CONGA Paper Review

By Buting Ma and Taeju Park

Paper Reference

Alizadeh, M. *et al.*, "CONGA: distributed congestion-aware load balancing for datacenters," *Proc. of ACM SIGCOMM '14*, 44(4):503-514, Oct. 2014.

Summary

This paper have introduced CONGA, which is a novel load balancing mechanism for datacenter networks. The CONGA adopts distributed load balancing mechanism to react the network link states as fast as possible and it is implemented in the network instead of transport layer to make deployment easy and to improve the performance of incast scenario. Also, the CONGA determines the path of flowlet by awaring the global link state and the awareness makes the CONGA resilient to the asymmetries such as a link failure. In order to amass the global link states, they add congestion metrics into the encapsulated header of packet and the metric is fed back to the source leaf. In order to evaluate the CONGA, the authors have done experiments extensively with a real hardware testbed, which consists of two racks (64 servers) and four switches in which CONGA is implemented. The experiments use two realistic workloads; enterprise workload and data-mining workload. The experimental results show that flow completion time (FCT) by CONGA is similar to the other scheme such as EMCP and MPTCP for both workload in case of the baseline. On the other hand, in case of the asymmetries, the FCT by CONGA is much better than the other schemes. The improvement is about 30% better than MPTCP 5x better than ECMP. Also, in case of the incast scenario, CONGA provides 2-8x better throughput than MPTCP. They also did large-scale simulations and the result of simulations is similar to result of testbed experiments.

Review

Strength

1. Implementing distributed load balancing in network using overlay encapsulation

Network load balancing intrinsically relies on the current state of the network. Thus, reacting the current network states significantly influence the overall performance. Because of distributed load balancing mechanism of CONGA, it can reflect the global path-congestion much faster than the centralized load balancing. The awesome point of this paper is that CONGA have achieved the distributed load balancing by implementing it in the network by using *leaf-to-leaf* scheme. As a result, CONGA overcomes the drawbacks of the host-based approach such as difficulty in deployment. To implement it in the network, each switch in the network updates the path congestion information by using the encapsulated header of packets. This updated information is fed back to the source leaf. Since this scheme conveys the path-wise congestion by only adding small amount of bits on the encapsulated header of transmitted packets, the overhead for transmitting the global state information is minimized.

2. Small-scale experiments with real hardware testbed, Large-scale simulations.

If they just did the experiments with real hardware testbed, then their contributions are somewhat limited because datacenter usually consists of large amount of servers. To complement their small scale experiment, they also did large scale simulations. The similar results of simulation to the result of the experiments shows that CONGA will outperforms the previous schemes even in the large-scale datacenters. Also, they have simulated various scenarios that is limited in the testbed. This additional simulation shows that CONGA will be useful in those kinds of scenarios.

3. Theoretical proof

Another strength of this paper is that it gives a theoretical proof on outcome. The other papers we read in the same field don’t do this, usually, the last parts of these papers are evaluations and analysis instead of theoretical proofs. The benefits of showing a theoretical proof is that we can compare what we have achieved to best possibility, and it will possibly give us a hint on how to improve the results. However, in this paper, what is given is the upper bound of traffic imbalance, the worst possibility, hence it may not be as valuable as a lower bound further analysis.

Weakness and Extensions

1.Realistic, but restricted experiment and analysis

The author of the paper claims that datacenters use very simple and regular topologies such as Leaf-Spine topology. Based on the claims, they only evaluate CONGA in that topology. However, without convincing all the datacenters use such a simple topology, evaluating CONGA only in the topology is not enough. It means that 3-hops or more-hops topologies have to be considered and CONGA have to be experimented in the complex topologies. With the experiment result, the effect of the number of hops also have to be analyzed.

2.Simple Selection of FB_LBTag

When CONGA conveys the amassed path-wise congestion to the source switch on the reverse way, the FB_LBTag is chosen in round-robin manner. Although the round-robin manner can deliver the congestion metric to source leaf fairly for each path, the round-robin cannot cope with the rapid state change of specific path as fast as possible. If the number of packets who traverse reverse way is small, then reacting the rapid state change is more delayed. Thus, how to select FB_LBTag to address the rapid change of link state based on the amassed link state is possible extension point.

3. Evaluation of responsiveness

In the beginning of this paper, it is argued that CONGA should respond to asymmetries such as a link failure in fast. However, in the evaluation section, although an asymmetric scenario is evaluated, it is evaluated statically. More works could be done to show the response behaviour of CONGA to a dynamic topology, for example, if a link fails, how long it would take for CONGA to re-balance the traffic load, and whether it achieves the goal of "seamlessly handle asymmetry"(with in 10s of microseconds).

Discussion

1. Structure of this paper

In the beginning, the paper presents the tree of engineering decision of CONGA and explained these decisions one by one. This is very helpful and motivating, because in this way, the reader can learn at very beginning what is the main concerns of CONGA. Then the paper details every aspect in the third part Design. Hence the paper looks well organized. Then follows evaluation, which is expected. At the end of this paper, in analysis, a theoretical proof is given, which I think is a good practise of paper on practical technology.

2. Comparison between theoretical and experimental result

At the end of this paper, it defines a new metric called traffic imbalance and a theoretical upper bound is proposed. However, in the evaluation part, this metric in not well reflected. Only in section 5.2.3, a metric called throughput imbalance is measured and presented, while I think these two are the same thing. More work could be done to show the difference between the theoretical upper bound and the experimental-measured value, e.g. show the line of theoretical upper bound in Figure 12. Maybe because this upper bound is too loose that it's meaningless to compare it to reality.

3. Combine with TIMELY

CONGA uses encapsulation to carry additional bits to detect and transfer information about flow congestion. What if use the same method of TIMELY that use RTT to detect congestion? Both of them are designed for Clos topology datacenter network. If RTT is used to detect the congestion, we may not need to encapsulate the packets at leaf switch.

Also, there are something in CONGA that can be adopted into TIMELY. An example would be Discounting Rate Estimater (DRE) algorithm. “It’s essentially a first-order low-pass filter” and is “similar to EWMA”, TIMELY could possibly try this rather than EWMA to see what benefits or drawbacks it will bring.