整合式帶內遙測技術於可程式化網路之數位孿生實驗平台規劃

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, 2503134, 0303118, tlliu}@niar.org.tw

摘要

本論文主要針對整合式帶內遙測技術與可程式化網路來進行數位孿生實驗平台進行規劃。帶內遙測技術主要是在網路資料封包中收集交換器資訊，讓網路管理監控可以在資料傳輸時同時完成。藉由帶內遙測的特性，可以改善傳統網路管理與監控技術的不足。另一方面，可程式化網路交換器能夠針對資料封包格式與傳輸資料封包的程序進行高度客製化操作，故為實作帶內遙測網路的良好解決方案。然而，由於 P4交換器的成本較高，在學研界網路尚未大量使用與部署。因此一般難以呈現大型可程式化實驗網路之效果。因此，我們擬針對整合式帶內遙測技術與可程式化網路之數位孿生實驗平台進行規劃。

關鍵詞：可程式化網路交換器，帶內遙測技術，數位孿生。

Abstract

In this paper, we demonstrate a digital twin system of the Programmable Protocol-independent Packet Processors (P4) network with the In-band Network Telemetry (INT) technology. The INT technology mainly gathers switch information and append to the data packets so that the process of network management and monitoring can be completed simultaneously when transferring data. Via the features of INT technology, network management and monitoring can be enhanced and improved. On the other hand, P4 switches can define and configure the format and operations of data packets with highly customized features. Therefore, P4 is suitable for implementing INT technology. However, a large-scale P4 network cannot be used and deployed due to the high cost of physical P4 switches so that the effectiveness cannot be realized and recognized. Therefore, we propose to give a plan on the development of the digital twin of the P4 INT platform.

Keywords: Programmable Protocol-independent Packet Processors, P4, In-band Network Telemetry, INT, Digital Twin.

1. 前言

近年來，由於人工智慧（AI）技術蓬勃發展，各類以 CPU 及 GPU 等具有高度運算能力為主的高端主機在 AI 發展的角色更顯重要。同時，為了讓 AI 技術能夠更加精準，龐大且足夠數量的訓練資料也是不可或缺的。因此，為了將遍布於世界各地的大資料與 AI 計算資源連結在一起，背後關鍵的主力是寬頻的高速傳輸網路。為了因應日益龐大且複雜的網路架構，網路效能量測與網路管理技術就顯得相當重要。

在傳統網路管理技術中，主要的監控工具是 NetFlow 或是 S-Flow。NetFlow 工具是針對每一個 Flow 為單位，亦即使用 7 個關鍵值來識別唯一流量：Source IP Address、Destination IP Address、Source Port、Destination Port、IP Protocol、入口介面（entry port）、服務類型（Type of Service）值。在上述所有欄位比對相同後，才會視為是同一個 Flow。

而 S-Flow 是一種取樣式（Sampling）的方法，針對每一個 Packet 來進行監控。然而，在 S-Flow 中如果針對每一個 Packet 均進行監控，會增加大量的網路 overhead；反之，如果以固定比例的取樣方式來針對網路封包進行監控，則會有遺失重要資訊的可能性。

另外，一旦網路出現故障，網管人員會採用 Ping 或 Traceroute 等等工具來進行連通性及路徑查詢。如果要測試線路的頻寬與網速，就必須使用 iPerf 與 SpeedTest 等等工具進行量測。

由於上述方式均需以額外的工具進行網管監控與效能量測。加上監控的方法必須在不增加網路負載與精細度間進行抉擇。因此，如果能夠在網路傳輸本身的封包進行網路效能監控與量測，將會提升整體效率與精細度。亦即能夠達成內部的營運管理維護（In-situ Operation, Administration, and Maintenance, In-situ OAM, IOAM）的模式，將會是較有效的作法。

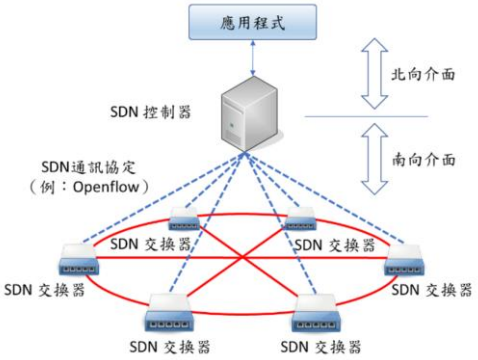

圖1. 軟體定義網路架構圖

近年來，由於開放式網路（Open Networking）的研究，軟體定義網路（Software Defined Network，SDN）的技術首先將網路交換器中的控制面（Control Plane）與資料面（Data Plane）分開，並將控制面的功能集中到另一部控制器（Controller）中。如圖1.所示，透過 OpenFlow 的通訊協定，而軟體定義網路之交換器只需要接受來自 Controller 的指令，做出轉送封包等等相對應的動作。因此，相關的網路管理及量測資訊可以集中由 Controller 或其他應用程式統一進行處理。然而，SDN 的解決方案僅僅只有在 Control Plane 能夠做到客製化。針對 Data Plane 的部份，卻沒有辦法進一步做到客製化。

