Understanding the Overhead of Large Layer 2 Data Center Networking: Measurement and Analysis of TRILL as an Example

Chunyang Lu[1], Zhixiong Jiang[1], Jinping Yu[2], Gan Zhang[3], Mingfu Li[2], Wei Liang[2], Jingping Bi[2]
[1]Information Technology Center
China National Petroleum Corporation
Beijing, 102200, China
[2]Institute of Computing Technology, Chinese Academy of Sciences
Beijing, 100190, China
[3]College of Electronics and Information Engineering
Nanjing University of Technology
Nanjing, 210000, China
{luchunyang, jiangzhixiong }@cnpc.com.cn, zgwonder@gmail.com, {yujinping, limingfu, liangwei, jpingbi}@ict.ac.cn

ABSTRACT: *Highly virtualized data center have placed many new and unique requirements on the networking fabric. Conventional network protocols limit the scale, latency, throughput of cloud networks. To solve these problems, the large layer 2 technology is proposed. The large layer 2 technology such as TRILL and SPB has obvious advantages in scalability, convergence and resource utilization compared with the traditional layer 2 technology. However, the disadvantages of employing the large layer 2 technology are barely addressed. To study the overhead of large layer 2 networking, we setup a data center environment with 6 core switches and conduct experiments to compare the performance of the representative large layer 2 protocol—TRILL and the typical layer 2 technology-STP. We also evaluate the intra-VLAN forwarding and inter-VLAN forwarding of TRILL. From our experiments, it is observed that the forwarding efficiency of TRILL is 30.16% lower than STP on average. It is also observed that the inter-VLAN traffic forwarding of TRILL averagely costs about 54% more time and consumes about 5% more CPU than intra-VLAN with a fixed packet generation rate. This is the first study towards unveiling the inefficiency of large layer 2 approach in data center networks as far as we know.*

Keywords: Large Layer 2 Technology, Data Center Optimization, TRILL

Received: 24 May 2013, Revised 19 June 2013, Accepted 26 June 2013

1. Introduction

Cloud Computing is proposed to share computing, storage and network resources between multiple independent tenants from one or multiple data centers.

The data center network fabric has great effect on the performance of cloud services. The network fabric is supposed to provide high-bandwidth, non-blocking and multipath communications within clouds. While computing and storage virtualization technologies have been extensively and widely exploited, there still exists critical inflection in the networking realm. On one hand, typical Layer 2 switching protocol, e.g. Spanning Tree Protocol (STP) [1], is used by classical Ethernet to ensure loop-free communication path. However, STP is notorious in that: (1) It blocks redundant ports and paths by creating a single path tree

and results in the waste of up to half of the aggregated available network bandwidth; (2) STP may lead to suboptimal routing path through the network and increase the end to end latency; (3) It may take extra-long time to recalculate the loop-free path and propagate the changes in the event of link failures. On the other hand, many conventional data center networks are based on typical layer 3 designs, e.g. the Clos network. Unlike the layer 2 STP network, VLAN is used to avoid broadcast loops and multiple load balanced paths are created by equal-cost multi-path routing (ECMP) [2]. While the advantages of typical layer 3 architecture are based on mature technologies, proven approaches as well as commodity equipment, the disadvantages are evident: (1) The native layer 2 domain segregated by VLAN or VLAG sub-network is relatively small, which limits the flexibility of virtual machine migration; (2) The separation of design, configuration and management of individual domain incurs tremendous complexities.

Therefore, new approaches are being proposed to address both the limitations of the layer 3 data center networking based on VLAN and ECMP [2], and the layer 2 design based on STP [1] such as Transparent Interconnection of Lots of Links (TRILL) [3] proposed by IETF, Shortest Path Bridging (SPB) [4] from IEEE 802.1aq standard, and proprietary protocols from vendors, e.g. FabricPath from Cisco [5], QFabric from Juniper [6], and VCS from Brocade [7]. All of these approaches involve some implementation of large layer 2 data center networking, which applies a form of layer 3 routing protocol to construct a uniform large layer 2 network domain at the link layer . There are many potential benefits of large layer 2 data center networking: (1) The introduction of shortest path routing to layer 2 could eliminate inefficient path; (2) It offers the ability to use multipath forwarding to spread traffic out among the available paths and decrease traffic congestion; (3) The use of a link state protocol versus spanning tree's distance vector algorithm decreases network convergence time in case of possible failure; (4) The flattened network fabric enhances the flexibility and scalability of live node migration.

