Intel® Omni-Path Architecture
Product Overview

Andrew Russell, HPC Solution Architect
May 2016
Legal Disclaimer

Copyright © 2016 Intel Corporation. All rights reserved.

Intel®, Xeon®, Xeon® Phi™, and Intel Atom™ are trademarks of Intel Corporation in the U.S. and/or other countries. *Other names and brands in this document may be claimed as the property of others.

Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

This document contains information on products in development and do not constitute Intel plan of record product roadmaps. All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. Designers must not rely on or finalize plans based on the information in this document or any features or characteristics described. Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. Contact your Intel representative to obtain Intel’s current plan of record product roadmaps.

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL’S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

A "Mission Critical Application" is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL’S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL AND ITS SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS’ FEES ARISING OUT OF, DIRECTLY OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENT IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.
WHO IS IN THE AUDIENCE?

- Fabric administrators
- MPI developers
- Knowledge of
  - InfiniBand
  - True Scale (aka QLogic InfiniBand)
RELEVANCE TO THIS AUDIENCE?

• Omni-Path is available on Cray CS Series Clusters
• Future?
Intel Fabrics over time

Forecast and Estimations, in Planning & Targets

Potential future options, subject to change without notice. Codenames.
All timeframes, features, products and dates are preliminary forecasts and subject to change without further notification.
Omni-Path (OPA): a quick introduction

• It's like InfiniBand?
  • Yes, very much like InfiniBand
  • Switches, adapters and cables
  • Verbs/RDMA and PSM APIs
  • LIDs and GUIDs, Fat-trees and Subnet Managers
  • Similar admin commands to True Scale

• But it's not InfiniBand
  • Link Layer is different – More functionality
  • So cannot be directly connected to InfiniBand
Omni-Path gen1 Architecture

OPA technology at a glance

- Enhanced Intel® True Scale host stack on new 100Gb hardware
  - Link layer from Cray* Aries:
    - Packet pre-emption and interleaving minimises the impact of large storage packets on latency sensitive MPI traffic
    - Error correction optimised for latency
    - Enhanced to 100Gb
      *Significant scalability benefits over InfiniBand* roadmap.
  - Host stack from True Scale
    - PSM: Connectionless tag-matching protocol
    - Proven scalable HPC platform
- Integration
  - Developing over time with each CPU generation

*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>
<th>Software</th>
<th>Cables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single port</td>
<td>1U Form Factor</td>
<td>QSFP-based</td>
<td>OEM custom designs</td>
<td>Open Source</td>
<td>Third Party Vendors</td>
</tr>
<tr>
<td>x8 and x16</td>
<td>24 and 48 port</td>
<td>192 and 768 port</td>
<td>HFI and Switch ASICs</td>
<td>Host Software and Fabric Manager</td>
<td>Passive Copper</td>
</tr>
<tr>
<td>x16</td>
<td>48-port Edge Switch</td>
<td>768-port Director Switch</td>
<td>HFI silicon</td>
<td>Up to 2 ports (50 GB/s total b/w)</td>
<td>Active Optical</td>
</tr>
<tr>
<td>Adapter</td>
<td>24-port Edge Switch</td>
<td>192-port Director Switch</td>
<td>Switch silicon</td>
<td>up to 48 ports (1200 GB/s total b/w)</td>
<td></td>
</tr>
<tr>
<td>(100 Gb/s)</td>
<td>(7U chassis)</td>
<td>(7U chassis)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x8 Adapter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(58 Gb/s)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></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.
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
Intel® OPA

Next generation
Additional integration, improvements, and features

Next Intel® Xeon® Phi™ processor (Knight Hill)
Future Intel® Xeon® processor (14nm)
Intel® Xeon Phi™ processor (Knights Landing)
Next-Generation Intel® Xeon® processor
Intel® Xeon® processor E5-2600 v3

Twinax Cable
Intel® XEON PHII inside
Intel® XEON PHII inside

INTEGRATION
INTEGRATION
What integration looks like

Xeon Phi: KNL-F

- PCIe carrier board, 2-port version (sideband cable and IFT connectors and cages on underside of the card)
- (2) Internal-to-Faceplate Processor (IFP) cable supporting two-ports
- Top view of card
- Bottom view of card
- QSFP sideband header (cable not shown) Connects to header on motherboard

EACH port requires:
(1) Internal Faceplate Transition (IFT) Connector
(1) IFT Cage Connector

Copyright © 2015, Intel Corporation.
The host sw stack, and PSM

Verbs-based
- Utilities
- Filesystems
- Etc

Applications
- OFED
- ULPs
- Using verbs drivers

Generic
- Verbs Provider / Driver
- Traditional InfiniBand HCA

PSM-based
- Utilities
- Filesystems
- Etc

Applications
- OFED
- ULPs
- Using PSM drivers

Adapter Specific
- Verbs Provider / Driver
- PSM2 Library

Wire Transports
- TrueScale HCA or Omni-Path HFI

PSM1 compatibility layer
INTEL® OPA LINK LEVEL INNOVATION STARTS HERE
LAYER 1.5: LINK TRANSFER LAYER

**InfiniBand**

- Application generates messages
- Message segmented in packets of up to Maximum Transfer Unit (MTU) size
  - MPI Message
    - 256 B → 4 KB
    - 256 B → 4 KB
  - Packets sent until entire message is transmitted

