TSVWG
INTERNET-DRAFT
Intended Status: Informational
Expires: April 12, 2019

Y. Li
X. Zhou
Huawei

October 9, 2018

Overlayed Path Segment Forwarding (OPSF) Problem Statement
draft-li-tsvwg-overlayed-path-segment-fwding-ps-00

Abstract

Various overlays are used in networks including WAN, enterprise campus and others. End to end path are divided into multiple segments some of which are overlay encapsulated to achieve better path selection, lower latency and so on. Traditional end-to-end transport layer is not very responding to microburst and non-congestive packet loss caused by the different characteristics of the path segments. With the potential transport enhancement for the existing or purposely created overlayed path segment, end to end throughput can be improved. This document illustrates the problems in some use cases and tries to inspire more about whether and how to solve them by introducing a reliable, efficient and non-intrusive transport forwarding over the overlayed path segment(s).

Table of Contents

1. Introduction

Overlay tunnels are widely deployed for various networks, including long haul WAN interconnection, enterprise wireless access networks, etc. End to end connection are normally broken into multiple path segments for different purposes, for instance, selecting a better overlay path over the WAN or deliver the packets over the heterogenous networks like enterprise access and core networks.

TCP-like transport layer provides end to end flow control and congestion control for path reliability and high throughput. Such an approach has the problems of slow congestion responding and non-congestive loss misinterpretation at the sender and does not achieve the optimal performance in certain cases.

Some of the problems have been well known over years. With new technologies are emerging like NFV (Network Function Virtualization) and various flexible overlay protocols, forwarding over the specific overlayed path segment(s) can be considered to be enhanced by providing a reliable and non-intrusive transport to improve the throughput to solve those problems.

This document illustrates the problems in some use cases and tries to inspire more about whether and how to solve them by introducing a reliable, efficient and non-intrusive transport forwarding for the overlayed path segment(s).

2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

OPSF: Overlayed Path Segment Forwarding

3. Use Cases and Problems

The following subsections presents use cases from different scenarios using overlay tunnels with a common need of higher performance and reliable overlayed path segment in best effort networks.

3.1 Microburst in Long Haul Network

Internet is a huge sized network of networks. The interconnections of end devices using this global network are normally provided by ISPs (Internet Service Provider). This ISP provided huge network is considered as traditional Internet. CSPs (Cloud Service Provider) are connecting their data centers using Internet or self-constructed networks/links. This expands Internet's infrastructure and together with the original ISP's infrastructure, forms the Internet underlay.

NFV further makes it easier to dynamically provision a new virtual node as a work load in a cloud for CPU/storage intensive functions. With the aid of various mechanisms such as kernel bypassing and Virtual IO, forwarding based on virtual node is becoming more and more effective. The interconnections among the purposely positioned virtual nodes and/or the existing nodes with vitalization functions potentially form "the Second Plane" or overlay of Internet. It is called the Cloud-Internet Overlay Network (CION) in this document.

CION makes use of overlay technologies to direct the traffic going to the specific path regardless the underlying physical topology to achieve better service delivery. Figure 1 shows an emerging multi-segment overlay over large geographic distances. It purposely creates or selects overlay nodes (ON) from clouds/Internet. Segment here is the virtual hop between two ONs. By directing the traffic to be forwarded along those virtual nodes rather than the default path, better delivery in terms of throughput and delay can be achieved. When a large number of potential virtual nodes are available, there is a high chance that a better path could be found [CRONets].

