Intel® Omni-Path Architecture

Product Overview

Andrew Russell, HPC Solution Architect

May 2016

Legal Disclaimer

Intel® Xeon®, Intel® 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 SUBSIDIARIES, 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.

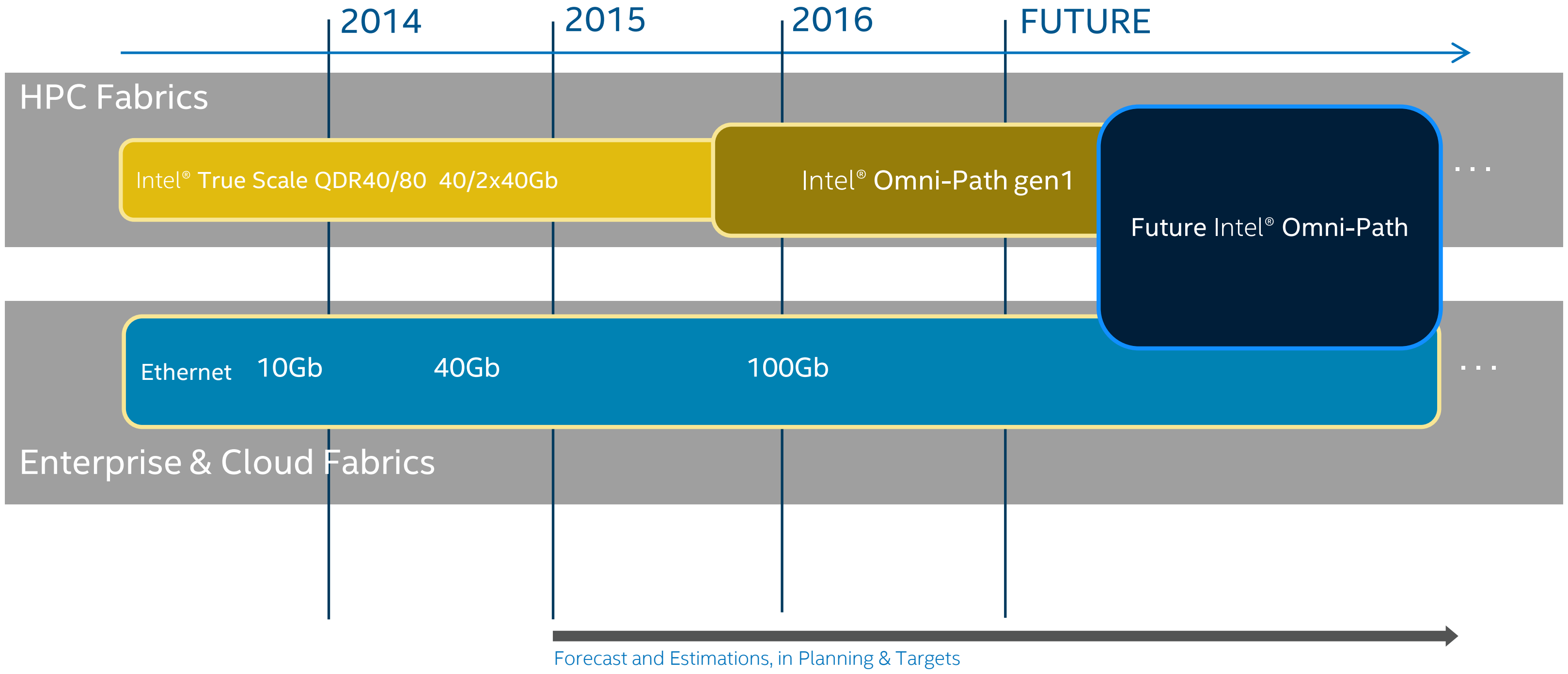
WHO IS IN THE AUDIENCE?

- 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


2014
2015
2016
FUTURE
HPC Fabrics
Intel® True Scale QDR40/80 40/2x40Gb
Intel® Omni-Path gen1
Future Intel® Omni-Path
. . .
Ethernet 10Gb
40Gb
100Gb
. . .
Enterprise & Cloud Fabrics
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

