RoCE
Status and Plans

November 2013

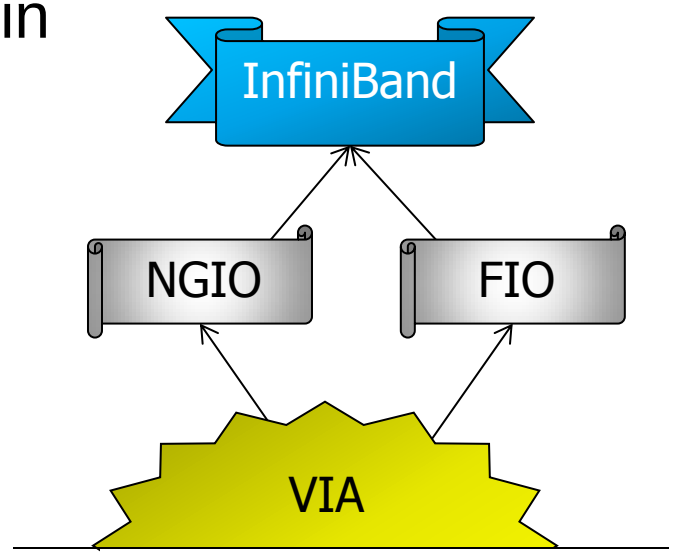
Disclaimer

Some of the information in this presentation represents technical development work currently underway in the InfiniBand Trade Association. As such, it has not yet been approved as an official IBTA standard and is still subject to changes.

The Origins

The InfiniBand Trade Association (IBTA) emerged in 1999 as the merger of competing RDMA proposals, rooted in the Virtual Interface Architecture.

- Message-orientation
- Memory semantic (RDMA read/write)
- Channel semantic (send/receive)
- Address translation
- Management infrastructure
- Verbs – a standard method for accessing the network



