台日間跨國多路徑帶內遙測實驗平台建置與測試規劃

周大源 胡乃元 曾惠敏 劉德隆
財團法人國家實驗研究院國家高速網路與計算中心
E-mail: {1203053, 2103081, 0303118, tlliu}@narlabs.org.tw

摘要

本論文展示一個台日之間跨國多路徑的可程式化網路（P4）實驗平台。在台灣的部份，我們在數個 TWAREN 節點佈建硬體式的 P4可程式化網路交換器並透過 TWAREN 的 VPLS 服務建置台灣的 P4可程式化實驗平台。在日本部份，NICT 在日本多個節點上佈建 BMv2軟體版本之 P4可程式化實驗網路平台。經由美國洛杉磯與新加坡之跨國連線，我們將台日兩端的 P4可程式化實驗平台互相介接，形成一個大型的 Layer 2之 cross-site 實驗網路平台。另一方面，我們也針對 cross-site 帶內遙測（INT）的方法，並針對 cross-site 封包寫入 metadata 資料過長之問題提出解決方案。未來將可針對多路徑傳輸方法提供路徑效能資訊，達成最佳化傳輸之目標。

關鍵詞：可程式化網路交換器、帶內遙測技術、跨站台效能量測

Abstract

This paper demonstrates a cross-border multi-path programmable network (P4) experimental platform between Taiwan and Japan. In Taiwan, we deployed hardware-based P4 programmable network switches on several TWAREN nodes and built Taiwan's P4 programmable experimental platform through TWAREN's VPLS service. In Japan, NICT deployed the BMv2 software version of the P4 programmable experimental network platform on multiple nodes in Japan. Through the cross-border connection between Los Angeles and Singapore, we interconnected the P4 programmable experimental platforms in Taiwan and Japan to form a large-scale Layer 2 cross-site experimental network platform. On the other hand, we also propose a solution for the cross-site in-band telemetry (INT) method and the problem of too long metadata data written in cross-site packets. In the future, path performance information will be provided for multi-path transmission methods to achieve the goal of optimized transmission.

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

1. 前言

近年來，人工智慧（AI）的時代來臨，各類以 CPU 及 GPU 為基礎的大型計算資源，是非常重要的基礎設施。除了計算資源之外，還需要儲存資源來進行輔助。為了要將分散在各地的計算資源與儲存資源串接在一起，其重要關鍵便是各種高速網路。

近幾十年來，網路骨幹技術蓬勃發展，而網路傳輸的技術也蓬勃發展。為了要達成網路介接的強韌性與穩定性，在多個站點之間往往建置多條的路徑可以彼此溝通。

同時，為了確保網路與各項服務能夠正常營運，網路維運單位會利用各種網管技術針對網路進行效能監控與效能量測等等。

傳統的網管技術大致上分為以下幾類，如 SNMP、NetFlow、Client-Server、Agent 等等。雖然傳統網管技術能夠提供網路管理者相關的運作與效能資訊。然而，傳統網管技術往往需要額外的設備或應用程式。在收集網路效能資訊時，往往得傳輸額外的封包到網路上，藉以獲得各種效能量測的結果。這樣的方式會增加網路傳輸額外的負擔。為了要克服傳統網管技術的缺點，需要一種更好的方法。

帶內遙測技術（In-band Network Telemetry, INT）是一種以資料封包本身為主的網路效能量測技術。帶內遙測技術可以利用封包的 metadata 欄位來進行效能資料的紀錄。相較於傳統網管技術傳遞大量量測封包的方式，帶內遙測技術對於網路頻寬的耗用將大幅度減少。

由於 INT 技術本身是針對網路封包進行處理，若以一般傳統網路技術來針對網路封包進行操作，會有相當程度之挑戰。換句話說，針對現行網際網路的架構，多數網路設備都已經實作固定之網路通訊協定。

