行政院國家科學委員會專題研究計畫　成果報告

子計畫一：寬頻網際網路中路由選徑技術與 QOS 訊務控制之研究設計(3/3)

計畫類別：整合型計畫
計畫編號：NSC91-2219-E-009-034-
執行期間：91 年 08 月 01 日至 92 年 07 月 31 日
執行單位：國立交通大學電信工程學系

計畫主持人：張仲儒

計畫參與人員：林立峰、黃鏗銘、黃慶喜、丁崇光、陳柏翰、鄭永宏、吳育葵、顏寧佑

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 16 日

計畫中文摘要

由於網際網路的蓬勃發展，使得網際網路上的訊務量急遽增加，且跨網域的比重也驟然提升甚多，除了最初且最基本的數據通訊之外，也加入了語音或影像等即時性類型的服務，故其需要高速的頻寬、有效的 QoS 運作架構與訊務控制機制來確保使用者的服務品質可以獲得保障。

為了解決寬頻的問題，本計畫擬從兩方面來著手：一是直接針對路由器本身著手，增加路由器路由選徑 (Routing) 的處理速度，即是要發展出高速的路由選徑機制；另一方面則是藉著改變架構，利用第二層網路的高速交換(Switch) 動作來取代第三層的 IP 位址路由選徑動作，使可以獲得等效於加快路由選徑速度的效果，我們可以將其視為一種虛擬的路由選徑(Virtual Routing) 方式，如 IETF 所提出的 MPLS(Multi-Protocol Label Switch) 技術。

在提供服務品質保證方面，目前 IETF 已經針對這方面（服務品質保證）的需求成立了相關的 Working Group。其中的 DiffServ 並不針對 per-flow 的訊務提供服務品質保證，而是將訊務分為幾種不同的等級種類(Class)，再對每一種類提供不同的服務品質。因此 DiffServ 在實際的運作方式上，必須在路由器上對訊務進行分類(Classify)的動作並施予不同的排程(Schedule)處理；同時也必須引入連線導向（電信）網路中訊務控制的概念。

所以在網際網路服務品質保證機制方面，我們將針對 DiffServ 的架構以及其中所必需的訊務控制，配合我們既有多年來在 ATM 網路上發展訊務控制機制的經驗，擬設計出適合於網際網路中，特別是具有 IntServ、DiffServ、或 MPLS 等相關 QoS 機制的網際網路環境之呼叫允諾控制法，對新進的連線作系統資源的確認，再決定是否接受此連線的訊務，以更進一步確保能提供各種類的訊務所要求的 QoS。在設計呼叫允諾控制法的同時，也投入標準中並未明確定義的使用參數控制機制的研究，對訊務進行監控以確保進入網路的連線訊務的合法性，維持呼叫允諾控制的正常運作。

最後我們將利用乏晰控制和類神經網路理論，選擇適當的乏晰控制器與類神經網路架構，發展出各相對應的乏晰和類神經訊務控制法則。並完成系統模擬程式的撰寫，利用電腦模擬的方式來驗證所獲得研究成果的正確性。

關鍵詞： 路由選徑、多協定標記交換、差異化服務／差別服務、訊務控制、呼叫允諾控制、使用參數控制

Abstract

With the blooming of Internet application services, the Internet traffic flow increases dramatically and the traffic flow from inter-network transmission also increases rapidly. Not only the basic data transmission but also some multimedia services (such as: voice, real-time video services) are also carried on the Internet. Henceforth, to provide the QoS-provisioning services, the larger bandwidth capacity, effective QoS-provisioning service framework, and traffic control mechanisms are the primary requirements in the design of the Internet.

For the issue of larger bandwidth capacity, 2 solutions are proposed in this project: faster routing and virtual routing. For the faster routing, a better routing path search scheme is proposed to speed up the performance of the router. The virtual routing technique is to apply the high speed layer-2 switching mechanism on the router to replace the original layer-3 routing mechanism, such as: IETF's MPLS (multi-Protocol Label Switch).

As to the QoS-provisioning services, the DiffServ (differentiated Service) service model is considered. An IETF Working Group is dedicated to the research of DiffServ service model. The goal of DiffServ is not to provide QoS-provisioning services on the pre-flow basis; instead, on the class basis. The DiffServ defines several service classes and each class has its own QoS requirements. Therefore, when the DiffServ services are provided on the Internet, the routers must classify all the input traffic and assign different scheduling priority level. Also, it is necessary to add the call control mechanism that has been well-developed in the connection-oriented network.

Broadband Network Lab has devoted to the research of the traffic control mechanism on the ATM network for a long time and gained some precious experiences. Based on those experiences, an appropriate call admission control (CAC) scheme is proposed for the traffic control of the DiffServ. Moreover, the proposed CAC scheme can also applied to other Internet QoS-provisioning services, such as: IntServ, DiffServ, MPLS and etc. On the receiving of new connection request, the CAC will check if there is available system resource for the new request. In such way, network can provide QoS guarantee for both the existing connections and new connection. Also, we will study the Internet usage parameter control (UPC) mechanism that is not clearly defined in the IETF specification. By monitoring the

traffic flow, the UPC can make the validation of the connection, such that the CAC will operate properly.

Finally, we will apply the fuzzy logic control and neural network mechanism on the CAC and UPC schemes. By choosing proper fuzzy logic controller and neural network architecture, we will propose the corresponding fuzzy and neural network traffic control rules. To verify the proposed schemes, we will build system simulation environment and simulate the schemes by software.

Keywords：Routing, Multi-Protocol Label Switch (MPLS), DiffServ, Call Admission Control (CAC), Usage Parameter Control (UPC)

目錄

一、計畫緣由及目的

從 1969 年發展至今的網際網路(Internet)，已經由最初實驗性的研究成果，在歷經幾次的變革之後普及至教育和商業環境；更由於其跨網路、跨地域（無遠弗屆）、極富有彈性(Flexibility) 與可能性(Possibility) 的特點，吸引了愈來愈多人投入此一新的網路世界，然而它的成功卻也加速顯現其發展的瓶頸。擁有超過兩千萬個節點及上億使用者的 Internet，必須進行大幅度的改造，才能進一步像電話一樣普及，事實上這些改造的研發工作也從未間斷。1996 年 10 月，全美三十四所大學宣佈合作建造 Internet2，加速此迫切的改造。1997 年 2 月克林頓(Clinton) 政府也提出新世代網際網路 NGI (Next Generation Internet) 五年計畫(1998-2003)，以配合延伸 Internet2 的構想。

架構出網際網路的網際網路協定(IP Protocol) 原本只是適用於數據 (Data) 通訊的第三層網路協定，並藉著路由器連接不同的網路／網域形成一非通訊連線導向(Connectionless) 的廣域網路。其目的，主要是希望透過一致的定址方式與有效率的路徑演算機制，達成不同網路中各通訊端點之間資料的傳輸、遞送，而此也已滿足一般數據資料通訊的需求。然而在網際網路蓬勃發展之際，也使得許多非單純數據傳輸的應用，例如語音或影像等具備即時 (Real-time) 傳輸要求的服務也採用 IP Protocol 進入網際網路中，再加上多媒體技術的突飛猛進與其服務的普及，使得對網路頻寬與服務品質的需求也相對應地提升，這是當初設計時所始料未及的，但也正是這些各式各樣的應用帶動了網際網路的蓬勃發展與目前的成功。所以，為了確保目前和未來可能的各項應用與服務能夠在網際網路上運作順暢甚或具有一定的品質，以維持網際網路的永續發展，有許多的改革動作也應運而生，其中最主要的即是在「高速頻寬傳輸」與「服務品質保證(QoS guarantee) 運作機制」上的研究。

在高速寬頻傳輸方面，除了在實體網路的傳輸鍊路(Link)傳輸頻寬與設備處理速度的提升之外，最重要的關鍵與瓶頸還是在於構成網際網路的核心設備－第三層路由器(Router) 的路由選徑(Routing) 處理速度。網際網路可以說是 OSI 通訊協定堆疊架構中第三層的網路系統，藉著路由器連接不同的（區域）網路／網域以形成廣域的網路系統。在以往網際網路與其上的應用仍未如今日如此普及之際，多數的訊務仍屬與區域子網路的範圍，藉由路由器連往其他網路的跨網路訊務並不多，子網路的內部訊務與跨網域訊務量呈現 80/20 的比值（80%為子網域訊務，20%為跨子網域訊務）。由於近年來網際網路上服務的多元化使得網路訊務量暴增，網路的流量不再遵守過去的 80/20 定律，而演變成 20/80 的分佈，使得路由器的負載量增加。再加上多媒體技術的突飛猛進與其服務的普及，使得這些跨網路的訊務也多屬於多媒體通訊的訊務，對頻寬與網路服務品質保證的需求也相對應地提升。這些現象皆會增加路由器的負擔，使得傳統路由器的效能成為

網路上的瓶頸。因此，提升路由器的處理速度以解決此問題成為一必然的趨勢。目前在這方面的研究可以分成兩大類：一是直接針對路由器本身著手，實際增加路由器路由選徑 (Routing) 的運算速度。這部分除了直接提高硬體運作平台速度的方式之外，即是要發展出有效率的路由選徑運算方法。傳統上主要是以純軟體操作方式的角度，來設計出一較佳的路由資訊的資料結構(Data Structure) 以及相對應的比對搜尋演算法(Algorithm)，而近幾年來，開始發展出硬體架構導向(Hardware-oriented) 觀念的路由選徑機制，即是期望所設計的路由比對搜尋演算法能夠適合以實際的硬體邏輯＆運算電路來予以實現，以較軟體操作程序高速的硬體運作方式來提升路由選徑的速度，亦即提升路由器的處理速度。在超高速 Gigabit 網路的路由器裡，假設網路每秒可傳送 1 Giga bits 的資料量，若在網路上平均每個封包的大小為 512 bits，則每個封包大約只允許 500 ns 的處理時間，這還不包括在封包進來路由器時的佇列延遲問題，而現今的路由器大多無法達到此一速度的要求。因此在本計畫中，我們將對既有的路由選徑的方式（路由表格查詢法，Routing Table Lookup）與前述幾種路由選徑速度提升的方法或架構進行研究，以提出符合 Gigabit 超高速網路中的高速路由選徑的方法。此外，隨著多樣化的即時(Real-time) 影音多媒體通訊服務的蓬勃發展，使得網際網路上的群播路由(Multicast Routing) 技術日益重要，以節省網路上所需流通的訊務資料量我們希望藉由上述在設計傳統 Unicast 路由選徑方法的技術以及經驗，能經由簡易的修改應對而快速地實現群播功能所需的群播路由(Multicast Routing) 機制。

另一類提升路由器處理速度的方式，則是藉著改變 IP 網路的運作架構，利用連結導向(Connection-oriented) 的第二層（電信）網路的高速交換(Switch) 動作來取代第三層非連結導向(Connectionless) 的 IP 協定網路的路由選徑動作，使可以獲得等效於在第三層 IP 網路上加快路由選徑運算速度的效果，我們也可以將其視為一種虛擬的路由選徑(Virtual Routing) 方式，如 IETF 所提出的多協定標記交換技術(Multi-Protocol Label Switch, MPLS) 技術。MPLS 可以視為是將 IP 路由器(Router) 建立在第二層網路的交換器(Switch) 上，或者可以說是將原本單純的 IP 路由器功能加以擴充，包含進第二層網路的交換器(Switch) 功能，並依照 IP 封包的終點位址，在 IP 封包上加上一個較短的、屬於第二層交換網路的交換標籤(label)，之後便能夠使此擴充功能的路由器直接透過其第二層的交換機能依照這個短標籤，而不是傳統的 IP Longest Prefix Matching 方式，迅速的交換 IP 封包，改善 IP 封包路由交換的性能。而這種結合傳統第三層 IP 路由選徑以及第二層高速交換機制的路由器可以稱之為交換路由器(Switch Router)；而此種 IP 封包路由交換技術也可稱之為 IP Switch 技術。

目前在以第二層交換網路技術實現的 MPLS 網路中，以 ATM (Asynchronous Transfer Mode) 網路被視為是最好的實現平台，採取結合 IP 與 ATM 交換網路的模式來提供最高性能的 IP 封包轉送能力，而其中用於取代 IP 路由選徑的交換用短標籤即是沿用 ATM 交換網路的路由識別標示 VPI/VCI (Virtual Path

Identifier/Virtual Circuit Identifier)，並以 ATM-LSR (Label Switching Router) 做為實際封包路由轉送的關鍵設備。ATM-LSR 的硬體架構主要是延襲自 ATM 交換機，並再擴充結合 IP 路由器的功能與 MPLS 網路所專有的標記交換協定(Label Switch Protocol) 等能力。

在 MPLS 網路的基本運作中，一個 Switch Router 處的一個交換用短標籤即是對應至一組｛Source IP (SIP), Destination IP (DIP)｝連線路由。此短標籤如同傳統交換網路的交換用短標籤一般，是屬於「區域性(Local)」以及「可重複利用(Re-usable)」的資源，以增加系統所能夠支援的同時在線(on-line)的 IP 連線路由數量，提升網路規模。為了能夠進一步充分利用有限的標籤資源，讓有限的標籤資源能支援更多的同時在線(on-line) IP 連線路由，提高 MPLS 網路規模的擴充性(Scalability)，一個重要的功能便是標記整合(Label-merge)機制：在一 Switch Router 中將多個前往相同目的網路或節點的 IP Route 轉換、對應(mapping)、整合成相同的短標籤，如此一個標籤便只對應至一個目的網路或節點｛Destination IP (DIP)｝。而必須注意的是，一旦將多個 IP Route 以相同的 Label 整合之後，原本各 IP Route 訊務流中的 Data 便無法再在此一標籤交換網路的層次予以分離，必須要至此一整合標籤的 IP Route 的終端節點上將 Data 還原至 IP 封包的層次後，才能夠再依據其不同的 Destination IP 資訊進行不同的路由轉送。而實現在以 ATM 技術為基礎的 MPLS 網路時，所對應的 Label-merge 技術即為 VC-merge：將對應至不同 IP Route 的交換短標籤 VPI/VCI 整合對應成同一個 VPI/VCI。但由於 ATM 網路的封包（稱為 Cell）容量較 IP 封包小，因此大多數的 IP 封包會被分割成多個 ATM Cell 來傳送，如此當啟動 VC-merge 功能時會產生問題：若在交換器處逕行將個別抵達的 ATM cell 的 VPI/VCI 標籤進行 VC-merge 的轉換與整合，原本分屬於不同 VPI/VCI route 中不同 IP 封包的多個 Cell，有可能會被以相同的 VPI/VCI 但交錯的(Interleaving) 順序送出，則導致在共同的目的節點處無法以傳統 ATM 的 cell 接收與重組機制將資料正確地還原成上層的 IP 封包，以順利地分離出經由 Label-merge 或 VC-merge 機制融合的訊務。一簡單的解決方案是 Frame-level Interleaving，即是在 ATM-LSR 交換機的 Input 端設置封包重組緩衝器(Reassembly Buffer, RB)，把屬於同一 IP 封包的多個 ATM cell 收集完整後，再進行 VC-merge 的 VPI/VCI 標籤轉換，並以連續輸出(back-to-back) 的方式將屬於同一 IP 封包的 cell 送出，待同屬一個 IP 的 cell 都送出後，再進行下一組屬於同一 IP 封包的 cell 輸出。然而，勢必將會因為重組緩衝器的設置而增加了 ATM-LSR 的記憶體需求，也可能影響資料傳輸的延遲(Delay)。因此在本計畫中，我們探討具有 VC-merge 能力的 ATM-LSR 交換機的性能：分析 VC-merge 的 ATM-LSR 所需的緩衝器和 cell blocking 機率之間的關係，以提供實際 ATM-LSR 設計上的參考，並和傳統不具有 VC-merge 能力的 ATM Switch 性能作比較，試著去探討具有 VC-merge 能力的 ATM-LSR 交換機需要比傳統的 ATM 交換機具備多少緩衝器資源，是否需提供額外的大量緩衝器以達到和傳統 ATM 交換機相同的 cell blocking 機率。

此外，近年來諸如語音、影像等即時性服務在網際網路上逐漸成為重要的網路應用型態，其需要仰賴寬頻高速的網路傳輸以維持良好的品質表現，而在此同時，MPLS 技術的提出的確適時為網際網路提供了高速且低延遲的訊務傳輸能力，但相對的，當高速的 MPLS 網路發生傳輸路徑錯誤或損壞的時候，往往也會造成更嚴重的影響（例如更大量的資料遺失），尤其對於即時性服務而言更是如此。因此我們針對 MPLS 網路，在其傳輸路徑保護(Protection) 以及發生錯誤(failure) 時之路徑回復(Path Recovery) 機制方面進行深入的研究，提出一套有效的路徑保護／回復機制，以便於 MPLS 網路傳輸路徑發生錯誤或損壞時還能夠維持部分基本的通訊，並可以**快速而正確**地恢復既有的通訊，降低 MPLS 網路上傳輸路徑錯誤或損壞所帶來的影響，減少封包遺失率，並期望能夠進一步達到動態**負載平衡**的附加效益，使系統資源做最佳的利用，如此便可有效的提高網路的資料輸出率(throughput)。此外，我們另一項著手的重點便是針對此機制發展出一套**系統化的方法**，讓業者在採用此方法時可以根據其需求與使用者付費原則，評估採用不同複雜度的運作形式與保護程度，在成本與演算法完整性之間取得一平衡點。最後我們以 ns-2 此套網路模擬軟體進行該路徑回復機制的效能評估與驗證，以貼近實際的運作狀況與結果。

提升網路頻寬或可稍微改善服務品質(QoS)，但由於既有的網際網路協定本質上是屬於盡力式(Best-effort)的服務，所以仍無法從根本上做有效地改進，必須要再配合其他的機制來達成服務品質保證，如此也才能夠確保頻寬獲得最**有效益**的利用。目前這仍是屬於新的研究領域，相關標準並未完備，而相關研究論文數量也不多，有許多值得研究的課題和空間。目前網際網路的標準組織 IETF(Internet Engineering Task Force) 已經針對這方面（服務品質保證）的需求成立了相關的 Working Group，制定了一些關於網際網路服務品質保證的訊務控制機制或運作架構等解決方案，例如：RSVP(Resource Reservation Protocol,資源保留協定)、IntServ(Integrated Service,整合服務)[23-26]、DiffServ(Differentiated Service, 差別服務)[28-33]、QoS Routing(服務品質路由選徑) ……等等。

IntServ 主要是想針對 per-flow 的單一訊務提供服務品質保證 [23]。在 IntServ 中，除了保留原有基本 Best-effort 方式的服務之外，另外定義了兩種新的、具有品質保證的服務方式：Guaranteed 以及 Controlled-Load 服務 [24-25]。引入電信網路通訊連線前呼叫允諾控制(CAC, Call Admission Control) 的觀念，其必須配合 RSVP 對 per-flow 做到在通訊之前，先根據其服務品質要求，在通訊路徑上的每一個路由器保留足夠的資源，來更進一步達成具有 QoS 的服務。雖然 IntServ 的架構(ISA, Integrated Service Architecture) 已經趨於成熟，但是由於其針對 per-flow 的訊務提供服務品質保證的特性，在原本即不具備 flow 識別功能的 IP 網路上則必須透過額外的 flow 識別分類機制來輔助，若要能夠支援足夠數量的同時在線(on-line) flow 的訊務流，則必須要有足夠大的 flow 識別記憶體或資料庫，並且 flow 識別機制的運作速度也不能太慢而影響整體速度，再加上

所有的訊務控制機制都必需要做到 per-flow 的處理方式，也使得網路設備的工作量與複雜度都增加，負荷(Loading)變重，因而導致其擴充性(Scalability) 不佳：當實行的網路規模不大（例如在一區域網路中）的時候還能夠維持服務品質，但是當實行的網路範圍擴大時，負擔便會急速增加而不易維護，故目前不適宜在廣域網路或骨幹網路上實施。

DiffServ [28] 可以說是在對 IP Network 上服務品質的迫切需求，以及希望能獲得儘早實現的壓力下應運而生的。DiffServ 可以視為是 IntServ 的改良版本，它並不針對 per-flow 的訊務提供服務品質保證，而是將訊務區分為**有限的**幾種不同的服務等級(Service Class)，而每一個服務等級即對應至一種訊務服務品質，然後只針對個別 Service Class 的整合訊務進行 QoS 的處理，而不再對其中單一的 per-flow 訊務進行處理，如此便能夠解決 IntServ 中 per-flow 處理方式的複雜度所造成的網路 Scability 受限的問題，加速具 QoS 保證的網際網路的實現。目前 DiffServ 相關的 RFC 在 Best-effort 之外共定義了 5 個 Service Class [28-33] 並可區分為兩大類，分別是 Expedited Forwarding (EF) Service [29-30] 以及 Assured Forwarding (AF) Service [31-32]。這兩大類的服務所要求的服務品質並不相同，其中 Expedited Forwarding (EF) 服務所欲達成的 QoS 是比較嚴格的，要求 Delay、Jitter、Loss 都必須獲得保障；而 Assured Forwarding (AF) 服務則可以容許較大的 Traffic Burst，故只要求 Loss。為實現 DiffServ 此網際網路 QoS 架構，IETF 也在相關的 RFC 中定義 DiffServ 網路設備應具備的元件以及其系統架構參考模型 [28]，其中以「訊務封包分類器(Packet Classifier)」與「訊務監控調節器(Traffic Conditioner)」為基礎關鍵元件，是為其他 QoS 訊務控制、處理機制（例如：CAC 連線允諾控制機制、Scheduling 排程控制、流量控制、壅塞控制等 Per-Hop Behavior）運作的基礎。與 IntServ 處相同的地方是，在原本即不具備 flow 識別功能的 IP 網路上則必須透過額外的 flow 識別分類機制來輔助，才可能進一步進行以 flow 為基礎的訊務控制機制並達成 QoS 服務品質保證的目的，而 Packet Classifier 即是負責此一 flow 識別分類機制的元件。而與 IntServ 中的 flow 識別分類機制不同之處在於，DiffServ 的 Packet Classifier 並非以「1 identification rule-to-1 flow」的方式用於鑑別出單一連線的 Traffic flow，而是以「multiple identification rules-to-1 service class flow」此含有後續匯整功能的識別方式，亦即「分類(Classification)」的方式，辨識並區隔出不同 Service Class 的訊務，以便接下來能夠透過不同的訊務處理方法而達到差別化服務品質的目的。Packet Classifier 的分類法則可以是多樣的，根據 IP 封包中不同的欄位（例如：Source IP、Destination IP、Transport-layer Port、ToS/DSCP……等）與其上記載的訊息，或單一欄位、或多欄位組合的條件方式進行訊務的分類。而其效能訴求主要是簡單而高速的運作速度，並期望具備較小的記憶體需求。

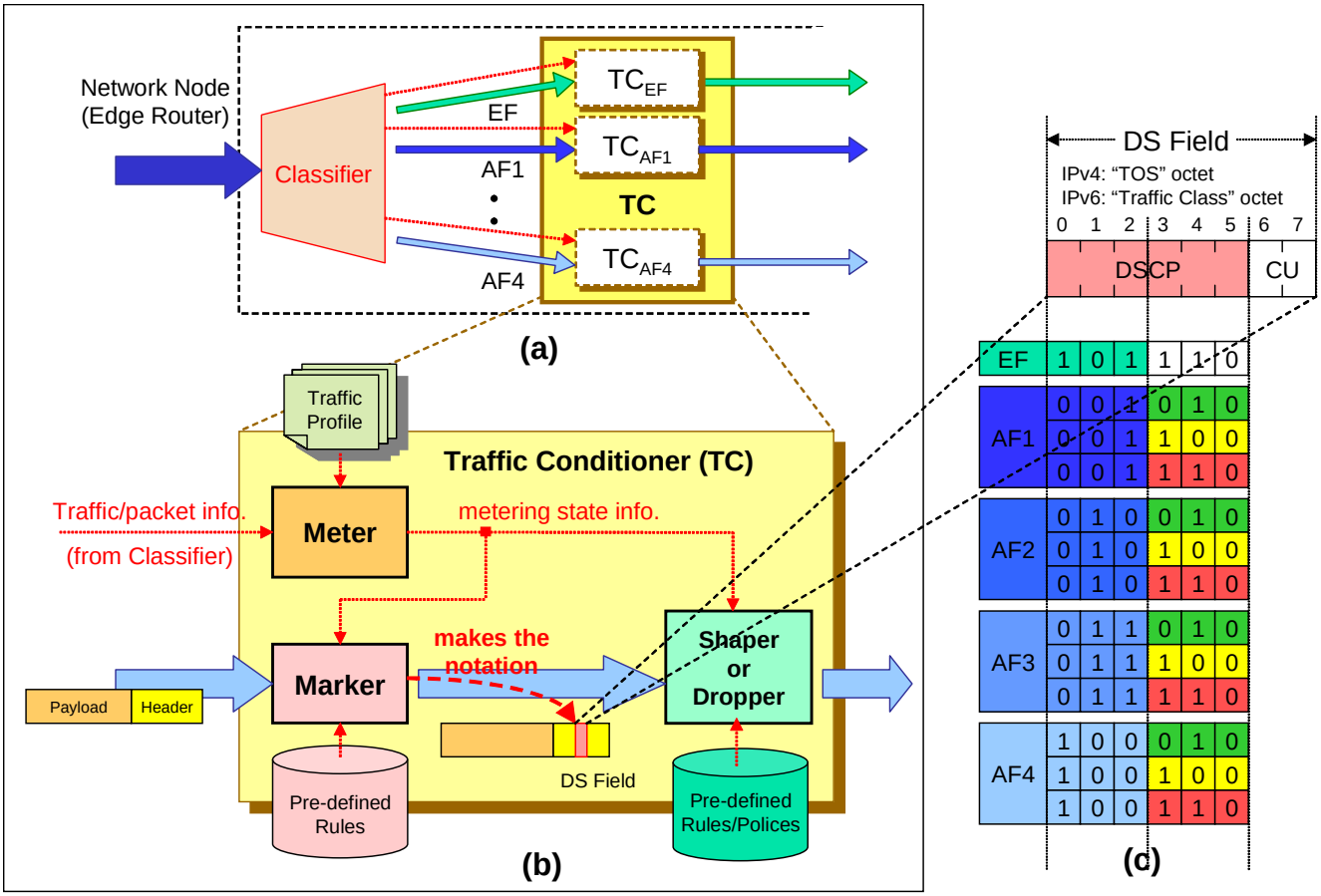


圖 0: (a) Classifier 與 Traffic Conditioner (TC) 之運作邏輯關係架構, (b) Traffic Conditioner 之功能架構圖, (c) DSCP 欄位與各 Service Class 標記代碼示意圖

