Advancing RDMA

OPENFABRICS ALLIANCE

A proposal for RDMA on Enhanced Ethernet

Paul Grun

SystemFabricWorks – pgrun@systemfabricworks.com

Objective: Accelerate the adoption of RDMA technology

Why bother?
I mean, who cares about low latency anyway?

RDMA – The Real Deal

RDMA means <u>applications talk to applications</u> which means:

- ➢ A competitive edge for the Enterprise user
  - Wall Street thrives on microseconds…predictable microseconds
- ➢ Reduced resource utilization; tThe IT guy buys less stuff. Period.
  - Less servers, less cables, less switches
  - And that means green
- ➢ Easier on the Enterprise IT budget
  - Scalability means grow as you go
- ➢ The IT guy can give his users much needed flexibility
  - Easy application deployment, easy application mobility

This is good for the Enterprise and his IT guy, good for the server vendor, good for the middleware and application provider, good for the hardware vendor…

The end user view

- RDMA delivers value propositions that are not available through any other communications paradigm
- These value propositions are compelling
  - Improved resource utilization, flexible resource allocation, low latency, scalability, unified fabric…
- Nevertheless, IT purchasing decisions are often driven by the lowest layers in the stack – the link and phy layers
  - Switches, cables, infrastructure and so forth

But RDMA isn’t about the wire!

Here’s the big idea:
Broaden the appeal of RDMA by lowering barriers to its adoption

(remember, RDMA is not about the wire)

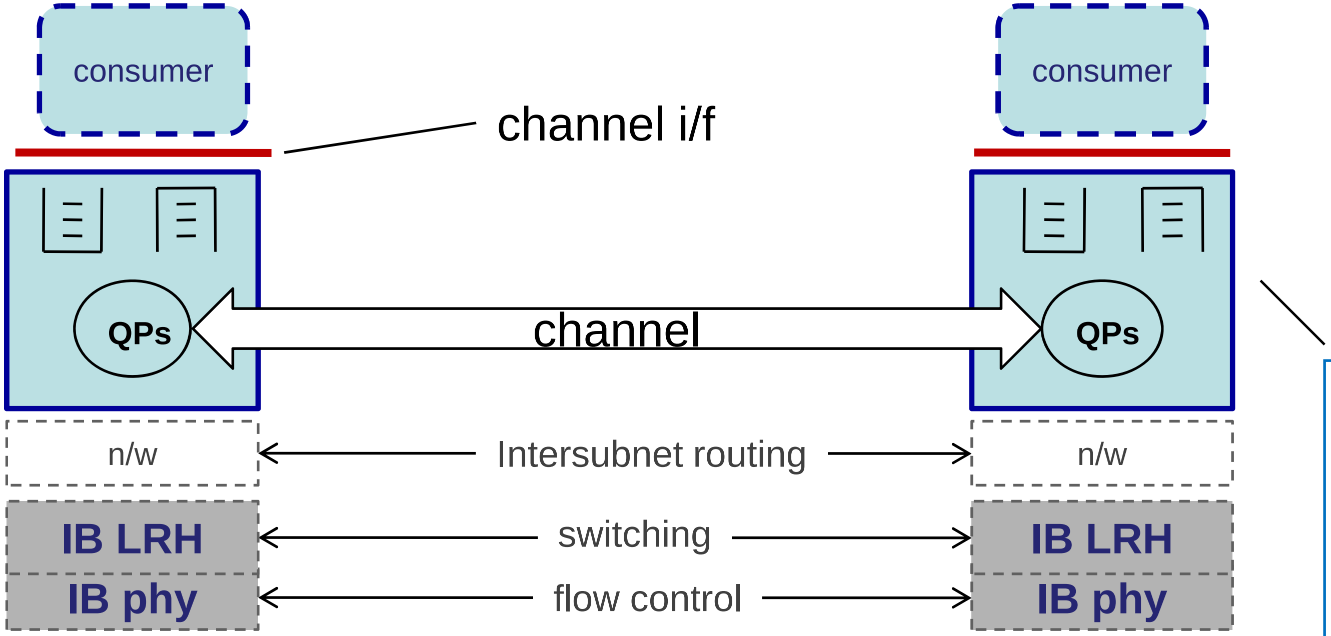
Proposed approach

1. Define & standardize RoEE – RDMA on Enhanced Ethernet

   RoEE defined to be a verbs compliant IB transport running over the emerging IEEE Converged Enhanced Ethernet standard