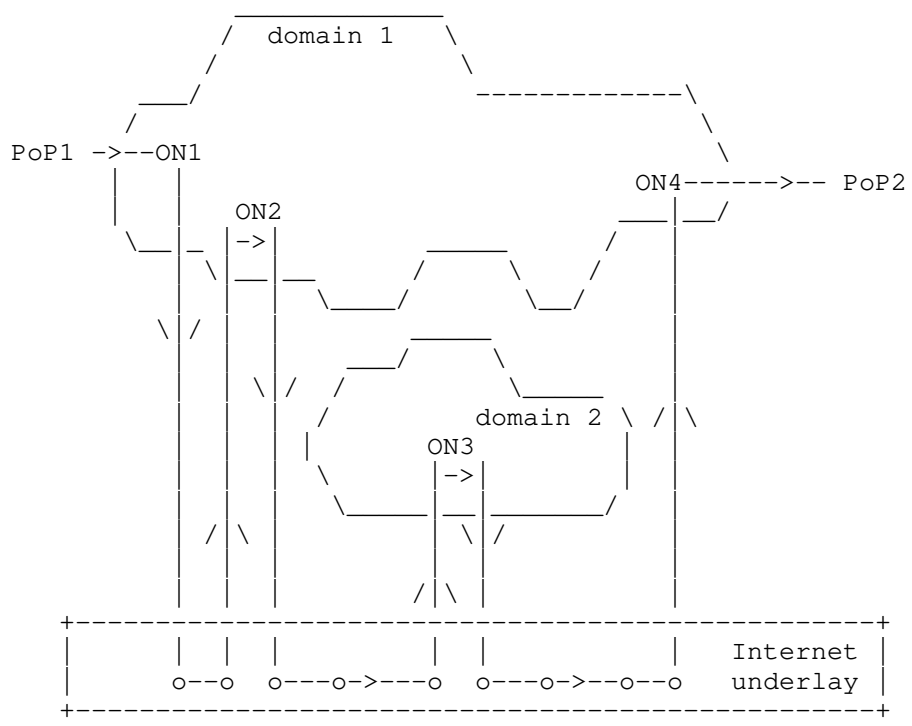

Figure 1. Cloud-Internet Overlay Network (CION)

Microburst is an unexpected data bursts within a very small time window (probably in micro-seconds). Some research shows microbursts happen even for underutilized link [BurstyAna]. The short spikes caused by microburst result in higher jitter and sometimes packet loss in a network. Such loss may trigger the congestion control like reducing the sending rate at the TCP sender as it exhibits the normal pattern of congestion loss in terms of duplicate acknowledgements and/or RTT increases. As microburst is extremely short, the packet loss caused by it is non-persistent and rather random. Therefore it does not necessarily require the sender to reduce its sending rate. Invoking the congestion control at the sender may unnecessarily make the average sending rate low and degrades the throughput in long haul CION. In addition, long haul transmission may take hundreds of milliseconds. The packet loss response at the sender to microburst over the long haul transmission is not timely. Sender's reaction does not really respond to the current instantaneous path situation.

Overlay nodes in the middle can potentially offer new possibilities,

e.g. retransmission over ONs, to better response to microbursts. Such enhancement can be enabled based on the individual overlayed path segment rather than on the entire end to end path to improve the response time and performance from the packet loss/re-order caused by microburst. Such enhancement should avoid racing with higher layer transport protocols.

3.2 Non-congestive Loss in WiFi Accessed Campus Overlay

Different path segments have different characteristics. The probabilities of packet loss over every and each segments have a large variance. The non-congestive packet loss usually occurs in some specific overlayed path segments. End to end TCP-like transport protocols do not take this factor into careful account. It assumes that packet loss for any reason is almost evenly distributed across the entire path, and adjusts the sender to accommodate the packet loss of the bottleneck segment. This results in non-optimal sending rate in some cases.

Figure 2 shows the WiFi accessed enterprise campus. AP connects to its edge switch normally using Cat5/5e twisted-pair cable which typically provides less than 10G bandwidth. The data packets are tunneled using various overlay mechanisms, like VXLAN [RFC7348], LISP [RFC6830] or CAPWAP [RFC5415]. Two edge switches use another overlay segment over campus core network to deliver the packets which provides more functions like policy enforcement and mobility enhancement. This overlay is usually over fiber which provides higher bandwidth.

