IETF Sidemeeting: Large Scale Data Center HPC/RDMA

Paul Congdon (Tallac Networks)

IETF Note Well

https://www.ietf.org/about/note-well/

This is a reminder of IETF policies in effect on various topics such as patents or code of conduct. It is only meant to point you in the right direction. Exceptions may apply. The IETF's patent policy and the definition of an IETF "contribution" and "participation" are set forth in BCP 79; please read it carefully.

As a reminder:
By participating in the IETF, you agree to follow IETF processes and policies.

If you are aware that any IETF contribution is covered by patents or patent applications that are owned or controlled by you or your sponsor, you must disclose that fact, or not participate in the discussion.

As a participant in or attendee to any IETF activity you acknowledge that written, audio, video, and photographic records of meetings may be made public.

Personal information that you provide to IETF will be handled in accordance with the IETF Privacy Statement.

As a participant or attendee, you agree to work respectfully with other participants; please contact the ombudsteam (https://www.ietf.org/contact/ombudsteam/) if you have questions or concerns about this.

Definitive information is in the documents listed below and other IETF BCPs. For advice, please talk to WG chairs or ADs:

- BCP 9 (Internet Standards Process)
- BCP 25 (Working Group processes)
- BCP 25 (Anti-Harassment Procedures)
- BCP 54 (Code of Conduct)
- BCP 78 (Copyright)
- BCP 79 (Patents, Participation)
- https://www.ietf.org/privacy-policy/ (Privacy Policy)

Join us for further discussion

- Side Meeting: Monday 8:30AM – 9:45AM – Notre Dame
  - NOTE on side meetings:
    - Open to all
    - Meeting minutes will be publicly posted
    - Not under NDA of any form
  - Remote participation is available:
    - https://zoom.us/j/294652109
    - Dial by your location
      - +1 669 900 6833 US (San Jose)
      - +1 646 876 9923 US (New York)
    - Meeting ID: 294 652 109
    - Find your local number: https://zoom.us/u/aeo5yUZXgm

Agenda

- Welcome – Paul Congdon – 5 mins
- Strategies to drastically improve congestion control in high performance data centers: next steps for RDMA - Jesus Escudero Sahuquillo (presenter) – 15 mins
- Discussion – 15 mins
- An Open Congestion Control Architecture with network cooperation for RDMA fabric - Yan Zhuang (presenter) – 15 mins
- Discussion – 15 mins
- Next steps – 10 mins

Strategies to drastically improve congestion control in high performance data centers: next steps for RDMA

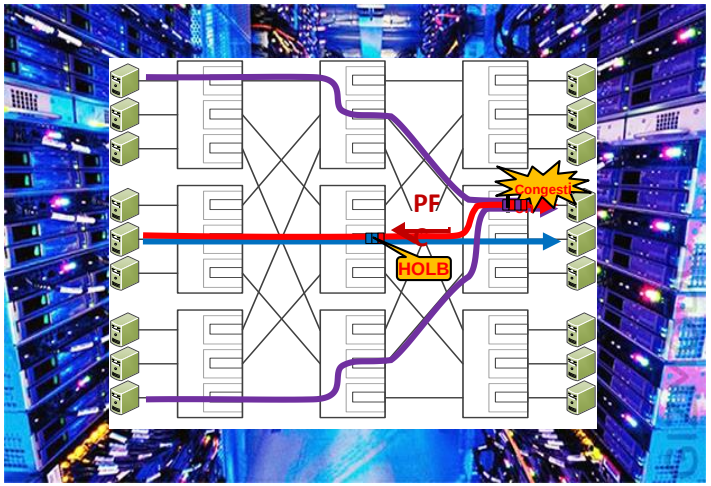
Paul Congdon (Tallac Networks), Jesus Escudero Sahuquillo (UCLM), Pedro Javier García (UCLM), Francisco J. Alfaro (UCLM), Francisco J. Quiles (UCLM)
and Jose Duato (UPV)

Motivation

Data center congestion is unique

The Internet

The High-Performance Data Centers



Data centers have…

- A much different bandwidth-delay product
- Different DCN switch implementations and buffer configurations from Routers
- More homogeneity with the network design and topology
- A high concentration of high-speed links, compute and storage
- Different traffic profiles with a higher degree of correlation
- Fewer management domains (typically a single management)

Congestion in the DCN environment is different than in the Internet

Motivation

Congestion in Datacenter Networks (DCNs)

- **Datacenter Use Cases** (OLDI services, Deep Learning, NVMeoF and Cloudification [Congdon18]), require convergent networks.
- **RDMA** for higher throughput and lower latency.
  - **Lossless or low loss**: Priority Flow Control (PFC).
- **Large DCNs** connecting thousands of server nodes:
  - Efficient topologies (rich path diversity and reduced diameter).
  - Efficient routing algorithms (load and path balancing).
- **Congestion dramatically threatens DCNs performance**, due to its negative effects: **HOL blocking**.

