附件一

行政院國家科學委員會補助專題研究計畫■ 成 果 報 告 □期中進度報告

異質網路環境之行動搜尋關鍵技術-子計畫一：

車用隨意網路存取控制與連結機制之研究 (3/3)

計畫類別：□ 個別型計畫　　■ 整合型計畫
計畫編號：NSC97-2221-E-009-049-MY3
執行期間：97 年 08 月 01 日至 100 年 07 月 31 日

計畫主持人：簡榮宏 教授
共同主持人：
計畫參與人員：鄭安凱、蔡嘉泰、曾蕙如、張家瑋、楊子興、黃志賢、黃淑盈、曾宇田、潘欣雅、彭冠傑、林良叡

成果報告類型(依經費核定清單規定繳交)：□精簡報告　■完整報告

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

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢
■涉及專利或其他智慧財產權，□一年■二年後可公開查詢

執行單位：國立交通大學資訊工程學系(所)

中　華　民　國　100　年　7　月　31　日

中文摘要：

本子計畫主要研究目標在於發展VANET 網路之存取控制協定(Media Access Control, MAC)及連結機制(Connectivity Support)。為到達這個目標，我們已針對車用隨於網路於安全性質服務、非安全性服務、以及通訊安全於所面臨的議題，提出解決的機制。在第一年的研究中，我們提出一個以功率控制為基礎的聯合碰撞避訊息廣播機制，並針對通訊全問題，設計了一混雜式對應基礎的金鑰協議協定，此這個協定不需任何檢查表就可在伺服器與使用者之間達到人工認證以及區段金鑰協議的能力。在第二年的計畫中，我們進一步對廣泛的安全應用，提出了一個以車輛密度為基礎的緊急訊息廣播機制VDEB，解決車間通訊所存在的延遲與傳送失敗問題，並利用WAVE/DRSC多重通道的特性，設計了多車道自由流動ETC系統的設計，提供高車流環境中穩定且快速的交易需求。在第三年的計畫中，我們對各種的非安全性應用，提出了一行動閘道器式繞徑協定，此協定結合了V2V與V2I兩種通訊模式的優點，並以部分車輛作為行動閘道器，延伸固定式式RSU的涵蓋範圍，以減少封包跳躍次數和連斷線的可能性。此外，針對通訊全問題，我們進一步提出了一個安全聚集訊息驗證機制簡稱SAMA。此機制是基於免憑證公開金鑰密碼系統，用以驗證車用隨意網路之突發緊急事件。經由效能的評估，此機制可有效地降低聚集訊息驗證所需之計算量。並驗證此一方法可有效地抵禦偽造攻擊，同時提供車輛之隱私保護。本計畫三年的成果，共包括三篇期刊論文以及六篇研討會論文發表，並培育了三位博士生以及七位碩士生。

關鍵詞： 車用隨意無線網路、聯合車輛碰撞避免、緊急訊息廣播、電子收費系統、多重通道架構、能源控制、行動式閘道、金鑰協議協定、安全訊息驗證

英文摘要：

The goal of subproject 1 is to provide the medium access control mechanism and connectivity support for vehicle ad-hoc networks. To achieve this goal, we have designed several mechanisms for safety services, non-safety services, and security issues in Vehicular Ad Hoc Network (VANET). In the first year, we proposed an efficient broadcast mechanism for the Cooperative Collision Avoidance (CCA) system using the power control technique, and designed a novel key agreement protocol based on chaotic maps to enhance the security. In the second year, we proposed a Vehicle-density-based Emergency Broadcast (VDEB) scheme to reduce the overhead and long delay in safety services. We also designed a multi-lane free-flow ETC system using the multi-channel character of the WAVE/DSRC. In the last year, we proposed a Mobile-gateway Routing Protocol (MGRP), for VANETs. The MGRP combines V2V and V2I communications, and utilizes certain vehicles as mobile gateways to improve the packet delivery ratios and reduce the transmission hop count. Moreover, we proposed a Secure Aggregated Message Authentication (SAMA) scheme in certificateless public key settings to validate emergency messages for VANETs. In the three-year project, we have totally published 3 journal papers and 6 conference papers, and fostered 3 Ph.D. students and 7 graduate students.

Keywords: Vehicle Ad-Hoc Network, Cooperative Collision Avoidance, Emergency Message Broadcasting, ETC, Multi-channel Architecture, Power Control, Key Agreement Protocol, Security Message Authentication

一、前言

隨著無線通訊網路技術的進步，加上電子元件價格漸漸下降以及行車安全問題越來越受重視，各國政府紛紛投入智慧型運輸系統(Intelligent Transport Systems, ITS)的研究[1-7]。ITS 為應用先進的電子、通信、資訊與感測等技術，以整合人、路、車的管理策略，其主要的目地為提供即時(real-time)的資訊，並增進運輸系統的安全、效率及舒適性，同時也減少交通對環境的衝擊。

車用隨意網路(Vehicle Ad-hoc Network, VANET)，為當前了為實現 ITS 所發展出重要網路架構[8]。在個架構下，每一個搭載無線通訊設備的車輛，都可以透過路旁的路側系統(Road Side Unit, RSU) 連線到 Server 端索取所需的資訊。而當車輛距離基地台太遠超出傳訊範圍時，也可透過其它的車輛幫忙轉送。換句話說，每一個車輛都可視為是一部行動無線路由器(Mobile Wireless Router)。因此，這樣的架構可大幅提升網路建置彈性，只要有車輛的地方，即可是 VANET 的涵蓋範圍。

車用隨意網路的應用主要可以分為兩類: (1) 安全性質服務(safety services) (2) 非安全性質服(non-safety services)。前者是在限定區域內提供一對多的緊急訊息廣播(one-to-all emergency message broadcasting)服務，例如電子煞車系統(electronic brake light)、車道切換輔助(lane changing assist)、路況報導(road condition reports)等，這類應用通常是與生命安全息息相關的(life critical)，因此訊息不只是要能成功的被接收還要在極短的時間內到達，以讓駕駛者有更多時間來對緊急事故作出反應。後者則是提供一對一資料路由(one-to-one data routing)或一對多資料廣播(one-to-all data broadcasting)服務，例如娛樂相關、路徑規劃等，這類應用著重於高穩定高頻寬的傳輸需求。除此之外，達到安全與非安全性質服的同時，車用網路本身的通訊安全性也是個重要的議題，如何安全、有效率、隱私地傳送資料，將影響車用網路於各種應用的可行性。

在安全性質服務部分，根據統計，每年超過數千萬的交通事故是因為車輛撞擊所引起的[9]。而導致車輛撞擊的因素因有許多，諸如車輛機械問題、天候狀況、行車時段等。其中，駕駛者的行為(driver behavior) 是最重要的因素。在行車過程當中，駕駛者若無法對緊急發生的事故(如: 落石、緊急煞車、車輛打滑) 即時作出煞車反應，則經常會造成一連串的車輛撞擊 (chain car collision)，也就是俗稱的連環車禍。VANET 網路的建構可大幅改善這樣的問題。當緊急事故發生時，前方車輛可將些事故的訊息，直接透過無線媒介的傳遞給後方的車輛，以避免中間車因為視線上阻隔的所帶來的延遲。此訊息可再藉由車輛上無線裝置的轉送，快速的傳遞給更後面的車輛，因此可以免除掉駕使者本身所需反應時間。這樣的系統稱作聯合碰撞避免(Cooperative Collision Avoidance, 簡稱 CCA) [10-14]。然而，當過多的訊息於無線媒介中傳送時，則會有嚴重干擾的關題(interference)，進而導致較大的傳輸延遲 (delivery delay)，這將直接威脅到駕駛人的安全。

更進一步來看，當車輛偵測到前方的碰撞事故，必需立即廣播出一個訊息來告知周邊的車輛，在這同時其中一台接收訊息的車輛會被選作轉送者(forwarder)來進行訊息的轉送。相同的，此轉送者會將訊息廣播出去並選擇下一個轉送者，這樣的動作會一直重覆直到此訊息已傳達到風險區的邊緣。由於訊息都只透過傳送者來廣播，如果沒有適當的選者轉送者，將會造成過多的車輛參與轉送，使得資料通道的競爭(channel contention)和資料封包碰

撞(packet collision)問題變的嚴重，而產生廣播風暴(broadcast storm)的問題，進而導致較高的延遲時間和傳輸失敗率。為了避免資料封包的碰撞，轉送者在接收到訊息後，通常會等待一段隨機的時間才進行轉送，這段等待時間(waiting time)也會增加傳輸的延遲，在最緊急的情況下(如上述的 CCA 系統)，所能容忍的傳輸延遲是非常低的。

在非安全性質服務部分，高速路電子收費系統(Electronic Toll Collection, ETC)為此類服務的重要應用之一，也是目前各國的趨勢[15, 16]。現行的高速公路電子收費模式分為單車道自由流動(Single-lane free-flow, SLFF)與多車道自由流動(Multi-lane free-flow, MLFF)兩種架構[17]。兩種架構都屬於自由流動模式，允許車輛自由通過收費口，不需設置柵欄迫使車輛停下來。不同點在於MLFF允許車輛任意切換車道不受閘門的限制，因此車輛可以較高的速度通過收費站，尤其是在車流量高的地區更為明顯。然而MLFF設計相對較為複雜，必需有穩定的通訊品質來進行高流量的收費交易。

不僅止於收費系統的應用，車用網路更日益收到各種商業用途的關注，如娛樂、遊程規劃、多媒體內容提供等。在車對車(Vehicle-to-Vehicle, V2V)模式中，每輛裝載車輛單元(On-Board Unit, OBU)的車輛都可透過多重跳躍的方式與其它車輛進行通訊。然而不同於傳統的無線網路，車用網路在傳遞資訊的同時，將面臨來至於高車速與車輛分佈變動所造成的影響，不易事先建立固定的路由路徑來傳送資料。由其在低車流密度的地區，更可能造成網路不連通的問題，而降低傳送的成功率。這樣的問題可透過車對基礎架構(Vehicle-to-Infrastructure, V2I)通訊模式的輔助來解決。在這個模式下，車輛可先將資料傳送到一個鄰近的路側單元(Roadside Unit, RSU)，RSU 再將資料透過骨幹網路轉送到目的地車輛，因此可提升通訊的品質與可靠性。然而，受限於有限的傳輸距離，V2I 通訊模式的輔助僅限於 RSU 所在的區域，在沒有 RSU 的區域，上述的問題仍然存在。

而在通訊安全部分，車用網路屬於公開性質的通訊網路，必需透過金鑰協議協定(Key Agreement Protocol)，讓多個通訊方可以協調出一把金鑰或交換資料為，以建立接下來的通訊建立區段金鑰 [18]，讓偷聽者無法猜測這把金論。然而，過去的方法無法提供通訊雙方的驗證，而且容易受到中介者攻擊(Man-in-the Middle Attack)。至 1990 年起，混雜式系統(Chaotic System) [19-24] 被使用於設計安全通訊協定之用，此系統可分為類比)與分散式數位兩種主要的方法，前者基於使用混雜式線路的混雜同步，後者則用作混雜式精片的產生。許多混雜式協定被提出[25, 26]，然而這些協定仍然有安全上的弱點，所需的驗證表可能會被竄改或竊取，並且需要較高的維護成本，此外，使用者可能會希望匿名地取得服務。

另一方面，由於車載隨意網路之網路範圍、車輛行進速度、車輛之相關地理位置，及其車輛之間連接的分散性，使得車用隨意網路通訊的安全問題隨之而生，特別是在車用隨意網路中，當行進車輛間有突發的緊急事件發生時，要如何快速以及有效地驗證此一緊急事件，是個值得探討的議題。

本計畫為總計畫「異質網路環境之行動搜尋關鍵技術」之第一子計畫，主要的目的為針對車用隨意網路於上述安全性質服務、非安全性服務、以及通訊安全於所面臨的議題，提供有效之存取控制與連結機制。

二、研究目的

本計畫主要的目標為提供車用隨意網路之存取控制與連結機制。達成這個目標，我們將針對車用隨意網路於安全性質服務、非安全性服務、以及通訊安全三個部分進行研究，各部分的研究目的詳述如下：

(1) 安全性質服務部分

- 車用隨意網路安全訊息廣播機制

安全訊息的傳送必需以廣播導向(broadcast oriented)的方式來設計。然而透過無線媒介進行廣播，經常會造成嚴重的干擾，甚至導致廣播風暴的現像(broadcast storming)，所以在設計上必需利用額外的資訊來降低這樣的問題。常用的方法是透過車輛上的衛星定位裝置(global positioning system, GPS)所取得的地理資訊來限制轉送的區域。如文獻[27][29]，在每一個安全訊息中夾帶傳送車輛的位置資訊，收端可依據和送端的距離決定傳送的優先順序，以避免同時轉送所帶來的干擾。而文獻[28][29]則是用送端相對收端的方向，來決定是否要進行傳送。文獻[30]更進一步利用地理資訊，估算出可能的威脅程度，使處在較高危險區域的車輛優先轉送。雖然現有的方法已利用地理資訊來減少過多轉送帶來的干擾，但是由於傳送功率是固定的，因此每一次傳送所造成的干擾範圍仍是無法降低。有鑑於此，在提出的方法中，我們利用功率控制的方式來真正降低干擾範圍。在此同時，我們利用地理資訊所估算的安全距離限止最低的功率，以確保所有潛在受到威脅的車輛都能收到安全訊息。此外，考慮到 VANET 網路高度的變動性與有限的頻寬，我們所提備的機制完全不需要任何拓撲的資訊與週期性資料的交換。

- 車輛密度為基礎的緊急訊息廣播機制

緊急訊息的廣播，可分為發送端導向(sender-oriented)與接收端導向(receiver-oriented)兩種架構。在發送端導向中[31-36]，傳送端利用正確的鄰居位置來選擇下一個轉送者。由於周邊節點的位置資訊是可取得的，每一個傳送端節點可直接選擇距離自己最遠的鄰居來減少轉送的點。另一方面，由於下一個轉送點已明確地被發送端指定，其它不是轉送點的節點不會參與競爭，因此訊息可在收到後直接地被送出，不需額外的等待時間。但這個架構必需仰賴高頻率且周期的控制訊息來即時地更新位置資訊，才能正確的選擇最遠的轉送點，因此有較高的控制負載(control overhead)，若是資料無法即時更新，更可能會選擇到一個已經脫離涵蓋範圍的車輛作為轉送點，而造成轉送失敗。而在接收端向導向的架構中[37-40]，轉送點是不會事先由發送端決定好，而是交由接收端自行去競爭傳送權，以選出下一個轉送點。然而，在稀疏的網路(sparse network)中，倒退時間可能會有過長的情況，而增加延遲時間。此外，在密集的網路(dense network)中，也可能會有數個節點計算出相近倒退時間的情況，造成較嚴重的干擾問題。為了解決上述問題，我們提出了一個以車輛密度為基礎的緊急訊息廣播機制。這個架構主要的概念是透過預估周邊車輛的密度，來減少參與競爭轉送者的數量，並同時降低轉送時所需的等待時間。這個架構也只需要有較少的控制負載，就能達成穩定的運作。

(2) 非安全性質服部分

- WAVE/DRSC MLFF 高速公路電子收費系統

在 ETC 整個系統架構，分成三個模組：扣款模組，執法模組，還有後端模組。在扣款模組方面，必需透過通訊技術完成扣款交易，這部分的技術主要是以專用短距通信技術(Dedicated Short Range Communication, DSRC)為主[9]，如 RFID、Infrared 等技術都可作為 DSRC 的媒介，然而各有其缺點，例如目前台灣所使用的 ETC 系統就是使用 Infrared 為主的 DSRC，由於 Infrared 的技術是壟斷的，每年必需支付給國外一筆龐大的權利金。相對的以車用環境無線取技術(Wireless access for vehicular environment, WAVE) [41]為主的 DSRC，由於國際標準已成立，將來將更具有前瞻性。此外 WAVE/DSRC 提供了穩定的多重通道技術，更適合處理 MLFF 這種高車流密度的網路環境。因此我們以 WAVE/DSRS 作為 MLFF ETC 扣款模組的通訊技術，並設法利用多重通道的特性，提升模組的能力與穩定性。目前國外的 MLFF 大部分都是讓每個車道都處於一個天線的涵蓋範圍，車輛進入這個範圍，執行扣款，這樣的方法是將每個天線都對應到某一個車道，會將接受的訊息跟後端結合去作比對，這樣的方法，車子在發生交易的範圍如果突然切換車道，可能會因為沒有即時更換頻道而導致交易失誤，此外要如何準確的將通道都準確的對應到某個車道，而且相鄰兩個車道也會有交疊的範圍。因此，我們將提出以數個全向式天線涵蓋完整的交易區域的多重通道架構，可以避免追縱車道所產生的問題。

- 行動閘道式車用隨意網路繞徑機制

至目前為止，已有許多使用固定式架構改善封包遞送的繞徑協定被提出。在 MPARP [42中，每台車輛備有一個 IEEE 802.11 及 IEEE 802.16 的網卡。當路徑存在時，車輛可直接與其它車輛使用 IEEE 802.11 模式溝通，若不在在，由基地台使用 IEEE 802.16 模式接手。在 RAR [43]以及 DDR [44]中，每個路段由兩台 RSUs 所包含，當一車輛要傳送到另一路段的車輛時，它先傳送封包到他所在路段的 RUS，之後封包會經由骨幹網路轉送到目的車輛。相同的，[45]所提出的繞徑是基於 RSU 所給的資訊來決定。雖然車用通訊可由 V2I 模式輔助，但有效範圍僅限於 RSU 存在的地區。因此在 MIBR 協定[46]中，提出了行動式閘道器(Mobile Gateway)的概念，以特定台車輛(如公車)可作為閘道器來轉送封包，由於特定車輛可裝載較大傳送範圍的天線(超過 300 公尺)，因此可提升遞送率與輸出。但是這些車輛的連通性將受限於行程和時段的限制。為了克服這個問題上述的限制，我們利用了部分的車輛作為行動閘道器車輛，所裝備的車上單元可透過 3G 或 IEEE 802.11 介面卡轉送資料封包。其它沒有 3G 介面卡的車輛可透過無線網路轉送封包到行動閘道器車輛，再使用 3G 介面卡轉送封包到閘道器控制器。最後，閘道器控制器將過鄰近目的地車輛的行動閘道器車輛轉送封包，可明顯地減少跳躍次數和連斷線的機率。

(3) 車用隨意網路通訊安全

- 車用隨意網路安全通訊金鑰協議協定

我們針對全安威脅以及私密議題，提出了一個混雜式對應基礎金鑰協議協定(Chaotic Maps-based Key Agreement Protocol)。這個協定不只可以解決現有的協定的弱點，同時可保留使用者的匿名性(Anonymity)。這個協定主要包含了下列的優點: (1) 可達到伺服器與使用者之間的人工認證; (2) 允許使用者匿名地與伺服器協區段金鑰; (3) 伺服器與使用者之間可產生保護接下來通訊的區段金鑰。此外，我們利用Petri nets [47]來驗証在攻擊者知道部分安全內容的情況下它進一步知道什麼。我分析解過指出所提出的協定可成功地防禦重覆攻擊(Replay Attacks)、偽造攻擊(Forgery Attacks)、以及竊取驗証攻擊(Stolen-verifier Attacks)。

- 車用隨意網路安全訊息整合認證機制

在安全訊息驗證方面，Raya 與 Hubaux [48]針對車用隨意網路的安全問題，於 2005 年提出系統性解決方法。隨後，許多用以增進安全、效率，與功能性之相關研究因應而生[49-60]。在 2008 年，Zhu [60]等人針對行車突發緊急事件的驗證程序，提出一個聚集訊息驗證機制。此機制植基於憑證公開金鑰密碼系統(certificate public key cryptography)，故此機制之聚集驗證包括憑證驗證與簽章驗證兩部份。為了有效地簡化傳統公開金鑰密碼系統所需之憑證管理，Shamir [61]於 1984 年提出以身份為基礎之公開金鑰密碼系統(ID-based public key cryptography，簡稱 ID-PKC)。此密碼系統中，使用者的公鑰是由使用者的身份識別碼(identity)推導產生。此系統存在一公正第三者(trusted third party，簡稱 TTP)，亦即私鑰產生者(private key generator，簡稱 PKG)，用以協助使用者產生其私鑰。故 ID-PKC 可能會衍生金鑰托管(key escrow)的問題。Al-Riyami 與 Paterson [62]為解決 ID-PKC 的金鑰托管問題，提出免憑證公開金鑰密碼系統(certificateless public key cryptography，簡稱 CL-PKC)的概念，在 CL-PKC 中，存在一公正第三者，亦即金鑰產生中心(key generation center，簡稱 KGC)，協助使用者產生部份使用者私鑰。使用者隨後自行產生一秘密資訊，再與部份使用者私鑰結合，產生完整者的使用者私鑰。因此，KGC 無法得知使用者的完整私鑰，即能有效解決 ID-PKC 之金鑰托管問題，亦同時能降低憑證的使用率。為改善上述問題，我們提出一個安全的聚集訊息驗證機制(secure aggregated message authentication scheme，簡稱 SAMA)。此機制是基於免憑證公開金鑰密碼系統(certificateless public key cryptography，簡稱 CL-PKC)，用以驗證車用隨意網路之突發緊急事件。經由效能的評估，此機制可有效地降低聚集訊息驗證所需之計算量。此外，我們利用 Petri nets 分析提出之機制，並驗證此一方法可有效地抵禦偽造攻擊(forgery attacks)，同時提供車輛之隱私保護(privacy protection)。

三、研究成果

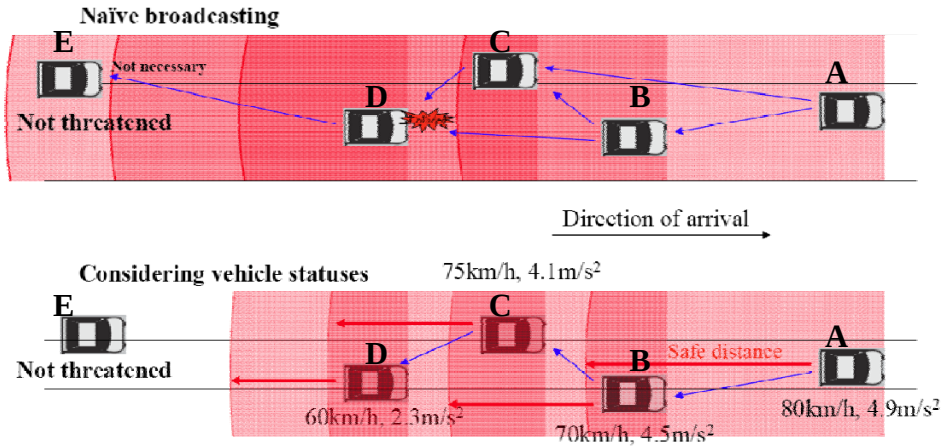
第一年

第一年的研究中，我們針對此網路中重要的安全應用(safety applications)，進行兩項主要的研究 (1)車用隨意網路安全訊息廣播機制; (2) 車用隨意網路安全訊息驗證機制。以下分別述之:

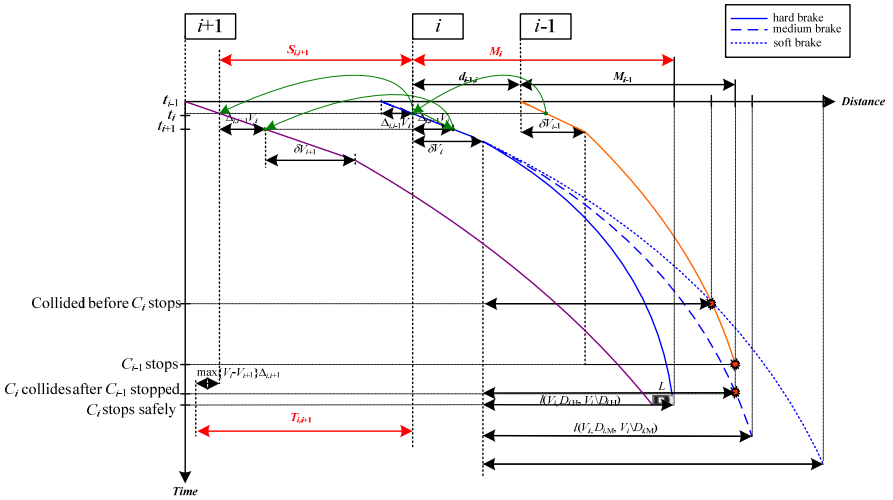
1. 車用隨意網路安全訊息廣播機制

我們提出利用功率控制的方式降低干擾範圍的廣播機制，並利用地理資訊所估算的安全距離限制最低的功率，以確保所有潛在受到威脅的車輛都能收到安全訊息。

此機制的主要概念，是利用功率控制(power control)的方式，來達到減少實體層干擾(physical interference)，以及限制廣播區域(broadcasting area)的效果。而傳送功率的調整是根據車輛與車輛之間的安全距離 (safe distance)。如圖一所示，在沒有功率控制的情況下(上圖)，每台車輛都必需以最大的功率進行傳輸，因此中間兩台車輛 (車輛 B C)將會收到前方車輛(車輛 A)所廣播的安全訊息，而這兩台車輛又必需將此訊息轉送給後方車輛，使得至後方車輛(車輛 D)同時收來相同的訊息，而導致干擾的發生。相較之下，我們的方法只將訊息傳送給少於安全距離內的車輛，較不會有多台車輛同時進行送或收的問題，因此可減少干擾的發生。這同時也能避免不受威脅的車輛收到此訊息的情況，如下圖，最左邊的車輛(車輛 E)已和前面的車陣保持足夠的距離，因此在我們的機制下將不會收到冗餘的訊息。



圖一: 訊息廣播使用功率控與未使用功率控制之比較



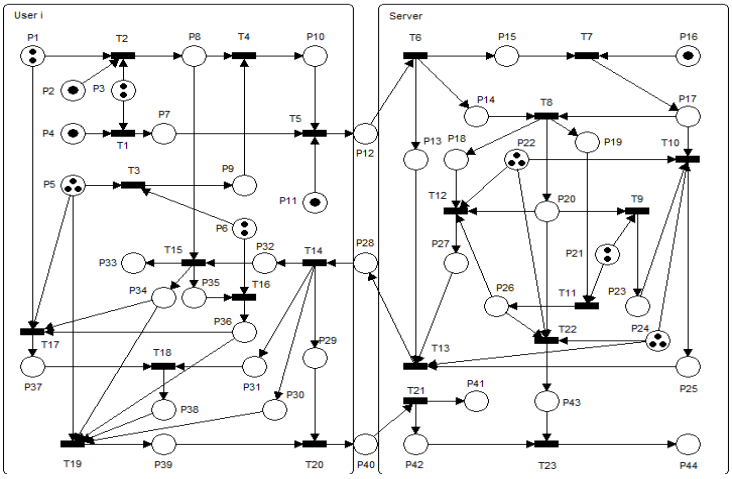
圖二: 車用隨意網路滑行距離、安全距離、及傳輸半徑之估算

如圖二所示，我們估算車輛 i-1 與車輛 i 之間的安全距離，並依此距離推算不同情況下的滑行距離 M_i。接著，根據此滑行距離 M_i，以及後方車輛的煞車力道 D_{i+1}，行車速度 V_{i+1}，和所需求的息訊傳遞時間，估算出在車輛移動下的所需保持的安全距離 $S_{i,i+1}$。最後，根據此安全距離 $S_{i,i+1}$，以及收到訊息前的車距變動，算出對車輛 i+1 所需要的最小傳輸半徑 $T_{i,i+1}$。

此外，考慮到 VANET 網路高度的變動性與有限的頻寬，我們所提出的機制完全不需要任何拓撲的資訊與週期性資料的交換

2. 車用隨意網路安全通訊金鑰協議協定

我們提出了一個混雜式對應基礎金鑰協議協定(Chaotic Maps-based Key Agreement Protocol)。這個協定不需任何檢查表就可在伺服器與使用者之間達到人工認證以及區段金鑰協議(session key agreement)的能力。不同於過去的金協議協定，伺服器不再需要任何的驗證表，而會先產生系統參數，包含Chebyshev polynomials$E(\cdot)$、$D(\cdot)$、以及$H(\cdot)$。假設一個新的使用者Ui 想要和伺服器建立通訊的區段金鑰。Ui會隨機選擇它的密碼PWi，並且透過一個安全的通道私下傳送 $(ID_i,H(PW_i))$到伺服器，當收到這個訊息後，伺服器將IDi和$H(PW_i)$由左到右並列到後面的訊息，並使單向雜湊函數$H(\cdot)$計算$H(IDi,H(PWi))$。之後，伺服器可計算出$Regi = H(IDi,H(PWi)) \oplus H(Ks)$，其中$Ks$是伺服器的私密金鑰，並透過一個安全的通道傳送$Regi$ 回Ui。



圖三: 所提出金鑰協議協定的Petri網路模型

我們使用如圖三的Petri nets來模組化及分析提出的協定，以驗證攻擊者是否能得到內容。結果如表一所示，我們的方法可有效抵抗重覆攻擊、偽造攻擊、以及竊取驗證攻擊，並且擁有相互驗證、使用者匿名、以及known-key安全性等優點。

表一: 安全性質比較

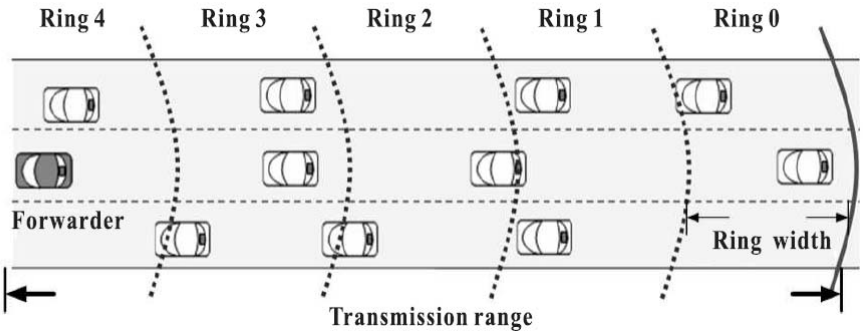
	Xiao et al.'s protocol [11]	**Yoon & Yoo's protocol [13]**	**Proposed protocol**
Replay attacks	Insecure	Secure	Secure
Forgery attacks	Insecure	Secure	Secure
Stolen-verifier attacks	Insecure	Insecure	Secure
Mutual authentication	Not provide	Provide	Provide
User anonymity	Not provide	Not provide	Provide
Known-key security	Provide	Provide	Provide

第二年

在第二年的研究中，我們兩項主要的研究: (1)車輛密度為基礎的緊急訊息廣播機制;(2)WAVE/DRSC MLFF 多高速公路電子收費系統。以下分別述之

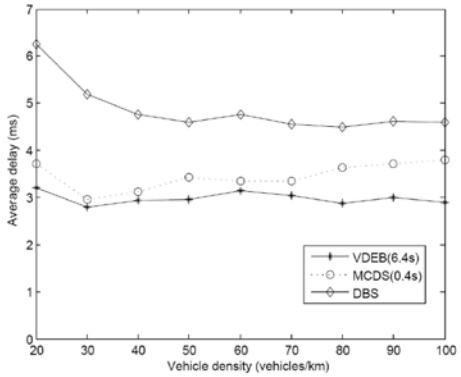
1. 車輛密度為基礎的緊急訊息廣播機制

為了解決接收端導向架構在稀疏網路下倒退時間過長的問題，我們提出一個以車輛密度為基礎的緊急訊息廣播機制(Vehicle-density-based emergency message broadcast, VDEB)，些機制主要採用接收端導向的架構，然而它同時結合了傳送端導向的優點，並設法消除兩種架構的缺點。VDEB將傳輸範圍切成數以送端點為圓心的環狀區帶，兩個環狀區帶所的間隔距離稱作環寬度(ring width)，環寬度是由發送端根據周圍車輛的密度所決定，發送端一旦決定了環寬度，就會將些訊息包含在廣播訊息中，並廣播給周邊的車輛，收到這個訊息的節點則會根據此環寬度，決定自己的倒退時間。如圖四所示，傳送範圍被均分為5個環狀區帶，每一個區帶會被指定一個時槽(time slot)，倒退時間則是隨著環狀區帶編號的增加而增加，因此處在最外圍區帶(ring 0)的車輛擁有最小的倒退時間。如果在ring 0時槽內沒有訊息被成功傳送，在ring 1區帶內的車輛則會進行傳送。相反的，如果任何車輛成功傳送了訊息，其它的車輛則會取消自己的倒退程序。



圖四: VDEB所切分的環狀區帶

我們透過車輛位的預測估計當前周圍鄰居的實際數量，傳送端則可依據此資訊決定最適當的環寬度。最大的環寬度出現在所有車輛都並排地行駛，而實際的環寬度，可從最小與最大的區間選取來隨機設定。當收到此環寬度的資訊，接收端可求出所有環狀區帶的數量，並根據本身到傳送端之間的距離，計算出所在的區帶，倒退時間(*BackoffTime*)則是以與區帶編號為等比求得。如圖五的實驗數據顯示，VBED整體所產生的平均延遲時間，比MCDS以及DBS兩種接收端導向的作法又更低。



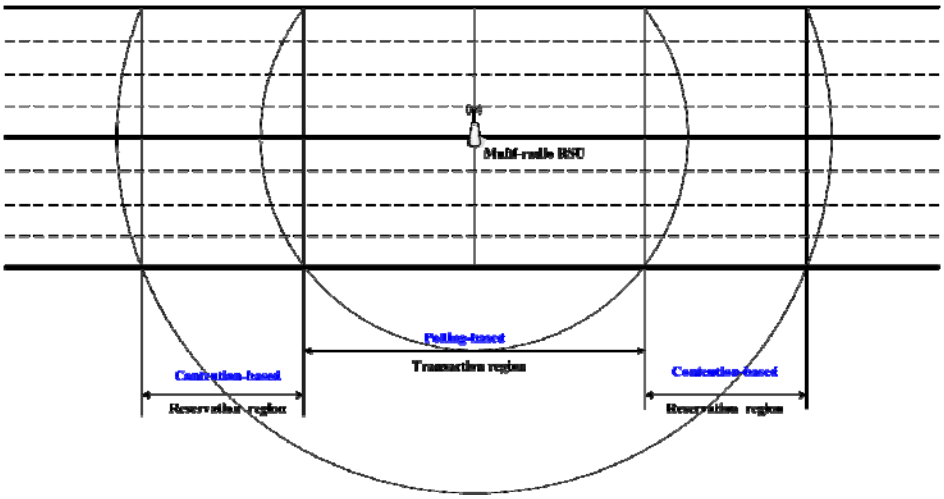
圖五: 平均延遲時間

2. WAVE/DRSC MLFF 高速公路電子收費系統

美國FCC在5.8GHz的頻帶上分配了75MHZ的頻寬作法WAVE的通訊。IEEE 802.11p則將此頻寬切分為七個相等頻寬通道(10MHz)，包含一個控制通道(control channel, CCH)和六個服務通道(service channels, SCHs)。CCH 區間時段下的SCHs是閒置的，這是因為現在的車用無線通訊裝置，基於成本的考量，大部分只使用單一的天線(single-radio)，因此無法同時利用CCH和SCHs的頻寬。然而相對於一般車輛，ETC收費站屬於大型公有建設，較無成本的限制。

因此我們所提出的ETC架構，主要是利用具有多重天線(multi-radio)的RSU作為收費站的通訊裝置，以利用CCH 區間時段下的SCHs頻寬作為ETC交易使用，這樣的好處是可以不被現有的其它WAVE服務所影響。Multi-radio RSU擁有與頻通同等數量的天線，因此可以同時監控所有的通道，並協調不同通道之間的交易，一般的車輛則仍可用單一的天線來降低成本。

如圖六，本架構將multi-radio RSU的傳送範圍切分為登入區(Reservation region)和交易區(transaction region)內外兩個層次，車輛進入reservation region時，以競爭基礎(contention-based)的方式向RSU保留一個SCH，而在進入transaction region時，則可以輪詢基礎(polling-based)的方式，在事先保留的SCH上跟RSU進行交易，以避免其它訊號的干擾，保證車輛在經過ETC收費站時能順利的完成交易。



圖六: Multi-radio RUS 的區域切分

當車輛進入到交易區，則會停止一切在CCH區間時的傳送，直到收到RSU的輪詢要求，才會與RSU在預約好的SCH上進行交易。一旦車輛完成交易或者駛離交易區，預約的SCH會再度被釋放出來，以供後面的車輛使用。除此之外，由於車輛進入到交易區後，會在CCH區間切換到不同的SCH，因此RSU必需負責收集這段時間在CCH上的訊息，並在CCH區間結束前彙整給所有在交易區內的車輛。

利用上述的流程，交易的負載可分散至不同的服務通道，避免單一控制通道的擁塞問題。區分登入區與交易區的架構，更可確保車輛在通過入收費站時，不會被其它的傳輸所干擾，以提升系統的可靠度。此外車輛在進行ETC交易的同時，仍可透過Multi-radio RSU的協助，來保證其它安全訊息的接收。因此我們預計這個架構，將可高度整合於未來的車用網路環境。

第三年

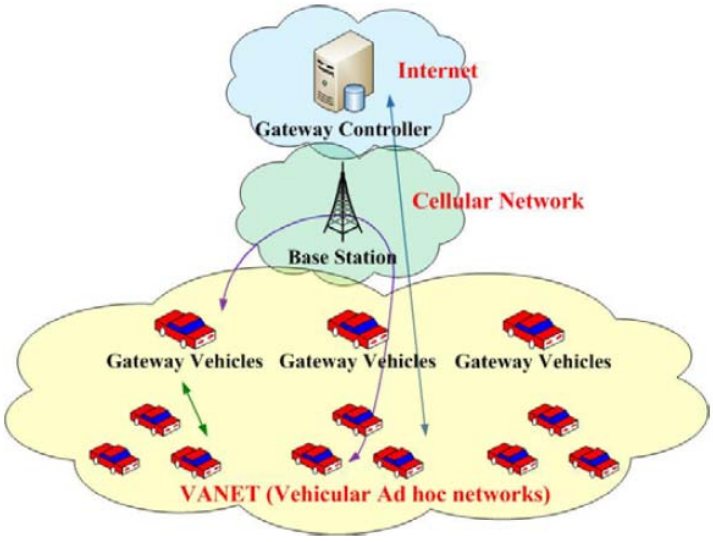
在第三年的研究中，我們兩項主要的研究: (1)車行動閘道式車用隨意網路繞徑機制;(2)車用隨意網路安全訊息整合認證機制。以下分別述之

1. 行動閘道式車用隨意網路繞徑機制

我們提出了一個以位置為基礎的繞徑協定，簡稱行動閘道式繞徑協定(Mobile-gateway routing protocol (MGRP)。MGRP 結合了車輛對車輛(V2V)以及車輛對路側(V2I)兩種通訊方式，並且利用一部分的車輛作為行動閘道器，以延伸固定式路側單元(RSU)的涵蓋範圍。每一個行動閘道器車輛上的車上單元(OBU)裝配有一張 IEEE 802.11 通訊介面卡以及一張 3G 網路介面卡。其它不具有 3G 通訊介面卡的車輛則以 IEEE 802.11 無線連結傳送封包到最近的行動閘道器。當收到一個封包時，行動閘道器再以 3G 通訊介面將此封包轉送到一個後端的閘道控制器(Gateway Controller)。這個閘道控制器會搜尋目標地置，並且決定一組與目的地相鄰的行動閘道器，最後將封包透過這行動閘道器轉送到目的地車輛去。

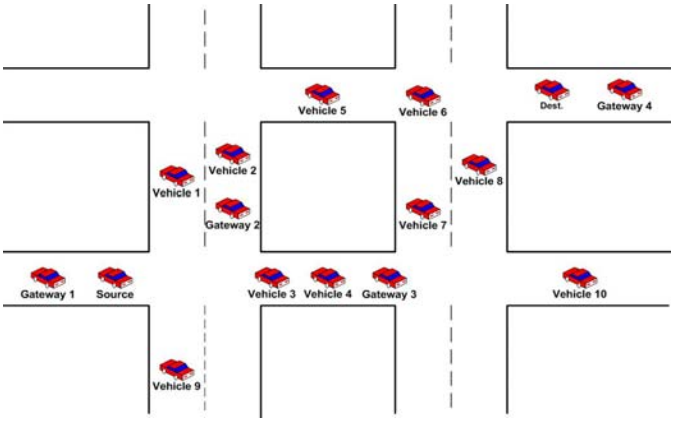
如圖七所示，我們利用部分車輛作為行動閘道器來取代傳統的路側單元。行每一個動閘道器上的車上單元閘道器裝備一張 IEEE 802.11 介面卡 以及一張 3G 介面卡。其中，3G 介面卡是用來與 cellular 網路的基地台(Base Station, BS)通訊，而 IEEE 802.11 介面卡則是用來與其它沒有 3G 介面卡的車輛通訊。當基地台接收到的行動閘道器車輛送來的資料封包，它將遞這個封包到一個閘道器控制器(Gateway Controller)。然後，這個閘道器控制器搜尋目的地車輛所在的位置，決定一組與目的車輛鄰近的行動閘道器車輛，並透過基地台傳送封包到每一個相鄰的行動閘道器車輛。最後，這些閘道器車輛將使用 IEEE 802.11 連節傳送資料封包到目的地車輛.

我們假設每一台車輛可透過全球衛星定位系統(Global Positioning system, GPS)獲得自己的位置，車速，以及行車方向。這些資訊 將週期性透過一個 Hello 訊息廣播到傳送範圍內鄰近的車輛。我們也假設每一台車輛將有數位地圖，以取得道路的交通流量狀況。此外，如果一台車輛 發現某路徑已中斷線，它將暫存任何接收到的資料封包，並傳送一個 RRER 封包到來源車輛以供選擇替代路徑之用。不同於傳統的繞徑協定，MGRP 限制存活時間值(Time-to life, TTL)在三次跳躍(hop)內.



圖七: 行動閘道式車用隨意網路架構

現在，我們描述其細部程序。類似於 AODV 協定，當一台車輛需要傳送封包時，它首先廣播一個 RREQ 封包到鄰近的車輛。當一個鄰居接收到此 RREQ 封包，如果它沒有可以到目的地車輛的繞徑路徑，它將重新廣播此路徑需求到其它鄰居。不同於 AODV，我們限制 RREQ 的 TTL 值在三次跳躍內。當一台車輛 接收這個 RREQ，他首先檢查跳躍次數是否仍然少於三次: 如果是，則這台車輛 將變成下一個轉送者來廣播此 RREQ 封包; 如果不是，則此車輛將丟棄此 RREQ 封包。如果目的地車輛的資訊存於繞徑表格中或行動閘道器車輛接收這個 RREQ 封包，它將傳送回一個 RREP 封包到來源車輛。更進一步，如果此車輛等待了一段時間而且沒有收到此 RREP 封包，它將重新廣播一個 RREQ 封包，並重覆上述程序。



圖八: 來源與目地車輛傳送通封包情境

在 MGRP 中，每一個車輛可行動閘道器遞送資料封包以減少傳送跳躍的次數，同時達到更可靠的通訊品質 quality。圖八顯示一台來源車輛 Source 如何(於圖左側)傳送封包到一台目的地車輛 Destination (於圖右側)。當來源車輛有封包要送給目的地車輛，它首先搜尋一台鄰近於自己的行動閘道器車輛，也就是 Gateway 1，以及傳送封包到此閘道器車輛。然後，Gateway 1 使用 3G 介面卡轉送此資料封包到一個基地台。當收到此封包，基地台遞送這個封包到一個後端的閘道器控制器，以搜尋目的地車輛所在的位置，以及傳送此封包到一組鄰近於目的地車輛的行動閘道器車輛，也就是 Gateway 4。最後，Gateway 4 將透過 IEEE 802.11 介面卡轉送封包到目的地車輛。若沒有 Gateway 1、Gateway 4、以及閘道器控制器的輔助，來源車輛將必需攜帶此封包直到遇到 Vehicle 1、Vehicle 3、或 Vehicle 9。相同的問題也會發生在下一台車輛攜帶這此封包的車輛上。在成功抵達目的地車輛前，上述的遞送程序可能導至較長的延遲時間。更糟的是，如果封包被轉送到 Vehicle 9，而沒有其它車輛與 Vehicle 9 連結的車輛可以傳送目的地，這個封包將會遺失，而導致不可靠的傳送。

當來源車輛廣播了 RREQ 封包去要求一個繞徑路徑後，會有三種可能的狀況發生。

- 狀況一: 除了來源車輛，其它車輛無法利即找到一台鄰近的車輛。在這個狀況下，此車輛將攜帶封包直到另一台出現在自己的傳送範圍為此。之後，它 將轉送此封包到那台車輛。

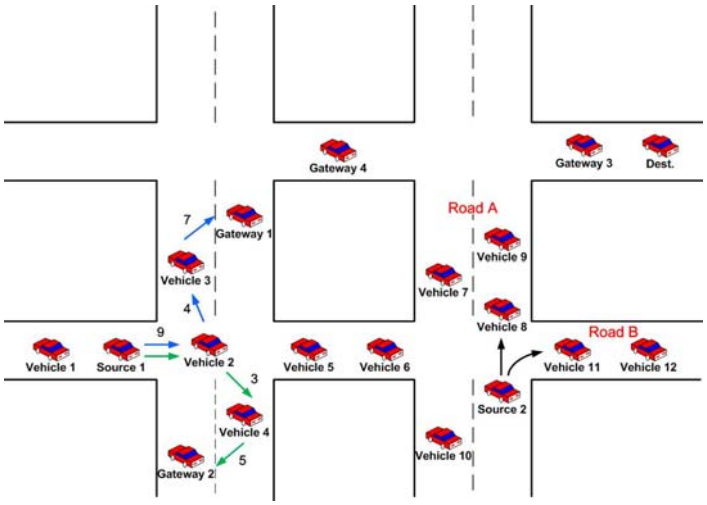
- 狀況二: 存在一台以上的鄰近車輛，但它們之中不存在一台可以在三次跳躍內到達目的地路徑的行動閘道器。在這個狀況下，來源車輛將根據各道路的車密度資訊(Road density)，為此封包決定一個轉送的方向。當此車輛位處於交叉路口時，它將選擇車密度較高的方的道路方向。如圖八顯示，那裡不存在任何繞徑路徑可從 Source 2 將資料

封包轉送到一台閘道器車輛或目的地車輛，而且 Road A 的車密度 (3 台車輛)高於 Road B (2 台車輛)。因此，Source 2 將轉送封包到 Road A 上的 Vehicle 8。此方法 可改進封包的遞送率(delivery ratio)，因為在每一修道路上的行動閘道器車輛與一般車的比率沒有明顯的差異的情況下，一個較高的道路車密度經常隱含了較高的機率可找到一台行動閘道器。

- 狀況三: 存在一條上的繞徑路徑可轉送封包到目的地或行動閘道器車輛。在這個狀況下，來源車輛需要選擇一個合適的繞徑路徑。MGRP 將選擇最可靠的路徑來轉送資料封包。路徑的可靠度是由繞徑的存活時間(routing lifetime)來評估。我們利用文獻[27]所提的連節存活時間(link lifetime)公式(如下)來預測車間連線的存活時間，其中 R 是每輛車的傳送範圍，D_{ij} 是車輛 i 以及車輛 j 的距離，V_i 是車輛 i 的車速，以及 V_j 車輛 j 的車速。繞徑的存活時間則是所有在些路徑上最小的連節存活時間:

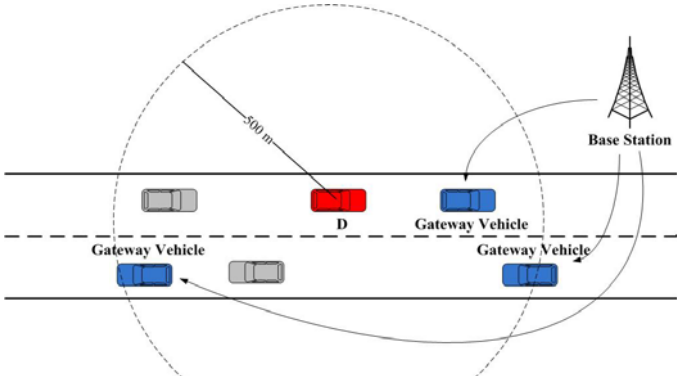
$$Link\_lifetime = \frac{R - D_{ij}}{V_i - V_j}$$

如圖九所示，存在 2 條路徑可從 Source 1 將轉送資料封包傳送到一台行動閘道器車輛上。第一修路徑是經由 Source 1→Vehicle 2→Vehicle 3→Gateway 1，以及第二路徑是經由 Source 1→Vehicle 2→Vehicle 4→Gateway 2。所以在第一條路徑的連節存活時間分別是 9s，4s，以及 7s，第二條路徑的連節存活時間則分別是 9s，3s 以及 5s。因此，第一以及第二條繞徑的存活時間分別是 4s 以及 3s。由此結果，Source 1 將選擇第一條路徑來轉送資料封包，因為他具有較長的路徑存活時間。要注意的是，當繞徑表格內所記綠到路徑的優先權(priority)比所選擇的路徑還高時，則選擇表格內的。



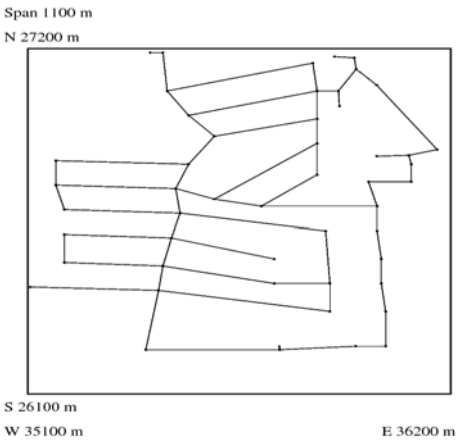
圖九: 封包傳送不同的情境

當資料封包被轉送到一個閘道器車輛後，此閘道器車輛將透過 3G 網路轉送封包到基地台以及閘道器控制器。閘道器控制器將選取鄰近目的地車輛的一組閘道器車輛作為轉送者。閘道器控制器會週期性地更新閘道器車輛所在的位置。閘道器控制器會依劇閘道器以及的地 車輛是否少於 500 公尺作為選取的考量。任何距離少於 500 公尺的閘道器車輛都會被選擇為轉送。然後封包將以 V2V 的型試遞送到目的地車輛，如此可以改進轉送資料封包到目的地車輛的成功率。然而，如果不存在閘道器車輛能轉送資料封包，則閘道器控制器將丟棄此資料封包。如圖十所示，三個閘道器車輛將接收來自於基地台的資料封包。



圖十: 閘道控制器轉送封包到一組少距離目地車輛於 500 公尺的閘道器車輛

接下來，我們使用 ns2 simulator (version 2.34) 評估 MGRP 的效能表現。我們比較 MGRP 以及傳統位置-基礎的繞徑協定 GPSR。並分析閘道器 車輛所佔的比率以及成功地遞送率之間的關係。



圖十一: 模擬街道圖

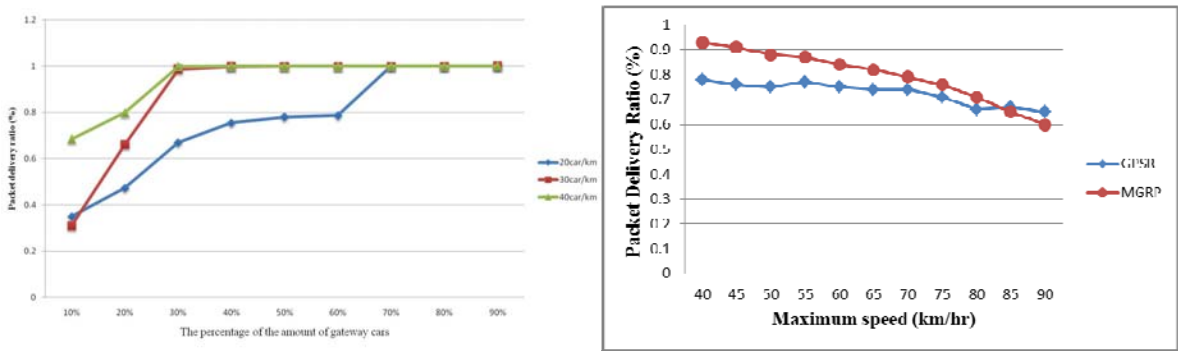
我們由 TIGER 資料庫(*Topologically Integrated Geographic Encoding 以及 Reference System*)。所取得的真實街道地圖進行模擬。測試 MGRP 在高速公路(highway)以及市區(urban)兩種情境。如圖五所示，市區街道是配置在一個 1100m*1100m 區域，包含 61 條道路以及 150 輛車。我們以 10 CBR 流量以及 512-byte 的封包大小作測試。表二詳列其它的模擬參數。

表二: 模擬參數

Parameter	Value
模擬 情境	高速公路/ Urban
速度 of 車輛 s	40–90 km/h
模擬 time	300 sec
Interval time of 資料遞送 y	0.5 sec
資料 封包 size	500 bytes
傳送 範圍	250 m

我們首先針對一條長達 4km 並且擁有 4 車道且雙向的高速公路情況進行評估，比較閘

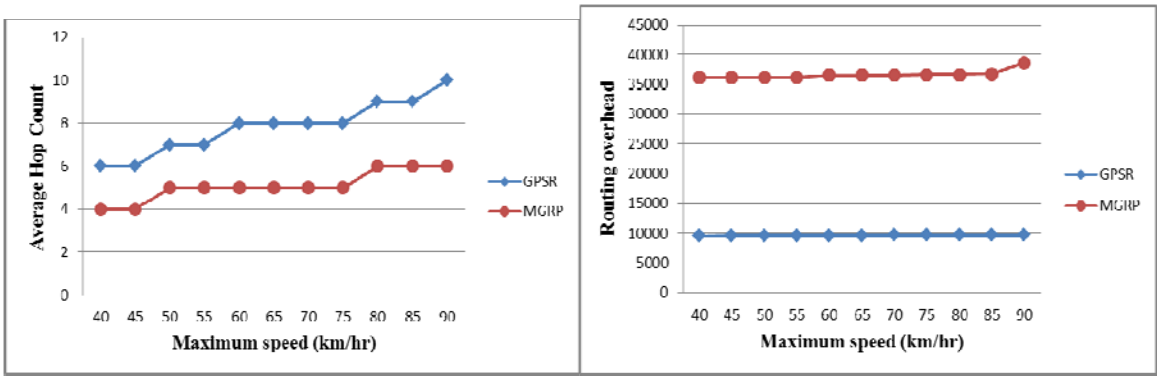
道器車輛比率以及封包遞送率之間的關係。模擬結果顯示在圖十二(a)，我們可看到當愈多的車輛為作行動閘道器時，封包遞送率明顯增加。即使在低車流密度情境下(20 車輛/km)，MGRP 可達 80%的封包遞送率，只要閘道器車輛超過 60%。反之，當閘道器車輛的比率為 10%時，封包遞送下降至 38%。此外，在低車流密度情境下，MGRP 需要至少 70%的閘道器車輛到達到 100%的遞送率。另一方面，結果顯示 MGRP 在中度(30 車輛 s/km) 以及高度(40 車輛 s/km)車密度情境下表現的更好。在這些狀態下，the MGRP 僅需要 30%的閘道器車輛就能達到近乎 100%的遞送率。也就是說，如果有 10 輛車在高速公路上，使用我們的協定僅需要 3 台車輛為行動閘道器，因此在這個網路架構下不需要過多的成本。



