INTEL® OMNI-PATH ARCHITECTURE TECHNOLOGY OVERVIEW

Mark S. Birrittella, Mark Debbage, Ram Huggahalli, James Kunz, Tom Lovett, Todd Rimmer, Keith D. Underwood, Robert C. Zak

Presented by Todd Rimmer – Intel® Omni-Path Architect

August 2015

THE NEW CENTER OF POSSIBILITY

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.

Copyright © 2015, Intel Corporation

OPTIMIZATION NOTICE

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

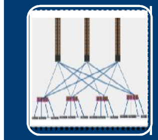
INTEL® OMNI-PATH ARCHITECTURE: FUNDAMENTAL GOALS[1]:

CPU/Fabric Integration

- Improved cost, power, and density
- Increased node bandwidth
- Reduced communication latency

Optimized Host Implementation

- High MPI message rate
- Low latency scalable architecture
- Complementary storage traffic support

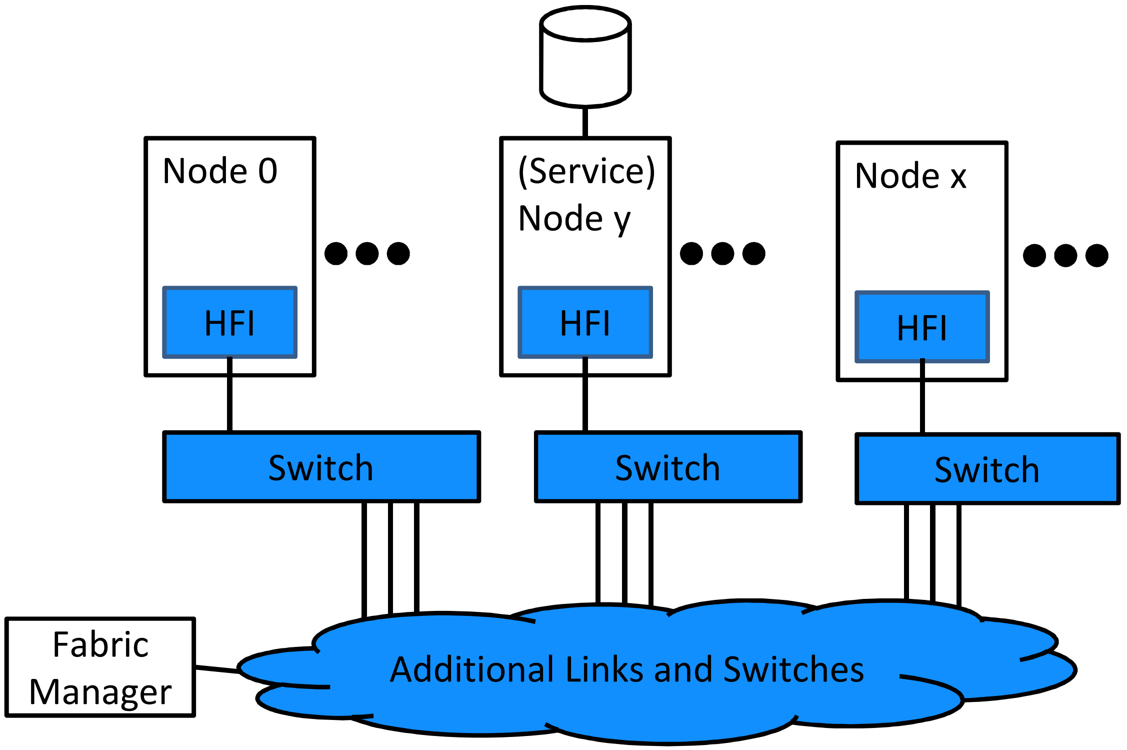


Enhanced Fabric Architecture

- Very low end-to-end latency
- Efficient transient error detection & correction
- Improved quality-of-service delivery
- Support extreme scalability, millions of nodes

[1] Performance goals are relative to Intel® True Scale 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 purchases.