為了要讓網路通訊能夠有更高的自由度，許多技術都針對網路進行客製化，並受到許多國際研網組織的關注。其中兩項是軟體定義網路與可程式化網路交換器技術。

軟體定義網路（Software Defined Network，SDN）架構如圖1所示。其技術是一種將交換器的控制面（control plane）與資料面（data plane）的部份分隔的技術。前者主要是將網路的控制功能集中在軟體為主的控制器上。後者則是在硬體的交換器上。兩者之間可以透過安全通道，並以特殊的協定（例如 OpenFlow），達成 SDN 控制器控制 SDN 交換器的功能。

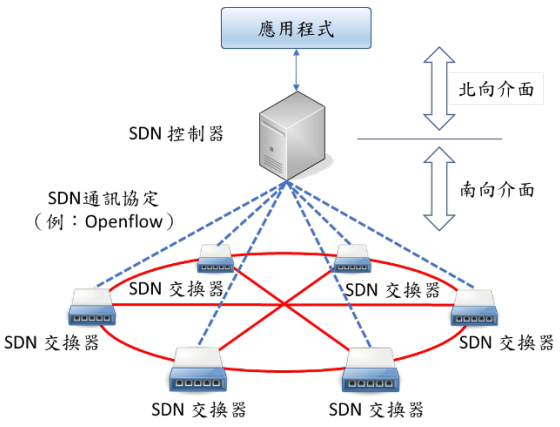

圖 1 軟體定義網路架構圖

然而，SDN 技術僅僅只能讓使用者對控制面

的部份進行功能客製化。對於資料面的部份，則沒有進一步的克制化的功能。因此，我們需要進一步的資料面客製化的技術。

可程式化網路交換器[1]（Programming protocol-independent packet processors, P4）是一種與協定無關的可程式化交換器。藉由撰寫 P4程式語言，使用者可以進一步針對資料封包的自行定義格式、以及資料封包傳輸的的行為等等定義相對應的處理程序，讓 P4交換器達成更高度客製化的特性。由於 P4技術本身能夠讓資料封包實現高度客製化的特性，讓 INT 技術在 P4網路平台上容易實現。

由於可程式化網路交換器技術具有高度客製化特性，世界上許多學研網路組織紛紛投入研究。在開放式網路基金會[2]（Open Networking Foundation, ONF）中亦有許多以 P4為基礎的延伸應用。

本中心自2021年以來便在台灣先進學術研究網路[3]（TaiWan Advanced Research and Education Network, TWAREN）上佈建多部 P4交換器。TWAREN 在國內與透過 TWAREN 所提供之 VPLS 服務，將各節點之 P4網路交換器串接起來，形成 P4實驗平台，並開放給學研界使用實驗性服務。TWAREN 的國內與國外連線架構分別如圖2與圖3所示。

圖 2 TWAREN 國內連線架構圖

圖 3 TWAREN 國際連線架構圖

由於 P4可程式化網路交換技術在日本也受到相當多的關注，許多學研界單位與企業組織亦有投入 P4可程式化網路技術之研究與開發，並定期召開 P4相關開發者大會。其中，情報通訊研究機構[4]（National Institute of Information and Communications Technology, NICT）也致力於可程式化交換網路技術研究，並於日本多地佈建 P4交換器。透過 NICT 所維運之 JGN-X 學術研究網路的連接，NICT 亦提供 P4可程式化實驗網路平台。自去年（2023年）起，本中心與 NICT 多次進行 P4可程式化網路技術進行交流，並且開放雙方之 P4可程式化網路實驗平台以進行試用。而在今年（2024年）本中心與 NICT 共同簽署合作備忘錄，將雙方之 P4可程式化實驗平台串連起來，真正達成跨國遠距之 P4 INT 實驗平台。平台的詳細資訊將於稍後的內容中介紹。

一般而言，P4 INT[5]的平台往往是侷限在單一組織，甚至是單一區域的環境中。針對跨國遠距的網域進行 INT 效能量測則少有著墨。而本中心與日本 NICT 之研究合作案則是針對跨國單一網域，但在彼此相距甚遠的 site 間進行 cross-site 之 P4 INT 技術研究。由於在單一 site 間多部主機進行 INT 資料較為單純。而在 cross-site 之間的 INT 收集流程也與單一 site 中的方式不同。

