IEEE INFOCOM 2021 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS) | 978-1-6654-0443-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/INFOCOMWKSHPS51825.2021.9484508

VXLAN-based INT: In-band Network Telemetry for Overlay Network Monitoring

Yan Zhang*, Tian Pan†, Yan Zheng*, Enge Song†, Tao Huang*† and Yunjie Liu*†
*Purple Mountain Laboratories, Nanjing 211111, China
†State Key Laboratory of Networking and Switching Technology, BUPT, Beijing 100876, China
{zhangyan, zhengyan}@pmlabs.com.cn, {pan, songenge, htao, liuyj}@bupt.edu.com

***Abstract*—Overlay network protocols, such as VXLAN, are leveraged to address the need for network multiplexing and resource isolation within public clouds to accommodate multiple tenants. Since overlay networks are much more complex than underlay networks, overlay network monitoring is more significant and challenging. In-band Network Telemetry (INT) can achieve fine-grained network monitoring by encapsulating data plane states into probe packets. However, as an underlying device-level primitive, INT cannot be directly applied to overlay network monitoring given underlay networks and overlay networks are generally transparent from each other. In this work, we propose VXLAN-based INT, a telemetry system for overlay network monitoring based on VXLAN. By inserting the INT metadata collected from the underlay devices into the VXLAN payload, we successfully build the real-time overlay-underlay association at the controller, through which, one can easily localize the root cause of overlay path congestion within a simple database lookup.**

I. Introduction

The public clouds provide on-demand and scalable computing and storage resources to multiple cloud buyers. To satisfy the resource isolation requirements from the cloud buyers, cloud vendors build virtual private clouds (VPCs) within the public clouds through network virtualization techniques such as VXLAN [1]. VXLAN is a framework for overlaying virtualized layer 2 networks over layer 3 networks and it uses a VXLAN network identifier (VNI) to identify a VXLAN segment. Only VMs within the same VXLAN segment can directly communicate with each other, which achieves the network isolation purpose. The overlay networks are generally more complex than the underlay networks since each physical server can host hundreds of VMs or dockers, all of which are the potential endpoints of some end-to-end overlay paths. The quality of service (QoS) of overlay networks is vital to public cloud vendors because the service-level agreements (SLAs) signed with their customers should always be guaranteed. However, network congestion and silent failures [2] occur from time to time in mega-scale data centers. Fast and precisely localizing the root cause of these network congestion or silent failures has become an urgent need of major cloud vendors.

Traditional network monitoring approaches, such as SNMP, are unable to meet the need of fine-grained monitoring of data center networks. In-band Network Telemetry (INT) can achieve high-precision monitoring by encapsulating data plane states into probe packets and reducing the frequent interaction with the controller during the state collection [3]. However, as an underlying device-level primitive, INT cannot be directly applied to overlay network monitoring given underlay networks and overlay networks are generally transparent from each other. That is, the underlay networks are unaware of the existence of the overlay networks, and vice versa. Due to this reason, when cloud buyers experience network problems, it is usually difficult to tell whether the problems are caused by underlay networks or overlay networks.

This work is supported by the National Key Research and Development Program of China (2019YFB1802600). Corresponding author: Tian Pan.

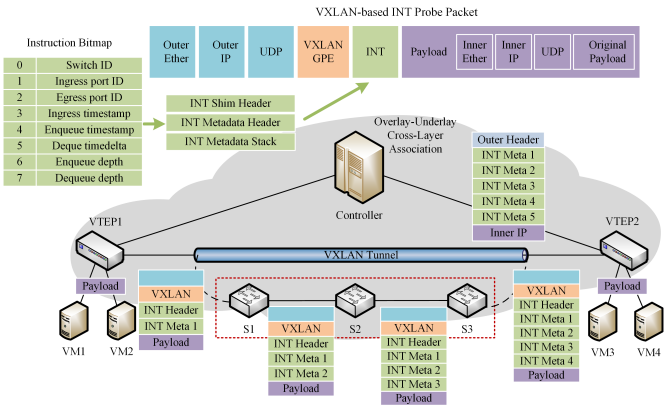

Fig. 1. Overlay network monitoring in VXLAN-based INT.

To ease the debugging of overlay networks, in this work, we propose VXLAN-based INT, a telemetry system dedicated to overlay network monitoring based on VXLAN. By inserting the INT metadata collected from the underlay devices into the VXLAN payload at the VTEP nodes, we successfully build the real-time overlay-underlay association at the controller, through which, one can easily localize the root cause of overlay path congestion by querying the corresponding underlay telemetry data using the overlay path information as the key.

II. VXLAN-based INT

Probe packet encapsulation. We define the probe packet encapsulation of the proposed VXLAN-based INT as shown in Fig. 1. The probe packet format is actually a combination of VXLAN and INT, consisting of Outer Ether header, Outer IP header, UDP header, VXLAN_GPE header, INT header and the Payload. Specifically, we use VXLAN_GPE header rather than the classic VXLAN header for INT information encapsulation. The value of the destination port in the UDP

978-1-6654-0443-3/21/$31.00 ©2021 IEEE

TABLE I
ASSOCIATING END-TO-END OVERLAY PATHS WITH COLLECTED UNDERLAY TELEMETRY RESULTS AT THE CONTROLLER.