ARCHITECTURE OVERVIEW



Omni-Path Components:

HFI – Host Fabric Interface

- Provide fabric connectivity for compute, service and management nodes

Switches

- Permit creation of various topologies to connect a scalable number of endpoints

Fabric Manager

- Provides centralized provisioning and monitoring of fabric resources

OMNI-PATH NETWORK LAYERS

Layer 1 – Physical Layer

- Leverages existing Ethernet and InfiniBand PHY standards

Layer 1.5 – Link Transfer Protocol

- Provides reliable delivery of Layer 2 packets, flow control and link control across a single link

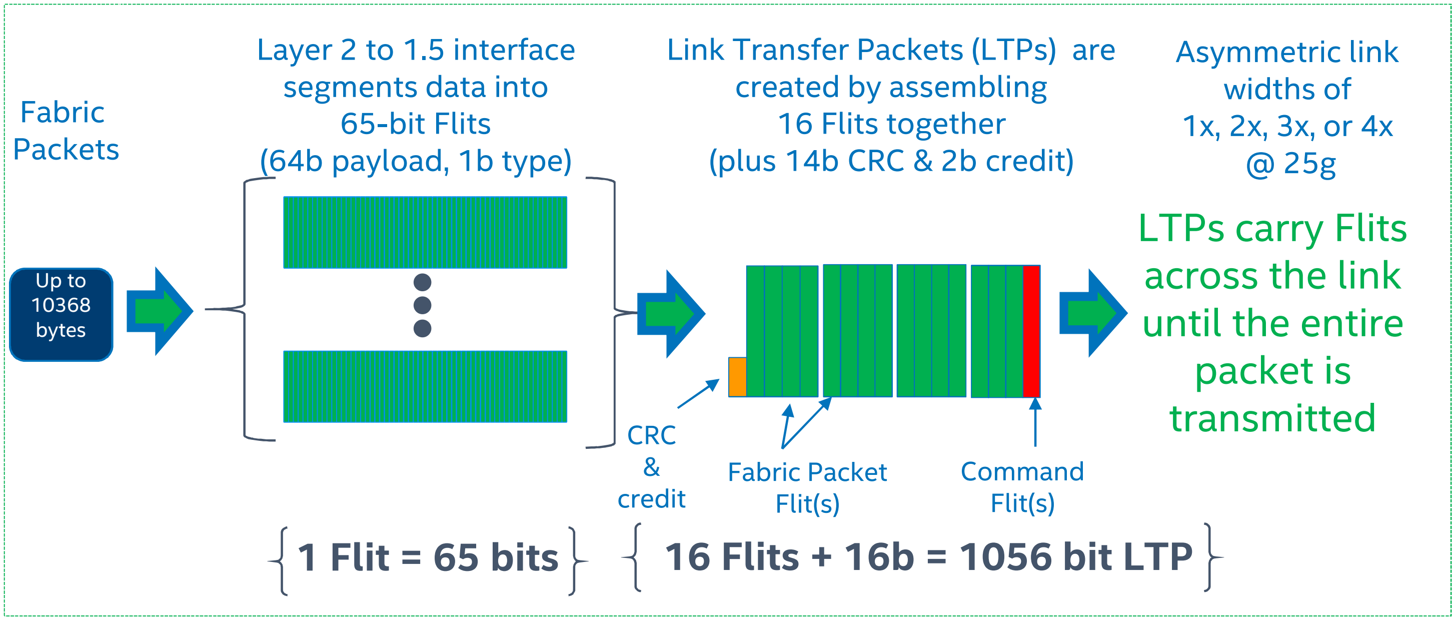
Layer 2 – Data Link Layer

- Provides fabric addressing, switching, resource allocation and partitioning support

Layers 4-7 – Transport to Application Layers

- Provide interfaces between software libraries and HFIs
- Leverages Open Fabrics as the fundamental software infrastructure

