



# **IBXoE**

## **IB Transport over Ethernet**

10/06/09

# Important Disclaimer

The following material represents technical development work currently underway in the InfiniBand Trade Association. As such, the work has not been approved by the IBTA and must be approved before it becomes an official IBTA document.

# Agenda

**IBTA and RDMA Overview**

IBXoE Motivation and Goals

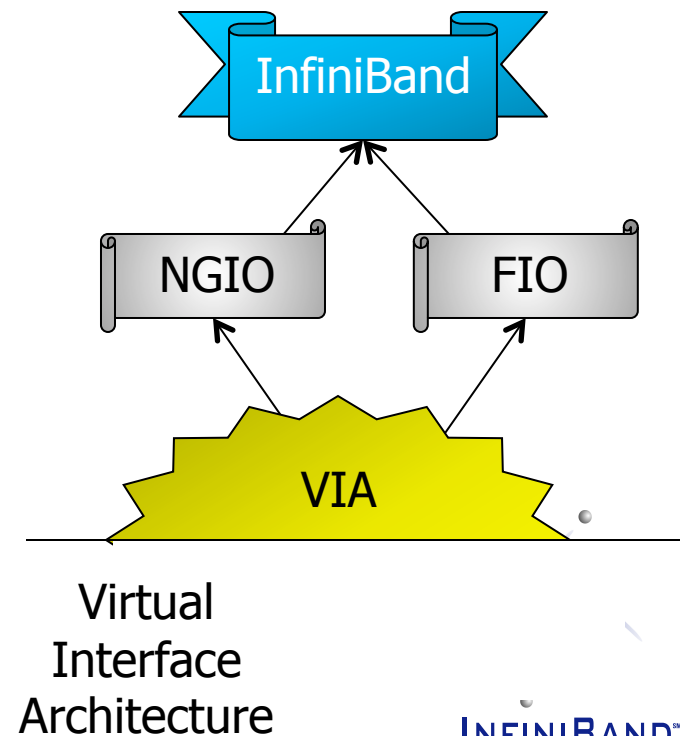
IBXoE Principles of Operation

Summary

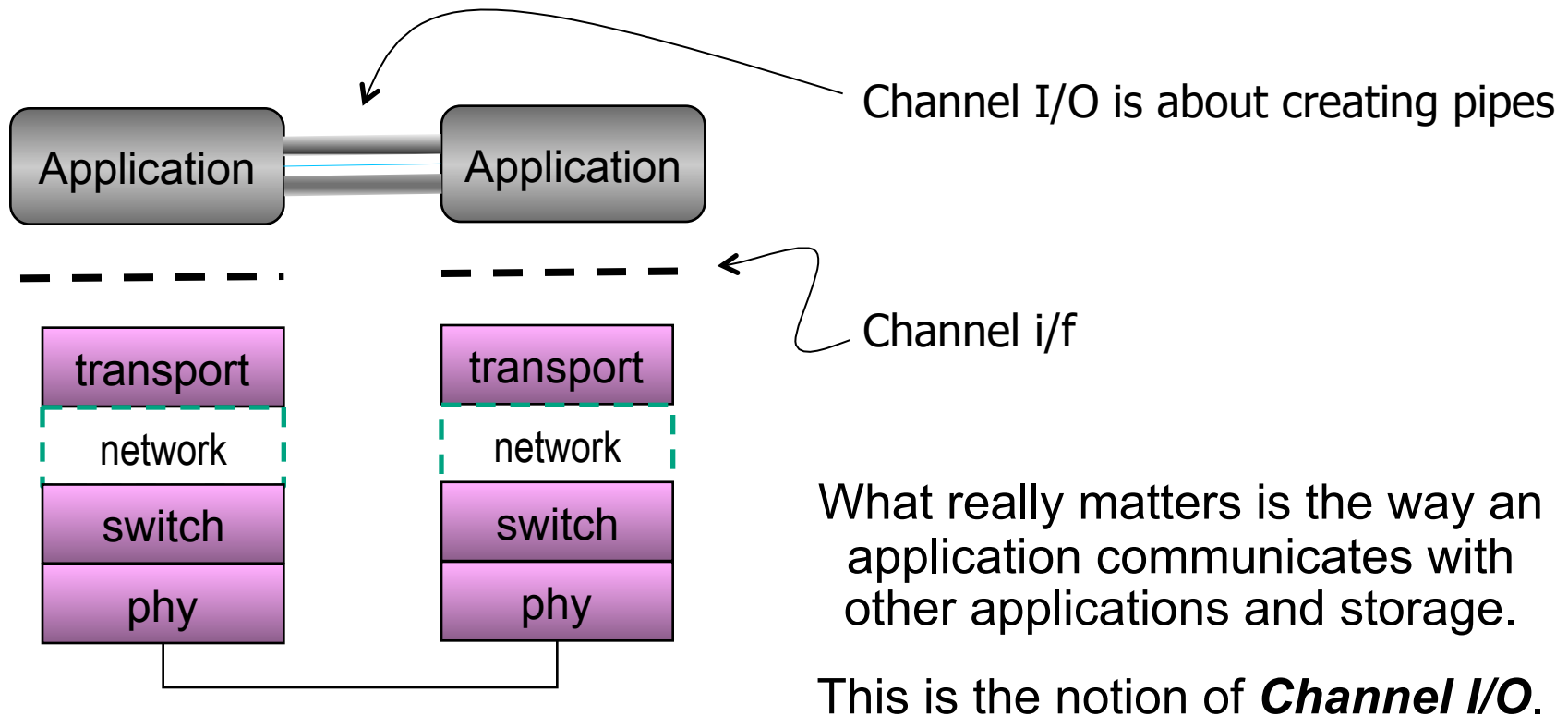
# Short history: InfiniBand Architecture

