13th ANNUAL WORKSHOP 2017

INTEL® OMNI-PATH STATUS, UPSTREAMING AND ONGOING WORK

Todd Rimmer, Omni-Path Lead Software Architect

Intel Corporation

March, 2017

LEGAL DISCLAIMERS

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.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The cost reduction scenarios described in this document are intended to enable you to get a better understanding of how the purchase of a given Intel product, combined with a number of situation-specific variables, might affect your future cost and savings. Circumstances will vary and there may be unaccounted-for costs related to the use and deployment of a given product. Nothing in this document should be interpreted as either a promise of or contract for a given level of costs.

Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.

Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families: Go to: Learn About Intel® Processor Numbers

All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or go to: http://www.intel.com/design/literature.htm

The High-Performance Linpack (HPL) benchmark is used in the Intel® FastFabrics toolset included in the Intel® Fabric Suite. The HPL product includes software developed at the University of Tennessee, Knoxville, Innovative Computing Libraries.

Intel, Intel Xeon, Intel Xeon Phi™ are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States or other countries.

AGENDA

- **Omni-Path Overview**
- **Omni-Path Innovations**
- **Omni-Path Upstreaming**
- **Omni-Path Ongoing Work**

OMNI-PATH OVERVIEW

- **Omni-Path is a new fabric technology**
  - Omni-Path is NOT InfiniBand
- **Significant deployments in Nov 2016 Top 500 list**
  - Clusters → 28
  - Flops → 43.7PF
  - Top10 → 1 system (#6)
  - Top15 → 2 system (#6, #12)
  - Top50 → 4 systems
  - Top100 → 10 systems
  - Xeon Efficiency → 88.5%
  - HPCG list → #3

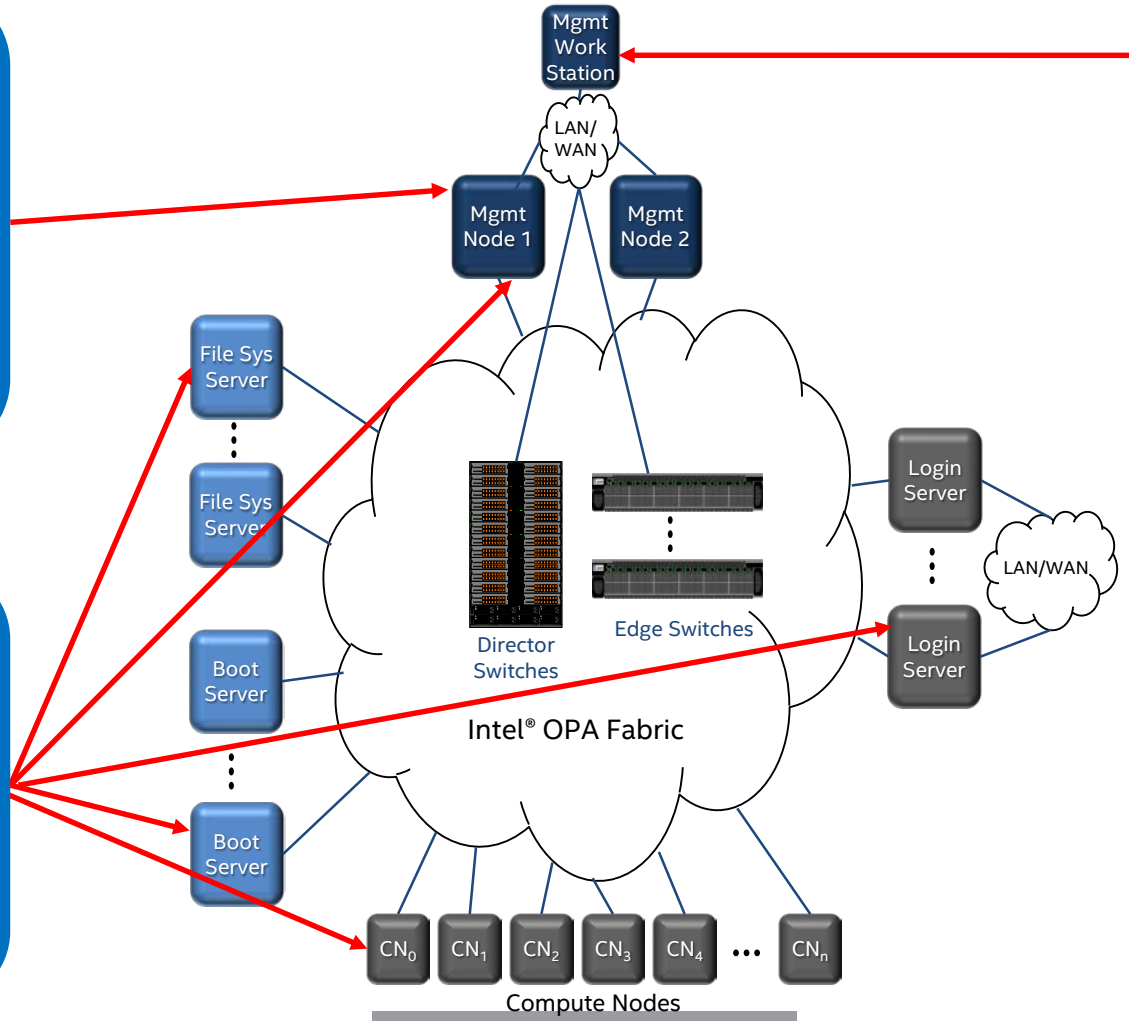
INTEL® OMNI-PATH HOST SOFTWARE COMPONENTS

