Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Real-Time Medium Access Control in Vehicular Ad Hoc Networks

Shu Zhang

Department of Computer Science and Statistics

University of Dublin, Trinity College

A thesis submitted for the degree of

Doctor of Philosophy

April 2013

Ph. D.
School of Computer Science & statistics

Declaration

I, the undersigned, declare that this work has not previously been submitted to this or any other University, and that unless otherwise stated, it is entirely my own work.

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this upon request.

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this upon request.

Shu Zhang

Dated: April 30, 2013

Acknowledgements

I would like to express my gratitude to Professor Vinny Cahill, with whose able guidance I could have worked out this thesis. He has offered me valuable ideas, suggestions and criticisms with his profound knowledge in computer science and rich research experience. I am also very grateful to Dr. Mikael Asplund, whose patient and meticulous guidance and invaluable suggestions are indispensable to the completion of this thesis.

Thanks are also due to my friends, Meng Leng Sin and Xue Zhou who gave me constant and unconditional support and company. I can not imagine how my life would have been in the past four years without you. Most importantly, I would like to thank my dad and mom for their support all the way from the very beginning of my PhD study. They are not always around me, but they are always in my heart.

Shu Zhang

Trinity College Dublin

April 2013

Abstract

The need to reduce the number of fatalities due to road traffic accidents as well as to improve the comfort and efficiency of travel has motivated the vision of vehicular ad hoc networks. Vehicles are envisaged to be able to communicate with each other to avoid imminent danger by exchanging safety messages, which are often safety-critical and time-sensitive. The success or failure of delivering such messages in a reliable and time-bounded manner is the key to the correct operation of safety applications and consequently the well-being of passengers.

The unique characteristics of vehicular ad hoc networks, namely high mobility, large and unbounded scale, potentially adverse channel conditions, and the diversity of network density pose huge challenges in achieving real-time communication in these networks. Previous medium access control protocols, whether contention-based or reservation-based, are best-effort protocols in nature, and cannot provide reliable and guaranteed medium access. A real-time medium access control protocol that can adapt to the fast-changing network topology is required to support the stringent communication requirements of those safety-critical applications in vehicular networks.

To circumvent the difficulties arising from the network dynamics, this thesis proposes a novel design paradigm, pre-scheduling, in order to allocate wireless resources in a proactive manner. Leveraging prediction algorithms, pre-scheduling identifies relevant participants in the network and negotiates resource allocations in the context of future network topology, which paves the way for designing a deterministic allocation algorithm with reservation guarantees. In addition, a new medium access control protocol, Real-time Reservation Protocol (RRP) has been devised which fully supports the pre-scheduling paradigm and achieves reliable and time-bounded medium access.

The contributions of this thesis are two-fold. First of all, the pre-scheduling concept is proposed which challenges the conventional paradigm that medium ac-

cess control regulates communication participants that are physically close to each other. In fact, to identify and negotiate transmission schedules, future neighbors are of more relevance in a fast-changing environment. Secondly, a new medium access control protocol, RRP is proposed which achieves deterministic and time-bounded resource allocation in dynamic networks. A node in the RRP protocol tracks, communicates and negotiates with future neighbors in the predicted network topology, and is guaranteed to be allocated a reliable transmission slot within a bounded time interval. In addition, a new application-level metric "staleness" is proposed to characterize perceived communication quality of service, and applied as a performance indicator in the evaluation study. Via extensive simulation experiments, the performance of RRP has been evaluated and compared with other benchmark protocols. The results show the feasibility of RRP in achieving reliable and time-bounded medium access in a vehicular environment.

Contents

List of Figures

Chapter 1

Introduction

This thesis presents a new medium access control protocol, real-time reservation protocol, to address the problem of providing real-time medium access control in vehicular ad hoc networks (VANETs).

A vehicular network is uniquely characterized by its fast-evolving network topology, which is resulted from the high node mobility of the network. The intuition to tackle the problem in an environment with a high level of volatility is to allocate resources in a proactive way, i.e., pre-scheduling. Leveraging this concept, the proposed RRP protocol predicts vehicles' positions in the forthcoming future and estimates the relationships among vehicles' transmitters in terms of their potential interferences with each other, based on which medium resources, i.e., time slots, are allocated. Simulation-based evaluation study has shown that, given certain assumptions, the RRP protocol can provide reliable one-hop broadcast communication, as well as time-bounded access to the wireless medium, which might be of value to safety-critical applications that are envisaged in VANETs.

This introductory chapter provides relevant background information on the vehicular networks and their challenges, the problem that this thesis strives to solve, and the contributions of this thesis, namely the pre-scheduling concept and a real-time medium access control protocol. An overview of the thesis is provided at the end of this chapter.

Table 1.1: Prospective applications in VANETs

Application Type	Aim	Example
Public safety application	Increase road safety	Cooperative forward collision warning. Traffic signal violation warning.
Traffic management application	Improve traffic flow Reduce congestion and travel time	Enhanced route guidance and navigation. Green light optimal speed advisory. Lane merging assistance.
Information / entertainment application	Provide data / voice services for passengers	Internet access. Tolling. Voice or instant messaging.

1.1 The Vision of Vehicular Ad Hoc Networks

A VANET is a special class of mobile ad hoc wireless networks that have emerged in recent years, and are widely considered as one of the mobile ad hoc network's real-life applications Moustafa & Zhang [2009]. Such a vehicular network is spontaneously formed among moving vehicles and stationary infrastructures that are equipped with wireless devices, enabling communications among vehicles themselves and those roadside equipments.

In the last decade, significant research efforts have been made in the field of VANETs. The major goals of these works are to increase transportation efficiency, reduce its impact on the environment, and, most importantly, to increase road safety. Several factors have stimulated the development in VANETs, e.g., the wide adoption of wireless local area networks (WLAN) and Global Positioning System (GPS) technology since the late 1990s; safety, comfort and environment concerns; rapidly growing numbers of vehicles; and the commitment to allocate wireless spectrum for vehicular wireless communication (generally in the 5.8/5.9 GHz band in the US and Europe) Hartenstein & Laberteaux [2008]. Prospective applications can be roughly categorized as: public safety, traffic management and information / entertainment applications. Some examples in each category are presented in Table 1.1.

Among the enormous number of applications, the Vehicle Safety Communications consortium VCS [2006] has identified 8 high-priority applications, with near-term and mid-term potential benefit. A brief description of these applications and a preliminary communication requirements analysis are summarized in Table 1.2.

Table 1.2: Preliminary application communication scenario requirements

Application	Description	Comm. Type	Trans. Mode	Min. Freq (Hz)	Lat. (msec)	Max. Comm Range (m)
Traffic Signal Violation Warning	To warn the driver to stop at the legally prescribed location if the traffic signal indicates a stop and it is predicted that the driver will be in violation	I2V One-way Point-to-multi point	Periodic	10	100	250
Curve Speed Warning	To aid the driver in negotiating curves at appropriate speeds	I2V One-way Point-to-multi point	Periodic	1	1000	200
Emergency Electronic Brake Lights	When a vehicle brakes hard, the Emergency Electronic Brake light application sends a message to other vehicles following behind	V2V One-way Point-to-multi point	Event-driven	10	100	300
Pre-Crash Sensing	Pre-crash sensing can be used to prepare for imminent, unavoidable collisions	V2V Two-way Point-to-point	Event-driven	50	20	50
Cooperative Forward Collision Warning	To aid the driver in avoiding or mitigating collisions with the rear-end of vehicles in the forward path of travel through driver notification or warning of the impending collision	V2V One-way Point-to-multipoint	Periodic	10	100	150
Left Turn Assistant	Provides information to drivers about oncoming traffic to help them make a left turn at a signalized intersection without a phasing left turn arrow	V2I and I2V One-way Point-to-multipoint	Periodic	10	100	300
Lane Change Warning	Provides a warning to the driver if an intended lane change may cause a crash with a nearby vehicle	V2V One-way Point-to-multipoint	Periodic	10	100	150
Stop Sign Movement Assistance	Provides a warning to a vehicle that is about to cross through an intersection after having stopped at a stop sign	V2I and I2V One-way Point-to-multipoint	Periodic	10	100	300

1.1.1 Unique Properties in VANETs

The vehicular network is a very unique type of mobile ad hoc network, with a number of challenging characteristics.

High mobility: Vehicles in VANETs move with relatively high speed, which is a distinctive property compared to other types of generic mobile ad hoc networks. The environment in which a VANET operates is very dynamic and include various scenarios. The relative speed of nodes could range from 0 to as high as 300 km/h on the highway. The implication of such high mobility is that the communication window in which two vehicles can talk could be as small as a few seconds. To establish a connection and complete the data exchange within such a short interval is very challenging.

Large scale: A vehicular network can potentially span over the entire road network of a city, which includes effectively unlimited number of vehicles. The scalability of any communication protocol in such a virtually unlimited network becomes a serious issue. In addition, the large scale of a vehicular network poses difficulties on attempts to conduct any real-world evaluations, as such an experiment at any meaningful scale in the VANETs implies huge logistical challenges.

Various network density: Similar to mobility, the density of a vehicular network can be extremely diverse. A country road with sporadic traffic is a completely different scenario from a 10-lane highway packed with vehicles during the rush hour. From a communication perspective, the protocol that works in a VANET is required to deliver the same level of QoS in both sparse and dense networks. Such a requirement calls for special attention in the design process to include mechanisms such as redundancy to ensure graceful degradation when the resources available in the network are reduced.

Adverse wireless channel: In terms of wireless communication, the vehicular environment is far from ideal. Multi-path fading caused by the environment such as the buildings, trees and vehicles themselves, as well as the Doppler Effect which stems from the high relative speed between vehicles all deteriorate the reliability of the wireless transmission. In addition, line-of-sight communication can not always be assumed in an urban scenario.

Nevertheless, there are a number of favorable conditions presented in VANETs

that could be exploited. First of all, power is generally not a constraint in VANETs, therefore there is no need to consider mechanisms to save energy, which is fundamentally different from wireless sensor networks. In addition, on-board computers are powerful in terms of their computing capabilities and available memory, which means that complex algorithms can be hosted and executed with sufficient performance in VANETs. Finally, vehicles in VANETs tend to have a very predictable trajectory that are usually limited to roads. Algorithms to predict the vehicle movement may be feasible and yield beneficial results.

1.1.2 VANETs vs. Cellular Solutions

With the proliferation of smart phones and the dazzling mobile applications that have emerged in recent years, people wonder whether the idea of VANETs is still relevant. Is it really necessary to form a distributed ad hoc network of vehicles, instead of simply connecting them via a base station just like smart phones?

Essentially, it is a traditional debate about the pros and cons of centralized versus distributed solutions. For starters, we can counter the argument with the single-point failure of the centralized solution and that the base station may be the bottle neck of the system. Furthermore, in the context of a safety application in VANETs, the distributed solution has unique merits as presented in the following.

The nature of safety applications in vehicular networks is to establish some level of mutual awareness with one's neighbors by periodically broadcasting localization information. Considering the number of messages generated per vehicle and the number of vehicles in a local area, there are potentially a huge amount of data, which only has local significance, being generated in the communication channel of a vehicular network. Therefore, to transmit such a huge amount of data away from its area of relevance to the base station, and transmit back to the relevant area is economically and technically dubious. In addition, other issues are of concern, such as the fact that the service provider of the base station is sensitive to the associated cost of such operations, the delay incurred during the long relay as well as the continuity of the service provided by the base station.

This thesis is based on the assumption that keeping communication local is a more appropriate approach for the envisaged safety applications in VANETs.

However, more research and experiments are needed to settle the dust on which technology is better suited for VANETs, and these two approaches are not necessarily mutually exclusive. For example, local communication is more suitable for safety applications as most information is generated and consumed in a local area, while wider range communication is needed for non-safety applications as most of them require the ability to access the Internet.

1.1.3 Summary

In this section, a brief introduction to the vehicular networks is presented, including the main objective, the prospective applications, properties and challenges of such a network. In addition, we discussed the necessity of the vehicular networks by a comparison with cellular-based solutions.

1.2 Problems Addressed by this Thesis

In the envisaged vehicular networks, severe accidents and casualties could potentially be avoided, provided that those safety applications are designed and operated properly. However, to achieve this target in a vehicular environment is a challenging task due to adverse characteristics of such a network and the stringent communication requirements of safety applications. For example, the proposed pre-crash sensing application requires communication latency to be less than 20ms VCS [2006].

For safety-critical applications that require reliable and timely message delivery, e.g., pre-crash sensing and forward collision warning, a real-time communication protocol is a prerequisite. As in such a system, the consequence of failing to meet the deadlines of message could be catastrophic. For example, a collision warning message that is lost during transmission could potentially mean the difference between life and death for the passengers on board.

Consequently, the value that the communication system offers is not just the ability to provide a communication channel in which messages are sent and received, but more importantly, the QoS of such a communication channel which is often characterized by, but not limited to, its ability to deliver messages in a

reliable and timely manner.

The main problem that this thesis addresses is therefore to design such a communication channel, or more specifically, a medium access control protocol, that supports reliable and time-bounded medium access in vehicular networks.

1.2.1 Scope and Assumptions of the Thesis

The scope of this thesis is therefore mainly focused on the design of a protocol that resides at the medium access control level in the OSI reference model Zimmermann [1980]. The proposed MAC protocol assumes that user messages arrive from layers above the MAC layer, and are delivered to the lower physical layer when the MAC protocol deems appropriate. However, the design of the MAC protocol incorporates techniques and algorithms that conventionally do not belong to the MAC layer, or undefined in the OSI reference model. For example, the proposed MAC protocol utilizes message forwarding algorithms, which is often categorized in a routing layer, or the modeling of wireless propagation process, or the problem of distributed resource scheduling.

In this thesis, a number of assumptions are made in the protocol design. Firstly, it is assumed that vehicles in VANETs are equipped with positioning devices, which provide up-to-date location information as well as speed and bearing information. In addition, all vehicles are equipped with wireless communication devices which are tuned to the same communication frequency, and the computing capability and available memory of each vehicle is sufficient to execute algorithms that are proposed in the thesis. Furthermore, clocks are synchronized for all vehicles in the network, and such synchronization information is provided by the positioning devices such as the GPS. Finally, the available bandwidth of the communication channel is not unlimited but is sufficient to support medium sharing among a reasonable number of vehicles in a local area.

1.2.2 Overview of Related Works

The topic of medium access control has been extensively studies in the literature. However, the study on the problem of achieving reliable and timely medium access in the context of vehicular networks is relative insufficient. MAC protocols that

are proposed in the literature can be generally divided into contention-based and reservation-based protocols. Nodes using the former approach contend with each other for the right to access the medium, which may result in possible transmission collisions and degraded reliability, especially in saturated channel conditions. On the other hand, reservation-based protocols provides reliable transmission in various load conditions, because medium access needs to be reserved in a collision-free way in advance. However, the reservation mechanism in these protocols may experience difficulties when the network topology constantly changes, such as the case in VANETs, which inevitably reduces the reliability of transmissions of these protocols.

In addition, the topic of reliable one-hop broadcast is an open issue in VANETs. Conventional mechanisms that improve the reliability of wireless communications such as Request-To-Send (RTS) and Clear-To-Send (CTS) messages 802.11 [1997], or busy tones Haas & Deng [2002] are not applicable in a broadcast settings.

Apart from the reliability issue in the related works, the topic of real-time medium access is rarely discussed in the VANET community. In addition, it is not widely acknowledged in the research community that a non-safety-critical communication system is fundamentally different from a safety-critical communication system, which may require a completely new design which focuses more on the QoS aspect of the system.

1.2.3 Summary

In this section, the main problems that this thesis aims to address are presented, which are to achieve reliable and time-bounded medium access. In addition, the scope of this thesis, the assumptions made in the design, and an overview of the related works are presented.

1.3 Contribution of this Thesis

This thesis addresses the problem of real-time medium access control in vehicular networks. The contributions of this thesis are two-fold. First of all, a new design scheme termed *pre-scheduling* is proposed which aims to allocate resources in the

context of future networks. Another contribution of this thesis is a new MAC protocol, RRP, which provides reliable and time-bounded medium access in the envisioned vehicular networks.

1.3.1 Pre-scheduling

By exploiting the predictability of vehicles in VANETs, the pre-scheduling scheme allocates resources among vehicles that are predicted to be in close proximity with each other, i.e., neighbors, in the estimated future networks. Consequently, when *prospective* neighbors move towards each other as time elapses and become *actual* neighbors, the process of allocating resources has completed already. By using such "prepare in advance" mechanism, the issues that arises from the dynamics of network topologies can be considerably alleviated, if not complete resolved.

1.3.2 Real-time Reservation Protocol

The proposed RRP protocol adopts a reservation-based approach, and is fully distributed. The architecture of the RRP protocol is composed of two layers, i.e., the lower *virtual cluster layer* and the upper *real-time scheduling layer.*

The virtual cluster layer identifies prospective neighboring nodes of the current node in the future network, and reports a list of such prospective neighbors to the real-time scheduling layer. The process of identifying future neighbors incorporates mechanisms such as mobility prediction, multi-hop messages forwarding, as well as topology-based interference analysis, which is a process that determines whether a candidate node is a neighbor or not to the current node in a given topology.

The real-time scheduling layer allocates time slots by exchanging negotiation messages between the current node and those that are identified as future neighbors. Priority numbers that pertain to each individual slot are used to determine the outcome of a slot contention. A slot contention, or slot reservation, is successful if and only if all other identified neighbors either give up, or have inferior priorities regarding this slot. As a result, exclusive access to a specific slot is achieved between the current node and its neighbors.

In addition to the collision-free slot contention mechanism, the real-time scheduling layer achieves time-bounded medium access. The intuition of such a reservation guarantee stems from the observation that if the maximum amount of resources that each neighbor can obtain is limited, then it is always possible for a node to obtain its fair share, provided that the total amount of resources is sufficient for all nodes. Following this intuition, the maximum interval between two reserved slots in RRP is bounded, which effectively results in time-bounded medium access delay.

1.3.3 Assumptions and Trade-offs

The property of reliable and time-bounded medium access that RRP provides depend on a number of assumptions and restrictions. First of all, the correct execution of the protocol relies on accuracy of the neighbor identification algorithm, which is a function of the accuracy of the message dissemination algorithm, node mobility prediction, and the interference estimation algorithm. In addition, provided that prospective neighbors can be accurately identified, the slot reservation guarantees that RRP claims also depends on the reliable and timely delivery of the negotiation messages between the current node and its prospective neighbors. If a decision has been made by another neighbor regarding a specific slot, but the negotiation message is lost or suffers a lengthy delay, the assumptions of sufficient total resources may not hold, and so is the reservation guarantee.

In the design and evaluation of the RRP protocol, the trade-off between the guaranteed properties and the efficiency of the protocol is often observed. For example, increasing the number of redundant messages that are forwarded may increase the reliability of the protocol by improving the accuracy of neighbor identification, but decreases the overall efficiency. In addition, by reducing the number of identified neighbors, the neighbor identification algorithm can increase the probability of obtaining a slot, but it also reduces the reliability of transmission, as collisions are more likely to occur among nodes that should have been identified as neighbors.

1.3.4 Evaluation

The performance of RRP has been evaluated and compared with 802.11p and RR-ALOHA protocol in a simulation-based study. The RR-ALOHA protocol is chosen because it is a well-studied slot allocation protocol which resembles RRP in a number of aspects. A new evaluation metric, staleness, is proposed in order to characterize the perceived communication QoS from an application's point of view. The measurement of staleness incorporates both transmission reliability and medium access delay in a synthesized manner, which is discussed in detail in Chapter 5.

Other metrics are also evaluated in the study, which includes, but not limited to, packet delivery rate, medium access delay, and throughput. The evaluation results show that the RRP protocol achieves better performance in most simulated scenarios compared to other protocols, in terms of those aforementioned metrics. In addition, the results demonstrated that reliable transmission and time-bounded medium access delay is feasible in vehicular networks, which may be of great value for envisaged safety-critical applications.

1.3.5 Summary

In this section, the contributions of this thesis is discussed, which are the introduction of a new design scheme termed pre-scheduling, as well as a MAC protocol that achieves reliable and time-bounded medium access. In addition, assumptions, restrictions and trade-offs of this protocol are presented. By a simulation-based evaluation, the performance of RRP is studied and compared with other protocols, which demonstrates the feasibility and usefulness of the proposed protocol.

1.4 Road Map of the Thesis

The thesis is structured as follows:
Chapter 2 presents the related works of MAC in wireless networks.
Chapter 3 presents the design of the RRP protocol.
Chapter 4 presents the implementation details of the RRP protocol.

Chapter 5 presents the evaluation of the RRP protocol.
Chapter 6 summarizes the thesis and discusses future work.

Chapter 2

Related Works

The problem of medium access control has been extensively studied in the past. However, to directly apply those protocols from previous literature to the vehicular environment may not yield promising results as vehicular networks possess unique characteristics and challenges. In this chapter, related works are presented regarding the problem of medium access control in both generic wireless network and vehicular networks, in the hope for discovering their intrinsic properties and investigate their suitability in vehicular environments.

2.1 Basic Protocols

This section focuses on a number of MAC protocols in the literature which laid the ground work in the area of medium access control in wireless communications. These pioneering protocols identify fundamental issues and challenges facing a MAC protocol and propose various solutions to solve or mitigate them. By investigating and comparing these basic protocols, an in-depth understanding can be obtained regarding a MAC protocol's goals, means, constraints and what's achievable. It is a crucial process for a further review on other more sophisticated MAC protocols and to design a MAC protocol of our own. In the following sections, these protocols are reviewed in a chronological manner, with interweaved examples of typical MAC protocol problems and their respective solutions.

In wireless communications, a radio transceiver can neither send and receive

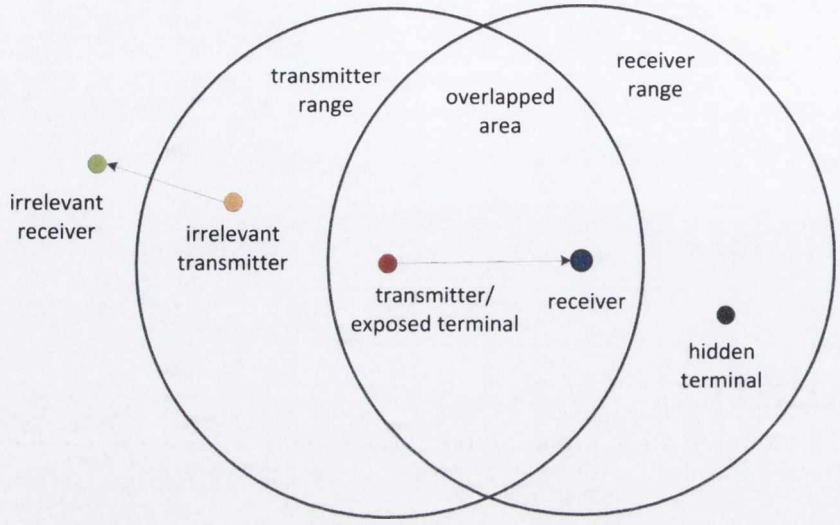

Figure 2.1: Hidden terminal and exposed terminal problem

at the same time, nor receive concurrent radio signals from more than one transmitters. Such a hardware constraint defines the most fundamental objective of any medium access control protocol, i.e., to avoid simultaneous arrivals of signals at a receiver, otherwise, a *collision* will occur. The most primitive MAC protocol in wireless communications is the ALOHA protocol Abramson [1970], in which nodes transmit whenever they desire. Due to high probabilities of collisions, the maximum throughput of ALOHA can only achieve 18.4% Tanenbaum [2002].

2.1.1 CSMA

Kleinrock & Tobagi [1975] proposed the Carrier Sensing Multiple Access (CSMA) protocol in which nodes verify an idle medium before transmitting. Carrier sensing reduces the chance of two or more nodes transmitting at the same time, thus significantly reduces the collision probability. However, due to the fact that carrier sensing occurs at the sender side as oppose to the receiver side, the sensed medium status is partial and is subject to error, which leads to the well-known hidden terminal and exposed terminal problem as depicted in Figure 2.1.

In Figure 2.1, node A's transmission to node B is subject to hidden terminal node C's interference, because node A cannot sense the on-going transmission that

occurred in a certain part of receiver B's neighborhood, i.e. the "receiver-only area". On the other hand, instead of sensing the medium near receiver node E, node F senses node A's transmission and is therefore unable to start a legitimate data transmission to node E. The consequence of these two problems are that the hidden terminal problem reduces the communication reliability while the exposed terminal problem reduces the efficiency of the wireless communication.

2.1.2 MACA

The Multiple Access Collision Avoidance (MACA) protocol was proposed to address the hidden and exposed problem Karn [1990]. Carrier sensing is not used in MACA, instead, two short signaling packets: Request-To-Send (RTS) and Clear-To-Send (CTS) were introduced. Before transmitting data packets, the sender sends a RTS message to the receiver which replies with a CTS message. Nodes that hear the CTS messages need to postpone their transmission.

After the RTS/CTS exchange, hidden terminal nodes receives the CTS message and back-off, which eliminates the hidden terminal problem. On the other hand, nodes hear only RTS without CTS know that they are exposed terminals and are thus allowed to access the medium, which eliminates the exposed terminal problem. However, the success exchange of RTS and CTS message cannot guarantee a collision-free data transmission, if the wireless propagation time is taken into account Garcia-Luna-Aceves & Fullmer [1999]. The authors argues that the transmission time of the RTS/CTS message needs to be twice the length of the maximum propagation delay in the network. In addition, by using the RTS and CTS message, the MACA protocol brings two other new problems, i.e., the RTS/CTS collision, and receiving/sending channel blockade.

In MACA, RTS and CTS messages are designed to protect the data transmission, but the RTS and CTS packets themselves are not protected. In fact, RTS/CTS messages are transmitted in an ALOHA fashion, which make them especially vulnerable to both 1-hop and 2-hop collisions. It has been reported that the performance of the MACA protocol degenerates to ALOHA if the hidden terminals are considered Haas & Deng [2002].

Another issue of the usage of RTS and CTS packets is that they may jam

the sending or receiving channels which are critical to establish another concurrent data transmission. This problem is often neglected due to its subtlety, but may significantly reduce the overall throughput. The sending channel blockade problem stems from the fact that the nodes that hear CTS are prohibited from sending during transmission. Nevertheless, they should not be prohibited from receiving in the meantime, but are unable to do so because their CTS message to grant another data session cannot be sent out.

For example in Figure 2.1, the data transmission from node D to C is legitimate with concurrent transmission between node A and B. However, when node D's RTS is received by node C, C cannot reply with a CTS to D due to the blocked sending channel after receiving node B's CTS. Consequently, the transmission from D to C cannot proceed until the transmission ends between node A and B. In addition, after several none-replied RTS messages to node C, node D's contention window grows exponentially, which put node D in an extremely inferior position in subsequent contentions.

The receiving channel blockade problem occurs when a node is in the shadow of an on-going transmission, and therefore cannot receive any packet including a replied CTS message to grant another concurrent data transmission. For example, in Figure 2.1, the data transmission from nod F to node E is also legitimate with concurrent transmission from node A to B. However, after node F sends out RTS to E, node E's CTS cannot be received by F due to the on-going transmission from node A. Consequently, the transmission from node F to E cannot proceed as well. In addition, same with the sending channel blockade problem, the contention window of node F also expands exponentially after several failed RTS attempt, and eventually becomes very uncompetitive in future contentions.

2.1.3 MACAW

The MACAW protocol Bharghavan *et al.* [1994] is an extension to the MACA protocol, with standard RTS/CTS mechanism and without the usage of carrier sensing. Due to the fact that carrier sensing is not used in MACAW, the RTS and CTS messages are not protected as in MACA. In MACAW, two additional control packets Data-Sending (DS) and Request-for-Request-to-Send (RRTS) are

introduced to addresses the sending and receiving channel blockade problem. These two additional packets however, do not solve the sending and receiving channel blockade problem but only mitigate their implication as discussed in the following.

In the receiving channel blockade scenario in Figure 2.1, node F continuously attempts to send RTS messages to node E during the data transmission period between node A and B, without knowing that it can never receive a CTS from node E. The DS message is introduced in this case to inform node F to stop transmitting these RTS messages to node E, which is not only meaningless but also self-destructive. Specifically, once node A receives a CTS from node B and is about to initiate data transmission, A send DS to node F containing the duration of its transmission, in which node F remain silent, and avoid sending any RTS message. It is worth noting that node F is still blocked from sending messages to node E, but its contention window will stop growing because of those failed RTS messages.

In the sending channel blockade problem as depicted in Figure 2.1, node D's contention window continuously grows, which makes it less competitive in subsequent contentions. In light of this, the RRTS message is introduce to let node C contend on behalf of node D. Specifically, after the transmission completes between node A and node B, node C broadcasts a RRTS packet. Any node that receives RRTS back-off immediately, which clears the way for node D and C to start their transmission.

2.1.4 802.11

The 802.11 protocol 802.11 [1997] which adopts the "Carrier Sensing Multiple Access/Collision Avoidance" (CSMA/CA) mechanism was introduced in 1997, which incorporates both RTS/CTS and carrier sensing mechanisms. In 802.11, due to the usage of RTS and CTS message, hidden terminal and exposed terminal problems are eliminated. However, the sending channel and receiving channel blockade problem still exist.

Nodes in 802.11 need to sense the channel before transmitting any RTS packet, which reduces the probability of RTS collisions and enhances its usability. How-

ever, due to the issues with the carrier sensing mechanism that the sensing only occurs at the sender side, RTS and CTS messages in 802.11 are still susceptible to 2-hop collisions.

2.1.5 DBTMA

The problem of the sending and receiving channel blockade problem reveals the dilemma between the blocking nature of a half-duplex radio and the need for a full-duplex channel to initiate a data transmission. In light of this, the Dual Busy Tone Multiple Access (DBTMA) protocol Haas & Deng [2002] utilizes two out-band control channel, Busy Tone Transmit (BTt) and Busy Tone Receive (BTr), in addition to the conventional data channel to solve these issues.

In DBTMA, the RTS sender turns on the BTt signal while transmitting the RTS, and turns it off when transmission is completed. After receiving the RTS packet, the receiver turns on the BTr signal to acknowledge the RTS sender. The BTr serves as a CTS message to back-off the receiver's 1-hop neighbors, thus the CTS message is not used in DBTMA. Nodes that attempt to initiate a RTS session need to sense idle channels of both BTt and BTr.

The DBTMA protocol solves the hidden terminal and exposed terminal problem by using carrier sensing, RTS message and busy tones. In addition, it also solves the sending and receiving channel blockade problem. In Figure 2.1 for example, node C whose sending channel is blocked by node B's CTS replies with a receiver busy tone to node D, and the data transmission may commence between node D and C. Likewise, node F can receive the receiver busy tone from node E, despite of node A's concurrent data transmission.

In addition, the use of busy tone in DBTMA eliminates the CTS collision problem, because an overlapped busy tone still indicates that the receiver is ready. However, the busy tone cannot prevent RTS from 2-hop collisions.

The use of busy tone has been proposes in other literature, such as the RI-BTMA Wu & Li [1987], in which only the receiver busy tone is used. In fact, the use of the sender busy tone in DBTMA is due to the fact that no carrier sensing is used in the data channel, where RTS packets are sent. If no carrier sensing is used in neither sender busy tone channel nor the data channel, the RTS messages

(also known as the "preamble" in RI-BTMA) are completely unprotected. Consequently, carrier sensing which protects the RTS message needs to be adopted in at least one of these two channels: sender busy tone channel or the data channel.

2.1.6 Summary and Analysis

In this section, a number of basic MAC protocols are introduced and analyzed, which address the fundamental problem of medium access control in wireless communications. ALOHA is introduced as the starting point, in which nodes transmit packets arbitrarily. The drawback of ALOHA is that packets may suffer from severe collisions. Subsequently, carrier sensing mechanism (CSMA) has been introduced to eliminate contentions that occur in the sender's vicinity. However, hidden and exposed terminal problem arise due to incorrect location where medium sensing is conducted.

In light of this, the pure RTS/CTS mechanism is proposed in MACA and MACAW. In these approaches, the hidden terminal and exposed terminal problem are resolved, but the RTS/CTS packets themselves are unprotected and are prone to collisions. Subsequently, the 802.11 protocol that combines the pure RTS/CTS mechanism and the carrier sensing approach are introduced. Although the blocking nature of RTS/CTS mechanism may cause starvation problem, which is addressed in MACAW and DBTMA, 802.11 protocol is generally effective in protecting transmissions from collisions, which is why it has been standardized and widely deployed.

The problems that are associated with theses protocols are summarized in Table 2.1, and compared in Table 2.2.

There are a number of issues that are relevant but not mentioned in the above discussions. For example, all aforementioned protocols and mechanisms are discussed in the context of unicast transmissions. These mechanisms may not be applicable in broadcast scenarios. For example, it is difficult to utilize the RTS/CTS in a broadcast scenario, because the returning CTS will collide at the RTS sender. In theory, a broadcast can be divided into a number of sequential unicast transmissions, in which RTS/CTS mechanism can be applied Tang & Gerla [2001]. However, the large delay incurred during the process to

Table 2.1: Possible issues of a unicast transmission in wireless communication

	Description	**Result**	**Data flow**	**Metrics affected**	**Possible solution**
Hidden terminal	C cannot sense the busy medium at B	Collision at receiver B	A → B	Correctness	RTS/CTS
Exposed terminal	F senses busy medium from A	F postpone talking to E	F → E	Efficiency	RTS/CTS
Sending channel block	C receives D's RTS but cannot reply CTS due to B's CTS embargo	1. D cannot talk to C 2. D's CW expanded (less competitive)	D → C	Efficiency Fairness	a. busy tone b. additional control packet (mitigation only)
Receiving channel block	F sends RTS to E, but E's CTS destroyed by A's transmission	1. F cannot talk to E 2. F's CW expanded (less competitive)	F → E	Efficiency Fairness	Ditto
RTS collision	RTS collide with 1 or 2-hop neighbors	Data transmission cannot proceed	A → B	Efficiency	carrier sensing (protect only 1-hop)
CTS collision	CTS collide with 1 or 2-hop neighbors	Data transmission cannot proceed	B → A	Efficiency	a. carrier sensing (protect only 1-hop) b. busy tone (protect both 1 and 2 hop)

Table 2.2: Comparison of basic MAC protocols

	CSMA (1975)	**MACA (1990)**	**MACAW (1994)**	**802.11 (1997)**	**DBTMA (2002)**
Basic operation	Carrier sensing	RTS/CTS	RTS/CTS + extra control packet	Carrier sensing + RTS/CTS	Carrier sensing + RTS + busy tone
Hidden terminal	Not solved	Solved	Solved	Solved	Solved
Exposed terminal	Not solved	Solved	Solved	Solved	Solved
Sending channel block	Does not exist	Exist	Exist, with mitigated consequence	Exist	Solved
Receiving channel block	Does not exist	Exist	Exist, with mitigated consequence	Exist	Solved
RTS collision	N/A	1-hop and 2-hop collision (severe)	1-hop and 2-hop collision (severe)	2-hop collision only	2-hop collision only
CTS collision	N/A	1-hop and 2-hop collision (severe)	1-hop and 2-hop collision (severe)	2-hop collision only	no collision

communicate with all neighbors for a single broadcast is a major concern.

Another issue of using RTS and CTS is the "long range interference" problem Xu *et al.* [2002]. Nodes that cannot successfully decode a CTS message from the data receiver may still interfere with the receiver due to long range interference. The authors therefore questions the effectiveness of the RTS/CTS mechanism.

2.2 802.11 Series Protocols

2.2.1 802.11 DCF

The 802.11 MAC protocol 802.11 [1997] is based on two coordination functions, namely the Distributed Coordination Function (DCF) and the Point Coordination Function (PCF) to determining when a station operating within a Basic Service Set (BSS) is allowed to transmit and receive frames via the wireless medium. The DCF is mandatory in 802.11 and is based on the CSMA/CA (Carrier Sensing Multiple Access with Collision Avoidance) mechanism, while the PCF is optional

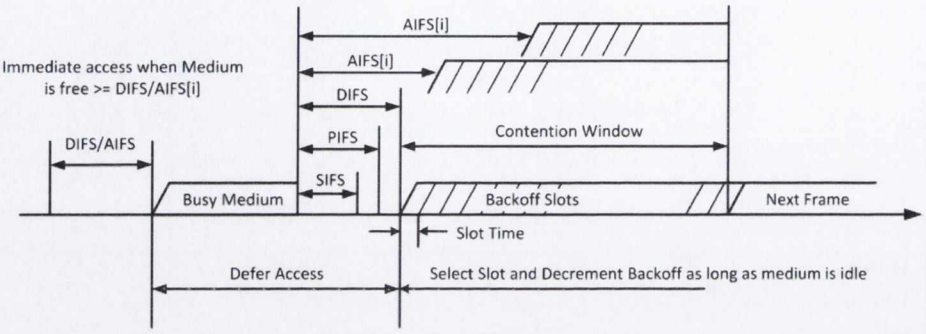

Figure 2.2: 802.11 DCF timing relationships

and is based on centralized poll-and-respond mechanism. The DCF is briefly introduced in the following as it is the basis for other 802.11 series protocols.

The transmission of a frame in DCF is subject to the CSMA/CA medium access mechanism, which is based on a local assessment of the channel status. When a frame arrives at the head of the transmission queue, if the channel is sensed idle at this point and during the following DIFS (DCF Inter-frame Space) time interval, the station can proceed with the transmission. On the contrary, if the channel is sensed busy, the station waits until the medium becomes idle then defers for a DIFS. If the medium is sensed idle during DIFS deference, the station starts a back-off procedure, which is designed to avoid collisions with stations that may be also waiting for the medium to be idle.

A back-off timer is generated randomly from a contention window within [0, CW], where CW is the size of the contention window. The back-off time is decremented by one when the medium is sensed idle in a slot. The back-off time is frozen if the medium is sensed busy, and is resumed after the medium has been sensed idle for another DIFS interval. If the back-off time reaches zero, the station is allowed to access the medium. If two or more stations finish the back-off procedures at the same time, they may transmit simultaneous, and a collision may occur. The timing of the DCF channel access is depicted in Figure 2.2.

In a unicast transmission, for each successfully received frame, the receiving station immediately replies with an acknowledgment frame (ACK) after a Short Inter-frame Space (SIFS), which is shorter than DIFS. Due to SIFS's shorter length compared to DIFS, the ACK packet has higher priority over the competing

transmissions from other stations. If the ACK frame is not received within a timeout period, the frame is retransmitted by the sender by entering another back-off procedure again. For each unsuccessful transmission, the CW is doubled until reaching a maximum value CW_{max}, which reduces collision probability when multiple stations are competing for the medium.

For each successful transmission, the CW is set to an initial value CW_{min} for the station. In addition, the station perform another DIFS deference and a random back-off, even if there is no additional frame to send. Such procedure is often referred to as "post" back-off, which ensures that the transmitting station will not have priority over any other waiting stations.

In DCF, RTS/CTS messages are used to address the hidden terminal and exposed terminal problem, as we discussed in Section 2.1.2. Any stations that receives the RTS or CTS message update their Network Allocation Vector (NAV) according to the duration field in the RTS and CTS message. Although the RTS/CTS mechanism significantly improves system performance, it is optional in DCF due to its large communication overhead, and is advised to be applied for only large data frames with size that exceeds the RTS-threshold.

It is worth mentioning that the MAC-level acknowledgment mechanism, i.e., the ACK message, as well as RTS/CTS mechanism are not applicable in a broadcast scenario. In addition, the contention window is not increased because the sending station cannot determine whether a broadcast transmission is successful or not, which further limits the adaptability of the DCF. In a vehicular network, non-unicast communication schemes such as broadcast and multi-cast are expected as primitive communication methods used by safety applications. Consequently, the usability and reliability of the 802.11 DCF in a VANET environment warrant further investigation.

2.2.2 802.11e

In 802.11 DCF, only best effort service is supported where all stations compete for the medium with the same priority. No differentiation mechanism is provided for applications with higher communication requirements, e.g., bandwidth, and end-to-end delay. In light of this, in the 802.11e protocol 802.11e [2005], a new

Table 2.3: Application priority to access category mappings

Application Priority	**Access Category (AC)**	**Designation (Informative)**
1	AC_BK	Background
2	AC_BK	Background
0	AC_BE	Best Effort
3	AC_BE	Best Effort
4	AC_VI	Video
5	AC_VI	Video
6	AC_VO	Voice
7	AC_VO	Voice

MAC function termed the hybrid coordination function (HCF) is proposed. HCF is composed of a contention-based medium access mechanism termed enhanced distributed channel access (EDCA), and a polling-based HCF controlled channel (HCCA).

EDCF aims to provided differentiated QoS by enhancing the contention-based DCF. Before entering the MAC layer, each packet is assigned with an application-specific priority category. At the MAC layer, EDCF implements four priority categories termed access categories (ACs), and each application priority category is mapped into a corresponding access category according to Table 2.3. Each AC maintains its own FIFO queue, which acts as an independent DCF contention entity with its own contention parameters, i.e., CW_{min}, CW_{max}, $AIFS$, and $TXOP_{limit}$ (which is the maximum time limit for a single transmission). The arbitrary IFS ($AIFS$) is a new type of IFS introduced in EDCF with length determined by:

$$AIFS[AC] = SIFS + AIFSN[AC] * slottime \tag{2.1}$$

where $AIFSN[AC]$ is called the arbitration IFS number, which is an integer greater than zero.

As depicted in Figure 2.3, four transmission queues are maintained in 802.11e, where each queue behaves as a single DCF entity. The purpose of using different contention parameters for different access categories is to reduce the medium access time for frames in high priority queues, i.e., realizing prioritized medium access. Note that in EDCF, when there is more than one AC queue that finishing

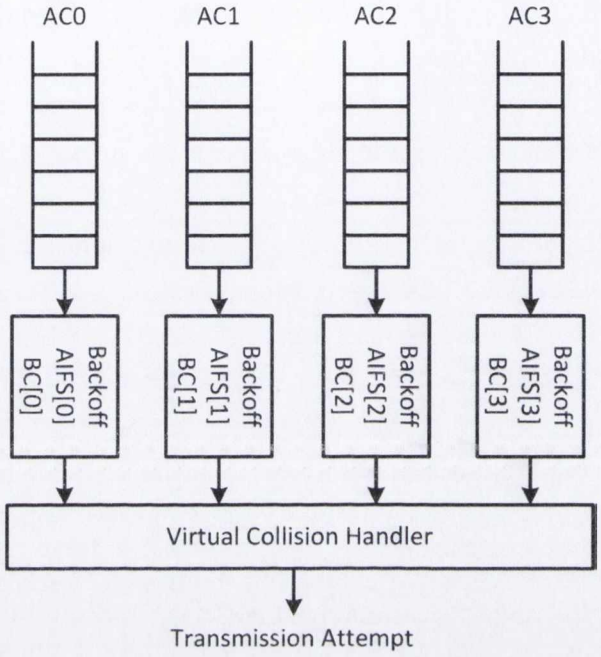

Figure 2.3: 802.11 EDCF virtual contention

the back-off at the same time, the highest priority AC is selected to transmit and other ACs perform a back-off as if experienced an actual collision. In other words, a frame in EDCS needs to compete internally before actually competing with frames from other stations.

2.2.3 802.11p and WAVE

Due to the enormous and worldwide success of the 802.11 standards, the IEEE aims to reuse this technology in the upcoming VANETs. With a few modifications, the new 802.11p protocol is expected to cater for new requirements in this new and challenging environment.

The IEEE WAVE (Wireless Access in Vehicular Environments) standard is currently under ratification. A WAVE system is a radio communication system intended to provide seamless, interoperable services to transportation. The standard is comprised of the following four components.

1609.1 [2006] - resource manager is a WAVE application that resides on a

Roadside Unit (RSU) or Onboard Unit (OBU). It is designed to "manage" OBU resources, such as memory, user interfaces and interfaces to other onboard equipments, on behalf of remote applications. OBU resources are abstracted and managed by a series of application independent commands (RM commands). The benefits of such abstraction are that OBU resources can be accessed in a consistent and interoperable manner for remote applications, and OBU is relieved from interpreting various application-specific operations.

1609.2 [2006] - specifies security services for the WAVE networking stack and for applications that are intended to run over that stack. Services include encryption using another party's public key and non-anonymous authentication.

1609.3 [2010] - networking services provide data delivery services between WAVE devices and management services to all layers. Data plane services include logical link control (LLC) for upper layers, IPv6, UDP and TCP support, and the WAVE Short Message protocol support (WSMP). Management plane services include: application registration, WBSS management, channel usage monitoring, and Received Channel Power Indicator (RCPI) polling.

1609.4 [2011] - multi-channel operation emphasizes the multi-channel coordination of the system. Such channel coordination interacts with 802.2 Logic Link Control (LLC) and 802.11p physical layer. Specific services provided by 1609.4 multi-channel operations includes: channel routing, user priority, channel coordination, and MSDU data transfer.

Among the above four components, WAVE 1609.3 (networking services) is the core of the entire WAVE architecture. As illustrated in Figure 2.4, there are two protocol stacks defined in the WAVE 1609.3 standard: IPv6 and the WSMP. WSMP allows applications to directly control physical layer parameters, e.g. channel number and transmit power, in order to minimize communication overhead. Another unique property of WAVE is the concept of WBSS (WAVE BSS), which is a lightweight version of BSS (Basic Service Set) in 802.11. Among others, the major modifications of WBSS are to remove authentication and reduce the number of handshakes for establishing connections in order to reduce communication overhead.