**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)
  - CRC: Cyclic Redundancy Check
  - Data Flit(s)
  - Control Flit(s) (optional)

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

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

8K/10K

1 Flit = 65 bits

16 Flits = LTP

1 Intel® OPA supports up to 8KB for MPI Traffic and 10KB MTU for Storage

CRC: Cyclic Redundancy Check

Intel Confidential

Intel and the Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries. Other names and brands may be claimed as the property of others. All products, dates, and figures are preliminary and are subject to change without any notice. Copyright © 2015, Intel Corporation.
Layer Innovation: Traffic Flow Optimization (TFO) - Enabled

Traffic Enters the Fabric

Host Ports

ISL Ports

Intel® Omni-Path Architecture
48 Radix Switch

Same Priority Packet C Transmits after Packet A Completes

Packet A Suspended to Send High Priority Packet B then Packet A Resumes to Completion

Packet A (VL0) – Low Priority
Storage Traffic
Packet B (VL1) – High Priority
MPI Traffic
Packet C (VL0) – Low Priority
Other Traffic

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

Packets Transiting the Same ISL
Packet A starts Transmitting 1ns prior to High Priority Packet B Arrives

Intel Confidential

Intel technologies’ features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration.
Traffic Flow Optimization (TFO): MPI Performance Results

- Qos Under Congested Link Conditions

**Traffic Flow Optimization (TFO): MPI Performance Results**

**Qos Under Congested Link Conditions**

*See Test Setup:*

- Server Configuration: Intel(R) Xeon(R) CPU E5-2699 v3 @ 2.30GHz, Turbo Disabled, Intel OPA 10.0.0.990.48 Software, RHEL 7.0, Kernel 3.10.0-123.el7.x86_64

**BW allocation 10%/80% - (Avg. 80 Iterations)**

**Relative Base MPI Latency**

- No Congestion
- High Priority MPI Traffic with Contention

**Based on preliminary Intel internal testing using two pre-production Intel® OPA edge switch (A0)es with one inter-switch link, comparing MPI latency over multiple iterations with varying bandwidth allocations for storage and MPI traffic over multiple virtual lanes, both with Traffic Flow Optimization enabled and disabled.**
STORAGE: CONNECTING TO NEW AND EXISTING SYSTEMS

NEW systems:
- Key HPC storage vendors will deliver Intel® OPA-based storage devices

Accessing storage in EXISTING systems:
- Multi-homed solution
  - Direct-attach Intel® OPA to existing file system server along with the existing fabric connection
- Router solution
  - Lustre: Supported via LNET Router
  - GPFS/NAS/Other: Supported via IP Router

"Implementing Storage in Intel® Omni-Path Architecture Fabrics" white paper available now (public link)
"Intel® Omni-Path Storage Router Design Guide" available now (ask for access)
Accessing Existing Storage

Router solution

Existing IB cluster

New OPA cluster

Login Nodes

Compute Nodes

Storage Servers

Routers

Multi-Homed Solution

Existing IB cluster

New OPA cluster

Login Nodes

Compute Nodes

Storage Servers

Key enabler:
Interfaces must co-exist
How It Used To Be...

Install manufacture's software, developed from OFED

- Software for the device comes from the manufacture of the device.
- No way to support coexistence of different devices.
- In fact – it is often very difficult to get coexistence of different devices working in the first place.
- **OS updates can break the device drivers!**

Installing M-OFED or IntelIB overwrites the fabric support provided in the distro.
...How We Do It Now

**Push support into the distro**

- Support for OPA is upstreamed to the Linux kernel.
- Red Hat back-port committed additions and changes to their current kernel.
- Software for one device does not overwrite support for another.
- Red Hat support any combination of devices whose software is in-distro.
- **OS updates can be applied safely**

Installing IntelOPA is [almost entirely] additive. It installs just those components not yet integrated into the distro, and overwrites little or nothing.
Cost advantages

Compute / interconnect cost ratio has changed

- Compute price/performance improvements continue unabated
- Current corresponding fabric metrics unable to keep pace as a percentage of total cluster costs which includes compute and storage

Challenge: Keeping fabric costs in check to free up cluster $$$ for increased compute and storage capability

Hardware Cost Estimate

- Up to 10% lower cluster cost mix than either FDR or EDR (at similar bandwidths)

More Compute = More FLOPS


Intel Confidential

Intel and the Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries. Other names and brands may be claimed as the property of others. All products, dates, and figures are preliminary and are subject to change without any notice. Copyright © 2015, Intel Corporation.
Intel® OPA Momentum is Building Quickly

Worldwide design wins keep rolling in

>100 OEM platform, switch, and adapters expected in 1H’16¹

**Worldwide Wins**

- Penguin: US DoE CTS-1
- Dell: TACC Stampede 1.5
- HPE: Pittsburgh Supercomputing
- Inspur: Qingdao, Tsinghua University
- Dell: NCAR, NASA, Uni Colorado
- Sugon: Beijing Academy of Science and Technology

**EMEA Wins**

- Clustervision: AEI Potsdam, University of Hull
- Dell: Wartsilla, University of Sheffield
- Lenovo: Cineca
- Cray: AWI, Juelich

- 280 nodes pre-stage “just worked”
- >800 nodes deployed in 3 days!
- Plus more that cannot be named at this time

¹ Source: Intel internal information