IETF-104 HotRFC and Side Meeting on Hyperscaler HPC/RDMA				
Date:		**April 23, 2019**		
Author(s):				
Name	**Affiliation**	**Address**	**Phone**	**email**
Paul Congdon	Huawei/Tallac			paul.congdon@tallac.com

Abstract

The included slides and draft meeting notes are originally from the HotRFC and side meeting on Hyperscalar HPC/RDMA held at the IETF-104 meeting in Prague during March 23 - 29, 2019. The original HotRFC slides are available on IETF repositories at: https://datatracker.ietf.org/meeting/104/materials/slides-104-hotrfc-10-hyperscale-hpc-and-rdma-01. The side meeting notes are included here as there is no known IETF repository for such notes.

Towards Hyperscale HPC & RDMA

Paul Congdon
(Tallac/Huawei)
paul.congdon@tallac.com

IETF-104 HotRFC

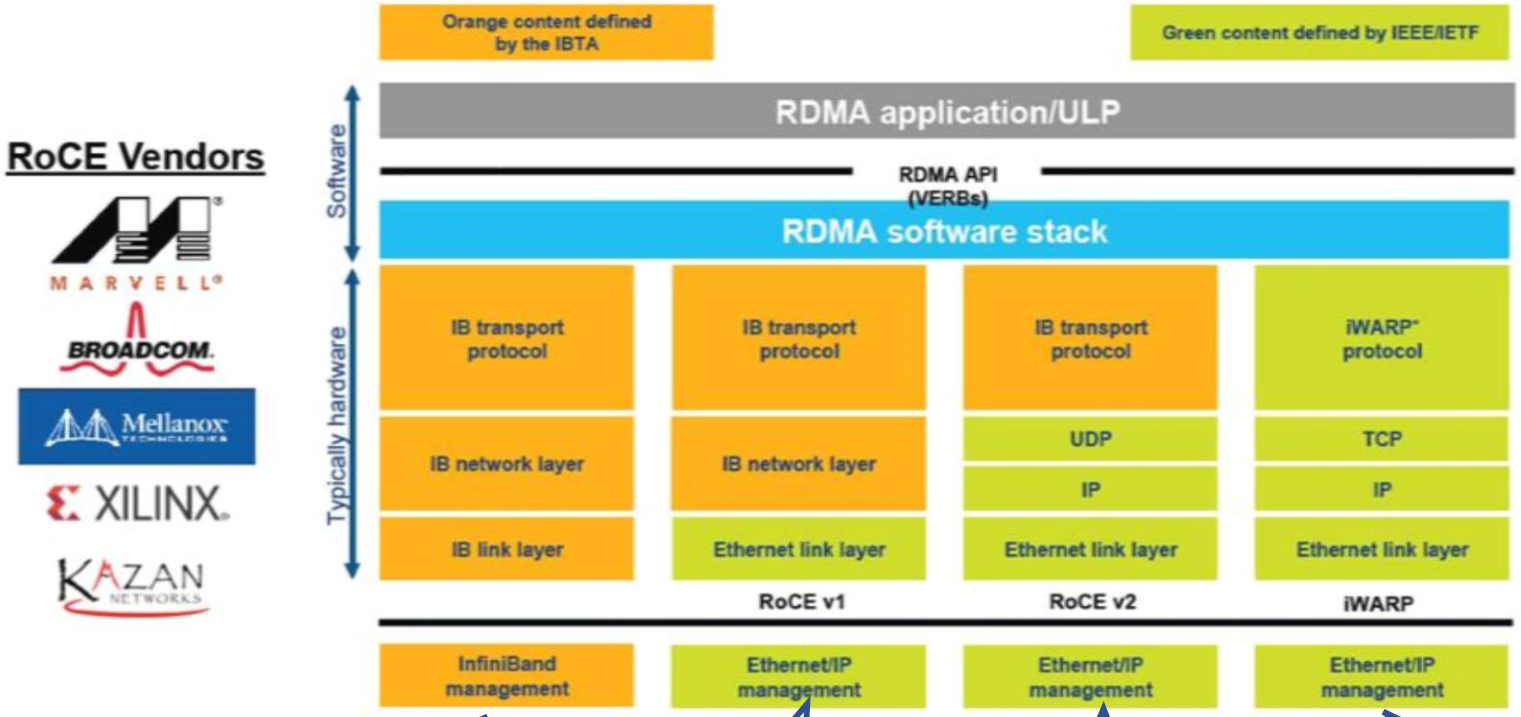
Current HPC/RDMA networks

"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

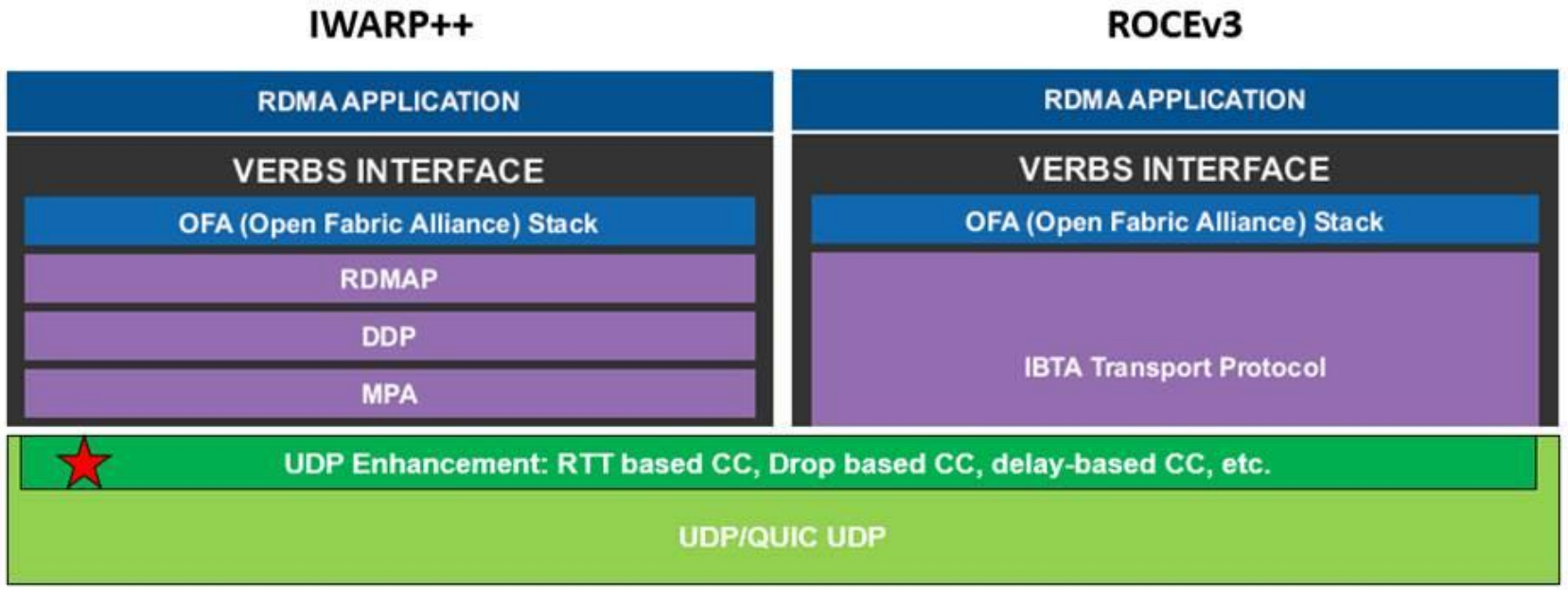
- The term “hyperscale” refers to a computer architecture’s ability to scale in order to respond to increasing demand.
- Goals
  - Common cloud scale infrastructure
  - Dynamic and automated provisioning
  - Diverse workload mix
  - Low latency, high throughput
- Suggestions have been made to scale RDMA/HPC
  - 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**



