TWAREN 可程式化實驗網路平台建置

周大源	黃文源	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	財團法人國家實驗研究院國家高速網路與計算中心	財團法人國家實驗研究院國家高速網路與計算中心	財團法人國家實驗研究院國家高速網路與計算中心	財團法人國家實驗研究院國家高速網路與計算中心
1203053@narlabs.org.tw	wunyuan@narlabs.org.tw	2103081@narlabs.org.tw	0303118@narlabs.org.tw	tlliu@narlabs.org.tw

摘要

本文介紹本中心於今年度（111年）在 TWAREN 學研網路上所建置的P4可程式化實驗網路平台。P4可程式化實驗網路，是軟體定義網路的延伸，針對網路封包格式，以及處理網路封包的相關程序均能以撰寫程式碼的方式來定義，比軟體定義網路具有更高的客製化特性。在開放網路社群中，有許多單位陸續投入P4可程式化網路的研究。然而，一般實驗室因為經費因素，往往只能針對軟體版本 BMv2 Model 進行研究。因此，本中心在 TWAREN 學研網路上擇取數個節點建置佈建多部 Tofino Model 之實體 P4可程式化網路交換器。透過 TWAREN 提供的 VPLS 串接，可形成大型遠距之 P4可程式化實驗平台，未來將可提供學研界申請使用。透過這樣的平台進行實驗，可以讓P4可程式化網路的技術在大型實體網路上得到實際的驗證。

關鍵詞：可程式化網路, 帶內遙測技術, TWAREN, P4, In-band Network Telemetry (INT)

I. 簡介

近年來，新一代的資訊技術蓬勃發展，包含人工智慧(Artificial Intelligence)、區塊鏈(Block Chain)、雲端運算(Cloud Computing) ，以及大數據(Big Data) 等等相關領域與應用與日遽增。在這些先進資訊技術，至關重要的是高速而穩定的網路建設。為了要能夠強化內部資訊傳遞能量與速率，並與國際介接，各國紛紛建置高速骨幹網路以串接各大學研界與業界組織。然而，為了要能夠管理規模日漸複雜的各類網路架構，網路的通訊協定與管理技術是相當重要的。

現行網路多數以TCP/IP通訊協定為基礎，將協定實作在硬體晶片中，並嵌入各類網路裝置（路由器、交換器、…等等）。為了要能夠與現存網路裝置溝通，並達到相容效果，許多功能都要遵循現有的通訊協定，能夠客製化的部份相當有限。對於網路維運或管理人員而言，一旦網路發生問題，若要進行除錯與故障排除，將會是非常困難又耗時的程序。

為了要讓網路架構能夠更加開放、讓工程師能夠高度參與網路裝置之運作，達成客製化的目的，軟體定義網路 (Software Defined Network, SDN) 的技術應運而生。這項技術主要源自於 Stanford 大學的一項計畫[1]。該項計畫提出新的網路設備溝通協定，稱為 OpenFlow。如圖1所示，SDN 的技術主要是將網路中的控制平面(control plane) 與資料平面(data plane)分開。前者與後者分別為 SDN 網路控制器與 SDN 交換器。