InfiniBand 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



# Channel I/O

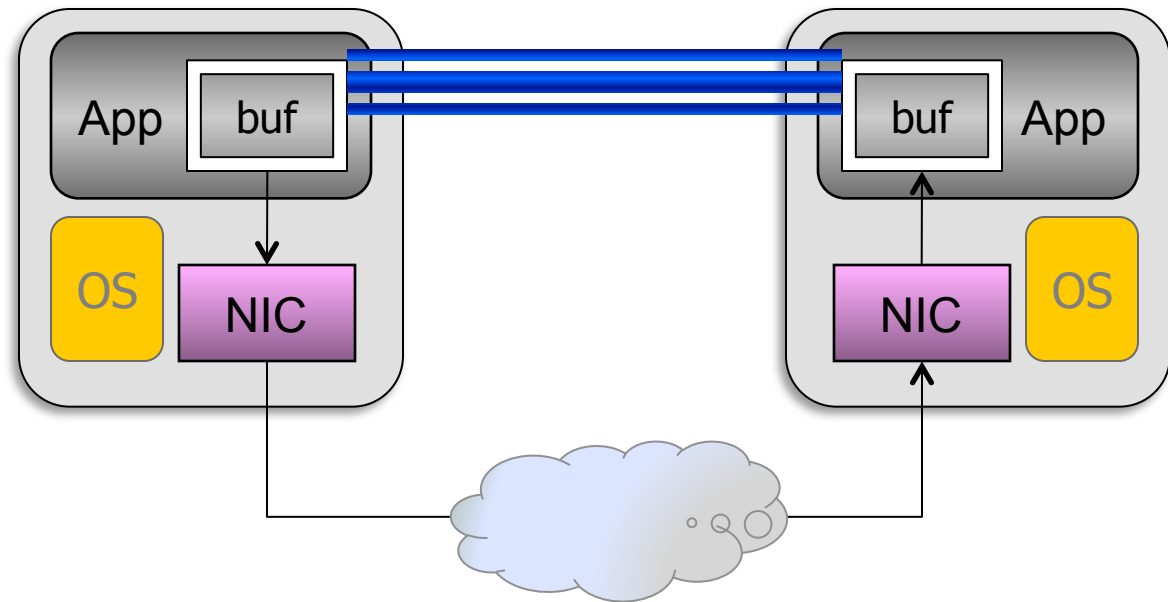


RDMA is the set of mechanisms used to create the channel

# Creating a channel: RDMA

RDMA connects virtual buffers which may be located in different physical address spaces...