Fabric Management Stack

- Runs on OPA-connected management nodes
- Initializes, configures and monitors the fabric routing, QoS, security, and performance
- Includes toolkit for TCO functions: configuration, monitoring, diags, and repair

Host Software Stack

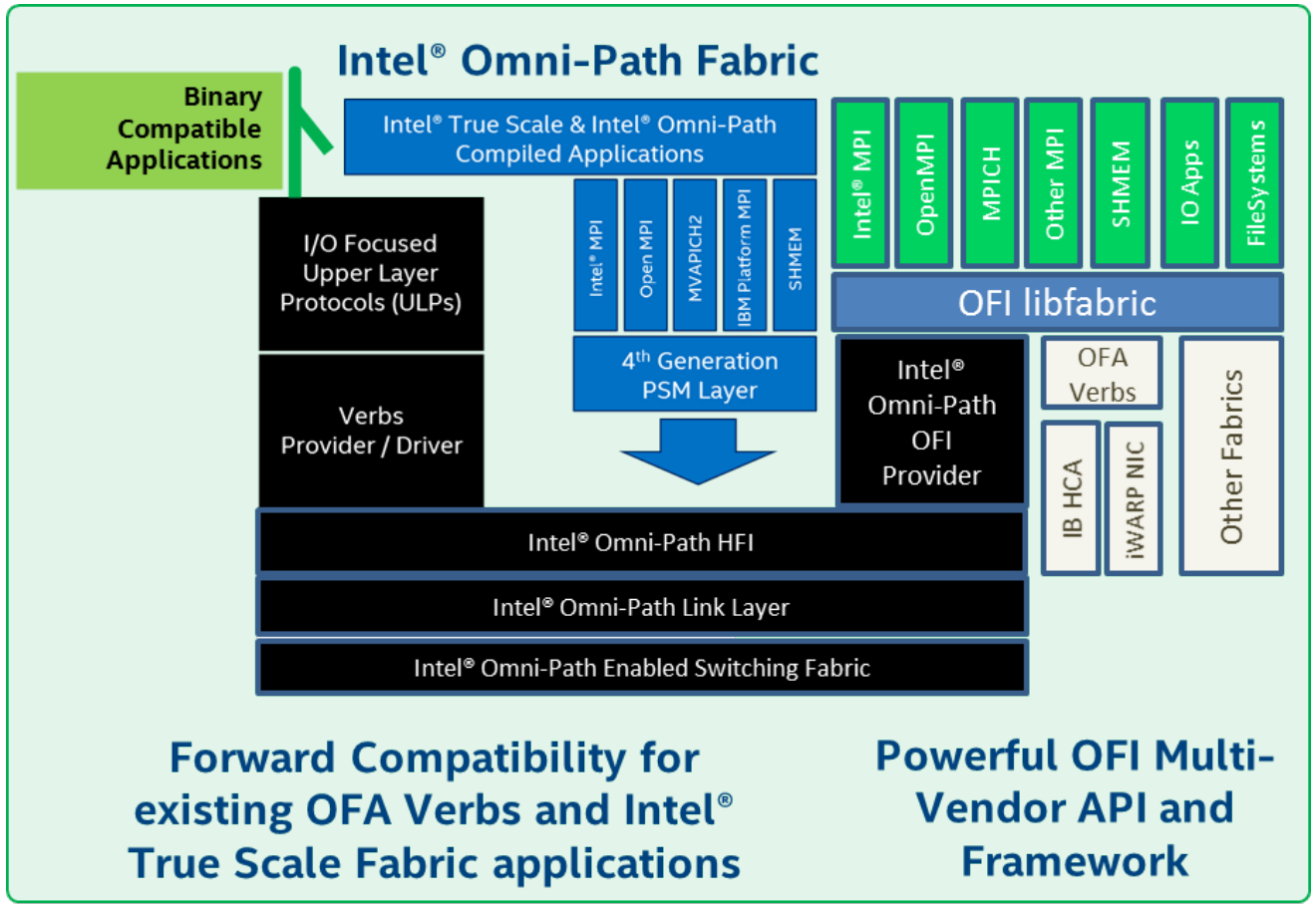
- Runs on all Intel® OPA-connected host nodes
- High performance, highly scalable MPI implementation and extensive set of upper layer protocols



