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

Tom Talpey

Microsoft Corporation

Paul Grun

System Fabric Works

Outline

- The paper
- The context of the paper
- RoCEE
- The road ahead
- Panel discussion

Summary of the Paper

- Remote Direct Memory Access is now mainstream over many operating systems, networks, and application upper layers
- Converged Enhanced Ethernet is a reality

- How can we bring them together?
- What will emerge as the protocol?

Remote Direct Memory Access - RDMA

- Peer-to-peer, memory-to-memory access
- Extremely low latency, low overhead
- High operation rate, high bandwidth

- Standard API (Open Fabrics Alliance Verbs)
- Ubiquitous Operating System support
- Multiple fabrics (Infiniband, iWARP, etc)
- Multiple upper layers and applications

Converged Enhanced Ethernet - CEE

- IEEE emerging standard, 10+ Gbps
- Key components:
  - 802.1Qbb priority-based flow control
  - 802.1Qaz Enhanced Transmission Selection and Data Center Bridge Exchange
  - 802.1Qau end-to-end congestion notification
- Taken together, enable a lossless network with differentiated classes of service
- Converges: IPC, network and storage traffic

RDMA Fabric deployment

- The Point Of Delivery (“POD”)
  - A “unit of computing”, with network
  - Datacenter rack / shipping container / etc
  - Cluster in a box
- High density, low diameter
  - Minimal latency, tightly managed
- CEE as a POD fabric (and beyond)
  - Lossless, congestion controlled, deterministic
- RDMA matches POD application semantics

RDMA Protocols: Our Requirements

1. Support the Open Fabrics Alliance software ecosystem
2. Have no significant barriers to adoption in the market
3. Be standards-based, preferably available in open source

RDMA Protocols: Our Criteria

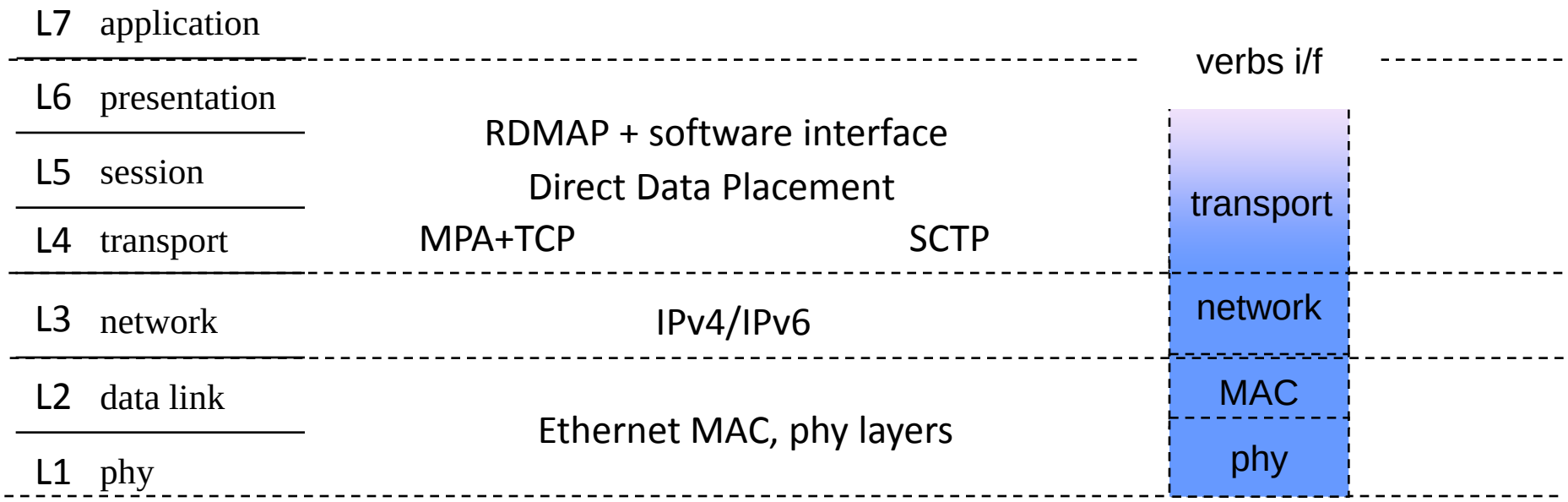
1. Degree of improvement to responsiveness
   - Use of endpoint and network resources
   - Optimal latency and throughput
   - Minimal latency variation (jitter)
2. Suitability for implementation in both **hardware** and **software**
   - Blended deployment
   - Cost-effective scalability