Motivation

Mitigating DCN Congestion [Garcia05][Garcia19]

- Congestion in the data center is dynamic (i.e. the congestion root can move)
- Roots of congestion can occur anywhere in the fabric (front, middle, back)
- There are two types of congestion depending on where the root is:
  - in-network
  - Incast
- Multiple roots can exist

Traditional solution	Strategy	Pros	Cons
ECMP Load-balancing	Avoid congestion by spreading flow on multiple paths	Exists and is easy	• Not congestion aware • Not flow-type aware • **Doesn't help incast congestion**
ECN	Adjust traffic injection by reacting to congestion signals from the network	Exists and is easy	• **Long reaction time in DCNs** • Limited information from the switch • Un(not-well)defined for non-TCP use
ECN + PFC (lossless)	Eliminate packet loss by signaling back pressure	Exists	• **Congestion spreading** → HoL blocking • **Hard to configure and tune**

Motivation

DCNs need low-latency, low-overhead, high-throughput and high-efficiency

In-common with the Internet is the trend to run more things over UDP…

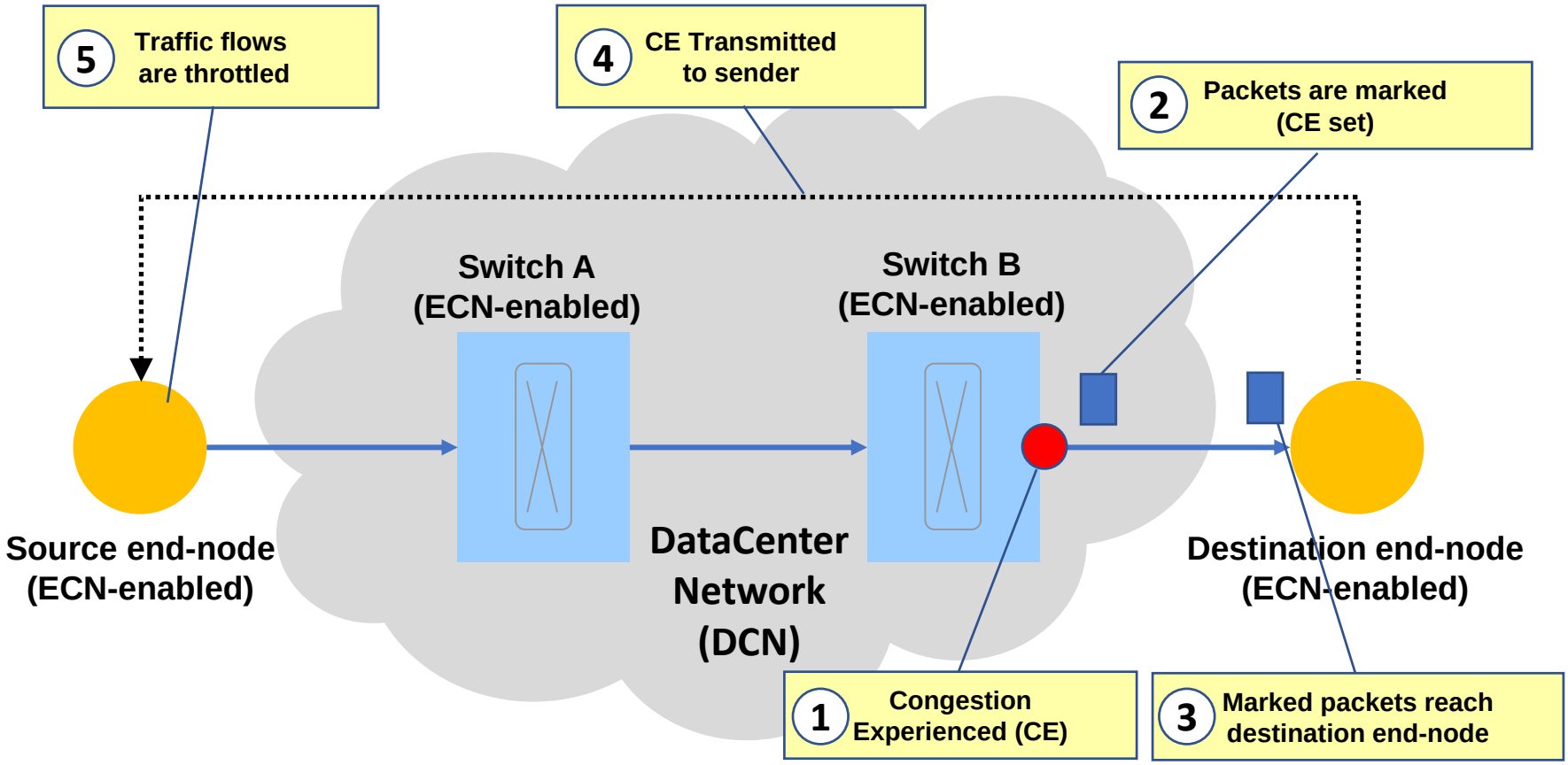
Would we benefit from some Quic-like (Quic-lite) data center transport with some DCCP-like congestion layer for the DCN?