本論文的組織架構如下。第2.節針對 P4與 INT 相關背景知識進行介紹。第3.節針對台日之間跨國之 P4實驗平台進行說明。而針對 cross-site 中的 INT 運作方式在第4.節中闡述。在最後一節中則陳述結論與未來工作。

2. 背景知識

2.1 可程式化網路交換器

圖4是 P4程式語言的 Pipeline。程式設計者可以使用 P4 語言定義處理資料封包的方式，並編譯產生一個 JSON 檔案以針對交換器晶片進行組態設定。程式設計者也可以使用 P4語言定義各種交換器、防火牆，或者負載平衡器、...等等裝置。

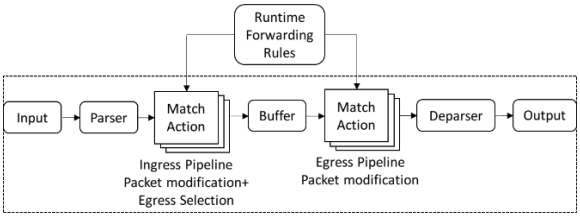

圖 4 P4 Pipeline

交換器在收到資料封包後，會經由 Parser 做剖析，得出偵測指令與需偵測的對象。資料封包稍後則進入 match + action 階段進行處理。此階段會進行 Ingress Pipeline 處理，得出相對應的 Egress，並修改資料封包。基於 Runtime Forwarding Rules，資料封包後輸出。

基礎的 P4交換器 Simple switch 的基礎程序如下：

```
V1Switch(
   MyParser(),
   MyVerifyChecksum(),
   MyIngress(),
   MyEgress(),
   MyComputeChecksum(),
   MyDeparser()
)main;
```

如上所示，基礎的 V1 model switch 就是經過上述的標準運作程序。在 MyParser()中主要是要針對封包進行剖析。而 MyVerifyChecksum()是用來驗證檢查碼是否正確。接下來 MyIngress()用以處理封包進入的 port 的相關程序。而 MyEgress()用以處理封包輸出 port 的相關程序。藉由 MyComputeChecksum()程序，可以針對封包的檢查碼進行更新。在所有對應的處置動作完成後，MyDeparser()會將最後結果包裝成對應的封包並傳到輸出的 port 中。

在 P4 可程式化交換器中，需要注意的是以下幾部份。

- Behavior Model(BMv2)：這個部份是用來描述硬體架構。
- P4 compiler：用來編譯 P4程式碼的工具。
- P4 runtime：這個是 P4程式碼的執行環境。

對於語法部份，P4官方網站也有提供 P4 cheat sheet[1]文件，讓程式開發者能夠針對重要的關鍵語法進行參考。

2.2 硬體版本之 P4可程式化交換器

目前有許多供應商實作硬體版本之 P4可程式化交換器，如 EdgeCore、Inventec、…等等。由於交換器硬體本身有共通之 Tofino/Tofino2 Switch ASIC，並搭配相關的 Board Support Packages (BSP)，開發者可藉由 P4程式碼開發並編譯出相對應的程式碼以針對硬體進行呼叫。

一般硬體版本的 Tofino/Tofino2 ASIC 交換器本身僅搭配基本的開放式網路安裝環境，如 Open Network Install Environment (ONIE)。有些供應商則會進一步提供交換器作業系統，如 Stratum、Sonic 等等開放網路社群軟體為基礎之作業系統。安裝作業系統後，交換器本身可具備基本網路交換功能。

然而，若要完整使用 P4可程式化交換網路功能，可以安裝 Open Network Linux (ONL)作業系統，並於作業系統中安裝 Intel Barefoot 之軟體開發環境 (Software Development Environment, SDE) 以進行開發。若要取得 Intel Barefoot 之 SDE 工具，則需要向 Intel 洽詢，並簽署 NDA 方能取得。