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

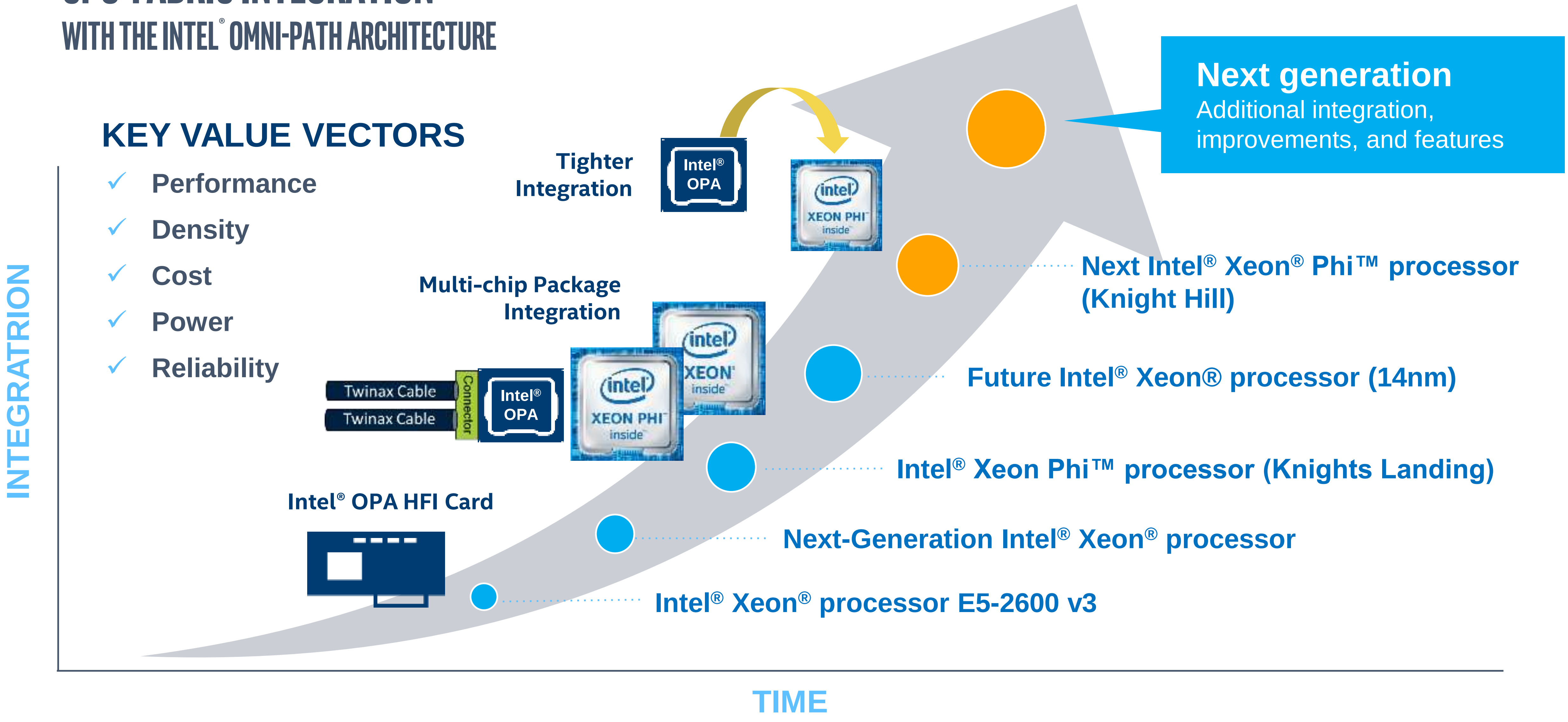
[1] Source: Intel internal information. Design win count based on OEM and HPC storage vendors who are planning to offer either Intel-branded or custom switch products, along with the total number of OEM platforms that are currently planned to support custom and/or standard Intel® OPA adapters. Design 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.

CPU-FABRIC INTEGRATION

WITH THE INTEL® OMNI-PATH ARCHITECTURE

KEY VALUE VECTORS

- ✓ Performance
- ✓ Density
- ✓ Cost
- ✓ Power
- ✓ Reliability



Next generation

Additional integration, improvements, and features

Tighter Integration

Multi-chip Package Integration

Intel® OPA HFI Card

Next Intel® Xeon® Phi™ processor (Knight Hill)

Future Intel® Xeon® processor (14nm)

Intel® Xeon Phi™ processor (Knights Landing)

Next-Generation Intel® Xeon® processor

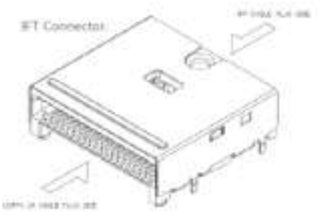
Intel® Xeon® processor E5-2600 v3

What integration looks like

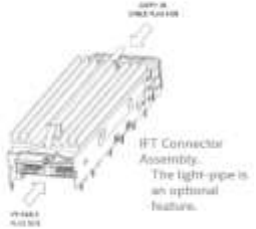
Xeon Phi: KNL-F

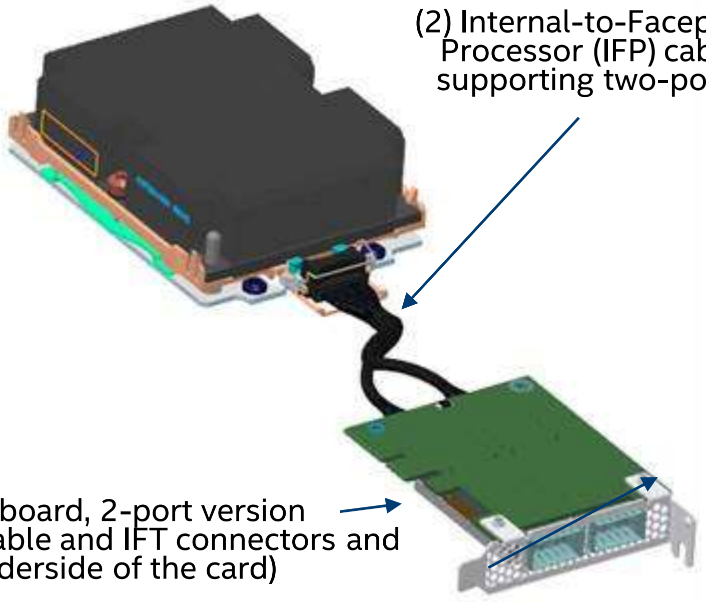
EACH port requires:

(1) Internal Faceplate Transition (IFT) Connector



(1) IFT Cage

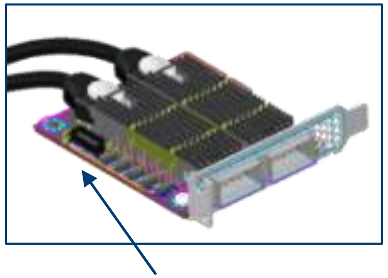




(2) Internal-to-Faceplate Processor (IFP) cable supporting two-ports

Bottom view of card

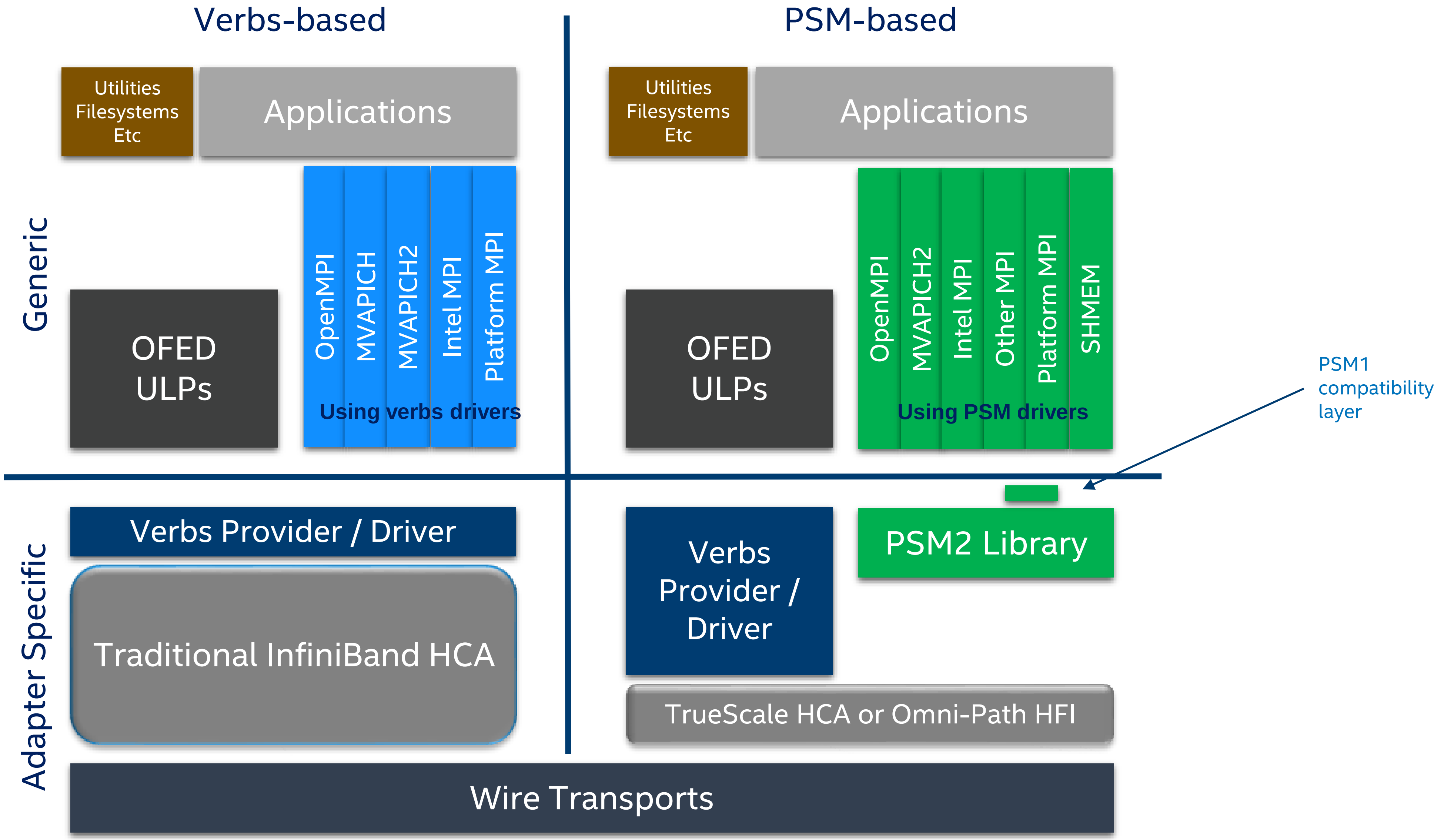
PCIe carrier board, 2-port version (sideband cable and IFT connectors and cages on underside of the card)