- Hardware offload-able (less emphasis on security and threading).
- Common congestion control targeting unique DCN congestion.
- In-DC-Network visibility, marking and signaling from switches.

…Leverage the IETF's expertise and not leave congestion control design to the applications

Problems with current CC

Explicit Congestion Notification (ECN) [RFC 3168]



Problems with current CC

Explicit Congestion Notification (ECN) [RFC 3169]

We identify the following **problems**:

- **Packets marking** is based on a queue occupancy threshold that triggers the congestion detection.
- **Long notification delays** between packets marking and the actual injection throttling.
- **Injection throttling** may be based on obsolete information due to congestion dynamics and long notification delays.
- **ECN does not directly approach HoL blocking:**
  - HoL blocking actually happening while congestion trees are throttled.

How can we improve it?

Augmenting ECN to enable Data Center focused UDP based congestion control:

- By providing **more detailed feedback from the switches and packet headers.**
- By **distinguishing in-network from incast congestion.**
- By **speeding up notifications.**
- By implementing **fast-response mechanisms in the switches.**

Some ideas to consider

open for discussion

- **More detailed feedback**
  - Switches indicate more details on congestion status.
  - Record accumulated packet delay in the packet headers and include this information in the notifications
- **Distinguish in-network from incast congestion**
  - Understand switch position in topology
  - Identify when congestion root appears
- **Speeding up congestion notifications**
  - Notifications directly from switches backwards to other switches and end-nodes.
- **Fast-response congestion mechanisms at switches**
  - Congestion Isolation (in progress – P802.1Qcz)

References

[Congdon18] Paul Congdon et al: **The Lossless Network for Data Centers**. NENDICA "Network Enhancements for the Next Decade" Industry Connections Activity, IEEE Standards Association, 2018.

[Garcia05] P. J. Garcia, J. Flich, J. Duato, I. Johnson, F. J. Quiles, and F. Naven, "**Dynamic Evolution of Congestion Trees: Analysis and Impact on Switch Architecture**," in High Performance Embedded Architectures and Compilers, ser. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, Nov. 2005, pp. 266–285.

[Garcia19] Pedro Javier Garcia, Jesus Escudero-Sahuquillo, Francisco J. Quiles and Jose Duato, "**Congestion Management for Ethernet-based Lossless DataCenter Networks**" DCN: 1-19-0012-00-Icne.

[Karol87] M. J. Karol, M. G. Hluchyj, S. P. Morgan, "Input versus output queuing on a space-division packet switch", *IEEE Trans. Commun.*, vol. COM-35, no. 12, pp. 1347-1356, Dec. 1987.

[RFC 3168] K. Ramakrishnan et al. **The Addition of Explicit Congestion Notification (ECN) to IP**. RFC 3168, Year 2001**:** https://tools.ietf.org/html/rfc3168.

[Congdon19Qcz] Paul Congdon: P802.1Qcz – Congestion Isolation. **Standard for Local and Metropolitan Area Networks — Bridges and Bridged Networks — Amendment: Congestion Isolation.** PAR approved 27 Sep 2018.

[Escudero11] Jesús Escudero-Sahuquillo, Ernst Gunnar Gran, Pedro Javier García, Jose Flich, Tor Skeie, Olav Lysne, Francisco J. Quiles, José Duato: **Combining Congested-Flow Isolation and Injection Throttling in HPC Interconnection Networks**. ICPP 2011: 662-672.