2. Drive development of the necessary software stack

IB Architecture

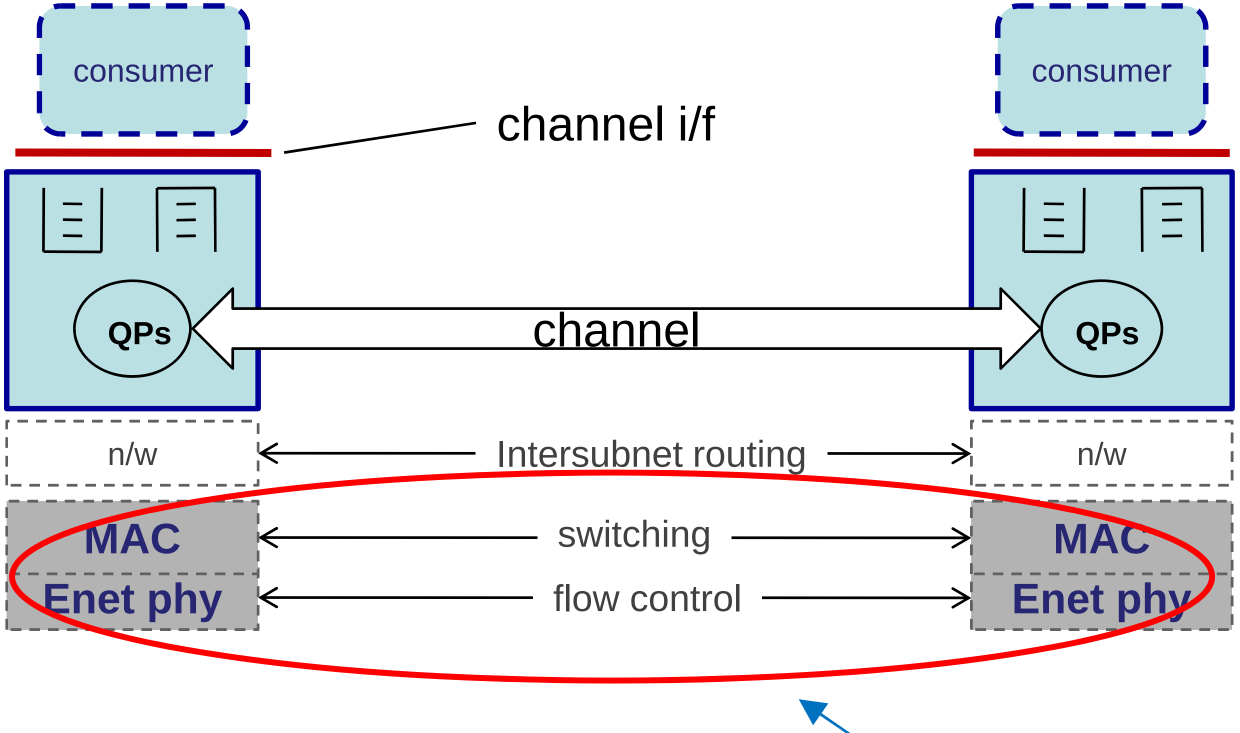


“RDMA in a box”…

InfiniBand Architecture includes:

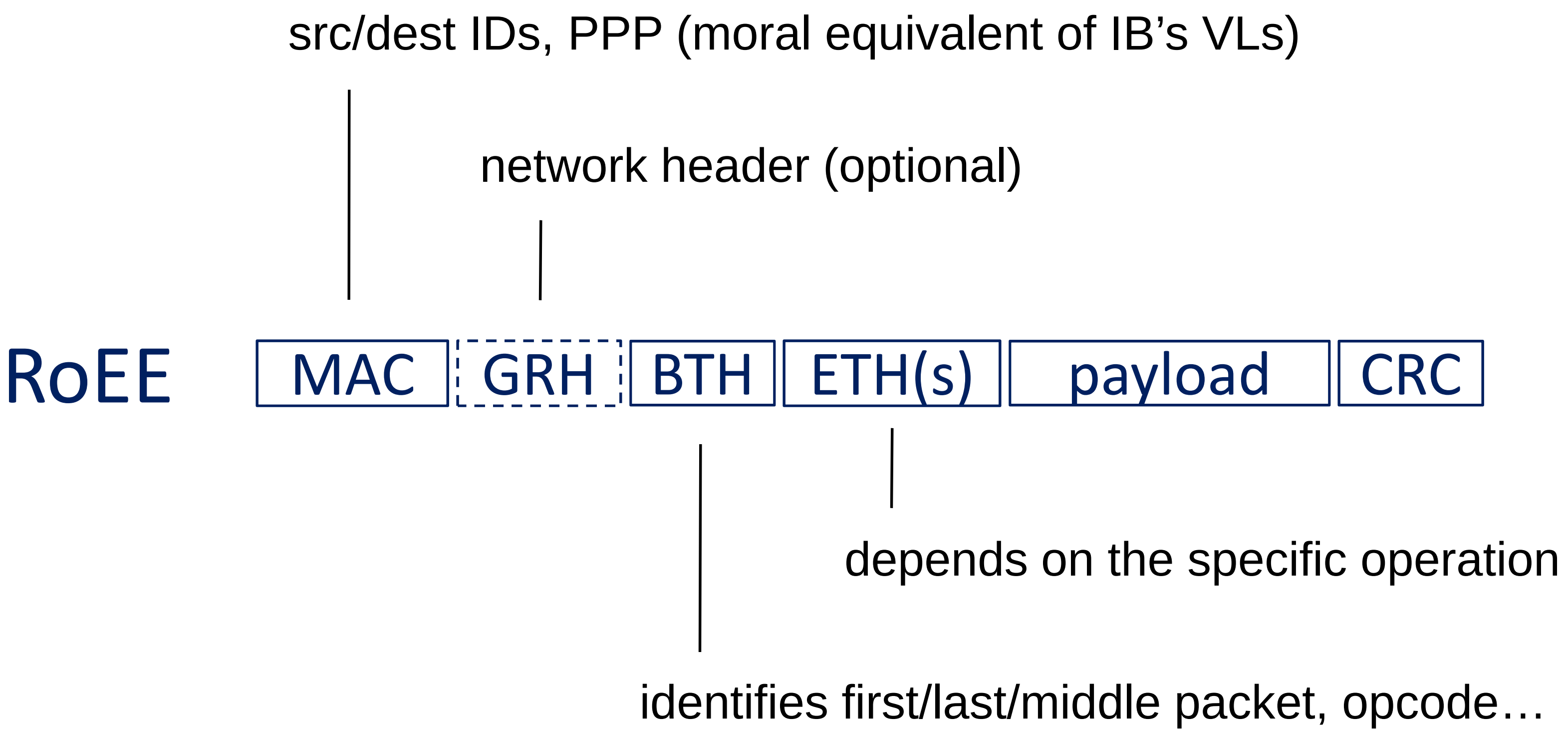
- a verbs-compliant *channel i/f*
- a *channel*, defined at each end by *queue pairs*
- a *transport service*
- network, link and phy layers, switches, cables, routers, mgmt, CM…

RoEE Architecture is…



"RDMA in a box"… …running on an Enet wire

On-the-wire packet format



src/dest IDs, PPP (moral equivalent of IB's VLs)

network header (optional)

RoEE

MAC

GRH

BTH

ETH(s)

payload

CRC

depends on the specific operation

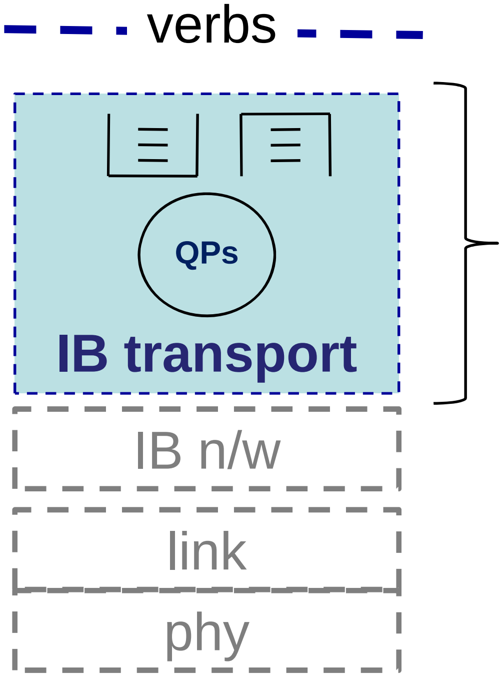
identifies first/last/middle packet, opcode…

A little more about the network layer

- IB assumes that routing between subnets is not always required and not always desirable
  - reduced latency, reduced jitter
  - therefore the network layer is optional

- IB uses a four tuple to demux traffic for delivery
  - SLID / DLID, SRC QPN / DEST QPN
  - This allows a simple linear lookup mechanism

The magic IB transport



Transport i/f: verbs-defined work queues, completion signaling, event mechanisms…

Transport: on-the-wire packet formats and protocols supporting a wide variety of services and operations

- ➢RC, UC, RD, UD, RAW services
- ➢Send/Receive, RDMA READ, RDMA Write, Atomic ops

“RDMA in a box:
Everything you could want in a verbs compliant transport”

Magic??