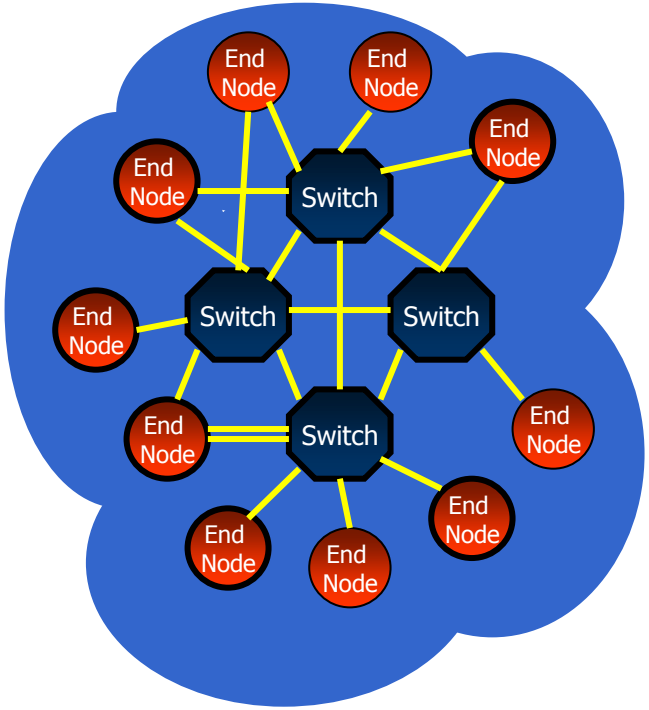
Virtual Interface Architecture

InfiniBand Specification 1.0 published in October 2000

INFINIBAND™ TRADE ASSOCIATION

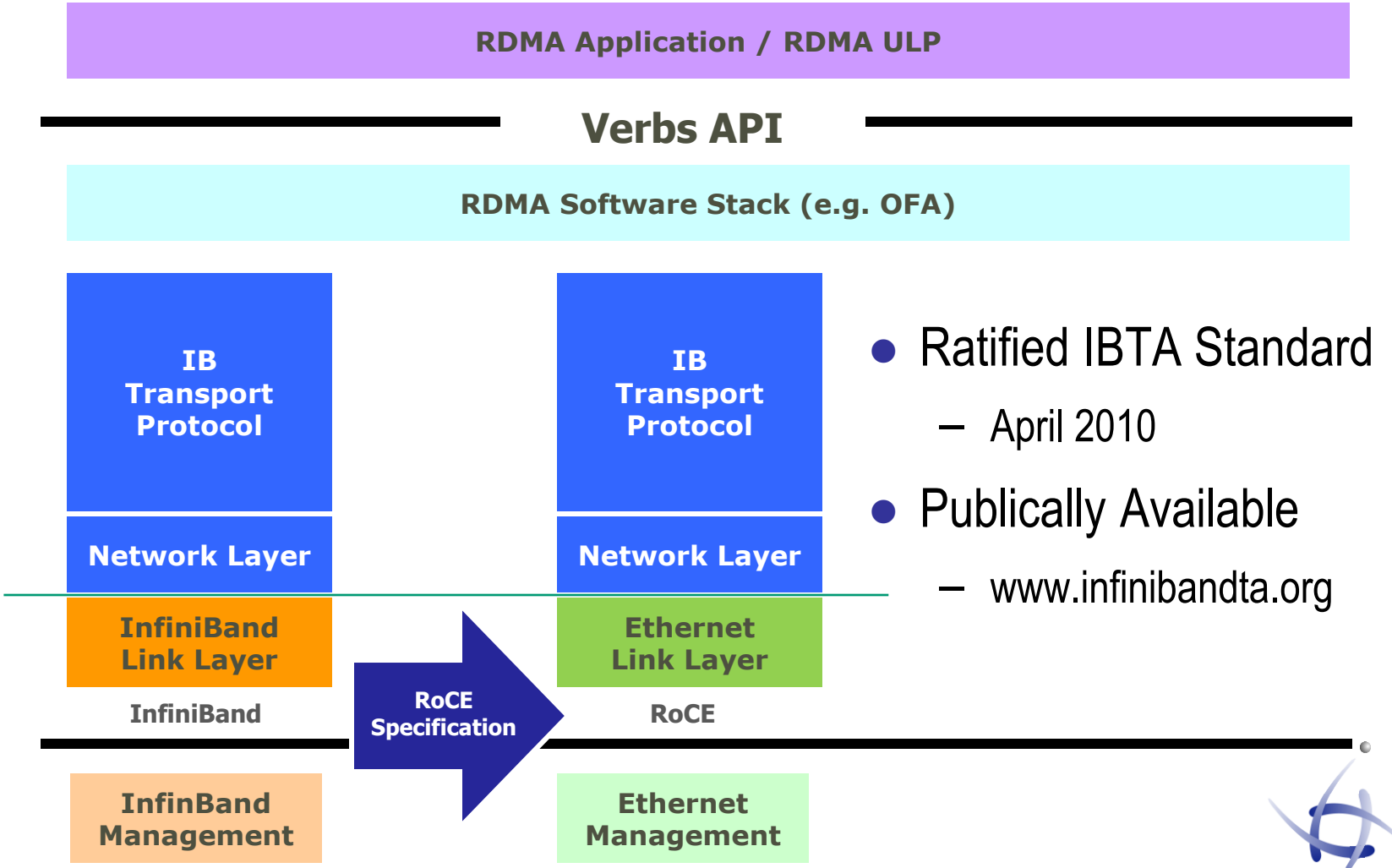
RDMA and the IBTA

- Core Value Proposition
  - User level IO
  - Zero Copy
  - Stack Offload
- Protocol Specification
  - HW Implementation Oriented
- Focus
  - High Performance Computing
  - Data Center Networks