為了要解決 Data Plane 無法進一步客製化的問題，可程式化協定獨立封包處理器（Programmable Protocol-independent Packet Processor, P4），又稱 P4 交換器的技術可針對資料封包的格式與資料傳輸

的行為作進一步的客製化，使得開放式網路的客製化與自由化的程度更加徹底。也由於 P4交換器本身高度客製化的特性，P4交換器特別適合實作 INT 技術，用以加強帶內遙測之監控技術。

由於P4交換器區分為軟體版本的 BMv2 Model 與硬體版本的 Tofino Model。BMv2主要是以軟體方式實作 P4的各種特性。而 Tofino Model 主要是透過 Board Support Package（BSP）的套件與供應商的硬體進行整合，透過特殊應用積體電路（Application Specific Integrated Circuit, ASIC）來發揮硬體的效能。不過，礙於 P4交換器本身硬體成本較為高昂，一般學術研究單位難以大量建置與部署 P4交換器網路。當網路交換器的特性需要以大量巨觀方式來衡量，便難以呈現其優點。舉例來說，在無線網路的 Media Access Control（MAC）Layer 中，改良式的 MAC 演算法往往需要在節點數量與網路連線量眾多時，並且進行長期的觀察，才能夠得到其效益。

數位孿生（Digital Twin）是一種以數位方法來建置真實世界系統的解決方案。透過數位方式建構模型，藉以真實反映出實際問題、真實運作情況。另一方面，由於網路連線與資料傳輸之行為往往是隨機程序（Stochastic Processes）。因此，在數為孿生的系統中仍需要客製化的隨機模擬參數來來建構此模型，使得模型本身更能夠反映出各種情境下的特性。

本論文主要的目的是發展出 P4與 INT 的數位孿生模型。藉由數位孿生模型，我們擬建構出：

- 大量 P4交換器之實驗網路
- 各種不同 topology 的網路情境
- 客製化封包格式與操作
- 各種不同路由策略
- 不同的傳輸與處理成本的路徑

本論文的組織架構如下。第2.節主要針對 P4的技術進行說明。在第3.節中，我們針對 INT 技術進行描述。在第4.節中我們提出 P4+INT 數位孿生系統的元件規劃。針對數位孿生的議題與挑戰會在第5.節中述明。而結論與未來工作會在第6.節中闡述。

2. P4 可程式化網路交換器與 P4程式語言

可程式化協定獨立封包處理器（Programmable Protocol-independent Packet Processor, P4），又稱 P4交換器。而用於開發 P4交換器使用的程式之語言，稱之為 P4語言[1]。P4語言的版本有分 $P4_{14}$與 $P4_{16}$。如圖2.所示，$P4_{14}$與 $P4_{16}$之間並不只是版本的差異，而是包含核心的函式庫與架構的配置規劃，使得 $P4_{16}$更容易讓使用者入門以開發更多客製化功能。