Hyperscale HPC/RDMA side meeting

IETF-104, Prague

Monday 3/25/2109 – 10AM-11AM

Title: Hyperscale HPC and RDMA

Traditional High-Performance Computing (HPC) and Remote Data Memory Access (RDMA) networks have been relatively small scale, custom, isolated network clusters involving careful tuning and manual configuration. Recent efforts have allowed RDMA to run over TCP via iWARP and UDP via RoCEv2, however, the networks still often remain isolated and relatively small scale. What would it take to run HPC and RDMA networks at hyperscale on cloud-style infrastructure? Research efforts have focused on addressing gaps in congestion management, scheduling incast traffic and improving the orchestration and manageability of supporting protocols, but standardization efforts appear stalled. What is next and what can be done in the IETF to support running HPC and RDMA at hyperscale?

Organizer/Host:

- Paul Congdon - Tallac/Huawei

Attendees:

- Richard Scheffenegger - NetApp
- David Black - Dell-EMC
- Yolanda Yu - Huawei
- Sowmini Varadhan - Microsoft
- Lars Eggert - NetApp
- Roni Even - Huawei
- Remy Lui - Huawei
- David Fan - Huawei

Notes:

1. The HotRFC advertisement of this meeting was on Sunday night. The slides are here:https://datatracker.ietf.org/meeting/104/materials/slides-104-hotrfc-10-hyperscale-hpc-and-rdma-01
2. RoCE defines the CNP message, but it fails to define how to process and generate it. It is believed that DCQCN is widely deployed. The Sigcomm paper describing DCQCN

is here: https://conferences.sigcomm.org/sigcomm/2015/pdf/papers/p523.pdf. This is the state of the art today.

3. There was another HotRFC talk with related topics by Roni. The slides are here:https://datatracker.ietf.org/meeting/104/materials/slides-104-hotrfc-5-fast-congestion-response-for-data-center-00. The idea is to send CNP with current bitrate information from the switches. There was concern about the validity of sending this information from switches. The proposal in the HotRFC wants to short-cut the CNP if possible and provide more information about the state of congestion in the network instead of just a binary signal from the receiver.

4. The RDMA stack slide from the HotRFC talk was used to kick off the conversation. It was pointed out that RoCEv1 is obsolete. Infiniband is often a specialized fabric that is found primarily in embedded uses (e.g., more than one of Dell EMC's storage arrays uses InfiniBand internally), High Performance Computing and High Frequency Trading. We assumed that RDMA networks are not hyperscalar today; they are not large scale and are not mixed with other traffic. The objective for the discussion is to understand what is preventing us from incorporating RDMA/HPC into hyperscale infrastructure and understand what changes might be required to achieve this; for example, RoCEv3 or DC-UDP transport for RDMA.

5. the DCQCN specification does not have interoperability and this currently it isn't seen as a huge issue because this is only used in back-end networks in homogenous environments. The customers aren't overly concerned about mixing and matching vendor components today.

6. NVMe-oF could likely change the private back-end issue. One configuration of NVMe-oF runs NVMe over RoCEv2 and the configuration will be mixed in with all sorts of other cloud traffic in a larger, public back-end. The single fabric concept is false with NVMe-oF over TCP and likely to be with NVMe-oF over RoCEv2.

7. It was discussed that the Go-Back-N for recovery in RoCE is not optimal, but it is being replaced by vendors in their latest solutions. iWarp has a problem with slow-start. It was clarified that iWarp doesn't say anything about slow-start and just uses TCP, leaving it up to the vendors. Improvements with TCP will be leveraged to iWarp. It was suggested that if the success if iWarp was only the fact that it is not using a modern congestion controller, this could be easily remedied.

8. Is the objective of this meeting to come up with a better transport for verbs? The intent is to identify what is preventing us from running RDMA workloads at very high scale.

9. A key difference between iWARP and RoCE is that RoCE/DCQCN has some reliable L2 requirements and iWARP does not require this, making no assumptions about the L2 network. RoCE makes assumptions about the L2 network. It was agreed that creating a large-scale lossless network is difficult and network admins hate PFC.