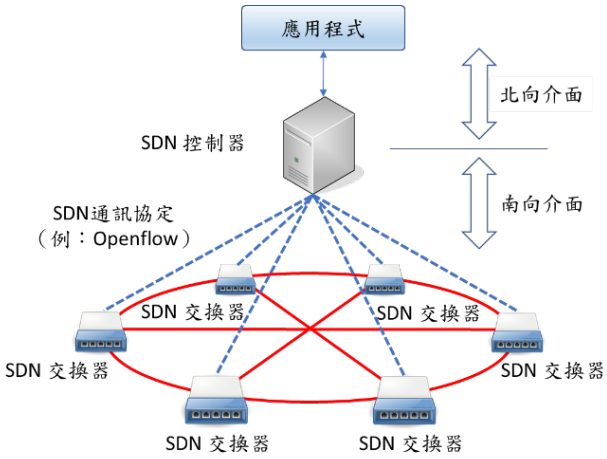

圖 1. 軟體定義網路架構圖

在 SDN 網路中，SDN 控制器通常是在另一部伺服器上，以軟體方式實作應用，透過獨立通道（一般為 TCP 或 SSL 連線）來控制多部 SDN 交換器，又稱為南向介面。在 SDN 控制器的部份，可以讓網路管理者透過 policy 或是 rule 設定，來控制資料傳輸的方式。而 SDN 交換器則是專注於資料的轉送(forwarding)。在每部 SDN 交換器會透過許多資料路徑互相連接，當收到用戶 host 傳來的資料封包時，會先查閱交換器本身內部的 flow table ，檢測是否有 match 的部份。若有 match 的 flow，則會直接套用該 flow 進行轉送。相反地，若是沒有 match 的 flow，就會透過獨立通道向控制器詢問資料處理的方法。因此，在SDN的網路中，只要控制器與交換器均遵循同樣的 SDN 通訊協定(例如 Openflow)，便可以讓控制器順利對交換器進行管控。如此一來，SDN 控制器與 SDN 交換器並不限制一定要同一種廠牌。再者，使用者可以透過北向介面的各種應用來針對 SDN 進行控制，達成控制面的客製化功能。

雖然網路管理者能夠針對控制平面的部份進行客製化管控，並能夠實作各類型的第三方應用(3rd -party

Application)與控制器溝通，但對於資料平面，亦即 SDN 交換器的客製化程度卻不足。為了要進一步針對資料傳輸的部份進行客製化與可程式化，可程式化交換器 (簡稱 P4) 的技術便應運而生。

P4可程式化網路交換器[2][3] (Programmable 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 P4的行為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。換句話說，程式開發者可以藉由程式語言的撰寫，打造更加符合需求的交換器。有關於 P4程式語言相關工具，可以參考 P4 Language 的網站 [2]，在網站上有詳細的安裝過程。一般使用者可以依照該資訊逐步安裝。另外，如果想要省略這個部份，也可以下載P4官方網站所提供的虛擬機器版本 tutorial 來進行測試。此部份之 P4交換器為軟體版本之 BMv2 Model。

P4交換器亦有開發硬體版本。在2017年，Barefoot Networks 公司開發出支援 P4 可程式化的交換器特殊應用晶片（Application Specific Integrated Circuit，ASIC）Tofino[4]，並於 2018 年改良為 Tofino 2[5]，亦可稱之為 Tofino Model 與 Tofino 2 Model。目前已有多家網路硬體公司推出商用的可程式化交換器，並採用 Tofino/Tofino 2 ASIC，讓網路開發者可以直接撰寫 P4 可程式化網路程式，讓網路功能具有高度客製化的特性。在2019 年 Barefoot Networks 加入 Intel，成為 Intel 的可程式化網路技術團隊[6]。

事實上，若針對軟體版本P4可程式化實驗網路進行實驗，最基礎的方式是利用 Mininet 搭配軟體版本的 P4 交換器來進行模擬實驗。利用同一部電腦即可模擬出具有多部 P4 可程式化交換器與多部客戶端的完整模擬網路。然而，這樣的環境無法完整反映出實際網路的情況。特別是在進行網路效能量測、帶內遙測（ In-band Network Telemetry）等等技術時，時沒辦法反映出與傳輸時間相關的真實數據。

然而，一般硬體版本的P4可程式化交換器價格較為昂貴，對於一般大專院校實驗室的經費而言往往較無法購買。因此，一般國內大專院校多以軟體版本的 BMv2 model 為主要研究目標。再者，基於這樣的動機，本中心擬以台灣先進學術研究網路為基礎，搭配硬體版本之P4可程式化交換器以建構P4可程式化交換器實驗網路。