至於 DiffServ 的另一個基礎關鍵元件－訊務監控調節器則是對進入 DiffServ 網路的訊務進行監控，確保其各 Service Class 訊務的統計特性皆能夠符合其 Traffic Profile 協議的條件。圖 0(a)所示即是訊務封包分類器(Packet Classifier)與訊務監控調節器(Traffic Conditioner)於一 DiffServ 網路設備中的運作邏輯關係架構。此外訊務監控調節器也會針對所監控的訊務，根據其符合 Traffic Profile 的程度狀況(Conforming Degree) 或是其封包內容的重要性，對其封包進行多層次的標記並對應至不同程度的 Dropping Precedence，以提供並輔助當網路壅塞情形發生而啟動 Packet Dropping 機制時，欲達成差異化 Packet Dropping 所需要的 Packet Dropping 優先順序的參考依據。目前 IETF 已對各 Service Class 的訊務定義了 3 個基本且必要的 Packet Dropping Precedence，分別具有 High、Medium、Low 三種 Dropping 機率，因此訊務監控調節器也將配合分別以紅色(Red)、黃色(Yellow)、綠色(Green) 三種顏色來進行標記。圖 0(b)與(c)分別顯示了一訊務監控調節器的功能架構圖，以及定義在 DSCP 處的 Service Class 與 Color（或稱

Dropping Precedence）的整合性標示定義。一網路訊務進入 DiffServ 的 Edge 設備時，會先由 Packet Classifier 進行分類以識別出其中各封包所屬的 Service Class，接下來訊務監控調節器中的 Meter 會對各 Service Class 的訊務封包進行統計並分別與其 Traffic Profile 比對，並將此結果傳送至 Marker；Marker 在得到 Meter 所做的訊務統計與其和 Traffic Profile 之間的比對數據後，會根據事先定義好的決策法則(Decision Rule) 進行訊務封包的實際標記動作，而此標記如同圖 0(c)所示，是包含了該訊務封包所屬的 Service Class 以及所對應 Dropping Precedence 的顏色標記。訊務監控調節機制的效能訴求，則是期望能夠達到在確保訊務特性符合 Traffic Profile 規範的同時，精確地讓系統資源獲得充分且最佳的利用，並進一步保障其中各 micro-flow 連線標記的公平性(Fairness)。由於 DiffServ 的各項訊務處理或 QoS 控制機制皆是以一 Service Class 的訊務流為基礎，即使能夠達成預定的訊務控制目標，卻也不能完全保障此時 Service Class 中各 micro-flow 連線所實際獲得的 QoS 能和控制目標相當，如此將導致 DiffServ 此項網際網路 QoS 架構方案無法獲得廣大的採用與實現。因此若能夠同時在 micro-flow 的連線標記公平性上有所改善，則將使各 micro-flow 實際所得到的網路資源和 QoS 也較為公平，如此才能夠有效**提升 DiffServ 網際網路 QoS 架構的可行性與使用效益**。

接下來我們將分別就本計畫在「高速頻寬傳輸」與「服務品質保證運作機制」上的各項具體研究成果，進行方法說明與成果討論，

二、研究方法、成果與討論

1. 高速的路由選徑機制

這部分主要是發展適合硬體實現(Hardware-oriented) 的高速路由選徑方法，期以硬體的運作方式加速路由選徑的運算速度，以滿足 Gigabit 超高速網路環境下以及未來更寬頻的網際網路應用的需求。

1-1. 高速單一路由選徑(Unicast Routing) 機制－階層分群解析架構

經過歸納與研究分析的結果，適合硬體實現的路由選徑方法應具備有下列的特性與概念：「固定資料長度」的資料運算動作，以及「規則化」的、「反覆運作」的處理程序(Process)，而「階層式分群解析」（或稱「多層次群組解析」）的方法即具備有上述的末兩項特點，再配合上以階層式的多元完全展開樹（Trie）來做為其將整個 IP 位址進行多層次的 IP 位址區段分群的參考架構後，便具備有「固定資料長度」的資料運算動作的條件，因而極適合於做為採用實際的硬體邏輯＆運算電路來實現的路由選徑機制。「階層式分群解析」的方式是：將整個 IP 位址進行多個層次的 IP 位址區段分群－先進行第一層次的較粗略分群，再依據實際路由表格中路由字首(Route Prefix) 的資訊，針對有需要做進一步細部分群解析的群組（即包含一個以上，對應至更小範圍 IP 位址區段的路由字首）進行下一層次更精細的分群展開, 如此反覆運作至每一分群中沒有對應至更小範圍 IP 位址區段的路由字首為止。接下來將每一層次的每一 IP 位址區段分群與該 IP 位址區段路由選徑的結果進行對應並以表格紀錄（「分群-路由結果」對應表格）：若為毋須再進行下一階段細部分群解析的群組，將必然可以對應至一個該 IP 位址區段的(共同)路由選徑結果（即封包的輸出埠(output port)）；若為需要再進一步細部分群解析的群組, 則可以對應至一個「必須進行下一層次更細部分群解析」的指示，並且指向連結至該進一層次的「分群-路由結果」對應表格。進行實際路由選徑的應用時，只要將所欲查詢的目的 IP 位址(Destination IP Address) 與各層次的 IP 位址區段分群進行比對，尋找其所屬的最細的分群，待確定目的 IP 屬於何層次的某一分群後，即可以由該層次的「分群-路由結果」對應表格直接查表得知其路由結果輸出埠。而此「比對」動作是規則地、次第從第一層分群開始，再視需要逐步往分群更精細的層次檢視、比對。由於是將路由選徑之搜尋演算動作化為「規則」化的、「反覆運作」的多層次比對與查表動作，因此已具備適合硬體實現的初步條件了。

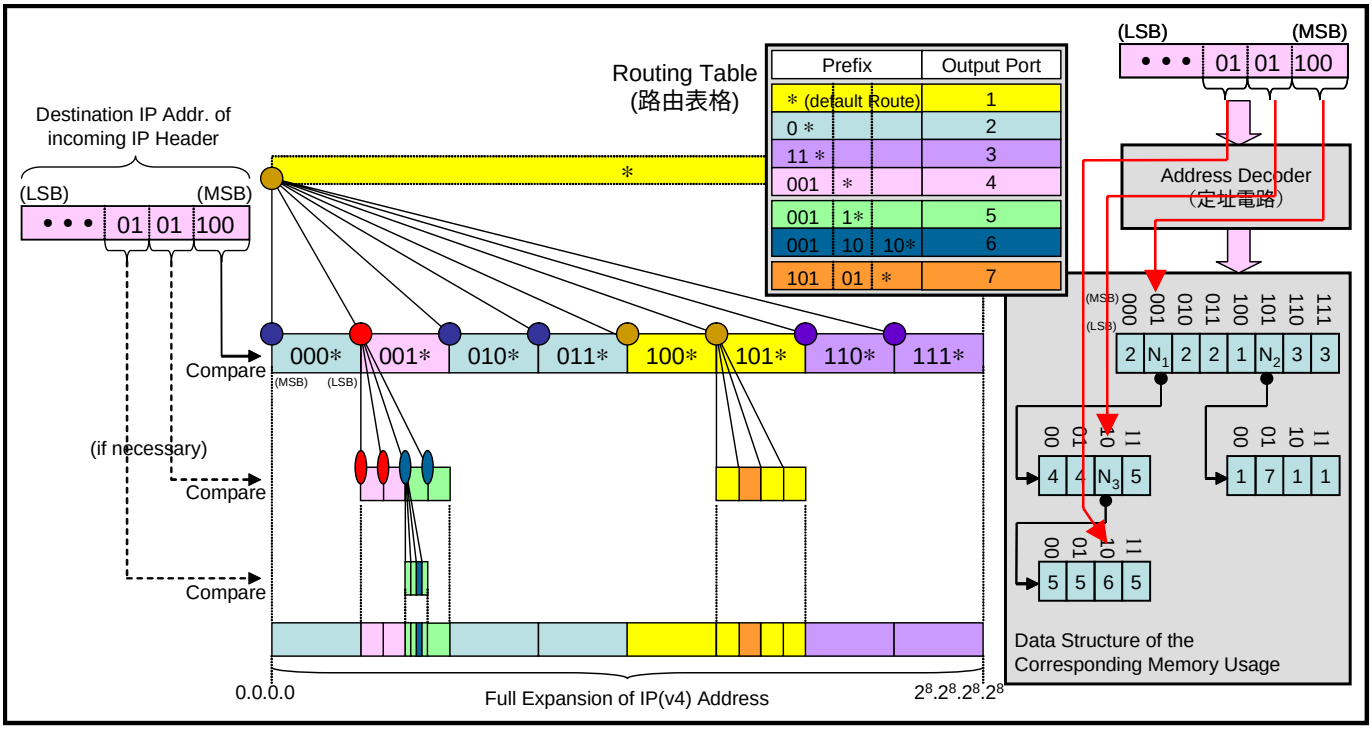


圖 0: Trie-based 階層式分群解析路由方法

因此我們將以「階層式分群解析」的方式為主，來發展適合硬體實現的高速路由選徑方法。而我們所提出的作法，是以階層式的多元完全展開樹（Trie）來做為「階層式分群解析」方法中，將整個 IP 位址進行多層次的 IP 位址區段分群的參考架構—也就是使每層次的 IP 位址區段分群恰好為一 N-bit ($1 \le N < 16$) 的完全展開樹(Trie)，如此一分群的 IP 位址區段範圍皆可以用一個 IP 位址字首(Prefix) 的形式來表示（例如：140.113/16）—再依據實際路由表格中路由字首(Route Prefix) 的資訊，進行實際階層式的多元完全展開樹的建置工作，在每一層次中，僅針對有需要做進一步細部分群解析的群組（即包含一個以上對應至更長路由字首的次 IP 位址區段）進行下一層次更精細的 N'-bits Trie ($1 \le N < 16$)的分群展開（以上的敘述可以參見圖 0 所示）。每一層次皆對應至一組表格，紀錄此層次中各分群 IP 位址區段(Segment) 及其所對應的路由選徑結果，或是必須進行下一層次更細部分群解析的指示。而針對有需要才進一步細部分群解析的方式也比起純粹的（階層式）多元完全展開樹方法大大減少所需的記憶體容量。實際應用於進行路由選徑查詢動作時，即是將所欲查詢封包的目的 IP 位址(Destination IP Address) 與各層次的分群進行比對：由第一層次分群開始，直至其所屬的分群不再有進一步的分群解析為止，此時只要查詢該層次的「分群-路由結果」對應表格即可得到該封包的路由輸出埠。而由於我們是以階層式的多元完全展開樹

（Trie），來做為其將整個 IP 位址區段進行多層次分群的參考架構，每層次的 IP 位址區段分群恰好為一 N-bit (1≤N<16) 完全展開樹(Trie)，因此可以如同圖 0 右半部份所示，進一步將目的 IP 位址在各層次的「分群比對」動作，轉換為 N-bit 固定長度的「定址」動作，而所定址到的記憶體內容即儲存該階層中該 IP 位址群組(Segment) 的路由選徑結果(output port)，或是指向儲存著下一層次分群解析的路由選徑結果的 Pointer，如此也相當於將原本分群比對之後的「分群-路由結果」表格的查詢動作也一併整合進來了。也就是說，只要透過反覆的（固定資料長度的）定址動作，即可以得到路由選徑的結果。至此，我們已將傳統路由選徑之搜尋演算動作，轉化為一套系統化的階層式路由資料結構，與一「規則」化的、「反覆運作」的多層次「定址」動作，而這些都可以採用實際的硬體邏輯＆運算電路來加以實現並獲得加速的效果，例如其中的定址動作便能夠採用定址電路來達成。此時 Routing 的速度則完全取決於「記憶體定址」的「存取次數」與「定址電路運作速度」。

為了決定實際運作時，所使用的實際階層分群（參數）的設定，包含階層數目與每一階層中的分群大小（因為每一階層是為一個 N-bit 的多元展開樹，所以這裡也相當於是在決定每一階層的 N 值大小，亦即多元展開樹的大小），以及所需要的記憶體容量大小，因此我們進一步探討 Routing Table 的特性，並再深入瞭解所設計方法的特點，以及兩者之間的關係。對於每一種路由選徑方法而言，不同的 Routing Table 資料皆會因為其中路由字首數據的不同，諸如路由字首數量多寡、分佈的型態／趨勢、分佈的量值大小差異，造成其不同的運作記憶體容量需求。在觀察我們所設計的路由選徑方法之後發現，對於一筆 Routing Table 資料而言，不同的階層分群方式會影響儲存路由結果的資料結構所需要的記憶體容量，這也表示我們可以藉由適當地設定階層分群方式，來獲得最小記憶體容量。透過初步的檢驗程序發現，增加分群階層數可以使每一階層的完全展開樹規模較小，相對上每一分群範圍較廣，因此可以較有效率地針對有需要再進一步分群解析的 Subnet 才進行下一階層的展開，所以可以相當有效地減少記憶體需求，然而卻有平均記憶體存取次數隨之遞增的缺點，因而我們必須在分群階層數與記憶體存取次數之間權衡一最佳點。此外觀察也發現，記憶體容量需求也並非隨著階層數的增加而永遠呈等速率地減少，而是會趨向一飽和值。因此在綜合考量階層數變化對於記憶體容量以及存取次數的效應之後，我們決定採用 5 層次的分群解析架構。

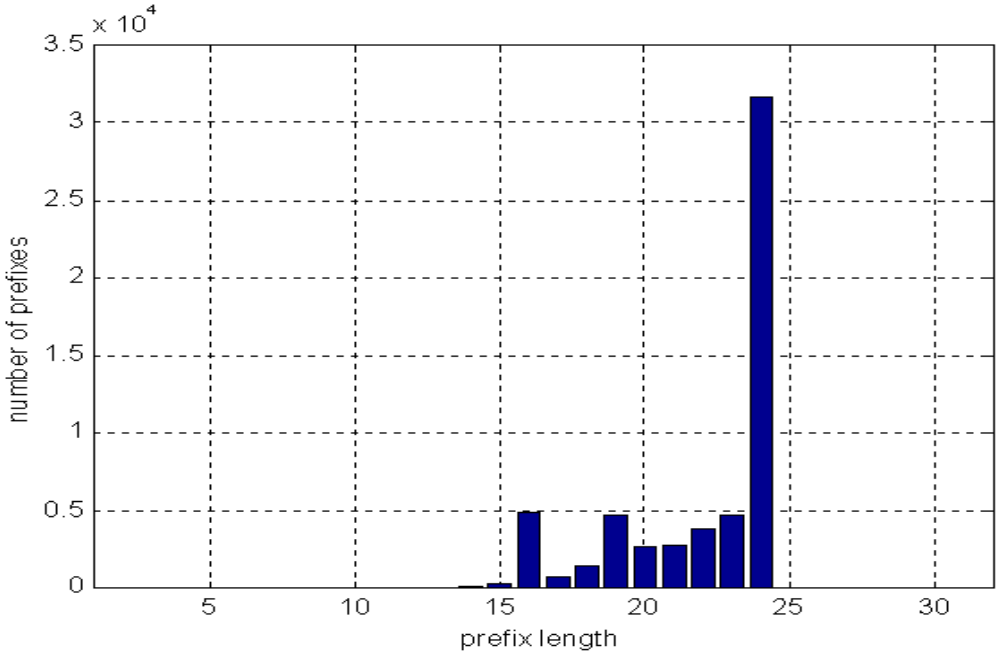


圖 0: Routing Table 中路由字首的分佈狀態圖

（本圖數據參考自 MaeEast 2000/07/18 的統計資料[4]）

最後我們更進一步地分析實際 Routing Table 中路由字首的分佈狀況，作為階層式多元完全展開樹各層次中分群的依據，期望能以對應路由字首分佈特性來最佳化的分群設計，獲得更有效率的記憶體使用。圖 0 為一般 Routing Table 中普遍的路由字首分佈狀態統計圖，從圖上可見大多數的路由字首長度集中在 16 bits 至 24 bits 之間，因此若是在此長度之間有一個階層分界展開點，便可以在此階層的展開中完成大多數的路由選徑查詢動作（如果實際的 IP 封包標頭的目的 IP 位址不特別集中在長度超過 24 bits 的路由字首的話)。而根據之前分群階層數考量與決策過程的經驗來看，如果在 1 至 15 bit 的字首長度之間定有分層展開點，則應該可以有較小的記憶體容量需求，然而卻也會因而增加平均記憶體存取次數。在同樣權衡記憶體大小與平均存取次數後之後，我們擬定以 16 bits 的路由字首長度做為第一分群階層的展開，期望以此展開長度在第一次的分群展開中便能夠完成大多數的路由選徑查詢，並進一步如圖 0 所示檢驗多種分群組合方式，最後我們擬定用於 Unicast Routing 的 32-bit IPv4 位址而言的最佳 IP 位址階層分群方式為「16-1-7-1-7」，因為此方式可以得到最佳的表現：平均記憶體存取次數較小，以及最小的記憶體需求。

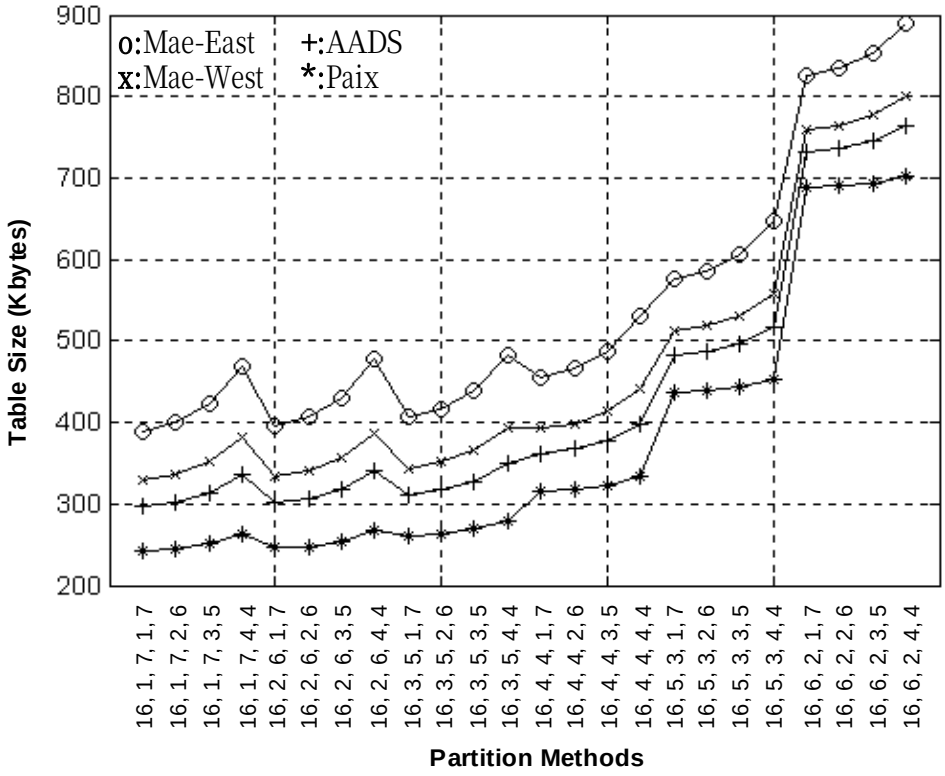


圖 0: 不同的 IP 位址 (32-bit IPv4) 階層分群方式及其所需之記憶體容量
（對不同 ISP/NSP 的 Routing Table 而言）

最後，我們採用一簡單的位元圖(Bit Map) 壓縮法(Compression Bit Map, CBM) 將儲存路由結果的資料結構做壓縮，以進一步減少所需的儲存記憶體空間. 而我們所設計的階層式分群解析法原本在運作架構特性上就相當適合模組化的運作, 可以將每一個階層的分群解析路由查詢動作都視為一獨立而完整的運作模組，因此透過適當的電路規劃與安排，可以達成硬體上超管線式(Pipeline) 平行多工架構的運作方式及其優點，當有 IP 路由選徑查詢進入第二階層後，便可以馬上接受下一 IP 路由查詢的要求，維持高度的路由查詢 Throughput，使每個封包路由選徑動作所要存取記憶體的平均次數減少至極致, 可達到相當於在一次的記憶體存取動作與時間, 便可以完成一筆 IP 路由選徑查詢的動作, 此時 Routing 的速度將幾乎完全取決於記憶體存取定址的硬體定址電路運作速度。

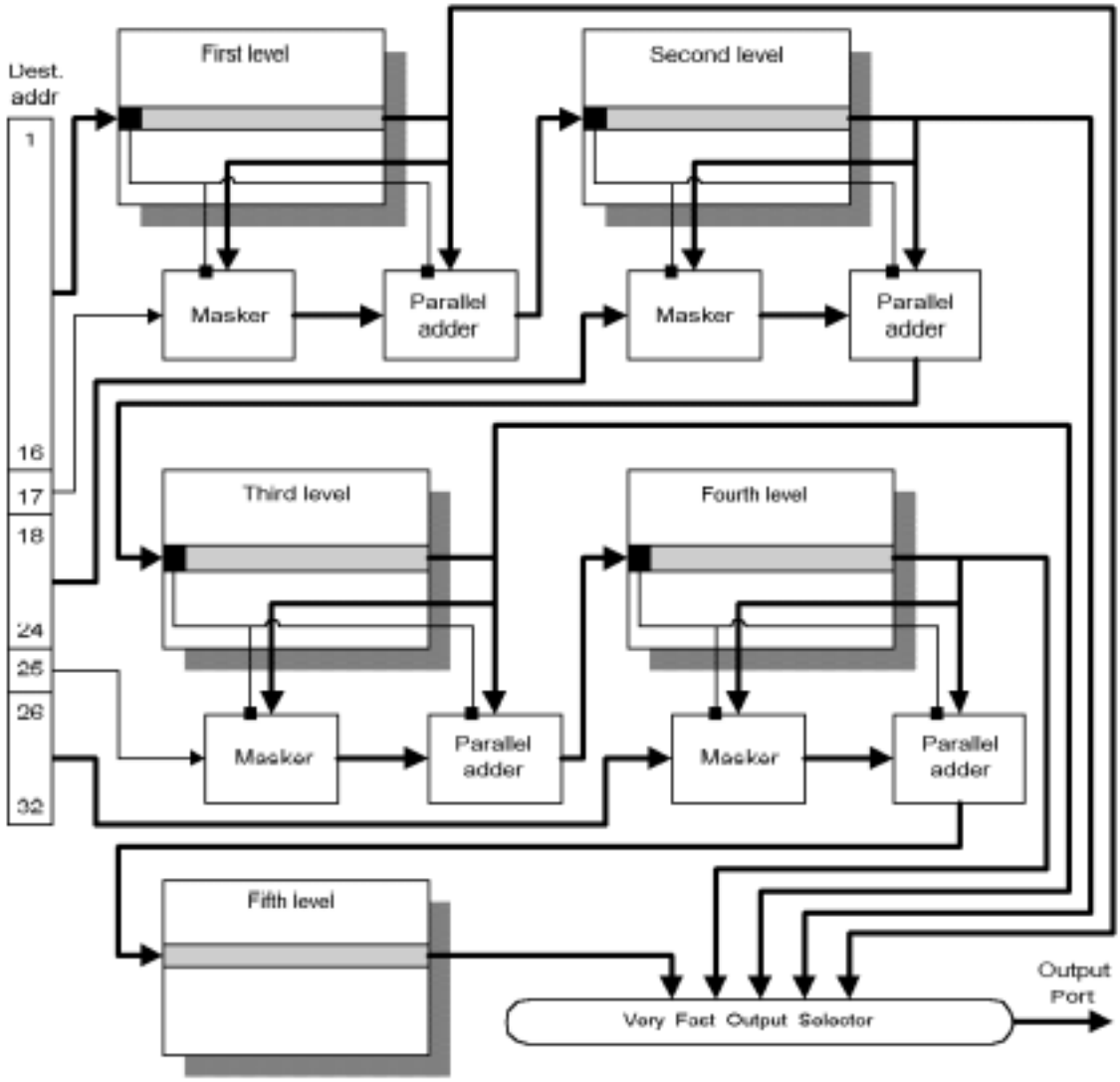


圖 0: 快速路由選徑機制的硬體邏輯架構

圖 0 所示即為我們所設計的快速路由選徑機制的硬體邏輯架構。從表 0 的模擬結果看來，目前我們所設計適合於 IPv4 Unicast 的高速路由選徑方法相較於其他類似概念的方法 [3]，僅需要不到 400 Kbytes 的記憶體空間，而且透過硬體上 Pipeline 平行多工的方式，可達到每個封包在平均一次記憶體定址存取的動作即可以獲得路由選徑結果。未來我們將考量實際商用化系統環境與規格（速度與記憶體容量需求）進行調整與最佳化設計，改進其路由表格資訊更新（Routing Table Update）時所必須對應的相關資料結構與內容的替換或更新，提出更有效率的方式。除此之外也擬配合次世代網際網路的潮流，進一步發展適用於 IPv6 環境的高效能路由選徑技術。

