行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※※
※　　　　　　　　　　　　　　　　　　　　　　　　※
※　寬頻網路網際網路品質保證（II）— 子計畫二：　※
※　使用於寬頻網際網路之 Gigabit 路由器與訊務管制　※
※　技術（II）　　　　　　　　　　　　　　　　　　※
※　　　　　　　　　　　　　　　　　　　　　　　　※
※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：□個別型計畫　　▒整合型計畫

計畫編號：NSC 89－2219－E－009－004－

執行期間：　88 年 08 月 01 日至 89 年 07 月 31 日

計畫主持人：李程輝　交通大學電信系 教授

共同主持人：

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：國立交通大學電信系

中　華　民　國　89　年　　　月　　　日

行政院國家科學委員會專題研究計畫成果報告

計畫編號：NSC 89-2219-E-009-004
執行期限：88 **年** 8 **月** 1 **日至** 89 **年** 7 **月** 31 **日**
主持人：李程輝　　交通大學電信系 教授
共同主持人：
計畫參與人員：

一、中文摘要

本計畫主要研究兩項大容量路由器的主要技術：交換機架構與利用硬體來做最長字首的比對(或稱為硬體路由)。為了建構一個真正大型的路由器，我們採用空間分割的架構如 crossbar**；並發展佇列管理和快速排程演算法來移除佇列前端擁塞**(HOL blocking)**問題，以改善路由器之傳輸率**(throughput)**。此研究主要針對可變長度封包設計一交換機架構，以減小輸出埠閒置時間來提高效率降低封包延遲。經由軟體模擬分析比較，其結果顯示固定長度封包設計的交換機並不適用於可變封包長度的網路環境。**

最長字首的比對技術是發展高速 IP **路由器一個重要瓶頸，為了達到有效快速的訊務分類，我們發展一套三級訊務區分演算法，來處理目的地位址、傳送端位址、目的地埠、傳送端埠與協定等五個欄位，並以硬體路由結構來加速路由決策。**

關鍵詞：IP **路由器，縱橫式交換系統，硬體路由，最長字首比對**

Abstract

In this project, we investigated two key technologies of developing high-capacity routers: switch architecture and longest prefix matching with hardware (or hardware routing). We employed space-division architecture, such as the crossbar, to build a real large-capacity router. In addition, we designed queue management and fast scheduling algorithms to (partially) remove head-of-line blocking in an attempt to improve the router's throughput. This study mainly focuses on designing a switch for variable-length packets to reduce idle time on output port in achieving better performance in terms of throughput and packet delay. Through simulation, the result showed that switches designed for fixed-length packets are inadequate for network environment of variable-length packets on the fly.

Longest prefix matching is a bottleneck in IP routing. Thus, to efficiently classify arriving packets, this study proposed a three-phase packet classification algorithm based on destination/source IP addresses, destination/source port numbers and protocol ID fields and designed efficient hardware routing schemes to speed up routing decision.

Keywords: IP router, crossbar switch, hardware routing, longest prefix matching

二、計劃緣由與目的

高效能網路交換機採用 crossbar **交換結構作為交換核心漸漸成為一種趨勢。而** crossbar **交換機需要一套交換排程演算法（**Switch scheduling algorithm**）來決定輸出入埠配對，目前存在的排程演算法，例如** PIM（Parallel Iteration Matching）、RRM（Round-robin matching）**與** SLIP（），**均是為固定長度封包所設計，雖然經由切割再重組的方式可以處理可變長度的封包**[1]**，卻可能造成輸出、入埠所需的緩衝空間與封包交換延遲的增加。本研究針對可變長度封包設計一交換機架構，以減小輸出埠閒置時間來提高效率，並降低封包延遲。**

高速的多欄位訊務區分器是現今發展服務保證網路最重要的核心技術之一。傳統的訊務區分器僅根據第三階層的標頭欄

位來分類，並不適用於較高階層之表頭欄位分類器；此外，網路頻寬與使用者需求的指數倍數成長使得路由器的傳輸率要求相對提高。為了支援多欄位分類與高傳輸率的需求，我們提出一個三級式多欄位訊務區分器演算法，透過 hashing、search tree、linear search 三個步驟，可以將 filter 集合有效分散開來，減少搜尋所需要的運算量，以達成高速訊務區分的目的。

