Analysis of Multipathing Techniques in Data Center Networks

Lyno Henrique G. Ferraz, Otto Carlos M. B. Duarte
Universidade Federal do Rio de Janeiro - GTA/COPPE - Rio de Janeiro, Brazil

***Abstract*—Data center network topologies rely on multiple paths to ensure reliability and performance. Meanwhile, data center applications present various communication patterns with conflicting demands that hamper network operation. To improve data center network utilization, forwarding mechanisms use multipathing techniques to configure the multiple paths and select which path to use. This paper investigates data center multipathing techniques for different communication patterns.**

I. Introduction

Data center offers a large amount of processing and storage resources for applications by aggregating capacity of a dedicated cluster of computers. The applications are distributed over the cluster, thus network plays an important role to ensure applications meet their needs [1]. The network must guarantee both bandwidth capacity and latency to all applications running on data centers, which requires substantial intracluster network capacity.

Data center network topologies improve network capacity using multi-rooted trees to achieve high aggregate bandwidth between hosts with multiples paths. Besides, the cloud computing environment has multiple tenants, each tenant with its own application. Thus, workloads are a priori unknown, so static resource allocation is insufcient, and as the applications run on commodity Operational Systems, the network must deliver high bandwidth without protocol or software changes. Therefore, an efcient forwarding mechanism that takes advantage of multiples paths for data center network is a challenge.

II. Data Center Communication Patterns

Cloud computing data centers run a variety of applications that result in various communication patterns, such as delay-insensitive large flows and delay-sensitive short flows. All flows are critical to data center networks, since the majority of flows are short, while a few large flows transfer most of bytes [2]. The communication patterns have different needs, and data centers fail to address all communication patterns at the same time [3].

Data center network topologies are designed to provide high aggregate bandwidth, path redundancy, and reliability. Topologies with link redundancy provide multiple paths for each pair of hosts that can improve overall transfer rate, but forwarding mechanisms typically use secondary paths for fault tolerance [4]. As delay-insensitive large flows contribute with most of traffic in data centers, forwarding mechanisms should optimize path assignment to achieve high network utilization. Though data center round-trip-times (RTT) are typically low, congestions cause a two order of magnitude RTT variations forming long tail distributions. Responses that take longer than a deadline, because of the long RTT, are usually discarded, which impacts the quality of the responses.

III. Multipathing Techniques

Several algorithms discover multiple paths based on minimum number of hops. Other proposals look for disjoint paths, creating multiple subtrees [1]. The forwarding mechanisms select a different path per flow. Equal Cost MultiPath (ECMP) distributes flows between available paths by using hashing functions. The randomized path selection neither considers network utilization nor flow sizes. Hence, two large flows may share the same path degrading overall switch utilization.

IV. Conclusion and Future Work

We are currently developing a simulator to evaluate the performance of the multipathing techniques for several communication patterns. We expect the results will highlight the main flaws and requirements of data center network. Our final goal is to provide elements to design a multipath forwarding mechanism that achieves high bandwidth and low latency for communication patterns, minimizing costs, and modifications in current data center fabric.

Acknowledgments

This work was supported by CNPq, CAPES, CTIC, FINEP and FAPERJ.

References

[1] M. Bari, R. Boutaba, R. Esteves, L. Granville, M. Podlesny, M. Rabbani, Q. Zhang, and M. Zhani, "Data center network virtualization: A survey," *Communications Surveys Tutorials, IEEE*, vol. 15, no. 2, pp. 909–928, 2013.

[2] T. Benson, A. Akella, and D. A. Maltz, "Network traffic characteristics of data centers in the wild," in *ACM SIGCOMM IMC'10*. ACM, 2010, pp. 267–280.

[3] D. Zats, T. Das, P. Mohan, D. Borthakur, and R. Katz, "Detail: reducing the flow completion time tail in datacenter networks," in *ACM SIGCOMM'12*. New York, NY, USA: ACM, 2012, pp. 139–150.

[4] M. Al-Fares, S. Radhakrishnan, B. Raghavan, N. Huang, and A. Vahdat, "Hedera: dynamic flow scheduling for data center networks," in *USENIX NSDI'10*, 2010, pp. 19–34.