Fabric Management GUI

- Runs on workstation with a local screen/keyboard
- Provides interactive GUI access to Fabric Management TCO features (configuration, monitoring, diagnostics, element management drill down)

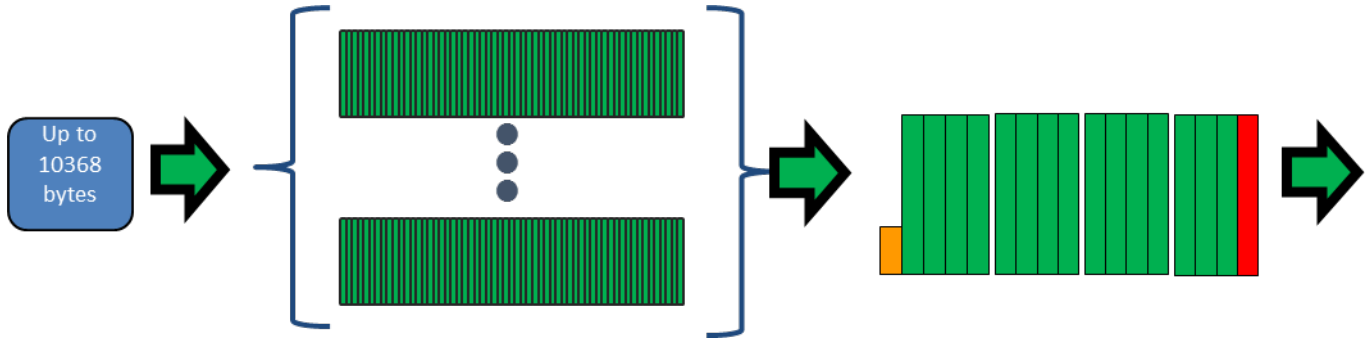
LAYER 4: TRANSPORT LAYER AND KEY SOFTWARE



- **Open Fabrics Interface (OFI) libfabric**
  - Framework providing an API applications and middleware can use for multiple vendors and L4 protocols
  - Allows vendor innovation w/o requiring application churn
- **Performance Scaled Messaging (PSM)**
  - API and corresponding L4 protocol designed for the needs of HPC
- **Open Fabrics Alliance verbs**
  - API and corresponding L4 protocol designed for RDMA IO

