HTTP 與 FTP 訊務在 TANET 骨幹上的分析與監控

劉人豪、 洪正雄、 張展肇、侯廷昭
中正大學 電機所通訊網路組
嘉義縣民雄鄉三興村 165 號
Email： m8989@cn.ee.ccu.edu.tw m9008@cn.ee.ccu.edu.tw
m9001@cn.ee.ccu.edu.tw tch@ee.ccu.edu.tw

摘要

隨著網路的興起,上網人數持續增加,在網路頻寬無法一直提昇的情況下,必須有一套精確的量測方法,分析與監控各單位的網路使用情況,網管人員可將蒐集到的資訊,作為制定相關政策的參考,達到資源有效分配的目的。傳統的測量方式通常採用 Netflow 或 MRTG 來做訊務的分析,可是這兩種方式最多只能分析到封包的第四層標頭,隨著網路上的應用愈來愈廣泛,這些資訊並不能滿足網管人員的需求。在本篇論文中,我們提出一種新的量測方法,並以雲嘉地區的區網中心為實驗環境,針對目前網路上最普遍的 HTTP 與 FTP 協定做進一步的分析與探討,透過該系統的幫忙,可彌補 MRTG 和 Netflow 在這方面的不足,另外對於打擊 FTP 地下站也非常有幫助。

關鍵字：網路監控、HTTP、FTP、雲嘉地區。

1. 前言

網際網路發展迅速,各式應用相繼推出,如網路遊戲與網路視訊等,由於配合中小學上網政策,再加上原先各大專院校的流量,導致 TANET 頻寬愈顯不足,雖然教育部積極提昇骨幹的頻寬,但仍趕不上使用者的需求。現有各大學之電算中心或各縣市的區網中心常利用 MRTG 或 Netflow 軟體,即時監控網路的流量,但我們知道現有各大專院校 FTP 地下站氾濫,幾乎佔據大部分的頻寬,現有的軟體又只能分析到第四層,所以無法進一步做防治。關於這個問題,我們會在第二節提出我們量測的方法,第三節介紹此一量測系統實際運作情況,第四節針對量測後的結果做詳細的探討與分析,最後我們做個結論。

2. TANET 骨幹訊務的量測

現有的量測方法有 MRTG/Netflow 等,但 MRTG 只可得知網路傳輸的總量,而 Netflow 為 Cisco 的專利技術,對於非 Cisco 的機器就無法利用此方式來觀察網路流量,而且兩者都有一個重大的缺點,就是無法觀察第七層的資訊,單單知道一個 IP 的流量,其實並沒有太大的意義,如果我們能夠知道實際第七層的傳輸內容,是 HTTP 或 FTP,瀏覽網頁的網址為何？FTP 抓取的檔案是什麼?是否為非法的軟體…等,相信對於網管人員來說會更有意義。針對這樣的需求,我們設計了一套架設在 OC-3 骨幹上的量測系統,透過我們的監控系統,針對 HTTP 與 FTP 封包做進一步的分析,了解使用者上網的網址以及抓取的檔案為何,達到線上即時監控的目的。由於本系統在設計之初即以輔助現有 MRTG/Netflow 為主要目的,所以現有 MRTG/Netflow 能得到的資訊,我們就不再做分析,用以減輕系統的負擔。

我們量測的方式是利用分光器與 CoralReef 軟體,擷取骨幹上的封包,再交由我們的程式做進一步的處理,量測的平台是架設在雲嘉地區 MOE 路由器與雲嘉地區連接至 TANET 的 ATM Switch（Cisco LS1010）之間,如圖 1 所示。