三、研究方法與成果

I.適用可變長度封包之 Crossbar 交換機

整個系統架構方塊圖顯示於圖 1。由網路進入交換機的封包首先由輸出入埠控制器（I/O Port Controller，如圖 2 所示）對標頭做處理，並將封包佇存在輸入埠的緩衝區（memory）內。每個輸出入埠控制器都有一組 P 通道（P_n in & P_n out）與交換核心（如圖 3 所示）連接，此通道是對交換核心下達指令以及傳送封包的途徑。交換核心藉由串連所有埠控制器的回應通道(RSP)送出連接請求指令。核心內的仲裁器用來決定連結的建立與否，仲裁結果同樣透過此通道傳出。群播封包則經由專屬的群播通道（MC）傳送到各個輸出入埠控制器。此外，還有一個控制通道（CTRL）作為中央處理器與輸出入埠之間溝通的管道。

為了使交換機有較高的運作效能，我們以輸出埠的角度來設計交換機；當輸出埠由忙碌變成閒置狀態時，立即向所有輸入埠要求送封包過來，除非都沒有以此輸出埠為目的的之封包進入交換機，或擁有送往此輸出埠封包的輸入埠正處於忙碌狀態，則輸出埠將經過一或數個仲裁週期後，開始送出封包。交換機以減少輸出埠閒置時間來提高較率。

而仲裁器以找出具最高優先權輸入埠的方式使交換機具有差別服務功能，並使用輪轉法避免飢饉現象。

以軟體模擬來分析比較 SLIP 交換機與本計劃所設計的交換機之封包延遲與輸入佇列長度，如圖 4（a）與（b）所示，相關參數設定為 s_size=16、MaxPktSize=1500、MinPktSize=64、PSN=5000、及 CellSize=64。結果顯示根據固定長度封包所設計的交換機並不適用於可變封包長度的網路環境。

II. 訊務區分器

（一）三級訊務區分演算法

- **第一級：**Hashing

選擇特定 bits 的排列將所有的 filter 區分開來。考慮於五個欄位共 104 個 bit 裡選 m 個 bit 出來，即可將 filters 所構成的 space 切割成 2^m 部分；理想上，每個部分至多有 $N/2^m$ 個 filter（其中 N 是 filter 總數），依此方法，將可大幅減少需要找尋的 filter 數目。我們採用的策略是於 Source IP address 與 Destination IP address 各取三個 bits，取一個 bit 以區分 UDP 與 TCP 的 protocol 欄位，總共 7 個 bits 來進行 Hashing 的運作，以期將 filter space 作最平均的分割。

- **第二級：**Search Tree

根據各個 filter 的 IP prefix 建立 search tree。建立順序為 Source IP address> Destination Ip address>Protocol 欄位；此外，為了提升速度，所建立的樹並非單純的二元樹，而是能夠一次檢查 m bits 所建立的 2^m-ary Search Tree。建構樹的過程中，為了避免建立整個 search tree 所浪費的記憶體空間，我們限定當節點的 filter 數超過一定值的時候，才建立下一層的節點，如此便可簡化建立 search tree 的複雜度。

- **第三級：**Conflict Check and Sorting

將位於每個部分的 filters 根據其 cost 來排序，以加快整體封包 search 的速度。然而，由於 IP address 是採用 longest prefix matching，即 mask 較少的 filter 擁有較高的 priority，因此可能造成 mask 較長的 filter 永遠不會被 match；此為 conflict problem。為了防止 conflict 發生，於排序之前，尚需執行 conflict check 運作，定義如下：1)若 mask 較長之 Filter i，其實際的優先權≤ mask 較短之 Filter j 的優先權，則移除 Filter i；2)若 mask 較長之 Filter i 的優先權> mask 較短之 Filter j 的優先權，則將此兩者的 costs 對調。完成 conflict check 之後，依 filters' cost 之高低排序，由於 filter 數目已