Top view of card

QSFP sideband header (cable not shown) Connects to header on motherboard

The host sw stack, and PSM



Verbs-based

PSM-based

Generic

Adapter Specific

Utilities Filesystems Etc

Applications

OFED ULPs

OpenMPI

MVAPICH

MVAPICH2

Intel MPI

Platform MPI

Using verbs drivers

Utilities Filesystems Etc

Applications

OFED ULPs

OpenMPI

MVAPICH2

Intel MPI

Other MPI

Platform MPI

SHMEM

Using PSM drivers

PSM1 compatibility layer

Verbs Provider / Driver

Traditional InfiniBand HCA

Verbs Provider / Driver

PSM2 Library

TrueScale HCA or Omni-Path HFI

Wire Transports

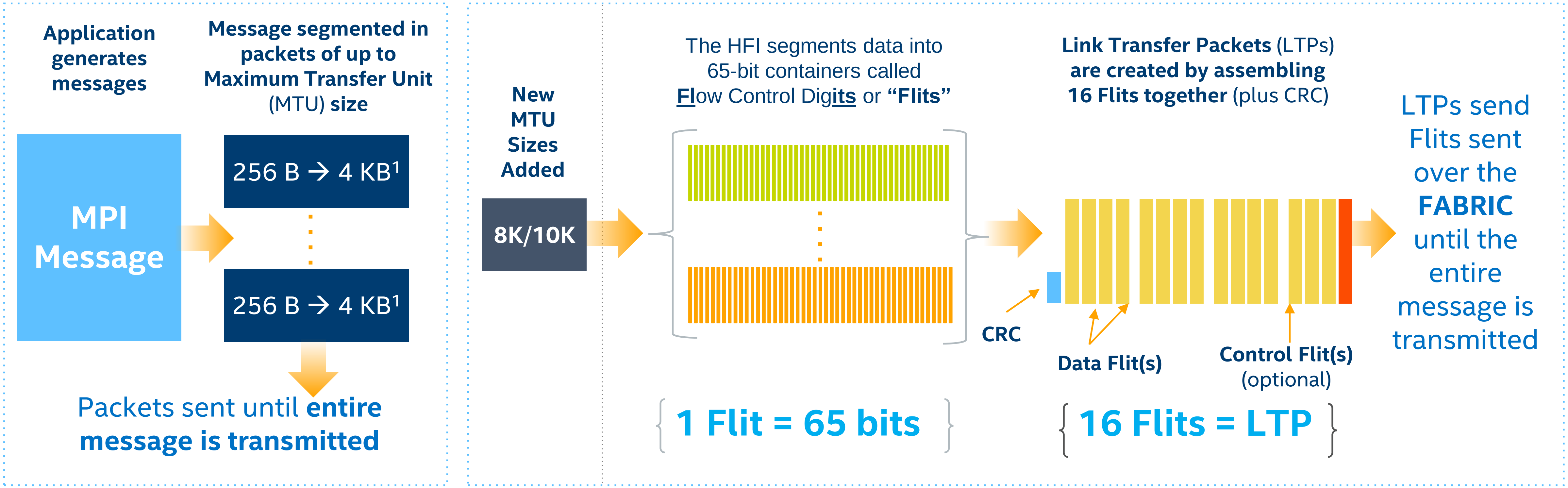
*Other names and brands may be claimed as the property of others.

INTEL® OPA LINK LEVEL INNOVATION STARTS HERE

LAYER 1.5: LINK TRANSFER LAYER

InfiniBand*

Intel® Omni-Path Fabric



Application generates messages

Message segmented in packets of up to Maximum Transfer Unit (MTU) **size**

MPI Message

256 B → 4 KB[1]

256 B → 4 KB[1]

Packets sent until **entire message is transmitted**

New MTU Sizes Added

8K/10K

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

Data Flit(s)

Control Flit(s) (optional)

1 Flit = 65 bits

16 Flits = LTP

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

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

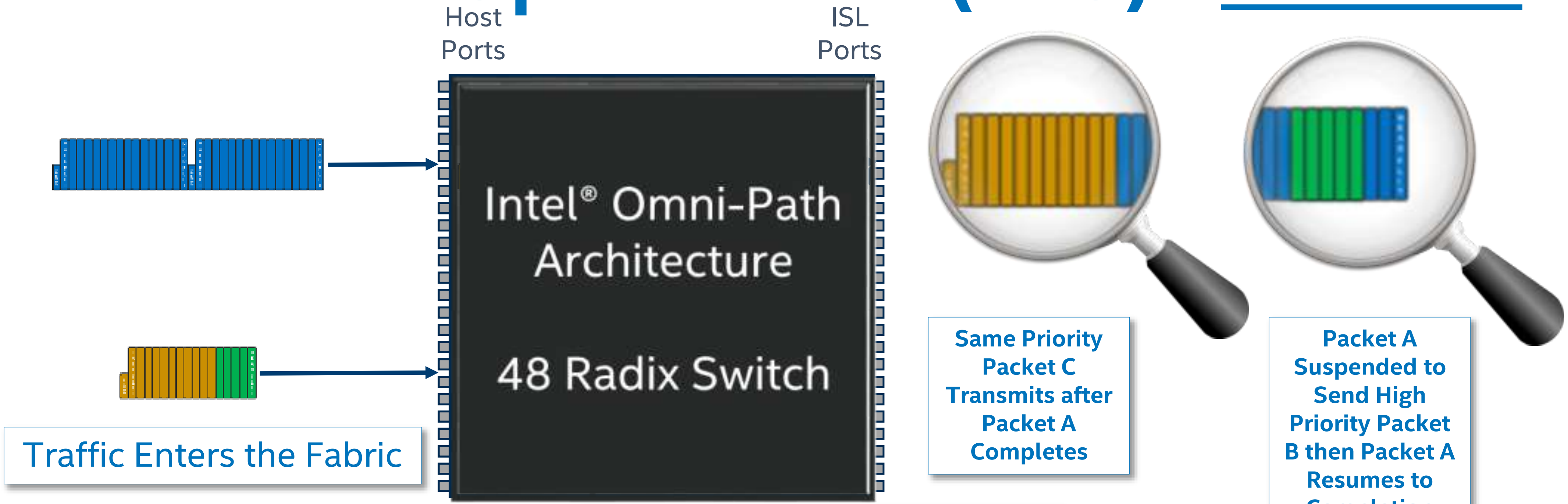
CRC: Cyclic Redundancy Check

Goals: Improved resiliency, performance, and consistent traffic movement

Intel Confidential

Layer Innovation:

Traffic Flow Optimization (TFO) - Enabled



Traffic Enters the Fabric

Same Priority Packet C Transmits after Packet A Completes

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

Packets Transiting the Same ISL

Packet A starts Transmitting 1ns prior to High Priority Packet B Arrives

🟦 Packet A (VL0) – Low Priority
Storage Traffic

🟩 Packet B (VL1) – High Priority
MPI Traffic

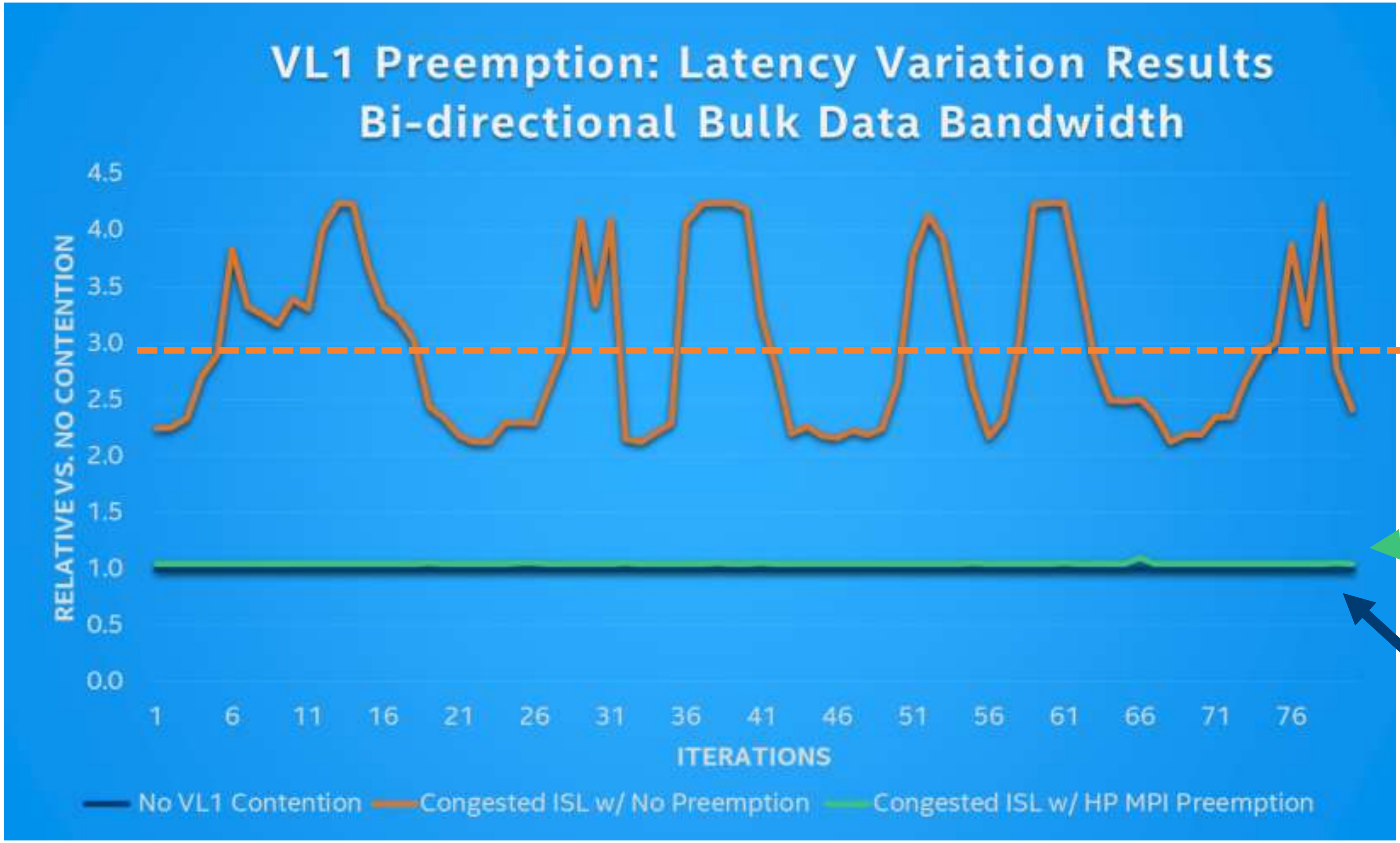
🟧 Packet C (VL0) – Low Priority
Other Traffic

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

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



MPI Job Running over an ISL Containing Bi-directional Bulk Data ~24.7GB on a Separate VL

No Prioritization with Data Contention

TFO Off (No Preemption) Average Latency

High Priority MPI Traffic with Contention

TFO Enabled (Preemption)

Relative Base MPI Latency – No Congestion

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.

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

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

STORAGE: CONNECTING TO NEW AND EXISTING SYSTEMS

NEW systems:

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

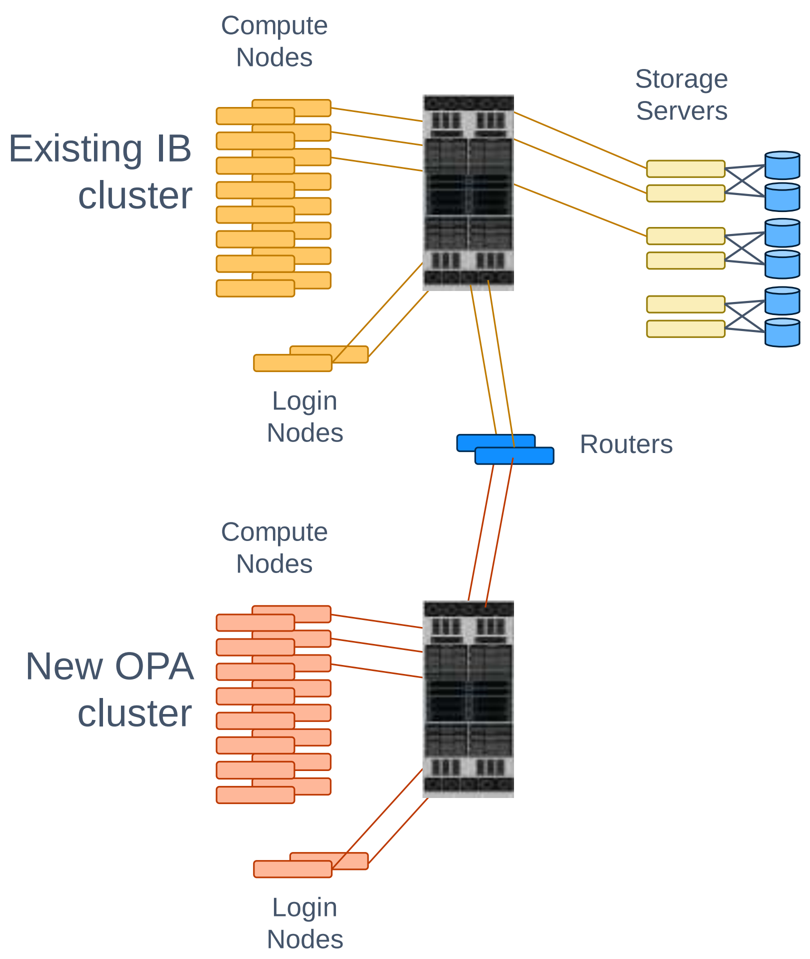
Accessing storage in EXISTING systems:

- 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

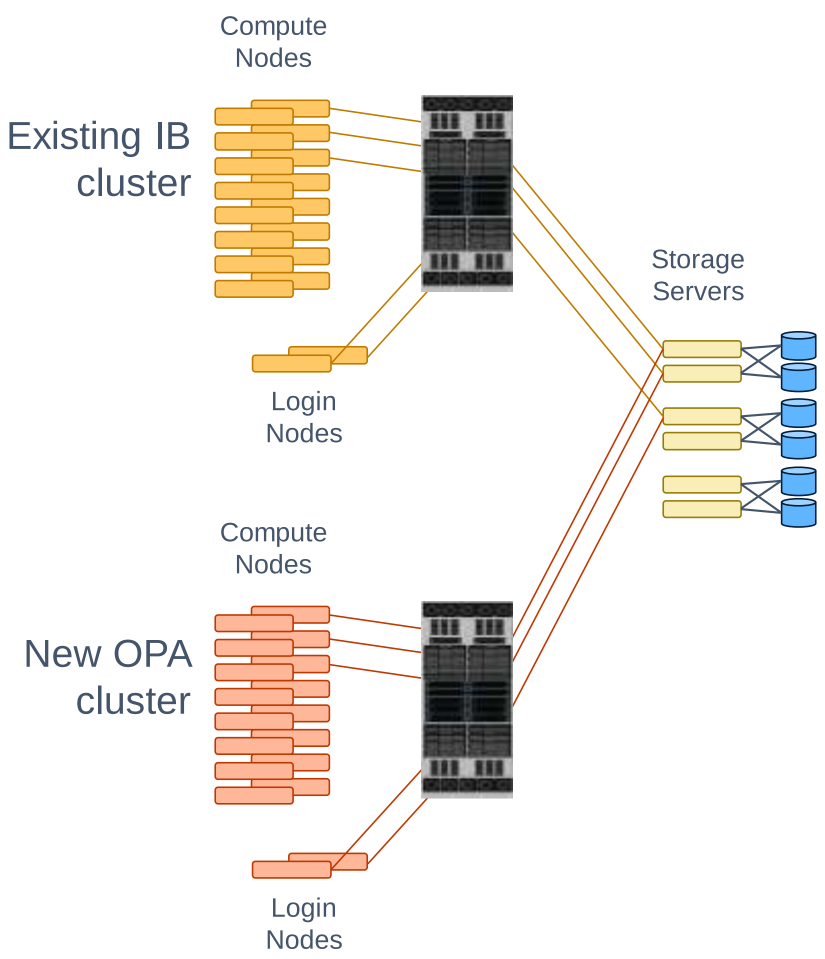
"Implementing Storage in Intel® Omni-Path Architecture Fabrics" white paper available now (*public link*)
"Intel® Omni-Path Storage Router Design Guide" available now (ask for access)