Even though the brilliant feature of the large layer 2 approaches and the vendor-ready status quo from the industry community, it has not yet been widely adopted both when constructing new data centers and upgrading the existing infrastructure. The expensive funds for upgrading infrastructures to support large layer 2 approaches may be a reason. Is large layer 2 good enough for future data center without considering the financial factors? In this study, we would like to unveil the shortcomings of large layer 2 technologies for data center networks through real infrastructure evaluation and quantitatively analysis, in order to inform potential adopters what price to pay by gaining the benefits of large layer 2. We choose TRILL [3] as the representative protocol under evaluation since TRILL has been standardized by IETF and supported by a broad range of vendors. However, we argue that the insights obtained from our study are applicable to similar large layer 2 technologies from other organizations and commercial vendors.

We conduct our study from the following two benchmarks: the forwarding efficiency and the overhead of network switches. We setup experiments to compare TRILL with the typical layer 2 technology-STP, according to the above benchmarks. Furthermore, we conduct experiments to evaluate TRILL both in intra-VLAN traffic forwarding and inter-VLAN traffic forwarding. By analyzing the results of these experiments, we will shed light on the overhead of TRILL. In addition, we will show some advices for the deployment of TRILL in data centers. This is the first preliminary study towards unveiling the inefficiencies of large layer 2 approach in data center networks as far as we know.

2. Related Works

Large layer 2 technology has being developed fast. Currently, the mainstream large layer 2 technologies are: TRILL [3], a standardized protocol proposed by IETF, FabricPath [5] proposed by Cisco, and SPB (Shortest Path Bridging) [4] proposed by IEEE. Each of these protocols is supported by many equipment manufacturers. We will describe them in detail in this section.

2.1 Large Layer 2 Technology

2.1.1 TRILL

TRILL is short for transparent interconnection of lots of links. TRILL is proposed to solve the problems caused by traditional Ethernet networks, such as the convergence problem of STP when there are frequent topology changes, and the bandwidth waste problem because of the single flow tree. With better performance, TRILL converges fast even when topology changes frequently. What's more, TRILL applies multi-path forwarding protocols which can take full use of bandwidth resources. The TRILL switches are usually called RBridges (Routing Bridges) [10]. TRILL forwards traffic flow based on link-state routing. Therefore, TRILL modifies the packet structure and adds two extra header to the Ethernet frames-the TRILL header which records the information of ingress RBridge and egress RBridge, and the outmost Ethernet header or the next-hop header which

records the information of the next-hop RBridge.

2.1.2 FabricPath

FabricPath can be regarded as the enhanced version of TRILL. It applies similar implementation mechanism as TRILL, so we don't dwell on it in this paper.

2.1.3 SPB

SPB (Shortest Path Bridging) is another large layer 2 technology. SPB forwards traffic flow based on IS-IS. However, unlike TRILL, SPB doesn't modify the Ethernet frames. It applies two forwarding strategy-SPBV (Shortest Path Bridging VID) and SPBM (Shortest Path Bridging MAC) which both reuse current Ethernet technology. SPBV is based on 802.1 ad/Q-in-Q [8] while SPBM is based on 802.1 ah [9].

2.2 Application of Large Layer 2 Technology

There have been a lot of efforts to improve TRILL or other large layer 2 technologies. The work described in [13] proposes a method to encode multi topology within TRILL data frames.

