IETF-105 Side Meeting:

Large Scale Data Center HPC/RDMA

Monday, July 22, 2019

8:30AM – 9:45AM

Attendees (who signed the blue sheet or where recognized):

First	Last	Affiliation
Hirochika	Asai	Preferred Networks / WIDE Project
David	Black	Dell
Randy	Bush	Arrcus
Xavier	de Foy	InterDigital
Jesús	Escudero	Universidad de Castilla-La Mancha
Roni	Even	Huawei
Randy	Haagens	Microsoft
Jianfei	He	Huawei
Russ	Housley	Vigil Security, LLC
Rachel	Huang	Huawei
Georgios	Karagiannis	Huawei Technologies Dusseldorf GmbH
Younghan	Kim	SSU
Kee-Cheon	Kim	
Wangbong	Lee	ETRI
Aini	Li	Huawei
Peng	Liu	China Mobile
Sonum	Mathur	Viasat
David	Melman	Marvell
Jan	Metzke	BSI
Tal	Mizrahi	Huawei
Yoshifumi	Nishda	GE Global Research / Keio Research Institute
Keyur	Patel	Arrcus
Fengwei	Qin	China Mobile
Richard	Scheffenegger	NetApp
Marcus	Sun	Huawei
KJ	Sun	
Sowmini	Varadhan	Microsoft
Stephan	Wenger	Tencent America
Hua Ru	Yang	Huawei
Hyunsik	Yang	IISTRC
Xiang	Yu	Huawei
Shuai	Zhao	Tencent

Yan	Zhuang	Huawei
Ning	Zong	Huawei

Details:

1. The meeting organizer, Paul Congdon, presented the IETF Note Well reminder and bashed the agenda. No changes to the agenda, but it was suggested to include additional technical approaches being pursued in the IETF such as LSVR.
2. The slide material presented at the meeting is available at: https://mentor.ieee.org/802.1/dcn/19/1-19-0061-01-ICne-ietf-sidemeeting.pdf
3. The slides from IETF-105 HotRFC that announced this side meeting are available at: https://datatracker.ietf.org/meeting/105/materials/slides-105-hotrfc-7-strategies-to-drastically-improve-congestion-control-in-high-performance-data-centers-next-steps-for-rdma-00
4. Jesús Escudero presented strategies to drastically improve congestion control in high performance DC and the next steps for RDMA. The slides are included in the above link. It was suggested to have an interactive discussion about the proposed solutions in the final slide – consider their feasibility and applicability to scaling RDMA and HPC networks.
5. David Black provided some background on RDMA protocols running over IP. He clarified that in practice, RDMA networks often use PFC, but in principle, it is not required for all transports. David provided the history of iWARP and RoCEv2 and the industry momentum behind RoCEv2. He indicated that iWARP is, perhaps, superior, especially when it comes to congestion control, but the industry has adopted RoCE. RoCEv2 is not an IETF protocol.
6. Randy Bush asked if we are aware of NDP, the best paper in Sigcomm 2017. Randy pointed out that IETF has a long history of doing such work as well as the research community and it would be nice to see work in this area. It was asked if there was any sensitivity to addressing such transports in the IETF
7. David Black asked if there were any NIC vendors in the audience. There was one in the crowd and one on the phone. The data rates in the data center require hardware offload support.
8. In the slides, it was suggested that perhaps a new UDP based transport for data centers would be interesting to consider. One of the key requirements would be that it would need to be hardware offload-able. The slides allude to a current trend to run more applications over UDP for low latency and high efficiency.
9. Keyur Patel points out that congestion occurs in switches and that there are already many switch vendors implementing proprietary extensions to address this. He would like to see a problem definition because people are currently building solutions but wonders what could be done beyond what the switch vendors are currently doing. Paul Congdon points out that we are a standards community and defining interoperable solutions is whole the point. Proprietary vendor solutions are not standards based nor necessarily interoperable between vendors.
10. The topic of being able to identify the type of congestion (e.g. incast verses in-network) could be valuable. The congestion mitigation approach might vary based upon where we are in the topology and what type of congestion the switch is experiencing. Knowing your position in the topology can be configured, but more

configuration is not desirable and could be error prone. An automated protocol to identify the location of a switch in the overall topology could be valuable.

11. One of the suggested improvements is to provide more congestion information within the headers. Jeff pointed out the use of overlay networks and their rich and extensible header can help. This is being used more and more now.
12. Roni Even points out that RoCE is driven by Mellanox. They were invited to attend this side meeting but did not attend. Roni felt the lack of attendance and participation is political and that Mellanox prefers to contribute to IBTA, which is a closed working group where they have control.
13. Yan Zhuang presented the material for two drafts: https://tools.ietf.org/html/draft-zhh-tsvwg-open-architecture-00 and https://tools.ietf.org/html/draft-yueven-tsvwg-dccm-requirements-00.html. The slides are included in the complete side meeting slide deck referenced above.
14. David Black asked if any switch in the network can directly communicate with a NIC? The answer is yes, in the proposed architecture, the switch will send messages back to the source. The increasing use of overlays creates a challenge here. RDMA doesn't currently do much with overlays, however, that may change. The headers the switch sees may not be the headers of the source/desk of the RDMA traffic. The messages sent back from the switch to the NIC need to have enough data for the source NIC to be able to unwrap and decode the overlay.
15. There was a question about the drafts and what is next for them. David Black, as TSVWG chair, expressed that the drafts are being presented in the side meeting to better understand their content and interest. Roni Even is asking where the appropriate place is to progress these drafts; iccrg or tsvwg.
16. Paul Congdon asked how the transport agnostic congestion signaling works in conjunction with a TCP based transport that also has congestion signals? Since the switch will be signaling directly to the NIC, it will be the responsibility of the source to combine and mix the signals appropriately.
17. There are some simulation results that show the effectiveness of the proposals in the drafts. David Black suggests the results should be shared with ICCRG because it has congestion control expertise and TSV looks to ICCRG for this guidance.
18. Paul Congdon closed the meeting with an observation that we have more people attending this side meeting, in part due to the favorable scheduling at the IETF. There is a request to create an IETF mailing list for this work. POST MEETING NOTE: the email list is created and is called rdma-cc-interest@ietf.org. The listserv sign-up and info is available at: https://www.ietf.org/mailman/listinfo/rdma-cc-interest
19. Jeff pointed out that NVME over RDMA solutions are coming and the NIC will no longer be the bottleneck, putting more pressure and congestion on the network. David Black, as author of NVME over Fabrics using TCP, points out that NVME over TCP (without RDMA) is new and generating a lot of interest.
20. Paul Congdon asked if there are other applications, besides RDMA, in the data center that might benefit from low-latency, high-throughput. Perhaps other control traffic (e.g. simple server-to-server REST APIs). Roni Even points out that we should look for other applications, but it will all depend on the tradeoffs and requirements.
21. The side meeting was adjourned approximately at 9:40AM