- No kernel buffer copies
- No OS context switch for data transfers
- Virtual-to-physical address translation in the NIC. Application accesses the NIC directly.
- RDMA R/W: initiating app targets a virtual buffer in the receiving end. Virtual addresses are carried over the network by the transport.
- Send/Receive: Sender targets a destination 'queue pair'; the destination buffer address is opaque to the sender.



...even across a network.

# How RDMA Works

Consumer

## Consumer:

- posts 'work requests' to a queue
- each work request represents a message...a unit of work

## Channel interface (verbs):

- allows the consumer to manage memory and access network services

Channel  
I/F provider

## IB Channel interface provider:

- Maintains work queues
- Manages address translation
- Provides completion and event mechanisms

IB  
Transport  
Layer

## IB Transport:

- Packetizes messages
- Provides transport service –
  - reliable/unreliable, connected/unconnected, datagram
- Implements RDMA protocol
  - send/receive, RDMA r/w, Atomics
- Implements end-to-end reliability
- Assures reliable delivery



# Agenda

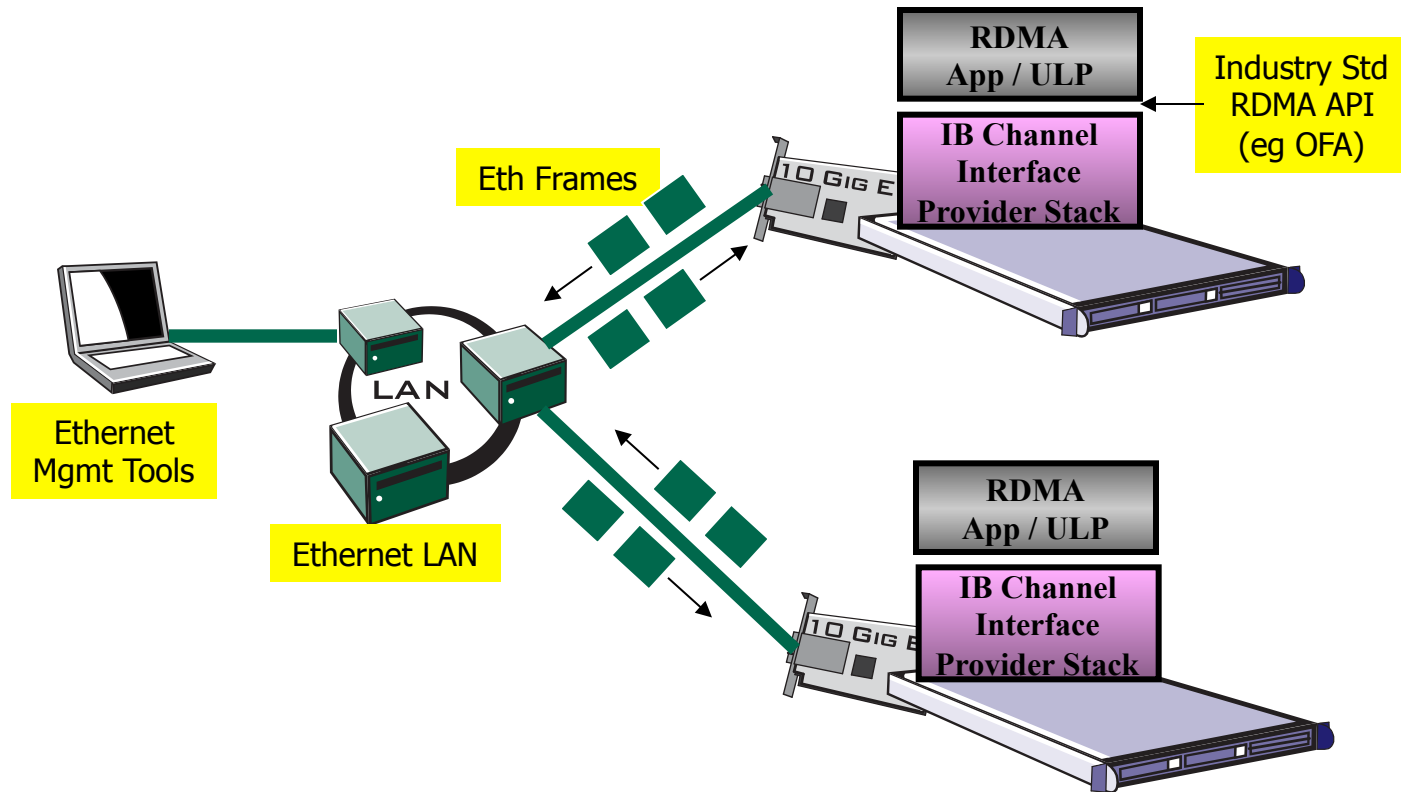
IBTA and RDMA Overview

**IBXoE Motivation and Goals**

IBXoE Principles of Operation

Summary

# IBXoE Goals Overview



# IBXoE Architectural goals

- Provide a robust RDMA solution over lossless L2 networks
- Leverage industry investment in RDMA
  - Preserve investment in development, testing and validation
  - Maintain compliance with industry standard RDMA verbs
  - Interoperate seamlessly with existing implementations
- Take full advantage of lossless Ethernet

# IBXoE Features

- ✓ Full support for RDMA verbs
- ✓ RDMA protocol + a rich set of transport features
- ✓ Memory Management
- ✓ Partitioning support
- ✓ Multicast
- ✓ QoS
- ✓ Congestion control
- ✓ Simple traffic engineering in the fabric

# IBXoE Transport

- ✓ Reliable / unreliable, connected / unconnected, datagram services, atomics
- ✓ Native message-oriented transport protocol
  - msg boundaries identified in on-the-wire packet format
  - simplifies delivery signaling
- ✓ Leverages the lossless wire
  - transport is simple(r) to build in h/w,
  - implementable in h/w or s/w
- ✓ Optimized for datacenter-sized fabrics
  - Linear lookup of MAC/QP

# DCB Link Layer Characteristics

- ✓ Lossless
  - 802.1Qbb priority-based flow control
  - 802.1Qau Congestion Notification
- ✓ Traffic classification (QoS)
  - 802.1Qaz Enhanced Transmission Service (ETS)
  - 802.1Qbb priority-based flow control
- ✓ Partitioning
  - VLANs