NAP (Number of Route Prefixes)	Enhanced Forwarding Table Size (Kbytes) with (*16, 1, 7, 1, 7*)	Forwarding Table Size (Kbytes) in [3]
Mae-East (57,701)	388	464
Mae-west (34,319)	350	438
AADS (30,705)	337	431
Paix (16,274)	270	357

表 0: 路由選徑方法所需記憶體容量之比較

綜合上述的內容來看，我們在高速路由選徑法之研究設計方面，藉由適當的查詢架構與路由表格結構的設計，快速而硬體架構導向的路由查詢演算法得以實現，配合上實際硬體化操作與超管線(Pipeline) 平行多工運作架構的設計，使得路由選徑查詢得以達到平均約一次記憶體存取動作即可完成的高通透率(Throughput)，可以向上支援至超高速乙太網路(Gigabit Ethernet) 甚或更高速網際網路頻寬的需求。根據階層式分群解析概念，配合壓縮化完全展開樹以及 Compression Bit Map 壓縮法而發展出來的路由選徑方法，確實能使路由選徑機制運作所需的記憶體容量降低許多，相對於平面展開式的壓縮方式，更小的記憶體容量亦更有利於硬體架構的設計與運作，甚至可以將記憶體與周邊相關的邏輯電路整合入同一單晶片中，成為一獨立的硬體路由選徑搜尋引擎，除了可以更大幅提升速度之外，也符合系統化晶片(System on Chip, SoC) 的發展潮流與趨勢。另外，我們也將所設計的方法針對 Routing Table 中，大量的路由字首長度大於 24 的此種特殊路由字首分佈情況做進一步的檢驗與分析，發現其所造成的記憶體容量需求成長的幅度相當低，並未隨之成比例地大量增加，從此結果也可見我們的方法還具有相當的空間擴展性的優勢。藉由對單一路由(Unicast Routing) 之選徑技術研究，我們可以很快進入多點群播路由(Multicast Routing) 及下一代網際網路 IPv6 之相關路由選徑技術的研究，甚至是封包分類器(Packet Classifier) 中複雜度更高、查詢參照資訊（欄位）更多的查表搜尋演算機制的設計。

1-2. 高速單一路由選徑(Unicast Routing) 機制－TCAM-based 架構

除了上述根據階層式分群解析概念，配合壓縮化完全展開樹而發展出來的硬體架構導向的路由選徑方法之外，另一類硬體化操作的路由選徑機制，即是著眼於因半導體製程技術的進步而開發出來的 TCAM (Ternary Content Addressable Memory) 記憶體特性，而採用其為基礎的路由選徑方法，可稱之為「直接 TCAM 路由比對方法 (Direct TCAM Match)」。因此在本計畫中，我們也提出一套採用 TCAM 記憶體特性的硬體化操作路由選徑機制。相較而言，如上一節中採取階層式分群解析概念，或以完全展開樹為基礎的這一類方法 [2, 3, 6]，相當於是將原本路由選徑所需要的查詢演算方式進行轉換，成為一些單純、規律的邏輯電路運作和記憶體存取動作的操作程序，之後只要將 IP 封包標頭的 Destination IP 位址輸入，透過此規則化的操作程序即可以獲得路由選徑的結果。由於並不是直接透過一般 IP 路由選徑搜尋演算方法的過程，而是間接地透過轉換後的規則化操作程序即可獲得等效於路由選徑搜尋演算法的結果，因此也被歸類稱為「間接路由查詢方法(Indirect Lookup)」。

隨著半導體技術的進步，記憶體的種類與功能也不斷推陳出新，除了運作速度與記憶容量之單位面積密度的提升與價格的下降之外，也從最初單純的資料儲存目的發展至以專屬或特殊應用為主的功能性記憶體，「內容定址記憶體（Content Addressable Memory, CAM）」即是做為「資料搜尋」用途的專門記憶體。其應用的方式為：記憶體中的每一個儲存單位存入的為某一應用的一筆候選內容資料以及該內容的相關聯數據資料，待該應用需要進行內容搜尋的時候，僅需將該內容輸入記憶體做為定址用途，CAM 記憶體自動會將內容比對吻合的該儲存單位的內容相關聯數據資料輸出。省去在傳統記憶體架構的操作程序中，必須自行將記憶體中的候選內容資料一一定址取出並分別比對，再將內容吻合項目的相關聯數據資料另外定址讀出所必須花費的時間。而具三元資料比對能力的 TCAM (Ternary CAM) 記憶體的提出，更是讓 CAM 記憶體的資料搜尋能力因為具備更彈性的應用方式而進一步提升：儲存的內容可以包含 don't care 萬用字元(＊)，因此可做到多對一的模糊化搜尋方式，也就是一筆候選內容可以包含多個可能性，輸入資料時不再需要完全吻合候選內容才會得到輸出，只要與候選內容近似，在其包含的可能性範圍內，就可以得到相對應的關聯數據輸出。這樣的資料搜尋應用方式與 IP 路由選徑有著相近似的運作方式（Routing Table 中每一筆路由字首資料皆可視為是以萬用字元對應至一 IP 位址區段，也就是多個 IP 位址，而同一 IP 位址可能會被多個路由字首所對應的 IP 位址區段範圍所涵蓋。當欲查詢某一 IP 位址的路由結果時，根據 Longest Prefix Match 的原則，即是將包含到此待查 IP 位址的所有路由字首都搜尋出來並比較其字首長度，而以最長字首所對應的路由結果做為此待查 IP 位址的路由選徑查詢結果），因而使得 TCAM 也開始被考慮應用在 IP 路由選徑的機制中。

以 TCAM 為基礎的 IP 路由選徑方法即是一個硬體化操作的路由選徑方式，具備有多方面的優點：由於屬於硬體的運作架構，所以運作的速度相當快；也因為其本身的操作特性即相當符合 IP 路由選徑搜尋演算模式，所以僅需要搭配相當簡單的周邊邏輯電路便能夠進行路由選徑查詢的應用，而運作所需要的資料結構的建立與更新速度也很簡單、迅速，可以直接使用 Routing Table 的路由字首資料而不需做任何轉換或處理的動作，所需要額外的記憶體容量也很小。然而其目前唯一、也屬重大的缺點是，價格仍然過高，使得其雖然具備執行路由選徑機制最佳且優秀的能力條件，但是真正商用化的路由器仍未見有採用其做為路由選徑機制的應用。在本研究中，我們提出了一個整合直接 TCAM 路由比對和間接路由查詢方法的路由選徑演算法，而設計的主要動機與概念便是：充分利用 TCAM 的特性並兼顧其價格缺點，整合間接路由查詢方法，並採用兩者平行處理、分工合作的概念，截長補短—以間接路由查詢方法彌補 TCAM 因價格高而數量不足，無法完全負擔路由選徑搜尋應用需求的缺點；利用少量的 TCAM 搭配 Priority 處理邏輯單元，來負擔部分（路由字首長度較長的）路由資訊的路由選徑應用，減少間接路由查詢方法所需負責的路由資訊數量，因而降低其記憶體需求。而此新的複合式路由選徑方法，最多只需要 2 次的記憶體查詢時間，便可以得到路由選徑的結果，而且也一併減少路由搜尋資料表的更新時間。其較細部的設計與運作程序如下面的內容所述。

令 l_i 和 h_i 表示路由器 (Router) 路由表格中第 p_i 筆路由字首的長度及其對應的路由器輸出埠。我們所設計採用 TCAM 記憶體特性的硬體化操作路由選徑方法的架構圖如圖 0 所示：上半部是為一個既有的間接路由查詢方法，處理 l_i 小於或等於 24 的路由字首 p_i，下半部即是直接 TCAM 路由比對方法，處理 l_i 大於 24 的路由字首 p_i。此兩部份在實際的操作中是為平行處理的運作方式，對於一筆輸入欲進行路由選徑查詢的 IP 位址，會同時被輸入至兩部份。若只有上半部分的 Indirect Lookup 有輸出，則 Selector 單元會將此結果直接做為自己的輸出，成為該待查 IP 路由選徑的最終結果；若上下兩部份都分別得到路由選徑的結果，則 Selector 單元將會因為 Longest Prefix Match 的路由選徑原則，而以下半部分 Direct TCAM Lookup 的結果做為自己的輸出，同時也表示是該待查 IP 路由選徑的最終結果。上半部的 Segment Table 是存放以 IP 位址的前 16 位元為第一分群解析階層展開後的每一分群 IP Segment 所對應的路由結果 Pointer；而邏輯處理單元(Logic Process Unit) 則是根據 Segment Table 所輸出的路由結果 Pointer，進一步指向 ADH (Associated Default Hop) 取得對應的路由選徑結果（也就是路由輸出埠），或是指向儲存著下一層次分群解析的路由選徑結果的 NHA (Next Hop Array)。而下半部的 1st、2nd、3rd 和 4th TCAM 表示 4 群 TCAMs 硬體單元，分別儲存並處理 Routing Table 中字首長度為 25 至 26 bits、27 至 28 bits、29 至 30 bits 和 31 至 32 bits 的路由字首，並依序具有由小至大的輸出優先權；Priority Resolve Unit 單元則如同一個 Filter，依 1st、2nd、3rd 和 4th TCAM 的實際輸出情形，選擇當中具最高優先權的 TCAM「有效」輸出做為自己的輸出，若四個

TCAM 皆無有效輸出，則 Priority Resolve Unit 便以一個預設輸出代替；Associated Memory 存放四個 TCAM 中所有（長度大於 24 bits 的）路由字首所對應的路由選徑結果 (也就是路由輸出埠)，並會根據 Priority Resolve Unit 的結果，輸出所對應的路由選徑結果；若 Priority Resolve Unit 的輸出是為預設輸出，則 Associated Memory 將不會有輸出。

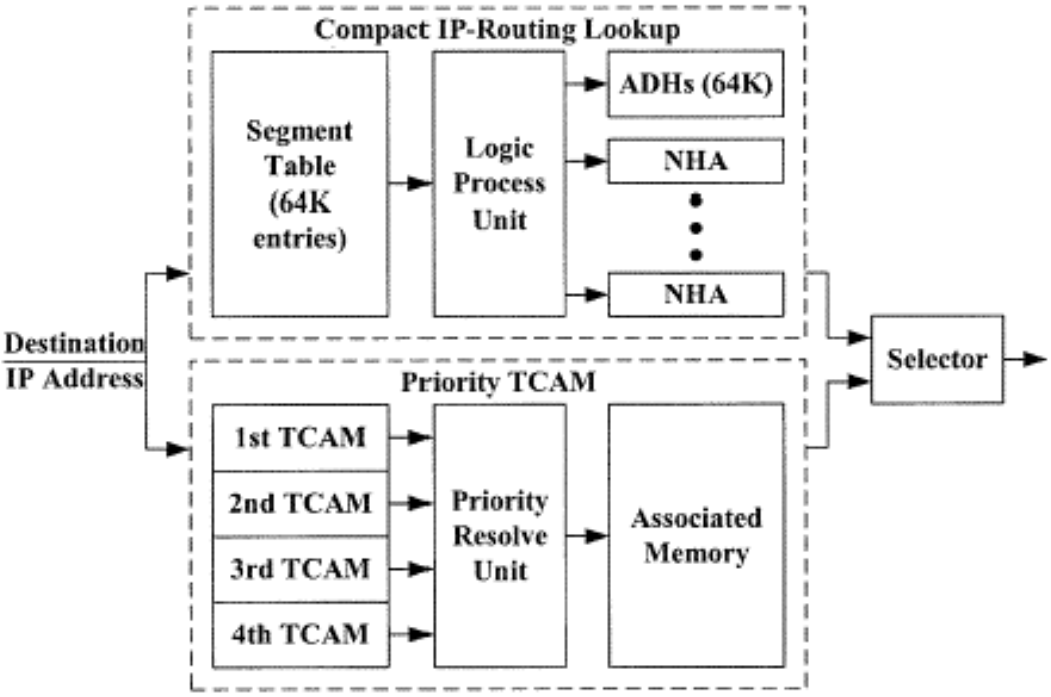


圖 0: Priority TCAM IP 路由選徑機制的功能架構圖

令 p_i (x, y)表示 pi 的 x 位元至 y 位元。表 0 顯示 IP segment 192.168 的對應路由表。因為 192.168 IP segment 的最大長度 l_i 為 32 bits。因此若利用傳統上單純的間接路由查詢方式 Huang's scheme [3] 來查詢的話，此 192.168 IP segment 需要 $2^{(32-16)}$ 個對應的 NHA 路由器輸出埠 entry。利用我們提出的複合式路由選徑查詢方法（如表 0 的最後一個欄位所示）：我們把 l_i 大於 24 bits 的路由字首 p_i, 用直接 TCAM 路由比對方法來處理；把 l_i 小於或等於 24 bits 的路由字首 p_i,才由間接路由查詢方法來處理。因此在間接路由查詢方法這部分只需要 $2^{(23-16)}$個（因為把 l_i 大於或等於 24 bits 的路由字首 p_i 排除，最大的 l_i 為 23 bits）對應的 NHA 路由器輸出埠 entry，大幅降低 NHA 路由器輸出埠 entry 的記憶體需求，同時也把最長的 IP 路由選徑查詢的時間降低在 2 次記憶體存取的時間內。

Entry i	Routing Entry P_i / l_1 / h_i	$p_i(x, y)$ x=17, y=24	Compact IP-Routing Lookup or Priority TCAM
0	192.168 / 16 / 0	xxxx xxxxb	Compact IP-Routing Lookup
1	192.168.20 / 22 / 1	0001 01xxb	Compact IP-Routing Lookup
2	192.168.84 / 22 / 2	0101 01xxb	Compact IP-Routing Lookup
3	192.168.68 / 23 / 3	0100 010xb	Compact IP-Routing Lookup
4	192.168.68.16 / 28 / 4	0100 0100b	Priority TCAM
5	192.168.68.16 / 32 / 5	0100 0100b	Priority TCAM

表 0: Routing prefixes of the 192.168 segment

不僅如此，我們還進一步提出一個「同值位元整合壓縮法(Common Bit Integration)」，來改善間接路由查詢方法的（NHA）記憶體需求。以表 0 中 entry 1 至 entry 3（l_i 小於或等 24 bits 的路由字首）為例，其第 17、19、21 和 22 位元是相同的，因此我們只需紀錄第 18、20 和 23 位元的 3-bit 的 pattern 即可，因此對應的 NHA 路由器輸出埠 entry 可進一步由原來的 $2^{(23-16)}$減少至 $2^{(23-16-4)} = 2^3$，更進一步減低所需要的記憶體大小。

圖 0 表示 TCAM 硬體單元的架構圖。TCAM 硬體單元由路由字首 p_i 暫存器，對應的 p_i mask bit pattern、32 個 3 位元比較器，和 1 個 32 位元的 AND 邏輯運算單元所構成。由於 TCAM 硬體單元完全是由硬體構成，因此以其為基礎的路由選徑方法比起同為硬體操作架構的間接路由查詢方法而言，仍是具有較快的運作速度。

以網際網路上運作的實際 IP 路由表為例，表 0 列出三個大型 ISP 處的實際 Routing Table 的資料，以及兩種具代表性的間接路由查詢方法（Huang's scheme 及 Chen's scheme）與我們提出的複合式路由選徑查詢方法對記憶體需求的比較。由表中可見，增加少量的 TCAM 硬體單元，所需的記憶體可大幅降低，不僅使 IP 路由輸出埠的查詢時間降低在 2 次的記憶體讀取的時間內，亦可降低更新搜尋的資料表的時間。

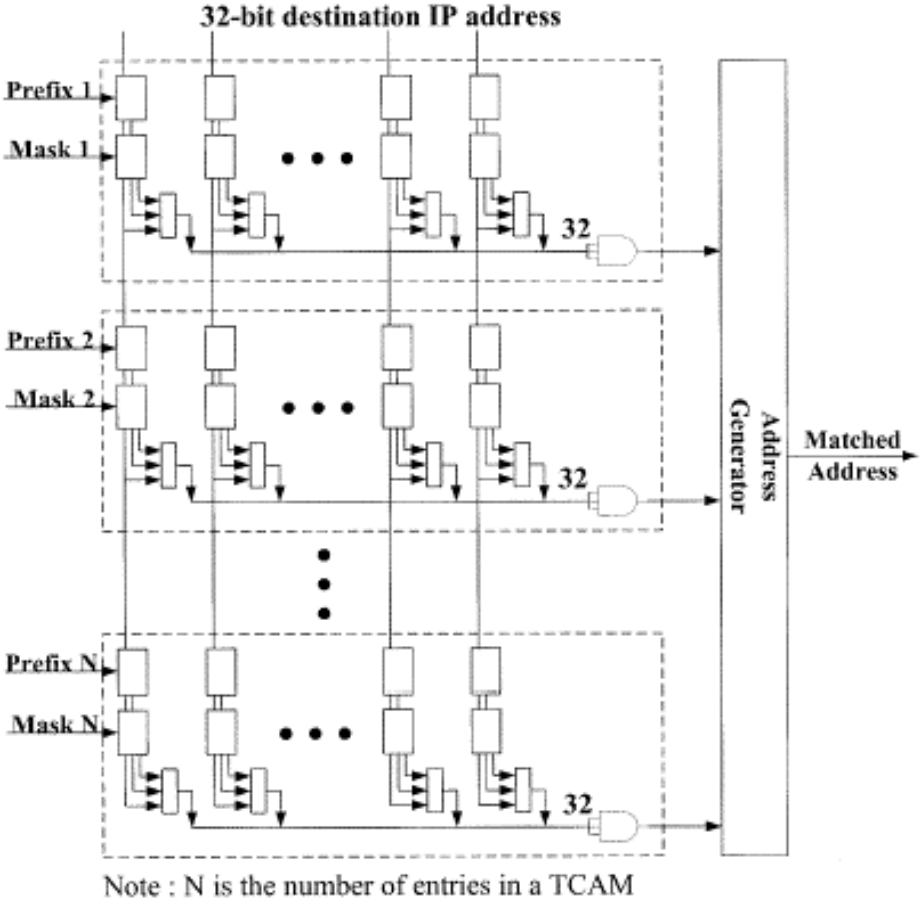


圖 0: TCAM 硬體單元的功能架構圖

	AADS	Mae-West	PAIX
Prefixes	33,931	37,523	18,569
Segments	5,813	6,126	3,571
Length＞24	431	433	443
Huang's [3]	599K	610K	507K
Chen's [6]	659K	781K	543K
Priority TCAM IP-Routing Lookup	423K +431 TCAM	463K +433 TCAM	377K +443 TCAM

表 0: 各種路由選徑方法所需記憶體容量之比較

1-3. 高速群播路由選徑(Multicast Routing) 機制

群播路由選徑(Multicast Routing) 與一般單一路由選徑(Unicast Routing) 的不同點在於：一個 Multicast IP 即代表著一個唯一的 Multicast Group，其包含（對應到）多個分佈在不同區域的主機(Host)，並非如其他（一般的）Unicast IP 多半是以一個連續的區段(Segment) 為單位，來指派、對應到某一區域的單一主機。所以，①在路由表格中，Multicast 的 Routing entry 是以（單一）完整長度的路由字首來呈現，與一般 Unicast 的 Routing entry 是以不完全長度的字首來代表一個連續 IP 位址區段的形式有所不同；②同時，其路由選徑結果也不再只是對應到單一的輸出埠，而可能是一「組」的多個輸出埠；③另外一個與 Unicast Routing 形式上較大的不同是，為了讓 Multicast 的運作更有彈性，可以自由地根據不同的資料發送端來設定不同的資料群播遞送的方式，Multicast Routing 採取多（雙）欄位資料搜尋比對的路由選徑動作，除了 Multicast 的 Destination IP 之外，也必須同時根據 Source IP 位址來決定路由結果輸出埠為何。因此，Multicast Routing 的關鍵技術是為一個高速的多欄位資料完全相符(Exactly Match) 的比對搜尋機制，並且能同時迅速地解析出所需要進行封包轉送的多個路由輸出埠。從高速單一路由選徑方法的設計經驗得知，內容定址記憶體 CAM 和 TCAM 本身即可視為一個簡單而易於實現的硬體架構的資料比對搜尋裝置，而 CAM 的運作方式即相當於一個資料完全相符(Exactly Match) 的比對搜尋機制，因此相當適合應用在

高速硬體化操作的 Multicast Routing 路由選徑機制的設計中；此外，有鑑於 Compression Bit Map (CBM) 的概念對於單一路由選徑機制運作所需的資料結構記憶體空間上的壓縮有很大的幫助，因此我們同樣將此概念應用在群播路由 (Multicast Routing) 的選徑機制設計上，以有效降低、控制路由表格的大小。綜合上述的考量，我們提出一套以 CAM 為基礎，並採用 CBM 技術為輔的高速硬體化操作的 Multicast Routing 路由選徑方法。此方法的群播路由查詢速度可以相當快，一次群播路由查詢只需要三次的記憶體存取次數(Memory Accesses，再加上管線式(Pipeline) 運作架構的可行性，也同樣能使得平均查詢次數降至一次的記憶體存取次數。其細部的設計與運作如下面的內容所述。

整個群播路由表格的組成基本上包含來源位址(Source Address)、目的位址 (Destination Address) 以及由此兩個位址所決定出唯一的一組輸出埠號碼(Output Port Numbers)。根據群播路由查詢法則，當路由器收到一個封包，發現其目的位址為群播網際網路位址(Multicast IP Address) 後，接著必需檢查其來源位址，確定這是此群播組(Multicast Group) 裡的成員，以及是哪一位成員所發出的封包，由這兩個位址便可決定出一組輸出埠號碼；每一個號碼表示路由器必須要複製一份此封包往這個輸出埠送。基本上，我們以底下所示的方式表示一個群播組和其所包括的成員：

Group G1: Sources H1, H4, H6, H7, H8.
Group G2: Sources H3.
Group G3: Sources H2, H5, H8.

