6C-05

A Study on Network Performance for Distributed Storage Systems

Luis Guillen[†1] Satoru Izumi[†1] Toru Abe[†1,†2] Takuo Suganuma[†1,†2] Hiroaki Muraoka[†3]

[†1]Graduate School of Information Sciences, Tohoku University [†2]Cyberscience Center, Tohoku University
[†3]Research Institute of Electrical Communication, Tohoku University

1 Introduction

Distributed Storage Systems (DSS) have gained increasing popularity since they provide highly available services by using replication techniques, being *replicas* data-copies stored in different nodes [2]. However, due to the high volume of data handled and bandwidth limitations, different problems might arise when using these techniques. Current solutions implement traditional network functionality which heavily relies on hardware capabilities and do not take advantage of state-of-the-art network paradigms.

To solve this, we propose the usage of Software Defined Networking (SDN) in DSS to achieve not only a higher overall throughput, but also to have a better control on the distribution mechanism. This paper shows preliminary experiments to confirm the limitations and problems of Link Aggregation (LAG) and briefly discuss SDN-based solutions.

2 Link Aggregation

One of the techniques to improve network performance is Link Aggregation[1], whose basic idea is that multiple physical links are combined into one logical bundle to provide two major benefits:

- Increased bandwidth, traffic can be balanced across the link members, and
- Resilient interconnection, so that the connection can still be active even if one or more link members have problems.

Until 2008, LAG was defined by the IEEE 802.1ad standard and later on superseded by the IEEE 802.1AX-2008 and IEEE802.1AX-2014 respectively. For a link to be member of an Aggregation Group, it needs to be configured with the same settings in terms of speed, duplex configurations, encapsulation type, etc. LAG can be manually or dynamically configured, however, each vendor has their own implementation.

The traffic is handled by using a deterministic Frame Distribution Algorithm (FDA), in which, depending on the vendor implementation, different parameters may be configured. For instance, the source/destination MAC address, IP address, L4 (TCP/UDP) port number, among others.

Based on the FDA each *flow* will be assigned to a member of the LAG, and all traffic for each flow will be placed (if possible) on the same link. However, that also means that traffic for a single flow cannot exceed the bandwidth of a single member, and since each node performs the calculation locally, the traffic for a single flow will not necessarily traverse the same link. Therefore, the overall performance is constrained by different factors to both logical and physical features.

3 Proposal

To solve the problem mentioned above, we propose the usage of a SDN-based network control method to improve the network performance on distributed scenarios, in which, by leveraging the control plane to software instead of vendor-specified logic, the overall performance will improve. The overview of the proposal is depicted in Fig. 1. This method selects a link to communicate with the storage dynamically depending on the network status. By using SDN technologies, we can realize flexible network management and control.

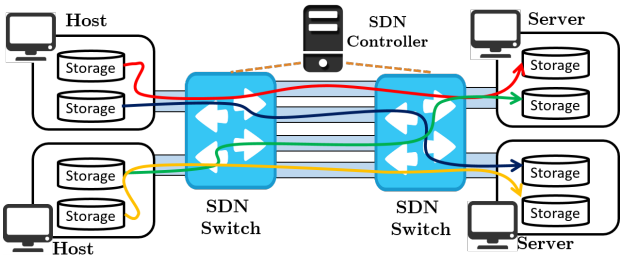

Figure 1: Overview of SDN-based Network Control Method

4 Preliminary Experiment

4.1 Overview

We conducted preliminary experiments to confirm limitations and problems of Link Aggregation and the effect of SDN-based solution using a real testbed. The experiment was conducted using 2 SDN switches as depicted in Fig. 2. Hosts are connected to switches using straight 1 Gbps cables, and switches are linked by (up to 8) crossover 100 Mbps cables.

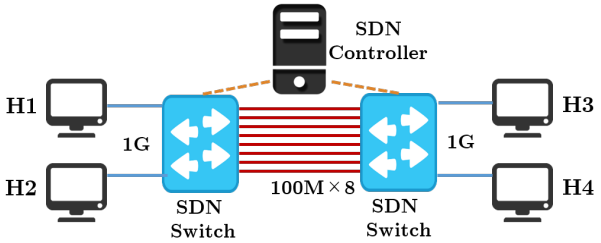

Figure 2: Evaluation Testbed

For the evaluation, traffic from H1 and H3 was generated to H2 and H4. To saturate the line, 25 parallel requests were sent per every enabled link to a TCP port on each of the Servers. This procedure was repeated 50 times and then we monitored the throughput by adding the combined result of each request. The results are described in the following subsections.