IEEE 802.11p [2010] is quite similar to the legacy 802.11 series. The 802.11p PHY layer adopts the OFDM modulation technologies from 802.11a with halved

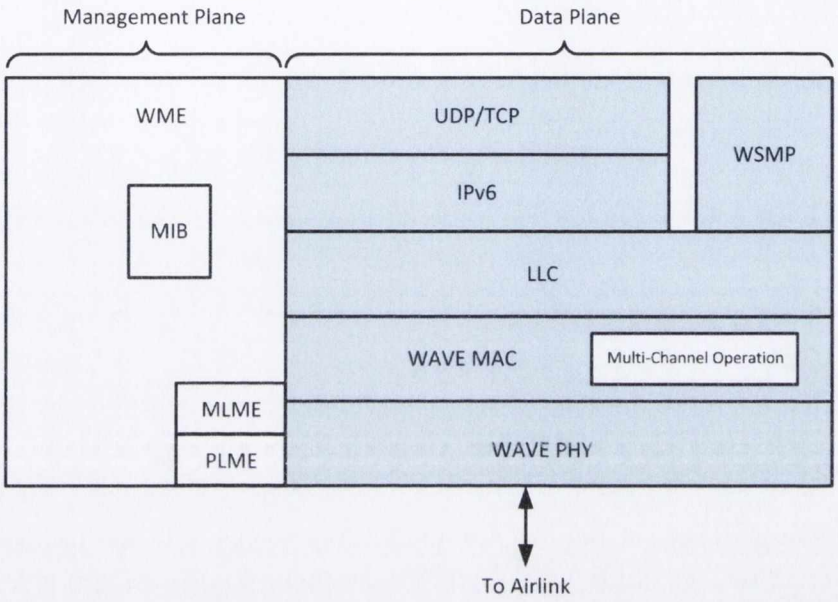

Figure 2.4: WAVE Protocol stack

channel bandwidth to increase reliability of radio transmission. The 802.11p MAC layer directly reuses the 802.11e MAC, which provisions for differentiated QoS. Therefore, safety critical messages in VANETs can be transmitted with higher priority. In terms of MAC and PHY layer, the comparison between 802.11p and 802.11a is summarized in Table 2.4 Moreno [2007].

Table 2.4: IEEE 802.11p and 802.11a PHY characteristics

Characteristic	**802.11p**	**802.11a**
Channel bandwidth	10 MHz	20 MHz
Data rates	3 to 27 Mbps	6 to 54 Mbps
Slot time	16 μs	9 μs
SIFS time	32 μs	16 μs
Channel switch time	$\leq$2048μs	N/A
Air propagation time	$<$ 4 μs	$\ll$ 1 μs
Preamble length	32 μs	20 μs
PLCP header length	8 μs	4 μs
CW_{min}	15	15
CW_{max}	1023	1023

2.2.4 Performance Analysis

The performance of the 802.11p protocol has been extensively evaluated in the literature, in both simulation and real-world environments Eichler [2007], Bai & Krishnan [2006]. Various metrics that characterize the communication quality have been examined in order to determine whether 802.11p meets the stringent requirements of vehicular environments. However, based on these studies, it is premature to conclude whether the 802.11p technology is the right choice for VANETs.

On one hand, both field evaluations and simulation-based studies confirm that 802.11p is able to provide a certain level of data delivery service in less competitive environments. For instance, the field evaluation reported in VCS [2006] shows that the performance of 802.11p is generally acceptable (with a high packet reception ratio) within 200 meters, which justifies 802.11p as a feasible technology in vehicular environments. In addition, compared with other types of MAC protocols, such as reservation-based protocols, the contention-based 802.11 series protocols require much less control overhead in terms of managing network topology or exchanging schedule tables.

On the other hand however, it has been massively criticized in the literature that the communication performance, including packet reception rate, channel access time and throughput drop significantly when the wireless channel becomes saturated Murray *et al.* [2008], Stibor *et al.* [2007]. As pointed out by Bilstrup *et al.* [2009], the contention-based CSMA/CA mechanism used in the 802.11 family is inherently flawed to provide predictable and reliable medium access. This could leads to degraded communication quality, such as unbounded delay or unbounded channel access time, which are particularly important for safety-critical applications.

In view of the large amount of evaluations conducted in the literature, it is reasonable to assume that there exists a tipping point of 802.11p. Below this point, the wireless channel is less competitive and the general performance of 802.11p is acceptable. Beyond this point however, the channel becomes more saturated and various performance metrics begin to deteriorate. If such a tipping point can be identified, then if the channel is heavily loaded, a non-contention-

based medium access strategy may be selected, and when channel load is light, 802.11p can still be used. The topic of performance evaluation of 802.11p is further discussed in Chapter 5.

2.3 Multi-channel Protocols

The use of multiple channels can substantially improve the performance of wireless networks Crichigno *et al.* [2008]. For example in 802.11p, seven data channels are available to provide differentiated data services, which increases the overall throughput. Nevertheless, the most challenging issue of multi-channel protocols is how to assign the channels efficiently while avoiding some of the pitfalls inherited in the multi-channel operations.

To illustrate the issues of the multi-channel protocols, a simple multi-channel protocol is presented here as an example. In the RDT protocol Shacham & King [1987], each node is equipped with one radio, which can be switched among multiple channels. In addition, each node is assigned with a "default" channel, and is aware of other node's default channel. If node A intends to talk to node B, node A switches to node B's default channel, and the data communication starts as usual.

There are a number of problems using this approach for a multi-channel protocol. These issues are mainly due to the fact that the receiver and the sender are not tuned to the same channel, and therefore cannot establish a correct data session. Specific problems are discussed in the following sections.

2.3.1 Issues of multi-channel MAC protocols

- Multi-channel Hidden Terminal Problem

The well-known hidden terminal problem can be alleviated by the use of RTS/CTS messages. However, in the multi-channel environment, a CTS message may not be received due to the fact that the intended CTS receiver is tuned to another channel. The absence of CTS message in this case is therefore called the "multi-channel hidden terminal" problem.

- Deafness problem

Similar to the multi-channel hidden terminal problem, the deafness problem is caused by the absent of RTS messages. In a multi-channel environment, a RTS message may not be received as the intended recipient is tuned to another channel. After a number of unsuccessful RTS trials, the RTS sender incorrectly concludes that the receiver is unreachable.

- Channel deadlock problem

Suppose that node A fails to communicate with one of its off-tuned neighbor B, and node B depends on another off-tuned node C, which in turn depends on node A. A circular dependency is formed among node A, B and C, which in this case, all become a deaf. Such a channel deadlock problem may significantly degrade system performance.

- Broadcast problem

Due to the fact that in multi-channel protocols, nodes may reside on different channels, a broadcast message sent on a specific channel can only reach a portion of the neighbors in the communication range. As broadcast is a primitive operation for many medium access control and routing protocols, the incapability of providing broadcast operation is a serious problem in multi-channel protocols.

2.3.2 Dedicated Control Channel Protocols

The principle of this type of multi-channel protocol is to use a dedicated channel to transmit control messages, e.g. RTS and CTS messages. Consequently, control messages are separated from the data messages. In the control channel, nodes negotiate which channels they are going used in the subsequent data transmissions. Later on, the data exchanges occur on the agreed data channels as depicted in Figure 2.5. The dedicated control channel approach is used in Jain *et al.* [2001] and Li *et al.* [2003].

The advantages of dedicated control channel approach are its simplicity to implement, and its ability to reuse 802.11 control messages. In addition, time synchronization is not required using this approach.

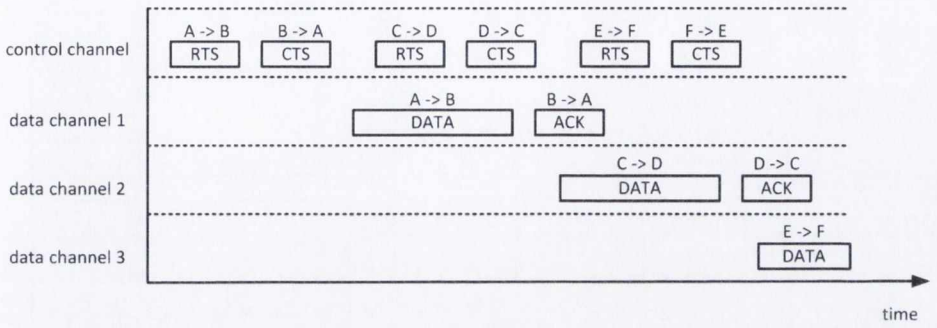

Figure 2.5: Procedure for dedicated control channel protocols

The disadvantages of dedicated control channel approach are the following: 1) multi-channel hidden terminal and deafness problem, 2) simultaneous assessment on the status of multiple channels is required, 3) it is necessary to assign an appropriate bandwidth between control channel and data channel (excessively large bandwidth in the control channel wastes resource, while excessively small bandwidth makes control channel a bottleneck), 4) broadcast is not supported.

2.3.3 Split Phase Protocols

In the dedicated control channel approach, more than one radio is needed to support simultaneous transmissions in both control and data channels. However, it is more cost effective to use one radio rather than multi-radio to support multi-channels. For a node with only one radio, the need for supporting both control and data channel can be realized using a time duplex method, i.e., by splitting the time into interleaving control phases and data phases. For splitting phase protocols, it is often assumed that the boundaries of the control and data phase are synchronized for all nodes in the network.

During the control phase, nodes negotiate the specific channels that are going to be used in the subsequent data phase. During the data phase, nodes exchanges messages on the previously agreed channels, as depicted in Figure 2.6.

So & Vaidya [2004] propose the Multi-channel MAC for Ad Hoc networks (MMAC) which set the control and data phase with fixed time intervals. During the control phase, which is termed ATIM window, channel reservation messages (ATIM, ATIM ACK and ATIM RES) are exchanged in a three-way hand shake

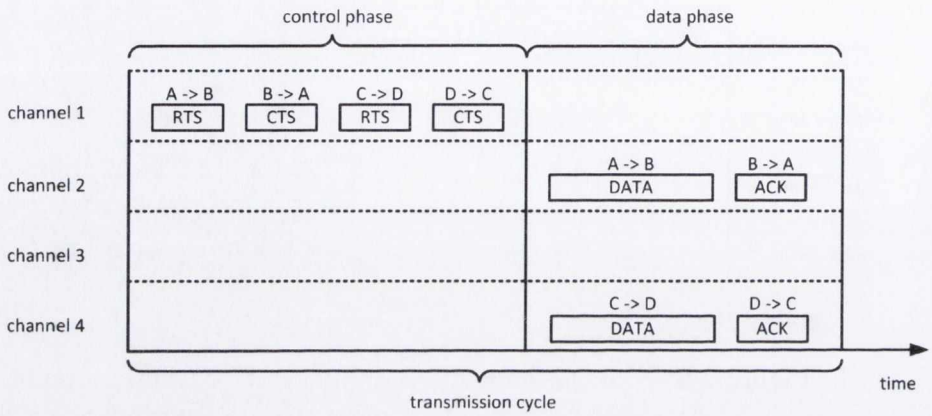

Figure 2.6: Procedure for split phase protocols

manner between a node pair to negotiate a specific data channel, and to notify their neighbors. During the data phase, nodes use selected data channels to send RTS/CTS message, before the actual data and ACK messages.

In Multichannel MAC Protocol (MAP) Chen *et al.* [2003], the length of the control phase is fixed as in MMAC, but the length of the data phase may vary depending on the scheduled transmission. The MAP protocol assumes that every node can hear each other during the contention reservation interval, i.e., the control phase, and therefore the reservation made in the control phase is known to every node. Under such an assumption, everyone makes the same decision on the transmission schedule, which aims to achieve maximum utility of channels and minimize the longest busy time in all channels. Nodes in MAP need to be strictly synchronized, i.e., receiving the exact same message, in order to create a consistent view on the current schedule. Such an assumption may be difficult if not impossible to be satisfied in a multi-hop environment.

Choi *et al.* [2003] propose a protocol that does not require synchronizing the control phase. Instead, a maximum length of channel usage is defined (called Maximum Transmission Time, MTT), and an ACK message following the data is sent on the control channel instead of the data channel. Consequently, by observing the control channel for a MTT time, if no ACK message is received, the channel is free. The drawback of this approach is that a node needs to wait for MTT time to determine the channel status, which is inefficient.

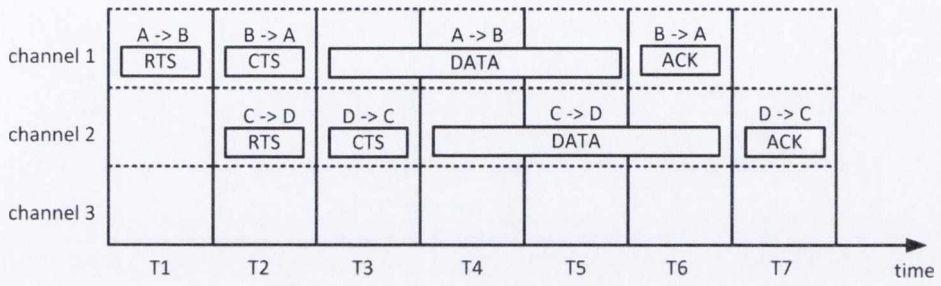

Figure 2.7: Procedure for hopping protocols

The split phase protocols have two advantages: 1) mitigate multi-channel hidden terminal and deafness problem, 2) support broadcast, as broadcast messages can be sent in the control phase. The disadvantages of split phase protocol are as follows: 1) tight synchronization is needed, but with less tighter requirement than common hopping, as channel switch is less frequent, 2) difficult to properly determine the proportion of control and data window, 3) constant monitoring of channel status is needed for channel selection, 4) channel switching takes time.

2.3.4 Hopping Protocols

Nodes in this category constantly switch, i.e., hop across a number of channels. Based on the hopping patterns, there are two major flavors: common hopping and parallel rendezvous.

In common hopping, such as CHMA Tzamaloukas & Garcia-Luna-Aceves [2000], all nodes follow a same hopping sequence. Time is divided into discrete intervals. Nodes that succeeded in a collision-avoidance handshake remain at the current channel, i.e., current hop, while all other nodes continue to follow the common hopping sequence. The basic operation of common hopping is shown in Figure 2.7. Nodes follow the hopping sequence C1, C2, C3. At time T1, all nodes are at hop C1, and node A and B successfully established a data connection. Therefore at time T2, node A and B will stay at hop C1, while all other nodes continue the hop sequence to C2.

In parallel rendezvous protocols such as McMAC Hoi-Sheung *et al.* [2007] and SSCH Bahl *et al.* [2004], nodes do not have a common hopping sequence. Instead, each node has its own individual sequence. Such individual hopping sequence or

hopping seed is exchanged among nodes. If a sender intends to send data, it hops to the receiver's hopping sequence, and starts to negotiate a data session. The data transmission is then occurred on the receiver's hopping channel.

The advantages of hopping protocols: 1) use all channels for data exchange, no control channel bottleneck, 2) simple decision making in selecting channel, no extra overhead for channel negotiation. The disadvantages of hopping protocols: 1) deafness and multi-channel hidden terminal problem, 2) channel switch delay penalty (channel dwell time should not be too small), 3) the hopping sequence needs to be synchronized consistently, which is difficult in multi-hop networks.

2.3.5 Conclusions and Suitability in VANETs

The use of multi-channel protocols in VANETs can improve the overall throughput, as more than one channel is used to transmit data. In addition, service differentiation can also be implemented by allocating data to different service channels.

However, as previously discussed, broadcast operation is not fully supported in multi-channel protocols, which makes such protocols less applicable in VANETs as a majority of safety applications requires broadcast. In addition, in order to achieve correct channel assignment, channel information such as the hopping sequence or the synchronization information needs to be maintained in a consistent manner. It is a challenging task to preserve such consistency in a multi-hop environment with fast-changing topology such as VANETs. Finally, to use multi-channel rather than one channel only improves the performance in terms of the system throughput. Other important metrics such as communication reliability and real-timeliness cannot easily be improved from multi-channel operations.

2.4 Topology-Transparent Scheduling Protocols

For most scheduling algorithms, it is implicitly assumed that a schedule needs to be recomputed if the network topology changes, which make these protocols "topology-dependent". The need for constant schedule adaptations may lead to large communication overhead and performance degradations. In a highly

dynamic environment, it is desirable if a schedule can be made independent of the network topology, i.e., topology transparent. In 1994, such an idea was proposed using mathematical properties of finite (Galois) fields Chlamtac & Farago [1994]. In this algorithm, a node can transmit a unicast message in a collision-free way for at least once in a frame. without knowing any specific network topology.

The core idea of the topology-transparent scheduling is as follows: by assigning a unique code to every node in the network, any two nodes can have only a finite number of collisions within a frame. If a message is transmitted multiple times in a frame, and the number of such retry is large enough to incorporate all possible collisions with the receiver's neighbors, then at least one retransmission can be made collision-free. In other words, even in the worst case scenario, a node still has at least one collision-free slot in each frame, regardless of the current network topology.

Before any specific description of this algorithm, a briefly introduction to the finite field (or Galois field) and its properties are presented in the following. First of all, a field is an algebraic object. The elements of a field can be added, subtracted, multiplied and divided with each other while the results are still elements of the field. A finite field is a field with a finite number of elements. For example, the Galois Field with q elements $GF(q) = \{0, 1, 2, 3, \ldots, q-1\}$ can be constructed if q is a prime number or prime power, i.e., $q = p * m$ (where p is a prime and m is an integer).

The $GF(q)$ has two operations: addition and multiplication modulo q. Further more, a number of polynomials $f(x)$ (mod q) with degree k can be constructed over $GF(q)$, meaning that all coefficients, domain and range of this polynomial are all elements in $GF(q)$. Since each coefficient has q possible values and the degree of the polynomial is k, there are q^{k+1} number of distinct polynomials over $GF(q)$, i.e., $f(x) = a_k x^k + a_{k-1} x^{k-1} + \ldots + a_1 x + a_0$. The most relevant property of the Galois Field to the scheduling problem is that the polynomial $f(x)$ (mod q) with degree k over $GF(q)$ has at most k distinct roots, which are also elements of $GF(q)$, i.e., between 0 and $q-1$. As a result, for two distinct polynomials over the same $GF(q)$, there are at most k common points, since $f_1(x) - f_2(x)$ is still a polynomial with at most k roots.

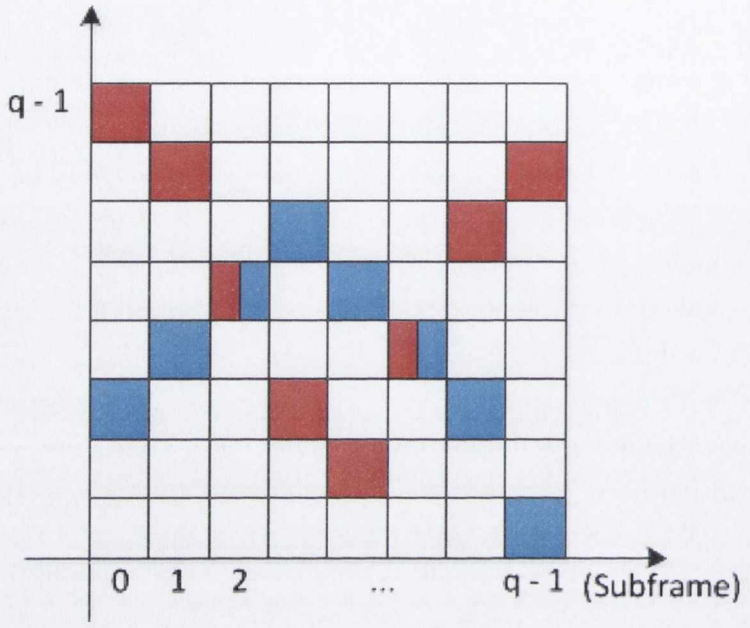

Figure 2.8: Mapping of a $GF(q)$ polynomial to $q * q$ slot space

2.4.1 Specific Topology-Transparent Scheduling Protocols

Chlamtac and Farago utilized this property of Galois field and mapped such polynomials into a $q * q$ slot space, i.e., $(q, f(q))$ as depicted in Figure 2.8. Suppose that there are two polynomials (blue and red) over $GF(q)$. Since any two polynomials with degree k has at most k common roots, the maximum number of common slots (slots with two colors in Figure 2.8) that two polynomial have is at most k as well. Consequently, a frame can be constructed to utilize this property. In the proposed algorithm, each frame is composed of q subframes (a vertical column in Figure 2.8) which has q slots. Every node selects a slot in each subframe, according to its assigned unique polynomial. According to previous discussion, the maximum number of common slots (i.e., collisions) that two nodes have is at most k. As a result, suppose that there is only one interfering neighbor to the message receiver, if a message is transmitted $k + 1$ times, it is guaranteed that this message can be received without collision for at least once in each frame.

Now consider the case that multiple neighbors exist for the receiver. Assume that the maximum nodal degree in a network is D, which implies that for every

message receiver there are at most D neighbors which may cause collisions. In the worst case scenario, a message can only collide with one neighbor in k subframes, and thus the total number of possible collided subframes for all D neighbors is $k * D$. Consequently, collision-free transmission which take all D neighbors into account can be achieved, if the number of subframes q is larger than $k * D$ (condition 1).

Another constraint for the collision-free transmission using Galois field is that the total number of available polynomials, or "codes", must exceeds the total number of nodes in the network, which is assumed to be known *a priori* as N. Consequently, each node is guaranteed be assigned a unique polynomial over $GF(q)$, and the number of available codes q^{k+1} of a $GF(q)$ must exceed N, i.e., $q^{k+1} \geq N$ (condition 2).

Given a network with its number of node N and maximum nodal degree k known, the proposed topology-transparent scheduling algorithm: Galois Radio Network Design (GRAND) works as follows:

1. search for a prime number q and an integer $k \geq 1$, that satisfies condition 1 and condition 2.

2. Assign each node in the network with a unique polynomial, which is constructed using q and k.

3. Each node calculates the slot position in each subframe, by calculating the value of the polynomial at each subframe. For example, for a node with assigned polynomial $x^2 + 2$, it uses the second slot in the first subframe as $f(0) = 2$.

The GRAND algorithm guarantees that a node has at least one collision-free unicast transmission slot in a frame, but its performance can be further optimized. For instance, the Modified Galois Design (MGD) protocol Cai *et al.* [2003] achieves better performance in terms of utility ratio and worst-case medium access delay, by reducing the frame length. In MGD, for a prime number q that satisfies $q \geq k * D + 1$ and $q^{k+1} \geq N$, the number of subframes p can be chosen between $k * D + 1$ and q as depicted in Figure 2.9, exploiting the fact that q is a prime number but p can be any integer. While retaining the "at least one

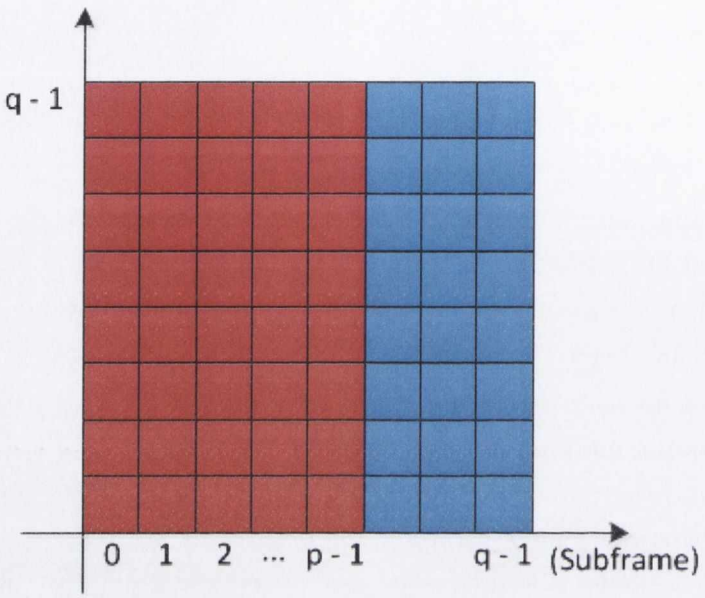

Figure 2.9: Choosing smaller subframe number to reduce frame length in MGD

collision-free slot" guarantee, the total frame length in MGD is now reduced to $p * q$, which is smaller then $q * q$ as in GRAND. In order to find the optimum q and k value to achieve minimum frame length (MFL), the authors rigorously proved that the MFL is obtained at $\lceil k_0 \rceil$ or $\lfloor k_0 \rfloor$, where k_0 is the unique positive root of equation $x * D + 1 = N/(x + 1)$.

The optimum q and k value can be decided by comparing the MFL at $\lceil k_0 \rceil$ and $\lfloor k_0 \rfloor$. Once q, k and p are finally decided, the $GF(q)$ and polynomials can be generated and distributed to all nodes in the network, same as the GRAND algorithm. The contribution of the MGD algorithm is the proof of the proposed method to calculate the optimum value of q and k, and the idea to use smaller number of subframes to reduce total frame length. However, MGD is still a unicast rather than broadcast scheduling algorithm as the authors claimed.

In the GRAND algorithm, one message is repeatedly sent in q subframes in order to guarantee that at least one retransmitted message can be received without a collision. As the goal being minimizing the frame length, it makes no sense to increase the q value after a proper q has been found. However, if the design strategy shifts from achieving collision-free message transmission to

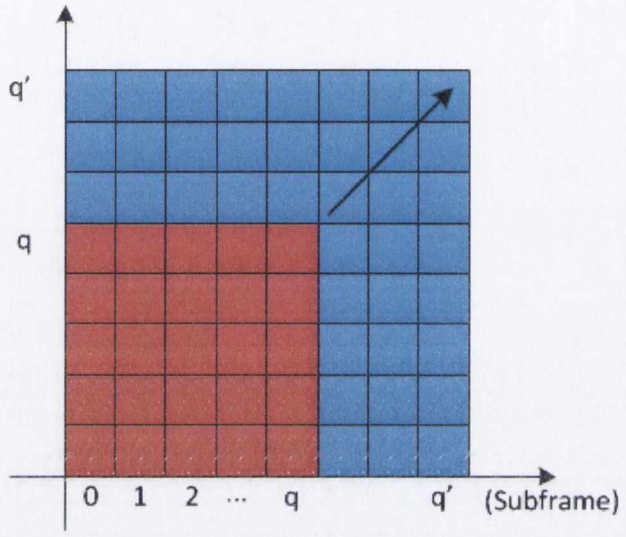

Figure 2.10: Increasing q value to maximize the number of collision-free subframes

transmit probabilistically with guaranteed minimum collision-free rate, then the idea of expanding the q value seems plausible Ju & Li [1998]. Given the Galois Field theory that two nodes can have at most k collisions in q subframes, the remaining $q - k$ subframes are collision-free subframes, and thus the q value can be made artificially large to incorporate more "collision-free" frames. Assume that the maximum nodal degree is D in the network, and if q different messages are sent by a node in a frame, at most $k * D$ of them will be collided, leaving $q - k * D$ collision-free messages. To maximize the number of collision-free subframes, the intuition is to increase the q value infinitely Figure 2.10. However, as q grows, the total frame length $q * q$ grows even faster, diluting the ratio of number of collision-free transmissions per frame.

Ju and V.O.K define the "minimum throughput" G_{min} as the ratio of the number of guaranteed successful transmissions in each frame length L, i.e., $G_{min} = (q - k * D)/q * q$, and proved that G_{min} has a maximal value when $q = 2 * k * D$. This algorithm (named "optimal algorithm") tends to have large q number and therefore, has a larger frame length. However, since the goal is to achieve maximized minimum throughput, the frame length is not the primary concern. It is worth emphasizing that the design philosophy of this algorithm is completely different from the original GRAND algorithm. Different messages are transmit-

ted in q subframes, and no collision guarantee is given to any of the messages. Although the algorithm transmit messages in a probabilistically manner, the minimum throughput can be guaranteed.

2.4.2 Analysis

A considerable amount of research efforts have been made since Chlamtac and Farago's initial work in 1994, which exploits the notion of topology-transparent scheduling. The idea of scheduling slots without knowing any topology specifics is a highly desirable idea, especially in a network with constant and fast changing topologies, e.g., VANETs. However, the topology-transparent (TT) approach has severe, if not insurmountable, flaws in the following three aspects:

First of all, the TT algorithm assume that each node in the network are assigned a unique (but relatively scarce) code before commence any data transmission. It is not clear how this objective is achieved, and the related time and communication cost are not mentioned in any previous works. In fact, if unique codes can be generated and disseminated in the entire network, why not just directly disseminate slot schedules? In addition, due to the scarcity of the code, the total number of nodes in the network must be estimated accurately. Otherwise, if nodes with identical code encounter, the collision-free property of the algorithm cannot be maintained. As a matter of fact, those nodes end up colliding all their transmissions with each other.

Secondly, the TT approach is highly inefficient with extremely large amount of redundant messages. In fact, the algorithm always assumes the worst case scenario and retransmits a packet multiple times. For example, in a network with maximum node degree of 10, it takes at least 10 retransmissions to send a unicast message, and it is impossible to know which one of them has been received successfully. If broadcast is taken into consider, the number of retransmissions is simply multiplied by the maximum number of node degree.

At last, as mentioned earlier, the accurate estimation of the two parameters N (total number of nodes) and D (maximum nodal degree) of the network is critical for the correctness of the TT algorithm. However, in a realistic network, these two parameters cannot be obtained easily. In fact, in an open network

in which nodes may join and leave, these two numbers are subject to constant change and are difficult to estimate. The correctness of the TT approach in such an environment is therefore highly questionable.

2.5 Distributed Scheduling Protocols

In this section, a number of distributed and reservation-based MAC protocols are reviewed. Protocols in this category use reservation mechanisms in accessing the medium which results in higher transmission reliability. However, most reservation-based protocols depends on the specific network topology, and therefore susceptible to network dynamics. In addition, the problem of reserving a time slot among a number of competitors is essentially a consensus problem. The issues and implications in achieving distributed consensus with dynamic participants and unreliable communication channel are discussed at the end of this section.

2.5.1 RR-ALOHA

The RR-ALOHA protocol Borgonovo *et al.* [2002] and Borgonovo *et al.* [2004] is a fully distributed, reservation-based MAC protocol that can dynamically establish reliable broadcast channels. In RR-ALOHA, nodes are divided into one or more clusters in which full connectivity is assumed. Such a cluster is called One-Hop (OH) and non-disjoint one-hop clusters with a common subset are referred to as Two-Hop (TH). An example is depicted in Figure 2.11 where 7 terminals formed three one-hop clusters and three two-hop clusters.

TDMA scheme is used in RR-ALOHA where time is sliced into equal-sized time slots, and a fixed number of slots form a frame. Within each time slot, the Frame Information (FI), which is a slot bitmap reflecting the world view of the message sender, is piggybacked along with data packets. A slot bitmap contains the status of each slot in a frame - either empty, or owned by a specific node. An example of FI is depicted in Figure 2.12. In RR-ALOHA, FIs received from one's 1-hop neighbors are rebroadcast again, thus effectively, slot allocations are disseminated in two hops. In Figure 2.11 for instance, node 2 which locates in the

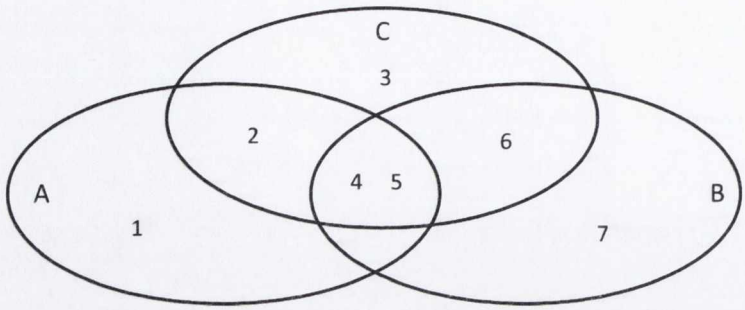

Figure 2.11: Cluster structure in RR-ALOHA

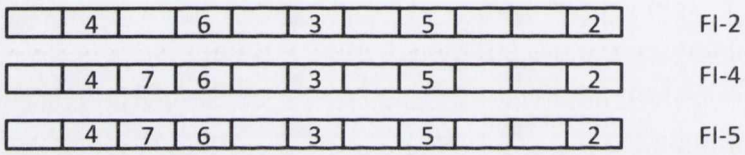

Figure 2.12: FI message received by terminal 1

overlapping region of OH A and C rebroadcasts the FI that it hears from both clusters, which effectively propagates slot allocations information from one cluster to another. In RR-ALOHA, these "bridging" nodes are critical in eliminating hidden terminal problems. For two or more clusters with an overlapped region, all cluster members should assign the slot consistently with those in the overlapped region, as their view of the network is more comprehensive.

To achieve collision-free slot allocation, there are two rules in RR-ALOHA, which are referred to as "environment probing" and "access attempt".

Environment probing: before a newly-joined node requests for an exclusive slot, it needs to probe the environment to familiarize itself with the existing slot allocation in the local region. In every non-empty slot, the probing node receives one FI message from the current transmitting node. After listening to a complete frame, the probing node receives FIs from all its 1-hop neighbors, and is able to generate its own FI.

For example, assume that node 1 tries to reserve a slot and starts to probe the environment. With a network structure depicted in Figure 2.11, node 1 receives FI message from node 2, 4 and 5, as depicted in Figure 2.12. All entries in the received FIs are exactly the same except in the third slot. Node 4 and 5

know node 7 since they are located in the overlapped region, while node 2 does not. Node 1 synchronizes its own FI with FI-4 and FI-5 following the rule that, even if there is only one neighbor marks a slot as occupied in its FI, the probing node should mark the slot as occupied as well. By marking slot usage based on nodes in the overlapped region, a consistent view of slot allocation can be achieved within 2-hops, and both 1-hop and 2-hop (hidden terminal) collisions are therefore avoided.

Access attempt: after a node obtains the current slot allocation within two hops, it attempts to transmit data in one of the empty slots. There are two possible outcomes of this action. During the next frame, if all 1-hop neighbors mark the attempted slot as being occupied by the requesting node, this slot is successfully reserved. On the other hand, if any of the 1-hop neighbor mark the attempted slot as empty or occupied by other nodes, an "access collision" has occurred and this slot is not reserved. A successful slot reservation ensures an exclusive allocation of this slot in two hops. Suppose that there is a rival node (either 1-hop or 2-hop) that attempts the same slot with the requesting node at the same time, a transmission collision, i.e., access attempt collision, will occur at the requesting node's 1-hop neighbors. This will cause nodes who experienced the collision to mark the attempted slot as empty in its FI, which implicitly reports a reservation failure to both competing nodes.

The basic operation of RR-ALOHA is summarized as follows. Any node that needs to reserve a slot has to be aware of the current slot allocation in its 2-hop neighborhood. Subsequently, it sends out a reservation request in an unused slot, and collects feedback from its 1-hop neighbors. If positive acknowledgments are received from all 1-hop neighbors, the attempted slot is successfully reserved and collision-free transmission in this slot can be guaranteed. Otherwise, the attempted slot is not reserved and the node has to start the process again.

However, the correctness, i.e., the collision-free property of the RR-ALOHA protocol is based on the assumption that the network topology remains static when the slot allocation is generated and used. If a node moves away from the location where the schedule is created, then collision-free slot access can not be guaranteed and a new slot allocation must be assigned. Based on this observation, RR-ALOHA is susceptible to topology changes which occur constantly in

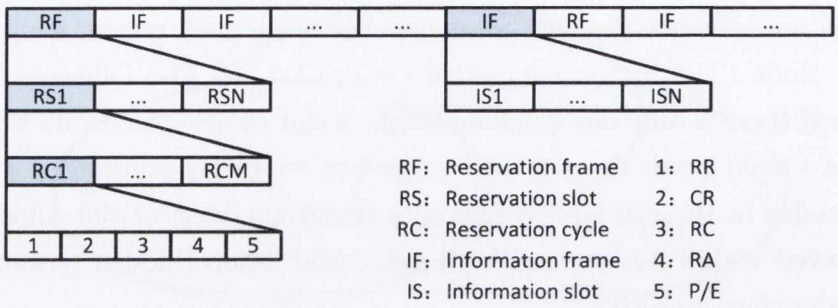

Figure 2.13: FPRP frame structure

vehicular networks.

2.5.2 FPRP

The Five-Phase Reservation Protocol (FPRP) Zhu & Corson [1998] is a fully-distributed and reservation-based MAC protocol. Reliable transmissions can be achieved without hidden terminal collisions. FPRP assumes that the capture effect does not exist, i.e., if two or more packets arrive at the same time then none of them can be received successfully. In addition, nodes have the capability to detect collisions. In FPRP, time is divided into frames, slots, cycles and mini slots, and synchronized among nodes. The frame hierarchy is depicted in Figure 2.13.

At the highest level, a super frame is composed of one Reservation Frame (RF) and a fixed number of Information Frame (IF). Both RF and IF are further divided into an equal number of Reservation Slot (RS) and Information Slot (IS), and there is a one-to-one mapping between a RS and an IS. The slot assignment agreed in the n-th RS will be applied in the corresponding n-th IS and repeated in subsequent IFs. On the third lever of the hierarchy, each reservation slot is composed of a fixed number of reservation cycles (RC). In each of the RC, there are five mini slots, i.e. five phases, which are used for slot contention.

Nodes in FPRP need to acquire a reservation slot in order to transmit in the corresponding information slot. There are a number of reservation cycles that can be used to serve this purpose. If the reservation attempt in the first cycle is not successful while no other node succeeds as well, a node may continue trying

Figure 2.14: FPRP deadlock scenario

the next cycle until the end of the reservation slot.

In each reservation cycle, nodes need to go through five phases. After these five phases, a successful node can use the slot exclusively and prohibit other nodes from attempting this slot in the following reservation cycles. In the first phase (Reservation Request), the Requesting Node (RN) broadcasts a slot request message with probability p. RN's 1-hop or 2-hop neighbors may also transmit slot request and cause attempt collisions. If such a collision occurs, RN's 1-hop neighbors reply a Collision Report (CR) to RN in the second phase (Collision Report), otherwise remain silent. In the third phase (Reservation Confirmation), if no CR is received in the previous phase, the RN broadcasts a reservation confirmation message to its 1-hop neighbors. In the forth phase (Reservation Acknowledgment), RN's 1-hop neighbors reply with an acknowledgment to the RN, while inform RN's 2-hop neighbors the successful slot reservation. The fifth phase (Packing / Elimination) is related to protocol efficiency and handling deadlock during slot contention.

- Comparison between FPRP and RR-ALOHA

The properties of the FPRP protocol are better understood if it is compared with RR-ALOHA. In the following discussion, three FPRP attributes that are different from RR-ALOHA are of interest: 1) the usage of collision report rather than positive acknowledgment in detecting access conflict. 2) four rather than two phases in slot reservation, 3) multiple reservation cycles. Interestingly, all these differences derive from a common root that, during reservation, nodes in RR-ALOHA have exclusive slots while nodes in FPRP do not. Exclusive reservation slots mean that reservation messages are transmitted in a collision-free manner, such as in RR-ALOHA; while in FPRP, nodes can only transmit reservation messages in a contention-based and collision-prone manner.

Such a difference has profound consequences in a broadcast scenario. Replies from 1-hop neighbors will collide with each other in FPRP, but can be successfully received in RR-ALOHA, which leads to FPRP's negative acknowledgment

(NACK) approach (where nodes only report to the requesting node if a collision is detected) as opposed to RR-ALOHA's positive acknowledgment approach. Under such a NACK scheme, the requesting node's 1-hop and 2-hop neighbor are unable to know whether a slot reservation is successful or not.

In addition, the requesting node can not guarantee that a reservation is successful, even if such a collision report is not received. This phenomenon is termed a "deadlock" as illustrated in Figure 2.14. Suppose that node A and B take exactly the same action, and both send a reservation request. Since nodes can not detect collision while transmitting, they both receive no collision reports and believe they have successfully reserved this slot and send out reservation confirmation in phase three. If node C and D does not exist, node A and B will not receive the expected reservation acknowledgment in phase four, thus understand that a dead lock is formed, and give up in this reservation cycle. However, if node C and D do exist, which are unable to hear B and A's transmissions, they will not report access collisions and reply a reservation acknowledgment in phase four. These acknowledgments make A and B erroneously believe that no deadlock exists and therefore use the same slot. In FPRP, this situation is partially amended by letting A and B transmits Elimination Packet (EP) in the fifth phase with a fixed probability of 0.5. Consequently, there is a possibility that in a deadlock situation, node A and B can discover each other and avoid using the same slot.

In RR-ALOHA, a node contends for a slot, only if it does not already have one; while in FPRP, all nodes need to contend for a slot at the beginning of a super frame. For a successful slot reservation in FPRP, it is implicitly assumed that the winning node is the only one that sends out the slot request message in its 2-hop ranges. Since every node needs to contend for a slot, the number of contenders for each slot in FPRP is much larger than in RR-ALOHA, which means higher probability of access collision. This is the reason why multiple reservation cycles are used in FPRP. As multiple reservation cycles exist for a single slot, a winning node need to prohibit its 1-hop and 2-hop neighbors from attempting this slot in subsequent reservation cycles. Otherwise, without knowing this result, other nodes may also reserve the slot successfully in subsequent cycles. The third and forth phase are introduced to inform the winner's 1-hop and 2-hop neighbors, as they are unable to know the results of the reservation, due to the NACK

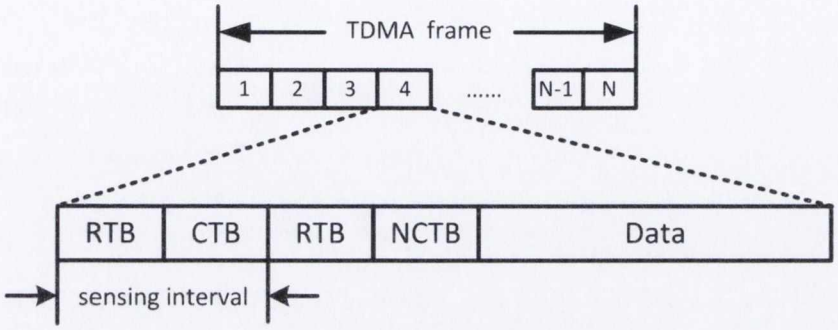

Figure 2.15: ABROAD frame structure

mechanism that are mentioned earlier.

Same with RR-ALOHA, FPRP is also susceptible to topology changes, and the correctness of the protocol is based on the assumption that topology does not change during the formation of the slot reservation. In addition, FPRP requires tight time synchronization as those mini slots are extremely short compared to the size of a data slot.

2.5.3 ABROAD, CATA and RBRP

Generally, the following protocols: ABROAD Chlamtac *et al.* [2000], CATA Tang & Garcia-Luna-Aceves [1999], and RBRP Marina *et al.* [2001] share a number of common properties with each other, therefore are discussed together. The most obvious similarity is that all these protocols are fully-distributed, and TDMA-based just like RR-ALOHA and FPRP. However, there are a number of subtle differences in terms of the ways that slot reservations are conducted, notified, and terminated. By analyzing the similarities and differences of these protocols, high-level objectives, available techniques, challenges and deficiencies of this particular type of MAC protocols can be better understood. In the following, a brief introduction of these three protocols is given, followed by the comparisons and analysis.

- ABROAD

In ABROAD, there are two types of slots: assigned and empty. A slot in a frame is composed of four mini control slots and one data slot as shown in

Figure 2.15. If a node has already reserved this slot (assigned), it broadcast a RTB (Request to Broadcast) in the first mini slot, and its 1-hop neighbors would broadcast a CTB (Clear to Broadcast) in the second mini slot. The purpose of this message exchange in the first two mini slots is to let a slot owner to back-off or block other possible contenders within two hops. This 2-hop slot ownership notification is seen in the third and fourth reservation phase in FPRP, and is equivalent to the busy tone mechanism used in the contention-based DBTMA. Later, the slot owner can transmit packets exclusively in this data slot and in subsequent frames.

If a node does not hear any transmission activity in the first two mini slots, it has the opportunity to contend for this slot in the next two mini slots (third and fourth). After sending a RTB message, if no NCTB, i.e., negative acknowledgment (NACK) message is received, this slot is successfully reserved, otherwise failed and the node tries again in the next available slot. The RTB and NCTB mini slot serve the same purpose as the first and second reservation phase in FPRP to let nodes to contend for an empty slot. If a deadlock scenario is not presented, simultaneous reservation attempts within two hops would not succeed because RTB messages would collide at one hop neighbors and being reported in the following NACK messages.

Based on their functionalities, the four mini slots in ABROAD can be divided into a notification part and a contention part. By comparing ABROAD with FPRP, it becomes clear that FPRP follows a contention - notification sequence while ABROAD is just the opposite, i.e., notification - contention. Intuitively, a slot should be contended first then the result can be notified (FPRP approach). However, this intuition only applies to a situation where nobody owns this slot. If the slot has already been reserved, the slot owner can block any possible contenders by notifying them before their contention begins (ABROAD approach).

Similar to RR-ALOHA, in ABROAD, slot contention occurs among nodes that do not possess any slot. The contention intensity is therefore mush lower in ABROAD than in FPRP, where contentions happen among all nodes at the beginning of a frame. This property of ABROAD is also preserved in CATA and RBRP.