圖十二: (a) 閘道器車輛與封包遞送率關係 (b) 最大車速與封包遞送率關係

現在，我們比較 MGRP 以及 GPSR 的表現。圖十二(b)顯示封包遞送率與最高車速度之間的關係，我們可看到雖然兩個協定的封包遞送率都隨著車速的增加而減少，我們的協定仍然表現的較好，因為 MGRP 利用了 3G 網路以及閘道器控制器來輔助封包轉送，使得因高速行種而造成的連節斷線可以被大量的避免。要注意的是，MGRP 的封包遞送率會在最大車速超過 85 km/h 時低於 GPSR，這是因為在高車速情境下 MGRP 必需經常維持繞徑表格，以至於它可能增加封包丟失的機會。

圖十三(a)顯示平均跳躍次數與最大車速之間的關系。結果顯示當車速上升時，不論是 MGRP 還是 GPSR，平均跳躍次數都會增加，但 MGRP 可以維持在 6 以內的最大跳躍次數，而 GPSR 則是 10 次，這因為我們使用 3G 網路來減少傳送 跳躍。此外，我們限制了來源車輛尋找閘道器或目的地車輛的跳躍次數。

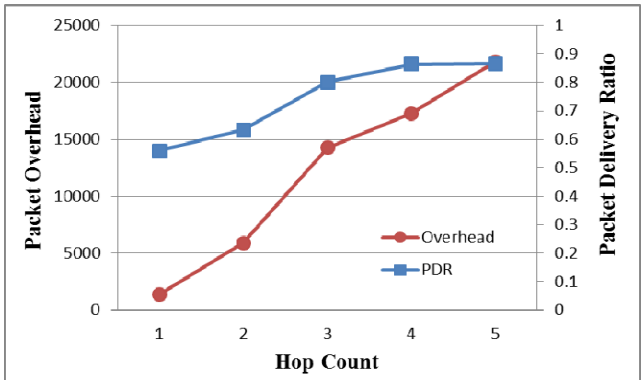


圖十三: (a) 平均跳躍次數與最大車速 (b) 路由成本與最大車速

圖十三(b)顯示繞徑成本(Overhead)與最大車速之間的關係。結果顯示，不論是 MGRP 還是 GPSR，封包成本都會隨著車速增加而增加，然而 MGRP 比 GPSR 有更高的封包成本，這是因為我們需要維持繞徑表格。然而，透過繞徑表格的建意，我們可避 GPSR 的區域最

大問題。更進一步，我們的方法可減少到目的地節點全部的跳躍次數。

圖十四顯示繞徑成本(不包含 Hello 訊息)以及封包遞送率與跳躍次數之間的關係。我們測試了 45km/h、65km/h、以及 85km/h 的平均車速。結果顯示當跳躍次數增加時，封包遞送率以及成本增加，這是因為當跳躍次數增加時，我們的協定有更高的成功地率能轉送封包到閘道器車輛，但同時也增加了封包成本。

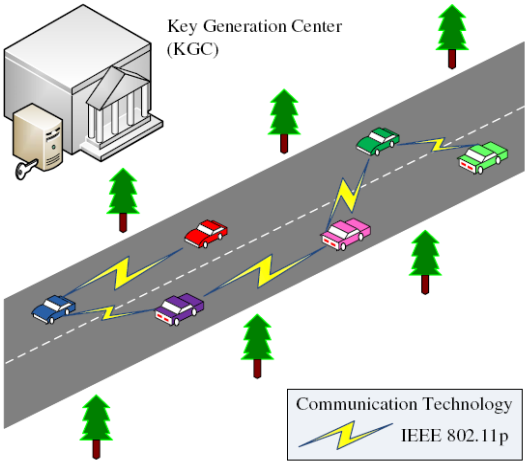


圖十四: 封包遞送率、路由成本、與跳躍次數

在第三年的計畫中，我們針對車用隨意網路，提出了一個以位置為基礎的繞徑協定，稱作行動閘道器繞徑協定(MGRP)。我們利用了部分的車輛作為行動閘道器車輛，所裝備的車上單元可透過 3G 或 IEEE 802.11 介面卡轉送資料封包。其它沒有 3G 介面卡的車輛可透過無線網路轉送封包到行動閘道器車輛，再使用 3G 介面卡轉送封包到閘道器控制器。最後，閘道器控制器將過鄰近目的地車輛的行動閘道器車輛轉送封包。我們設計的繞徑協定適用於這樣的混合式網路架構，並明顯地減少跳躍次數和連斷線的機率。模擬結果 MGRP 確實表現的比傳統位置-基礎繞的徑協定 GPSR 要更好。

2. 車用隨意網路安全訊息整合認證機制

SAMA 完整的系統架構如圖所示，其運作於車用隨意網路之車輛的車間通訊 IVC，並假設是在無固定路邊設備的協助下進行資料的傳遞。車間媒介層通訊部分可適用基於定義在 IEEE 802.11p 使用 5.9 GHz 頻段的專用短距通訊技術 DSRC。本機制存在一 KGC，用以建置系統初始參數的設定與協助車輛產生部份私鑰。



圖十五: SAMA 機制的系統架構

在系統需求部分，包含訊息認證(message authentication)、有條件式隱私保護(conditional privacy preservation)、和可追縱性(traceability)，此機制主要分為下列四個階段：

系統設置階段(System Setup)

KGC 產生與公開系統參數，以及定義了緊急安全報告(security emergency report, SER)格式。 除此之外，SER 的格式也是由 KGC 定義，對一件緊急事件 *Ei*，車輛 *Vj 產生一個如下的 SERi:*

$$SER_j^i = (Type_i, Loc_i, PID_j, Time_j^i, Sig_j^i, PK_j)$$

對特定的事件E_i，相關的SERs將共享相同的$Type_i$ 以及 Loc_i。詳細的定義如表三所列

表三: SER符號定義

Symbol	Definition
KGC	A key generation center
$\mathcal{V}_j$	The j-th vehicle
ID_j	A real-identity of the vehicle $\mathcal{V}_j$
PID_j	A pseudo-identity of the vehicle $\mathcal{V}_j$
$\mathbb{G}_1$	A cyclic additive group
$\mathbb{G}_2$	A cyclic multiplicative group
q	The order of the groups $\mathbb{G}_1$ and $\mathbb{G}_2$
e	$e : \mathbb{G}_1 \times \mathbb{G}_1 \to \mathbb{G}_2$
s	A master key of the KGC
P_{pub}	A public key of the KGC
(x_j, D_j)	A private key of the vehicle $\mathcal{V}_j$
PK_j	A public key of the vehicle $\mathcal{V}_j$
$\mathcal{E}_i$	The emergency event i
SER_j^i	The secure emergency report generated by the vehicle $\mathcal{V}_j$ for the emergency event $\mathcal{E}_i$
$Type_i$	The type of the emergency event $\mathcal{E}_i$
Loc_i	The location where the emergency event $\mathcal{E}_i$ takes place
$Time_j^i$	The time when the vehicle $\mathcal{V}_j$ makes the report on the emergency event $\mathcal{E}_i$
Sig_j^i	The signature generated by the vehicle $\mathcal{V}_j$ on the emergency event $\mathcal{E}_i$
$Enc(\cdot)$	A secure symmetric encryption algorithm (13)
$Dec(\cdot)$	A secure symmetric decryption algorithm (13)
$H_1(\cdot)$	A hash function such as $H_1 : \{0,1\}^* \to \mathbb{G}_1$
$H_2(\cdot)$	A hash function such as $H_2 : \{0,1\}^* \to \mathbb{G}_1$
$H_3(\cdot)$	A hash function such as $H_3 : \{0,1\}^* \to \mathbb{Z}_q^*$
$\|$	Message concatenation operation

註冊階段(Registration)

車輛透過向 KGC 註冊，以取得部份私鑰，隨後自行選取一秘密資訊，與其部份私鑰結合，以產生完整的車輛私鑰，同時車輛亦自行計算其相對應之公鑰。

1. V_j 經由一個存的安全通道傳送 ID_j 到KGC;
2. 當收到ID_j, KGC 先檢查他自已，如果ID_j 是有效的(valid), KGC 會使用master key s 去加密ID_j 到一個虛擬的PID_j，如下:

$$PID_j = Enc_s(ID_j)$$

3. KGC產生如下部分的私密金鑰D_j，其中$Q_j = H_1(PID_j)$:

$$D_j = sQ_j$$

4. KGC 經由一個安全通道傳送虛擬PID_j 和部分私密金鑰D_j 回V_j;
5. 在收到PID_j和D_j後，V_j 選擇一個隨機數字$x_j \square Z_q$, 設定他的完整私密金鑰成$(x_j ; D_j)$，並且計算如下的共開金鑰PK_j:

$$PK_j = x_j P$$

簽章產生階段(Signature Generation)

當一個緊急事件E_i被車輛V_j偵測到時，且內容是($Type_i$; Loc_i; $Time^i_j$)，V_j產生一個的SER如下:

1. V_j計算一對如下的(W_i; S_j)，其中W_i是事件狀態的雜湊值，S_j是事件狀態結果車輛V_j's 的虛擬碼及公開金鑰:

$$W_i = H_2(Type_i \| Loc_i)$$

$$S_j = H_3(Type_i \| Loc_i \| Time^i_j \| PID_j \| PK_j)$$

2. 當有了私鑰(x_j ;D_j), V_j產生如下的簽章(W_i; S_j)

$$Sig^i_j = D_j S_j + x_j W_i$$

因此，($Type_i$; Loc_i; PID_j ; $Time^i_j$; Sig^i_j ; PK_j) 組成一個SER宣告。之後，V_j廣播SER^i_j到他的鄰居。給定SER^i_j = ($Type_i$; Loc_i; PID_j ; $Time^i_j$; Sig^i_j ; PK_j), 一個單一的SER驗證可以被如下的驗證者(verifier)執行:

1. 驗證者首先計算一組(Q_j ;W_i; S_j)如下:

$$Q_j = H_1(PID_j)$$

$$W_i = H_2(Type_i \| Loc_i)$$

$$S_j = H_3(Type_i \| Loc_i \| Time^i_j \| PID_j \| PK_j)$$

2. 之後，驗證者依下檢查簽章的有效性:

$$e(Sig^i_j, P) \stackrel{?}{=} e(Q_j S_j, P_{pub}) e(W_i, PK_j)$$

如果這個式子成立，簽章就能被接受。

聚集授權階段(Aggregated Authentication)

聚集授權包含聚集簽章(signature aggregation)批次驗證(batch verification)，詳細的步驟如下:

聚集簽章: 對一個特定的緊急事件E_i，任何車輛如以當作一個聚集簽章產生者，稱作aggregator，這個aggregator可以聚集一組擁有相同事件宣告($Type_i$;Loc_i)的獨立簽章。如圖十六所示，給定n個SERs，其中SER^i_j = ($Type_i$; Loc_i; PID_j ; $Time^i_j$; Sig^i_j ; PK_j)， $Vj(1 \leq j \leq n)$，aggregator可獲得一個如下的SER_{agg}:

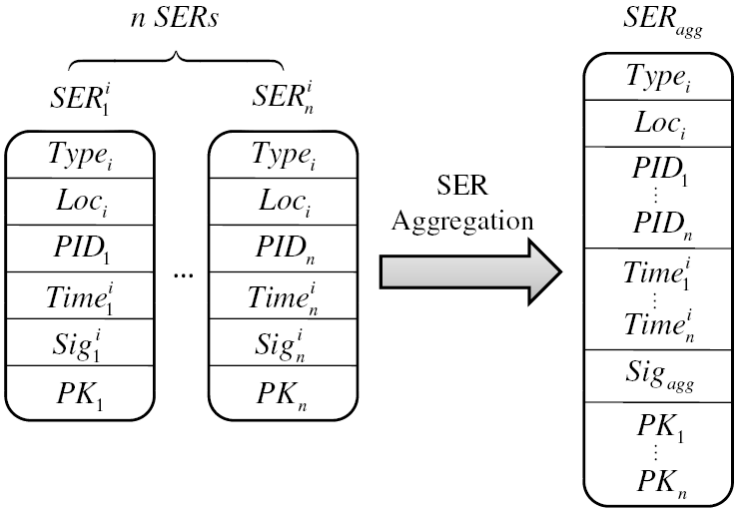
$$\begin{aligned} SER_{agg} = (&Type_i, Loc_i, PID_1, PID_2, \ldots, PID_n, \\ &Time^i_1, Time^i_2, \ldots, Time^i_n, \\ &Sig^i_1, Sig^i_2, \ldots, Sig^i_n, \\ &PK_1, PK_2, \ldots, PK_n) \end{aligned}$$

然後aggregator會產生一個如下的Sig_{agg}:

$$\begin{aligned} Sig_{agg} &= \sum_{j=1}^{n} Sig^i_j \\ &= \sum_{j=1}^{n} (D_j S_j + x_j W_i) \end{aligned}$$

現在，aggregator 依下獲得SER_{agg}:

$$\begin{aligned} SER_{agg} = (&Type_i, Loc_i, PID_1, PID_2, \ldots, PID_n, \\ &Time^i_1, Time^i_2, \ldots, Time^i_n, \\ &Sig_{agg}, PK_1, PK_2, \ldots, PK_n) \end{aligned}$$



圖十六: SER產生程序

批次驗證: 給定一個聚集簽章Sig_{agg}和SER_{agg}，aggregator 計算一組(Qj ;Wi; Sj):

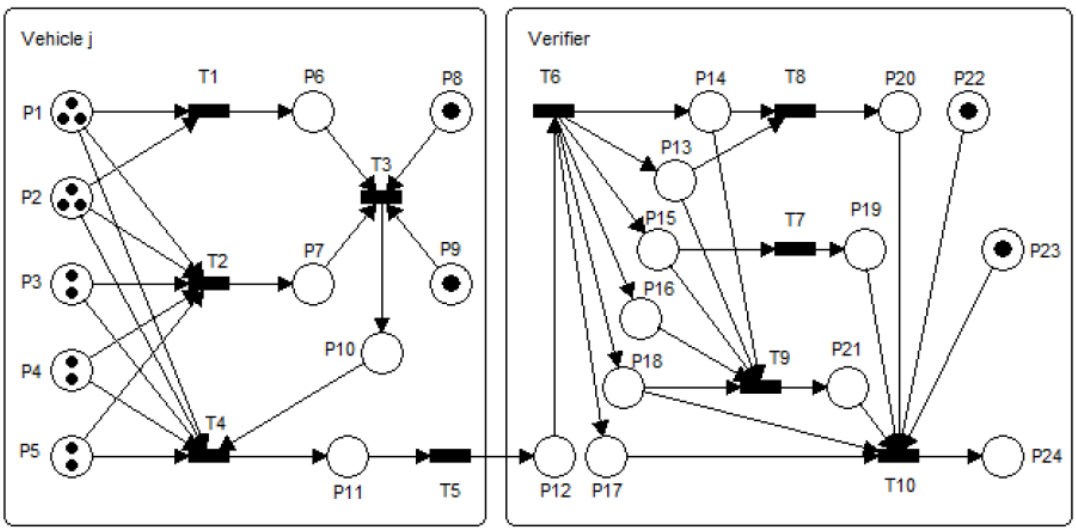
$$\begin{aligned} SER_{agg} = & (Type_i, Loc_i, PID_1, PID_2, \ldots, PID_n, \\ & Time_1^i, Time_2^i, \ldots, Time_n^i, \\ & Sig_{agg}, PK_1, PK_2, \ldots, PK_n) \end{aligned}$$

之後，aggregator檢查聚集簽章的有效性如下:

$$Q_j = H_1(PID_j)$$
$$W_i = H_2(Type_i \| Loc_i)$$
$$S_j = H_3(Type_i \| Loc_i \| Time_j^i \| PID_j \| PK_j)$$

如果式子成立，聚集簽章可被接受。

我們使用 Petri nets 分析所提出機制之資訊流，並可藉此驗證此機制可抵禦偽造攻擊，同時提供車輛之隱私保護，詳細說明可於附錄查詢。本機制之 Petri net 模型如圖十七所示，其中的 place 與 transition 定義分別如表四與表五所示。



圖十七: SAMA 機制之 Petri net 模型

在效能評估方面，如表六所示，我們利用計算量作為效能評估的指標。如表七所示，當驗證 n 個 SERs 時，Zhu 等人所提之機制需要 5 次群數對(bilinear pairings)的計算，以驗

證聚集簽章與憑證；由於本機制採用免憑證公開金鑰密碼系統，故僅需要 3 次群數對的計算來驗證聚集簽章，故本機制可有效地降低聚集訊息驗證之計算量。

表四、相關 Places 之定義

Place	**定義**	**Place**	**定義**
P_1	$Type_i$	P_{13}	$Type_i$
P_2	Loc_i	P_{14}	Loc_i
P_3	ID_j	P_{15}	ID_j
P_4	$Time_j^i$	P_{16}	$Time_j^i$
P_5	PK_j	P_{17}	Sig_j^i
P_6	W_i	P_{18}	PK_j
P_7	S_j	P_{19}	Q_j
P_8	D_j	P_{20}	W_i
P_9	x_j	P_{21}	S_j
P_{10}	Sig_j^i	P_{22}	P
P_{11}	SER_j^i	P_{23}	P_{pub}
P_{12}	SER_j^i	P_{24}	驗證成功資訊

表五、相關 Transitions 之定義

Transition	**定義**	**Transition**	**定義**
T_1	計算 W_i	T_6	分解 SER_j^i
T_2	計算 S_j	T_7	計算 Q_j
T_3	計算 Sig_j^i	T_8	計算 W_i
T_4	建構 SER_j^i	T_9	計算 S_j
T_5	傳送 SER_j^i	T_{10}	驗證 $e(Sig_j^i, P) \stackrel{?}{=} e(Q_j S_j, P_{pub}) e(W_i, PK_j)$

表六、效能評估參數定義

符號	**定義**
T_H	執行一次單向雜湊函數所需之時間
T_E	執行一次指數運算所需之時間
T_P	執行一次群數對運算所需之時間
T_M	執行一次橢圓曲線點乘積運算所需之時間
T_A	執行一次橢圓曲線點加法運算所需之時間

表七、聚集訊息驗證機制效能比較表

階段 \ 方法	Zhu 等人提出之機制	我們所提出之 SAMA 機制
註冊階段	$1T_H + 2T_E$	$1T_H + 2T_M$
SER 產生階段	$3T_H + 2T_E + 2T_M$	$2T_H + 2T_M + 1T_A$
單一 SER 驗證	$4T_H + 1T_E + 5T_P$	$3T_H + 3T_P + 1T_M$
SER 聚集	$2(n\text{-}1)T_M$	$(n\text{-}1)T_A$
SER 批次驗證	$(n+3)T_H + nT_E + 5T_P + 4(n\text{-}1)T_M$	$(2n+1)T_H + 3T_P + nT_M + 2(n\text{-}1)T_A$

四、計畫成果

本計畫為總計畫「異質網路環境之行動搜尋關鍵技術」之第一子計畫，主要的目的為提供車用隨意網路之存取控制與連結機制。為了解決上述車用網路於安全性質服務、非安全性服務、以及通訊安全於所面臨的議題，本子計畫已分三年執行，並完成下列研究成果:

在第一年的研究中: (1) 提出一個以功率控制為基礎的聯合碰撞避訊息廣播機制 PC-CCA。此機制透過功率控制減少實體層中訊息互相的干擾，而功率的調整則是根據車輛之間所需的安全距離來加以設計。考慮到 VANET 網路高度的變動性與有限的頻寬，我們所設計的機制完全不需要任何拓撲的資訊與週期性資料的交換。(2) 針對通訊全問題，設計了一混雜式對應基礎的金鑰協議協定。這個協定不需任何檢查表就可在伺服器與使用者之間達到人工認證以及區段金鑰協議的能力

在第二年的計畫中: (1) 進一步對廣泛的安全應用所存在的延遲與失敗率問題，提出了一個以車輛密度為基礎的緊急訊息廣播機制 VDEB。這個架構主要的概念是透過預估周邊車輛的密度，來減少參與競爭轉送者的數量，並同時降低轉送時所需的等待時間。這個架構也只需要有較少的控制負載，就能達成穩定的運作。(2) 利用 WAVE/DRSC 多重通道的特性，設計了多車道自由流動 ETC 系統的設計，提供高車流環境中穩定且快速的交易需求。

在第三年的計畫中: (1) 我們對各種的非安全性應用，提出了一個以位置為基礎的繞徑協定，稱作行動閘道器繞徑協定 MGRP。此協定結合了 V2V 與 V2I 兩種通訊模式的優點，並以部分車輛作為行動閘道器，延伸固定式式 RSU 的涵蓋範圍，以減少封包跳躍次數和連斷線的可能性。(2) 針對通訊全問題，我們進一步提出了一個安全聚集訊息驗證機制簡稱 SAMA。此機制是基於免憑證公開金鑰密碼系統，用以驗證車用隨意網路之突發緊急事件。經由效能的評估，此機制可有效地降低聚集訊息驗證所需之計算量。並驗證此一方法可有效地抵禦偽造攻擊，同時提供車輛之隱私保護。

本計畫三年的成果, 共包括三篇期刊論文以及六篇研討會論文，並培育了三位博士生以及七位碩士。詳細的論文發表請參見下表。

期刊論文	
1	Andy An-Kai Jeng, Rong-Hong Jan, “Adaptive Topology Control for Mobile Ad Hoc Networks,” accepted and to appear in *IEEE Transactions on Parallel and Distributed Systems*, 2011.
2	Andy An-Kai Jeng, Rong-Hong Jan, Chi-Yu Li, and Chien Chen, “Release-time-based multi-channel MAC protocol for wireless mesh networks,” *Computer Networks*, vol. 55, no. 9, 2011, pp. 2176-2195.
3	Chia-Tai Tsai, Rong-Hong Jan, Chien Chen, "Optimal modulation and coding scheme allocation of scalable video multicast over IEEE 802.16e networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2011, no. 33, July 2011.
研討會論文	
1	Huei-Ru Tseng, Rong-Hong Jan, and Wuu Yang, ”A chaotic maps-based key

	agreement protocol that preserves user anonymity,"*Proceedings of the IEEE International Conference on Communications* (*ICC* 2009), 2009 June
2	Andy An-Kai Jeng, Rong-Hong Jan, Chien Chen, and Tsun-Chieh Cheng, "Efficient broadcast mechanism for cooperative collision avoidance using power control," *Proceedings of the 10th International Symposium on Pervasive Systems, Algorithms and Networks* (*I-SPAN* 2009), 2009 December.
3	Yu-Tian Tseng, Rong-Hong Jan, Chien Chen, Chu-Fu Wang and Hsia-Hsin Li "A vehicle-density-based forwarding scheme for emergency message broadcast in VANETs", *Proceedings of the 3rd IEEE International Workshop on Intelligent Vehicular Networks* (*InVeNet*), 2010 November
4	Hsin-Ya Pan, Rong-Hong Jan, Andy An-Kai. Jeng, Cien Chen, and Huei-Ru Tseng, "Mobile-gateway routing for vehicular networks", *Proceedings of the 8th IEEE Asia Pacific Wireless Communication Symposium* (APWCS 2011), Singapore, Aug. 2011.
5	Huei-Ru Tseng, Rong-Hong Jan, Wuu Yang, and Emery Jou, "A secure aggregated message authentication scheme for vehicular ad hoc networks," *Proceedings of the 18th World Congress on Intelligent Transport Systems*, Orlando, USA, Oct. 2011. (Best Paper Award)
6	Ren-Jhong Liu, Kuochen Wang, Rong-Hong Jan, Yhu-Jyh Hu, and Tien-Hsiung Ku, "An efficient cluster-based data dissemination scheme in wireless sensor networks," *Proceeding of the 73rd IEEE Vehicular Technology Conference* (VTC), 2011, Spring

人才培育	
碩士	7 名
博士	3 名

五、參考文獻

[1]. Vehicle Safety Communications Consortium, http://www-nrd.nhtsa.dot.gov.

[2]. Dedicated Short Range Communications Project, http://www.leearmstrong.comIDSRC.

[3]. The Pre VENT Project, http://www.prevent-ip.org.

[4]. Car2Car Communication Consortium, http://www.car-to-car.org.

[5]. Internet ITS Consortium, http://www.internetits.org.

[6]. The NOW: Network on Wheels Project, http://twww.network-on-wheels.de.

[7]. ITS Taiwan, http://www.its-taiwan.org.tw.

[8]. National Center for Statistics and Analysis, "Traffic Safety Facts 2003", Report DOT HS 809 767 Nat'l. Highway Traffic Safety Admin., U.S. Dot, Washington, DC, 2004.

[9]. ASTM E22213-03, "Standard specification for telecommunication and information exchange between roadside and vehicle ytem – 5GHz band dedicated short range communications (DSRC) MAC and PHY specifications," *ATM International*, July, 2003.

[10]. X. Yang, J. Liu, F. Zhao, N.H. Vaidya, "A vehicle-to-vehicle communication protocol for cooperative collision warning", In Proc. of 1st Annual International Conference on Mobile and Ubiquitous Systems, pp. 114-123, 2004.

[11]. S. Biswas, R. Tatchikou, and F. Dion, "Vehicle-to-vehicle wireless communication protocols for enhancing highway traffic safety", *IEEE Communications Magazine*, Vol. 44, No. 1, pp. 535-547, 2006.

[12]. N. Wisitpongphan, O.K. Tonguz, J.S. Parikh, P. Mudalige, F. Bai, and V. Saderar, "Broadcast storm mitigation techniques in vehicular ad hoc networks", *IEEE Wireless Communications*, Vol. 14, no. 6, pp. 84-94, 2007.

[13]. S. Biswas, R. Tatchikou, and F. Dion, "Vehicle-to-vehicle wireless communication protocols for enhancing highway traffic safety", *IEEE Communications Magazine*, Vol. 44, No. 1, pp. 535-547, 2006.

[14]. F. Yu, and S. Biwas, "Impacts of radio access protocols on cooperative vehicle collision avoidance in urban traffic intersections", *Journal of Communications*, 2008.

[15]. Electronic Toll Collection Market Analysis & Technology Update, Transport Technology Pulish Co., 1999.

[16]. Jianling Li, David Gillen and Joy Dahlgren, "Benefit-cost evaluation of the electronic toll collection system: a compreshensive framework and application", *Transportation Research Board 78th Annual Meeting*, Jan. 10-14, 1999, Washington, D.C..

[17]. W.-Y. Shieh, W.-H. Lee, S.-L. Tung, and C.-D. Ho, "A novel architecture for multilane-free-flow electronic-toll-collection systems in the millimeter-wave range", IEEE Transactions on Intelligent Transportation System, Vol. 6, no.3, 2005, pp. 294-301.

[18]. W. Diffie and M. Hellman, "New directions in cryptography," *IEEE Transactions on Information Theory*, vol. 22, no. 6, Nov. 1976, pp. 644-654.

[19]. F. Dachselt and W. Schwarz, "Chaos and cryptography," IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, vol. 48, no. 12, Dec. 2001, pp. 1498-1509.

[20]. L. Kocarev, "Chaos-based cryptography: a brief overview," IEEE Circuits and Systems

Magazine, vol. 1, no. 3, 2001, pp. 6-21.
[21]. L. M. Pecora and T. L. Carroll, ”Synchronization in chaotic systems,” Physical Review Letters, vol. 64, no. 8, Feb. 1990, pp. 821-824.
[22]. J. Fridrich, ”Symmetric ciphers based on two-dimensional chaotic maps,” International Journal of Bifurcation and Chaos, vol. 8, no. 6, Jun. 1998, pp. 1259-1284.
[23]. L. M. Pecora and T. L. Carroll, ”Driving systems with chaotic signals,” Physical Review A, vol. 44, no. 4, Aug. 1991, pp. 2374-2383.
[24]. K. W. Wong, ”A fast chaotic cryptographic scheme with dynamic lookup table,” Physics Letters A, vol. 298, no. 4, Jun. 2002, pp. 238-242.
[25]. D. Xiao, X. Liao, and S. Deng, ”A novel key agreement protocol based on chaotic maps,” Information Sciences, vol. 177, no. 4, Feb. 2007, pp. 1136-1142.
[26]. E. J. Yoon and K. Y. Yoo, ”A new key agreement protocol based on chaotic maps,” In Proceedings of The Second KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications (KES-AMSTA ’08), Mar. 2008, pp. 897-906.
[27]. X. Yang, J. Liu, F. Zhao, N.H. Vaidya, “A vehicle-to-vehicle communication protocol for cooperative collision warning”, In Proc. of 1st Annual International Conference on Mobile and Ubiquitous Systems, pp. 114-123, 2004.
[28]. S. Biswas, R. Tatchikou, and F. Dion, “Vehicle-to-vehicle wireless communication protocols for enhancing highway traffic safety”, *IEEE Communications Magazine*, Vol. 44, No. 1, pp. 535-547, 2006.
[29]. N. Wisitpongphan, O.K. Tonguz, J.S. Parikh, P. Mudalige, F. Bai, and V. Saderar, “Broadcast storm mitigation techniques in vehicular ad hoc networks”, *IEEE Wireless Communications*, Vol. 14, no. 6, pp. 84-94, 2007.
[30]. F. Yu, and S. Biwas, “Impacts of radio access protocols on cooperative vehicle collision avoidance in urban traffic intersections”, *Journal of Communications*, 2008.
[31]. C. Zhang, R. Lu, P. H. Ho, and A. Chen, “A location privacy preserving authentication scheme in vehicular networks,” In *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC 2008)*, Mar. 2008, pp. 2543-2548.
[32]. X. Lin, R. Lu, C. Zhang, H. Zhu, P. H. Ho, and X. Shen, “Security in vehicular ad hoc networks,” *IEEE Communications Magazine*, vol. 46, no. 4, Apr. 2008, pp. 88-95.
[33]. C. Langley, R. Lucas, and H. Fu, “Key management in vehicular ad-hoc networks,” In *Proceedings of the IEEE International Conference on Electro/Information Technology (EIT 2008)*, May 2008, pp. 223-226.
[34]. N. W. Wang, Y. M. Huang, and W. M. Chen, “A novel secure communication scheme in vehicular ad hoc networks,” *Computer Communications*, vol. 31, no. 12, Jul. 2008, pp. 2827-2837.
[35]. C. T. Li, M. S. Hwang, and Y. P. Chu, “A secure and efficient communication scheme with authenticated key establishment and privacy preserving for vehicular ad hoc networks,” *Computer Communications*, vol. 31, no. 12, Jul. 2008, pp. 2803-2814.
[36]. C. I. Fan, R. H. Hsu, and C. H. Tseng, “Pairing-based message authentication scheme with privacy protection in vehicular ad hoc networks,” In *Proceedings of the International*

Conference on Mobile Technology, Applications, and Systems, Sep. 2008.

[37]. C. Chiasserini, R. Gaeta,M. Garetto,M. Gribaudo, andM. Sereno, Efficient broadcasting of safety messages in multihop vehicular networks, In Proc. of the 20th International Parallel and Distributed Processing Symposium (IPDPS 2006), pp.8, 2006.

[38]. S. Yu and G. Cho, A selective flooding method for propagating emergency messages in vehicle safety communications, In Proc. of the International Conference on Hybrid Information Technology (ICHIT 2006), Vol. 2, pp.556-561, 2006.

[39]. Y.T. Yang and L.D. Chou, Position-based adaptive broadcast for inter-vehicle communications, In Proc. of the IEEE International Conference on Communications (ICC 2008), pp.410-414, 2008.

[40]. H. Alshaer and E. Horlait, An optimized adaptive broadcast scheme for intervehicle communication, In Proc. of the IEEE Vehicular Technology Conference (VTC 2005-Spring), Vol. 5, pp.2840-2844, 2005.

[41]. "IEEE 1609.4 Trial-Use Standard for wireless accesses in vehicular environments (WAVE) – multi-channel operation" IEEE Vehicular Technology Society, October, 2006.

[42]. C. C. Hung, H. Chan, and E. H. K Wu, "Mobility pattern aware routing for heterogeneous vehicular networks," Wireless Communications and Networking Conference (WCNC), Apr. 2008, pp. 2200-2205.

[43]. Y. Peng, Z. Abichar, and J. M. Chang, "Roadside-aided routing (RAR) in vehicular networks," IEEE International Conference on Communications (ICC), Jun. 2006, pp. 3602-3607.

[44]. R. He, H. Rutagemwa, and X. Shen, "Differentiated reliable routing in hybrid vehicular ad-hoc networks," IEEE International Conference on Communications (ICC), May 2008, pp. 2353-2358.

[45]. N. Brahmi, L. Boukhatem, N. Boukhatem, M. Boussedjra, N. D. Nuy, H. Labiod, and J. Mouzna, "End-to-end routing through a hybrid ad hoc architecture for V2V and V2I communications," Ad Hoc Networking Workshop (Med-Hoc-Net), Jun. 2010, pp. 1-8.

[46]. J. Luo, X. Gu, T. Zhao, and W. Yan, "A mobile infrastructure based VANET routing protocol in the urban environment," IEEE International Conference on Communications and Mobile Computing (CMC), Apr. 2010, pp. 432-437.

[47]. C. A. Petri, "Kommunikation mit Automaten," Ph. D. Thesis, University of Bonn, 1962.

[48]. M. Raya and J. P. Hubaux, "The security of vehicular ad hoc networks," In *Proceedings of the 3rd ACM Workshop on Security of Ad hoc and Sensor Networks (SASN 2005)*, Nov. 2005.

[49]. M. Raya, A. Aziz, and J. P. Hubaux, "Efficient secure aggregation in VANETs," In *Proceedings of the 3rd International Workshop on Vehicular Ad hoc Networks (VANETs 2006)*, Sep. 2006, pp. 67-75.

[50]. J. Nikodem and M. Nikodem, "Secure and scalable communication in vehicle ad hoc networks," In *Proceedings of the International Conference on Computer Aided System Theory (EUROCAST 2007)*, Feb. 2007, pp. 1167-1174.

[51]. M. Raya and J. P. Hubaux, "Securing vehicular ad hoc networks," *Journal of Computer*

Security, vol. 15, no. 1, 2007, pp. 39-68.

[52]. C. Zhang, R. Lu, P. H. Ho, and A. Chen, "A location privacy preserving authentication scheme in vehicular networks," In *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC 2008)*, Mar. 2008, pp. 2543-2548.

[53]. X. Lin, R. Lu, C. Zhang, H. Zhu, P. H. Ho, and X. Shen, "Security in vehicular ad hoc networks," *IEEE Communications Magazine*, vol. 46, no. 4, Apr. 2008, pp. 88-95.

[54]. C. Langley, R. Lucas, and H. Fu, "Key management in vehicular ad-hoc networks," In *Proceedings of the IEEE International Conference on Electro/Information Technology (EIT 2008)*, May 2008, pp. 223-226.

[55]. N. W. Wang, Y. M. Huang, and W. M. Chen, "A novel secure communication scheme in vehicular ad hoc networks," *Computer Communications*, vol. 31, no. 12, Jul. 2008, pp. 2827-2837.

[56]. C. T. Li, M. S. Hwang, and Y. P. Chu, "A secure and efficient communication scheme with authenticated key establishment and privacy preserving for vehicular ad hoc networks," *Computer Communications*, vol. 31, no. 12, Jul. 2008, pp. 2803-2814.

[57]. C. I. Fan, R. H. Hsu, and C. H. Tseng, "Pairing-based message authentication scheme with privacy protection in vehicular ad hoc networks," In *Proceedings of the International Conference on Mobile Technology, Applications, and Systems*, Sep. 2008.

[58]. M. Burmester, E. Magkos, and V. Chrissikopoulos, "Strengthening privacy protection in VANETs," In *Proceedings of the IEEE International Conference on Wireless and Mobile Computing (WIMOB 2008)*, Oct. 2008, pp. 508-513.

[59]. C. Zhang, X. Lin, R. Lu, P. H. Ho, and X. Shen, "An efficient message authentication scheme for vehicular communications," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 6, Nov. 2008, pp. 3357-3368.

[60]. H. Zhu, X. Lin, R. Lu, P. H. Ho, and X. Shen, "AEMA: An aggregated emergency message authentication scheme for enhancing the security of vehicular ad hoc networks," In *Proceedings of the IEEE International Conference on Communications (ICC 2008)*, May 2008, pp. 1436-1440.

[61]. A. Shamir, "Identity based cryptosystems and signature schemes," In *Proceedings of the Advances in Cryptology (Crypto 1984)*, 1984, pp. 47-53.

[62]. S. Al-Riyami and K. Paterson, "Certificateless public key cryptography," In *Proceedings of the ASIACRYPT*, 2003, pp. 452-473.

六、附錄

Adaptive Topology Control for Mobile Ad Hoc Networks

Andy An-Kai Jeng and Rong-Hong Jan, *Senior Member*, *IEEE Computer Society*

Abstract—In MANETs, mobile devices are usually powered by batteries with limited energy supplies. Topology control is a promising approach, which conserves energy by either reducing transmission power for each node or preserving energy-efficient routes for the entire network. However, there is empirically a trade-off between the energy efficiency of the nodes and routes in a topology. Besides, it may consume considerable energy to maintain the topology due to node mobility. In this paper, we propose an adaptive topology control protocol for mobile nodes. The protocol allows each node to decide whether to support energy-efficient routing or conserve its own energy. Moreover, it can drastically shrink the broadcasting power of beacon messages for mobile nodes. We prove that any reconstruction and change of broadcasting radius converge in four and five beacon intervals, respectively. The experimental results show that our protocol can significantly reduce the total energy consumption for each successfully transmitted packet, and prolong the life times of nodes, especially in high mobility environments.

Index Terms—Mobile ad hoc network, topology control, energy-efficient protocol, distributed system.

1 Introduction

The continuing developments in mobile ad hoc networks (MANETs) have led to many available applications in commercial, military, and educational areas. Nodes can communicate through wireless carries without any wired connection, thereby enhancing conventional deployment. However, mobile devices are usually powered by limited energy supplies, where a continuing recharging could be hardly attainable. Hence, a substantial body of research has been devoted to conserving energy in MANETs.

The *topology control* is an important approach to conserving energy [1], which aims at determining a set of wireless links among nodes so as to achieve certain energy-efficient properties. Generally speaking, it can reduce the energy consumption in two ways.

1. *Reduce energy consumption of nodes*. In wireless networks, the power required to transmit from one node to another is considerable, and could be exponentially grown by their distance [2]. Thus, to conserve a node's energy, the *transmission radius* should be confined to cover closer neighbors only in the underlying topology. On the other hand, nodes are responsible for relaying messages in MANETs. If the loads are overly concentrated on a certain node, the node's energy could be quickly drained out. Keeping a lower *node degree* (the number of links connected to a node) can prevent a node from relaying for too many sources [3].
2. *Reduce energy consumption of routes*. In MANETs, communication is typically conducted by relaying messages through some paths. During the relaying process, each node on a path should transmit at sufficient power to cover the next hop. Therefore, the path with smaller total transmission power, called an *energy-efficient route*, should be preserved for any possible communication pair while controlling the topology.

Overall, the energy efficiencies of nodes and routes are equally important. The living time of an individual node can be prolonged, if the node degree (transmission radius) is reduced to consume less energy. Moreover, the total energy consumption for the global wide communication can be saved by preserving more energy-efficient routes. Nevertheless, there is empirically a trade-off. To reduce the transmission radius or node degree, some links constituting an energy-efficient route could be sacrificed.

To address the trade-off, we have proposed a flexible structure, called the *r-neighborhood graph*, in our recent studies [4], [5], [6]. As shown in Fig. 1, given two nodes u and v, and a parameter $0 \leq r \leq 1$, we define the region (the shaded area) intersected by two open disks centered, respectively, at u and v with the radius of their distance $d(u, v)$ and an open disk centered at the middle point m with the radius $l = (d(u, v)/2)(1 + 2r^2)^{1/2}$ as the *r-neighborhood region* of u and v, denoted as $NR_r(u, v)$. The r-neighborhood graph of a set of nodes V, denoted as $NG_r(V)$, consists of an edge uv if and only if $NR_r(u, v)$ contains no other node in V.

The energy consumption between nodes and routes in this graph can be balanced by adjusting the parameter r. By increasing r, the radius and degree of each node become smaller. On the contrary, more energy-efficient routes can be found by reducing r. In particular, when $r = 1$ the node degree is not greater than 6, and the optimal energy-efficient routes are preserved when $r = 0$. More importantly, each node can asynchronously determine its links in this graph using the positions of its one-hop neighbors. In other words, the construction is *fully distributed* and *localized*.

• *The authors are with the Department of Computer Science, National Chiao Tung University, Hsinchu 300, Taiwan, R.O.C. E-mail: {andyjeng, rhjan}@cis.nctu.edu.tw.*

Manuscript received 15 July 2010; revised 13 Dec. 2010; accepted 7 Feb. 2011; published online 18 Feb. 2011.
Recommended for acceptance by P. Santi.
For information on obtaining reprints of this article, please send e-mail to: tpds@computer.org, and reference IEEECS Log Number TPDS-2010-07-0426.
Digital Object Identifier no. 10.1109/TPDS.2011.68.

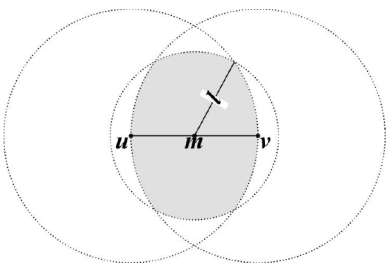


Fig. 1. The r-neighborhood region of u and v.

However, the r-neighborhood graph was primarily designed for stationary nodes. When applied to mobile environments, more attention should be paid to nodes' mobility. Besides, for theoretical interest, the power consumption model was simplified in our previous works. The simplification may overlook some facts in reality. For these reasons, our goal is to extend the concept of the r-neighborhood graph to a more realistic network. To achieve this purpose, we make the following contributions in this paper.

1. *Generalized power consumption model*. In [4] and [5], we assumed the power consumed at the receiver is negligible. Besides, the path loss is specific to free space environments, where no obstacle or reflection exists. In this paper, we generalize the r-neighborhood graph to a more realistic power consumption model proposed by Rodoplu and Meng [7]. Both the receiving cost and the general path loss exponent are considered in this model.
2. *Extended parameter set*. Although the energy consumption can be adjusted through the parameter r, the desired value of r would be varied for different nodes. For example, a node with less energy would prefer a larger r to reduce its own transmission radius or node degree, while a smaller r would be preferred, if the node has surplus energy to perform relaying for other communication pairs. In other words, an identical r cannot provide the most appropriate settings for all nodes. Therefore, we extend the r-neighborhood graph so that each node u has the flexibility to configure its own r_u.
3. *Energy-efficient maintenance protocol*. To maintain the topology for mobile nodes, each node has to periodically broadcast a beacon to denote its new position. It may consume considerable energy, if the broadcasting power is large. We design an energy-efficient maintenance protocol, named the *Adaptive Neighborhood Graph-based Topology Control (ANGTC)*. The *ANGTC* can drastically shrink the broadcasting power for each periodic beacon. Moreover, we prove that any reconstruction can be done in 4Δ, where Δ is the beacon interval.
4. *Adaptive configuration rule*. In [6], we turned the value of r to find the minimal energy consumption using simulation. But there was no discussion about how to adjust r in a decentralized matter. Moreover, the settings of different r_u's could be more complicated. Therefore, this paper proposes an adaptive configuration rule inside the ANGTC to configure the parameter r_u for each node u. The rule aims at achieving balanced energy consumption between nodes and routes, and improving the stability of the topology.

For a detailed introduction to the r-neighborhood graph and its challenges in mobile environments, readers can refer to Appendices A.1, A.2, and A.3, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPDS.2011.68.

The rest of this paper is organized as follows: Section 2 specifies the network model and measurements. Section 3 defines the graphic structures and analyzes their properties. The protocol and configuration rule are investigated in Sections 4 and 5, respectively. Section 6 presents a series of simulation results. Concluding remarks are given in the last section. The proof of any property shown in this paper can be found in Appendix F, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPDS.2011.68.

2 Network Model and Measurements

Given a deployment region $\aleph$, a set V of n nodes is distributed on $\aleph$. Each node $u \in V$ can obtain its location $Loc(u)$ on $\aleph$ using a lower power GPS. Besides, the power consumption follows the *path loss model* [2]. More specifically, let $p_{max}(u)$ denote the maximum transmission power of a node u. Node u can transmit to another node v only if $td(u,v)^\alpha \le p_{max}(u)$, where $d(u,v)$ is the euclidean distance between u and v, α is an exponent depending on the environment [2], and t is the predetection threshold (in mW) at the receiver side, $t > 0$. The network can be represented as a digraph $G_{max}(V)$, where a directed edge $uv \in G_{max}(V)$ if and only if $td(u,v)^\alpha \le p_{max}(u)$. In addition, node v needs additional c power to receive from u. Therefore, the least power required for a transmission from u to v in this model is $c + td(u,v)^\alpha$ [7].

Generally speaking, the topology control is to determine a subgraph of $G_{max}(V)$. Consider a controlled topology $G(V)$. The *transmission radius* and *degree* of a node u in $G(V)$ are defined, respectively, as

$$T_u(G(V)) = max_{uv \in G(V)} d(u,v); \tag{1}$$

$$D_u(G(V)) = |\{v \in V | uv \in G(V)\}|. \tag{2}$$

Let $\pi(u,v) = v_0 v_1 \cdots v_{h-1} v_h$ denote a path connecting two nodes u and v, where $v_0 = u$ and $v_h = v$. The *total transmission power* required to relay on $\pi(u,v)$ is

$$P(\pi(u,v)) = \sum_{i=1}^{h} [c + td(v_{i-1}, v_i)^\alpha]. \tag{3}$$

In worse-case situations, the energy efficiency of nodes is measured by the *maximum node degree*, $D_{max}(G(V))$, and the energy efficiency of routes being preserved is measured by the *power stretch factor* [3]

$$\rho(G) = max_{uv \in V} \frac{P(\pi^*_{G(V)}(u,v))}{P(\pi^*_{G_{max}(V)}(u,v))},$$

where $\pi^*_{G(V)}(u,v)$ is the path with the least total transmission power between u and v in $G(V)$.

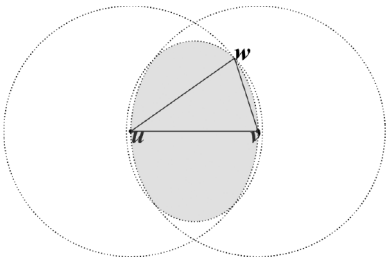


Fig. 2. The general r-neighborhood region of u and v.

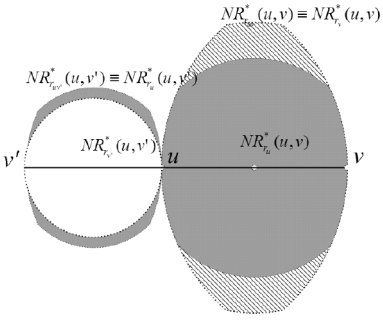


Fig. 3. General f_r-neighborhood graph of nodes v', u, and v, where $r_{v'} < r_u < r_v$.

3 Graphic Structures

In this section, we first generalize the r-neighborhood graph to a more realistic power consumption model and variant r_u's. Then, an equivalent structure is defined to facilitate the design of an energy-efficient maintenance protocol in the next section.

3.1 Generalization

First of all, we define the following region under the power consumption model $P(uv) = c + td(u,v)^\alpha$, for any two nodes u and v.

Definition 1. *Given two nodes u and v on $\aleph$, and $0 \le r \le 1$, the general r-neighborhood region of u and v is defined as*

$$NR^*_r(u,v) = \left\{ \begin{array}{l} x \in \aleph : d(u,x) < d(u,v), \\ d(v,x) < d(u,v), \\ P(uxv) < P(uv)(1+r^\alpha) \end{array} \right\}.$$

Fig. 2 shows the general r-neighborhood region (the shaded regions) of two nodes u and v. Compared with Fig. 1, we can see that $NR_r(u,v)$ and $NR^*_r(u,v)$ are only diverse in their third conditions, which correspond to an inner circle and an inner ellipse, respectively. The condition indicates that the total power required by relaying through any node w in $NR^*_r(u,v)$ (i.e., $\pi(u,v) = uwv$) is not worse than $(1+r^\alpha)$ times of a direct transmission from u to v.

Based on this region, the graph with variant r_u's is defined as follows.

Definition 2. *Given a set V of n nodes on $\aleph$, a set $f_r : \{r_{v_1}, r_{v_2}, \ldots, r_{v_n}\}$, $0 \le r_{v_i} \le 1$, the general f_r-neighborhood graph of V, denoted as $NG^*_{f_r}(V)$, has an edge uv if and only if $uv \in G_{max}(V)$ and there is no other node w such that $Loc(w) \in NR^*_{r_{uv}}(u,v)$, where $r_{uv} = max\{r_u, r_v\}$.*

A three-node example of $NG^*_{f_r}(V)$ is depicted in Fig. 3. The regions with respect to r_u, r_v, and $r_{v'}$ are filled with gray, twill, and white, respectively. We can see that the gray area determines edge uv' because $r_{v'} < r_u$, while edge uv is determined by the twilled area because $r_u < r_v$. That is, the presence of an edge is now determined by the larger one of the two sides, instead of an identical r.

Let $NG^*_{f_r|r_u=r_0}$ represent the case where the parameter of a node u is fixed on a ratio r_0. We have the following monotonic property with respect to each r_u (see Appendix F.1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPDS.2011.68 for the proof).

Property 1. *Given a set V of n nodes on $\aleph$, for any $f_r : \{r_{v_1}, r_{v_2}, \ldots, r_{v_n}\}$, and $0 \le r_1 \le r_2 \le 1$,*

1. $D_u(NG^*_{f_r|r_u=r_2}) \le D_u(NG^*_{f_r|r_u=r_1})$, $\forall u \in V$;
2. $T_u(NG^*_{f_r|r_u=r_2}) \le T_u(NG^*_{f_r|r_u=r_1})$, $\forall u \in V$;
3. $P(\pi^*_{NG^*_{f_r|r_u=r_1}}(s,t)) \le P(\pi^*_{NG^*_{f_r|r_u=r_2}}(s,t))$, $\forall s,t \in V$.

We can see that no matter what the values of other parameters in f_r are taken, a node u can conserve its own energy by choosing a larger r_u, i.e., reducing its $D_u(NG^*_{f_r}(V))$ and $T_u(NG^*_{f_r}(V))$. On the contrary, if node u has sufficient energy, it can just choose a smaller r_u to support more energy-efficient routing, i.e., reducing $P(\pi^*_{NG^*_{f_r}(V)}(s,t))$, for any possible communication pair of s and t.

Let $NG^*_r(V)$ denote the case where $r_u = r$ for any $u \in V$. The worst-case performance for an identical r is presented below (proven in Appendix F.2, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPDS.2011.68).

Property 2. *Given a set V of n nodes on $\aleph$, for any $0 \le r \le 1$,*

1. $D_{max}(NG^*_r(V)) \le \lceil \pi / \sin^{-1}(r/2) \rceil$, *if* $c = 0$;
2. $\rho(NG^*_r(V)) \le 1 + r^\alpha(n-2)$.

Property 2(2) shows that the graph preserves the upper bound of the power stretch factor as proven in [8], even if it is now defined under a general power consumption model. On the other hand, the node degree's bound in [8] is also preserved in Property 2(1), but it is restricted to the case where the receiving cost is negligible, i.e., $c = 0$. However, the new structure can result in a much lower degree and shorter transmission radius for $c > 0$ in an average sense, especially when the path loss exponent α is large (see the numerical results in Appendix D.1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPDS.2011.68).

For the case of variant r_u's, the two upper bounds in Property 2 are clearly determined by the largest and smallest r_u's in f_r, denoted as r_{min} and r_{max}, respectively. Therefore, it is easy to infer that

$$D_{max}(NG^*_{f_r}(V)) \le \lceil \pi / \sin^{-1}(r_{r_{min}}/2) \rceil, \text{if } c = 0,$$

and

$$\rho(NG^*_{f_r}(V)) \le 1 + r^\alpha_{max}(n-2).$$

Finally, the property below shows that our structure is symmetric and connected (the proof is in Appendix F.3, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPDS.2011.68). A topology is symmetric if the presence of an edge implies that its inverse exists, which is important, since the designs of many network primitives, such as collision avoidance, would become very complicated if links are asymmetric. Moreover, connectivity is unquestionably the prerequisite in any network.

Property 3. *Given a set V of n nodes on $\aleph$, for any $f_r : \{r_{v_1}, r_{v_2}, \ldots, r_{v_n}\}$,*

1. *$NG^*_{f_r}(V)$ is symmetric;*
2. *$NG^*_{f_r}(V)$ is connected.*

The relationship between $NG_r(V)$ and $NG^*_{f_r}(V)$ is as follows (see Appendix F.4, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPDS.2011.68, for the proof).

Property 4. *Given a set V of n nodes on $\aleph$, for any $0 \le r \le 1$, if $c = 0$, $\alpha = 2$, and $r_u = r$ for any $u \in V$,*

$$NG^*_{f_r}(V) \equiv NG_r(V).$$

We can see that the two structures are equivalent when $\alpha = 2$ and $c = 0$. In other words, $NG^*_{f_r}(V)$ is a general structure of $NG_r(V)$ in terms of α and c.

3.2 Equalization

Now, we present an equivalent structure of the general f_r-neighborhood graph, called the general *f_r-enclosed graph*. The basic idea is borrowed from the *enclosed graph*, proposed by Rodoplu and Meng [7].

First, we define a duality[1] of the general r-neighborhood region.

Definition 3. *Given two nodes u and w on $\aleph$, $0 \le r \le 1$, $\alpha \ge 2$, the general r-relaying region of u and w is defined by*

$$RR^*_r(u,w) = \left\{ \begin{array}{l} x \in \aleph : d(u,w) < d(u,x), \\ d(w,x) < d(u,x), \\ P(uwx) < P(ux)(1 + r^\alpha) \end{array} \right\}.$$

A region, enclosed by the complements of the general r-relaying regions, is given as follows.

Definition 4. *Given a set V of nodes on $\aleph$, the general r-enclosed region of a node u is defined by*

$$ER^*_r(u) = \bigcap_{uw \in G_{max}(V)} \{\aleph \cap R_{max}(u) - RR^*_r(u,w)\}.$$

Based on the region, the graph is defined below.

Definition 5. *Given a set V of n nodes on $\aleph$, a set $f_r : \{r_{v_1}, r_{v_2}, \ldots, r_{v_n}\}$, $0 \le r_{v_i} \le 1$, the general f_r-enclosed graph of V, denoted as $EG^*_{f_r}(V)$, has an edge uv if and only if $uv \in G_{max}(V)$ and $Loc(v) \in ER^*_{r_{uv}}(u)$, where $r_{uv} = \max\{r_u, r_v\}$.*

1. See Appendix B.1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPDS.2011.68, for two comparisons that explain the dual relationship between the two regions.

