The Implementation of High Performance Data Transfer over TWAREN International Connectivity

楊哲男　古立其　劉德隆
國家高速網路與計算中心
{yangcn, lku, tlliu}@narlabs.org.tw

摘要

大資料分析及深度學習人工智慧相關技術的逐漸成熟，成為推動科學研究方法向前演化的又一巨大動力。為了呼應大資料研究對於巨量網路頻寬的需求，台灣的學術網路骨幹於2016年升級至100G，使各 GigaPOP 大學獲得高達100G 的直接連網頻寬。然而傳統的網路傳輸方法無法克服頻寬延遲的瓶頸，僅能使用大頻寬網路的極小部份。本研究在 TWAREN 骨幹佈建資料傳輸節點以克服上述瓶頸，充份發揮大頻寬優勢。本文將敘述資料傳輸節點之相關技術，並實際於 TWAREN 長距離之跨洋國際頻寬上透過專屬資料傳輸主機進行大資料傳輸之實作及效能分析。

關鍵詞：DTN、TWAREN、大資料傳輸、網路效能。

Abstract

The blooming of big data science and deep learning AI technology provides a tremendous driving force to scientific researches. To address the resulting needs for huge network bandwidth, the academic backbone and the available bandwidth of its GigaPOP universities have been upgraded to 100G in 2016. However the traditional way of network transmission can only utilize a small fraction of such bandwidth. In order to overcome this bottleneck, this study has implemented the data transfer nodes on TWAREN backbone to fully take the advantage of the 100G backbone. This paper introduces this technology, the implementation of it in TWAREN international circuit and the analysis of the performance gain it has achieved.

Keywords: DTN, TWAREN, Bulk Data Transfer, network performance.

1. 前言

當大量的資料想快速地透過網路在兩點之間傳輸，需要同時具備多種條件的配合方可以達到此目的，首先是高速的骨幹網路，再來是快速的儲存設施，以及高效能的傳輸工具。但往往受限於某種條件，使用者端會利用交通運輸之方式運送磁碟至遠地，然後在拷貝至當地之儲存設備。如何讓使用者方便快速的傳輸檔案將是一大挑戰。各國際學術研究網路皆投入許多經費提昇骨幹及對外頻寬，也帶動許多科學計算領域的發展，包含高能物理(High Energy Physics)、天文物理(Astrophysics)、生物資訊(Bioinformatics)及地球科學(Earth Science)等，這些科學資料利用高品質網路進行跨單位與跨國際間的傳送，創造了豐碩的科學研究成果。此外，隨著大資料分析及人工智慧的蓬勃發展，巨量資料的搬移之需求日益增加，已有許多研究[1][2]在探討如何快速的搬移資料。

美國能源部網路(ESnet)在2010年發展出Science DMZ[3]，其目的為如何讓科學資料能夠最佳化的在廣域網路設傳輸的設計方法。Science DMZ 整合了三大重要關鍵因素，來達成快速傳遞資料之目的，包含了適合於 High-performance 應用之專用網路、擁有專屬的資料傳輸機器以及良好的效能量測機制，可隨時量測點對點間之網路，以確保網路品質良好。其中專屬的資料傳輸機器我們稱之為之資料傳輸節點(Data Transfer Node, DTN)。在本文中，將敘述建置資料傳輸節點之相關技術，並透過建置於 TWAREN 跨洋間之資料傳輸節點進行大資料傳輸測試，並與傳統之傳輸方式進行比較分析。在第二章節中我們將介紹 DTN 之相關技術及建置 DTN 時所需注意事項。在第三章節中我們將介紹 TWAREN[4]現有之國際骨幹。在第四章節中我們將描述本文之測試架構，並在第六章節中說明測試結果分析，最後為本文結論。

2. Data Transfer Node

DTN 在美國能源部[5]的定義為:一種在廣域網路傳輸上為了得到更好的資料傳輸效能所建立的系統。DTN 通常是以 PC-based 為主的 Linux 伺服器所組成，此外為了達到內部傳輸資料快速之目的，此伺服器通常無其他一般用途之服務，例如網頁服務、多媒體服務或一般文書處理，本章節中我們將詳細介紹建置 DTN 時所需注意之事項。

High-performance 通常指的是10Gbps 以上之網路效能。目前 TWAREN 骨幹已升級至100Gbps，因此連線單位連接至骨幹10Gbps 的連線可輕易達到，但要達到10Gbps 之檔案傳輸效能，則需要透過網路參數優化以及優良的網路品質才可達成。