Gx 表示目的位址，Hx 表示來源位址。根據這個法則，我們把路由器上經由路由協定(Routing Protocol) 所獲知的來源位址和目的位址（這裡特別是針對對應至 Multicast Group 的 Multicast IP 位址）作成一個二維空間的棋盤狀對應表，此表格的縱軸表示來源位址，橫軸表示目地位址，而兩軸的交叉點基本上有一個圓圈，黑色圓圈表示這是一個有效的群播組和組員的關係，（群播組(Multicast Group), 組員(Membership Source)），白色圓圈則表示是一個無效的群播組和組員的關係，如圖 0(a) 所示。將這些圖形資訊做一個二進位數字編碼，黑色圓圈為 1，白色圓圈為 0，便可形成如圖 0(b) 的一串二進位數字串列。接著我們將此種 Bit Map 的資訊編碼方式也同樣運用在路由輸出埠組上，我們將每一個有效的（Multicast Group, Membership Source）關係所對應的輸出埠號碼組作以 Next Hop Bit Map Stream 的形式來表示，每一筆 Next Hop Bit Map Stream 的每個位元依序對應到一個輸出埠，而以其位元值表示是否要複製一份封包往這個輸出埠送，位元 1 表示需要，位元 0 表示不需要，如圖 0(c) 表格中每一個 Entry 所示。最後再將所有有效的 (Multicast Group, Membership Source) 所對應到的 Next Hop Bit Map Stream 以一個如圖 0(c) 所示的 Next Hop Bit Map Stream Array 陣列來依序置放，做為查詢之用。

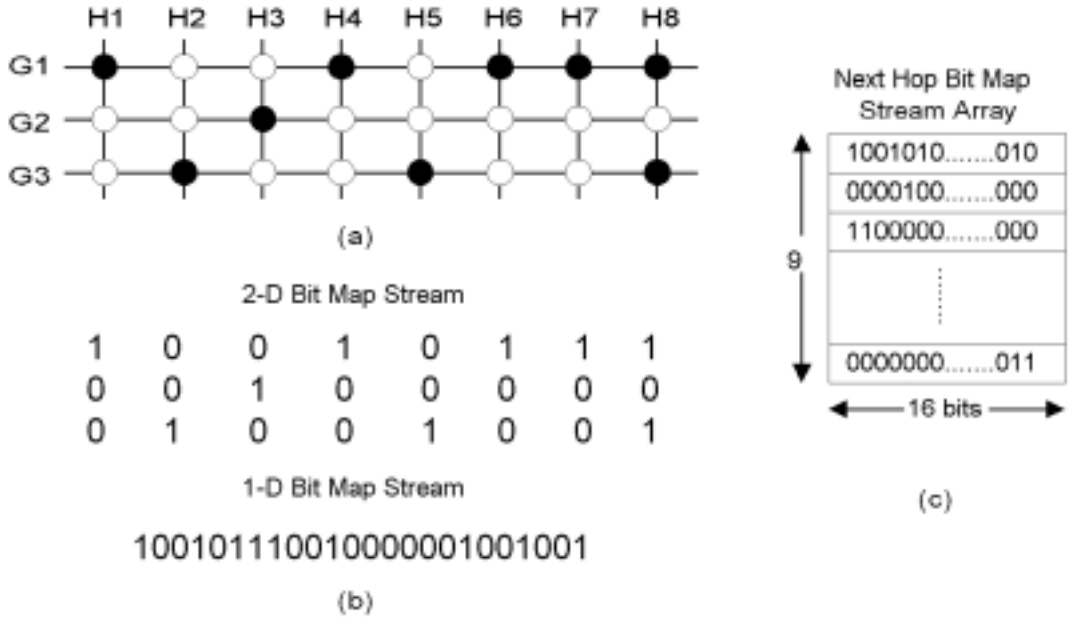


圖 0: (a) The grid to represent the existence information of the (group, source) pairs, (b) The pair bit map streams, (c) The next hop bit map stream array.

進行 Multicast Routing 路由選徑查詢時，先要根據封包標頭的（Source IP, Destination IP）資訊在這個二維空間棋盤狀對應表中尋找其對應的圓圈的幾何位置，接著再從棋盤最左上邊的第一個交會點開始計算到此對應位置前位元值為 1 的個數是多少，此數目即代表從棋盤狀對應表左上角第一個交會點算起有效的（Multicast Group, Membership Source）關係組數，將這個數目對應到存有輸出埠資訊的 Next Hop Bit Map Stream Array 陣列中的位置，如此便可快速取得所需要的路由資訊。為使路由速度更快，在查詢由來源位址和目的位址所組成的幾何位置資訊上，採用「內容定址記憶體 Content Addressable Memory (CAM)」的方式來實作，並將原本圖 0(b) 的二進位數字串列作一個簡單切割以利實作上匯流排的資料寬度需求，形成如圖 0 的系統架構圖；這個架構適用於管線運作方式，因此也可大幅提高整體路由查詢的資料輸出率(Throughput)。

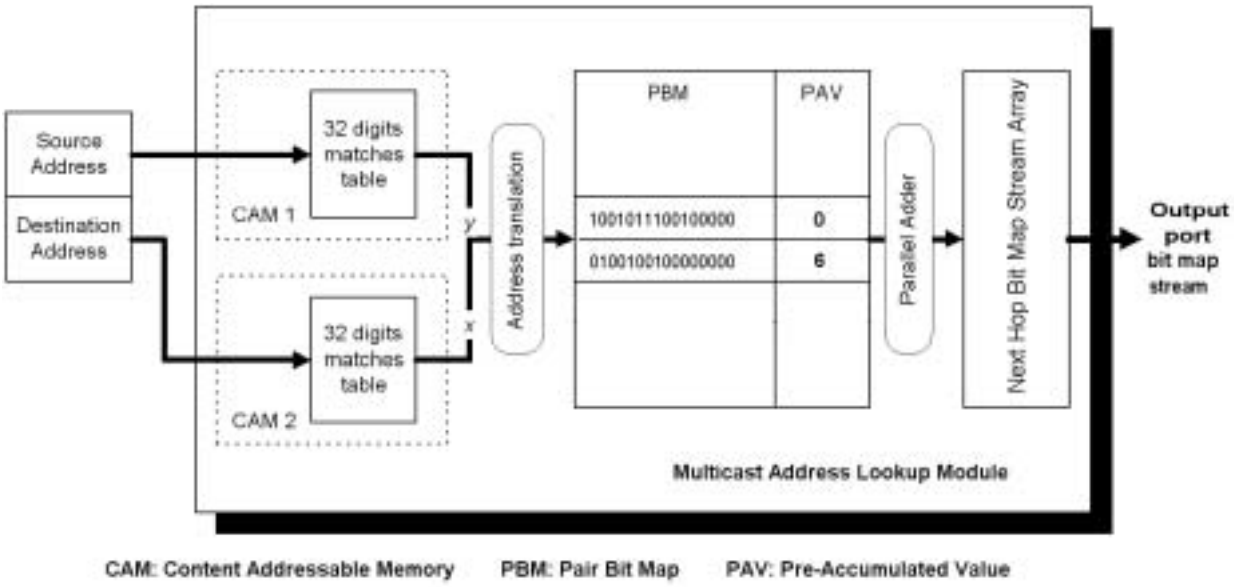


圖 0: The Dual CAM-based Multicast Routing Structure

2. MPLS 網路之 VC-Merge 機制的效能分析

在此，我們主要是設計一套有效的系統效能分析演算方法，針對 VC-merge 的 ATM-LSR 所需的緩衝器和 cell blocking 機率之間的關係進行分析，試著去探討具有 VC-merge 能力的 ATM-LSR 交換機需要比傳統的 ATM 交換機具備多少緩衝器資源，這其中包含了 Frame-level Interleaving 機制所需要的 ATM cell 重組緩衝器以及 ATM-LSR 原本既有的輸出緩衝器(Output Buffer, OB)。

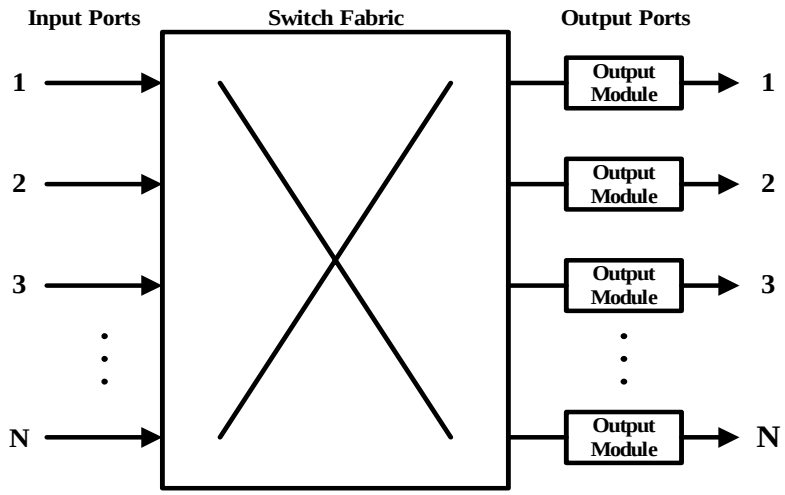


圖 0: Block diagram of a VC-merging capable ATM-LSR

圖 0 顯示一個具有 VC-merge 能力的 ATM-LSR 交換器架構;包含了 N 個輸入/N 個輸出的 ATM cell 交換單元及 N 個輸出模組(Output Module) 單元。圖 0 進一步顯示輸出模組的架構，包含了 M 個重組緩衝器(Reassembly Buffer)、具 VC-merge 功能方塊、S 個依服務等級不同的輸出緩衝器(Output Buffer) 及 ATM cell 服務的輸出排序等。

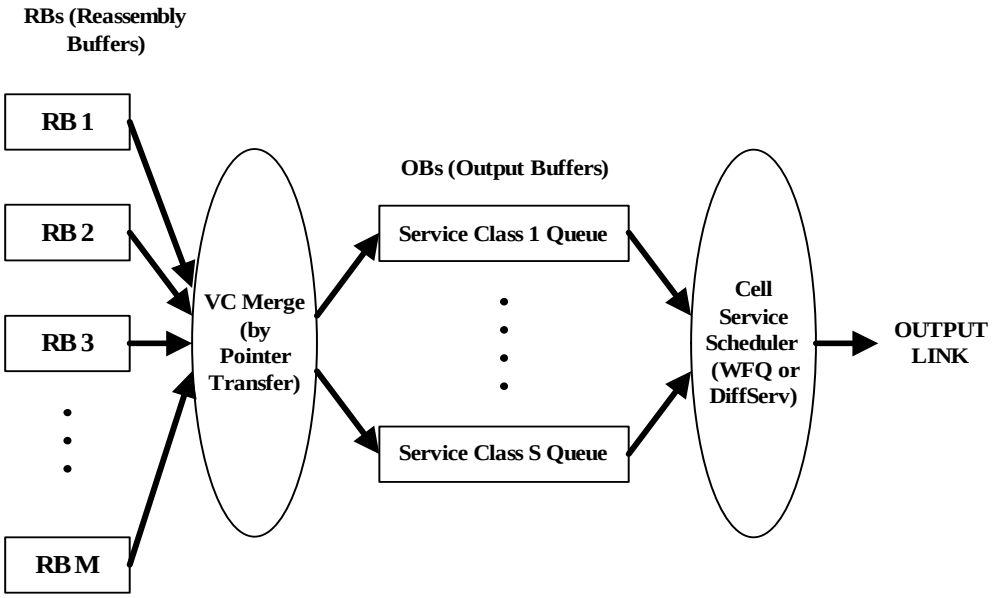


圖 0: Block diagram of the output module

圖 0(d) 為傳統不具 VC-merge 能力的 ATM-LSR 交換機形式，即便是欲前往相同目的網路或節點的 IP Route，其相對應的 ATM cell 在交換後仍是採用不同的 VPI/VCI 以區隔出其為分屬不同 IP Route 的訊務流，如此即便是對應至不同 IP Route 中 IP 封包的 ATM cell 有任意交錯傳輸的現象也沒有關係;圖 0(a)、(b)及(c) 顯示三種不同型態的 VC-merge 形式：(a)是為 Full VC-merge，對於欲傳輸至相同目的網路或節點的 IP Route，將其所屬的 ATM cell 皆轉換成一致的 VPI/VCI 進行傳輸，(b)和(c)則為在進行 VC-merge 時，除了一致的目的網路或節點之外，額外考慮 IP Route 不同的服務品質或其他屬性，將欲前往相同目的網路或節點而且屬性相同的 IP Route，其所屬的 ATM cell 才會在交換後轉換成一致的 VPI/VCI 進行傳輸，此種 VC-merge 方式則稱為 Partial VC-merge。在此種 VC-merge 方式中，VC-merge 後不同 VPI/VCI 的 ATM cell 可以任意交錯傳輸沒關係，但是相同 VPI/VCI 的 ATM cell 仍是必須要遵循 Frame-level interleaving 機制，不可與對應至不同 IP Route 中某一 IP 封包的 ATM cell 交錯。圖 0 中的縮小數字表示起始的 VCI，底線代表 ATM EOM (End of Message) cell。例如 $\underline{5}_2$ 代表 EOM cell，原先 VCI=2 經過 ATM-LSR 交換機後轉成 VCI=5。

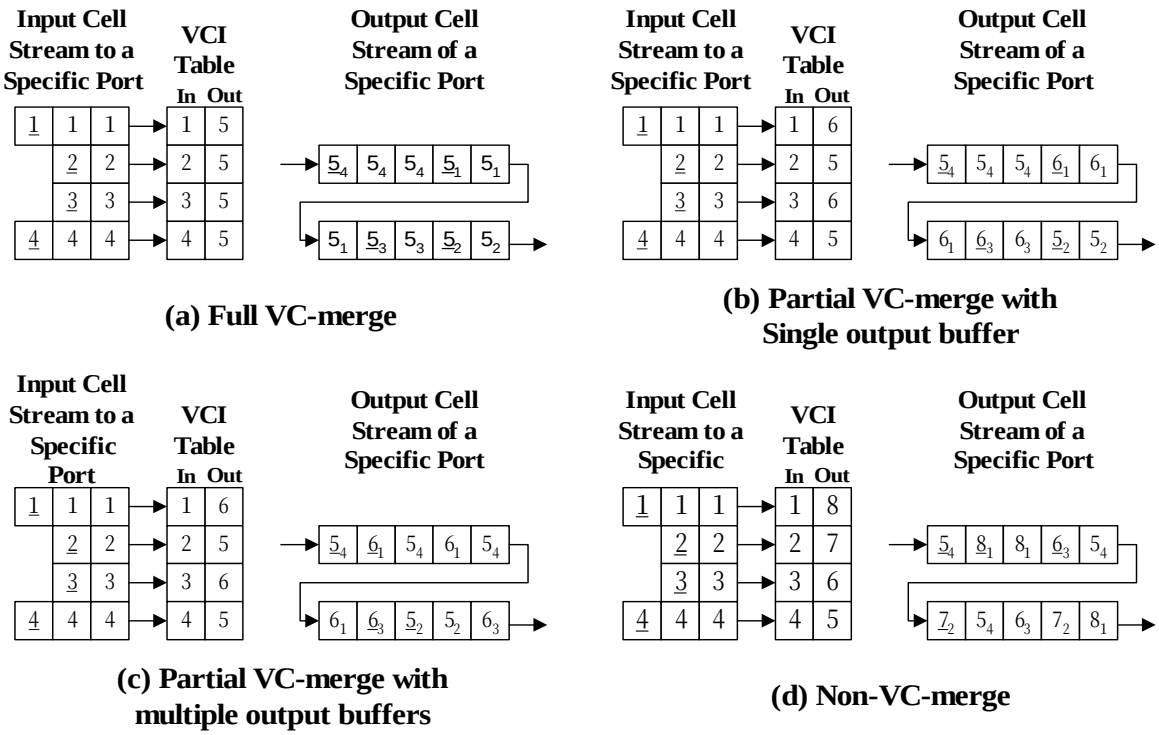


圖 0: Three types of VC-merging and one type of non-VC-merging

我們所設計的分析方法主要是參考論文[7]的方法並加以改進：在我們提出的系統中，放寬[4]的 Input Model，可允許 Cell-interleaving 的 Input Pattern，並且允許 Partial VC-merge 的運作方式。我們建立一個排隊理論數學模式 D-BMAP/D/1 來分析輸出模組的機率分佈。D-BMAP/D/1 的排隊理論模式為：

- M 個 IP Stream 可模擬成 M 個 ON-OFF sources，ON 的時間和 OFF 的時間為幾何分佈(Geometrical Distribution)；
- 在 ON 的時段裡，IP Stream 會產生 1 個 ATM cell 的機率為 r。在 OFF 的時段裡，IP Stream 不會產生任何一個 1 個 ATM cell。參數(1/r)代表 ATM cell interleaving 的程度；
- 輸出服務模式為每個 ATM time slot 處理完 1 個 ATM cell。

研究結果顯示，相較於論文[7] D-BMAP (Discrete – Batch Markovian Arrival Process)/D/1 的排隊理論模式，我們所設計的分析方法其計算複雜度可由 $O(M^4)$ 減少為(M^2)，其中 M 為所模擬的 ON-OFF Source 的個數，並且更進一步多引進了一個參數 r，來描述 IP 封包的 interleaving 程度，使得數學模型更接近實際的情況，所得的分析結果也較為接近實際。複雜度較低的結果，使得對較大的 Buffer Size 分析也能夠得到較高的 Cell Loss Prob.的準確度(尤其 Cell Loss Prob.通常低於 10^{-6} 以下)。我們也利用了 Moment-Generation Function 的理論方法來近似 D-BMAP/D/1 的排隊理論模式，求得一個精確度蠻高的、接近前述所設計的分析方法結果的輸出緩衝器(Output Buffer) 中 ATM cell 數量分佈的數學近似方程式，可更快速計算 Cell Loss Prob.。分析及模擬結果顯示 ATM-LSR 交換機需要具備比傳統 ATM 交換機多 50-70%的緩衝器資源來支援 ATM cell 重新組合使用，以及避免 VC-merge 後無法在目的網路或節點處分離出不同 IP Route 訊務的問題。

圖 0 顯示透過數學分析及電腦模擬一個支援 VC-merge 功能的 ATM-LSR 其總緩衝器 overflow 機率分佈的結果。可以發現，我們的數學分析結果和電腦模擬結果相當一致。由圖 0 我們可得到以下結論：

1. 具有 VC-merge 能力的 ATM-LSR 交換機較傳統 ATM 交換機需要較大的緩衝器資源。如圖 0 中顯示在 overflow 機率為 10^{-5} 次方的假設下，傳統 ATM 交換機需要約 390 個 ATM cell 緩衝器;ATM-LSR 交換機需要至少 520 個 ATM cell 緩衝器。
2. 在細胞交錯 cell-interleaving 愈嚴重的情況下(r 參數愈小)，需要更多的重組緩衝器來重組 IP 封包，因此需要更多的緩衝器資源。在一般的 cell-interleaving 程度的訊務下，約需要比傳統 ATM 交換機增加 50%-70%的緩衝器資源。

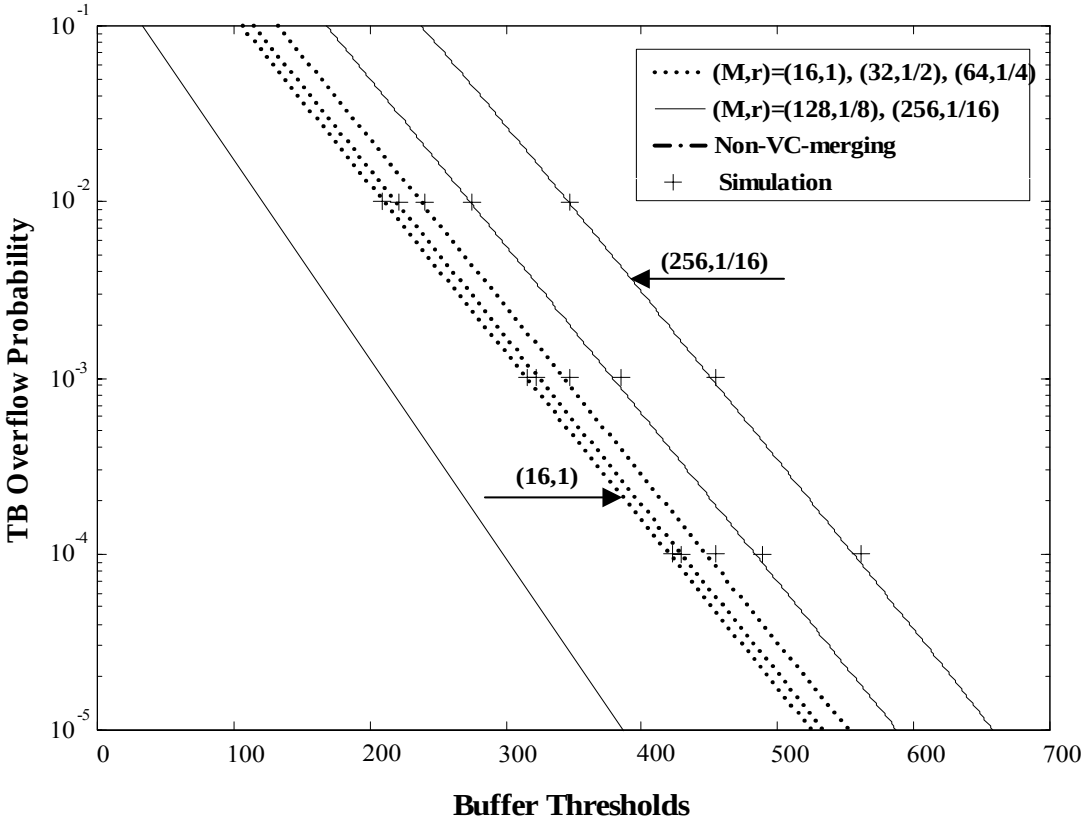


圖 0: Overflow probability of the total buffer versus buffer threshold

3. MPLS 網路之路徑保護(Protection)及快速回復(Path Recovery) 機制

這部分的研究主要在於提出一套有效的 MPLS 網路之路徑保護／回復機制，以便於 MPLS 網路傳輸路徑發生錯誤或損壞時還能夠維持部分基本的通訊，並可以**快速而正確地**地恢復既有的通訊，降低 MPLS 網路上傳輸路徑錯誤或損壞所帶來的影響，減少封包遺失率，並期望能夠進一步達到動態**負載平衡**的附加效益，使系統資源做最佳的利用，如此便可有效的提高網路的資料輸出率(throughput)。

Recovery Model	Backup Path Type	Backup Path Establish Point	Switch -over Processing	Recovery Time	Resource Utilization/ Optimization
Re-Routing	Establish-On-Demand (Simple-Dynamic, Shortest-Dynamic)	After Fault	Local	Slowest	High
	Pre-Qualified			Medium	Medium
Protection Switch	1+1 (Haskin Algorithm, Makam Algorithm)	Before Fault	Head-end	Fast	Lowest
	1:1, 1:n, m:n*			Fast	Low

* m：n = (Backup path number)：(Working path number)

表 0: 兩種 Path Protection/Recovery Model 之比較

一個完整的路徑保護／回復機制應包含有幾個要素：單一或多條的工作路徑(Working path) 與備援路徑(Backup path)、路徑狀態監控／錯誤偵測和警告通知機制、以及路徑回復程序啟動時，切換到 Backup path 的機制。IETF 提出了兩種路徑保護／回復機制的運作模型(Model) [12, 9]，也可做為許多 Path Protection/Recovery 方法的兩大分類依據，其主要是根據 Backup path 建立的時間點做為區隔：如表 0 所示，在路徑發生錯誤之前即預先規劃並建立好 Backup path 的方式稱之為「Protection Switch Model」；相對的，在路徑發生錯誤之後才開始建立適當 Backup path 的方式便稱為「Re-Routing Model」。「Protection Switch Model」由於採取事先規劃並建立好 Backup path 的方式，因此當發生路徑錯誤時可以迅速地切換至 Backup path 以回復正常通訊，然而由於相關的系統頻寬資源也會是在建立 Backup path 時即被保留，等待路徑發生錯誤時可以立即切換專

用，造成系統資源的使用效率最低，因為需要預先保留做為備援專用的相當數量的資源在一般時候是閒置的。也由於必須耗費較多系統資源，因此 Backup path 多半僅在靠近通訊兩方的頭端(Head-end) 網路節點建立。「Re-Routing Model」的傳統方式（如 Simple-Dynamic 和 Shortest-Dynamic）是當路徑錯誤時才開始搜尋並建立適當的 Backup path，因此路徑回復（切換至 Backup path）的速度較慢，但也因為不事先保留資源所以可以讓系統資源做最有效的利用，而且可以在距離路徑錯誤當地(Local) 的最近或較近的節點（LSR）即進行路徑回復，不需要再回到頭端節點才來進行，對路徑錯誤的反應速度較快且靈活度佳。其在路徑錯誤的 Local 節點建立至通訊終端的 Backup path 的方式，也對於路徑回復後因路徑錯誤影響而需要重送的資料量減少。而分類上仍歸屬於「Re-Routing Model」的 Pre-Qualified 方式則其實是介於兩類方法間的機制，兼具兩者的優點，其關鍵在於：以 Re-Routing Model 為主，但是引用 Protection Switch Model 預先建立 Backup path 的觀念，在路徑錯誤發生之前先搜尋、規劃好可作為備援 Backup path 的數個候選路徑，一旦路徑錯誤發生時即能夠透過簡單的資源保留與設定的動作，便可迅速地建立 Backup path 並恢復原有的通訊。由於並未在事先建立 Backup path 並保留資源，僅是預先搜尋、規劃候選 Backup path 而已，因此資源仍可做充分的利用而不浪費，並且也可以預先在 Working path 沿線的各個 LSR 節點上分別搜尋、規劃至通訊終端節點的 Backup path，保有 Re-Routing Model 中在路徑錯誤的最近幾個 LSR 處即可以快速反應的優點。如此再配合上前述當錯誤發生時可節省搜尋時間並快速建立 Backup path 的優點，更加使 Path Recovery 的速度獲得更大幅度的提升。