- ✓ Native message-oriented transport protocol
  - →msg boundaries identified in on-the-wire packet format
  - →simplifies delivery signaling
- ✓ Leverages the lossless wire
  - →transport designed for a lossless wire is simple(r) to build in h/w
  - →no need for slow start
- ✓ Optimized for datacenter-sized fabrics
  - →Network header is supported, but optional
  - →Linear lookup of LID/QP vs IP address
- ✓ Rich set of transport services
  - →Reliable / unreliable connected services, atomics…

Q: So what changed?

A: Ethernet!

	802.1	IB	CEE
Lossless	No	Yes	Yes
Classes of service	No	Yes	Yes
Congestion management	No	Yes	Yes

These are precisely the features required by the IB transport for efficient operation

Standards

- IEEE defines Ethernet standards. RoEE requires no modification to the emerging CEE standard
- IBTA defines a proven RDMA transport layer. Adapting it to run over the emerging CEE link/phy layers is not a difficult extension to the existing standard
- OFA builds on those standards by providing an open source, verbs compliant implementation based on the RoEE standard

Defining the RoEE standard

- On-the-wire protocol
- Switch discovery and programming
  - centralized vs distributed switch management?
- Connection management protocol
  - CM? ARP?
- Supported network topologies, network layer
  - GRH? IPv6?
- Multiplexing RoEE and IP over a single CEE NIC

In consultation with IEEE and/or IETF and/or OFA

Next steps

- IBTA: Create an Annex to the IBTA standard defining RoEE
- OFA: Create the necessary s/w stack

Thanks!

Paul Grun
Chief Scientist
SystemFabricWorks – Fabric Computing that Works
pgrun@systemfabricworks.com

BACKUP

Converged Enhanced Ethernet

- Consists of three related IEEE standards
  - 802.1Qau Congestion Notification, 802.1Qbb Priority-based flow control, 802.1Qaz Enhanced Transmission Selection, DCBX
- Primary target is on supporting a converged fabric
  - Driven to some extent by FCoE
- That target can be significantly expanded by complementing CEE with a native RDMA transport
- The combination of CEE + a native RDMA transport can deliver:
  - Better resource utilization, vastly improved performance, flexible resource allocation, improved ‘green’ footprint, low latency, scalability, natural virtualization…
  - …to a broad range of standard upper layer protocols

Key elements of CEE

Introduces a lossless wire via link level flow control & congestion control. The combination of these two is intended to reduce or eliminate the incidence of dropped packets, except in the case of error.

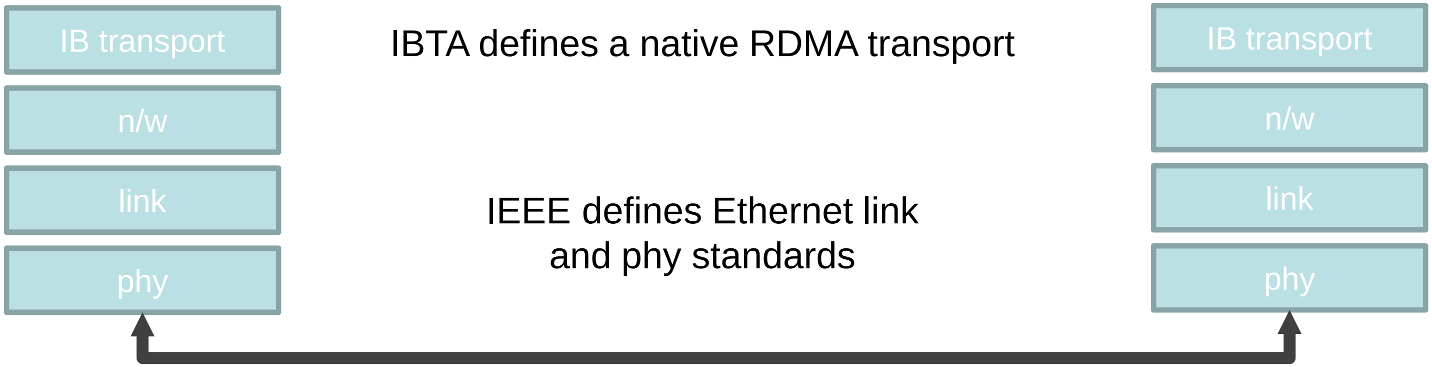
This crucial change means that it is no longer necessary for the transport layer to manage routinely dropped packets. This is critical to transport performance.