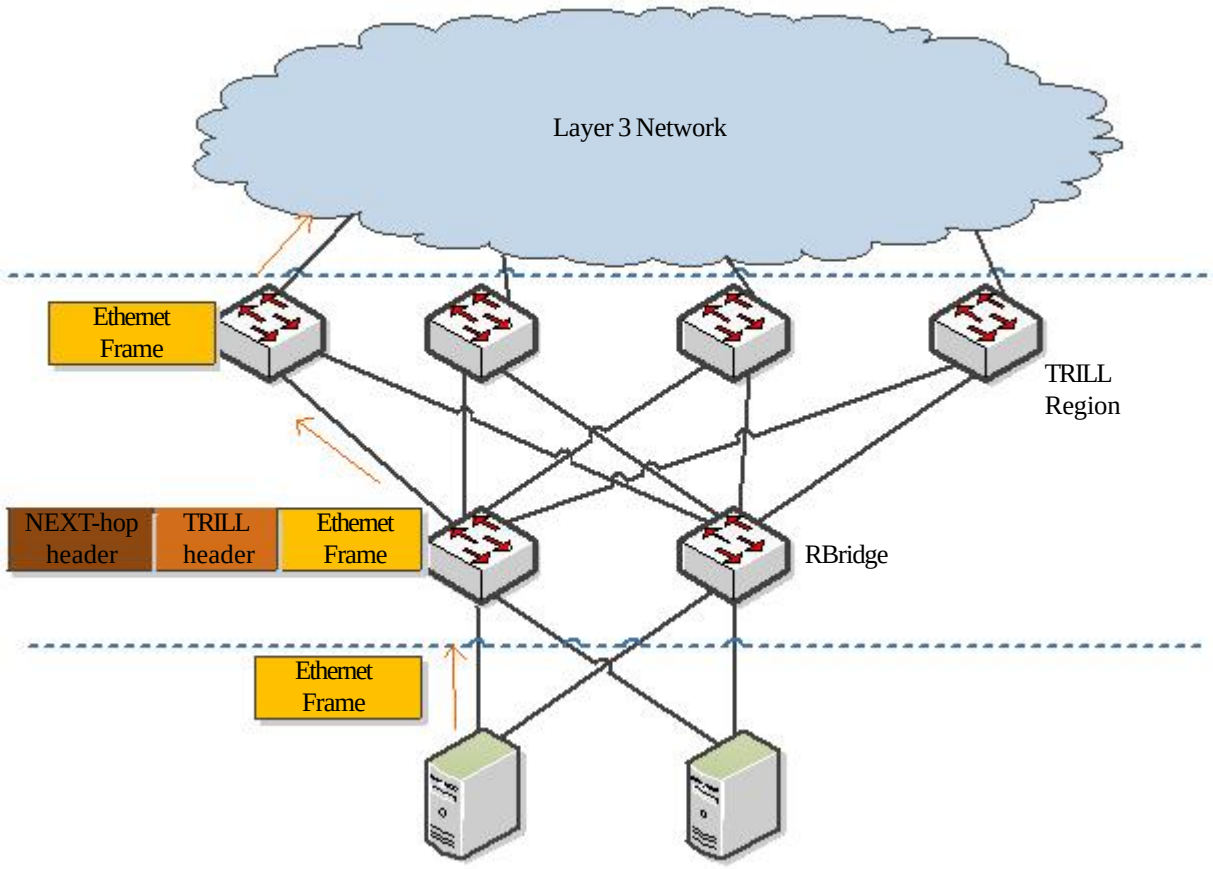

Figure 1. A typical TRILL deployment

Reference [14] describes how to use BGP (Border Gateway Protocol) to connect disparate TRILL-based data centers or TRILL-based networks. However, there is no work to address the disadvantage of this technology in practical application as far as we know.

3. Preliminaries

We will discuss the basic measurement setup, goals and metrics in our experiments in this section.

3.1 TRILL Deployment

We setup a typical TRILL deployment environment which is shown in Figure 1. All switches in the TRILL area are RBridges which implement TRILL and run IS-IS protocol. Note that there may be ordinary switches which don't support TRILL between servers and RBridges, but we only describe the crucial part of TRILL in Figure 1 for simplicity.

