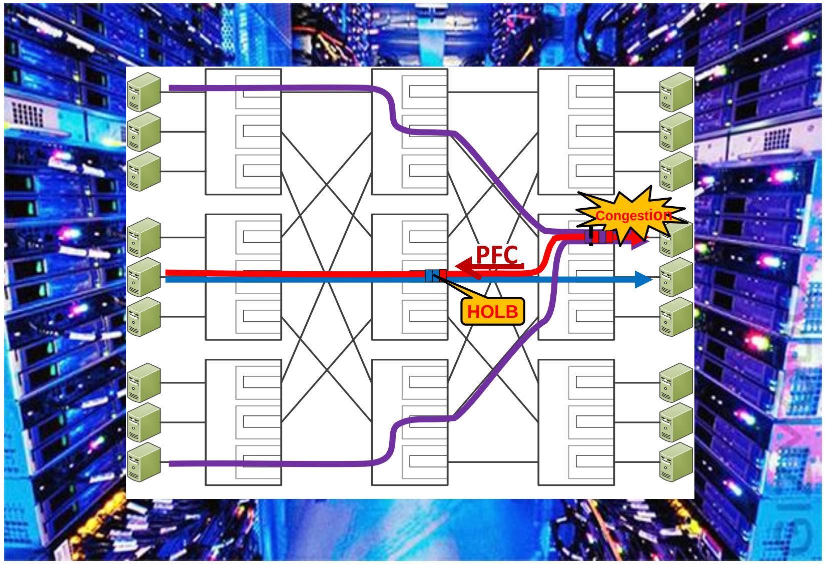
Strategies to drastically improve congestion control in high performance data centers: next steps for RDMA

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

Data center congestion is unique

The Internet

The High-Performance Data Centers



Data centers have…

- A much different bandwidth-delay product
- Different switch implementations and buffer configurations than Internet 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

DCN needs low-latency, low-overhead, high-efficiency, high-throughput

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

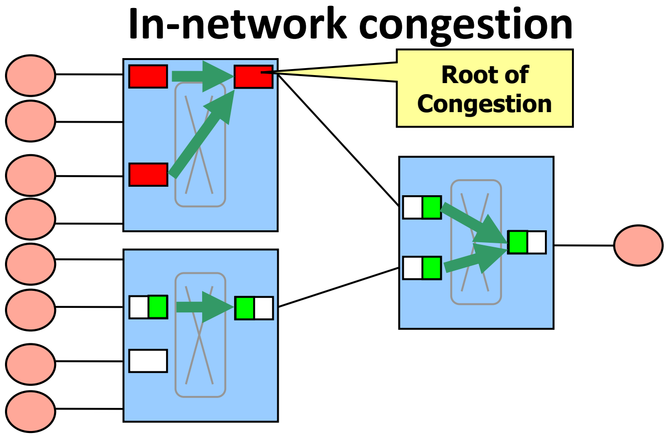
Would we benefit from some Quic-like (Quic-lite) data center transport with some DCCP-like congestion layer for the DCN?

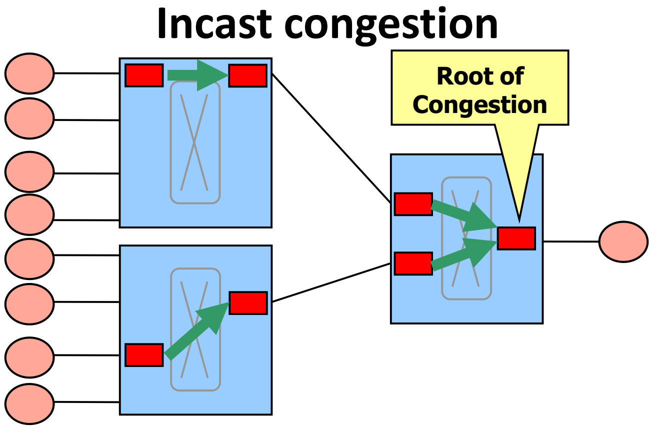
- Hardware offload-able (less emphasis on security and threading)
- Common congestion control targeting unique DCN congestion
- In-DC-Network visibility, marking and signaling from switches

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

Data Center Congestion and Current Solutions

Congestion tree dynamics [Garcia05][Garcia19]





Current solution	Pros	Cons
ECMP Load-balancing	• Exists and is easy	• Not congestion aware • Not flow-type aware • Doesn’t help incast congestion
ECN	• Exists and is easy	• Long reaction time in DCN • Limited information from the switch • Un(not-well)defined for non-TCP use
ECN + PFC (lossless)	• Exists	• Congestion spreading • Hard to configure and tune

Ideas to improve current situation

Augment 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.**

Let’s discuss technical approach and feasibility of these improvements…

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

- Request to start a non-wg IETF mailing list

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)