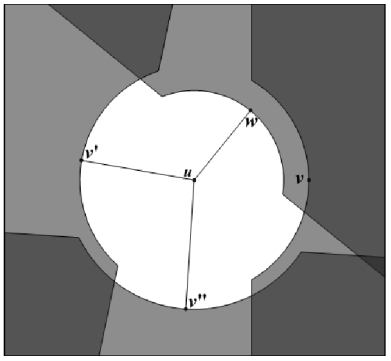


Fig. 4. General f_r-enclosed graph of a node u, where $c = 0$, $\alpha = 2$, and r_u, r_v, $r_{v'}$, $r_{v''}$ are all set as 1.

Fig. 4 shows the f_r-enclosed region of a node u (white area), which is enclosed by the four r-relaying regions (dark areas) of u with surrounding nodes v, v', v'', and w (the darker areas are overlapped by two or more regions). We can see that a node has a link from u if and only if it is located in the f_r-enclosed region of u. To see how r_u changes the shapes of our defined regions, readers can refer to Appendix B.2, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPDS.2011.68, for further illustration.

Now we show that the two structures are equivalent (See Appendix F.5, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPDS.2011.68, for the proof).

Property 5. *Given a set V of n nodes on $\aleph$, for any $f_r : \{r_{v_1}, r_{v_2}, \ldots, r_{v_n}\}$, $0 \le r_{v_i} \le 1$,*

$$NG^*_{f_r}(V) \equiv EG^*_{f_r}(V).$$

4 Energy-Efficient Maintenance

Based on the equivalence in Property 5, in this section, we design an energy-efficient maintenance protocol for the general f_r-enclosed graph.

4.1 The ANGTC Protocol

The main idea of this protocol is to utilize the information partially received from nearby nodes to confine the broadcasting radiuses of subsequent beacons.

In every time interval of Δ, each node broadcasts a beacon at a certain radius to nearby nodes. Consider a node u. Let S_u denote the set of nodes detected by u during the previous Δ time. Similar to Definition 4, we define $ER^*_r(u|S_u)$ as the general r-enclosed region of u based on nodes in S_u, i.e.,

$$ER^*_r(u|S_u) = \bigcap_{w \in S_u} \{\aleph \cap R_{max}(u) - RR^*_r(w,v)\}. \quad (4)$$

The set of nodes in S_u being enclosed by node u in $EG^*_{f_r}(V)$ is specified as

$$N_u = \{v \in S_u | Loc(v) \in ER^*_{r_{uv}}(u|S_u)\}. \quad (5)$$

In addition, we denote λ_u as the least radius covering $ER^*_{r_u}(u|S_u)$, i.e.,

$$\lambda_u = \max\{d(u,x)|x \in ER^*_{r_u}(u|S_u)\}. \quad (6)$$

With the definition in (6), for any node v in S_u, if u is within the radius λ_v of v which covers $ER^*_r(v|S_v)$, then v will be included in a nodes set B_u. This is,

$$B_u = max\{v \in S_u | d(u,v) < \lambda_v\}. \quad (7)$$

The radius covering all nodes in B_u is specified by

$$\chi_u = max\{d(u,v)|v \in B_u\}. \quad (8)$$

Then, in the next maintenance process, node u will broadcast its current position $Loc(u)$, λ_u, and r_u using the beacon to nearby nodes at the radius M_u, where

$$M_u = max\{\lambda_u, \chi_u\}. \quad (9)$$

On the other hand, if node u cannot receive the beacon from a node v over a beacon interval Δ, the information about v will be discarded. In other words, for every maintenance process, the broadcasting radius M_u of u is adjusted to cover both the area in $ER^*_r(S_u)$ and all nodes in B_u, depending on the information received during the previous Δ interval. For every Δ time, the value of r_u will be reconfigured and broadcasted to nearby nodes along with the periodic beacon message. The protocol, named ANGTC, is now presented below.

ANGTC PROTOCOL

$N_u = \{\}$; $S_u = \{\}$; $B_u = \{\}$;
For every Δ time
  reconfigure r_u;
  $N_u = \{v \in S_u | Loc(v) \in ER^*_{r_{uv}}(u|S_u)\}$;
  $\lambda_u = \max\{d(u,x)|x \in ER^*_{r_u}(u|S_u)\}$;
  $\chi_u = \max\{d(u,v)|v \in B_u\}$;
  $M_u = \max\{\lambda_u, \chi_u\}$;
  Broadcast ($Loc(u)$, λ_u, r_u) at the radius M_u;
Upon receiving ($Loc(v)$, λ_v, r_u) from a node v,
  $S_u = S_u + \{v\}$;
  If $d(u,v) < \lambda_v$, $B_u = B_u + \{v\}$;
Upon not receiving from v over Δ time,
  $S_u = S_u - \{v\}$; $B_u = B_u - \{v\}$.

An example of the ANGTC protocol is elaborately illustrated in Appendix C, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPDS.2011.68.

4.2 Correctness and Convergency

Now, we discuss the correctness and convergency of the ANGTC protocol. We show that after nodes' placement changes, the neighbor set N_u and maintenance radius M_u of each node u can be correctly recalculated and converged to a stable status in constant time.

We assume that the propagation delay and computation time are relatively small in comparison with Δ and the starting points of every time interval among nodes are aligned (i.e., a synchronous network). Without loss of generality, we consider the case of $\Delta = 1$ henceforth. In addition, we assume that the configuration of each r_u is temporarily fixed before the radius M_u is converged. Let X^t stand for the status of a variable X at time t, and $N_u(G)$ denote the neighbor set of a node u in a graph G. The results are shown in the following property (proven in Appendix F.6, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPDS.2011.68).

Property 6. *Given a set V of nodes on $\aleph$, a placement change occurs during $[t-1,t]$, i.e., $\exists u \in V$, $Loc^{t-1}(u) \neq Loc^t(u)$, and there is no further change after t, i.e., $\forall u \in V$, $k > 0$, $Loc^t(u) = Loc^{t+k}(u)$. If the network is synchronous, and the parameter r_u of each $u \in V$ is fixed after time t,*

1. $N_u(EG^*_{f_r}(V)) \subseteq N_u^{t+2}$;
2. $N_u(EG^*_{f_r}(V)) = N_u^{t+k}$, *for any* $k \geq 3$;
3. $M_u^{t+4} = M_u^{t+k}$, *for any* $k > 4$.

Property 6(1) indicates that the ANGTC protocol can find out all links in $EG^*_{f_r}(V)$ in 3Δ after a change occurs. Besides, Properties 6(2) and (3) show that the correct neighbor set N_u and maintenance radius M_u converge in 4Δ and 5Δ, respectively. The convergency of M_u is important, because the continuing change in broadcast radius will incur additional energy expense and latency for power switching.

For an asynchronous network, after the change during $[t-1,t]$, each node u can receive updated positions from nearby nodes before $t+2$. Hence, the converged time is postponed by at most Δ, i.e., 5Δ and 6Δ for N_u and M_u, respectively.

About the communication cost, as shown in Property 3(1), since our graph is inherently symmetric, nodes are not required to exchange their neighbor lists. Thus, each message has only constant bits.

4.3 Further Power Shrinking

Although the ANGTC can reduce the beacon power, the radius M_u could be too large to cover some nodes which are not essential for the operations. Therefore, we attempt to further shrink the radius.

Consider two nodes u and v. Let $ER^*_{r_v}(v|B_u)$ denote the enclosed region of v based on nodes in B_u, i.e.,

$$ER^*_r(v|B_u) = \bigcap_{w \in B_u} \{\aleph \cap R_{max}(u) - RR^*_r(v,w)\}.$$

We define

$$B'_u = \{v \in B_u | ER^*_{r_{uv}}(v|B_u) \neq ER^*_{r_{uv}}(v|B_u + \{u\})\},$$

and

$$\chi'_u = max\{d(u,v)|v \in B'_u\}.$$

Because B'_u is a subset of B_u, χ'_u must be smaller than (or at most equal to) χ_u.

Below, we show that Property 6 is still preserved when χ_u is replaced by χ'_u in the ANGTC. Since the definition of λ_u is not changed, it is sufficient to prove the property below (see Appendix F.7, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPDS.2011.68, for the proof).

Property 7. *Given a set V of nodes on $\aleph$, for any $u \in V$ and $w \in S^*_u$, $u \in B'_w$.*

We have also conducted a numeric study for our protocol. The results show that the ANGTC can significantly reduce the average transmission radius by 5 to 60 percent. The power shrinking mechanism can further shrink the radius up to 10 percent. The detailed results can be found in Appendix D.2, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPDS.2011.68.

5 Adaptive Configuration Rule

In this section, we propose an adaptive configuration rule for the ANGTC protocol. We first analyze how an individual r_u affects the overall energy efficiency from the following three points.

1. *Energy efficiency of nodes versus Energy efficiency of routes*. No matter what the values of other parameters in f_r are, Property 1 has shown that a smaller r_u can always lead to an overall improvement in the energy efficiency of routes, and a node u can conserve its own energy by simply turning up its r_u.
2. *High mobility versus Low mobility*. If a node moves frequently, its links are unstable, which in turn costs more energy for route reconstruction, and deteriorates the quality of the established routes. In this case, the node should keep a lower degree to reduce its dependency on nearby nodes by turning up its r_u. On the contrary, if a node has lower mobility, it should turn down its r_u to construct more reliable routes.
3. *Topology maintenance power*. In the ANGTC protocol, the broadcasting radius M_u is not fixed. As shown in Property 8 (proven in Appendix F.8, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPDS.2011.68), it could be varied by the configuration of r_u.

Property 8. *In the ANGTC protocol, for each node u, the broadcasting radius M_u is decreased by r_u.*

Concluding the above observations, we have two principles for adjusting r_u:

1. If a node u has sufficient energy or rarely moves, it should connect with more neighbors to improve the energy efficiency as well as the stability of routes. In this case, a smaller r_u is preferred.
2. If a node has insufficient energy or moves frequently, a smaller r_u that leads to a lower node degree, transmission radius, and beacon power is desired.

Accordingly, the adaptive configuration rule is characterized as follows:

$$r_u = \left(1 - \frac{Energy_u}{Energy_{Full}}\right) w_{e,u} + \left(\frac{Mobility_u}{Mobility_{Max}}\right) w_{s,u}, \qquad (10)$$

where $Energy_u$ and $Mobility_u$ stand for the residual energy and current mobility of node u, and $Energy_{Full}$ and $Mobility_{Max}$ represent the full energy level and the maximum mobility level, respectively. In addition, we adjust the impact from residual energy and mobility level by two weights $w_{e,u} = 1 - (Energy_u/Energy_{Full})$ and $w_{s,u} = 1 - w_{e,u}$. At the initial stage, since nodes have little deviation in their residual energy (assuming the initial energy is equal), the mobility dominates the value of r_u. As time goes by, the impact from node's energy will become more and more significant.

6 Experiments

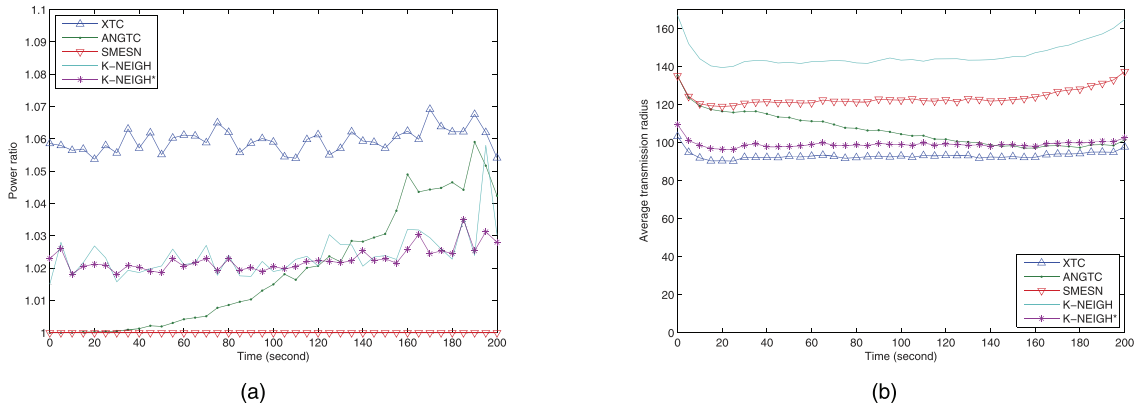
In this section, we compare the ANGTC with existing topology control protocols for mobile nodes using the ns2 simulator [8]. For each test case, we simulated 50 networks, each with 100 nodes uniformly placed on a 1,000 meters square region. Each node has a maximum transmission radius of 500 meters and is initiated by 0.5 Joules. The 802.11b MAC is used for link-layer contention. We modify the DSDV routing protocol [9] such that packets are conveyed on the least-energy path. The connections of CBR traffic are established for 20 distinct source-destination pairs, and the packet size is 256 bytes. The energy cost consists of all network operations during the simulation. The mobility pattern is based on the random-way point model. We test three speed intervals of [0, 5], [0, 15], and [0, 30] m/s, to imitate *low-speed*, *middle-speed*, and *high-speed* circumstances, respectively. In addition, the pause time of each node is randomly taken from [0, 5] s. Each run lasts 200 s.

For comparison, we also implemented the following protocols in the ns2: the SMECM [10] is considered appropriate for conserving route energy. It preserves the least-energy path (i.e., $\rho = 1$) for any node pair. On the other hand, the XTC [11] is considered appropriate for conserving nodes' energy. It confines node degrees to within 6, with connectivity guaranteed. The K-NEIGH [12] is considered resilient to node mobility. It only requires nodes to identify their K-closest neighbors instead of their precise positions. A pruning stage is proposed in [12] to revoke redundant links. We denote this version as K-NEIGH*. Here, we take $K = 9$, which is the least value for the topology to be connected with a probability of 0.95 [12]. For more details, readers can refer to Appendix E, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPDS.2011.68, where we provide a review of related protocols.

Figs. 5a and 5b report the power ratio and average transmission radius for stationary nodes. The power ratio, defined as

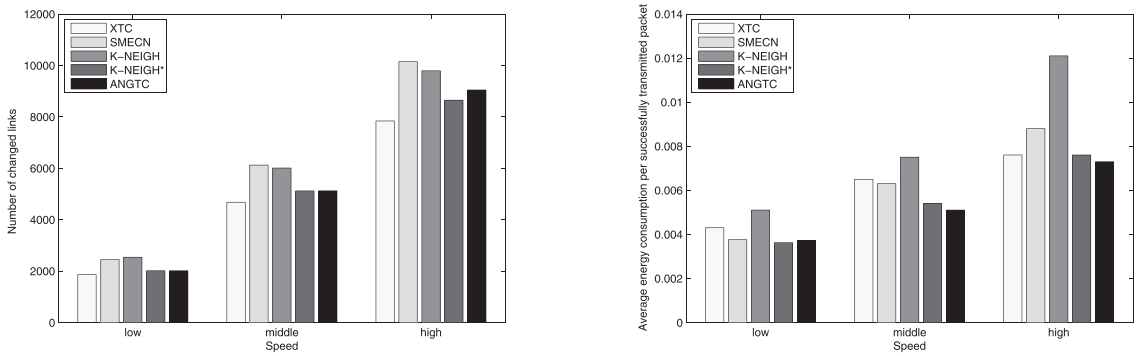
$$\frac{\sum_{u,v \in V} p(\pi^*_G(u,v))}{\sum_{u,v \in V} p(\pi^*_{G_{max}}(u,v))},$$

measures the average energy efficiency of the routes. We can see that the SMECN always preserves the optimal routes, but it compensates for a larger transmission radius. On the other hand, the power ratio of the XTC is about six percent larger than the optimal, but it has a much smaller transmission radius. The radius of K-NEIGH can be reduced significantly in the pruning stage, but it is still slightly above than that of the XTC. However, these



(a) (b)

Fig. 5. (a) Energy efficiency of routes. (b) Energy efficiency of nodes.



(a) (b)

Fig. 6. (a) Number of changed links. (b) Overall energy efficiency of communication.

protocols have no flexibility as the energy level goes down with time. In contrast, the ANGTC allows nodes to adjust their ways of conserving energy. At the beginning, since each node u has full energy, according to (10), the parameter r_u is close to 0, which forces node u to support energy-efficient routing. As time goes by, r_u is gradually raised to 1 so that node u can reduce the radius to conserve its own energy.

Fig. 6a shows the number of changed links. It measures the stability of a topology under nodes' mobility. The XTC has the least link changes, since each node only has to maintain no more than six links. The K-NEIGH* also performs well, because keeping the order of nodes is much easier than keeping the precise positions. Although the ANGTC uses positions and has more links than XTC, the stability is nearly at the same level for both of them. The reason is that our configuration rule can adaptively reduce to the degree of a node if the node moves frequently.

Fig. 6b reports the average energy consumption per successfully transmitted packet. It measures the overall energy consumption for a communication, including routing, route reconstruction, and retransmissions. In low-speed networks, since links change rarely, the energy efficiency of routes is relatively important. In this case, the SMECN is suitable. On the contrary, the XTC performs well in high-speed networks, because a large portion of energy may be consumed for advertising the link's changes. Since the ANGTC can change link status according to node mobility, it accommodates well in both cases. Our protocol, however, can perform even better. One possible reason is that the K-NEIGH* (and K-NEIGH) is only connected in a probability sense, while the general f_r-neighborhood graph always guarantees connectivity. Hence, our protocol requires less energy for retransmissions before a packet arrives successfully.

The number of living nodes with middle speed is drawn in Fig. 7 (the results are almost the same for the other two speeds). Even though the transmission radius and node degree of the ANGTC are not lowest, it still outperforms the others. This is because the enhancement of routes can also reduce the energy expenditure of nodes. In other words, the nodes' energy synergically is conserved in these two ways.

Fig. 8 shows the ratio of the maintenance radius (power) to the maximum transmission radius (power) The results show that our protocol requires no more than 50 percent of power to maintain the topology. With this improvement, the power can be further reduced by over 20 percent. Notice that the radius (power) steadily decreases along with the depletion of node energy and slightly increases when some nodes are exhausted.

7 Conclusions

In this paper, we have generalized the r-neighborhood graph into a more realistic power consumption model with independent parameter r_u to each node u. For mobile nodes, we have also proposed an energy-efficient maintenance protocol to reduce the beacon power. It has been proven that any reconstruction and power change can coverage in

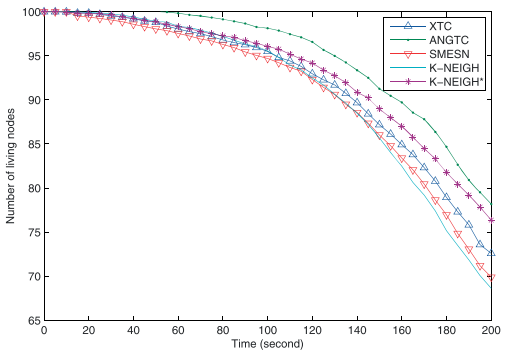


Fig. 7. Number of living nodes.

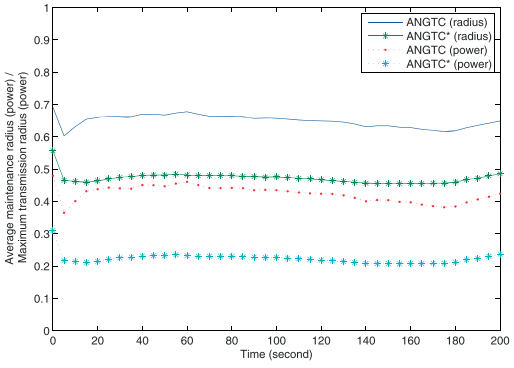


Fig. 8. Maintenance radius and power.

four and five beacon intervals. Finally, an adaptive configuration rule is given to configure the parameter for each node based on the node's mobility and energy levels. Experimental results show that our protocol has significantly reduced the overall energy consumption and network lifetime. For future research, a node may lose important information to construct the graph if a collision occurs. It is, thus, worthwhile to design a collision avoidance mechanism for the ANGTC protocol.

ACKNOWLEDGMENTS

This research was supported in part by the National Science Council (NSC) of the ROC, under grants NSC97-2221-009-049-MY3 and NSC99-2218-E-009-007.

REFERENCES

[1] P. Santi, "Topology Control in Wireless Ad Hoc and Sensor Networks," *ACM Computing Survey,* vol. 37, no. 2, pp. 164-194, 2005.
[2] L. Feeney, "An Energy Consumption Model for Performance Analysis of Routing Protocols for Mobile Ad Hoc Networks," *ACM J. Mobile Networks and Applications,* vol. 6, no. 3, pp. 239-249, 2001.
[3] X.Y. Li, P.J. Wan, and Y. Wang, "Power Efficient and Sparse Spanner for Wireless Ad Hoc Networks," *Proc. 10th Int'l Conf. Computer Comm. and Networks,* pp. 564-567, 2001.
[4] A.A.K. Jeng and R.H. Jan, "An Adjustable Structure for Topology Control in Wireless Ad Hoc Networks," *Proc. Int'l Conf. Wireless Network Comm. and Mobile Computing,* 2005.
[5] A.A.K. Jeng and R.H. Jan, "The R-Neighborhood Graph: an Adjustable Structure for Topology Control in Wireless Ad Hoc Networks," *IEEE Trans. Parallel and Distributed Systems,* vol. 18, no. 4, pp. 536-549, Apr. 2007.
[6] A.A.K. Jeng and R.H. Jan, "An Adaptive Topology Control Scheme for Energy-Efficient Routing in Mobile Ad Hoc Networks," *Proc. IEEE Wireless Comm. and Networking Conf.,* 2007.
[7] V. Rodoplu and T.H. Meng, "Minimum Energy Mobile Wireless Networks," *IEEE J. Selected Area in Comm.,* vol. 17, no. 8, pp. 1333-1344, Aug. 1999.
[8] Ns2 Simulator: http://www.isi.edu/nsnam/ns/, 2011.
[9] C.E. Perkins and P. Bhagwat, "Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers," *Proc. ACM SIGCOMM,* vol. 24, no. 4, pp. 234-244, 1994.
[10] L. Li and J.Y. Halpern, "Minimum-Energy Mobile Wireless Networks Revisited," *Proc. IEEE Int'l Conf. Comm.,* vol. 1, pp. 274-286, 2001.
[11] R. Wattenhofer and A. Zollinger, "XTC: A Practical Topology Control for Ad-Hoc Networks," *Proc. 18th Parallel and Distributed Processing Symp.,* pp. 26-30, 2004.
[12] D.M. Blough, M. Leoncini, G. Resta, and P. Santi, "The K-Neighbors Approach to Interference Bounded and Symmetric Topology Control in Ad Hoc Networks," *IEEE Trans. Mobile Computing,* vol. 5, no. 9, pp. 1267-1282, Sept. 2006.

Andy An-Kai Jeng received the BS degree in statistics from Tamkang University, Taiwan, in 2001, the MS degree in management information systems from the National Chi Nan University, Taiwan, in 2003, and the PhD degree in the computer science from National Chiao Tung University, Taiwan, in 2007, where he is currently a postdoctoral researcher. His research interests include wireless networks, distributed algorithm design and analysis, scheduling theory, and operations research.

Rong-Hong Jan received the BS and MS degrees in industrial engineering and the PhD degree in computer science from the National Tsing Hua University, Taiwan, in 1979, 1983, and 1987, respectively. From 1991-1992, he was a visiting associate professor in the Department of Computer Science, University of Maryland, College Park. He joined the Department of Computer and Information Science, National Chiao Tung University, in 1987, where he is currently a professor. His research interests include wireless networks, mobile computing, distributed systems, network reliability, and operations research. He is a senior member of the IEEE Computer Society.

Computer Networks 55 (2011) 2176–2195

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Release-time-based multi-channel MAC protocol for wireless mesh networks ☆

Andy An-Kai Jeng, Rong-Hong Jan *, Chi-Yu Li, Chien Chen

Department of Computer Science, National Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu 30005, Taiwan

ARTICLE INFO

Article history:
Received 21 August 2009
Received in revised form 12 July 2010
Accepted 28 February 2011
Available online 6 March 2011
Responsible Editor: L. Lenzini

Keywords:
Wireless mesh network
Multi-channel multi-interface
Medium access control
Channel selection

ABSTRACT

The wireless mesh network (WMN) has been considered one of the most promising techniques for extending broadband access to the last mile. In order to utilize multiple channels to increase the throughput in WMNs, a variety of multi-channel MAC (MMAC) protocols have been proposed in the literature. In particular, the *dedicated control channel* (*DCC*) *approach* can greatly simply many design issues in multi-channel environments by using a common control channel to exchange control signals. On the other hand, it allows each sender–receiver pair to dynamically select a data channel for their data transmission in an on-demand matter. However, the common control channel would become a bottleneck of the overall performance. Besides, the selection of data channels would be highly related to the final throughput. In this paper, we propose a new MMAC protocol, named the *release-time-based MMAC* (*RTBM*) to overcome the control channel bottleneck and data channel selection problems in the *DCC* approach. The *RTBM* consists of three major components: (1) *Control initiation-time predication* (*CIP*); (2) *Dynamic data-flow control* (*DDC*); (3) *Enhanced channel selection* (*ECS*). The *CIP* can predict a proper initiation time for each control process to reduce control overhead. The *DDC* can dynamically adjust the flow of each data transmission to fully exploit the channel bandwidth. The *ECS* can achieve a higher reusability of data channels to further enhance the throughput. Simulation results show that the *RTBM* can substantially improve the throughput in both single-hop and multi-hop networks.

1. Introduction

The wireless mesh network (WMN) has been considered one of the most promising techniques for extending broadband access to the last mile [1]. The WMN consists of a set of mesh access points (MAPs). A mobile station can access the network by connecting to a nearby MAP. Each MAP acts as a wireless router to forward traffic hop-by-hop to destinations. Thus, by deploying in such a fashion, a backhaul network can easily be built up without wired connection. Moreover, the network capacity can be substantially improved by using multiple channels. The IEEE 802.11b/g and 802.11a standards provide up to 3 and 12 orthogonal channels, respectively, in 2.4 GHz and 5 GHz spectrums. Nodes within the interference range of each other can transmit on different channels simultaneously to increase the throughput.

In order to utilize multiple channels in WMNs, a key issue is to design a *multi-channel medium access control* (*MMAC*) *protocol* [2] to handle operations at the data-link layer. However, due to the limitation that a wireless card (for most commercial

☆ This research was supported in part by the National Science Council, Taiwan, ROC, under grants NSC97-2219-009-006, and 97-2221-E-009-049-MY3.
* Corresponding author. Tel.: +886 3 5712121x31539.
E-mail address: rhjan@cis.nctu.du.tw (R.-H. Jan).

1389-1286/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.comnet.2011.02.017

devices) cannot perform on two different channels at a time, the design of many essential mechanisms, such as contention resolution, hidden terminal avoidance, and broadcasting support, could be much more difficult in comparison with a single-channel MAC (e.g. IEEE 802.11 DCF).

To overcome the above limitation, a prominent approach, called the *dedicated control channel* (*DCC*), was proposed [3–10]. In this approach, each node is equipped with a *control interface* and a *data interface*. The control interface is permanently fixed on a common channel (called *control channel*) for sensing or exchanging control signals. On the other hand, the data interface can switch among the remaining channels (called *data channels*) for data transmission. As shown in Fig. 1, if node u intends to transmit to node v, it first initiates a control process with v on the control channel to coordinate a data channel. Then, nodes u and v can communicate on that channel. Since the control channel is shared by all nodes, other competitors in the interference range of u or v can be aware of the coordination. Besides, a link-layer broadcast can be easily achieved by emitting on the control channel. Most importantly, since each node has an interface dedicated to listen to the control channel, nodes can exchange control signals without any time synchronization mechanism.

However, designing a *DCC*-based MMAC protocol confronts two major challenges:

(1) *Control channel bottleneck problem*: As described above, using a common control channel can greatly simplify the design of a MMAC. Nonetheless, if too many nodes contend on it, the control channel would become a bottleneck of the overall performance. As shown in Fig. 2(a), three sender–receiver pairs $\{u, v\}$, $\{x, y\}$, and $\{w, z\}$ are coordinating on the control channel ch_0. Besides, there are three data channels with bandwidth B. Ideally, the throughput can achieve $3B$ if each pair transmits on a different data channel. However, since the time required for a control process is about a half of a data transmission in this example, at most two data channels can be utilized at the same time, resulting in a lower throughput of $2B$. In other words, the throughput is saturated by the control channel's bandwidth. The bottleneck problem will become more serious as the number of data channels, data rates, or node's density increases [3].

(2) *Data channel selection problem*: The *DCC* approach allows each sender–receiver pair to select a data channel in an on-demand matter. But, if the selection strategy were not carefully designed in concern with the reusability of channels, the throughput would be lower. As shown in Fig. 2, assume that the sets of channels which are free to nodes y, x, u, and v are as specified in Fig. 2(b). Clearly, the transmissions from y to x and from u to v can be active simultaneously on different channels. But, if nodes u and v do not consider channel statuses at nearby nodes, they may select ch_2 in prior to nodes y and x, further degrading the throughput from $2B$ to B.

In this paper, we propose a new MMAC protocol to resolve the two challenges in the *DCC* approach, composing of the following three components:(1) *Control initiation-time prediction* (*CIP*) reduces the control overhead by properly predicting the initiation time of each control process to avoid unsuccessful channel coordination; (2) *Dynamic data-flow control* (*DDC*) dynamically adjusts the amount of flow in each data transmission to balance the congestion in the control and data

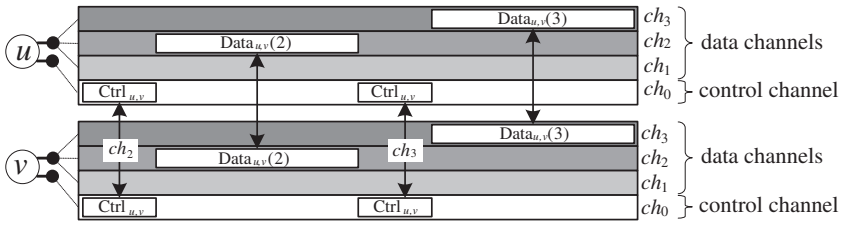


Fig. 1. Dedicated control channel approach.

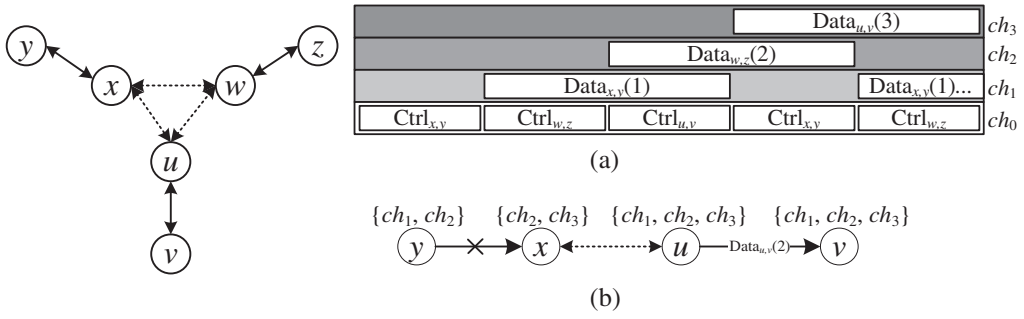


Fig. 2. (a) Control channel bottleneck problem and (b) data channel selection problem.

channels. (3) *Enhanced channel selection* (*ECS*) improves the reusability of data channels by selecting a channel that adds the least total delay to the starting times of the transmissions at nearby nodes. The design of these components is primarily based on exploiting the release times of channels and interfaces. Hence, we named this protocol the *Release-Time-Based MMAC* (*RTBM*).

The rest of this article is organized as follows: In Section 2, we compare different types of MMAC approaches and review existing protocols related to the *DCC* approach. Section 3 describes the basic operation of the *RTBM* protocol. The detailed design of each component is present in Section 4. In Section 5, we conduct a series of simulations to evaluate the performance. Concluding remarks are given in the last section.

2. MMAC approaches and related works

This section consists of two parts. First, we compare the pros and cons of different MMAC approaches. Next, existing protocols related to the *DCC* approach are reviewed.

2.1. MMAC approaches

A variety of MMAC protocols has been proposed in the literature. According to the way of coordinating data channels [1,2,23], existing protocols can be classified into the *channel fixed*, *receiver based*, *hybrid*, *spite control phase*, *dedicated control channel, and single/parallel rendezvous* approaches.

In the *channel fixed approach* [11,12], each interface is fixed on a channel permanently or for a long period of time. If a node *A* wants to communicate with a neighboring node *B*, they must have some interfaces fixed on the same channel. Then, A can send packet (e.g. RTS/CTS/DATA) directly to *B*, without additional control overheads to find a common channel. Besides, since the channels are fixed, contention within each group of interfaces on the same channel can be resolved by a standard MAC (e.g. 802.11 DCF). However, the fixed structure also limits the ability of using diverse channels. The number of channels that can be used by a node is limited by the number of its interfaces. In addition, if there is no common channel shared by two adjacent nodes, their traffic has to be relay through a longer path.

In contrast, the *receiver based approach* allows each node to utilize diverse channels with a single interface [3,13]. Each node is specified a channel in advance. For communication, a node u can connect with any nearby node v by just turning its interface the channel specified to v. However, since nodes are always fixed on the specified channels, some control signals may lose. As shown in Fig. 3(a), nodes *A*, *B*, *C*, and *D* are specified ch_1, ch_2, ch_2, and ch_3, respectively. At the beginning, node *A* intents to transmit to *B*. So, it exchanges the RTS-CTS with *B* using *B*'s channel. However, at the moment, node *C* has turned its interface to ch_3 for transmitting to *D*. Thus, node *C* cannot be aware of the CTS from *B*. As a result, after the current transmission, node *C* may content for ch_2 and incur a collision at *B*. It is the so called *multi-channel hidden terminal problem* [2]. Even worse, the lost signals may cause link failure. In Fig. 3(b), node *C* is sending requests to *B* using *B*'s channel. But, node *B* has turned its interface to ch_1 so that it cannot be aware of the request from *C*. Thus, node C may keep sending the RTSs until it falsely concludes that its link to *B* has broken. It is the so-called *deafness problem* [2].

The *hybrid approach* employs two interfaces to overcome the problems in both channel fixed and receiver based approaches [23,24]. One interface is fixed on a specified channel for receiving tasks, and the other interface can be dynamically switched on the channel of the fixed interface of the intended receiver. In this way a node can utilized diverse channel via its switchable interface and keep trace of control packets via its fixed interface. Moreover, singe channels of fixed interfaces are rarely changes, a channel switching can be made immediately without any coordination. However, this approach has two drawbacks. First, although a node can utilize more diverse channels, not limited by the number of its interface, the commu-

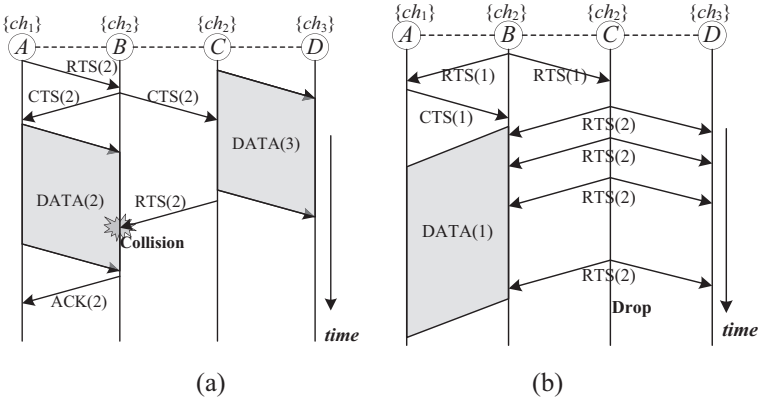


Fig. 3. Receiver based approach (a) multi-channel hidden terminal problem and (b) deafness problem.

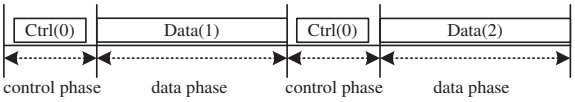


Fig. 4. Split control phase approach.

nication channel between two nodes is still fixed, which is not flexible when the channel of the intended receiver is highly interfered. Second, a link-layer broadcast is hard to be implemented. A node should broadcast the same message on all channels to ensure the delivery.

The *split control phase approach* divides each beacon interval into a *control phase* and a *data phase* [14,15]. As shown in Fig. 4, all nodes periodically rendezvous to a common channel in the control phase to sense carriers and coordinate channels for the subsequent transmissions in the data phase. In this approach, nodes have the most flexibility to use diverse channels during the data phase. Besides, the exchange of control packets and link-layer broadcast can be easily achieved during the control phase. Nonetheless, the phase alignment relies on strict time synchronization so that it is hard to implement in practice.

The *dedicated control channel approach* does not require any time synchronization mechanism. With the control interface, each node can access the control channel at anytime. However, as mentioned in Section 1, the throughput could be limited by the bandwidth of control channel and influenced by the selection of data channels. Besides, the dynamic channel selection means that more control overheads are required for the channel coordination. Hence, in this paper we focus on solving these challenges.

Finally, in the *common rendezvous approach* [16], nodes not exchanging data hop through all channels synchronously. A pair of nodes stops hopping as soon as they make an agreement for transmission and rejoin the common hopping pattern subsequently after transmission ends. In comparison with the *DCC*, this approach can make use of all the channels for data exchange and requires only one interface per node. However, it does not resolve the control bottleneck problem since only one pair of nodes can make an agreement at the same time. The *parallel rendezvous approach* [17,18] overcomes this problem by assigning different hopping sequences to nodes. In SSCH [17], each node picks multiple sequences and follows them in a time-multiplexed manner. When node *A* wants to talk to node *B*, *A* waits until it is on the same channel as *B*. The McMAC [18] improves SSCH by allowing a sender to temporally hop to the sequence of the receiver to avoid waiting. The major problem in these rendezvous approaches is that they may incur large switching delay for channel hopping, and each node requires synchronization mechanisms to track the hopping sequence of the others. Besides, a link-layer broadcast is hard to be implemented in a hopping fashion.

2.2. Existing protocols for DCC approach

The concept of using separated channels or special devices (e.g. busy tones) to improve medium access control has been extensively studied in works such as [19–22], but these researches consider only one data channel, i.e. designed for single-channel MAC. The first MMAC protocol under the *DCC* approach was presented in [3] and named the *dynamic channel assignment* (*DCA*). In this protocol, each node maintains a *free channel list* (*FCL*) to record unused data channels. As a node u intends to transmit to a node v, it first sends a RTS to v carrying its *FCL*. Then, node v compares the received *FCL* with its own *FCL* to select a common free channel. If there is any, the selected channel will be replied to u using a CTS. Once received the CTS, node u emits a reserve-to-send (RES) to inhibit other nodes from using the same channel. Meanwhile, nodes u and v can start to exchange the DATA and ACK frames on the selected channel. Integration with the power control technique was proposed in [4].

Compared with the IEEE 802.11 DCF, the *DCA* needs an additional control frame (e.g. RES) to reserve the channel selected at the receiver's side. To avoid such overhead, the protocol in [5] suggests that each sender can firstly propose a free channel in the RTS. If the channel is free to the receiver, the data transmission can be started immediately upon received the CTS. Otherwise, it follows the same way in [3]. A similar protocol appears in [6], where the proposed channel in the RTS will be replied by a reply-to-RTS (RRTS) that indicates whether the channel is free or not. The negotiation will continue until a common free channel is found. Nonetheless, these protocols may spend more control bandwidth if the proposed channel is not accepted by the receiver.

Another way to relieve the bottleneck problem is by using multiple control channels. Koubaa [7] showed that the number of control channels required to achieve the maximal throughput is a function of the available channels and packet size. For instance, with 12 channels and packet size of 1024 bytes, providing three control channels is optimal. Likewise, the protocol in [8] employs an extra channel for replaying ACKs to improve the channel reusability. By replying the ACK in a separated channel, a sender can be active simultaneously with its hidden terminals. Although using multiple control channels is beneficial, coordinating on different channels could be more complicated.

The channel selection strategy in the *DCA* [3] is simply to find a communicable channel. Each sender–receiver pair coordinates a channel that is free at the two sides. If there are multiple choices, one channel will be chose at random. The strategy was slightly improved in [9,10]. In [9], the channel with the least received power will be chosen to avoid potential interference. Similarly, in [10], the most robust channel will be selected according to the carrier-to-interference ratio. In other words, [9,10] are concerned about not only the communicability, but also the *quality* of the selected channel.

3. Protocol description

The *RTBM* is primarily based on the *DCA* protocol [3]. Similarly to the *DCA*, each sender–receiver pair has to exchange the RTS-CTS-RES sequence on the control channel to coordinate and reserve a data channel for the subsequent transmission of the DATA and ACK messages. The *RTBM* further incorporates with the following components to resolve the control channel bottleneck and data channel selection problems in the *DCC* approach:

(1) *Control initial-time prediction* (*CIP*): In the *DCA* [3], each sender has to initiate a control process (e.g. the RES-CTS-RES) with the intended receiver to coordinate a data channel that is free at two sides. If the coordination succeeds, the data transmission (e.g. the DATA-ACK) can be started on the selected channel. However, if the coordination fails, the expensed control bandwidth and time (including backoff timer, transmission time, inter-frame spaces, propagation delay, and processing time) are wasted. The *CIP* can avoid unsuccessful channel coordination by properly predicting the initiation time of each control process. The prediction will jointly consider the channels and interfaces statuses at the sender and receiver.
(2) *Dynamic data-flow control*: (*DDC*): As shown in Fig. 2(b), the bottleneck problem from the common control channel would lower down the utilization of data channels. To breakthrough this limitation, an intuitive way is to expend the *data flow* (i.e. the number of packets to be sent) for each control process to increase the bandwidth usage [20]. However, if too many packets were transmitted with only a few control processes, the data channels could be occupied by some node pairs for a long period of time, which may instead incur unfairness problem to other competitors who are intended to access the data channels. Ideally, the above problems can be optimally solved by adjusting the data flow such that both the control and data channels are fully utilized. For example, in Fig. 2(b), if node u sent 1.5 times of packets to node v by each control process, all channels can be fully exploited. However, due to the dynamicity in wireless environments, such as the variation in network traffic or topology, the optimal setting would be varied from time to time. The *DDC* component can dynamically adjust the amount of flow in each data transmission to balance the congestion in the control and data channels.
(3) *Enhanced channel selection*: (*ECS*): The selection strategy in the *DCA* [3] is simply to find a communicable data channel that is free to sender and receiver. If there are multiple choices, one channel will be chose at random. Succeeding works in [9,10] also consider the quality of the selected channel, but none of them concerns the influence to nearby transmissions. The *ECS* can improve the reusability of data channels. Each pair will cooperatively coordinate a free data channel that will add the least total delay to the starting times of potential transmissions at nearby nodes. In other words, the selected channel is not only communicable, but also has the least influence to the transmission opportunities of nearby nodes.

In the following, we first define the statuses and symbols used in this protocol. Next, the basic operation is described. We will focus here on how the *RTBM* interacts with the three components and defined statuses. The detail design of each component will be presented in the next section.

3.1. Statuses and symbols

We assume that the network has a set of orthogonal channels $\mathscr{H} = \{h | h = 0, 1, 2, \ldots, H\}$. Each node has a control interface and a data interface. The first channel ($h = 0$) is for control purpose, and the remaining channels ($h = 1, 2, \ldots, H$) can be used for data transmission. A channel (or interface) is *released* when it can be used for a transmission or reception. Each node u has the following statues:

- $ch\_rel\_time(u, h)$: the *release time* of the hth channel at u, $h = 0, 1, \ldots, H$;
- $if\_rel\_time(u)$: the *release time* of the data interface at u.

In addition, each node u maintains a *channel release time table* (CRT_u) and an *interface release time vector* (IRV_u) to record the channels' and interfaces' statuses at nearby nodes. The fields $CRT_u(v, h)$ and $IRV_u(v)$, respectively, keep track of the latest statuses about the $ch\_rel\_time(v, h)$ and $if\_rel\_time(v)$ for each neighboring node v. Moreover, a separate queue is created for each 1-hop destination. The purpose is to avoid head-of-line blocking if two or more packets need to be sent for the same node. Other time symbols that will be used in our protocol are listed in Table 1.

3.2. Basic operation

The basic operation of the *RTBM* is illustrated in Fig. 5. Consider node u. Let us see how node u operates with other nodes in this protocol. In the beginning, node u creates a separate queue for nodes a and v. To avoid starvation, when some packets arrived in queues the destination with the oldest packet will be chosen as the target. Assume that node u has decided to transmit to node v. First of all, it applies the *CIP* to predict a *control initiation time* for v, denoted as $ctrl\_ini\_time(u, v)$. The

Table 1
Time symbols used in the *RTBM* protocol.

Symbols	Meanings
T_{curr}	Current time of a node
T_{DATA_i}	Time to transmit the i-th data packet in a queue
T_{RTS}	Time to transmit a RTS frame
T_{CTS}	Time to transmit a CTS frame
T_{RES}	Time to transmit a RES frame
T_{ACK}	Time to transmit an ACK frame
BF	Remaining backoff time
τ	Maximal propagation delay

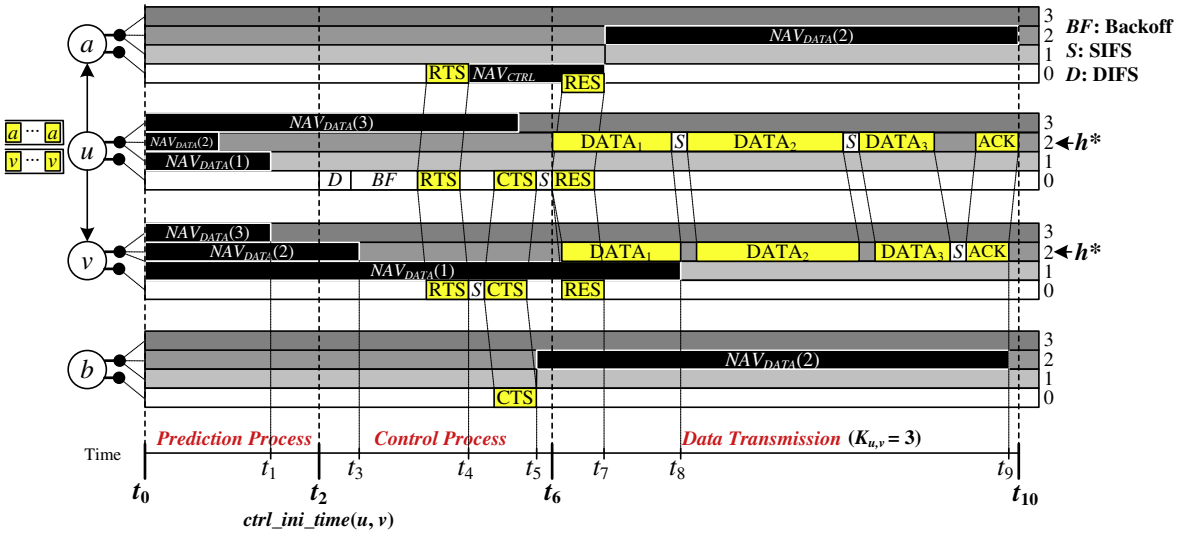


Fig. 5. Basic operation of the *RTBM* protocol.

time indicates when node u can successfully initiate a control process with v. As shown in Fig. 5, $ctrl\_ini\_time(u, v) = t_2$. The prediction process will continue before the control process is actually started.

At $ctrl\_ini\_time(u, v)$, node u starts the following control process. It first applies the *DDC* to determine the number of data frames to be sent for v, denoted as $K_{u,v}$. According to the $K_{u,v}$, the network allocation vector required to exchange the DATA and ACK messages can be set as

$$NAV_{DATA} = \sum_{i=1}^{K_{u,v}} (T_{DATA_i} + SIFS) + T_{ACK} + 2\tau,$$

where T_{DATA_i} is the time to transmit the ith data packet in the queue for v. As shown in Fig. 5, $K_{u,v} = 3$ and $NAV_{DATA} = t_{10} - t_6$. So, node u will transmit the first three packets to v during t_6 to t_{10}. Next, node u applies the *ECS* to evaluate the cost for transmitting on each data channel h. The cost, denoted as $\Delta_u(h)$, is the total increment to the starting times of possible transmissions at u's nearby nodes as if node u transmitted on channel h for a period of NAV_{DATA}. Then, following the IEEE 802.11 backoff mechanism, if there was no carrier on the control channel in a *DIFS* plus the *BF*, node u sends a RTS to v, containing the NAV_{DATA} and the $\Delta_u(h)$, for any $h = 1, 2, \ldots, H$.

Once received the RTS, node v also applies the *ECS* to evaluate the cost $\Delta_v(h)$, for each data channel h, i.e. the total increment to the starting times of possible transmissions at v's nearby nodes as if node v transmitted on channel h for a period of $NAV_{DATA} - \tau$. Combining the costs evaluated from the two sides, the total cost for communicating between u and v on a channel h, denoted as $\Delta_{u,v}(h)$, is the sum of the $\Delta_u(h)$ and $\Delta_v(h)$. The data channel that will be released to u and v and has the least $\Delta_{u,v}(h)$ will be selected. If such a channel can be found, denoted as h^*, node v has to update the following statues for the forthcoming data transmission on h^* such that

$$if\_rel\_time(v) = ch\_rel\_time(v, h^*) = T_{curr} + 2SIFS + T_{CTS} + NAV_{DATA}.$$

For example, at $T_{curr} = t_4$, node v sets $if\_rel\_time(v) = t_9$, and $ch\_rel\_time(v, h^*) = ch\_rel\_time(v, 2) = t_9$. Besides, node v sets two timers as follows in its CTS.

$$if\_duration(v) = \max\{0, if\_rel\_time(v) - T_{curr} - (SIFS + T_{CTS} + \tau)\},$$

$$ch\_duration(v,h) = \max\{0, ch\_rel\_time(v,h) - T_{curr} - (SIFS + T_{CTS} + \tau)\}, \quad h = 1,2,\ldots,H.$$

The $if\_duration(v)(ch\_duration(v,h))$ indicates the amount of time the data interface (data channel h) of node v will be reserved. For example, at $T_{curr} = t_4$, node v sets

$$if\_duration(v) = \max\{0, t_9 - t_4 - (SIFS + T_{CTS} + \tau)\} = t_9 - t_5,$$

$$ch\_duration(v,1) = \max\{0, t_8 - t_4 - (SIFS + T_{CTS} + \tau)\} = t_8 - t_5$$

$$ch\_duration(v,2) = \max\{0, t_9 - t_4 - (SIF + T_{CTS} + \tau)\} = t_9 - t_5,$$

$$ch\_duration(v,3) = \max\{0, t_1 - t_4 - (SIFS + T_{CTS} + \tau)\} = 0.$$

After waiting a *SIFS*, node v replies a CTS to u, containing the h^* (if there is any), $if\_duration(v)$, and $ch\_duration(v,h)$, for any $h = 1, 2, \ldots, H$.

Once received the CTS, if a channel h^* was indicated, node u also updates

$$if\_rel\_time(u) = ch\_rel\_time(u,h^*) = T_{curr} + ch\_duration(v,h^*) + \tau,$$

for the forthcoming data transmission on h^* and sets two timers as follows in its RES.

$$if\_duration(u) = \max\{0, if\_rel\_time(u) - T_{curr} - (SIFS + T_{RES} + \tau)\},$$

$$ch\_duration(u,h) = max\{0, ch\_rel\_time(u,h) - T_{curr} - (SIFS + T_{RES} + \tau)\}, \quad h = 1,2,\ldots,H.$$

After waiting a *SIFS*, the h^* is rebroadcasted to nearby nodes along with a RES to reserve the channel. Meanwhile, node u can start to send $K_{u,v}$ packets to node v via channel h^*. Finally, node v replies an ACK to node u at the end of the data transmission.

On the other hand, once an irrelevant node x (e.g. nodes a and b) received the CTS or RES from a node y (e.g. nodes u and v), if a channel, i.e. h^*, was indicated inside, node x has to inhibit itself from using the same data channel by setting

$$ch\_rel\_time(x,h^*) = \max\{ch\_rel\_time(x,h^*), T_{curr} + ch\_duration(y,h^*)\}.$$

Note that the NAV_{DATA} has been implicated in both the $ch\_duration(v,h^*)$ and $ch\_duration(u,h^*)$. Hence, it is neither in the CTS nor in the RES.

Moreover, to prevent nodes from disrupting the control process between u and v, an extra *NAV* is employed on the control channel. More precisely, once an irrelevant node x (e.g. node a) received the RTS from u, it has to block its control channel for a period of time by setting

$$ch\_rel\_time(x,0) = T_{curr} + NAV_{CTRL},$$

where $NAV_{CTRL} = 2SIFS + T_{CTS} + T_{RES} + 2\tau$, specifying the amount of time the control channel will be for the RTS-CTS-RES dialogue.

Lastly, to maintain the statuses of nearby nodes, once a node x (e.g. nodes a, b, u, and v) receives a CTS or RES from a node y (e.g. nodes u and v), it has to update its *channel release time table* (CRT_x) and its *interface release time vector* (IRV_x) as follows.

$$IRV_x(V) = T_{curr} + if\_duration(y),$$

$$CRT_x(y,h) = T_{curr} + ch\_duration(y,h), \quad \text{for any} \quad h = 1,2,\ldots,H.$$

The above processes are in normal situation which could be interrupted in some circumstances. First, node u detected some control signals before the *BF* expires. In this case, the RTS will not be sent out by u. Second, the RTS or CTS was collided or the control channel of v was blocked by a NAV_{CTRL}. In this case, node u will not receive the CTS from v within the timeout period of $SIFS + T_{CTS} + 2\tau$. Third, there was no data channel h^* in indicated in the CTS from v. In this case, the data transmission cannot be started by u. When some of these cases happened, the following process should be done: If the retry limited is not reached, node u has to terminate the current control process, apply again the *CIP*, and restart the next attempt at the new $ctrl\_ini\_time(u, v)$; otherwise, node u has to abort any process with v an select a new target from its queues.

4. Components design

In this section, we present the detail design of the three components in our protocol.