著眼於兩種 Path Protection/Recovery Model 各有其優缺點，因此我們嘗試結合兩邊的優點，以屬性介於兩者之間的 Re-Routing Model 作法中 Pre-Qualified 類型的網路路徑錯誤復原演算法為主，提出一套高速而最佳的路徑保護／回復機制的方法。然而在實際的操作上，由於 Pre-Qualified 方法對於 Backup path 的形式與數量並沒有明確的定義或限制，因此我們可以，也有必要根據 Pre-Qualified 方法的原則，設計、定義出一套適當並有效率的實際運作方式。如同之前內容所提，除了在通訊的起迄的兩終端 LSR 節點之間於通訊開始之前預先規劃 Backup path 之外，我們也在 Working path 沿線的各個 LSR 節點上分別搜尋至通訊終端節點的 Backup path，然而此種 Backup path 的搜尋動作並不需要在通訊開始之前就進行完畢，可以在通訊開始之後才進行，以避免過多通訊前的程序延遲了通訊開始的時間，而且也不要求每一個沿線 LSR 節點都必須要找到 Backup path。若發生路徑錯誤時其最近一 LSR 節點無法或尚未搜尋到 Backup path，便可通知其上游(Up-stream) LSR 節點進行 Backup path 的建立與通訊的恢復；若此上游 LSR 節點亦尚未搜尋到適當的 Backup path，便可以再往其上游的 LSR 節點進行通知，直到有上游 LSR 已經找到 Backup path，或是到最上游的通訊終端 LSR 節點處為止（這裡必定有 Backup path，是通訊開始前即預先搜尋、規劃好的）。由於在搜尋到 Backup back 之後，並不需要在路徑錯誤發生前對其所需的資源加以保留，然

而又必須確保當路徑錯誤發生時 Backup path 的「可用性(Availability)」—路徑連線正常且仍有足夠資源可立即建立 Backup path，因此必須要有一套有效的 Backup path 可用性的監控、維護方法。最簡單的方式即是以一套週期性的信令(Signaling) 方式，對於預先選擇、規劃好的 Backup path 的通訊路徑連線狀況以及剩餘資源情形進行持續的監控，一旦可用資源不足建立 Backup path 的需求，或在此路徑上亦同樣發生路徑錯誤，便重新搜尋可用的其他路徑作為 Backup path，以保持 Backup path 在 Working path 路徑錯誤時的可即時使用性。此外也可以進一步對於 Backup path 所需的資源先進行軟性保留(Soft Reservation)，亦即對 Backup path 所需要的網路資源有進行預先保留的動作，然而與 Protection Switch Model 不同的是，此保留的資源在 Backup path 尚未真正建立使用時，仍可允許被其它的通訊利用而不因此造成資源的浪費（特別是在有 traffic priority 的網路中，priority 等級低於目前連線的通訊而言），直有當 Working path 路徑發生錯誤而必須啟用 Backup path 時，便以優先使用權的身份將此預先軟性保留的資源收回，以快速而順利地建立 Backup path。在此方式中，Backup path 狀態監控機制僅需針對其路徑連線狀況進行偵測即可，省去可用資源方面的監控，如此可降低監控機制的複雜度與系統資源負擔(overhead)，也使所選定的 Backup path 穩定度增加，較不容易時常需要進行重新搜尋、更換（路徑），而 Backup path 也僅需要一個簡單的通知性(Notification) 信令經過 Backup path 上的所有 LSR 之後即可以馬上建立使用，有助於備援切換速度的提升。而唯一的代價是，當 Backup path 建立並進行切換時，對於原來位於 Backup path 上其他通訊的影響較大，尤其是 Traffic priority 等級低於此連線的通訊。另外，Backup path 的可用性監控、維護方法還可以根據單一節點所建立的 Backup path 數目發展出多種變形，例如：一開始即選定多個候選 Backup path，並對所有的候選 Backup path 都進行監控，確保有需要時至少有一個 Backup path 能立即使用；或是仍僅對其中一個進行監控，若剩餘資源不足或路徑錯誤再換另一個，如此只是節省重新搜尋的時間而已。一般來說，對於比較重要的 Backup path（例如如通訊開始之前，在兩通訊終端 LSR 節點之間必須預先規劃的 Backup path），可以採用軟性資源保留的方式，以較穩定的 Backup path 條件確保其立即可用性；而其餘的 Backup path 便可以採用最簡單的監控方式，即時可用性雖然會較 Soft Resource Reservation 方式稍差，也會有較多的監控資源負擔(overhead)，但是當路徑錯誤發生時，對於 Backup path 上其他通訊的影響則是最小的。

在確定 Backup path 的實際運作方式之後，我們設計採用網路第三層中之 OSPF 繞徑協定做為搜尋適當 Working path 與 Backup path 的基礎 [15, 16]，並利用 MPLS 中既有之 E-RSVP 通訊協定做為持續偵測 Backup path 線路狀態與資源使用情況的可用性監控機制 [18-20。此外，還針對 Working path 著手，參考 Protection model 中 m:n 方式的觀念，利用「多路徑傳送」方式所帶來空間上的多樣性(Diversity)，來減少當錯誤發生時封包的遺失率，這也就是主動的 Path Protection 的能力，並同時也在此多條 Working path 上進行動態負載平衡(Load

balancing) 機制 [17]，充份利用 Working path 資源，提升網路資料輸出率 (Throughput)。

我們所設計的路徑保護／回復機制的運作步驟摘要如下：首先在通訊之前，先透過第三層網路的 OSPF 繞徑協定，配合我們對於通訊路徑品質（例如：頻寬、傳輸延遲、封包遺失率等）的需求，進行自通訊起始 LSR 節點處至通訊終端 LSR 節點處的 Working path 與 Backup path 的選擇。原則上我們是以 N（$N \geq 1$）條 Working path 與 M（$M \geq 1$）條 Backup path 做為通訊的基礎。接下來，便需要在所選擇的 M 條 Working path 上進行 Load balancing 的規劃：根據各 Working path 的 (Bottleneck) BW 與 Delay 來適當地分配其上傳輸的 Traffic 的比例，以確保能夠獲得最佳的通訊品質與 Throughput。當規劃好 Load balancing 之後，便可以開始進行通訊資料的傳輸動作。而在通訊開始進行的同時，除了通訊起迄兩終端 LSR 節點之間 Backup path 的監控機制之外，位於通訊路徑（Working path）上各 LSR 處至通訊終端 LSR 節點處 Backup path 的搜尋與監控機制也開始運作：原則上我們會在 Working path 沿線的各 LSR 同樣以 OSPF 繞徑協定配合我們對於通訊路徑品質的需求，來找尋一條最佳的 Backup path，若能順利找到此 Backup path，便將該路徑資料存在 LSR 的 Database 之中，接下來則以監控機制確保其可用性，透過現有 MPLS 本身的 E-RSVP Signaling Protocol 以 Monitoring 的方法來代替實際上的資源保留與佔用，根據 RFC2205 的定義， E-RSVP 每隔 30 秒會重新檢查每個 LSR 節點與鏈路(Link) 的 State 狀況 [21, 22]，包含是否有 Node 或是 Link 被增加、移除或發生錯誤，以及其頻寬的使用與相關的 QoS 參數條件是否滿足需求等等。如此便可決定此 Backup path 是否可以繼續保留或需要重新搜尋，如結果為後者，則將再回到以 OSPF 搜尋 Backup path 的步驟。而當 Working path 發生路徑錯誤時，系統即在最近一個且 Database 具有 Backup path 紀錄的上游 LSR 處，進行 Backup path 的建立與訊務的切換。

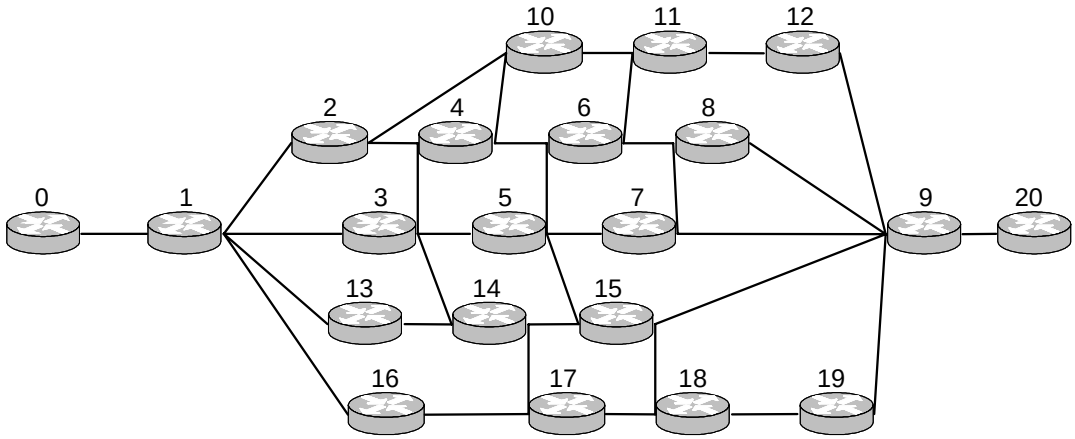
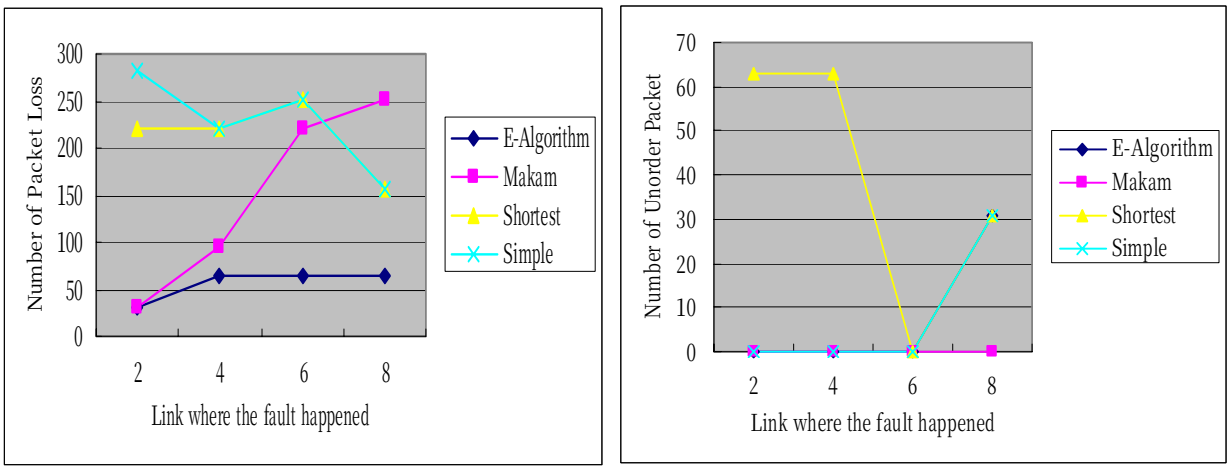


圖 0: Network configuration for NS-2 simulation

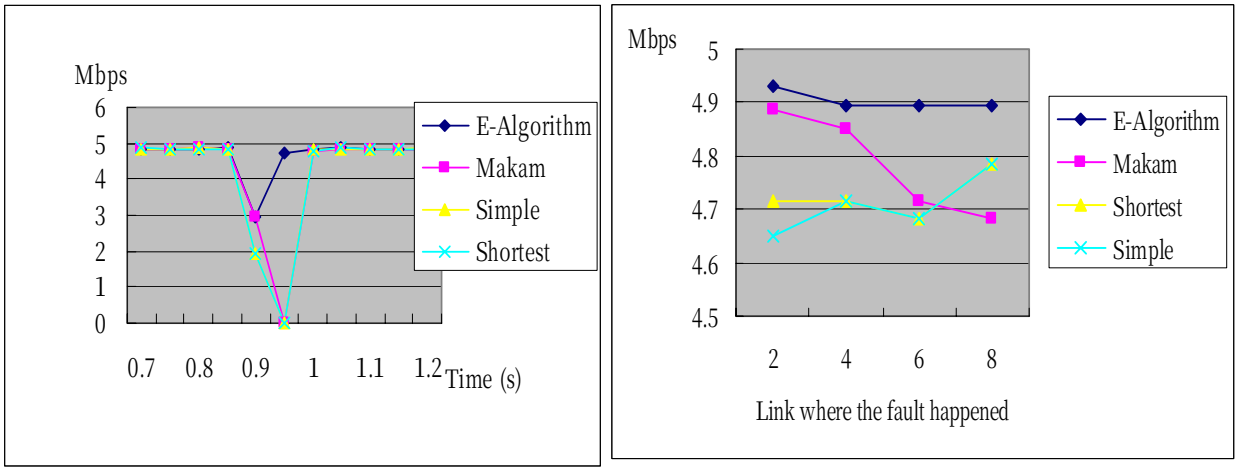
我們採用 NS-2 這一套網路模擬軟體平台，來對所設計的 Path Protection/Recovery 方法進行系統模擬與效能評估、驗證的工作，並與其他方法進行比較。模擬驗證時所採取的網路架構如圖 0 所示：通訊的起迄終端節點分別為 LSR0 與 LSR20，其中每個節點間的鏈路(Link) 頻寬為 10Mbps，並且在通訊的啟始節點 LSR0 處產生 5Mbps 的 CBR/UDP Traffic 進入網路中。圖 0(a) 是從 Working path 上幾處鏈路損壞(Link Failure)所導致的路徑錯誤狀況中，來比較各 Path Protection/Recovery 方法在封包遺失率上的表現。從圖上可以可明顯的看出，我們所設計的方法－E-Algorithm－在 Packet Loss 上的確比其他的方法有較好的結果。因為我們以預先規劃的、離路徑錯誤處最近節點的 Backup path 資料庫的存取來取代重新搜尋 Backup path 的時間，因此 Packet Loss 便少很多；而 Makam 的方法則是必須透過 Signaling 方式告知頭端網路節點 LSR1 網路錯誤的狀況並進行 Backup path 的切換，因此一旦錯誤發生點離入口處的 LSR 愈遠，則通知的時間愈長，丟掉的 Packet 便愈多。Re-Routing Model 中的兩種方法（Simple-Dynamic 和 Shortest-Dynamic）因為是在路徑錯誤之後才搜尋 Backup path，因此 Packet Loss 數量普遍較前兩者為高。Simple-Dynamic 則是因為在 Local 處尋找直接至目的地端的其他路徑做為 Backup path，因此路徑錯誤發生點離目的地端愈近，丟掉的封包數便愈少。在 Shortest-Dynamic 的方面則是因為利用 IP 層繞徑(Routing) 的方式將資料 Bypass 地繞過錯誤發生點至下一個 LSR 繼續傳輸，所以基本上與錯誤發生點沒有太大的關係，但是若路徑錯誤發生點離目的地端愈近，丟掉的封包數也是會減少。在圖 0(b) 我們可以發現，只有 Shortest-Dynamic 方式會因為採用 IP 層繞徑(Routing) Bypass 的方法而容易發生 Packet disordering 的封包亂序情況，而在最後則是因為網路 Topology 的關係才使得多數的方式都容易產生 Packet disordering。



(a) **(b)**

圖 0: Performance comparison on (a) Packet loss and (b) Packet disorder

圖 0(a) 顯示的是當 Working path 上 LSR5 與 LSR7 之間的 Link6 發生 Link Failure 前後，位於通訊終端處所見到的 Throughput 的「(時間) 暫態圖」。由此圖可以看到，Link Failure 發生在 0.8 秒的時候，而我們的機制在 Link Failure 導致的 Throughput 衰減量比起其他的方法都小許多，並不會使 Throughput 降到零，衰減速度也因為 Multipath 多重路徑的特性而較其他方法為緩，在錯誤發生到回復機制的運行過程中，我們依舊可以維持 High Throughput 的狀態，因而能在最短時間內即恢復正常的傳送資料輸出率。反觀其他的方法衰減速度較快，加上又都沒辦法很快完成接替備援的工作，導致 Throghput 皆會在某段時間內降到 0。因此在圖 0(b) 的平均 Throughput 比較圖看來，無論 Link Failure 發生於 Working path 的何處，我們的方法都可以獲得最高的 Throughput 值，而且這些值皆可維持在一個相當好而穩定的水準；相較之下，其他的方法則會因為 Link Failure 發生地點的不同而使 Throughput 的值有所差異。



(a) **(b)**

圖 0: Overflow probability of the total buffer versus buffer threshold

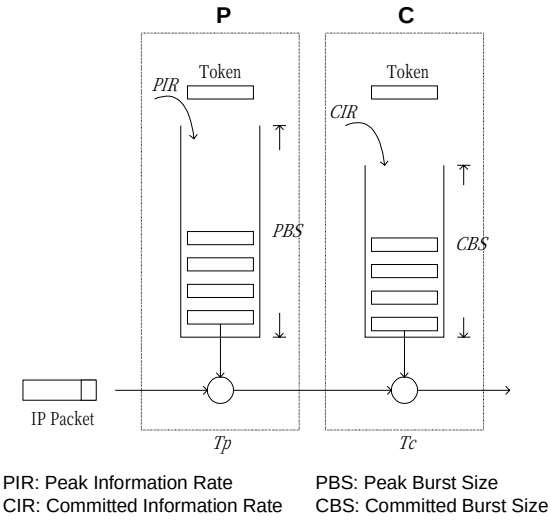
綜合上述的內容來看，我們在 Path Protection/Recovery 機制的設計上，利用每個 LSR 上預先規劃的 Backup path 的 Database 來幫助我們達成快速路徑回復的目的，取代了以往錯誤發生後才開始搜尋 Backup path 或是錯誤發生前先保留資源建立 Backup path 的缺點：藉由 Re-Routing Model 方法的使用，不但排除了 Protection Switch Model 此類作法中預先建立 Backup path 並保留資源而造成資源浪費的缺點，引入 Protection Switch Model 預先規劃 Backup path 的概念，加速了既有 Re-Routing Model 作法的 Recovery time，而且仍保有 Re-Routing Model 中在最近路徑錯誤的節點馬上進行切換的優點，因而能夠快速且有效地降低網路錯誤發生時之影響時間與所需付出的代價。然而此種 Path Protection/Recovery 機制需要一套有效的 Backup path 監控機制來確保 Backup path 可立即使用的特性，因此我們也配合 MPLS 網路既有的 E-RSVP Signaling Protocol 來進行 Backup path 狀態的監控。此外，同時選擇多條 Working path 進行資料傳輸，不僅享有多路徑傳送下 Path protection 的優點，即使有其中一條 Working path 發生問題，也不會在瞬間完全中斷資料的傳輸，維持通訊的可靠性與降低路徑錯誤的影響；同時可藉由適當的 Load balancing 機制的設計，獲得 Throughput 與傳輸品質提升的優點。最後，我們的方法可以讓 ISP 業者很自由而彈性地根據其服務品質的需求與使用者付費原則，評估採用不同複雜度的運作形式，在成本與演算法完整性之間權衡一平衡點。

4. DiffServ 網路中精確的訊務監控調節(Traffic Conditioner)機制

這部分的研究主要在於提出一兼具訊務監控精確度與網路資源使用效率，並進一步達成 micro-flow 微訊務流標記公平性(Fairness) 的訊務監控調節機制(Traffic Conditioner, TC)：在確保訊務特性符合 Traffic Profile 的規範下，精確地讓系統資源獲得充分且最佳的利用，並進一步保障其中各 micro-flow 連線所實際獲得的 QoS，如此才能夠**提升 DiffServ 此網際網路 QoS 架構的可行性與使用效益**，並做為其他 QoS 訊務控制機制（例如：CAC 連線允諾控制機制、Scheduling 排程控制、流量控制、壅塞控制等 Per-Hop Behavior）運作的基礎。IETF 在所建議的 DiffServ 網路 QoS 架構中，也如同 ITU-T 於 ATM 網路定義了 GCRA 的 UPC 訊務監控調節機制的參考模型一般，建議了三種訊務監控調節機制模型，分別是 Single Rate Three Color Marker (SRTCM) [34]、Two Rate Three Color Marker (TRTCM) [35] 與 Time Sliding Window Three Color Marker (TSWTCM) [36]。在初步的研究分析並參考過去我們在 ATM 網路上的發展經驗後，我們選擇以 TRTCM 方式的訊務監控調節機制做為基礎模型，來進行更高效能的 DiffServ 網際網路之訊務監控調節機制的發展與設計。

如圖 0 所示，TRTCM 是藉由兩級的 Token Bucket 造成多階的訊務封包的合法性等級(Conforming Degree)（有別於傳統僅有合法與否的兩級判斷），將進來的每一種服務等級(Service Class)的訊務封包皆根據其運作機制再區分並標記成 DiffServ 所定義的三種等級：綠色(Green)、黃色(Yellow)、紅色(Red)的封包出去，並期望輸出後的訊務特性能夠符合系統對於各服務等級(Service Class) 的訊務所定義的 Traffic Profile 之規範。經過初步的驗證與分析程序後發現，TRTCM 的確是一個相當簡單而又實用的訊務監控調節方法，然而畢竟其只是一個參考模型（作法），所以在繼續深入探究之後發現仍有許多值得改進之處，其中一個重要的缺點為：當 TRTCM 運作在 Color-aware 模式下時，如果實際進入到 TRTCM 的訊務中，黃色封包與綠色封包訊務的總量違反 Traffic Profile 的協定超量使用，而且黃色封包訊務量對綠色封包訊務量的比值超過 Traffic Profile 中的比值時，則會因為共用第一級 Peak Rate Token Bucket 產生交互作用的關係，使得進來的合法綠色封包訊務量會嚴重受到黃色封包訊務量的影響，使得 Output 出去的綠色封包訊務明顯的減少且黃色封包訊務則超過 Traffic Profile 所定義之值，而不符合事先在 TRTCM 所設定的訊務合約。雖然在黃色與綠色封包訊務的總量上符合規範，但是讓許多綠色封包因為在第一級 Token Bucket 未有足夠的 Token 即被判定標記成違法的紅色封包，而將其原本該享有的 Token 轉讓給超量的黃色封包取得而達成黃色與綠色封包訊務總量的合法，實屬不恰當的設計，而且單獨的黃色封包訊務也會明顯違反 Traffic Profile 的規範。著眼於此，為了提供更精準的

監測控制，對於上述的問題，我們基於「保護高 Conforming Degree 的綠色封包訊務不受 Conforming Degree 較低的黃色封包訊務影響」的原則，在原有的 TRTCM 運作機制作了細部的修改及調整，增加了綠色封包訊務的保護機制。綠色封包訊務保護機制的關鍵在於，允許第一級的 Token Bucket 進行「Token 預借」（或稱「Token 超量使用」）的機制：當綠色封包抵達時，若第一級 Token Bucket 中沒有足夠的 Token 數量可使用（←即上述被超量的黃色封包訊務所影響），但在第二級的 Committed Rate Token Bucket 中仍有足夠的 Token，則仍然將該綠色封包輸出為綠色封包，並在第一級的 Bucket 中進行 Token 的預借機制，也就是允許第一級的 Token Bucket 中的 Token 數量呈現負值。如此在上述的超量黃色封包訊務的情形發生時，將可以保障綠色封包訊務不受其影響而取得該享有的 Token，而超量的 Input 黃色封包訊務將會因為第一級 Toekn Bucket 的允許呈現負值狀態，而在一段時間後將其輸出的黃色封包訊務下降至符合 Traffic Profile 所規範的範圍內。由實驗數據中[48] 可得知，修改過後的 TRTCM 比原有的機制可以提供更精確而優良的訊務監測控制：在多種訊務組合情境下都可以維持並確保 Output 的黃色與綠色封包訊務的總量合法，同時也確保 Output 的綠色與黃色封包訊務必定分別符合其在 Traffic Profile 的規範，並保護綠色封包訊務能充分享有其應有的 Token 資源與網路頻寬；除此之外，所 Output 的訊務皆能夠呈現較原始 TRTCM 的設計更逼近理想值的最佳**訊務監控精確度**的狀態。



Input	Bucket Condition/State		Output	Token consumption	
	T_{num_p}	T_{num_c}			
Green	$T_{num\_p} \geq P_{size}$	$T_{num\_c} \geq P_{size}$	Green	T_{num_p}, T_{num_c}	
Green	$T_{num\_p} \geq P_{size}$	$T_{num\_c} < P_{size}$	Yellow	T_{num_p}	Demotion
Green	$T_{num\_p} < P_{size}$	--	Red	(none)	
Yellow	$T_{num\_p} \geq P_{size}$	--	Yellow	T_{num_p}	
Yellow	$T_{num\_p} < P_{size}$	--	Red	(none)	Demotion
Red	--	--	Red	(none)	