As described in [10], TRILL adds two new headers to original Ethernet frames-the TRILL header and the next-hop header. TRILL needs to query several tables to find the next-hop RBridge and modify the next-hop header. Therefore, the switches need to consume extra resource to support TRILL and consequently cause the overhead to the forwarding efficiency. We measure the performance of TRILL with two metrics: the forwarding efficiency and the overhead of switches.

3.2 Basic Measurement Setup

In order to effectively evaluate TRILL, we conduct experiments both to compare TRILL with STP and to measure the performance of TRILL itself in different environments such as inter-VLAN and intra-VLAN. We establish a data center with 6 core switches as RBridges and the Spirent Test Center to simulate more than 1,000 servers. The measurement setup is shown in Figure 2. The switches in the shaded region are the H3C S10500 series [15] switches which support both STP and TRILL. The bandwidth of

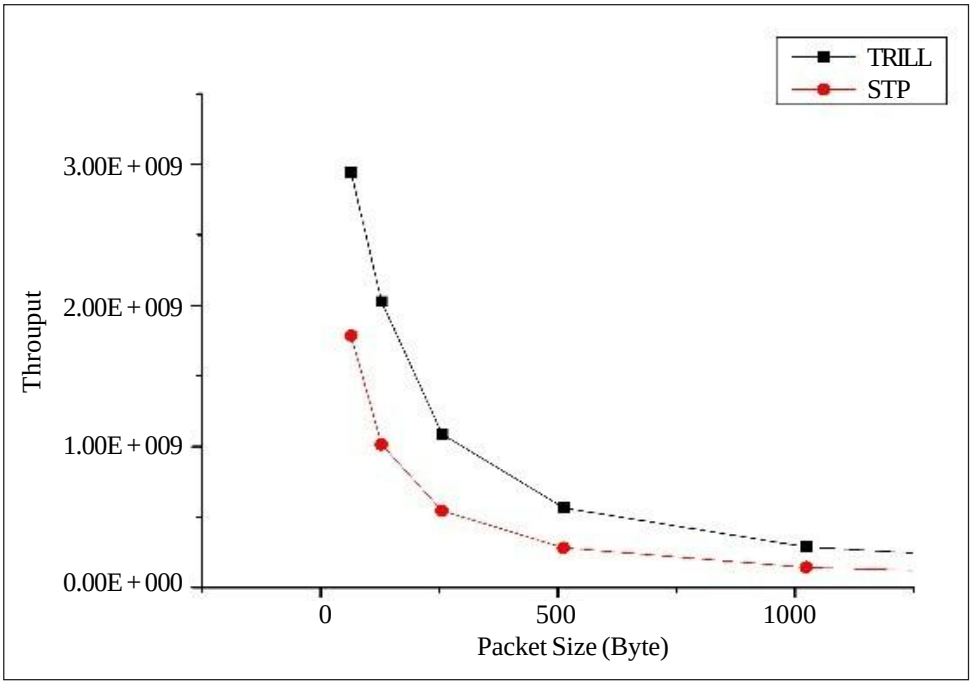

Figure 2. Throughput of TRILL and STP in 30 seconds with 8Gbps packet generation rate

Packet size	Latency (TRILL/STP) unit: μs				
	8 *Gbps*	**16 *Gbps***	**24 *Gbps***	**32 *Gbps***	**40 *Gbps***
64 *bytes*	2.76/2.13	2.88/—	2.86/—	2.86/—	2.82/—
128 *bytes*	2.87/2.25	2.83/—	2.89/—	2.85/—	2.81/—
256 *bytes*	3.10/2.36	3.12/—	3.01/—	3.13/—	3.16/—
512 *bytes*	3.58/2.45	3.57/—	3.52/—	3.57/—	3.57/—
1024 *bytes*	4.57/3.68	4.52/—	4.54/—	4.54/—	4.55/—
1518 *bytes*	5.51/4.51	5.53/—	5.55/—	5.66/—	5.68/—