10. One of the criticisms of iWarp is suffering from TCP slow-start. Most agreed that the real issue is that iWarp is just relying on standardized congestion control and not something specific for the DCN. There is currently work going making TCP more applicable to the DCN. It appears that RoCE is doing the same – re- inventing SACK and

focusing on issues in the DCN, but currently there is little interoperability amongst vendors.
11. Is DCQCN doing a 'me-too' to DCTCP? There is experience with certain NICs that are limited and were unable to generate CNPs and needed to assert PFC to avoid dropping.
12. The direction in RoCE networks is to rely heavily upon the switch setting ECN so the end-point can generate CNPs and only rely on PFC as a backstop. The goal is to use end-to-end CC before PFC fires. This was agreed, but there is experience in practice that NICs were too limited to realize this view. The need to use PFC from the NIC comes at connection start-up and once things reach a steady state communication runs smoothly. We all agreed that we wouldn't want to write standards to work around deficient implementations.
13. There was a question regarding the scope of work is we are considering. For example, there is work from Sigcomm regarding header clipping (i.e. NDP, another NSDI paper from Cambridge, a cell switched DCN fabric). What is in scope? Since RoCE is an IBTA protocol, we will likely not get their attention.
14. There was belief that some NIC vendors see DCQCN as a value add and have no desire to standardize anything related to this. It is believed that there is no desire to standardized enhanced congestion control by RDMA market leaders. The industry has been able to get away with keeping things proprietary because of the closed, back-end nature of the network. NVMeoF will change this because of the heterogenous nature of the deployments. The serves and storage of NVMeoF will come from different vendors.
15. One problem with iWARP is that there is only one big player in the space, and they have a limited product portfolio. It was pointed out that the IETF could work on iWARP because it is from the IETF, but RoCE is IBTA. It is rumored that DCTCP for iWARP is being pursued by one vendor. This is a positive direction because DCTCP is right for the DCN job. The difference between DCTCP and TCP is that DCTCP wants to empty buffers and TCP wants to fill them. RoCE vs iWARP is like VHS vs Beta – iWARP is better technology but is losing in the market.
16. It was pointed out that we need hardware offloads for performance and considering things like Quic as a transport for RDMA would be difficult because it is hard to implement in hardware. It is too soon to define a Quic-NIC with hardware offload because the spec is still changing. In addition, the multiple streams inside a Quic connection are going to be challenging for hardware.
17. There was a discussion about the security model for RoCE. This is always an after though. IPSec and others need offloads, but the acceleration in the stack often violates security principals in order to allow offloads. AES offload is needed for some storage protocols.
18. Quic wants to use TLS security. NVMeoF over TCP uses TLS 1.2. The NVMe-oF over TCP spec does not (yet) contain the updates to also use TLS 1.3. TLS offload is an issue because of the control plane and data plane need to be handled differently, but this violates security in the stack. It was asked if people actually run security protocols within the data center backends? People tended to agree that, no, they are not doing much of this currently.