INFINIBAND™ TRADE ASSOCIATION

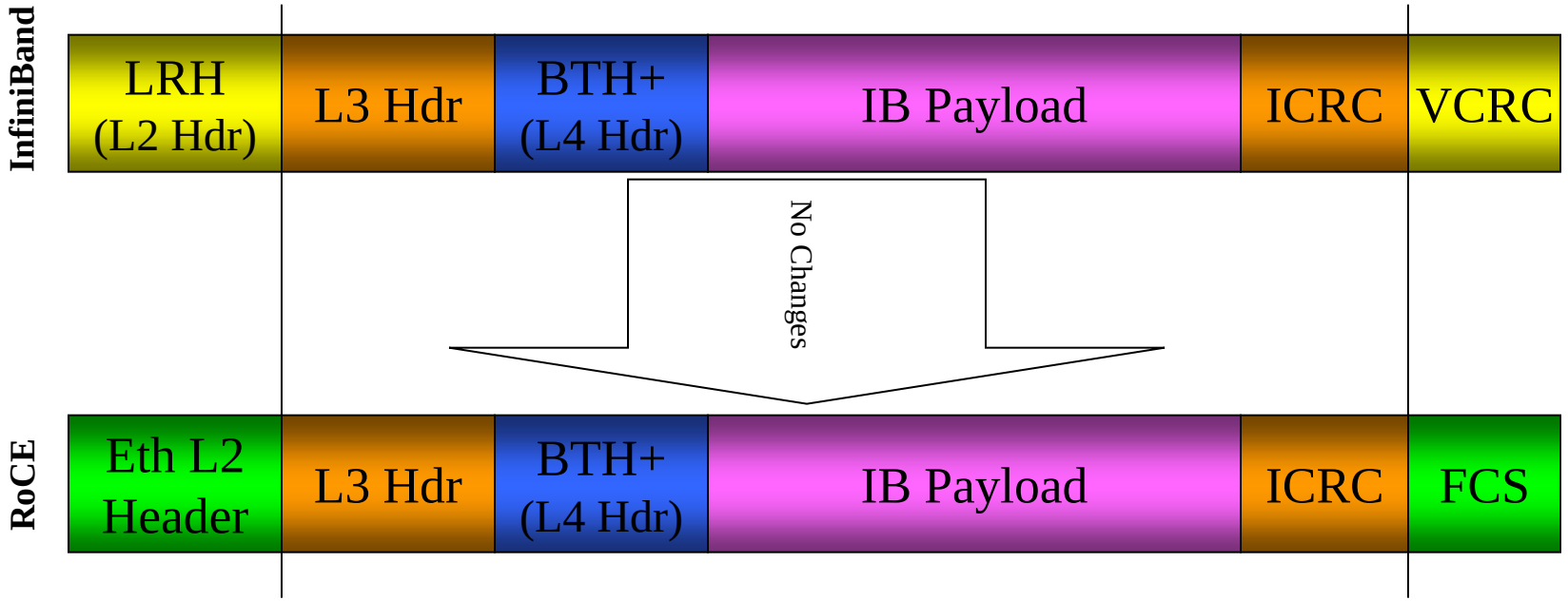
The RoCE Protocol Stack



- Ratified IBTA Standard
  - April 2010
- Publically Available
  - www.infinibandta.org

INFINIBAND™ TRADE ASSOCIATION

RoCE Packet Format Overview



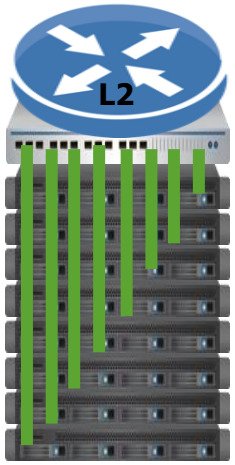
INFINIBAND™ TRADE ASSOCIATION

RoCE Highlights

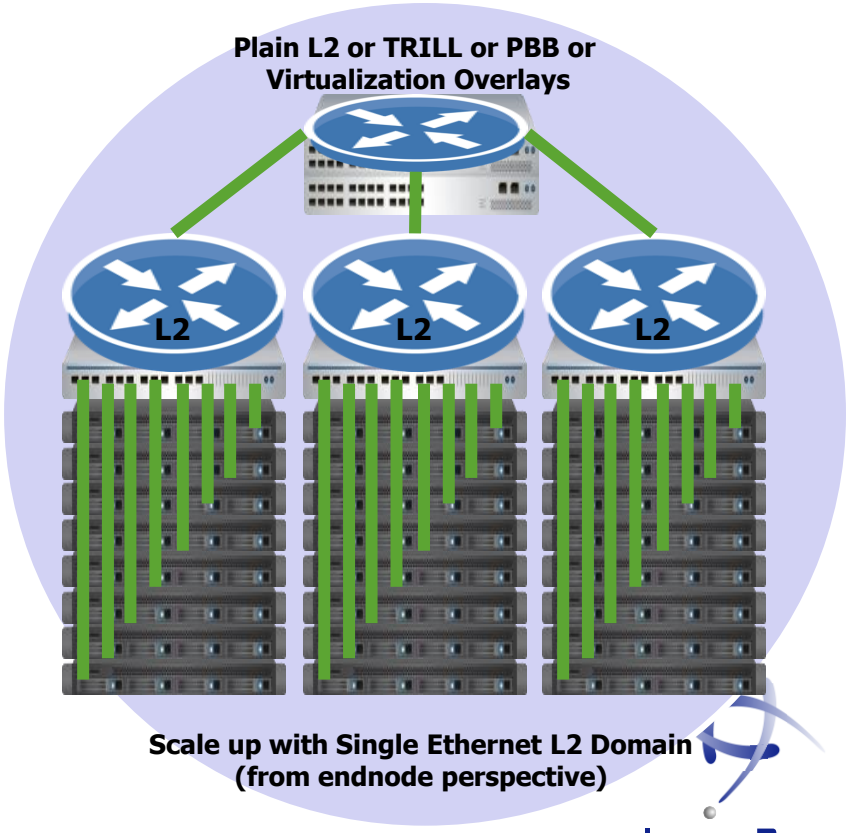
- RDMA Verbs API
  - Transparent to Applications/ULPs
- Ethernet Management Practices
- Purpose Built IB-RDMA Transport Protocol
  - Connected Services (RDMA and Send/Recv)
  - Datagram Services
  - Atomic Operations
  - User Level Multicast
- User Level IO Access / Kernel Bypass / Zero Copy
- Ethertype Based Traffic De-multiplexing (Converged NICs)
- Stateless Traffic Identification (Switch/Fabric Monitoring, ACLs)