LAYER 1.5: LINK TRANSFER LAYER



Fabric Packets

Up to 10368 bytes

Layer 2 to 1.5 interface segments data into 65-bit Flits (64b payload, 1b type)

Link Transfer Packets (LTPs) are created by assembling 16 Flits together (plus 14b CRC & 2b credit)

Asymmetric link widths of 1x, 2x, 3x, or 4x @ 25g

LTPs carry Flits across the link until the entire packet is transmitted

{1 Flit = 65 bits} {16 Flits + 16b = 1056 bit LTP}

Fabric Packet Flits and Command Flits may be mixed in an LTP

Command Flits can carry flow control credits or other link control commands

LTP=128B data, 4B overhead -> 64/66

Link error detection and replay occurs in units of LTPs

LTPs implicitly acknowledged (no overhead)

Retransmission requests via Null LTPs which carry replay Command Flits

CRC: Cyclic Redundancy Check

CAPABILITIES ENABLED BY LAYER 1.5 ARCHITECTURE

	Description	Benefits
Traffic Flow Optimization	▪ Flits from different packets on different VLs can be interleaved ▪ Optimizes Quality of Service (QoS) in mixed traffic environments, such as storage & 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 recovery of transmission errors on an Intel® OPA link with low latency for both corrupted and uncorrupted packets ▪ Resends 1056-bit LTPs rather than entire packet	▪ 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. ▪ Enables service at appropriate time.

[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. Link: www.ieee802.org/3/bj/public/jan13/ran_3bj_01a_0113.pdf

VIRTUAL LANES & CREDIT MANAGEMENT

Up to 31 data VLs and 1 management VL

- Receiver implements a single buffer pool for all VLs

Transmitter manages receiver buffer space usage

- Dedicated space for each VL
- Shared space shared by all VLs
- FM can dynamically reconfigure buffer allocation

Credit Return

- 2 bits per LTP, 4 sequential LTPs yield 8b credit return message
- Explicit command flit may return credits for 16 VLs in 1 flit

Credit Return is reliable via LTP Packet Integrity Protection mechanisms

LAYER 2: LINK LAYER

Supports 24 bit fabric addresses

Allows up to 10KB of L4 payload, 10368 byte maximum packet

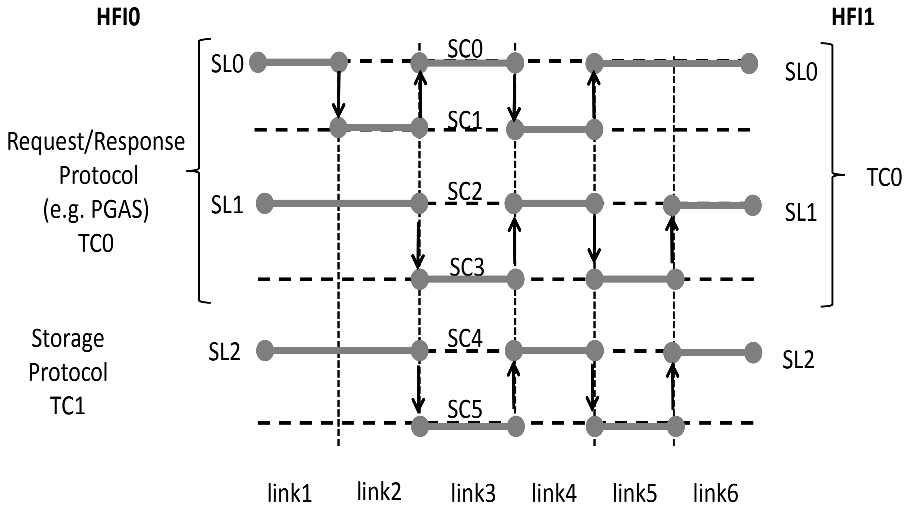
QoS features built on top of VLs and Service Channels (SCs)

Congestion Management

- Adaptive Routing
- Dispersive Routing
- Explicit Congestion Notification

Isolation via Partitioning

QUALITY OF SERVICE



FM allocates and configures TCs, SLs, SCs and VLs based on sysadmin vFabric input

QoS Architectural Elements:

vFabric – Virtual Fabric

Syadmin view. Intersection of a set of fabric ports and one or more application protocols along with a specified set of QoS and security policies.