非常少，一般的線性排序就可相當快速地完成，故此，我們採用氣泡排序法（bubble sort）。

（二）搜尋程序

1）根據 Hashing 特定位置的 bits，找到對應 search tree 的根（root）。

2）依照 Source IP、Destination IP 的順序, 利用 prefix 尋找節點, 當節點上的 filter 數目不為零時，表示路徑搜尋完成。

3）以 Linear Search 的方式尋找符合條件的 filter，第一個符合條件的 filter 即為 best matching filter。

（三）演算法分析

假設 filter 總數為 n、hashing key 為 b bits、2^m-array search tree 的 depth 為 $h(n)$，則平均每個節點所含的 filter 數目為 $n/2^{mh(n/2^b)}$，因此 search 一個 filter 所需的複雜度約為：$O\left(h(\frac{n}{2^b})\right)+O\left(\frac{n}{2^{mh(n/2^b)}}\right)$。

建構 2^m-ary search tree 時，可能會根據 IP prefix 作複製的動作。假設 depth=K 時，檢查的 prefix 恰小於 address mask，即 $m\times K<Address\ Mask<m\times(K+1)$，則 depth=$K$+1 時，必須複製 $[(K+1)\times m\text{-}Address\ Mask]\times 2$ 個 filters。所以當最後 depth=P（$P>K+1$）時，總共需要複製 $\left(2^m\right)^{P-(K+1)}[(K+1)\times m\text{-}Address\ Mask]\times 2=2^{m(P-K-1)+1}[(K+1)\times m\text{-}Address\ Mask]$。現假設 n 個 filters 中，有 i 個 filters 需要複製，其複製起始的 depth 以 K_i 表示，address mask 為 $aMask_i$，則最後 search 所需的運算複雜度約為

$$O\left(h(n/2^b)\right)+O\left(\frac{n-i+\sum_{a=1}^{i}2^{m(h(n)-K_i-1)+1}[(K_i+1)\times m-aMask_i]}{2^{mh(n/2^b)}}\right)。$$

四、結論

本年度計劃延續第一年的研究成果並擴充交換機的功能，具有第三層硬體路由技術、第四層防火牆功能、支援群播功能、port trunking 以及八種優先權處理。交換機架構採用 crossbar，並發展佇列管理和快速排程演算法來移除佇列前端擁塞問題，進而提昇路由器的傳輸率。特別是本交換機適用於可變長度封包的網路環境，以減小輸出埠閒置時間來提高效率降低封包延遲。並經由分析結果驗證固定長度封包設計的交換機並不適用於可變封包長度的網路環境。

此外，我們設計了一個三級式多欄位訊務區分器演算法，透過 hashing、search tree、linear search 三個步驟，可以將 filter 集合有效分散開來，減少搜尋所需要的運算量，以達成高速訊務區分的目的；並分析了此演算法的複雜度。

五、參考文獻

[1]. McKeown, N. “The iSLIP Scheduling Algorithm for Input-queued Switches,” in *IEEE/ACM Transactions on Networking*, Vol. 7 , April 1999.

[2]. Karol, M.; Hluchyj, M.; and Morgan, S. “Input versus Output Queueing on a Space Division Switch,” in *IEEE Transactions of Communications*, vol. 35, pp. 1347-1356, 1987.

[3]. V. Srinivasan, S. Suri, and G. Varghese. “Packet Classification Using Tuple Space Search,” in *Proceedings of ACM SIGCOMM’99*, Sept. 1999.

[4]. V. Srinivasan, G. Varghese, S. Suri, and M. Waldvogel. “Fast and Scalable Layer Four Switching,” in *Proceedings of ACM SIGCOMM’99*, Sept. 1999.

[5]. P. Gupta and N. Mckeown. “Packet Classification on Multiple Fields,” in *Proceedings of ACM SIGCOMM’99*, Sept. 1999.

[6]. Thomas Y.C. Woo. “A Modular Approach to Packet Classification Algorithms and Results,” in *Proceedings of INFOCOM 2000*.

[7]. Wei-Che Chen. “Design and Implementation of Traffic Classification for Layer3 Switch,” July, 1999.

[8]. M.Buddhikot, S. Suri, and M. Waldvogel. “Space Decomposition Techniques for Fast Layer-4 Switching,” in *Proceedings of IFIP Workshop on Protocols for High Speed Networks*, Salem, Massachusetts, August 1999.