若 DTN 未來需要支援到40G 以上之等級，主機須至少有 PCIe Gen3以上之介面，新一代的主機通常皆已支援。此外為了能安裝多顆 disk 或 SSD 於單一主機上，最好能安裝24個(以上)之磁碟槽。包含主機 Bios、CPU、IRQ、儲存、網路、檔案系統、應用軟體等。

2.1 儲存(Storage)

如**圖1** 所示，DTN 系統之儲存可分為本地端儲存空間(Local storage)及網路儲存空間(Networked storage)。本地端儲存空間為了達到快速及提高可用性，一般會用磁碟陣列(RAID)來達成，例如以

RAID5、RAID6或 RAID10。在磁碟的選擇上又可分為使用傳統 SATA 或 SAS 傳統磁碟，或是使用 SSD 固態硬碟方式。選擇 SATA 或 SAS 傳統磁碟的好處是較為便宜及容量較高，缺點是讀寫速度較慢，但使用多顆磁碟構成一個磁碟陣列可改善讀寫速度問題。SSD 固態硬碟之讀寫速度較傳統硬碟快，但是其價格也比較高。SSD 固態硬碟有多種介面形式，例如 NVMe PCIe 卡片形式，這是目前最快的 SSD 種類，每一張 SSD 卡之讀或寫速度可超過1GB/s 以上之速度，但因每台主機之 PCIe 插槽有限，因此一般形式之主機無法同時插多張 NVMe PCIe SSD 卡。NVMe SSD 亦有像傳統硬碟大小的形式，且讀寫速度亦可超過1GB/s 以上，使用此種 SSD 也會占用主機 PCIe lane 之資源，但可透過轉接之方式，讓多張 SSD 卡同時使用一個 PCIe lane 資源，因此有些伺服器上單一台主機可同時插上24張 NVMe SSD。使用多張 NVMe SSD 再透過磁碟陣列之方式可以達到每秒40Gb/s 甚至是100Gb/s 以上之速度。

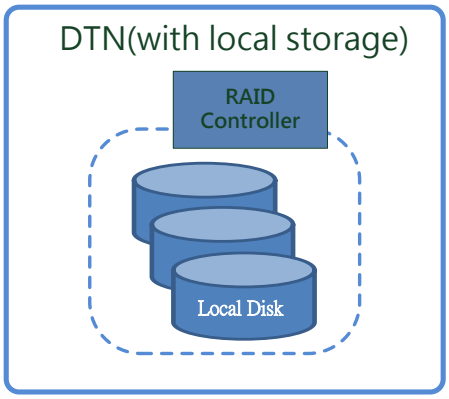


圖 1 DTN Local Storage

如**圖 2**所示，網路儲存空間可透過高速 Ethernet 或是 InfiniBand 連至其他高容量之高速平行檔案系統作為外部儲存空間，例如 NFS(Network File System)[6]、GPFS(General Parallel File System)[7]或是 Lustre[8]檔案系統。因為空間及其使用上的彈性，網路儲存空間是大部分 DTN 平台所使用之方式，但若只是單純的作為短暫儲存之用或是想快速地做資料搬移，則可使用本地端儲存空間做為資料暫存之用。

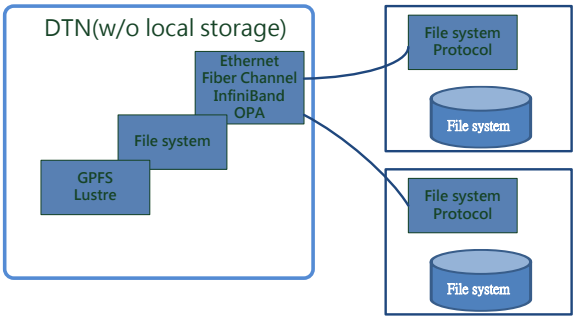


圖 2 DTN Network Storage

2.2 網路(Network)

在目前的 DTN 環境上所使用的網路通常為 10Gbs 以上頻寬，甚至是40Gb/s 以上之廣域網路環境，部分國際研究網路還透過 VPN 或是 Virtual Circuit(VC)技術規劃專用傳輸骨幹，作為點對點間之傳輸骨幹，此一方式不僅保證傳輸品質，更提升了傳輸之安全性。

2.3 主機(Chassis)

