Towards Hyperscale HPC & RDMA

Paul Congdon

(Tallac/Huawei)

paul.congdon@tallac.com

IETF-104 HotRFC

Current HPC/RDMA networks

"Future datacenters of all kinds will be built like high performance computers," said Nvidia CEO Jensen Huang

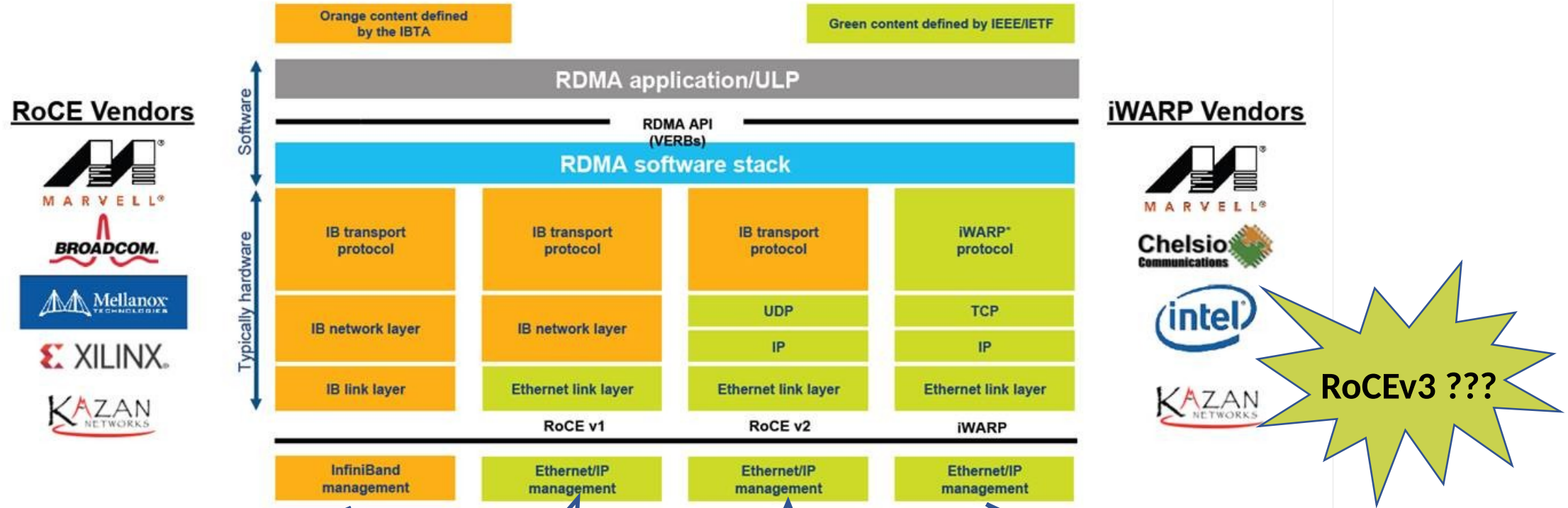
- Traditionally, HPC runs over custom lossless technologies
  - Infiniband
  - Link Layer Credit-based Flow Control

- More recently designed to run over IP infrastructure
  - iWARP (IETF RFC 5040 – RFC 5044, RFC 6580, RFC 6581, RFC 7306)
  - RoCEv2 (https://www.infinibandta.org/)

- The results produced by these networks are mainstream through the integration of *artificial intelligence*, *machine learning*, *data analytics and data science workloads*

RoCE vs. iWARP Network Stack Differences

SNIA ESF | ETHERNET STORAGE



Separate Network, Not Ethernet/IP

Not Route-able, L2 Data Center, Complex L2 Congestion Control (QCN)

Incomplete Congestion Control, reliance on L2 PFC

Unspecified TCP tweaks, TCP HW NIC, Slow Start

What does it mean to be Hyperscale

- The term “hyperscale” refers to a computer architecture’s ability to scale in order to respond to increasing demand.
- Goals
  - Common cloud scale infrastructure
  - Dynamic and automated provisioning
  - Diverse workload mix
  - Low latency, high throughput
- Suggestions have been made to scale RDMA/HPC
  - RDMA over commodity Ethernet at scale, SIGCOMM 2016
  - iWARP Redefined: Scalable Connectionless Communication over High-Speed Ethernet, 2010 International Conference on High Performance Computing
  - Tuning ECN for Data Center Networks, CoNEXT '12
  - Revisiting Network Support for RDMA, SIGCOMM 2018
  - https://datatracker.ietf.org/doc/draft-chen-iccrg-rocev3-cm-requirements/
    - RoCEv3 = Improved retransmission strategy
      Improved congestion control mechanism (RTT, credit, ECN)
      Finer grain load balancing with looser re-ordering requirements

What if scenarios for Hyperscale HPC

- What if networks didn’t have to be lossless, but just very low loss?
- What if iWARP was run over Enhanced UDP instead of TCP?
- What if congestion management was fully defined for RDMA?

Can we hyperscale HPC?

- **Side Meeting: Monday 10AM Room: Tyrolka**