- CATA

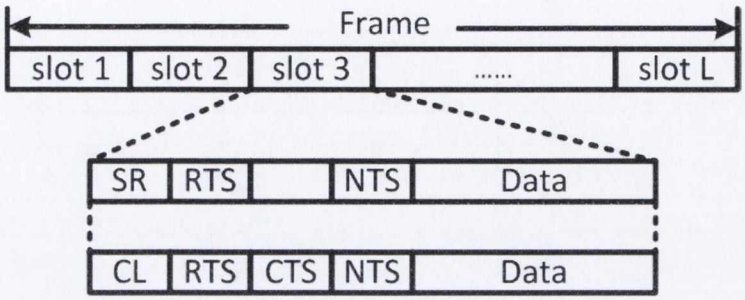

Figure 2.16: CATA frame structure

In CATA, nodes contend and notify their slot reservations in a very similar way as ABROAD, therefore the frame structure is quite similar to ABROAD as depicted in Figure 2.16.

CATA supports both broadcast and unicast transmissions by using different slot structures. In a broadcast scenario (upper slot structure in Figure 2.16), nodes that receive data in the same slot in previous frame(s), i.e. slot owner's 1-hop neighbor, send out a Slot Reservation (SR) packet in the first mini slot, which is used to back-off potential contenders on behalf of the slot owner. Nodes that do not hear any activity in the SR mini slot can contend for this slot in mini slot two and four. Collision-free reservation is achieved via conventional NACK mechanism. A contention winner can use the slot for subsequent frames until the end of its desired transmission.

The slot reservation mechanism in CATA is also composed of the same two functionalities - contention and notification, as in ABROAD. The unique property of CATA is that the first part of the notification, i.e., 1-hop notification, disappears, leaving only the 2-hop notification part (SR). The notification to the slot owner's 1-hop neighbors has been implicitly embedded in the data slot of the previous frame. Nodes that receive a data packet (equivalent to the 1-hop notification) are obliged to broadcast a SR message in the same slot during the next frame. Therefore, an explicit 1-hop notification is not necessary.

- RBRP

In RBRP, a number of new concepts and approaches are proposed. The most significant one is the introduction of a 1-hop slot table. Upon a successful

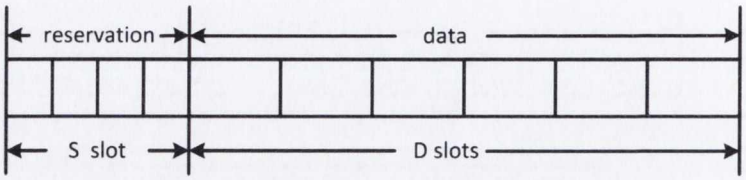

Figure 2.17: RBRP frame structure

reservation, 1-hop neighbors of the winning node mark the slot with the winner's ID. If this slot is ever contested by another node in the following frame, 1-hop neighbors can NACK such attempt according to the slot table, in addition to the conventional mechanism to detect simultaneous attempts. In other words, nodes in RBRP can veto a slot request if either collisions are detected, or the requested slot has already been assigned to another node.

With the help of the slot table, the explicit notification process, which involves sending messages to neighbors within two hops, is no longer needed prior to each slot usage. The only exception is when the slot is reserved for the first time and the winner still needs to notify its 1-hop neighbors. Another benefit of using slot table is the decoupling of the one-to-one mapping between the reservation slot and data slot as explicitly or implicitly specified in ABROAD, CATA and FPRP. Nodes in RBRP are able to contend any data slot in any reservation slot as depicted in Figure 2.17, which means a more flexible slot allocation mechanism similar to RR-ALOHA.

The second innovation of RBRP is to multiply the first contention phase in order to eliminate the deadlock possibility. As discussed earlier, the so-called deadlock situation occurs when two nodes without a common neighbor copycat each others' behavior, and are unable to discovery the other side, which leads to a conflicting reservation. This problem is tackled by adding another phase (the fifth) in FPRP, based on the idea that deadlocked nodes may discover each other via probabilistic beaconing. The same idea is applied in RBRP by expanding the single contention request slot (the first part of the contention) into M mini slots, where each contender uses K of them to transmit a same reservation request. Unless two deadlocked nodes choose exactly the same K slots in M reservation slots, they will eventually discover each other and realize a deadlock situation.

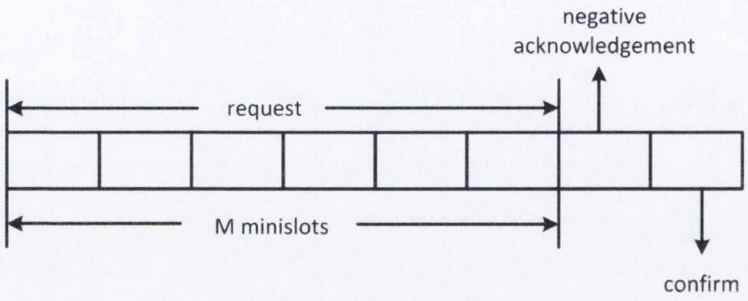

Figure 2.18: RBRP reservation slot structure

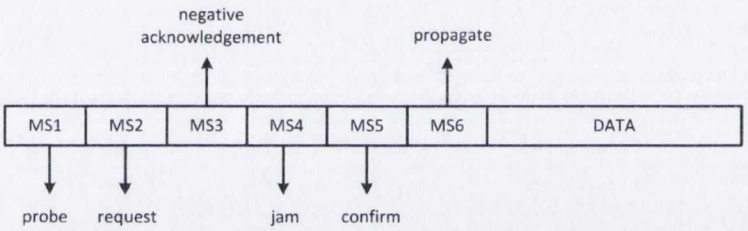

Figure 2.19: RBRP data slot structure

As depicted in Figure 2.18, M mini slots and the NACK slot constitute the contention part of a reservation protocol. The subsequent confirmation mini slot is the first part of the conventional notification process, and is also the only part needed in RBRP. If a node wins a contention, it sends out a notification to its 1-hop neighbors in the confirmation mini slot. Neighbors update the corresponding slot table, and block any slot reservation attempt in the following frames. A slot owner can use its data slot in the following frames without notifying its 2-hop neighbors, which is similar to ABROAD and CATA. If a data slot is not used in a frame, it becomes available for other nodes to contend.

The third unique design of RBRP is the introduction of a protection mechanism from node mobility. The basic idea is to insert a re-confirmation process inside a data slot prior to transmitting in the reserved data slot. Specifically, as depicted in Figure 2.19, slot owner engages in a complete contend (1-3 mini slots) and notify process (5-6 mini slots) before broadcasting data. These two processes are standard operation procedures that are previously discussed. The only unique point is the use of an extra slot (mini slot 1) in which probabilistic beaconing mechanism is used to reduce deadlock probability. If a slot owner does

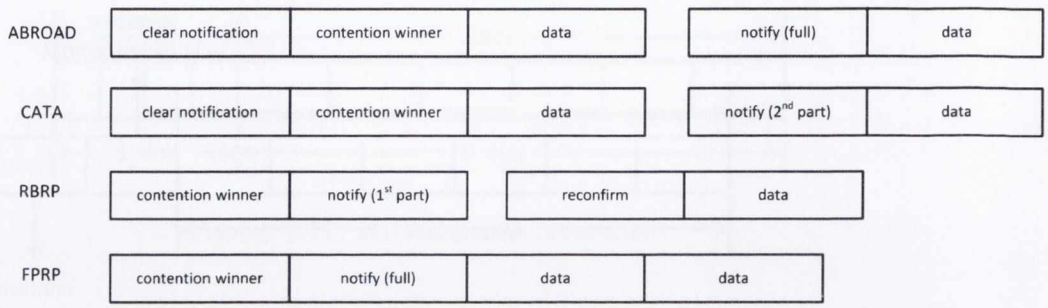

Figure 2.20: Comparison of distributed protocols

not receive a NACK in mini slot 3, it notifies 2-hop neighbors and transmits; otherwise gives up this data slot and contend for another one later.

The ABROAD, CATA, RBRP and FPRP protocol are summarized in Figure 2.20. The focus is on the conditions to successfully reserve a slot, and the actions taken in the subsequent frames. Possible conditions are abstracted as: 1. clear of notification, i.e., absent of any slot notification from others, and 2. contention winner, i.e., no NACK is received.

In addition, possible actions are abstracted as: 1. notify (full / 1st part / 2nd part), i.e., distribute slot notification in both 1-hop and 2-hop / 1-hop only / 2-hop only, 2. reconfirm, i.e., only apply to the reconfirmation process in RBRP, and 3. data, i.e., transmit data in the reserved slot.

The common properties that are shared by ABROAD, CATA, RBRP, FPRP and RR-ALOHA are also summarized in Table 2.5.

2.5.4 NAMA and SEEDEX

NAMA Bao & Garcia-Luna-Aceves [2001] is a distributed TDMA-based protocol which uses an implicit slot contention method. It is different from aforementioned protocols such as FPRP and RBRP in a way that the slot contention is conducted without sending explicit slot request messages. In NAMA, by broadcasting IDs periodically within 2 hops, nodes achieve complete 2-hop neighbor awareness. Before a new node joins the neighborhood, it listens for a whole frame to acquaint itself with current neighbors in the vicinity.

The implicit slot contention is achieved via the Neighborhood-aware Con-

Table 2.5: Comparison of FPRP, ABROAD, CATA, RBRP and RR-ALOHA

	FPRP	**ABROAD**	**CATA**	**RBRP**	**RR-ALOHA**
Private slots	No	Yes	Yes	Yes	Yes
Contention intensity	High	Low	Low	Low	Low
Contention mechanism	NACK	NACK	NACK	NACK	ACK
Notification mechanism	One-off, Full	Consecutive, Full	Consecutive, 2nd Part	One-off, 1st Part	One-off, No need
Deadlock handling	5th Phase	Not specified	Not specified	Duplicate request slots	No need
Reservation and data slot mapping	Fixed	Fixed	Fixed	Flexible	Flexible
Mobility consideration	No	No	No	Verification	No
Slot termination condition	Fixed cycle	End of Tx	End of Tx	End of Tx	End of Tx
Channel assumption	Ideal	Ideal	Ideal	Ideal	Ideal

tention Resolution (NCR) mechanism. In short, NCR locally calculates all neighbors' priority at any specific slot, and if a node has the highest among all neighbors, then it can transmit, otherwise it listens. Specifically, for a given slot t, which is also known as a contention context in NAMA, the priority value that a node k has is randomly but deterministically decided by: $Rand\ (k + t) + k$; where $+$ is the concatenation operation, and $Rand$ is a pseudo-random number generator. The priority value is solely decided by the slot number and the node ID in NCR, which means that an entire time schedule of any specific node can be calculated if this node's ID is known. As a result, with all nodes in the network understand their 2-hop neighbors' ID, their priorities and slot allocations can be calculated and collision-free transmission can be achieved.

NAMA essentially transforms the problem of allocating slots into allocating sequencing node IDs in a 2-hop neighborhood. Although the problem is simplified by eliminating the "contention" part of a distributed MAC protocol, the downside is that all slots are allocated statically, and it is not possible to alter the slot allocation during run time to support QoS or real-time communication.

In addition, to achieve consistent 2-hop awareness is not trivial, especially when the network topology is dynamic. In order to transmit legitimately, a new node must be recognized and acknowledged by all of its 2-hop neighbors. For a newly-joined node, if such unanimous recognition is not guaranteed, collisions are

always possible between the joining node and those who do not recognize this new member. It is more complicated when more than one nodes are joining or leaving at the same time. In summary, the correct operation of NAMA depends on the consistent 2-hop membership, which may be susceptible to network dynamics in VANETs.

The SEEDEX protocol Rozovsky & Kumar [2001] is a unicast and probabilistic MAC protocol. It cannot guarantee the collision-free transmission, but only reduces the collision probability. In SEEDEX, a node transmits according to its schedule which is generated from a seed. In such a schedule, a node either listens (L state) or transmits with probability p (PT state). SEEDEX resembles NAMA in a way that random seeds, rather than the actual transmission schedules are disseminated in 2-hops. By knowing the seeds of all 2-hop neighbors, a node also knows how many neighbors of its intended recipients are in the PT state. Based on this information, the node alters its transmission probability in order to reduce the collision probability.

The seeds that SEEDEX distributes cannot provide deterministic slot allocation as in NAMA. The number of possible rivals in the receiver's neighborhood, which is derived from the received seeds can only be used to reduce the collision probability rather than complete elimination. In addition, SEEDEX requires consistent and up-to-date 2-hop information in the neighborhood, which is difficult and costly to maintain.

2.5.5 STDMA

The Self-organizing Time Division Multiple Access (STDMA) protocol Bilstrup *et al.* [2009] is a decentralized, TDMA-based scheme, specifically designed for VANETs. Time is divided into slots, and time synchronization is achieved via external positioning system, e.g. GPS or Galileo. Nodes in STDMA distribute positioning information to their 1-hop neighbors periodically, which is also used for slot allocation. The STDMA algorithm has been standardized in the Automatic Identification System (AIS), which is widely applied in the shipping industry.

The basic procedure to reserve a slot in STDMA involves four phases: initialization, network entry, first frame, and continuous operation. In the initialization

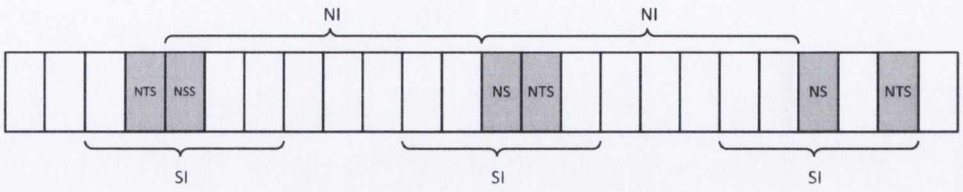

Figure 2.21: STDMA frame structure

phase, a node listens to the channel for an entire frame, in order to obtain the current slot assignment. In the network entry phase, a node locally decides which slots in a frame it intends to use. This process determines a number of specific slots in the frame as depicted in Figure 2.21.

For the first step, a node decides its Nominal Increment (NI) interval, which is the node's desired transmission interval. Subsequently, the Nominal Start Slot (NSS) is randomly selected between the current slot and the current slot plus NI. NSS is effectively a starting reference point for all following slots. The subsequent Nominal Slot is located at NI slots away. Both NSS and NS can be thought of as the "desired" slots that a node intends to reserve.

In order to evenly distribute slot reservation, the actual reservation slot - Nominal Transmission Slot (NTS) is randomly chosen around NSS and NS, and bounded by Selection Interval (SI), which is 20% of NI. If the chosen NTS has been occupied by its 1-hop neighbor, then the closest free slot in SI is chosen instead. If all slots in SI are occupied, the slot used by the node that is geographically located furthest from the reserving node is chosen. This is how the position information can be used to facilitate slot allocation in STDMA.

When the first NTS is due, a node enters the first frame phase. In this phase, future NS and NTS will be decided in this frame, and subsequent frames as well. Each decided NTS slot has a randomly decided integer (ranging from 3 to 8) specifying the number of frames after which this NTS will "expire". By then, new NTS will be randomly chosen again within SI. This special arrangement is designed to address the node mobility issue in VANETs. After the first frame, a node enters the last phase: continuous operation, where a node operates using determined NTS to transmit packets and choosing new NTS when they expire.

- Comparison with RR-ALOHA

STDMA resembles to RR-ALOHA in a way that nodes in both protocols need to listen to the channel for an entire frame, and maintain a slot allocation table locally. However, STDMA provides no collision-free guarantee for any reserved slot, since there is no unsuccessful reservation in STDMA. In addition, hidden terminal collisions, i.e., 2-hop collisions are not considered in STDMA at all. Further comparison between these two protocols are presented below, focusing on 1) listen phase, 2) empty slot access collision, and 3) collision-free transmission guarantee in reserved slots.

Listen phase: nodes in STDMA receive and obtain 1-hop slot allocation, while nodes in RR-ALOHA maintain slot allocations in 2-hops.

Empty slot access collision: In RR-ALOHA, no reservation can succeed if more than one attempt is simultaneously made within 2-hops. In STDMA, both 1-hop and 2-hop nodes can reserve an empty slot at the same time.

Collision-free transmission guarantee in reserved slots: In RR-ALOHA within a 2-hop range, 1) in the slot contention phase, no one can successfully reserve the same slot with another node, and 2) in the transmission phase, no one would attempt to access a slot that has been reserved by another node, via the 2-hop slot bitmap dissemination. In STDMA, 1) simultaneous access to an empty slot is possible for both 1-hop and 2-hop nodes, and 2) 2-hop nodes may attempt to use a reserved slot, due to the lack of 2-hop slot allocation knowledge. Only 1-hop simultaneous access can be avoided in STDMA via 1-hop slot allocation dissemination.

Compared with RR-ALOHA, the bright side of STDMA is that slot reservation is properly terminated while no explicit rules are applied in RR-ALOHA in this regard. In addition, the slot usage in STDMA is more flexible as there are multiple slots in a frame. Finally, mobility is considered in STDMA via re-selecting transmission slots after certain periods.

2.5.6 MS-ALOHA

The MS-ALOHA protocol Scopigno & Cozzetti [2009] and its prior version RR-ALOHA+ Cozzetti & Scopigno [2009], are both extensions to the RR-ALOHA protocol. The improvement focuses on three aspects: 1) taking node mobility into

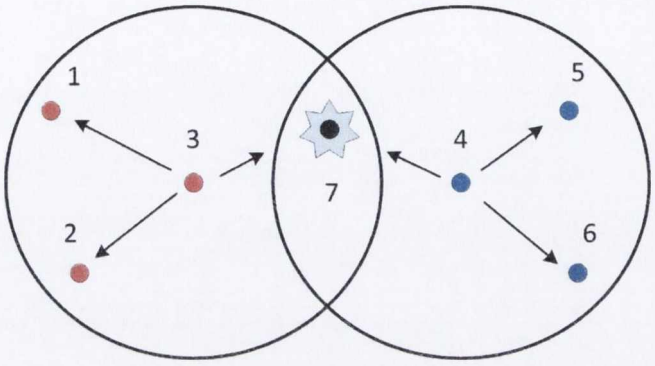

Figure 2.22: RR-ALOHA inconsistency problem

consideration, slots are reserved and used for only one frame, and then discarded, 2) the maximum number of hops that FI messages can propagate is bounded, and 3) MS-ALOHA proposes an extension of the busy bit in the FI message, in order to distinguish a free slot from a collided slot.

In RR-ALOHA, when an access attempt fails, the requesting node's 1-hop neighbors may end up in an inconsistent state. For example in Figure 2.22, the simultaneous reservation request from node 3 and 4 collided at node 7. This event causes node 3, 4 and 7 to mark the slot as "free". But for the requesting node's other 1-hop neighbors, e.g., node 1, 2 and node 5, 6, since they are unable to know the result of the reservation, they still mark the slot as reserved by node 3 and 4 respectively.

Such an incorrect slot status may be propagated in the network and causing further inconsistent and incorrect slot status. For example, the FI message from node 6 which marks the slot as reserved by node 4, may be propagated to node 7. Since node 7 believes that this slot is "free", it accepts the new allocation information from node 6 and further propagated it to node 3, 1 and 2. Because node 1, 2 and 3 previously mark the slot reserved by node 3, they now realize that a collision occurred and mark the slot as free. In the end, the slot status perceived by the network members gradually deviates from the reality, which may pollute further slot reservations.

The authors of the MS-ALOHA protocol believed that the FI message which contains the slot assignment should not be propagated beyond 2 hops. In MSA-

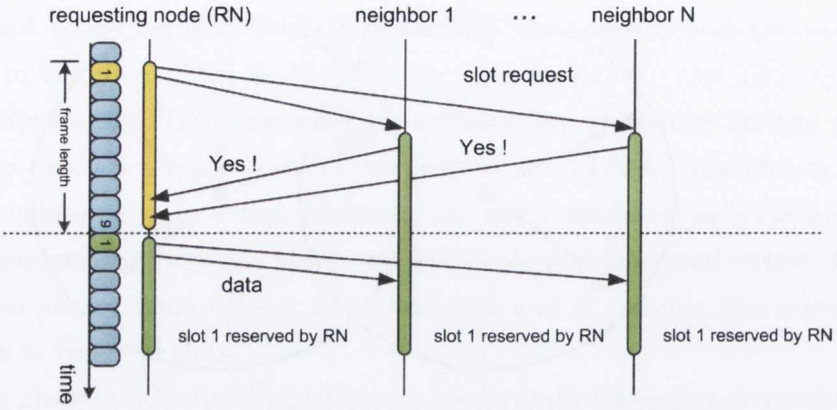

Figure 2.23: RR-ALOHA inconsistent slot status problem (a)

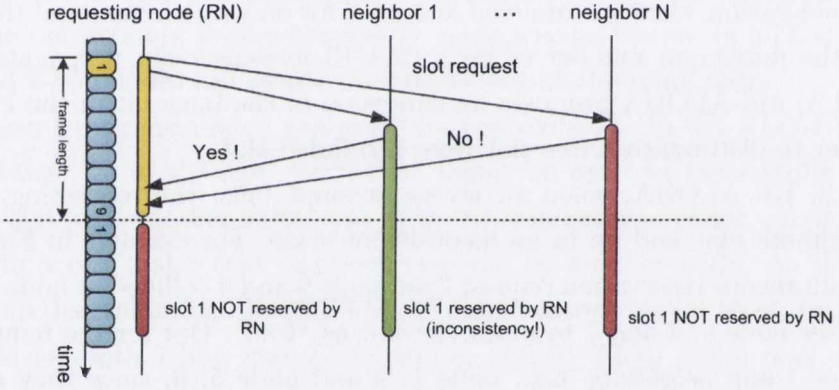

Figure 2.24: RR-ALOHA inconsistent slot status problem (b)

LOHA, a *de facto* hop count is added to each entry of the FI message, which effectively bounds their propagation. In addition, MS-ALOHA extends the slot status bit from 1 (represents reserved / free) to 2 (free / reserved / collision), in order to distinguish a genuine "free" and a "collision" slot status. With this amendment, node 7 in Figure 2.22 (with status "collision") would now disregard the slot status received from node 6 (with status "reserved by node 4") and stop propagating incorrect slot status further.

The amendment proposed by MS-ALOHA actually reveals a more profound issue in RR-ALOHA - the inconsistent slot status caused by unsuccessful slot reservations. As depicted in Figure 2.23, the RR-ALOHA protocol works correctly if the slot reservation request is successful, i.e., every 1-hop and 2-hop neighbors of

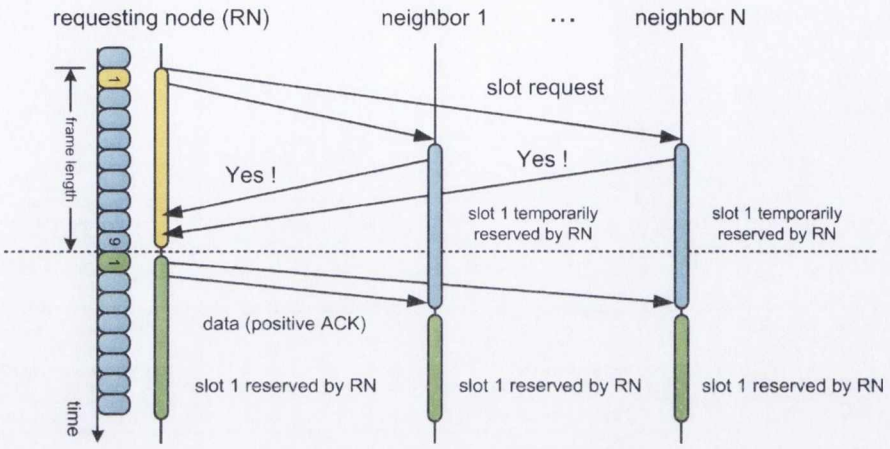

Figure 2.25: RR-ALOHA inconsistent problem 2-phase solution (a)

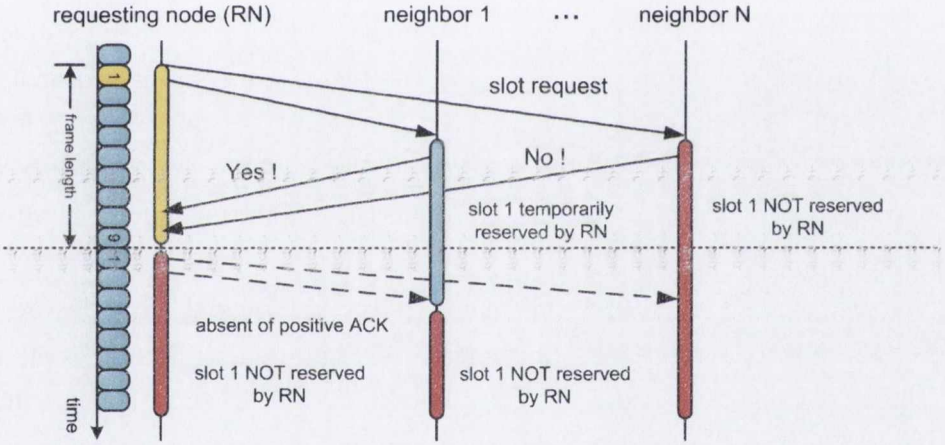

Figure 2.26: RR-ALOHA inconsistent problem 2-phase solution (b)

the requesting node (RN) assign the slot consistently. However, if the slot request is not successful, the opinion of the slot allocation among the neighbors begins to divide as depicted in Figure 2.24. 1-hop neighbors that do not experience an access collision mark the slot as reserved by RN, while those that detected collisions mark the slot as "free" (RR-ALOHA), or "collision" (MS-ALOHA). Such an inconsistent slot status will cause further slot allocation confusions and compromises the correctness of the protocol.

In RR-ALOHA, the slot bitmap is used to achieve 2-hop consensus, which guarantees collision-free broadcast, but the inconsistent status problem is not

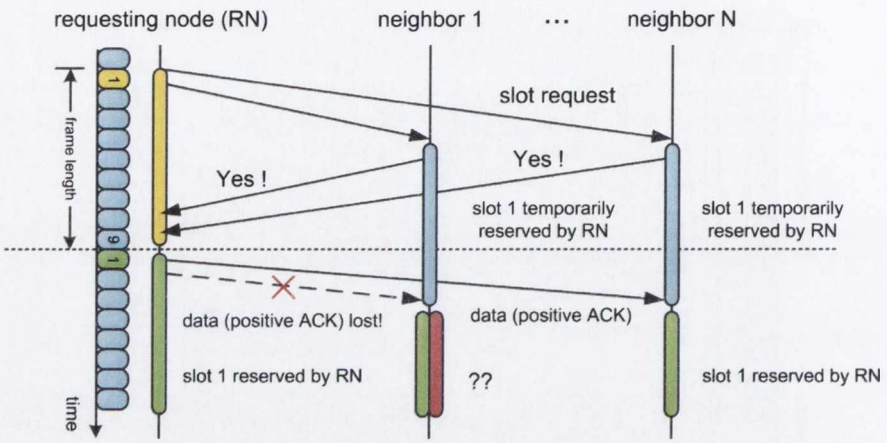

Figure 2.27: RR-ALOHA inconsistent problem 2-phase solution (c)

properly considered in the design. Actually, the inconsistency problem in RR-ALOHA might be mitigated by using the two-phase commit protocol, provided that nodes do not fail and messages do not get lost. As illustrated in Figure 2.25, when a node acknowledges a slot reservation, it temporarily (rather than permanently) marks the slot as reserved by the requesting node. If all neighbors agree on this slot request, the requesting node sends out its data, i.e., a positive acknowledgment, in the second phase (Figure 2.25), and all nodes permanently mark the slot as reserved. As depicted in Figure 2.26, if there exists any node that does not acknowledge the slot allocation, data message is not sent from the requesting node, and all neighbors, regardless of their respective status on the current slot (not reserved or temporarily reserved), consistently mark the slot as not reserved.

However, the two-phase commit solution cannot cope with the scenario where messages can get lost. For example in Figure 2.27, if the data message, i.e., positive acknowledgment from a successful requesting node is lost, the neighbor cannot distinguish between the scenario where "the slot reservation is successful but the acknowledgment message is lost", from "the slot allocation is unsuccessful". Any action taken by such a confused neighbor may cause further inconsistency in the neighborhood. In fact, in an environment with unreliable communication, i.e., messages can be lost, consensus cannot be achieved between two or more entities (the two generals' problem). The intuition is that the last message to

achieve a consensus might be lost and needs to be acknowledged, thus there is no such "last message".

If the constraint is lifted to allow node failure, the "three-phase commit" protocol need be used to ensure a consistent slot status, which is even more complicated and costly than the two-phase commit solution. It is worth pointing out that, generally speaking, it is very costly, if not impossible to achieve a consistent state in a distributed environment, especially when communications are not always reliable. The efforts to generate and maintain a 2-hop slot bitmap in RR-ALOHA is time-consuming, message-intensive, and without guaranteed correctness, which is probably not a good option in VANETS.

In essence, the purpose of the 2-hop slot bitmap as well as other "notification" mechanisms in protocols, such as FPRP, RBRP and CATA is to blockade one and two hop neighbors of the winning node from attempting for the same slot in subsequent frames. This is an important requirement for protocols in which slot allocations are repeated in more than one frame. Protocols of this kind use various mechanisms e.g., 2-hop slot bitmap in RR-ALOHA, to achieve this goal, i.e., to explicitly notifies the ownership of a reserved slot, and to force potential nodes to give up any further attempt. The *de facto* busy tone used in FPRP, ABROAD and CATA blocks potential rivals' attempt on previously reserved slot, while in RBRP, 1-hop neighbors of the slot owner use the 1-hop slot table to effectively veto any reservation request from 2-hop neighbors.

Nevertheless, if a slot is contested and used for only once, then there is no need for such a notification process or to maintain a consistent state among a node's neighbors. The "memoryless" slot status resembles the way how the contention-based protocols use the medium, and is especially desirable in a network with unstable links and dynamic network topologies. It is also the rationale behind the proposed MAC protocol in this thesis.

2.6 Summary

In this chapter, a large number of MAC protocols in the domain of wireless communications are reviewed. Starting with some of the pioneering protocols, such as MACA and FAMA, this chapter presented a broad and in-depth analysis

on 802.11p, topology-transparent scheduling protocols, as well as a wide range of schedule-based protocols.

For those safety-related applications that are envisaged in the future VANETs, the reliability and timeliness of medium access are of paramount importance. On this regard however, it is demonstrated in the review that both contention-based and schedule-based MAC protocols are inherently flawed, and are unable to satisfy the communication requirements specified by these applications.

Nevertheless, by reviewing the previous literatures, valuable insights are obtained with respect to the problem of medium access control. As wireless medium is a location-specific resource, a better understanding of the surroundings of a vehicle, i.e., being context-awareness in terms of the number of neighbors or the network topology in its vicinity, can potentially increase the effectiveness and efficiency of a MAC protocol. In addition, the predictability of vehicle trajectories is a unique advantage of VAENTs, which can be exploited to extend the concept of context-awareness into the forthcoming future. Based on these insights and observations, a new medium access control protocol is proposed in this thesis and is discussed in details in Chapter 3.

Chapter 3

Design

As discussed in Chapter 2, the approach used by the MAC protocols in mobile ad hoc networks can be divided into two categories: contention-based and reservation-based. For contention-based protocols such as 802.11p, the state of the environment e.g., the number of neighbors and their transmission schedules are not maintained in each node. These protocols are therefore simpler to implement and are less susceptible to the network dynamics. However, the down side of the contention-based approach is that it is intrinsically probabilistic with non-deterministic results. It is difficult, if not impossible, to provide guarantees on the perceived communication QoS, such as packet delivery ratio or end-to-end delay, which is vital for safety applications in VANETs.

Compared to the contention-based approach, the reservation-based approach, e.g., RR-ALOHA, is a "heavy-weight" method for allocating wireless resources. In such an approach, the status of one's neighbors is usually maintained and exchanged with considerable communication overhead. However, the benefit of the extra effort is improved communication reliability, as well as deterministic and predictable access to resources, which makes reservation-based approach more favorable for safety applications with tight QoS requirements.

However, conventional reservation-based MAC protocols have difficulties in VANETs because of their dependency on the specifics of the network topology, which changes at a rapid pace in vehicular environments. Although reservations can be rescheduled and updated according to the changed topology, "schedule gaps", in which old schedules are not applicable while the new schedule is not

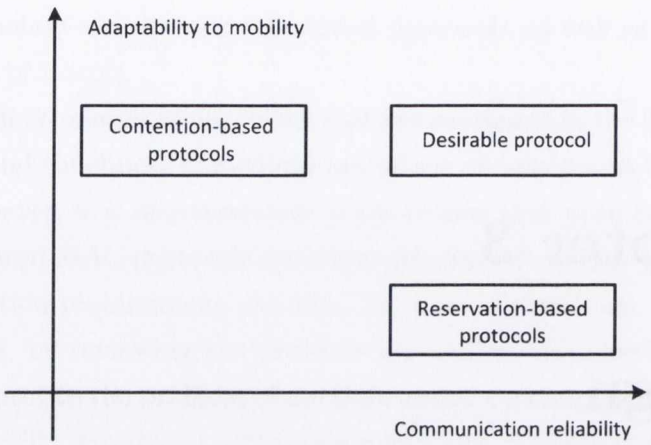

Figure 3.1: Protocols's adaptability to mobility and communication reliability

ready, are inevitable.

In Figure 3.1, a qualitatively comparison is given regarding the contention-based and reservation-based protocols in terms of their abilities to handle mobility and their abilities to provide communication QoS. Generally, contention-based protocols, e.g., 802.11, are less affected by topology changes of the network, but is also less capable of providing reliable and guaranteed communication. On the contrary, reservation-based protocols, e.g., RR-ALOHA, are able to provided reliable and deterministic communication, but have difficulties coping with network dynamics. A desired protocol for VANETs would ideally possess both properties, i.e., the capability to provide high level of communication QoS, and the capability to adapt dynamically when network topology changes.

As the non-deterministic nature of the contention-based approach is fundamental and difficult to circumvent, we believe that the reservation-based approach has the most potential and is therefore chosen as our design basis. The difficulty of applying conventional reservation-based protocols directly in vehicular networks is that the transmission schedules that are generated based on the current network cannot keep up with the fast-changing environment, which may cause undesirable schedule gaps.

In this thesis, A proactive scheduling mechanism, i.e., pre-scheduling is proposed to address this issue. The defining property of a pre-scheduling scheme is that the resources are allocated based on the future rather than the current

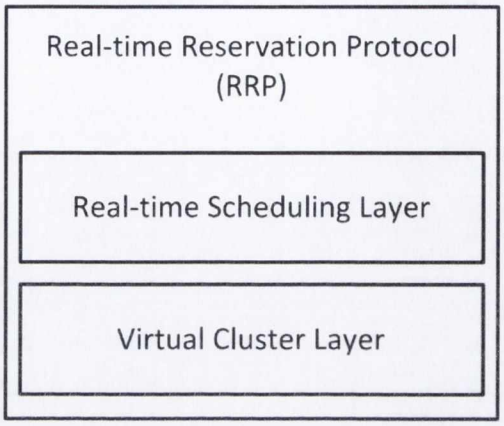

Figure 3.2: RRP architecture

network topology. A new MAC protocol termed *Real-time Reservation Protocol (RRP)* is proposed in the thesis based on such a concept. It is demonstrated later in the thesis that if the wireless resources are sufficiently abundant in a local area, and the wireless channel conditions conform with a statistical distribution, it is feasible for RRP to provide reliable communication (with over 90% packet delivery rate), and time-bounded medium access delay in a vehicular environment. As we discussed in Chapter 1, such properties, i.e., reliable and time-bounded medium access are of great importance to the successful operation of most safety-critical applications proposed in VANETs.

In the rest of this chapter, the concept of pre-scheduling and an overview of the RRP protocol are presented in Section 3.1. The detailed design of the RRP protocol is described in Section 3.2 and Section 3.3, which introduces the lower and the upper layer in the RRP architecture respectively. Conclusions and discussions of this chapter is presented in Section 3.4.

3.1 Overview of RRP

The real-time reservation protocol aims to provide reliable and real-time medium access control in VANETs with a *pre-scheduling* design philosophy. The fundamental task of a reservation-based MAC protocol is to allocate resources, e.g., time slots, among neighbors, which can be roughly defined as nodes that are

physically close to each other. In a dynamic VANET environment, constantly and rapidly changing neighbor relations makes scheduling difficult. To counter such adversity, the key concept of pre-scheduling is to schedule resources with future neighbors rather than current neighbors. There are two steps involved in this process: 1) a node first identifies its future neighbors, and 2) a node then negotiates future slot allocations with these identified future neighbors. The advantage of the pre-scheduling approach is that nodes are given a period of "preparation time" to perform slot allocation which will be used in the future. Such a period of preparation time is not available in conventional reservation-based protocols, and is therefore the defining characteristics of the proposed protocol. In the pre-scheduling scheme, each slot is allocated prior to its actual usage, and slots do not form repetitive frames.

One of the key requirements in the pre-scheduling scheme is to identify one's future neighbors in advance. However, there are a number of challenges need to be addressed to achieve this goal. For instance: a) How to define future neighbors? On one hand, future neighbors are not necessarily currently located in one's vicinity, and on the other hand, one's current neighbors are not necessarily future neighbors. b) How to exchange messages with them? and c) How to allocate slots with them?

In this thesis, the first two tasks are addressed by the lower part, while the third task is addressed by the upper part of a two-tier architecture, as depicted in Figure 3.2. The lower part, which is termed virtual cluster layer, manages the dynamics of the network, while the upper part, termed real-time scheduling layer, manages slot allocation. Such a design follows the principle of separation of concerns, which exempts the upper scheduling layer from worrying about the specifics of the network structure, and reduces the complexity of the scheduling algorithm.

In this architecture, the real-time scheduling layer only needs to focus on the slot allocation among identified neighbors, which makes it easier to achieve certain properties such as mutual exclusion and guaranteed reservation delay. The upper real-time scheduling layer are not aware of the information regarding the specifics of neighbors such as their position, velocity, or transmission power and so on, nor the exact time when a prospective neighbor becomes a current one.

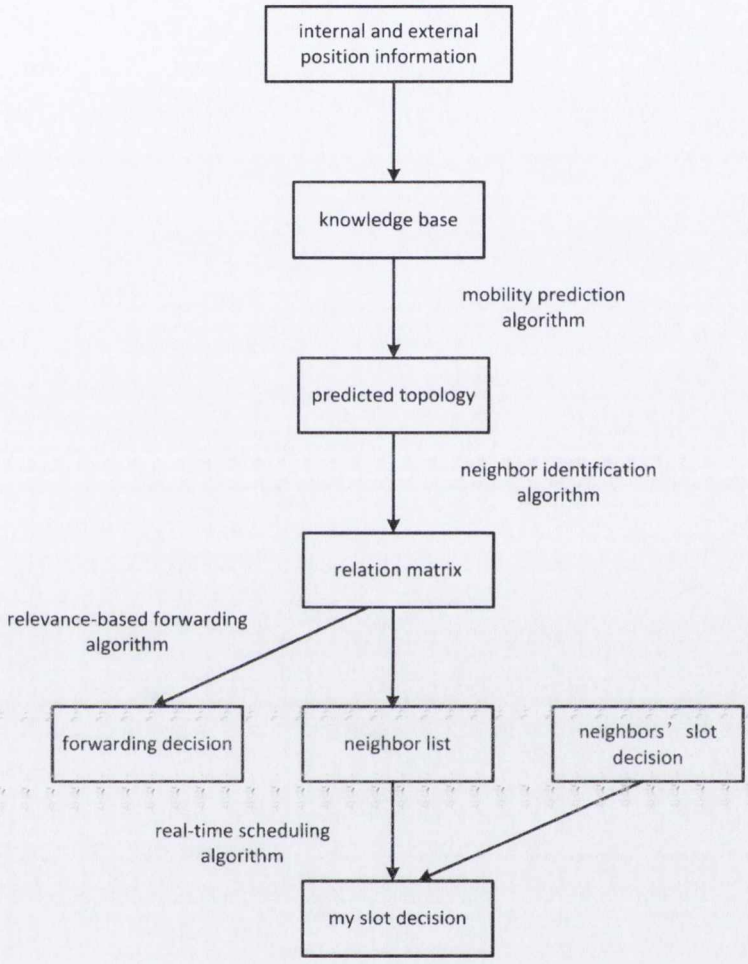

Figure 3.3: RRP design architecture

As far as the scheduling layer is concerned, the lower layer provides a neighbor list and a communication interface to send and receive messages with these neighbors. Consequently, the slot allocation algorithm is effectively transparent to the details of the underlining network topology.

The principle of separation of concerns in the RRP architecture is mainly achieved by the lower layer, which creates an appearance that all neighbors are always located in a cluster from the upper layer's point of view. Since such a cluster is not a physical cluster but an abstraction, the lower layer is termed "virtual cluster layer". The output of the lower layer is a list of future neighbors,

and an interface to send and receive messages for them. The specific tasks of this layer are to identify a future neighbor, which involves obtaining all the specifics of a node's kinetic information and transmission parameters, and to disseminate messages among these neighbors in order to create communication channels for the upper layer.

The main data structures and algorithms of the RRP protocol are depicted in the flow diagram in Figure 3.3. The algorithms in RRP are called when a new event occurs, such as a newly-arrived message, or an elapsed slot boundary. The RRP interacts with upper application layer by providing a number of reserved slots, in which an application can send and receive data.

In the RRP protocol, each node disseminates and receives positioning messages from its neighbors, and stores such information in a local repository termed the "knowledge base". Using mobility prediction algorithms, the predicted topology is estimated and is used as input for the neighbor identification algorithm. The algorithm calculates the neighbor relations among nodes in the knowledge base and uses the results in a message forwarding algorithm, which determines whether a received message should be relayed further.

In addition, the neighbor identification algorithm provides a neighbor list to the upper scheduling layer. Once the scheduling layer receives the neighbor list, the slot allocation algorithm determines whether a specific slot can be reserved according to the current node's own intention as well as other neighbors' decisions. The final slot decision, i.e., which slot is successfully reserved, is notified to the application layer above the RRP.

Algorithms used in both virtual cluster layer and the real-time scheduling layer are discussed in detail in the following sections.

3.2 Virtual Cluster Layer

This section focuses on the design issues related to the virtual cluster layer. The virtual cluster layer resides at the lower part of the RRP design architecture as depicted in Figure 3.2, with the main objective to identify and communicate with future neighbors. This task is divided into three subtasks: neighbor identification, relevance-based forwarding and node mobility prediction, which are discussed in

detail in Section 3.2.1, Section 3.2.2 and Section 3.2.3 respectively. Conclusions of this section and the mechanism to interact with the scheduling layer is presented in Section 3.2.4.

3.2.1 Neighbor Identification

As far as a MAC protocol is concerned, a "neighbor" may cause interference and should not transmit simultaneously with the node in question. However, the specific definition of a neighbor is not clearly defined. For instance, it is not defined how to measure such interference inflicted on the node of interest, nor how to determine a threshold for the interference above which a node is categorized as a neighbor.

Suppose that a node has a number of intended receivers if it sends out a broadcast message. The neighbors of this node is defined as those that have the potential to cause the Signal-to-Noise Ratio (SNR) to drop on the intended receivers to a certain extend. It is worth noting that in the above definition, a node may be categorized as a neighbor as long as it has the "capability", or the "potential" to cause interference, without "actually" transmitting any signal and causing any interference.

The fundamental goal of a MAC protocol is to regulate the use of the wireless medium, and to avoid interference between neighbors. Consequently, to recognize the identity and location of one's neighbor may benefit the MAC protocol to avoid simultaneous transmissions.

- How are neighbors identified?

A number of methods are available to identify a neighbor. These methods can be divided into two categories: real-world testing and theoretical estimation. For instance, the carrier sensing mechanism used in 802.11 is a type of real-world testing method to detect the existence of any neighbors in the vicinity. Such a method however, cannot detect 2-hop neighbors, which causes the notorious "hidden terminal problem". Other methods proposed in the literature are able to identify neighbors in the 2-hop range by broadcasting probe messages. If there is a 1-hop or 2-hops neighbor that also transmits at the same time with the probe

sender, the probe message suffers a collision and is not received. By collecting the feedback from all 1-hop receivers, the probe message sender is able to deduce the existence of any neighbors in its 2-hop proximity, and create transmission schedules to avoid colliding with these neighbors.

The theoretical estimation approach usually appears in the reservation-based MAC protocols. It is often assumed that the network topology is fully known and the transmission range is simplified as a radius. Generally, nodes that are located within 2-hops of a given node are considered as neighbors, while nodes beyond the 2-hop boundary are non-neighbors. Such "2-hop" estimation of neighbors simplifies the neighbor identification process, but the binary identification is obviously not realistic.

The proposed proactive scheduling scheme has a unique requirement regarding neighbor identification: future neighbors are of interest as well as the current neighbors. Note that such a future neighbor may not be a current neighbor when a neighbor identification procedure is invoked, therefore the real-world testing approach cannot be used to identify such non-current future neighbors. Based on this observation, the theoretical estimation approach is chosen in this thesis to identify a potential future neighbor.

- Neighbor Index - Basic concept

The main problem facing the neighbor identification process is that which metric should be used to quantitatively define a neighbor. To address this question, the definition of a neighbor is revisited in the following.

