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（III）— 子計畫一： ※
※ 使用於寬頻網際網路之 Gigabit 路由器與訊務管制 ※
※ 技術（III） ※
※ ※
※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：□個別型計畫 ▒整合型計畫

計畫編號：NSC 89-2219-E-009-028-

執行期間： 89 年 08 月 01 日至 90 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 90 年 10 月 31 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-009-028-
執行期限：89 年 8 月 1 日至 90 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

本計畫主要研究大容量路由器的核心技術並加以實現。在前兩年的計劃中，我們已經完成了第三層路由器的雛形設計，其中包括了資料路徑模組、佇列管理模組、訊務區分器模組等等。本年度計劃將持續研究高速路由器的關鍵技術，包含了排程技術的開發與交換技術的研究。

在排程技術方面，於實作複雜度的考量之下，第一版的排程器採用了 SCFQ (Self-Clocked Fair Queueing)**機制，但其延遲的保證上並不如** WFQ (Weighted fair queueing)**來的理想，因此我們研究了** FFQ (frame-based fair queueing)**的設計考量與實作，以期改善訊務的延遲與公平性。**

在交換技術的研究方面，在雛型機上使用的佇列管理模組是採用共享式記憶體架構，其交換容量會受限於記憶體的速度，而輸入佇列(input queueing)**架構對記憶體的需求較低，但會有佇列前端擁塞** (head-of-line blocking)**的問題，因此我們研究混合輸入輸出佇列的架構，證明了它可以完全仿效一個輸出佇列的交換機，並探討** CIOQ (combined input and output-queued) **交換機在輸出埠** buffer **有限的情況下的行為。**

關鍵詞：路由器、訊務排程、交換機架構

Abstract

In this project, we design and develop high-capacity routers. We have developed a prototype router, which consists of a data path module, a queueing module, a classifier, and a scheduler in last two years. In this sub-project, we continue developing two key technologies including scheduling and switching.

In the prototype, we implemented SCFQ mechanism due to the consideration of implementation complexity. However, the delay bound and fairness of SCFQ is not as good as WFQ. In new design, we implement the FFQ to improve the performance of the scheduler.

In the queueing management of the prototype, we adopted the shared memory architecture. It has a disadvantage that its capacity is limited for switch with many ports because of heavy memory access. It is possible to develop a large-scale input queueing system, but there exists a head-of-line blocking problem. In this sub-project, we investigate the CIOQ architecture. We have developed a algorithm to emulate an output-queueing switch and evaluate the performance of CIOQ with finite buffers.

Keywords: Router, traffic scheduling, switch architecture

二、計劃緣由與目的

未來的網路服務除了滿足頻寬需求外，網路更必須保證服務品質(Quality of Service, QoS)**，因此在路由器中除了要能把封包加以分類之外，還要對於不同的封包由排程器來處理，以符合不同使用者各自的要求。在排程的技術上，複雜度與效能一直是很難兼顧的，如** GPS、WFQ **雖然擁有好的公平性與** delay bound**，但是，要維持準確的系統時間的需要複雜度高的硬體才能達成；**SCFQ **降低了系統時間的複雜度，卻付出了** delay bound **作為代價，其系**

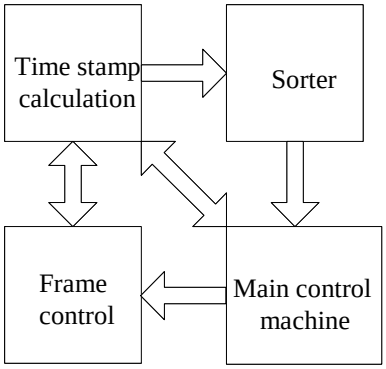
統表現與系統複雜度成正比。因此我們選擇了 FFQ 作為硬體實踐的演算法，因為它保證的效能與 WFQ 相同, 但是大大的減低了實踐的複雜度。