圖 0: TRTCM 之系統架構圖與運作邏輯表

除了前述的綠色封包訊務保護機制外，在參考 [37] 此篇論文的內容後，我們也進一步進行「封包 Conforming Degree 升級(Promotion)」與「micro-flow 微訊務流公平性(Fairness)」的可行性研究與方法設計，完成一兼具優良的訊務監控精確度與網路資源使用效率，並可達成 micro-flow 微訊務流公平性(Fairness)的

「增強型 TRTCM 流量監控調節器」。其中在「封包 Conforming Degree 升級(Promotion)」機制的研究方面，最初的設計動機是由於：原始 TRTCM 或 SRTCM 等 DiffServ 所定義的訊務監控調節機制只是以確保 Output 訊務特性符合 Traffic Profile 的規範為最高原則，是故考量的功能僅有當訊務違反 Traffic Profile 時對其封包進行的 Conforming Degree 的降級(Demotion) 標記，然而此單調的功能卻可能導致當一訊務流經過愈多個監控調節器後，會有較高的機率使得愈多的封包因為經過的某些網路資源不足而被降級，如此一但當網路資源恢復或是充裕的時候，仍無法使訊務回復到其原有的特性與資源使用率；而且只要後來若又面臨到網路資源不足的情況，則便可能因為 Conforming Degree 不高而面臨被首先丟棄的情形。因此，我們希望藉由「封包 Conforming Degree 升級」的機制使得當網路剩餘資源夠多的時候，可以將 Yellow Traffic 升級為 Green 等級，甚至是將 Red Traffic 升級為 Yellow Traffic，如此除了可以彌補有些 Green Traffic 通過某些 Congestion 區域而被判定成 Yellow，而造成之後都不再能獲得完全服務品質保障的情形，也可以在系統資源足夠的情況下增加系統資源使用效率（Utilization）。圖 0 所示即是整合增加「綠色封包訊務保護機制」與「封包 Conforming Degree 升級」機制後的 TRTCM 運作流程圖，主要的設計概念是在 Token Bucket 上設定一 Promotion 的 Threshold（如圖 0 上的 τ 所示），一旦 Token 數量超過此門檻值即表示網路資源充足，便可以啟動封包 Conforming Degree 的升級標記動作。由實驗數據中[48] 證實，此增強功能的 TRTCM 的確能夠進一步使網路資源獲得充分的利用，得到最佳的 Traffic Throughput。

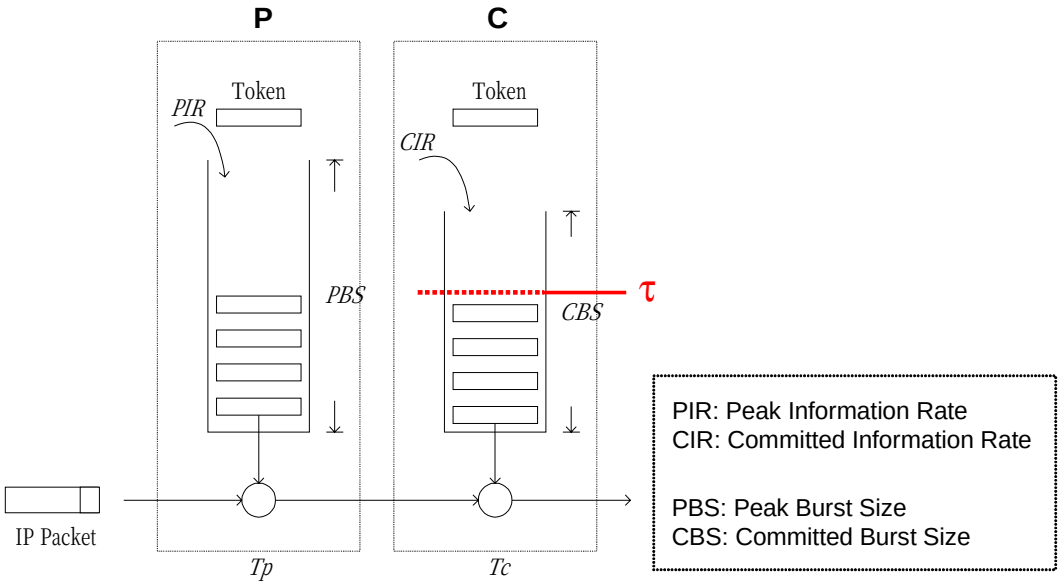


圖 0: 具備 Conforming Degree Promotion 功能之 TRTCM 系統架構圖

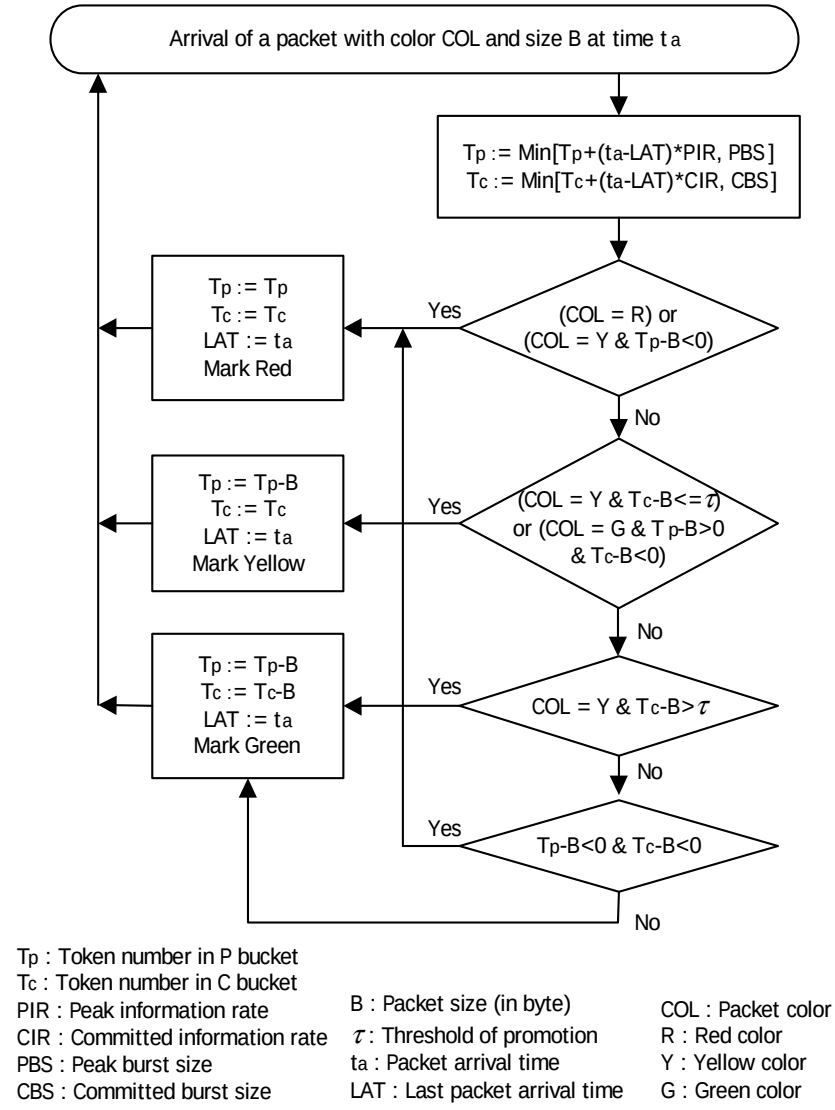
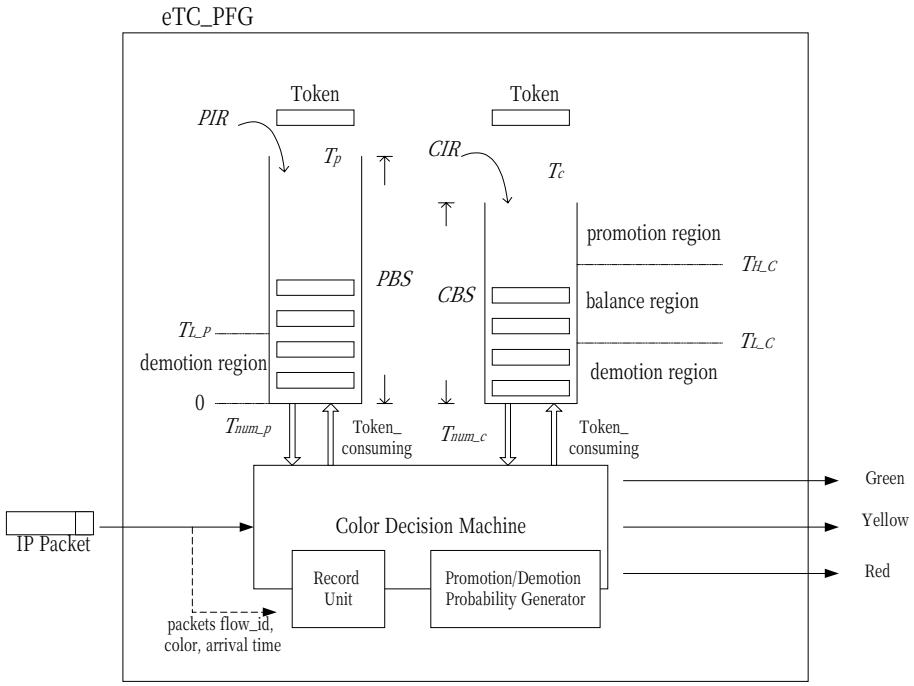


圖 0: 整合 Green Traffic Protection 功能與 Yellow Traffic Promotion 功能之 TRTCM 運作流程圖

除了前述透過新增的「綠色封包訊務保護機制」與「封包 Conforming Degree 升級機制」達成訊務監控精確度與網路資源使用效率的有效改進之外，我們也更進一步在 micro-flow 微訊務流 Conforming Degree 標記公平性(Fairness) 的機制方面進行研究與設計，以期在同一 Service Class 訊務流中的各個 micro-flow 之網路資源使用達到最佳的公平狀態，保障各連線所實際獲得的 QoS，並充分利用系統資源。由於不同的 Traffic Source 特性（例如週期性傳送或是網路擁塞敏感性(Congestion Sensitive)）的交互影響下，使得在同一個服務等級(Service Class) 中的多個微訊務流，卻可能會產生封包的 Conforming Degree 標記不公平的情形，

如此將使屬於同一 Service Class，原本應具備相近 QoS 的各訊務流的實際 QoS 表現差異過大，因而有必要針對 micro-flow 的 Conforming Degree 標記公平性進行控制與改進。實務上欲達成訊務監控調節標記的 Fairness，則是必須在進行封包 Conforming Degree 升級(Promotion) 或降級(Demotion)等標記變更動作的時候, 再根據所欲達成的Fairness目標, 適當地決定是否真正執行該封包Conforming Degree 升級或降級動作來達成。在參考相關論文資料並深入的探討後發現，在運作概念上屬於 Hard-decision 方式的 TRTCM 的 Conforming Degree 標記法並不容易做到標記公平性的功能，因此我們首先針對前述所發展出來的增強型 TRTCM，引進屬於 Soft-decision 方式的「隨機早期偵測(Random Early Detection, RED)」的概念與方法加以改進，在 Bucket 未達到極限狀態（空了 or 滿了）時即提前進行處理，根據目前在 Bucket 中的 Token 數量與其佔有空間的比例，以一機率成分來進行封包 Conforming Degree 降級或升級的標記，而非只是如過去傳統上僅根據目前 Bucket 中的 Token 數量或是與 Threshold 的高低關係, 就進行標記；同時仍維持我們原本設計上於 Color-aware 運作模式下封包 Conforming Degree 升級(Promotion) 與綠色封包訊務流保護機制的功能：當 Bucket Size 低於某一 Lower Threshold 值時，即進入 Demotion-phase 運作模式，也就是訊務監控調節器原始的 Conforming Degree 降級標記功能，並啟動 RED 運作方式計算一 Demotion Prob. Factor, 將進來的 Green 或 Yellow Packet 分別根據此 Demotion 機率值降級成 Yellow 或 Red Packet；當 Bucket Size 高於某一 Higher Threshold 值時，即進入屬於新增功能的 Promotion-phase 並仍採用 RED 運作方式計算一 Promotion Prob. Factor，將進來的 Yellow Packet 升級成 Green Packet。

在引入 RED 概念的運作方式改進後, 便可以容易地進行 Conforming Degree 標記公平性機制的設計工作。我們所設計的 Conforming Degree 標記公平性機制主要的概念為：首先在訊務監控調節器上新增一訊務流量統計的附屬輔助裝置，配合 DiffServ 架構中定義在該訊務監控調節器前端的封包分類器(Packet Classifier) 來識別 Incoming 封包所屬的 micro-flow，持續計算並紀錄該訊務監控調節器所處理的 Service Class 訊務流中，每一 micro-flow 的各種 Conforming Degree 封包的流量統計值；接下來當系統進入 Demotion-phase 或 Promotion-phase，並根據 RED 概念的運作方式計算出一適用於該 Service Class 訊務流整體的 Demotion 或 Promotion 機率值時，並不是如同上述一般將此機率值直接作用在此時抵達的封包上, 而是必須進一步根據前述輔助的流量統計裝置所計算的各 micro-flow 的流量狀況，以及所定義的公平性原則（在此我們以 Max-Min Fairness 為基本的 Fairness 原則），藉由適當的權重(Weighting) 調整該機率值, 使抵達的各 micro-flow 的封包享有不同的升降級機率的手段, 來達成各 micro-flow 的最高 Conforming Degree 的綠色與次 Conforming Degree 的黃色封包訊務流彼此間標記的公平性。圖 0 所示即為所設計的具備 Conforming Degree 標記公平性並整合「綠色封包訊務保護機制」與「封包 Conforming Degree 升級」機制的增強型 TRTCM 的系統架構圖與其運作邏輯表。



Input	Bucket Condition / State		Output	Conditional probability	Token consumption
	T_{num_p}	T_{num_c}			
Green	--	$T_{num\_c} \geq T_{L\_C}$	Green	1	T_{num_p}, T_{num_c}
	$T_{num\_p} \geq T_{L\_P}$	$T_{num\_c} < T_{L\_C}$	Green	$1-P_{demo\_g}(J)$	T_{num_p}, T_{num_c}
			Yellow	$P_{demo\_g}(J)$	T_{num_c}
	$T_{num\_p} < T_{L\_P}$	$T_{num\_c} < T_{L\_C}$	Green	$1-P_{demo\_g}(J)$	T_{num_p} T_{num_c},
			Yellow	$P_{demo\_g}(J)*(1-P_{demo\_y}(J))$	T_{num_c}
			Red	$P_{demo\_g}(J)*P_{demo\_y}(J)$	(none)
Yellow	$T_{num\_p} \geq T_{L\_P}$	$T_{num\_c} \geq T_{H\_C}$	Green	$P_{prom}(J)$	T_{num_p} T_{num_c}
			Yellow	$1-P_{prom}(J)$	T_{num_p}
	$T_{num\_p} \geq T_{L\_P}$	$T_{num\_c} < T_{H\_C}$	Yellow	1	T_{num_p}
	$T_{num\_p} < T_{L\_P}$	$T_{num\_c} \geq T_{H\_C}$	Green	$P_{prom}(J)$	T_{num_p} T_{num_c}
			Yellow	$(1-P_{prom}(J))*(1-P_{demo\_y}(J))$	T_{num_p}
			Red	$(1-P_{prom}(J))*P_{demo\_y}(J)$	(none)
	$T_{num\_p} < T_{L\_P}$	$T_{num\_c} < T_{H\_C}$	Yellow	$1-P_{demo\_y}(J)$	T_{num_p}
			Red	$P_{demo\_y}(J)$	(none)
Red	--	--	Red	1	(none)

圖 0: 增強型 TRTCM 之系統架構圖與運作邏輯表

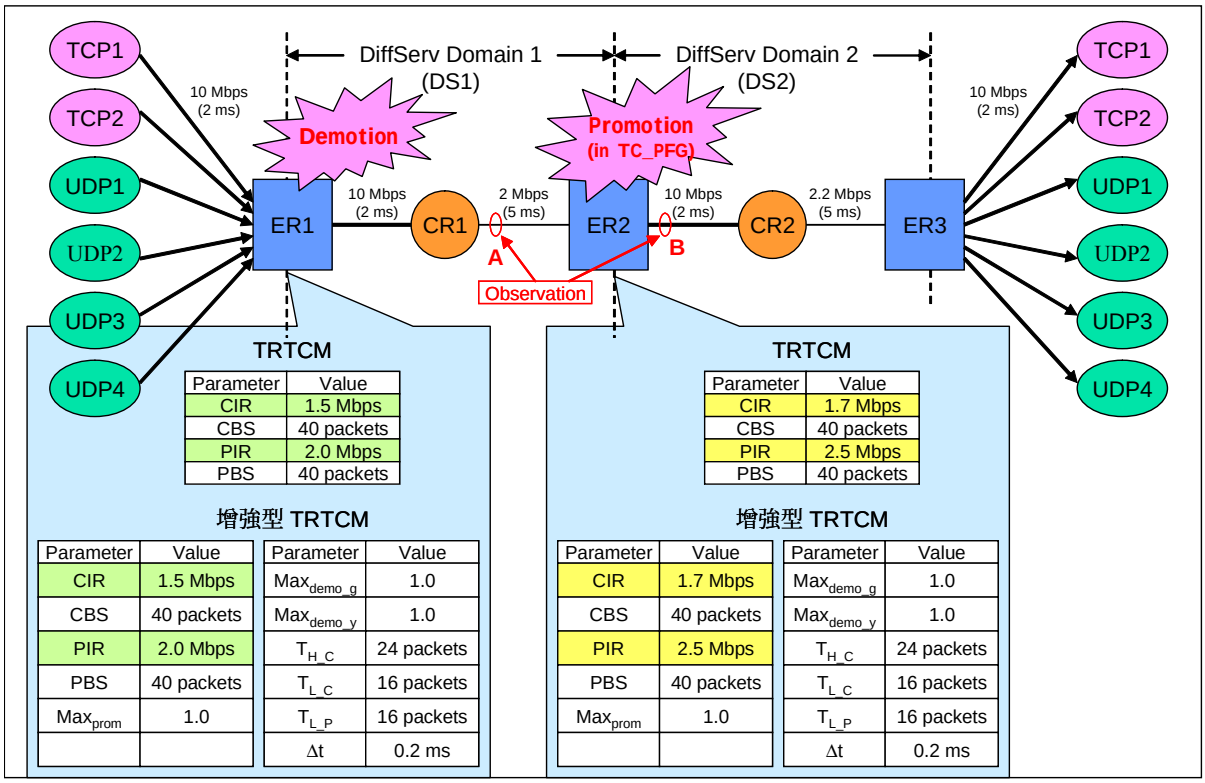
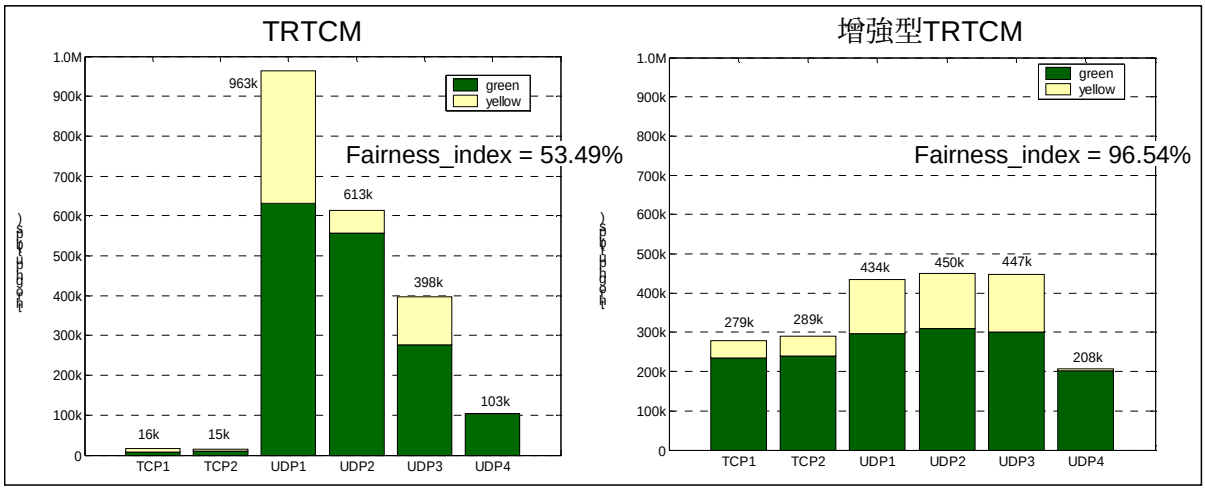
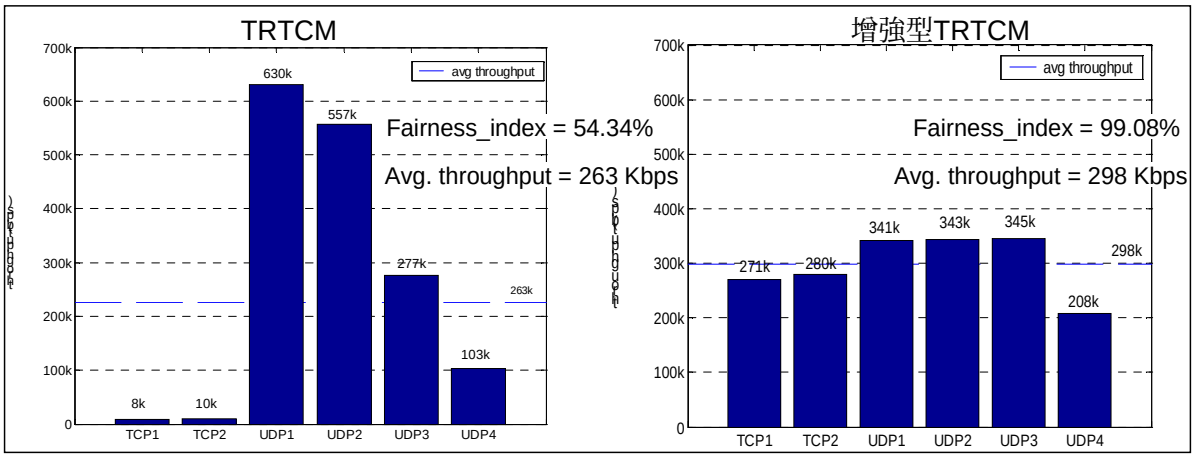


圖 0: 增強型 TRTCM 效能驗證之模擬網路架構與參數設定

圖 0 是為進行模擬驗證時所採用的網路架構與訊務源模型（分別是兩個 TCP 與四個 UDP 連線），並在其中 A、B 兩處分別觀察訊務經過 Demotion 與 Promotion 作用後的狀態。模擬結果如 圖 0 所示，無論是在哪一個觀察點所得到的 micro-flow 訊務流量分佈狀態，都可以很明顯地發現，增強型 TRTCM 在標記公平性方面皆較原始 TRTCM 有著顯著的改善：面對不同特性的 Traffic Source 組合，增強型 TRTCM 的 Output micro-flows 的流量分配較為平均，保障了屬於 Congestion Sentitive 的 TCP 連線的網路資源與 QoS 不易受其他連線的影響，並使原本流量較大的 UDP1 連線不會因而獨佔較多的網路資源，而是與其他連線皆保持相近的標記與資源使用比例，達成標記公平的目標。除此之外，在圖 0(b) 的數據也顯示，增強型 TRTCM 亦能夠由於較公平而精確的監控標記，而獲得較佳的訊務 Throughput 值。



(a)



(b)

圖 0: 增強型 TRTCM 與原始 TRTCM 方法之效能比較：(a) 於模擬網路架構 A 點之訊務流量分佈；(b) 於模擬網路架構 B 點之訊務流量分佈

5. 高速 IP 封包分類(Packet Classification) 機制

藉由發展硬體架構導向／硬體化操作的高速 IP 路由選徑機制的經驗，我們也在原本規劃的計畫項目與目標之外，新增投入高速 IP 封包分類器(Classifier)的研究設計。要達到具備服務品質保證或差異化服務(DiffServ) 品質的網際網路服務，其第一步驟就是要進行訊務流(Traffic Flow) 的區隔(Isolation)，如此才能針對不同的訊務流及其不同的服務品質和資源需求，施以相對應的訊務處理或控制機制，達成其各別差異化的服務品質目標。而根據封包中各個欄位的資訊，對封包做不同服務等級分類的封包分類器即是在本質上並不提供訊務流識別功能

