行政院國家科學委員會專題研究計畫　成果報告

子計畫三：網際網路差異化服務之訊務分類與排程研究設計

(3/3)

計畫類別：整合型計畫
計畫編號：NSC91-2219-E-009-036-
執行期間：91 年 08 月 01 日至 92 年 07 月 31 日
執行單位：國立交通大學資訊工程學系

計畫主持人：陳耀宗

報告類型：完整報告

處理方式：本計畫可公開查詢

中　華　民　國 93 年 2 月 3 日

行政院國家科學委員會補助專題研究計畫執行進度報告

※※※※※※※※※※※※※※※※※※※※※※※※※
※　　　　　　　　　　　　　　　　　　　　　　　※
※　網際網路差異化服務之訊務分類與排程研究設計　※
※　Study and Design for Traffic Classification and　※
※　Scheduling of Differentiated Services on Internet ※
※　　　　　　　　　　　　　　　　　　　　　　　※
※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：□個別型計畫　　☑整合型計畫

計畫編號：NSC－91－2219－E－009－036

執行期間：91 年 8 月 1 日至 92 年 7 月 31 日

計畫主持人：陳耀宗

計畫參與人員：詹益禎、郭國承、張君瑋、何凱元

本成果報告包括以下應繳交之附件：

□赴國外出差或研習心得報告一份

□赴大陸地區出差或研習心得報告一份

□出席國際學術會議心得報告及發表之論文各一份

□國際合作研究計畫國外研究報告書一份

執行單位：國立交通大學資訊工程系所

中　華　民　國　九 十 二　年　九　月　三十　日

「寬頻網際網路端對端技術之研究」子計畫三
「網際網路差異化服務之訊務分類與排程研究設計」
Study and Design for Traffic Classification and Scheduling of Differentiated Services on Internet

計畫編號：NSC 91-2219-E-009-036
執行期限：91 年 8 月 1 日至 92 年 7 月 31 日
主持人：陳耀宗教授　國立交通大學資訊工程學系暨研究所
計畫參與人員：詹益禎、郭國承、張君瑋、何凱元

一、中文摘要

差異化服務(Diffserv)是近年來吸引高度關切的網際網路服務技術，目的在於避開非常複雜的 ATM 訊務系統，使用較簡化的方法以提供差異化的服務。差異化服務的路由器架構基本上是分為兩類，其一為邊界路由器，另一則為核心路由器。在邊界路由器需要依據許多的方法來區分出不同的資料流，並分析各資料流是否符合所要求的流量特徵，而後送至不同的服務動作區(Per Hop Behavior; PHB)來提供傳送服務，而核心路由器則是用簡化了的分類器，只依據特別的標號來決定所該送往的服務動作區。在此架構之下，邊界路由器將會變的複雜，但速度不需要像核心路由器要求那麼高。我們有兩個方向的研究發展興趣，其一是在邊界路由器所特有的多欄位分類器(Multi Field Classfier)設計，另一則為負責供應差別服務的排程器(Scheduler)設計。多欄位分類器的設計有兩大方向，一者為直接的 hash 方法，分析資料流所適用的演算法及參數設定，另一則為多欄位最佳吻合(Best match)的方法研究。而第二的研究發展方向則為差異化服務專用排程器的設計，將各種已發展成熟的，或自行研發的排程器設計導入差異化服務之中。

我們將用系統模擬以驗證並比較出最適合之設計，並利用各種軟硬體方式以實現所完成之設計。

關鍵詞：差異化服務，邊界路由器，核心路由器，服務動作區，多欄位分類器，最佳吻合

英文摘要

Differentiated Service (DiffServ) is a fast growing Internet service technology in recent years. The purpose of DiffServ is to avoid very complicated ATM signaling, and use a simpler method to provide differentiated services. The router architecture for differentiated services can be categorized into two types. One is the edge router, and the other is the core router. An edge router needs to distinguish differentiated data flows based on many considerations, and it analyzes each data flow to verify whether the flow conforms to the flow characteristics, then it sends the packet of the flow to its associated PHB to obtain the required forwarding service. While a core router uses a simplified classifier. It forwards the packet to the associated PHB according only to a specific marking. Generally speaking, edge router will be more complicated than the core router, but it does not require very high speed as that in the core router. Under such infrastructure, we organize two research topics based on or interest. The first is to investigate and design a multi-field classifier for the edge router, and the other is to study and design the scheduling mechanism for differentiated services. Regarding the multi-field classifier, we have two directions, one is to use hash mechanism and find a proper algorithm and the associated parameter setting. The second approach is to study the best match mechanism. The second topic is to design a specific scheduler for differentiated services. We will use those matured scheduling

schemes as well as our newly developed scheduler for differentiated services.

