COC: Hierarchical Coflow Ordering for WAN Bandwidth Optimization in Inter-Data Center

Jingxuan Zhang[1,2], Y. Richard Yang[2,3]

[1]Tongji University, [2]Yale University, [3]Peng Cheng Laboratory

Problem

More and more applications involving large-scale data analytics tend to send coflows among inter-data centers. Existing SD-WAN traffic engineering systems [4, 5, 6] cannot handle coflows very well. To enable the SD-WAN controller to optimize performance of tenant-level coflow applications, there are three challenges:

- *Coordination of flow-level schedule:* It is hard for the SD-WAN controller to coordinate the flow-level schedule of each tenant, as the SD-WAN can only control the tenant-level traffic on the end hosts. The flow-level traffic are controlled by tenants themselves.
- *Heterogeneity of coflow schedulers:* Adjusting with all the complexity, the tenant may adopt different flow scheduling algorithms. Thus, it is hard for the SD-WAN controller to estimate the performance of tenants' applications.
- *Oscillation of rate control:* Most of popular coflow schedulers [2, 1] control ordering rather than rates of coflows. The flow-level rate limits may oscillate frequently. If the SD-WAN controller still use rate control for bandwidth allocation, the oscillation will lead to scalability issue for centralized coordination.

Hierarchical Coflow Ordering

- *Tenant-level coflow ordering:*
  - Each tenant can run its own coflow ordering algorithm using We reuse its tenant-level coflow coordinator.
  - Each tenant report its coflow information and ordering to the global coflow coordinator.
- *Coflow ordering composition:*
  - From reported coflow information and ordering, the global coflow coordinator maintains virtual output queues for each tenant.
  - global coflow coordinator composes the virtual output queues by computing tenant ordering for each output queue.
- *Virtual transport layer:*
  - maintains the meta data (tenant-level priority, coflow id, target egress port, etc.) of each tenant flow,
  - and provides callback functions to start the flow transmission.

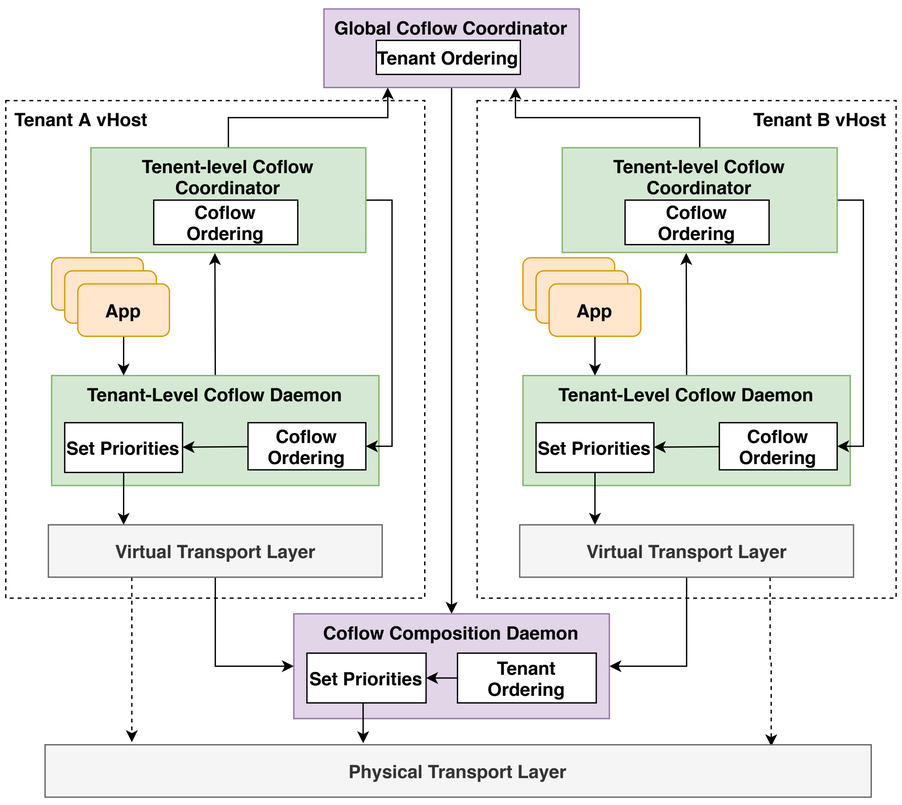

Figure 1: Workflow of inter-data center hierarchical coflow ordering coordination.