According to our previous discussion, the key property of a neighbor is its capability to cause interference with another node's broadcast. To demonstrate this idea more clearly, we visualize the packet reception probability in a wireless environment for those transmissions in Figure 3.4. In this diagram, the packet reception probability is calculated using deterministic propagation models and represented by different colors at various locations with regard to the transmitter. Specifically, the distance between two senders are 1000 meters and 200 meters in the upper and lower diagram respectively. The transmission power is 20 mW, packet size is 4k bits, and the bit rate is 6Mbps. The packet reception probabilities

are represented by the following colors: white: 100%, green: 90% - 100%, yellow: 80% - 90%, red: 50% - 80%, grey: 1% - 50%, and black: less than 1%.

As illustrated in Figure 3.4, when two transmitters move towards each other, their capability to deliver messages to their respective neighbors gradually diminishes, as the white area shrinks substantially. From the distance at which these two nodes do not interfere with each other at all (Figure 3.4: upper) to the distance at which they have a clear and negative impact on each other (Figure 3.4: lower), there is no "tipping point" such that their relationship flips from "non-neighbor" to "neighbor" with regard to each other. The diagram clearly shows that the influence of a neighbor is a progressive and non-binary property, and the tipping point is usually not at the 2-hop boundary as assumed by most reservation-based MAC protocols.

- Neighbor Index - Definition

In this thesis, a "neighbor index" is proposed as a metric to characterize the severity of interference between two nodes that transmit simultaneously. A higher neighbor index means higher interference from the neighbor, which indicates a stronger "neighbor". Neighbor index is a non-binary metric, as the SNR degradation is continuous. In addition, it can describe the current as well as the predicted future relation between two nodes.

The value of neighbor index characterizes whether two nodes can co-exist without causing interference on intended receivers. A large neighbor index indicates high probability of interference and message loss, therefore simultaneous transmissions by such node pairs should be avoided. The essence of the neighbor index is to calculate the message reception probability loss, by comparing the reception probability with and without the presence of the interfering node.

Suppose that we have a signal node s and a receiving node i, and know the following information: the distance $d(s, i)$ between s and i, transmission power of the signal node $P(s)$, and all other relevant parameter set H, such as modulation methods and packet size. Using the deterministic two-ray ground propagation model, given that the transmission is free from any interference from other nodes, the probability that a message of a certain size sent from signal node s can be

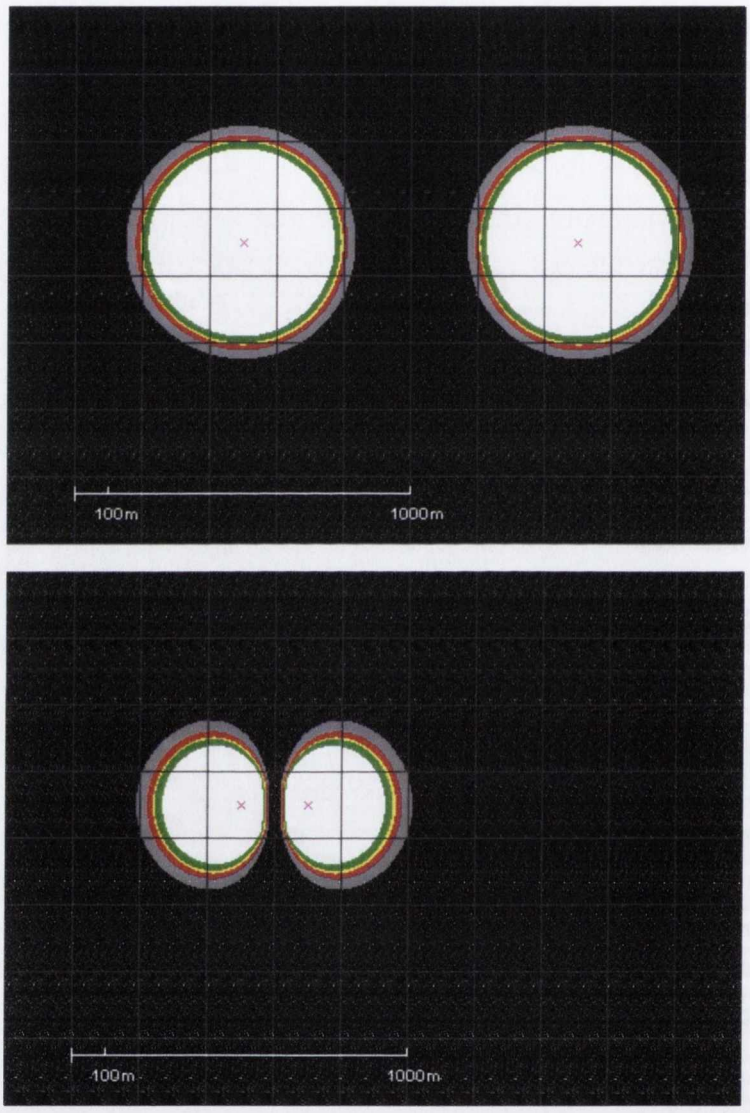

Figure 3.4: Visualization of neighbor interference on packet reception probability

successfully received at receiving node i is represented as: $p_{i-free}(d(s,i), P(s), H)$, i.e., the probability of node i receiving a message from node s without interference.

Now suppose that an interfering node n is near the signal node s and interferes with receiving node i. We assume that the distance between the interfering node n and the receiver i is $d(n,i)$, the transmission power of the interfering node n is $P(n)$, and all other related parameters are the same (H). The probability of

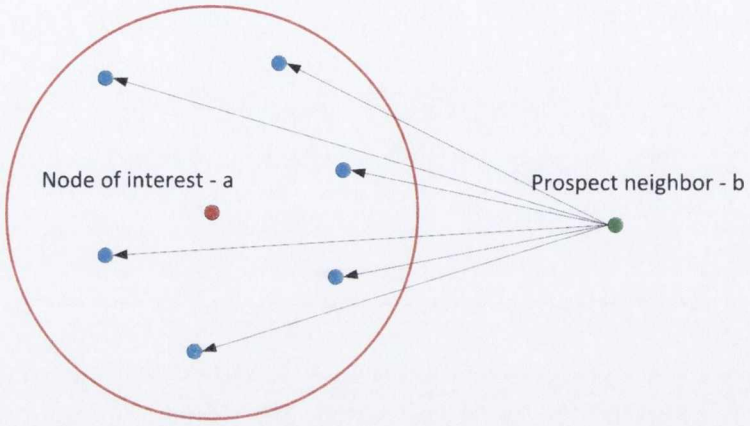

Figure 3.5: Neighbor index calculation

a message being received at r under interference from node n is represented as: $p_{i-interference}(d(s,i), d(n,i), P(s), P(n), H)$.

Assume that node s's broadcast message can be received by a set of recipients $R = \{r_1 \ldots r_m\}$ provided that no interfering node exists. The neighbor index between node s and n is defined as:

$$NeighborIndex(s,n) = \sum_{i=1}^{m} p_{i-free} - p_{i-interference} \tag{3.1}$$

As illustrated in Figure 3.5, the essence of the neighbor index is to calculate the cumulative reception probability loss caused by the interfering neighbor. A stronger interfering signal or a larger number of recipients both increase the neighbor index value, indicating a more "dangerous" interfering node facing the signal node. Nodes with high neighbor index, i.e., dangerous nodes, are identified as neighbors and are excluded from using the same time slots.

The neighbor index metric is a numerical characterization of two nodes' relationship. The calculation of neighbor index takes the information of the node pair as well as other nodes in the vicinity regarding their distance, transmission power and other parameters, to determine the potential "loss" caused by the neighbor. Due to the fact that the neighbor index takes more parameters into account, the identification process is more realistic and accurate than the static 2-hop radius classification method. In addition, it can be used to predict the rela-

tionship between two nodes, as long as their trajectory are known. This property is particularly useful for our proactive scheduling approach.

The output of the neighbor index calculation is a numerical value. The neighbor identification process utilizes a threshold as the last step in determining if a node is a neighbor or not.

3.2.2 Relevance-Based Forwarding

To clarify our description, we first define the concept of message forwarding and distinguish it from message broadcasting. We refer to "message broadcasting" as a mechanism to propagate messages within the radio range of a node, which is a primitive operation in wireless communication. On the other hand, "message forwarding" is defined as the propagation of messages over a larger area, which usually spans several radio hops.

In generic mobile ad hoc networks, message forwarding is often seen as the building block for routing protocols. For example, routing protocols such as Dynamic Source Routing (DSR) Johnson & Maltz [1996] and Ad hoc On Demand Distance Vector (AODV) Perkins & Royer [1999] rely on message forwarding to establish routes. In vehicular networks, information obtained from on-board sensors such as GPS positions and velocities is useful not only for the vehicle itself, but also important for other vehicles in its proximity. Such information can facilitate vehicles in close vicinity to adapt their behavior, which makes message forwarding a natural choice for a communication paradigm in a VANET.

Message forwarding is particularly relevant and important in the realization of the "virtual cluster" mechanism. The cornerstone of implementing a virtual cluster is to identify and communicate with future neighbors, as if they are physically located in an actual cluster. However, the future neighbors may be located far away in the physical world. Consequently, virtual clustering needs a message forwarding mechanism to transfer information such as the position, velocity, transmission power, or slot usage among vehicles that are physically outside of broadcast range of the sender and the intended receiver.

- Message forwarding in the literature

In the following, we introduce a number of message forwarding schemes from the literature. The simplest way of forwarding a message is called simple flooding Ho *et al.* [1999]. It starts with a source node broadcasting a message and all nodes that receive this message rebroadcast it only once. Simple flooding can achieve reliable message forwarding Ho *et al.* [1999], but with a large number of duplicate messages that congests the network and causes what is known as the "broadcast storm problem" Tseng *et al.* [2002].

The probabilistic flooding mechanism is similar to the simple flooding, except that a node only rebroadcasts the received message probabilistically. For example, Tseng *et al.* [2002] propose to select rebroadcast nodes randomly in order to reduce the number of redundant messages. In a counter-based scheme Tseng *et al.* [2002], the probability of rebroadcasting a message is determined by a counter, which records the number of times that the same message has been received. If the value of the counter is below a threshold, it indicates that the node is more likely to reach additional area with a rebroadcast. Such a counter-based scheme is simple but can adapt to local topology in the sense that, in dense networks fewer nodes will rebroadcast, while in sparse networks more nodes will rebroadcast.

Position-based message forwarding utilizes the position of the forwarding node to make forwarding decisions Tseng *et al.* [2002] and Briesemeister *et al.* [2000]. If the distance between the forwarding node and the previous forwarder of this message exceeds a threshold, the message is rebroadcast. Further more, the additional covered distance Tseng *et al.* [2002] and Tonguz *et al.* [2007] as well as the novelty of the forwarded message Wegener *et al.* [2007] can also be utilized as a forwarding criteria.

The aforementioned message forwarding schemes do not answer a fundamental question: how far should a message be propagated? For a real-world network such as a VANET, it is not feasible to propagate a message to the entire network, as some message forwarding protocols suggest, or simply mandate a fixed range, beyond which the message should stop being forwarded onwards as suggested by some position-based forwarding schemes.

- Message relevance

To address the questions above, a novel message forwarding scheme is pro-

posed which relies on the remaining "relevance" of a message to make forwarding decisions. Prior to the detailed description of the new scheme, we discuss the concept of message relevance in the context of vehicular networks.

A fundamental question regarding message forwarding in general is: for a specific message, who should receive it? To answer the question, a concept called "relevance-zone", which is proposed for safety applications in VANETs, is of particular interest. As far as a vehicular safety application is concerned, a warning message that indicates a road hazard, such as an emergency breaking notification, is only relevant to nearby vehicles. In other words, each warning message in a vehicular network has a unique and geographic relevance zone. The warning message is of little or no use for vehicles beyond the relevance zone, and thus a warning message forwarding algorithm should terminate the message forwarding process at such a boundary.

Inspired by this idea, we observe that for a MAC protocol, a message containing the sender's position, velocity and slot usage has its relevance zone as well. The information contained in the message is important to identify future neighbors and schedule time slots, if and only if the recipient of the message is a potential neighbor of the message sender. From another perspective, a message has its highest relevance to its immediate neighbors, and as the message is forwarded away from the sender, its relevance gradually diminishes. Whenever a forwarding node decides that the message has no relevance to its potential recipients, this message has reached its relevance boundary and should not be forwarded again.

Based on the description of the message relevance concept, the proposed relevance-based forwarding mechanism is described as follows: whenever a message is received, a node determines the relevance of the message with respect to the prospective rebroadcast recipients. If the message is relevant to at least one recipient, the message is rebroadcast, otherwise it is discarded.

Consequently, the question is how to determine the relevance of a message with respect to prospective receiving node. In this thesis, the relevance of a message is characterized by the neighbor index between the source of the message and the prospective receiver. The rationale is that, from a receiver's point of view, a message coming from its identified neighbor should be received and therefore by

definition, such a message is "relevant".

After the calculation of the neighbor index, if the message forwarder believes that at least one prospective message recipient (assume that the forwarder rebroadcasts this message) will be a neighbor of the message originator in the future and therefore should receive this message, the forwarder rebroadcasts this message, otherwise the message is discarded. In the proposed relevance-based forwarding scheme, all messages will eventually reach their relevance boundaries and become irrelevant after being propagated from their sources. In other words, the dissemination of messages is geographically bounded via relevance-based forwarding.

In the proposed relevance-based forwarding algorithm, positioning messages and scheduling messages are bundled together if they are created by the same node, and are assembled, transmitted, received and interpreted as a whole during their lifetime. This rationale is based on the observation that, as both messages describe the message source, either both messages are "relevant" to the recipient, or none of them are. Therefore, it is unnecessary to forward position-related and schedule-related messages separately.

- Relevance-based forwarding

In the relevance-based forwarding algorithm, it is assumed that, by receiving messages from others, a node is aware of the position of its neighbors (including 1-hop, 2-hop and beyond if necessary), as well as their transmission power and other transceiver-related parameters. Consequently, it is possible for a node to calculate the following:

1. a set of nodes that also received the message from the last forwarder
2. a set of nodes that will receive the message if the node itself rebroadcasts
3. the neighbor index between the message source (not the last forwarder) and the prospective recipients if it rebroadcasts

Upon receiving a message, a node initially adds all nodes in its communication range into an "intended recipients list". Following a step-by-step process of

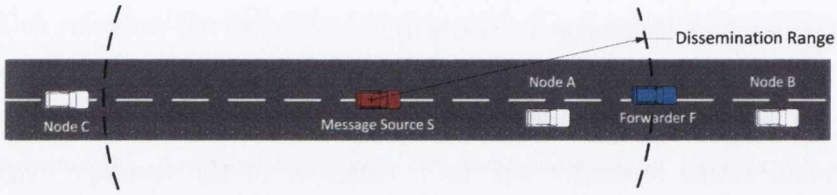

Figure 3.6: Disseminating a message further from its source

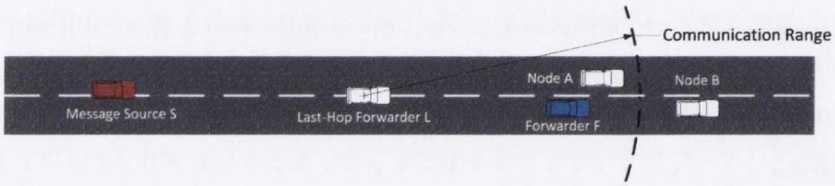

Figure 3.7: Avoid re-sending a message to nodes that have previously received

the relevance-based forwarding algorithm, nodes that are deemed unnecessary to receive this message are removed from this list. At the end of the elimination process, if there is at least one node in the intended recipients list, the message is rebroadcast, otherwise, the message is discarded. In the following, the three-step elimination process is briefly introduced, followed by a pseudo-code specification of the relevance-based forwarding algorithm.

Step 1: propagate further. Candidates: all nodes in the communication range of the forwarder. Goal: propagating the message away from the message source. Example: As depicted in Figure 3.6, the forwarder F should forward the message from source S to node B, but not node A or Node C.

Step 2: Eliminate already informed recipients. Candidates: all remaining nodes in the intended recipient list after step 1. Goal: eliminate nodes that have already received the message. Example: As depicted in Figure 3.7, the forwarder F should relay the message from source S to node B rather than node A, since it may have already received it from the last-hop forwarder L.

Step 3: Eliminate irrelevant recipients. Candidates: all remaining nodes in the intended recipient list after step 2. Goal: eliminating recipients that are irrelevant to the message source. Example: As depicted in Figure 3.8, the forwarder F should relay the message from source S to node A rather than node B, since

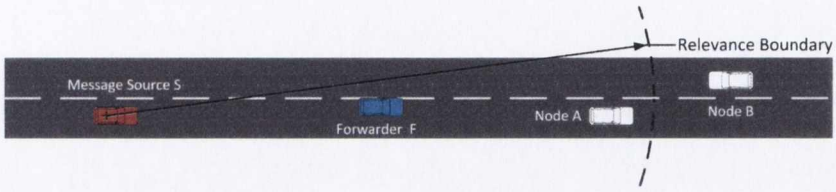

Figure 3.8: Bound the message dissemination area by message relevance

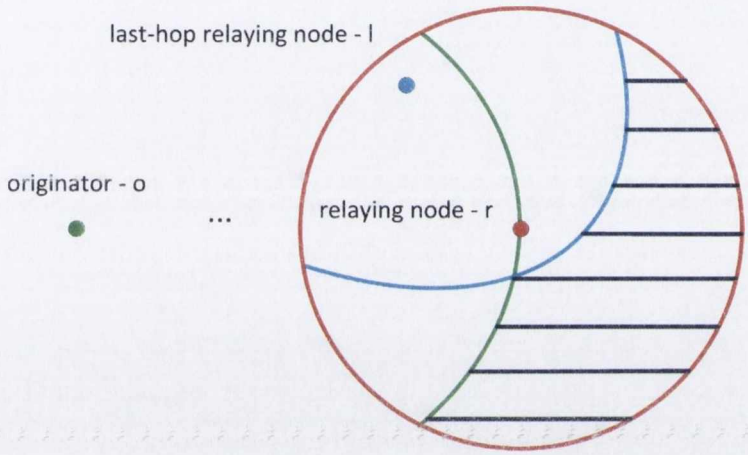

Figure 3.9: Message forwarding scenario

source S' status (position and slot reservation) is deemed irrelevant to node B.

In the following, the relevance-based forwarding algorithm is presented. As illustrated in an example scenario in Figure 3.9, the relaying node r receives a beacon message m, which originates from node o, from it's last-hop relaying node l. Now it needs to decide whether or not to rebroadcast this message. Before the detailed description of the forwarding algorithm, some notations and assumptions are outlined as follows:

The set of nodes that are located within node $i's$ broadcast range at time Δt are defined as: $B(i, \Delta t)$, $\Delta t \in (t_0, t_0 + max\ prediction\ interval)$, t_0 is the current time. Distance between nodes i and j is $d(i, j)$. Neighbor identification function $f(i, j, \Delta t)$: f determines whether two nodes i, j are neighbors at a predicted time Δt.

It is assumed that, for the relaying node r, the following information is known:

1) $B(r)$, $B(l)$ (since nodes in $B(l)$ are within 2-hops of node r)

2) $d(o,i)$, $i \in B(r)$ (based on propagated location information)

3) $f(o,i,\Delta t)$, $i \in B(r)$

Require: Assumption 1-3

Ensure: decide whether to rebroadcast message m

$R \leftarrow \emptyset$ ▷ initialize rebroadcast beneficiary node set R

for *each node* n_i $inB(r)$ **do**

if $d(o,i) < d(o,r)$ **then**

continue; ▷ node i is eliminated, as it is closer to the source than the relaying node

end if

if $n_i \in B(l)$ **then**

continue; ▷ node i is eliminate, as it has already received m

end if

if $f(o,i,\Delta t) < neighboridentificationthreshold$ **then**

continue; ▷ node i is eliminated, as it is not a neighbor of source node o in the foreseeable future

end if

$R \leftarrow n_i$ ▷ add node i to R

end for

if $R \neq \emptyset$ **then**

Rebroadcast m

else

Discard m

end if

The relevance-based forwarding algorithm has the following properties: 1) bounded dissemination area, 2) forwarding of messages towards message boundary (i.e., no bouncing of the message), and 3) duplicated message suppression to avoid broadcast storm problem.

3.2.3 Neighbor Mobility Prediction

In vehicular ad-hoc networks, nodes rarely move with complete randomness. Therefore, by exploiting the mobile node's non-random mobility pattern, it is

possible to predict the future state of the network topology Su *et al.* [2000]. A variety of protocols may benefit from knowing the future network topology, by adapting themselves to the network dynamics. The topic of mobility prediction has been extensively addressed in the literature, and is often regarded as a lower-layer service provided to upper-layer applications.

- Mobility prediction input

In order to predict the future position of a node in a VANET, four types of information are potentially useful: scenario context, driver intention and road restriction information, mobility model, and historical data such as trajectory and received signal strength.

The scenario context refers to the background information that is known as common sense or known a priori. For example, in a vehicular environment, the speed, acceleration and deceleration of a car are bounded by a maximum limit. The traffic scenario, e.g., urban or highway may determine if a car can maintain a constant speed for a relatively long time or not. In addition, it is impossible for a vehicle to pass through another vehicle within a single lane. These kinds of information may not lead to precise prediction results, but it can help us to narrow down the possibilities.

As for the driver intention and road restriction information, they are not widely adopted in mobility prediction in the VANET community yet, since such information is difficult and costly to access. However, its potential to increase prediction accuracy is tremendous. For instance, if a digital map and the driver's intention is known, it would be much easier to predict which way a car will go at a junction.

The topic of the mobility model has been extensively studied in the literature with various models being proposed. Two major categories are identified: single node mobility models and group mobility models. Representative single node models include: random walk, random waypoint, random direction, boundless simulation area, Gauss-Markov, and the Manhattan grid models. Group mobility model includes column, nomadic community, pursue, reference point, and reference velocity models. A comprehensive survey can be found in Camp *et al.* [2002] and Harri *et al.* [2009].

It is of vital importance for a mobility model to capture the essence of a moving node. In vehicular networks however, the validity of the above models are in doubt due to their generic nature. The evaluations conducted in Harri *et al.* [2009] confirmed that using any of the aforementioned mobility models produce completely useless results. A mobility model that can capture the genuine vehicle behavior is highly desirable for mobility prediction. However, it is difficult if not impossible to construct such a model for prediction, considering the driver's involvement and other factors such as road restrictions. For the time being, it seems that the only feasible mobility models in vehicular networks are simple ones: straight line, curve line with uniform curvature, constant speed, constant acceleration or deceleration.

Historical data refers to information regarding trajectories, positions, or received signal strength that is either observed by the node itself or provided by other nodes. In many proposed mobility prediction schemes, "hello" messages are periodically broadcast in a vicinity, in order to provide historical data in the neighborhood.

- Mobility prediction paradigms

Currently, the process of trajectory prediction is mainly based on previous records. A number of issues need to be considered:

- Is the prediction scheme based on coordinates, velocity, acceleration, or received signal strength? Or a combination of the above?
- If the previous records are received periodically, how many such records are used for prediction? Only one, two, five, or more?
- Among the records that are received at different times in the past, are they equally treated, or with a different weight, e.g. more focused on recent data?
- Which mathematical model is used to extrapolate future values based on previous records? A simple linear model, or a more advanced linear first-order autoregressive model, or Markov- chain?

Table 3.1: Mobility prediction paradigms

	Linear Trajectory	**Polynomial Trajectory**	**Weighted Velocity**	**Linear First Order**
RSSI Information	not considered	not considered	not considered	not considered
Coordinates	considered	considered	considered	considered
Velocity	assumed constant	assumed constant	varied	varied
Acceleration	not considered	not considered	not considered	considered
Number of records	2	3+	2	1
Weight of records	Equal	Equal	Weighted	N/A
Extrapolation model	Linear	Polynomial	Linear	First order autoregressive
Result type	Deterministic	Deterministic	Deterministic	Deterministic

- Are those predicted results deterministic or probabilistic?

Linear trajectory prediction is the simplest approach. The idea is to use two historic points on the trajectory to extrapolate the future point. The speed and direction of the node is assumed constant, thus a node follows a linear trajectory. A more complicated method utilizes more points on the trajectory, and predicts the next point using polynomial extrapolation. Therefore, a node may follow a non-linear trajectory which is much more realistic. Another approach Sharma *et al.* [2004] predicts a node's future position by using two previous velocities, where a tunable coefficient α is introduced to regulate the weight in between. Venkateswaran *et al.* Venkateswaran *et al.* [2005a] utilize a vector to describe a node's state, which includes coordinates, velocity and acceleration. The future state of a node can be calculated by multiplying the current state vector with a matrix A, and then adding another noise matrix Q. Matrices A and Q are both estimated based on training data. A summary of these mobility prediction paradigms are listed in Table 3.1.

- Mobility prediction inaccuracy problem

The inaccuracy of mobility prediction results from a variety of aspects which can be categorized into three types based on their origin. First of all, the location information obtained from sensors, e.g., GPS may have already been distorted at

the source. Secondly, in the transmission phase, packets which include geographic information could be lost or delayed, which may result in unavailable or outdated location information. Lastly, during the information processing phase, a number of prediction errors may occur resulting from an over-simplified mobility model, which is termed "realism error" Härri *et al.* [2008]. For instance, the linear trajectory and constant acceleration that is assumed in many mobility model. In addition, certain unpredictable events, such as human behavior will cause any mobility prediction scheme to fail. In this thesis, the consequences and implications of prediction errors in vehicular networks with various speeds are evaluated in Chapter 5.

- Applications of mobility prediction

The idea of mobility prediction has been applied in many routing protocols. Mobility prediction algorithms estimate the future trajectories of mobile nodes, in order to be aware of, and take proactive actions regarding the future network dynamics. In one-hop routing protocols, good data forwarders can be selected by evaluating prospective forwarders' movements; while in multi-hop routing protocols, mobility prediction is used to estimate the expiration time of a link as well as an entire route.

For example, as an one-hop routing protocol, Distributed Mobility-Aware Route Selection (DMARS) Abhyankar & Agrawal [2002] uses mobility vector to reflect a node's movement and its connectivity with its neighbors. Nodes with a shorter mobility vector magnitude are chosen as the data forwarding node as they are less likely to move drastically in the future. In a novel relaying node selection algorithm Kinetic Multipoint Relaying (KMPR)Härri *et al.* [2008], a node estimates all of its neighbors' Link Expiration Times (LET), which are the durations for which that two nodes are able to communicate with each other. If a link to a particular neighbor is predicted to break in the future, the degree of the node in question will be reduced by one. Consequently, the integration of the dynamic node degree with respect to time can be calculated as the "kinetic nodal degree". Nodes with higher kinetic nodal degree are chosen as relaying nodes, as they are likely to have more neighbors in the near future.

In multi-hop routing protocols, mobility prediction algorithms are mainly used to estimate a route's life time. As pointed out by Lee *et al.* [1999], as far as a specific multi-hop link is concerned, the Route Expiration Time (RET) depends on the weakest link on the route, i.e., the one with the lowest LET. In light of this observation, in Flow Oriented Routing Protocol Lee *et al.* [1999], the route expiration time is determined by the predicted least LET of the entire route. The estimation of route expiration time can also be used as a criterion when creating new routes. For example, in Distance Vector with Mobility Prediction (DV-MP) Su [2000], the route expiration time is included as an element in the routing table which helps to construct stable routes with longer survival times.

Mobility prediction is also applied in a number of clustering algorithms. In dynamic networks with fasting changing topologies, it is desirable for cluster algorithms to group nodes with mobility similarities, in order to form a more stable cluster structure and extend the resident time of each cluster member. With mobility prediction, a node can estimate the time that their neighbors are reachable and use such estimated resident time as an admission criteria when establishing a cluster.

In MOBIC Basu *et al.* [2001], nodes are grouped together based on their speed and direction, and the most "common" node is selected as the cluster head, which is relatively stable with respect to its neighbors. If a group mobility pattern does not exist however, it is difficult to construct a cluster based on common velocity and direction directly. An interesting algorithm is proposed in Wang & Li [2002] to address this issue. In this protocol, the physical space, in which nodes are actually located is converted to a velocity space, where nodes are represented given their velocity in both x and y axis. It is much easier to group nodes in the velocity space by using pattern recognition techniques, and any partitions in physical space can be predicted by recognizing different clusters in the velocity space, since mixed nodes will separate if their velocities are different.

A more advanced approach estimates the communication window of a node's neighbors as the cluster admission criteria. In Venkateswaran *et al.* [2005b] and McDonald & Znati [1999] for example, assuming that 2-hop neighbor knowledge, such as node ID, location and speed is known, a variety of mobility prediction schemes such as LET, or Linear First Order Autoregressive Model are used to

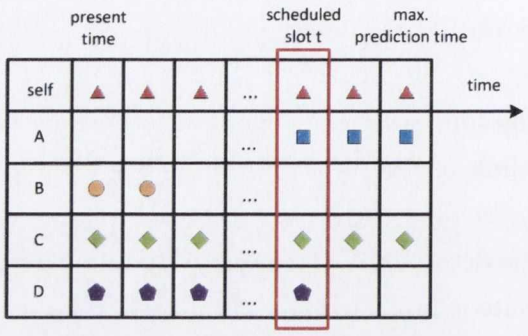

Figure 3.10: Neighbor table

calculate the communication window of a particular neighbor in the future. Based on the estimation, qualified neighbors which present long-term willingness to remain in the communication range will be chosen as cluster members.

- The approach used in RRP

In our design, no off-the-shelf mobility model is used. Nodes in the simulation environment in our evaluation study follow the driver model provided by the traffic simulator VISSIM, and they can accelerate, decelerate or change lanes. We use a simple dead reckoning algorithm to estimate a node's current and future position, based on the node's last known position and velocity.

3.2.4 Interactions with the Scheduling Layer

In virtual cluster layer, future neighbors of a node are identified by calculating the neighbor index between the node in question and its prospective neighbors in the predicted future network. An example of identified future neighbors is illustrated in Figure 3.10. In this diagram, if a box is filled with a shape, it means that a node is an identified neighbor to the current node, otherwise not. For example, node a is predicted to join the neighborhood of the current node, node b and d are leaving the current node, and node c will remain in the neighborhood of the current node.

3.3 Scheduling Layer

Based upon the information provided by the virtual cluster layer, the scheduling layer's responsibility is to allocate slots in a way that exclusive access to a slot is guaranteed. In addition, in the context of vehicular safety applications, the slot allocation algorithm requires certain properties such as fairness of allocation, adaptivity to the dynamic environment, and deterministic slot reservation. The above demanding requirements pose great challenges on the design of a slot allocation algorithm. Prior to the discussion of the proposed slot allocation protocol, some preliminaries regarding its goals, rules and assumptions are provided in the following.

The slot allocation algorithm takes input from the lower virtual cluster layer, and outputs the reserved slots (if any) represented by their slot numbers to the upper layer above the RRP protocol. In other words, the virtual cluster layer provides a list of estimated neighbors in a period of predicted future to the scheduling layer. For a specific slot, by running the slot allocation algorithm, a node eventually decides whether it can use this slot or not. Note that a node can only determine its own slot usage; it can not, and should not decide other nodes' slot usage.

The allocation algorithm must ensure collision-free behavior such that if the node successfully reserved a slot, none of its identified neighbors will use the same slot as well. In addition, the algorithm also needs to ensure consistency such that if a reservation decision is made, all of the deciders' neighbors draw the same conclusion. To satisfy the two requirements above, a general rule of slot allocation is applied. In order for a node to successfully reserve a slot, it needs to obtain explicit endorsement, i.e., positive acknowledgment, from all of its neighbors (as identified by the virtual cluster layer). However, if a node does not wish or fails to reserve a slot, the node does not have the obligation to know which neighbor has successfully reserved this slot.

The proposed slot allocation algorithm is fully-distributed. For each individual node, the whole network is composed of "me" (the node itself) and the "others" (the rest of the nodes in the network). The idea of the algorithm is actually quite simple: if a node obtains endorsement from all of its neighbors, it guarantees a

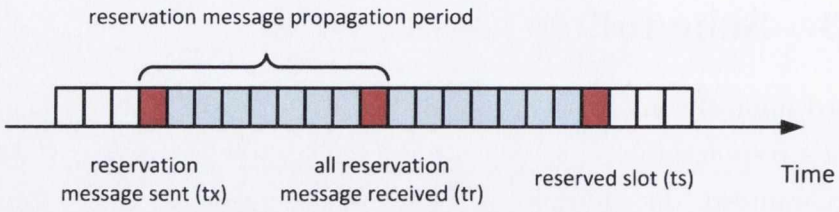

Figure 3.11: Timing of reservation messages

collision-free slot; otherwise, it simply waives this slot and does not care about it anymore. Based on this logic, the status of the slot is simply "win" or "lose" from a particular node's perspective. In this thesis, this type of consensus between a node and its neighbors is termed "weak consensus", which is different from a "strong consensus" where a node keeps tracks of all slot allocations, even for those slots that the node is not interested in.

The slot allocation algorithm uses priorities to arbitrate the slot contention. For each slot, a node generates a random priority and disseminates this number to its neighbors. If a node has the highest priority among all of its neighbors, it wins this slot. Based on the assumptions and rules that are discussed so far, a simple slot allocation protocol is presented in the following, which achieves collision-free and consistent slot allocation (but not guaranteed slot reservation).

3.3.1 A Simple Slot Allocation Protocol

In this section, a simple slot allocation protocol is presented. The slot allocation rules are as follows: for a specific slot, if a node has the highest priority among its neighbors, it wins the contention and successfully reserves this slot; otherwise, the node loses this slot. As a node quits the contention for this slot immediately after realizing the existence of a higher priority node, this approach is termed "non-sticky" reservation. It is worth mentioning that, such a "non-sticky" behavior is often seen in reservation-based MAC protocols, such as RR-ALOHA and FPRP. In these two protocols, the attempt for a slot is immediately lost when a concurrent attempt has been detected. The non-sticky approach guarantees that the node with the highest priority number among its neighbors always wins the contention, and preserves the collision-free property of the slot.

Table 3.2: Annotations of variables in simple slot allocation algorithm

Description	Annotation
Target node	n
Target slot	s
Neighbor set	$N(n,s)$ at slot s, of node n
Priority number	$p(n,s)$ at slot s, of node n
Neighbor's priority number	$p_i(n_i,s)$ at slot s, of node n_i, $n_i \in N(n,s)$
Priority number set of node n's neighbors	$P(n,s)$ at slot s
Slot reservation information sent time	t_x
Slot reservation information all received	t_r
Time of slot s	t_s

Suppose that at time slot s, the neighbors that node n identifies are $n_1(n,s)$, $n_2(n,s)$, ...which constitute the "neighbor set" $N(n,s)$. We use n_1, n_2, ...to represent node $n's$ neighbors at slot s for short in the following description.

To guarantee a collision-free transmission at slot s, node n needs to obtain explicit consensus from all neighbors in neighbor set $N(n,s)$. For each slot, each node randomly and independently generates a priority number. The priority for node n at slot s is denoted as $p(n,s)$. For node n's neighbor n_i, the priority number at time s is denoted as p_i.

The priority is disseminated as part of the reservation information, and is sent out in advance at time t_x which is prior to time t_s of slot s as illustrated in Figure 3.11. This is the key idea of pre-scheduling, that information which is used for reserving slots is disseminated in advance. At time t_r, node n receives the priority from all nodes in $N(n,s)$, which is denoted as priority number set for node n at slot s, $P(n,s)$. The notations is listed in Table 3.2. The simple slot allocation algorithm is illustrated as follows:

Ensure: whether node n can use slot t

1: **if** $\forall p_i \in P(n,t), p > p_i$ **then**
2: *n wins slot t*
3: **else**
4: *n loses slot t*
5: **end if**

There are two major drawbacks of this approach:

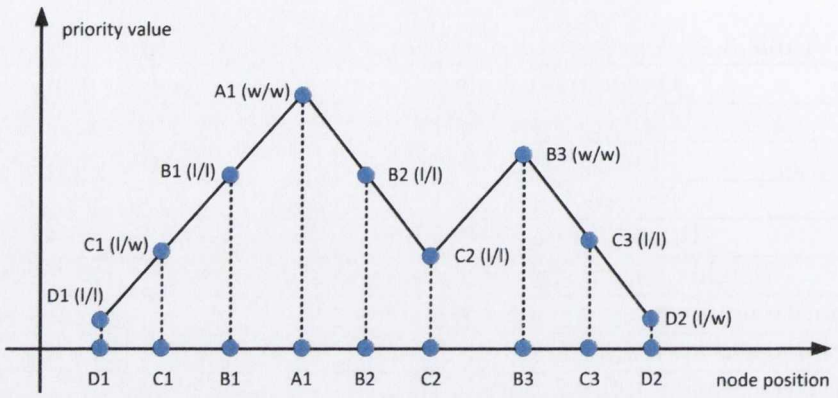

Figure 3.12: Simple and time-bounded slot allocation in a 1-dimension network (simple/time-bounded)

1) This approach may lead to inefficient slot allocation, because the node with higher priority numbers may not actually use the slot as it loses the competition to another neighboring node. An example is presented in Figure 3.12. Assume that a number of nodes are positioned in a one-dimensional space (x-axis), and their respective priority values are plotted on the y-axis. It can be observed that only two nodes at the top of the "mountain" can use a slot using the simple slot allocation algorithm, which is much less than the maximum number of nodes that can potentially use the slot concurrently.

2) On average, the probability that a node wins a slot is $1/(N+1)$, where N is the average number of neighbors. However, due to the fact that a slot is reserved independently from other adjacent slots, the time interval between two consecutive successful slots cannot be bounded. For safety applications with a communication pattern of periodic beaconing, a burst of successfully reserved slots cannot improve the quality of service, as most broadcast messages are similar to each other and carry no new information. On the other hand, a long period of time without a reserved slot means an interruption to communication which may significantly reduce the perceived QoS of communication.

3.3.2 Time-bounded Slot Allocation Protocol

Although the protocol proposed in the previous section is easy to implement, it is inefficient as it cannot provide guarantees on slot reservations. In this protocol, the longest period between two consecutive slot reservations is unbounded, and a node may experience unbounded time to access the medium. As discussed in Chapter 1, timeliness of message delivery is critical for safety applications, which calls for a different slot allocation protocol.

In this thesis, a time-bounded slot allocation protocol is proposed which aims to fulfill the requirements that the maximum time interval between two consecutive slots is bounded. The proposed new protocol a) introduces an admission control mechanism that caps the maximum rate a node can reserve slots, and b) proposes a new slot reservation scheme termed "sticky reservation". These two new ideas are discussed in the following sections.

3.3.2.1 Sticky Reservation Concept

In conventional slot allocation protocols such as the simple protocol described in Section 3.3.1, when a node recognizes the existence of a node with higher priority (i.e., a rival), it waives this slot immediately. In sticky reservation, as the name suggests, the node in question does not waive this slot, but adds this rival to a "pending list" and waits for its decision. Consequently, for a particular node, a slot may have a number of pending nodes attached, which indicates that the usage of the slot has not been decided.

In sticky reservations, the eventual status of a slot is determined not only by the priority numbers, but also the decisions of other rivals. A node gives up on a slot if and only if one of the nodes on the pending list successfully reserves this slot and explicitly informs the node in question via messages. On the other hand, if the received message indicates that some nodes in the pending list give up on this slot, these nodes are removed from the pending list. If the pending list becomes empty, the node in question wins this slot.

Following such a slot reservation scheme, a node's decision triggers its neighbors' decision which triggers their neighbors and so on. Considering the network at a particular slot, nodes with the highest priority number in their neighborhood

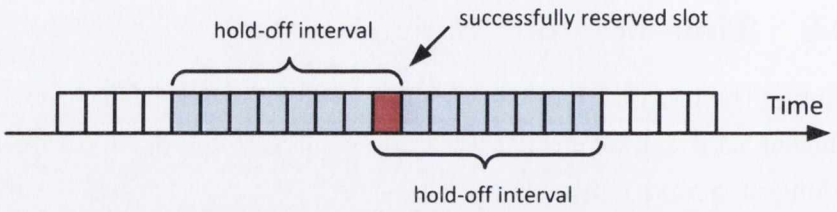

Figure 3.13: Hold-off interval concept

are the first ones to decide. Subsequently, the neighbors of the first decider also decide and the decision process "ripples" towards other parts of the network.

The sticky reservations deterministically arbitrate the slot usage among nodes, i.e., no neighbor will use the same slot. In addition, this scheme guarantees progress as all nodes in the network will eventually decide. The most important property of sticky reservation is that, a node compete for all slots, and a slot is waived only when a node has to. This property is vital for satisfying the slot reservation guarantee as will be discussed in later sections.

3.3.2.2 Hold-off Interval Concept

In wireless communications, time slots used in a local area are scarce resources. Any slot allocation algorithm that does not prohibit nodes from monopolizing the resources faces the possibility of running out of resources and being unable to meet the allocation guarantee. In light of this, an admission control mechanism is proposed in this thesis in order to cap the resources that a node is allowed to reserve, which is critical in achieving guaranteed slot allocation.

For a slot that is successfully reserved, a "hold-off" interval is proposed which prohibits the winning node from using any other adjacent slots within that period, as depicted in Figure 3.13. All slots within that interval are considered lost, regardless of their prior status. The "tightness" of the admission control mechanism is determined by the length of the hold-off interval. If slots are in high demand in an area where a large number of rivals exist, the length of the hold-off period is longer, indicating a lower portion of the usable resource for any particular node.

In the proposed protocol, the length of the hold-off interval is influenced by the node density in a local area. If the number of neighbors increases, a node extends

its hold-off interval, and vice versa. In the following, the logic that guarantees the slot allocation is presented first, which is followed by the specific algorithms and proof.

For simplicity of the description, it is assumed that for all nodes in the network, the hold-off interval is set to a sufficiently large number, which is at least greater than the number of neighbors that any node has. Suppose that a particular node has N neighbors, and there is an arbitrary slot fragment that has $N+1$ slots waiting to be reserved. Consequently, the hold-off period is larger than N, and thus none of the neighbors, including the node in question can reserve twice in this particular slot fragment.

Suppose that the node in question tries to reserve a slot in this slot fragment (with size $N+1$). In the worst case scenario, all its N neighbors have successfully reserved a slot in the slot fragment, but due to the fact that none of them can reserve twice, there is at least $N+1-N$, i.e. one slot left unreserved. Because the node in question is using the sticky reservation approach, it has not given up on the remaining slot yet. Thus, the node in question is guaranteed to reserve a slot.

To summarize, the key to the time-bounded slot allocation algorithm is to make sure that there is always enough slots left, and nodes do not give up slots voluntarily.

3.3.2.3 Formal Description of the Algorithm

Some preliminaries and assumptions regarding the algorithm are presented below. There are two types of decisions regarding a specific slot: win(w) and lose(l). A pending slot is not considered as decided as its status is not final. When a node decides, it can not change its mind later. The decision is included in a message and sent out to all of its neighbors. The slot decision message has the following format: Msg(decider ID, decision, slot number), e.g. $Msg(n, w, s)$, which indicates that node n has won slot s. Neighbors with higher priority are included in a *superior neighbor set*. The rest of the notations are described in Table 3.3. The time-bounded slot allocation algorithm is described as follows:

Require: The assumptions are the same with algorithm 1

Table 3.3: Annotations of variables in time-bounded slot allocation algorithm

Description	Annotation
Decisions	$win(w)$, $lose(l)$
Target node	n
Target slot	s
Superior neighbor set	$R(n,s)$ at slot s, of node n
Neighbor set	$N(n,s)$ at slot s, of node n
Priority number	$p(n,s)$ at slot s, of node n
Neighbor's priority number	$p_i(n_i,s)$ at slot s, of node n_i, $n_i \in N(n,s)$
Message	$Msg(n,w/l,s)$
Hold-off interval	h
Set of slots in hold-off interval of slot s	$H(s,h)$

Ensure: whether node n reserves slot s successfully or not

1: $R \leftarrow \emptyset$
2: **if** $\forall p_i(n_i,s), n_i \in N(n,s), p(n,s) > p_i(n_i,s)$ **then**
3: node n wins slot s
4: node n sends $Msg(n,w,s)$ to n_i, $n_i \in N(n,s)$
5: **for all** slot s_i in $H(s,h)$ **do**
6: node n loses slot s_i, $s_i \in H(s,h)$
7: node n sends $Msg(n,l,s_i)$ to n_i, $n_i \in N(n,s)$, $s_i \in H(s,h)$
8: **end for**
9: **else**
10: **if** $p(n,s) <= p_i(n_i,s), n_i \in N(n,s)$ **then**
11: $R(n,s) \leftarrow n_i$
12: **end if**
13: node n waits for decision from nodes in R
14: **if** node n receives $Msg(n_i,w,s)$, $n_i \in R$ **then**
15: node n loses slot s
16: node n sends $Msg(n,l,s)$ to n_i, $n_i \in N(n,s)$
17: **end if**
18: **if** node n receives $Msg(n_i,l,s)$, $n_i \in R$ **then**
19: $R \leftarrow R - n_i$
20: **if** $R == \emptyset$ **then**

21: repeat line 3-8
22: **end if**
23: **end if**
24: **end if**

In line 1, the superior neighbor set R is initialized as empty for node n at slot s. Line 2-8 describes the case that the target node n has the highest priority among all of its neighbors, and wins the slot. In addition to marking the winning slot as reserved, node n waives all slots in the hold-off interval, as shown between line 5 and line 8. The winning slot decision and lost slot decisions are sent out in line 4 and 7 respectively.