Ability to segment traffic into classes, (“virtual lanes”)

a. Historically, Enet used spanning tree to avoid deadlocks. IB, uses virtual lanes to guarantee deadlock free topologies. With the emergence of ‘virtual lanes’ in Enet, it becomes possible to build arbitrary topologies.
b. The existence of ‘virtual lanes’ means that Ethernet is now capable of supporting various forms of traffic engineering.

Taken together, these changes mean CEE is layer suitable for supporting an IB RDMA transport.

RoEE – why a new standard?



IBTA defines a native RDMA transport

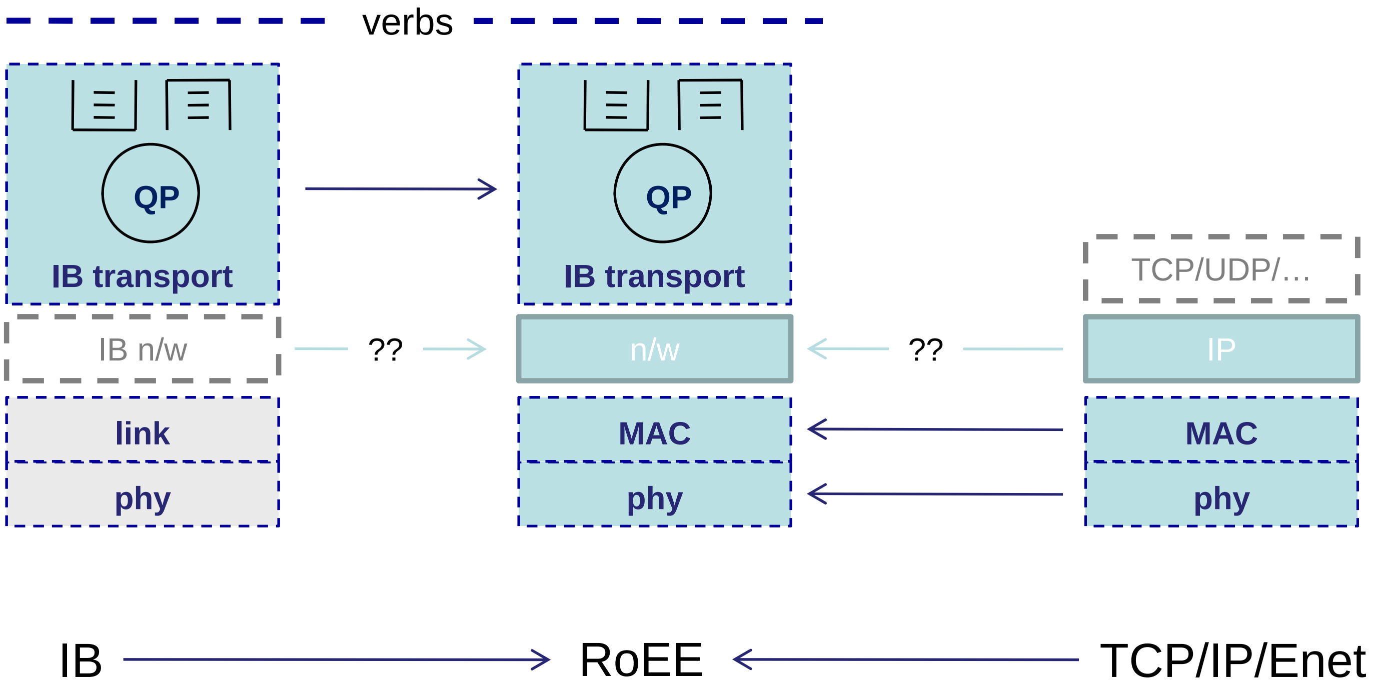
IEEE defines Ethernet link
and phy standards

Gap: there is no standard defining a wire protocol for a native RDMA transport over an Ethernet wire

To be done:
- on-the-wire protocol for native RDMA over CEE
- connection management
- supported topologies

As usual, the IBTA does not define an implementation. It is anticipated that both h/w and s/w implementations will emerge.