[Rocher17] Jose Rocher-Gonzalez, Jesús Escudero-Sahuquillo, Pedro Javier García, Francisco J. Quiles: **On the Impact of Routing Algorithms in the Effectiveness of Queuing Schemes in High-Performance Interconnection Networks**. Hot Interconnects 2017: 65-72.

[Escudero19] Jesús Escudero-Sahuquillo, Pedro Javier García, Francisco J. Quiles, José Duato: **P802.1Qcz interworking with other data center technologies**. IEEE 802.1 Plenary Meeting, San Diego, CA, USA July 8, 2018 (cz-escudero-sahuquillo-ci-internetworking-0718-v1.pdf)

An Open Congestion Control Architecture with network cooperation for RDMA fabric

draft-zhh-tsvwg-open-architecture-00
draft-yueven-tsvwg-dccm-requirements-00

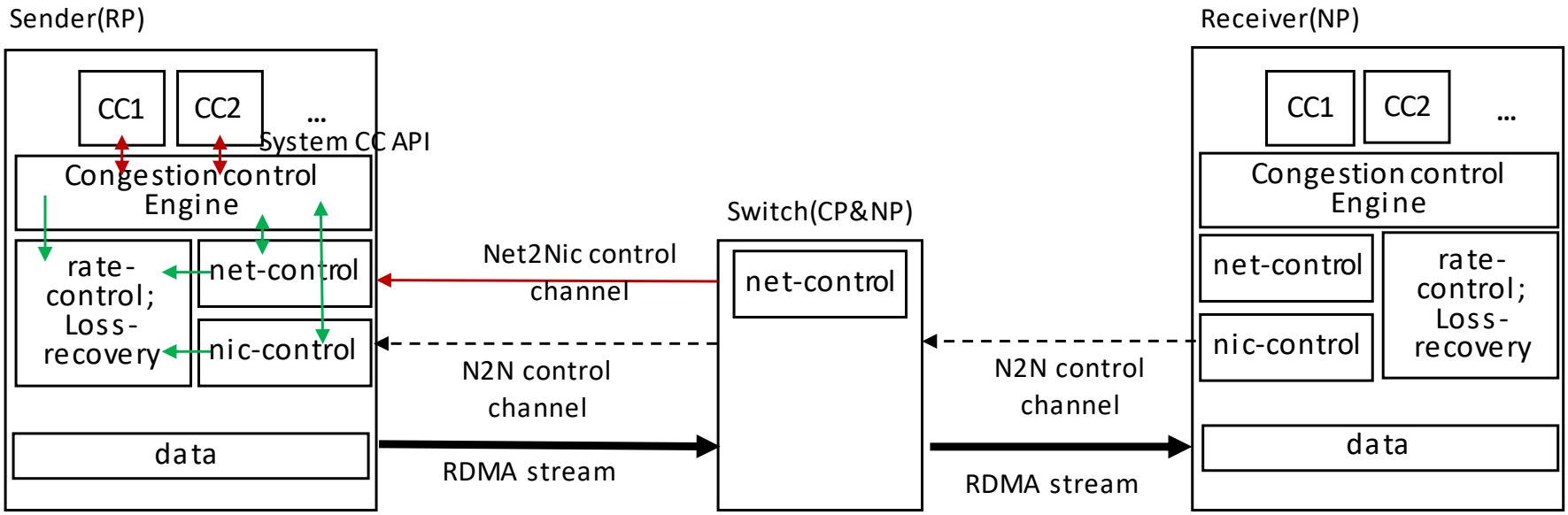
IETF 105, Montreal, Canada

Yan Zhuang (presenter),Rachel Huang
Yu Xiang, Roni Even
Huawei Technologies

An open congestion control architecture with network cooperation for RDMA fabric

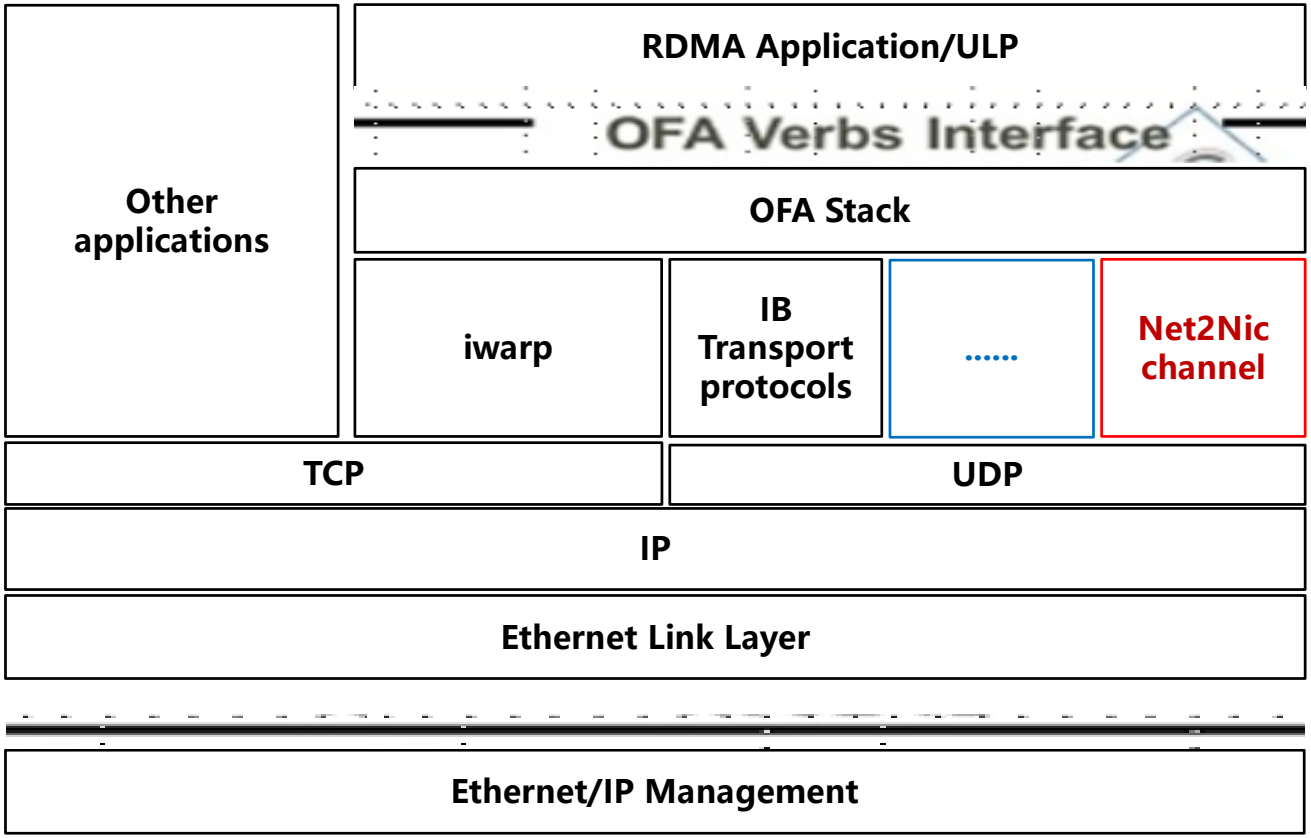
- **Scope**
  - Managed datacenter networks
  - RDMA traffics for applications, such as HPC and storage….requiring low latency, high throughput…
- **Motivation, requirements and use cases**
  - Incast traffic cause severe congestion in the data center network.
  - Mixture of RDMA traffic and TCP traffics effects each other.
  - More efficient and effective congestion controls are needed to support the scalability and high performance.
- **Objectives**
  - Define an open congestion architecture with network cooperation to enable more effective congestion controls for RDMA fabrics.

Open Architecture Overview



- Open to network cooperation