[9]. S. T. Chuang, A. Goel, N. McKeown, and B. Prabhakar, “Matching Output Queueing witch a Combined Input Output Queued Switch,” *Technical Report*, CSL-TR-98-758, April 1998.

[10]. W. Doeringer, G. Karjoth, and M.Nassehi,

"Routing on Longest Matching Prefixes," in IEEE/ACM Transactions on Networking, vol. 4, Feb. 1996.

[11]. D. C. Stephens and H. Zhang, "Exact Emulation of an Output Queueing Switch by a Combined Input Output Queueing Switch," in *Proceedings of IWQOS'98*.

[12]. P. Krishna, N. Patel, A. Charney, and R. Simcoe, "On the Speedup Required for Work-Conserving Crossbar Switches," in *Proceedings of IWQOS'98*.

[13]. N. McKeown, "Scheduling Algorithms for Input-queued Cell Switches," *Ph.D. dissertation*, Univ. California, Berkeley, May 1995.

[14]. A. Charny, P. Krishna, N. Patel, and R.simcoe, "Algorithms for Providing Bandwidth and Delay Guarantees in Input-Buffered Crossbar with Speedup," in *Proceedings of IWQOS'98*.

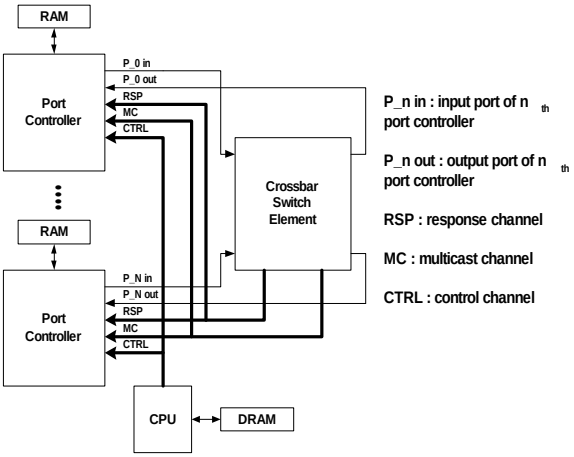

圖 1、系統架構

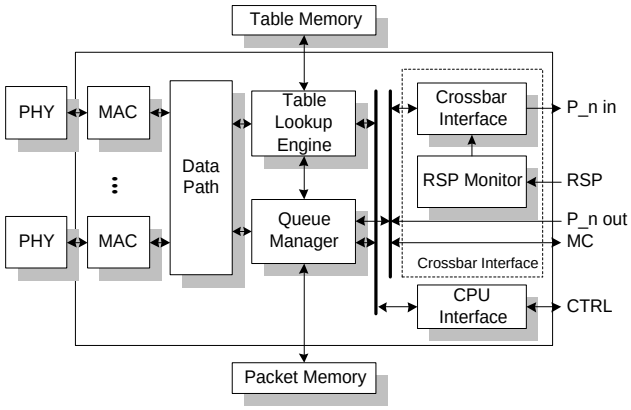

圖 2、埠控制器功能方塊圖

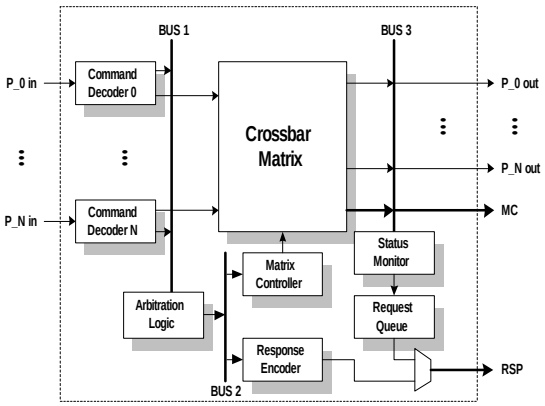

圖 3、交換核心功能區塊圖

(a)

(b)

圖 4.8 SLIP 交換機與我們的交換機在不同負載量之封包平均延遲與輸入佇列長度的比較：（a）平均延遲 （b）平均輸入佇列長度。