Lines 9 to 24 represent the case that the target node depends on at least one other node's decision. Node n add nodes that it depends on to its superior node list and waits for their decisions (line 10-13). If the decision arrives and a superior node has won the slot, the target node has lost this slot and sends out a lost decision (line 14-17). On the other hand, if the decision from a superior node is to give up on this slot, the superior node is removed from the superior neighbor set, i.e., as if the target node had been "promoted".

If the superior neighbor set has no elements, i.e., all superior neighbors have given up on the target slot, the target node wins this slot (line 20-22). The target node then follows the same procedure where a node has the highest priority number and win the slot immediately.

3.3.2.4 Slot Reservation Guarantee

For a node with N neighbors, the best reservation guarantee that can be achieved is to get one slot in every $N + 1$ slots. This is the best case scenario of any achievable guarantee for $N + 1$ slots with $N + 1$ competitors. As far as the node in question is concerned, the guarantee is met if two conditions are satisfied: 1) no neighbor uses more than one slot during the $N + 1$ slot fragment, i.e., "protected fragment", and 2) the node in question sticks to all slots in the protected fragment and never voluntarily gives up.

Due to the fact that the nodes may have different local environments and thus have different protected fragment lengths with each other, to achieve this

guarantee in all local areas, a node needs to 1) knows the length of the protected fragment of all its neighbors, and 2) prohibits itself from using more than one slot within any neighbors protected fragment.

The critical part of achieving the slot reservation guarantee is that nodes set their hold-off intervals according to the neighbor count of their neighbors. Specifically, each node estimates the maximum number of neighbors in its vicinity, and exchanges this information with its neighbors. By doing so, each node is informed of the neighbor count of all of its neighbors. A node sets its hold-off interval to the largest neighbor count it received, which prohibits itself from using excessive slots and violating the slot reservation guarantee of the neighbor with the largest neighbor count.

Assume that for an arbitrary node n, the number of neighbors at time t, is denoted as $|N(n,t)|$. In addition, value l is defined as the *one* + maximum value of $|N(n,t)|$, where t is the time range between present time t_0 and the max predictable time t_{max}. In other words, l represents the maximum number of neighbors that node n will have in the predicted future plus one. The rest of the annotations are described in Table 3.4. Prior to the description of the reservation guarantee, some assumptions are listed as follows:

1) For node n, an arbitrary interval is defined which is composed of l consecutive slots. This piece of slot fragment is called "protected fragment" of node n, and is a property of a particular node.

2) Node n propagates its l value to all of its neighbors and also receives their l values respectively. We use l_i to represent the value received from neighbor i of node n.

3) Node n sets its hold-off length h to the maximum value of l_i that is received from the neighbors, to make sure it does not use more than one slot in the protected fragment of any of its neighbors.

Theorem 1: *node n is guaranteed to reserve at least one slot in its protected fragment (with length l)*

Proof. The hold-off value from node $n's$ neighbor i is denoted as h_i. According to assumption 3, we have $h_i >= l$.This means, in node $n's$ protected fragment (with length l), for all of its neighbors, each neighbor can use no more than one

Table 3.4: Annotations of variables in proving slot reservation guarantee

Description	Annotation
Target node	n
Present time	t_0
Maximum prediction time	t_{max}
Neighbor set at time t, of node n	$N(n,t)$
Number of nodes in neighbor set $N(n,t)$	$\|N(n,t)\|$
Max $\{\|N(n,t_0)\|, \ldots, \|N(n,t_{max})\|\} + 1$	l
Size of protected fragment	l
Received l from neighbor node n_i	l_i
Max l received from all neighbors n_i	l_{max}
Hold-off interval	h

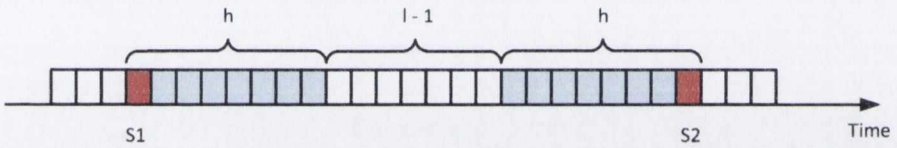

Figure 3.14: Reservation guarantee

slot. In the worst case scenario, according to our definition of l, that node n has up to $l-1$ neighbors, and each of them has taken one slot. Therefore, there is at least one slot left. As node n is still sticking to the remaining slot, and because no other neighbor declares victory on that slot, node n eventually obtains this slot.

Thus, at least one slot can be successfully reserved in an arbitrary node's protected fragment. □

Theorem 2: *For an arbitrary node n with hold-off interval h and l, the reservation interval, i.e., the distance d between two consecutive successfully reserved slot s_1 and s_2 is bounded by:* $h <= d <= 2h + l - 2$

Proof. Based on the definition of hold-off interval length, we have $d >= h$.

Consider the scenario which is illustrated in Figure 3.14. Two consecutively reserved slots s_1 and s_2 are located on each end of two slot fragments. The hold-off interval is marked as the shadowed area, and the slot fragment in the middle (which is defined as the free fragment) is composed of a number of consecutive

slots that are located outside of the hold-off interval, and do not contain a reserved slot. If the length of the free fragment $f \geq l$, then it is guaranteed by Theorem 1 that a new slot (s_3) will be successfully reserved for node n within f. It contradicts our assumption that slot s_1 and s_2 are two consecutive successfully reserved slots. Thus, the length of free fragment cannot exceed $l - 1$.

Therefore, we have $d \leq 2h + l - 2$ □

To summarize, the time-bounded slot allocation protocol provides reservation guarantee that the maximum interval between two consecutive reserved slots are bounded, which means that the medium access delay can be bounded. The real-timeliness property of the protocol is critical for ensuring the QoS of the safety applications in VANETs. A thorough evaluation of the protocol will be discussed in Chapter 5.

3.4 Summary

This chapter described the design of the real-time reservation protocol that supports reliable and time-bounded medium access. The key idea to achieve these properties in a vehicular environment is the pre-scheduling concept. In the RRP protocol, a two-tier architecture is proposed. The lower virtual cluster layer identifies future neighbors, while the upper real-time scheduling layer allocates slots to achieve guaranteed medium access. In this chapter, detailed descriptions are provided for the specific algorithms in both virtual cluster layer and the real-time scheduling layer, which explains the process of achieving reliable and time-bounded medium access in a vehicular environment.

Chapter 4

Implementation

This chapter describes the implementation details of the RRP protocol which aims to provide reliable and time-bounded medium access in a vehicular environment. Since the main algorithms of RRP are presented in Chapter 3, the focus of this chapter is on the specifics of these algorithms, e.g., the data structures, events, messages, and interactions between components. This chapter starts with the description of the protocol implementation architecture, which gives an overview of the protocol and describes the general functionality of each component. The implementation details of each component are further presented in the rest of this chapter.

4.1 System Architecture

The architecture of the RRP protocol is illustrated in Figure 4.1, where messages are passed around between functional components. The two logical layers: the lower virtual cluster layer and the upper real-time scheduling layer are encapsulated in components termed the "compatibility manager" (CM) and the "slot manager" (SM) respectively. Specifically, the SM receives the neighbor list and all slot-related messages from CM, and it also contains the slot allocation algorithm. When a slot that is reserved by the current node is due, SM notifies the broadcast manager to send out buffered messages. In RRP, CM represents the virtual cluster layer, which maintains the knowledge base (which stores informa-

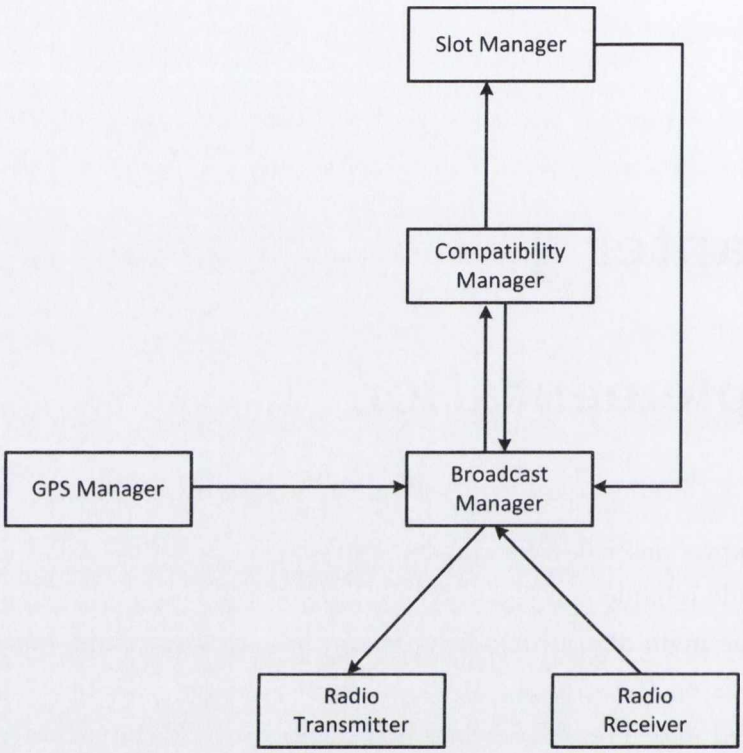

Figure 4.1: RRP protocol architecture

tion regarding other nodes), identifies neighbors and makes forwarding decisions as described in Chapter 3.

The main objective of the broadcast manager (BM) is to buffer outbound messages. BM passes received messages from the radio receiver to CM, or broadcasts messages that are stored in the buffer to the radio transceiver when a reserved slot is due. The GPS manager is a simple data source that periodically sends the current position messages to the BM. The radio transceiver and receiver represents the wireless interface.

4.2 Compatibility Manager

The main responsibility of the CM component is to identify neighbor relations. As the name suggests, this component calculates the neighbor index, which characterizes the neighbor relations, based on received information from other nodes,

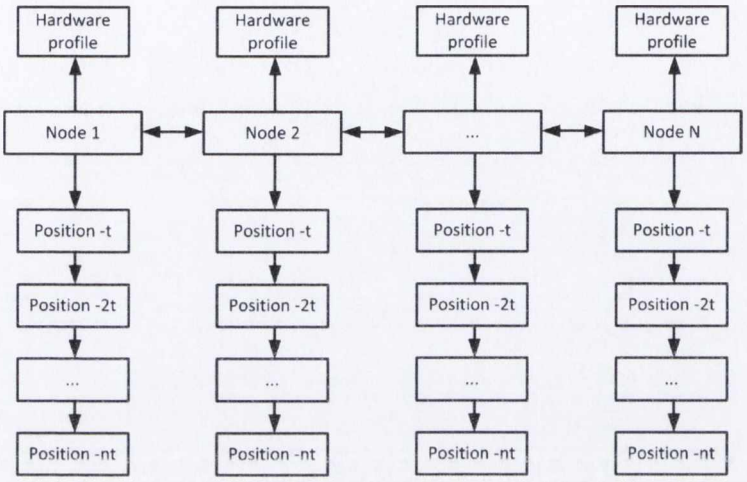

Figure 4.2: Knowledge base data structure

such as position and transmission power, and provides the results of such calculations to other components in the system. The compatibility manager keeps track of the nodes that are in the vicinity of the current node, and stores such information in a data structure called the "knowledge base". Note that the number of nodes in the knowledge base is usually larger than the number of one's neighbors, in order to track any node that has the potential to become a neighbor.

After receiving messages from an internal source, i.e., a GPS device, or from other nodes via the broadcast manager, the knowledge base is updated to reflect the current network status that the current node understands. Given the node's information in the knowledge base, the fundamental and the most important task of CM is to calculate the neighbor index between any two nodes, and make the decision as to whether they are neighbors. The results are the input for the scheduling algorithm in the slot manager as well as the forwarding process in this component. If a message is deemed necessary to be forwarded, it is send back to the broadcast manager, otherwise it is discarded.

4.2.1 Knowledge Base

The knowledge base is a storage of information regarding a node itself and the nodes in its vicinity. The information includes: node identifier, hardware re-

lated information, e.g., the transmission power or the modulation method used in transmission, a list of last known position and speed measurements, and other parameters regarding slot scheduling. The knowledge base is internally implemented as a double linked list as shown in Figure 4.2. The structure of each node element is shown in Listing 4.1.

Listing 4.1: Element data structure in knowledge base

```
struct knowledge_node_s{
    int ID;
    int number_of_entry;

    knowledge_node_hardware* knowledge_node_hardware_profile;
    knowledge_node_dynamic* knowledge_node_dynamic_header;
    knowledge_node_dynamic* knowledge_node_dynamic_tail;
    topology_entry* topology_entry_header;

    struct knowledge_node_s* previous_node;
    struct knowledge_node_s* next_node;
    //other parameters
    //...
};
typedef struct knowledge_node_s knowledge_node;
```

Listing 4.2: Hardware profile data structure in knowledge base

```
typedef struct knowledge_node_hardware_s{
    //transmitter information
    double transmitter_frequency;
    double transmitter_bandwidth;
    double transmitter_data_rate;
    double transmitter_power;
    double transmitter_antenna_gain_in_DB;
    double transmitter_antenna_height;
    const char* transmitter_modulation;

    //receiver information
    double receiver_frequency;
    double receiver_bandwidth;
    double receiver_data_rate;
```

```
    double receiver_antenna_gain_in_DB;
    double receiver_antenna_height;
    const char* receiver_modulation;
    double receiver_noise_factor;
    int packet_size;
}knowledge_node_hardware;
```

For each node stored in the knowledge base, the hardware related information is stored in a "hardware profile", which is shown in Listing 4.2. The parameters stored in the hardware profile are used to calculate the message reception probability, which is a crucial parameter in determining the neighbor relations.

For each node, a list of elements of type "node dynamic" is attached, which stores geographic information of a node's past. (See Listing 4.3). The previously received position-related information is used by the mobility prediction algorithm, which is also a necessary step in determining the neighbor relationship.

Listing 4.3: Node dynamic data structure in knowledge base

```
struct knowledge_node_dynamic_s{
    short ID;
    short sequence_number;
    double time_stamp;
    double x_position;
    double y_position;
    double speed;
    double orientation;

    struct knowledge_node_dynamic_s * next_entry;
};
typedef struct knowledge_node_dynamic_s knowledge_node_dynamic;
```

The information stored in the knowledge base is updated whenever a new message is received from the broadcast manager, whether the source of the message is internal or external. If the node is not in the knowledge base, a new element is created; while if a node is deemed irrelevant to the current node, e.g. too far away or not heard from for an excessively long time, the node is removed. When a new message, either internal or external, is received from the broadcast manager, the dynamic field of a node is updated accordingly. The format of a

origin node ID (16 bits)	sequence number (16 bits)	time stamp (32 bits)
x position (32 bits)		y position (32 bits)
speed (32 bits)		orientation (32 bits)

Figure 4.3: Position information

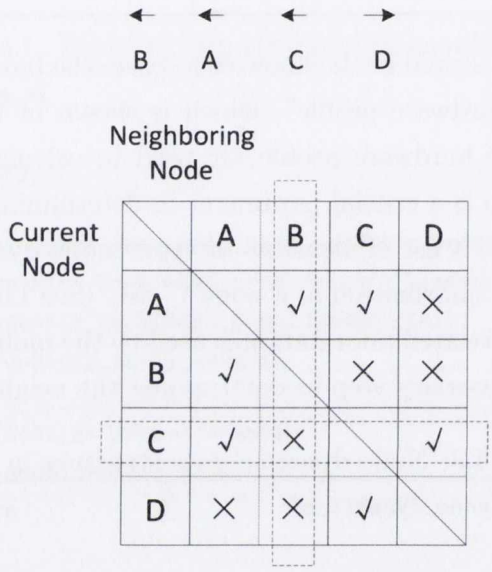

Figure 4.4: Neighbor relation matrix

geographic information packet is depicted in Figure 4.3.

4.2.2 Rehearsal Operation

Based on the node information stored in the knowledge base, whether a node is a neighbor of another node can be decided using the algorithm discussed in Section 3.2. The API of the neighbor identification process is shown in Listing 4.4. For each node pair in the knowledge base, their neighbor relations can be expressed as "neighbor" or "non-neighbor". Consequently, at any specific time with a corresponding network topology, a neighbor relation matrix can be constructed, and a node is able to determine the neighbor relation between any node pair by looking up in such a matrix.

For example in Figure 4.4, the network topology is shown in the upper part

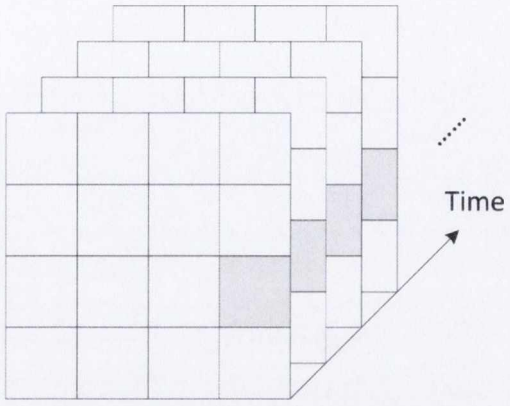

Figure 4.5: Composite relation matrix

of the diagram, as node A, B, C and D are 1-hop away from each other. In the corresponding relation matrix below, a tick represents a neighbor relation and a cross represents a non-neighbor relation between a signal node (with ID in the row) and its potential interfering node (with ID in the column). For instance, node A and node D are neighbors of node C, but node B is not a neighbor of node C.

Listing 4.4: Neighbor identification API

```
int neighbor_identification(unsigned short protected_node_ID, unsigned short
    target_node_ID, int prediction_level, int* result);
```

The neighbor relation matrix is the main product of the CM component. Note that such a matrix characterizes the estimated neighbor relationship at a specific time in the future. In other words, it is a "snapshot" of the relationship among the current node's neighbors. If the neighbor relation matrix is created at multiple time instants in the future, a composite neighbor relation matrix is created as illustrated in Figure 4.5.

Each layer of the composite relation matrix is a snapshot of the predicted future. In other words, the composite matrix can be seen as a "rehearsal" of how the current network may evolve in the near future. During such a rehearsal, the neighbor relation of any node pair as time progresses (Figure ??) can be seen as a section of the composite matrix in Figure 4.5.

The rehearsal concept reflects RRP's intention to predict the future networks

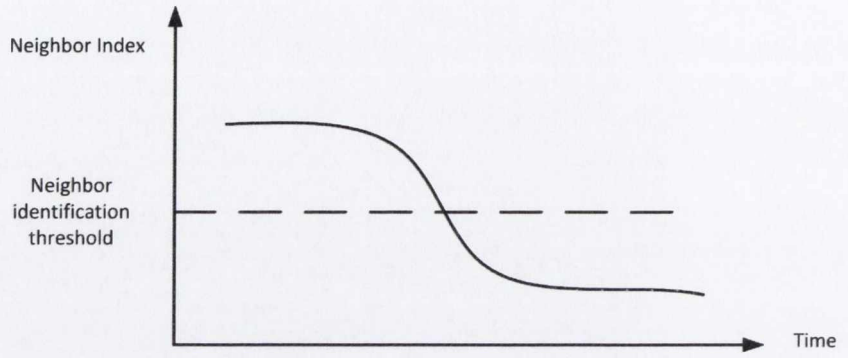

Figure 4.6: Neighbor identification with respect to time

in order to make preparations in advance. For the sake of simplicity, in the implementation, only the latest layer of the composite relation matrix is used, which represents the furthest predictable time instant. A particular relation matrix has two purposes: to generate a neighbor list for the scheduling layer, and to make relevance-based forwarding decisions.

As shown in Figure 4.4, a row represents the relationship between the current node and all other nodes in the knowledge base. The list of nodes that are identified as neighbors are sent to the SM component after each rehearsal operation. For the message forwarding algorithm, it is required to determine the neighbor relations between the message source and the potential recipients, which is recorded in a column of the relation matrix.

4.2.3 Message Handling

Another important functionality provided by the CM component is the implementation of the relevance-based message forwarding, which is described in details in Section 3.2.2. A message received from the broadcast manager first goes through a classifier to determine whether it is new or not. Each message contains a source node identifier and a unique sequence number, which is used for the differentiation. An old message is the one that has been received by the current node previously. The content of the old message has already been reflected in the node's internal knowledge base.

For a new message, RRP follows a three-step procedure of the relevance-based

message forwarding. The algorithm ensures that messages propagate further from the source, and any potential recipients that may benefit from a rebroadcast are stored on a list termed "beneficiary list". Eventually, a decision is made on whether or not to rebroadcast this message. The decision, along with the attached beneficiary list and the message itself are sent back to BM, which are stored in BM's buffer if a rebroadcast is deemed necessary.

The reception of an old message indicates that this message has been broadcast again in the current node's vicinity, which may reach a number of new recipients. If such new recipients are also on the beneficiary list of a buffered message, then there is no need to forward this message to them again. Therefore, such beneficiary nodes are removed from the beneficiary list of this message. If the beneficiary list becomes empty, the rebroadcast of this message is canceled.

4.2.4 Interactions with Other Components

The interactions between the CM and other components are depicted in Figure 4.7. There are three types of messages received by the CM: internal position messages, external new messages and external old messages. An internal message contain only position information without the slot information part, and is handled differently from external messages. When an internal position message arrives, the CM updates entries in the knowledge base accordingly.

When a new external message is received, CM first extracts the slot information part and sends it directly to SM. Subsequently, the CM updates the knowledge base with the position information from the message, and calculates the latest predicted topology as well as the neighbor relation matrix. Next, based on the updated relation matrix, the CM generates the latest neighbor list which is immediately sent to the SM, as well as making the forwarding decision. If a rebroadcast for the received external new message is deemed necessary, the CM sends back the message to the BM, along with the beneficiary list for the message.

When an old external message is received, it is not necessary for the CM to update the knowledge base, as this message has been received and processed before. The only responsibility of the CM is to eliminate certain beneficiary nodes of the message, as they may have received this message during the last broadcast.

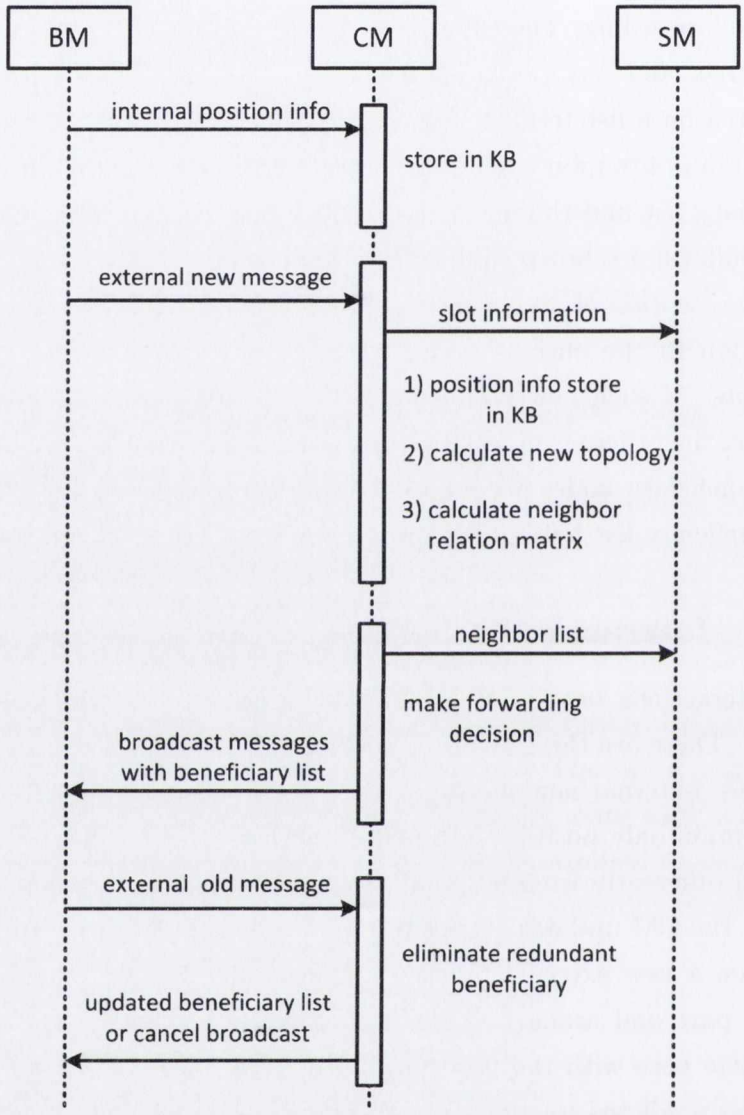

Figure 4.7: CM interface with other components

The updated beneficiary list (same or reduced) is sent back to BM, and if the list is empty this message is canceled for rebroadcast.

4.3 Slot Manager

The main objective of the SM is to schedule slots among one's neighbors, given the neighbor list provided by the CM. The slot scheduling algorithm does not aim to achieve complete consensus among all neighbors, i.e., a node does not need to know which neighbor owns a specific slot, instead it only needs to know whether it owns this slot itself. Consequently, there are three possible status for a single slot: won (owns this slot), lost (gives up on this slot), and pending (whether the node owns or gives up on this slot is yet to be determined).

Listing 4.5: Slot data structure

```
struct slot_entry_s{
    unsigned short slot_number;
    unsigned char slot_status; // pending, lose, win
    unsigned short attached_node_number;

    //list of pending nodes
    slot_attached_node_entry * attached_node_header;
    slot_attached_node_entry * attached_node_tail;

    //connection with other slots
    struct slot_entry_s * next_slot_entry;
    struct slot_entry_s * previous_slot_entry;

    unsigned char deliver_times; //number of times that the status of this
        slot being sent out
    unsigned short once_won; // record if this slot has won in the history,
        but now become lost, for new born slot

    unsigned short is_inquiry_msg_delivered; // indicate whether an inquiry
        message has been sent out
    unsigned short hold_off_length_per_slot; // slot-specific hold-off value

    //..
};
typedef struct slot_entry_s slot_entry;
```

In the SM component, a list is maintained to keep track of the status of each slot as depicted in Listing 4.5. The slots start from the current slot, which represents the present time, and stretches until the maximum prediction time in the future, which is the furthest instant in time at which a slot can be reserved. For each slot, the ownership status is stored in addition to other relevant information such as the slot number and whether the slot status has been sent out or not. In addition, for each slot, a list of pending nodes whose decisions the current node is waiting for are maintained.

As time progresses, a new slot is added to the end of the slot list, and the current slot is removed. The newly added slot is checked to determine its status, i.e., won, lost, or pending. When another node's decision arrives, the status of this slot changes accordingly. If a node on the pending node list wins, the current slot is lost, otherwise if that node gives up, that node is removed from the pending node list of this slot. If all pending nodes are removed, the current node changes the slot's status to won.

4.3.1 Slot Priority Number

Previously, it is assumed that for each slot, a priority number is used to arbitrate the slot contention. However, how to generate and how to make nodes mutually aware of such priority numbers are not specified. In the following, some desired properties of the priority numbers are discussed in general, followed by the priority generation mechanism that is used in the protocol.

A simple priority can be generated randomly within a specific range, e.g., between 0 and 100. However, disseminating these random numbers in the network consumes large quantity of valuable resources, and is less efficient because the random numbers *per se* are meaningless. Ideally,a "seed" of some sort, which can be used to generate these priority numbers, is circulated in the network rather than the priority numbers themselves. If such a "seed" is mutually understood in the network, a node is able to "calculate" another node's priority at any time in the future.

In addition, it is also desirable that the generated priority number presents some level of fairness and can adapt dynamically in the environment. Further-

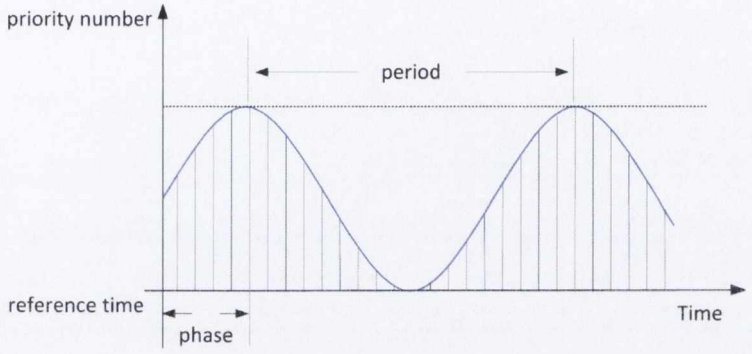

Figure 4.8: Priority number generation

more, the "seed clash" scenario should be avoided where different nodes have the same seed and generate identical priority numbers.

In RRP, a sine function is used to generate priority numbers. Specifically, for a specific node, the priority number at a specific time slot is calculated as the value of a sine function given the specific time, as depicted in Figure 4.8. The sine function has three parameters: amplitude, period and phase. The amplitude of the function is constant, e.g., 100, while the value of period and phase are node specific.

As a property of the sine function, the value of a node's priority number wanes and waxes periodically. A node's priority number reaches its peak value once in a period, when it has a high probability of beating other nodes. The length of a node's period is determined according to the number of its neighbors in the vicinity. Therefore, in a dense network, a node has a longer period, which means that it takes longer for a node to win a contention.

To avoid having the same priorities, nodes use different phases for the sine function. A node's phase is defined as the offset between the peak slot of the node and a common reference slot. Technically, it is possible that two or more nodes have identical period and phase with each other, however, considering the size of period and phase, the probability of seed clash is low. In addition, even if a seed clash did happen, the correctness of the scheduling algorithm is not affected.

The period and phase of a node is disseminated as a seed among nodes. A node's period often changes according to the node density in the local area. By knowing each other's phase value, nodes can adjust their own phase and

avoid other's peak slot, which reduces the number of messages needed in slot reservations.

4.3.2 Slot Life Cycle

In this section, the life cycle of a slot is described, which starts from when the slot is created until its allocation has been finalized. As mentioned in earlier sections, there are three possible statuses of a slot: won, lost and pending, where the first two statuses are considered a decision. Once the decision has been made, a node cannot change its mind, and should notify other neighbors of its decision via messages.

The algorithm for reserving a slot is described in Section 3.3.2, and is composed of three phases. The first phase starts when a slot is created and added to the end of the slot list. The status of the new slot is decided immediately, and this process is called "new slot arbitration". If the slot status is pending after the new slot arbitration process, it enters the "pending slot" phase, during which it waits for decisions from the pending nodes of this slot. If decisions are not received, e.g., lost during transmission, or the pending node remains undecided, the slot enters the last phase called the "terminal stage". During this stage, an inquiry message is sent to those pending nodes as an explicit way to request the decision, in case the decision has been made but lost during transmission. This mechanism serves as a "back up" method to improve the performance of the protocol. The specifics of these three phases are described in detail in the following:

When the time passes over the boundary of a slot, a new slot with certain priority is added to the slot list. with its status undefined. The first task of the new slot arbitration process is to check whether this slot is within any winning slot's hold-off interval. If so, the slot is not allowed to be used by the current node, and is considered lost immediately. If the slot is not within any hold-off interval, the priority of this slot is compared with the priority of all other neighbors for this slot. If the current node has the highest number, it wins the slot and sends out the decision. Otherwise, the neighbors that have a higher priority are added to the pending node list.

If a slot has at least one pending node after the new slot arbitration process,

Table 4.1: Description of slot state transition

Event Number	Description
1	slot within hold-off area of current node
2	slot has the highest priority among all neighbors
3	slot does not have the highest priority among all neighbors
4	received decision that a pending node has won
5a	received decision that a pending node has lost
5b	slot within hold-off area of other node's won slot
6	no pending node left
7	time up lost

the slot enters the pending phase as it waits for the decisions from those pending nodes. As the rule of the slot competition mandates, if a win decision is received from any of the pending nodes, a decision is made and the slot is lost. On the other hand, if a slot decision is received which indicates that a pending node has given up on this slot, then this pending node is removed from the pending list. In other words, the current node is no longer depending on this node, and this process is referred to as a "promotion". If a slot has en empty pending node list after a number of such promotions, a win decision is made on this slot.

If the slot remains undecided, i.e., has at least one pending node after a certain period of time, it is reasonable to doubt that the decision message may have been lost during transmission. Therefore, an inquiry message is sent to the pending node whenever the node has the opportunity to send messages. The receiving node should reply with its decision of this slot. This mechanism improves nodes' decision ratio by increasing the delivery probability of the slot decision messages.

If a slot still remains undecided, due to reasons such as inquiry reply not being received, or the pending node is undecided, the node probabilistically gives up this slot, in order to let other nodes which may depend on the decision of the current node proceed. This mechanism is another enhancement to improve the slot decision ratio.

4.3.3 Slot Transition Diagram

The slot life cycle is summarized in Figure 4.9 and Table 4.1.

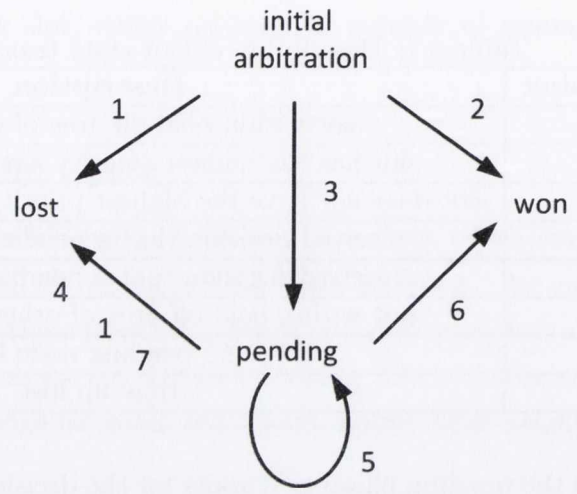

Figure 4.9: Slot state transition

4.3.4 SM Input and Output

The interactions between the SM and other components are depicted in Figure 4.10.

The SM handles two types of events: incoming message events, and time-triggered events. For the first category, there are two types of messages that come from CM: updated neighbor list messages and received slot information. The neighbor list is the main output from CM, and is stored locally in SM. Slot information is received from neighboring nodes regarding their slot decisions, based on which local slot status is updated.

There are two types of time-triggered events: new slot due events and my slot due events. The new slot due event is triggered when time passes a slot boundary and a new slot arrives. The SM adds a new slot to the slot queue and removes the oldest slot from the queue. In addition, the SM arbitrates the status of the new slot as described before. The my slot due event is triggered when a slot that is reserved by the current slot arrives. It is the opportunity for the current node to transmit and therefore the SM assembles the slot decisions of the current node and sends them to the BM. In addition, the SM notifies the BM that the slot is due and broadcasts all buffered messages.

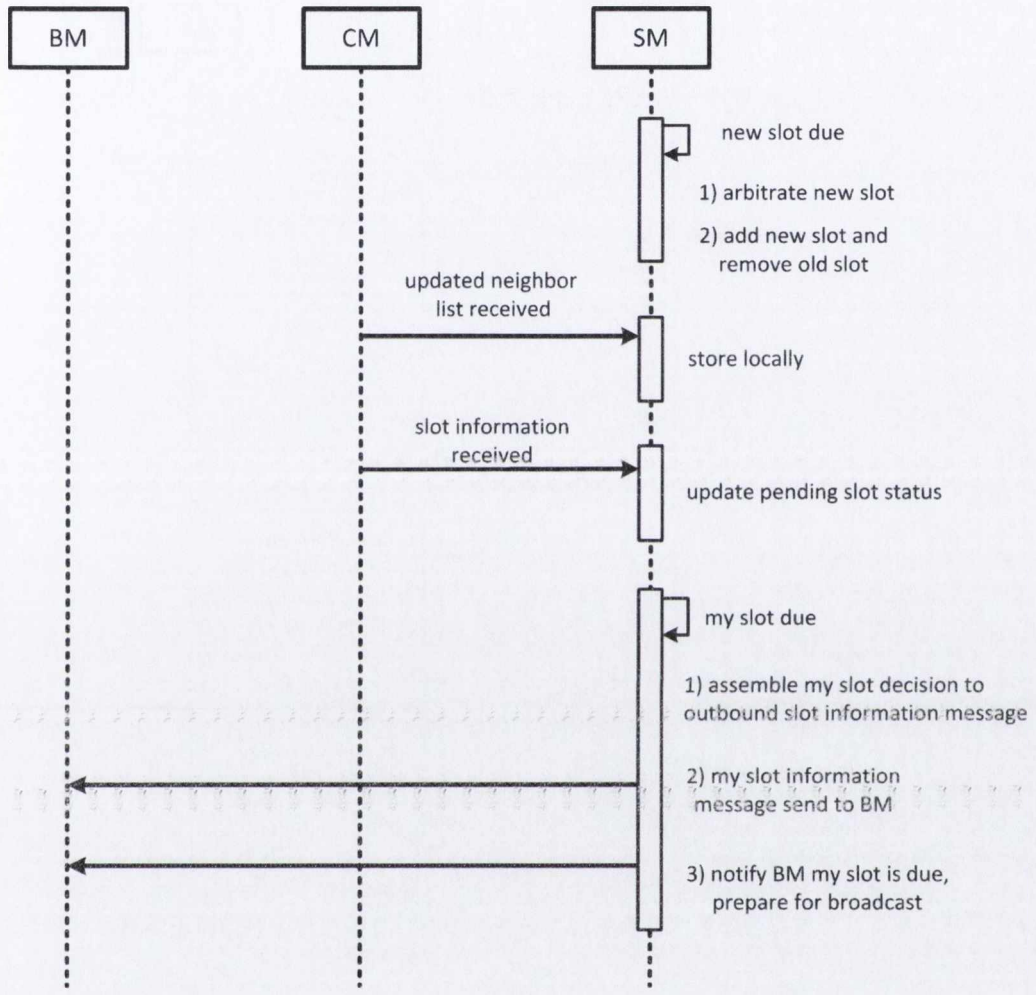

Figure 4.10: SM's interaction with other components

4.4 Broadcast Manager and GPS Manager

The BM component primarily serves three objectives: initial processing of incoming messages from the radio, buffering and updating the outbound messages, and broadcasting the messages when one's own slot is due. The main data structure maintained in BM is therefore a message queue.

When a message arrives from the CM (either an internal or external message) which needs to be forwarded, BM temporarily stores them in the message queue and waits for the current node's reserved slot, in which messages are sent. As

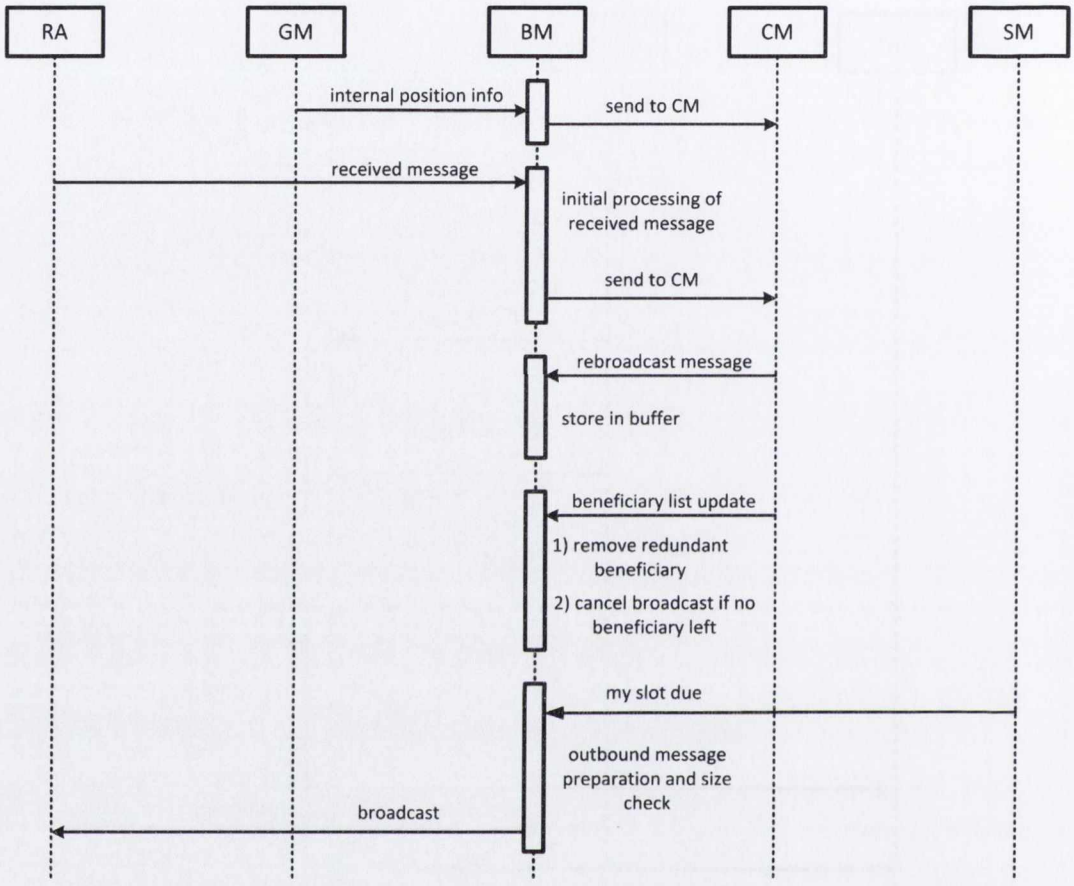

Figure 4.11: BM's interaction with other components

described in previous sections, each message that needs to be forwarded has a list of beneficiary nodes. If CM decides to eliminate some nodes from the beneficiary list, it is for the BM to remove these nodes. If the beneficiary node list becomes empty, BM deletes this message from the buffer.

When a message is received from the radio component, an initial screening is conducted to remove messages that are from the current node itself, as well as any duplicated messages. Subsequently, the received message is forwarded to CM. If the message is coming from the GPS manager, BM forwards the message to CM directly.

When SM notifies BM that the current node's own slot is due, BM flushes all messages in the queue to the radio component. It is worth noting that BM

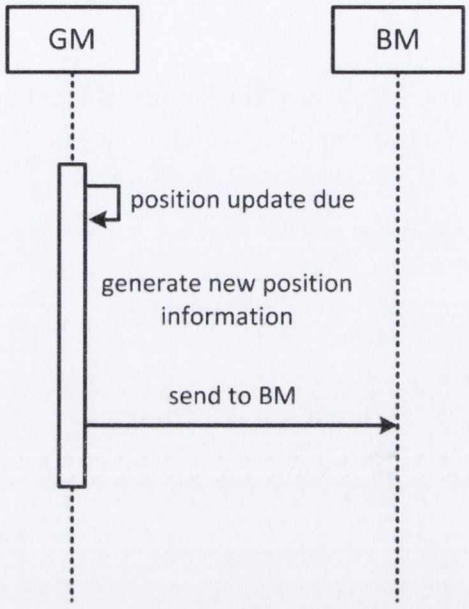

Figure 4.12: GM's interaction with other components

needs to make sure that the size of the total outbound message does not exceed the maximum packet size that can be squeezed in a single slot.

The interaction between BM and other components are depicted in Figure 4.11.

The GPS manager emulates a GPS device that can provide the current node with localization information. It periodically (typically every 100ms) generates the position and velocity information of the current node, and sends to BM. Each message is tagged with a unique sequence number. From the implementation point of view, the position of the current node can be readily and accurately obtained from the simulation environment. To emulate a real-world device, deviations of the position and velocity are artificially added.

The interaction between the GPS manager and other components are depicted in Figure 4.12.

4.5 Summary

In this chapter, the implementation details of the RRP protocol is presented, in terms of the data structure, message flows, and interactions among different components. An extensive evaluation of the RRP protocol is presented in Chapter 5.

Chapter 5

Evaluation

In this chapter, the performance of the proposed RRP protocol is evaluated and compared with two benchmark protocols: 802.11p 802.11p [2010] and RR-ALOHA Borgonovo *et al.* [2004]. The primary goal of the evaluation study is to investigate the suitability of the proposed RRP protocol, and compare it with those benchmark protocols in a simulated but realistic environment. Conducted in a variety of scenarios and contexts, the study aims to reveal the intrinsic characteristics of RRP, 802.11p and RR-ALOHA; compare and analyze their performance in terms of the communication QoS, and shed some light on their applicability in the envisaged vehicular networks. The evaluation study can help achieving not only a better understanding of the strengths and weaknesses of this protocol, but may also identify possible improvements for future works.

As a control group against RRP, 802.11p is chosen because of its quasi-standard status in vehicular communications, and its worldwide success as a MAC protocol. However, there is wide spread skepticism regarding 802.11p's applicability in vehicular environment, which warrants the need to further investigate its performance. As a counter proposal to the contention-based approaches, a reservation-based protocol is also essential to the evaluation study. The RR-ALOHA protocol is chosen because it represents a large number of reservation-based protocols which share similar characteristics, advantages and limitations with RRP. The similarities and differences between RR-ALOHA and RRP will be our focus in the following study.

This chapter starts with the description of some preliminaries of the evalu-

ation, e.g., the metrics and parameters used, the methodology to conduct the experiments, as well as the mobility and physical model adopted in the simulation. The main evaluation part starts with protocol-specific evaluation of RRP, 802.11p and RR-ALOHA, then followed by a comparative study among those three with extensive analysis. This chapter ends with a summary of the results and final conclusions of the evaluation.

5.1 Evaluation Methodology

5.1.1 Evaluation Metrics and Parameters

For a MAC protocol, the goal of achieving reliable transmission and achieving short medium access delay are not necessarily compatible. A reliable MAC protocol may have a large medium access delay, while a fast responsive protocol, i.e., with low medium access delay, may experience high packet loss rate. For safety applications in VANETs, both properties are required, i.e., the MAC protocol needs to be both reliable and with a short (preferably bounded) medium access delay. However, there is no single metric that is readily available to characterize both properties at the same time.

