Cross-Site INT over Heterogeneous P4 Testbeds between Japan and Taiwan

Ta-Yuan Chou [1], Nai-Yuan Hu [1], Hui-Min Tseng [1], Te-Lung Liu[1], Shuji Ishii [2], and Hidehisa Nagano[2]*

[1] *National Center for High-Performance Computing, National Applied Research Laboratories, Tainan, Taiwan*
[2] *National Institute of Information and Communications Technology, Tokyo, Japan*
**tlliu@niar.org.tw*

Keywords: P4, In-band Network Telemetry, Cross-Site Performance Measurement

Abstract

This paper demonstrates the cross-site heterogeneous P4-INT testbeds between NCHC of Taiwan and NICT of Japan. NCHC and NICT have constructed Tofino-based and BMv2-based testbeds, respectively. To explore P4 technology together, we interconnect the testbeds via the multiple submarine cables through Los Angeles and Singapore. To monitor the statistics of the international testbed, we apply the INT technology for collecting real-time data. The proposed mechanism can help us explore various routing optimization algorithms to improve network performance in the future.

1 Introduction

In recent years, the era of artificial intelligence (AI) has arrived, and various large-scale computing resources based on CPU/GPU have become very important infrastructure. To integrate computing resources and storage resources spreading across different locations, various high-speed networks are extremely important. Also, it is also very important to monitor and manage the network status to ensure various service level agreements (SLAs).

Traditional network management technologies can be roughly divided into the following categories, such as SNMP, NetFlow, Client-Server, Agent, etc. These technologies often require additional equipment, applications, and additional packets to obtain various performance measurement results. This approach will add additional burden to network transmission. To overcome the shortcomings of traditional network management technology, a better method is needed.

To allow network communications to have a higher degree of freedom, many technologies have been customized for the Internet and have attracted the attention of many international network research organizations. Two of them are software-defined networking and programmable network switch technology.

The Software Defined Network (SDN) technology[1] is a technology that separates the control plane and data plane of the switch. The former mainly concentrates the control functions of the network on a software-based controller. The latter is on the hardware switch. They can communicate through a secure channel and use special protocols (such as OpenFlow) to achieve the function of the SDN controller controlling the SDN switch. However, SDN technology only allows users to customize the functions of the control plane.

Programming protocol-independent packet processors (P4) [2] can provide customized mechanism using programming language. Due to the features of high customization, various academic organizations have joined to study this topic. In Open Networking Foundation (ONF) [3], there are lots of applications based upon P4.

In-band Network Telemetry (INT) is a network performance measurement technology that mainly uses data packets themselves. It can use the metadata field of the packet to record performance data. Compared with traditional network management technology that transmits lots of measurement packets, in-band telemetry technology will significantly reduce the consumption of network bandwidth.

Since INT technology itself processes network packets, it would be quite challenging to use conventional network technology to operate on network packets. In other words, for the current Internet architecture, most network devices have already implemented fixed network communication protocols.

Taking advantage of the features of highly customization of P4, the functions of INT can be implemented on P4 platforms. Thus, P4 and INT can be combined as a complete solution to yield the ability of network monitoring.

In 2021, NCHC has deployed multiple P4 switches on the Taiwan Advanced Research and Education Network (TWAREN)[4]. TWAREN in Taiwan and through the VPLS service provided by TWAREN, the P4 network switches of each node are connected in series to form a P4 experimental platform, which is open to the academic and research community for experimental service.

Since P4 programmable network switching technology has also received considerable attention in Japan, many academic and research institutions and corporate organizations have also invested in the research and development of P4 programmable network technology and regularly hold P4-related developer conferences. Among them, the National Institute of Information and Communications Technology [5] (NICT) is also committed to the research of programmable switching network technology and has deployed P4 switches

in many locations in Japan. NICT also provides a P4 programmable experimental network platform through the connection to the JGN-X academic research network maintained by NICT.

In general, P4 INT[6] platforms are often limited to a single organization or even a single region. In our previous research, we also study the P4 Testbed[7]. However, there is little research on INT effectiveness measurement in cross-border and long-distance domains. The research cooperation between NCHC and NICT of Japan is aimed at conducting cross-site P4 INT technology research between sites that are far away from each other in a single transnational domain. Because it is relatively simple to conduct INT data among multiple hosts in a single site. The INT collection process between cross-sites is also different from that within a single site.

The organization of this paper is as follows. Section 2 introduces the background knowledge related to P4 and INT. Section 3 explains the cross-border P4 experimental platform between Taiwan and Japan. The measurement operation of INT in cross-site is described in Section 4. The last section states the conclusions and future work.