19. NVMe over TCP is just pure NVMe over TCP, no RDMA, no queue pairs, etc. This is expected to take advantage of common TCP offloads, but was intended to support software data paths. NVMe does not use RDMA, verbs, queue pairs, so it is a different model. Full TCP offload is highly rare and found in something like iWARP NICs, but accelerations like segmentation, checksum, interrupt coalescing, etc are common. NVMe over TCP takes advantage of these but doesn't do the data placement seen in iWARP and RoCE, so it will run slower, but will take advantage of commodity CPUs.
20. Can DCTCP and DCQCN be mix together and how will it perform? It is rumored that mixing the two will not work well because DCTCP would get starved. Certainly mixing regular TCP with DCQCN could be an issue. The belief is that RoCEv2 vendors are not interested in sharing traffic classes in the network. There is some belief that DCTCP and RoCE+DCQCN would likely work ok together because they both have similar end-to-end CC approaches, but the feeling is that no RoCE vendors are interested in seeing this. Again, NVMEoF will likely change this because they will certainly mix workloads.
21. If slow start is a problem for iWARP, it will be a problem for NVMe over TCP. The TCP stack is not specified by NVMe over TCP, and the expectation is that the TCP stack in Linux will just be leveraged. CPU utilization will be the tradeoff for NVMe over TCP, but it is assumed this is acceptable for large DCN operators. It may be a better financial tradeoff to get heterogenous interoperability from software implementations than pay for homogenous hardware-based network infrastructure and solutions just for RDMA.
22. There is a catch to just using CPU cycles for storage and messaging transport; you can't rent or sell those CPU cycles to tenants. So, there is an advantage for hyperscalars to use customer hardware and accelerators, but it must scale and fit into their model. NOTE: AWS recently purchased Annapurna to build Nitro, a custom processor for integrated NICs and doing accelerators. So, the hyperscalars can consider custom hardware, but it seems they would prefer to own it and can't be locked into a vendor by it.
23. What about a Quic type of transport for iWARP? Many of the techniques in Quic for congestion control are not specific to Quic, but rather good lessons learned that could be applied. Considering something that is closer to Quic rather than build something entirely new might be preferred. The biggest challenge is that Quic is still evolving. This is another motivation for using software stacks like NVMe over TCP – because things change quicker than hardware can be built.
24. There was a belief that it would be a couple of years away to consider a UDP transport for iWARP. It takes a long time to bake this stuff into hardware. Intel just supported crypto offload, so getting a new transport will take time. It was agreed that the data path needs to be in hardware for RDMA. This is one of the reasons why RoCE moved faster than iWARP. A Quic datapath in hardware will be hard, partly because of the crypto. This will be a smart NIC, which has the same issues as previous smart NICs – cost/complexity. People prefer dumb, fast and cheap NICs. This is what NVMe over TCP is leveraging.
25. What can the IETF do in this space if iWARP has fallen out of favor? We are missing the NIC vendors here to voice what needs to be done. Some felt that there aren't any

real large-scale problems to solve in the RoCE or iWARP congestion control area. There are small tweaks that can be done, but no large-scale problem that warrants a significant standardization effort. Others feel that iWARP is not successful enough to try to improve it. RoCE isn't an IETF protocol and it isn't likely the IBTA will come to the IETF for help. Congestion control for RoCE should be part of the IBTA protocol, and at a minimum the IBTA should work with IETF experts on congestion control. If we want to create a generalized congestion control algorithm for UDP, we should go talk to the rmcat group. This is because rmcat is an example UDP use case, but clearly a different traffic pattern. UDP options could be used to communicate congestion information. Without considering the use case in detail a generalized UDP congestion control is a 'boil the ocean' exercise.

26. RDMA traffic can be somewhat bursty and the state of the network can change quickly as a result. It will be difficult for a congestion controller to adapt, thus some kind of network support is needed. In addition, it is very easy to create an incast problem in the storage world. The current RDMA vendors come from more of a compute cluster background and the traffic patterns are different from storage, so they haven't dealt with incast as much.

27. Is there any work in the IETF to specifically address the incast issue? The NDP work is related, but not an IETF standardization activity. There is a Broadcom solution that is doing something like a cell switched solution using 256-byte cells. This has a link layer support requirement. What is the layer-3 support for incast? Unclear

28. For next steps, it was believed that we can't really do anything with RoCE. One can go into IBTA. Could we have a joint meeting with IBTA to talk about congestion control? If IBTA doesn't want to come to IETF, why would IETF go there? If IBTA doesn't believe they have a problem and doesn't believe IETF can help, there isn't any reason to engage with IBTA. There is some history within IBTA over standardization of DCQCN. It has not yet happened. IBTA is a vendor driven organization.

29. Ideally, customers would demand standardization and action, however, the customers aren't doing this. As discussed, many of them are going off and implementing their own solutions (e.g. AWS with Nitro).

30. It was suggested to revisit this topic every so often because things continue to change – NVMe over Fabrics as an example. It was agreed that NIC vendors and customers would be good additions to a future meeting. Is Montreal too soon for another side meeting? Perhaps, but if we could expand the participation to include end-users and NIC vendors this could make sense.