RoCE and L2 Datacenters

- RoCE is a L2 Protocol
  - RDMA within a Single Ethernet L2 Domain



RoCE within a Single Rack



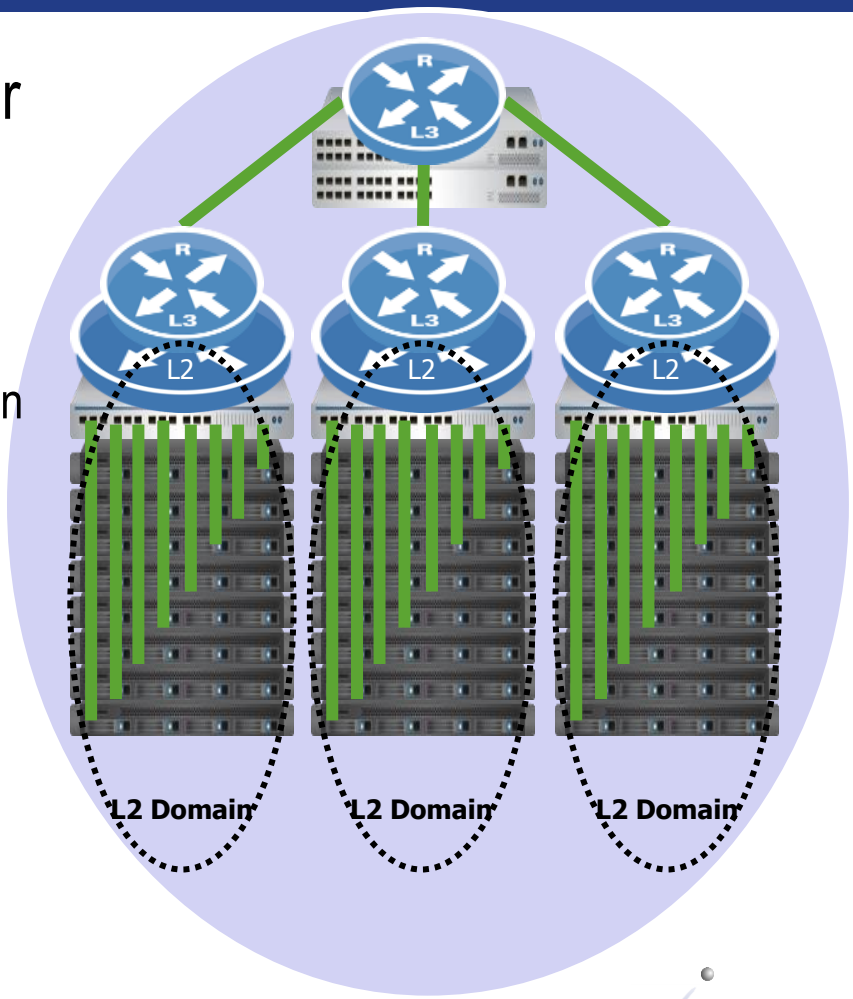
Scale up with Single Ethernet L2 Domain
(from endnode perspective)

RoCE L2 Networks

- Lossless
  - Demystifying “Lossless” – Packets **CAN** be lost in a lossless network and that is OK! (e.g. Alpha Particle Corruption)
  - Link Level Flow Control
- Lossy and Lossless Coexistence in Converged Networks
  - RDMA Traffic on Lossless priorities (PFC)
  - Legacy Traffic on Lossy priorities
- Congestion Management and QoS (Ethernet)
  - Applies to Data Center Traffic in general (not just RDMA)
  - Increasingly Relevant for Converged Networking Scenarios

RoCEv2 - Routable RoCE

- One Common Class of L3 Datacenter
  - Rack defines Ethernet L2 Domain
  - TOR Device
    - L2 Switch for intra-rack communication
    - L3 (IP) Router for inter-rack communication
  - other topologies also apply
- Customer Demand
  - RoCE Across Racks
  - Focus on Data Center Networks
- RoCEv2
  - IBTA Initiated Technical Work to define Routable RoCE



INFINIBAND™ TRADE ASSOCIATION

Some Other Related Topics

- Support for Virtualization
- Multi-Tenant Networks
- Congestion Management

- Possible Areas of Collaboration with IETF

Thank You!

Questions?