Traffic Class (TC)

A group of SLs for use by a transport layer or application. Multiple SLs may be used for separation of control vs bulk data or L4 protocol deadlock avoidance

Service Level (SL)

End to end identification of a QoS level. Lowest level concept exposed to L4 and applications. SL2SC and SC2SL mappings occur in endpoints

Service Channel (SC)

Only QoS field in packets. Differentiate packets as they pass through the fabric. A Service Level may use multiple SCs to avoid topology deadlock via hop by hop changes in SC

Virtual Lane (VL)

Per link credit management and separation. SC2VL tables at each port control mapping. VL Arbitration and packet preemption tables control flit scheduling

CONGESTION MANAGEMENT

Distributed Switch Based Adaptive Routing

- Every Switch ASIC analyzes congestion and adjusts routes
- Works well for applications with bursty or consistent traffic patterns
- Mechanism reduces impacts to transports by limiting frequency of adjustment

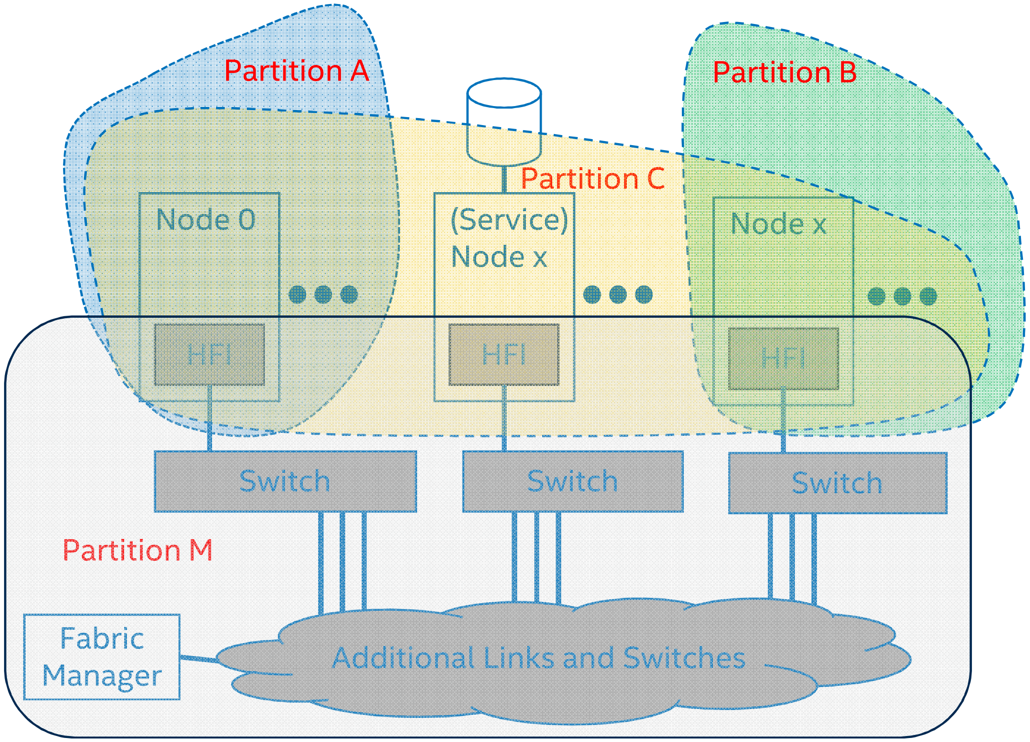
Dispersive Routing

- Probabilistic distribution of traffic
  - Across multiple routes and/or multiple virtual lanes
  - PSM acquires multiple routes and sprays traffic across those routes
- Leverages multi-pathing within fabric