4.1. Control initiation-time prediction

The *CIP* is designed to reduce unsuccessful channel coordination that would incur redundant control overhead. It is achieved by properly predicting the initiation time of each control process, based on the information about the release times

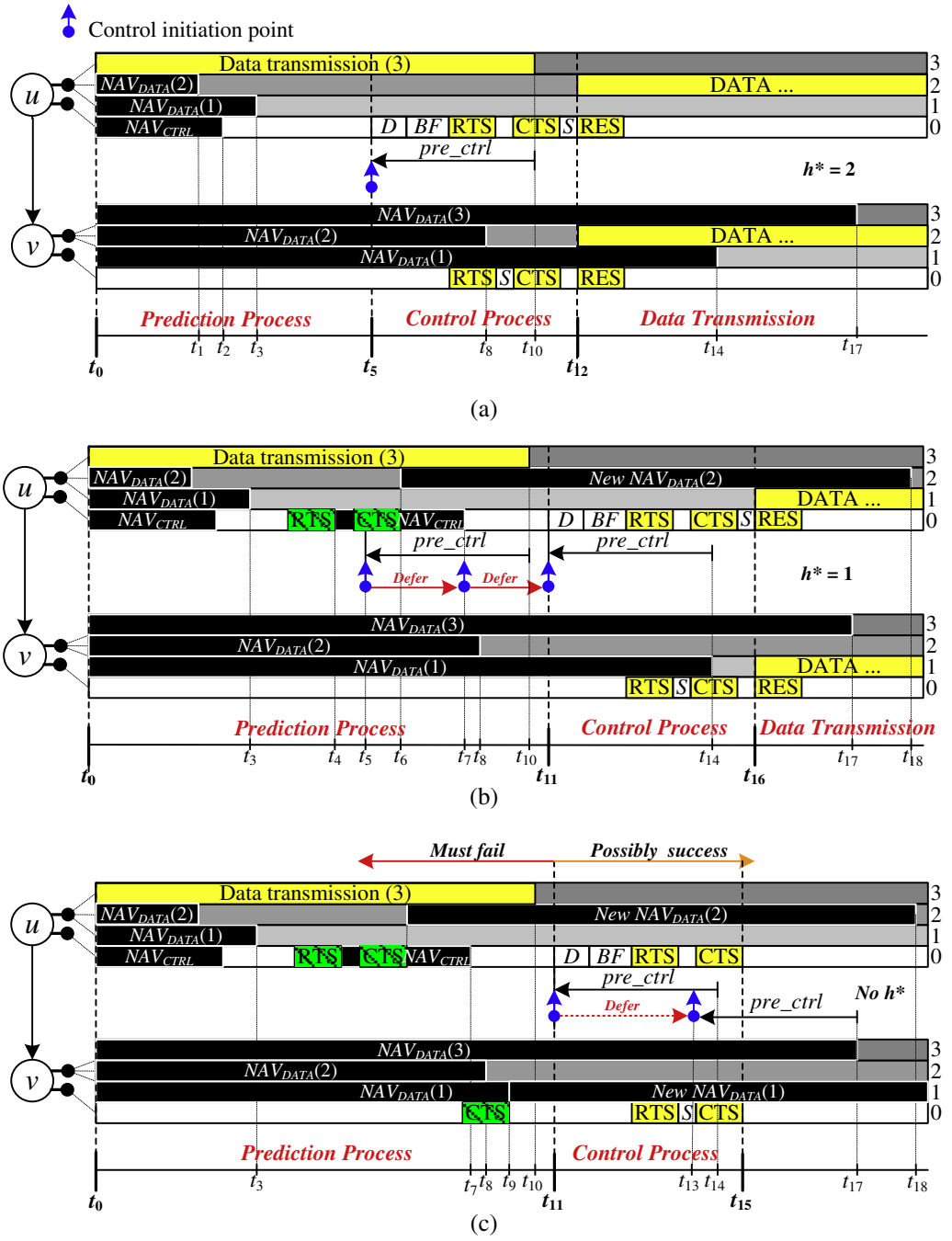


Fig. 6. Control initiation prediction process (a) without deferral, (b) with deferral and (c) invalid scenario.

of channels and interfaces at sender and receiver. To explain the idea, let us see the example in Fig. 6(a).[1] In the beginning, node u has some packets for v at t_0. At the moment, the release times of ch_1, ch_2, and ch_3 at u and v are (t_3, t_1, t_{10}) and (t_{14}, t_8, t_{17}), respectively. We know that two nodes can communicate with each other only if at least one data channel has been released to both of them. Therefore, node u cannot send any data frame to v before $t_8 = \min\{\max\{t_3, t_{14}\}, \max\{t_1, t_8\}, \max\{t_{10}, t_{17}\}\}$, i.e., the release time of ch_2. Moreover, no data frame can be transmitted until the data interfaces of both u and v have been released. With the two considerations, we define the *link release time* of u and v as

[1] To simplify our presentation, the propagation τ will not be drawn in the hereafter figures.

$$link\_rel\_time(u,v) = \max\left\{\begin{array}{l}\min\limits_{h=1,\ldots,H}\left\{\max\left\{\begin{array}{l}ch\_rel\_time(u,h),\\ CRT_u(v,h)\end{array}\right\}\right\},\\ if\_rel\_time(u),\\ IRV_u(v)\end{array}\right\},$$

which means that all resources required for communicating on the virtual link between u and v will be released at *link_rel_time*(u, v). In Fig. 6(a), the link release time of u and v at t_0 is $t_{10} = \max\{t_8, t_{10}, t_0\}$. Clearly, it is the earliest possible time that u can start a data transmission with v.

With the observation, now we discuss when node u can initiate the control process with v. In the *DCC* approach, since control frames (e.g. RTS/CTS/RES) and data frames (e.g. DATA/ACK) are exchanged using different interfaces and channels, a control process can be started earlier before the virtual link (data channels and data interfaces) is released. Besides, the data frames can be sent out once node u has received the CTS from v for a *SIFS*. Therefore, if node u intends to transmit at the earliest time t_{10}, it should initiate the control process in advance at $t_{10} - (DIFS + BF + T_{RTS} + T_{CTS} + 2SIFS + 2\tau)$. However, the backoff timer *BF* should be removed from the composition of $(DIFS + BF + T_{RTS} + T_{CTS} + 2SIFS + 2\tau)$. Otherwise, other senders waiting for the same resource (u's data interface) may send their RTSs at the same time with u's RTS, and thus incur collision. In addition, node u cannot perform any control process until the control channel is released. Combining these facts, the *control initiation time of u and v* is defined as

$$ctrl\_ini\_time(u,v) = \max\left\{\begin{array}{l}link\_rel\_time(u,v) - pre\_ctrl,\\ ch\_rel\_time(u,0)\end{array}\right\},$$

where $pre\_ctrl = DIFS + T_{RTS} + T_{CTS} + 2SIFS + 2\tau$, denoting the preprocessing time. In this example, node u can apply this function at t_0 and predict that the initiation time is $t_5 = \max\{t_{10} - pre\_ctrl, t_2\}$.

Note that, before the control process is actually started, the predicted time could be updated if the *ch_rel_time*$(u, 0)$ or *link_rel_time*(u, v) is changed. When such event occurs, the control process is deferred to the updated initiation time. For example, in Fig. 6(b), node u received a RTS at t_4 that changes the *ch_rel_time*$(u, 0)$ to t_7. So, the *ctrl_ini_time*(u, v) is deferred from t_5 to t_7. Furthermore, node u received a CTS at t_6, indicating that ch_2 will be released at t_{18}. Hence, the *link_rel_time*(u, v) is changed from t_{10} to t_{14}. Accordingly, the *ctrl_ini_time*(u, v) is further deferred from t_7 to t_{11}.

In addition, since the records in the CRT_u and IRV_u are not always the newest, the predicted time could be invalided. As shown in Fig. 6(c), node v received a CTS with an updated *ch_duration*$(v, 1)$ at t_9 and the *ch_rel_time*$(v, 1)$ is changed accordingly. Normally, the control initiation time should be deferred from t_{11} to t_{13} to response to the change. But, since node v does not send any control frame during t_9 to t_{11}, the change is invisible to u. As a result, node u will still initiate the control process at t_{11} and the control process may fail. However, the *CIP* does ensure that any control process initiated before t_{11} must fail, because for any channel h (or interface), if it is not released at $CRT_u(v, h)$ (or $IRV_u(v)$) to u, it is also not released at *ch_rel_time*(v, h) (or *if_rel_time*(v)) to v. Thus, the *CIP* can help nodes to avoid unsuccessful channel coordination.

4.2. Dynamic data-flow control

The *DDC* can dynamically adjust the number of data packets being sent in each data transmission to balance the congestion in the control and data channels. The *DCC* maintains a variable $K_{u,v}$ for each 1-hop node v specifying the number of data packets that will be sent in the next transmission from nodes u to v. $K_{u,v}$ is initiated as 1 and will be dynamically adjusted according to the idle statuses on the control and data interfaces.

Take a look at Fig. 7 to explain this idea. There are three sender–receiver pairs of $\{u, v\}$, $\{x, y\}$, and $\{w, z\}$ with data flows sustained during $[t_0, t_3]$, $[t_1, t_3]$, and $[t_1, t_2]$, respectively. Besides, nodes u and v are in the interference range of nodes x and w,

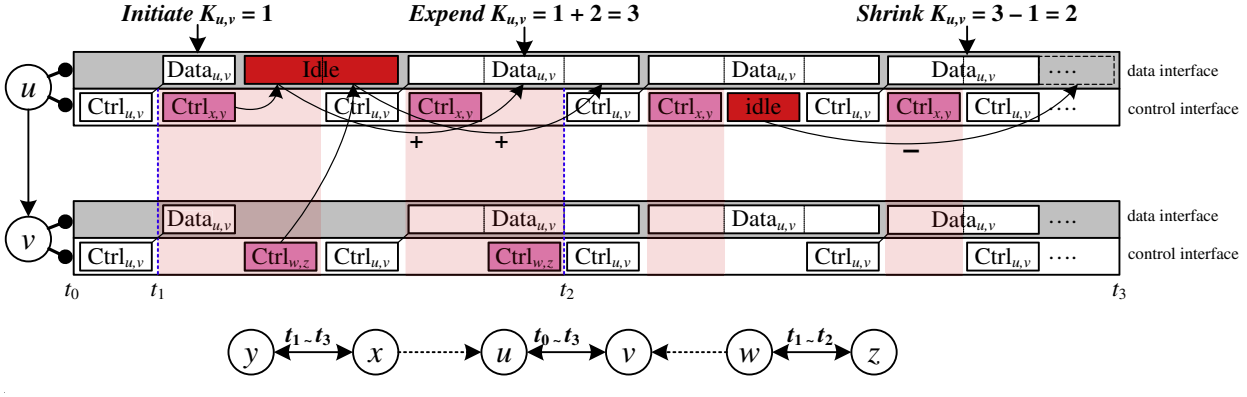


Fig. 7. Basic concept of the *DDC* component.

receptively. Assuming that the time required for a control process is equal to the time required for sending one data packet, we show how *DDC* adjusts $K_{u,v}$. In the beginning, $K_{u,v} = 1$. Thus, node u sends one data packet to v. Then, node u intends to send the next packet to v. However, at this moment, the control processes for other two flows have started, so node u cannot initiate a control process immediately with v. Consequently, u's data interface experiences a period of idle status before the next data transmission starts. In order to mitigate such influence from control channel, the *DDC* expends $K_{u,v}$ in proportion to the idle time experienced from u's data interface (i.e. $K_{u,v} = 1 + 2 = 3$) to increase the utilization of data channels for the subsequent data transmission to v. As shown in this example, by the end of the second data transmission, the next data transmission can be started immediately without any idling. Nonetheless, during the third data transmission, since the data flow between w and z has finished, the control channel becomes less congested at v, resulting in a period of idle status in u's control interface before the next control process starts. In order to reflect such fact, the *DCC* shrinks $K_{u,v}$ in proportion to the idle time experimented from u's control interface (i.e. $K_{u,v} = 3 - 1 = 2$). Finally, $K_{u,v}$ converges to 2 and the congestion in control and data channels are balanced. The process will continue if any further change occurred.

In summary, $K_{u,v}$ is adjusted according to two idle timers: the *data interface idle time*, denoted as $dif\_idle\_time(u, v)$, and the *control interface idle time*, denoted as $cif\_idle\_time(u, v)$. Therefore, below we give a more realistic example to show how an idle status occurs in the *RTBM*. Then, the two idle timers and adjusting rule are formally defined. Lastly, we provide an online algorithm for efficient maintenance.

In Fig. 8(a), node u has some packets for v at t_0. At the moment, node u finds that the link release time $link\_rel\_time(u, v) = t_4$ and control initial time $ctrl\_rel\_time(u, v) = t_4 - pre\_ctrl = t_2$. Ideally, if there is no contention before t_2, the data transmission can start at $t_5 = t_2 + pre\_ctrl + BF$ (recall that *BF* is not in the *pre_ctrl*). However, at t_1 an irrelevant RTS is detected, so node u has to update $ctrl\_ini\_time(u, v) = \max\{t_4 - pre\_ctrl, t_1 + NAV_{CTRL}\} = \max\{t_2, t_3\} = t_3$. Consequently, the data transmission should be deferred from t_5 to $t_7 = t_3 + pre\_ctrl + BF$. The deferral will result in a period of idle status from t_5 to t_7 on u's data interface. Therefore, we can estimate that $dif\_dile\_time(u, v) = t_7 - t_5$ at this time point.

Then, the control process starts at t_3 (see Fig. 7(b)). During the *BF*, another RTS is detected, so node u has to suspend the current process and update the new control initiation time to t_8. Consequently, the data transmission should be further deferred from t_7 to $t_{11} = t_8 + pre\_ctrl + BF$, resulting in a longer period of idle status on u's data interface from $[t_5, t_7]$ to $[t_5, t_{11}]$. However, during the sub period $[t_6, t_9]$, node u cannot transmit any data to node v even if the control channel is not congested, since all data channels (ch_1 and ch_3 of node u and ch_2 of node v) are blocked by some NAVs in this time interval. Thus, the $dif\_idle\_time(u, v)$ should be set as $dif\_idle\_time(u, v) = (t_{11} - t_5) - (t_9 - t_6)$ to reflect the deferral purely incurred by the control channel congestion. We call the time interval $[t_6, t_9]$ the *non-idle time*.

After a *DIFS* and *BF*, node u sends a RTS and waits for the CTS from v (see Fig. 8(c)). However, before the RTS arrived, the control channel of v has been blocked by another RTS so that u cannot obtain any reply from v before the end of the time out period at t_{10}. As a result, the starting time of the data transmission is further deferred from t_{11} to t_{12}. That is, the $dif\_idle\_time(u, v)$ should be updated as $(t_{12} - t_5) - (t_9 - t_6)$. This example shows that the idle status of u's data interface could be incurred by the control congestion at both node u and node v.

Continuing the example, in Fig. 8(d), the data transmission starts at t_{12} and ends at $t_{17} = t_{12} + NAV_{DATA}$. During this period, the control interface of u is released after sending the RES and reserved for the next control process after $t_{16} = t_{17} - pre\_ctrl$. In addition, the control channel is contended during the sub-period from t_{14} to t_{15}. Therefore, the control interface idle time is the sum of the two periods from t_{13} to t_{14} and from t_{15} to t_{16}, i.e. $cif\_idle\_time(u, v) = (t_{16} - t_{15}) + (t_{14} - t_{13})$.

Now, we formally define the two idle timers based on the above observations. Let $data\_tx\_time^*(u, v)$ and $data\_tx\_time(u, v)$ stand for the *earliest* and the *actual* starting times of a data transmission from u to v, respectively. At any time t, the data interface of u is *idle* if and only if

(i) $data\_tx\_time^*(u, v) \leqslant t < data\_tx\_time(u, v)$;
(ii) $link\_rel\_time(u, v) < t$.

For a transmission from u to v, $dif\_idle\_time(u, v)$ is the total time satisfying (i) and (ii). In Fig. 8, $data\_tx\_time^*(u, v) = t_5$, and $data\_tx\_time(u, v) = t_7$, t_{11}, and t_{12}, respectively, in Figs. 8(a)–(c). The period of time not satisfying (ii) is $[t_6, t_9]$, i.e. the non-idle time. Similarly, at any time t, the control interface of u is *idle* if and only if

(i) $data\_tx\_time(u, v) + T_{RES} + \tau < t < data\_tx\_time(u, v) + NAV_{DATA} - pre\_ctrl$;
(ii) $ch\_rel\_time(u, 0) - T_{type} < t$, where *type* is the type of the received control frame.

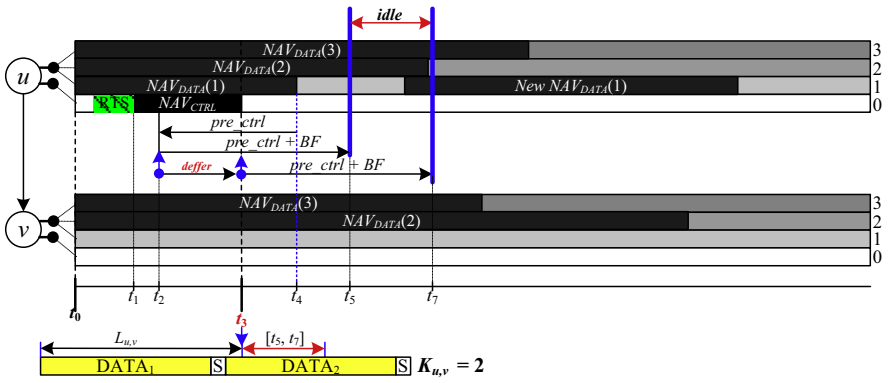
For a transmission from u to v, the $cif\_idle\_time(u, v)$ is the total time satisfying (iii) and (iv).

With the two idle timers, at the start point of each control process (i.e. t_3, t_8, and t_{10}), $K_{u,v}$ can be updated as

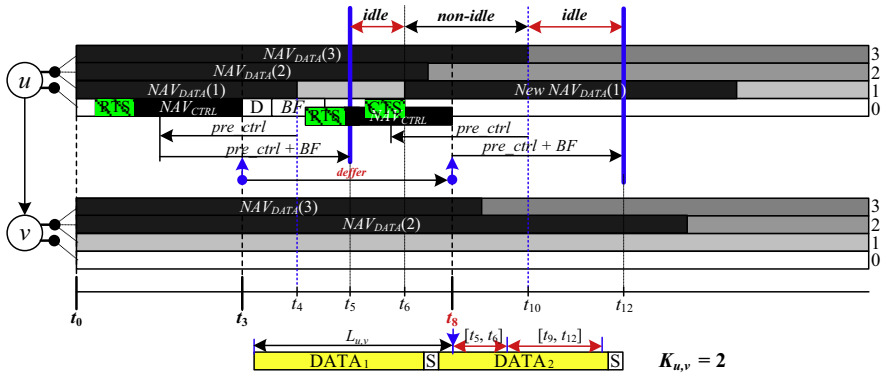
$$K_{u,v} = \min_{1 \leqslant k \leqslant Q_v} \left\{ \sum_{i=1}^{k} (T_{DATA_i} + SIFS) \geqslant L_{u,v} + dif\_dile\_time(u, v) \right\},$$

where Q_v is the number of data packets remaining for v, and $L_{u,v}$ is the time of the previous data transmission. As shown in Fig. 8(c), $K_{u,v} = 3$ at t_{10}, since

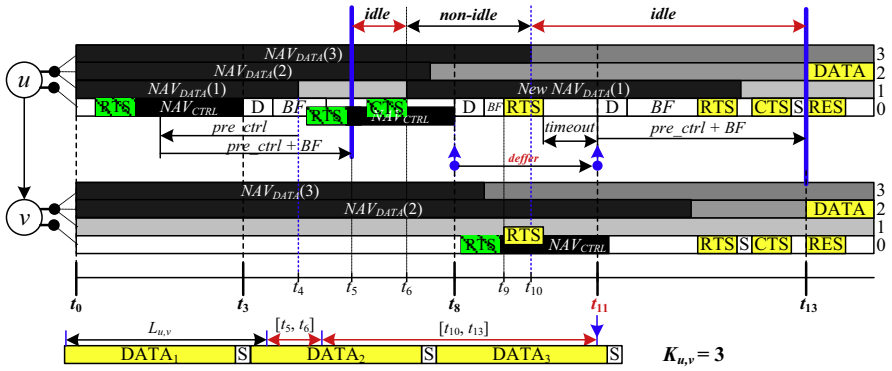
$$(T_{DATA_1} + SIFS) + (T_{DATA_2} + SIFS) + (T_{DATA_3} + SIFS) \geqslant L_{u,v} + (t_{12} - t_9) + (t_6 - t_5).$$



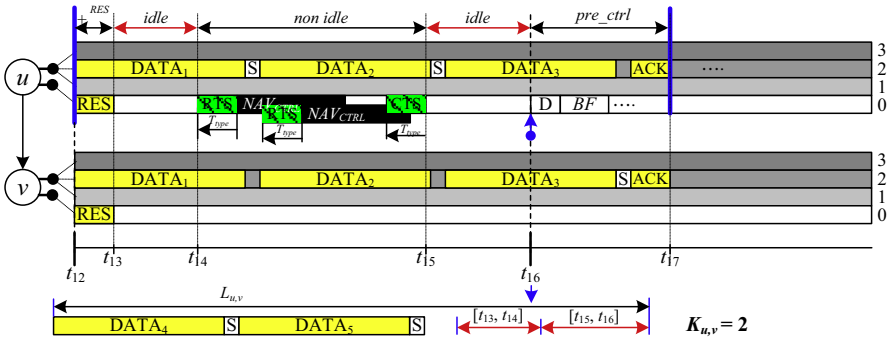
(a)



(b)



(c)



(d)

Fig. 8. Dynamic Data Aggregation (a) expending at t_3, (b) expending at t_8, (c) expending at t_{10} and (d) shrinking at t_{17}.

On the other hand, at the end point of each data transmission (i.e. t_{17}), $K_{u,v}$ can be updated

$$K_{u,v} = \max_{1 \leqslant k \leqslant Q_v} \left\{ \sum_{i=1}^{k} (T_{DATA_i} + SIFS) < L_{u,v} - cif\_dile\_time(u, v) \right\}.$$

Note that $L_{u,v}$ should be updated according to the time of the previous data transmission when used. As shown in Fig. 8(d), $L_{u,v} = t_{17} - t_{12}$, and $K_{u,v} = 2$ since

$$(T_{DATA_4} + SIFS) + (T_{DATA_5} + SIFS) < L_{u,v} - (t_{16} - t_{15}) + (t_{14} - t_{13}).$$

In accordant with the above definitions, an algorithm is presented below to maintain the $K_{u,v}$. Additional time symbols used in this algorithm are listed in Table 2. The corresponding values for Fig. 8 are shown in Table 3. The algorithm is designed in an online fashion and takes only constant time in each step. Thus, it is very efficient and practical.

Online algorithm: *DDC* Component
Initial:

$K_{u,v} = 1$; $L_{u,v} = T_{DATAi} + SIFS + T_{ACK} + 2\tau$;

Whenever node v is chosen as a target at T_{curr}*:*

$ctrl\_ini\_time^*(u, v) = \max\{T_{curr}, link\_rel\_time(u, v) - pre\_ctrl\}$;
$data\_tx\_time^*(u, v) = ctrl\_ini\_time^*(u, v) + pre\_ctrl + BF$;
$dif\_nid\_end(u, v) = data\_tx\_time^*(u, v)$;
$dif\_nid\_time(u, v) = 0$;

Before ctrl_ini_time(u, v):

Once the $link\_rel\_time(u, v)$ is deferred at T_{curr}, update
$dif\_nid\_time(u, v) = dif\_nid\_time(u, v) + link\_rel\_time(u, v) - \max\{T_{curr}, dif\_nid\_end(u, v)\}$;
$dif\_nid\_end(u, v) = \max\{data\_tx\_time^*(u, v), link\_rel\_time(u, v)\}$;

At ctrl_ini_time(u, v):

$data\_tx\_time(u, v) = ctrl\_ini\_time(u) + pre\_ctrl + BF$;
$dif\_idle\_time(u, v) = data\_tx\_time(u, v) - data\_tx\_time^*(u, v) - dif\_nid\_time(u, v)$;
$K_{u,v} = \min_{1 \leqslant k \leqslant Q_v} \{\sum_{i=1}^{k} (T_{DATA_i} + SIFS) \geqslant L_{u,v} + dif\_dile\_time(u, v)\}$;

At data_tx_time(u, v):

$cif\_nid\_time(u, v) = 0$;
$cif\_nid\_end(u, v) = data\_tx\_time(u, v) + T_{RES} + \tau$;

Before data_tx_time(u, v) + $NAV_{DATA} - pre\_ctrl$*:*

Once the $ch\_rel\_time(u, 0)$ is deferred at T_{curr}, update
$cif\_nid\_time(u, v) = cif\_nid\_time(u, v) + ch\_rel\_time(u, 0) - \max\{T_{curr} - T_{type}, cif\_nid\_end(u, v)\}$;
$cif\_nid\_end(u, v) = \max\{data\_tx\_time(u, v) + \sigma_{res}, ch\_rel\_time(u, 0)\}$;

At data_tx_time(u, v) + $NAV_{DATA} - pre\_ctrl$*:*

$cif\_idle\_time(u, v) = NAV_{DATA} - pre\_ctrl - cif\_nid\_time(u, v)$;
$L_{u,v} = \sum_{i=1}^{K_{u,v}} (T_{DATA_i} + SIFS) + T_{ACK} + 2\tau$;
$K_{u,v} = \max_{1 \leqslant k \leqslant Q_v} \{\sum_{i=1}^{k} (T_{DATA_i} + SIFS) < L_{u,v} - dif\_dile\_time(u, v)\}$;

4.3. Enhanced channel selection strategy

The *ECS* aims at improving the reusability of data channels. The main idea is based on exploiting the release times of channels and interfaces at neighboring nodes to select a free data channel that will cause the least influence to nearby transmissions.

Table 2
Time symbols in the online algorithm of the *DDC* component.

Status	Meaning
*ctrl_ini_time**(*u*, *v*)	Earliest control initiation time with *v*
dif_nid_time(*u*, *v*)	Non-idle time of *u*'s data interface before a data transmission to *v*
dif_nid_end(*u*, *v*)	End point of the non-idle status of *u*'s data interface before a data transmission to *v*
cif_nid_time(*u*, *v*)	Non-idle time of *u*'s control interface during a data transmission to *v*

Table 3
Corresponding values of the used statues for Fig. 8.

*ctrl_ini_time**(*u*, *v*); *ctrl_ini_time*(*u*, *v*)	*data_tx_time**(*u*, *v*); *data_tx_time*(*u*, *v*)	*dif_nid_time*(*u*, *v*); *cif_nid_time*(*u*, *v*)	*dif_nid_end*(*u*, *v*); *cif_nid_end*(*u*, *v*)	*dif_idle_time*(*u*, *v*) *cif_nid_time*(*u*, *v*)	$K_{u,v}$
t_2	t_5	0	t_5	0	1
t_3	t_7	0	t_5	$(t_7 - t_5)$	2
t_8	t_{11}	$(t_9 - t_6)$	t_9	$(t_{11} - t_5) - (t_9 - t_6)$	2
t_9	t_{12}	$(t_9 - t_6)$	t_9	$(t_{12} - t_5) - (t_9 - t_6)$	3
t_{17}	–	$(t_{15} - t_{14})$	t_{15}	$(t_{16} - t_{13}) - (t_{15} - t_{14})$	2

The concept is shown in Fig. 9. Nodes *a* and *c* are in the interference range of node *u*, and nodes *b* and *d* are in the interference range of node *v*. According to the specified NAVs, nodes *a*, *b*, *c*, and *d* cannot transmit to any other node, respectively, before t_6, t_9, t_2, and t_7, since prior to these time points, there is neither an interface nor a channel released (note that the interface of node *d* will be released by t_7). In other words, the earliest possible time to start a transmission from *a*, *b*, *c*, and *d* are t_6, t_9, t_2, and t_7, respectively. Assume that nodes *u* and *v* have decided to communicate on ch_1. See what happen to nearby nodes. Since ch_1 is not the first channel released to node *a*, node *a* can still transmit at t_6. Likewise, the earliest possible starting time of node *d* remains t_7. However, to nodes *c* and *d*, since ch_1 is no longer the first channel released to them, they cannot start any transmission before ch_2 and ch_3 are released to them at t_8 and t_9, respectively. As a result, the earliest possible starting times of *b* and *c* are increased (postponed). Our goal is to coordinate a data channel that is

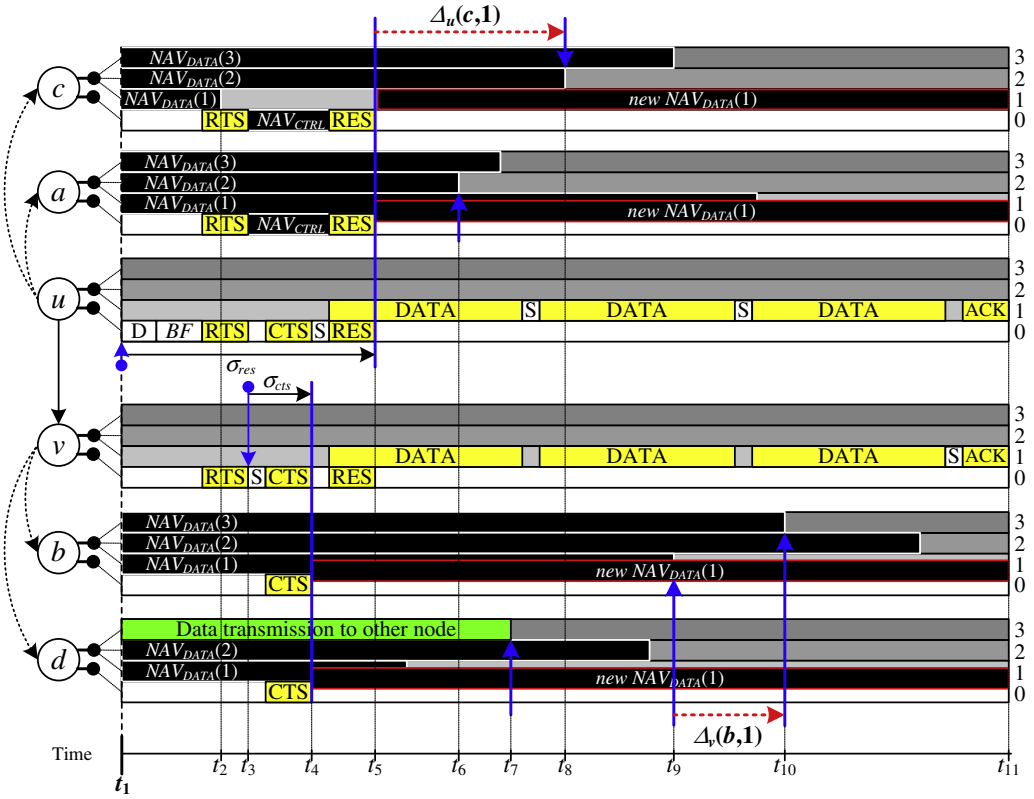


Fig. 9. Increments to starting times of possible transmissions at nearby nodes of *u* and *v* if selecting *ch*1.

not only free to the sender and receiver themselves, but also adds the least total delay to the starting times of possible transmissions at nearby nodes.

Now, we formally describe the *ECS* below. Consider a sender u and a receiver v. As the control process is initiated at *ctrl_ini_time*(u, v), node u firstly identifies a list of data channels that will be released to itself at the expected starting time of the transmission to v as follows.

$$FCL(u) = \{h \epsilon \mathscr{H} \mid ch\_rel\_time(u,h) \leqslant ctrl\_ini\_time(u,v) + pre\_ctrl + BF\}.$$

Note that by definition of the *CIP*, $FCL(u)$ must be non-empty at *ctrl_ini_time*(u, v). In addition, let N_u denote the set of nodes in u's interference range. For each $w \in N_u - \{v\}$, node u calculates the time that at least one data channel will be released to w as

$$CR_u(w) = \min\{CRT_u(w,h) |\ h \epsilon \mathscr{H}\}.$$

The $CR_u(w)$ is called the *critical channel release time* of w, and the first data channel released to w is called the *critical channel* of w. Combining with the interface release time of w in IRV_u, we define

$$NR_u(w) = \max\{CR_u(w),\ IRV_u(w)\}.$$

It is called the *node release time* of w. Clearly, node w cannot start any transmission before $NR_u(w)$.

Next, node u evaluates the cost for transmitting on each $h \in FCL(u)$. For each $w \in N_u - \{v\}$ and $h \in FCL(u)$, if nodes decided to transmit on channel h for a period of NAV_{DATA}, the $CR_u(w)$ could be enlarged (at least equally), and the new critical release time of w can be formulated as

$$CR_u^+(w|h) = \min\left\{ \begin{array}{l} \max\left\{ \begin{array}{l} CRT_u(w,h), \\ T_{curr} + (pre\_ctrl + BF) + NAV_{DATA} \end{array} \right\}, \\ \min\left\{CRT_u(w,h') | h' \epsilon \mathscr{H} - \{h\}\right\} \end{array} \right\}.$$

The above equation indicates that the original $CR_u(w)$ could be replaced by the release time of another channel $h' \in \mathscr{H} - \{h\}$, if the release time of the original critical channel of w is enlarged so that it is no longer critical to w. As shown in Fig. 9, at t_1, $CR_u(c) = t_2$, but $CR_u^+(c|1) = t_8$ The gap between t_2 and t_8 is due to the fact that the critical channel of c will be altered from ch_1 to ch_2 if u transmits on ch_1. Similarly, the corresponding node release time of w can be wrote as

$$NR_u^+(w|h) = \max\{CR_u^+(w|h), IRV_u(w)\}.$$

Using these terms, the increment to the node release time of w resulted from transmitting on channel h can be characterized as

$$\Delta_u(w,h) = \max\{NR_u^+(w|h), T_{curr} + \sigma_{res}\} - \max\{NR_u(w), T_{curr} + \sigma_{res}\},$$

where $\sigma_{res} = pre\_ctrl + BF + T_{RTS} + \tau$, denoting the duration before other nodes can receive the RES from u. Notice that the $\Delta_u(w,h)$ neglects the increment before *ctrl_ini_time*(u, v) + σ_{res}, since the transmission is not influential to w before the RES is received by w. As shown in Fig. 9, although $NR_u(c) = CR_u(c) = t_2$ and $NR_u^+(c|1) = CR_u^+(c|1) = t_8$, since the RES will arrive at t_5, the $\Delta_u(c,1)$ is $(t_8 - t_5)$ instead of $(t_8 - t_2)$. Accordingly, if node u transmits on channel h, the total increment to the nodes release time of all neighboring nodes can be defined by

$$\Delta_u(h) = \sum_{w \epsilon N_u - \{v\}} \Delta_u(w,h).$$

The $\Delta_u(h)$ will be sent to v along with a RTS frame for any $h \in FCL(u)$.

When node v received the RTS, it performs the same evaluation for each $h \in FCL(v) \cap FCL(u)$ and $w \in N_v - \{u\} - N_u$, where

$$FCL(v) = \{h \epsilon \mathscr{H} | ch\_rel\_time(v,h) \leqslant T_{curr} + 2SIFS + T_{CTS} + \tau\}.$$

That is, node v calculates

$$CR_v^+(w|h) = \min\left\{ \begin{array}{l} \max\left\{ \begin{array}{l} CRT_v(w,h), \\ T_{curr} + (\sigma_{cts} + SIFS) + (NAV_{DATA} - \tau) \end{array} \right\}, \\ \min\left\{CRT_v(w,h') | h' \epsilon H - \{h\}\right\} \end{array} \right\},$$

$$\Delta_v(w,h) = \max\{NR_v^+(w|h), T_{curr} + \sigma_{cts}\} - \max\{NR_v(w), T_{curr} + \sigma_{cts}\},$$

and

$$\Delta_v(h) = \sum_{w \epsilon N_v - \{u\} - N_u} \Delta_v(w,h),$$

where $\sigma_{cts} = SIFS + T_{CTS} + \tau$, denoting the duration before other nodes can receive the CTS from u.

Then, combining the transmission costs from the two sides, the total cost for communicating between u and v on a channel h can be defined as

$$\Delta_{u,v}(h) = \Delta_u(h) + \Delta_v(h).$$

The $\Delta_{u,v}(h)$ is the total increment to the node release time of all neighbors of u and/or v. In Fig. 9, $\Delta_{u,v}(1) = \Delta_u(1) + \Delta_v(1) = \Delta_u(a, 1) + \Delta_u(c, 1) + \Delta_v(b, 1) + \Delta_v(d, 1) = (t_5 - t_5) + (t_8 - t_5) + (t_8 - t_{10}) + (t_7 - t_7) = (t_8 - t_5) + (t_8 - t_{10})$. $= \Delta_u(a, 1) + \Delta_u(c, 1) + \Delta_v(b, 1) + \Delta_v(d, 1) = (t_5 - t_5) + (t_8 - t_5) + (t_8 - t_{10}) + (t_7 - t_7) = (t_8 - t_5) + (t_8 - t_{10})$. Similarly, we can find the $\Delta_{u,v}(2)$ and $\Delta_{u,v}(3)$. Finally, a channel $h \in FCL(v) \cap FCL(u)$ that has the least $\Delta_{u,v}(h)$ will be chosen as the communication channel, i.e. the h^*. The selected h^* will be sent back to u using the CTS. By definition, the h^* will be released to both u and v and has the least total increment to the node release times (starting times of possible transmissions) of potentially interfered nodes.

5. Simulations

In this section, we conduct simulations using the ns2 simulator. In order to evaluate how much performance gain each of the three components obtains, we have implemented the following versions for the *RTBM*:

- *RTBM* (*CIP*) consists of *CIP* only, has no flow control (i.e. $K_{u,v} = 1$), and follows the random channel selection strategy in [3].
- *RTBM* (*CIP* + *DDC*) is akin to the first version except that *DDC* is applied.
- *RTBM* (*CIP* + *DDC* + *ECS*) employs all designed components.

Note that *DDC* cannot be applied without *CIP*, since $K_{u,v}$ depends on the link release and control initiation times updated from *CIP*. Besides, if $K_{u,v}$ was always fixed on 1, the discrepancy between different data channels h's would be insignificant in terms of their communication costs $\Delta_{u,v}(h)$'s. Hence, we do not individually test *DDC* and/or *ECS* for the *RTBM*.

In addition, due to transmission failures or delays in updating control messages, the values stored in CRT_u and IRV_u could be stale or incomplete. The inaccuracy may lead to a series of invalid decisions in the *RTBM* (as shown in Fig. 6(c)). For this reason, we provide an ideal version, denoted as *RTBM**(*CIP* + *DDC* + *ECS*), where each node can directly access the information in its neighbors, to show how the performance degrades due to inaccurate information.

On the other hand, we compare the *RTBM* with the representative *DCC*-based protocol (*DCA*) in [3] and the parallel rendezvous protocol (McMAC) in [18]. The primary goal of McMAC is also to overcome the control bottleneck problem. The difference is that McMAC disperses control traffic across all available channels by allowing a sender switch to the hopping sequence of the intended receiver with a probability p. Here, we set p as the default setting in [18]. For all of our evaluations, the IEEE 802.11 DCF serves as the baseline protocol.

For each network under test, we generate 100 static nodes uniformly distributed on a 1500 m $\times$ 1500 m region. Each node is equipped with two interfaces, each with a default transmission range of 250 m. The interference range is assumed to be twice the transmission range. On the other hand, we provide 12 orthogonal channels. One of them is used for control purposes and the others are used for data transmission. The default channel bit rate is 11 Mbps.

On top of each generated network, we consider the throughput under two circumstances. In *single-hop communication*, we establish UDP flows with a variable bit rate over 200 random node pairs (i.e. each node communicates with four 1-hop nodes on average). Each communication pair is confined within the transmission range of each other. This case is intended to evaluate the link-layer performance. In *multi-hop communication*, we establish UDP flows with a constant bit rate over 20 randomly chosen node pairs. All sources and destinations are distinct nodes. Each packet is forwarded along with a shortest path in terms of the hop count. In both cases, the (average) packet arrival rate is 1 Mbps and the packet length is 1024 bytes. The throughput is defined as the total size of data packets successfully received by all destinations divided by the simulation time (100 s). Besides, each node u maintains a 50-packet queue for any 1-hop node v, i.e. $Q_v \leqslant 50$.

Table 4 lists the frame sizes of each compared protocol under $H = 11$. In addition to those in IEEE 802.11 DCF specifications (i.e. 20 bytes for RTS, 14 bytes for CTS, and 14 bytes for ACK), the DCA needs a 2-byte bitmap to carry the free channel list in RTS, and 2 bytes to indicate the selected channel and a parameter in CTS (see [3]). Thus, the frame sizes of RTS and CTS in DCA are 22 bytes (i.e. 22 = 20 + 2) and 16 bytes (i.e. 16 = 14 + 2), respectively. About the RTBM, its RTS additionally needs

Table 4
Frame sizes under $H = 11$ and system values of DSSS PHY.

	802.11 DCF/McMAC	*DCA*	*RTBM*
RTS length	20 bytes	22 bytes	44 bytes
CTS length	14 bytes	16 bytes	39 bytes
RES length	–	16 bytes	39 bytes
ACK length	14 bytes		
Propagation delay	5 μs		
Backoff slot time	20 μs		
SIFS	10 μs		
DIFS	50 μs		

2 bytes to encode the communication $\Delta_u(h)$ for each data channel h. Thus the size of RTS under $H = 11$ is 44 bytes (i.e. $44 = 20 + 2 + 2 \times 11$). Moreover, the CTS and RES needs 2 bytes for each duration of *if_duration*(v) and *ch_duration*(u, h), and 1 byte for the selected channel h^* (see Section 3.2). Thus, the sizes of the later two frames under $H = 11$ are 39 bytes (i.e. $39 = 14 + 2 + 2 \times 11 + 1$). Other system values specified for the DSSS PHY are also listed in Table 4. Although the *RTBM* needs larger frame sizes, as shown later, the benefit from our components can wholly compensate for such a drawback. The above contexts are the default settings. We also vary some parameters one at a time to assess the performance under different scenarios. Any result point of the following figures is averaged from 20 networks.

Fig. 10 reports the throughput versa the number of data channels (H). In comparison with the single-channel 802.11 DCF, the *DCA* can use multiple data channels to increase the throughput. When $H = 2$, it achieves a 28.98% gain in single-hop case and a 44.72% gain in multi-hop case. However, the improvement becomes very limited as more channels are provided. With 11 channels, the *DCA* obtains merely a 40.08% and 79.59% gain in the two cases, respectively. The results indicate that only a small portion of data channels were utilized due to the bottleneck in control channel.

The bottleneck problem can be mitigated by *CIP*. As shown in Fig. 10, the *RTBM*(*CIP*) has 4–42% throughput increments over the *DCA*. The reason is that *CIP* can prevent nodes from sending redundant RTSs and thus avoid unnecessary blocking from NAV_{CTRL}. Notice that *CIP* is particularly useful for small H, because a coordination process may fail when two nodes have no free data channels in common, and there are relatively fewer common channels when H is small. Nonetheless, *CIP* cannot entirely break through the limitation of the control channel, since a coordination process is still required before sending each data packet.

The *DDC* can substantially resolve this problem. As shown in Fig. 10, *RTBM*(*CIP+DDC*) has improved the baseline throughput by over 2.26 times in single-hop case and 3.49 times in multi-hop case, with just one additional data channel, i.e. $H = 2$. Moreover, for $H = 4$, it can further gain 2.86 times and 5.17 times performance in the two cases. However, since the traffic load is not quite heavy in default, the throughput reaches the limit when $H \geqslant 6$. As shown in Fig. 11, by pressuring nodes

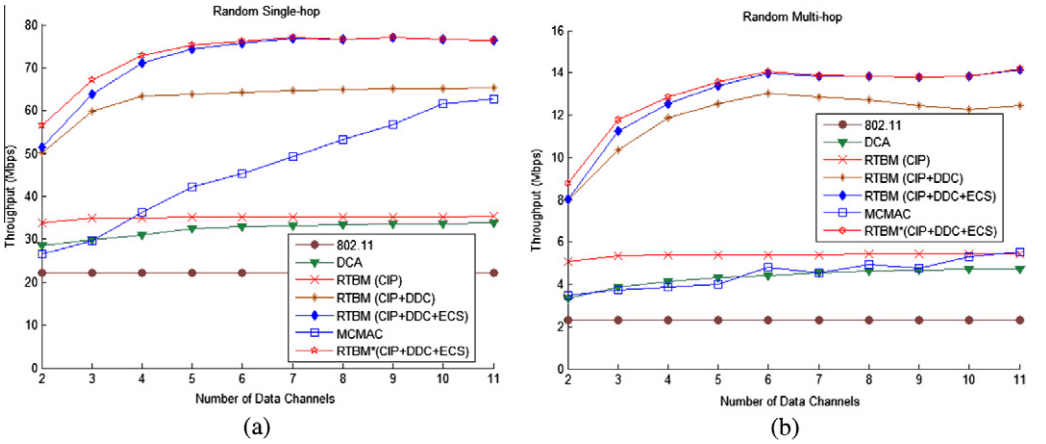


Fig. 10. Variation in the number of data channels.

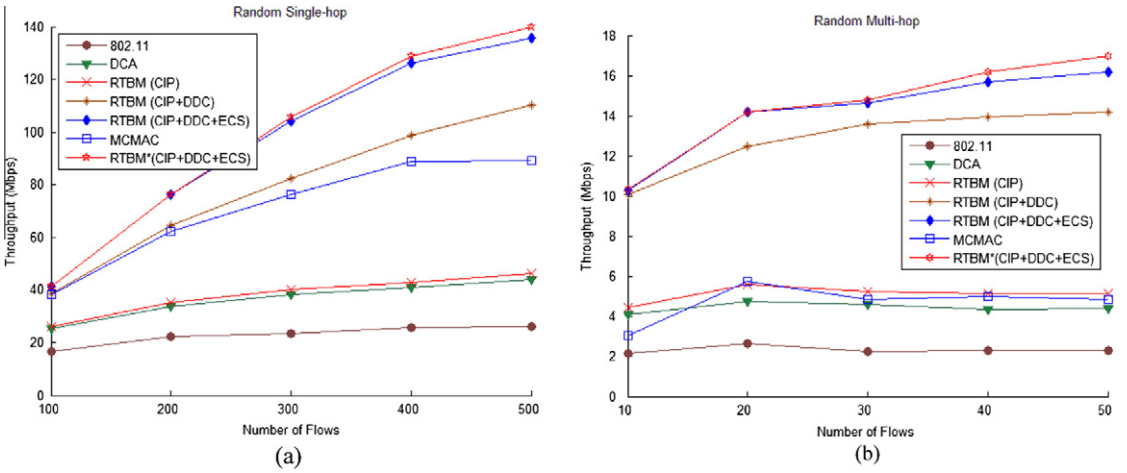


Fig. 11. Variation in the number of flows ($H = 11$).

with more data flows, the throughput is almost linear to the number of flows under $H = 11$. Contrarily, other protocols without *DDC* even degrade the throughput as the number of flows exceeds a certain limit.

The bottleneck problem is also resolved by McMAC. As shown in Fig. 10(a), the throughput of McMAC increases significantly when H raises. However, when H is small, McMAC does not perform well (even worse than the original *DCC* when $H = 2$). One reason is that the congestion in each channel is still heavy for a small H. In contrast, *DDC* can dynamically adjust the number of data packets being sent to balance the congestion in both control and data channels no matter how many channels are provided. Another reason is that the *DCC*-based approach can separate control and data traffic into different channels so that control and data signals are not interfering to each other. Note that although *RTBM* needs one additional channel for control purposes, we can see that the throughput of *RTBM*(*CIP* + *DDC*) with h channels is higher than that of McMAC with $h + 1$ channels in all cases. Moreover, McMAC has no significant improvement in a multi-hop case, because a node may frequently depart to different hopping sequences to forward messages, which however are always not visible to other nodes in the interference range, incurring a serious collision problem. In contrast, *RTBM* allows each node to acquire the status of other nodes and channels at any time.

The *ECS* can further enhance the throughput. As shown in Fig. 10, *RTBM*(*CIP* + *DDC* + *ECS*) shows clear gains over *RTBM*(*CIP* + *DDC*), especially when more data channels are provided. One observation explains this tendency: When H is large, each sender u and receiver v have more choices to find a data channel h which has a lower cost $\Delta_{u,v}(h)$. Particularly, under 11 data channels and 500 flows, *ECS* additionally contributes 26.37% of throughputs in single-hop case (15.41% in multi-hop case).

In Fig. 12, we vary the channel bit rate (R) to show how raw channel capacity relates to the throughput. As R is raised from 0.5 Mbps to 11 Mbps, we can see that 802.11 DCF gains approximately 9–10 times throughput, but there are relatively lower gains (no more than 4 times) for *DCA* and *RTMB*(*CIP*). The lower growth rates are caused by the following reason: with a higher channel bit rate, the data transmission is faster, which implies that a node may initiate more control processes within a shorter period of time. As a result, contention in the control channel becomes more intensive. *RTBM* can get rid of such limitations by using *DDC*. As shown in Fig. 12, *RTMB*(*CIP* + *DDC* + *ECS*) obtains 8.22 times throughput in the single-hop case and 9.39 times throughput in the multi-hop case as long as R is raised from 0.5 Mbps to 11 Mbps, which are very close to those achieved by 802.11 DCF, where no control bottleneck exists. The reason for this is that *DDC* can dynamically send multiple data frames to prolong the data transmission whenever the control channel is congested, i.e. data interface's idle time rises. Besides, *ECS* can provide more enhancements when R is large, because the greater the channel bit rate is, the more free data channels exist, providing more choices for finding the lower communication cost.

Fig. 13 demonstrates the impact of increasing transmission (interference) range We can see that the throughputs of 802.11 DCF, *DCA*, and *RTBM*(*CIP*) degrade drastically as the range increases. In contrast, *RTBM*(*CIP* + *DDC*) has only suffered a slight decrement in its throughput, since *DDC* can dynamically adapt to any increasing interference in the control channel. Moreover, *ECS* can help nodes to achieve better channel reusability. Hence, *RTBM*(*CIP* + *DDC* + *ECS*) is almost not affected by this factor. Surprisingly, the throughput can be even better when the range is more than 250 m. This phenomenon possibly stems from the fact that a larger transmission (interference) range can avoid the presence of hidden terminal nodes in the control channel.

Now let us compare with the ideal version. We can see that the throughput of *RTBM*(*CIP* + *DDC* + *ECS*) is very close to that of *RTBM**(*CIP* + *DDC* + *ECS*) especially when H is large (see Fig. 10), because the change of a channel release time can be less frequent if more other channels are offered. Besides, the throughput degradation is very small even under a heavy loading (see Fig. 11). Therefore, the inaccuracy of information has a little impact to our protocol.

Fig. 14 shows the normalized control overhead for each protocol (the total size of control messages being sent divided by that of the 802.11 DCF) over different number of flows. Consider the single-hop case: By comparing Fig. 11(a), we can see

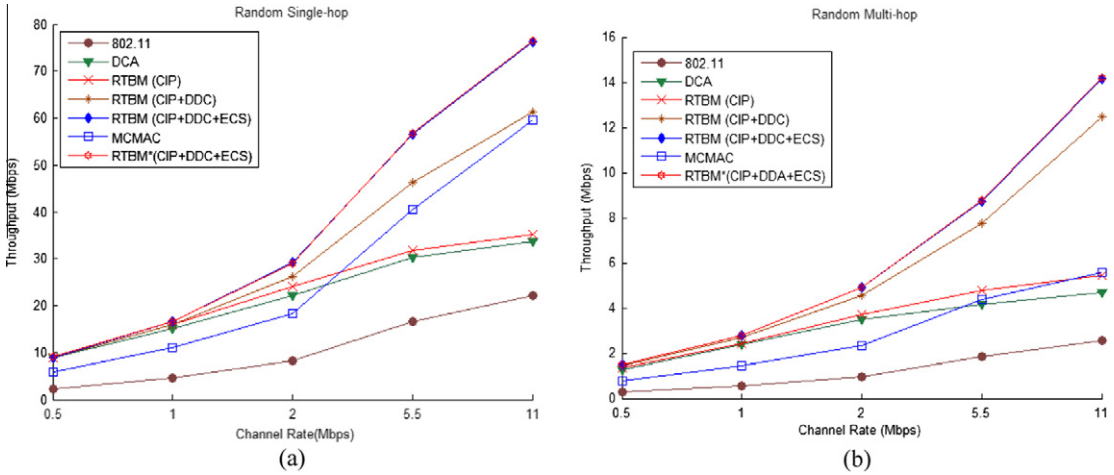


Fig. 12. Variation in channel bit rate ($H = 11$).

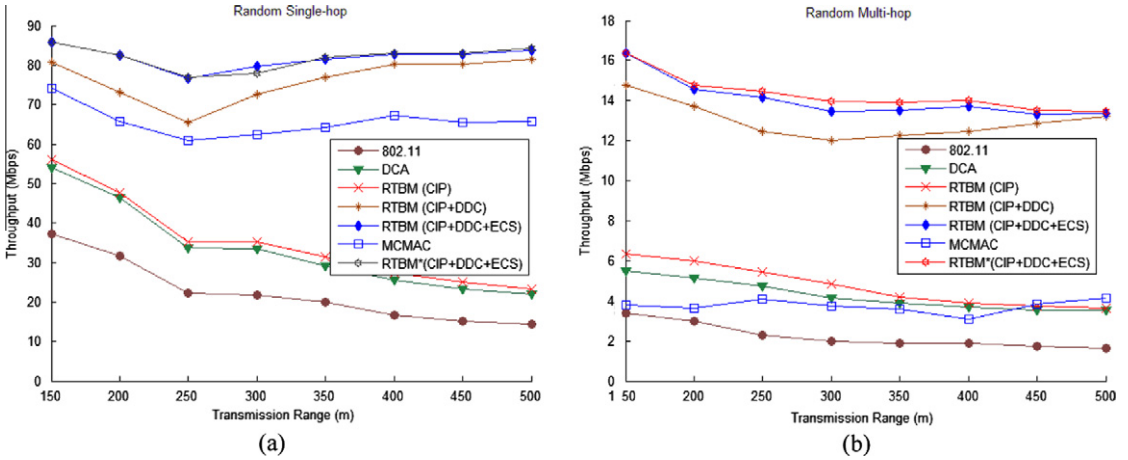


Fig. 13. Variation in transmission range ($H = 11$).

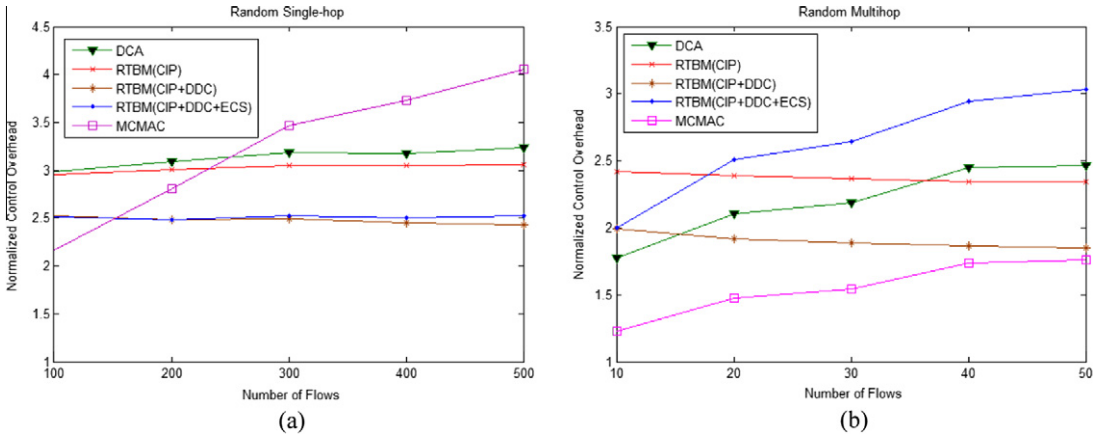


Fig. 14. Normalized control overhead ($H = 11$).