We use a popular simulation package to perform the system evaluation and to find out the most proper design in both classifier and scheduler. Then we would like to realize our design through all possible hardware and software solutions.

Keywords: Differentiated services, edge router, core router, per hop behavior, multi field classifier, best match

二、緣由與目的

網際網路的迅速發展，不但讓它成了一個全球性的商業通訊架構，網際網路技術也成為工商產業以及個人生活之通訊技術基礎。各式各樣之應用，從傳統的文字數位資料傳輸到現今之整合視訊音訊之分散式多媒體應用，都透過同一網際網路架構與機制以運作。由於網際網路在傳統上是針對非即時性之資料傳送提供盡力而為(Best Effort) 之服務，對於傳統之電子郵件(E-mail)、FTP 以及 Telnet 等服務，Best Effort 服務運作的很好。但自 1993 World Wide Web 發展問世以來，網際網路使用者每年成幾何級數增加，加上微處理器運算能力大幅提昇，對視訊音訊之即時處理也游刃有餘，加之光纖傳輸容量之躍增，各種以往認為遙不可及之複雜應用，如今確可以一一實現。不過，面對各式各樣的網際網路應用，若依然以同樣盡力而為方法去面對變化多端的多媒體訊務，很顯然的可以看出其中之不適當性。

上述之各種新的應用與它們所突顯出的 Best Efforts 服務之不適當性，挑戰著當初 TCP/IP 之端點與端點訊務控制之設計目標。因此，各種與服務品質(Quality of Services;QOS) 相關之服務模式與因應之訊務控制方發陸陸續續的被提出。在網際網路工程任務小組(Internet Engineering Task Force; IETF) 這邊，提出了兩種與 QOS 有關之架構: 一為整體服務(Integrated Services; IntServ)，另一為差異化服務(Differentiated Services; DiffServ)。整體服務係以每個資料流為基礎提供端點對端點之 QOS 服務，每個資料流皆可向網路要求一特定程度之服務，譬如最大之端點與端點之延遲，或最小之資料傳輸率等。網路可以根據當時可用資源之狀況，答應或拒絕每個資料流之要求。整體服務有三個主要功能: 允諾控制(Admission Control)，封包傳送機制(Packet Forwarding Mechanisms)與資源保留協定(Resource Reservation Protocol; RSVP)。不過整體服務有兩個主要問題: 一為大規模化(Scalability)之問題，另一為管理之問題。未來在 Gigabit 或甚至 Terabit 網路傳輸中，核心路由器面對上百萬的資訊流是可能的。而整體服務架構可能需要同時管理如此龐大的資訊流，其任務之艱鉅複雜可想而知。除此之外，必須所有在資料流所經路徑上的每個子網路都支援整體服務，才能達到所要求的保證，否則服務品質將無法預期。

由上述可知，整體服務所需求之架構，在實現上有其不易克服之困難；而另外一種差異化服務，雖然也非十分令人滿意，但至少在大規模化(Scalability)與管理架構上，問題較小。在目前以 IP 為主之網際網路上，較有實現之可能。差異化服務之論點，不再針對每一筆資料流做 QOS 之處理，而是將訊務分類、集中，然後處理這些訊務匯整 (Traffic Aggregates)。相較於整體服務，差異化服務之發展較晚，複雜度較小，不過它至今仍在演進中，許多相關技術仍未明朗化。雖然如此，鑑於網際網路上服務品質之需求迫切，仍有大量的研究發展人力投入。依目前之情況，差異化服務之研究，分成兩大方向: 一稱為絕對服務差異，另一稱為相對服務差異。前者著重絕對的效能表現，其在核心路由器上之做法與整體服務類似。只不過它所處理的，已不再是每一筆來自端點的資料流，而是匯整資料流(Flow Aggregates)。它也不做動態的資源保留。至於後者，IETF 訂出了一些服務模式以確保不同服務品質間相對的優先次序，同時也可依此訂定差異化服務收費之標準。

眾所周知網際網路將為二十一世紀帶來產業與生活的重大變革，它的涵蓋面甚廣，舉凡商業、教育、醫學、娛樂、人際溝通、政府等之運作都將受到影響。但現

今網際網路有其能，亦有其所不能。譬如說現今之網際網路電話，其通話品質就不如傳統之電話。差異化服務並不一定能澈底保證網際網路上之服務品質，而且其相關技術尚有許多未解決之問題。不過網路服務品質需要以代價換取，要得到較佳之品質，其相對之付出代價也高，而如何有效而系統化的將變異極大的訊務設計出壹套分類之方法，並依此分類設計相對應之排程器，即為本計畫之目的。