台灣先進學術研究網路 (TaiWan Advanced Research and Education Network) 是串接國內外多個學研單位的寬頻網路。TWAREN 本身是由國網中心維運管理，提供多種7x24小時服務。不但肩負起各大學研單位間的大資料高速傳輸，本身亦能夠架設許多網路相關的實驗測試平台。在105年度，本中心進行 TWAREN 100G 骨幹升級作業[7]。圖 2 為 TWAREN 100G 線路架構，包含5個骨幹(Core)節點與12個區域網路(GigaPop)節點。在5個主節點之間線路頻寬均為100G，而區域網路節點之間則為50G。更多 TWAREN 相關資訊可以參閱 TWAREN 網站[8]。

圖2. TWAREN 100G 網路架構圖

為了更進一步地研究軟體定義網路技術，本團隊持續在 TWAREN 上建置國內大型的 SDN 實驗場域，亦即在4個骨幹節點與12個區網節點佈建 SDN 交換器。其間之資料路徑由TWAREN骨幹的VPLS線路達成。TWAREN SDN 網路架構如圖3所示。透過開放使用實驗性質服務方式，以實際使用案例的方式驗證 SDN 應用，例如：

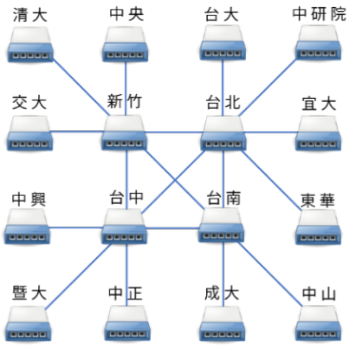

圖 3. TWAREN 虛擬專用連線 topology

- 本中心台中分部與美國西北大學 iCAIR 實驗室之間的 DTN 實驗線路。
- 與國立交通大學共同進行 SDN-IP 網路介接，並與日本、韓國等多個學研單位進行連線。
- 106年8月，本中心提供 TWAREN SDN 虛擬專用連線，與使用與中華電信研究院、工研院、資策會等單位一同進行世大運影音轉播[9][10]。本團隊亦開發一套 Web UI 版本的虛擬網路供裝系統[11]，並進行高速傳輸測試[12]。

基於上述經驗，本中心據以建立未來將進一步地於

TWAREN 之骨幹節點與區網節點擇取數個節點來進行硬體可程式化交換器佈建工作，提供國內大型遠距線路實驗場域。

本論文的主要內容如下。第 II 節介紹可程式化交換器與帶內遙測技術。第 III 節會針對我們今年建置的 TWAREN 可程式化實驗網路的整體架構進行進行解說。在第 IV 節中展示資源預約系統，提供國內學研界單位申請使用。而最後一節則是結論與未來展望。

II. 可程式化網路交換器與帶內遙測技術簡介

A. BMv2

可程式化交換器[2][3] (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 P4的行為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。

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。

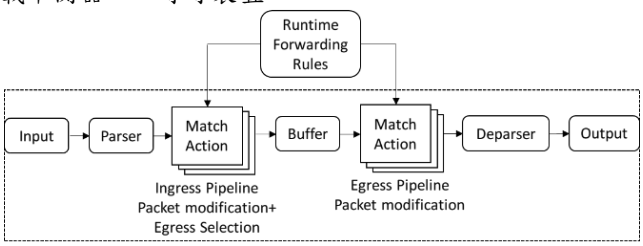

圖4. P4 Pipeline

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，封包轉送至 Egress 進行 match + action，並再修改資料封包後輸出。

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[13]文件，讓程式開發者能夠針對重要的關鍵語法進行參考。