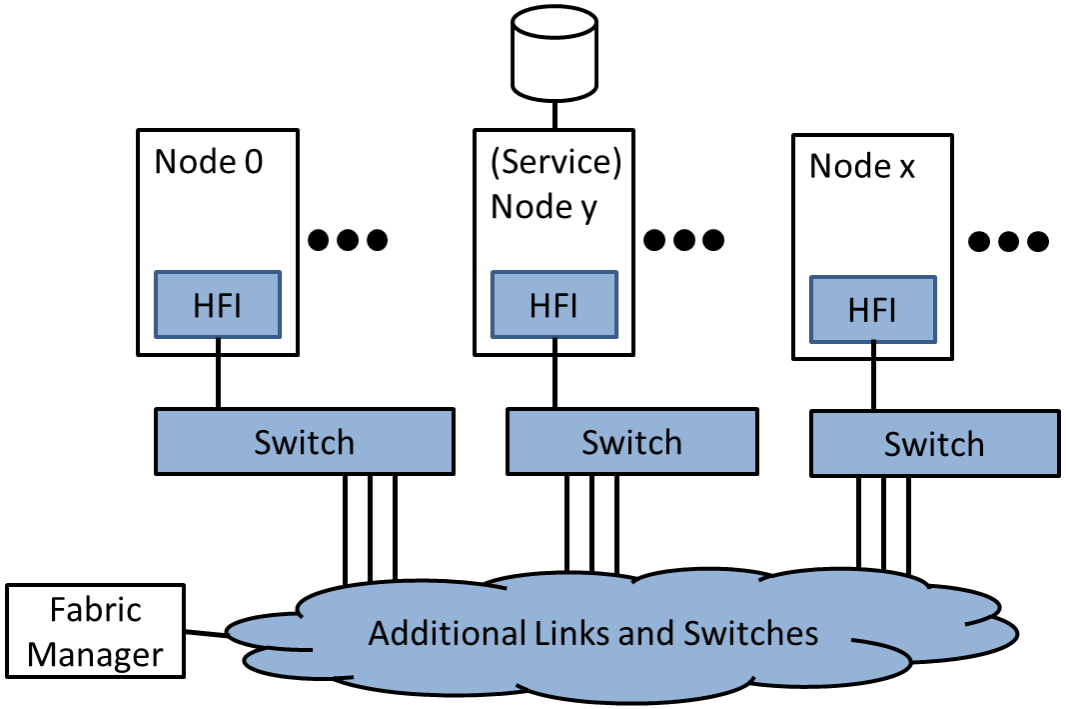
OMNI-PATH LINK LAYER INNOVATIONS

- 8K and 10K MTUs
- Up to 31 VLs and 32 SLs
- LIDs extended to 24 bits
- Packet Integrity Protection - link level retry
- Dynamic Lane Scaling - link width downgrade
- Traffic Flow Optimization – packet preemption
- SCs – topology deadlock avoidance



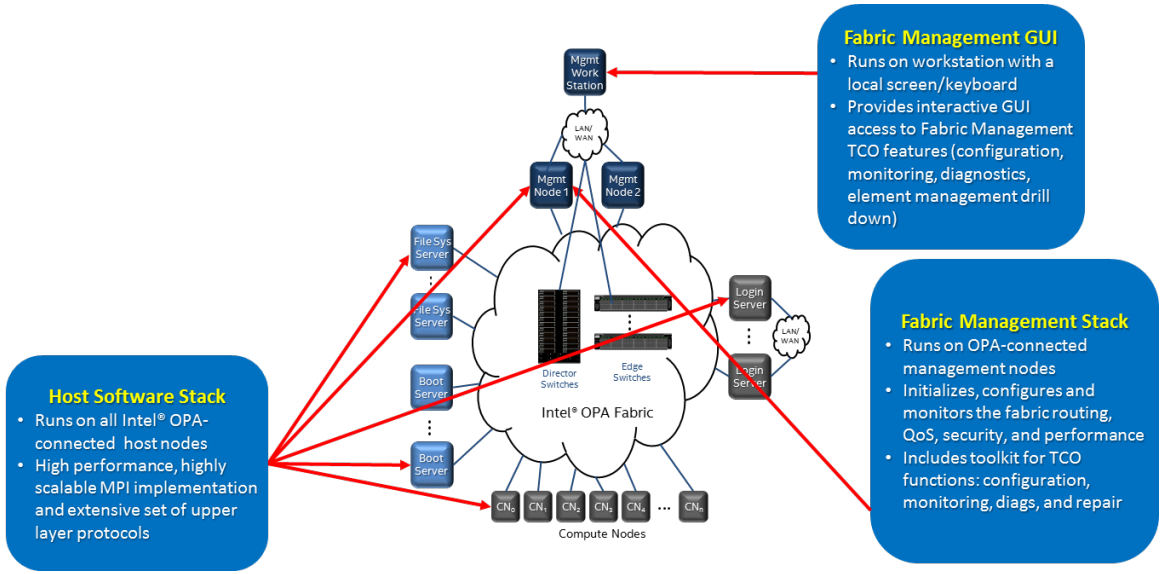
FABRIC MANAGEMENT
Scalability for Exascale

- **Optimized management protocols**
  - 2K MAD packets
    - Allowing more efficient SMA, PMA, SA, PA protocols
    - New attributes, attribute layouts, and fields
  - P-Key based security for SMA packets
  - New and extended PMA counters
  - Architected PA protocol
- **Optimized SA clients**
  - ibacm plug-ins, ib_sa use of ibacm
- **Many new HW features, diagnostics, controls**
  - Drives new SMA and PMA attributes, fields, attribute layouts
- **IB compatibility for key name services SA queries**
  - PathRecord
  - MulticastMemberRecord
  - ServiceRecord
  - NodeRecord
  - Notices



