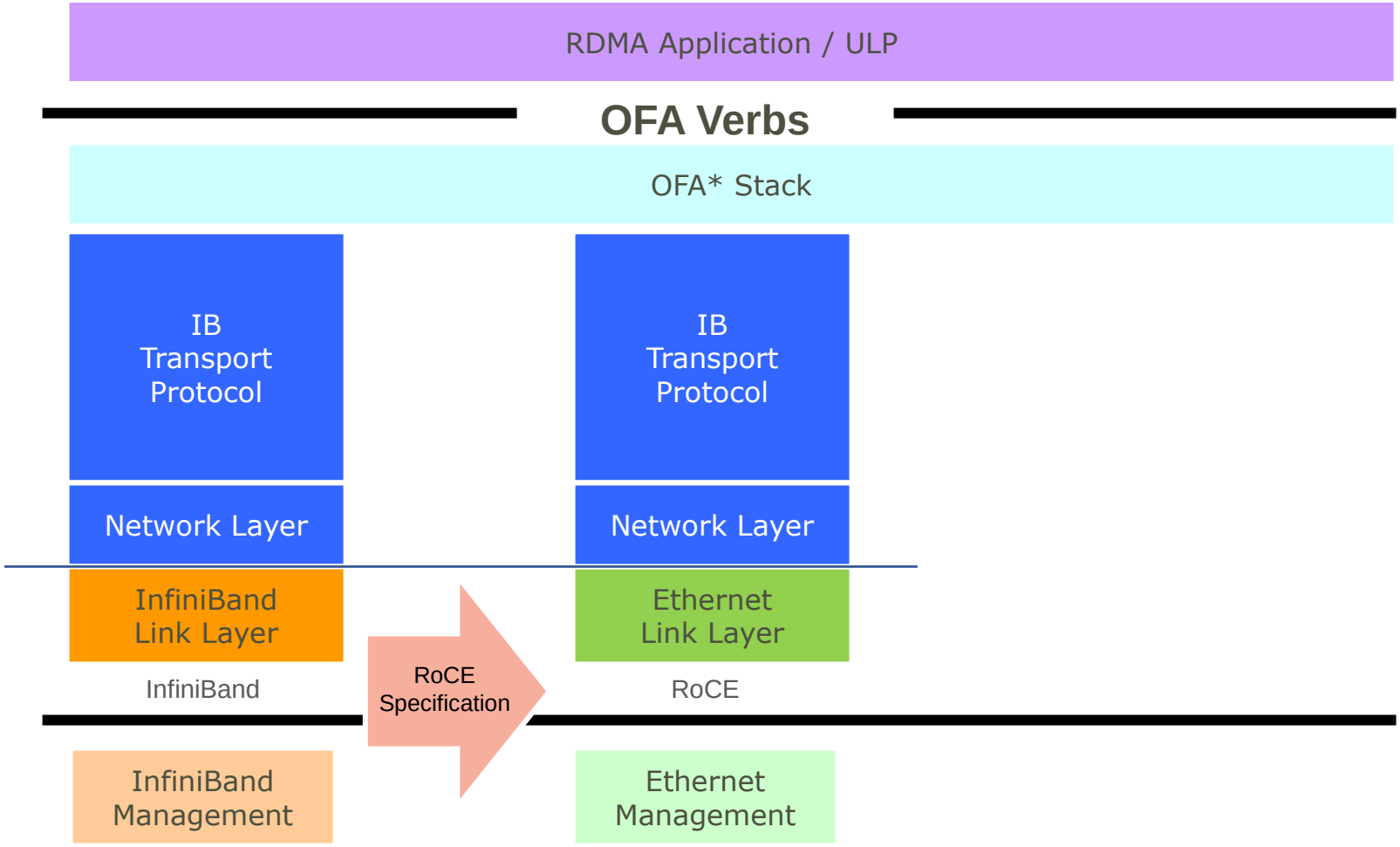
OPENFABRICS ALLIANCE

10TH ANNUAL INTERNATIONAL DEVELOPER WORKSHOP