Table 1. Latency in TRILL/STP Environment

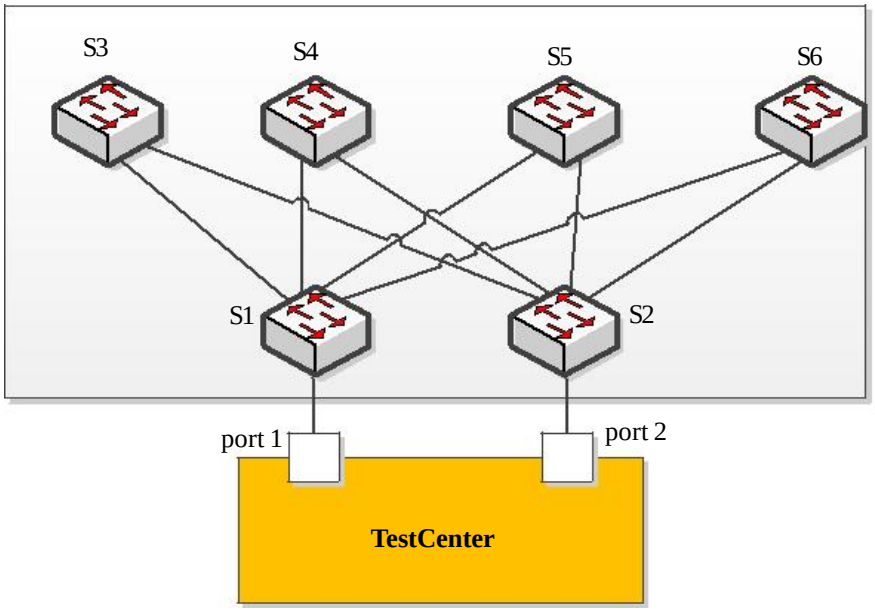

Figure 3. Measurement setup architecture

links connecting the test center and *S*1/*S*2 is 40Gbps while the capacity of links connecting the switches are all 10Gbps. We measure the latency and throughput of packet forwarding since they are the basic indicators to evaluate the forwarding efficiency. We also measure the CPU utilization of the ingress and egress RBridges because these switches need to encapsulate and decapsulate the data frames. These metrics are especially important because they can reveal the overhead of the networking equipment itself as well as the cloud applications in TRILL environment.

4. Performance Measurement I: TRILL vs. STP

In order to compare the forwarding efficiency of TRILL with STP, we setup experiments in unicast environment in the architecture shown in Figure 2.

4.1 Experiment Setup

All the switches are in the same VLAN. We enable TRILL and STP respectively in the shaded area to measure the forwarding efficiency. In each experiments, the test center simulates servers and sends packets from *port*1 to *port*2 in a fixed rate. There are two paramters: the packet size and the packet generation rate. The packet size is one of the impact factors of latency. And the packet generation rate decides how long to send a packet and if it is larger than the link capacity, the packet loss will happen. For example, STP will lose packets when packet generation rate is larger than 10Gbps in the environment showed by Figure 2. We don't show the loss rate, because it is apparently that the packet loss rate of STP will be higher than TRILL, since STP will block nearly half bi-section bandwidth while TRILL can utilize all possible path. Packets may choose one of the four paths: *S*1–>*S*3/*S*4/*S*5/*S*6–>*S*2. For each packet, we record the time it leaves *port*1 and the time it is received totally by *port*2. In addition, we also record the number of packets that *port*2 received. Therefore, the latency is computed bysr $T_r - T_s$ where T_s is the time that *port*1 starts sending the first byte of packet and T_r is the time that *port*2 finishes receiving the last byte of packet. And the throughput is the number of packets that *port*2 receives in unit time.

4.2 Results & Analysis

The latency for different sizes of packets in different packet generation rates is shown in Table 1. Because the capacity of links in the shaded region in Figure 2 is 10Gbps, STP will lose most packets when the packet generation rate excesses 10Gbps. Therefore, it is impossible to compute the latency under these situations. The results in Table 1 show that the forwarding efficiency of TRILL is averagely 30.16% lower than STP when the packet generation rate is 8Gbps. Therefore, using TRILL is not good for services which are sensitive to the response time, for example financial business applications.