由於 DTN 會使用到高速網卡以及 PCIe NVMe SSD，這些使用 PCIe lane 資源之硬體至少需支援 PCIe Gen3以上之版本方可滿足其高速需求，新一代的主機通常皆已支援。此外對於資料傳輸之需求，CPU 之時脈(clock rate)也是相當重要，在 DTN 系統中 CPU 時脈比 CPU 核心數量更為重要。在記憶體的部分則是建議至少有32GB 以上，且越多越好。

2.4 系統調整(Tuning)

DTN 之系統調整在整個傳輸平台中佔有相當重要之工作。當 DTN 系統有了適當之硬體設備，為了發揮最佳之傳輸效能，必須進行系統調整，其中包含了主機 Bios、CPU、IRQ、儲存、網路、檔案系統及應用軟體等，當中網路部分的調整最為重要。一般所常用的 TCP 網路傳輸控制協定，已經成為大檔案資料於網路傳輸之瓶頸，若有需要使用 TCP 作為傳輸協定，必須做 TCP tuning 之工作，以增加傳輸效能，包含了 window size 值、MTU 以及 congestion control Algorithms 等等，不同的作業系統有不同的設定方式。在 DTN 傳輸平台間，MTU size 必須支援 jumbo frame，方可發揮大頻寬之優勢。在[3]有詳細說明如何進行調整。在經過系統調整之狀況下，其傳輸速度可超過原有速度之2至3倍以上。

2.5 檔案傳輸工具

在一般情況下，使用匿名式的 ftp 或是 http 方式做檔案傳輸的情形是最多的，此種方法若用於傳輸一個較小的檔案其速度跟時間還可以接受，但若是傳送一個較大的檔案時，通常需要較長的時間才可以傳完，其效能往往不能被使用者接受。其他使用一些需要帳號密碼的工具，例如 lftp、sftp、scp 或 rsycn，這些工具有的是認證時需要加密，有的是連傳輸過程都需要加密，因為加密需要額外的資源來做處理，因此也常發現其效能跟實際的可用頻寬相比，其所得到之結果也是不盡理想的，特別時當 RTT 較高時，影響特別的大。因此在 DTN 傳輸工具的判斷上可以朝幾個方向來當作選擇傳輸工具的依據：

- 選擇可以平行傳輸(parallel streaming)的工具：
  當傳輸工具可以將檔案做切割來平行傳輸，或是可以同時間傳送多個檔案的方法，都可以增加 TCP 傳輸的效能，一般都建議以4個 streams 平行傳輸為最佳。
- 選擇可以修改、調大 buffer size 的工具：
  由於作業系統的 window size 值會影響傳輸工具裡的 buffer size 值，若我們調整完作業系統之 window size 值，但沒有調整或是無法調整傳輸工具之 buffer size 值的話，傳輸效能還是會受到影響

的。目前已有許多傳輸工具可以選擇 buffer size 值，例如 scp、rsync、GridFTP、bbftp、bbcp…等等。其中 scp[9]，全名為 Secure Copy，是一種以 Secure Shell(SSH)協定為基礎，並讓資料可以在兩台主機上加密傳輸的工具。它是一個強大的工具，可以有效且安全的將資料從遠端的檔案複製到本地端，在 Linux 或是 windows 作業系統皆有支援之工具來使用。scp 預設是無法修改 buffer size 值的，但經過目前已有之 hpn-scp 版本，它是利用修改 openssh 的原始碼，除了增大了預設的 buffer size 外，亦可直接輸入自訂之 buffer size 值，大幅增加 scp 工具的傳輸效能。

- 選擇可以不加密的工具：

在 Linux 作業系統最常被使用的傳輸工具就是 scp 及 sftp，最主要是因為他支援了傳輸加密，也就是可以確保傳輸的過程中之安全性，但這往往會犧牲了傳輸的效能，也因此若我們已經確保欲傳送的資料之機密性沒那麼高時，我們可以建議使用非加密的傳輸工具。若欲使用 scp 的話，目前亦有需安全認證，但可不加密的傳輸方法。

- 選擇利用 UDP 或是 UDT 傳輸協定的工具：

因為 TCP 的一些防止擁塞的機制，造成 TCP 成為了大檔案傳輸的瓶頸，因此選擇 UDP 為基礎的傳輸工具是一項不錯的選擇，但因為 UDP 為一種不可靠的傳輸協定，因此後來有人撰寫了另一種傳輸協定，稱為 UDT(Udp-based Data Transfer)。他是一種支援可靠的傳輸工具，但又有 UDP 協定傳輸快速的優點，因此相當適合用大檔案資料傳輸、雲端運算或是遠距離，高延遲的網路下使用。