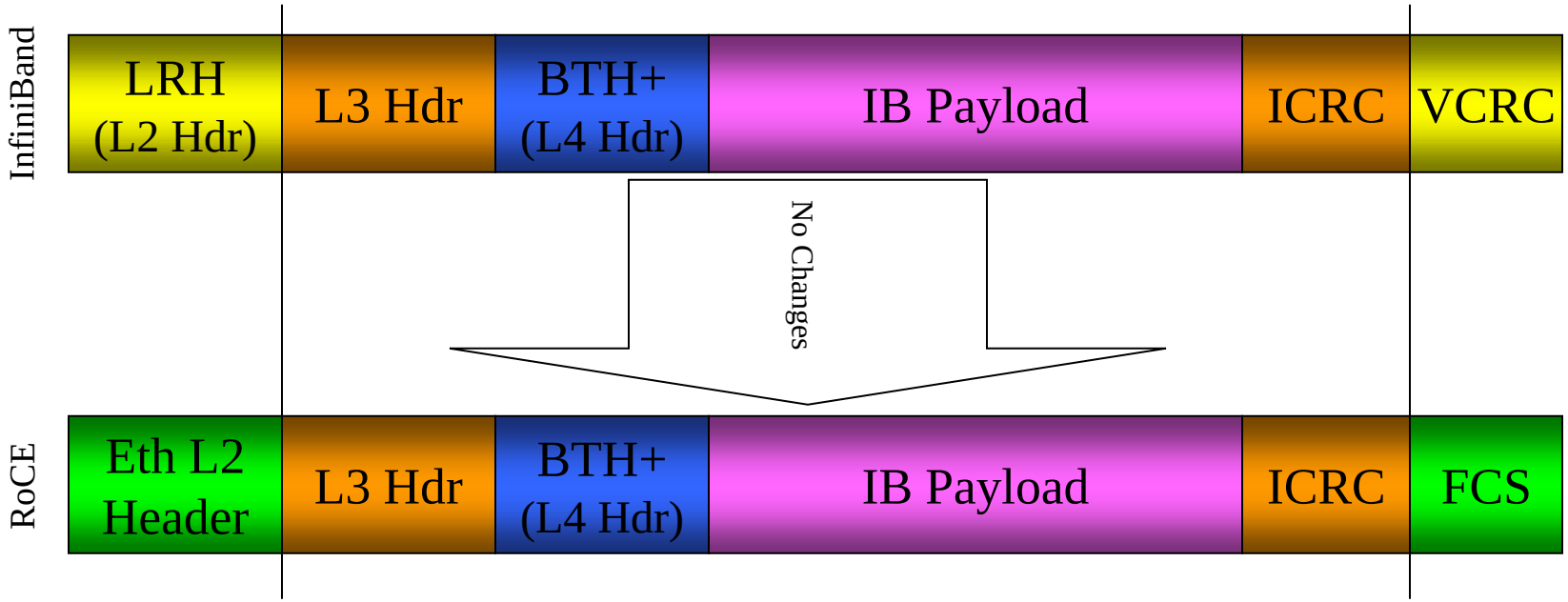
RoCEv2
Update from the IBTA

#OFADevWorkshop

Introduction – RoCE (v1)

