Intel® Omni-Path 4.8 Tbps Switch ASIC and Platform

HotChips 2016

Agenda

- Omni-Path goals and architectural overview
- Switch features and high level block diagram
- Packet Forwarding
- Switch internal bandwidth over-provisioning
- Traffic Flow Optimization (Preemption)
- Congestion
- Innovative dense platform packaging
- Performance

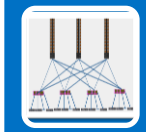
Intel® Omni-Path Architecture: Fundamental GOALS[1]:

CPU/Fabric Integration

- Improved per port 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

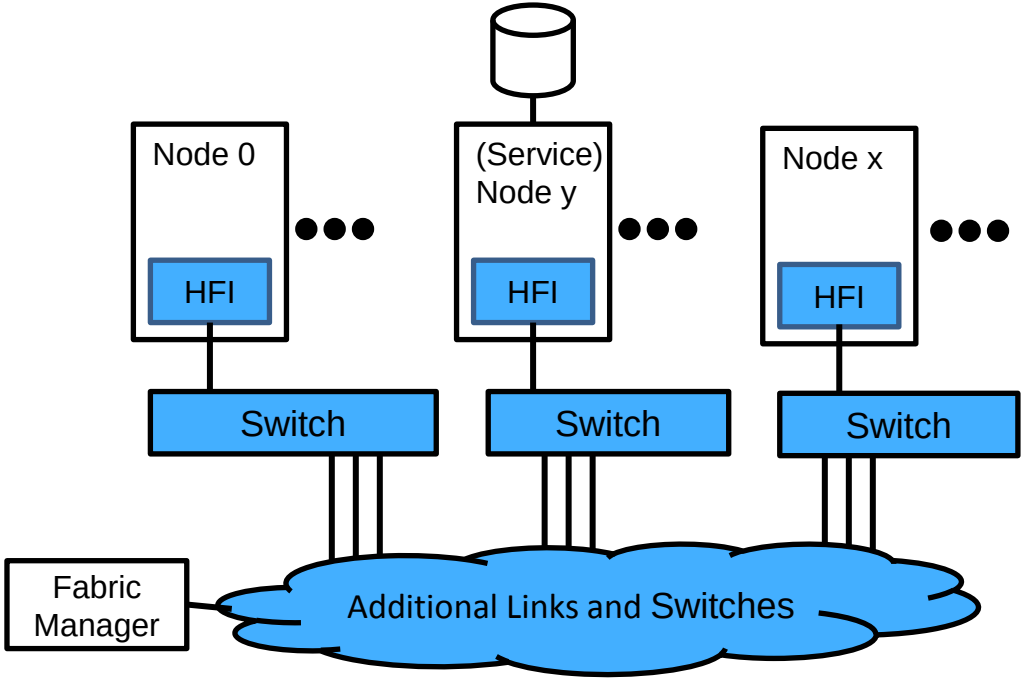


Enhanced Fabric Architecture

- 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

Switch Architecture Features

- Switch speeds and feeds:
  - 48 ports
  - Integrated 25Gb/s long-reach SerDes links (192 SerDes)
  - 100 Gbps per port, Intel® OPA link layer
  - Switching via Unicast URT and Multicast MRT
- Ports 48:1
  - 8 Data VLs and 1 Management VL
  - Packet size ranging from 16B to 10KB + 128 bytes
  - Packet preemption for Traffic Flow Optimization (TFO)
- Port 0
  - Switch management port
  - On chip micro-controller (MCU)
  - PCIe interface for optional CPU, supports in-band PCIe based switch management
  - I2C interfaces
    - MCU firmware/configuration, field upgradable firmware
    - Baseboard management
- Data integrity
  - Internal SECDED ECC data path protection
  - Link-level CRC for Packet Integrity Protection (PIP)
  - Dynamic Lane Scaling (DLS)