In this thesis, staleness is proposed as the single metric to characterize the perceived communication QoS of applications in a vehicular environment, which incorporates both transmission reliability and medium access delay. The staleness idea stems from the observation that, in vehicular networks the foundation of most safety-critical applications is to obtain the whereabouts of one's neighbors. By periodic and frequent beaconing, each vehicle maintains real-time context awareness in its vicinity. If a vehicle's communication layer fails to deliver the positioning information from another adjacent vehicle for a predefined time, emergency actions will be triggered by the vehicle's safety applications, which could lead to a severe degradation in offered services.

Consequently, safety-critical applications in VANETs have a unique communication requirement - the connectivity between vehicles must be constantly preserved VCS [2006]. For safety applications, the "black out period" during which positioning information is unavailable, needs to be bounded within an acceptable

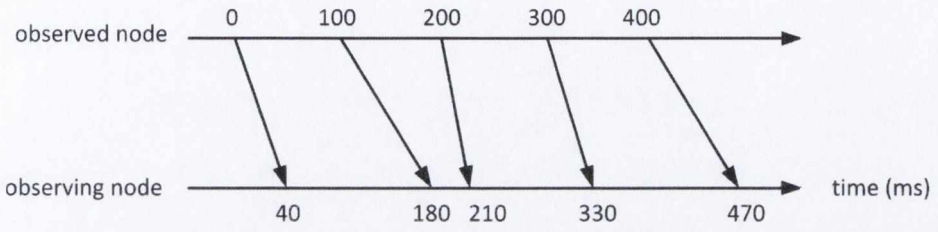

Figure 5.1: An example of observing staleness between two nodes

range, otherwise the overall performance of such a communication system drops drastically. From a different perspective, the staleness concept can be interpreted as the "anxiety level" of a vehicle regarding its neighbor's position. A higher staleness value of a neighboring vehicle indicates a longer black out period of that vehicle, which makes the current vehicle more "anxious".

Suppose that an observed node (A) sends out beacon messages periodically, which are received by a nearby observing node (B) with random communication delays, as illustrated in Figure 5.1. The staleness value is continuously measured by the observing node. The last received message at the observing node B from observed node A is denoted as m_n, among all received messages (m_i, i from 0, 1, 2 ... n). The creation time of the message m_n at observed node A is T_{mn}, and the present time is T_c. We define staleness between observed node A and observing node B in Equation 5.2. The respective staleness value of the example is depicted in Figure 5.2.

$$S(A, B) = T_c - T_{mn} \tag{5.1}$$

Note that the staleness $S(A, B)$ increases with time when no new message is received, and drops to D_{mn} when a new message is received, where D_{mn} is the communication delay of the newly arrived message between the observed and the observing node.

In the evaluation, the staleness is periodically sampled between a given node and its neighbors within its desired communication range. The mean value of the staleness is measured, as well as the 90%, 95%, and 99% percentiles, which quantifies the deviation of the measured staleness. Both mean and deviation of the staleness are important metrics in quantifying the communication QoS perceived

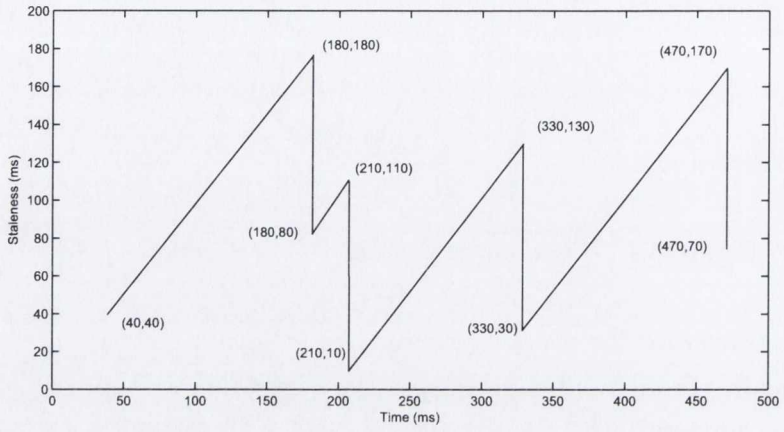

Figure 5.2: The calculation of staleness between two nodes

by a safety application. However, it is worth noting that, safety applications are very sensitive to large staleness deviations as staleness spikes drastically deteriorate communication quality.

It can be observed from the definition of staleness and the example in Figure 5.2 that, if an expected message is not received, or received with a long delay, the overall staleness will increase. In other words, staleness takes into account both packet collision and medium access delay.

A similar idea which is termed "T-window reliability" has been proposed by Bai & Krishnan [2006]. It is defined as: "the probability of successfully receiving at least one single packet from neighbor vehicles during the tolerance time window T". There are similarities between T-window reliability and staleness, but the main difference is that staleness incorporates the medium access delay while T-window does not.

In addition to staleness, a number of "traditional" metrics that measure communication quality are included in the evaluation study as well, which includes Packet Delivery Rate (PDR), medium access delay, reservation interval and throughput.

Unlike unicast, it is not straightforward to determine whether a broadcast is successful or not, as some neighbors of the sender receive the packet while others do not. In this chapter, it is assumed that for each broadcast, the packet sender

has a "desired broadcast range" (DBR), which is determined by the transmission power and the sender's location. Only those nodes that are within the DBR are of interest for this particular broadcast. As the broadcast commences, there are three and only three possible outcomes for any node within the DBR. It either a) receives the packet successfully, b) dropps the packet due to interference (i.e. collision-induced drop), or c) drops the packet due to a weak signal (i.e. non-collision-induced drop). For a specific broadcast, the packet delivery rate is defined as:

$$PDR = N_{success}/N_{total} \tag{5.2}$$

where $N_{success}$ is the number of nodes that successfully received the broadcast in DBR, and the N_{total} is the total number of nodes in DBR.

Other evaluation metrics that are used in this chapter are defined as follows. Medium access delay is measured as the duration between when a packet arrives at the MAC layer to its actual transmission. The reservation interval measures the time interval between two consecutive reserved slots, and is only applicable to reservation-based protocols, i.e., RRP and RR-ALOHA. Throughput is defined as the amount of bits sent or received during a unit of time at a specific node. It is measured at the node level (rather than network level) for both senders and the receivers.

5.1.2 Evaluation Approach

The above metrics are evaluated in a variety of scenarios, which have different settings and configuration parameters. Among these parameters, some are protocol specific, such as the contention window size in 802.11p. and the frame size in RR-ALOHA, while others are generic and used for all protocols, such as vehicle speed, beacon generation rate, and vehicle density. Due to the number of evaluation metrics and parameters and the combination of these two categories, it is obviously intractable to evaluate all possible combinations of metrics and parameters. Consequently, the evaluation study is divided into two parts: a protocol-specific evaluation and the comparison of all protocols.

In the first part, each protocol of interest is evaluated and analyzed. Metrics

and parameters are selectively chosen in order to reveal the characteristics of that specific protocol. Using this evaluation methodology, it is unnecessary to exhaustively explore the parameters space, while highlighting the most relevant properties of a protocol. In the second part of the evaluation, all three protocols are measured and evaluated with the same set of rules, metrics and parameters, which creates a fair basis for comparison.

Prior to the detailed evaluation of the protocols, the models used in the evaluation study, including the mobility model and physical layer model are discussed in Section 5.2 and Section 5.3 respectively.

5.2 Mobility Modeling

A good mobility model captures the essence of vehicle movements in the real-world, while eliminating irrelevant details to reduce simulation complexity. Selecting an appropriate mobility model is crucial to constructing a realistic simulation environment. In the literature, there are three major approaches that are proposed to simulate vehicle movements: a) dynamic mobility models, b) interlinked mobility models, and c) trace-based mobility models. In the following, each of these three approaches and their simulation environment is briefly introduced.

5.2.1 Dynamic Mobility Models

A large number of dynamic mobility models have been proposed, mainly in the MANET community Camp *et al.* [2002], such as random walk, random way point and Manhattan Grid model. These models are termed dynamic because nodes that comply with these models may dynamically, and in most cases randomly, change their trajectory during simulation. According to predefined rules, a node may stop, change speed, or change direction when certain triggering conditions are met, such as a time or distance limit.

Another type of dynamic mobility model controls a group, instead of a single node, e.g., the nomadic community mobility model Sánchez & Manzoni [2001], and the reference point group mobility model Hong *et al.* [1999]. Group mobility

models often have a reference point or reference velocity, which other nodes follow. A following node may have its own autonomic behavior, i.e., deviating from the common reference to a certain degree. For example, in the reference velocity group mobility model, a node has a base vector, a deviation vector, and a deviation factor which regulates the degree that a node can deviate from the reference velocity.

5.2.2 Interlinked Mobility Models

Different from generic mobile ad hoc networks, nodes in vehicular networks rarely move with complete randomness, and are confined within a specific geographic area, e.g., a segment of road, or within a junction. The specific mobility pattern of a vehicle makes the aforementioned all-purpose mobility models inapplicable in VANETs.

Traffic simulators such as VISSIM PTV [2004] can provide realistic and high fidelity vehicle trajectories, and are often used by civil engineers to study traffic patterns on highways, bridges, or junctions. Interestingly, traffic simulators have drawn the attention of researchers in the vehicular communication community, and efforts have been made to interlink traffic and network simulators, and create two-way real-time data flows in between Sommer & Dressler [2008].

To achieve co-simulation between a traffic simulator and a network simulator, it is necessary to exchange control and data messages between them, e.g., a vehicle's current position, speed, and received messages. As a result, synchronization points during which data exchange takes place are defined during the co-simulation. Based on the behavior during the synchronization process, four different approaches are proposed in the literature.

Approach 1 (Figure 5.3): At each synchronization point, vehicles in both simulators are synchronized at the same position. Later, the traffic simulator moves first and feeds the new results, e.g., the coordinates of vehicles, to the network simulator. After receiving this update, the network simulator sets the new coordinates directly, which synchronizes itself with the traffic simulator again. The drawbacks of this approach are that not all network simulators support such "teleport" operations, e.g., NS-2 Simulator [1989], and the sudden movement in

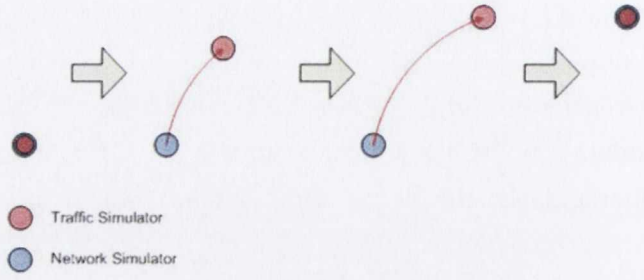

Figure 5.3: Inter-linking simulators approach 1

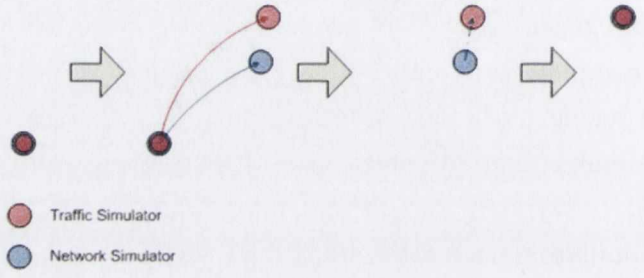

Figure 5.4: Inter-linking simulators approach 2

the network simulator, i.e., vehicles halt at each synchronization point, which may cause unrealistic behavior such as messages being sent from identical positions.

Approach 2 (Figure 5.4): Similar to approach one, vehicle positions are calibrated at synchronization points. Subsequently, the traffic simulator estimates the vehicle's next location and notifies the network simulator. Then parallel simulation is executed since both simulators are aware of the vehicle's next position. At the next synchronization point, the vehicle's position in traffic simulator, which is the correct one, overwrites the estimated position in the network simulator. Due to the estimation error, the vehicle in the network simulator therefore appears to have experienced a short teleport, but not a long teleport as in approach 1. One of the drawbacks of this approach is the same with approach 1, i.e., the unrealistic behavior caused by the teleport. In addition, due to the parallel nature of the co-simulation, the delivery of messages from the traffic simulator to the network simulator must be postponed to the next synchronization point.

Approach 3 (Figure 5.5): This approach is similar to approach 2, with the

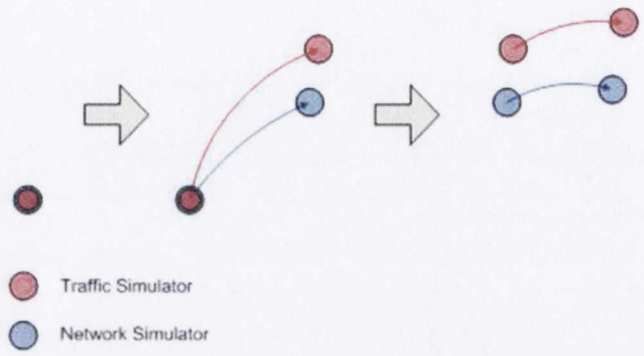

Figure 5.5: Inter-linking simulators approach 3

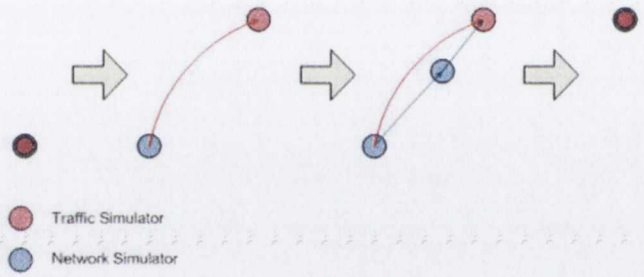

Figure 5.6: Inter-linking simulators approach 4

only difference being that vehicles do not calibrate to the same position at each synchronization point. The traffic simulator only sends the next estimated position to network simulator and the deviations between those two simulators are never corrected. The rationale of such an approach is to avoid the inconsistency caused by the teleport. The drawback is that there will be no chance for the vehicles in the network simulator to be aligned with the traffic simulator ever again.

Approach 4 (Figure 5.6): The essence of this approach is similar to approach 1 in the sense that traffic simulator always precedes the network simulator and provides vehicles' actual position, instead of an estimation. The traffic simulator also informs the network simulator regarding the messages that need to be delivered during this step and the timing of such messages. Subsequently, an exact "replay" is performed in the network simulator. This approach is effectively a higher resolution approach, compared to approach 1, as a vehicle travels

Figure 5.7: A segment of the M50 highway in Dublin city

smoothly via multiple positions rather than a single jump, and events such as broadcasting a message are performed at each intermediate position, rather than at the synchronization points.

5.2.3 Trace-based Mobility Model

Although inter-linking simulators provides high fidelity in evaluating protocols in VANETs, the associated communication overhead and implementation complexity are of concerns for researchers. In fact, if the trajectory of a vehicle in the traffic simulator is not affected by the messages received from the communication layer, i.e., a vehicle travels along a message-independent trajectory, there is no need to synchronize these two simulators in real-time. Hence a low-cost trace-based approach is more appropriate in this case.

In the trace file approach, traffic simulator generates a trace file that records the trajectories of all vehicles during the entire simulation. Taking the generated trace file as input, the network simulator manipulates vehicles accordingly. In this approach, simulations are executed in the traffic simulator and network simulator separately, which eliminates the communication overhead in linking two simulators, reduces implementation complexity, and accelerates the simulation speed.

5.2.4 Simulation Environment

In the evaluation study, the trace-file approach is used. The traffic simulator, VISSIM, generates realistic vehicle traces, which are imported and replayed in the network simulator OPNET Modeler [2009]. The scenario simulated in VISSIM

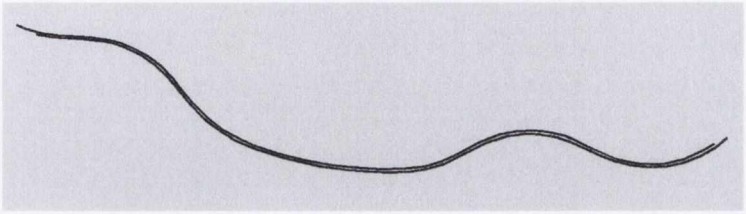

Figure 5.8: Simulated highway segment in VISSIM

is composed of a 4.75 kilometer segment of a highway in the suburb of Dublin city as depicted in Figure 5.7 and Figure 5.8. On each direction of the highway there are three lanes. Cars are inserted at both ends of the highway following a Possion process. The speed and the throughput of the vehicle are configurable in VISSIM, which creates a variety of scenarios represented by customized trace files. In VISSIM, vehicles follow a driver model that may accelerate, decelerate or change lanes according to the front vehicles or certain specified rules.

The trace file generated from VISSIM is a simple text-based, formatted file. Each vehicle that participates in the traffic simulation has its own trace file, which is mapped to a corresponding vehicle in OPNET. A trace file includes the following information: the entering and leaving time of the vehicle in the simulation, coordinates at each sampling instance, and the sampling interval. An example is given in Figure 5.9.

To coordinate the vehicle movement, a centralized mobility controller is implemented in OPNET. At the beginning of the simulation, the mobility controller reads in the trace files for all vehicles that will participate in the simulation, and activates and deactivates them according to their subsequent entering and leaving times. When the simulation starts, the mobility controller periodically reads each vehicle's new position from their respective trace files, and manipulates them directly in OPNET.

5.3 Physical Layer Modeling

In this section, models with regard to the physical layer used in the OPNET simulator are described. The OPNET's pipeline stages, which is a procedure to model radio transmission is introduced first. Next, the adopted propagation

```
Entering time: 0 s
Sample interval: 0.1 s
X-position, Y-position (meters):
218.836195915823,909.806768708813
217.214461043352,910.098940668913
215.592726170881,910.391112629008
213.970991298409,910.6832845891
212.349256425938,910.975456549191
.
.
.
189.644968211323,915.06586399043
188.023233338851,915.358035950518
186.401498466378,915.650207910606
184.779763593906,915.942379870695
183.158028721433,916.234551830783
Leaving time: 2.20000000000005 s
```

Figure 5.9: A example of the trace file

model, interference model and signal reception model, which play a vital role in determining the evaluation results, are presented.

5.3.1 Introduction of Pipeline Stages in OPNET

In OPNET, wireless communication is modeled by a mechanism called the "radio transceiver pipeline". Such a mechanism is provided by OPNET's built-in model library, but can be customized and modified based on users' specific needs. From the transmitter module to the receiver module, the radio transceiver pipeline (pipeline for short thereafter) in the middle is composed of 14 stages as depicted in Figure 5.10, most of which need to be executed for each transmitter-receiver pair due to the broadcast nature of wireless communication. Each stage provides a specific functionality in determining the final result of a wireless transmission, i.e., accept or reject, between a sender and receiver pair. If an early rejection has been decided, later stages are skipped to avoid unnecessary computation.

In the following, a brief introduction to these stages is given, and special attention is paid to stage 7 "received power", stage 8 "interference noise", and stage 11 "bit error rate". Detailed functionality and specification for the other

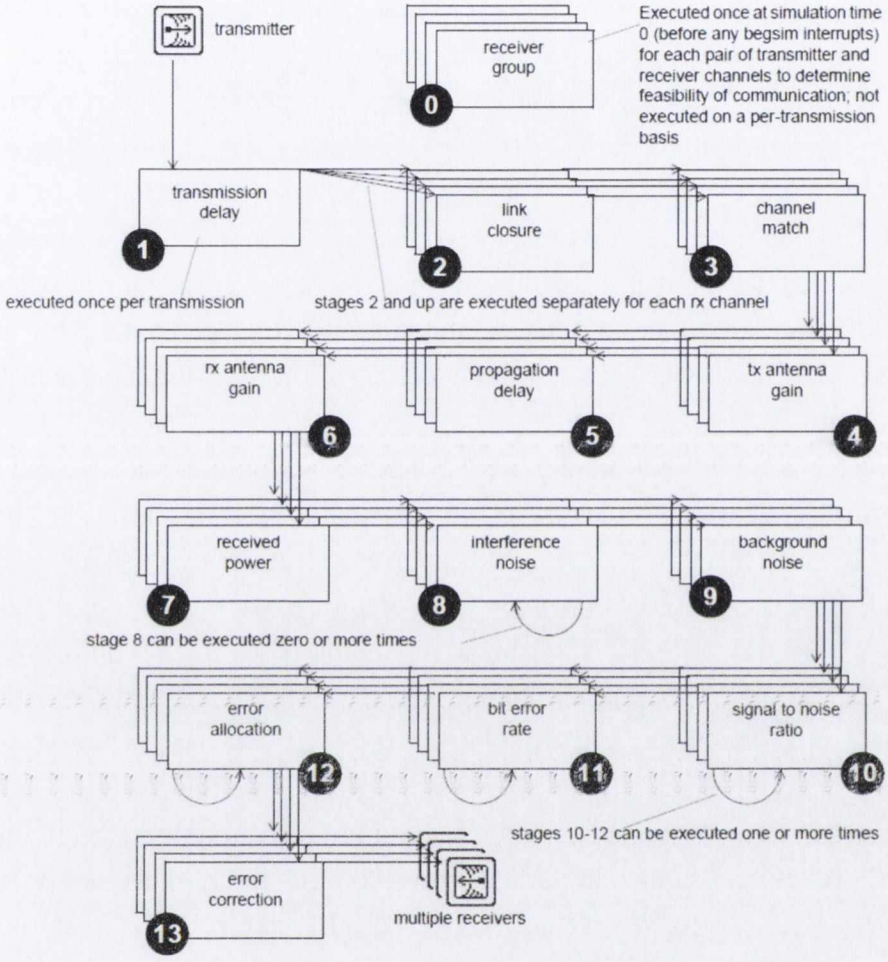

Figure 5.10: Pipeline stages in OPNET

stages can be found in the OPNET documentation: "Wireless - Radio Transceiver Pipeline".

The modeling of a wireless transmission, i.e., broadcasting a packet on the wireless medium, starts when a packet is delivered to the wireless transmitter module from the upper layers. In the first a few stages (before stage 7), the eligibility of a possible communication channel between transmitter and receiver is verified (stage 0, 2 and 3). Possible reasons for disqualifying a possible communication channel include: disjunctive frequency bands, physical separation, antenna nulls, etc. In addition to the eligibility tests, other parameters which

are used in subsequent stages are calculated such as transmission delay (stage 1), propagation delay (stage 5) and antenna gains (stage 4 and stage 6).

Stage 7 computes the received power of the arriving packet, which takes into account the communication distance, transmitter power, antenna gain and signal propagation model. The received power is a key indicator as to whether the information in the packet can be captured by the receiver. There exists a number of propagation models in the literature, such as free space model, two-ray ground model, log-normal model, and Nakagami model. By default, OPNET uses the free space model. A detailed introduction to these propagation models is discussed in Section 5.3.2.

The correct reception of a packet depends not only on the received power, but also the concurrent transmissions in the vicinity of the receiver, i.e. interference. The collective impact of the interference on the received signal is modeled in stage 8. The specific implementation of interference modeling in OPNET and other related work regarding this issue are discussed in Section 5.3.3.

In stage 9, the background noise, which usually stems from thermal or galactic noise, and emissions from neighboring radios is calculated. Based on the computed received signal power, interference noise power and background noise, the Signal to Noise Ratio (SNR) is calculated in stage 10. Note that in OPNET, the SNR value is dynamically updated whenever the interference power varies.

Based on the SNR at the receiver, the Bit Error Rate (BER) is calculated in stage 11, and used in stage 12 to estimate the specific number of errors in a packet segment. The decision to accept or reject a packet based on the number of errors and the ability to correct the errors is modeled in stage 13. The detailed process of deriving BER from SNR is introduced in Section 5.3.4.

5.3.2 Propagation Model

Theoretically, the received signal strength can be exactly calculated if all elements in the environment such as the buildings, trees, pedestrians and other geographic features are precisely known. Although significant efforts e.g., using ray tracing techniques, have been made towards this direction, such an approach is computationally too expensive. A more affordable approach is to use geo-

graphic information, such as the distance between transmitter and receiver to model signal attenuation on a large scale, and tune the result with small-scale fading models, which are often derived from empirical statistics.

Of all large-scale fading models, the free space model is the most basic one. The free space model assumes an ideal environment, in which the path loss, and the received power is given in Equation 5.3 and Equation 5.4 respectively:

$$\gamma = \frac{\lambda^2}{16\pi^2 r^2} \tag{5.3}$$

where γ is the path loss, λ is the wavelength of the transmission signal, and r is the distance between the transmitter and the receiver.

$$P_{rx} = P_{tx} G_{tx} \gamma G_{rx} \tag{5.4}$$

where P_{rx} is the received power, P_{tx} is the transmission power, G_{tx} is the transmitter antenna gain and G_{rx} is the receiver antenna gain.

A more realistic propagation model that involves the reflection from the ground is the two-ray ground model. At short distances (within critical distance), the received power attenuates at a rate inversely proportional to r^2, and diminishes at r^4 thereafter. The critical distance d_c is given in Equation 5.5:

$$d_c = \frac{4\pi h_t h_r}{\lambda} \tag{5.5}$$

where h_t and h_r are transmitter and receiver antenna hight.

In the two-ray ground model there are two slops that represent two attenuation regions which are divided by the critical distance. A generalization of the two-ray ground model allows the slopes to be parameters, with an additional random variable representing small-scale fading. The receiver power of the log-normal model is given in Equation 5.6, and Equation 5.7:

$$P(r) = \begin{cases} P(d_0) - 10\gamma_1 \log_{10}\left(\frac{r}{d_0}\right) + X_{\sigma 1} & \text{if } d_0 \leq r \leq d_c \\ P(d_0) - 10\gamma_1 \log_{10}\left(\frac{d_c}{d_0}\right) - 10\gamma_2 \log_{10}\left(\frac{r}{d_c}\right) + X_{\sigma 2} & \text{if } r > d_c \end{cases} \tag{5.6}$$

where γ_1 and γ_2 are the path loss exponents in the two regions, and σ_1 and σ_2

are the standard deviations of the random variables $X_{\sigma 1}$ and $X_{\sigma 2}$, which have normal distribution describing the small-scale fading. d_0 is a reference distance at which the received power $P(d_0)$ complies with the free space model. In our study, $d_0 = 1m$, $\gamma_0 = 2$, and $\lambda = c/f \approx 5.0847$ cm , we have:

$$P(d_0) = -10\gamma_0 \log_{10}\left(\frac{4\pi d_0}{\lambda}\right) \tag{5.7}$$

Small-scale fading is caused by constructive and destructive interference between multi-path components. Based on whether or not there is a line-of-sight (LOS), the Rayleigh and Rician distribution are used to model small-scale fading. The Nakagami distribution can model both types of distributions and is demonstrated to be the best fit with the data collected from real world experiments Rubio *et al.* [2007], and therefore is chosen in our study. The following Equation 5.8 describes the Nakagami probability density function (PDF) of the received signal amplitude x:

$$f(x; m, \Omega) = \frac{2m^m x^{2m-1}}{\Omega^m \Gamma(m)} exp\left(-\frac{mx^2}{\Omega}\right) \tag{5.8}$$

where m is a shape parameter of the Nakagami distribution that defines the fading intensity, $\Omega = E[x^2]$ is an estimate of the average received power at a certain distance, and Γ is the Gamma function.

By varying the m parameter, the Nakagami model simulates various fading severities. For $m = 1$, Nakagami describes Rayleigh distribution, which represent non-LOS communication, and when m > 1, Nakagami describes a Rician distribution (with LOS). In the following evaluation, Ω is set with results obtained from the two-ray ground model, and the m parameter is varied to model different fading scenarios.

5.3.3 Interference Model

Among many existing interference models, the additive interference model and capture threshold model are the most commonly used. For example, additive interference is used in OPNET and GloMoSim Gerla *et al.* [1999], and the capture threshold model is used in NS-2. In the additive interference model, a received

packet is decoded by considering the sum of all other on-going transmissions and environmental noises. The SNR calculated in the additive interference model is given in Equation 5.8:

$$SNR_s = \frac{P_s}{P_n + \sum_{i \in I \setminus \{s\}} P_i} \tag{5.9}$$

where P_s is the received power from the signal, P_n is the power of back ground noise, I denotes the set of on-going transmissions, and P_i is the received power from transmission i.

The packet reception condition for the additive interference model is given in Equation 5.10:

$$SNR_s > SNR_{threshold} \quad \text{throughout the transmission of s} \tag{5.10}$$

where the $SNR_{threshold}$ is a threshold for SNR that depends on the specific modulation method and coding scheme.

In the capture threshold model, the received signal is compared to only one of the concurrent interference signals, and the packet reception condition for the capture threshold model is given in Equation 5.11 and Equation 5.12:

$$P_s > P_n \quad \text{and} \tag{5.11}$$

$$P_s > P_i \quad \forall \text{ i} \in \text{I} \setminus \text{s} \tag{5.12}$$

Compared to the additive interference model, there are two major drawbacks of the capture threshold model. First of all, the capture threshold model considers the impact of only one interfering signal on the receiving signal, and categorizes such an interfering signal as either not causing a collision ("benign") or causing a collision ("malign"). In certain conditions, a collection of interfering signals which are harmless individually may have a cumulative and adverse impact on the receiving signal, which cannot be accurately modeled.

In addition, in the capture threshold model, the receiving signal can be decoded successfully, as long as its signal strength is higher than the interfering signal with even a slightest margin. However, due to the existence of environ-

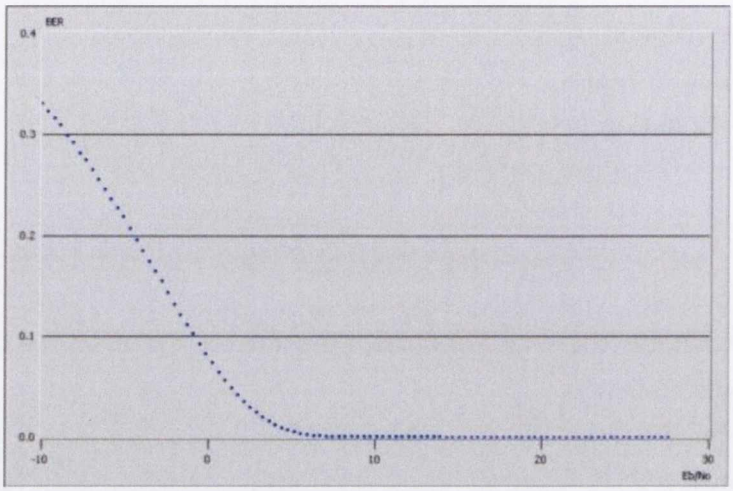

Figure 5.11: Mapping from SNR to BER using Quadrature Phase-Shift Keying (QPSK) modulation scheme in OPNET

mental noise, such an assumption do not hold when the receiving signal strength is weak and comparable to the background noise. In Equation 5.11 and Equation 5.12 for example, suppose that $P_s > P_n$, $P_s \ / \ P_n \approx 1$, and $P_n = P_i$. The packet can be decoded successfully in the capture threshold model, but not in the additive interference model as background noise becomes more prominent.

In OPNET, the additive interference model is used with the additional feature that, not only all arriving interferences are considered, the temporal differences of the arriving interfering signals are considered as well. Each new arriving interference triggers a recalculation on the receiving packet's SNR. In addition, there is a channel locking mechanism used in the OPENT implementation to avoid the possibility of successfully decoding two overlapped packets simultaneously.

5.3.4 Signal Reception Model

The signal reception model determines whether to accept or reject an arriving packet based on its SNR. In a simple threshold-based approach, if the receiving SNR exceeds a predefined threshold "SNRT", the packet is accepted, otherwise rejected. A more complicated approach, which is adopted in OPNET, maps receiving SNR to a specific Bit Error Rate. Such a mapping from SNR to BER

(Figure 5.11) is derived from empirical statistics and is a practical approximation from a networking point of view. Consequently, the success of decoding a packet is probabilistic and depends on the SNR, packet length, and specific modulation method used for transmission.

5.4 RRP Specific Evaluation

As described in Chapter 3 and Chapter 4, the proposed RRP protocol adopts a two-tier architecture where the lower layer manages mobility while the upper layer allocates time slots. To better understand the characteristics of RRP, a detailed performance evaluation is conducted with regard to the virtual cluster and the real-time scheduling layers separately, then their respective impact on RRP's overall performance is investigated. The evaluation study of RRP as a single protocol is presented in Section 5.7, where RRP is also compared with 802.11p and RR-ALOHA.

The lower virtual cluster layer and the upper real-time scheduling layer are evaluated in Section 5.4.1 and Section 5.4.2 respectively in substantial details. A general review regarding the functionality of a specific layer is presented first, which is followed by an introduction to the evaluation methodology. The metrics and parameters used in the evaluation are presented and discussed, after which the evaluation results are depicted and elaborated. Conclusions are presented at the end of each subsection.

5.4.1 Virtual Cluster Layer Evaluation

The virtual cluster layer is composed of three major components: mobility prediction, message dissemination, and neighbor identification. Since the neighbor identification function produces the outputs of the virtual cluster layer, it is the focus of our evaluation. In the following section, the metrics that characterize the performance of a neighbor identification procedure are discussed first, which is followed by the analysis of the parameters that may influence its performance. The evaluation of the virtual cluster layer itself is presented in Section 5.4.1.2, and the impact of the virtual cluster layer on the overall RRP is presented in

Section 5.4.1.3.

5.4.1.1 Evaluation Metrics and Parameters

What are the key performance indexes for the neighbor identification procedure?

Ideally, to evaluate the performance of a neighbor identification procedure is simply to compare the *actual* neighbors with the *identified* neighbors. However, the definition of an actual neighbor is ambiguous. An actual neighbor can be classified by one of the two categories: 1) a neighbor with certain properties, e.g. the distance to the node of interest, regardless of its actual behavior, e.g., causing a collision or not, or 2) defined purely based on its actual behavior.

There are two drawbacks for the second category above. First of all, due to the fluctuation of the received signal strength, the truth of a neighbor identification result is skewed by the uncertainty of the radio propagation. For instance, by such a definition, a very close node is not an actual neighbor if its actual transmission is weak, while on the contrary, a very distant node with a strong signal will be categorized as an actual neighbor. Secondly, because the correctness of a neighbor identification algorithm is only known when a node actually transmits, the evaluation of a neighbor identification algorithm depends on the specifics of the data traffic. In an extreme case for example, if a node does not transmit at all, it is impossible to know the performance of the node's neighbor identification algorithm.

Considering the above factors, the former approach is chosen in the evaluation study. The actual neighbors are defined as the set of nodes that are located within 2-hops of a node in question. A hop is defined as the deterministic communication range, within which a message with a specific size can be received with above 95% probability. The actual collisions are also measured and monitored, and cross referenced with the neighbor identification results. To summarize, the neighbor identification procedure is evaluated by two methods in the evaluation study:

1) Analyze the discrepancy between the *identified* neighbors and the *actual* neighbors. It is worth noting that a misidentification of this kind does not necessarily lead to a collision, as long as the misidentified node and the node in question do not "actually" transmit simultaneously.

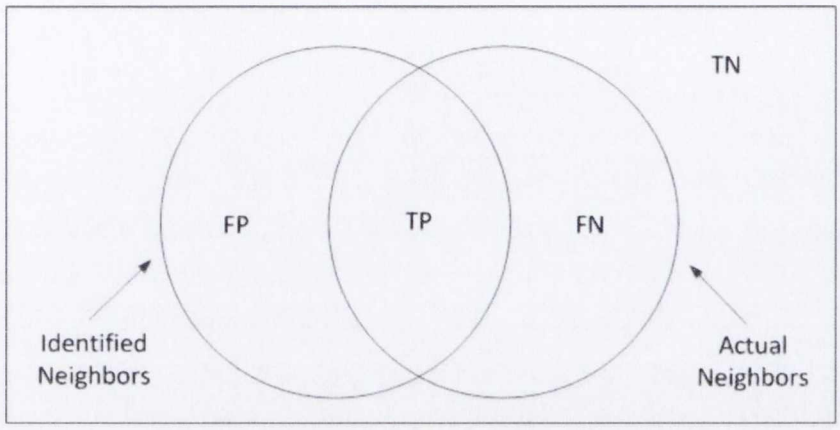

Figure 5.12: Possible results of a neighbor identification procedure

2) Analyze each transmission collision that actually occurred, and examine the neighbor identification results that are generated by these two colliding nodes. Since a collision *actually* happened, those misidentifications are recorded and analyzed as they cause a degradation in the virtual cluster layer's performance.

The comparison between identified and actual neighbors result in the following four possible outcomes, as illustrated in Figure 5.12:

- True Positive (TP): An actual neighbor that is identified as a neighbor.
- True Negative (TN): An actual non-neighbor that is identified as a non-neighbor.
- False Positive (FP): An actual non-neighbor that is identified as a neighbor.
- False Negative (FN): An actual neighbor that is identified as a non-neighbor.

For the neighbor identification evaluation, two metrics are of interest: 1) Sensitivity (Equation 5.13) - the percentage of true positives in all actual neighbors, i.e., the probability that actual neighbors can be identified. and 2) Positive Predictive Value (Equation 5.14) - the percentage of true positives in all identified neighbors. i.e., the probability an identified neighbor is an actual neighbor.

$$Sensitivity : TPR = \frac{TP}{TP + FN} \tag{5.13}$$

$$PositivePredictiveValue : PPV = \frac{TP}{TP + FP} \tag{5.14}$$

A false positive unnecessarily identifies a node as a neighbor, which leads to resource waste and decreased efficiency. A false negative on the other hand, incorrectly identifies an actual node as a non-neighbor, which may potentially cause an actual collision. Consequently, a neighbor identification procedure with higher sensitive value, i.e., less FN, has higher communication reliability, while a higher positive predictive value, i.e., less FP, has better efficiency.

However, to achieve both communication reliability and efficiency at the same time is difficult. A risky neighbor identification which inclines to ignore neighbors may increase efficiency but decreases reliability. On the contrary, a conservative neighbor identification improves reliability by identifying more neighbors, but is less efficient. The trade-off between the risky and conservative approach will be evaluated and discussed in the following sections.

Another method to evaluate neighbor identification performance is to analyze the overall collision probability and the specific cause of a collision. It is assumed that collisions from more than two nodes are rare and thus not considered. For any collision that occurred between a reference node A, and the colliding node B, there are three possible scenarios: 1) node A does not know B (type-1 collision), 2) node A knows B, and identifies B as neighbor (type-2 collision), 3) node A knows B, and identified B as non-neighbor (type-3 collision). Type-1 collision are caused by message forwarding error or a fluctuation in signal strength, which results in insufficient knowledge regarding a specific neighbor and is thus irrelevant to neighbor identification. Type-2 collision is due to non-reciprocal neighbor identification, i.e., the colliding node does not identify the node of interest. Type-3 collision represents a neighbor misidentification, which is the focus of the evaluation.

In general, the collision probability is one of the performance indicators of the virtual cluster layer, as a high collision probability may indicate an inaccurate neighbor identification. However, it is worth noting that such inference is inconclusive as an increased collision probability may be attributed to other possibilities, such as channel variation or a flawed scheduling algorithm in which two

identified neighbors incorrectly use the same slot.

In addition to collision probability, the percentage of misidentification-caused collision (type-3 collision) is measured, which is also an indication of the virtual cluster layer's performance. For example, higher percentage of misidentification-caused collision results from less accurate neighbor identification.

What parameters may influence the performance of the neighbor identification process?

In the following evaluation study, three parameters: vehicle speed, Nakagami-m, and assumed communication range are chosen to investigate their impact on the neighbor identification performance.

Vehicle speed may influence the mobility prediction accuracy, which may degrade the virtual cluster performance. In the study, the vehicle speed is configured between 20 - 140 km/h.

The Nakagami-m is a parameter that tunes the signal strength variations in the wireless channel, which may affect the collision probability. A higher m value means less deviation in received signal strength. In the evaluation, m is configured from 0.5 to 2, which covers the fading severity scenario from very heavy fading (non-line-of-sight) to very low fading.

The assumed communication range is a key parameter in the neighbor identification procedure. A large assumed communication range makes virtual cluster more conservative, as more nodes are identified as neighbors, while a small one makes virtual cluster more risky.

5.4.1.2 Virtual Cluster Layer Performance Evaluation

The performance study of the virtual cluster layer is divided into three parts: the vehicle speed study, channel characteristics study, and the assumed communication range study. In all three sections, the vehicle density is fixed at 80 veh/km, and the beacon interval at 100 ms. The goal is to investigate how virtual cluster layer's performance, including sensitivity, positive predictive value, general collision probability, and percentage of misidentification-caused collision are affected by these three parameters. Note that for presentation clarity, the results for sensitivity and PPV are plotted in a single diagram. The collision probability and

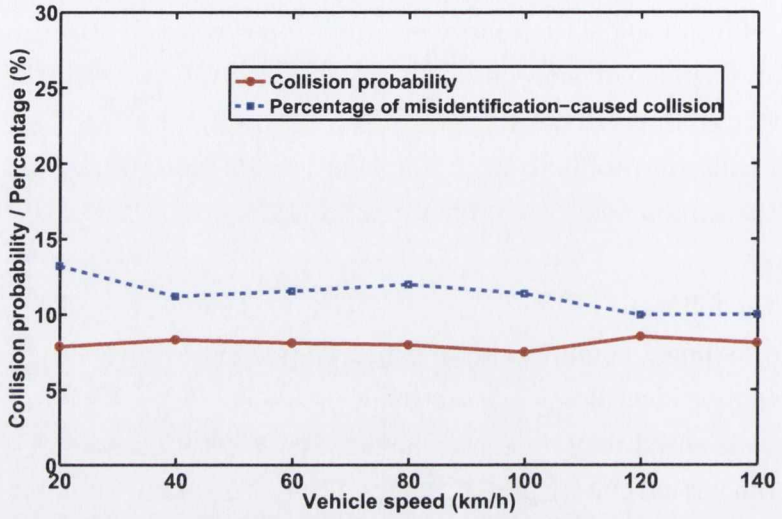

Figure 5.13: Collision probability vs. vehicle speed

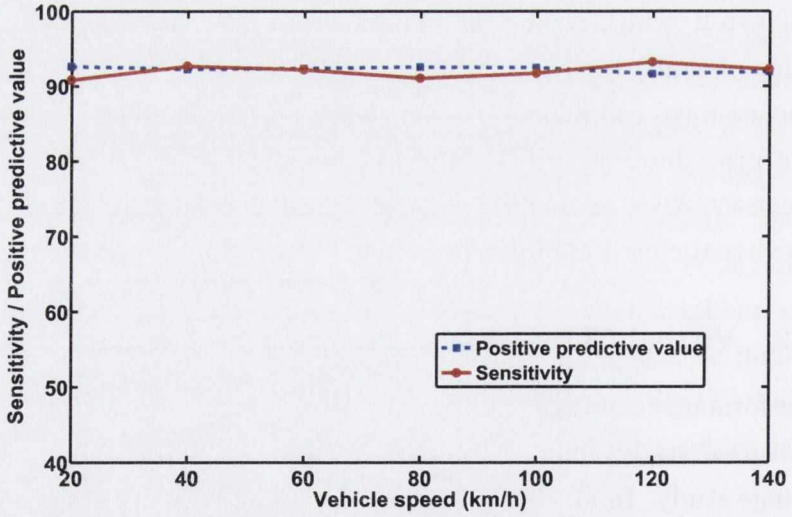

Figure 5.14: Sensitivity and PPV vs. vehicle speed

percentage of misidentification-caused collision are plotted together as well.

- Speed Study

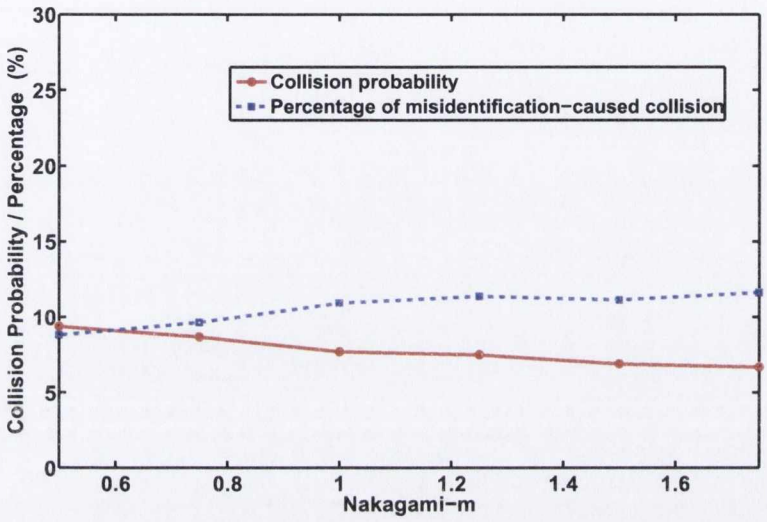

Figure 5.15: Collision probability vs. Nakagami-m parameter

As shown in Figure 5.13 and Figure 5.14, the performance of virtual cluster in terms of collision probability, the percentage of misidentification, sensitivity and PPV remains stable as the vehicle speed varies. The results demonstrate that the virtual cluster layer can adapt well in a mobile environment by predicting vehicles' positions and using relevance-based forwarding. The sensitivity remains above 90% of all vehicle speeds, which indicates that over 90 percent of the neighbors are successfully identified by the neighbor identification procedure.