Accessing Existing Storage

Router solution



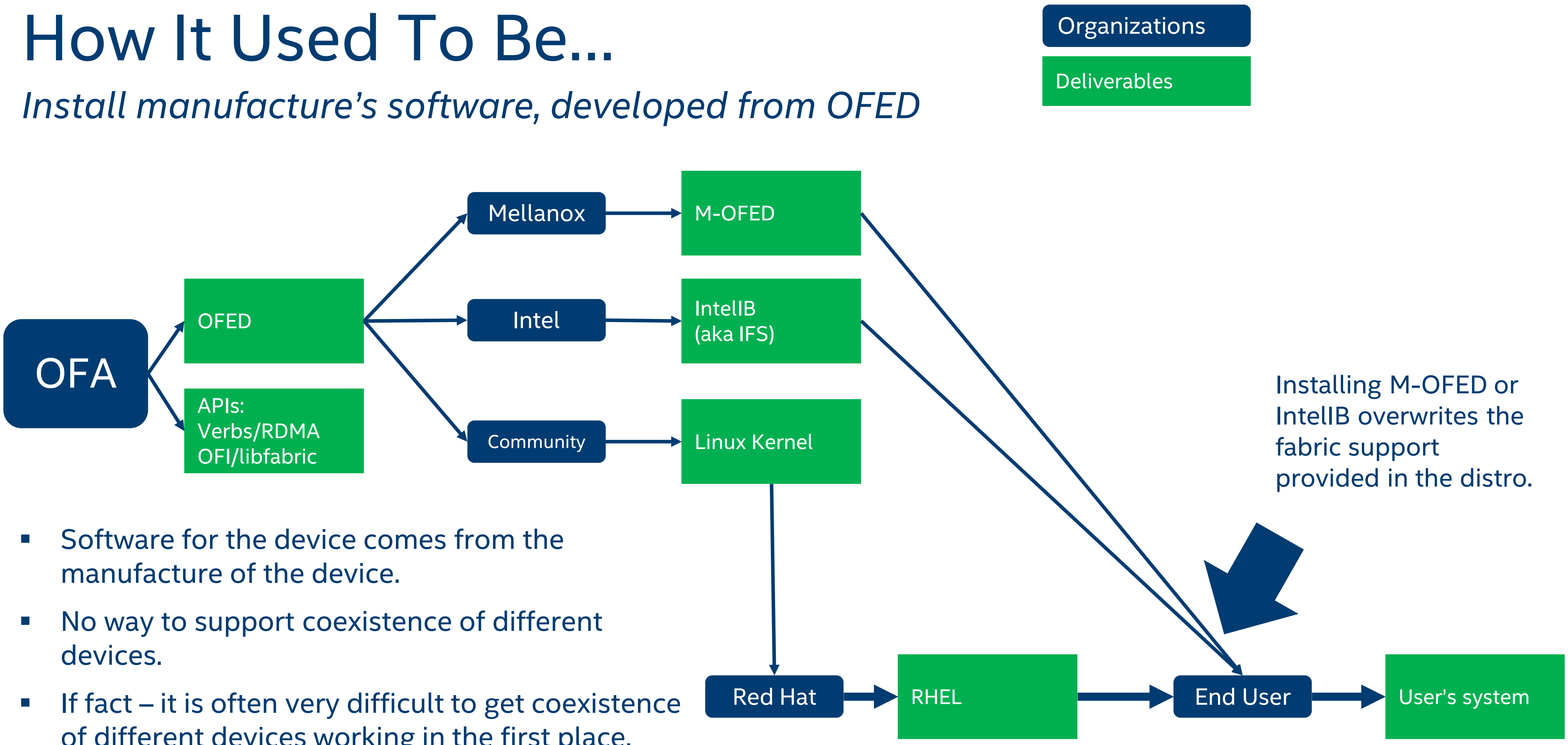
Multi-Homed Solution



Key enabler:
Interfaces must co-exist

How It Used To Be…

Install manufacture's software, developed from OFED

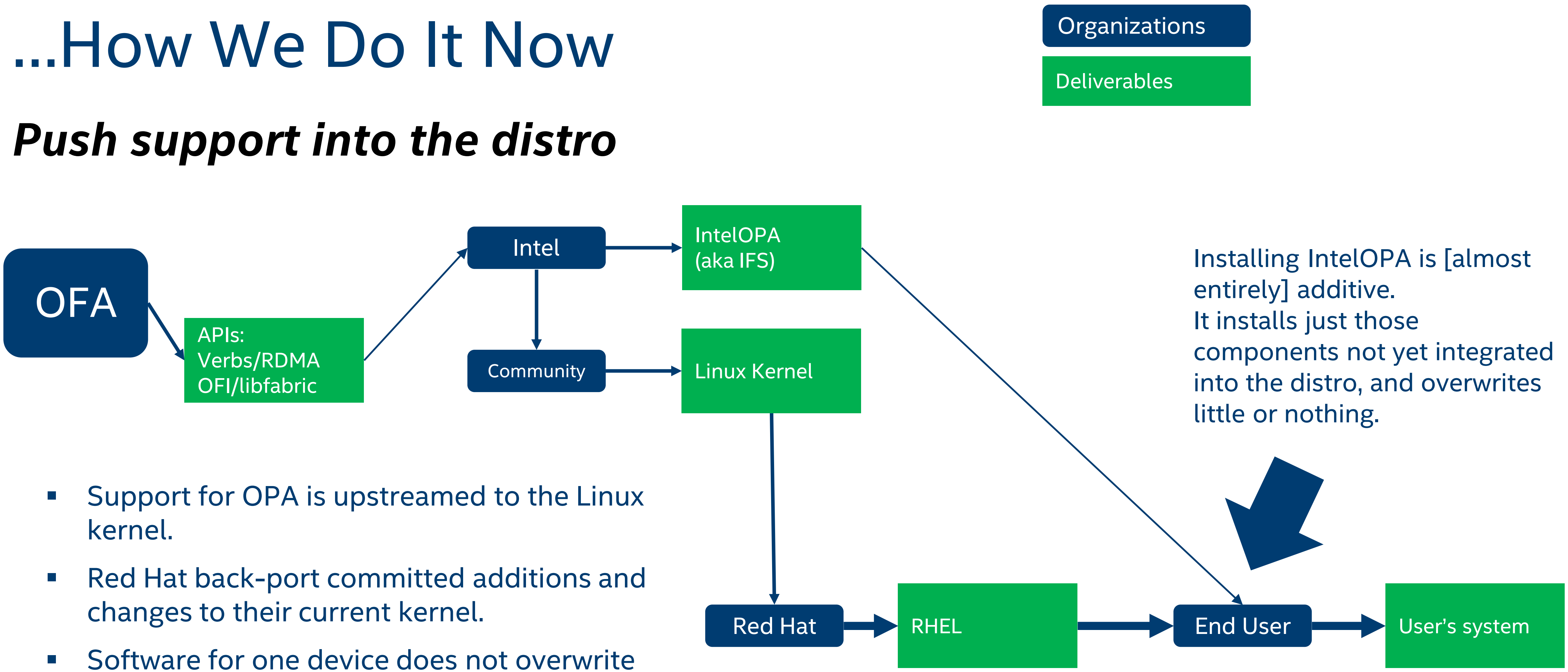


Installing M-OFED or IntelIB overwrites the fabric support provided in the distro.