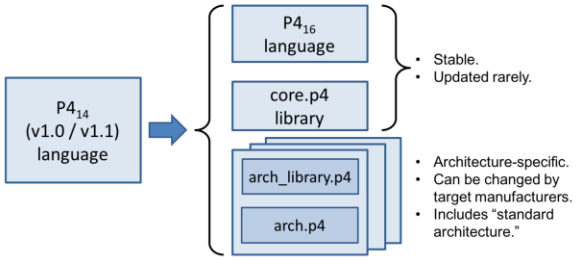

圖2. $P4_{14}$與 $P4_{16}$的差異

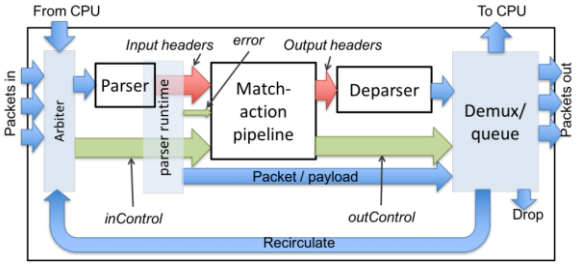

圖3. 典型的 P4交換器架構（資料來源：p4.org 官方規格文件）

如圖3.所示，在P4程式被執行的狀態下，當封包進入P4交換器時，會首先經過Parser進行剖析，並進行到 Match-Action 的 Pipeline 當中進行處理。換句話說，交換器會針對封包本身的來源 IP、目的 IP、入口 port、出口 port 等等與內部 flow table 進行比對，如果相符（Match）則進行相對應的動作（Action）。接著，封包會經過 Deparser 進行包裝，並針對剛剛 Match-Action 的結果將封包傳送到指定的出口 port。

P4交換器有區分為 BMv2 Model 與硬體版的 Tofino Model。在2025年年初，Intel 公司將 Tofino Model 的 Source Code 釋出，讓更多學研界的研究人員可以據此投入 P4相關應用之研究開發。

相關資訊可以在 P4的官方網站取得。其中：

- P4 語言：定義 P4語言的規格
- P4 Portable NIC Architecture (PNA)
- P4 Portable Switch Architecture (PSA)
- P4 Runtime：用以定義 P4交換器的 Control Plane。
- In-Band Network Telemetry：帶內遙測網路的架構
- Telemetry Format：帶內遙測的格式，於下一節中詳述。

3. 帶內遙測技術

而帶內遙測技術（In-band Network Telemetry，INT）[2]則是一種以資料封包為主，將在資料封包傳輸時便可以同時收集各大節點的資訊。INT 能儘量做到在不增加網路額外負擔的前提下，進行網路資訊收集與效能量測。由於需要針對資料封包進行處理，因此採用 P4交換器來搭配 INT 技術，是相當適合的組合。

在 INT 技術當中，交換器可以區分為 INT Source、INT Transit，以及 INT Sink。其中，在 INT Source 中一般是針對資料封包附加指令或節點資訊

的起始點。INT Transit 是負責轉送並附加自身資訊。而 INT Sink 則是用以將資料拋出到區域網路內的 Monitor 或 Collector 的結束點。不過，在 INT 當中，節點本身可以有兩種以上不同的角色。

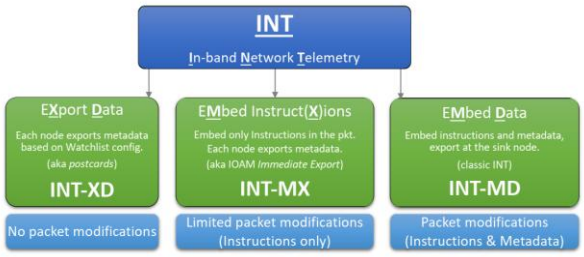

圖4. In-band Network Telemetry 的三種模式（資料來源：p4.org 官方規格文件）

由於在 INT 帶內遙測技術當中，會有 INT-MD、INT-MX，以及 INT-MD 三種模式。搭配上述三種不同的角色，其操作方式如下所述：

- 匯出資料模式（Export Data INT-XD）：各大 INT 節點依據他們所屬 Flow 的 Watchlist 所配置的指令，直接將 Metadata 匯出至 dataplane 的監控系統。在這種狀況下，完全不需要針對資料封包進行修改。這種方式也稱為明信片模式（Postcard mode），亦即各大節點透過類似明信片的方式將資訊寄回給監控系統。
- 內嵌指令模式（Embed Instructions INT-MX）：從 INT Source 將 INT 指令內嵌到封包中，接著 INT Source、INT Transit，以及 INT Sink 均會依據封包內嵌之指令將 Metadata 傳送到監控系統中。最後 INT Sink 會將指令自封包中抽離，並將純粹的資料封包傳送到接收端。在這種模式下，資料封包的大小並不會因為途中傳送的 Transit 個數增加而膨脹。
- 內嵌資料模式（Embed Data，INT-MD）：在這個模式下，INT 的指令與 metadata 都會被寫入封包當中。這是最典型的方式：INT Source：將指令嵌入，而 INT Source 和 Transit 將 metadata 嵌入到封包中，然後 INT Sink：會從資料包中剝離指令和 metadata，並「選擇性地」將資料傳送到監控系統。在此模式下，封包的修改程度最大，同時最大限度地減少了監控系統整理來自多個 INT 節點報告的 overhead。