B. 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。

C. 帶內遙測技術

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。

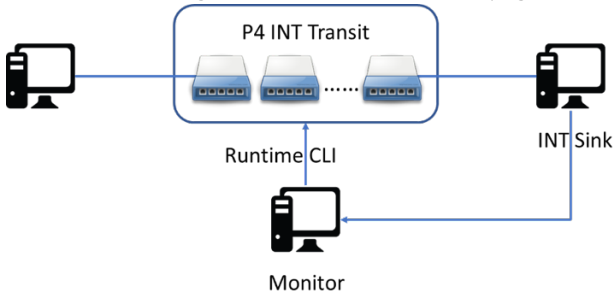

圖5. 典型的 P4 INT 框架

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。

III. 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。

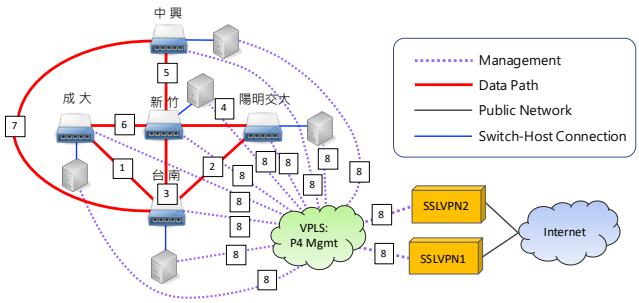

圖6. P4 可程式化交換器實驗網路

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。

表 I 各節點與 TWAREN 介接介面

編號	起點	迄點	類型
1	台南	成大	10G SR
2	台南	陽明交大	10G SR
3	新竹	台南	10G SR
4	新竹	陽明交大	10G SR
5	新竹	中興	10G SR
6	新竹	成大	10G SR
7	台南	中興	10G SR
8	台南	SSLVPN	1G RJ-45
8	陽明交大	SSLVPN	1G RJ-45
8	成大	SSLVPN	1G RJ-45
8	新竹	SSLVPN	1G RJ-45
8	中興	SSLVPN	1G RJ-45
8	SSLVPN		1G RJ-45

如表 I 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。

IV. 資源預約系統

本節展示我們針對 TWAREN 可程式化交換器實驗網路之預約系統。

由於P4可程式化實驗網路的相關主機是單工且是獨佔式平台，因此若有國內學研單位需要使用 P4可程式化實驗網路平台時，需要透過預約系統進行獨佔式預約。預約作業流程大致區分為以下幾個步驟。

- 由使用單位提出申請：填寫申請人姓名、所屬之使用單位名稱、聯絡電話、聯絡用的 E-mail、希望使用哪些主機、預期使用時段。除此之外，使用單位亦須提出計劃書，或者簽署 TWAREN 學術網路切結書，確保申請資源用於學術用途。
- 系統寄發通知：系統寄發 E-mail，一方面是讓使用者掌握一份申請紀錄。而另一方面則是寄發通知 E-mail 給本中心系統管理者，說明目前已有新申請之使用者，等待處理。這樣，本中心管理者便能透過聯絡資訊與使用單位聯絡。
- 由本中心小組審核：確認並安排可用時間與可用主機，為使用者準備連線測試環境。
- 寄發通知：SSLVPN 程式（或連線用金鑰）、使用者帳號密碼、相關 IP 位址、可使用時間等等。

