8.1.0 dynamic link. Intel Fortran Compiler 17.0.0.20160512. Intel® MPI Library 2017.0.2.0.0 (OFED 2.0.0.0 (OFED)). Mellanox EDR ConnectX-4 Single Port Rev 3 MCX455A HCA. Mellanox SB7700 - 36 Port EDR Infiniband switch.

- LST-DYNA MPF R8.1.0 dynamic link. Intel Fortran Compiler 17.0.0.20160512. Intel® MPI Library 2017.0.2.0.0 (OFED 2.0.0.0 (OFED)). Mellanox EDR ConnectX-4 Single Port Rev 3 MCX455A HCA. Mellanox SB7700 - 36 Port EDR Infiniband switch.