2.3 帶內網路遙測技術

網路遙測（Network Telemetry）是一種較新的網路資料蒐集技術，而遙測是對網路資訊進行遠端搜集和處理的自動化過程。網路遙測和傳統網路量測的比較上，前者被廣泛認為比後者於了解網路狀態方面，具有更好的可擴充性、準確度、覆蓋範圍和效能。

帶內遙測技術 (In-band Network Telemetry, INT) 則是多項網路遙測技術的一種新案例，近年來受到了學術界和業界的廣泛關注。將 packet forwarding 與網路量測相結合，其主要是利用於路徑上的交換器內將搜集的網路狀態資訊插入封包之中的方式來達到測量的目

INT 是一種以 data-plane 為主，收集與回報 data plane 網路狀態的框架。這種架構不需要 control plane 介入。相較傳統的網管技術需要額外指令來進行網路狀態監控，INT 是將包含偵測用指令的 header 加入資料封包的 metadata 欄位中。這樣不會額外造成網路的負擔。這樣的好處就是：由於網路狀態資訊是附在資料封包內，因此當網路封包量愈大、網路狀態更新的頻率愈高。帶在 INT 的架構上，有三種元件：

- INT Source：INT 來源，發出資料封包。
- INT Transit：INT 中繼點，用以將資料記錄至 metadata 中或執行 INT 指令。
- INT Sink：INT 終點，將資料封包中的 metadata 取出。

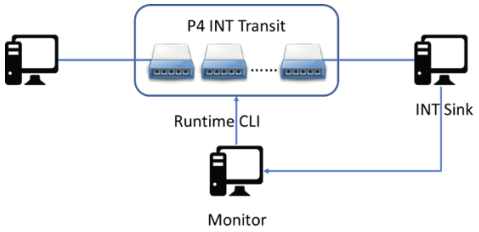

圖 5 典型的 P4 INT 框架

一般而言，INT 區分為以下三種方式：

- INT-XD（eXport 資料）：INT 節點根據在其流監視清單中配置的 INT 指令，直接將 metadata 從其資料平面匯出到監控系統。無需對資料包進行修改。
- INT-MX（內嵌指令）：INT Source 節點將 INT 指令嵌入封包頭中，然後 INT Source、每個 INT Transit 和 INT Sink 依照嵌入的指令將元資料直接傳送到監控系統在封包中。
- INT-MD（eMbed 資料）：在此模式下，INT 指令和元資料都會寫入資料包中。這是經典的逐跳 INT，其中 INT Source 嵌入指令，INT Source & Transit 嵌入 metadata，以及 INT Sink 從資料封包中分解指令和聚合 metadata，並有選擇性地將數據發送到監控系統。在此模式下，封包修改最多，同時最大限度地減少了監控系統整理來自多個 INT 節點的資料的空間消耗。

在 P4 switch 中，亦有針對 INT 進行實作。網路裝置在收到這些封包後，就會把偵測指令所指定的資訊寫入資料封包中。圖5. 為 P4 INT 框架示意圖。INT框架包含 INT Source、INT Sink，以及 INT Transit。一般在 INT Source 中會建構一個包含 INT

偵測指令的 header。而 INT Transit 即為在資料傳輸路徑中每一部支援 INT 的裝置。當 INT Transit 收到資料封包時，會將 INT 偵測指令所對應的狀態資訊寫入資料封包中。而前述的資訊會收集到 INT Sink，並會傳給 monitor。Monitor 所收集到的網路狀態資訊可以直接傳送給 data-plane 以使用，或者是進一步轉送給 control-plane 進行分析。

在使用 P4 INT 時可以自行定義並蒐集任何交換器內部的資訊，目前 P4 INT 的官方規範中提供了幾種可以使用的 Metadata，其中大多數可以直接透過 P4 定義的 Standard Metadata 直接從設備中取得：