4.2 Link Aggregation Evaluation

Based on the recommended configuration for a *many to many topology* in [1], we used as a hash-key the *source and destination MAC Address* as the FDA.

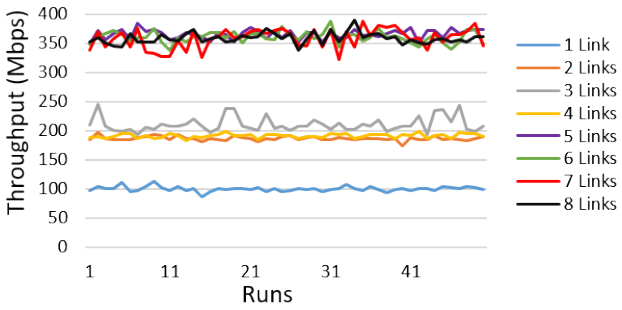

Figure 3: Throughput of each host (Link Aggregation)

As shown in Fig. 3, the throughput is staggered and even if 8 links were available only 4 of them were used, reaching a maximum of around 380 Mbps.

4.3 SDN-based Evaluation

For this preliminary experiment we used a naïve approach, in which, a SDN controller installed two flows on each of the switches based on the ingress port and the TCP port destination so that traffic is balanced among all available links.

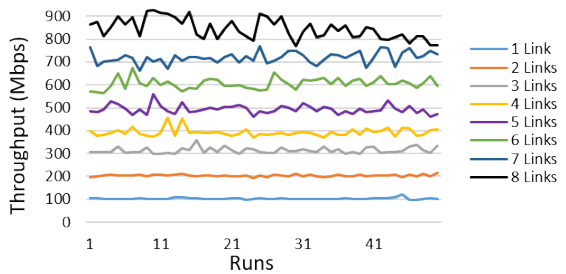

Figure 4: Throughput of each host (SDN-based Solution)

As observed in Fig. 4, by this approach we obtained a steady linear increase on the throughput, reaching a maximum of 850 Mbps. However, since SDN virtually allows full control of network functionality, it leaves room for several improvements.

4.4 Results Comparison

Fig. 5 shows the maximum combined average throughput of both approaches, it is noteworthy to mention that LAG shows a staggered behavior, however, after extensively testing different configurations, even in the best case scenario the performance would be equal or even lower than the naïve version adopted for the SDN-solution.

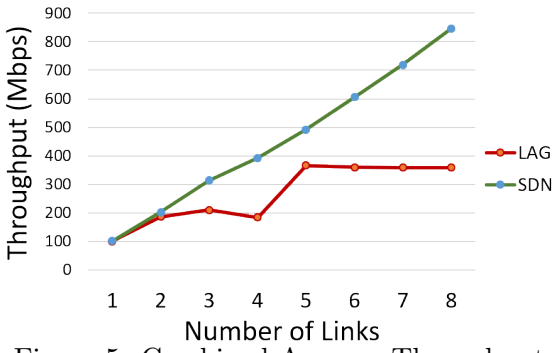

Figure 5: Combined Average Throughput

5 Conclusions

In this paper, we evaluated two solutions: Link Aggregation and a SDN-based solution that might contribute to tackle issues such as the effective data transfer in DSS. Preliminary results show that the SDN-based solution offers more advantages in terms of extensibility, programmability, fail tolerance and high performance with a steady increase of the throughput.

As future work, we design and implement dynamic link selection algorithms in our proposed method.

Acknowledgments: This work is supported as "Research and Development on Highly-functional and Highly-available Information Storage Technology", sponsored by the MEXT Japanese Government.

References

[1] "IEEE Standard for Local and metropolitan area networksLink Aggregation". In: *IEEE Std. 802.1AX-2014* (2014), p. 200.

[2] Ying Liu et al. "ProRenaTa: Proactive and Reactive Tuning to Scale a Distributed Storage System". In: *15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid 2015), pp. 453-464*. 2015.