- Nakagami Study

In Figure 5.15, the overall collision probability drops from 9% to about 6% as the channel fading becomes less severe. As fading alleviates, the number of interference-caused collision reduces, which leads to the increase of the percentage of misidentification-caused collision as depicted in Figure 5.15.

The sensitivity and PPV are both not affected by the changes in the channel conditions as shown in Figure 5.16. This phenomena is due to the fact that both actual neighbors and identified neighbors are defined by their relative distance, transmission power, and other parameters that are determined prior to the actual

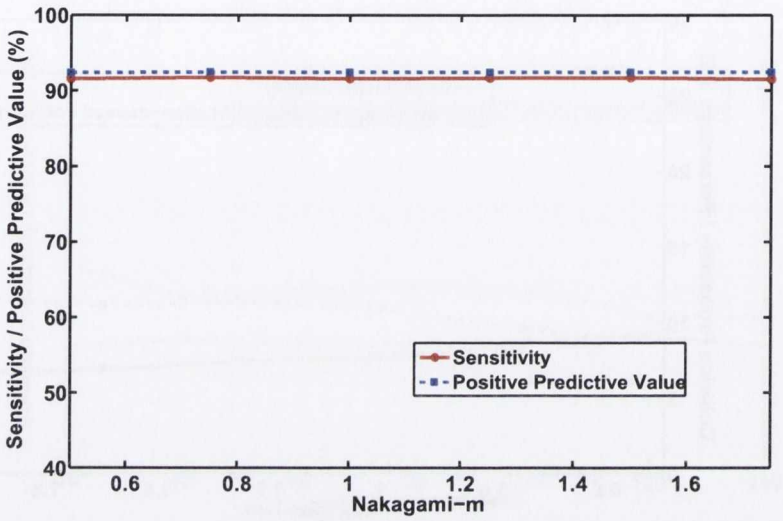

Figure 5.16: Sensitivity and PPV vs. Nakagami-m parameter

transmission. Thus the sensitivity and PPV are independent of the actual channel condition.

- Assumed Communication Range Study

When the virtual cluster layer becomes more risky and takes less neighbors into consideration, i.e., smaller assumed communication range, the performance of virtual cluster layer in terms of the communication reliability deteriorates, as both general collision probability, and the number of misidentification-caused collision increases as illustrated in Figure 5.17.

The degraded performance of the neighbor identification procedure when adopting the risky approach is also reflected in the sensitivity measurement as depicted in Figure 5.18. The sensitivity measure drops from 90% to as low as 40%, indicating that only 40% of the actual neighbors can be correctly identified. However, the reduced number of identified neighbors makes identified neighbors more likely to be an actual neighbor, i.e., an increased PPV value as depicted in Figure 5.18.

Based on the above comparison and discussions, it can be observed that among the three chosen parameters: speed, Nakagami-m and assumed communication

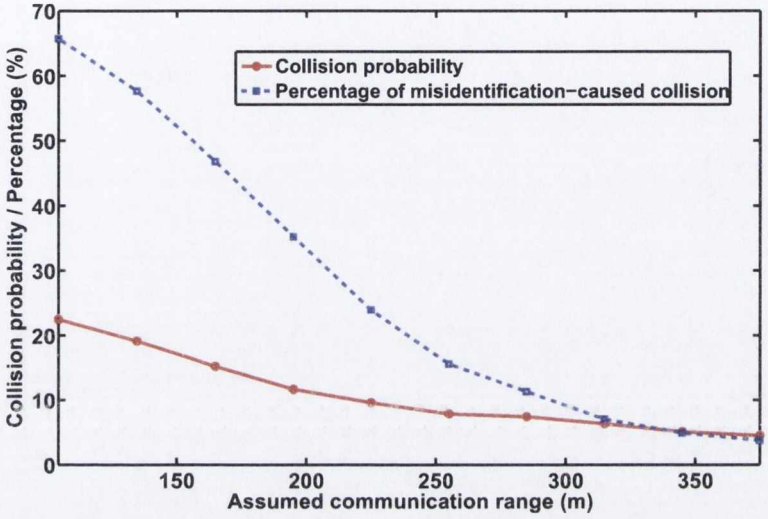

Figure 5.17: Collision probability vs. assumed communication range

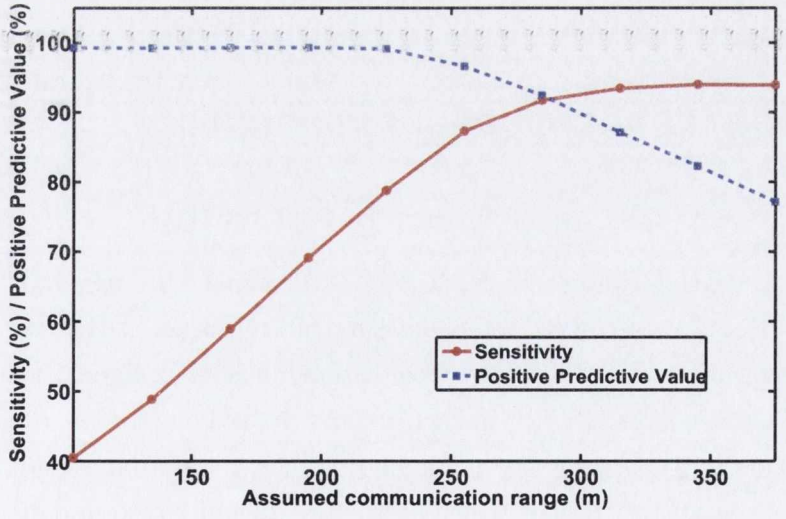

Figure 5.18: Sensitivity and PPV vs. assumed communication range

range, the last parameter, which determines how risky the virtual cluster layer behaves, has a greater impact on the overall performance of the virtual cluster layer.

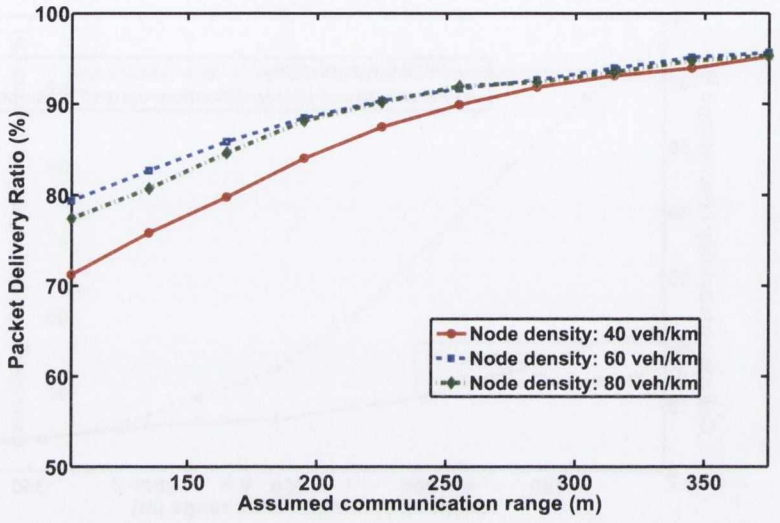

Figure 5.19: PDR vs. various assumed communication range

Consequently, in the next section, the assumed communication range is selected as the primary parameter, which will be tuned between risky and conservative, in the study of the virtual cluster layer's impact on the overall performance of RRP.

5.4.1.3 Virtual Cluster Layer's Impact on RRP

The goal of this evaluation is to study the impact of the virtual cluster layer on the overall protocol, which is measured in terms of PDR, reservation interval and mean staleness. The assumed communication range is tuned from risky virtual cluster layer (125 m) to conservative virtual cluster layer (375 m), and the actual communication range configured in the evaluation is approximately 250 m. In the study, node densities are also varied from 40, 60 and 80 vehicle/km for comparison.

For a specific node density, being risky makes the communication less reliable as PDR drops from above 90% to below 80%, as depicted in Figure 5.19. This result is expected since a risky virtual cluster layer reduces the sensitivity, i.e., more actual neighbors are neglected and not considered in the slot reservation

process. It is worth noting that when the virtual cluster layer is risky, the PDR in 40 vehicle/km is lower than in the 60 and 80 vehicle/km scenario, for the following reasons.

By adopting the message forwarding mechanism in RRP, the awareness range (in which high PDR transmissions are scheduled) is larger than the assumed communication range (typically more than twice as large). Consequently, when the assumed communication range shrinks, the awareness range also shrinks but may still be larger than, or covering a large portion of, the actual transmission range.

For convenience, assume that the size of the awareness range equals to the size of the actual transmission range, and the assumed communication range is sufficiently smaller than the awareness range as well as the actual transmission range. Ideally, if the message forwarding mechanism works perfectly, no collision will occur within the actual transmission range, i.e., no 1-hop collision. However, since nodes beyond actual transmission range are not considered, hidden terminal problems, i.e., 2-hop collisions are still possible, which decreases the PDR in the actual transmission range. The reduced PDR leads to a *chain reaction*, i.e., the less reliable forwarding mechanism will make a collision even more likely.

Consequently, given that the assumed communication range is small, which implies less reliable communication, the PDR in sparse networks is further degraded by the dysfunctional message forwarding mechanism as shown in the 40 vehicle/km case in Figure 5.19.

In terms of reservation interval, for the same node density, a more conservative virtual cluster layer identifies more neighbors, which means longer time to reserve a slot, i.e., higher reservation interval as depicted in Figure 5.20. For various node densities, it follows the same principle that a larger number of neighbors leads to a longer time to reserve a slot.

In Figure 5.21, the staleness is evaluated with respect to the assumed communication range. In the 40 vehicle/km node density scenario, the "U" shape is quite obvious: overly conservative (large assumed communication range) leads to more reliable transmission, but more neighbors need to be considered, which leads to longer reservation interval and larger staleness. On the other hand, overly risky neighbor identification may reduce the number of neighbors being

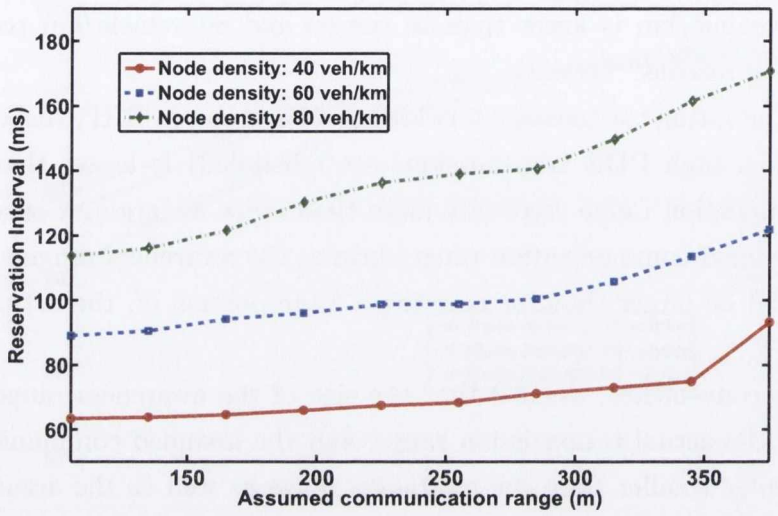

Figure 5.20: Reservation interval vs. various assumed communication range

considered during slot allocation, but also increases the collision probability and reduces the reliability of transmission, which deteriorates the overall performance resulted in larger staleness observed.

In the 60 and 80 vehicle/km node density scenarios, being overly conservative substantially increases staleness, which is similar to the 40 vehicle/km scenario. However, when being overly risky, the staleness does not increase as expected. The reason is similar to the analysis regarding the PDR, that the built-in 2-hop forwarding mechanism in RRP compensates for some of the lost messages suffered from hidden terminals, which improves the performance in terms of staleness. This result demonstrates the flexibility of RRP in certain circumstances, e.g in dense networks, that a risky neighbor identification approach can be adopted to boost overall performance without substantial penalty. However, to achieve a QoS guarantee in all scenarios, it is necessary to follow standard RRP procedures.

5.4.1.4 Conclusion

In this section, three parameters: vehicle speed, channel fading parameter and assumed communication range are chosen to determine the most prominent param-

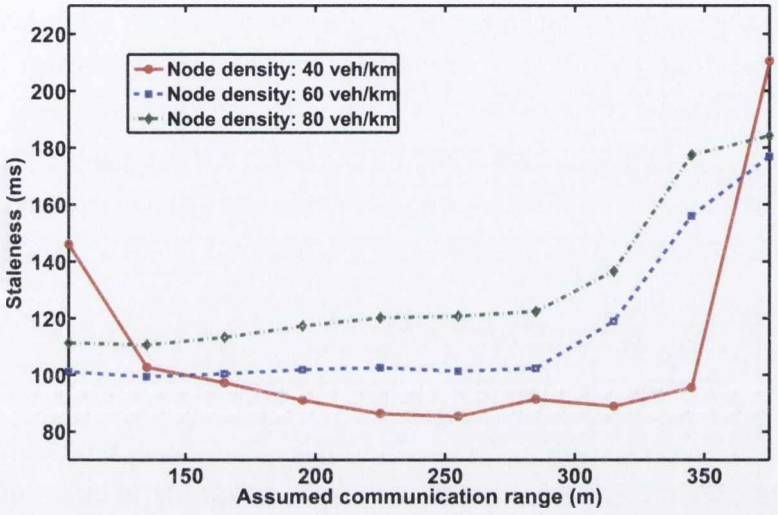

Figure 5.21: Mean staleness vs. various assumed communication range

eters that may influence the virtual cluster layer's performance. The evaluation results show that the virtual cluster layer's performance is largely not affected by vehicle speed and only marginally affected by the Nakgami-m parameter. The results also show that the assumed communication range, which effectively determines the number of neighbors that the virtual cluster layer identifies, has a major impact on both the virtual cluster's performance and the overall RRP performance. An overly risky virtual cluster layer, which identifies fewer neighbors, or a overly conservative virtual cluster both have a negative impact on the RRP performance in terms of staleness. Consequently, given the actual communication range, to set an appropriate assumed communication range is critical in order to preserve the reservation interval guarantee claimed by RRP.

5.4.2 Scheduling Layer Evaluation

The responsibility of the scheduling layer is to reserve slots for communication with one's identified neighbors, based on the information provided by the virtual cluster layer. To evaluate the scheduling layer as a standalone layer, parameters in the virtual cluster layer are configured as constant. The performance of the

scheduling layer is evaluated on three levels: the slot level, the reservation interval level and the staleness level. The specific metrics used to characterize the scheduling layer performance on these three levels are described in Section 5.4.2.1.

Among many parameters that may influence the scheduling layer's performance, the hold-off interval and the allowed control message size are the most prominent parameters and are chosen in our study. The evaluation of scheduling layer performance given various hold-interval and allowed control message sizes is presented in Section 5.4.2.2 and Section 5.4.2.3 respectively.

5.4.2.1 Evaluation Metrics and Parameters

A slot is the most basic element of an allocation algorithm, and the evaluation study focuses on three aspects of a slot allocation: 1) indisputable decisions on slot allocation 2) analysis of undecided slots, and 3) the probability of delivering a slot allocation decision.

Firstly, the slot allocation decision made by the scheduling algorithm needs to be indisputable, i.e., none of the mutually-identified neighbors uses the same slot. In our evaluation study, the indisputable property is verified by analyzing the cause of each collision, and searching for the existence of a collision that is caused by mutually-identified neighbors.

In the slot allocation algorithm, the slot usage decision is made given internal and external information. However, if the information is not sufficient for a node to make a decision when the deadline for the slot is due, a undecided slot is considered lost. Undecided slots degrade the efficiency of slot allocation, which further compromises the reservation guarantees.

In the performance evaluation, the percentage of undecided slots is studied as well as their reasons for being undecided. A slot remains undecided due to three possible causes: 1) a pending node itself is not decided, 2) a pending node is decided but the decision message is not sent (possibly due to lack of a transmission opportunity), and 3) a pending node is decided, with its decision sent, but the message is lost during transmission. In the following evaluation, the percentage of each case is provided with analysis on their implications.

The ability to deliver slot decision messages is critical to the performance of

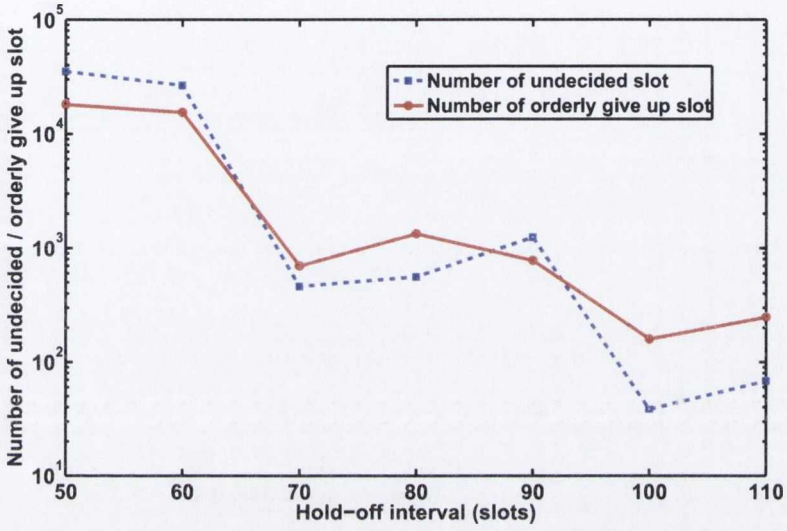

Figure 5.22: Number of undecided and orderly give up slots vs. hold-off interval

the scheduling algorithm. Low decision delivery rates lower reservation success rates, which increases the reservation interval and degrades the overall performance. Provided that a node decides on a slot, the probability that its neighboring nodes, which also consider this node as a neighbor, can successfully receive this decision is measured in the study.

The reservation interval which represents the duration between two consecutive reserved slots is measured in terms of its mean and distribution. The value of reservation interval is categorized into three classes: fast reservation ($<$ 1 maximum hold-off interval), slow reservation (1-3 maximum hold-off intervals), and beyond bound reservation ($>$ 3 maximum hold-off intervals). Staleness is measured as a high-level metric in determining the scheduling algorithm's performance.

5.4.2.2 Hold-off Interval Effect on Scheduling Layer Performance

In the slot scheduling algorithm, the hold-off interval is dynamically determined based on the number of neighbors. In the evaluation, the hold-off interval is fixed in each simulation setting in order to observe the impact of choosing different

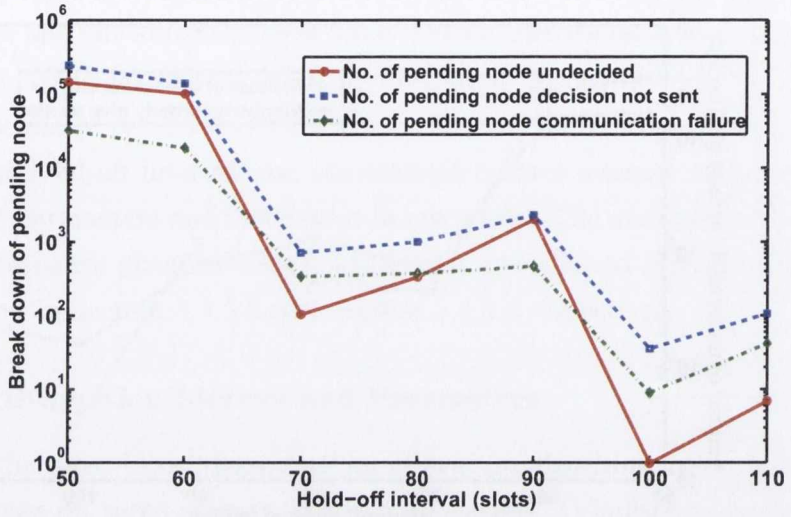

Figure 5.23: Breakdown of pending nodes vs. hold-off interval

hold-off intervals on RRP performance.

Firstly, the evaluation study focuses on the impact of various hold-off intervals on individual slots, in terms of the number of undecided, and orderly given up slots during the whole period of simulation. The node density is set at 80 vehicle/km. As illustrated in Figure 5.22, a small hold-off interval leads to an abrupt increase in the number of undecided slots. Excessively small hold-off intervals encourages nodes to reserve more, which increases the number of pending nodes of a given slot. As a result, slot reservation becomes more difficult, and a large number of slots remain undecided. Based on the above analysis, the hold-off interval should always remain sufficiently large for a specific node density.

The breakdown of the pending nodes attached to each undecided slot is studied in Figure 5.23. In general, the results demonstrate similar trends as the undecided slots, that the number of pending nodes increases significantly as hold-off interval shrinks. As for the specific reasons for a pending slot, the "decision message not sent" is the primary reason why a slot cannot be decided.

The overall performance of RRP in terms of average Reservation Interval (RI), percentage of unbounded RI, and average staleness, is studied with respect

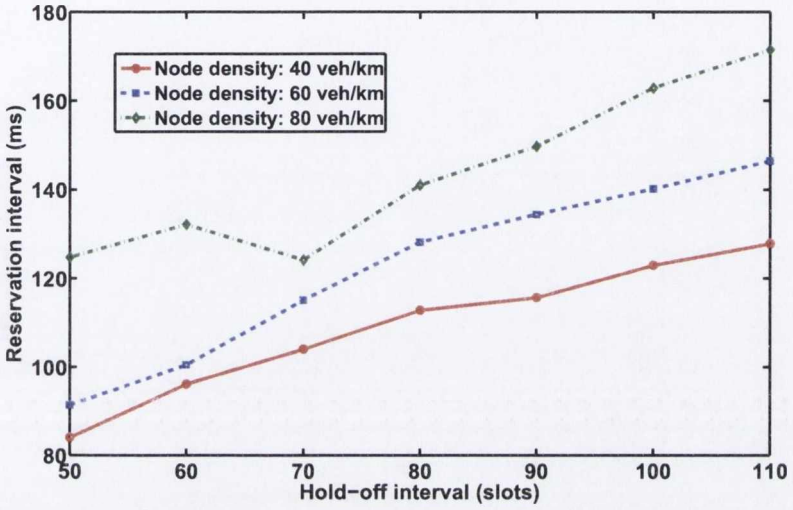

Figure 5.24: Reservation interval vs. hold-off interval

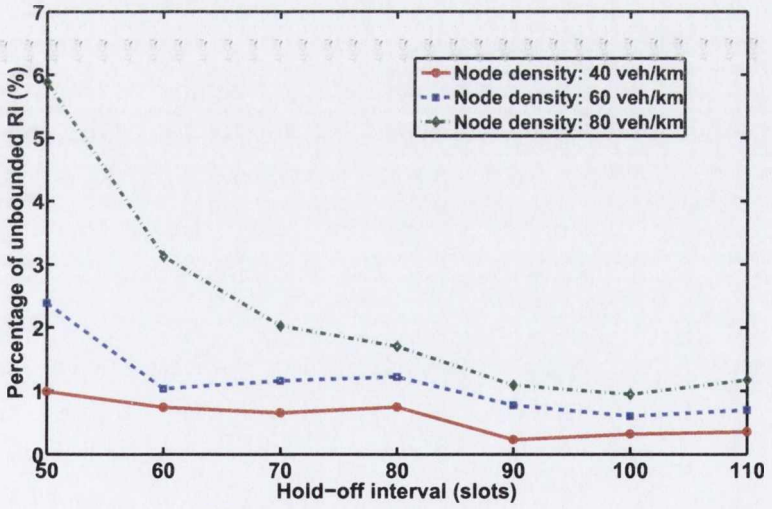

Figure 5.25: Beyond bound reservation interval vs. hold-off interval

to various hold-off intervals in different node densities. As depicted in Figure 5.24, the average RI in all node densities grows as hold-off interval increases, since a

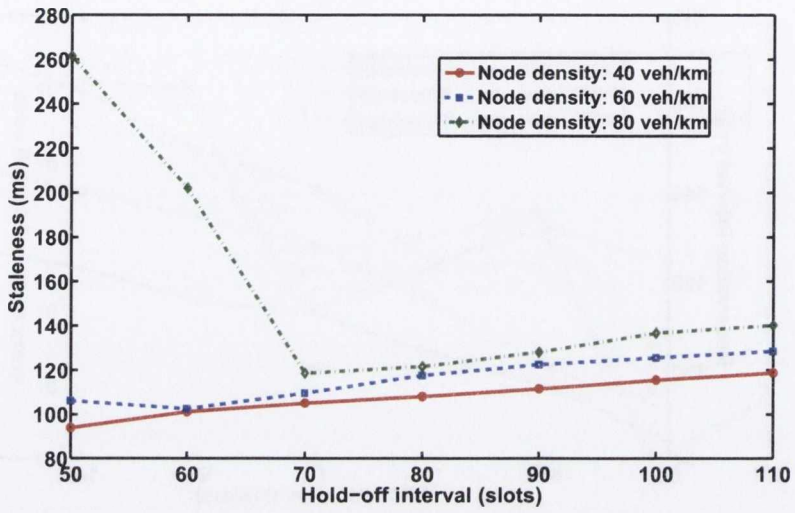

Figure 5.26: Mean staleness vs. hold-off interval

node is prohibited from reserving another slot during a longer period. In addition, the RI in high-density scenarios is larger than in low-density scenarios because it takes a longer time to compete with a larger number of neighbors.

It is worth noting in the high node density case (80 vehicle/km) that the RI remains high when the hold-off interval drops, as compared to other node density cases. This can be explained that in high density environments, the higher number of neighbors rather than the hold-off interval, becomes the dominant factor that prohibits nodes from reserving more slots.

The percentage of unbounded RI is directly affected by the hold-off interval as depicted in Figure 5.25. Given a specific node density, it grows quickly if the hold-off interval drops, especially in dense networks. Nevertheless, expanding hold-off interval does not help reducing the percentage of unbounded RI, which indicates that with sufficient resources, it is still possible to have unbounded reservation intervals.

As for average staleness, the general observation in Figure 5.26 is that the performance in sparse networks is better than in dense networks, as more contenders exist in the neighborhood. It is observed that the average staleness drops when

the hold-off intervals decreases. This is due to the fact that, in a given network, excessively large hold-off interval unnecessarily limits nodes from reserving more slots. However, when the hold-off interval decreases to a certain tipping point, the average staleness begin to pick up quickly, e.g., in 80 vehicle/km scenario. The reason is similar to our previous analysis that excessively small hold-off intervals significantly intensifies the contention among neighbors in a local area which leads to more undecided slots.

From the analysis above, it is observed that for a specific network density, there exist an optimal hold-off interval which takes into account the number of neighbors in the local area. If the hold-off interval is set excessively high, slots are wasted and the efficiency is low. If it is set excessively low, contentions will be intense which makes reserving a slot difficult. In our algorithm, the hold-off interval is set based on the current environment and is adjusted dynamically to the changes of node density in the local area.

5.4.2.3 Allowed Control Message Size Effect on Scheduling Layer Performance

The allowed control message size is the size limit on the proportion of control message versus payload message. Given the current configuration, the maximum packet size in a slot is 9000 bits, and therefore, the allowed control message size is tuned between 1000 and 9000 bits.

Compared to the hold-off interval study, the number of undecided slots and the break down of possible reasons show similar trend as depicted in Figure 5.27 and Figure 5.28. The undecided slot issue becomes severe when the allowed control message size is below 5000 bits. Without adequate bandwidth, the reservation decision messages are less likely to be sent out and received, which causes more slots to remain undecided.

The above rationale is confirmed by Figure 5.29, in which the decision message delivery success rate is measured against the allowed control message size. In all node density scenarios, 5000 bits is the tipping point for a good delivery rate ($> 98\%$), and 3000 bits is the threshold beyond which the delivery rate drops dramatically. The results demonstrate that in order to maintain an acceptable

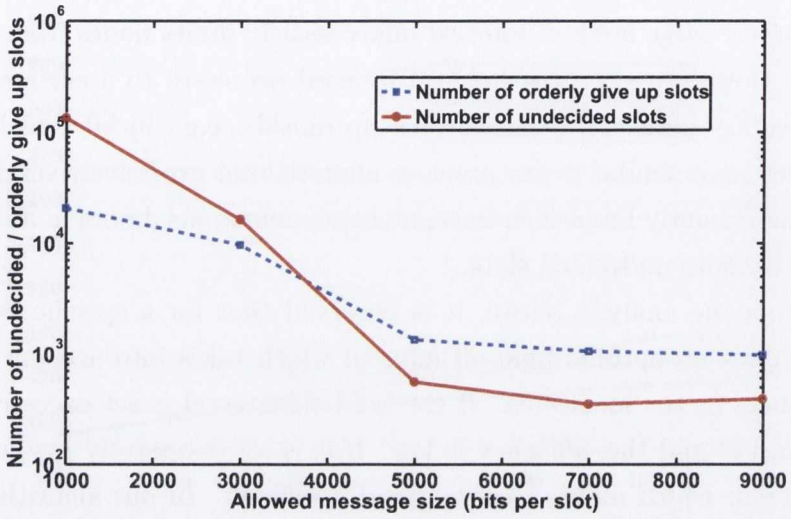

Figure 5.27: Number of undecided and orderly given up slots vs. allowed control message size

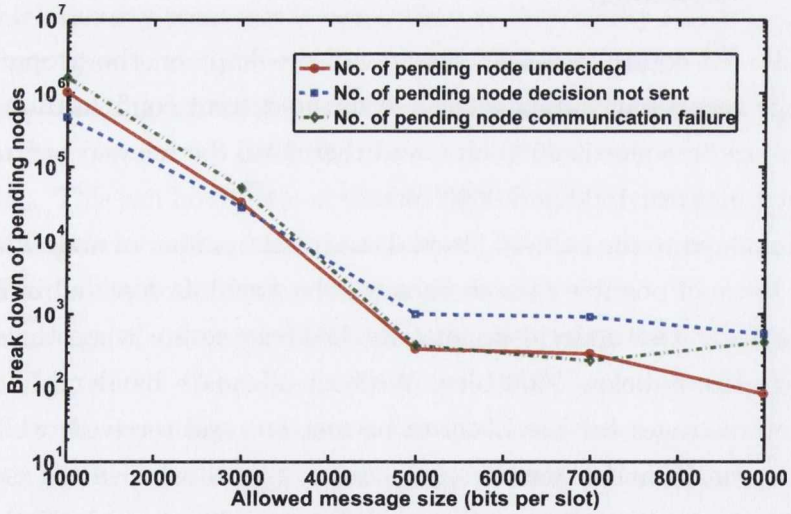

Figure 5.28: Breakdown of pending nodes vs. allowed message size

performance in RRP, it is vital to maintain a sufficient bandwidth for control messages.

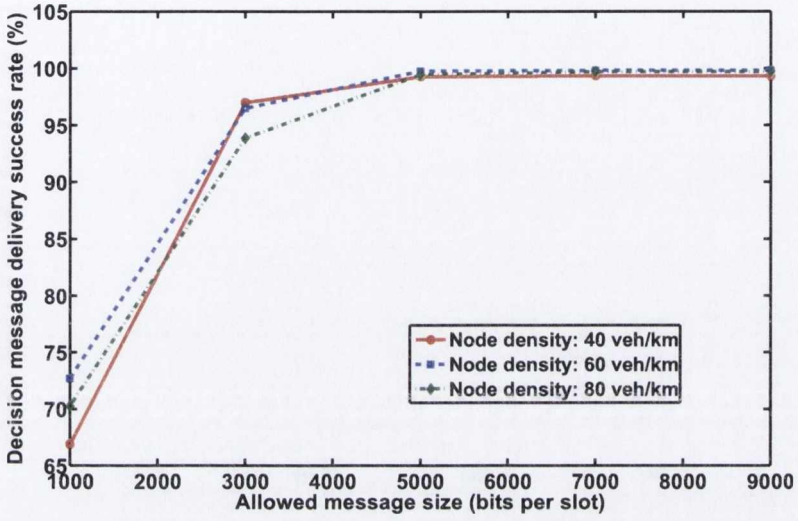

Figure 5.29: Decision message delivery success rate vs. allowed message size

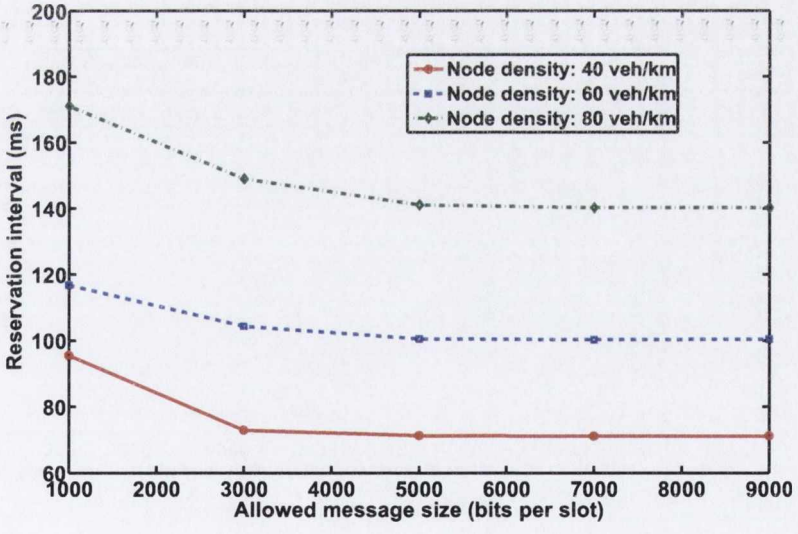

Figure 5.30: Reservation interval vs. allowed message size

The impact of allowed message size on the overall performance is illustrated in Figure 5.30, Figure 5.31, and Figure 5.32. If the allowed control message

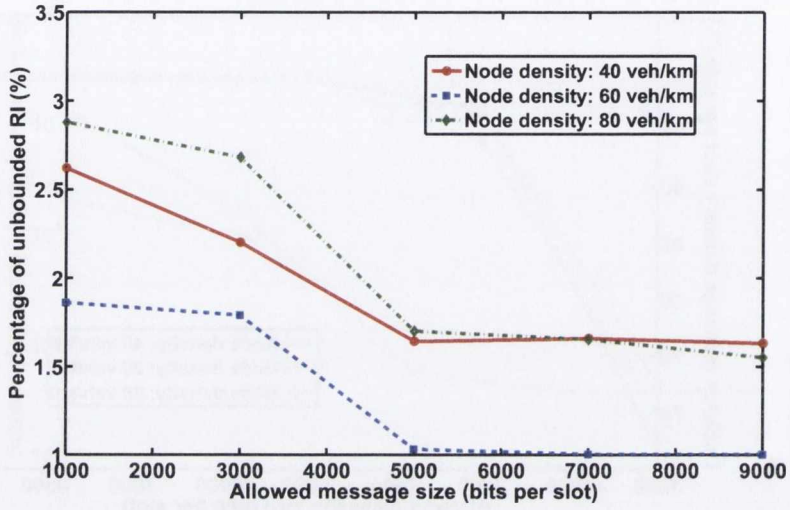

Figure 5.31: Beyond bound RI vs. allowed message size

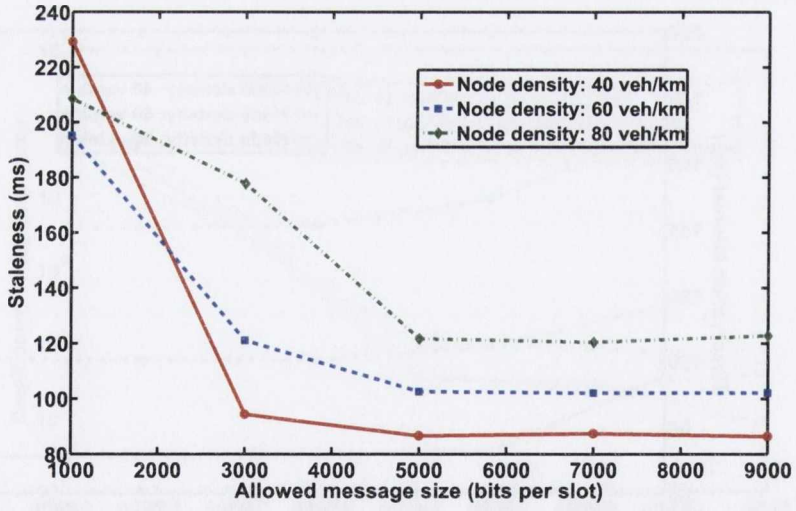

Figure 5.32: Mean staleness vs. allowed message size

size is below 5000 bits, the overall performance, including average reservation interval, percentage of unbounded reservation interval and average staleness begin

to dip dramatically. In terms of average staleness, it is worth noting that lower density scenarios are more tolerant to insufficient control message size as depicted in Figure 5.32, because less reservation information needs to be delivered. For example, in the 40 vehicle/km scenario, the average staleness remains low until the size limit drops below 3000 bits; while in high node density case, 5000 bits is a the tipping point for the staleness to drop.

5.5 RR-ALOHA Specific Evaluation

As a reservation-based MAC protocol, RR-ALOHA shares a lot of similarities with RRP. For example, time is divided into slots, and each slot needs to be contested among neighbors to ensure collision-free medium access. However, due to its reservation procedure, RR-ALOHA suffers two major drawbacks. First of all, the collisions occurred during slot reservation, i.e., access collisions, substantially reduce the probability of successfully reserving a slot, especially when the number of neighbors exceeds the size of a frame. Secondly, the dynamics of the network topology erodes the agreed slot reservations, which causes collisions during actual transmissions. In the following evaluation, these two drawbacks are the focus of the study, and the RR-ALOHA performance is measured in terms of generic metrics such as PDR, RI and staleness, in addition to a number of RR-ALOHA specific metrics, e.g., reservation success rate and reservation delay. In the following, the study of the access collision problem is presented in Section 5.5.1, and the study of dynamic topology on RR-ALOHA performance is presented in Section 5.5.2.

5.5.1 Effect of Frame Size on RR-ALOHA Performance

In this section, the impact of the access collision problem on RR-ALOHA performance is evaluated. The performance is characterized by reservation success rate, reservation interval, PDR, and staleness. By tuning the frame size in a given network, the severity of access collision and its impact on RR-ALOHA's overall performance are evaluated and analyzed.

There are three possible outcomes for a reservation attempt in RR-ALOHA:

success, failure due to omission (some neighbors fail to respond), and failure due to inconsistency (inconsistent reply from neighbors). In Figure 5.33, the respective percentages of these possibilities are plotted with respect to the frame size. In a specific node density scenario (60 vehicle/km), the reservation success rate is less than 50% when the frame size is below 120 slots, and the omission failure and inconsistency failure are above 20% and 30% respectively. Beyond 120 slot frame size, the reservation success rate increases and the failure rate drops significantly. To summarize, given a specific node density, the frame length needs to be sufficiently large to maintain a high reservation success rate in RR-ALOHA.

The consequence of a low reservation success rate is also investigated by comparing the number of actual 1-hop neighbors with the number stored in the local knowledge base. As depicted in Figure 5.34, when the frame size is small, the number of known 1-hop neighbor is significantly less than the actual numbers, which suggests that some 1-hop neighbor's existence are unknown. This is due to the fact that these missing nodes (which are referred to as "black hole" nodes) are unable to reserve a slot to broadcast their existence to their neighbors. To summarize, excessively small frame size in a given node density creates "black

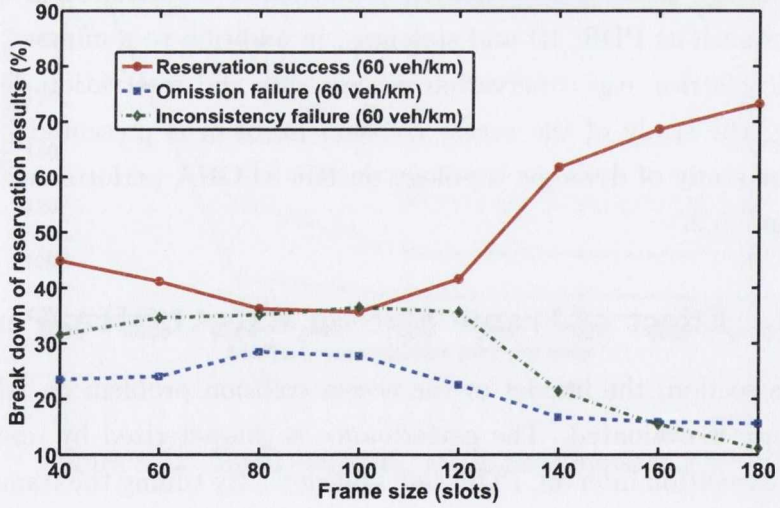

Figure 5.33: Reservation success rate vs. frame size

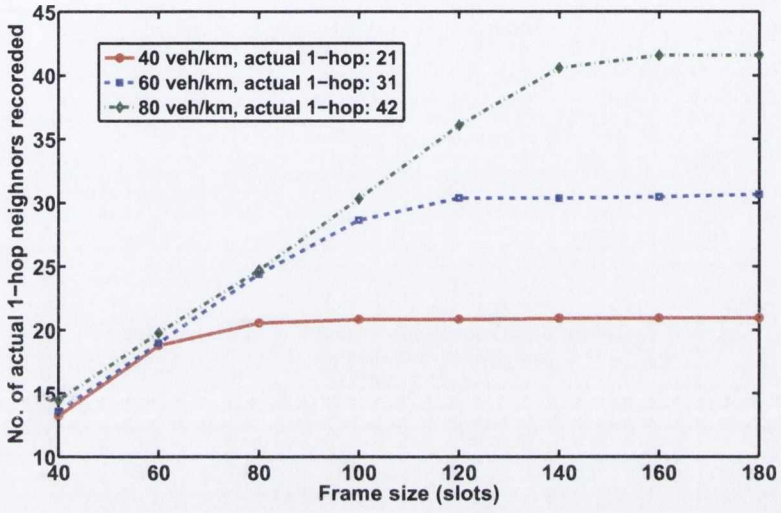

Figure 5.34: Number of 1-hop neighbors recorded in knowledge base vs. frame size

hole", nodes whose existence is unknown to their neighbors. The impact of such black hole nodes are further analyzed in subsequent sections.

In Figure 5.35, the PDR is measured with respect to the frame size given various node densities. In general, RR-ALOHA can deliver messages reliably with over 90% PDR. However, it is interesting to observe that a "V" shape of PDR in all node densities. The reason is that when the frame size is less than what is required for the specific node density, black hole nodes appear, which do not participate in the reservation contest, and thus the PDR becomes higher. When the frame size increases, the number of available slots also increases which improves PDR.

In Figure 5.36, the average RI is measured with respect to various frame length in three node densities. When the frame size exceeds a density-specific threshold, RI starts to grow linearly. This is due to the fact that the enlarged frame size proportionally increases the time interval between two reserved slots. However, when the frame size falls below the threshold, the RI remains stable as RI is now determined by the number of neighbor in one's neighborhood.