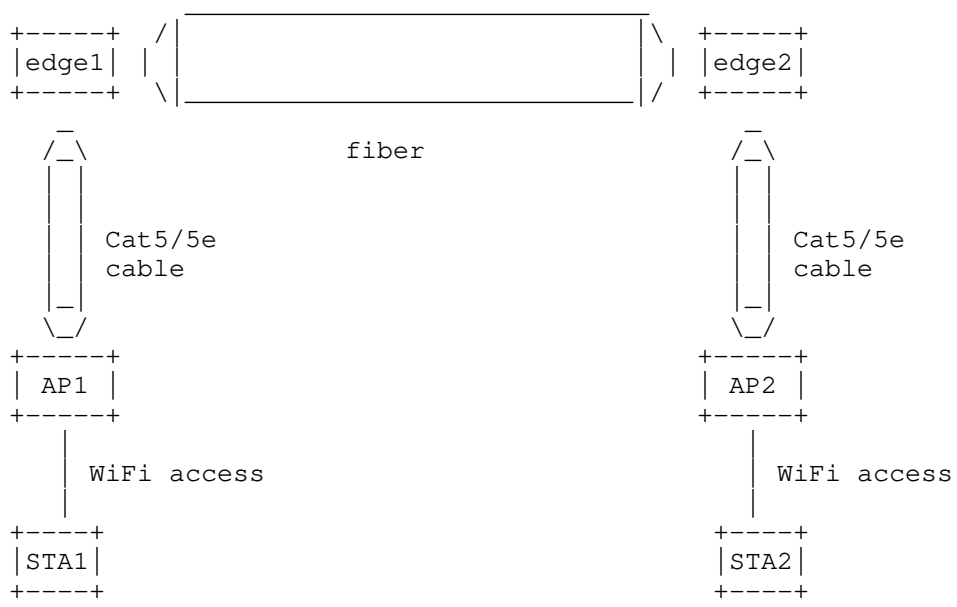

Figure 2. WiFi accessed Campus Overlay Network

Cat5/5e cables, especially UTP (Unshielded twisted pair), are susceptible to distance, interference, and bundling. The environment and the way they are deployed cause drastic changes in random loss rate. The overlay tunnel running over it will have more transmission unreliability than the overlay running on the fiber. Current transport layer is not able to identify such specific problematic segment and simply leaves it for the end to end congestion control to handle it so that the sender may be kept at a lower sending rate and the throughput is not optimal.

In addition to the uplink of the AP, the non-congestive packet loss generated by the wireless access link itself accounts for the largest proportion in the end-to-end path. Wifi access is affected by fades, interference, attenuation and corruption. Non-congestive loss is common. Its link layer has mechanisms to do the packet recovery. However the number of local link layer retransmission is usually based on the empirical value or the static configuration. When the value is not properly chosen, the TCP sender can be unnecessarily exposed by the random packet loss and reduce the sending rate. It is hard to make the link layer frame recovery work in concert with the current end to end transport layer.

3.3 Higher Reliability and Low Latency for Interactive Application

Mobile gaming and VoIP like application normally can not tolerate a retransmission even over a path segment. When two divergent overlay segments are available like shown in figure 3 for path from ON1 to ON2, purposely duplicating packets over two segments provides more reliability. Two disjoint segments can usually be obtained by measuring to find segments with very low mathematical correlation in latency change. When the number of overlay nodes is large, it is easy to find such disjoint segments. Random node or memory failure may also benefit from the duplicating packets over disjoint segments.

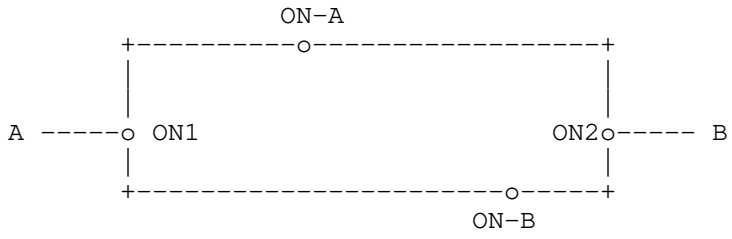

Figure 3. Multiple Overlayed Path Segments for Higher Reliability

4. Features to be Considered for OPSF (Overlayed Path Segment Forwarding)

The diagram shown in Figure 4 illustrates a typical scenario with an overlayed path segment. Transport layer provide the end to end flow control between two end host. When an overlayed path segment exists or is purposely created between two overlay nodes, an enhanced forwarding over that segment can potentially solve some problems of end to end transport performance issues and at the same time provides more reliability and flexibility to traffic path.