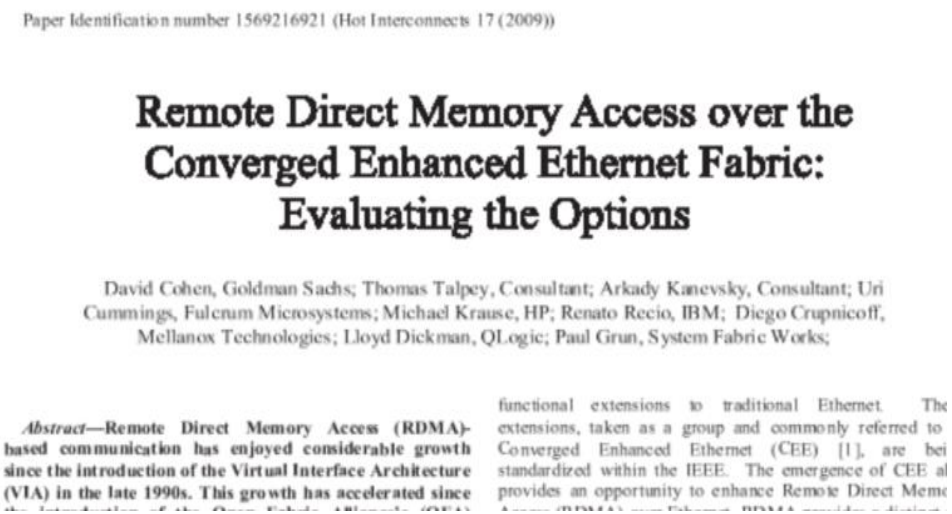


The RoCE(v1) Packet Format



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



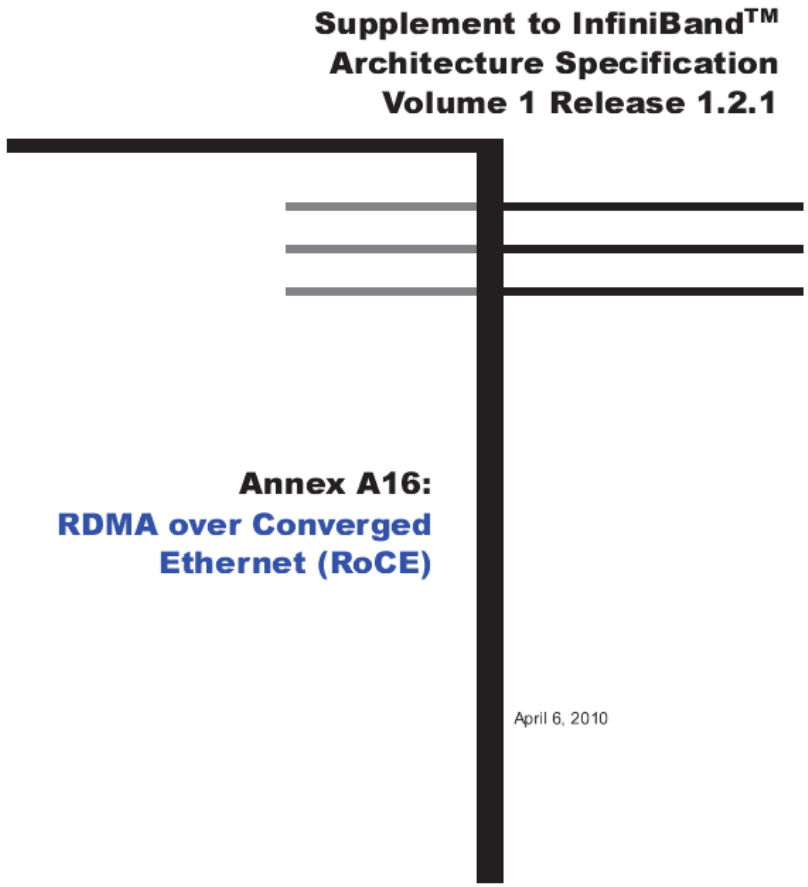
Paper Identification number 1569216921 (Hot Interconnects 17 (2009))

Remote Direct Memory Access over the Converged Enhanced Ethernet Fabric: Evaluating the Options

David Cohen, Goldman Sachs; Thomas Talpey, Consultant; Arkady Kanevsky, Consultant; Uri Cummings, Fulcrum Microsystems; Michael Krause, HP; Renato Recio, IBM; Diego Crupnicoff, Mellanox Technologies; Lloyd Dickman, QLogic; Paul Grun, System Fabric Works;

Abstract—**Remote Direct Memory Access (RDMA)-based communication has enjoyed considerable growth since the introduction of the Virtual Interface Architecture (VIA) in the late 1990s. This growth has accelerated since**

functional extensions to traditional Ethernet. The extensions, taken as a group and commonly referred to Converged Enhanced Ethernet (CEE) [1], are bei standardized within the IEEE. The emergence of CEE al provides an opportunity to enhance Remote Direct Meme



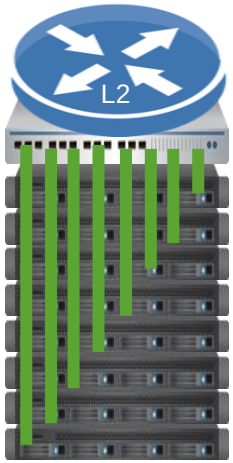
Supplement to InfiniBand™ Architecture Specification Volume 1 Release 1.2.1

Annex A16: RDMA over Converged Ethernet (RoCE)