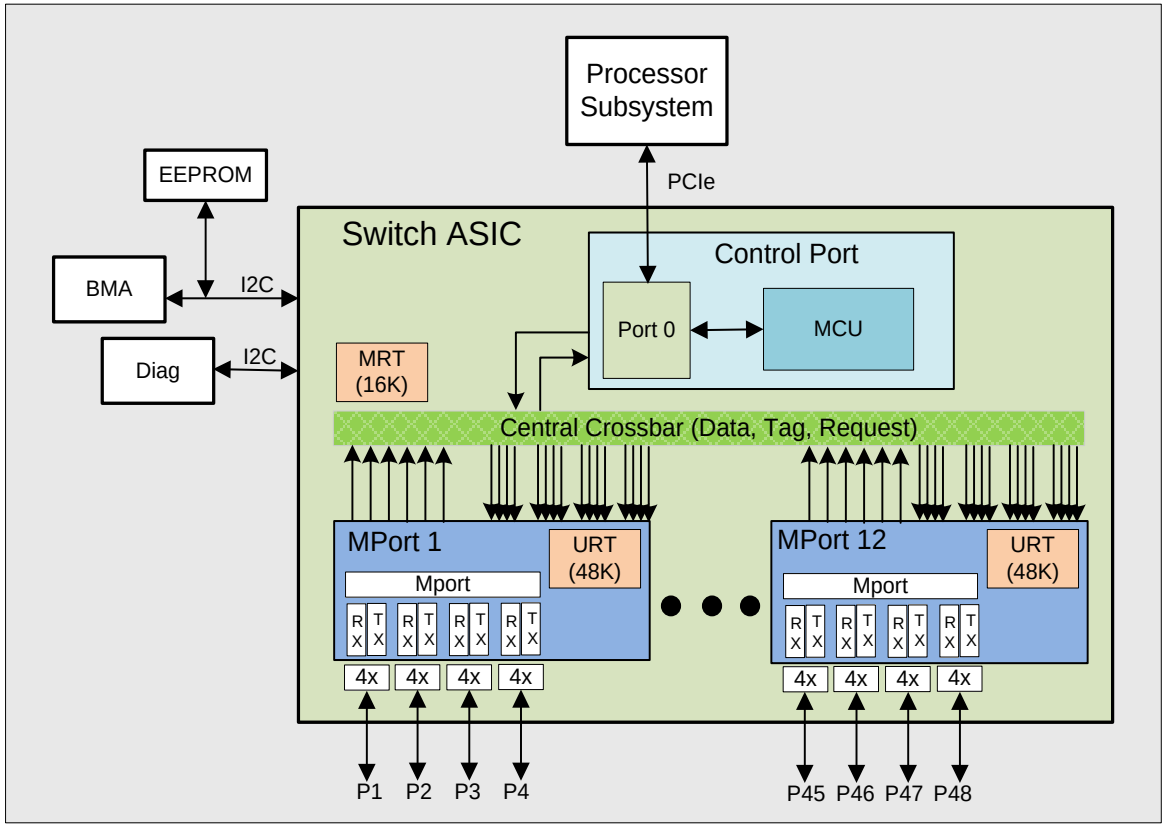
Switch ASIC

Atributes

- 12 Mports
- Each Mport contains 4 ports
- Each Mport has a copy of the URT
- One shared MRT
- Control port supporting port 0 and switch management firmware
- Central crossbar connects all Mports and Port 0 together.

Key Points

- Ports are grouped together in Mports to share logic and reduce the number of connections to the central crossbar.
- Full copy URT in each Mport guarantee's the predictable and low access times for destination lookup's when scaling fabric.

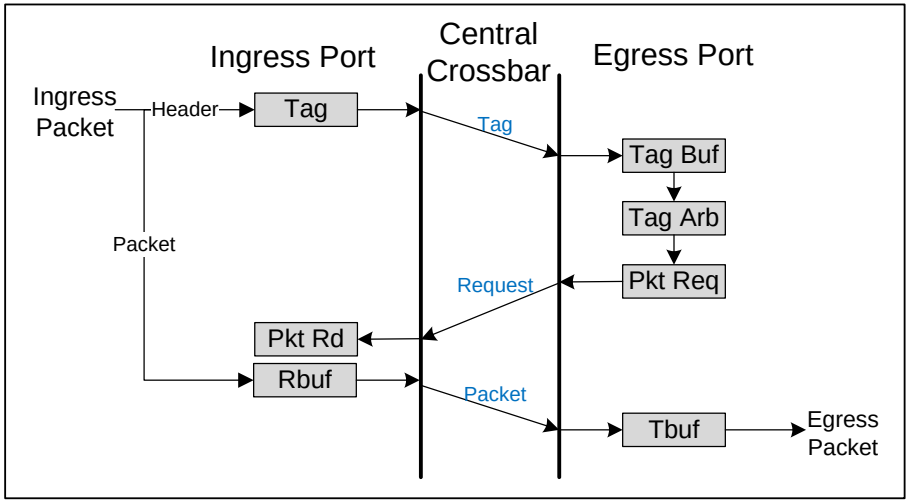


Packet Forwarding

Description

- Switch supports a packet “pull” architecture.
- Ingress packets are written into a receive buffer, header data is processed by a tag controller.
- Tag controller performs URT or MRT destination port lookup and forms a packet tag.
- Packet tag is distributed to one egress port if unicast or multiple egress ports if multicast.
- Egress ports organize packet tags from all ingress ports into VL work-load fifo’s (Tag Buffer).
- Egress ports arbitrates and select the next packet to transmit.
- Egress ports issue packet requests back to the ingress ports.
- Ingress ports read packet from receive buffer and send it to the requesting egress port (s).
- If multicast, multiple egress ports can be requesting the packet at the same time, a single packet is sent to all requesting egress ports.