- Switch identifier: 交換機的唯一 ID。
- Ingress port ID: 接收 INT 封包的 port ID。
- Ingress timestamp: 設備接收到 INT 封包時的本地時間戳記。
- Egress port ID: INT 送出封包的 port ID。
- Hop latency: INT 封包在設備中傳輸的延遲。
- Egress port TX Link utilization：送出 INT 封包的 port 當前的使用率。
- Queue occupancy：INT 封包在設備中傳送時觀察到 Queue 中已儲存的流量。
- Queue congestion status：當前 Queue 的壅塞狀態。

然而，INT 也存在一些缺點：首先，隨著傳輸 hop 數的增加，包括遙測 metadata 在內的總資料包大小可能會超過網路 MTU，因此無法插入更多 metadata；此外，如果網路所涵蓋的地理區域較大，INT 收集器可能會從遠距離的交換器接收 metadata，會導致時間同步和傳播延遲問題，因此在大規模網路上部署 INT 非常困難。

3. 台日間跨國 P4實驗平台

本節針對台日間跨國 P4實驗平台進行介紹，分別是本中心所建置之 P4可程式化實驗平台、NICT 所建置之 P4可程式化實驗平台，以及雙邊介接之架構。

3.1 TWAREN 可程式化實驗網路平台架構

本節介紹 TWAREN 可程式化實驗網路平台之架構。如圖6.所示，我們選定本中心新竹本部、台南分部、成功大學、陽明交通大學，以及中興大學等等節點各設置一部 P4交換器與一部伺服器。在各大節點的 P4 交換器與伺服器間會有 Switch-Host Connection，伺服器得以藉由 P4交換器進行傳送與接收資料，故可視為 P4交換器之用戶端，如圖6.中 P4交換器與 Server 間之實心細線。

而在各大節點之間，我們利用 TWAREN 所提供的 VPLS 連線進行介接，亦即為P4交換器之間的 Data Path，其 topology 如圖6.之實心粗線所示。

為了減少來自外部 Internet 的網路攻擊流量，我們設置兩部實體 SSLVPN，藉以提供用戶進行身分驗證，並可在用戶端裝置取得 private IP 位址。

由圖6.可知，我們的 TWAREN P4 可程式化實驗網路以國網中心新竹本部、台南分部為核心，與陽明交大、成大，還有中興均透過 TWAREN VPLS 直接進行介接。另一方面，陽明交大、成大、中興則與兩核心節點介接。目前第一階段僅以國網中心台南分部、陽明交大，以及成大共三節點投入研究。在先前的研究中，我們也針對 P4 實驗平台開發預約系統[6][7][8]。

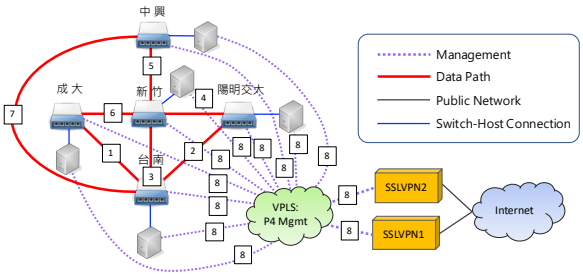

圖 6 P4 可程式化交換器實驗網路

如圖6所示，編號欄位代表所需要的 VPLS，起點與終點欄位則註明介接之兩端節點。介面類型有區分為10G SR 與 1G RJ-45。編號1~7的部份是交換器間資料傳輸之路徑，故為10GSR。編號8的部份是基於 SSLVPN 轉換 private IP 位址與 public IP 位址之用。由於 SSLVPN 部份僅供使用者登入管控使用，因此並不需要使用10G 高頻寬的介面，故以1G RJ-45介面即可。

由於目前本中心所購置之硬體版 P4可程式化交換器僅有100G/40G 之介面，因此需要使用一分四之分光器 (Breakout) 將 40G 訊號輸出為 4*10G 線路，即可與 TWAREN 設備之10G 介面進行介接。相關 Data Path 部份均以10G 線路進行介接。