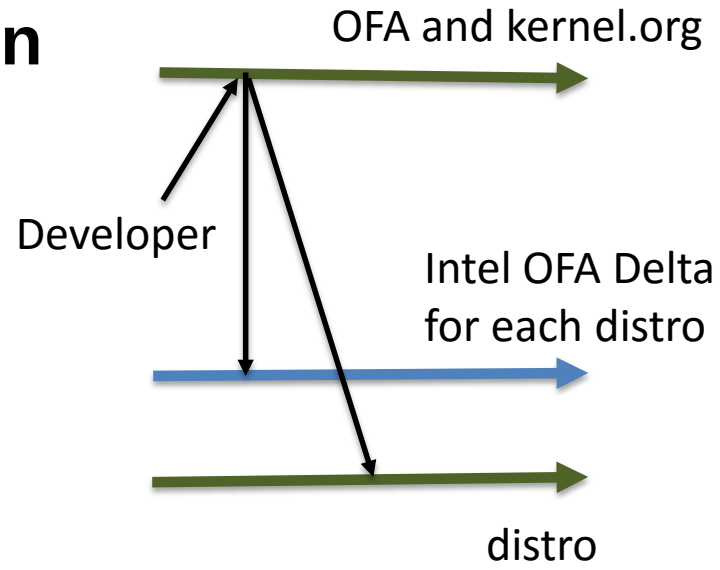
OMNI-PATH OPEN SOURCE STRATEGY

- **ALL OPA specific host software is open sourced and upstreamed**
  - Drivers
  - Providers/libraries
  - Management software, including FM GUI
  - OFA enhancements
    - 2K MADs, extended LIDs, etc
  - Included in RHEL 7.3 and SLES 12 sp2