Packet Flow Diagram

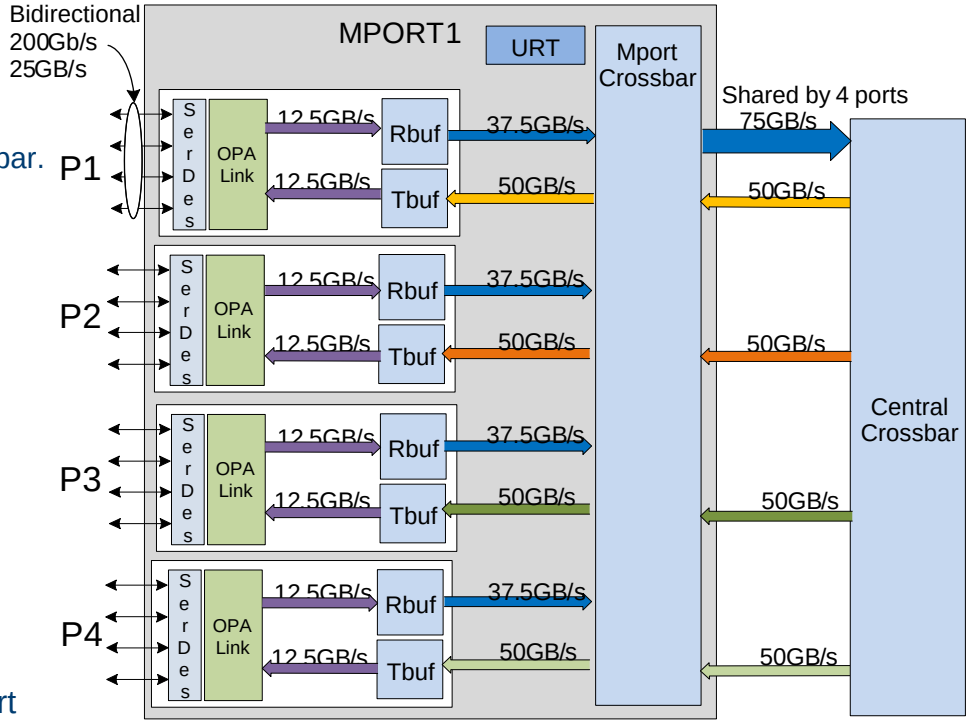


Key Points

- Egress ports are the master, view of all packets waiting to be transmitted, packets selected are committed to the wire.
- Bandwidth, QoS and congestion are managed and tracked by the egress port.

Switch Mport

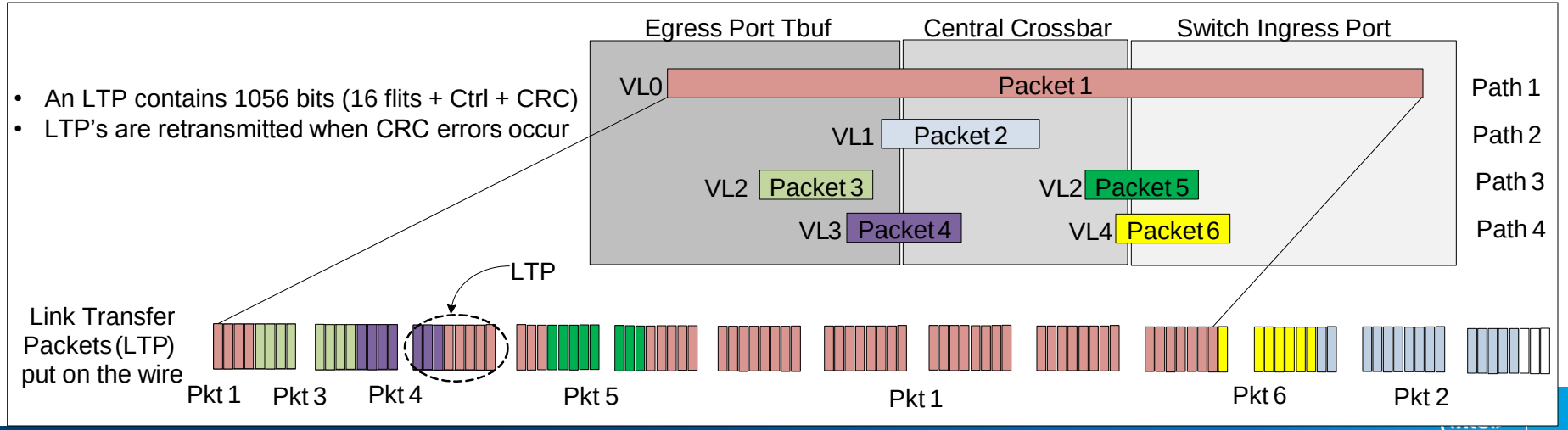
- **Mport attributes**
  - 12 Mports, 4 ports per Mport.
  - Ports within Mport are connected via Mport crossbar
  - Mports are connected together via a central crossbar
  - Ports source up to 37.5GB/s of packet data to the Mport crossbar.
  - Mport crossbar provides 75GB/s of packet data to the central crossbar (supports the four ports).
  - Ports sink up to 50GB/s packet data from the central crossbar.
- **Innovative internal bandwidth over-provisioning**
  - 3x over-provisioned at Rbuf exit
  - 1.5x over-provisioned at Mport exit
  - 4x over-provisioned at Tbuf entry
  - 2.75x over-provisioned at central crossbar (3.3TB/s of bi-directional packet data bandwidth)
  - Over-provisioning reduces latency variation, improves port bandwidth, improves packet performance and recovery time when egress port congestion is encountered and released.