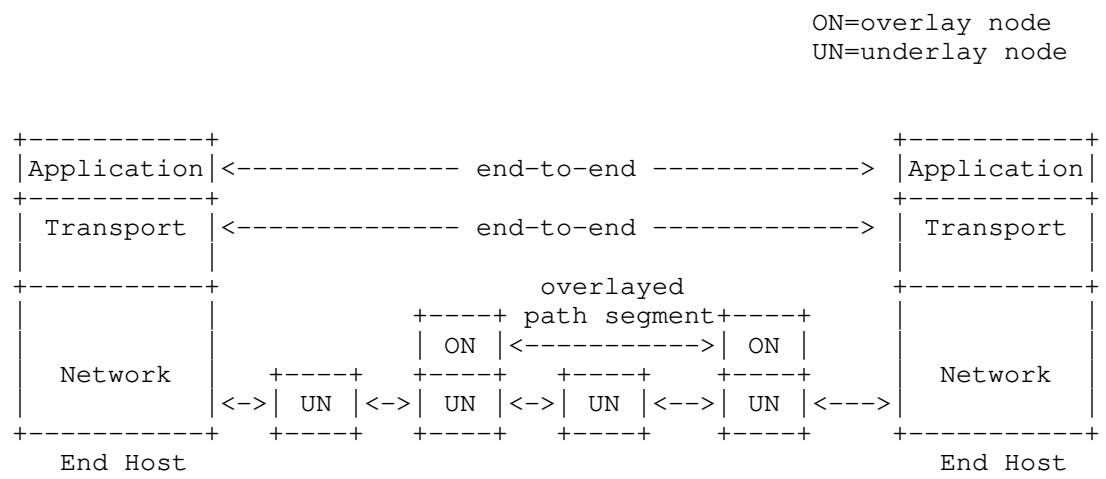

Figure 4. A Simple Overlayed Path Segment Forwarding Usage Scenario

Features need more investigations include,

- Enhancement for the overlayed path segment forwarding/transport, like retransmission, FEC(forward error correction), duplicating packet over the segments, lightweighted congestion control, etc. When the segment is a small portion of the whole end to end path, the retransmission over it has more benefit. Retransmission over the path segment has to be carefully designed to avoid the racing condition with the upper layer. The segment enabled retransmission may measure the segment RTT by itself to determine the appropriate retransmission attempts. On the other hand, the upper layers including the applications can indicate the credit as the safe band time that allows for the overlayed path segment to do the retransmission. At the same time, the persistent congestion caused packet loss should be exposed to the upper transport layer, so that the sender's congestion control can work properly. The timing of activation of the enhancement scheme, parameters such as the threshold setting of retransmission are worthy of further determination.

- Measurement based path selection for better performance, backup or load balancing. Overlay nodes have to be continuously monitored in order to find one or more appropriate overlayed paths. Such measurement can be in-band or out of band of data packets. When more than one overlayed segment with the same ingress and egress are used,

it has to be determined how the traffic are split and merged.

5. Security Considerations

TBD

6. IANA Considerations

No IANA action is required.

7. References

7.1 Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

7.2 Informative References

[RFC4122] Leach, P., Mealling, M., and R. Salz, "A Universally Unique IDentifier (UUID) URN Namespace", RFC 4122, July 2005.

[RFC5415] Calhoun, P., Ed., Montemurro, M., Ed., and D. Stanley, Ed., "Control And Provisioning of Wireless Access Points (CAPWAP) Protocol Specification", RFC 5415, March 2009.

[BurstyAna] Chung S., Agrawal D., Kim M., Hong J., and Park K. "Analysis of bursty packet loss characteristics on underutilized links using SNMP", IEEE/IFIP E2EMON, 2004.

[CRONets] Cai, C. X., Le, F., Sun, X., Xie, G. G., Jamjoom, H., and Campbell, R. H. CRONets: Cloud-Routed Overlay Networks. In 36th International Conference on Distributed Computing Systems (ICDCS) (2016), IEEE, pp. 67--77.

[RFC6830] Farinacci, D., Fuller, V., Meyer, D., and D. Lewis, "The Locator/ID Separation Protocol (LISP)", RFC 6830, January 2013.

[RFC7348] Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for

Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, August 2014.

Authors' Addresses

Yizhou Li
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China

Phone: +86-25-56624584
EMail: liyizhou@huawei.com

Xingwang Zhou
Huawei Technologies
101 Software Avenue,
Nanjing 210012
China

EMail: zhouxingwang@huawei.com