Explicit Congestion Notification Protocol

- Reduces impact of hot spots due to oversubscribed endpoints
- Packet marking by switches as congestion trees form
- Destination HFI returns a backward notification to HFI source
- Source HFI reduces bandwidth of packets to that destination

PARTITIONING



Service Nodes are Full Members of C

Other nodes are Limited members of C

Every fabric packet is associated with one partition

Isolates a group of endpoints for all types of traffic

Individual endpoint can be Full or Limited member of a given partition

- Full may talk to any member of partition
- Limited may only talk to full members of partition
- Allows shared services

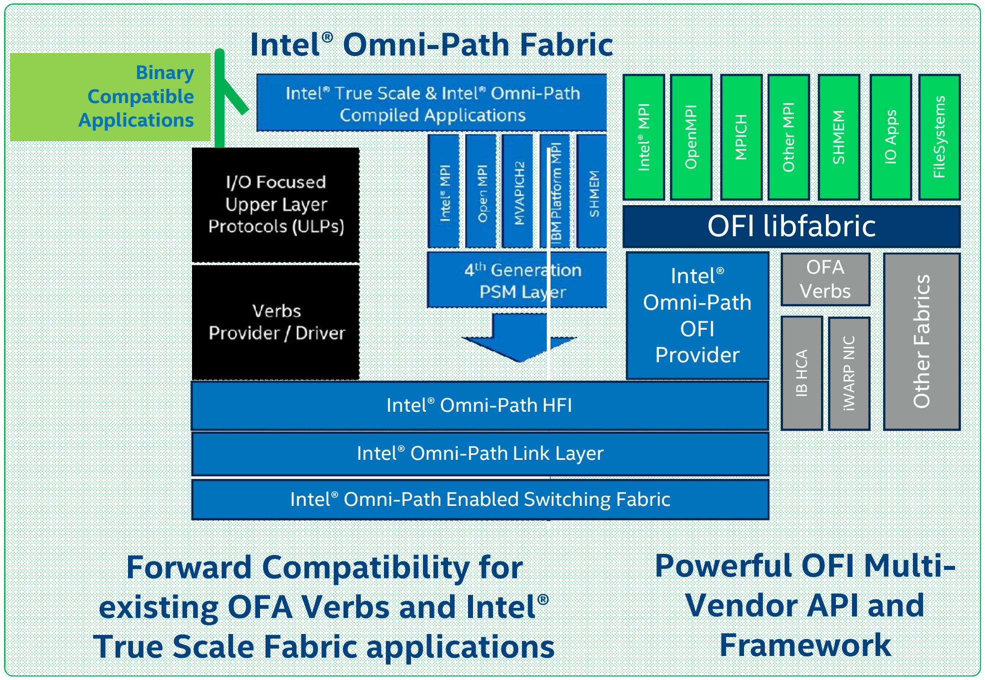
A management partition is defined

- All endpoints are members
- Only management nodes are full members of this partition

Partitions enforced by switch at HFI-SW link

FM Creates and configures all partitions

LAYER 4: TRANSPORT LAYER AND KEY SOFTWARE



Forward Compatibility for existing OFA Verbs and Intel® True Scale Fabric applications

Powerful OFI Multi-Vendor API and Framework

Performance Scaled Messaging (PSM)

- API and corresponding L4 protocol designed for the needs of HPC

Open Fabrics Interface (OFI) libfabric

- General purpose framework providing an API applications and middleware can use for multiple vendors and L4 protocols

Open Fabrics Alliance Verbs

- API and corresponding L4 protocol designed for RDMA IO

FIRST GENERATION INTEL® OMNI-PATH PRODUCT FAMILY

On track for Q4'15 introduction

Host Fabric Interface (HFI)