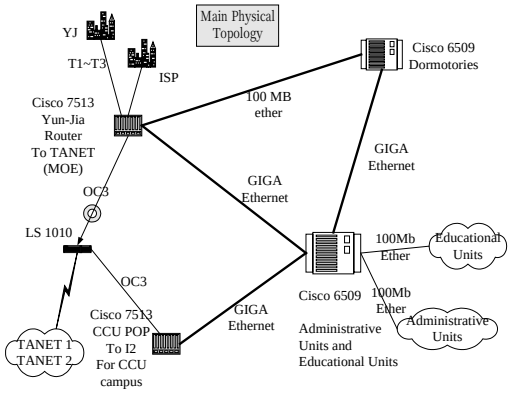

圖 1. 中正校園網路架構圖

圖 2 為量測系統的架構圖,當 CoralReef 把封包從骨幹擷取出來後,首先透過我們的程式做分析,觀察是否為 TCP Traffic,因為從先前觀察的數據可得知,網路上 90%的應用為 TCP 連線,如常見的 HTTP/FTP/E-mail 等,所以我們忽略 UDP 封包,以減輕系統的負荷,接著我們利用 Port Number 來決定 AP 層的應用為何,如果目的埠為 80 或 8080 則判定為 HTTP 封包,其餘埠號則先假設為 FTP Flow,最後交給相關函式做處理。在 HTTP 訊務方面,我們分析 HTTP 的 Request 封包,得知 Client 欲連接的伺服器網址,至於 Response 封包我們就不做分析,用以減輕機器的負擔,況且有 Request 才有 Response,有無分析 Response 並無太大的影響。而在 FTP 方面,由於地下站氾濫,單純分析 Port 20/21 並無法準確得知 FTP 的交通量究竟有多大,所以我們的做法是將所有封包都做檢查,只要封包中有 FTP 才有的關鍵字,即可確定此連線為 FTP,最後將擷取到的資訊儲存到 MySQL 資料庫中。至於採用 MySQL 的原因有兩點,一、MySQL 支援完整的 SQL 語法,二、MySQL 為免費的軟體,所以透過 SQL Language 可大大減化程式碼的複雜度且提高程式的可讀性,將來想新增功能或維護上也比較方便。

系統架構圖中關於 SNMP Agent 與 Mail Server 部分,前者在機器發生狀況或程式無故中斷時,可以主動發出 SNMP Trap Message 通知網管人員做處理,把系統閒置的時間降到最低。另外也可在某 IP 的使用量到達上限,讓監控系統主動發出 SNMP Set Message,做遠端設定 Router 的動作,藉由修改 Router 中的 ARP Table,使該 IP 無法繼續使用網路,達到主動監控與犯罪防治的目的,不過這部分牽扯到管理權則的問題,無法得知 Router 的設定密碼,所以僅有構想但未實作。至於 Mail 部分,量測系統會選擇網路離峰時間自動將先前擷取下來的資訊做統計,並將結果寄給網管人員,減輕網管人員的負擔。

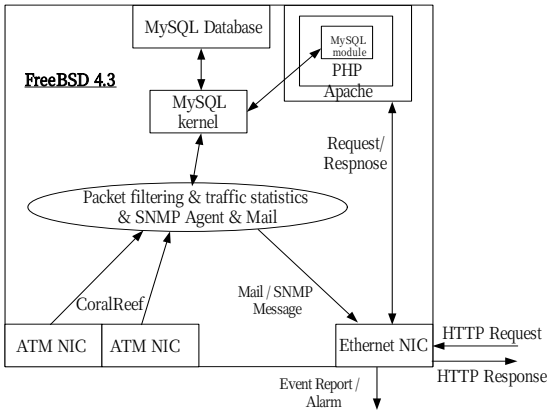

圖 2. 量測系統架構圖

在一切都就緒後,網管人員想從遠端觀察蒐集結果時,只要先連結我們設計的網頁,透過該網頁即可與 MySQL 資料庫做連結,並將結果以網頁方式呈現,操作簡單、無須任何的學習。網管人員可以從這些資訊判斷出 FTP 站台的合法性,如果放置的檔案為版權軟體或 MP3 音樂檔,即可通知相關單位處理。

3. 系統實際運作情況

本系統屬教育部委辦計畫,目標是讓網管人員能掌握網路實際運作情況,主要是針對網路上最常見的 HTTP 與 FTP 做進一步的分析,從圖 3、圖 4 可觀察中正大學、雲林科技大學、嘉義大學、虎尾技術學院、吳鳳技術學院、環球技術學院及其他 ISP 業者在 HTTP 與 FTP 協定上的頻寬使用情況。