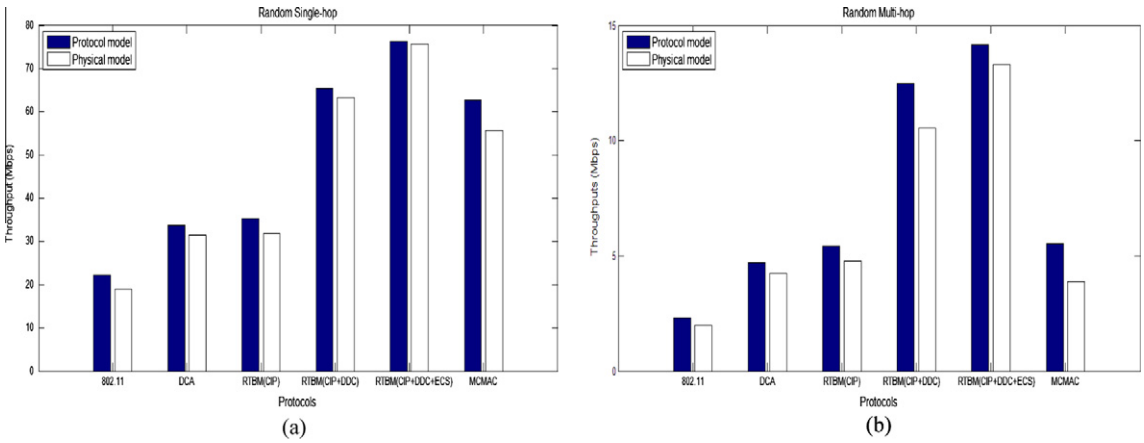


Fig. 15. Throughput under protocol vs. physical interference model ($H = 11$).

that the control overhead of McMAC goes up in accompany with more data throughput, because McMAC can spread control loading over different channels. By contrast, the throughput of *DCA* is saturated by a small number of flows, since the control overhead is congested in a single channel. Compared with *DCA*, although *RTBM* needs larger control frames sizes, it incurs

lower control overhead and achieves higher throughput in the three versions, because *CIP* can offer a better utilization of the control channel and *DDC* can increase the number of data packets being sent followed by each control process. Consider the multi-hop case: McMAC has the lowest control overhead and lower throughput, while *RTBM*(*CIP + DDC + ECS*) has the highest control overhead and throughput. The results indicate that the impact from data channels could be more serious in a multi-hop case.

Fig. 15 shows the throughout under protocol interference model versa physical interference model based on SINR. We can see that our protocol has only a slight degradation under the physical model especially when *ECS* is applied, while McMAC has a larger degradation in throughput. The reason is that *ECS* is designed to increase the reusability of data channels so that the interference problem is less serious under both of the two models. Note that the throughput degradation becomes more significant in multi-hop case for each protocol because in multi-hop case the physical model not only incurs more serious interference among different communication pairs (i.e. the inter-flow interference) but also worsens the interference between hops of the same flow (i.e. the inter-hop interference). Besides, for a transmission, there are more hidden terminal nodes in both control and data channels when multi-hop flows exist.

6. Conclusion

In this paper, we have proposed a release-time-based MMAC protocol to overcome the control channel bottleneck problem and the dynamic channel selection problem for the *DCC* approach. Three components have been investigated and incorporated into this protocol. The control initiation-time predation can reduce the redundant control overhead by properly predicting the control point to increase the chance for a successful coordination. The dynamic data-flow control can dynamically adjust the number of data packets to be sent according to the real-time condition of both the control and data interfaces. The enhanced channel selection can gain better channel reusability by selecting the channel that has the least influence to the transmission starting times of nearby nodes. Simulation results have shown that our protocol with these components achieves significant improvement in comparison with previous works, especially when the number of channels, data frame size and data rates of data channels are large. Both McMAC and *RTBM* with *DCC* can overcome the control channel problem. However, *RTBM* can provide higher throughput due to channel selection strategy. Besides, *RTBM* is a more suitable to a complex environment where multi-hop transmission and frequent link-layer broadcasting exist.

For future research, it is worthy to evaluate the performance of mobile nodes. In this circumstance, the quality of a channel selection strategy could be more influential to the final results. In addition, the bottleneck problem can be further alleviated if a candidate control channel is used, but the design is expected to be more complicated.

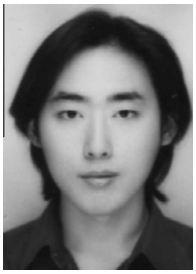
Acknowledgements

The authors would like to thank Yu-Hsiang Cheng for this helpful assistance in conducting the experimental results.

Reference

[1] I.-F. Akyildiz, X. Wang, Welin Wang, Wireless mesh networks: a survey, Computer Networks 47 (2005) 445–487.
[2] J. Mo, H.-S. So, J. Walrand, Comparison of multichannel MAC protocols, IEEE Transactions on Mobile Computing 7 (1) (2008) 50–65.
[3] S.-L. Wu, Y. Lin, Y.-C. Tseng, J.-P. Sheu, A new multi-channel MAC protocol with on-demand channel assignment for multi-hop mobile ad hoc networks, in: Proc. of Int'l Symp. on Parallel Architectures, Algorithms and Networks, 2000, pp. 232–237.
[4] S.-L. Wu, Y. Lin, Y.-C. Tseng, J.-P. Sheu, A multi-channel MAC protocol with power control for multi-hop mobile ad hoc networks, Computer Journal 45 (1) (2002) 101–110.
[5] P.-J. Wu, C.-N. Lee, On-demand connection-oriented multi-channel MAC protocol for ad-hoc network, in: Proc. of 3rd IEEE Conf. Society Conf. on Sensor and Ad Hoc Networks, 2006, pp. 621–625.
[6] W.-C. Hung, K.-L.-E. Law, A. Leon-Garcia, A dynamic multi-channel mac for ad-hoc LAN, in: Proc. of 21th Biennial Symp. On Communications, 2002, pp. 31–35.
[7] H. Koubaa, Fairness-enhance multiple control channels MAC for ad hoc networks, in: Proc. of 62th IEEE Vehicular Technology Conference, 2005, pp. 1504–1508.
[8] M. Benveniste, Z. Tao, Performance evaluation of a medium access control protocol for IEEE 802.11s mesh networks, in: Proc. of IEEE Sarnoff Symp., 2006, pp. 1–5.
[9] J. Zhang. Y. Wang, J. Wang, DCC-MAC: A new MAC protocol for ad-hoc networks based on dual control channel, in: Proc. of 25th IEEE Int'l Symp. on Personal Indoor and Mobile Radio Communication, 2003, pp. 1341–1345.
[10] N. Jain, S.-R. Das, A. Nasipuri, A multichannel CSMA MAC protocol with receiver-based channel selection for multihop wireless networks, in: Proc. of 10th Int'l Conf. on Computer Communications and Networks, 2001, pp. 432–439.
[11] A. Raniwala, K. Gopalan, T.C. Chiueh, Centralized channel assignment and routing algorithm for multi-channel wireless mesh networks, ACM Mobile Computing and Communications Review 8 (2) (2004) 50–55.
[12] A.-K. Das, H.-M.-K Alazemi, R. Vijayakumar, S. Roy, Optimization models for fixed channel assignment in wireless mesh networks with multiple radios, in: Proc. of 2nd Annual IEEE Communications Society Conf. on Sensor and Ad Hoc Networks, 2005, pp. 463–474.
[13] J. Crichigno, M.-Y. Wu, W. Shu, Protocols and architectures for channel assignment in wireless mesh networks, Ad Hoc Networks 6 (7) (2007) 1051–1077.
[14] J. So, N.-H. Vaidya, Multi-channel MAC for ad hoc networks: handling multi-channel hidden terminals using a single transceiver, in: Prof. of 5th ACM Int'l Symp. on Mobile Ad Hoc Networking and Computing, 2004, pp. 222–233.
[15] J. Chen, S.-T. Sheu, C.-A. Yang, A new multichannel access protocol for IEEE 802.11 ad hoc wireless LANs, in: Proc. of 14th IEEE Int'l Symp. on Personal, Indoor and Mobile Radio Communications, 2003, pp. 2291–2296.
[16] Z.-J. Haas, J. Deng, Dual busy tone multiple access (DBTMA)-a multiple access control scheme for ad hoc networks, IEEE Transactions on Communications 5 (6) (2000) 975.
[17] S.-L. Wu, Y.-C. Tseng, J.-P. Sheu, Intelligent medium access control for mobile ad hoc networks with busy tones and power control, IEEE Journal on Selected Areas in Communications 18 (9) (2000) 1647–1657.
[18] A. Tzamaloukas, J.J. Carcia-Luna-Aceves, Channel-hopping multiple access, in: Proc. of IEEE Int'l Conf. Comm. (ICC'00), June, 2000.

[19] Y. Li, H. Wu, N.-F. Tzeng, D. Perkins, M. Bayoumi, MAC-SCC: a medium access control with separate control for reconfigurable multi-hop wireless networks, IEEE Transactions on Wireless Communications 5 (7) (2006) 1805–1817.
[20] P. Kyasanur, J. Padhye, P. Bahl, On the efficacy of separating control and data into different frequency bands, in: Proc. of Broadband Networks, 2005, pp. 646–655.
[21] P. Bahl, R. Chandra, J. Dunagan, SSCH: slotted seeded channel hopping for capacity improvement in IEEE 802.11 ad hoc wireless networks, in: Proc. of ACM MobiCom, September 2004.
[22] H.-S. So, J. Walrand, J. Mo, McMAC: a parallel rendezvous multi-channel mac protocol, in: Proc. of IEEE Wireless Communications and Networking Conference, 2007, pp. 334–339.
[23] P. Kyasanur, J. So, C. Chereddi, N.H. Vaidya, Multichannel mesh network: challenges and protocols, IEEE Wireless Communications 10 (2) (2006) 30–36.
[24] A.A.K. Jeng, R.H. Jan, Role and channel assignment for wireless mesh networks using hybrid approach, Computer Networks 53 (12) (2009) 2225–2240.



Andy An-Kai Jeng received the B.S. degree in statistics from Tamkang University, Taiwan, in 2001, the M.S. degree in management information systems from National Chi Nan University, Taiwan, in 2003, and Ph.D degree in the computer science from National Chiao Tung University, Taiwan, in 2007, where is currently a post-doctoral researcher. His research interests include wireless networks, distributed algorithm design and analysis, scheduling theory, and operations research.

Rong-Hong Jan received the B.S. and M.S. degrees in Industrial Engineering, and the Ph.D. degree in Computer Science from National Tsing Hua University, Taiwan, in 1979, 1983, and 1987, respectively. He joined the Department of Computer and Information Science, National Chiao Tung University, in 1987, where he is currently a Professor. During 1991–1992, he was a Visiting Associate Professor in the Department of Computer Science, University of Maryland, College Park, MD. His research interests include wireless networks, mobile computing, distributed systems, network reliability, and operations research.

Chi-Yu Li received the B.S. degree in Computer and Information Science from National Chiao Tung University, Taiwan, in 2004 and M.S. degree in Computer and Information Science from National Chiao Tung University, Taiwan, in 2006. His research interests include WMN (Wireless Mesh Network) and wireless MAC protocols.

Chien Chen received his B.S degree in Computer Engineering from National Chiao Tung University in 1982 and the M.S. and Ph.D. degrees in Computer Engineering from University of Southern California and Stevens Institute of Technologies in 1990 and 1996. Dr Chen hold a Chief Architect and Director of Switch Architecture position in Terapower Inc., which is a terabit switching fabric SoC startup in San Jose, before joining National Chiao Tung University as an Assistant Professor in August 2002. Prior to joining Terapower Inc., he is a key member in Coree Network, responsible for a next-generation IP/MPLS switch architecture design. He joined Lucent Technologies, Bell Labs, NJ, in 1996 as a Member of Technical Staff, where he led the research in the area of ATM/IP switch fabric design, traffic management, and traffic engineering requirements. His current research interests include wireless ad-hoc and sensor networks, switch performance modeling, and DWDM optical networks.

Optimal Modulation and Coding Scheme Allocation of Scalable Video Multicast over IEEE 802.16e Networks[1]

Chia-Tai Tsai, Rong-Hong Jan[2], Chien Chen
Department of Computer Science
National Chiao Tung University
1001 University Road, Hsinchu, 300, Taiwan, R.O.C.
{tai, rhjan, chienchen}@cs.nctu.edu.tw

Abstract: With the rapid development of wireless communication technology and the rapid increase in demand for network bandwidth, IEEE 802.16e is an emerging network technique that has been deployed in many metropolises. In addition to the features of high data rate and large coverage, it also enables scalable video multicasting, which is a potentially promising application, over an IEEE 802.16e network. How to optimally assign the modulation and coding scheme (MCS) of the scalable video stream for the mobile subscriber stations to improve spectral efficiency and maximize utility is a crucial task. We formulate this MCS assignment problem as an optimization problem, called the total utility maximization problem (TUMP). This paper transforms the TUMP into a precedence constraint knapsack problem, which is an NP-complete problem. Then, a branch and bound method, which is based on two dominance rules and a lower bound, is presented to solve the TUMP. The simulation results show that the proposed branch and bound method can find the optimal solution efficiently.

Keywords: Adaptive modulation and coding, Branch and bound algorithm, IEEE 802.16e, Resource allocation, Scalable video coding.

[1] This paper was supported in part by the National Science Council of the ROC, under Grants NSC -97-2221-E-009-049-MY3.
[2] Corresponding Author; Fax: 886-3-5721490.

1. Introduction

With the popularity of wireless networks, the need for network bandwidth is growing rapidly. In order to provide high quality service, various categories of broadband wireless network techniques, e.g., IEEE 802.16e (or WiMAX, Worldwide Interoperability for Microwave Access) and 3GPP LTE, have been proposed. Among these techniques, IEEE 802.16e is an emerging network technique and has been deployed in many metropolises (e.g., Chicago, Las Vegas, Seattle, Taipei and so forth [1][2]). It provides mobile users with a high data rate (up to 75 Mbps) and a large coverage range (up to a radius of 10 miles) [3][4][5]. In addition, it also enables new classes of real-time video services, such as IPTV services, video streaming services, and live TV telecasts, which require a large transmission bandwidth, and need identical content to be delivered to several mobile stations. The most efficient way to provide such services is to use wireless multicasting, sending one copy of the video stream to multiple subscriber stations via a shared multicast channel, instead of sending multiple copies via several dedicated channels [6]. In this way, wireless multicasting can reduce bandwidth consumption significantly.

IEEE 802.16e supports a variety of modulation and coding schemes (MCSs), such as QPSK, 16QAM, and 64QAM, and allows these schemes to change on a burst-by-burst basis per link, depending on channel conditions [3][4][5]. Adaptive modulation and coding (AMC) is a term used in wireless communications to denote the matching of the modulation and coding to the channel condition for each subscriber station. It is widely applied to wireless networks. For example, the IEEE 802.16e base station (BS) can assign an appropriate MCS to each mobile subscriber station (MSS) based on its channel quality. This can be done by having the MSS advises its downlink channel quality indicator to the BS. The BS scheduler can take into account the channel quality of the MSSs and assign an appropriate MCS for each of them so that the throughput is maximized.

Due to the mobility (i.e. the ability to move within the coverage area) of an MSS, the signal-to-noise ratio (SNR) from the BS may become degraded (i.e. the MSS could be in poor channel condition at some time). The adaptation strategy for the MSS with the worst channel condition will cause the data rate to be low, especially when the multicast group size is large [7]. For example, as shown in Figure 1, the BS chooses QPSK, the most conservative and robust MCS, to

accommodate all MSSs in the multicast group, even if there are some MSSs (e.g. MSS1, MSS2, and MSS3) that can be accommodated with a higher data rate MCS (e.g. 16QAM). That is, the multicast data rate is determined by the MSS which has the worst channel condition (e.g. MSS 4). As a result, the spectral efficiency tends to be poor.

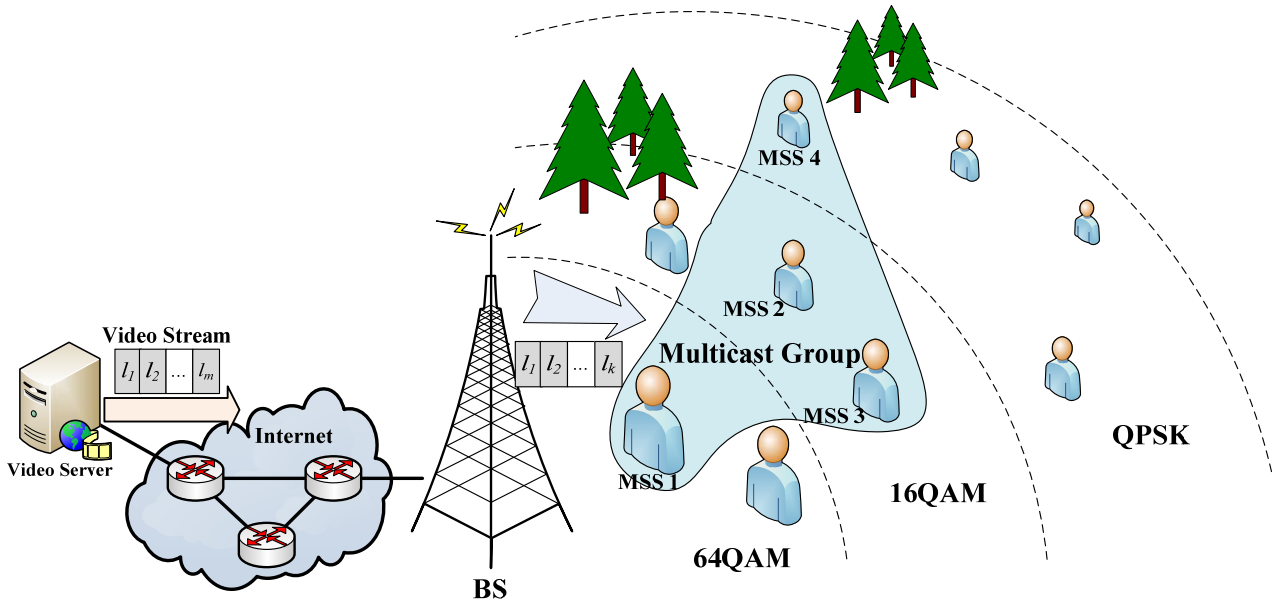


Figure 1: The video multicast network environment over IEEE 802.16e networks.

The Scalable Video Coding (SVC) scheme [8] allows for the delivery of a decodable and presentable quality of the video depending on the MSS's channel quality. The SVC scheme divides a video stream into one base layer and several enhancement layers [8]. The base layer provides a basic video quality, frame rate, and resolution of the video, and the enhancement layers can refine the video quality, frame rate, and resolution. Figure 2 shows the video quality under various combinations of video layers. The more video layers an MSS receives, the better video quality it can get. In this paper, we apply the utility [9][10] to measure the satisfaction degree of the video quality that the MSS received.

(a) Only one base layer

(b) One base layer and one enhancement layer

(c) One base layer and two enhancement layers

Figure 2: The video quality for the MSS under various numbers of video layers (The video, foreman, is downloaded from the video trace library [11]).

In wireless networks, because the air resources are limited and shared by all receivers, organizing the layering structure of a video stream and assigning the appropriate MCS for each video layer to maximize the total utility is a crucial task[12]-[22]. Formally, the problem can be stated as follows: consider a video multicasting network having a scalable video stream V consisting of m video layers $L = \{l_1, l_2, \ldots, l_m\}$ and adaptive modulation and coding scheme consisting of n MCSs $\{M_1, M_2, \ldots, M_n\}$. The BS chooses a layering structure (i.e. selecting a set of video layers L' from L), which will multicast to the MSSs, and determines an appropriate MCS for each video layer in L' such that the total utility is maximized subject to a bandwidth constraint.

In this paper, we formulate the MCS assignment of the layering structure as a total utility maximization problem (TUMP). This paper transforms the TUMP into a precedence constraint knapsack problem, which is an NP-complete problem [23]. The precedence-constraint knapsack problem is a generalization of the knapsack problem, which includes the constraint on the packed order of the items. For example, if item i precedes item j, then item j can only be packed into the knapsack if item i is already packed into the knapsack. Because the solution space of the problem TUMP consists of a large number of fruitless candidates, a branch and bound method which is based on two dominance rules and a lower bound is presented to solve the TUMP. The simulation results show that the proposed branch and bound method can find the optimal solution efficiently. Because the optimal solution can be found with just a little computation time, the proposed method is suitable

for MCS assignment in a scalable video multicast over IEEE 802.16e networks.

This paper is organized as follows. In section 2, we describe and formulate the TUMP problem. We transform the TUMP into a precedence constraint knapsack problem and propose a branch and bound method to solve the TUMP in section 3. The experimental results are given in section 4. Finally, we conclude this paper in section 5.

2. Problem Description

2.1. Statement of the problem

In this paper, we consider a video multicast network environment over an IEEE 802.16e network as shown in Figure 1. The MSSs can access the Internet through the BS. The ranging process occurs when an MSS joins the network and updates periodically; hence the BS can obtain the link quality of each MSS [3][4][5]. Suppose that there is a set of MSSs joined to a multicast group and subscribing to a scalable video stream V consisting of m video layers $L = \{l_1, l_2, \ldots, l_m\}$. The video server delivers V to the BS through the Internet. The BS has n MCSs $\{M_1, M_2, \ldots, M_n\}$. It takes each MSS's channel quality and the number of available time slots into account before organizing the layering structure. If the number of available time slots is not large enough, then the BS has to choose a set of feasible video layers L' from L and determine an appropriate MCS for each video layer in L'. Our goal is to maximize the total utility under a bandwidth constraint.

2.2. Model and Notations

Based on the specification of IEEE 802.16e [3][4][5], each frame consists of subchannels and OFDMA symbols. For the down link frame, a time slot, the minimum allocable resource unit, includes two consecutive OFDMA symbols in a subchannel [3][4][5]. Let S be the number of the available time slots allocated to the video stream. The MCSs, $M_1, M_2, \ldots M_n$, are sorted in ascending order from the lowest data rate (i.e., the most robust) MCS to the highest data rate MCS. Let r_j be the data rate (bytes per time slot) of M_j, $j = 1, 2, \ldots, n$, and $r_1 \le r_2 \le \ldots \le r_n$. For example, as shown in Figure 1, the BS supports three MCSs QPSK, 16QAM, and 64QAM, i.e. M_1 = QPSK, M_2 = 16QAM, and M_3 = 64QAM.

Suppose that an MSS receives a set of video layers $L' = \{l_1, l_2, \ldots, l_k, l_x, l_y, \ldots, l_z\}$ from a BS where $k+1 < x < y < z$. Note that an enhancement layer, say layer l_k, can be used to refine the video quality only when the MSS has received all the lower layers, i.e., $l_1, l_2, \ldots, l_{k-1}$[14]. Therefore, in this example, the maximum number of consecutive video layers of L' is k. Then, we say that the received enhancement layers $l_2, l_3, \ldots, l_k$ are the *valid* video layers for refining the video quality. The invalid video layer (e.g. l_x, l_y, or l_z) will be discarded by the MSS.

In order to determine the satisfaction degree of the video quality for an MSS, a relative measure of satisfaction, called *utility*, is used in [12]- [22]. Figure 3 is an example of the utility function for an MSS under various numbers of video layers [10]. When an additional video layer is received, the utility is increased and the MSS can experience the additional satisfaction. Because the attenuation is caused by shadowing or slow fading in the wireless communication, the utility function is often assumed to be log-normally distributed [24]. Let $Util(i)$ be the utility of an MSS when it has received i valid video layers. Let δ_i be the additional utility when the MSS received the i^{th} video layer, $i = 1, 2, \ldots, m$. Then, δ_i can be calculated as follows:

$$\delta_i = Util(i) - Util(i-1). \quad (1)$$

Note that $Util(0) = 0$. Thus, the additional utility of the base layer, δ_1, equals $Util(1)$. Table 1 lists the utility and additional utility of an MSS under various numbers of video layers (e.g., $m = 5$).

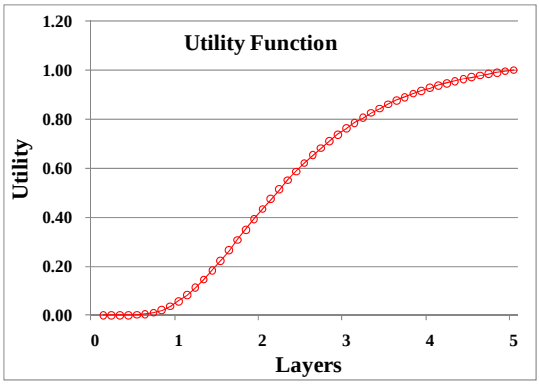


Figure 3: Utility function under various numbers of video layers.

Table 1: Utility and additional utility of an MSS under various numbers of video layers.

	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$
$Util(i)$	0.06	0.43	0.76	0.93	1
δ_i	0.06	0.37	0.33	0.17	0.07

Let u_j be the number of MSSs which can receive the video stream encoded by M_j. The number of MSSs at lower MCSs (e.g., QPSK) is greater than that at higher MCSs (e.g. 64QAM), i.e. $u_1 \geq u_2 \geq \cdots \geq u_j$. For example, Table 2 lists the set of MCSs which can be accepted by the MSSs in the multicast group as shown in Figure 1. From Table 2, we can find $u_1 = 4$, $u_2 = 3$, and $u_3 = 1$.

Table 2: The set of MCSs which can be accepted by the MSSs in the multicast group.

	The set of MCSs that can be received by the MSS
MSS1	$\{M_1, M_2, M_3\}$
MSS2	$\{M_1, M_2\}$
MSS3	$\{M_1, M_2\}$
MSS4	$\{M_1\}$

Let w_{ij} be the amount of utility when the video layer l_i is encoded by M_j. We can compute w_{ij} by $w_{ij} = \delta_i u_j$, $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$. Note that $w_{i1} \geq w_{i2} \geq \cdots \geq w_{ij}$, $i = 1, 2, \ldots, m$ because $u_1 \geq u_2 \geq \cdots \geq u_j$. In addition, suppose that the video layer l_i contains λ_i bytes, $i = 1, 2, \ldots, m$. The number of time slots t_{ij} required to transmit the layer l_i using MCS M_j can be computed by

$$t_{ij} = \left\lceil \frac{\lambda_i}{r_j} \right\rceil, \text{where } i = 1, 2, ..., m \text{ and } j = 1, 2, ..., n. \quad (2)$$

2.3. Problem Formulation

The optimal MCS assignment for scalable video multicast can be mathematically stated as follows.

Problem TUMP:

Maximize $$z = \sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij}x_{ij} \quad (3)$$

Subject to $$\sum_{i=1}^{m}\sum_{j=1}^{n} t_{ij}x_{ij} \le S \quad (4)$$

$$\sum_{j=1}^{n} x_{ij} \le 1,\ i = 1, 2, ..., m \quad (5)$$

$$\sum_{j=1}^{n} x_{i-1j} - \sum_{j=1}^{n} x_{ij} \ge 0\ ,\ i = 2, 3, \ldots, m \quad (6)$$

$$x_{ij} = 0 \text{ or } 1,\ i = 1, 2, \cdots m \text{ and } j = 1, 2, \cdots n\ . \quad (7)$$

This is a 0-1 integer programming problem. x_{ij} is the decision variable where x_{ij} = 1 indicates that video layer l_i is encoded by M_j; otherwise, x_{ij} = 0. Constraint (4) ensures that the sum of the required time slots cannot exceed S. Constraint (5) limits a video layer to being encoded by only one MCS at the same time. In order to avoid sending the invalid video layer, constraint (6) ensures that the video layer l_i can only be encoded if the video layer l_{i-1} has been encoded.

3. The Solution Method

In this section, we first transform the TUMP into a precedence constraint knapsack problem, which is a well-known NP-complete problem [23]. Then, we propose a branch and bound algorithm for solving the TUMP problem.

3.1. Problem Hardness

We convert the inequality constraint of the TUMP problem (equation (5)) to the equality constraint by introducing a set of slack variables X, where $X = \{x_{1n+1}, x_{2n+1}, \ldots, x_{mn+1}\}$. For all i, $x_{i\,n+1}$ is defined as

$$x_{in+1} = \begin{cases} 0, & \text{if the video } l_i \text{ is encoded by } M_j \\ 1, & \text{otherwise} \end{cases} \quad (8)$$

That is, $x_{in+1} = 1 - \sum_{j=1}^{n} x_{ij}$, where i = 1, 2, …, m. For all i, let $w_{i\,n+1}$ = 0 and $t_{i\,n+1}$ = 0. We can rewrite

equations (3), (4), and (5) as follows.

$$z = \sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij}x_{ij} = \sum_{i=1}^{m}\left(w_{i1}x_{i1} + w_{i2}x_{i2} \cdots + w_{in+1}x_{in+1}\right) = \sum_{i=1}^{m}\sum_{j=1}^{n+1} w_{ij}x_{ij} \tag{9}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} t_{ij}x_{ij} = \sum_{i=1}^{m}\left(t_{i1}x_{i1} + t_{i2}x_{i2} \cdots + t_{in+1}x_{in+1}\right) = \sum_{i=1}^{m}\sum_{j=1}^{n+1} t_{ij}x_{ij} \le S \tag{10}$$

$$\sum_{j=1}^{n} x_{ij} + x_{in+1} = \sum_{j=1}^{n+1} x_{ij} = 1,\ i = 1, 2, \ldots, m. \tag{11}$$

From equation (6), we know that $\sum_{j=1}^{n} x_{i-1j} \ge \sum_{j=1}^{n} x_{ij}$. Note that $\sum_{j=1}^{n} x_{i-1j} + x_{i-1n+1} = \sum_{j=1}^{n} x_{ij} + x_{in+1} = 1$. Thus, equation (6) can be transformed as follows.

$$x_{i-1n+1} \le x_{in+1},\ i = 2, 3, \ldots, m. \tag{12}$$

Therefore, the TUMP problem can be transformed as follows:

Problem TUMP1:

Maximize $$z = \sum_{i=1}^{m}\sum_{j=1}^{n+1} w_{ij}x_{ij} \tag{13}$$

Subject to $$\sum_{i=1}^{m}\sum_{j=1}^{n+1} t_{ij}x_{ij} \le S \tag{14}$$

$$\sum_{j=1}^{n+1} x_{ij} = 1, \text{ where } i = 1, 2, \ldots, m \tag{15}$$

$$x_{1n+1} \le x_{2n+1} \le \cdots \le x_{mn+1} \tag{16}$$

$$x_{ij} = 0 \text{ or } 1,\ i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, n+1 \tag{17}$$

Note that the above problem TUMP1 is equivalent to the precedence constraint knapsack problem [23], which is an NP-complete problem.

3.2. Branch and Bound Algorithm

In this section, we propose a branch and bound algorithm, which is commonly used to solve integer programming problems [25], for solving the TUMP problem. Obviously, the solution space of

TUMP may consist of all 2^{mn} combinations of the mn binary variables. However, we can apply the multiple choice constraints (5) and the precedence constraints (6) to reduce the solution space to $\binom{m+n}{n}$ combinations. Figure 4 shows a possible tree organization for the case m = 4 and n = 3. We call such a tree a combinatorial tree. The links are labeled by possible choices of M_j for l_i (i.e., x_{ij} = 1). For example, links from the root (level-0) node to level-1 nodes specify that each of x_{1j}, j = 1, 2, …, n, is selected and set to 1. The links from the level-i node, pointed to by the link with label x_{ij}=1, to level-(i+1) nodes are labeled by x_{i+1j} = 1, x_{i+1j+1} = 1, …, or x_{i+1n} = 1 due to the precedence constraints. For example, there are only two links from node 13 at level-2, pointed to by the link with label x_{22} = 1, to the level-3 nodes 14 and 17. They are labeled x_{32} = 1 and x_{33} = 1, respectively. Thus, the solution space is defined by all paths from the root node to any node in the tree. The possible paths are () (this corresponds to the empty path from the root to itself); (x_{11} = 1); (x_{11} = 1, x_{21} = 1); (x_{11} = 1, x_{21} = 1, x_{31} = 1); (x_{11} = 1, x_{21} = 1, x_{31} = 1, x_{41} = 1); (x_{11} = 1, x_{21} = 1, x_{31} = 1, x_{42} = 1); (x_{11} = 1, x_{21} = 1, x_{31} = 1, x_{43} = 1); (x_{11} = 1, x_{21} = 1, x_{32} = 1); (x_{11} = 1, x_{21} = 1, x_{32} = 1, x_{42} = 1); etc. The path $(x_{1y_1} = 1, x_{2y_2} = 1, \ldots, x_{iy_i} = 1)$ defines a possible solution that $x_{1y_1} = 1, x_{2y_2} = 1, \ldots, x_{iy_i} = 1$ and the others x_{ij} equals zero. There are $\binom{m+n}{n} = \binom{3+4}{4}$ = 35 nodes in Figure 4. That is, there are 35 possible combinations for selecting M_j, j = 1, 2, 3 for l_i, i = 1, 2, 3, 4.

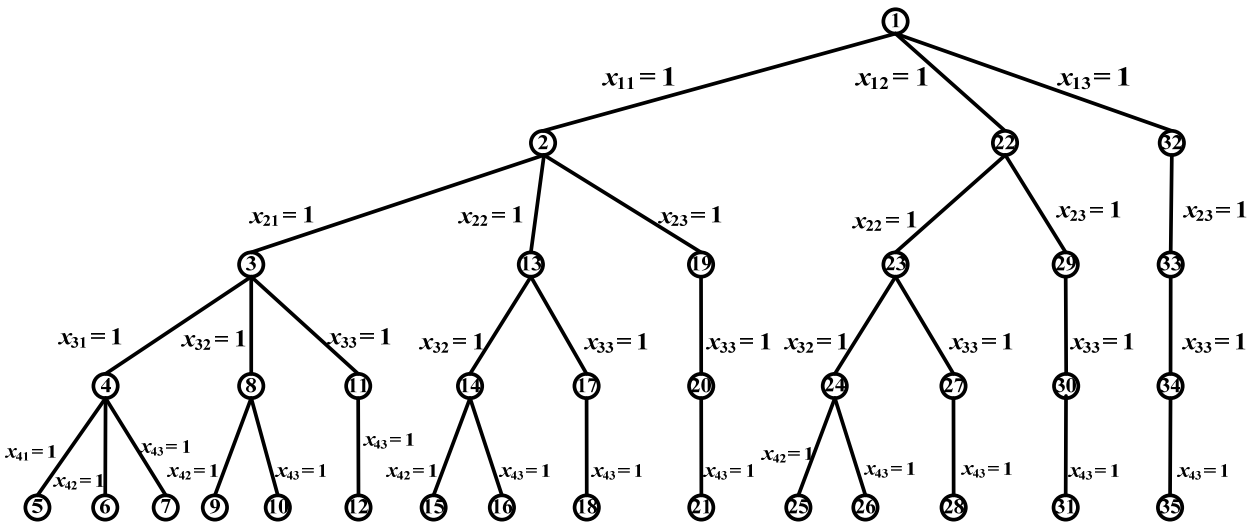


Figure 4: The combinatorial tree where m =4 and n = 3.

To find an optimal solution, we do not consider all combinations, since it is time-consuming.

We apply the greatest utility branch and bound algorithm to find the optimal solution by traversing only a small portion of the combinatorial tree. The branch and bound method has three decision rules that provide the method for:

1. Estimation of the upper bound of the objective function (i.e., total utility) at every node of the combinatorial tree.
2. Feasibility test at each node.
3. Selecting the next live node for branching, and terminating the algorithm.

3.2.1. Estimation of the upper bound of the objective function at each node

Let p be the current node in the combinatorial tree and ($x_{1y_1}=1, x_{2y_2}=1, ..., x_{iy_i}=1$) be the path from the root to the node p. Let $f(p)$ be the total utility received at node p (i.e., $f(p) = w_{1y_1}+w_{2y_2}+...+w_{iy_i}$). Let $g(p)$ be the maximum total utility that appears in the solutions generated from node p.

$$g(p) = f(p) + \sum_{k=i+1}^{m} w_{ky_i} . \tag{18}$$

Equation (18) results from $w_{ky_i} \geq w_{ky_{i+1}} \geq \cdots \geq w_{ky_m}$, $k = i+1, i+2, \ldots, m$.

3.2.2. Feasibility test at each node

Whenever a node is visited, the feasibility test, asking for the required number of time slots which cannot exceed S (see constraint (4)), is applied. Let p be the visiting node in the tree and ($x_{1y_1}=1, x_{2y_2}=1, ..., x_{iy_i}=1$) be the path from the root to the node p. Thus, the total number of time slots consumed so far can be computed by $h(p) = t_{1y_1}+t_{2y_2}+...+t_{iy_i}$. If $h(p) \leq S$, node p is feasible; otherwise, node p is infeasible.

3.2.3. Selection of a branching node and termination condition

To handle the generation of the combinatorial tree, a data structure (live-node list) records all live nodes that are waiting to be branched. Initially, the child nodes of the root node are generated and added to the live-node list. The search strategy of the branch and bound algorithm is the greatest

utility first. That is, the node, say p, selected for next branching is the live node whose $g(p)$ is the greatest among all the nodes in the live-node list. If node p is feasible, then the child nodes of p are added to the live-node list. For example, if node 3 is feasible and selected for branching, then three nodes, 4, 8, and 11 are generated and added to the live-node list (see Figure 4).

Traversal of the combinatorial tree starts at the root node and stops when the live-node list is empty. In addition, a lower bound of total utility (LT) is associated with the branch and bound algorithm. $LT = 0$, initially, and is updated to be max (LT, $f(u)$) whenever a feasible node u is reached. If node p satisfies $g(p) \leq LT$ (i.e. the maximum total utility of node p is smaller than or equal to the lower bound total utility of the current optimal solution), then it is bounded since further branching from p does not lead to a better solution. If node p is infeasible, then it is bounded since further branching from p does not lead to a feasible solution. When any branch is terminated, the next live-node is chosen by the greatest utility policy. If the live-node list becomes empty, the optimal solution is defined by the path from the root to the node w with $f(w) = LT$. Optimal utility LT is the output of Algorithm 1.

Algorithm 1: Branch and bound algorithm for solving the problem TUMP

```
1  Initialize the live-node list to be empty;
2  Put root node v1 on the live-node list;
3  Set f(v1) := 0;
4  Set LT := 0;
5  while live-node list is not empty do
6  begin
7      choose node p with the greatest value of g(p) from the live-node list;
8      Set G := 0;
9      if h(p) > S then
10         remove node p from the live-node list;
11     else begin
12         Put the child nodes of node p into set G;
13         for each node u in G do
14             begin
15                 if g(u) > LT then
16                     set max (LT, f(u));
17             end;
18             insert node u into the live-node list;
19         end;
20     remove node p from the live-node list;
21     end;
22 end;
23 output the answer: node w and the optimal value g(w) := LT;
```

4. Numerical Example and Results

4.1. A Numerical Example

Consider an example of a scalable video with four video layers (i.e., l_1, l_2, l_3, l_4). The BS supports three MCSs (i.e., M_1, M_2, M_3). Suppose that $[\delta_i] = \begin{bmatrix}0.4 & 0.3 & 0.2 & 0.1\end{bmatrix}^T$, $[u_j] = \begin{bmatrix}7 & 3 & 2\end{bmatrix}$ and $[r_j] = \begin{bmatrix}48 & 96 & 192\end{bmatrix}$(bits per time slot). We assume that each video layer has the same size, λ = 192 bits per frame; that is, $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \lambda$. We then assume that the number of available time slots S = 21. The number of required time slots $[t_{ij}]$ and the total utility $[w_{ij}]$ can be found as follows

$$[t_{ij}] = \begin{bmatrix}8 & 4 & 2\\ 8 & 4 & 2\\ 8 & 4 & 2\\ 8 & 4 & 2\end{bmatrix},$$

$$[w_{ij}] = \begin{bmatrix} 2.8 & 1.2 & 0.8 \\ 2.1 & 0.9 & 0.6 \\ 1.4 & 0.6 & 0.4 \\ 0.7 & 0.3 & 0.2 \end{bmatrix}.$$

First, as shown in Figure 5, the algorithm checks if node 1 is a feasible node or not. Because $h(1) = 0$ which is smaller than 21, node 1 is a feasible node. The current total utility is $f(1) = 0$. Then, the algorithm adds nodes 2, 22, and 32 to the live-node list and computes $g(2)$, $g(22)$, and $g(32)$. By equation (18), we obtain:

$g(2) = f(2) + w_{21} + w_{31} + w_{41} = w_{11} + w_{21} + w_{31} + w_{41} = 2.8 + 2.1 + 1.4 + 0.7 = 7$,

$g(22) = f(22) + w_{22} + w_{32} + w_{42} = w_{12} + w_{22} + w_{32} + w_{42} = 1.2 + 0.9 + 0.6 + 0.3 = 3$,

$g(32) = f(32) + w_{23} + w_{33} + w_{43} = w_{13} + w_{23} + w_{31} + w_{41} = 0.8 + 0.6 + 0.4 + 0.2 = 2$.

Since $g(2) = 7$ is the greatest value among nodes 2, 22, and 32, the algorithm chooses node 2 for branching.

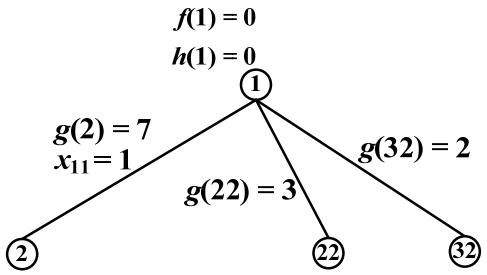


Figure 5: The algorithm chooses node 2 for branching.

Next, the algorithm checks the feasibility of node 2 (see Figure 6). Because $h(2) = t_{11} = 8 < 21$, node 2 is a feasible node. The current total utility is $f(2) = w_{11} = 2.8$. Then, the algorithm adds nodes 3, 13, and 19 to the live-node list. Because $g(3) = f(3) + w_{31} + w_{41} = (w_{11} + w_{21}) + w_{31} + w_{41} = (2.8 + 2.1) + 1.4 + 0.7 = 7$ is the greatest value among nodes 3, 13, and 19, it chooses node 3 for branching.

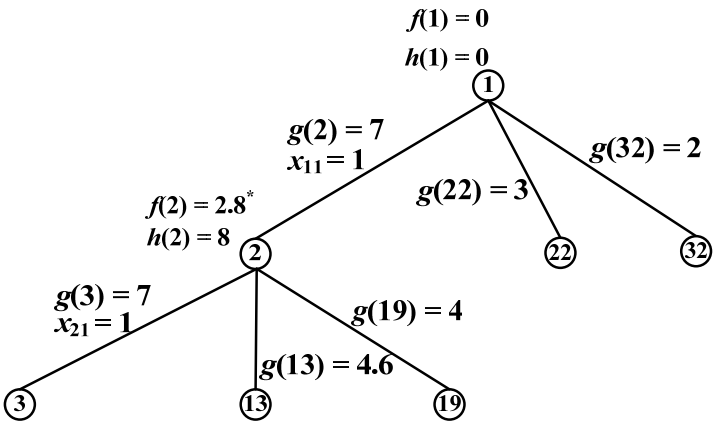


Figure 6: The algorithm chooses node 3 for branching and the current optimal solution is (x_{11} = 1) and current total utility is z^* = f(2) = 2.8.

Because h(4) = t_{11} + t_{21} + t_{31} = 24 > 21, node 4 is infeasible and gets killed (or bounded). By the same method, the algorithm chooses node 8 for branching (see Figure 7). Since h(9) = 24 > 21 and h(10) = 22 > 21, nodes 9 and 10 get killed. The algorithm finds the next node for branching from the live-node list. Since $g(p)$, p = 11, 13, 19, 22, 23, which are smaller than or equal to LT = f(8) = 5.5, nodes 11, 13, 19, 22, and 32 are bounded. Now, the live-node list is empty and then the algorithm will be terminated. The maximum utility answer node is node 8. It has a utility of 5.5. That is, the optimal solution is (x_{11} = 1, x_{21} = 1, x_{32} = 1). The video layers l_1, l_2, and l_3 are selected to be delivered and are encoded by M_1, M_1, and M_2, respectively.

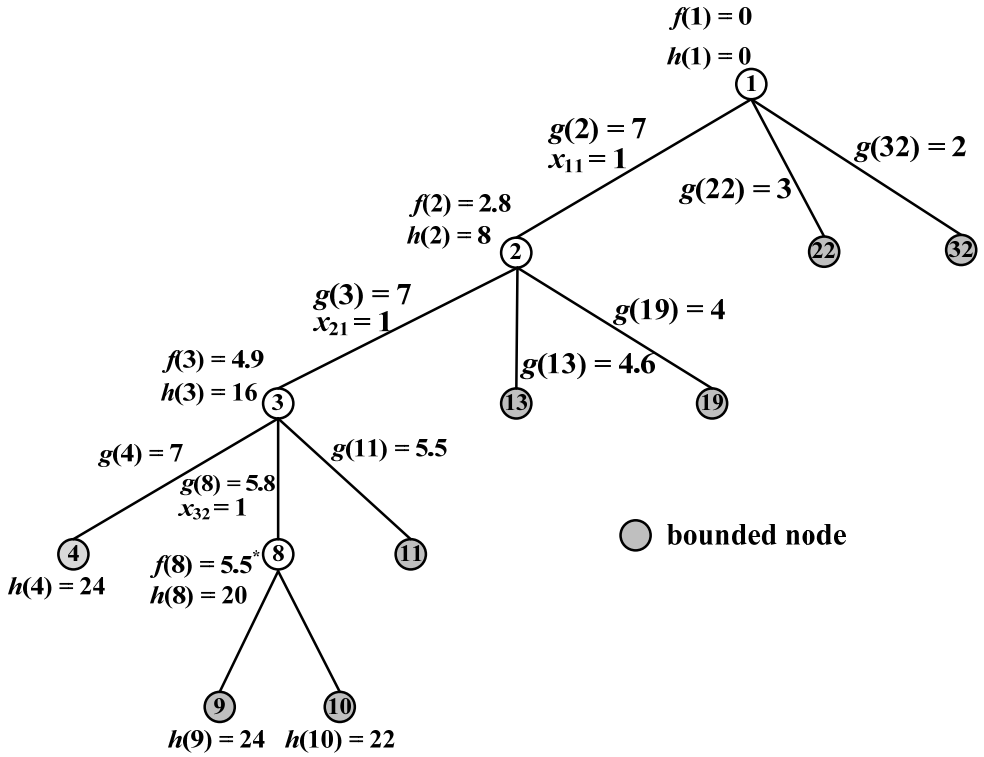


Figure 7: The optimal solution of a video stream is (x_{11} = 1, x_{21} = 1, x_{32} =1) and optimal total utility $z^* = f(8) = 5.5$.

4.2. Experimental Results

We have conducted simulations to demonstrate how effective the proposed mathematical model is. The simulation ran on a BS with 100 MSSs which were randomly placed within a cell. The coverage area of the BS was divided into six rings, P_1, P_2, …, and P_6 as shown in Figure 8. Six types of MCS as in the IEEE 802.16e standard [3][4][5] were used (i.e., n = 6). The MSS in rings P_1, P_2, …, and P_6 can be accommodated with MCS sets {M_1, M_2, M_3, M_4, M_5, M_6}, {M_1, M_2, M_3, M_4, M_5}, …, and {M_1}, respectively. The video stream was divided into one base layer and six enhancement layers (i.e., m = 7). The utility function was assumed to be log-normally distributed due to the attenuation caused by shadowing or slow fading in the wireless communication. The shape parameter and the scale parameter of the utility function were set to 1.5 and 0.5, respectively (see Figure 3) [10].

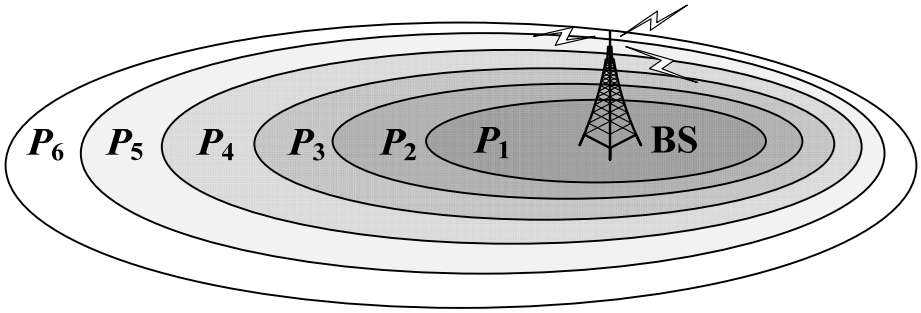


Figure 8: The coverage area of the BS with six rings.

Three assigning MCS methods were considered in the simulation:

1) The naive method: it chooses the highest MCS, which can be received by all MSSs in the multicast group, to encode the video layers, and allocates the available timeslots to the video layers one by one until the remaining timeslots cannot accommodate the next layer.
2) The uniform method [27]: it chooses the highest MCS, which can be received by all MSSs in the multicast group, to encode the base layer. Next, the uniform algorithm chooses the MCS which covers at least 60% of the MSSs in the multicast group to encode the enhancement layers.
3) The proposed method: it solves the TUMP problem to find the optimal MCS for each video layer by the branch and bound algorithm.

The total utility values achieved by the naive method, the uniform method and the proposed method are denoted by X_{naive}, X_{uni} and X_{opt}, respectively. The comparisons among X_{naive}, X_{uni} and X_{opt} are made (shown in Figure 9). Each data point in Figure 9 is the average over 10 runs. The results show that the total utility values X_{opt} are greater than X_{uni} or X_{naive}. The gaps among X_{naive}, X_{uni} and X_{opt} are larger when the available bandwidth is in the range of 1500 to 3000 timeslots/sec.

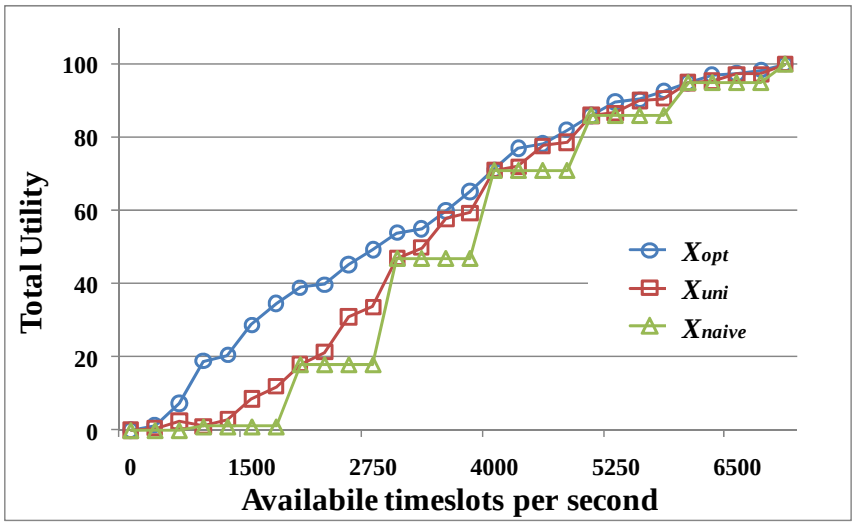


Figure 9: The utility of the optimal solution, the uniform algorithm, and the naive algorithm with different available timeslots per second.

Figure 10 shows one sample of the simulation results for the optimal algorithm and the uniform algorithm with the number of available timeslots S = 2500. As shown in Figure 10(a), for both algorithms, the MSS can receive more video layers when it is more closed to the BS. However, the numbers of video layers delivered by the optimal algorithm to the MSS in all rings except ring P_6 are greater than or equal to the numbers of video layers delivered by the uniform algorithm. Similarly, from Figure 10(b), note that the utility values achieved by the optimal algorithm are greater than or equal to the values achieved by the uniform algorithm for all rings except ring P_6. In this sample, the numbers of the MSSs for rings P_1, P_2, P_3, P_4, P_5, and P_6 were 3, 5, 42, 7, 10, and 33, respectively. The total utility achieved by the proposed algorithm was 40.92 (= $(3+5+42)\times 0.71 + 7\times 0.47 + 10\times 0.18 + 33\times 0.01$), while it achieved by the uniform algorithm was 34.53 (= $(3+5+42+7)\times 0.47 + (10+33)\times 0.18$). The optimal algorithm shows its benefit.

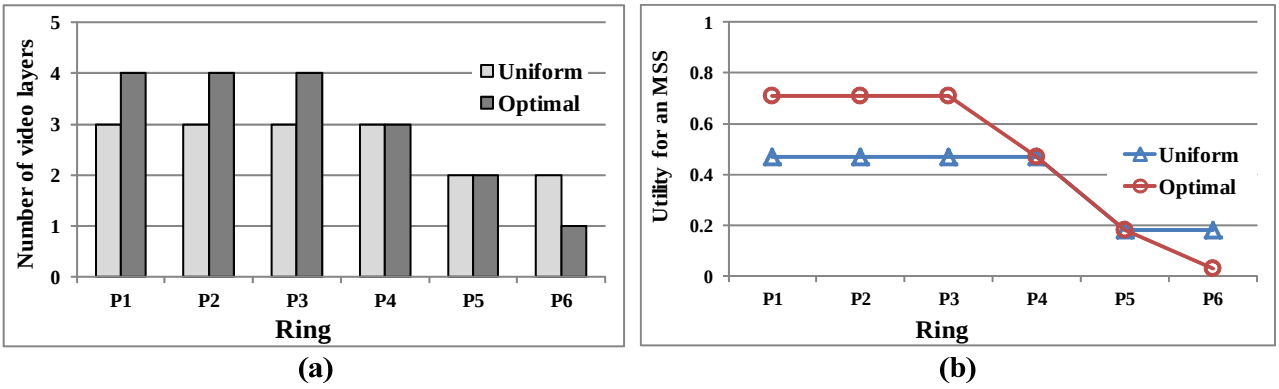


Figure 10: The number of video layers that an MSS can receive and the utility of an MSS under various rings when the available timeslots (S) equal to 2500.

On the other hand, we also present the computational experiments to show the effectiveness of the branch and bound algorithm. The real execution times of the algorithm depend on the number of video layers (m), the number of MCSs (n), and the number of available time slots (S). The experiments were conducted on a desktop PC with an Intel Core 2 Duo 1.6GHz processor and 2 GB memories. The operating system was Windows XP. The programs were coded in C and are available from the corresponding author upon request.

The simulation also ran on a BS with 100 MSSs which were randomly placed. We assume the frame duration is 5 *ms*. Each MSS subscribes a scalable video, in which the video rate is 320 kbps (i.e., 1.6 kb per frame). The video rate is a measure of the rate of information content in a video stream. The video is divided equally across the number of video layers. The simulation results are summarized in Table 3, which includes the number of nodes generated, the number of computations of $f(p)$, and the execution time (CPU time). Table 3 shows that Algorithm 1 appreciably reduces the number of nodes generated and the number of unnecessary tries for infeasible nodes. For example, if you apply an exhaustive search to a problem with size (m, n) = (10, 6), then the total number of nodes is $\binom{10+6}{6} = 8008 \approx 2^{14}$. However, the number of nodes generated by Algorithm 1 is only 202 < 2^8 (see Table 3) for (m, n, S) = (10, 6, 2×10^3) because we apply the branch and bound approach. It takes 107 tries to compute $f(p)$ for obtaining the MCS assignment of the layering structure, which is much less than 8008. This means that the dominance rules (i.e., feasibility test and utility bound) can be used to discard the most infeasible and unnecessary nodes before computing $f(p)$. This reduces the

computational time significantly.