2. Background Knowledge

2.1 Software-based BMv2 Model

Fig. 1 is the Pipeline of the P4 programming language. Programmers can use the P4 language to define how to process data packets and compile it to generate a JSON file to configure the BMv2 Model. Programmers can also use the P4 language to define various switches, firewalls, load balancers, etc.

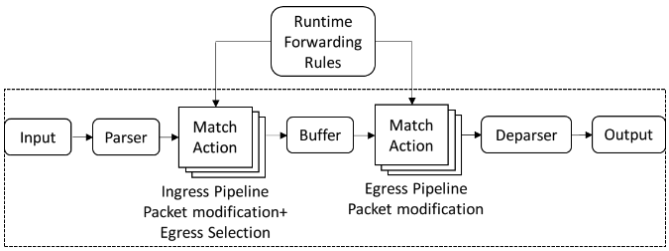

Fig. 1 P4 Pipeline of BMv2

After receiving the data packet, the switch will analyze it through the Parser to obtain the detection instructions and the object to be detected. The data packet then enters the match + action phase for processing. In this stage, the Ingress Pipeline is processed to obtain the corresponding Egress and modify the data packet. Based on Runtime Forwarding Rules, data is packaged and output.

2.2 Hardware-based Tofino Model

In addition to software-based BMv2 Model, there are many suppliers who have implemented hardware versions of P4 programmable switches, such as EdgeCore, Inventec, etc. Since the switch hardware itself has a common Tofino/Tofino2 Switch ASIC and is paired with related Board Support Packages (BSP), developers can develop and compile the corresponding code through P4 code to call the hardware. To fully utilize the P4 programmable switching network function, one can install the Open Network Linux (ONL) operating system and install Intel Barefoot's Software Development Environment (SDE) in the operating system for development.

2.3 In-band Network Telemetry

In-band Network Telemetry (INT) is a new example of multiple network telemetry technologies and has received extensive attention from academia and industry in recent years. Combining packet forwarding with network measurement mainly uses the method of inserting the collected network status information into the packet in the switch on the path to achieve the measurement purpose.

INT is a data-plane-based framework that collects and reports data plane network status. This architecture does not require the intervention of the control plane. Compared to traditional network management technologies that require additional commands to monitor network status, INT adds a header containing detection commands to the metadata field of the data packet. The advantage of this is that since the network status information is attached to the data packet, the larger the network packet volume, the higher the frequency of network status updates. There are three components in the INT architecture.

- INT Source: INT source, sending data packets.
- INT Transit: INT relay point, used to record data into metadata or execute INT commands.
- INT Sink: INT endpoint, extracts metadata from data packets.

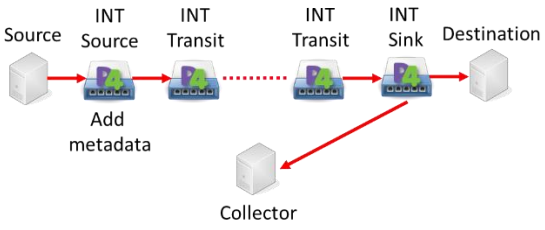

Fig. 2 Schematic diagram of the P4 INT framework

As shown in Fig. 2, when the INT source receives the packet from the Source host, it will add metadata. Next, when the INT Transits receive the packet, it will also append the metadata to the packet, then send the packet to the next switch. Finally, when the INT Sink receives the packet, it will send the fetching metadata to the collector, and send the data packet to the Destination host.

3. Cross-Site INT over Heterogeneous P4 Testbeds

Fig 3 is the topology of the cross-site P4-INT testbed. As shown in Fig 3, the left-hand side and the right-hand side represent the testbeds of NICT and NCHC, respectively.

These testbeds are connected via submarine cables through Los Angels (LA) and Singapore (SG). Hence, this can form a multipath cross-site testbeds.

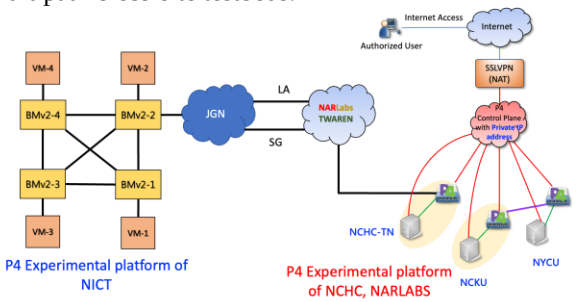

Fig. 3 Domestic topology of TWAREN in Taiwan

Also, as shown in Fig. 3, since NICT and NCHC use switches in BMv2 and Tofino models respectively, it can be viewed as an integration of heterogeneous testbeds. To make the heterogeneous testbeds can communicate with each other, the P4 programs should be modified to be compatible on each platform.