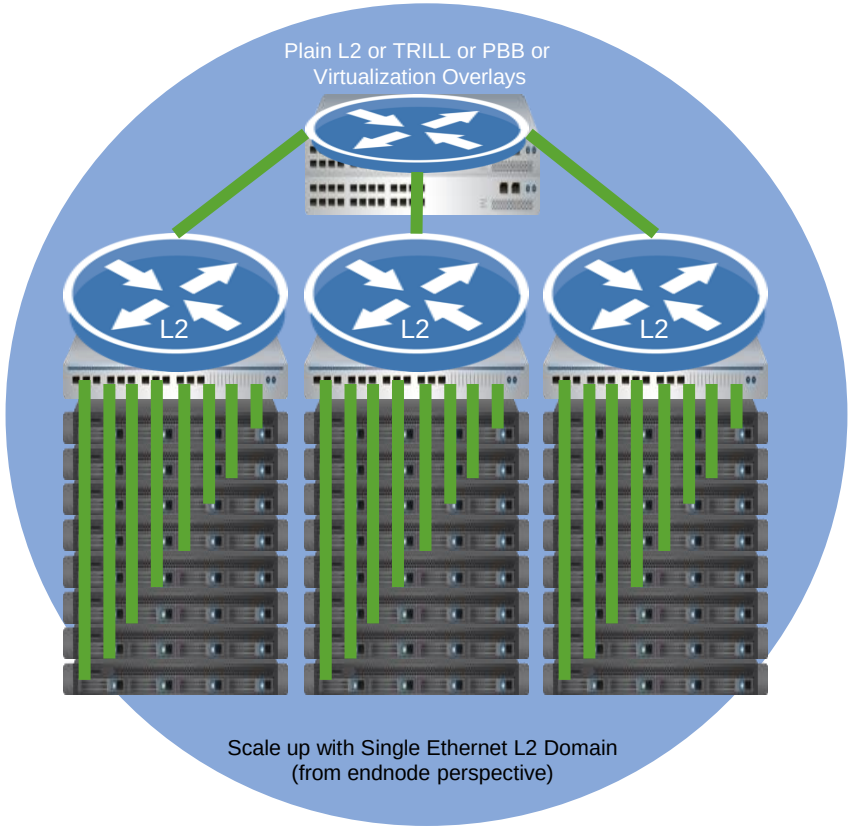
April 6, 2010

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



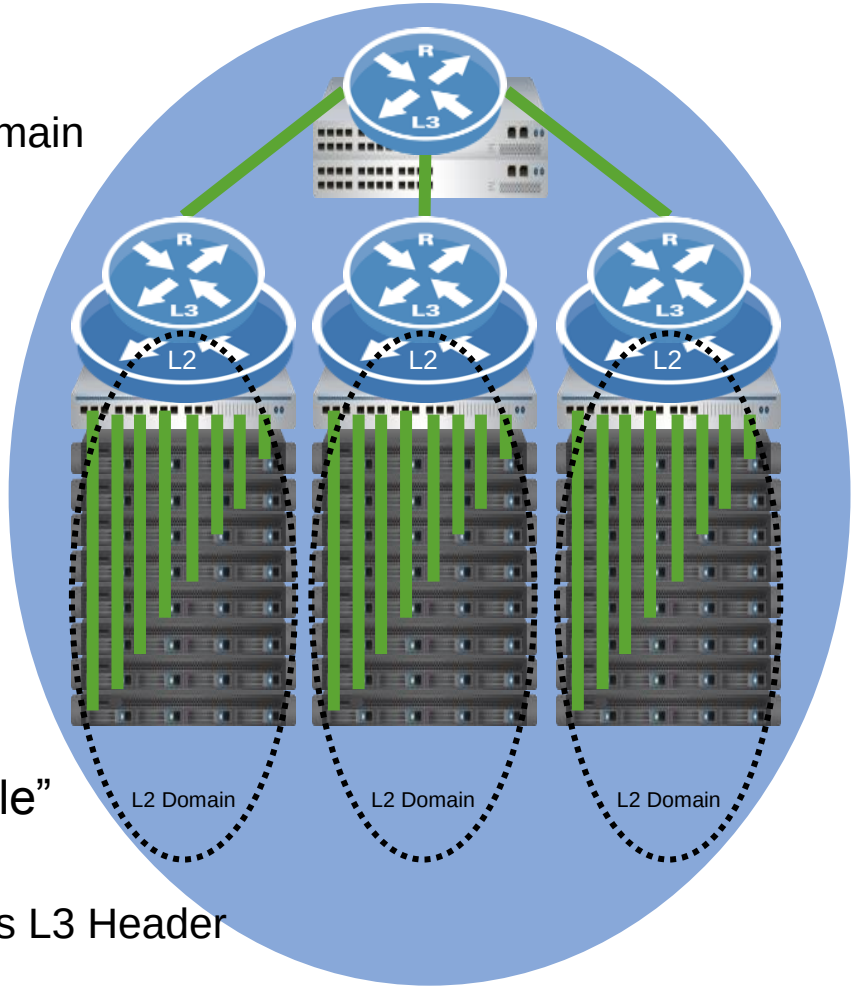