From Table 3, the computational time to determine the optimal MCS assignment for the video layering structure is less than 75.604 μs. The computational time is small enough. Thus, the branch and bound method is effective and suitable for BS to determine the video layering structure and MCS assignment for IEEE 802.16e network multicast.

Table 3: The simulation results under various numbers of MCSs, video layers, and available time slots.

m	$n = 3$				$n = 6$			
	S	Computations of $f(p)$	Nodes generated	CPU time (μ sec.)	S	Computations of $f(p)$	Nodes generated	CPU time (μ sec.)
2	2×10^3	3	6	0.842	2×10^3	1	4	0.914
	4×10^3	1	3	0.634	4×10^3	1	6	1.138
	6×10^3	2	5	0.756	6×10^3	2	11	1.442
	8×10^3	2	6	0.817	8×10^3	2	12	1.618
4	2×10^3	9	9	2.928	2×10^3	6	14	4.190
	4×10^3	3	7	1.727	4×10^3	3	15	3.038
	6×10^3	4	11	2.021	6×10^3	4	22	3.500
	8×10^3	4	12	2.105	8×10^3	4	24	3.580
6	2×10^3	19	19	6.655	2×10^3	22	43	14.027
	4×10^3	4	10	2.813	4×10^3	5	22	5.220
	6×10^3	5	15	3.583	6×10^3	5	30	6.286
	8×10^3	6	18	3.831	8×10^3	6	36	6.632
8	2×10^3	22	25	9.673	2×10^3	47	96	32.430
	4×10^3	7	15	4.955	4×10^3	6	29	8.512
	6×10^3	7	21	5.583	6×10^3	7	42	9.869
	8×10^3	8	24	5.980	8×10^3	8	48	10.357
10	2×10^3	40	43	17.61	2×10^3	107	202	75.604
	4×10^3	10	20	7.397	4×10^3	13	50	16.035
	6×10^3	10	27	8.210	6×10^3	10	55	15.239
	8×10^3	10	30	8.225	8×10^3	10	60	14.747

5. Conclusion

In this paper, we consider an optimal MCS assignment problem which improves spectral efficiency and maximizes total utility for the scalable video multicast in IEEE 802.16e networks. We propose a branch and bound algorithm to find an optimal solution for this problem. In the experiment, it was shown that the proposed method performs well compared to the uniform method and the naïve method. The computation time of the proposed branch and bound algorithm is very small. Thus, our

proposed method is suitable for BS to determine the video layering structure and the MCS assignment in the IEEE 802.16e network multicast.

Because of the Doppler Effect, when an MSS is moving, the MSS's velocity causes a shift in the frequency of the signal transmitted along each signal path. It causes fast fading of the received signal for an MSS. Thus, the BS will encode the video layers by the more conservative and robust MCS for the moving MSS. Therefore, the video quality for an MSS when it stands at a point is better than that when it moves within the same region. Looking ahead, considering mobility of MSS for the MCS assignment problem might be interesting future work.

References

[1] 4G coverage, Sprint, http://www.sprint.com/.

[2] VMAX, http://www.vmax.net.tw/.

[3] IEEE computer society and IEEE microwave theory and techniques society, IEEE Standard for Local and metropolitan area networks Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems, IEEE standard 802.16e-2005, Dec. 2005.

[4] IEEE computer society and IEEE microwave theory and techniques society, IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems, IEEE standard 802.16-2004, June 2004.

[5] J. G. Andrews, A. Ghosh, and R. Muhamed, *Fundamentals of WiMAX: Understanding Broadband Wireless Networking*, 1st Edition, Prentice Hall, New Jersey, USA, Feb. 27, 2007.

[6] M. Hauge and Ø. Kure, "Multicast in 3G Networks: Employment of Existing IP Multicast Protocols in UMTS," *International Workshop on Wireless Mobile Multimedia* (*WoWMoM*), Sep. 2002, Atlanta, USA.

[7] N, Jindal and Z.Q. Luo, "Capacity Limits of Multiple Antenna Multicast," *IEEE International Symposium on Information Theory* (*ISIT*), July 2006, Seattle, USA.

[8] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 17, no. 9, Sept. 2007, pp. 1103–1129.

[9] S. Shenker, "Fundamental Design Issues for the Future Internet," *IEEE Journal on Selected Areas in Communications*, Vol. 13, No. 7, Sep. 1995, pp. 1176–1188.

[10] L. Shi, C. Liu, and B. Liu, "Network Utility Maximization for Triple-play Services," *Computer Communications*, Vol. 31, Feb. 2008, pp. 2257–2269.

[11] Video trace library, Arizona State University, http://trace.eas.asu.edu/.

[12] J. Liu, B. Li, Y. T. Hou, and I. Chlamtac, "Dynamic Layering and Bandwidth Allocation for Multisession Video Broadcasting with General Utility Functions," *The 22th IEEE International Conference on Computer Communications* (*IEEE INFOCOM*), Apr. 2003, San Francisco, USA.

[13] W. H. Kuo, T. Liu, and W. Liao, "Utility-based Resource Allocation For Layer-encoded IPTV Multicast in IEEE 802.16 (WiMAX) Wireless Networks," *IEEE International Conference on Communications* (*ICC*), June 2007, Glasgow, Scotland.

[14] P. Li, H. Zhang, B. Zhao, and S. Rangarajan, “Scalable Video Multicast in Multi-carrier Wireless Data Systems,” *The 17th IEEE International Conference on Network Protocols* (*ICNP*), Oct. 2009, Princeton, USA.

[15] H. Chi, C. Lin, Y. Chen, and C. Chen, “Optimal Rate Allocation for Scalable Video Multicast over WiMAX,” *IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2008, Seattle, USA.

[16] J. Shi, D Qu, and G. Zhu, “Utility Maximization of Layered Video Multicasting for Wireless Systems with Adaptive Modulation and Coding,” *IEEE International Conference on Communications* (*ICC*), June 2006, Istanbul, Turkey.

[17] S. Deb, S. Jaiswal, and K. Nagaraj, “Real-time Video Multicast in WiMAX Networks,” *The 27th IEEE Conference on Computer Communications* (*IEEE INFOCOM*), Apr. 2008, Phoenix, USA.

[18] C. Huang, P. Wu, S. Lin, and J. Hwang, “Layered Video Resource Allocation in Mobile WiMAX Using Opportunistic Multicasting,” *IEEE Wireless Communication and Networking Conference* (*WCNC*), Apr. 2009, Budapest, Hungary.

[19] C. S. Hwang and Y. Kim, “An Adaptive Modulation Method for Multicast Communications of Hierarchical Data in Wireless Networks,” *IEEE International Conference on Communications* (*ICC*), Apr. 2002, New York, USA.

[20] M. Shabany, K. Navaie, and E.S. Sousa1, “A Utility-Based Downlink Radio Resource Allocation for Multiservice Cellular DS-CDMA Networks,” *EURASIP Journal on Wireless Communications and Networking*, Vol. 2007, Issue 1, Jan. 2007.

[21] J. Kim and D. Cho, “Enhanced Adaptive Modulation and Coding Schemes Based on Multiple Channel Reportings for Wireless Multicast Systems,” *IEEE Vehicular Technology Conference* (*VTC* 2005 Fall), Sep. 2005, Dallas, USA.

[22] H. Wang, H. P. Schwefel, and T. S. Toftrgaard, “History-based Adaptive Modulation for a Downlink Multicast Channel in OFDMA systems,” *IEEE Wireless Communication and Networking Conference* (*WCNC*), Mar. 2008, Las Vegas, USA.

[23] H. Kellerer, U. Pferschy, and D. Psdinger, *Knapsack Problem*, Springer-Verlag, Berlin, 2004.

[24] W. Nelson, *Applied life data analysis*, John Wiley & Sons, Toronto, New York, USA, 1982, pp. 32–36.

[25] R. Breu and C. Burdet, “Branch and Bound Experiments in 0-1 Programming,” *Mathematical Programming Study 2*, 1974, pp. 1 – 50.

[26] R. Granfinkal and G. Nemhauser, *Integer Programming*, Wiley, London, 1972.

[27] A. M. C. Correia, J. C. M. Silva, N. M. B. Souto, L.A. C. Silva, A. B. Boal, and A. B. Soares, “Multi-Resolution Broadcast/Multicast Systems for MBMS,” *IEEE Transaction on Broadcasting*, Vol. 53, no. 1, pp. 224 – 234, Mar. 2007.

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2009 proceedings

A Chaotic Maps-based Key Agreement Protocol that Preserves User Anonymity

Huei-Ru Tseng, Rong-Hong Jan, and Wuu Yang
Department of Computer Science
National Chiao Tung University
Hsinchu, Taiwan 30010
{hueiru, rhjan, wuuyang}@cs.nctu.edu.tw

***Abstract*—A key agreement protocol is a protocol whereby two or more communicating parties can agree on a key or exchange information over an open communication network in such a way that both of them agree on the established session keys for use in subsequent communications. Recently, several key agreement protocols based on chaotic maps are proposed. These protocols require a verification table to verify the legitimacy of a user. Since this approach clearly incurs the risk of tampering and the cost of managing the table and suffers from the stolen-verifier attack, we propose a novel key agreement protocol based on chaotic maps to enhance the security. The proposed protocol not only achieves mutual authentication without verification tables, but also allows users to anonymously interact with the server. Moreover, security of the proposed protocol is modelled and analyzed with Petri nets. Our analysis shows that the proposed protocol can successfully defend replay attacks, forgery attacks, and stolen-verifier attacks.**

***Index Terms*—Key agreement protocol, Chaotic maps, Stolen-verifier attacks, Anonymity, Petri nets.**

I. Introduction

A key agreement protocol is a protocol whereby two or more communicating parties can agree on a key or exchange information over an open communication network in such a way that both of them agree on the established session keys for use in subsequent communications. In 1976, Diffie and Hellman invented the first key agreement protocol [1], in which two parties jointly exponentiate a generator with random numbers, in such a way that an eavesdropper has no way of guessing the key. However, their protocol does not provide authentication of the communicating parties, and is thus vulnerable to the man-in-the-middle attacks. Since then, a variety of secure key agreement protocols have been developed to prevent man-in-the-middle and related attacks.

Since the 1990s, chaotic systems [2-7] have been used to design secure communication protocols. Two main approaches to the use of chaotic systems in designing communication protocols are analog and discrete digital. The former is based on chaos synchronization using chaotic circuits, and the latter is designed for generating chaotic ciphers.

This work was supported by the National Science Council, Taiwan, Republic of China, under grant NSC 97-2221-E-009-048-MY3, NSC 97-2221-E-009-049-MY3, and NSC 96-2628-E-009-014-MY3.

In 2003, Kocarev and Tasev [8] proposed a public-key encryption algorithm based on Chebyshev chaotic maps [9] as its semi-group properties meet the cryptographic requirements. However, Bergamo et al. [10] proved that Kocarev and Tasev's protocol [8] is insecure since an adversary can efficiently recover the plaintext from a given ciphertext. Later, in order to address Bergamo et al.'s attack [10], Xiao et al. proposed a novel key agreement protocol [11]. Recently, Han [12] pointed out that Xiao et al.'s protocol [11] is still insecure against their new attacks that can hinder the user and the server from establishing a session key even though the adversary cannot obtain any private information from the communicating parties. In 2008, Yoon and Yoo [13] proposed a new key agreement protocol based on chaotic maps that can resist Han et al.'s developed attacks [12] and off-line password guessing attacks, and can reduce the numbers of communication rounds.

However, these protocols [11, 13] still have several security weaknesses. In these protocols, the server needs a verification table. The verification table could be tampered or stolen and there is the cost of managing the table. In addition, users would wish to obtain services anonymously.

Taking the security threats and privacy issues into consideration, we propose a chaotic maps-based key agreement protocol that not only fixes these weaknesses, but also aims to preserve user anonymity. The crucial merits of the proposed protocol include: (1) it achieves mutual authentication between a server and a user; (2) it allows users to anonymously interact with the server to agree on session keys; (3) a server and a user can generate sessions keys for protecting the subsequent communications. Moreover, Petri nets [14] may be used to infer what an attacker could know if he happens to know certain items in the security protocol. We used Petri nets in the security analysis of the proposed protocol. Our analysis shows that the proposed protocol can successfully defend replay attacks, forgery attacks, and stolen-verifier attacks.

The rest of this paper is organized as follows: In Section 2, we state the definitions of Chebyshev chaotic map and introduce the hash function based on chaotic maps. Next, our proposed protocol is presented in Section 3. Then, we shall analyze our proposed protocol, show that our protocol can resist several attacks, and provide a comparative study with

978-1-4244-3435-0/09/$25.00 ©2009 IEEE

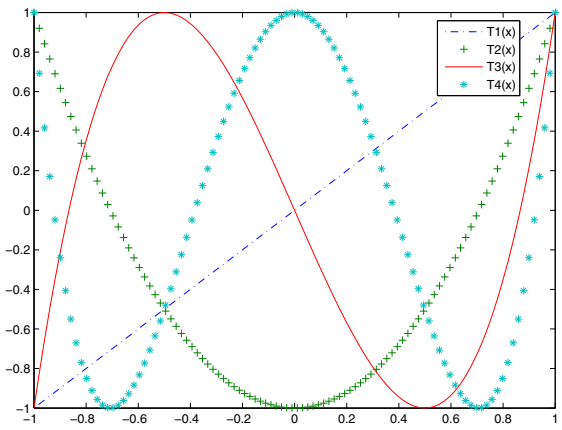


Fig. 1. Chebyshev polynomials

other key agreement protocols in Section 4. Finally, we will conclude our paper in Section 5.

II. Preliminaries

In this section, we define Chebyshev chaotic maps and introduce the hash functions based on chaotic maps.

A. Chebyshev Chaotic Maps

Chebyshev polynomial [9] and its properties [8, 11, 13] are described as follows.

Definition 1. *The Chebyshev polynomial $T_n(x)$ is a polynomial in x of degree n, defined by the following relation:*

$$T_n(x) = \cos n\theta, \text{ where } x = \cos\theta \ (-1 \le x \le 1) \quad (1)$$

With Definition 1, the recurrence relation of $T_n(x)$ is defined as:

$$T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x), \text{for any } n \ge 2, \quad (2)$$

together with the initial conditions $T_0(x) = 1, T_1(x) = x$.

Some examples of Chebyshev polynomials are shown as follows: (see Figure 1)

$$T_2(x) = 2x^2 - 1 \quad (3)$$
$$T_3(x) = 4x^3 - 3x \quad (4)$$
$$T_4(x) = 8x^4 - 8x^2 + 1 \quad (5)$$

Chebyshev polynomials have two important properties [8, 11, 13]: the semi-group property and the chaotic property.

- The semi-group property:

$$\begin{aligned} T_r(T_s(x)) &= \cos(r\cos^{-1}(\cos(s\cos^{-1}(x)))) \\ &= \cos(rs\cos^{-1}(x)) \\ &= T_{sr}(x) \\ &= T_s(T_r(x)) \end{aligned} \quad (6)$$

- The chaotic property: If the degree $n > 1$, Chebyshev polynomial map: $T_n : [-1, 1] \to [-1, 1]$ of degree n is a chaotic map with its invariant density $f^*(x) = \frac{1}{\pi\sqrt{1-x^2}}$ for Lyapunov exponent $\lambda = \ln n > 0$.

B. Hash Functions based on Chaotic Maps

The hash function used in the previous key agreement protocols [11, 13] is based on the following chaotic one-way hash function [15]. A one-dimension piecewise linear chaotic system is defined as:

$$X(t+1) = F(X(t), P) \quad (7)$$

where $F(u, P) =$

$$\begin{cases} u/P & \text{if } 0 \le u < P, \\ (u-P)/(0.5-P) & \text{if } P \le u < 0.5, \\ (1-u-P)/(0.5-P) & \text{if } 0.5 \le u < 1-P, \\ (1-u)/P & \text{if } 1-P \le u \le 1, \end{cases}$$

where $X \in [0, 1]$ and $P \in (0, 0.5)$. X_i is the chaining variable, where $0 \le i \le 3N$. X_0 is an initial value of the chaining variable and is chosen from $(0, 1)$.

Given a pending message M, H_0 is a constant which is chosen from $(0, 1)$. The 3-unit iterations—1st to N-th, $(N+1)$-th to $2N$-th, $(2N+1)$-th to $3N$-th—ensure that each bit of the final hash value will be related to all bits of the message. The following is a brief referring to how to generate the hash value:

- The pending message M is translated to the corresponding ASCII numbers, then by means of linear transform, these ASCII numbers are mapped into an array C whose length N is the number of characters in the message and whose elements are numbers in $[0, 1]$.
- The iteration process is as follows:
  1) 1st: $P_1 = (C_1 + H_0)/4 \in [0, 0.5), X_1 = F(X_0, P_1) \in [0, 1]$;
  2) 2nd to N-th: $P_i = (C_i + X_{i-1})/4 \in [0, 0.5), X_i = F(X_{i-1}, P_i) \in [0, 1]$;
  3) $(N+1)$-th: $P_{N+1} = (C_N + X_N)/4 \in [0, 0.5), X_{N+1} = F(X_N, P_{N+1}) \in [0, 1]$;
  4) $(N+2)$-th to $2N$-th: $P_i = (C_{2N-i+1} + X_{i-1})/4 \in [0, 0.5), X_i = F(X_{i-1}, P_i) \in [0, 1]$;
  5) $(2N+1)$-th: $P_{2N+1} = (C_1 + H_0)/4 \in [0, 0.5), X_{2N+1} = F(X_{2N}, P_{2N+1}) \in [0, 1]$;
  6) $(2N+2)$-th to $3N$-th: $P_i = (C_{i-2N} + X_{i-1})/4 \in [0, 0.5), X_i = F(X_{i-1}, P_i) \in [0, 1]$.
- Next, X_N, X_{2N}, X_{3N} are transformed to the corresponding binary format, and 40, 40, 48 bits after the decimal point are extracted, respectively, and are juxtaposed from left to right to form a 128-bit hash value.

For more details, the reader is referred to [15].

III. Proposed Key Agreement Protocol

In this section, we propose a chaotic maps-based key agreement protocol. The proposed protocol does not require a verification table while achieving both mutual authentication and session key agreement between a server and a user. We list the notations used in this paper in Table I.

Different from the previous key agreement protocols [11, 13] where the server and user i share the hash value $h_{PW} =$

TABLE I
NOTATIONS

Symbol	Definition
U_i	User i
ID_i	User i's identity
PW_i	User i's password
K_s	The server's private key
sn	The session number
$H(\cdot)$	A one-way hash function based on chaotic maps
$E(\cdot)$	A symmetric key encryption algorithm
$D(\cdot)$	A symmetric key decryption algorithm
SK_i	The session key constructed by the server and user i
$\oplus$	The exclusive-or (XOR) operation

$H(ID_i, PW_i)$, the server does not require any verification table in the proposed protocol. Before performing the key agreement protocol, the server first publishes system parameters including Chebyshev polynomials, $E(\cdot)$, $D(\cdot)$, and $H(\cdot)$. Suppose a new user U_i with the identity ID_i wants to communicate with a server for establishing session keys. U_i randomly chooses his password PW_i and sends the pair $(ID_i, H(PW_i))$ to the server in person or through an existing secure channel. Upon receiving the message, the server juxtaposes ID_i and $H(PW_i)$ from left to right as the pending message, and uses the one-way hash function $H(\cdot)$ to compute $H(ID_i, H(PW_i))$. Then the server computes Reg_i as follows:

$$Reg_i = H(ID_i, H(PW_i)) \oplus H(K_s) \tag{8}$$

where K_s is the server's private key.

After that, the server transmits Reg_i back to U_i over a secure channel. Note that U_i has to keep Reg_i secret.

The details of the proposed key agreement protocol are presented as follows.

1) $U_i \rightarrow Server : \{sn, R_i, C_1\}$
   U_i first chooses three random numbers r_i, r, and v, where $r_i \in [-1, 1]$ is the seed x of the Chebyshev polynomial of degree r and v is a nonce. Next, U_i computes the pair (R_i, K_i) as follows.

$$R_i = Reg_i \oplus H(v) \tag{9}$$

$$K_i = H(ID_i, H(PW_i)) \oplus H(v) \tag{10}$$

   Then U_i encrypts ID_i, r_i, and $T_r(x)$ with K_i:

$$C_1 = E_{K_i}(ID_i, r_i, T_r(x)) \tag{11}$$

   Finally, U_i transmits sn, R_i, and C_1 to the server, where sn is the session number.

2) $Server \rightarrow U_i : \{sn, ID_s, C_2, AU_s\}$
   Upon receiving the message, the server computes $K_i = R_i \oplus H(K_s)$, and extracts ID_i, r_i, and $T_r(x)$ from C_1 with K_i. The server first checks the validity of ID_i, and then chooses two random numbers s and r_t, where s is the degree of the Chebyshev polynomial and r_t is a nonce. Next, the server computes the pair (C_2, SK_i) as follows.

$$C_2 = E_{K_i}(ID_s, r_t, T_s(x)) \tag{12}$$

$$SK_i = T_s(T_r(x)) = T_{rs}(x) \tag{13}$$

   Finally, the server computes the authentication value AU_s and sends sn, ID_s, C_2, and AU_s back to U_i.

$$AU_s = H(ID_i, r_i, r_t, SK_i) \tag{14}$$

3) $U_i \rightarrow Server : \{sn, AU_i\}$
   After receiving the message, U_i extracts ID_s, r_t, and $T_s(x)$ from C_2 with K_i. Next, U_i computes the pair (SK_i, AU'_s) as follows.

$$SK_i = T_r(T_s(x)) = T_{rs}(x) \tag{15}$$

$$AU'_s = H(ID_i, r_i, r_t, SK_i) \tag{16}$$

   Then U_i checks whether AU_s and AU'_s are equal. If so, the identity of the server is authenticated. Next, U_i computes AU_i as follows.

$$AU_i = H(ID_s, r_i, r_t, SK_i) \tag{17}$$

   Finally, U_i sends sn and AU_i back to the server.

4) After receiving sn and AU_i, the server computes AU'_i as follows.

$$AU'_i = H(ID_s, r_i, r_t, SK_i) \tag{18}$$

   Then the server checks whether AU_i and AU'_i are equal. If so, the identity of U_i is authenticated.

After mutual authentication and key agreement between U_i and the server, SK_i is used as a shared session key.

IV. ANALYSIS OF OUR SCHEME

In this section, we show that our protocol can resist several notorious attacks. In addition, we provide a comparative study with other key agreement protocols.

A. Security Analysis

We first use Petri nets [14] to model and analyze the proposed protocol. Next, security properties of our protocol will be specified.

1) Petri Net Model: We used a Petri net to model our security protocol. The formal definition of a Petri net [16] is listed in Table II. Petri nets are composed from graphical symbols designating places (shown as circles), transitions (shown as rectangles), and directed arcs (shown as arrows). The places denote (atomic and composite) data items. The transitions denote decryption or decomposition operations. Arcs run between places and transitions.

When a transition fires, a composite data item is decomposed or decrypted, resulting in one or more simpler data items. Since we assume an open network environment, all data items in the transmitted messages are assumed to be public, and are known to the attacker. There will be tokens in the places representing the data items in the transmitted messages initially. From this initial marking, we can infer what an attacker can know eventually. Furthermore, we can also experiment what an attacker can know if he knows additional data items from other sources. The Petri net model

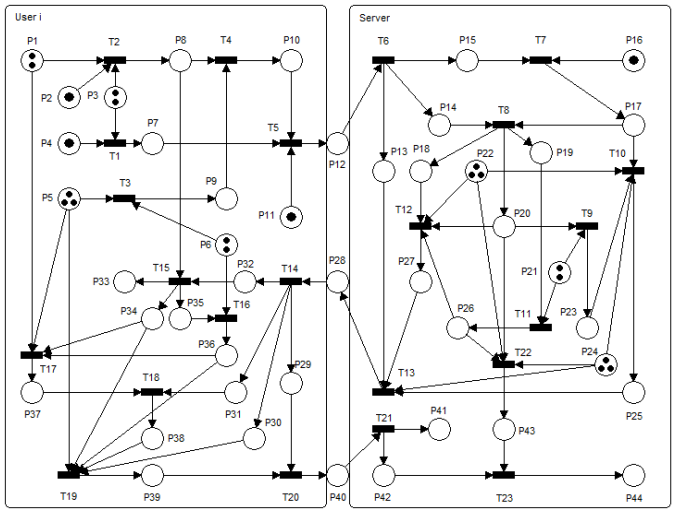


Fig. 2. A Petri net model of the proposed key agreement protocol

TABLE II
FORMAL DEFINITION OF A PETRI NET

A Petri net is a 5-tuple, $PN = (P, T, F, W, M_0)$ where:
$P = \{P_1, P_2, \cdots, P_m\}$ is a finite set of places,
$T = \{T_1, T_2, \cdots, T_n\}$ is a finite set of transitions,
$F \subseteq (P \times T) \cup (T \times P)$ is a set of arcs (flow relation),
$W : F \rightarrow \{1, 2, 3, \cdots\}$ is a weight function,
$M_0 : P \rightarrow \{0, 1, 2, 3, \cdots\}$ is the initial marking,
$P \cap T = \emptyset$ and $P \cup T \neq \emptyset$.
A Petri net structure $N = (P, T, F, W)$ without any specific initial marking is denoted by N.
A Petri net with the given initial marking is denoted by (N, M_0).

is illustrated in Figure 2. The definitions of the places and transitions used in this model are listed in Table III and Table IV, respectively. The model is simulated with the HPSim Petri net simulation tool [17].

2) Security Properties: The security of the proposed protocol is based on the difficulty of the discrete logarithm problem (DLP) and the Diffie-Hellman problem (DHP), which are believed to be unsolvable in polynomial time. We first specify the mathematical difficult problems [13] used in this paper.

Definition 2. *The discrete logarithm problem (DLP) is defined as follows: given an element α, find the integer r such that $T_r(x) = \alpha$.*

Definition 3. *The Diffie-Hellman problem (DHP) is defined as follows: given $T_r(x)$ and $T_s(x)$, find $T_{rs}(x)$.*

Now we show that our protocol can resist replay attacks, forgery attacks, and stolen-verifier attacks, and also analyze the following security properties: mutual authentication, user anonymity, and known-key security.

Theorem 1. *The proposed protocol can resist a replay attack.*

Proof. Assume an adversary A eavesdrops the messages $\{sn, R_i, C_1\}$ and $\{sn, AU_i\}$ sent by U_i and replays them to log in to the system in a later session. Upon receiving the replay message, the server computes $K_i = R_i \oplus H(K_s)$, and extracts ID_i, r_i, and $T_r(x)$ from C_1 with K_i. The server first checks the validity of ID_i, and then chooses two

TABLE III
DEFINITIONS OF PLACES

Place	Definition	Place	Definition
P_1	ID_i	P_{23}	$T_s(x)$
P_2	$H(PW_i)$	P_{24}	ID_s
P_3	$H(v)$	P_{25}	C_2
P_4	Reg_i	P_{26}	SK_i
P_5	r_i	P_{27}	AU_s
P_6	r	P_{28}	$Packet\{sn, ID_s, C_2, AU_s\}$
P_7	R_i	P_{29}	sn
P_8	K_i	P_{30}	ID_s
P_9	$T_r(x)$	P_{31}	AU_s
P_{10}	C_1	P_{32}	C_2
P_{11}	sn	P_{33}	ID_s
P_{12}	$Pakcet\{sn, R_i, C_1\}$	P_{34}	r_t
P_{13}	sn	P_{35}	$T_s(x)$
P_{14}	C_1	P_{36}	SK_i
P_{15}	R_i	P_{37}	AU'_s
P_{16}	$H(K_s)$	P_{38}	Success verification message
P_{17}	K_i	P_{39}	AU_i
P_{18}	ID_i	P_{40}	$Packet\{sn, AU_i\}$
P_{19}	$T_r(x)$	P_{41}	sn
P_{20}	r_i	P_{42}	AU_i
P_{21}	s	P_{43}	AU'_i
P_{22}	r_t	P_{44}	Success verification message

TABLE IV
DEFINITIONS OF TRANSITIONS

Trans.	Definition	Trans.	Definition
T_1	Perform XOR operation to compute R_i	T_{13}	Transmit $\{sn, ID_s, C_2, AU_s\}$
T_2	Compute K_i	T_{14}	Split the packet
T_3	Compute $T_r(x)$	T_{15}	Decrypt C_2 with K_i
T_4	Encrypt $\{ID_i, r_i, T_r(x)\}$ with K_i	T_{16}	Compute SK_i
T_5	Transmit $\{sn, R_i, C_1\}$	T_{17}	Compute AU'_s
T_6	Split the packet	T_{18}	Check $AU_s \stackrel{?}{=} AU'_s$
T_7	Perform XOR operation to compute K_i	T_{19}	Compute AU_i
T_8	Decrypt C_1 with K_i	T_{20}	Transmit $\{sn, AU_i\}$
T_9	Compute $T_s(x)$	T_{21}	Split the packet
T_{10}	Encrypt $\{ID_s, r_t, T_s(x)\}$ with K_i	T_{22}	Compute AU'_i
T_{11}	Compute SK_i	T_{23}	Check $AU_i \stackrel{?}{=} AU'_i$
T_{12}	Compute AU_s		

random numbers s^* and r_t^*. Next, the server computes the pair (C_2^*, SK_i^*) as follows.

$$C_2^* = E_{K_i}(ID_s, r_t^*, T_{s^*}(x)) \tag{19}$$

$$SK_i^* = T_{s^*}(T_r(x)) = T_{rs^*}(x) \tag{20}$$

Finally, the server computes the authentication value AU_s^* and sends sn, ID_s, C_2^*, and AU_s^* back to A.

$$AU_s^* = H(ID_i, r_i, r_t^*, SK_i^*) \tag{21}$$

After receiving the message, A has to transmit $\{sn, AU_i^*\}$ back to the server. However, A cannot just replay the message AU_i directly since the random number r_t and the session key SK_i embedded in AU_i are different from r_t^* and SK_i^* in this

session. As shown in Figure 2, computing AU_i is defined in transition T_{19}, which has five input places, P_5, P_{30}, P_{34}, P_{36}, and P_{38}. Place P_{34} is the value of r_t and place P_{36} is the value of SK_i. Because having no idea about r_t^* and SK_i^*, the adversary cannot launch a replay attack. □

Theorem 2. *The proposed protocol can resist a forgery attack.*

Proof. If an adversary A wants to impersonate U_i, A has to create a valid authentication value AU_i^*. Assume A eavesdrops the message $\{sn, R_i, C_1\}$ sent by U_i and uses it to log in to the system in a later session. Upon receiving the message, the server computes $K_i = R_i \oplus H(K_s)$, and extracts ID_i, r_i, and $T_r(x)$ from C_1 with K_i. The server first checks the validity of ID_i, and then chooses two random numbers s^* and r_t^*. Next, the server computes the pair (C_2^*, SK_i^*) as follows.

$$C_2^* = E_{K_i}(ID_s, r_t^*, T_{s^*}(x)) \tag{22}$$

$$SK_i^* = T_{s^*}(T_r(x)) = T_{rs^*}(x) \tag{23}$$

Finally, the server computes the authentication value AU_s^* and sends sn, ID_s, C_2^*, and AU_s^* back to A.

$$AU_s^* = H(ID_i, r_i, r_t^*, SK_i^*) \tag{24}$$

However, A cannot compute a correct authentication value $AU_i^* = H(ID_s, r_i, r_t^*, SK_i^*)$ unless A can obtain K_i to get ID_i, r_i, and $T_r(x)$ by decrypting C_1 and get ID_s, r_t^*, and $T_{s^*}(x)$ by decrypting C_2^*, and also derive r from $T_r(x)$ to compute SK_i^*. Based on the difficulty of DLP, it is computationally infeasible to compute r from $T_r(x)$. As shown in Figure 2, computing SK_i^* is defined in transition T_{16}, which has two input places, P_6 and P_{35}. Place P_6 is the value of r. Because having no idea about K_i and SK_i^*, the adversary cannot compute a valid authentication value and hence cannot launch a forgery attack. □

Theorem 3. *The proposed protocol can resist a stolen-verifier attack.*

Proof. The stolen-verifier attack means that an adversary who steals the password-verifier from the server can use it directly to masquerade as a legitimate user in an authentication run. Different from the previous key agreement protocols [11, 13] where the server and user i shared the hash value $h_{PW} = H(ID_i, PW_i)$, the server does not require any verification table in the proposed protocol. Since the proposed protocol does not require a verification table, the proposed protocol can prevent the stolen-verifier attack. □

Theorem 4. *The proposed protocol can provide mutual authentication.*

Proof. The security of the session key is based on the difficulty of DLP and DHP, which are believed to be unsolvable in polynomial time. Using equation (6), the session key between the server and U_i is established as follows:

$$SK_i = T_r(T_s(x)) = T_{rs}(x) = T_s(T_r(x)) \tag{25}$$

As shown in Figure 2, computing a session key SK_i is defined in transition T_{16} and transition T_{11}. Therefore, U_i and the server can use the session key SK_i in subsequent communications. □

TABLE V
COMPARISON OF SECURITY PROPERTIES

	Xiao et al.'s protocol [11]	**Yoon & Yoo's protocol [13]**	**Proposed protocol**
Replay attacks	Insecure	Secure	Secure
Forgery attacks	Insecure	Secure	Secure
Stolen-verifier attacks	Insecure	Insecure	Secure
Mutual authentication	Not provide	Provide	Provide
User anonymity	Not provide	Not provide	Provide
Known-key security	Provide	Provide	Provide

Theorem 5. *The proposed protocol can provide user anonymity.*

Proof. If an adversary A eavesdrops the messages, he cannot extract the user's identity from the ciphertext $C_1 = E_{K_i}(ID_i, r_i, T_r(x))$ since it is encrypted with K_i, which is unknown to the adversary. In addition, due to the use of the nonce, the messages submitted to the server are different in each session. As shown in Figure 2, decrypting C_1 is defined in transition T_8, which has two input places, P_{14} and P_{17}. Place P_{17} is the value of K_i, which is only known to the user and the server. Hence, it is difficult for the adversary to discover a user's identity. Clearly, the proposed protocol can provide user anonymity. □

Theorem 6. *The proposed protocol can provide known-key security.*

Proof. Known-key security means that the compromise of a session key will not lead to further compromise of other secret keys or session keys. Even if a session key SK_i is revealed to an adversary, he still cannot derive other session keys since they are generated from the random numbers r and s. Hence, the proposed protocol can achieve known-key security. □

We summarized the security properties of key agreement protocols in Table V.

B. Efficiency Analysis

In this section, we examine the performance of our proposed protocol. The evaluation parameters are defined in Table VI. The performance comparison among the proposed protocol, Xiao et al.'s protocol [11], and Yoon & Yoo's protocol [13] is presented in Table VII. We use the computational overhead as the metric to evaluate the performance of key agreement protocols. We can see from Table VII that the computations among these protocols are very similar. The only difference is that the proposed protocol takes few more XOR operations and hash operations for each user and the server, due to fixing the security weaknesses in Xiao et al.'s protocol [11] and Yoon and Yoo's protocol [13] and preserving user anonymity.

V. CONCLUSIONS

We propose a chaotic maps-based key agreement protocol that not only fixes the weaknesses of the existing chaotic maps-based key agreement protocols [11, 13], but also aims to preserve user anonymity. The crucial merits of the proposed

TABLE VI
EVALUATION PARAMETERS

Symbol	Definition
T_X	Time for performing an XOR operation
T_H	Time for performing a one-way hash function based on chaotic maps
T_E	Time for performing a symmetric encryption operation
T_D	Time for performing a symmetric decryption operation
T_{CM}	Time for performing a Chebyshev chaotic map operation

TABLE VII
PERFORMANCE COMPARISON OF CHAOTIC MAPS-BASED KEY AGREEMENT PROTOCOLS

	Xiao et al.'s protocol [11]	**Yoon & Yoo's protocol [13]**	**Proposed protocol**
Per user	$1T_H + 1T_E + 1T_D + 2T_{CM}$	$2T_H + 1T_E + 1T_D + 2T_{CM}$	$2T_X + 5T_H + 1T_E + 1T_D + 2T_{CM}$
The server	$1T_H + 1T_E + 1T_D + 2T_{CM}$	$2T_H + 1T_E + 1T_D + 2T_{CM}$	$1T_X + 3T_H + 1T_E + 1T_D + 2T_{CM}$

protocol include: (1) it achieves mutual authentication between a server and a user; (2) it allows users to anonymously interact with the server to agree on session keys; (3) a server and a user can generate sessions keys. Moreover, we used Petri nets in the security analysis of the proposed protocol. Our analysis shows that the proposed protocol can successfully defend replay attacks, forgery attacks, and stolen-verifier attacks.

REFERENCES

[1] W. Diffie and M. Hellman, "New directions in cryptography," *IEEE Transactions on Information Theory*, vol. 22, no. 6, Nov. 1976, pp. 644-654.

[2] F. Dachselt and W. Schwarz, "Chaos and cryptography," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 48, no. 12, Dec. 2001, pp. 1498-1509.

[3] L. Kocarev, "Chaos-based cryptography: a brief overview," *IEEE Circuits and Systems Magazine*, vol. 1, no. 3, 2001, pp. 6-21.

[4] L. M. Pecora and T. L. Carroll, "Synchronization in chaotic systems," *Physical Review Letters*, vol. 64, no. 8, Feb. 1990, pp. 821-824.

[5] J. Fridrich, "Symmetric ciphers based on two-dimensional chaotic maps," *International Journal of Bifurcation and Chaos*, vol. 8, no. 6, Jun. 1998, pp. 1259-1284.

[6] L. M. Pecora and T. L. Carroll, "Driving systems with chaotic signals," *Physical Review A*, vol. 44, no. 4, Aug. 1991, pp. 2374-2383.

[7] K. W. Wong, "A fast chaotic cryptographic scheme with dynamic look-up table," *Physics Letters A*, vol. 298, no. 4, Jun. 2002, pp. 238-242.

[8] L. Kocarev and Z. Tasev, "Public-key encryption based on Chebyshev maps," In *Proceedings of the International Symposium on Circuits and Systems (ISCAS '03)*, vol. 3, May 2003, pp. III-28-III-31.

[9] J. C. Mason and D. C. Handscomb, *Chebyshev polynomials*, Chapman & Hall/CRC, Boca Raton, Florida, 2003.

[10] P. Bergamo, P. D'Arco, A. Santis, and L. Kocarev, "Security of public-key cryptosystems based on Chebyshev polynomials," *IEEE Transactions on Circuits and Systems-I*, vol. 52, no. 7, Jul. 2005, pp. 1382-1393.

[11] D. Xiao, X. Liao, and S. Deng, "A novel key agreement protocol based on chaotic maps," *Information Sciences*, vol. 177, no. 4, Feb. 2007, pp. 1136-1142.

[12] S. Han, "Security of a key agreement protocol based on chaotic maps," *Chaos, Solitons & Fractals*, vol. 38, no. 3, Nov. 2008, pp. 764-768.

[13] E. J. Yoon and K. Y. Yoo, "A new key agreement protocol based on chaotic maps," In *Proceedings of The Second KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications (KES-AMSTA '08)*, Mar. 2008, pp. 897-906.

[14] C. A. Petri, "Kommunikation mit Automaten," Ph. D. Thesis, University of Bonn, 1962.

[15] D. Xiao, X. Liao, and S. Deng, "One-way hash function construction based on chaotic map with changeable-parameter," *Chaos, Solitons & Fractals*, vol. 24, no. 1, Apr. 2005, pp. 65-71.

[16] T. Murata, "Petri nets: Properties, analysis and applications," *Proceedings of the IEEE*, vol. 77, no. 4, Apr. 1989, pp. 541-580.

[17] HPSim 1.1 Petri nets simulation tool, copyright© 1999-2002 Henryk Anschuetz.

A Vehicle-density-based Forwarding Scheme for Emergency Message Broadcasts in VANETs

Yu-Tian Tseng, Rong-Hong Jan, Chien Chen
Department of Computer Science
National Chiao Tung University
Hsinchu, Taiwan 30010
{herrowei, rhjan, chienchen}@cs.nctu.edu.tw

Chu-Fu Wang
Department of Computer Science
National Pingtung University of Education
Pingtung, Taiwan 900
cfwang@mail.npue.edu.tw

Hsia-Hsin Li
Industrial Technology Research Institute
Hsinchu, Taiwan 31040
hhli@itri.org.tw

***Abstract*—With the extension of wireless technology, vehicular ad hoc networks (VANETs) provide general data transmission services and emergency warning services. To rescue more drivers from being involved in an emergency event on the road, fast emergency message propagation is an important issue in VANET studies. There are two types of multi-hop broadcasting forwarder selection schemes for emergency broadcasting, known as sender-oriented schemes and receiver-oriented schemes. The sender-oriented schemes periodically maintain neighbor information in order to choose the best forwarder before broadcasting the message, while the receiver-oriented schemes distributed elect the forwarders. In this paper, we propose a vehicle-density-based emergency broadcast (VDEB) scheme to solve the problem of high overhead in sender-oriented schemes, and long delay in receiver-oriented schemes. The simulation results show that our VDEB scheme can achieve better performance with low delay and little overhead.**

***Index Terms*—vehicular ad hoc networks; emergency message broadcast; inter-vehicle communications.**

I. Introduction

The applications of vehicular ad hoc networks (VANETs) can be classified into two main categories: general data routing services and safety applications. The general data routing provides one-to-one data routing or one-to-all data broadcasting for services, such as entertainment, route planning, and communications. The data transmission requirement of this type of service is reliability; that is to say the packets should be successfully received by the receivers. The safety applications provide one-to-all emergency message broadcasting for receivers in a predefined region, such as, electronic brake lights, lane changing assistance, and road condition reports. These applications are usually life critical. Therefore, the data should not only be successfully received by the receivers, but be received in a very short time to provide the driver with more reaction time. For the most urgent situations, e.g. vehicle collisions, the limit of propagation time of the emergency message is extremely low. Some research focuses on Cooperative Collision Avoidance (CCA) [1][2] to broadcast collision avoidance messages in a very short latency in order to save as many victim vehicles as possible.

This work was supported in part by the National Science Council, Taiwan, Republic of China, under grant NSC 97-2221-E-009-048-MY3 and NSC 97-2221-E-009-049-MY3.

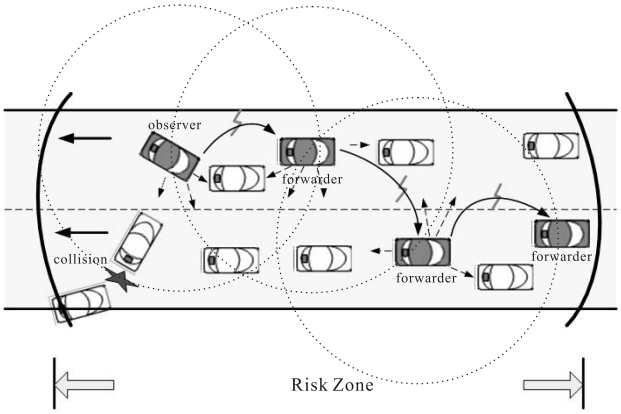


Fig. 1. An example illustrating the emergency message forwarding scheme for safety application in VANET.

In this paper, we focus on the broadcasting of emergency messages for the safety applications. When an emergency event occurs, an emergency message (or messages) will be generated by an observer vehicle and be broadcast. Since the message might not be relevant to all areas, the emergency message is only broadcast within a limited region, called the risk zone (usually several kilometers). Generally the one-hop broadcast range is just several hundreds of meters and cannot cover the entire risk zone; thus multi-hop broadcasting is required during the communication. Figure 1 shows an emergency message forwarding scheme. As the observer detects an automobile accident occurring, it immediately broadcasts an emergency message to inform all its neighbors; meanwhile one of the observer's neighbors (the forwarder) will be selected to play the role of message relaying. Similarly, the new chosen forwarder will broadcast the message and also select the next forwarder for message relaying. The above tasks continue until the message reaches the edge of the risk zone. Since the involved vehicles of message broadcasting are only the forwarders, the broadcast storm problem can then be relieved. However, without appropriate forwarder node selection, there will be many nodes rebroadcasting the message. A large number of receiving nodes broadcasting the same message will create the problems of channel contention and serious packet collisions. Rebroadcasting due to transmission failure

978-1-4244-7489-9/10/$26.00 ©2010 IEEE

makes the problem even worse. In order to ensure all the vehicles in the risk zone can receive the emergency message in time, we consider two important ways to reduce the message propagation latency. The key point is to reduce the hop counts needed to propagate the message to the entire risk zone. This can be done by choosing the forwarders such that the hop distance is maximized. Moreover, the additional waiting time before rebroadcasting the message should also be as small as possible to ensure low latency. In this paper, a vehicle-density-based emergency broadcast scheme called the VDEB is proposed to reduce the message propagation latency with little overhead. The rest of this paper is organized as follows. In section 2, we will review the related works about emergency message forwarding schemes. Then, our proposed VDEB protocol will be introduced in section 3. Section 4 will show the simulation results. The concluding remarks are given in the last section.

II. Related works for emergency message forwarding schemes

Several research papers have proposed methods for emergency message forwarding [3-12]. These methods can be categorized into two types: sender-oriented schemes and receiver-oriented schemes. In the following, we will briefly describe the related works for each type of forwarding scheme, respectively. The discussion related to the drawbacks of each is given at the end of this section.

A. The sender-oriented schemes

Sender-oriented schemes [3-8] use accurate neighbor positions to select the best forwarder node (usually the farthest node from the sender) claiming minimum hop count and no additional waiting time. This approach uses high frequency beacon messages or handshaking mechanisms to choose a single node as the forwarder. Among the papers [3-8] concerning the sender-oriented scheme, [3-5] are backbone-based methods. They try to create and maintain stable clusters (or virtual backbones) to reduce the overhead of the backbone structure maintenance. The emergency messages are rebroadcast by the farthest vehicle of each cluster head. The papers [6-8] are MCDS-like (minimum connected dominating set like) methods. They use periodic beacon messages to acquire the vehicle ID of the farthest vehicle in the transmission range of a vehicle, and try to maintain an MCDS structure as the backbone. The vehicle nodes located in the backbone will be responsible for message broadcasting. Due to the requirement of frequent beacon message exchanges, this type of method encounters the problem of high overhead.

B. The receiver-oriented schemes

The receiver-oriented schemes [9-12] use contention to automatically elect the forwarder(s) in a distributed fashion. Contrasting with the sender-oriented schemes, the sender does not assign the forwarder in the broadcast message. All the one hop receivers concerning the emergency event enter the contention phase after receiving the message in the receiver-oriented schemes. They calculate a waiting time, and wait this time period before rebroadcasting the message. The node with the shortest waiting time may rebroadcast the message first. All the other nodes overhearing this broadcast will cancel their broadcasting process. The parameters for waiting time calculation can be the distance between the sender and receiver, vehicle velocity, vehicle's moving direction, number of nodes, and so on.

Note that the vehicle with the larger waiting time will not necessarily rebroadcast the message first. Statistically, the furthest node from the sender will have the highest chance of rebroadcasting the emergency message first. However, in a sparse network, the waiting time still has a high chance of being long. Moreover, several vehicles may have a similar contention window size. As the vehicle density increases, the situation of multiple rebroadcasting also could occur. This might cause network congestion and collisions.

Both types of the existing message forwarding schemes have their drawbacks. The sender-oriented schemes mainly rely on accurate position data to select the forwarder. To maintain accurate data, the broadcasting period of the beacon messages should be very short, and the overhead will therefore become large. In addition, inaccurate data may cause the sender to select a forwarder outside the transmission range and thus the forwarding process will fail. On the other hand, the latency of the receiver-oriented schemes usually gets much longer in a sparse network. The waiting time is usually reversely proportional to the distance between the sender and the receiver. The forwarders in a sparse network have a higher chance of being far from the border of the sender's transmission range because the inter-vehicle spaces are larger than in a dense network.

III. The vehicle-density-based emergency broadcasting scheme (VDEB)

The considered scenario of our method is in a highway environment. We assume all of the vehicles on the road are equipped with a positioning device such as GPS to acquire their own positions. In addition, to detect an emergency event, the vehicles should also be equipped with several sensors for vehicle velocity, etc. For the transmission of packets, each vehicle is equipped with the WAVE DSRC device. The VDEB is a receiver-oriented forwarding method; however, it synthesizes the main ideas of the sender-oriented scheme and tries to reduce the influence of the drawbacks of both methods. The basic idea is that, in order to overcome the problem of a long waiting time in a sparse network for the receiver-oriented scheme, the VDEB adopts a ring-based approach to shorten the waiting time. The aim of the ring-based approach is to partition the transmission range of the current forwarder into multiple concentric rings. The gap between two adjacent rings is called the ring width. The computation of the ring width is performed at the forwarder. As the ring width is determined, it will be broadcast to notify the neighbors of the forwarder. The receiver can then compute the waiting time according to the

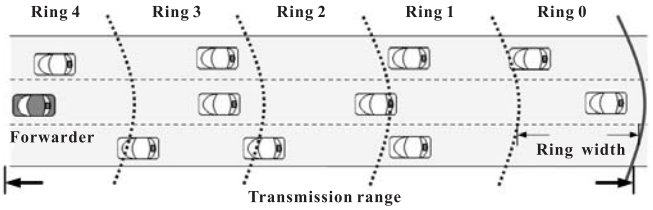


Fig. 2. The ring-based approach of the VDEB.

Vehicle ID	Position	Velocity	Timestamp

Fig. 3. The hello message format.

Source ID	Event position	Event description	Event emergency level	RZ length	Current forwarder ID	Current forwarder position	Ring width

Fig. 4. The emergency message format.

received ring width. The method of ring width computation and the waiting time computation will be discussed later.

For the example in Figure 2, the transmission range is partitioned into 5 rings. Each ring is assigned a timeslot. The length of the waiting time is increased for each ring number. The vehicles in the outermost ring, say ring 0, have the shortest waiting time. If no successful transmissions are done by the vehicles in ring 0, the vehicles in ring 1 will forward the message. If any vehicle successfully broadcasts the message, other vehicles will cancel their waiting process since the forwarding is done without collisions. Usually, if there are vehicles in ring i, the vehicles in ring number greater than and equal to $i+1$ will not rebroadcast the message. Note that, if two (or more) vehicles in the same ring come out, then their messages would collide. In this case, they can use an 802.11 like backoff mechanism to resolve their collisions.

In order to obtain the ring width, the forwarder has to maintain a neighbor table. The main objective of maintaining the neighbor information is to determine the vehicle density in the transmission range of the current forwarder, not to keep track of each vehicle's correct position. Besides, we adopt a vehicle position prediction mechanism in the VDEB forwarding scheme. Thus the frequency of the beacon message exchange for maintaining the neighbor table is less compared with in the traditional sender-oriented scheme. In the following, the details of the VDEB scheme are introduced. First, the neighbor table maintenance method is described. Then the emergency message format and the ring width determination method are given. The complete operations of the sender (the forwarder) and the receiver of the VDEB are given in the last subsection.

A. Maintenance of the neighbor table

In our approach, the purpose of the beacon message is to inform other neighboring vehicles to roughly estimate the future position of vehicles in their neighbor table. Then it can be used to count the number of neighboring vehicles around one vehicle (called vehicle density in this paper). The beacon messages are called hello messages in this paper based on their functionality. The format of the hello message is given in Figure 3. The first and the second fields of the hello message denote the vehicle's identity and its position (x_0, y_0), respectively. The third field is the vehicle's velocity (v_x, v_y) when the message is generated. The final field is the time stamp t_0. With the position information, velocity, and the time stamp, the receivers can roughly estimate the current position (x, y) of the sender using the following equation.

$$(x, y) = (x_0, y_0) + (t - t_0) \times (v_x, v_y) \quad (1)$$

where t is the current time. Note that, the hello messages in sender-oriented schemes usually have position information and the timestamp but not velocity. In this type of scheme, vehicles only care about the presence of neighboring vehicles. The accuracy of the position information is reached by frequently transmitting the hello message. The validity of the neighbor information is verified by checking it in the neighbor table to see if its lifetime has expired or not. In other words, if the same vehicle information is received, the position information and the timestamp will be updated; otherwise the entry for this vehicle will just be removed. By employing the velocity of the neighboring vehicle to estimate its position after a period of time, every vehicle can roughly estimate the number of neighboring vehicles in its transmission range. If the current position of a vehicle is not found in its transmission range, then it will not be a neighboring vehicle.

B. The emergency message format and the ring formation

Generally, an emergency message has the information about the event, such as the position of the event, the event description, the event emergency level, and the risk zone (RZ) size of the event. To reflect the characteristics of an emergency event, we consider that the emergency message should be rebroadcast until it reaches the edge of the risk zone. Besides the above information, we also add the IDs of the source and the current forwarder, the position of the current forwarder, and the ring width to the emergency message. The detailed emergency message format is shown in Figure 4.

Note that the ring width is included in the message for receivers to calculate their waiting time. The position of the current forwarder for receivers is used to determine in which ring they are located. We assume that the vehicles have identical inter-vehicle distance with its immediate front and back vehicles. Under this assumption, we can easily determine the maximum and minimum ring width as follows. The maximum ring width appears in the case that all the vehicles drive side by side (see Figure 5(a)). Hence the maximum ring width can be calculated by:

$$MaxRingWidth = \frac{R \times l}{N}, \quad (2)$$

where R is the transmission range radius, l is the number of lanes, and N is the number of vehicles in the transmission range. Consider the same case of inter-vehicle distance; the

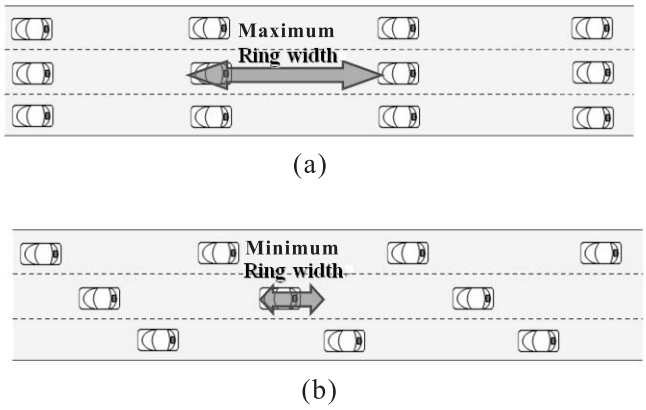


Fig. 5. The possible cases of the maximum and minimum ring widths.

minimum ring width appears when the vehicle distribution is as shown in Figure 5(b). The vehicles in the second and third lanes partition the inter-vehicle distance of the first lane into three equal parts. The minimum ring width can be calculated by:

$$MinRingWidth = \frac{R}{N}, \tag{3}$$

The ring width can be randomly set to be the value in $[MinRingWidth, MaxRingWidth]$ and then the emergency message can be broadcast to the neighbors of the current forwarder.