的網際網路上實現訊務流區隔的主要機制。進行封包分類時，會有一封包分類表格(Packet Classification Table) 紀錄封包的分類規則(Packet Classification Rule)或稱為封包過濾法則(Filter)：設定各類別於封包欄位上的資訊條件，以及分類後各類別訊務對應的控制機制或封包處理動作（例如：Dropping or Buffering 等）。而封包分類器即是根據封包各欄位的資訊與表格中的分類規則進行比對，將封包進行區隔，並將此資訊傳遞至後續的封包處理單位，以執行分類規則所指定的對應封包處理動作。因此，Packet Classification 可以視為是一個多欄位的資料比對搜尋機制。而由於 Packet Classification Rule 主要是由網路管理人員根據網路管理策略(Policy) 或其他經營上的目標或安全性考量而設定的，而且可以隨時根據需求進行增修或刪除以配合新的策略或目標，因而往往在多次異動之後，容易產生所設定的新分類規則會和舊有的規則發生衝突或自相矛盾的情形，特別是一個封包分類表中的分類規則數常常可以多達上百條或甚至千條的情況下，就更不容易單純以人力來檢查各分類規則彼此間是否有所衝突。因此一個完整的封包分類解決方案(Solution) 也應該包含封包分類之前，對於封包分類規則衝突與否的檢查與排除的機制，以便使真正執行封包分類時能夠達成預期的分類處理結果。在此，我們提出一個以 TCAM 為基礎的高速硬體化操作的封包分類機制架構，藉由對 TCAM 邏輯電路的修改與重新設計得到一增強功能的 TCAM (Enhanced TCAM, e_TCAM)，除了能以包含萬用字元(Wildcard) 的三元形式 (Ternary State) 儲存搜尋比對用的候選資料，並與輸入的二元資料進行比對搜尋之外，也能夠同樣以三元的形式來輸入待比對資料，並進行我們新定義的「雙三元資料比對邏輯運算(Dual Ternary Data Comparison Logic)」。以此增強型 TCAM 為基礎搭配適當的周邊邏輯電路（例如優先權解析電路），便能夠同時運用於高速封包分類或做為快速的封包分類規則衝突偵測機制。

由發展前述高速硬體架構導向（硬體化操作）的路由選徑方法的設計經驗得知，內容定址記憶體 CAM 和 TCAM 本身即可視為一個簡單而易於實現的硬體架構的資料比對搜尋裝置，其中 CAM 的運作方式相當於一個資料完全相符(Exactly Match) 的比對搜尋器機制，而 TCAM 則可視為是群組資料相符(Group-match) 的比對搜尋器，同時兩者還有一個好處，即是真正用於搜尋比對動作的候選資料和其資料結構與人們邏輯概念上理解的運作形式相同，無須做進一步的轉換，如此有助於方便資料的維護更新以及操作者的理解。因此在分析 Packet Classification 的運作流程，進而瞭解 Packet Classification 的關鍵技術是為一個多欄位的資料部分相符／群組資料相符 (Partially-match/Group-match) 的比對搜尋機制後，我們初步即決定採取以 TCAM 為基礎的方式，來進行高速硬體化操作的封包分類機制的方法與架構設計。而在決定系統基礎關鍵架構之後，接下來便是如何達成封包分類規則衝突偵測與排除的功能。我們發現，所有封包分類規則的衝突狀況大致可以歸納為三類，分別是「無衝突(Conflict free)」、「子集合形式衝突(Subset Conflict)」、以及「範圍重疊衝突(Overlapping Conflict)」。圖圖 0 所示即為一個採取雙欄位(2-field)**且為字首(Prefix)形式**參考資訊的封包分類規

則下的一組範例封包分類表格，以及前述的三種分類規則衝突狀況的示意圖，在此可以以一個二維的平面圖形來表示。

An example of classification table with two-tuple filters

Filter	Field-1	Field-2	Action
R1	01**	1***	10Mbps
R2	1***	1***	10Mbps
R3	101*	110*	Deny
R4	11**	00**	100Mbps
R5	0***	001*	1Mbps
R6	001*	0***	100Mbps

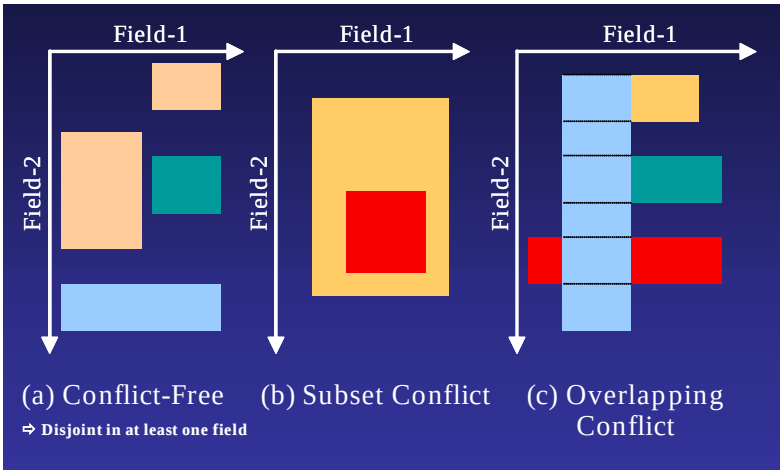


圖 0: 封包分類表格範例與封包分類規則衝突狀態示意圖：
(a)無衝突, (b)子集合形式衝突, (c)範圍重疊衝突

雖然這僅是一個二維的示意圖，不過我們已經可以據此得到一個適用於更多欄位狀況下的重要結論，那就是：一個分類規則會對應到該多維展開空間中的部分範圍，只要在任一維度上兩個規則所對應的空間範圍並沒有任何接觸或重疊，則此兩規則必為無衝突的狀態；反之則必然有分類規則互相衝突或為冗餘的情況發生，而至於究竟是衝突還是冗餘的情形，又或者是哪一種形式的衝突，則必須要再進一步檢驗兩規則在該完全展開空間中每一維度的重疊關係，以及其相對應的封包處理方式（如圖 0 的封包分類表格中 Action 欄位所示）才能夠決定。不過我們已經可以由此得知，欲判斷封包分類規則彼此間的關係（為矛盾衝突或是冗餘），則首先最基本的工作便是要檢驗各分類規則對應的空間範圍是否有重疊的情況發生。

一個直覺、簡單而又快速的偵測的方式是：利用現成既有的以 TCAM 為基礎的封包分類方法，將某一規則所對應空間範圍中的每一個成員(Element) 依序都作為該 TCAM 封包分類器的輸入資料，必且進行分類搜尋比對動作，若是在原來的規則之外還有其他的分類規則被觸發(trigger)，則表示此元素也同時符合另一規則的條件而屬於它，亦即是說，兩分類規則所對應的範圍有重疊的狀況；若是沒有其他的規則被觸發，則表示此規則所對應的空間範圍是獨立的，沒有和其他規則發生重疊的狀況。此檢測方式由於直接採用既有 TCAM 基礎的封包分類方法，因而延續了其硬體化操作的高速運作優點，已經可以滿足快速偵測的目標了。不過由於其必須將每一規則的每一個成員都以此方式輸入操作一遍，才能

夠完成所有規則間的對應空間範圍重疊偵測，在分類規則數量增加、每一規則對應空間範圍變大、或是重疊情形多且範圍大的情況下，仍會大幅降低整體的偵測速度以及偵測的效率。進一步的研究後發現，若是能採取以一個分類規則即為一個 TCAM 基礎的封包分類器的輸入比對單元，而非以該規則的每一個成員為輸入比對單元，則可以大幅降低所有分類規則重疊偵測所需的資料比對次數和時間。不過這牽涉到之前所沒有（定義）過的特殊群組相符比對(Group-match) 機制，也就是輸入的待搜尋比對資料亦可以是一個群組的形式（在表示法上即是包含了萬用字元＊的三元資料形式），因此必須要定義一個新的群組相符比對(Group-match) 的邏輯。圖 0 左上所示，即是我們為採用位元序列(Bit sequence)資訊表示法環境下的此種新群組相符比對機制所設計和定義的「雙三元資料比對邏輯」；而圖 0 的其餘部分則是將圖 0 的範例分類表格中的分類規則，採用新定義的比對邏輯進行運算的範例。我們由此可發現，若分類規則所對應的空間範圍有重疊，則比對的結果必然為「All 1's」的狀態；而且此一新的資料比對邏輯與 TCAM 所使用的（單）三元資料比對邏輯不同的地方僅在於，此新邏輯多定義了萬用字元與萬用字元相比較的成分。

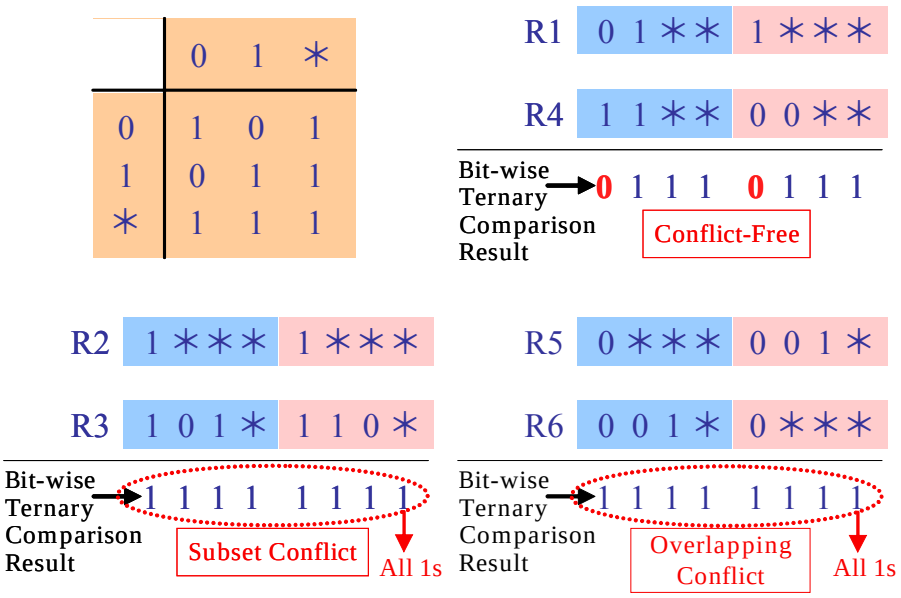


圖 0: 雙三元資料比對邏輯與各狀況下的應用結果

因此我們提出一個新的 TCAM 的邏輯電路，針對傳統的 TCAM 邏輯電路進行修改與重新設計，加入三元資料的輸入能力以及新的「雙三元資料比對邏輯」，並將其命名為「增強型 TCAM (enhanced TCAM, e_TCAM)」。此一新的 TCAM 除了保留了原有 TCAM 的邏輯運算能力（(單）三元資料比對）以及應用範圍（群

組資料相符比對搜尋）之外，還能夠用作為新的群組相符比對搜尋器，進而可以據此實現有效率的分類規則重疊偵測方法。圖 0 即是採用增強型 TCAM 為基礎所設計，兼具封包分類機制與分類規則衝突偵測功能的高速雙欄位(2-field)參考資訊之封包分類平台。將我們所設計的增強型 TCAM 搭配適當的周邊邏輯電路（例如：優先權解析電路），便能夠同時運用於高速封包分類或做為快速的封包分類規則衝突偵測裝置。

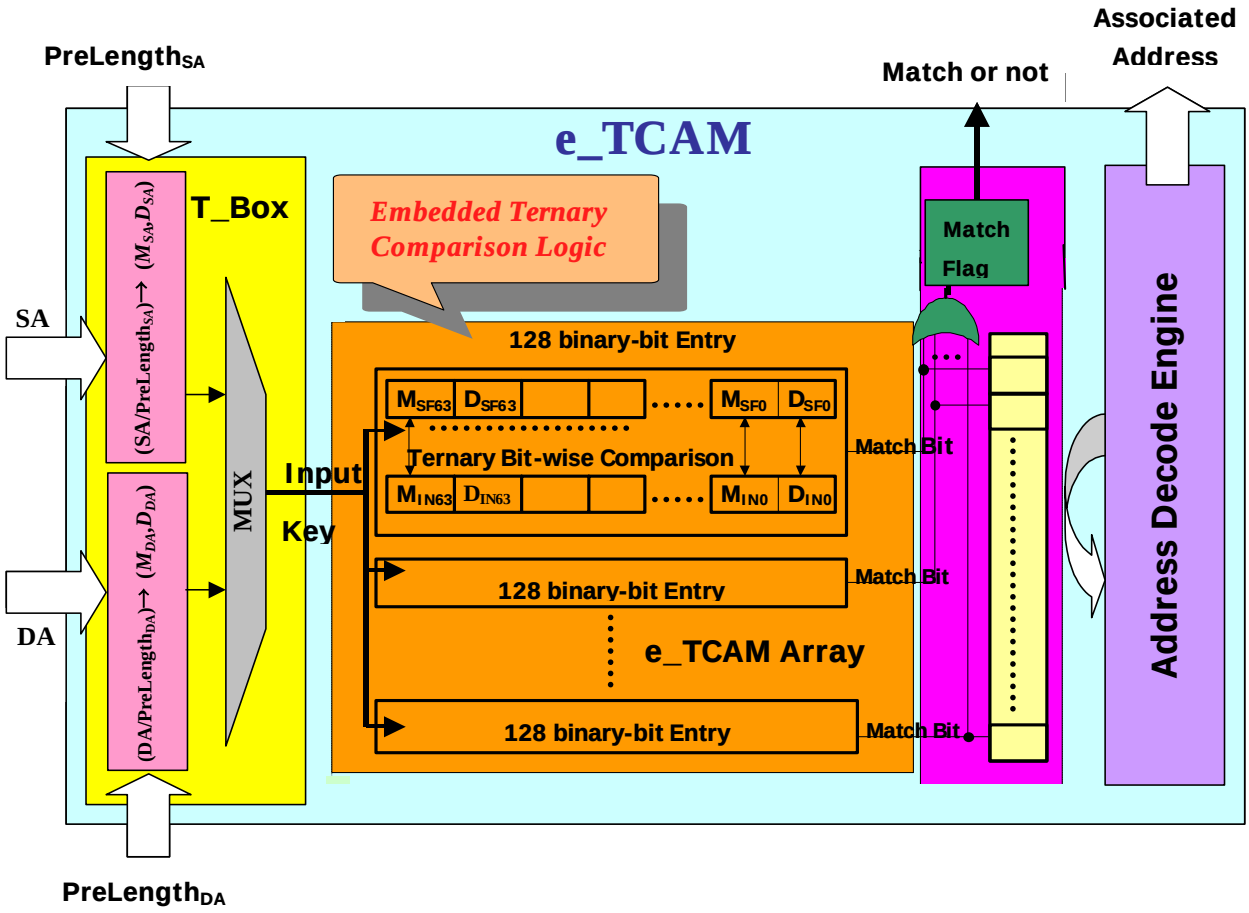


圖 0: 增強型 TCAM(e_TCAM)基礎之高速硬體運作形式的封包分類器暨分類規則衝突檢測器

三、參考文獻

[1] F. Baker, Ed. “Requirements for IP version 4 routers,” *IETF RFC 1812*, June 1995.

[2] P. Gupta, S. Lin, and N. Mckeown, “Routing lookups in hardware at memory access speeds,” *Proc. IEEE INFOCOM’98*, San Francisco, USA, pp. 1240-1247, Mar. 1998.

[3] N. Huang, and S. Zhao, “A novel IP-routing lookup scheme and hardware architecture for multigigabit switching routers,” *IEEE Journal on Selected Areas in Communications*, Vol. 7, No. 6, pp. 1093-1104, June 1999.

[4] Michigan University and Merit Network Internet Performance Measurement and Analysis (IPMA) Project. Available WWW: http://nic.merit.edu/~ipma/

[5] B. Lampson, V. Srinivasan, and G. Varghese, “IP lookups using multiway and multicolumn search,” *IEEE/ACM Trans. Networking*, Vol. 7, No. 3, pp. 324–334, June 1999.

[6] P. C. Wang, C. T. Chan, and Y. C. Chen, “A fast IP routing lookup scheme,” *IEEE Commun. Letter*, vol. 5, pp. 125–127, Mar. 2001.

[7] I. Widjada and A. I. Elwalid, “Performance issues in VC-merging capable switches for multiprotocol label switching,” *IEEE J. Sel. Areas Commun*., vol. 17, pp. 178-1189, June 1999.

[8] V. Makam, C. Huang, K. Owens and V. Sharma, "MPLS: Much Potential Leading to Somewhere? An Assessment of QoS and Protection in MPLS," Proceedings of MPLS'99, Paris, June, 1999.

[9] S. Makam, V. Sharma, K. Owens, C. Huang, F. Hellstrand, J. Weil, L. Andersson, B. Jamoussi, B. Cain, S. Civanlar, and A. Chiu, “Framework for MPLS-based recovery,” *IETF Internet Draft*, Aug. 2000.

[10] C. Huang, V. Sharma, S. Makam and K. Owens, “Extensions to RSVP-TE for MPLS Protection,” IETF contribution, work in progress, June, 2000.

[11] C. Huang, V. Sharma, S. Makam and K. Owens, “A Path Protection/Restoration Mechanism for MPLS Networks,” IETF contribution, work in progress, July, 2000.

[12] K. Owens, V. Sharma, S. Makam, B. Mack-Crane, and C. Huang, “A path protection/restoration mechanism for MPLS network,” *IETF Internet Draft*, Jan. 2002.

[13] C.Huang, V. Sharma, K. Owens and S. Makam, “Building Reliable MPLS Networks Using a Path Protection Mechanism,” *IEEE Communications Magazine*, pp. 3-8, March 2002.

[14] V. Sharma, and F. Hellstrand, “Framework for Multi-Protocol Label Switching (MPLS)-based Recovery,” *IETF RFC-3469*, Feb. 2003.

[15] C. Villamizar, “MPLS Optimized Multipath (MPLS-OMP),” *Internet-Draft*, 1999.

[16] C. Villamizar, “OSPF Optimized Multipath (OSPF-OMP),” *Internet-Draft*, 1999.

[17] K. Long, Z. Zhang, and S. Cheng, “Load balancing algorithms in MPLS traffic engineering,” *IEEE Workshop on High Performance Switching andRouting (HPSR) 2001*, pp.175-179, 2001.

[18] L. Berger, D. Gan, G. Swallow, P. Pan, F. Tommasi, and S. Molendini, “RSVP refresh overhead reduction extensions,” *IETF RFC 2961*, April 2001.

[19] D. Awduche, L. Berger, D. Gan, T. Li, V. Srinivasan, and G. Swallow, “RSVP-TE: extensions to RSVP for LSP tunnels,” *IETF RFC 3209*, Dec. 2001.

[20] D. Awduche, A. Hannan, and X. Xiao, “Applicability statement for extensions to RSVP for LSP-tunnels,” *IETF RFC 3210*, Dec. 2001.

[21] J. M. Chung, “Analysis of MPLS traffic engineering,” *Proc. IEEE Midwest Symposium on Circuits and Systems 2000*, Vol.2, pp.550-553, 2000.

[22] P. Brittain, A. Farrel, “MPLS traffic engineering: A choice of signaling protocols,” *Data Connection Limited*, Jan. 17, 2000.

[23] R. Braden, D. Clark, and S. Shenker, “Integrated Services in the Internet Architecture: an Overview,” *IETF RFC-1633*, June 1994.

[24] J. Wroclawski, “Specification of the Controlled-Load Network Element Service,” *Internet RFC 2211*, September 1997.

[25] S. Shenker, C. Partridge, and R. Guerin, “Specification of Guaranteed Quality of Service,” *Internet RFC 2212*, September 1997.

[26] S. Shenker and J. Wroclawski, “General Characterization Parameters for Integrated Service Network Elements,” *Internet RFC 2215*, September 1997.

[27] S. Shenker and J. Wroclawski, “Network Element Service Specification Template,” *Internet RFC 2216*, September 1997.

[28] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, “An Architecture for Differentiated Services,” *IETF RFC 2475*, Dec. 1998.

[29] V. Jacobson, K. Nichols, and K. Poduri, “An Expedited Forwarding PHB,” Internet Draft, draft-ietf-diffserv-phb-ef-01.txt, November 1998.

[30] V. Jacobson, K. Nichols, and K. Poduri, “An Expedited Forwarding PHB,” *IETF RFC 2598*, June 1999.

[31] J. Heinanen, F. Baker, W. Weiss, J. Wroclawski, “Assured Forwarding PHB Group,” *Internet Draft*, draft-ietf-diffserv-af-03.txt, November 1998.

[32] J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski, “Assured Forwarding PHB Group,” *IETF RFC 2597*, June 1999.

[33] K. Nichols, S. Blake, F. Baker, and D. Black, “Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers,” *IETF RFC 2474*, Dec. 1998.

[34] J. Heinanen, and R. Guerin, “A Single Rate Three Color Marker,” *IETF RFC 2697*, Sept. 1999.

[35] J. Heinanen and R. Guerin, “A Two Rate Three Color Marker,” *IETF RFC 2698*, Sept. 1999.

[36] W. Fang, N. Seddigh, and B. Nandy, “A Time Sliding Window Three Colour Marker (TSWTCM),” *IETF RFC 2859*, June 2000.

[37] F. Wang, P. Mohapatra, S. Mukherjee, D. Bushmitch, “A random early demotion and promotion marker for assured services,” *IEEE Journal on Selected Areas in Commun*., vol.18, no.12, pp. 2640-2650, Dec. 2000.

[38] V. Srinivasan, G. Varghese, S. Suri, and M. Waldvogel, "Fast scalable level four switching," *Proc. ACM Sigcomm*, Sept. 1998.

[39] V. Srinivasan, S Suri, and G. Varghese, "Packet Classification using Tuple Space Search," *Proc. ACM Sigcomm*, pp. 135-146, Sept. 1999.

[40] T. V. Lakshman and D. Stiliadis, "High-speed Policy-based Packet Forwarding using Efficient Multi-dimensional Range Matching," *Proc. ACM Sigcomm*, pp. 191-202, Sept 1998.

[41] P. Gupta and N. McKeown, "Packet Classification on Multiple Field," *Proc. ACM Sigcomm, Comp. Commun*. Rev., vol. 29, no. 4, pp.147-160, Sept. 1999.

[42] V. Srinivasan et al.," Fast and Scalable Layer-4 Switching," *Proc. ACM Sigcomm*, pp. 203-14, Sept 1998.

[43] M. Uga and K. Shiomoto, "A Novel Ultra High-speed Multi-layer Table Lookup Method using TCAM for Differentiated Service in the Internet," *IEEE Workshop on HPSR*, 2001, pp. 240 -244.

[44] P. Gupta and N. McKeown, "Algorithms for Packet Classification," *IEEE Network Special Issue*, vol. 15, no. 2, pp. 24-32, March/April 2001.

[45] A. Hari, S. Suri, and G. Parulkar, "Detecting and Resolving Packet Filter Conflicts," *Proc. IEEE Infocom*, vol. 3, pp.1203-1212, March 2000.

[46] F. Baboescu, and G. Varghese," Fast and Scalable Conflict Detection for Packet Classifiers," *Proc. IEEE ICNP*, pp.270-279, 2002.

[47] Keng-Ming Huang（黃鏗銘）, “An Effective IP Address Lookup Scheme: Unicast and Multicast（高效能之 IP 位址查詢技術：單點廣播與多點廣播）,” M.S. thesis（碩士論文）, Department of Communication Engineering, National Chiao Tung University, Taiwan, 2001.

[48] Ching-Shi Huang（黃慶喜）, “Further Investigation of TS-UPC and New Metering Algorithms for DiffServ（訊務塑型器-使用參數控制的深入研究與差異性服務網路新的量測機制）,” M.S. thesis（碩士論文）, Department of Communication Engineering, National Chiao Tung University, Taiwan, 2001.

[49] Chuang-Kuang Ting（丁崇光）, “The Load Balancing Dispatcher Using Fuzzy/ANFIS Technique in Heterogeneous Servers Environment（在異質伺服器環境下利用乏晰／適應性類神經乏晰技術之負載平衡分配器）,” M.S. thesis（碩士論文）, Department of Communication Engineering, National Chiao Tung University, Taiwan, 2001.

[50] Po-Han Chen（陳柏翰）, “Traffic Reduction and Path Recovery Algorithm for Real-time Services（即時性服務之流量縮減與路徑復原演算法）,” M.S. thesis（碩士論文）, Department of Communication Engineering, National Chiao Tung University, Taiwan, 2002.

[51] Yung-Hung Cheng（鄭永宏）, “The Traffic Conditioner improving Bandwidth Allocation for DiffServ Networks（在差異性服務網路中具有改善頻寬配置功能的訊務調節器）,” M.S. thesis（碩士論文）, Department of Communication Engineering, National Chiao Tung University, Taiwan, 2002.

[52] Yu-Kuey Wu（吳育葵）, “The Scheduling Algorithms for Proportional Delay Differentiated Services（提供比例式延遲差異性服務相關排程演算法之研究）,” M.S. thesis（碩士論文）, Department of Communication Engineering, National Chiao Tung University, Taiwan, 2002.