- 將許多小檔案壓縮成一個檔案做傳輸：

檔案傳輸若需要傳遞許多個小檔案時，建議可將所有檔案包裝或是壓縮成一個大檔案來做傳輸，其效能可增加許多。

在本文中我們將使用 GridFTP 及 FDT 做為測試工具。GridFTP[10]是由標準 FTP 所延伸的一種傳輸工具，由 Argonne 國家實驗室（ANL）所開發的通訊協定。它被定義在 Globus toolkit 底下的一個子工具，當初的目的是希望能在格網(grid)應用下，能夠提供可靠且高效能的檔案傳輸。GridFTP 具備數種優勢勝過其他資料傳輸系統，不同於 FTP 這種傳輸應用程式，GridFTP 使用多重資料通道來加速傳輸速度，且亦支援手動修改 TCP buffer size 大小，以獲取最大 TCP 使用效益。由於 GridFTP 預設必須搭載 CA 認證作為伺服器端及用戶端之間的認證使用，其建置步驟相當複雜，因此我們重新編譯原始碼，使其能支援 ssh，利用 openssh 來做為身分認證即可，但傳輸過程不加密。另一套傳輸工具為 FDT，全名是 Fast Data Transfer[11]，是由瑞士的 CERN 所開發的一種有效率的傳輸工具，它是由 java 程式語言開發，因此可以跨平台使用，使用者只要擁有 Java 執行環境（Java Runtime Environment, JRE）即可執行。FDT 由於免安裝，只要直接下載 fdt.jar 這個檔案，即可執行。此外，它支援平行傳輸，並且可以指定 buffer size 之大小，因此可以大幅增加傳輸效率。FDT 是一種主從式架構的傳輸，檔案方向由使用者端(client)流向服務端(server)。

3. TWAREN 國際骨幹

為因應國內各級學校網路教育、教學研究、國際交流等國際網際網路(Internet)連線服務需求，國網中心自 TWAREN 建置以來，經由國際骨幹電路的規劃設計，提供國內研究網路與國際學術研究網間相互合作、交流之平台。TWAREN 國際骨幹主要負責跨太平洋海纜與歐美國研究網路串接[12]，其架構如下**圖 3**，國內端接取點包含新竹與台北，美國端接取點包含洛杉磯、芝加哥與紐約，架構設計本身即具備備援能力，單一線路中斷並不會影響台灣國際骨幹連線之服務。

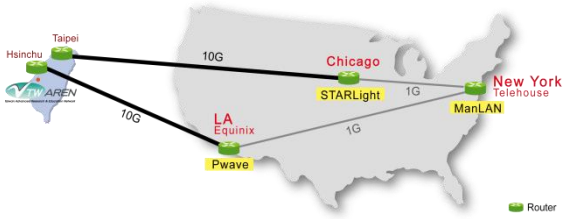


圖 3 TWAREN 國際骨幹網路架構圖

目前跨洋海纜部份包含台北-芝加哥、新竹-洛杉磯 10G 電路各 1 路，而美國國內陸纜包含芝加哥-紐約 、洛杉磯-紐約1G 電路各 1 路，主要將美國端落地點芝加哥、紐約與洛杉磯三點串連起來，分散單點或單路電路中斷的風險。

4. DTN 測試環境

圖4 為本文之 TWAREN 長距離之跨洋國際頻寬測試架構圖，芝加哥至台北主機兩點間之來回延遲時間(RTT)為186ms，頻寬為10Gbps，這種長距離傳輸對於 TCP 傳輸來說是個挑戰。我們於 TWAREN 台北主節點及美國芝加哥節點各架設一台 DTN 主機。其中兩台主機之規格如下:

- 芝加哥 DTN 主機規格
  - Chassis: Dell R730
  - CPU: Intel Xeon E5-2643 v4 3.4GHz
  - RAM: 64GB DDR4
  - HDD: SAS 1.2TB 10K RPM x 7
  - NIC: Intel X520 10Gb + 1Gb
  - OS: Centos 6.10

- 台北 DTN 主機規格
  - Chassis: Dell R730
  - CPU: Intel Xeon E5-2620 v3 2.40GHz
  - RAM: 32GB DDR4
  - SSD: OCZ PCIe SSD x 1
  - NIC: Intel X710 10Gb + 1GB
  - OS: Centos 6.10