長久以來，因為前端壅塞限制傳輸率的關係，交換系統的設計大都避免採用輸入佇列(input-queued)架構。但是輸出佇列架構或共享記憶體架構的交換系統容量受限於記憶體頻寬，而目前記憶體頻寬尚無法與傳輸速率匹配，因此，隨訊務量急速增加的結果，使得輸入佇列架構受到重視。降低前端壅塞的方法主要可從兩方面下手：(1)在輸入埠中，不再採用先進先出(FIFO)的架構, 而是到相同輸出埠的封包形成一個佇列, 然後搭配 maximal 配對方式，可以達到接近 100%傳輸率的效果；(2)將縱橫式架構加速(speedup)，使其速度超過輸入/輸出線(input/output link), 因為加速的關係，所以輸出埠也需要緩衝器(buffer)，因此變成一個 combined input and output-queued (CIOQ)的架構。在本計劃中，我們證明了 CIOQ 能夠完全仿效一個輸出佇列的架構，並探討了它在輸出緩衝器有限的情況下的考量。

三、研究方法與成果

I.FFQ 排程技術的實現

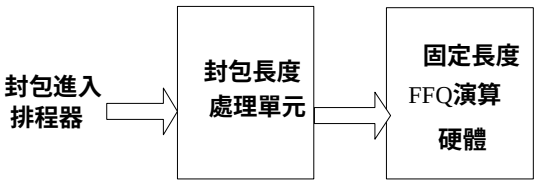
首先我們分析固定長度的 FFQ 演算法加以分析，其方塊圖如圖一所示有四個主要的工作區塊：主要控制單元、時間戳記的計算單元、frame 的控制單元、以及時間戳記的排序單元。當封包進出公平排程器是由主要控制單元來控制, frame 的控制單元則是依照封包進出的狀況以及 frame counter 的變化來決定是否更新系統時間或者是作 frame 的更新。時間戳記的計算單元則是依據現在的系統時間、類別的結束時間以及依據進來的封包的比重計算出它的結束伺服時間，並且判別此封包的結束伺服時間是否超越了現在的 frame, 並加以標記。算完了結束伺服時間的封包便被送到時間戳記的排序單元，依據時間戳記的排序單元加以排序並等待被伺服。



圖一、功能方塊圖

對於一些參數在硬體的實現上我們必須加以設計，如時間標籤格式、frame 大小之決定、比重之決定、與矩陣計數的簡化，之後, 將系統的 state machine 與 I/Q 介面定義出來。

在可變封包長度的環境下，結束伺服時間(finish time)時間可能不是整數，因此需要應用到浮點運算除法，但浮點運算是非常複雜的動作，所以比較好的解決方式是對算法加以簡化，以降低實踐上的複雜度。簡化型可變長度 FFQ 算法概念在於利用單位量化(quantization)的方式來表示浮點數字，也就是說以下的所有時間戳記(time stamp)的表示法再也不是絕對時間表示法，而是以某個量來作基本單位的相對表示法。我們訂定 64 個位元組的傳送時間為其時間基本單位。因此在封包長度也需要以 64 位元組為單位來表示, 因此我們可以利用固定長度 FFQ 的硬體來實現可變長度的 FFQ 如圖二所示。



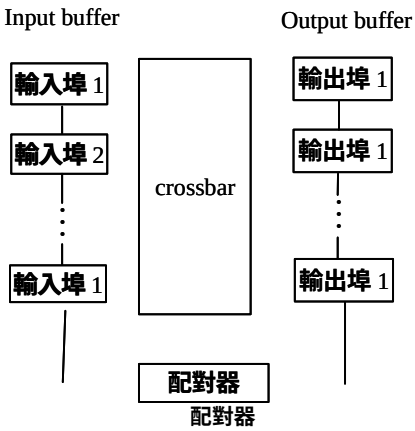
圖二、可變長度 FFQ 演算法的架構

II. CIOQ 交換架構

（一）仿效 output queueing

利用 CIOQ 的架構, 在 corssbar 加速兩

倍的情況下，我們不僅可以消除 HOL 的困擾, 並發展出 LCF/MUF 演算法來仿效一個 output queueuing 交換機。這裡仿效的意思是對相同的輸入訊務，每個封包離開交換機的時間及順序與輸出佇列架構相同，如此就可達到與輸出佇列交換系統相同的效能，也就是說只要這個被仿效的輸出佇列交換系統能達到的服務品質，這個 CIOQ 交換系統也能達到。此外，這個被仿效的輸出佇列交換系統並不限定使用何種服務排程(service scheduling)方法。