RDMA Protocol Candidates

- RDDP over TCP/IP
  - Today's iWARP, several implementations
- RDDP over SCTP/IP
  - Specified by iWARP but not commercially implemented
- RoCEE
  - The subject of this paper

RDDP layering

- A set of RDMA protocols running on an Internet stack
- Compliant with the verbs API and s/w stacks defined by OFA*



* OFA = Open Fabrics Alliance

RDDP/TCP – "iWARP"

- IETF standard RDDP/TCP specification
- Strongly layered
  - Requires addition of MPA framing layer for record marking, integrity (CRC32c) and optional "markers"
- Utilizes existing infrastructure seamlessly
  - Carried by network as any other TCP ULP
- Relies on TCP offload for performance
  - Hardware-based solution

RDDP/TCP

- Meets requirements
  - OFA well-supported, highly standards compliant
- Mixed meeting of criteria
  - Uncompelling as a software implementation
  - Vulnerable to jitter (TCP processing)
- Good choice for:
  - Larger Delay-Bandwidth pipelines (> ~10^6)
  - RDMA traffic leaving the POD
  - Overcoming significant error/reordering network conditions

RDDP/SCTP - Future

- IETF RDDP standard, but rarely implemented
- Requires infrastructure to support SCTP
  - Lightly supported and deployed in datacenter
- Potential advantages
  - Reduced offload state
  - Avoids expensive (to software) MPA layer

RDDP/SCTP

- Meets requirements
  - OFA supportable, standards compliant
- Mixed meeting of criteria
  - Uncompelling as a software implementation
    - But better than RDDP/TCP, due to MPA omission and message/packet framing
  - Minimal infrastructure support
    - SCTP adoption unclear
    - Lack of compelling reasons to change from TCP

RoCEE

- RDMA over Converged Enhanced Ethernet
- Reuse of existing InfiniBand protocols

Disclaimer

The name ‘RoCEE’ (RDMA over Converged Enhanced Ethernet), is a working name.

You might hear me say RoXE, RoE, RDMAoE, IBXoE, IBXE or any other of a host of equally obscure names.

As technologists, I am certain that we will all exercise with the utmost diligence our responsibility to come up with a new and interesting name.

It just hasn’t happened yet.

“Ibex over e”

RoCEE

RDMA over Ethernet is designed to allow…

… the deployment of RDMA semantics on lossless Ethernet fabrics…

… by running the IB transport protocol using Ethernet frames.

RDMA over Ethernet packets consist of standard Ethernet frames with an IEEE assigned Ethertype, a GRH, *unmodified* IB transport headers and payload