# Agenda

IBTA and RDMA Overview

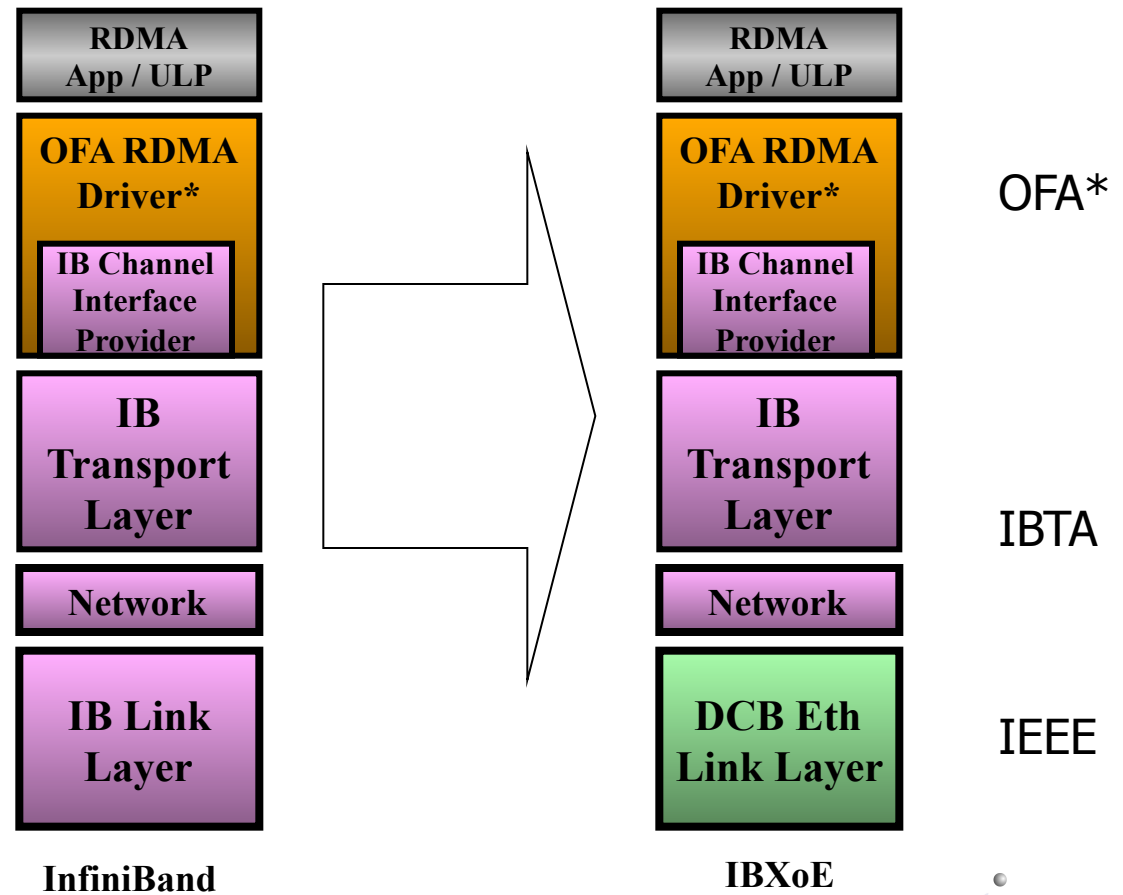
IBXoE Motivation and Goals

**IBXoE Principles of Operation**

Summary

# IBXoE Protocol Stack Overview