Egress Port Packet Arbitration

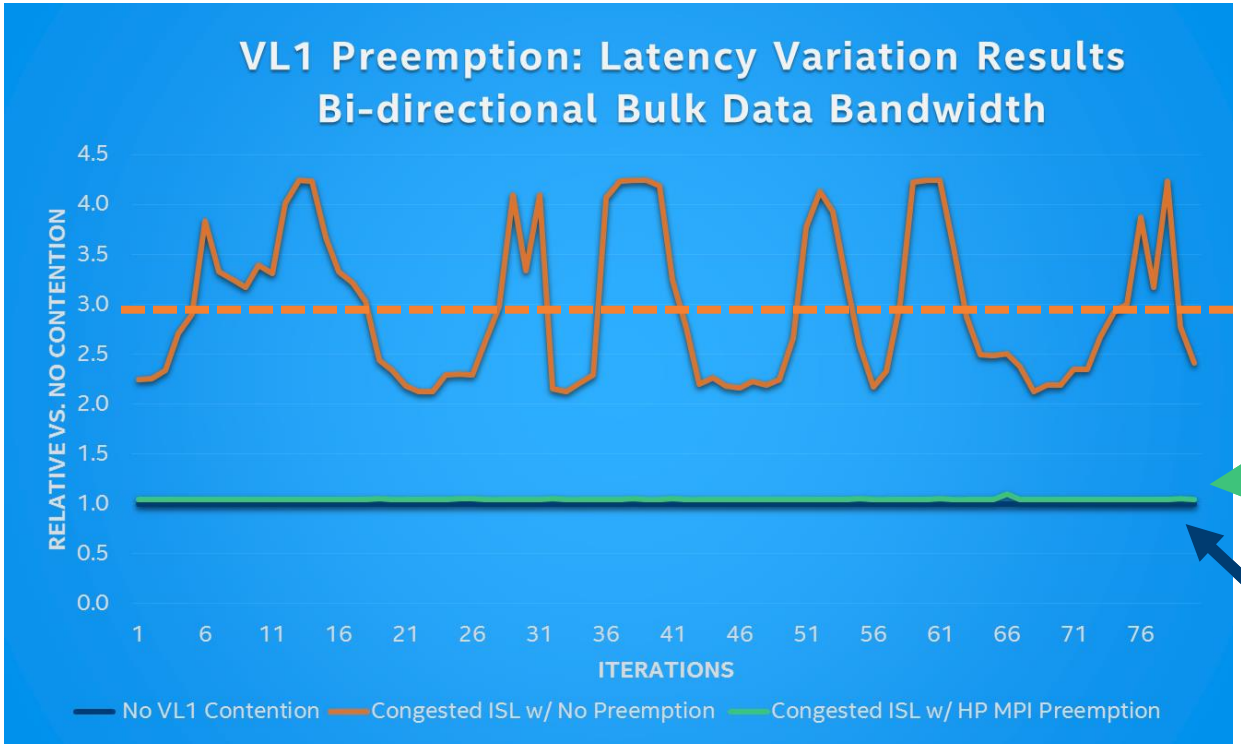
- Egress arbitration selects the next packet segment to put on the wire. Arbitration keeps track packets inflight (up to 8 based on packet size) and organizes packet data segments to send.
- Data segments are flits (8 bytes), with flit based flow control.
- The following diagram is an example (5 packets inflight).
  - (low) Large messages, storage traffic, bursty, less latency sensitive (VL0, VL1)
  - (high) Medium messages, somewhat latency sensitive (VL4)
  - (pre) Small messages, MPI compute traffic, bursty, very latency sensitive (VL2, VL3 – enabled for preempt VL0)
- Packet preemption enables BW fairness and deterministic latency