The possible reason why TRILL is less efficient for forwarding packet is that it needs to encapsulate the TRILL header and the next-hop header in the ingress RBridge and decapsulate these headers in the egress RBridge. The extra encapsulation and decapsulation work obviously increase the latency of a frame.

Figure 3 shows that the throughput of TRILL and STP in 30 seconds with 8Gbps packet generation rate. The figure shows that the throughput of TRILL is larger than STP. But as the size of packets grows, the throughput of TRILL and STP becomes close. It is argued that TRILL doesn't perform much better than STP for big packets. Therefore, for elephant flows, TRILL may cause congestion and lose packets just as STP does. Even though TRILL is better than STP in many aspects, it is still not good enough for every kind of services running in data centers. We argue the possible reason why the throughput of TRILL and STP becomes close as the packet size grows is as follows: equal-cost multipath (ECMP) [2] has been used as the de facto routing algorithm in TRILL. In ECMP, flows (as identified by the TCP 5-tuple) between a given pair of servers are routed through one of the path using hashing. However, because not all flows are identical in their size (or their duration), this simple scheme is not sufficient to prevent the possible congestion in the network.

Packet size	**Latency (Intra-VLAN/inter-VLAN) unit: µs**					**Average ratio (%)**
	8 *Gbps*	**16 *Gbps***	**24 *Gbps***	**32 *Gbps***	**40 *Gbps***	
64 *Bytes*	2.79/4.56	2.81/4.63	2.82/4.65	2.83/4.61	2.84/4.64	63.9
128 *Bytes*	2.89/4.71	2.88/4.70	2.85/4.67	2.85/4.69	2.88/4.70	63.6
256 *Bytes*	3.11/4.97	3.12/4.97	3.09/4.98	3.10/4.97	3.11/4.99	60.2
512 *Bytes*	3.55/5.53	3.54/5.52	3.57/5.53	3.57/5.54	3.59/5.55	55.3
1024 *Bytes*	4.52/6.64	4.54/6.66	4.54/6.66	4.55/6.64	4.57/6.67	46.4
1518 *Bytes*	5.53/7.66	5.57/7.67	5.58/7.69	5.60/7.71	5.61/7.72	37.9
Average ratio (%)	54.6	54.6	54.9	54.4	54.2	—