RoCE within a Single Rack



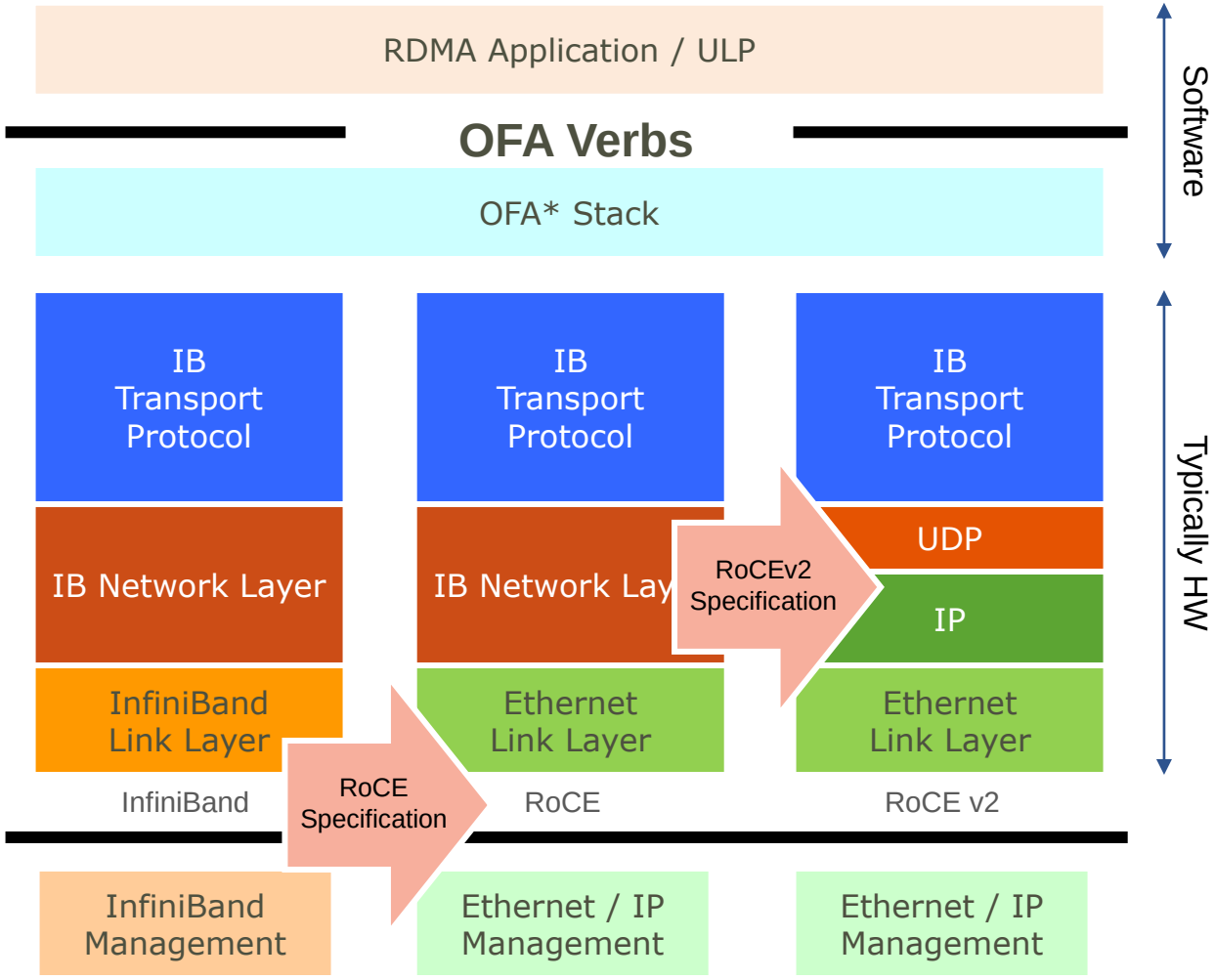
Scale up with Single Ethernet L2 Domain
(from endnode perspective)