- Open to congestion control algorithms deployment and management

Protocol Stack Overview



Solution should be RDMA transport agnostic.

Open for Network Cooperation

- **What?**
  - Net-control module inside network nodes (e.g. switches) can signal back to senders' NIC directly, and further incorporated into NICs' transmit control.
- **Why?**
  - **Fast Convergence**: reduce the CC feedback/control time.
  - **Accurate congestion awareness:** as congestion point, network aware of the degree of the ongoing and expected congestion and can requests for proper moderation of the selected flows.
- **How?**
  - A Net2Nic control channel can be used to collect congestion information from the network nodes to be further incorporated to the congestion control of sender NICs.

Open for Congestion control deployment and management

- **What?**
  - Deploy/manage congestion control algorithms in a common way regardless of the detailed hardware implementation.
- **Why?**
  - **More flexibility:** Traffic patterns may differ in CC choices.
  - **Easy to deployment in HW**: New CC algorithms are suggested to be implemented in hardware easily.
- **How?**
  - A system CC interface is provided to the operators to deploy CCs through a common platform and then be mapped to local actions/functions.
  - Local functions related to congestion controls can be implemented as function blocks (in hardware) and interact with each other through internal interfaces to achieve the final congestion controls.

Next Step

- Solicit more feedbacks/comments/interests on this open architecture.