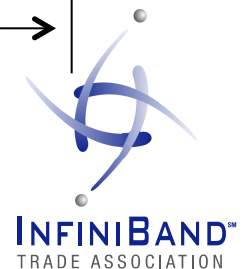
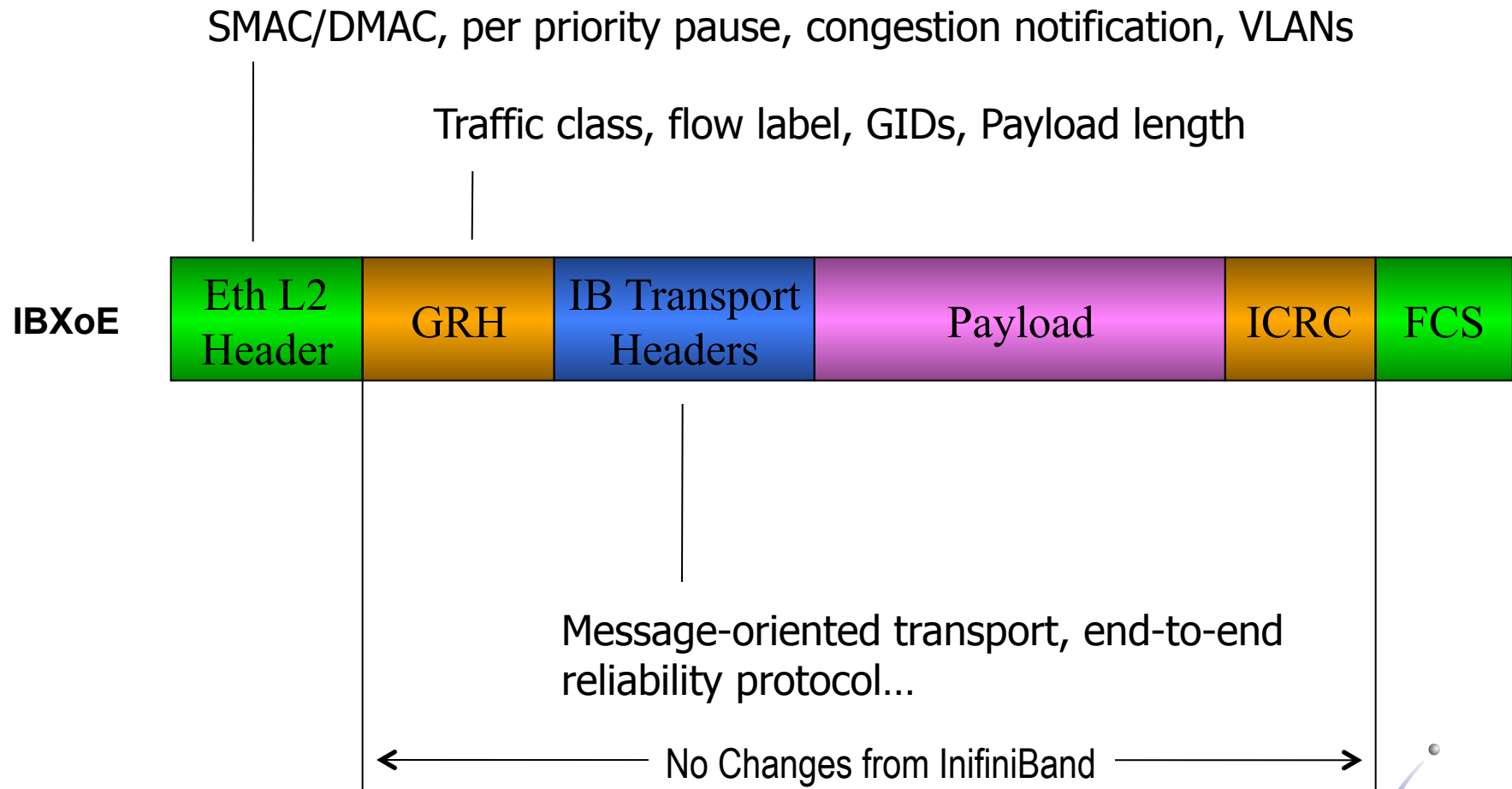
IBXoE makes IB's efficient transport available to applications (user, kernel) using a familiar Ethernet wire.