圖 3. 各校使用 WWW 的情況

圖 4. 各校使用 FTP 的情況

本系統也可得知目前使用 WWW 以及使用 FTP 下載檔案的前十名 IP (圖 5、圖 6)。

圖 5 . 使用 WWW 前十名的 IP

圖 6. 使用 FTP 下載檔案前十名的 IP

在 FTP 分析方面,本系統會統計出前二十名可疑的 FTP 站 (如圖 7),網管人員只要點選其中任何一個 IP 就可以查出該主機曾經被下載過的檔案,從中即可判斷出網站的合法性,圖 8 顯示該主機上放置大量 MP3 歌曲供人下載,顯然違反教育部既定政策。由於本系統是監看所有的 TCP Port,而非傳統的 Port 20/21,一旦啟動本系統監控後,非法的 FTP 網站將無所遁形。

圖 7. 主要的可疑 FTP 網站

圖 8. 持有非法的 MP3 歌曲

本系統也可顯示某 IP 的 FTP 使用情況,如下載過的檔案名稱或各檔案類型所佔比例等。據我們這陣子的觀察,在 TANET 骨幹使用 FTP 協定來傳送檔案,前三名分別為 video、zip、mp3,每日傳輸量甚至高達數百 GB,結果頗為驚人(如圖 9、圖 10)。

圖 9. 使用 FTP 的情況

圖 10. 下載的檔案

在 HTTP 分析方面,我們可以選擇來源端或目的端的 IP 來做查詢,透過該選項可以查詢某 IP 在 WWW 上所瀏覽過的網頁資訊、各式應用所佔比例以及瀏覽網站之網址等 (圖 11、圖 12)。

圖 11. 根據 Sourec 端的 IP 查

圖 12. 該 IP 瀏覽過之網址

由於系統會選擇離峰時間自動做當日資訊的統計與備份動作,所以透過網頁可立即查詢過去一週來所擷取到的資料,或者是更早之前的統計資訊(如圖 13、圖 14)。

圖 13. 過去各校使用 WWW 的流量

圖 14. 過去各校使用 FTP 的流量

最後,網路管理人員可以透過專屬的網頁來管理資料庫,不需要在文字模式下輸入一大堆的指令,透過人性化的網頁設計使網路管理人員能輕鬆地掌握資料庫的運作情況(如圖 15)。

圖 15. 網路管理人員專用

4. 分析與討論

4.1 HTTP

我們分析封包第七層的 Header,觀察 GET 關鍵字後面所帶資訊,我們捨棄 Content Type 關鍵字來決定檔案型態的方式,直接觀察 URL 整個字串,擷取出檔案的副檔名來決定檔案的型態,就我們的觀察,檔案型態的前三名如下表所示。

表 1. HTTP Content_type 分布情形

Content Type	Percent (%)
GIF	17.721
JPG	15.240
Html	13.177

由表 1 可觀查出網頁設計仍以 GIF 佔大宗,由於一個 HTML 網頁通常使用大量的圖片來美化網頁,而 GIF 圖檔最為常見且支援漸層顯示的功能,不用等整張圖檔下載完畢就能先顯示圖片的概廓,所以深受網頁設計者的喜愛。

在綜合統計上,我們以 Htm、Php、Asp、Jsp、Cgi 代表網頁資料, Doc、Txt、Pdf 表文字資料, Gif、Jpg、Ttf 為圖形資料，剩餘的檔案類型我們就不額外做分類，全部歸類於其他這個欄位。我們從表 2 可觀察出 HTTP Request 仍以圖片(GIF/JPG)佔大宗,另外由於許多網站上放置共享軟體供人下載(Ex: Toget、史萊姆…等)，再加上新興網路技術愈來愈發達(Ex: Flash), 所以其他這個欄位佔了將近一半的比例。由此可知，多媒體與檔案傳輸在 Web 上的應用愈來愈廣泛。