4 Measurement Operation

As mentioned in the previous introduction to in-band telemetry technology, INT is divided into INT Source, INT Transit, and INT Sink. The data packet is sent from the INT Source, and after the INT Transit collects relevant information, the metadata in the data packet is decrypted at the INT Sink. This is not a problem for a single site. However, different problems will arise for cross-site.

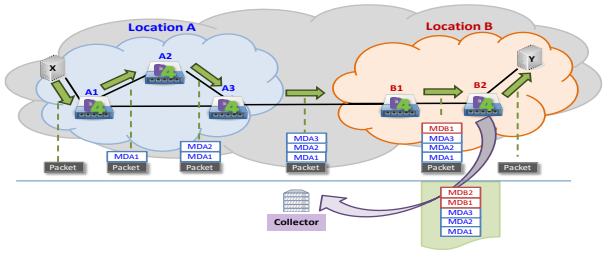

Fig. 6 Concept of the cross-site P$-INT testbed

As shown in Figure 6, assume that this is a cross-site network architecture with Location A on the left and Location B on the right. The data packets transmitted in Location B will continuously add relevant data to the metadata field of the data packets when passing through each INT Transit device. Due to the network MTU limitation, the relevant data must be temporarily stored in the collector corresponding to Location B. For Location A, the same transmission method will be used and the problem of exceeding the MTU will be encountered. The corresponding data can also be stored in the collector of Location A.

Next, if the data is to be transmitted from Location A to Location B, when the data packet reaches the exit of Location A, the relevant metadata will be thrown out, so the packet that continues to be transmitted to Location B is the original data packet.

If the above method is used, cross-site metadata cannot be successfully transferred to another site. Therefore, one of the solutions is to allow the INT Sinks of Location A and Location B to read each other's Collectors, as shown in Fig 7.

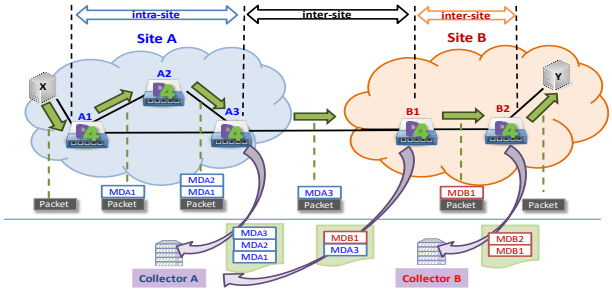

Fig. 7 Domestic topology of TWAREN in Taiwan

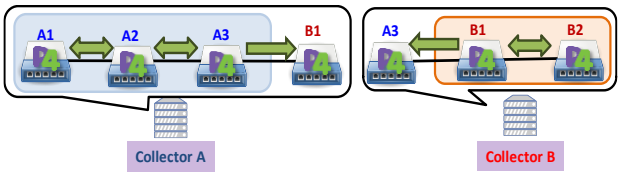

Fig. 8 Domestic topology of TWAREN in Taiwan

As shown in Fig. 8 , conceptually, the scope that Collector A can cover is all nodes at Location A and the INT Sink nodes at Location B. The scope that Collector B can cover is all nodes in Location B and the INT Sink node in Location A.

5 Conclusion

This paper presents the P4 programmable testbeds cross Taiwan and Japan. Through two international lines in Los Angeles and Singapore, two testbeds can be integrate as a heterogeneous and multipath testbed. In the future, we will conduct field tests on large-scale multinational P4 experimental platforms and develop cross-site INT methods to provide more performance measurement solutions.

References

[1] Software Defined Network Definitions, https://opennetworking.org/sdn-definition/

[2] P4 Language, https://p4.org/

[3] Open Networking Foundation, https://opennetworking.org/

[4] TWAREN, https://www.twaren.net/

[5] National Institute of Information and Communications Technology, NICT , https://www.nict.go.jp/

[6] P4 INT Specification, https://p4.org/p4-spec/docs/INT_v2_1.pdf

[7] Wun-Yuan Huang , Ta-Yuan Chou ,Nai-Yuan Hu, Hui-Min Tseng, and Te-Lung Liu, Design and Building of P4 Programmable Network Testbed and Reservation System on TWAREN," 2023 International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan 2023), Pingtung, Taiwan, July 2023