End-to-End Overlay Path	Collected Underlay INT Metadata [Switch, Outgoing Port, Per-Hop Latency]
10.0.1.10-10.0.2.10	[S1 P2 155]
10.0.1.10-10.0.3.10	[S1 P3 149] [S9 P2 187] [S2 P1 255]
10.0.1.10-10.0.4.10	[S1 P3 200] [S9 P2 167] [S2 P2 201]
10.0.1.10-10.0.5.10	[S1 P3 198] [S9 P3 101] [S17 P2 312] [S11 P1 175] [S3 P1 200]
10.0.1.10-10.0.6.10	[S1 P3 166] [S9 P3 121] [S17 P2 300] [S11 P1 193] [S3 P2 180]
10.0.1.10-10.0.7.10	[S1 P3 198] [S9 P3 101] [S17 P2 312] [S11 P2 175] [S4 P1 167]
10.0.1.10-10.0.8.10	[S1 P3 177] [S9 P3 115] [S17 P2 272] [S11 P2 196] [S4 P2 188]

header is 4790 for VXLAN_GPE. In VXLAN_GPE header, the VNI field is used to identify different VXLAN segments in the overlay network and the value of the encapsulation protocol (*i.e.*, Next Protocol) is 0x05 for indicating that the following protocol is INT. The INT header consists of INT shim header, INT metadata header and INT metadata stack. In INT shim header, the length field indicates the total length of the INT header. In INT metadata header, the Instruction Bitmap field is an 8-bit bitmap, and each bit represents a type of INT metadata to collect that can be customized by network operators. The INT metadata stack accommodates the hop-by-hop INT metadata collected from the underlay network.

INT source. The INT source can be embedded in the source VXLAN tunnel endpoint (*i.e.*, VTEP). The INT source is responsible for spawning the VXLAN-based INT probe packets at the first hop of the monitoring path, either by periodically labelling INT onto user traffic (Fig. 1) or by generating/mirroring separate probe packets. For the latter case, the VM addresses can be obtained from the original user packets. Besides, as the first hop on the monitoring path, the INT source also needs to insert its own INT metadata (INT Meta 1) into the probe packet before forwarding the packet.

INT transit. The INT transit is responsible for pushing its local data plane states into the INT metadata stack according to the Instruction Bitmap before forwarding the packet. The INT transit should parse the VXLAN_GPE header to get the INT information but should be unaware of the user payload.

INT sink. The INT sink can be embedded in the destination VTEP. It receives the probe packets with INT metadata, extracts the user payload, sends it to the destination VMs, and uploads the INT metadata as well as the inner VM IPs to the central controller. Notice that we only upload the inner VM IPs rather than the entire user payload for privacy protection.

Controller. On receiving the VM IPs at the two endpoints of the overlay paths as well as the INT metadata collected from the underlay devices along the overlay paths, the controller can build the overlay-underlay cross-layer association as shown in Table I for fast overlay network troubleshooting. As the telemetry data is uploaded in a high frequency, the mapping relationship can be maintained and updated in real time.

III. EVALUATION

Experiment setup. We conduct experiments on an x86 machine with Ubuntu 16.04 OS, i7-9700K CPU@3.0GHz, and 16GB memory. The network emulation system is built with Mininet and BMv2 to evaluate VXLAN-based INT, which is available at the git repository [4]. Specifically, VTEPs are implemented on BMv2 with customized VXLAN encapsulation and decapsulation logic with INT. We evaluate the telemetry system on a fat-tree topology as shown in Fig. 2.

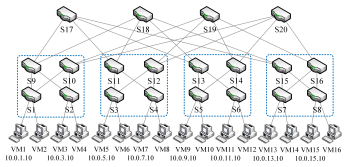

Fig. 2. The fat-tree network topology in the experiment.

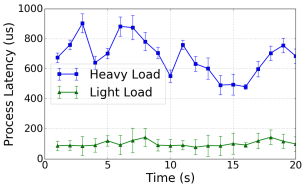

Fig. 3. Processing latency of VXLAN-based INT under heavy/light load.

Telemetry results. The INT metadata collected from the underlay devices along the end-to-end overlay paths are recorded into the local MySQL database at the controller as shown in Table I, which shows the mapping relationship between each end-to-end overlay path (*i.e.*, the key) and the corresponding INT metadata collected from the underlay network (*i.e.*, the value). For example, when VM1 communicates with VM3, the overlay path will pass through S1, S9 and S2 at the underlay network. The INT probe packets will collect the information such as switch ID, egress port ID and egress forwarding latency from each physical switch along the path. For example, for switch S9, the egress port is P2 and the latency is 187us. When overlay path congestion occurs, the network operator can easily localize the underlay choke point with the mapping table. Fig. 3 shows the single-hop forwarding latency under heavy load and light load. For BMv2, the latency is about 150us under light load and 400-1000us under heavy load.

IV. CONCLUSION

In this work, we propose VXLAN-based INT to meet the need of overlay network monitoring. Specifically, we design the probe packet encapsulation, define the forwarding behaviors of the source, transit and sink devices and elaborate the telemetry database at the controller. By associating overlay path information with underlay telemetry data, we can profile the overlay network status through a simple database lookup.

REFERENCES

[1] M. Mahalingam, D. G. Dutt, K. Duda, P. Agarwal, L. Kreeger, T. Sridhar, M. Bursell, and C. Wright, "Virtual extensible local area network (vxlan): A framework for overlaying virtualized layer 2 networks over layer 3 networks." *RFC*, vol. 7348, pp. 1–22, 2014.

[2] C. Jia, T. Pan, Z. Bian, X. Lin, E. Song, C. Xu, T. Huang, and Y. Liu, "Rapid detection and localization of gray failures in data centers via in-band network telemetry," in *NOMS 2020-2020 IEEE/IFIP Network Operations and Management Symposium*. IEEE, 2020, pp. 1–9.

[3] T. Pan, E. Song, Z. Bian, X. Lin, X. Peng, J. Zhang, T. Huang, B. Liu, and Y. Liu, "Int-path: Towards optimal path planning for in-band network-wide telemetry," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2019, pp. 487–495.

[4] "Vxlan-based int repository," https://github.com/graytower/INT_Overlay, 2020.