表 2. HTTP Content Type 統計表

Content Type	Percent (%)
網頁方面	18.99
圖片方面	32.96
文字	2.9
其他	45.15

在 URL Request 方面,由於我們的系統在統計上是以網頁為基本單位,我們統計的方式是觀察哪個網頁被瀏覽的次數最多,而非以網站為單位,所以一些大型的網站(如奇摩),雖然連線次數很多,但因為分散在不同的網頁,所以平均起來並不會特別高。我們從 HTTP Request Top 10 中可觀察到一個現象,由於大專院校學生喜愛上 BBS 站發表言論或欣賞文章,而台大學生所發展的 KKMan 網路瀏覽器,因為功能強大且為免費軟體,深受學生族群的喜愛,在其最新版的 KKMan 程式中,右上角有一小塊

的廣告區,程式會不定期的更新廣告內容,所以我們可以看到有非常多的連線都連結到某特定網頁下載廣告內容,從每日 Http request Top 10 統計來看,通常位居前三名,可見 KKman 這套軟體設計十分成功,使用者人數也相當多。

4.2 FTP

由於 FTP 協定設計上的關係,Client 與 Server 在傳送多個檔案時會建立許多 TCP 的連線,而且 Port Number 也不固定,因此我們的監控程式針對每個 Port 都做分析,表 3 是我們統計過後的資訊。

表 3. FTP File Type 統計表

File Type	Percent (%)
影音	78.94
圖片	1.37
文字	0.18
壓縮檔	6.86
MP3	3.67
其他	8.98

從上表可明顯觀察出,FTP 大部分以傳送影音資料為主,我們將副檔名為 mpg、mpeg、dat、avi、asf、rm 歸類為影音方面,其傳送的容量佔總傳送量的 75% ~ 80%,結果可說是非常的驚人。這樣的統計結果代表著影音方面的資料由於容量都非常大,所以不適合用 E-mail/HTTP 等方式傳送,而支援檔案續傳功能,傳送速度又快的 FTP 協定,自然被大家所採用。而這類影音檔案的內容,大部分為院線片偷拍版與色情影片,結果令人憂心。

就我們觀察的結果可歸納成以下兩點：

(1) 觀察所有的 Port Number 所得 FTP 使用量為 Netflow 所得結果的 8.5 ~ 9 倍之多,可見 FTP 協定的使用量非常驚人。
(2) 網路上 FTP 被下載量前十名的網站,其網路流量佔全部 FTP 流量的 1/4,可見這類大型站台佔據大部分的使用頻寬,如果要讓網路有更好的連線品質,這類大型的站台應該優先處理。

5. 結論

本篇論文主要提出量測 HTTP 與 FTP 訊務的方法,我們以雲嘉地區的網路流量來測試我們實作的系統,透過我們的系統,可以得知 HTTP 與 FTP 訊務的細部資訊,從單一個 IP 查詢、各流量前十名 IP 列表、可疑 FTP 地下站列表,到最後的歷史資料查詢。配合先前已有資訊,如教育部的 TANet 骨幹流量列表、雲嘉區網流量統計、及中正大學的 Proxy 及 FTP 流量統計圖,將可提供網管人員最豐富的參考資料。而在 FTP 量測方面,先前研究文獻通常僅將 Port 20/21 的訊務當作 FTP,忽略 FTP 地下站的封包,透過我們這套系統,以人性化的網頁介面與分析列表,執行全天候的網路監控,時時監控 HTTP/FTP 的傳輸情況,相信對於掃蕩校園內 FTP 地下站會有很大的幫助,更期望能因此紓解 TANET 骨幹塞車的情況。

6. 參考文獻

[1]. K. Claffy, “Coral Reef 3.3.2”, Caida.

[2]. J. Castagnetto, H. Rawat, S. Schumann, C. Scollo, D.Veliath, "Professional PHP Programming", Wrox, 2000.

[3]. A. S. Tanenbaum, “Computer Networks”, Prentice Hall.

[4]. W. W. Gay, “Linux Socket Programming”, QUE.

[5]. http://www.mysql.com online help.