圖三、CIOQ **架構圖**

圖三所示是 CIOQ 的架構圖，中間是一個 crossbar switch，輸入與輸出均有 buffer，而每個 phase，corssbar 可以將輸入埠的資料送到輸出埠，而究竟是哪個輸入部會送到哪個輸出埠，則由配對器來決定，因此，為了增加 CIOQ 交換系統的效能，如何有效的配對是一個關鍵。本計劃提出新的配對法與架構，其中使用的配對方式稱為 Least Cushion First / Most Urgent First (LCF/MUF)，定義出緊急的程度，然後將最緊急的資料優先配對。經過證明，在 crossbar 加速兩倍的情況下，CIOQ 可以仿效任意的輸出佇列交換系統。

(二)有限 Output buffer 之分析

在 CIOQ 架構下，每個輸出埠都有自己的 buffer，因此我們需要了解在有限 buffer 的情況下，效能降低的情形。經過模擬實驗，對於 uniform 的訊務，就算是 16X16 的交換機，輸出的 buffer 只要有八個 cell 就可以達到很好的效果。至於 correlated 的訊務，假設其 burst length 為 L，在負載為 0.8 的情況下，則需要 3L 至 4L 的 buffer 才可以達到不錯的效果。

四、結論

本年度計劃延續前兩年的研究成果並擴充交換機的功能。首先我們改進了排程模組，利用實現 FFQ 演算法的方式來改進舊有的 SCFQ 機制，使得公平性與延遲的保證都有改善。在可變封包長度的環境下，利用量化的概念，達到硬體重複使用，使得開發的時間得以降低。

此外，我們進行了交換架構的研究，針對 CIOQ 架構加以分析，並提出 LCF/MUF 演算法，證明出它可以使 CIOQ 交換機仿效一個 output queueing 交換機，並分析了它在有限輸出 buffer 的情況下，其效能降低的情況。

五、參考文獻

[1] Hui Ahang, “Service Discipline for Guaranteed Performance Service in Packet -Switching Networks,” Proceedings of IEEE, vol. 83, No.10, Oct., 1995.

[2] A.K. Parekh and R.G. Gallager, “A Generalized Processor Sharing Approach to Flow Control – The Single Node case,” Proc. INFOCOM ’92, vol. 2, May 1992, pp. 915-24.

[3] J.C.R. Bennett and Hui Zhang, “WF^2Q: Worst-case Fair Weighted Fair Queueing,” in Proc. IEEEINFOCOM ’96, San Francisco, CA, Mar. 1996, pp. 120-128.

[4] S.Jamaloddin Golestani, “A self-clocked Fair Queueing Scheme for Broadband Applications,” in Proc. IEEE INFOCOM ’94, Toronto, CA, June 1994, pp. 636-646.

[5] Dimitrios Stiliadis and Anujan Varma, “Frame-based Fair Queueing: a New Traffic Scheduling algorithm for Packet-Switch Networks,” Tech. Rep.UCSC-CRL-95-39, July 18, 1995.

[6] Dimitrios Stiliadis and Anujan Varma, “Latency-rate servers：A general model

for analysis of traffic scheduler algorithms,”Tech. Rep. UCSC-CRL-95-38, U.C. Santa Cruz, Dept. of computer Engineering, July 1995.

[7] David A. Patterson, John L. Hennessy, "Computer Organization & Design, The Hardware/Software Interface," 2nd Edition, Morgan Kaufmann, 1998.

[8] A. Varma and D. Stiliadis, “Hardware Implementation of Fair Queuing Algorithms for Asynchronous Traffic Mode Networks,” IEEE Communications Magazine, vol. 35, no. 12, Dec. 1997, pp. 54-68.

[9] B. Prabhakar and N. McKeown, “On the Speedup Required for Combined Input and Output Queued Switching,” Computer Systems Lab, Technical Report CSL-TR-97-738, Stanford University.

[10] Karol, M. Hluchyj, and S.Morgan, “Input Versus Output Queueing on a Space Division Switch,” IEEE Trans. Commun., vol. 35, pp. 1347-1763, Dec.

[11] N. McKeown, V. Anantharam, and J. Walrand, “Achieving 100% Throughput in an Input-queued Switch,” Proc. IEEE INFOCOM'96, pp. 296-302.