3.2 NICT 之 P4實驗平台架構

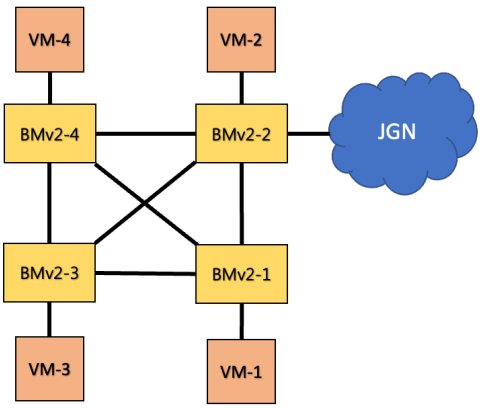

圖7 NICT 之 P4實驗平台架構

如圖7所示，NICT 於四個節點分別佈署四部 BMv2軟體版本的 P4交換器，透過 JGN 網路對外連線。

3.3 跨國連線架構

如圖8所示，左半邊為日本 NICT 的 P4實驗平台，透過 JGN 網路對外連接。右半邊為本中心之

P4實驗平台，透過 TWAREN 網路對外連線。現階段而言，雙邊實驗平台已經透過 LA 介接。未來，我們也將透過新加坡進行介接。由於本中心提供 Tofino 平台，而 NICT 提供 BMv2之平台。未來仍需要在兩種不同的平台之間進行連線測試，並用不同的方法實作 INT。

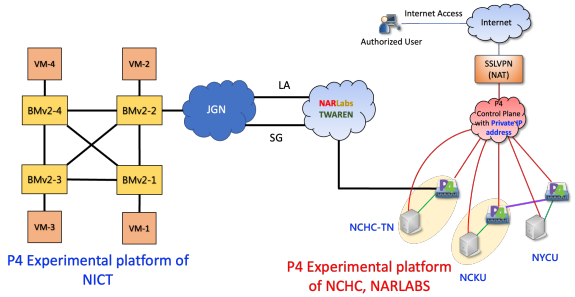

圖 8 台日跨國 P4與 INT 實驗平台架構

```
nchclab@nchclab-prop42:~$ ping 192.168.1.1
PING 192.168.1.1 (192.168.1.1) 56(84) bytes of data.
64 bytes from 192.168.1.1: icmp_seq=1 ttl=64 time=261 ms
64 bytes from 192.168.1.1: icmp_seq=2 ttl=64 time=261 ms
64 bytes from 192.168.1.1: icmp_seq=3 ttl=64 time=261 ms
64 bytes from 192.168.1.1: icmp_seq=4 ttl=64 time=261 ms
64 bytes from 192.168.1.1: icmp_seq=5 ttl=64 time=261 ms
64 bytes from 192.168.1.1: icmp_seq=6 ttl=64 time=261 ms
```

圖 9 以 ping 指令進行時間測試

如圖9所示，我們初步以 ping 指令，在台南節點之 P4實驗平台（綁定 IP 位址192.168.1.2）來針對 NICT 之 P4平台（綁定 IP 位址192.168.1.1）進行連通測試，時間為261ms。

4. Cross-Site 效能量測方案

如前所述，本論文主要針對是 cross-site 之間長距離大規模的網路進行 INT 研究。這部份對於具有分佈於多各地裡區域的企業組織或大型學研單位等等都相當重要。然而，這部份的文獻相當稀少，因此，本計畫將針對這部份進行研究。我們將在跨國環境進行網路實驗平台的建置，並在實驗平台上進行 P4與 INT 的驗證實驗，發展傳輸方法。