HFI
ASIC

"Wolf River" (HFI) Silicon
2 x 100 Gbps, 50 GB/sec Fabric Bandwidth

Switch

Switch
ASIC

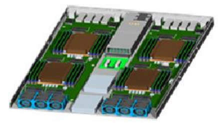
"Prairie River" Switch Silicon
48 ports, 9.6Tb/s, 1200 GB/sec Fabric Bandwidth

Software

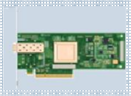
Intel® Fabric Suite
[based on OFA with Intel® Omni-Path Architecture support]

Cables

Passive Copper & Active Optical Cable (AOC)

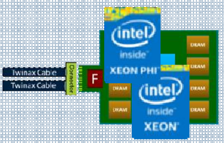


Custom Mezz & PCIe Cards

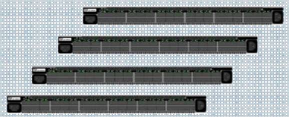


Standard PCIe Board[1]
[code name Chippewa Forest]

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

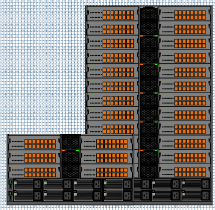


Intel® Xeon® processor and Intel® Xeon Phi™ coprocessor with integrated Host Fabric Interface (HFI)



Intel® Omni-Path Edge Switch[1]
[code name Eldorado Forest]

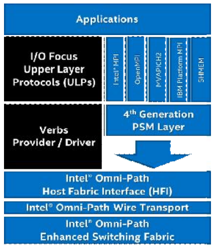
- 24- and 48-port switches
- 1U form factor

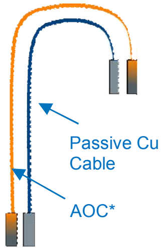


Intel® Omni-Path Director Class Switch[1]
[code name Sawtooth Forest]

- 192- and 768-port switches
- 7U and 20U form factor

Custom Switches


Applications
I/O Focus Upper Layer Protocols (ULPs)
Intel® MPI
OpenMPI
MVAPICH2
IBM Platform MPI
SHMEM
Verbs Provider / Driver
4th Generation PSM Layer
Intel® Omni-Path Host Fabric Interface (HFI)
Intel® Omni-Path Wire Transport
Intel® Omni-Path Enhanced Switching Fabric

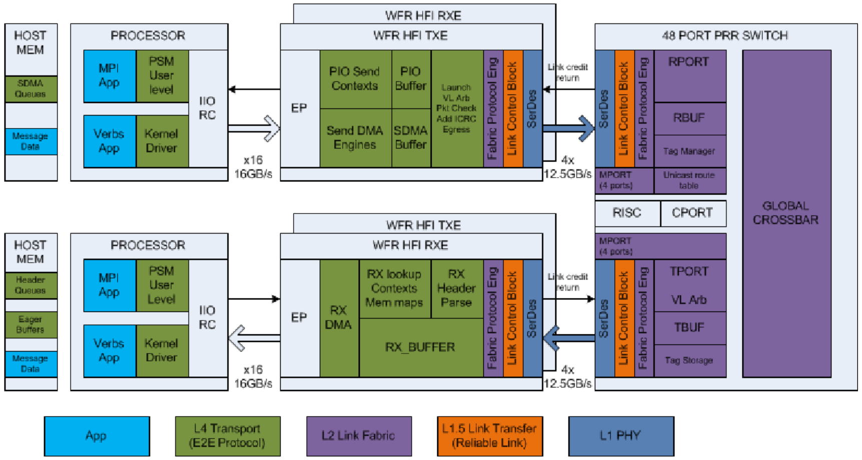


Potential future options, subject to change without notice.
All timeframes, features, products and dates are preliminary forecasts and subject to change without further notice.