[53] Ning-You Yan（顏寧佑）, “Packet Classification and Traffic Marking Algorithms for DiffServ Networks（差異性服務網路中封包分類與訊務標示方法之討論）,” M.S. thesis（碩士論文）, Department of Communication Engineering, National Chiao Tung University, Taiwan, 2003.

[54] P. C. Lin and C. J. Chang, “A Priority TCAM IP-Routing Lookup Scheme,” *IEEE Communication Letters*, Vol. 7, No. 7, pp. 337-339, July 2003.

[55] K. M. Huang and C. J. Chang, “A Fast Multicast IP-Routing Lookup Scheme,” *IEEE Communications Letters*, Vol. 7, No. 3, pp. 133-135, March 2003.

[56] P. C. Lin and C. J. Chang, “Analysis of Buffer Requirement for ATM-LSRs with Partial VC-Merging Capability,” *IEICE Transactions on Communications*, Vol. E-85B, No. 6, pp. 1115-1123, June 2002. (NSC 88-2213-E-009-088, NSC 89-2213-E-009-105)

[57] P. H. Chen and C. J. Chang, “An Enhanced Path Recovery Algorithm for MPLS Networks,” *Proc. of National Symposium on Telecommunications (NST’2002)*, Nantou, Taiwan, Dec. 2002, Session. NET-1-6.

[58] C. J. Chang, Y. H. Cheng, and L. F. Lin, “The Traffic Conditioner with Promotion and Fairness Guarantee Schemes for DiffServ Networks,” *Proc. of ICC 2003*, Anchorage, Alaska, USA, May 11-15, 2003, Vol. 1, pp. 238-242.

[59] F. Akujobi, I. Lambadaris, R. Makkar, N. Seddigh, and B. Nandy, "BECN for congestion control in TCP/IP networks: study and comparative evaluation," *Proc. IEEE GLOBECOM'02*, pp. 2588-2593, Nov. 2002.

[60] Yi-Cheng Chan, Chia-Tai Chan, and Yaw-Chung Chen, “An Enhanced Congestion Avoidance Mechanism for TCP Vegas,” *IEEE Communication Letters*, Vol. 7, No. 7, pp.343-345, July 2003

四、計畫成果自評

本子計畫的研究主題是在寬頻網際網路的環境中，為達成頻寬與 QoS 服務品質保證的各項訊務控制或運作機制之研究設計，針對寬頻網際網路中的路由選徑（包含 Unicast Routing 與 Multicast Routing）、寬頻傳輸技術與多項 QoS 相關之控制運作機制，皆提出有效、且效能更佳的設計或解決方案，在三年的執行時間中，計畫參與人員分別針對各項研究主題的各個層面進行探討，除了瞭解現有方法與研究的不足之處，也追蹤並研讀最新發表的相關研究論文或技術標準，進而提出相對的改進方案。

在高速路由選徑機制方面，我們分別發展出「階層式分群解析架構」與「TCAM-based 架構」兩種適合於硬體邏輯電路實現的路由選徑方法，對於每一個 IP 封包的路由查詢，可達成平均在 1 至 2 次的記憶體存取(Memory Access) 動作內即可以獲得路由結果，大幅提昇網際網路中路由選徑的處理速度，而配合硬體邏輯電路的運作方式，可以滿足 Gigabit 超高速網路環境下以及未來更寬頻的網際網路應用的需求；並且可以將相關電路整合成一硬體路由搜尋引擎的模組形式，更方便商用化路由功能產品的使用與成本的降低。同時透過適當的運作架構與資料結構的安排和設計，也兼具所需記憶體容量小的優點，此優點將有助於進一步將前述硬體路由搜尋引擎轉化為系統晶片(SoC) 的可行性（因為已經可以將運作所需的記憶體一併整合入單一晶片中），或是在傳統的路由器架構中能夠以相同的記憶體容量支援更大規模網際網路（例如：骨幹級網路）的訊務路由工作。另外我們也對於可降低網路訊務流量的網路「群播(Multicast)」機制所需的特殊路由方式－群播路由(Multicast Routing) 進行研究設計，在瞭解 Multicast Routing 的關鍵技術是為一個多欄位資料完全相符(Exactly Match) 的比對搜尋機制後，我們提出以 CAM 此一支援資料完全相符比對搜尋的記憶體為基礎的高速群播路由選徑方法，可在 3 次的記憶體存取動作便完成一筆群播路由的查詢，再加上管線式(Pipeline) 運作架構的可行性，也同樣能使得平均查詢次數降至一次的記憶體存取次數。

在寬頻傳輸技術方面，我們著眼於 MPLS 網路的相關議題。首先，對於以 ATM 技術為基礎的 MPLS 網路之 VC-merge 機制，提出 Partial VC-merge 運作模式下 ATM-LSR 交換路由器記憶體容量需求之快速且精確的分析方法：不僅較傳統分析方式降低分析過程的複雜度（可由 $O(M^4)$ 減少為 $O(M^2)$），也因此能夠對較大 Buffer 的分析得到較高的 Cell Loss Prob.的準確度；而我們所採用的 Input 訊務的數學模型更接近於實際的情況，所得的結果也較為接近真實，所以此分析方法極適合用於評估實際 ATM-LSR 交換路由器所需的記憶體緩衝區容量(Buffer Size)，讓交換路由器的系統資源做更有效率的應用。其次，我們也提出一 MPLS 網路路徑損壞或錯誤發生的保護及快速回復(Recovery) 機制，讓傳輸路徑發生錯

誤或損壞時還能夠維持部分基本的通訊，並可以快速而正確地地恢復既有的通訊，降低高速 MPLS 網路上傳輸路徑錯誤或損壞所造成的影響，減少封包遺失率，並能夠進一步達到動態負載平衡的附加效益，使系統資源做最佳的利用，因而有效提高 MPLS 網路的資料輸出率(throughput)。

在支援 QoS 保證的訊務處理與控制機制方面，我們著重在 IETF 所建議的 DiffServ 網際網路 QoS 架構之基礎關鍵性元件的研究與設計，包含封包分類器(Classifier) 與訊務監控調節器(Traffic Conditioner)。在封包分類器(Classifier) 的研究與設計方面，我們參考之前採用硬體邏輯電路實現的高速路由選徑方法的發展經驗，提出以 TCAM 為基礎的高速硬體化操作之封包分類機制架構，並進一步修改 TCAM 既有的邏輯運算電路，使其具備雙三元資料比對邏輯運算的能力，在搭配適當的周邊邏輯電路後，便能夠同時兼具高速封包分類器以及高效能封包分類規則衝突偵測裝置的功能。在訊務監控調節器(Traffic Conditioner) 的研究設計方面，提出一兼具訊務監控精確度與網路資源使用效率，並進一步達成 micro-flow 微訊務流**之 Cell Dropping Precedence** 標記公平性(Fairness) 的訊務監控調節機制：在確保訊務特性符合 Traffic Profile 的規範下，精確地讓系統資源獲得充分且最佳的利用，並進一步保障其中各 micro-flow 連線所實際獲得的 QoS，如此才能夠**提升 DiffServ 此網際網路 QoS 架構的可行性與使用效益**，並做為其他 QoS 訊務控制機制（例如：CAC 連線允諾控制機制、Scheduling 排程控制、流量控制、壅塞控制等 Per-Hop Behavior）運作的基礎。

此外，特別值得注意的是，在路由選徑機制（包含 Unicast 和 Multicast）以及封包分類器的設計上，都具有以硬體的邏輯電路運作架構為主的設計目標，以期利用硬體的快速運作特點來得到速度上的提升。對此，我們除了開發設計自創(有)的硬體邏輯架構和電路之外，也思考採用現有功能接近的數位邏輯電路或晶片為基礎的方式來進行設計，以收事半而功倍之效，在符合設計目標的同時，也因為採用實際的邏輯電路元件為基礎而具備即時實用的能力和優點（亦即實際應用的可行性）。而採用 CAM 或是 TCAM 為基礎所設計的資料搜尋比對方法，除了具有硬體快速運作的優點之外，其真正用於搜尋比對動作的候選資料和資料的結構與人們邏輯概念上理解的運作形式相同，無須做進一步的轉換，如此有助於方便資料的維護更新以及操作者的理解。目前雖然以 CAM/TCAM 為主的方法或許還有 CAM/TCAM 相關晶片單價仍高、以及搜尋比對的候選資料容量仍不算非常大的實用上限制，不過可以透過多個晶片整合使用的方式擴大容量，未來預期在應用範圍擴大以及需求量增加後，將可以因為大量製造而降低成本。再加上晶片設計和生產相關技術的益發進步與成熟（例如：製造良率的提升），也將有助於成本進一步的下降和效能的提升，這包含運作速度和容量的提升。屆時我們也預期，採用 CAM/TCAM 為基礎的資料搜尋比對機制將會是最經濟而有效益的方式。

綜合這三年的研究成果，我們認為研究內容與原計畫緊密結合，已達成預

期的計畫目標。此外，本計畫的研究成果已整理成相關的學術論文投稿，且為國際性學術期刊所接受並刊出，或為國際會議所接受並於相關議程中進行發表 [54-58]。另一方面，我們的研究成果也具備相當高的實用價值，除了擬進行專利的申請外，亦可進一步與產業界合作或技術移轉進行實際產品的應用與開發，落實學術研究成果。綜此觀之，我們於本計畫執行的研究成果兼具有學術成就與實用價值，可謂相當成功而豐富。

附件一

國際合作研究計畫國外研究報告書

●計畫名稱： 寬頻網際網路端對端技術之研究
(End-to-End Techniques for Broadband Internet)

●合作機構：加拿大・渥太華・Carleton 大學寬頻網路實驗室

●執行時間：2002 年 7 月 26 日　　2002 年 8 月 25 日

●我方研究人員： 張仲儒 教授（第一子計畫暨總計畫主持人）、
陳耀宗 教授（第三子計畫主持人）、
林立峰、詹益禎（博士班研究生）

●對方計畫主持人： Prof. Changcheng Huang（黃長城 教授）、
Prof. Ioannis Lambadaris

●國際合作研究計畫之緣起和必要性：

本國際合作研究計畫案，是於原三年期國科會整合型計畫「寬頻網際網路端對端技術之研究(End-to-End Techniques for Broadband Internet)」執行的第二年度，進行計畫變更為國際合作研究計畫，與加拿大渥太華 Carleton 大學工學院的寬頻網路實驗室(BNL) 合作，共同投入下一代支援服務品質(QoS) 之寬頻網際網路相關技術的研究。

在我們原本的三年期整合型計畫中，主要的目標即是針對達成寬頻網際網路所需的有線或無線寬頻傳輸技術，以及頻寬與 QoS 服務品質保證的各項端對端訊務控制或運作機制進行研究設計，提出有效、且效能更佳的解決方案，或是從事其效能分析評估與改進的工作；此外，亦以一子計畫投入網際網路中具有自我類似(Self-similar) 性質之真實網路訊務的研究，瞭解其特性和分析其成因，如此將有助於 QoS 訊務控制機制或通訊協定於實務應用上的最佳化設計。此子計畫並擬提出一有效率的 Self-similar 訊務源模型與訊務產生方法，可做為所開發的各項控制機制或架構之效能模擬分析、評估之用，或進一步做為其設計的考量或改進的依據。

而 Carleton 大學寬頻網路實驗室在其實驗室計畫主持人 Prof. Changcheng Huang（黃長城教授）與 Prof. Ioannis Lambadaris 兩位教授的指導和帶領下，過去在寬頻網路之訊務源模型、快速（網路）模擬技術、壅塞控制機制、與光纖網路之架構與傳輸技術等方面已有相當豐富的研究成果，甚至在其中的 Self-similar 自我類似特性的訊務與快速網路模擬技術上更是有世界領先的卓越表現。近幾年來，該實驗室研究團隊則開始投入網際網路上的服務品質保證相關技術的研究，

在支援 QoS 服務品質保證的訊務控制機制、網際網路群播(Multicast) 技術的應用和相關運作控制機制的開發和設計、甚至是未來全光化高速光網路與現有 IP 網際網路之間的網路互連運作(Interworking) 或網路整合的問題都加以探討並提出解決方案，其中亦延續在 Self-similar 特性訊務源的研究（成果），輔助各項控制機制或架構的設計，確保達到實務應用上的最佳效能。

由此觀之，雙方研究團隊在相關研究領域上均有相當且相近的背景與實力（經驗），同時對於目前和未來規劃的研究課題上亦有共同的目標與興趣，因此非常適合以共同研究合作的方式交流彼此的經驗和成果，除了可促進雙方在既有研究題目的進度與研究領域的深度之外，也可由不同視野和經驗的激盪，達到提升雙方在其他相關研究領域的廣度的雙贏效益。此外，Carleton 大學系統及電腦工程學系寬頻網路實驗室，長期與國際通訊大廠 Nortel Networks 或其他業界公司有密切的合作關係，在寬頻網路訊務控制機制、訊務模式、與光纖通訊網路等已有多年的設計實作心得及研究成果。著眼於本研究團隊大多的研究成果僅止於理論分析或電腦模擬驗證，較缺乏實際系統或網路環境的試煉，因此期望透過本計畫雙方的合作研究，能達到理論與實務並重，以及實際問題思考經驗與理論推導另類方向的激盪討論。

●國外研究之執行：

於本三年期整合性計畫執行之第二年度開始，進行計畫研究人員赴國外合作單位訪問研究之行程，並以兩個年度共四位教授與四位博士班研究生的人次來執行。第一年度的訪問研究於 2002 年 7 至 8 月間順利執行完畢，由張仲儒（子計畫一暨總計畫主持人）與陳耀宗（子計畫三主持人）二位教授帶領博士班林立峰和詹益禎二位研究生，赴加拿大渥太華 Carleton 大學進行為期一個月的國外（出訪）研究交流行程。然而由於研究主力仍是以學生部分的交流為主，因此二位教授的部分實際僅規劃十天的行程，主要是希望能夠透過雙方計畫主持人會面的機會，擬定此次國外研究的實際進行方式，並討論後續在雙方跨國合作計畫上的執行模式、細節與計畫預期成果。待二位教授返國後，兩位博士生仍留在合作的寬頻網路實驗室與該實驗室成員共同研究，學習並交流彼此的經驗與成果，至一個月後才結束此次國外研究行程返國。第二年度（原整合型計畫第三年度）的國際合作研究部分，亦分別規劃有兩位教授與兩位博士班研究生進行國外（訪問）研究的行程，由於原訂於本年度出訪的陳伯寧教授（子計畫二主持人）臨時無法成行，在行文國科會變更出訪教授名單後，則仍維持由第一年度的張仲儒與陳耀宗二位教授帶領另二位博士生來進行。第二年度所核准的研究時程較第一年度增加為三個月，但教授部分仍維持十天，除了用於雙方目前研究進度與成果的報告外，也仍希望透過見面的機會，進行上一年度計畫執行問題的檢討，同時有鑑於此年度已為計畫執行的最後一年度，故也期望能針對計畫結束後，如何延續此一

國際合作研究關係、維持雙方持續而穩定的學術交流，討論出可行的進行方式。然而原擬於 2003 年 4、5 月間出發的第二年度國外研究行程，卻因為當時國際間爆發了 SARS 的流行，而我國亦成為 SARS 疫區，且加拿大方面升高對我防疫與檢疫的標準，所有入境人員皆必須進行十天自我或強制隔離的安全防疫措施，相關手續繁瑣而不便，在與加方 Carleton 大學計畫主持人聯繫討論後，決定取消第二年度國際合作計畫的國外研究行程，並行文國科會告知此一訊息且同時申請變更計畫執行。

由於 SARS 的影響，導致本國際合作研究計畫之國外研究部分僅於原整合型計畫之第二年度執行，而且所核准的研究時程較短（教授十天，博士班研究生一個月），為了在短時間內能有一定的成果呈現，在雙方計畫主持人會議討論後決定採取的策略是，先針對各自手邊已略具雛形的研究進行報告及討論，在對彼此的研究專長與目前正在進行的研究題目有所瞭解後，再由雙方計畫主持人討論決定具體的合作研究題目，原則上是採取雙方皆有興趣也最好是正在進行中的研究題目為主。因此接下來便以兩天的時間進行雙方的研究專題報告，題目的範圍涵蓋了網路標準的二、三，四層。

在我方的報告中，由陳耀宗教授與其博士班研究生詹益禎提出一個關於網際網路 TCP 通訊協定於網路壅塞狀態下的問題，即是當 TCP 的 ACK 路徑上若發生擁塞，會因為對於整個通訊連線的 RTT 值的估算錯誤而做出不恰當的 Congestion Window 的調整，如此將對於 TCP 的控制效能有嚴重的影響。理論上 ACK 路徑上的擁塞不應該造成 Data Packet 路徑上的效能減低，然而實際上的運作結果的確會發生這樣的問題，特別是對於一個新提出的 TCP 版本「Vegas」的影響最為嚴重，因此當前正著手進行 TCP Vegas 上此一課題的研究，以期提出簡單而有效的解決方案。而張仲儒教授與博士班研究生林立峰則針對本身實驗室過去在寬頻網路上各項支援 QoS 服務品質保證的訊務控制機制的研究，以及本整合型計畫第一年度的研究進度與成果，連同目前正在研究的課題，進行介紹和說明，主要是著重在連線允諾控制(Call Admission Control, CAC) 機制、訊務監控調節(Traffic Conditioner) 機制、採用硬體邏輯電路架構為主的高速路由選徑方法、以及 MPLS 寬頻傳輸技術網路之路徑錯誤保護暨快速回復機制…等方面。而有鑑於 Carleton 大學方面的計畫主持人黃長城教授在 MPLS 網路之可靠性研究領域亦長期投入，並已有相當成熟的研究經驗和成果 [8-13]，其中包括參與網際網路標準組織 IETF 在相關技術的建議標準—RFC-3469 [14] 的制訂工作，以及恰於當年度國際期刊 IEEE Communication Magazine 中有相關的、MPLS 網路之通訊路徑回復機制的論文發表 [13]，因此我們也特別針對此方面的研究進度做較為深入的報告，以期雙方能在此問題上有進一步探討與合作的機會。而 MPLS 網路之路徑回復機制的起因與問題主要是在於，雖然藉由 MPLS 技術的提出的確適時為網際網路提供了高速且低延遲的訊務傳輸能力，但相對的，當高速的 MPLS 網路發生傳輸路徑錯誤或損壞的時候，往往也會造成更嚴重的影響（例如

更大量的資料遺失)，尤其對於即時性服務而言更是如此。因此有必要提出一套有效的路徑保護(Protection) 或回復(Path Recovery) 機制，以便當 MPLS 網路傳輸路徑發生錯誤(failure) 或損壞的時候，還能夠維持部分基本的通訊，並可以**快速而正確**地恢復既有的通訊，降低 MPLS 網路上傳輸路徑錯誤或損壞所帶來的影響，減少封包遺失率，如此便可有效的提高網路的資料輸出率(throughput)。

在 Carleton 大學方面，則是由計畫主持人及寬頻網路實驗室其他研究成員進行實驗室研究成果介紹，以及幾項進行中或已完成之研究專題報告，包含：採取分波多工(Wavelength Devision Multiplexing, WDM) 技術之光纖網路中，資料傳輸用光波長的最佳化分配機制、IEEE 802.17 標準的 RPR 環狀光網路中，光網路節點之資料傳輸多工緩衝器的效能分析與設計、以及網際網路中 TCP 壅塞控制機制…等方面的具體研究成果。其中特別值得注意的是，在 TCP 通訊協定的壅塞控制機制方面，研究生 Frank Akujobi 提出了一個有別於傳統 TCP 壅塞控制的方式，不再只是單純地由網路上一通訊連線的兩個端點來進行而已，而是希望可以利用連線的封包路徑上所通過的路由器的協助，能夠進一步提高 TCP 的效能。其作法是，當路由器發現網路開始壅塞時，路由器便透過主動發出 ICMP 訊息(BECN) 的方式，通知經過它的所有 TCP 連線的 Source 端網路已開始壅塞，使 TCP 的 Source 端及早對壅塞做出反映。經過不斷的討論及修正，最後經由網路模擬器(NS) 的模擬，結果證實達到了預期的效果。此研究結果已發表在國際會議論文中，如 [59] 所示。

根據前述專題報告會議的交流與討論的結果，雙方決定以「TCP 通訊協定之壅塞控制機制」和「MPLS 網路之路徑錯誤保護暨快速回復機制」此兩項議題為此次合作研究的主要項目，但也保持在其他研究議題方面的討論和經驗交流。

●國外研究之具體成果：

在赴 Carleton 大學寬頻網路實驗室進行為期一個月的合作研究行程，以及返國後我研究團隊與對方持續透過 E-mail 進行討論後，在上述兩項具體的合作研究課題上皆獲得初步的研究成果。在「TCP 通訊協定之壅塞控制機制的設計」方面，針對 TCP Vegas 在 ACK 路徑壅塞時的效能下降問題，擬採取的解決方式是，在 TCP 連線的 receiver 端對每一個 ACK 打上時間標記(timestamp)，source 端經過某種運算，分別計算出封包在 Data Packet 路徑上以及 ACK 路徑上的佇列延遲時間(queuing delay) 以瞭解網路壅塞的狀況，藉著區別出網路擁塞的方向後，我們所提出的控制方法便可透過較正確的 RTT 值估算而有效的改進 TCP Vegas 在 ACK 路徑上發生擁塞時的效能，提升訊務流量(Throughput)。此研究結果已發表在期刊論文中，如 [60] 所示。在「MPLS 網路之路徑錯誤保護暨快速回復機制」方面，我們提出一個結合「Re-Routing Model」與「Protection Switch

Model」兩種方式的優點高速而最佳的路徑保護／回復機制的方法，利用每個 LSR 上預先規劃的 Backup path 的 Database 來幫助我們達成快速路徑回復的目的，取代了以往錯誤發生後才開始搜尋 Backup path 或是錯誤發生前先保留資源建立 Backup path 的缺點：藉由 Re-Routing Model 方法的使用，不但排除了 Protection Switch Model 此類作法中預先建立 Backup path 並保留資源而造成資源浪費的缺點，引入 Protection Switch Model 預先規劃 Backup path 的概念，加速了既有 Re-Routing Model 作法的 Recovery time，而且仍保有 Re-Routing Model 中在最近路徑錯誤的節點馬上進行切換的優點，因而能夠快速且有效地降低網路錯誤發生時之影響時間與所需付出的代價。而且在原訂的研究目標之外，我們還能夠透過所設計的機制具備多路徑傳輸的能力達到動態**負載平衡**的附加效益，使系統資源做最佳的利用，如此可再更進一步提高網路的資料輸出率(throughput)。目前此部分的研究成果已整理為論文投稿 [57]，並已在 2002 年於暨南大學所舉辦的全國電信會議(NST 2002) 中進行發表。

而在上述的兩項具體合作研究成果之外，在本計畫其他的研究項目中，有許多具有實用價值的設計亦有對方的協助，達成國際合作計畫預期的理論與實務並重的目標！

●國外研究心得與建議：

經過短短四星期的交流及合作討論，合作雙方藉由腦力激盪的過程，都獲得了新的研究思維，也認識到對方研究態度上的長處，對於本研究團隊日後的研究應有不小的正面助益。同時藉著此次合作研究關係的建立，將有助於未來在長距離大型網際網路（特別是跨國性長距離網際網路）上通訊與訊務控制相關機制的分析與研究的發展，雙方可實際建立一跨國性網際網路實驗平台，彼此互為對方在此實驗平台兩端的研究助理，將所發展的研究成果在此實際的跨國性大型網際網路上進行測試、驗證，所得結果將可進一步改善我們的設計，亦可對於我們研究成果的正確性與實用性提供強而有力的證明。

由於此計畫是本研究團隊第一次的國際性合作研究案的執行，經驗仍不太足夠，所以在實際出訪時間的選擇上恰好正是當地暑假的開始，因此留在學校的學生人數並不多，互相交流機會比較少，再加上此期間也多為教授們的休假、休息時間，因此在原本合作的教授外，也未能再和更多當地的教授認識與交流，為本次國外合作較為可惜之處。另外在核准出訪的時間長度方面亦顯不足，許多研究交流與成果多有賴於返國後繼續透過 E-mail 進行後續的討論來達成，雖然這也可以是一種計畫執行的方式與考驗，不過面對面的對談、討論仍是比較有助於問題的釐清，對於研究工作而言是比較有效率的方式。期望在未來的國際合作研究計畫裡可以針對這些部分加以改進，延長出訪國外研究的時間，建議至少為三

個月的時程，避免過短的時間將可能導致赴國外研究人員在開始習慣當地的生活模式與作息規律，並可以進入實質的研究工作時即屆返國日期並必須暫停研究工作的缺點了。而較長時間的合作與交流，也較有助於建立更佳深厚而長久的友誼與合作研究關係。