Table 2. Latency in Intra-VLAN/INTER-VLAN Network

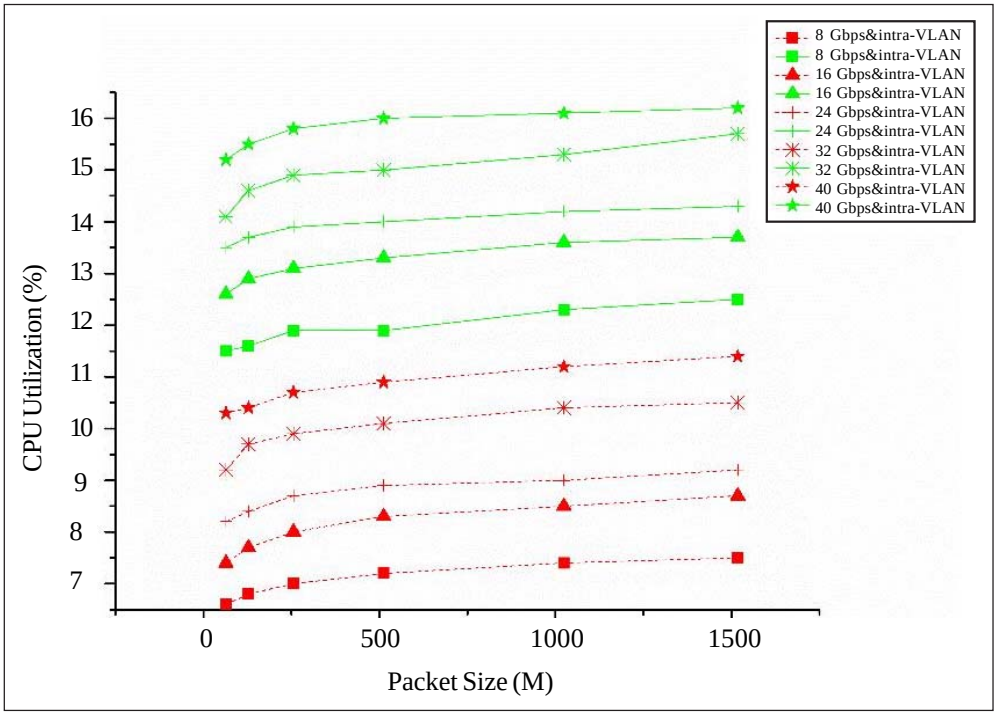

Figure 4. CPU utilization of edge switches for different packet generation rate both inter-VLAN and intra-VLAN

5. Performance Measurement II: INTRA-VLAN vs. INTER-VLAN

In order to compare the performance of TRILL in intra-VLAN scenario with inter-VLAN scenario, we also measure TRILL in the test environment shown in Figure 2.

5.1 Experiment Setup

In this experiment, we focus on the cost that packets travel across VLAN in TRILL area. The switches in the shaded area are all running TRILL. In each round of experiments, the test center simulates servers and sending packets from *port*1 to *port*2 in a fixed rate. We setup two experimental scenarios. One is that *S*1~*S*6 are in the same VLAN. The other is that the switches are divided into two VLANs. One contains *S*1, *S*3, *S*5 while the other contains *S*2, *S*4, *S*6. There are also two parameters as defined in the previous section: the packet size and the packet generation rate. We record the time each packet spends on traveling from *port*1 to *port*2. Because it will consume some CPU when packets travel across VLANs, we also record the CPU utilization of edge switches under both inter-VLAN scenario and intra-VLAN scenario. Therefore, the two metrics are the latency and CPU utilization of edge switches.

5.2 Result & Analysis

The latency in both inter-VLAN and intra-VLAN scenarios are shown in Table II. Table II shows the latency in different packet sizes and different packet generation rates for both intra-VLAN and inter-VLAN packet forwarding. The average ratio is computed as follows:

- Average ratio for each column:

$$\sum_{p = 64 \sim 1518} \left(\frac{L_{inter}}{L_{intra}} - 1 \right) \Big/ 6$$

- Average ratio for each column:

$$\sum_{p = 8 \sim 40} \left(\frac{L_{inter}}{L_{intra}} - 1 \right) \Big/ 5$$

where p denotes the packet size, r denotes the packet generation rate, L_{inter} and L_{intra} denote the latency of inter-VLAN packet forwarding and intra-VLAN packet forwarding respectively.

Results in Table 2 show that the latency in inter-VLAN scenario is much larger than that in intra-VLAN scenario. When the packet generation rate is fixed and the packet size is variable, inter-VLAN packet forwarding spends 54.54% more time on average than intra-VLAN packet forwarding. When the packet size is fixed and the packet generation rate is variable, inter-VLAN packet forwarding averagely spends 54.55% more time than intra-VLAN packet forwarding. This means when users access to inter-VLAN services, the response time may be very large.

The CPU utilization of edge switches in both inter-VLAN scenario and intra-VLAN scenario is shown in Figure 4. The red dash lines describe the intra-VLAN CPU utilization of edge switches in different packet generation rate. And the green solid lines describe the inter-VLAN CPU utilization of edge switches in different packet generation rate. The lines with the same symbol type are CPU utilization of inter-VLAN scenario and intra-VLAN scenario with the same packet generation rate. Results show that inter-VLAN packet forwarding consumes about 5% more CPU than intra-VLAN packet forwarding with a fixed packet generation rate. It is acceptable when there are not many services in data centers. However, if the load of data center is enormous, the increased 5% CPU of inter-VLAN packet forwarding may affect the performance of services running in the data center.

The reason why inter-VLAN packet forwarding spends more time and consumes more CPU is because when it comes to inter-VLAN forwarding, TRILL has to travel across the layer 3 routing gateway. It naturally increases the time spending on forwarding.