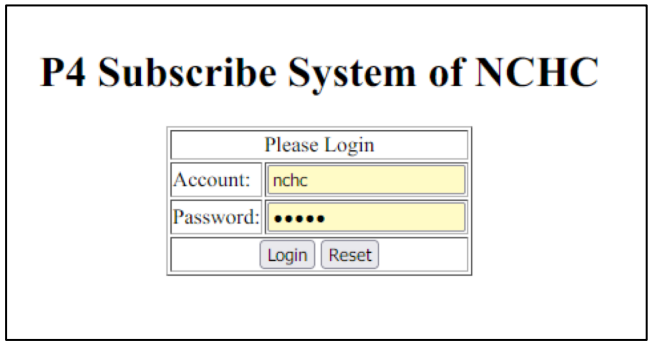

圖7. 帳號密碼驗證

圖7.至圖9.是我們著手開發之預約系統。如圖7.所示，學研單位在通過審核、接到本中心通之後，將得到一組使用者帳號與密碼，可登入本預約系統進行常態性預約。

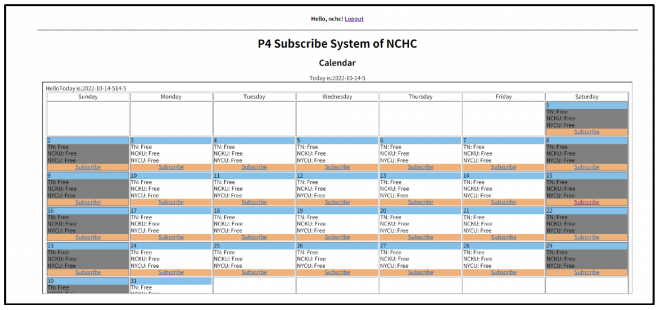

圖8. 以月曆形式顯示目前使用狀況

如圖8.所示，在使用者成功登入系統後，會以月曆方式顯示日期與可用資源。使用者可以依據需求，並挑選未被佔用的時段以進行預約。

圖9. 資源預約介面

如圖9.所示，使用者選定某時段進入後，則可視其需要以勾選所需之節點進行預約。在按下 Subscribe 後，將由本中心進行管理分配，確保各時段為合法申請之使用者獨佔使用。而在使用者使用完畢後，我們也會將 P4交換器的環境恢復為預設狀態，以便下一位使用者進行使用。

V. 結論與未來展望

本論文說明國網中心於今年度佈建於 TWAREN 上的實體可程式化實驗網路平台。現階段我們選定於國網中心台南分部、陽明交通大學、成功大學、國網中心新竹本部，以及中興大學。透過 TWAREN 所提供的 VPLS 服務，我們將各節點的可程式化網路交換器串接起來，打造實體之網路實驗平台。未來我們將開放這個可程式化網路交換器平台供學研界申請租用，期能藉由此平台為國內學研界增進可程式化實驗網路的研究能量。

參考文獻

[1] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, J. Turner, "OpenFlow: enabling innovation in campus networks," ACM SIGCOMM Computer Communication Review, vol. 38, no. 2, pp. 69-74, April 2008.

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

[3] A. Sivaraman, C. Kim, R. Krishnamoorthy, A.t Dixit, and M. Budiu, "DC.p4: Programming the Forwarding Plane of a Data-Center Switch," Proceedings of the 1st ACM SIGCOMM Symposium on Software Defined Networking Research, Article No. 2, 2015.

[4] Barefoot Networks looks to redefine ASIC in network device design with its Tofino chip, https://www.techrepublic.com/article/barefoot-networks-looks-to-redefine-asic-in-network-device-design-with-its-tofino-chip/

[5] Barefoot Networks Debuts Tofino 2, Using 7nm Technology, https://www.sdxcentral.com/articles/news/barefoot-networks-debuts-tofino-2-using-7nm-technology/2018/12/

[6] 英特爾收購 SDN 及晶片新創業者 Barefoot Networks, https://www.ithome.com.tw/news/131215

[7] 台灣 100G 教育學術研究網路正式啟用 http://technews.tw/2016/10/06/taiwan-100g-education-network/

[8] 台灣先進學術研究網路(TaiWan Advanced Research and Education Network, TWAREN), https://www.twaren.net/

[9] 無線轉播無限精彩 世大運賽事試驗轉播 精彩賽事不累格 打造未來直播新境界 https://www.cna.com.tw/postwrite/Detail/219194.aspx

[10] 周大源, 胡仁維, 黃文源, 劉德隆,TWAREN SDN 虛擬專用連線管理系統," TANet2017集，台中，2017年10月

[11] 周大源, 胡仁維, 黃文源, 劉德隆, "WebGUI 版本虛擬網路供裝系統," 2016 年雲端與大數據研討會, 2016

[12] 周大源, 楊哲男, 古立其, 劉德隆, " TWAREN SDN 虛擬專用連線之高速傳輸應用," TANet2016論文集，花蓮，2016年10月

[13] P4 Language Cheat Sheet, https://p4.org/assets/P4WS_2018/p4-cheat-sheet.pdf