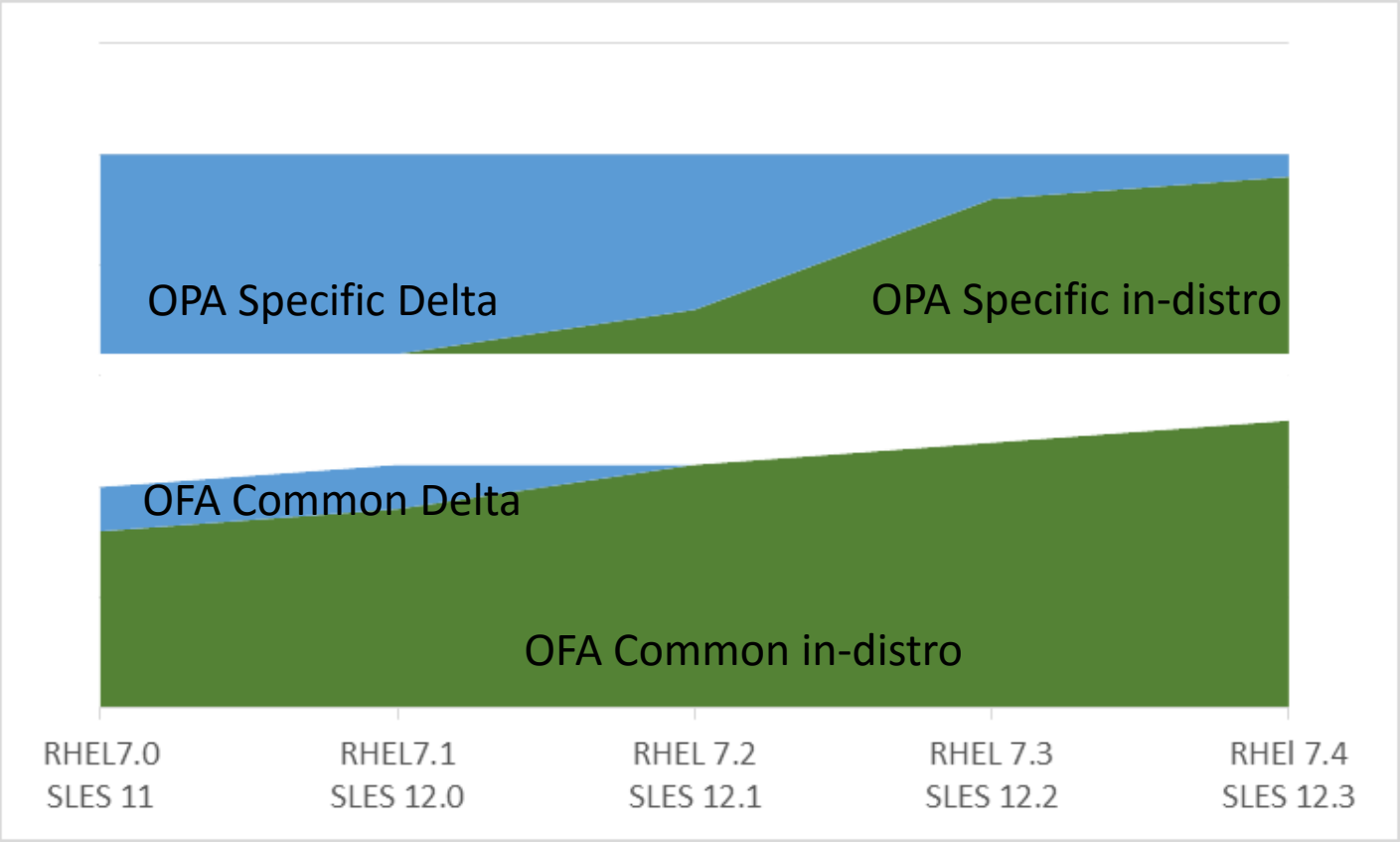
OMNI-PATH OPEN SOURCE STRATEGY

- **OFA enhancements via community collaboration**
  - Ensuring ongoing app interop and multi-vendor OFA
- **Older distros supported via "delta package"**
  - Intel shipped additions to standard distro
  - Minimal footprint of changes
  - Surgically add OPA enablement
    - Without changing fundamental OFA rev in distro



- **Delta package used for new OPA features and optimizations**
  - Rapid delivery of OPA optimizations, fixes, special features
  - src.rpms and open source on github/01org
- **Create proper API abstractions to allow vendor innovation**
  - OFI/Libfabric
  - Allow vendor innovation in drivers/libraries without OFA API churn

INTEL OFA DELTA



- **Minimize distro changes**
- **Avoid replacing common OFA code in distro**
  - Changes driven through standard open processes
  - OPA support fully in current distros
- **OPA specific value adds**
  - New features not yet in distro
  - Items where Intel is Maintainer
  - HFI Driver, OPA libraries/providers, OPA Mgmt
  - All open sourced and upsteamed

CHALLENGES

- **OFA core and verbs are very InfiniBand oriented**
  - Even forced design decisions in OPA HW architecture
  - No forward compatible way for apps to identify OPA cards/features
  - Dependency on IB specific enums, such as speed, rate, MTU, address ranges
- **Support for extended LIDs not ideal**
  - Didn't want to break existing APIs/ABIs
  - Forced to overload fields in ibv_global_route to carry extended LID fields
- **Difficult to get extended MTU into some components**
  - Don't want to break existing APIs nor ABIs
  - IPoIB UD mode on OPA limited to 4K MTU
  - Apps that use SA PathRecord or RDMA CM can be given 8K MTU
    - Works if pass through as given, issues if try to display or interpret
  - Assorted verbs benchmarks limited to 4K MTU
- **Mgmt Interfaces assume IB MADs**
  - example: SRP needed IB PortInfo records from SA to find targets with DMA
  - example: SMPs can't be sent from a UD QP

ONGOING WORK

- **Continued OFA and fabric scalability work**
- **Raw packet snoop/injection interface to driver**
  - Enable wireshark, fabric HW testing, fabric simulation, HW emulation
- **Improved trace and debug features**
  - Historical traces for “after incident” analysis without log clutter
- **SA RMPP stack scalability**
  - Timeouts too large, need way to timeout a large query quickly when no response
  - Get ALL SA queries through a common code path (ibacm and plugins)
- **Introduction of vNIC driver**
  - Ethernet over fabric, optimized for Omni-Path
  - New Ethernet Management Agent (EMA) in hosts
  - Allow use of vNIC IP addresses as fabric address in RDMACM, ibacm

SUMMARY

- **Intel® Omni-Path is a new fabric**
  - In production and installed at many sites, including top500 sites
- **Omni-Path is not InfiniBand, many HW & SW innovations**
  - Extended LIDs, 2K MADs, TFO, etc
- **Omni-Path is open sourced and integrated into OFA**
  - Some challenges, but now upstream and in standard distros
  - Delta distribution model to support older distros
- **Omni-Path work is ongoing**
  - Community collaboration

13th ANNUAL WORKSHOP 2017

THANK YOU

Todd Rimmer, Omni-Path Lead Software Architect

Intel Corporation