6. Discussion

TRILL offers distinct advantages to the development of future data center networks. However, TRILL also has some disadvantages which will affect the performance of the services and devices in data centers. The insights obtained from our measurements are

as follows:

• According to our experiments, the forwarding efficiency of TRILL is averagely 30.16% lower than STP. Especially, TRILL will spend about 54% more time when forwarding packets across VLAN compared with intra-VLAN forwarding.

• Under the same condition, the throughput of TRILL is close to STP for big frames. It means that when elephant flows happen, TRILL may cause congestion and packet loss as STP, which seriously affects the performance of data centers.

• TRILL consumes about 5% more CPU resources for inter-VLAN forwarding than intra-VLAN forwarding, which will affect the performance of current service and other services in data centers.

In order to take advantages of TRILL, we propose the following advises for TRILL:

• Because the TRILL header already contains the information of source and destination RBridges, and the TRILL area is typically segregated from traditional Ethernet network, the outmost next-hop header can be eliminated to improve the forwarding efficiency.
• The multi-path mechanism can base on sub-flow level like MPTCP [11] and FLARE [12] to alleviate the congestion caused by elephant flows.

• Administers should try to design large-scale VLAN to reduce inter-VLAN communication.

7. Conclusion

The large layer 2 network is the trend of future data center establishment. Even though it has many advantages compared with traditional Ethernet network, it also has its own shortages. From our study, we find that TRILL has the following shortages: 1) It is less efficient for packet forwarding than STP in unicast network; 2) When it comes to elephant flow, TRILL still have the risk of congestion and packet loss; 3) TRILL is inefficient for inter-VLAN communication. To alleviate the effect of these shortcomings, we suggest that TRILL need to be improved and potential adopters should design large-scale VLAN when they deploy TRILL in their data centers.

8. Future Works

In our future works, we will measure the performance of TRILL in multicast environment. In addition, we will analyze the effect that TRILL has on the upper applications such as web services and MapReduce services in depth.

References

[1] STP: http://en.wikipedia.org/wiki/Spanning_Tree_Protocol.

[2] Thaler, D., Hopps, C. (2000). Multipath Issues in Unicast and Multicast Next-Hop Selection, RFC2991, November.

[3] Touch, J., Perlman, R. (2009). Transparent Interconnection of Lots of Links (TRILL): Problem and Applicability Statement RFC5556, May.

[4] Seaman, M. (2005). Shortest Path Bridging.

[5] FabricPath: http://www.cisco.com/en/US/netsol/ns1151/index.html.

[6] QFabric: http://www.juniper.net/us/en/dm/datacenter/.

[7] VCS: http://www.brocade.com/solutions-technology/technology/vcs-technology/index.page.

[8] Q-in-Q: http://www.cisco.com/en/US/docs/ios/lanswitch/configuration/guide/lsw_ieee_802.1q.html.

[9] 802.1ah: http://www.ieee802.org/1/pages/802.1ah.html.

[10] Perlman, R., Eastlake, D. (2011). Routing Bridges (RBridges): Base Protocol Specification, RFC6325, July .

[11] Ford, A., Raiciu, C., Handley, M. (2009). Tcp extensions for multipath operation with multiple addresses, Internet-draft, IETF, Oct.

[12] Sinha, S., Kandula, S., Katabi, D. (2004). Harnessing TCPs Burstiness using Flowlet Switching, in 3rd ACM SIGCOMM Workshop on Hot Topics in Networks (HotNets), San Diego, CA, November.

[13] Senevirathne, T., Banerjee, A., Aldrin, S., Nimmu, N. (2012). Multi Topology Encoding within TRILL Data Frames, June.

[14] Perlman, R., Bhikkaji, B., Venkataswami, B. V., Mahadevan, R., Sundaram, S., Swamy, N. P. Connecting Disparate TRILL-based Data Center/PBB/Campus sites using BGP, February.

[15] H3C 10500 series switches: http://www.h3c.com/portal/Technical_Support___Documents/Technical_Documents/Switches/H3C_S10500_Series_Switches/.