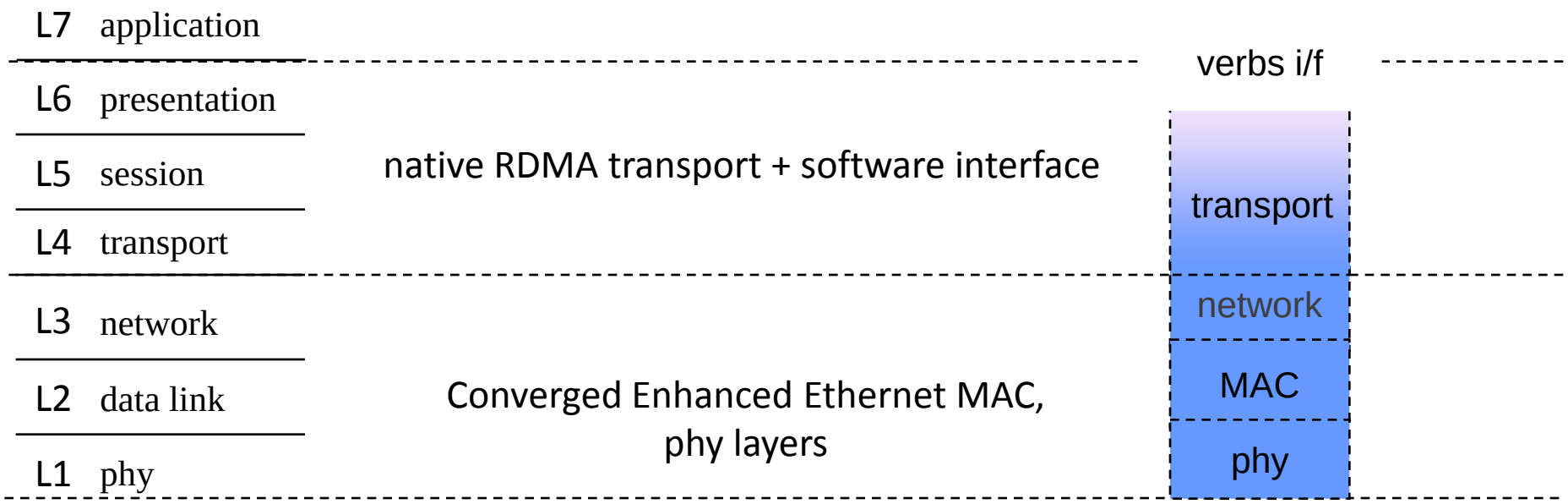
RDMA over Ethernet encodes IP addresses into its GIDs and resolves MAC addresses using the host IP stack. For multicast GIDs, standard IP to MAC mappings apply.

IB subnet management and SA services are not required for RDMAoE operation; Ethernet management practices are used instead.

RoCEE

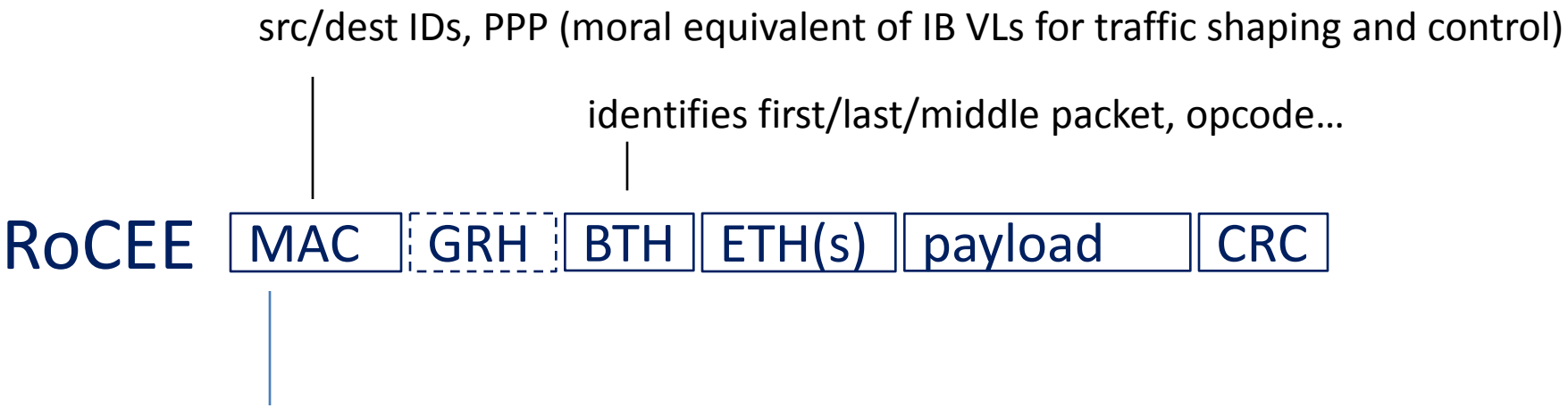
RDMA over (Converged Enhanced) Ethernet

- A native RDMA protocol running on an Ethernet fabric
- Compliant with the verbs API and s/w stacks defined by OFA*