\*OFA: Open Fabrics Alliance.

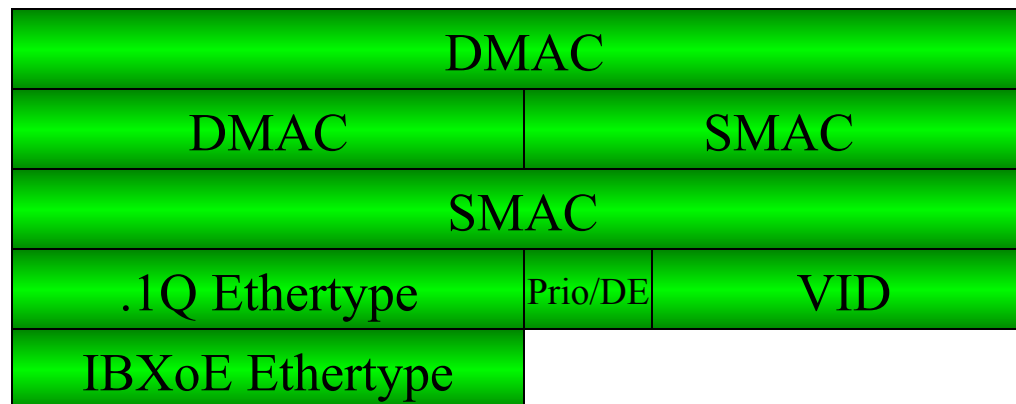
OFA Driver is shown as an example. The solution is intended to operate in conjunction with any driver that implements the IB channel interface

# IBXoE Packet Format



# IBXoE L2 Header

- DMAC,SMAC replace DLID,SLID
- .1Q header priority field replaces SL
- .1Q header VLAN ID field allows for L2 partitioning
- IEEE Assigned Ethertype 0x8915 for IBXoE



# IBXoE L2 Addresses and Forwarding Management

	<b>IBXoE</b>	<b>InfiniBand</b>
<b>L2 Address Assignment</b>	<b>Burned In (or Eth managed)</b>	<b>IB Subnet Manager</b>
<b>L2 Topology Discovery</b>	<b>STP (or Eth managed)</b>	
<b>Switch FDB Configuration</b>	<b>Learning (or Eth managed)</b>	