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

...How We Do It Now

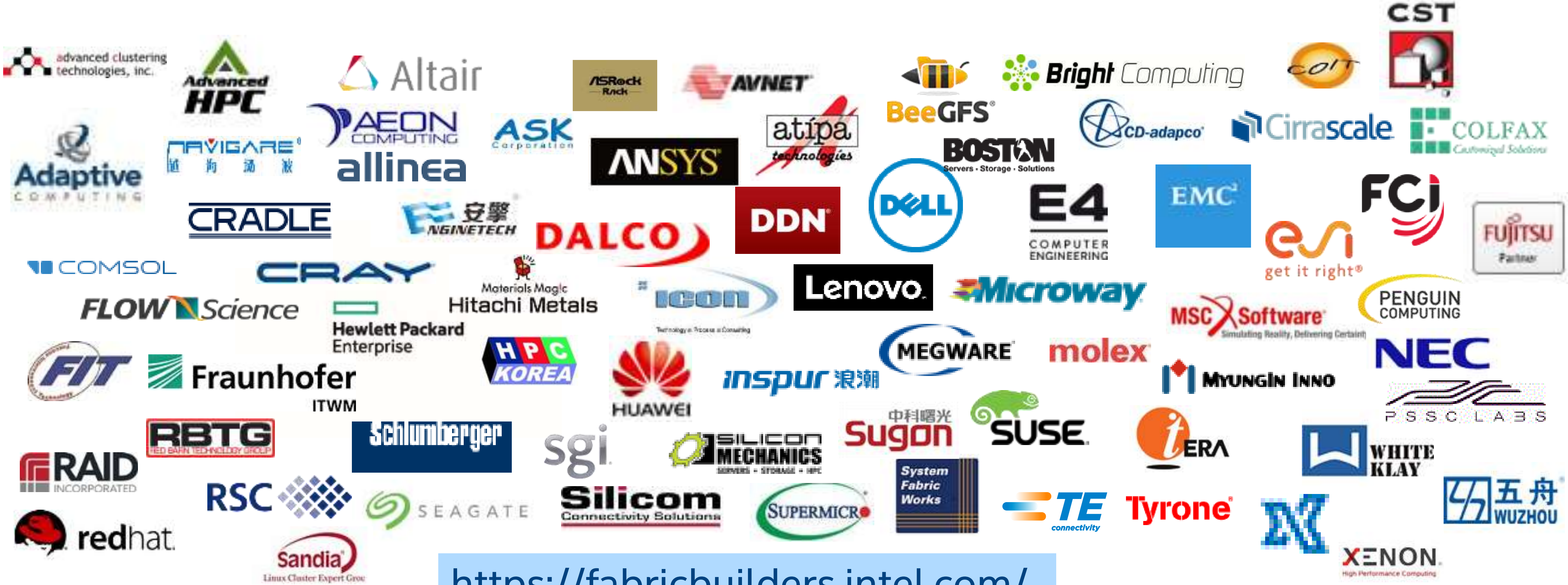
Push support into the distro



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

- 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***

Acceptance: Intel® Fabric Builders



https://fabricbuilders.intel.com/

Last update November 13, 2015

Cost advantages

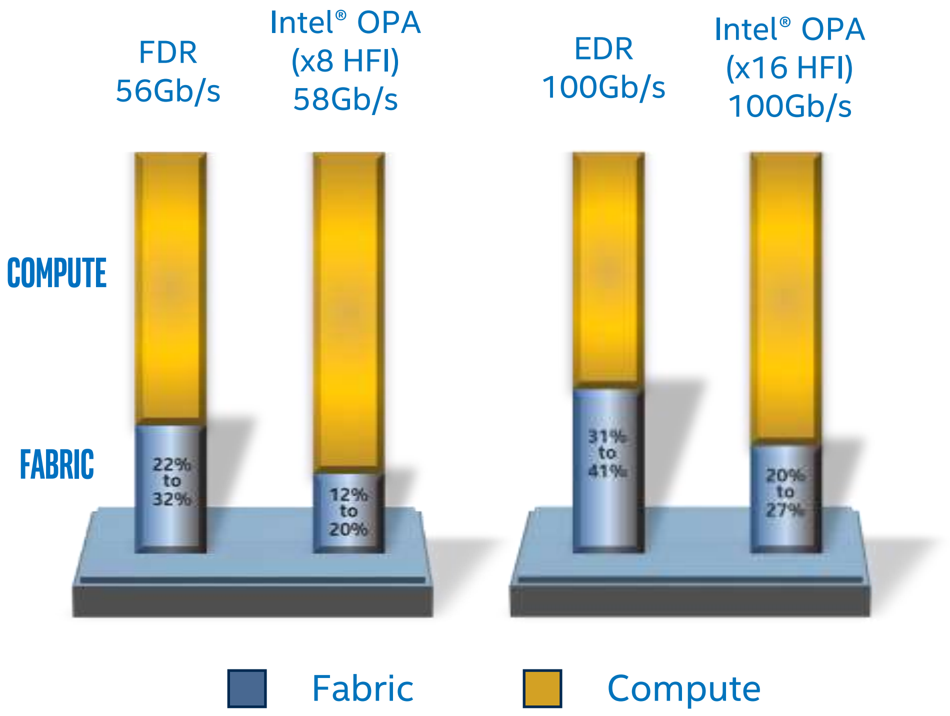
Compute / interconnect cost ratio has changed

- 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

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

Hardware Cost Estimate[1]



More Compute = More FLOPS

[1] 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® OPA Momentum is Building Quickly

Worldwide design wins keep rolling in

>100 OEM platform, switch, and adapters expected in 1H’16[1]

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

>800 nodes deployed in 3 days!

280 nodes pre-stage “just worked”

EMEA Wins

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

- Plus more that cannot be named at this time

[1] Source: Intel internal information

intel®