Coflow Ordering Composition Algorithm

- At each epoch, for each output queue of each endhost, COC find the tenant with the maximum welfare and let it go first
- The welfare of a tenant is defined as the estimated Coflow Completion Time (CCT) of the tenant multiplying the number of coflows of the tenant.
- Fig. 2 shows the example of coflow ordering composition using the MWTF algorithm.

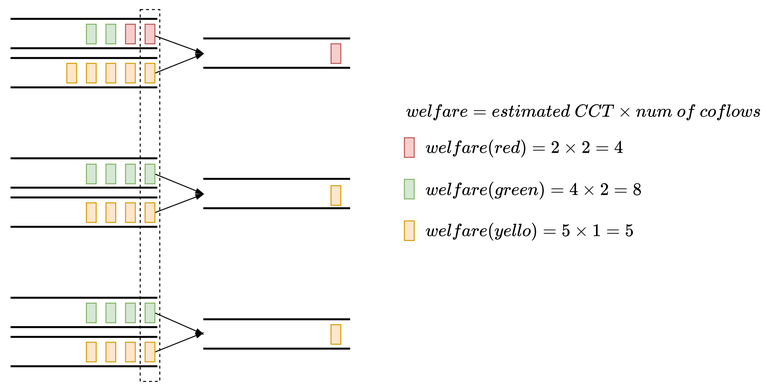

Figure 2: Example of coflow ordering composition using MWTF.

Preliminary Result

- Preliminary result based on a simple inter-DC setting with two DC fabrics connected via a single WAN link.
- 4k coflow traces synthesized from the Facebook workload [3] and randomly separated into three tenants.
- To compare with BwE [5], MWTF reduces 20% the coflow completion time for each tenant on average.

Discussions

- The scalability with increasing numbers of tenants and (co)flows have not been studied yet.
- How to guarantee strategy-proof should also be studied.

References

[1] S. Agarwal, S. Rajakrishnan, A. Narayan, R. Agarwal, D. Shmoys, and A. Vahdat.
Sincronia: near-optimal network design for coflows.
In *Proceedings of SIGCOMM '18*, pages 16–29, New York, New York, USA, 2018. ACM Press.

[2] M. Alizadeh, S. Yang, M. Sharif, S. Katti, N. McKeown, B. Prabhakar, and S. Shenker.
Pfabric: Minimal near-optimal datacenter transport.
SIGCOMM Comput. Commun. Rev., 43(4):435–446, Aug. 2013.

[3] M. Chowdhury, Y. Zhong, and I. Stoica.
Efficient coflow scheduling with varys.
SIGCOMM Comput. Commun. Rev., 44(4):443–454, Aug. 2014.

[4] C.-Y. Hong and et al.
Achieving high utilization with software-driven WAN.
In *Proceedings of SIGCOMM '13*, page 15, New York, New York, USA, 2013. ACM Press.

[5] A. Kumar and et al.
BwE: Flexible, Hierarchical Bandwidth Allocation for WAN Distributed Computing.
In *Proceedings of SIGCOMM '15*, volume 45, pages 1–14, New York, New York, USA, 2015. ACM Press.

[6] B. Schlinker and et al.
Engineering Egress with Edge Fabric: Steering Oceans of Content to the World.
In *Proceedings of SIGCOMM '17*, pages 418–431, New York, New York, USA, 2017. ACM Press.