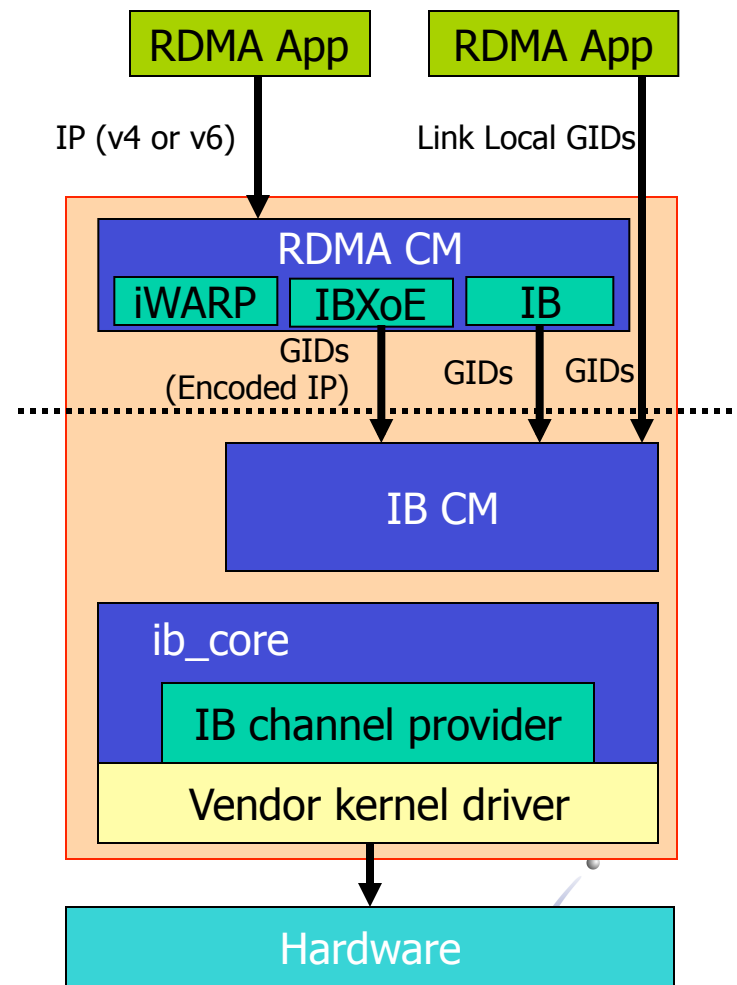
- IB SM/SA replaced by standard Ethernet methods
- Supports any topology supported by Ethernet

# L2 Feature Mapping

	<b>IBXoE</b>	<b>InfiniBand</b>
<b>Link Layer Flow Control</b>	<b>802.1Qbb Priority Based Flow Control (PFC)</b>	<b>IB L2 Flow Control</b>
<b>QoS</b>	<b>802.1Qaz Enhanced Transmission Selection (ETS) .1Q prio field - Eth Network Management</b>	<b>VL Arbitration (+ IB SA / QoS Manager)</b>
<b>Congestion Management</b>	<b>802.1Qau Congestion Management</b>	<b>IB Congestion Management</b>
<b>Configuration</b>	<b>DCBX / SNMP MIBs</b>	<b>SMI/GSI</b>

# IBXoE Addressing

- IBXoE uses the IB Communication Management (CM) protocol
  - Connection Establishment
  - UD Service Resolution (SIDR)
- IBXoE addressing is GID based
  - GIDs are IPv6-like L3 addresses
  - IBXoE L3 Header (GRH) is IPv6-like
- Transport-agnostic RDMA application addressing is IP based (e.g. OFA RDMA CM)
  - IBXoE GIDs can carry the corresponding App Level IP addresses
  - Address Resolution and Interface Binding implemented as RDMA CM Service
- Support for non-RDMA CM apps through Link Local GIDs
  - Encode L2 Address



# IBXoE Management Framework Overview

- No SM/SA/SMI/SMA
  - No QP0
- QP1 preserved for CM

	InfiniBand	IBXoE
L2 Performance Monitoring	IB GSI	SNMP/RMON MIBs
Baseboard Management		SNMP/RMON MIBs
Device Management		SNMP/RMON MIBs
SNMP Tunneling		N/A
...		...
Communication Management	IB CM	IB CM

# Agenda

IBTA and RDMA Overview

IBXoE Motivation and Goals

IBXoE Principles of Operation

Summary

# Summary, next steps

- ❑ Lossless Ethernet complements IB verbs + IB transport
- ❑ IB transport
  - is a native RDMA transport
  - enjoys widespread adoption among users of RDMA technology
  - strong industry support
  - major investments in technology development, validation and deployment
- ❑ IBXoE Annex under development in IBTA