由本計畫的參考文獻上，可以看出國外已有相當多的人力投入這方面之相關研究。國內也有不少學術單位或科學園區公司也已著手此方面之研發。雖然說，在網際網路演進之過程中，差異化服務並不必然是達成 QOS 的最佳方法，不過基於我們多年來對 TCP/IP 訊務控制、緩衝區管理、與排程器之研究經驗，經由本計畫相關之深入研究與軟硬體設計發展，我們將深信會有可觀之結果。

本計畫為寬頻網際網路端對端技術之研究總計畫下之第三子計畫。另外之第一子計畫題目為真實網路封包流量自我類似性質之成因探討與統計模式建立，主要在探討網際網路中具有自我類似(Self-Similar)特性的真實網路訊務，瞭解其特性並可作為其它子計畫架構設計與訊務控制之依據。第二子計畫題目寬頻網際網路中路由選徑技術與 QoS 訊務控制之研究設計，由於差異化服務之目的即為達到不同之服務品質, 因此與本子計畫有非常密切之關係。

其它二個子計畫雖各有所司，但關係十分密切，執行過程中將定期召開工作檢討會，彼此溝通，交換研究心得；且本研究群將以謹慎務實的態度來執行各項研究工作，在經費上並沒有申請過當的設備費用，且本計畫較屬學術研究性質，故所需設備均編列在相關的子計畫中，申請單位也將在空間及行政措施上給予最大的配合。

三、研究方法與成果

在這一年的研究中，我們設計了一個新的封包分類演算法，它能有效的將進入路由器的封包分類至所屬的過濾器，並針對該過濾器指定的動作採取行動。

封包分類的動作基本上不外乎擷取封包中的資訊，與資料庫中的各個欄位做比對，以找出最佳比對的過濾器，並採取所指定的動作。目前已有相當多的論文在探討這樣延伸的問題，可以利用硬體的方式實做，也可利用軟體的方式進行模擬。在眾多方法中，令我們感興趣的是一個稱為朗訊位元向量（Lucent Bit Vector）的方式，這個方法原本是利用硬體的方式來探討封包分類的問題，但之後的一些論文則是開始利用軟體模擬的方式來研究該方法是否有改進的空間。由於此方法已經被路由器設計大廠朗訊採用，所以可以說是已被驗證可實際運用於目前的路由器中，如何改進這個方法便成為一個很有趣的問題。

我們著眼於朗訊位元向量方法耗費大量的儲存空間來維持位元向量，所以我們利用建立子樹的觀念來節省儲存位元向量的空間，不同於朗訊位元向量的方法，我們僅在子樹的根節點儲存位元向量（Bit Vector），將此子樹含有位元向量的所有資訊都整合起來儲存於根節點的位元向量中，如此一來，位元向量所耗費的儲存空間會隨著子樹中所包含子節點個數的增多而節省其儲存空間。之外，由於我們觀察到只利用這樣單純的觀念並沒有完全解決封包分類的問題，而會導致所謂錯誤比對（false match）的情形發生。為了解決錯誤比對的情況，我們在事先處理過濾器資料庫（filter database）時，會將資料庫做分析，並將封包分類分離成兩個階段來處理，一個採取原本位元向量的方式，另一個則採取線性搜尋的方法來比對。

以上的設計導因於我們的兩點觀察，第一，會發生錯誤比對的情形是由於我們採取子樹的概念來節省儲存空間，原本分屬於不同分支的資訊會因而整合到根節點中，使得在執行位元向量比對時找出錯誤的位置，所以我們在事先處理過濾器資料庫時，會把這些會發生錯誤比對的過濾器給分離成另一個資料庫，僅對於那些不會發生錯誤比對的產生位元向量，這樣可以得到的第一個優點是位元向量的長度大大減少，如此也是減少最差情形的記憶體存

取次數；第二，事實上，在過濾器資料庫含有的過濾器數量很少時，利用複雜方式的演算法所得到的效能並不會比簡單的線性搜尋方式來的快速。由於我們可以控制在第二階段要比對的過濾器數量大小，所以在第二階段利用線性搜尋的方式便可以達到極快的速度。以下即開始說明我們的方法是如何運作以及實驗的結果。