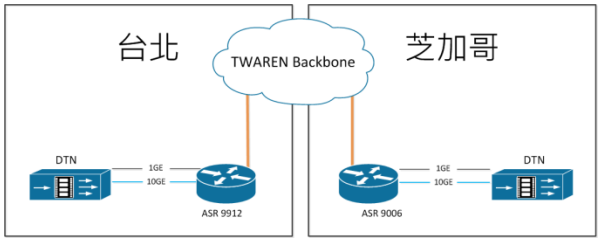


圖 4 跨洋 DTN 測試架構圖

為了確保每台主機之 TCP 效能，我們於作業系統下修改 windows size 大小，在 Centos 作業系統下，我們直接修改/etc/sysctl.conf 之設定檔即可，相關設定檔如**圖5**。我們將每個網路介面之傳送佇列數設定為10000，並將每個網路介面代號之 IRQ CPU Affinity 手動綁定在同一個 CPU 核心上，以避免 NUMA 問題[13]。

```
# allow testing with buffers up to 128MB
net.core.rmem_max = 134217728
net.core.wmem_max = 134217728
# increase Linux autotuning TCP buffer limit
to 64MB
net.ipv4.tcp_rmem = 4096 87380 67108864
net.ipv4.tcp_wmem = 4096 65536 67108864
net.core.netdev_max_backlog = 30000
net.ipv4.tcp_no_metrics_save = 1
net.ipv4.tcp_congestion_control = htcp
```

圖 5 TCP 相關優化參數

本次測試包含以下測試:

- 從芝加哥利用 scp 工具傳送100GB 檔案至台北主節點之/dev/null(Disk to Memory)。
- 從芝加哥利用 GridFTP 工具傳送100GB 檔案至台北主節點之/dev/null(Disk to Memory)。
- 從芝加哥利用 FDT 工具傳送100GB 檔案至台北主節點之/dev/null(Disk to Memory)。
- 每個傳輸工具分別測試兩次。

測試步驟:

- 確認點對點間之 MTU 是否全程支援 jumbo frame:

由於 MTU jumbo frame 對於傳輸效能影響相當大，因此首先必須透過 ping 指令可測試點對點間是否之支援 jumbo frame，**圖6** 顯示以8972 bytes 之 ICMP 封包，可以順利得到回應。

```
[sun1@tp-server1 ~]$ ping -s 8972 -M do chi-server1
PING chi-server1  8972(9000) bytes of data.
8980 bytes from chi-server1 : icmp_seq=1 ttl=61 time=186
ms
8980 bytes from chi-server1: icmp_seq=2 ttl=61 time=186
ms
8980 bytes from chi-server1 : icmp_seq=3 ttl=61 time=186
ms
```

圖 6 MTU Jumbo Frame 測試

- 使用 GridFTP 測試100GB 檔案傳輸效能:

如**圖7** 所示，我們從芝加哥主機透過 globus-url-copy 指令傳送資料至台北主機之/dev/null 裡。測試結果約為每秒900MB/s 之傳輸速度，傳遞一個100GB 之檔案約花兩分鐘之時間。若以一篇藍光光碟資料量約為40GB 來計算，傳遞一片藍光工疊資料所需的時間約為50秒。

```
[sun1@tp-server1 ~]$ globus-url-copy -vb  -p 10 -sync  \
 sshftp://sun1@chi-server1/home/100GB   file:///dev/null
      Source: sshftp://sun1@chi-server1/home/
Dest:   file:///dev/
 100GB  ->  null
 104857600000 bytes      880.28 MB/sec avg      963.19
MB/sec inst
```

圖 7 使用 gridftp 傳輸測試

- 使用 scp 測試100GB 檔案傳輸效能:

如**圖8**所示，我們從芝加哥主機透過 scp 指令傳送資料至台北主機之/dev/null 裡。測試結果約為每秒10.5MB/s 之傳輸速度，傳遞一個100GB 之檔案約花兩個半小時之時間。

```
[sun1@ tp-server1 ~]$ scp sun1@chi-server1:/home/100GB
100GB
1% 1665MB  10.5MB/s 2:35:58 ETA
```

圖 8 使用 scp 傳輸測試

- 使用 FDT 測試100GB 檔案傳輸效能:

如**圖9** 所示，我們從芝加哥主機透過 FDT 之 java 指令傳送資料至台北主機之/dev/null 裡。測試結果約為每秒7.2Gb/s(約918MB/s)之傳輸速度，傳遞一個100GB 之檔案約花兩分鐘之時間。