The Need for (IP) Routable RoCE

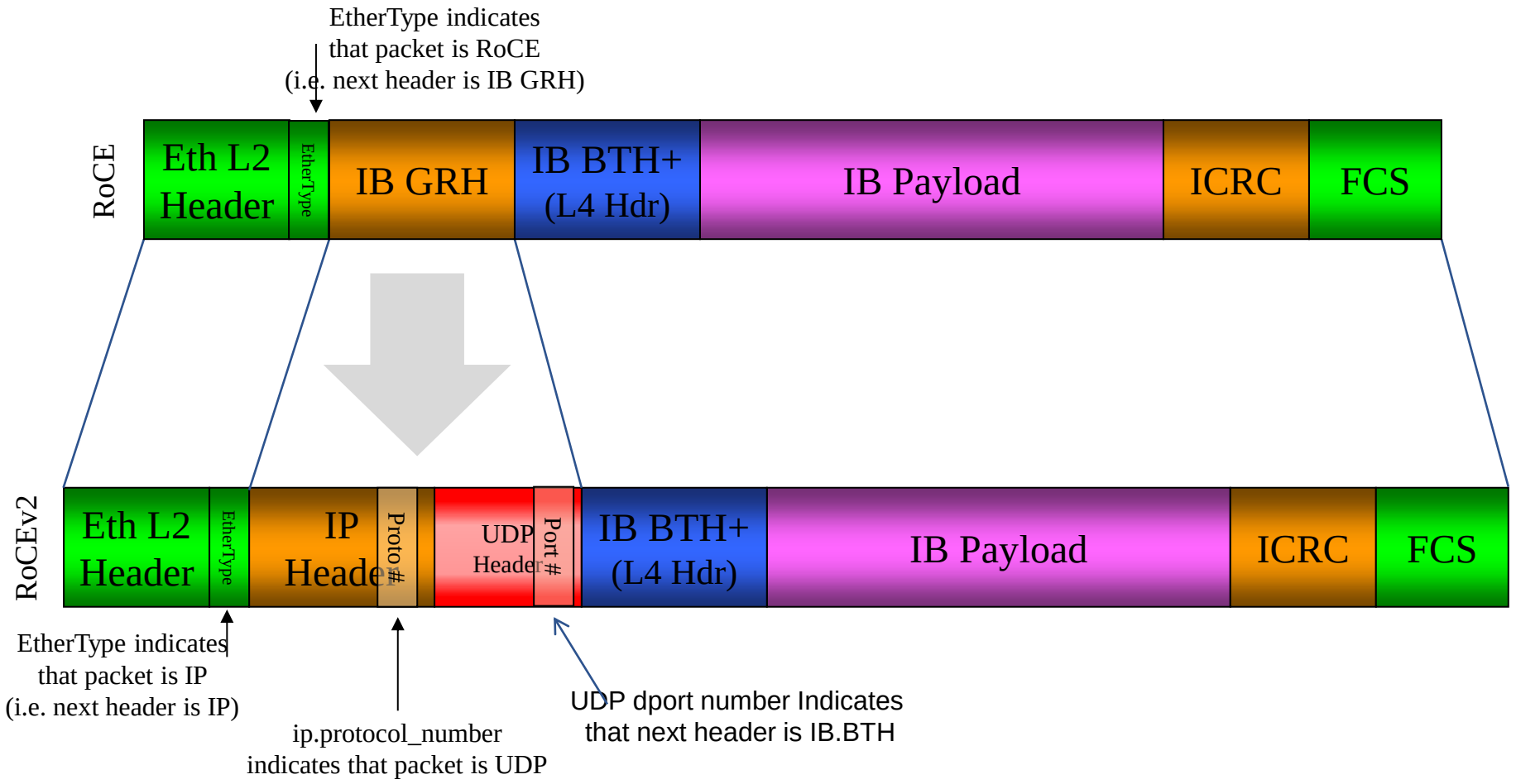
- A Common Class of L3 Datacenter
  - Nodes within a Rack share a Ethernet L2 Domain
  - TOR Device
    - L2 Switch for intra-rack communication
    - L3 (IP) Router for inter-rack communication
  - other topologies also apply
- Customer Demand
  - RDMA Across Racks
    - i.e. Across IP Subnets
  - Focus 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



RoCEv2 - IP Routable Packet Format



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

OPENFABRICS ALLIANCE

10TH ANNUAL INTERNATIONAL DEVELOPER WORKSHOP