1. 我們的方法採用原本朗訊位元向量的方式來建立位元向量，其建立的方式是，若是過濾器資料庫中有 1000 個過濾器，那麼我們的位元向量的大小便是 1000 個位元(bit)，在我們的假設中，過濾器資料庫已經按照優先權的順序加以排序，越高優先權的過濾器放在資料庫的開始，相對應來說，越低位元位置所代表的便是越高的優先權，例如：位元 1 的優先權比位元 5 的優先權來的高。在進行位元向量的比對時，我們是利用交叉（intersection）的方式來比對，所以只要交叉比對出來的值等於 1，那麼該比對便可停止，此位元的位置也能直接對應到相對應的過濾器位址。
2. 為了節省儲存位元向量的位址空間，如前所述，我們利用建立子樹的觀念來建立位元向量，建立的方式如下：

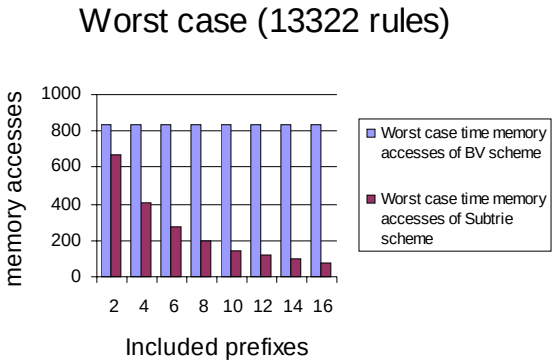
```
GenerateSubtrieRoot(Cur_Node, Cur_Depth)
BEGIN
    IF ( ChildWithPrefix(Cur_Node) == Max_Prefix_Num )
        GenerateSubtrieRoot(Cur_Node);
    ELSE
        SubtrieConstructor(Cur_Node->left, Cur_Depth+1);
        SubtrieConstructor(Cur_Node->right, Cur_Depth+1);
        IF ( ChildWithPrefix(Cur_Node ) >=
        Max_Prefix_Num )
            GenerateSubtrieRoot(Cur_Node);
        ELSE IF ( (Cur_Node == Trie_Root) &&
                (ChildWithPrefix(Cur_Node) !=
                Max_Included_Prefix_Num) )
            GenerateSubtrieRoot(Current_Node);
END
```

利用從下而上（bottom-up）的建構方式，根據目前這個節點所包含的子樹中有多少個帶著 prefix 的節點以決定是否將此節點設成該子樹的根節點。藉由這樣的方式建構子樹的根節點，其目的是在有效的控制根節點維護的資訊量，也就是控制整合到根節點的位元向量的數目。整合的數目越多，位元向量中為 1 的位元便會增加，使得錯誤比對的機會大增，這個部分我們以第二階段的線性比對做為改進。

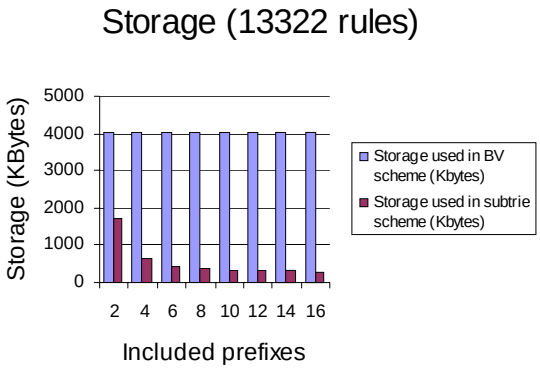
3. 建立完所有的根節點位置時，我們會給予每個根節點一個相對應的辨識碼，同時去修改過濾器資料庫，對於每個維度去更新相對應的辨識碼，接著搜尋整個過濾器資料庫，將所有維度辨識碼的集合相同之過濾器集中放置到第二階段處理，例如：以來源端位址的辨識碼為 2，目的端位址的辨識碼為，那麼我們便會去搜尋整個資料庫，將所有來源端位址的辨識碼是 2 和目的地位址的辨識碼是 3 的集合起來，放置到第二階段的資料庫。若是某個辨識碼的集合僅有自己而沒有其他人與其相同，那麼此過濾器一定不會與其他的過濾器產生錯誤比對的狀況，所以我們將此過濾器放置在第一階段處理。
4. 根據步驟 3 處理的結果，我們將第一階段的過濾器資料庫依照位元向量的方式來建構其位元向量，在這裡可以注意到的是因為我們在事先處理時已經將原本的過濾器資料庫分成兩個部分來考量，所以此時在第一個階段的位元向量比起朗訊位元向量會短小許多，這也可以視為我們的方法在基本上比朗訊位元向量的方式擁有較少的記憶體存取次數。
5. 在實際執行我們演算法的部分，如同原本朗訊位元向量的方式，我們會針對每個維度都進行位址搜尋的步驟，這個部分可以利用現有的 IPv4 以及 IPv6 快速位址搜尋資料結構，來加快此步驟的速度。接著，找出每個維度所符合的子樹根節點後，我們首先對每個維度的根節點做一次記憶體存取，讀出存於該節點的辨識碼，檢查該辨識碼的集合是否存在第二階段的辨識碼集合中，若是，則我們可直接跳過第一階段的比對，直接進入第二階段做處理，可以這麼做的原因是，由於我們已經先做了事先處理的步驟，所以可直接在這裡便直接決定是否跳躍進第二階段；若否，那麼代表該