C. The emergency message broadcasting scheme

The VDEB uses hello messages to periodically exchange the basic information between any two vehicles. Based on the information in the received hello message, each vehicle can maintain its own neighbor table. For the current forwarder, the vehicle density can be easily obtained from its neighbor table and then the ring width can be determined. As the ring width is estimated, it will be added to the emergency message and then be broadcast to the one-hop neighbors. Upon receiving an emergency message from a vehicle, it firstly calculates its ring number by the ring width and the position of the forwarder given in the message. The waiting time of ring 0 is 0 SIFS time. The waiting time of the other rings is proportional to their ring number as follows:

$$WaitingTime = SlotTime \times RingNumber \tag{4}$$

When a vehicle waiting in the waiting process overhears any vehicle forwarding an emergency message, the waiting process is canceled. As the next forwarder is elected, the message relaying will be continued until the border of the risk zone is reached. The complete operations of a forwarder and a receiver of the VDEB are given in Figures 6 and 7, respectively.

Procedure forwarder-emergency-message-forwarding;
Output: emergency message (EM);
{
  if (the current vehicle is the observer) {
    Add event-related information into the EM; }
  Compute the vehicle density N;
  /* Determine the ring width value */
    $MaxRingWidth = \frac{R \times l}{N}$,
    $MinRingWidth = \frac{R}{N}$, where R is the transmission range,
    l is the number of lanes;
    RingWidth= randomly depicts any value in
    $[MinRingWidth, MaxRingWidth]$;
  Add RingWidth into the EM;
  Broadcast(EM);
}

Fig. 6. The complete procedure for forwarder-emergency-message-forwarding.

Procedure receiver-emergency-message-forwarding;
Input: emergency message (EM);
{
  Retrieve RingWidth value from the EM;
  RingNumber=Ring-Number-Determination(RingWidth, current forwarder's position);
  WaitingTime=SlotTime × RingNumber;
  Perform the Waiting process;
    if ((the waiting timer expired) and (does not overhear EM from neighbor)) {
      Set itself to be the forwarder;
      call procedure forwarder-emergency-message-forwarding;
    }
    else cancel the waiting process;
}

Fig. 7. The complete procedure for receiver-emergency-message-forwarding.

IV. Simulation results

The purpose of our simulations was to compare the performance of our proposed VDEB with the other conventional emergency message forwarding schemes. The compared methods are the DBS and the MCDS schemes, which belong to the receiver-oriented scheme and the sender-oriented scheme, respectively. There are two parts to our simulations. Since a better choice of a hello interval (HI) will greatly affect the performance of sender-oriented schemes, thus at first we conducted simulations to show the tradeoff between the efficiency and the overhead of the MCDS and the VDEB, and then to determine the best value of the hello intervals for both methods. The second part of the simulations was to conduct performance comparisons between our proposed VDEB and the DBS and MCDS schemes. The performance evaluations were rely on NS2. In this section, the simulation environment setup is firstly described. The numerical results are then given.

A. The simulation environment setup

The simulation scenario is an 8km, 3 lane, single direction section of highway. The risk zone of an emergency event is $1km$. We vary the hello interval from $0.2s$ to $3.2s$ for the MCDS scheme. The VDEB is also simulated with the hello interval from $3.2s$ to $25.6s$ The mobility of vehicles is evaluated by two speed scenarios: the low speed scenario has

a speed of [20km/h, 70km/h], while the high speed scenario has a speed of [70km/h, 120km/h]. The vehicle density varies from a sparse to a dense road environment based on the safety distance with respect to the vehicles' speed. Since the performance results for both scenarios (the low speed and the high speed) are quite similar, thus we only demonstrate the results of the high speed scenario. The complete simulation setting is summarized in Table 1.

Table 1: List for the simulation parameter settings.

Simulation scenario	highway
Simulation area	$8km \times 3$ lanes
Risk zone size	$1km$
Transmission range	$250m$
MCDS hello interval	$0.2s$, $0.4s$, $0.8s$, $1.6s$, $3.2s$
VDEB hello interval	$3.2s$, $6.4s$, $12.8s$, $25.6s$
Simulation time	$150s$
Vehicle density	$[20, 130]$ vehicles/km
Vehicle speed	$[20, 70], [70, 120]$ km/h
Congestion window	CW_{min} (32 SIFS)
Slot time	64 SIFS

B. The numerical results

1) The simulations for hello interval setting: Figure 8 shows the average delay for the MCDS scheme as the hello interval (HI) varies. Clearly, the shorter the hello interval we set, the lower average delay we have. However, a shorter hello interval will disseminate many hello messages in the network, which will cause the maintenance to be too costly. As shown in Figure 8, the performance of the MCDS when the hello interval is set to 0.4 is a proper choice, since the average delay is under $4ms$ for any vehicle density. Similarly, Figure 9 gives the average delay of the VDEB scheme as the hello message varies. As shown in Figure 9, the impact of varying the hello interval in the VDEB scheme is so small that it can be neglected. To achieve a similar delay performance to the MCDS scheme, we can use larger hello intervals in our VDEB scheme, which will result in less network overhead consumption. The hello interval of VDEB will therefore be set to 6.4 for the second part of the simulation.

2) The simulation results for performance metrics comparison: Figures 10 and 11 show the average delay and the farthest delay simulation results for the MCDS, VDEB, and DBS. The farthest delay is defined as the average of the delay time in which the emergency message reaches the farthest receiver. From Figures 10 and 11, we learnt that the average delay and the farthest delay of DBS is the worst compared to the MCDS and VDEB schemes. In addition, our proposed VDEB scheme outperforms the other schemes. Figure 12 gives the results of the average retransmission comparison. The DBS also performs poorly in the number of retransmissions. As shown in Figure 12, the number of retransmissions of DBS increases when the vehicle density gets higher. This is because there are more vehicles with a similar waiting time if the vehicle density gets higher. In the MCDS scheme, because the next forwarder is chosen before broadcasting the message, it will not cause multiple forwarders. Our VDEB scheme is a receiver-oriented method, so the multiple-forwarder problem will occur. Regardless of the vehicle density, the number of average retransmissions is not greater than 15 for both the MCDS and VDEB schemes.

The comparison results for the number of total transmitted hello messages are summarized in Figure 13. When the vehicle density is 130 vehicles per kilometer, the number of total transmitted hello messages in MCDS (hello interval is 0.4 seconds) is $390,000$, while in VDEB (hello interval is 6.4 seconds), there are only $24,375$ hello messages transmitted. Thus the VDEB consumes very low overhead to obtain even better performance than MCDS at the expense of a few more retransmissions.

V. Conclusion

In this paper, we propose a vehicle-density-based broadcast scheme, called VDEB, for emergency message forwarding. Because the sender-oriented schemes have the drawback of high overhead, and the receiver-oriented schemes cause longer delays, our VDEB scheme resolves both of these problems and provides lower delay and lower overhead for emergency message broadcasting. In the VDEB scheme, a receiver-oriented contention mechanism is adopted with the vehicle density measurement component. Vehicle density can help to reduce the number of retransmissions of messages, and avoid the situation that the delay time increases if the forwarder is not far enough away. In addition, the number of forwarders is limited by the ring width which is estimated by the vehicle density and neighbor information. In the VDEB scheme, the number of retransmissions is not proportional to the vehicle density. The simulation results show that our VDEB scheme provides a good delay performance with reasonable overhead that does not hurt the general data services in VANETs.

References

[1] K.V.N. Kavitha, A. Bagubali, and L. Shalini, V2V wireless communication protocol for rear-end collision avoidance on highways with stringent propagation delay, In Proc. of the ARTCom 2009, pp.661-663.

[2] S. Biswas, R. Tatchikou, and F. Dion, Vehicle-to-vehicle wireless communication protocols for enhancing highway traffic safety, IEEE Communications Magazine, Vol. 44, No. 1, pp.74-82, 2006.

[3] M. Durresi, A. Durresi, and L. Barolli, Emergency broadcast protocol for inter-vehicle communications, In Proc. of the ICPADS 2005, Vol. 2, pp.402-406.

[4] L. Bononi and M.D. Felice, A cross layered MAC and clustering scheme for efficient broadcast in VANETs, In Proc. of the IEEE MASS 2007, pp.1-8.

[5] T. Taleb, K. Ooi, and K. Hashimoto, An efficient collision avoidance strategy for ITS systems, In Proc. of the IEEE WCNC 2008, pp.2212-2217.

[6] L.O. Djedid, N. Lagraa, M. Yagoubi, and K. Tahari, Adaptation of the MCDS broadcasting protocol to VANET safety applications, In Proc. of the IIT 2008, pp.534-538.

[7] Z. Shi, F. Liu, and S. Xu, Novel relay scheme based on traffic type in vehicular networks, In Proc. of the ICICTA 2008, Vol. 2, pp.392-397.

[8] P. Lai, X. Wang, N. Lu, and F. Liu, A reliable broadcast routing scheme based on mobility prediction for VANET, In Proc. of the IEEE Intelligent Vehicles Symposium, pp.1083-1087, 2009.

[9] C. Chiasserini, R. Gaeta, M. Garetto, M. Gribaudo, and M. Sereno, Efficient broadcasting of safety messages in multihop vehicular networks, In Proc. of the 20th IPDPS 2006, pp.8.

[10] S. Yu and G. Cho, A selective flooding method for propagating emergency messages in vehicle safety communications, In Proc. of the ICHIT 2006, Vol. 2, pp.556-561.

[11] Y.T. Yang and L.D. Chou, Position-based adaptive broadcast for inter-vehicle communications, In Proc. of the IEEE ICC 2008, pp.410-414.

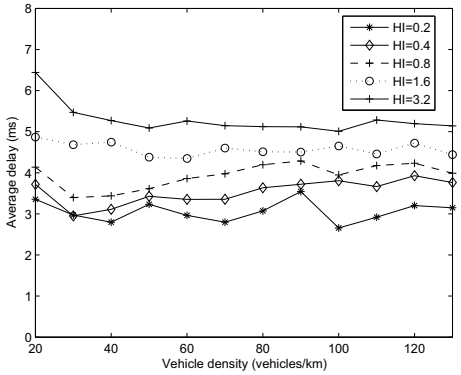


Fig. 8. The average delay comparisons of the different hello intervals of MCDS.

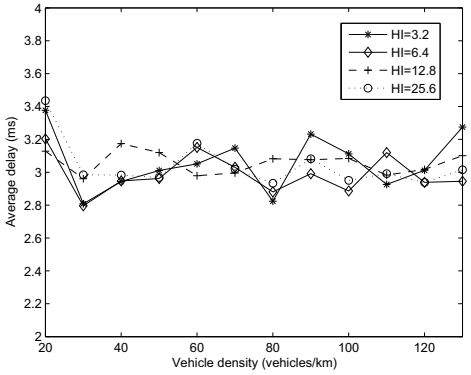


Fig. 9. The average delay comparisons of the different hello intervals of VDEB.

[12] H. Alshaer and E. Horlait, An optimized adaptive broadcast scheme for inter-vehicle communication, In Proc. of the IEEE VTC 2005, Vol. 5, pp.2840-2844.

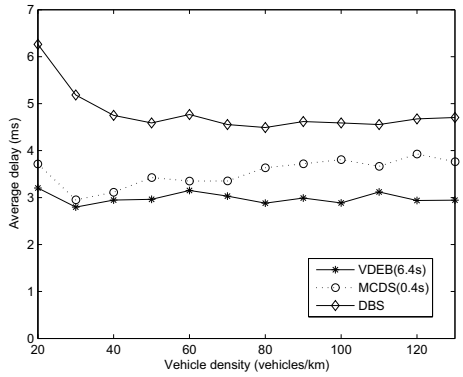


Fig. 10. Performance comparisons of the average delay.

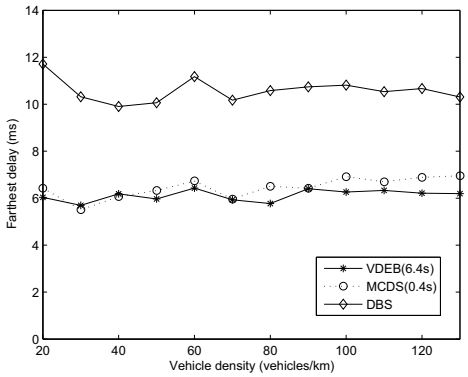


Fig. 11. Performance comparisons of the farthest delay.

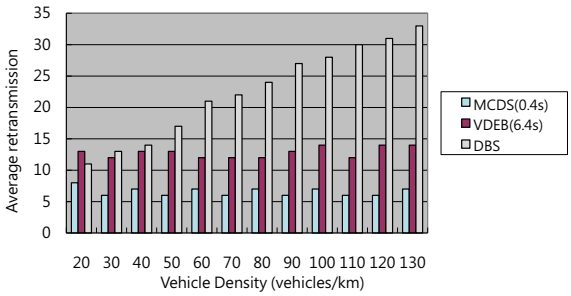


Fig. 12. Performance comparisons of the average retransmissions.

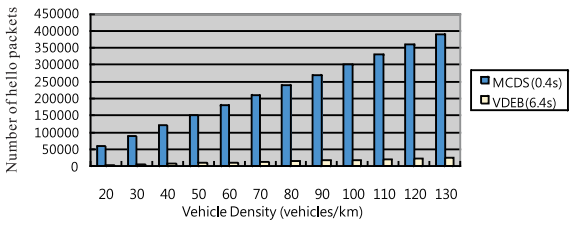


Fig. 13. Performance comparisons of the overhead.

Mobile-Gateway Routing for Vehicular Networks[1]

Hsin-Ya Pan, Rong-Hong Jan[2], Andy An-Kai Jeng, and Chien Chen
Department of Computer Science
National Chiao Tung University
Hsinchu, 30010, Taiwan
{hypan, rhjan, andyjeng, chienchen}@cs.nctu.edu.tw

Huei-Ru Tseng
Information and Communications Research Laboratories
Industrial Technology Research Institute
Hsinchu, 31040, Taiwan
hueiru@itri.org.tw

Abstract—**Development of vehicular ad hoc networks (VANETs) has drawn intensive attention in recent years. Designing routing protocols for vehicle-to-vehicle (V2V) communication in VANET may suffer from frequent link change and disconnection. Vehicle-to-infrastructure (V2I) communication can overcome the challenge by relaying packets through the backbone network, but is limited to those areas where a RSU exists. In this paper, we present a position-based routing protocol, named mobile-gateway routing protocol (MGRP), for VANETs. The MGRP combines V2V and V2I communications, and utilizes certain vehicles as mobile gateways. Each mobile gateway connects with a base station through a 3G interface and communicates with other vehicles without the 3G interface through an IEEE 802.11 interface. Upon receiving packets from a vehicle, the mobile gateway forwards the packets to a gateway controller via the base station. The gateway controller then searches the position of the destination vehicle and determines a set of gateway vehicles close by the destination to forward the packets. Simulation results show that the MGRP can significantly improve the packet delivery ratios and reduce the transmission hop count.**

Keywords- vehicular ad hoc network; vehicle-to-vehicle; vehicle-to-infrastructure; position-based routing protocol;

I. INTRODUCTION

Vehicular Ad Hoc Networks (VANETs) have received increasing attention from the research and industrial communities recently. Many valuable applications, such as entertainments, trip planning, and accident avoidance, have been envisioned in VANETs.

Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) are two major types of communications in VANET. Each vehicle equipped with an On-Board Unit (OBU) can either transmit hop-by-hop to the destination using V2V communication or transmit to a Roadside Unit (RSU) using V2I communication.

Different from traditional wireless networks, designing a routing protocol for V2V communication is more challenging. The network topology may change rapidly due to the high speed characteristic of vehicles. Thus, a proactive routing, such as DSDV [1] that pre-establishes shortest paths between nodes, is not appropriate. In other words, a packet could be transmitted in a longer hop in V2V communication. Moreover, the disconnection problem may happen at the areas of low traffic density, further degrading the packet delivery ratio.

The challenges in V2V can be overcome by the support of V2I communication. A vehicle can firstly transmit packets to a RSU. The RSU connects to the backbone network and thus can forward packets for vehicles using a more efficient and reliable way. However, due to the limited transmission ranges of OBUs, the support of V2I communication is only restricted to those areas where a RSU is reachable. In other places, the above challenges still exist.

In this paper, we propose a position-based routing protocol, named *mobile-gateway routing protocol* (MGRP) for VANETs. The MGRP combines V2V and V2I communication, and utilizes certain vehicles as mobile gateways to extend the coverage of fixed RSUs. The OBU on each mobile gateway is equipped with an IEEE 802.11 interface and a 3G interface. The other vehicles without the 3G interface can forward packets through 802.11 links to the nearest mobile gateway. Upon receiving a packet, the mobile gateway forwards the packet to a gateway controller through the 3G interface. The gateway controller then searches the position of the destination vehicle and determines a set of gateway vehicles close by the destination to forward the packet.

Simulation results show that the MGRP can significantly improve the packet delivery ratios and reduce the transmission hop counts. In other words, the proposed protocol can provide vehicles with more instantaneous services. We also investigated the percentage of gateway vehicles to find the appropriate ratio to guarantee the successful delivery ratio.

The rest of the paper is organized as follows. In Section II, we review the previous studies and related works. Then, the system model, assumptions and detailed description of MGRP are presented in Section III. In Section IV, we evaluate the proposed approach by simulation and compare with the GPSR routing protocol. Finally, a conclusion is given in Section V.

II. RELATED WORK

Research on V2I communication can be divided into two categories: One is that the RSU just plays the role of packet storage but does not provide the function of transmission, such as SADV [2]. The other focuses on utilizing the pre-established RSU for transmission. When a vehicle enters the transmission range of a RSU, it will start to send or receive the data packet to the RSU.

[1] This paper was supported in part by the National Science Council of the ROC, under Grant NSC-97-2221-E-009-049-MY3
[2] Corresponding Author; Fax: 886-3-5721490

A number of routing protocols that use fixed infrastructures to improve the packet delivery have been proposed. In MPARP [3], each vehicle is equipped with an IEEE 802.11 and an IEEE 802.16 interfaces. When routes exist, vehicles can communicate directly with each other using the IEEE 802.11 mode; otherwise, their communication will be taken over by a base station using the IEEE 802.16 mode. In RAR [4] and DDR [5], each section of the road is embraced by two RSUs. When a vehicle has some packets for another vehicle on a different section, it transmits to one of the RSU on which it is located. Then, the packets will be forwarded to the destination through the backbone network. Similarly, the routing protocol in [6] tries to make a proper decision on whether to broadcast or to use end-to-end transmission based on the information provided by RSUs.

Although vehicular communication can be assisted by the support of V2I model, the advantage is restricted to those areas where fixed infrastructures exist. To overcome it, the MIBR protocol in [7] introduces the concept of mobile gateways. It employs each bus as a mobile gateway to forward packets for vehicles. Because buses have fixed travel routes and can be equipped with radios of larger transmission ranges (over 300m), it is beneficial to improve the delivery radio and throughput. However, the connectivity between buses is still limited to the period of bus schedules and the covered region of bus routes.

To conquer the above limitations, this paper uses certain vehicles, e.g. taxi, in which 3G infrastructures are added to their OBUs, as mobile gateways. Other vehicles without the 3G infrastructures can deliver packets to destinations through those gateway vehicles. Because the coverage of 3G infrastructure is large enough to cover the whole area, the proposed protocol can significantly reduce the hop count and improve the packet delivery ratio.

III. MOBILE GATEWAY ROUTING PROTOCOL

In this section, we first introduce the architecture and assumptions of the mobile gateway routing. Then, the MGRP routing protocol is presented in details.

A. The Architecture of Mobile Gateway Routing

As shown in Fig. 1, we utilize certain vehicles as mobile gateways to substitute traditional RSUs. The OBU on each gateway vehicle is equipped with an IEEE 802.11 interface and a 3G interface. The 3G interface is used to communicate with a base station in a cellular network and the 802.11 interface is used to communicate with other vehicles without the 3G interface. When the base station received data packets from a gateway vehicle, it will deliver the packets to a gateway controller. The gateway controller then searches the position of the destination, determines a set of gateway vehicles close by the destination, and sends packets to each of the chosen gateway vehicles via the based station. Finally, those gateway vehicles will transmit the data packets to the destination with IEEE 802.11 links.

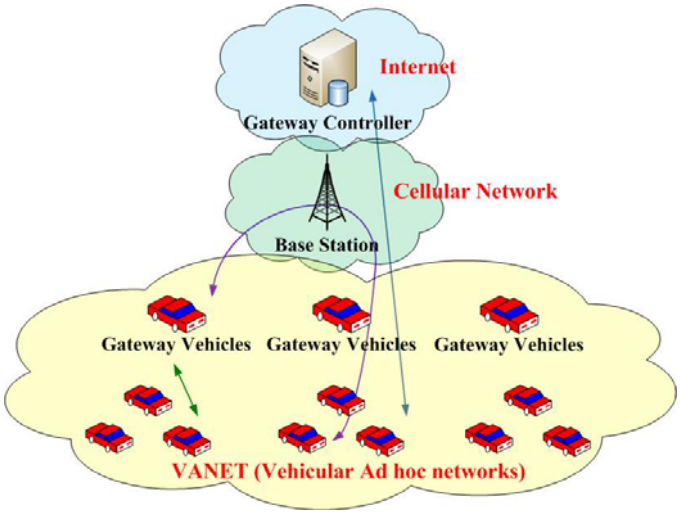


Figure 1: Architecture of mobile gateway routing.

B. Assumptions

We assume that each vehicle can obtain its position, velocity, and direction through a global positioning system (GPS) equipped on the vehicle. This information will be periodically broadcasted to nearby vehicles within the transmission range using a hello message. We also assume that a digital map with traffic load condition of roads is installed in each vehicle. Besides, if a vehicle founds that the route has been broken, it will buffer any received data packet and send a RRER packet to the source vehicle for selecting an alternative route. Different from ordinary routing protocols, such as AODV [8], the MGRP limits the time-to life (TTL) value to three hops.

C. MGRP Routing Protocol

In MGRP, each vehicle can deliver data packets via a mobile gateway to decrease the transmission hop count and to achieve more reliable communication quality. Fig. 2 shows how a source vehicle (left hand side) sends the packet to a destination vehicle (right hand side). When the source vehicle has some packets for the destination vehicle, it first searches a mobile gateway closest to itself, i.e. Gateway1, and sends packets to the gateway vehicle. Then, Gateway1 forwards the data packets to a base station using the 3G interface. Upon receiving the packets, the base station delivers the packets to the back-end gateway controller in order to search the position of the destination vehicle and transmits the packets to a set of gateway vehicles close by the destination vehicle, i.e. Gateway4. Finally, Gateway 4 will forward the packets to the destination vehicle via the 802.11 interface. Without the assist from Gateway1, Gateway4, and the gateway controller, the source vehicle has to carry the packets for a while until it meets vehicle1, vehicle 3, or vehicle9. The same problem happens on the next vehicle carrying those packets. The above relaying process may cause a longer delay time before reaching the destination vehicle. Even worse, if the packet is forwarded to vehicle9 and there is no further vehicle connecting vehicle9 to the destination, the packet will lose, causing an unreliable transmission.

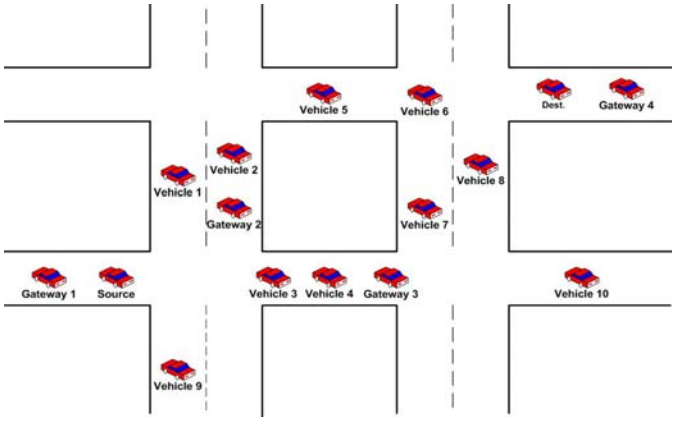


Figure 2: Scenario of the source vehicle sending packets to a destination vehicle

Now we descript a detail processes. Similar to the AODV protocol, when a vehicle needs to send packets, it first broadcasts a RREQ packet to the neighboring vehicles. Once a neighbor received the RREQ packet, if it has no routing path to the destination, it will rebroadcast the route request to other neighbors. Different from the AODV, the TTL of RREQ is limited to three hops in the MGRP. Once a vehicle receives the RREQ, it first checks whether the hop count is still less than three. If so, the vehicle will become the next forwarder to rebroadcast the RREQ packet; otherwise, the vehicle will drop the RREQ packets. If the information of the destination vehicle was in its routing table or the gateway vehicle receives the RREQ packet, it will send back a RREP packet to the source vehicle. Furthermore, if the vehicle waits for a while and does not receive the RREP packet, it will rebroadcast the RREQ packet and repeat above steps.

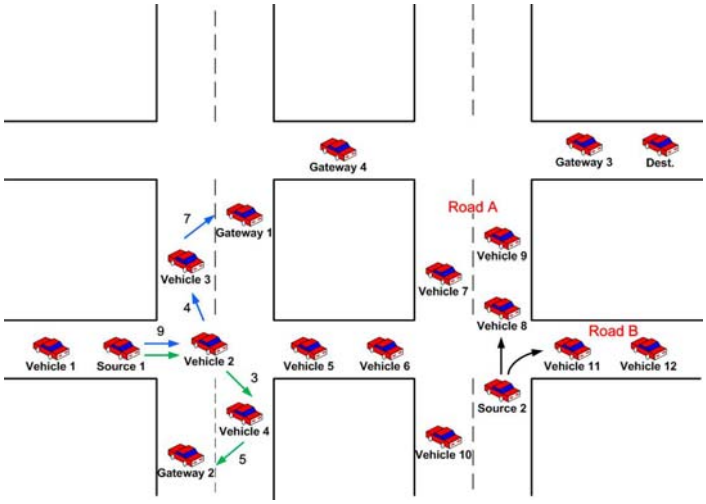


Figure 3: Packet forwarding in different scenarios

After a source vehicle broadcasts the RREQ packet to require a routing path, there are three situations may happened.

The first is that there is no other vehicle beside the source vehicle or cannot find a neighboring vehicle instantly. In this case, the vehicle carries the packets until a vehicle appears within its transmission range. Then, it will forward the packets to the vehicle.

The second situation is that there are more than one neighboring vehicles, but none of them having a path to a mobile gateway or destination vehicle in three hops. In this case, the source vehicle determines a forwarding direction for the packets according to the road density information. When the vehicle located at an intersection, it will select the direction which has the highest road density. As shown in Fig. 3, there is no routing path that can forward data packet from source 2 to a gateway vehicle or destination vehicle, and the density of road A (3 vehicles) is higher than that of road B (2 vehicles). So, source 2 forwards the packets to vehicle 8 which is on road A. This method can improve the packet delivery ratio, because a higher road density usually implies a higher probability of finding a mobile gateway if the percentage of gateway vehicles to ordination vehicles on each road has no significant difference.

The third situation is that there are more than one routing paths that can forward to destinations or gateway vehicles. In this case, the source vehicle needs to select a suitable route path. The MGRP will select the most reliable path to forward the data packets. The reliability of routing path is evaluated by the lifetime of the route. We utilize the following link lifetime formula proposed in [3] to predict the inter-vehicle lifetime,

$$Link\_lifetime = \frac{R - D_{ij}}{V_i - V_j},$$

where R is the transmission range of each vehicle, D_{ij} is the distance between vehicle i and vehicle j, V_i : the velocity of vehicle i, and V_j is the velocity of vehicle j. The lifetime of a routing path is the smallest link lifetime on this path.

As shown in Fig. 3, there are two routes can forward data packets from Source1 to a gateway vehicle. The first path is through Source1→Vehicle2→Vehicle3→Gateway1, and the second path is Source1→Vehicle2→Vehicle4→Gateway2. The lifetimes of links on the first route path are 9s, 4s and 7s, and lifetimes of links on the second route are 9s, 3s and 5s. So, the route lifetimes of the first and second paths are 4s and 3s respectively. As a result, Source 1 will select the first route to forward data packets, which has a longer route lifetime.

Note that if the routing table has recorded the routing path to the destination vehicle, it has the higher priority to select this routing path for transmission.

After the data packets are forwarded to a gateway vehicle, the gateway vehicle will forward these packets to the base station via 3G network and the gateway controller. The gateway controller will choose gateway vehicles nearby the destination vehicle as forwarders. The gateway controller periodically updates the gateway vehicles' position. The forwarding decision of the gateway controller server depends on whether the distance between gateway vehicles and the destination vehicle is less than 500 meters. If there are many gateway vehicles' distance less than 500 meters, all of those gateway vehicles will be selected as the forwarders. And the packets will be delivered to the destination vehicle via V2V. It could enhance the probability of successfully forwarding data packets to the destination vehicle. However, if there is no gateway vehicle forwarding data packets to the destination vehicle, the gateway controller will drop the data packets. As

shown in Fig. 4, three gateway vehicles will receive the data packets from the base station.

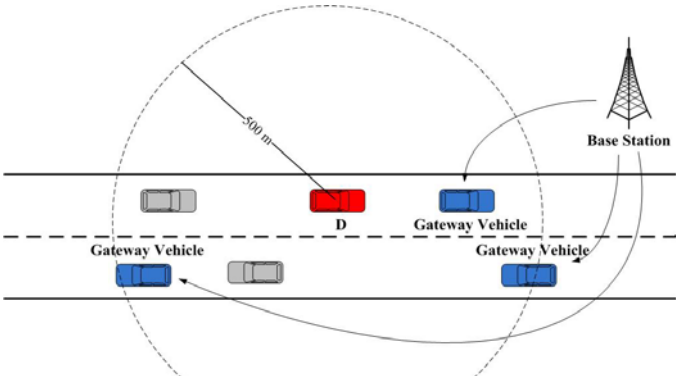


Figure 4: Gateway controller forwards packets to a set of gateway vehicles which are less than 500m to the destination

IV. Simulation and Performance Evaluation

In this section, we evaluate the performance of MGRP using ns2 simulator [9] (version 2.34). We compare MGRP with the traditional position-based routing protocol GPSR [10] routing protocol, and analyze the relationship between the percentage of gateway vehicles and the success delivery ratio.

We perform the simulation on a real street map, captured from TIGER database (Topologically Integrated Geographic Encoding and Reference System) [11]. We simulated the MGRP within two scenarios, highway and urban. The urban street layout is within a 1100m*1100m area as shown in Fig. 5. There are totally sixty-one roads and 150 vehicles. We offer 10 CBR flows and the packet size is 512-byte. The simulation parameters are summarized in Table 1.

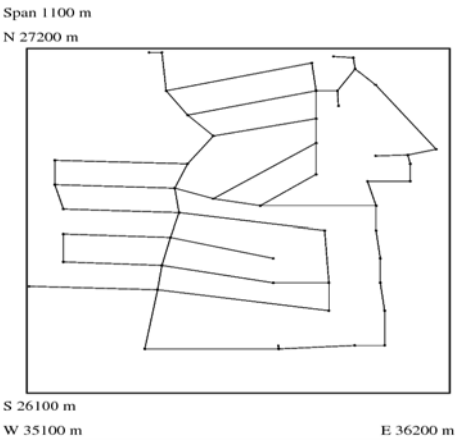


Figure 5: Simulation street layout

TABLE 1. Simulation parameters

Parameter	Value
Simulation scenario	Highway/ Urban
Speed of vehicles	40–90 km/h
Simulation time	300 sec
Interval time of data delivery	0.5 sec
Data packet size	500 bytes
Transmission range	250 m

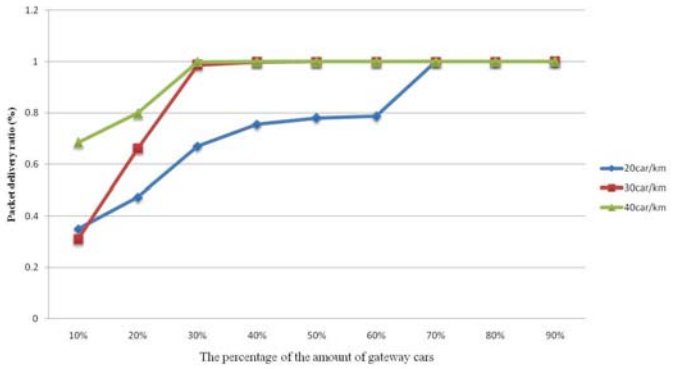


Figure 6: The percentage of gateway vehicles vs. packet delivery ratio

We first evaluate the relationship between the percentage of gateway vehicles and the packet delivery ratio for a scenario of 4km highway with four lanes and double directions. The results are shown in Fig. 6. We can see that when more vehicles play the role as mobile gateways, the packet delivery ratio increases significantly. Even in low density scenario (20 vehicles/km), the MGRP can achieve 80% packet delivery ratio if the percentage of the gateway vehicles is over 60%. When the percentage of gateway vehicles is 10%, the packet delivery ratio is down to 38%. Besides, the MGRP needs at least 70% gateway vehicles to reach the 100% delivery ratio in the low density scenario. On the other hand, the result shows that it can perform better in middle (30 vehicles/km) and high density scenarios (40 vehicles/km). In these cases, the MGRP needs just 30% of gateway vehicles to achieve nearly 100% delivery ratio. That is, if there are ten vehicles in the highway scenario, using our protocol just needs three vehicles to equip the OBU devices. Therefore, it does not need too much cost in this network architecture.

Now, we compare the performance of MGRP and GPSR. Fig. 7 shows the packet delivery ratio versa the maximum speed. We can see that although the packet delivery ratios of both protocols decrease as long as the velocity of vehicles rises, our protocol still perform better, because the MGRP utilizes the 3G network and gateway controller to assist packet forwarding so that link disconnect due to high mobility can be greatly avoided. Notice that the packet delivery ratio of MGRP is lower than GPSR when vehicles' velocity is over 85 km/h. The reason is that MGRP has to frequently maintain the routing table in high velocity scenarios. As a result, it may raise the opportunity of packets lost.

Fig. 8 shows the average hop count to the maximum speed. The results reveal that the average hop count increases when the velocity of vehicles rises regardless of MGRP or GPSR. The MGRP can keep count within 6 while the maximum hop count in GPSR is 10. It is because of the fact that we use 3G network to reduce the transmission hops. In addition, we limit the hop counts when the source vehicle intends to find a route to the gateway vehicle or destination vehicle.

Fig. 9 shows the routing overhead versa the maximum speed. The results show that the packet overhead increases when the velocity of vehicles rises regardless of MGRP or GPSR. The packet overhead in MGRP is more than GPSR because we need to maintain the routing table. However, by

establishing the routing table we can avoid the local maximum problem in GPSR. Furthermore, our method can decrease the total hop count to the destination node.

Fig. 10 shows the routing overhead (without hello message) and the packet delivery ratio versa the hop count between source vehicle to gateway vehicle. We test average velocities of 45km/h, 65km/h and 85km/h. The results show that the packet delivery ratio and overhead increase when hop count rises, because our protocol has more success rate to forward packets to gateway vehicles the hop count is raised. And it also raises the packet overhead.

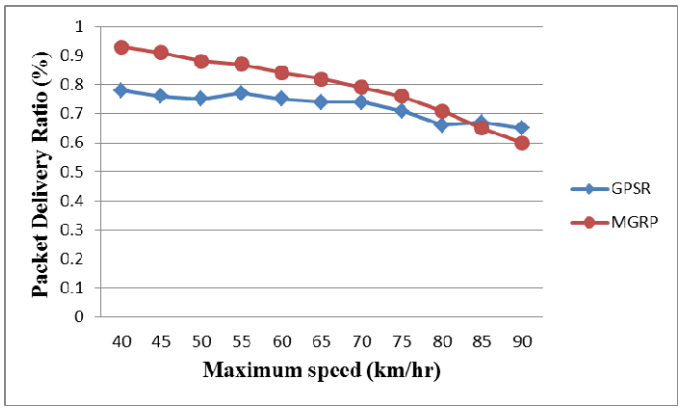


Figure 7: Packet delivery ratio vs. maximum node speed

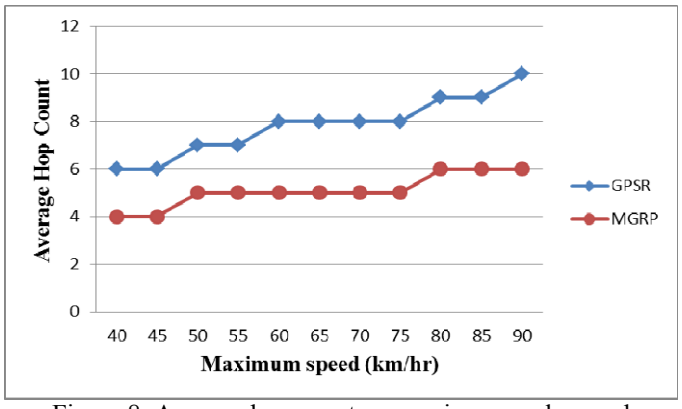


Figure 8: Average hop count vs. maximum node speed

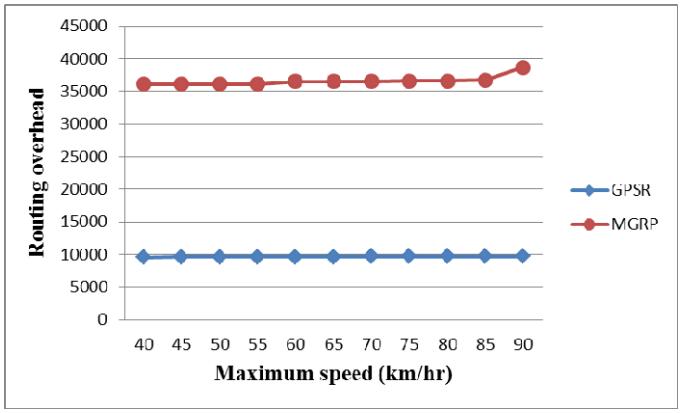


Figure 9: Routing overhead vs. maximum node speed

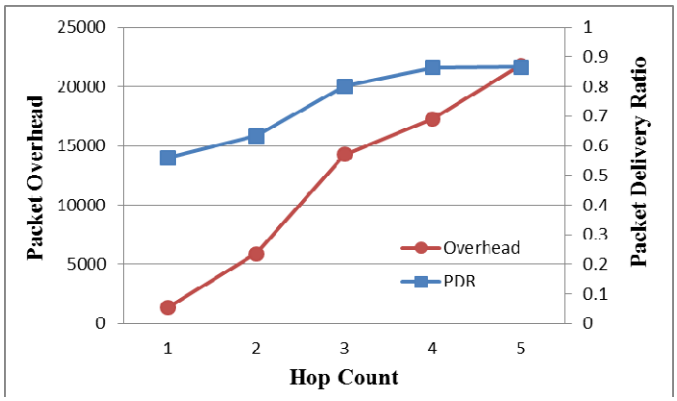


Figure 10: Packet delivery ratio and overhead vs. hop count

V. CONCLUSIONS

In this paper, the position-based routing protocol, called mobile gateways routing protocol, (MGRP) has been proposed for vehicular ad hoc networks. We utilize certain vehicles as mobile gateway vehicles equipped with the OBU which can forward the data packets through interfaces of 3G or IEEE 802.11. Other vehicles without 3G interface can forward the packets through wireless network to mobile gateway vehicles, then using 3G interface to forward packets to the gateway controller. Finally, the gateway controller will forward the packets via mobile gateway vehicles nearby the destination vehicle. We design the routing protocol suitable for this hybrid network architecture and it decreases the total hop counts and the probability of links disconnection obviously. The simulation results show that MGRP performs better than the traditional position-based routing protocol GPSR.

REFERENCES

[1] Charles E.Perkins and Pravin Bhagwat. “Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers,” *SIGCOMM Comput. Commun.*, 1994, Rev., 24(4):234–244.

[2] Y. Ding and L. Xiao, “SADV: Static-node-assisted adaptive data dissemination in vehicular networks,” *Vehicular Technology*, vol. 59, no. 5, Jun. 2010, pp. 2445-2455.

[3] C. C. Hung, H. Chan, and E. H. K Wu, “Mobility pattern aware routing for heterogeneous vehicular networks,” *Wireless Communications and Networking Conference (WCNC)*, Apr. 2008, pp. 2200-2205.

[4] Y. Peng, Z. Abichar, and J. M. Chang, “Roadside-aided routing (RAR) in vehicular networks,” *IEEE International Conference on Communications (ICC)*, Jun. 2006, pp. 3602-3607.

[5] R. He, H. Rutagemwa, and X. Shen, “Differentiated reliable routing in hybrid vehicular ad-hoc networks,” *IEEE International Conference on Communications (ICC)*, May 2008, pp. 2353-2358.

[6] N. Brahmi, L. Boukhatem, N. Boukhatem, M. Boussedjra, N. D. Nuy, H. Labiod, and J. Mouzna, “End-to-end routing through a hybrid ad hoc architecture for V2V and V2I communications,” *Ad Hoc Networking Workshop (Med-Hoc-Net)*, Jun. 2010, pp. 1-8.

[7] J. Luo, X. Gu, T. Zhao, and W. Yan, “A mobile infrastructure based VANET routing protocol in the urban environment,” *IEEE International Conference on Communications and Mobile Computing (CMC)*, Apr. 2010, pp. 432-437.

[8] C. Perkin, E. Belding-Royer, and S. Das, “Ad hoc on-demand distance vector (AODV) routing,” *IETF Experimental RFC*, MANET working group, RFC 3561, Jul. 2003.

[9] Network simulator. ns-2. http://www.isi.edu/nsnam/ns, 2011.

[10] B. Karp and H. T. Kung, “GPSR: Greedy perimeter stateless routing for wireless networks,” *ACM/IEEE International Conference on Mobile Computing*, Aug. 2000, pp. 243-254.

[11] TIGER. http://www.census.gov/geo/www/tiger/, 2011.

A SECURE AGGREGATED MESSAGE AUTHENTICATION SCHEME FOR VEHICULAR AD HOC NETWORKS

Huei-Ru Tseng
Information and Communications Research Laboratories
Industrial Technology Research Institute
Chutung, Hsinchu, 31040, Taiwan
hueiru@itri.org.tw

Rong-Hong Jan, Wuu Yang
Department of Computer Science
National Chiao Tung University
Hsinchu, 30010, Taiwan
{rhjan, wuuyang}@cs.nctu.edu.tw

Emery Jou
Networks and Multimedia Institute
Institute for Information Industry
Taipei, 10574, Taiwan
emeryjou@iii.org.tw

Abstract

Vehicular ad hoc networks (VANETs) are an emerging area of interest for the security community. Due to the scale of the network, the speed of the vehicles, their geographic positions, and the very sporadic connectivity between them, security issues of VANETs are very challenging, especially on how to ensure the authenticity of emergency messages efficiently. In this paper, we propose a secure aggregated message authentication (SAMA) scheme in certificateless public key settings to validate emergency messages for VANETs. We make use of aggregation and batch verification techniques for emergency message verification to reduce the computation overhead. Moreover, the SAMA scheme is modelled and analyzed with Petri nets. Our analysis shows that the SAMA scheme can successfully defend forgery attacks and ensure the conditional privacy preservation and traceability of vehicles.

Keywords: Vehicular ad hoc networks, Authentication, Petri nets, Conditional privacy preservation, Traceability

INTRODUCTION

Vehicular ad hoc networks (VANETs) can be divided into inter-vehicle communications (IVC) and roadside-to-vehicle communications (RVC) that require roadside unit (RSU) equipment. The main goal of VANETs is to achieve safety and comfort for passengers. In VANETs, each vehicle equipped with an on-board unit (OBU) can receive and relay messages through the wireless network without predefined or centralized infrastructure. Vehicle-collision warning, road sign alarms, and in-place traffic view will give the driver essential tools to decide the best path along the way.

Due to the scale of the network, the speed of the vehicles, their geographic positions, and the very sporadic connectivity between them, security issues of VANETs are very

challenging. To tackle the security problems, Raya and Hubaux (1) proposed the first solution in a systematic and quantified way for VANETs in 2005. Thereafter, various security mechanisms (2; 3; 4; 5; 6; 7; 8) have been proposed to improve security, efficiency, and functionality in VANETs.

To ensure the authenticity of emergency messages efficiently is also an important security issue for VANETs. In 2008, Zhu et al. (8) proposed an aggregated emergency message authentication (AEMA) scheme to validate an emergency event. The scheme makes use of aggregation and batch verification techniques to reduce the computation overhead. Zhu et al.'s scheme (8) is based on certificate-based public key cryptography. Therefore, aggregation and batch verification in Zhu et al.'s scheme (8) have two parts, certificates and signatures.

In order to simplify the certificate management as in traditional public key infrastructure (PKI), Shamir (9) proposed identity-based public key cryptography (ID-PKC) in 1984. In ID-PKC, an entity's public key is derived directly from certain aspects of its identity. Private keys are generated for entities by a trusted third party called a private key generator (PKG). Therefore, the entity's private key fully depends on its public known identity and the master secret owned by the PKG. Obviously, ID-PKC suffers from the key escrow problem, i.e., the dishonest PKG can forge the signature of any entity; meanwhile, the entity can deny the signature actually signed by itself.

To overcome the key escrow problem of ID-PKC, Al-Riyami and Paterson (10) proposed a new paradigm called certificateless public key cryptography (CL-PKC). In CL-PKC, a trusted third party called a key generation center (KGC) helps the entity to compute a partial private key from the entity's identity and the KGC's master key. The entity then combines the partial private key with a secret value to generate its actual private key. Thus, the entity's private key is not available to the KGC. The entity's public key is also computed from the KGC's public parameters together with the entity's secret value. The CL-PKC scheme overcomes the key escrow problem in ID-PKC and does not require the use of certificates to guarantee the authenticity of public keys.

In this paper, we propose a secure aggregated message authentication (SAMA) scheme in certificateless public key settings to validate emergency messages in VANETs. The proposed SAMA scheme enhances Zhang and Zhang's scheme (11) to reduce the computation overhead. In the SAMA scheme, the entity makes use of its partial private key generated by the KGC and the private key chosen by itself to generate the signatures on the emergency messages. Due to the characteristics of CL-PKC, the SAMA scheme only needs signature aggregation and batch verification. Compared to Zhu et al.'s scheme (8), the SAMA scheme achieves more efficient authentication on emergency messages.

Privacy preservation is another important security requirement for VANETs, where the source privacy of the emergency message is envisioned to emerge as a critical security issue since privacy-sensitive information, such as the driver's name, position, and driving route, could be jeopardized (4). Therefore, how to preserve the privacy of vehicles is regarded as a fundamental security requirement in VANET communications. However, a malicious driver may abuse the privacy protection by damaging the regular driving environment, such as escaping from the investigation when he involved in a dispute event of emergency

messages. Therefore, the privacy preservation in VANETs should be conditional, i.e., senders are anonymous to receivers while traceable by the KGC, namely conditional privacy preservation (4). With traceability, once a dispute occurs to the emergency message, the KGC can reveal the identities of the vehicles.

Moreover, Petri nets (12) may be used to infer what an attacker could know if he happens to know certain items in the security protocol. We used Petri nets in the security analysis of the proposed scheme. Our analysis shows that the proposed scheme can successfully defend forgery attacks and ensure the conditional privacy preservation and traceability of the vehicles.

The rest of this paper is organized as follows: In Section II, we state the concept of bilinear pairings and introduce the mathematical problems used in this paper. Next, the proposed SAMA scheme is presented in Section III. Then, we shall present the security analysis of our SAMA scheme and provide a performance comparison with other aggregated signature schemes in Section IV. Finally, we will conclude our paper in Section V.

PRELIMINARIES

In this section, we first briefly state the concepts of bilinear pairings and introduce the mathematical problems needed for our proof of security. The notations with their meanings throughout this paper are listed in Table 1.

Bilinear Pairings

Let $\mathbb{G}_1$ be a cyclic additive group of 160-bit prime order q and $\mathbb{G}_2$ be a cyclic multiplicative group of the same order. A map $e : \mathbb{G}_1 \times \mathbb{G}_1 \to \mathbb{G}_2$ is called a bilinear map if it satisfies the following properties:

1. Bilinearity: $e(Q, W + Z) = e(Q, W)e(Q, Z)$ and $e(Q + W, Z) = e(Q, Z)e(W, Z)$, for all $Q, W, Z \in \mathbb{G}_1$.
2. Non-degeneracy: There exists $P, Q \in \mathbb{G}_1$ such that $e(P, Q) \neq 1$.
3. Computable: There is an efficient algorithm to compute $e(P, Q)$ for $P, Q \in \mathbb{G}_1$.

Mathematical Problems

Now we specify the mathematical difficult problems used in this paper as follows.

Definition 1. Discrete Logarithm Problem (DLP). Given a prime p, a generator g of $\mathbb{Z}_p^*$, and an element $\beta \in \mathbb{Z}_p^*$, the DLP is to find the integer α, $0 \leq \alpha \leq p - 2$, such that $g^\alpha \equiv \beta \pmod{p}$.

Definition 2. Elliptic Curve Discrete Logarithm Problem (ECDLP). Given a group $\mathbb{G}_1$ of prime order q, two elements P and Q, the ECDLP is to find an integer $l \in \mathbb{Z}_q^*$, such that $Q = lP$ whenever such an integer exists.

Table 1: Notations.

Symbol	Definition
KGC	A key generation center
$\mathcal{V}_j$	The j-th vehicle
ID_j	A real-identity of the vehicle $\mathcal{V}_j$
PID_j	A pseudo-identity of the vehicle $\mathcal{V}_j$
$\mathbb{G}_1$	A cyclic additive group
$\mathbb{G}_2$	A cyclic multiplicative group
q	The order of the groups $\mathbb{G}_1$ and $\mathbb{G}_2$
e	$e : \mathbb{G}_1 \times \mathbb{G}_1 \to \mathbb{G}_2$
s	A master key of the KGC
P_{pub}	A public key of the KGC
(x_j, D_j)	A private key of the vehicle $\mathcal{V}_j$
PK_j	A public key of the vehicle $\mathcal{V}_j$
$\mathcal{E}_i$	The emergency event i
SER_j^i	The secure emergency report generated by the vehicle $\mathcal{V}_j$ for the emergency event $\mathcal{E}_i$
$Type_i$	The type of the emergency event $\mathcal{E}_i$
Loc_i	The location where the emergency event $\mathcal{E}_i$ takes place
$Time_j^i$	The time when the vehicle $\mathcal{V}_j$ makes the report on the emergency event $\mathcal{E}_i$
Sig_j^i	The signature generated by the vehicle $\mathcal{V}_j$ on the emergency event $\mathcal{E}_i$
$Enc(\cdot)$	A secure symmetric encryption algorithm (13)
$Dec(\cdot)$	A secure symmetric decryption algorithm (13)
$H_1(\cdot)$	A hash function such as $H_1 : \{0,1\}^* \to \mathbb{G}_1$
$H_2(\cdot)$	A hash function such as $H_2 : \{0,1\}^* \to \mathbb{G}_1$
$H_3(\cdot)$	A hash function such as $H_3 : \{0,1\}^* \to \mathbb{Z}_q^*$
$\|$	Message concatenation operation

THE PROPOSED SCHEME

In this section, the SAMA scheme is presented. The scheme is divided into four phases: system setup, registration, signature generation, and aggregated authentication.

System Model

Assume the inter-vehicle communication (IVC) in VANETs without any presence of fixed infrastructure such as access points (APs), road side units (RSU), and satellite communication for assisting in data propagation. The medium used for communication among vehicles is based on 5.9 GHz Dedicated Short Range Communications (DSRC) protocol identified as IEEE 802.11p (14). We assume that there is a KGC which is in charge of generating a vehicle's partial private key. The full private key is finally generated by the vehicle that makes use of the partial private key obtained from the KGC and the secret information chosen by itself. The system model is illustrated in Figure 1.

Security Requirements

The inter-vehicle communication in VANETs is subject to the security requirements: message authentication, conditional privacy preservation, and traceability.

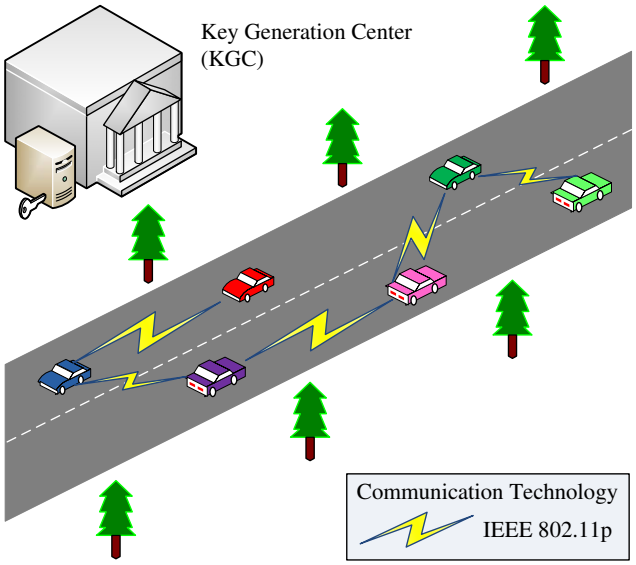


Figure 1: The system model of the SAMA scheme.

- **Message authentication** Similar to wireless sensor networks (15; 16), the major threat that can target specifically VANET aggregation schemes is that of false information dissemination, where attacker's goal is to make the vehicles to accept false emergency reports. Therefore, emergency messages from vehicles have to be authenticated to confirm that they are indeed sent unaltered.

- **Conditional privacy preservation** Privacy preservation is regarded as a fundamental security requirement in VANET communications since overhearing privacy-sensitive information could happen frequently. However, privacy protection may be abused by malicious drivers. Therefore, conditional privacy preservation should be provided in VANETs, i.e., senders are anonymous to receivers while traceable by the KGC, such that the identities can be uniquely revealed by the KGC under exceptional cases.

- **Traceability** The KGC should have the ability to retrieve a vehicle's real-identity from its pseudo-identity once a dispute occurs to the emergency message.

System Setup

Prior to the network deployment, the KGC sets up the system parameters as follows:

1. Let $\mathbb{G}_1$ be a cyclic additive group generated by P with a prime order q, and $\mathbb{G}_2$ be a cyclic multiplicative group of the same order. Let $e : \mathbb{G}_1 \times \mathbb{G}_1 \rightarrow \mathbb{G}_2$ be a bilinear map.

2. The KGC first chooses a random number $s \in \mathbb{Z}_q^*$ as its master key and sets $P_{pub} = sP$ as its public key.

3. The KGC defines hash functions $H_1 : \{0,1\}^* \rightarrow \mathbb{G}_1$, $H_2 : \{0,1\}^* \rightarrow \mathbb{G}_1$, and $H_3 : \{0,1\}^* \rightarrow \mathbb{Z}_q^*$, and a secure symmetric encryption algorithm $Enc(\cdot)$ (13).

4. The KGC publishes the system parameters $(\mathbb{G}_1, \mathbb{G}_2, e, P, P_{pub}, H_1(\cdot), H_2(\cdot), H_3(\cdot), Enc(\cdot))$.