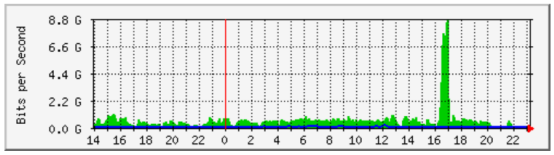
```
[sun1@tp-server1 ~]$ java -jar fdt.jar -c chi-server1 -pull
/home/100GB -d /dev/null
FDT [ 0.24.0-201512041353 ] STARTED ...

Net In: 6.166 Gb/s      Avg: 7.626 Gb/s 100.00% ( 00s )
```

圖 9使用 FDT 傳輸測試

5. 測試結果

由於本文測試之跨洋骨幹為現有 TWAREN 提供服務之網路環境，因此骨幹上還有其他網路流量。**圖10** 為測試過程中之 MRTG 流量圖，可以明顯發現最高流量達8.8Gb/s。**表1** 為三種傳輸工具之測試結果，明顯發現可以使用有平行傳輸功能以及可調整 buffer size 之傳輸工具其傳輸效能遠遠超過 scp 達到快90倍。數據顯示若用於長距離且大頻寬之檔案傳輸時，scp 不是一套適合的工具。此外 GridFTP 與 FDT 測試之結果大致相同，都是很適合用於 DTN 之傳輸工具。由測試結果得知，透過選擇適當之傳輸工具在應用於調教過後之 DTN 系統，確實可發揮高品質之傳輸效能，對於有大資料傳輸需求之海量資料確實可發揮效益。

`Daily' Graph (5 Minute Average)



	Max	Average	Current
CHI9006->TP9006	8741.8 Mb/s (87.4%)	571.9 Mb/s (5.7%)	76.2 Mb/s (0.8%)
TP9006->CHI9006	179.3 Mb/s (1.8%)	28.8 Mb/s (0.3%)	13.6 Mb/s (0.1%)

圖 10 台北至芝加哥流量圖

表1檔案傳輸測試結果

	第一次測試	第二次測試
SCP	84Mb/s	84Mb/s
GridFTP	7.042Gb/s	7.167Gb/s
FDT	7.626 Gb/s	7.26Gb/s

6. 結論

當頻寬越來越大時，選擇適當的傳輸工具以及建立良好的傳輸環境將可大幅提升傳輸效能。我們透過跨洋之 TWAREN 國際頻寬遠從台北至美國芝加哥透過 DTN，驗證能在短時間內傳遞海量資料。未來若同時結合 MPTCP 及 Open Flow 應可創造一個可自動繞境，又可達到容錯(fault tolerance)及負載平衡(load balance)之網路。透過不同技術之結合，將可提供給需要使用海量資料傳輸之將可又快又穩之網路。

參考文獻

[1] Dart, E.; Rotman, L.; Tierney, B.; Hester, M.; Zurawski, J. The Science DMZ: A Network Design Pattern for Data-Intensive Science. In Proceedings of the SC'13 International Conference on High Performance Computing, Networking, Storage and Analysis, Denver, CO, USA, 17–21 November 2013.

[2] Toward a smart data transfer node, ZhengchunLiu, Rajkumar Kettimuthua, Ian Foster, Peter H.Beckmana, Future Generation Computer Systems Volume 89, December 2018, Pages 10-18.

[3] Science DMZ Network Architecture. Available online: http://fasterdata.es.net/science-dmz/

[4] Taiwan Advanced Research and Education Network (TWAREN), http://www.twaren.net/

[5] Energy Sciences Network, http://es.net/

[6] NFS, https://en.wikipedia.org/wiki/Network_File_System

[7] GPFS, https://en.wikipedia.org/wiki/IBM_Spectrum_Scale

[8] Lustre, https://en.wikipedia.org/wiki/Lustre

[9] Secure copy, http://en.wikipedia.org/wiki/Secure_copy

[10] GridFTP, http://www.globus.org/toolkit/docs/latest-stable/GridFTP/

[11] FDT, http://monalisa.cern.ch/FDT/

[12] 林書呈, 先進研究網路發展趨勢與 TWAREN 國際骨幹未來展望, TANet2015.

[13] Collin McCurdy, Jeffrey Vetter, Memphis: Finding and fixing NUMA-related performance problems on multi-core platforms. 2010 IEEE International Symposium on Performance Analysis of Systems & Software (ISPASS)