4. P4-INT 數位孿生平台規劃

在 P4-INT 的數位孿生平台中，假設有一個網路 N 包含 switch 的集合S、host 的集合$\boldsymbol{H}$，以及連結的集合$\boldsymbol{L}$，則$\boldsymbol{N} = \{\boldsymbol{S}, \boldsymbol{H}, \boldsymbol{L}\}$為一個集合。

在集合 L 中分成兩種不同的 link，假設是lss和lsh。其中每一個 link：

$$lss = \{lssID, s_{src}, s_{dest}, cost\}$$

表示 switch 與 switch 之間的 link。其中$lssID$是 link 的識別碼，s_{src}, s_{dest}均為 switch 集合中的元素，且分別表示來源交換器與目標交換器。而 cost 則表示 switch 之間的連通成本。另外，為了要表示交換器之間的連通性與連通成本，我們可以採用 adjacent matrix：

$$M = \begin{bmatrix} m_{0,0} & \cdots & m_{0,p} \\ \vdots & \ddots & \vdots \\ m_{p,0} & \cdots & m_{p,p} \end{bmatrix}$$

其中每一個$m_{i,j}$: $\begin{cases} = 0 & not\ connected \\ > 0 & cost \end{cases}$

利用上述的 adjacent matrix 就可以定義出網路中交換器之間連通成本或者是不連通。

而在每一部 switch 中，假設 switch 本身有個 port，則針對 switch s 可以定義：

$$s = (swID, \boldsymbol{P}, src, transit, sink)$$

其中，$swID$ 是 switch 的識別碼。$\boldsymbol{P}$ 是一個 port 的集合，用以表示每個 port 的頻寬數量。而 src、$transit$ 以及 $sink$ 的數值為0或1，分別代表其在 INT 中的 source、transit 與 sink 的角色。如果該數值為1則表示此 switch 有扮演此一角色，反之則無。

而在網路的每一部 host 中，假設 host 本身都介接在 switch 上，則針對 host h 可以定義：

$$h = (hostID, swID, ipAddress, macAddress)$$

其中 $hostID$ 是 host 本身的識別碼，$swID$ 是 host 所介接的交換器。而 ipAddress 與 macAddress 分別是是 host 本身的 IP 位址與 MAC 位址。

另外一種 link，記為：

$$lsh = \{lshID, swID, hostID, sportID, hportID\}$$

其中$lshID$是 link 的識別碼，$swID$是 switch 的識別碼，$hostID$是 host 的識別碼。$sportID$和$hportID$分別是 switch 與 host 上的 port 的編號，用以記載 switch 與 host 分別透過哪一個 port 介接。

5. 開發數位孿生平台之議題與挑戰

5.1 多工模擬

在數位孿生平台中最大的挑戰就是多工模擬的部份。由於網路當中的傳輸是隨機且同時發生，但一般的模擬程式僅能以單一時間軸的方式順序模擬一系列的資料流傳書。

本研究擬採用動態調整單位時間長短的概念來模擬網路內資料傳輸的現象。在這樣的方式中，我們可以假設在同一個單位發生許多傳輸的 data flow，或者僅僅只有單一 data flow。而在同一單位時間內發生的事件則以隨機方式決定優先順序。

5.2 隨機分佈模型

由於網路中的 data flow 是隨機發生，因此必須套用不同的隨機分佈模型到我們的實驗平台當中，如 Normal distribution、Poisson distribution 等等，用以模擬更多的隨機情境。

5.3 動態相鄰矩陣

在第4.節中，我們提到用相鄰矩陣的方式來實作交換器之間連線的成本。不過，在實驗平台實際運作後，便會遇到另外一個挑戰，就是相鄰矩陣中的連線成本數值會隨時改變。

6. 結論與未來工作

在本論文中，我們針對能夠改善傳統網路管

理、監控與效能量測的INT技術進行介紹。在傳輸資料的過程中，就可以同時收集交換器的資訊。而P4可程式化網路交換器具有Data-Plane客製化的特性，相當適合實作 INT 技術。為了要針對大型P4-INT 實驗網路進行模擬，我們展示數位孿生的構想與規劃。未來我們擬真正實作出實際數位孿生的模型，用以針對 P4-INT 網路進行模擬，並以本中心建置的 P4實驗網路進行實驗，用以驗證數位孿生模型與實驗系統之間的真實性與還原度。

參考文獻

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

[2] In-band Network Telemetry, https://github.com/p4lang/p4-applications/blob/master/telemetry/specs/INT.mdk