RoCEv2
Update from the IBTA

#OFADevWorkshop
Introduction – RoCE (v1)

- RDMA Application / ULP
- OFA Verbs
- OFA* Stack
- IB Transport Protocol
- Network Layer
- InfiniBand Link Layer
- Ethernet Link Layer
- RoCE
- InfiniBand
- Ethernet Management
- InfiniBand Management

RoCEv2 Update from the IBTA
The RoCE(v1) Packet Format

InfiniBand (L2Hdr) LRH L3 Hdr BTH+ (L4 Hdr) IB Payload ICRC VCRC

RoCE Eth L2 Header L3 Hdr BTH+ (L4 Hdr) IB Payload ICRC FCS

No Changes
RoCE(v1) Timeline

- May 2009 - IBTA forms RoCE WG
- July 2009 – RoCE Prototype Available
- Aug 2009 – Position paper on RoCE at HOTI
- April 2010 - Ratified IBTA Standard
- October 2010 – RoCE in upstream Linux
RoCE(v1) and L2-Datacenters

- RoCE is a L2 Protocol
  - ("flat") L2 Eth Topology
    - L2 Extensions are Supported
      - TRILL
      - Provider Networks (PB, PBB, etc)
      - Virtualization Overlays
    - RDMA within an Ethernet L2 Domain
      - FCoE is similar
The Need for (IP) Routable RoCE

- A Common Class of L3 Datacenter
  - Nodes within a Rack share a Ethernet L2 Domain
  - TOR Device
    - L2 Switch for intra-rack communication
    - L3 (IP) Router for inter-rack communication
  - other topologies also apply

- Customer Demand
  - RDMA Across Racks
    - i.e. Across IP Subnets
  - Focus on Data Center Networks

- RoCE is Already “close” to being “IP Routable”
  - RoCE Address in API is IP
  - RoCE Packet Format / Wire Protocol Includes L3 Header
RoCEv2 – A Straightforward Extension

OFA Verbs

IB Transport Protocol

IB Network Layer

InfiniBand Link Layer

InfiniBand Management

IB Transport Protocol

IB Network Layer

Ethernet Link Layer

Ethernet / IP Management

InfiniBand

RoCE

RoCE v2

UDP

IP

OFA* Stack

RDMA Application / ULP

Software

Typically HW

RoCEv2 Update from the IBTA
RoCEv2 - IP Routable Packet Format

EtherType indicates that packet is RoCE (i.e. next header is IB GRH)

EtherType indicates that packet is IP (i.e. next header is IP)

udp.dport number indicates that next header is IB.BTH

ip.protocol_number indicates that packet is UDP
RoCEv2 Highlights

- **Contained Change**
  - Clean L3 Replacement
  - Strict Layering Preserved
  - Generated and Consumed Below the Channel Interface (i.e. the “API”)

- **Transparent to Applications**
  - No Application Software Changes
    - Current RoCE API is already IP L3 based

- **Transparent to Underlying Network Infrastructure**
  - Mainstream processing on L2 Ethernet Switches / L3 IP Routers

- **Additional Benefits of the RoCEv2 Approach (some examples)**
  - Traditional Network Management Tools Apply
  - ACLs (Metering, Accounting, Firewalling)
  - IGMP Snooping for optimized Multicast
  - Network Monitoring Tools
Status Update

• IBTA SC Requested Technical Work on RoCEv2 (Nov 2013)

• IBTA Invited to give Presentation at November 2013 IETF Meeting
  – IBTA Starting Definition of IP Routable RoCE
  – Well Received by the IETF Community

• IBTA IBXoE Working Group Re-Activated to Specify RoCEv2
  – Started Work in December 2013

• Initial RoCEv2 Specification Draft
  – Undergoing IBXoE WG Review

• Next Steps
Thank You