RoEE - what it is



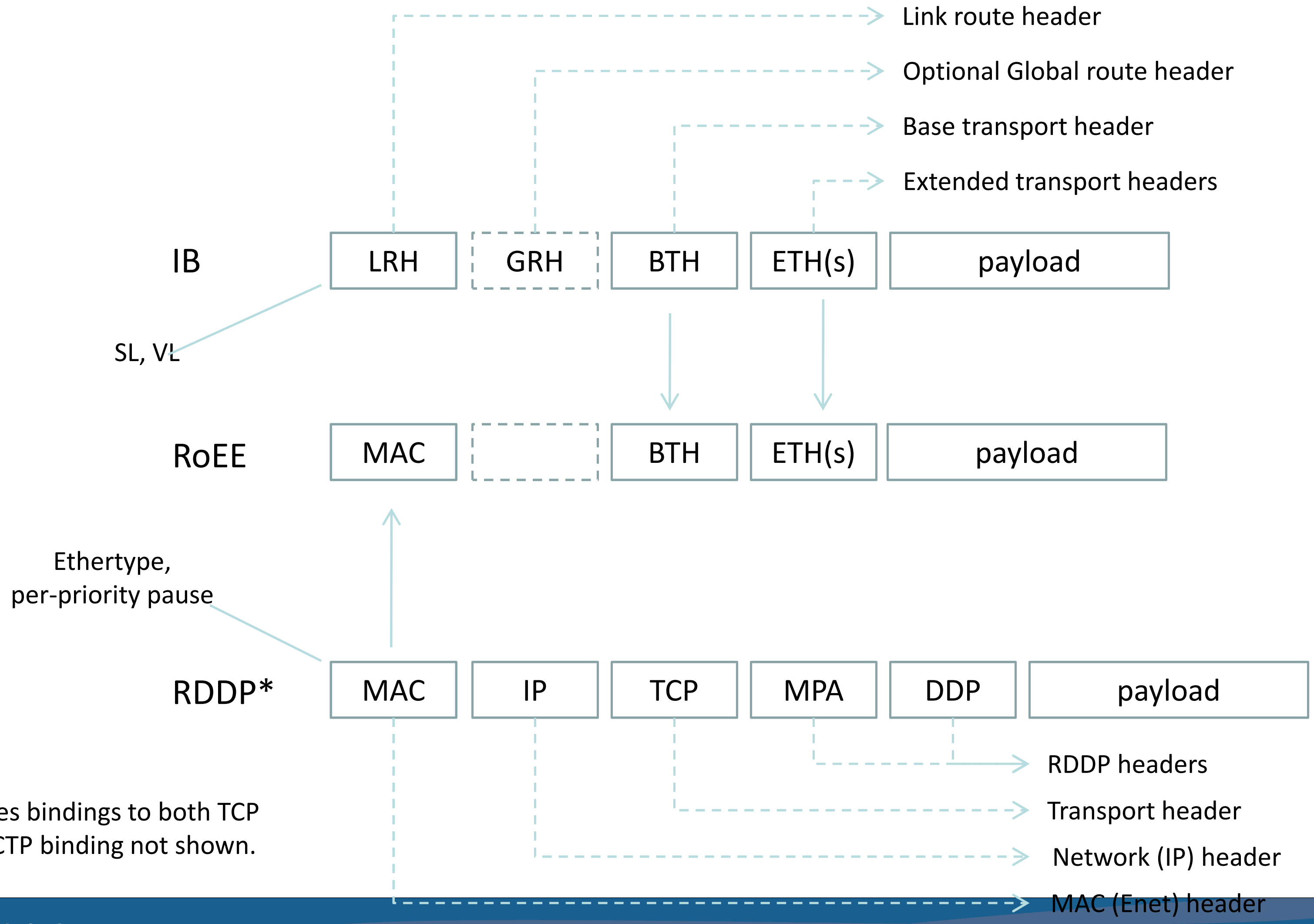
RoEE makes IB's efficient transport available to applications (user, kernel) using a familiar Ethernet-based wire.

No changes are expected at the verbs layer or the application interface.

A technical view

- The RDMA transport paradigm depends on a required set of characteristics at the lower levels in the stack
  - No packet dropping, support for multiple classes of traffic, arbitrary topologies and so on
- Those characteristics were not available in Ethernet
  - Until very recently

Comparative header formats



*RDDP defines bindings to both TCP and SCTP. SCTP binding not shown.