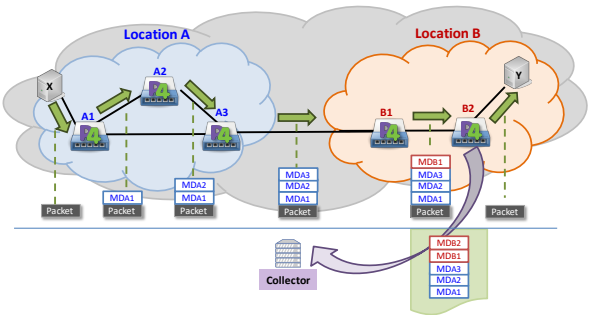

圖 10 跨國 cross-site INT 運作方式

在先前對帶內遙測技術的介紹中有提到，因為 INT 有區分為 INT Source、INT Transit，以及 INT Sink。資料封包從 INT Source 發出，經由 INT Transit 收集相關資訊後，在 INT Sink 將資料封包中的 metadata 解開。這對於單一 site 而言並無任何問題。然而，針對 cross-site 而言，將產生不同的問題。

如圖10所示，假設這是一個跨區域的 cross-site 網路架構，左邊為 Location A，右邊為 Location B。在 Location B 中所傳輸的資料封包在經過每一部 INT Transit 裝置時會不斷將相關資料疊加至資料封包的 metadata 欄位中。礙於網路 MTU 限制，必須先將相關資料額外暫存至 Location B 所對應的 collector 中。而針對 Location A 而言，也會有相同的傳輸方式、亦會遇到超出 MTU 的問題，亦可將相對應的資料儲存至 Location A 的 collector 中。

接下來，倘若資料要由 Location A 傳輸至 Location B，當資料封包抵達 Location A 的出口時，相關 metadata 資料就會被拋出，故繼續傳至 Location B 的封包是原始的資料封包。

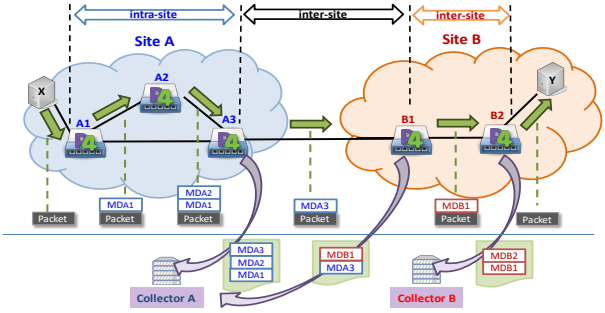

圖 11 跨站台（Cross-Site）INT 之情境

如果以上述方式運作，則跨 site 的 metadata 無法順利傳至另外一個 site。因此，其中之一的解決方案便是讓 Location A 與 Location B 的 INT Sink 能夠互相讀取對端的 Collector，如圖11所示。

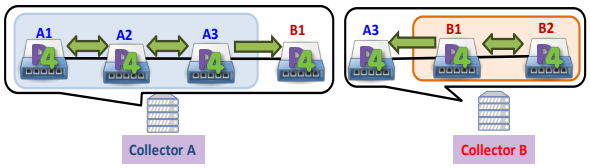

圖 12 雙向資料封包傳輸後所收集到的 INT metadata

如圖12所示，以概念而言，Collector A 能夠涵蓋的 scope 為 Location A 的所有節點及 Location B 的 INT Sink 節點。而 Collector B 能夠涵蓋的 scope 為 Location B 的所有節點及 Location A 的 INT Sink 節點。

5. 結論與未來工作

本論文展示一個橫跨台灣與日本之間的 P4可程式化實驗網路平台。透過經由美國洛杉磯與新加坡的兩條國際線路，將本中心與日本 NICT 所維運的 JGN-X 網路之 P4實驗平台介接起來，形成跨國 P4跨國實驗平台。未來我們將針對大型跨國之 P4實驗平台進行實地場域測試，並發展 cross-site 的 INT 方法，提供更多效能量測的解決方案。

參考文獻

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

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

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

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

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

[6] 周大源, 黃文源, 胡乃元, 曾惠敏, 劉德隆, “TWAREN 可程式化實驗網路平台建置,” TANet2022研討會, 桃園, 2022 年12月

[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

[8] 周大源, 黃文源, 胡乃元, 曾惠敏, 劉德隆, “TWAREN 可程式化實驗網路平台預約系統前端建置, ” TANet2023研討會, 台北, 2022 年11月