[1] Will be available as both a reference design and Intel-branded product

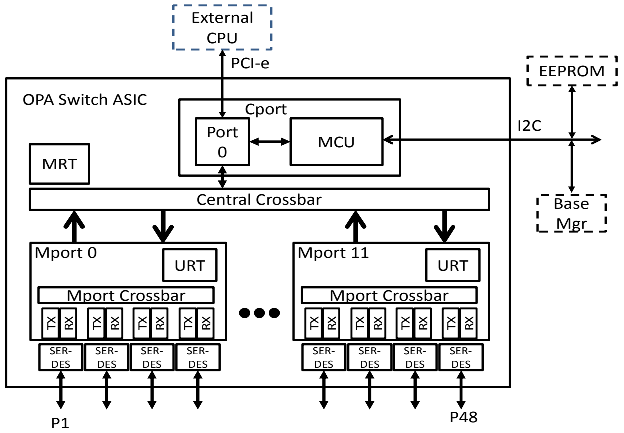
WFR HFI ARCHITECTURE FEATURES

- WFR ASIC has 1 or 2 HFIs with:
  - 100 Gbps fabric, Intel® OPA link layer
  - PCIe v3.0x16 host interface
- Host on-load architecture
- Send side:
  - Packet store and forward
  - 1 or more send contexts per CPU core
  - Multiple SDMA engines
  - Automatic header generation (AHG)
- Receive side:
  - Packet cut-through to reduce latency
  - 1 or more receive contexts per CPU core
  - Receive side mapping (RSM)
  - Eager delivery – host memory FIFO
  - Expected TID – direct data placement, DDP



- Data integrity: highly reliable E2E
  - Internal SECDED ECC data path protection
  - Link-level CRC: 14-bits
  - Packet Integrity Protection
  - End-to-end ICRC: 32 bits
  - KDETH HCRC: 16 bits
  - 16-bit job key and 31-bit PSN

PRR SWITCH ARCHITECTURE FEATURES



- 48 port ASIC
  - 100 Gbps fabric, Intel® OPA link layer
- Over provisioned hierarchical cross bar
  - 12 Mports, each consisting of 4 OPA 100 Gbps ports
  - Switching via Unicast URT and Multicast MRT
- Port Logic:
  - 8 VLs
  - Packet Preemption
  - VL arbitration
- Port 0:
  - Switch management port
  - Permits in-band and PCIe based switch management
  - On-chip micro-controller (MCU)
  - PCIe interface for optional external CPU
  - I2C interface
    - MCU firmware/config eeprom access
    - baseboard management

- Data integrity:
  - Internal ECC and parity data path protection
  - Link-level CRC: 14-bits
  - Packet Integrity Protection

FIRST GENERATION PRELIMINARY PERFORMANCE RESULTS:

	Intel True Scale	Intel OPA
SERDES Rate (Gbps)	10	25.78
Peak Port Bandwidth (Gbps)	32	100
HFI Message Rate (Million Messages per second)	35[3]	160[1]
Switch Ports	36	48
Switch Packet Rate (Million Packets per Second)	42[3]	195[1]
Switch Latency (ns)	165-175[3]	100-110[2]

[1] Based on Intel projections for Wolf River and Prairie River maximum messaging rates.

[2] Latency based on 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.

[3] 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 purchases.

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. Copyright © 2015, Intel Corporation.

SUMMARY

Intel® Omni-Path Architecture introduces a multi-generational fabric

- Designed to scale to needs to high end HPC
- And meet the needs of commercial data centers

Advanced Link Layer Features

- Reliability & pervasive EEC to meet large scale system reliability needs
- Packet preemption enables BW fairness and low latency jitter

Existing Software Ecosystem preserved

- New OFA OFI API designed for semantic match & allows HW innovation

First Gen HW available to partners now, targeting 4Q15 introduction

- 100Gbps links, 160M msg/sec[1], Switch latency < 110ns[2]

[1] Based on Intel projections for Wolf River and Prairie River maximum messaging rates.

[2] Latency based on 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.

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. Copyright © 2015, Intel Corporation.

THANK YOU