Packet preemption for Traffic Flow Optimization (TFO)



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.

Congestion Management

Congestion Detection

- Egress ports contain VL work-load fifo's, these fifo's represent switch wide ingress packets waiting to be transmitted
  - VL work-load packet counters
  - VL work-load packet flit counters
  - VL queue wait counters
  - Congested DLID counters
  - Wire utilization counters
- This information along with remote credit can be used to detect if the egress port is a victim or root of congestion

Explicit Congestion Notification Protocol (FECN, BECN)

- 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

Medium Grain Adaptive Routing

- Medium grain adaptive routing (MGAR) function can alter DLID paths every 100's ms to several seconds based on link state and congestion without software intervention.
- Every Switch ASIC analyzes congestion and adjusts DLID routes.
- Hardware mechanisms partner with firmware to alter the path of the DLID in response to change conditions such as congestion and link failures. These alternate paths are pre-configured by software allowing each switch to perform MGAR functions independently.
- By allowing selected packets to circumvent congested ports, AR will allow better use of STL network resources and increase network performance.

Omni-Path Gen 1 Platform

Challenges

- How to implement a 1U top of rack switch with 48 ports on faceplate
- Implementing 25Gb/s channel to enable 3 meter DAC interconnect and …
  - Without using embedded re-drivers (added power)
  - Without requiring FEC (latency penalty)
  - Achieving corrected BER < 3e-29 (using OPA link level retry, PIP)

- Implementing director class switch
  - Without using embedded re-drivers (added power)
  - Without requiring FEC (latency penalty)
  - Achieving corrected BER < 3e-29 (using OPA link level retry, PIP)

Intel® Omni-Path Edge Switch 100 Series

- 48 x 100Gb/s port Switch,1U
- Up to 9.6 Tb/s aggregate BW
- 100-110ns Switch latency

- **Innovative dense packaging**
  - 3 row QSFP28 achieves 48 x 100Gb/s ports in 1U
- **Innovative recessed I/O faceplate**
  - minimizes PCB trace loss, achieves excellent signal integrity and cable length
  - supports >3m x30AWG and >5m x26AWG copper cables
  - no re-timers on any ports
  - no FEC latency penalty for bit error detection
- **Power efficient**
  - no re-timers inside

Intel® Omni-Path Director Class Switch 100 Series

768 x 100Gb/s port Switch

20U

192 x 100Gb/s port Switch

7U

- Scales in 32 x100Gb/s port increments
- Scales up to 153.6 Tb/s aggregate BW
- 300-330ns 2-tier Switch latency

- **Innovative, dense packaging**
  - 2-tier Fat Tree internal topology, 768 x 100Gb/s QSFP28 ports in 20U chassis.
  - Two chassis per Rack – 307.2 Tb/s aggregate BW per Rack
- **Excellent signal integrity**
  - No re-timers on internal or external ports
  - No FEC latency penalty for bit error detection.
  - Supports >3m x30AWG and >5m x26AWG copper cables
- **Power efficient**
  - no optics or re-timers inside

Latency, Bandwidth, and Message Rate

Intel® Xeon® processor E5-2699 v3 & E5-2699 v4
Intel® Omni-Path Architecture (Intel® OPA)

Metric	E5-2699 v3[1]	E5-2699 v4[2]
Latency (one-way, 1 switch, 8B) [ns]	910	910
Bandwidth (1 rank per node, 1 port, uni-dir, 1MB) [GB/s]	12.3	12.3
Bandwidth (1 rank per node, 1 port, bi-dir, 1MB) [GB/s]	24.5	24.5
Message Rate (max ranks per node, uni-dir, 8B) [Mmps]	112.0	141.1
Message Rate (max ranks per node, bi-dir, 8B) [Mmps]	137.8	172.5

Near linear scaling of message rate with added cores on successive Intel® Xeon® processors

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
2. Intel® Xeon® processor E5-2699 v4 2.20 GHz 22 cores


intel®
experience
what's inside™

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