* OFA = Open Fabrics Alliance

On-the-wire packet format

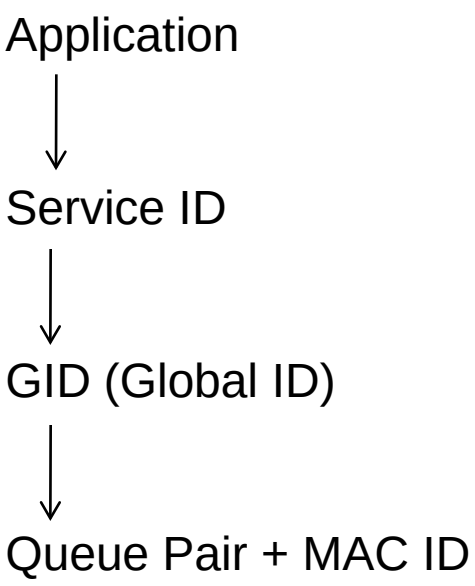


Ethertype field (0x8915) uniquely identifies an RDMA network

Distinguishing the network type at Layer 2 enables traffic classification and shaping to occur in the fabric.

This is immensely helpful in a unified fabric.

Simplified Addressing



Application references a service

Service ID resolves to a GID

GID resolves to a (QPn* + MAC ID)

Packet delivery to the application is based on a simple lookup of the QP number

Packet switching in the fabric and message delivery in the end point are based on the simple Queue Pair number / MAC ID tuple

*QPn = queue pair number

So, what changed?

The emergence of CEE is the precipitating event…

	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

RoCEE characteristics

- ✓ Complements the unified fabric
  - → Ability to recognize RDMA at Layer 2 allows traffic engineering in the fabric
- ✓ Optimized for datacenter compute models
  - → Fast efficient linear lookup of LID/QP vs IP address
  - → Highly efficient message-based model
- ✓ A native message-based transport
  - →Rich set of transport services – reliable/unreliable, connected, unconnected, atomics…

Why RoCEE?

RDMA means applications talk to applications which means:

- A competitive edge for the Enterprise user
  - Wall Street thrives on microseconds…predictable microseconds
- Reduced resource utilization; The IT guy buys less stuff. Period.
  - Less servers, less cables, less switches
- 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

✓ RoCEE brings these same values to the Ethernet market

CONCLUSION

RoCEE protocols

- New RDMA stack framable on raw Converged Enhanced Ethernet, employing proven RDMA technologies and protocols
- Stack may be compellingly supported in end nodes by multiple implementation approaches, including pure software.
- Network framing of RDMA traffic enables the fabric to apply traffic criteria per-packet and per-flow

Future

- Standardization
- Implementation
- Application and Upper Layer adoption
- Enterprise adoption

PANEL DISCUSSION

Question to the Panel

- Has the time come to promote Ethernet as a low latency, high throughput fabric interconnect?

Discussion

Abstract: The answer to this question hinges on support for RDMA-based Communication. Motivated by FCoE requirements, enhancements to the Ethernet standard transform this unreliable media into one that can support lossless transmissions. This has motivated three distinct camps to come forward with conflicting proposals for an RDMA transport. The three options are:

(i) Remote Direct Data Placement (RDDP) over TCP,

(ii) RDMA over Converged Enhanced Ethernet (RoCEE), and

(iii) Data Center RDMA Protocol (DCRP).

All three see the host I/O adaptor and the message framing protocols as being at the heart of the challenge. Where they differ, however, is hotly contested.

- Moderator:
  - David Cohen, EMC
- Panel members:
  - Dwight Barron, HP
  - Ásgeir Eiriksson, Chelsio
  - Patrick Geoffray, Myricom
  - Gopal Hegde, Cisco
  - Michael Kagan, Mellanox