Additionally, the format of a security emergency report (SER) (8) is also defined by the KGC. For an emergency event $\mathcal{E}_i$, the vehicle V_j generates a SER_j^i as follows:

$$SER_j^i = (Type_i, Loc_i, PID_j, Time_j^i, Sig_j^i, PK_j) \tag{1}$$

Note that for a specific emergency event $\mathcal{E}_i$, it is assumed that the relevant SERs will share the same $Type_i$ and Loc_i. For the detailed definition of each component of a SER, please refer to Table 1.

Registration

Prior to join the VANET, each vehicle has to register to the KGC. Suppose a new vehicle $\mathcal{V}_j$ with the identity ID_j wants to register with a KGC for IVC services. The details are presented as follows.

1. $\mathcal{V}_j$ sends ID_j to the KGC through an existing secure channel.
2. Upon receiving ID_j, the KGC first checks its validity. If ID_j is valid, the KGC uses the master key s to encrypt the real-identity ID_j into a pseudo-identity PID_j as follows.
$$PID_j = Enc_s(ID_j) \tag{2}$$
3. The KGC generates the partial private key D_j as follows.
$$D_j = sQ_j \tag{3}$$
where $Q_j = H_1(PID_j)$.
4. The KGC sends the pseudo-identity PID_j and the partial private key D_j back to $\mathcal{V}_j$ over a secure channel.
5. After receiving PID_j and D_j, V_j chooses a random number $x_j \in \mathbb{Z}_q^*$, sets its full private key as (x_j, D_j), and computes its public key PK_j as follows.
$$PK_j = x_j P \tag{4}$$

Signature Generation

When an emergency event $\mathcal{E}_i$ is sensed by the vehicle j and the observation is $(Type_i, Loc_i, Time_j^i)$, $\mathcal{V}_j$ generates a SER as follows.

1. $\mathcal{V}_j$ computes a pair (W_i, S_j) as follows.
$$W_i = H_2(Type_i \| Loc_i) \tag{5}$$
$$S_j = H_3(Type_i \| Loc_i \| Time_j^i \| PID_j \| PK_j) \tag{6}$$
where W_i is the hash value of the event statement and S_j is the hash value of the event statement binding the vehicle $\mathcal{V}_j$'s pseudo-identity and public key.
2. With the private key (x_j, D_j), $\mathcal{V}_j$ generates the signature Sig_j^i on (W_i, S_j) as follows.
$$Sig_j^i = D_j S_j + x_j W_i \tag{7}$$

Thus, $(Type_i, Loc_i, PID_j, Time_j^i, Sig_j^i, PK_j)$ constitutes a SER claim. After that, $\mathcal{V}_j$ broadcasts SER_j^i to its neighbors.

Given $SER_j^i = (Type_i, Loc_i, PID_j, Time_j^i, Sig_j^i, PK_j)$, a single SER verification can be performed by a verifier as follows.

1. The verifier first computes a triple (Q_j, W_i, S_j) as follows.

$$Q_j = H_1(PID_j) \tag{8}$$

$$W_i = H_2(Type_i \| Loc_i) \tag{9}$$

$$S_j = H_3(Type_i \| Loc_i \| Time_j^i \| PID_j \| PK_j) \tag{10}$$

2. After that, the verifier checks the validity of the signature as follows.

$$e(Sig_j^i, P) \stackrel{?}{=} e(Q_j S_j, P_{pub}) e(W_i, PK_j) \tag{11}$$

If equation (11) holds, the signature is accepted. The correctness of equation (11) can be checked as follows:

$$\begin{aligned} e(Sig_j^i, P) &= e(D_j S_j + x_j W_i, P) \\ &= e(D_j S_j, P) e(x_j W_i, P) \\ &= e(sQ_j S_j, P) e(W_i, x_j P) \\ &= e(Q_j S_j, sP) e(W_i, x_j P) \\ &= e(Q_j S_j, P_{pub}) e(W_i, PK_j) \end{aligned} \tag{12}$$

Aggregated Authentication

Aggregated authentication consists of two parts, signature aggregation and batch verification. The detailed procedures are presented as below.

- **Signature aggregation** For a specific emergency event $\mathcal{E}_i$, any vehicle can act as an aggregate signature generator, namely aggregator, who can aggregate a collection of individual signatures that have the same event statement, $Type_i$ and Loc_i. Given n SERs, where $SER_j^i = (Type_i, Loc_i, PID_j, Time_j^i, Sig_j^i, PK_j)$ by $\mathcal{V}_j (1 \leq j \leq n)$, the aggregator can obtain SER_{agg} as follows.

$$\begin{aligned} SER_{agg} = (&Type_i, Loc_i, PID_1, PID_2, \ldots, PID_n, \\ &Time_1^i, Time_2^i, \ldots, Time_n^i, \\ &Sig_1^i, Sig_2^i, \ldots, Sig_n^i, \\ &PK_1, PK_2, \ldots, PK_n) \end{aligned} \tag{13}$$

Then the aggregator computes Sig_{agg} as follows.

$$\begin{aligned} Sig_{agg} &= \sum_{j=1}^{n} Sig_j^i \\ &= \sum_{j=1}^{n} (D_j S_j + x_j W_i) \end{aligned} \tag{14}$$

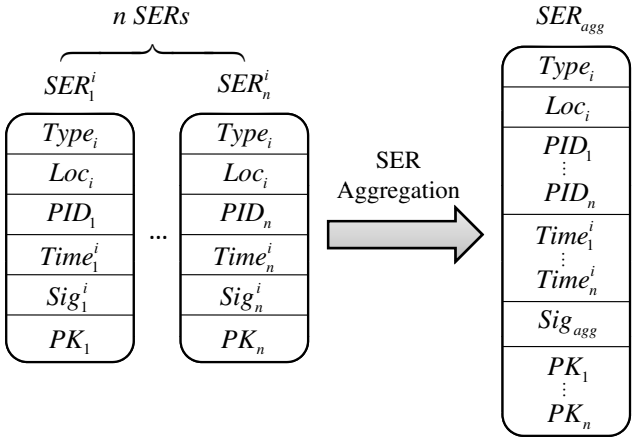


Figure 2: SER aggregation.

Now the aggregator obtains SER_{agg} as follows.

$$\begin{aligned} SER_{agg} &= (Type_i, Loc_i, PID_1, PID_2, \ldots, PID_n, \\ &\quad Time_1^i, Time_2^i, \ldots, Time_n^i, \\ &\quad Sig_{agg}, PK_1, PK_2, \ldots, PK_n) \end{aligned} \tag{15}$$

The aggregation procedure is illustrated in Figure 2.

- **Batch verification** Given the aggregate signature Sig_{agg} and the message set SER_{agg}, the aggregator computes a triple (Q_j, W_i, S_j) for $1 \leq j \leq n$ as follows.

$$Q_j = H_1(PID_j) \tag{16}$$

$$W_i = H_2(Type_i \| Loc_i) \tag{17}$$

$$S_j = H_3(Type_i \| Loc_i \| Time_j^i \| PID_j \| PK_j) \tag{18}$$

After that, the aggregator checks the validity of the aggregate signature as follows.

$$e(Sig_{agg}, P) \stackrel{?}{=} e(\sum_{j=1}^{n} Q_j S_j, P_{pub}) e(W_i, \sum_{j=1}^{n} PK_j) \tag{19}$$

If equation (19) holds, the aggregate signature is accepted. The correctness of equation (19) can be checked as follows:

$$\begin{aligned} e(Sig_{agg}, P) &= e(\sum_{j=1}^{n} (D_j S_j + x_j W_i), P) \\ &= e(\sum_{j=1}^{n} sQ_j S_j, P) e(\sum_{j=1}^{n} x_j W_i, P) \\ &= e(\sum_{j=1}^{n} Q_j S_j, sP) e(W_i, \sum_{j=1}^{n} x_j P) \\ &= e(\sum_{j=1}^{n} Q_j S_j, P_{pub}) e(W_i, \sum_{j=1}^{n} PK_j) \end{aligned} \tag{20}$$

Table 2: Formal definition of a Petri net.

A Petri net is a 5-tuple, (P, T, F, W, M_0) where:

$P = \{P_1, P_2, \cdots, P_m\}$ is a finite set of places,

$T = \{T_1, T_2, \cdots, T_n\}$ is a finite set of transitions,

$F \subseteq (P \times T) \cup (T \times P)$ is a set of arcs (flow relation),

$W : F \to \{1, 2, 3, \cdots\}$ is a weight function,

$M_0 : P \to \{0, 1, 2, 3, \cdots\}$ is the initial marking,

$P \cap T = \emptyset$ and $P \cup T \neq \emptyset$.

A Petri net structure $N = (P, T, F, W)$ without any specific initial marking is denoted by N.

A Petri net with the given initial marking is denoted by (N, M_0).

ANALYSIS OF THE SAMA SCHEME

In this section, we show that the SAMA scheme can resist forgery attacks and ensure the conditional privacy preservation and traceability of vehicles. In addition, we provide a performance comparison with other aggregate signature schemes for VANETs.

Security Analysis

We shall use Petri nets (12) to model and analyze the proposed scheme. Next, security properties of our scheme will be specified.

Petri Net Model

We used a Petri net to model the SER generation of the SAMA scheme. The formal definition of a Petri net (17) is listed in Table 2. Petri nets are composed from graphical symbols designating places (shown as circles), transitions (shown as rectangles), and directed arcs (shown as arrows). The places denote (atomic and composite) data items. The transitions denote decryption or decomposition operations. The directed arcs run between places and transitions.

When a transition fires, a composite data item is decomposed or decrypted, resulting in one or more simpler data items. Since we assume an open network environment, all data items in the transmitted messages are assumed to be public, and are known to the attacker. There will be tokens in the places representing the data items in the transmitted messages initially. From this initial marking, we can infer what an attacker can know eventually. Furthermore, we can also experiment what an attacker can know if he knows additional data items from other sources. The Petri net model is illustrated in Figure 3. The definitions of the places and transitions used in this model are listed in Table 3 and Table 4, respectively. We use the HPSim Petri net tool (18) to model our proposed scheme.

Security Properties

We now analyze the security properties of our scheme. The security of the proposed scheme is based on the difficulty of ECDLP, which is believed infeasible to solve in polynomial time. We will show that our scheme can resist forgery attacks and ensure the conditional privacy preservation and traceability of vehicles.

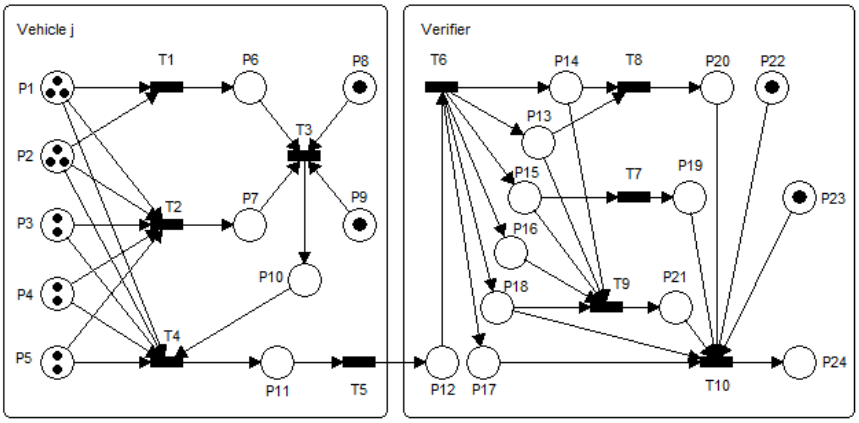


Figure 3: A Petri net model of the SER generation of the SAMA scheme.

Table 3: Definitions of places.

Place	Definition	Place	Definition
P_1	$Type_i$	P_{13}	$Type_i$
P_2	Loc_i	P_{14}	Loc_i
P_3	PID_j	P_{15}	PID_j
P_4	$Time_j^i$	P_{16}	$Time_j^i$
P_5	PK_j	P_{17}	Sig_j^i
P_6	W_i	P_{18}	PK_j
P_7	S_j	P_{19}	Q_j
P_8	D_j	P_{20}	W_i
P_9	x_j	P_{21}	S_j
P_{10}	Sig_j^i	P_{22}	P
P_{11}	SER_j^i	P_{23}	P_{pub}
P_{12}	SER_j^i	P_{24}	Success verification message

Theorem 1. *The proposed scheme can resist a forgery attack.*

Proof. If an adversary $\mathcal{A}$ wants to forge the vehicle $\mathcal{V}_j$ to produce a valid signature on the emergency event $\mathcal{E}_i$; according to the SER generation phase, $\mathcal{A}$ first computes a pair (W_i^*, S_j^*) as follows.

$$W_i^* = H_2(Type_i^* \| Loc_i^*) \tag{21}$$

$$S_j^* = H_3(Type_i^* \| Loc_i^* \| Time_j^{i*} \| PID_j \| PK_j) \tag{22}$$

After that, $\mathcal{A}$ generates the signature Sig_j^{i*}, where

$$Sig_j^{i*} = D_j S_j^* + x_j W_i^* \tag{23}$$

However, $\mathcal{A}$ cannot compute a valid signature unless $\mathcal{A}$ can obtain D_j and also derive x_j from PK_j. Based on the difficulty of ECDLP, it is computationally infeasible to compute x_j from PK_j. As shown in Figure 3, computing Sig_j^i is defined in transition T_3, which has four input places, P_6, P_7, P_8, and P_9. Place P_8 is the value of D_j and place P_9 is the value of x_j. Because having no idea about D_j and x_j, $\mathcal{A}$ cannot compute a valid signature and hence cannot launch a forgery attack. □

Table 4: Definitions of transitions.

Trans.	Definition	Trans.	Definition
T_1	Compute W_i	T_6	Split SER_j^i
T_2	Compute S_j	T_7	Compute Q_j
T_3	Compute Sig_j^i	T_8	Compute W_i
T_4	Constitute SER_j^i	T_9	Compute S_j
T_5	Transmit SER_j^i	T_{10}	Check $e(Sig_j^i, P) \stackrel{?}{=} e(Q_j S_j, P_{pub}) e(W_i, PK_j)$

Theorem 2. *The proposed scheme can ensure the conditional privacy preservation of vehicles.*

Proof. In the SAMA scheme, we propose to use pseudo-identities to preserve the identity privacy of witness vehicles. Since the vehicle $\mathcal{V}_j$ uses the pseudo-identity PID_j during its communication with other vehicles, the real-identity ID_j is protected. As shown in Figure 3, constituting SER_j^i and broadcasting SER_j^i to the verifier are defined in transition T_4 and T_5, respectively. Transition T_4 has six input places, P_1, P_2, P_3, P_4, P_5, and P_{10}. Place P_3 is the value of PID_j. However, only the KGC has the ability to trace the real-identity from the pseudo-identity PID_j. Hence, the conditional privacy preservation can be satisfied in the proposed scheme. □

Theorem 3. *The proposed scheme can provide the traceability of vehicles.*

Proof. Given the pseudo-identity PID_j, only the KGC, with the master key s, can trace the real-identity as follows.

$$\begin{aligned} Dec_s(PID_j) &= Dec_s(Enc_s(ID_j)) \\ &= ID_j \end{aligned} \tag{24}$$

Therefore, once a dispute occurs to the emergency message, the KGC has the ability to reveal the real-identity of the vehicle from the disputed message, in which the traceability can be achieved. □

Performance Evaluation

We use the computation and communication overhead as the metric to evaluate the performance of the proposed SAMA scheme. The evaluation parameters are defined in Table 5. The performance comparison between Zhu et al.'s scheme (8) and the SAMA scheme is presented in Table 6. According to the implementation results in (19), which observes processing time (in milliseconds) for an MNT curve of embedding degree $k = 6$ and 160-bit q, running on an Intel Pentium IV 3.0 GHz machine, T_P is 4.5 ms and T_M is 0.6 ms. Therefore, elliptic curve point multiplication operations are much cheaper in comparison to pairing operations.

From Table 6, Zhu et al.'s scheme (8) requires five pairings for verifying n distinct signatures and certificates; however, in the SAMA scheme, it requires only three pairings for verifying n distinct signatures without certificates. Therefore, our proposed scheme achieves better time efficiency than Zhu et al.'s scheme (8).

Table 5: Evaluation parameters.

Symbol	Definition
T_H	Time for performing a one-way hash function
T_E	Time for performing an exponentiation operation
T_P	Time for performing a bilinear pairing operation
T_M	Time for performing an elliptic curve point multiplication operation
T_A	Time for performing an elliptic curve point addition operation
T_{ENC}	Time for performing a symmetric encryption operation

Table 6: Performance comparison of aggregate signature schemes for VANETs.

	Zhu et al.'s scheme (8)	SAMA scheme
Registration	$1T_H + 2T_E$	$1T_H + 2T_M + 1T_{ENC}$
Sig. generation	$3T_H + 2T_E + 2T_M$	$2T_H + 2T_M + 1T_A$
Sig. verification	$4T_H + 1T_E + 5T_P$	$3T_H + 3T_P + 1T_M$
Sig. aggregation	$2(n-1)T_M$	$(n-1)T_A$
Batch verification	$(n+3)T_H + nT_E + 5T_P + 4(n-1)T_M$	$(2n+1)T_H + 3T_P + nT_M + 2(n-1)T_A$

Table 7: Broadcasting message format from a vehicle to its neighbors.

Component	$Type_i$	Loc_i	PID_j	$Time_j^i$	Sig_j^i	PK_j
Size (Bytes)	8	8	8	8	40	40

The communication overhead is in terms of the following aspect: the overhead incurred by broadcasting a SER from a vehicle to other vehicles within its transmission range. In our analysis, we assume the size of the element in $\mathbb{G}_1$ is 160-bit. The approximated length of the SER is shown in Table 7. $Type_i, Loc_i, PID_j$, and $Time_j^i$ each costs 8 bytes. The fifth part is the 40-byte signature on the emergency event and the last part is the public key of the vehicle, which also costs 40-byte. Thus, the communication overhead incurred by broadcasting a SER from a vehicle to its neighbors is 112 bytes.

CONCLUSIONS

In this paper, we propose a secure aggregated message authentication (SAMA) scheme based on bilinear pairings for VANETs. The SAMA scheme makes use of aggregation and batch verification techniques for emergency message verification to reduce the computation overhead. Compared to Zhu et al.'s scheme (8), the SAMA scheme achieves more efficient authentication on emergency messages. Moreover, we used Petri nets in the security analysis of the proposed scheme. Our analysis shows that the proposed scheme can successfully defend forgery attacks and ensure the conditional privacy preservation and traceability of vehicles.

ACKNOWLEDGEMENTS

This work was supported by the National Science Council, Taiwan, Republic of China, under grant NSC 97-2221-E-009-048-MY3, NSC 97-2221-E-009-049-MY3, and NSC 96-2628-E-009-014-MY3.

REFERENCES

(1) M. Raya and J. P. Hubaux, "The security of vehicular ad hoc networks," in *Proceedings of the 3rd ACM Workshop on Security of Ad hoc and Sensor Networks (SASN 2005)*, Nov. 2005.

(2) M. Raya, A. Aziz, and J. P. Hubaux, "Efficient secure aggregation in VANETs," in *Proceedings of the 3rd International Workshop on Vehicular Ad hoc Networks (VANETs 2006)*, Sep. 2006, pp. 67–75.

(3) M. Raya and J. P. Hubaux, "Securing vehicular ad hoc networks," *Journal of Computer Security*, vol. 15, no. 1, 2007, pp. 39–68.

(4) X. Lin, R. Lu, C. Zhang, H. Zhu, P. H. Ho, and X. Shen, "Security in vehicular ad hoc networks," *IEEE Communications Magazine*, vol. 46, no. 4, Apr. 2008, pp. 88–95.

(5) N. W. Wang, Y. M. Huang, and W. M. Chen, "A novel secure communication scheme in vehicular ad hoc networks," *Computer Communications*, vol. 31, no. 12, Jul. 2008, pp. 2827–2837,

(6) C. T. Li, M. S. Hwang, and Y. P. Chu, "A secure and efficient communication scheme with authenticated key establishment and privacy preserving for vehicular ad hoc networks," *Computer Communications*, vol. 31, no. 12, Jul. 2008, pp. 2803–2814.

(7) C. Zhang, X. Lin, R. Lu, P. H. Ho, and X. Shen, "An efficient message authentication scheme for vehicular communications," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 6, Nov. 2008, pp. 3357–3368.

(8) H. Zhu, X. Lin, R. Lu, P. H. Ho, and X. Shen, "AEMA: An aggregated emergency message authentication scheme for enhancing the security of vehicular ad hoc networks," in *Proceedings of the IEEE International Conference on Communications (ICC 2008)*, May 2008, pp. 1436–1440.

(9) A. Shamir, "Identity based cryptosystems and signature schemes," in *Proceedings of the Advances in Cryptology*, 1984, pp. 47–53.

(10) S. Al-Riyami and K. Paterson, "Certificateless public key cryptography," in *Proceedings of the ASIACRYPT*, 2003, pp. 452–473.

(11) L. Zhang and F. Zhang, "A new certificateless aggregate signature scheme," *Computer Communications*, vol. 32, no. 6, Apr. 2009, pp. 1079–1085.

(12) C. A. Petri, *Kommunikation mit automaten.* PhD thesis, University of Bonn, 1962.

(13) D. R. Stinson, *Cryptography: Theory and practice.* Boca Raton, FL: Chapman & Hall/CRC, 2006.

(14) "IEEE 802.11p, Amendment 6: Wireless access in vehicular environments (WAVE)," 2010.

(15) L. Hu and D. Evans, "Secure aggregation for wireless networks," in *Proceedings of the 2003 Symposium on Applications and the Internet Workshops (SAINT 2003)*, Jan. 2003, pp. 384–391.

(16) Y. Yang, X. Wang, S. Zhu, and G. Cao, "SDAP: A secure hop-by-hop data aggregation protocol for sensor networks," *ACM Transactions on Information and System Security*, vol. 11, no. 4, Jul. 2008.

(17) T. Murata, "Petri nets: Properties, analysis and applications," *Proceedings of the IEEE*, vol. 77, no. 4, Apr. 1989, pp. 541–580.

(18) "HPSim 1.1 Petri nets simulation tool, copyright© 1999-2002 Henryk Anschuetz."

(19) M. Scott, "Efficient implementation of crytographic pairings," [Online]. Available: http://ecrypt-ss07.rhul.ac.uk/Slides/Thursday/mscott-samos07.pdf.

參加「2011 IEEE 73rd VTC」國際會議報告

簡榮宏

一、參加會議經過

本次的 2011 IEEE 73rd VTC (2011 IEEE 73rd Vehicular Technology Conference) 於匈牙利布達佩斯舉行。本次會議共有 1164 篇論文投稿，接受 423 篇論文在會議中口頭發表與及 173 篇論文海報發表。因此安排了 86 場 oral sessions 與 9 場 poster 與展示的 sessions。除此之外，另安排有 5 個 workshops: (1) International Workshop on Self-Organizing Networks; (2) 2nd Green Wireless Communications and Networks Workshop; (3) Cognitive radio and Cooperative strategies for POWER saving; (4) Broadband Femtocell Technologies; (5) 1st International Workshop on Cross-Layer Operation Aided Multimedia Streaming。以及 6 場 tutorials: (1) Towards Holistic Green Communications and Networking; (2) Cognitive radio based on UWB technology – a perfect binomial; (3) Participatory Sensing: Crowdsourcing Data from Mobile Smartphones in Urban Spaces; (4) Cooperative active and passive localization and tracking: fundamental limits and UWB case study; (5) Low-Complexity Algorithms for Large-MIMO Detection; (6) Mobility models and social networks。是個相當盛大的國際會議。

第一天 5/15 是大會安排的 5 workshops 與 6 場 tutorials 首先登場。

第二天 5/16 早上 8:30-10:00 有一個 Plenary 演講，由 Dr. Magnus Frodigh, (Director Wireless Access Networks, Ericsson Research, Ericsson)主講。他的講題為「Navigating the Mobile Data Growth – Research Challenges」。他強調在行動網路中，並不是只有增加資料傳輸速率和容量的挑戰，隨著網路的複雜性增加，如何利用自主組織之性能(self organizing features)來減少的營運成本，更是一項挑戰。我們發表的論文「An Efficient Cluster-based Data Dissemination Scheme in Wireless Sensor Networks」被安排在 Routing 的 session，於上午 10:30-12:00 發表，與會有人對於如何做 clustering data dissemination 有興趣。

第三天 5/17，早上 8:30-10:00 有一個 Panel 的討論，主題為「Wireless Futures…」，主要是探討無線通訊研究未來的研究方向及其可能的發展。

第四天 5/18，早上 8:30-10:00 的 Panel 討論「The Networked, Plugged Smart Vehicle」，主要討論未來的智慧車輛，這項技術須同時結合行動網際網路與交通兩大關鍵元素，才有可能成功。

二、與會心得

(1) VTC (Vehicular Technology Conference)是車載相關技術的國際會議，每年分

春季與秋季各舉行一次，也是車載網路技術的主要論壇，與會的人員討論與發問都相當踴躍，可學到不少研究經驗。

(2) 有關「Green Radio Network」也是此次會議的焦點之一，未來在無線通訊網路的編碼與規約的設計上，除了傳統考慮 bits/second 的效能之外，Joules/bit 亦是一個重要的指標。

三、攜回資料

(1) Proceedings of 2011 IEEE 73rd VTC 光碟一片。

Rong-Hong Jan <ronghong.jan@gmail.com>

[TrackChair] Congratulations, your VTC'11S paper was accepted

1 封郵件

Lajos Hanzo <lh@ecs.soton.ac.uk> **2011年1月10日上午1:05**
收件者: arvin10211021@hotmail.com, kwang@cs.nctu.edu.tw, rhjan@cs.nctu.edu.tw, yhu@cs.nctu.edu.tw, 21203@cch.org.tw

Regarding the following paper:

Title: An Efficient Cluster-based Data Dissemination Scheme in Wireless Sensor Networks
Paper: http://vtc2011spring.trackchair.com/paper/48219

Dear Colleague,

On behalf of the Technical Program Committee, I am pleased to inform you that the above paper has been conditionally accepted for <ORAL/POSTER> presentation at <CONFERENCE> in <LOCATION>.

All submitted papers have been thoroughly independently reviewed.

This paper is accepted contingent upon: 1) your completion and submission of the camera-ready final version of the technical manuscript for the paper, 2) receipt of the signed copyright form, and 3) your pre-registration for the conference. All of these items must be done by <FINAL_SUBMISSION_DATE>.

Additionally, the IEEE Vehicular Technology Society requires that each accepted paper be presented in-person at the conference site according to the schedule published. It reserves the right to exclude from distribution on IEEE Xplore any paper not presented on-site. If none of the authors are able to attend, by a qualified surrogate may present the paper, and registrations may be transferred free of charge.

Camera ready copy must be submitted to the IEEE Conference eXpress Publishing system, as they will be producing the final proceedings.

**** do NOT upload final manuscripts to TrackChair ****

Full details can be found on the conference web site at http://www.vtc2011spring.org/final-submission.php

Reviews of your paper can be found on the TrackChair site at the URL at the top of this email, in the "Actions" sidebar, under "Reviews". Please take the reviewers comments into account where appropriate in preparing your final manuscript.

Note that it is the VT Society policy that each paper must have either a full IEEE member or non-member registration (not a student registration). Authors submitting more than one paper must pay a full registration for the first paper, plus an extra paper charge for each additional paper. If you only have a paper in a workshop, you may register just for the workshop if you do not wish to attend the main conference. Your registration must be completed before you will be able to upload your manuscript to the IEEE Conference eXpress Publishing site.

Thank you for submitting your paper to <CONFERENCE> - we look forward to a reunion with you in the vibrant city of Budapest.

<CHAIR>
<CONFERENCE> Chair

--
Lajos Hanzo, University of Southampton
http://www.trackchair.com/account/1955
This message was sent using TrackChair.

An Efficient Cluster-based Data Dissemination Scheme in Wireless Sensor Networks

Ren-Jhong Liu, Kuochen Wang
Rong-Hong Jan and Yuh-Jyh Hu

Department of Computer Science
National Chiao Tung University
{arvin1021.cs97g, kwang, rhjan, yhu}@cs.nctu.edu.tw

Tien-Hsiung Ku

Department of Anesthesiology
Changhua Christian Hospital
21203@cch.org.tw

[1] ***Abstract*— Existing data dissemination protocols adopted flooding to propagate interests and find forwarding paths in wireless sensor networks (WSNs), which cause large energy consumption. To relieve this problem, we propose an *efficient cluster-based data dissemination* (ECDD) scheme. In the proposed ECDD, besides piggyback control information into interests to perform on-demand passive clustering, we also use control information to set each node a hop count for assisting a node to select next forwarding node with the least hop count to the sink. In this way, a shortest path to forward sensed data back to the sink can be found. Simulation results show that the proposed ECDD is 61.30% better than DD, a classical approach, and 22.33% better than ELPC in terms of average dissipated energy. Furthermore, our approach is 57.45% and 23.49% better than DD and ELPC, respectively, in terms of average delay. The proposed ECDD is feasible for applications of long-term monitoring and real-time response, such as a community health care system and rescue operation in a disaster area.**

Keywords **- Cluster-based, data dissemination, piggybacked control information, wireless sensor network.**

I. INTRODUCTION

A wireless sensor network (WSN) consists of a large number of small sensor nodes with low cost. Typically, sensor nodes are deployed throughout an area to sense events that we are interested in. The sink will send request messages to sensor nodes if the sink is interested in some types of events. The request messages which the sink sends to the sensor nodes are called *interests*. When sensor nodes sense event data that match the interests, the sensed data will be forwarded back to the sink through multi-hop paths. Because sensor nodes have the characteristics of low cost and well adaptability to environments, the WSN technology can be applied to various domains, such as wildlife habitats, disaster management, emergency response, ubiquitous computing environments, asset tracking, healthcare, and manufacturing process flow [2]. Nevertheless, the capabilities of a sensor node, such as battery power, computing, signal range and storage space, are limited. Thus, efficient resource management of a sensor node is essential, especially for energy management. The energy consumption of sensor nodes is greatly affected by the forwarding path selection mechanisms in data dissemination protocols. Therefore, designing an energy-efficient forwarding path selection mechanism becomes an important research issue in WSNs [1].

[1] The support by the National Science Council under Grants NSC 97-2221-E-009-049-MY3, NSC 99-2218-E-009-002 and NSC 99-2221-E-009-081-MY3 is gratefully acknowledged.

II. RELATED WORK

Directed Diffusion (DD) [3][4] is a classical data-centric data dissemination scheme in WSNs. The sink repeatedly diffuses interests to inform each sensor node what kinds of events the sink wants. When a sensor node senses events which match the interests and the forwarding path is not decided, the sensor node sends exploratory data messages along all possible forwarding paths toward the sink. After the sink received the exploratory data messages, the sink chooses a path with the lowest delay, which is called a reinforced forwarding path, according to the received exploratory data messages. The sink notifies the sensor node with this path using a reinforcement message, and the sensor node uses this reinforced forwarding path to forward sensed event data back to the sink instead of flooding. In DD, there are two flooding problems. First, the sink periodically floods interests in order to notify sensor nodes to send sensed event data that match the interests [5]. Second, in order to find and maintain reinforced forwarding paths, sensor nodes repeatedly send exploratory data messages by flooding [13]. Because of these two flooding problems, the energy of sensor nodes will be exhausted rapidly.

Energy Level-based Passive Clustering (ELPC) [10] modifies a passive clustering mechanism for building and maintaining the cluster formation. ELPC also combines DD so that the consumed energy balancing of the sensor nodes can be achieved [10]. It adds additional two parameters, the node energy level and the network energy level, in order to achieve balanced energy consumption in each node. The node energy level represents the total energy consumption of a sensor node and the network energy level represents an energy threshold. Cluster heads or border nodes change their node statuses when their node energy levels are higher than the predefined network energy level [10]. Each sensor node has chances to be a cluster head or a border node so that the energy consumption of each node will be balanced and the network lifetime will be prolonged. ELPC consumes more energy and has more delay due to flooding exploratory data messages for selecting a data forwarding path.

III. PROPOSED EFFICIENT CLUSTER-BASED DATA DISSEMINATION SCHEME

In this section, we propose an Efficient Cluster-based Data Dissemination (ECDD) scheme and describe details of our

proposed approach. The proposed ECDD is based on a passive clustering mechanism to build and maintain clusters and uses a *first declaration wins* [8][9] mechanism to select cluster heads. With the *first declaration wins* mechanism, a node which first claims to be the cluster head will become the cluster head of its clustered area (in the radio coverage) and manages nodes in this clustered area [8][9]. The proposed ECDD does not use waiting periods (to make sure all the neighbors have been checked) [8], which is used in all other weight-driven clustering mechanisms, to select the best node to be the cluster head [8]. Once the cluster formation is finished, all nodes can be categorized into three types: *cluster head*, *cluster member*, and *border node*. A cluster head has the responsibility to manage cluster members and border nodes in its clustered area. A border node is a node which is in the communication range of two or more different clusters. The main tasks of a cluster member are receiving interests from its cluster head and forwarding sensed data that match the received interests to its cluster head. We make the following assumptions in the proposed ECDD:

- The sink has unlimited memory, processing capability and rechargeable battery. That is, the energy of the sink can be recharged anytime [14].
- All sensor nodes besides the sink have the same processing capability.
- Initially, the hop count of the sink is set to 0 and the hop count of each node other than the sink is set to the total number of sensor nodes in the sensor field.
- The cluster formation is complete before any data gathering and transmission begins [5].

A. Piggybacked control information in an interest

The piggybacked control information in the interests is used to set the hop count of each node for selecting an efficient data dissemination path and an energy threshold of each node for reconstructing clusters. In the proposed ECDD, we extend the piggybacked control information proposed by [11] by adding two more fields: HOP_COUNT and ENERGY_THRESHOLD, as shown in Figure 1.

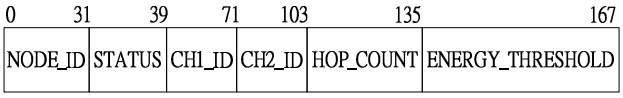


Figure 1. Piggybacked control information in an interest.

Once a sensor node receives an interest, it updates a *neighbor information table* according to the piggybacked control information in the interest. The neighbor information table contains the following information: *NEIGHBOR_ID*, *NEIGHBOR_STATUS*, and *NEIGHBOR_HOP_COUNT*. NEIGHBOR_ID is used to identify which neighbor is referred by this information record. NEIGHBOR_STATUS keeps track of which type of this neighbor is. NEIGHBOR_HOP_COUNT represents the total hop counts from this neighbor to the sink. According to the neighbor information table, a sensor node selects the next forwarding node with the least hop count, and thus a shortest data dissemination path to forward sensed data back to the sink can be found.

B. Cluster formation and data dissemination path selection

If a cluster-based sensor network has not been initialized, the cluster formation mechanism will be performed. In the proposed ECDD algorithm, we design a novel cluster formation mechanism which utilizes hop count information in each sensor node. After the cluster formation mechanism is finished, the data dissemination path selection process can make use of the updated hop count information in each sensor node to select a data dissemination path from source to sink with low delay.

1) Cluster formation

Initially, the hop count of each node is set to the total number of sensor nodes in the sensor field. When an interest is propagated into the sensor field for the first time, the cluster formation mechanism will be performed, and the interest will be propagated to all sensor nodes in the sensor field. Before the interest is propagated into the sensor field, the HOP_COUNT field in a piggybacked interest will be set to the value of the hop count stored in the sink plus one. When a sensor node receives the interest, the sensor node will compare its hop count with the value stored in the HOP_COUNT field of the received interest. If the hop count value stored in the sensor node is larger than the value of the HOP_COUNT field in the received interest, the sensor node updates its hop count by the value stored in the HOP_COUNT field of the received interest. Otherwise, the interest will be discarded. Then, the sensor node checks whether there are more interests coming in or not. If the sensor node receives other interests, the sensor node must compare the hop count value with the value of the HOP_COUNT field in the recently received interests again to decide whether the hop count value needs to be updated or not. If the sensor node does not receive any more interests, it replaces the value of the HOP_COUNT field in the received interest with the value of the sensor node's hop count plus 1. Finally, the sensor node sends the updated interest to other sensor nodes. When an interest is diffused throughout the entire sensor field, each node will be set a new hop count.

Figure 2 illustrates flowcharts of the actions taken by cluster heads, cluster members, and border nodes. Figure 2(a) shows the actions taken by a cluster head. After the cluster formation is finished, a cluster head will piggyback its ID, status, and the hop count value plus 1 into the NODE_ID, STATUS, and HOP_COUNT fields in the received interest. After finishing the update of the received interest, the cluster head then rebroadcasts the updated interest to its cluster members and border nodes. Figure 2(b) illustrates the actions taken by a cluster member. If a cluster member receives a modified interest, the cluster member will replace its hop count with the HOP_COUNT in the modified interest. The cluster member also stores NODE_ID, STATUS and HOP_COUNT minus 1 in the modified interest to the neighbor information table.

Figure 2(c) depicts the actions taken by a border node. If a border node receives a modified interest, the border node will compare its hop count with the HOP_COUNT of the modified interest. If the HOP_COUNT in the modified interest is larger than the border node's hop count, the border node stores NODE_ID, STATUS, and HOP_COUNT minus 1 in the modified interest to the corresponding fields in the neighbor information table, then discards the interest. Otherwise, the

border node will replace its hop count with the HOP_COUNT of the modified interest. Besides, the border node also stores NODE_ID, STATUS, and HOP_COUNT minus 1 in the modified interest to the corresponding fields in the neighbor information table. No matter the border node updates its hop count or not, it updates the received interest according to its NODE_ID, STATUS, CH1_ID, CH2_ID, and HOP_COUNT, and sends the updated interest to each cluster head that sends the modified interest to the border node.

2) Data dissemination path selection

In this phase, we want to find a shortest data dissemination path to forward sensed data back to the sink after the cluster formation is finished. If a sensor node senses an event that matches an interest, the sensor node will become a source and periodically forwards the sensed data to its cluster head. After the cluster head receives sensed data, the cluster head checks whether the next hop is the sink or not. If the next hop is the sink, the sensed data will be forwarded to the sink directly. Otherwise, the cluster head looks up its neighbor information table to select a border node with the least hop count. The cluster head then forwards the sensed data to the selected border node. After the selected border node receives the sensed data, it also looks up its neighbor information table to find a cluster head with the least hop count, which connects to the selected border node. The border node then forwards the sensed data to the selected cluster head. The sensed data will be forwarded hop by hop between cluster heads and border nodes until the sensed data arrive at the sink.

Figure 3 shows an example of how the sensed data being transferred toward the sink. Because node R senses an event that matches an interest, node R becomes a source and forwards sensed data to it cluster head, which is node K. When node K receives the sensed data, node K checks whether the next hop is the sink or not. If the next hop is not the sink, node K looks up its neighbor information table to select a border node with the least hop count among nodes J, F, and L, which is node F. When node F receives the sensed data, node F also looks up its neighbor information table to select a cluster head with the least hop count among nodes B, I, M, and K, which is node B. When node B receives the sensed data, node B forwards the sensed data to the sink because the next hop of node B is the sink.

C. Route maintenance mechanism

In the proposed ECDD, besides finding an efficient data dissemination path to reduce the energy consumption of sensor nodes, we also need to consider a situation that an data dissemination path may change if some of the sensor nodes become invalid due to energy depletion. In order to deal with the invalidation of a data dissemination path caused by energy depletion, we piggyback ENERGY_THRESHOLD, which is the energy threshold of a sensor node, into an interest. The value of ENERGY_THRESHOLD is controlled by the sink. When the sink wants to send an interest to a sensor field, the sink sets an energy threshold value to the ENERGY_THRESHOLD field of an interest. Therefore, when the residual energy of a sensor node is below the ENERGY_THRESHOLD, the sensor node sends a *reconstruction message* to notify the sink that the cluster must be reconstructed. After the sink receives the reconstruction message, the reconstruction of the cluster formation will be initiated by the sink.

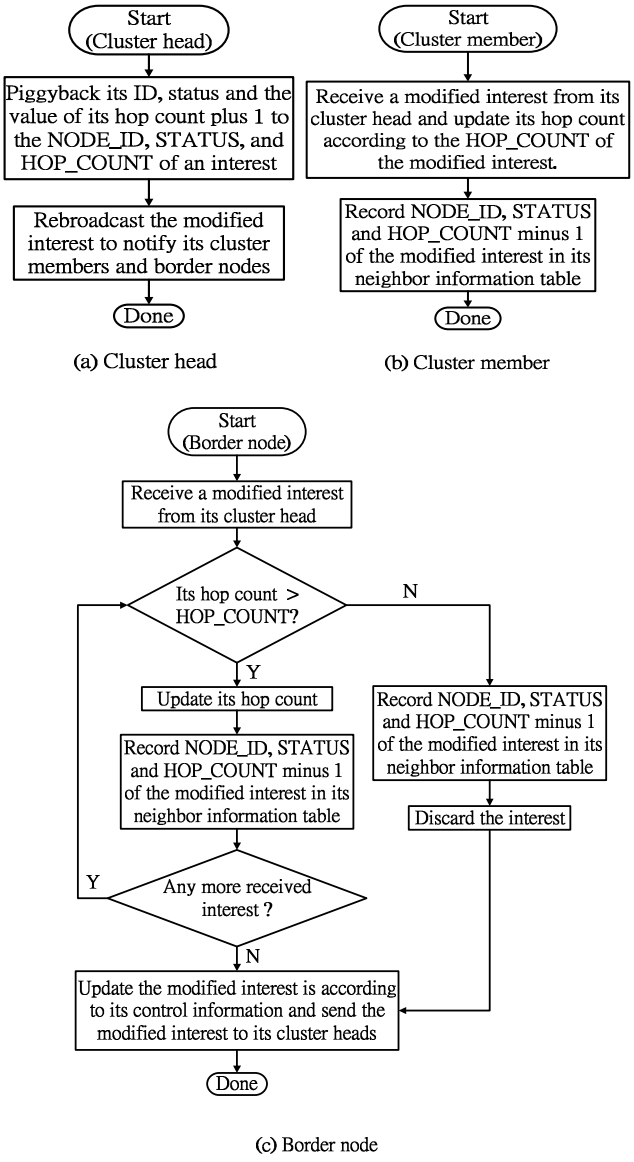


Figure 2. Flowcharts of the actions taken by (a) cluster heads, (b) cluster members, and (c) border nodes after the cluster formation is finished.

IV. PERFORMANCE EVALUATION

A. Simulation environment and parameters

Sensor nodes are randomly placed in a 160 x 160 m^2. The transmission range of each sensor node is 40 m. In the simulations, we set a single sink and five sources. The sink is located in the bottom left corner of the sensor field and the five sources are randomly selected from the nodes in the sensor field. The sink sends interests every 20 seconds, and the sources send data messages every 2 seconds after they receive an interest [11]. The data messages are 64-byte long for DD and ECDD and 72-byte long for ELPC. The interests are

57-byte, 49-byte and 36-byte long for ECDD, ELPC, and DD, respectively. The initial energy of a sensor node is 10 *J*. The transmit power is 0.66 *W* and the receive power is 0.395 *W*. The simulation time is 1000 seconds. We evaluate each scheme with the following parameters: *average dissipated energy*, *average delay*, *throughput*, and *the time till the first node death,* which are defined as follows.

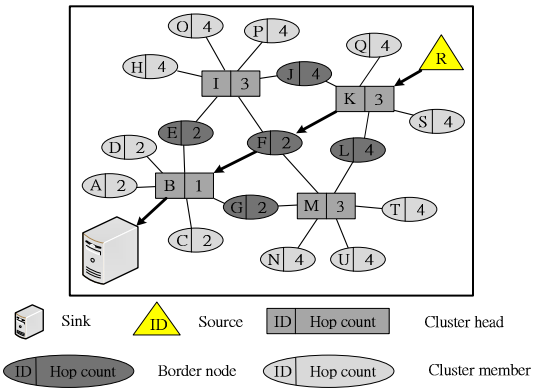


Figure 3. An example of how sensed data being transferred toward the sink.

- *Average dissipated energy*

The average dissipated energy (E_{avg}) measures the ratio of total dissipated energy of all nodes in the network to the number of distinct events seen by the sink [3][4], as defined in equation (1).

$$E_{avg} = \frac{\sum_{k=1}^{m}(Ei_k - Ef_k)}{n} \quad (1)$$

where Ei_k is the initial energy of the k^{th} node, Ef_k is the final energy of the k^{th} node, m is the total number of sensor nodes, and n is the total number of events.

- *Average delay*

The average delay (D_{avg}) measures the average transmission time of events, which indicates the average latency between a source that transmits an event and the sink that receives the event [14]. The definition of D_{avg} is as follows:

$$D_{avg} = \frac{\sum_{i=1}^{n}(Ts_i - Te_i)}{n} \quad (2)$$

where Ts_i is the timestamp when the interest of the i^{th} event is transmitted from the sink, Te_i is the timestamp when the i^{th} event is received by the sink, and n is the total number of events.

- *Throughput*

This metric indicates the performance of data dissemination by the sensor network.

$$Throughput = \frac{Successfully\ received\ data\ packets}{Total\ elapse\ time} \quad (3)$$

- *The time till the first node death (FND)* [6]

The FND represents the timestamp that the first sensor node dies due to the depletion of energy [6]. It indicates whether the energy consumption of nodes in the sensor field is balanced or not.

B. Comparison of the proposed ECDD with DD and ELPC

Figure 4 shows the average dissipated energy of DD, ELPC and ECDD. Because the ELPC protocol only floods interests among cluster heads and border nodes, it consumes less energy than DD. However, the ELPC protocol also floods the exploratory data messages to find a forwarding path, which causes additional energy consumption. The energy consumption of ECDD is the lowest because ECDD uses a cluster-based scheme to avoid the interests flooding problem and uses hop count information to avoid the exploratory data message flooding problem. Simulation results show that ECDD is 61.30% and 22.33% better than DD and ELPC, respectively, in terms of average dissipated energy.

Figure 5 shows the average delay of each approach. In ELPC, the average delay is better than that of DD because ELPC uses clusters to reduce the number of flooded exploratory data messages. Although the number of flooded exploratory data messages in ELPC is decreased, the flooding problem still exists and it causes additional overheads. The proposed ECDD has the lowest average delay because we use hop count information to further eliminate flooding of exploratory data messages. Simulation results in Figure 5 show that in terms of the average delay, the proposed ECDD is 57.45% and 23.49% better than DD and ELPC, respectively.

Figure 6 shows the comparison of throughput. Although ELPC reduces the degree of the interest flooding problem by using clusters, which makes ELPC better than DD, ELPC still must flood exploratory data messages among cluster heads and border nodes to find a forwarding path. By piggybacking additional control information into interests, ECDD can establish clusters by sending interests from the sink. Thus, it has the highest throughput. Simulation results show that the proposed ECDD is 55.54% and 15.85% better than DD and ELPC, in terms of throughput.

Figure 7 Figure 7 shows the comparison of the time till that the first node death (FND) among the proposed ECDD, DD, and ELPC. Because both ELPC and ECDD use a cluster-based mechanism, cluster heads and border nodes will deplete their energy faster than cluster members. In spite of the energy consumption of each node is balanced in ELPC, the amount of exploratory data messages will increase when the network scale increases. Thus, it causes cluster heads and border nodes to deplete their energy faster than that of the proposed ECDD. Simulation results show that the proposed ECDD is 71.62% and 24.40% better than DD and ELPC, respectively, in terms of FND.

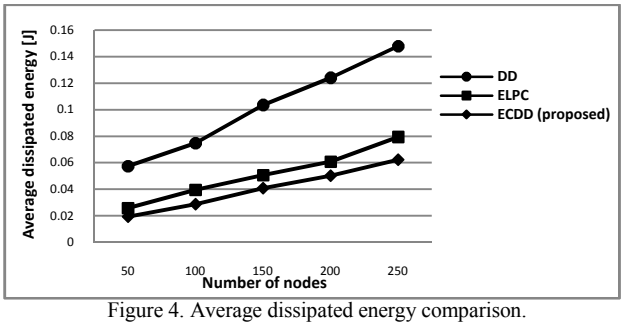


Figure 4. Average dissipated energy comparison.

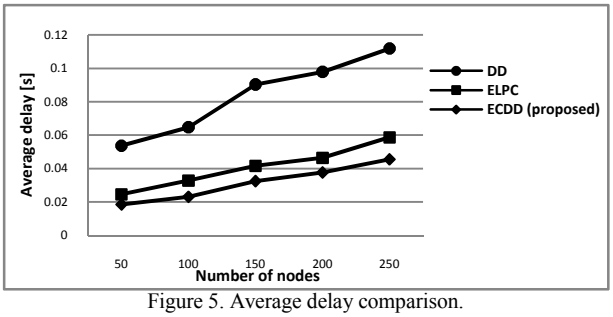


Figure 5. Average delay comparison.

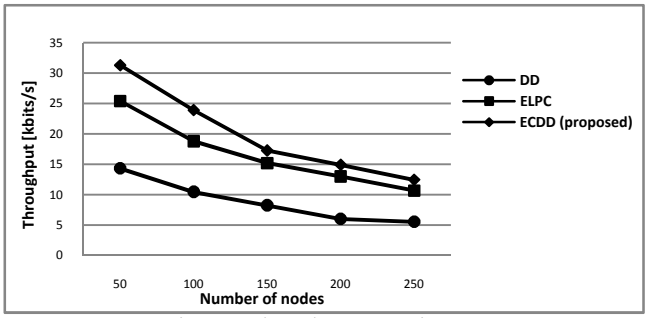


Figure 6. Throughput comparison.

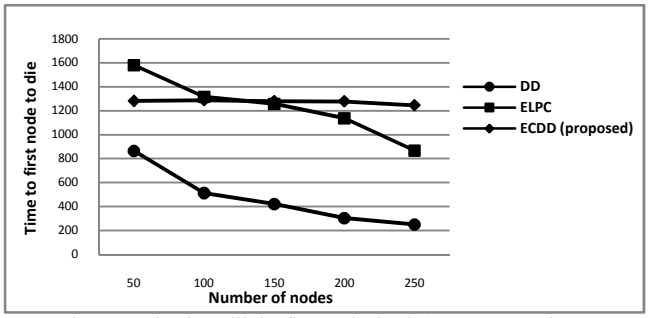


Figure 7. The time till the first node death (FND) comparison.

V. CONCLUSION

We have presented an efficient cluster-based data dissemination (ECDD) scheme that piggybacks hop count information in an interest to each node. The hop count information can assist a node to select the next forwarding node with the least hop count. In this way, without flooding exploratory data messages, a shortest data dissemination path to forward sensed data back to the sink can be found, while saving energy and reducing delay. Simulation results have shown that ECDD is 61.30% and 22.33% better than DD and ELPC in terms of average dissipated energy, respectively. Furthermore, ECDD is 57.45% and 23.49% better than DD and ELPC in terms of average delay, respectively. Besides, ECDD is 55.54% and 15.85% better than DD and ELPC in terms of throughput, respectively. In addition, we also piggyback an energy threshold into an interest to balance the energy consumption of sensor nodes and prolong network lifetime. Simulation results have also shown that, in terms of the time till the first node death (FND), the proposed ECDD is 71.62% and 24.40% better than DD and ELPC, respectively.

REFERENCES

[1] I.F. Akyildiz, S. Weilian Su, Y. Sankarasubramaniam, E. Cayirci., "A Survey on Sensor Networks," *IEEE Commun. Mag.*, Vol. 40, Issue 8, pp. 102-114, Aug. 2002.

[2] D. Culler, D. Estrin, M. Srivastava, "Guest Editors' Introduction: Overview of Sensor Networks," *IEEE Computer*, Vol. 37, Issue 8, pp. 41- 49, Aug. 2004.

[3] C. Intanagonwiwat, R. Govindan, D. Estrin, "Directed Diffusion: a Scalable and Robust Communication Paradigm for Sensor Networks," in *Proceedings of the ACM International Conference on Mobile Computing and Networking*, pp. 56-67, Aug. 2000.

[4] C. Intanagonwiwat, R. Govindan, D. Estrin , J. Heidemann, F. C. Silva, "Directed Diffusion for Wireless Sensor Networking," *IEEE/ACM Trans. Netw.*, Vol. 11 , Issue 1, pp. 2-16, Nov. 2003.

[5] Y. Cui , J. Cao, "An Improved Directed Diffusion for Wireless Sensor Networks," in *Proceedings of the IEEE International Conference on Wireless Communications, Networking and Mobile Computing*, pp. 2380-2383, Sept. 2007.

[6] A. Ozgovde, C. Ersoy, "WCOT: A Realistic Lifetime Metric for the Performance Evaluation of Wireless Sensor Networks," in *Proceedings of the IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communication*, pp. 1-5, Sept. 2007.

[7] X. Liu, F. Li, H. Kuang, X. Wu, "The Study of Directed Diffusion Routing Protocol Based on Clustering for Wireless Sensor Networks," in *Proceedings of the IEEE 6th World Congress on Intelligent Control and Automation*, Vol. 1, pp. 5120-5124, Oct. 2006.

[8] J. K. Taek, M. Gerla, V. K. Varma, M. Barton, T. R. Hsing, "Efficient Flooding with Passive Clustering - an Overhead-Free Selective Forward Mechanism for Ad Hoc/Sensor Network," in *Proceedings of the IEEE/ACM MobiHoc*, Vol. 91, Issue 8, pp. 1210-1220, Aug. 2003.

[9] M. Gerla, T. J. KOWN, G. Pei, "On Demand Routing in Large Ad Hoc Wireless Networks with Passive Clustering," in *Proceedings of the IEEE Wireless Communications and Networking*, vol. 1, pp. 100-105, Aug. 2002.

[10] H. Zeghilet, N. Badache, M. Maimour, "Energy Efficient Cluster-based Routing in Wireless Sensor Networks," in *Proceedings of the IEEE Symposium on Computers and Communications*, pp. 701-704, Aug. 2009.

[11] V. Handziski, A. Kopke, H. Karl, C. Frank, W.Drytkiewicz, "Improving the Energy Efficiency of Directed Diffusion Using Passive Clustering," in *Proceedings of 1st European Workshop in Wireless Sensor Networks*, Vol. 2920, pp. 172-187, 2004.

[12] Y. Zhang, L. Wang, "A Comparative Performance Analysis of Data Dissemination Protocols in Wireless Sensor Networks," in *Proceedings of the IEEE 7th World Congress on Intelligent Control and Automation*, pp. 6669-6674, June 2008.

[13] A. Booranawong, W. Teerapabkajomdet, "Reduction of Exploratory Data Messages on Directed Diffusion in Mobile Wireless Sensor Networks," in *Proceedings of the IEEE 6th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, Vol. 2, pp. 996-999, May 2009.

[14] E. Lee, S. Park, F. Yu, Y. Choi, M. S. Jin, S. H. Kim, "A Predictable Mobility-based Data Dissemination Protocol for Wireless Sensor Networks," in *Proceedings of the IEEE 22nd International Conference on Advanced Information Networking and Applications*, pp. 741-747, March 2008.