辨識碼不在第二階段中，所以我們可以很明確的決定進入第一階段的位元向量比對。

6. 下面兩個圖表是我們的實驗結果之一，在這個結果中，我們的過濾器資料庫是兩個維度的資料庫，包含來源端位址和目的端位址兩個欄位。實驗中，我們假設路由器一次記憶體存取能夠讀取出 32 個位元，所以對於朗訊位元向量的方式來說，其最差的記憶體存取次數是（過濾器個數/一次記憶體存取的位元數），結果如下圖：



可以發現我們方法的記憶體存取數量會因為子樹包含的節點數量增多而變小。在這裡我們會對最差的狀況有興趣的原因是，由於過濾器資料庫的型態會根據網路的特性而改變，以致於若探討平均的記憶體存取數時，其變動的因素遠大於最差的記憶體存取數，所以若是能提供一個在最差狀況下能夠表現優良的封包分類演算法，代表的是該演算法在平均的狀況下也會有不錯的效能。下圖是我們演算法所耗費的儲存空間，可以清楚的發現依然遠小於朗訊位元向量的儲存空間。



四、結論與討論

為了符合使用者日益變化的需求，網路服務提供者必須適應使用者的要求提供服務，因此在路由器的設計中也必須考量到更多的層面，而不僅是侷限於單一的目的地位址搜尋。解析封包更多的欄位進行分析，快速決定封包所屬的過濾器並針對封包採取動作，是未來路由器不可或缺的功能。所以如何應用快速，正確的多欄位封包分類演算法，將會是未來路由器研發設計時一個重要的考量。

我們在本子計畫一年的研究期間，提出一個可實際使用在未來路由器設計的演算法，藉由實驗結果的呈現，得知我們可以依據實際網路的需求和路由器本身的硬體限制，調整演算法以得到最佳的效能，同時適應各種不同的分類器特性，這是與其他方法最大的不同之處。

五、參考文獻

[1] V. Srinivasan, G. Varghese, S. Suri, and M. Waldvogel, “Fast and Scalable Layer Four Switching,” Proc. ACM SIGCOMM ‘98.

[2] M. Waldvogel, G. Varghese, “Scalable High Speed IP Routing Lookups,” Proc. ACM SIGCOMM ’97, Sept. 1997, pp. 25-36.

[3] B. Lampson, V. Srinivasan, and G. Varghese, “IP Lookups Using Multiway and Multicolumn Search,” Proc. IEEE INFOCOM ’98, Apr. 1998, pp.1248-56.

[4] T. V. Lakshman and D. Stidifialis, “High Speed Policy-based Packet Forwarding Using Efficient Multi-dimensional Range Matching,” Proc. ACM SIGCOMM ’98,

Sept. 1998.

[5] A. Feldman and S. Muthukrishnan, “Tradeoffs for Packet Classification,” Proc. IEEE INFOCOM ’00, Mar. 2000, pp.397-413.

[6] P. Gupta and N. McKeown, “Packet Classification on Multiple Fields,” Proc. ACM SIGCOMM ’99, Sept. 1999.

[7] F. Baboescu and G. Varghese, “Scalable Packet Classification,” Proc. ACM SIGCOMM ’01, Aug. 2001.

[8] Ji Li, Haiyang Liu, and Karen Sollins, “Scalable Packet Classification Using Bit Vector Aggregating and Folding,” Proc. ACM SIGCOMM ’02, Aug. 2002.

[9] V. Srinivasan, S. Suri, and G. Varghese, “Packet Classification using Tuple Space Search,” Proc. ACM SIGCOMM ’99.

[10] P. Gupta and McKeown, “Classification Using Hierarchical Intelligent Cuttings,” Proc. Hot Interconnects VII, Aug. 1999.;also available in IEEE Micro, vol. 20. no. 1, Jan./Feb. 2000, pp. 34-41.

[11] Abilene NetFlow Nightly Reports, http://www.itec.oar.net/abilene-netflow