As depicted in Figure 5.37, when the frame size shrinks below a density-specific

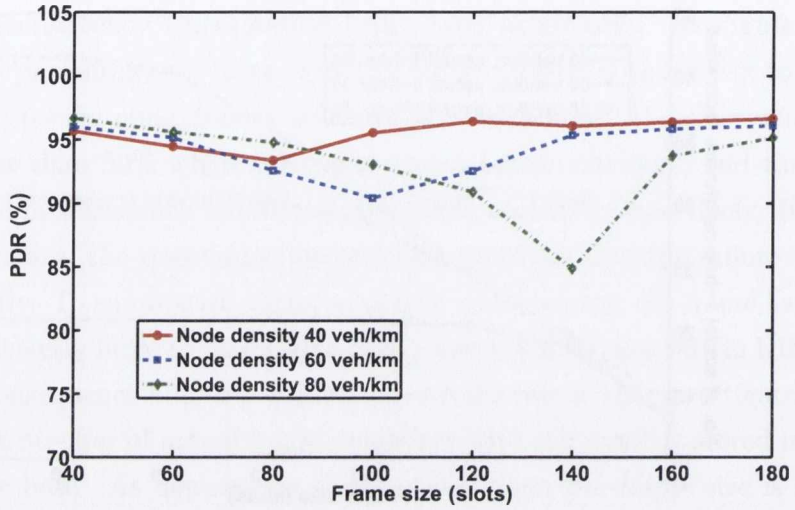

Figure 5.35: PDR vs. frame size

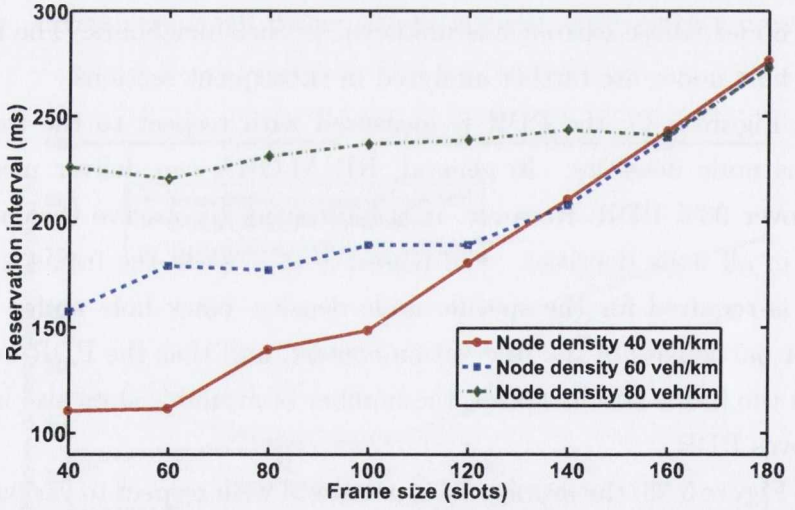

Figure 5.36: Reservation interval vs. frame size

threshold, the black hole nodes significantly increase the number of unbounded RI. These black hole nodes also have a significant impact on average staleness,

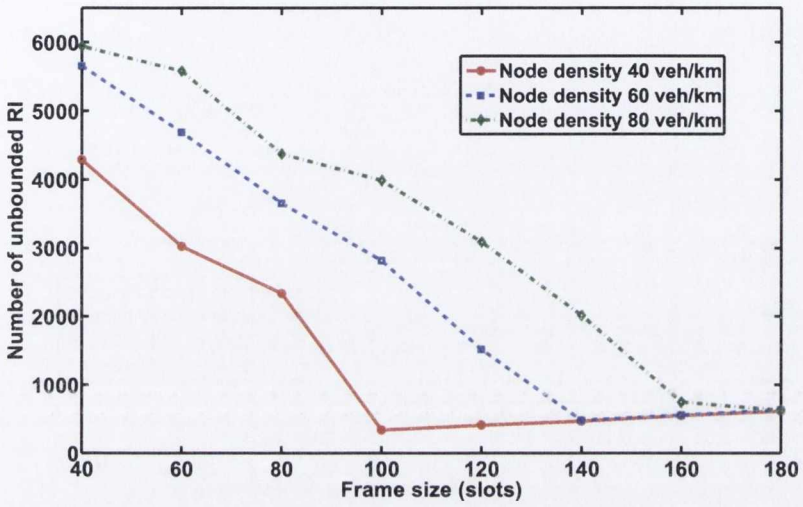

Figure 5.37: Beyond bound reservation interval vs. frame size

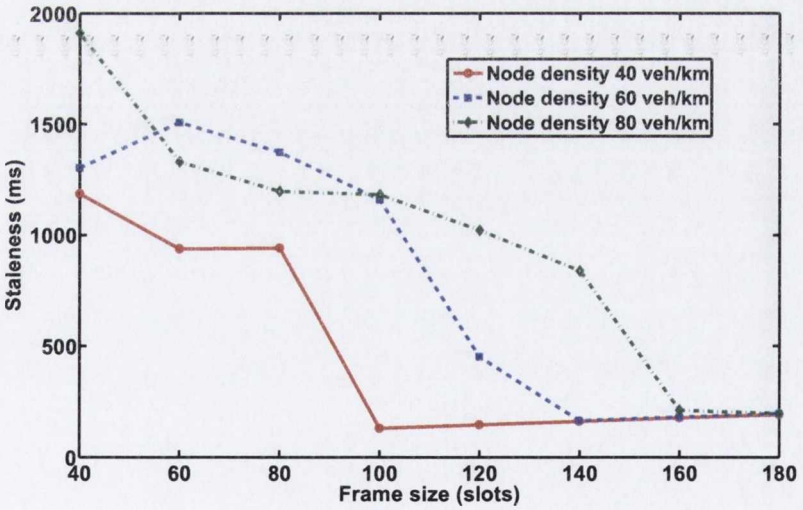

Figure 5.38: Mean staleness vs. frame size

as illustrated in Figure 5.38. The long-term absence of a node makes the measurement of staleness in all neighboring nodes grow rapidly, which significantly

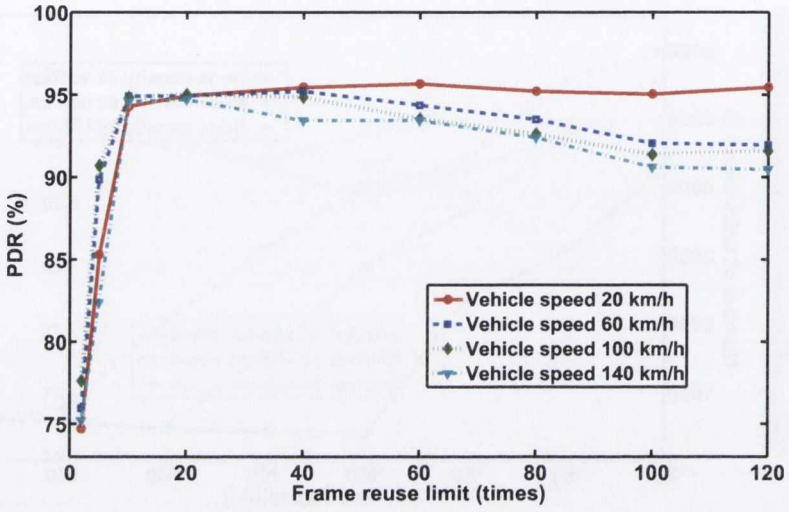

Figure 5.39: PDR vs. Frame reuse limit

increase the average staleness.

To summarize, in RR-ALOHA, it is very important to set an appropriate frame size according to the local node density. Otherwise, the reservation success rate begins to dip which might create a large number of black hole nodes that do not possess any slots. Such black hole nodes dramatically deteriorate the overall performance of RR-ALOHA including the number of unbounded RI and staleness. However, to set a fixed and universal frame length is practically infeasible in a network with various node densities, which is a huge obstacle for frame-based protocols to be applied in VANETs.

5.5.2 Effect of Vehicle Speed on RR-ALOHA Performance

In order to adapt to the rapidly changing network topology, RR-ALOHA can be modified to refresh its slot reservations. A frame reuse count can be introduced, which is the maximum number of frames in which a reserved slot can be reused. A smaller reuse count, i.e., frequent refresh, can update the slot allocation according to the network dynamics, but increases the cost and overhead during slot reservation. On the other hand, large reuse count, i.e., less frequent refresh,

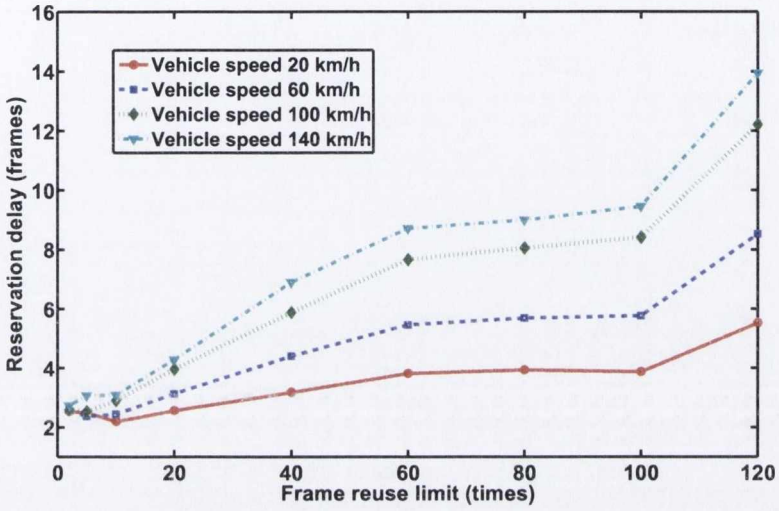

Figure 5.40: Reservation delay vs. Frame reuse limit

reduces the slot reservation cost, but the schedule may not be adaptive enough to the changing environment.

In Figure 5.39, the PDR is measured with respect to the frame reuse count, given various vehicle speeds. When the frame reuse count is less than 10, the PDR falls because most slots need to be contested before use, which causes a large number of access collisions. When the frame reuse limit increases beyond the tipping point, a lot more slots are used without reservation, therefore increases the PDR. However, the PDR begins to drop when the frame reuse limit continue to grow, as node mobility begins to erode the correctness of scheduled slots. For various vehicle speeds, it is also illustrated in Figure 5.39 that higher vehicle speed has a more negative impact on the PDR.

To successfully reserve a slot, a node needs to wait a number of frames to send out the trial message and wait another number of frames to be confirmed. The reservation delay and the trial message delay (in terms of frames) are depicted in Figure 5.40 and Figure 5.41 respectively. In general, both types of delays increase as frame reuse limit grows. This is due to the fact that, larger frame reuse limit makes nodes hold reserved slots for a longer time, which means less available slots

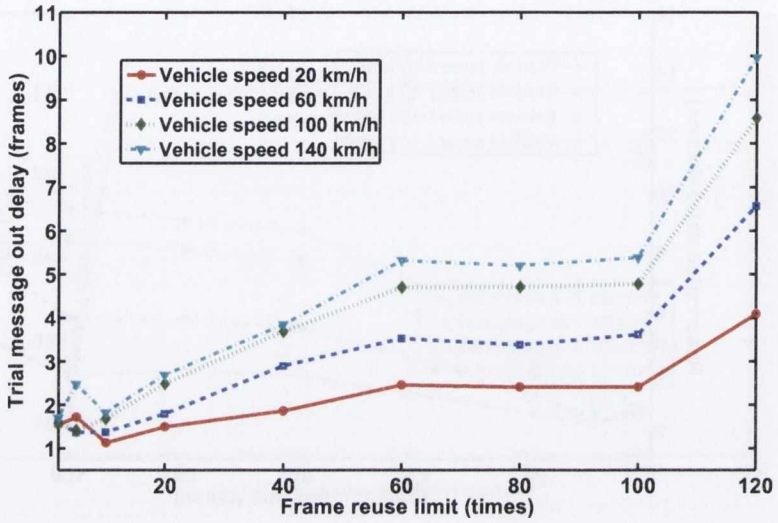

Figure 5.41: Reservation trial message out delay vs. Frame reuse limit

for other nodes to choose from, and subsequently send their trial messages.

As depicted in Figure 5.40, the reservation delay grows as the vehicle speed increases. It is because a node in a fast moving network knows more slot usage than in a slow network, which makes it more difficult to find a clear slot to send a trial message. To summarize, higher node mobility has a negative impact on the reservation delay in RR-ALOHA.

The average staleness of RR-ALOHA is measured with respect to the frame reuse limit in various vehicle speeds. As depicted in Figure 5.42, larger frame reuse counts increase average staleness, as reserving a slot becomes more difficult as discussed earlier. The vehicle speed also has a direct impact on average staleness as faster vehicle speed leads to a larger staleness, especially when the frame reuse limit is large. This is because higher vehicle speed decrease PDR, and increases the waiting time to reserve a slot, which eventually leads to a higher staleness.

Based on the evaluation study above, two major challenges in RR-ALOHA are identified which prevents it from being applied in VANETs directly. The first one is how to dynamically determine an appropriate frame reuse limit based on local vehicle speed. Frequent frame refresh may lead to access collisions and

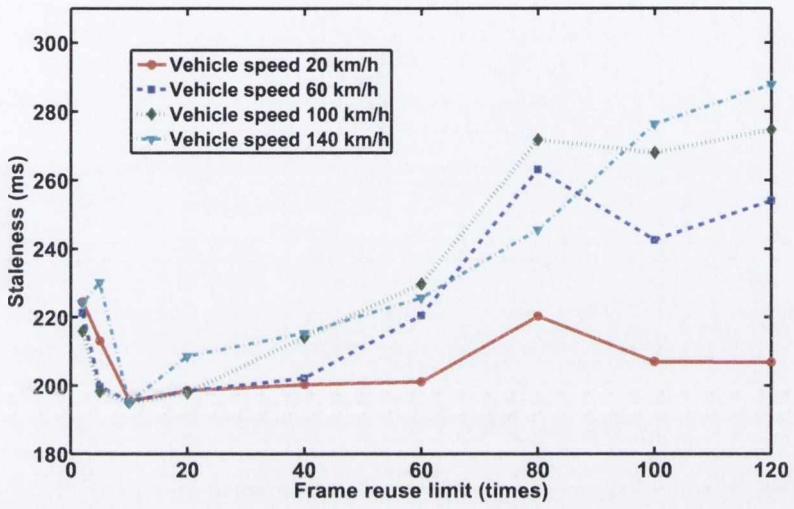

Figure 5.42: Mean staleness vs. Frame reuse limit

making reservation more difficult, while insufficient frame refresh corrupts the established schedule. In addition, it is challenging to maintain a consistent view on neighbors' frame reuse limit, which is subject to change.

The second challenge relates to the determination of an appropriate frame length in a local area. As shown in the diagrams above, insufficient frame length may create black hole nodes which deteriorates the overall performance significantly. On the other hand, excessively large frame length is inefficient and not practical for time sensitive medium access. Further more, nodes with heterogeneous frame length cannot co-exist in RR-ALOHA, which means that it is very difficult to apply it in a real-world environment, where frame length depends on local node density.

5.6 802.11p Specific Evaluation

In this section, the impact of vehicle speed, channel characteristics and contention window size on the performance of 802.11p are studied in terms of PDR, medium access delay and average staleness. The evaluation study is therefore divided into

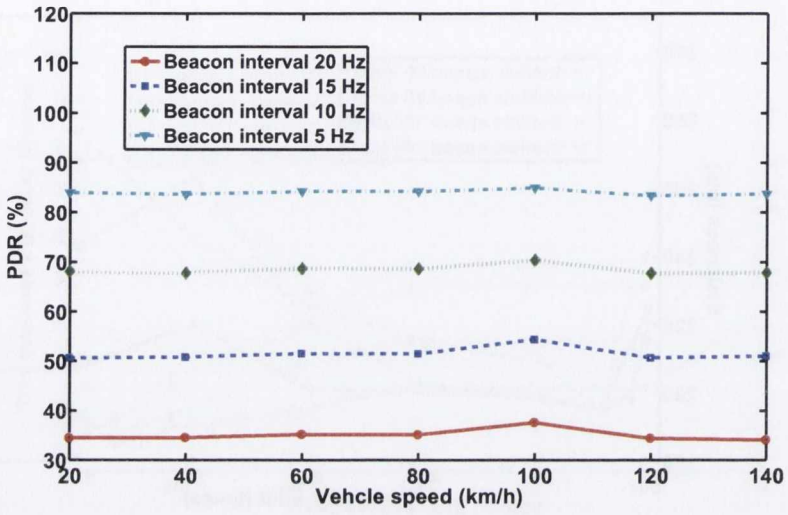

Figure 5.43: PDR vs. vehicle speed

three parts: the vehicle speed study in Section 5.6.1, the channel characteristics study in Section 5.6.2, and finally, the contention window study in Section 5.6.3.

5.6.1 Effect of Vehicle Speed on 802.11p Performance

In this study, the vehicle speed is varied from 20 to 140 km/h, and the PDR, medium access delay and staleness are measured with node density configured at 80 vehicle/km. As depicted in Figure 5.43, Figure 5.44, Figure 5.45. The general trend in these three diagrams is that none of these performance indexes are affected by the vehicle speed, which confirmed our previous expectation that contention-based protocols are largely immune to mobility.

In the study, the performance of 802.11p is also measured with respect to various beacon intervals, which creates a different offered load in each speed scenario. For PDR and medium access delay in general, increasing the offered load (more frequent beacon interval) leads to worse performance, as PDR begin to drop and medium access delay starts to grow. For staleness, the best performance, i.e., lowest staleness value is obtained at 10 Hz rather than 5 Hz, because an excessively low beacon rate increases the delay of receiving new beacons.

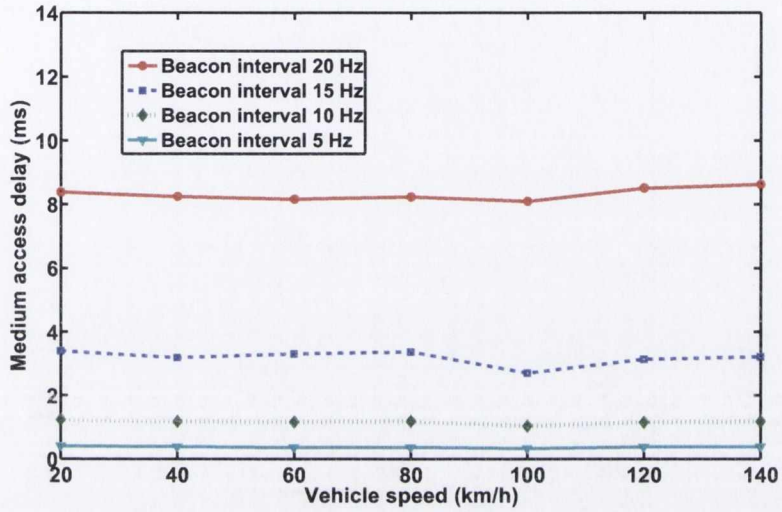

Figure 5.44: Medium access delay vs. vehicle speed

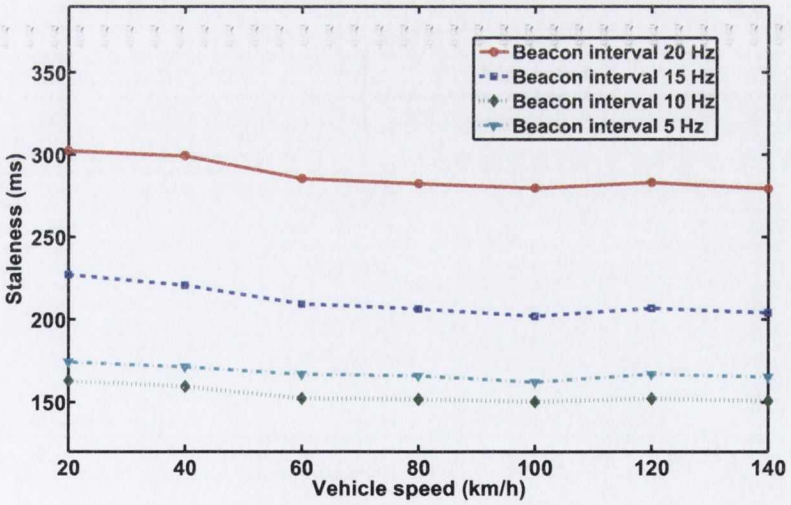

Figure 5.45: Mean staleness vs. vehicle speed

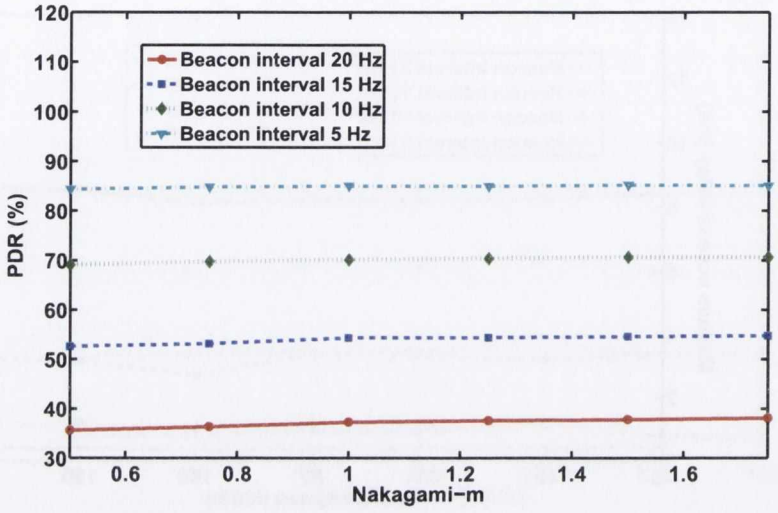

Figure 5.46: PDR vs. Nakagami-m

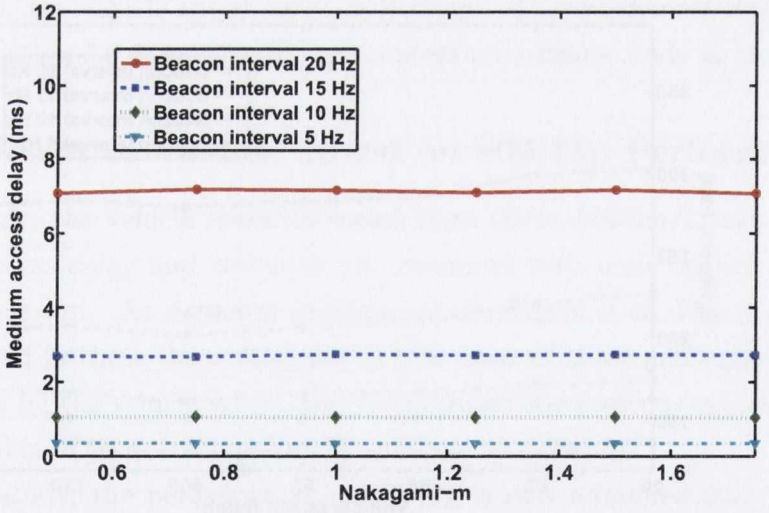

Figure 5.47: Medium access delay vs. Nakagami-m

5.6.2 Effect of Wireless Channel Characteristics on 802.11p Performance

In this section, the Nakagami-m parameter is tuned to create various channel characteristics, and the results are depicted in Figure 5.46, Figure 5.47, and Fig-

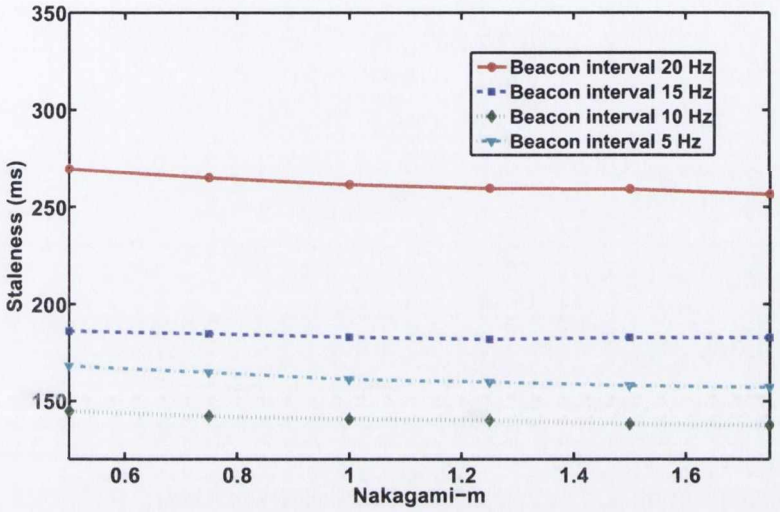

Figure 5.48: Mean staleness vs. Nakagami-m

ure 5.48 for PDR, medium access delay and mean staleness respectively. Similar to the speed evaluation, 802.11p does not seem to be affected by the channel fluctuation.

5.6.3 Effect of Contention Window Size on 802.11p Performance

The impact of the contention window size on 802.11p performance is studied in this section. As far as PDR is concerned, the expansion of contention window does not affect the PDR when the offered load is not heavy, as depicted in Figure 5.49. The PDR is improved only when the channel is heavily congested in the 20 Hz beacon rate scenario. In this particular case, the expansion of the contention window relieves the congestion in the channel and thus improves PDR.

The impact of contention window size on the medium access delay is depicted in Figure 5.50. A larger contention window increases the medium access delay, as the time wasted on waiting for the back-off slot grows. When the contention window expands to a large value, e.g., > 100 slots, the medium access delay grows

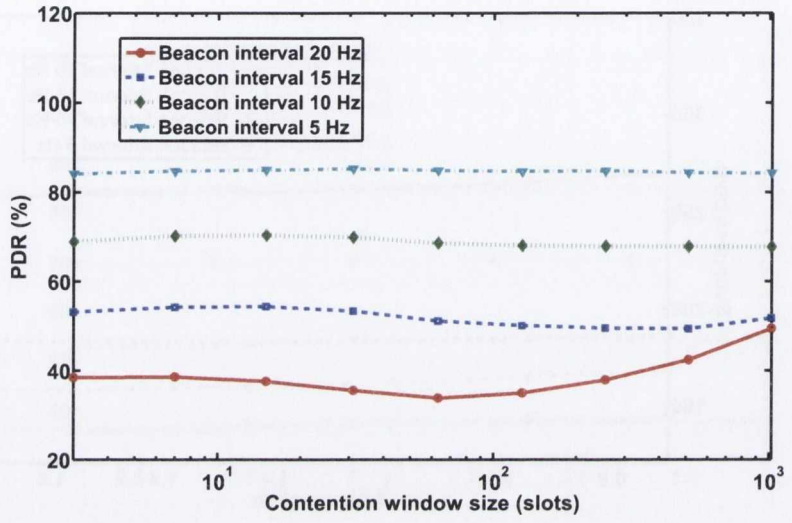

Figure 5.49: PDR vs. Contention window size

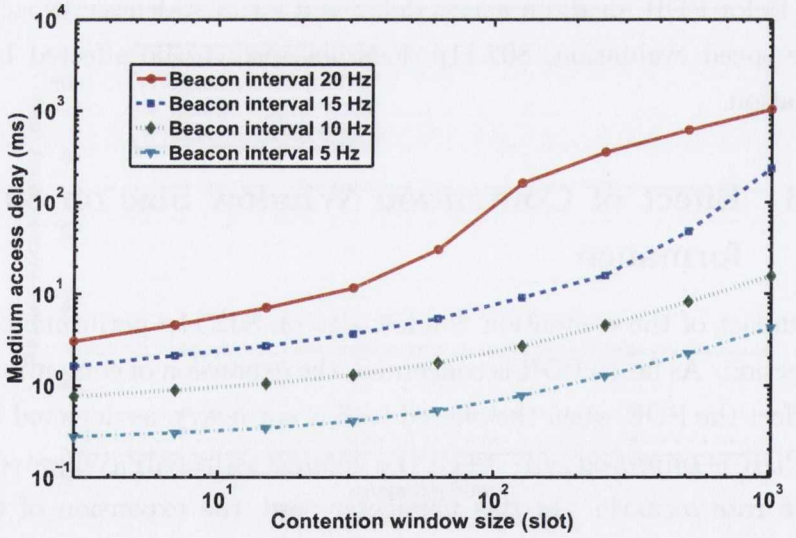

Figure 5.50: Medium access delay vs. Contention window size

significantly. For example in 20 Hz case, the medium access delay is larger than 100ms when the contention window exceeds 100 slots.

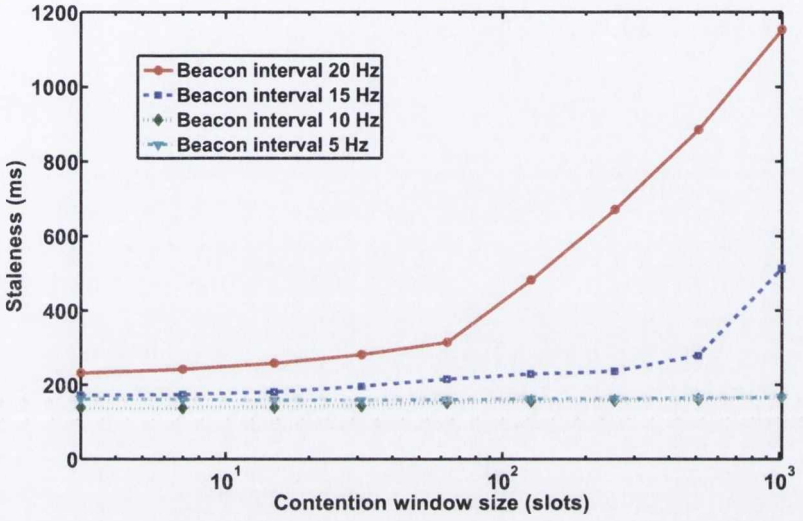

Figure 5.51: Mean staleness vs. Contention window size

The impact of contention window size on staleness also depends on the offered load as depicted in Figure 5.51. When the channel is not heavily saturated, e.g., in 5 Hz and 10 Hz beacon rate cases, the impact of the contention window on mean staleness is negligible. However, when the channel begin to saturate, the expansion of the contention window has a devastating effect on staleness, as the medium access delay begins to soar. This is because when 802.11p loses its advantage in maintaining low medium access delay, its disadvantage of having low PDR makes the communication quality, in terms of staleness, unacceptably poor.

5.7 Comparison and Analysis

In this section, the performance of 802.11p, RR-ALOHA and RRP is compared in order to study their strengths, weaknesses and their suitability in VANETs. The metrics used in the evaluation are the key performance indicators for a communication protocol, which include PDR, medium access delay, reservation interval, sender and receiver throughput, and the mean and deviation of staleness. In

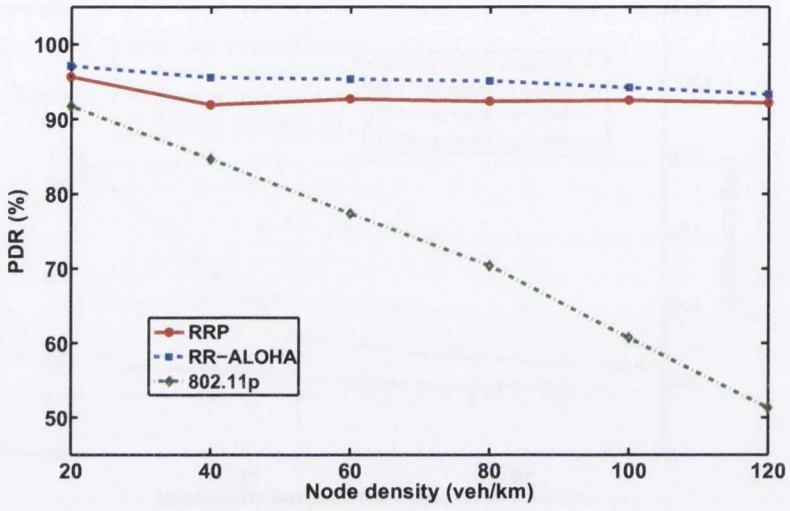

Figure 5.52: PDR vs. Node density

the comparison study, two parameters, i.e., node density and beacon interval are varied in order to simulate various load conditions, and those key performance indicators are measured and analyzed. In the following, each section evaluates and compares a specific performance metric among the three protocols. At the end of this section, the evaluation is concluded with the discussion of the suitability of each protocol in vehicular networks.

5.7.1 Packet Delivery Rate

In Figure 5.52, the PDR is measured in various node density scenarios from 20 vehicle/km to 120 vehicle/km. In general, the reservation-based protocol RRP and RR-ALOHA are able to maintain a high level of PDR as node density increases. On the contrary, contention-based 802.11p performs poorly when node density increases and the contention among neighbors become more severe. 802.11p's PDR drops to as low as 50 percent in the 120 vehicle/km scenario, meaning that almost half of the broadcast messages will be lost during a single transmission.

The PDR is also measured with various beacon intervals, which is the time interval between consecutive beacons generated by each vehicle. Effectively, re-

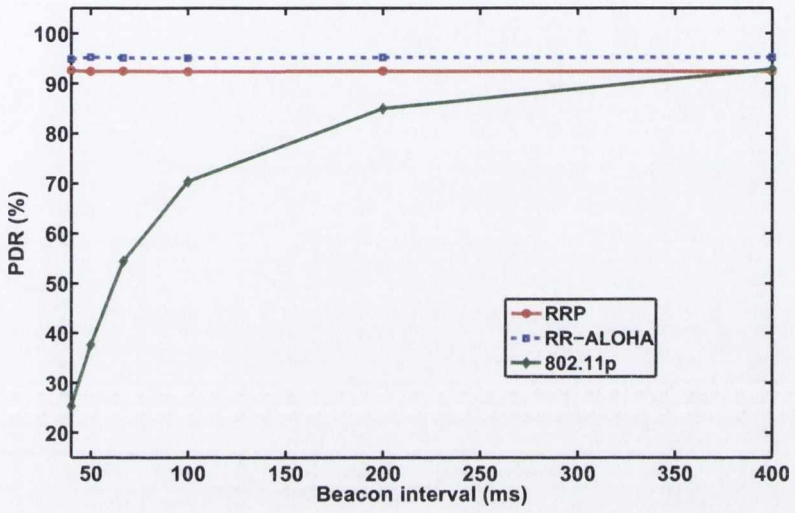

Figure 5.53: PDR vs. Beacon interval

ducing beacon intervals increases offered load on the channel, which is the same as increasing the node density. The results depicted in Figure 5.53 reveals similar trends in reservation-based protocols and contention-based protocols. As the beacon interval decreases and thus the load on the channel increases, contention intensifies in 802.11p networks and the PDR drops significantly. When the beacon interval reduces to less than 100 ms, the 802.11p channel becomes saturated and the PDR deteriorates to less than 50 percent. On the contrary, RRP and RR-ALOHA are not affected by the beacon interval changes and maintain a high level of PDR.

Based on the study of PDR vs. offered load, it is observed that the reservation-based protocols preserve a high level of reliable transmission regardless of the offered load, while the contention-based protocol's performance significantly deteriorates when the channel becomes congested. This is due to the fact that transmissions are explicitly reserved in RRP and RR-ALOHA, but contested and unprotected in 802.11p. The advantage of the reservation approach is not obvious when the contention is light, but becomes apparent in challenging load conditions.

In reservation-based protocols, when the demand for resource increases as

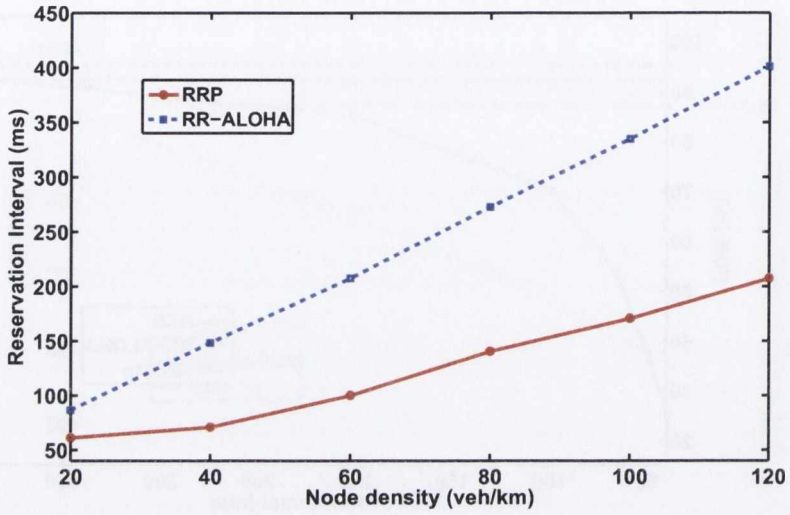

Figure 5.54: Reservation interval vs. Node density

the contention intensifies, the reservation requests are more difficult to be satisfied, which effectively restrains the actual transmission rate. On the contrary, contention-based protocols do not consider the actual available resources and transmit without rate control, which results in increased collisions and reduced transmission reliability.

5.7.2 Reservation Interval and Medium Access Delay

The RI measures the duration in between two consecutive reserved slots, and it only applies to reservation-based protocols, i.e., RRP and RR-ALOHA. RI is another indicator in addition to medium access delay showing how often the medium can be accessed. In Figure 5.54, RI is measured with respect to node density, with a fixed beacon interval at 100 ms. As the node density increases, the number of contenders grows, which makes it harder to reserve a slot in both RRP and RR-ALOHA. Therefore the distance between successfully reserved slots, i.e., RI, increases along with node density.

In Figure 5.55, RI is irrelevant to various beacon intervals for both RRP and RR-ALOHA, while the node density is fixed at 80 vehicle/km. As a matter of

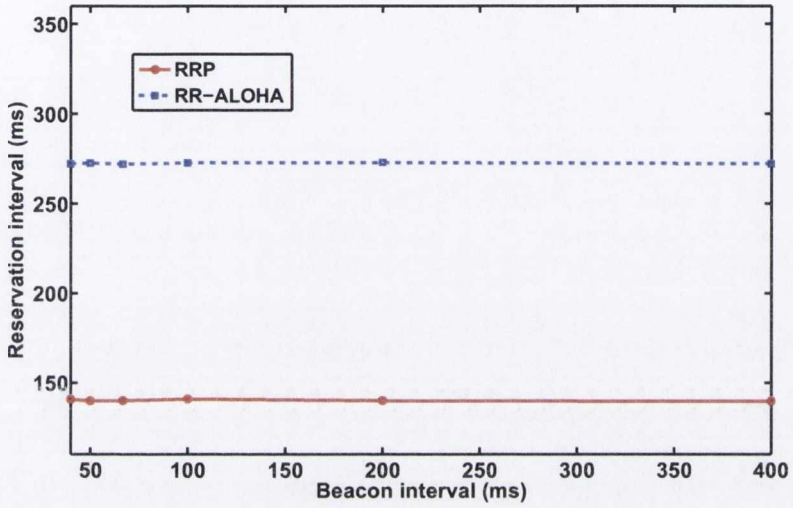

Figure 5.55: Reservation interval vs. Beacon interval

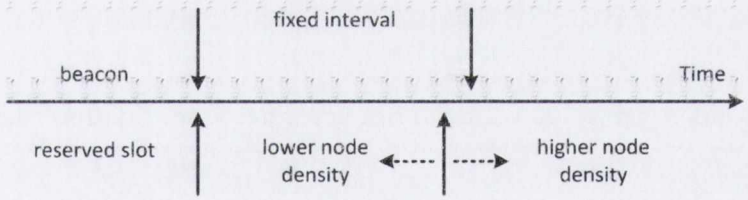

Figure 5.56: Representation of node density, RI and beacon interval in various node density scenarios

fact, RI reflects the amount of bandwidth that is available in a specific network environment, which is only environment-specific and is independent of the offered load.

In Figure 5.54 and Figure 5.55, at the same configuration conditions, the RI of RRP is always smaller than RR-ALOHA's, which indicates that RRP has higher efficiency in terms of allocating resources than RR-ALOHA.

The medium access delay is measured as the time interval between a packet arriving at RRP until its eventual broadcast to the medium. Due to the periodic nature of safety messages in VANETs, a beacon that is buffered for a long time will be overwritten by a new beacon and the medium access delay of the erased

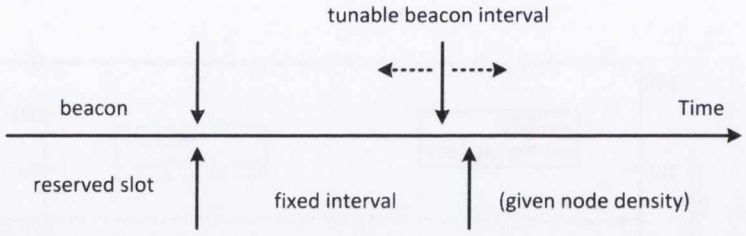

Figure 5.57: Representation of node density, RI and beacon interval in various beacon interval scenarios

beacon is not counted in the evaluation since the message is never actually sent out.

The medium access delay of RRP, RR-ALOHA and 802.11p are evaluated and compared with respect to node density and beacon interval. In Figure 5.58, the beacon interval is fixed at 100 ms while node density varies, and in Figure 5.59, the node density is fixed at 80 vehicle/km while beacon interval varies. Prior to the description of the evaluation results, some observations regarding medium access delay, beacon interval and reservation intervals for reservation-based protocols are presented as follows:

The relationship between the beacon interval and RI has implications on the MAC delay, which are illustrated in Figure 5.56 and Figure 5.57. Two scenarios are of particular interest: 1) RI is considerably smaller than beacon interval, and 2) beacon interval is considerably smaller than RI.

In the first case, as the parameter arrow moves to the left in Figure 5.56, or the parameter arrow moves to the right in Figure 5.57, for each beacon message, it is very likely that the message will be sent within a short period of time (considering the length of the beacon interval). Consequently, the medium access delay of the beacon message largely depends on the length of the RI, and is approximately half of the RI on average.

In the second case, as the parameter arrow moves to the right in Figure 5.56, or the parameter arrow moves to the left in Figure 5.57, the number of beacon messages overwhelms the available number of slots, which means that a large number of beacons are overwritten by subsequent beacons and are not included in the measurement of medium access delay. For each reserved slot, the beacon message being transmitted has a very short buffering time (considering the length

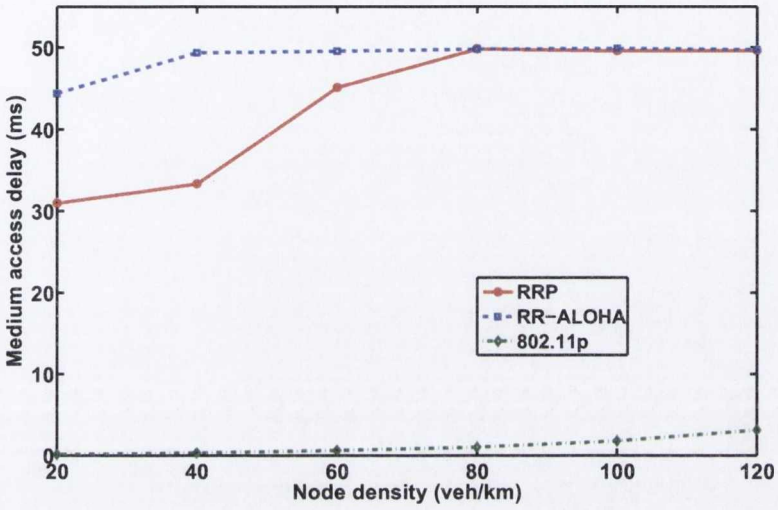

Figure 5.58: Medium access delay vs. Node density

of the RI) due to the rapid beacon overwrites. Consequently, the medium access delay of the beacon messages largely depends on the length of the beacon interval, which is roughly about half on average.

The performance of RRP, RR-ALOHA and 802.11p are compared in terms of medium access delay in Figure 5.58, and Figure 5.59. The medium access delay of contention-based 802.11p is significantly smaller than reservation-based RRP and RR-ALOHA due to their differences in medium access mechanism. 802.11p only grows marginally in extreme node densities while RRP and RR-ALOHA maintains a high level of medium access delay.

It is interesting to observe the value of medium access delay by cross referencing the relationship between RI and beacon interval, as is discussed previously. For instance, for RRP in Figure 5.58, the RI has approximately the same length compared to beacon interval (fixed at 100 ms) when the node density is about 60 vehicle/km according to Figure 5.54. If the node density decreases, i.e., moving the arrow in Figure 5.56 leftwards, the medium access delay is dominated by the RI, which shows a near linear decrease. On the other hand, if the node density increases, i.e., moving the arrow in Figure 5.56 rightwards, the medium access

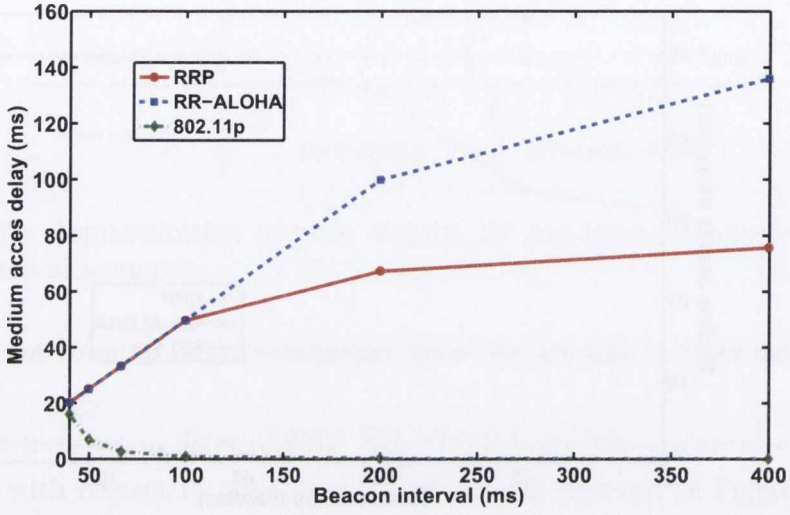

Figure 5.59: Medium access delay vs. Beacon interval

delay is dominated by the beacon interval, which is fixed at 100 ms. This explains why the medium access delay is capped at half of the beacon interval at 50 ms.

In Figure 5.59, the RI and beacon interval are approximately the same when the beacon interval is about 140 ms, according to Figure 5.55. For smaller beacon intervals, beacon interval dominates medium access delay which makes medium access shrink linearly. For larger beacon intervals, RI dominates medium access delay, which caps it at half of RI, i.e., approximately 70 ms for RRP and 140 ms for RR-ALOHA.

5.7.3 Sender Throughput and Receiver Throughput

The sender throughput are measured with respect to node density for RRP, RR-ALOHA and 802.11p in Figure 5.60. The beacon interval is fixed at 100 ms. As node density increases, RI of RRP and RR-ALOHA increases as we discussed in Section 5.7.2, which effectively reduces the sender throughput. For 802.11p however, packets are sent to the medium regardless of the environment which is reflected by a constant sender throughput with constant beacon interval.

The receiver throughput is the throughput experienced by a single receiver,

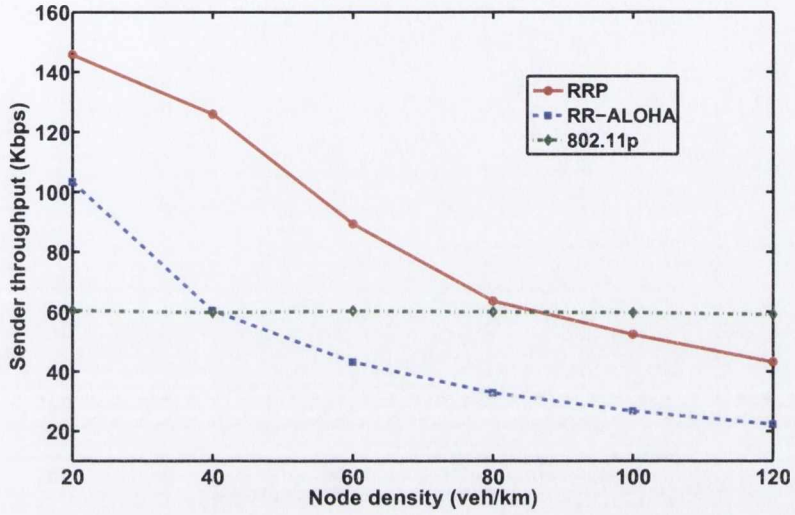

Figure 5.60: Sender throughput vs. Node density

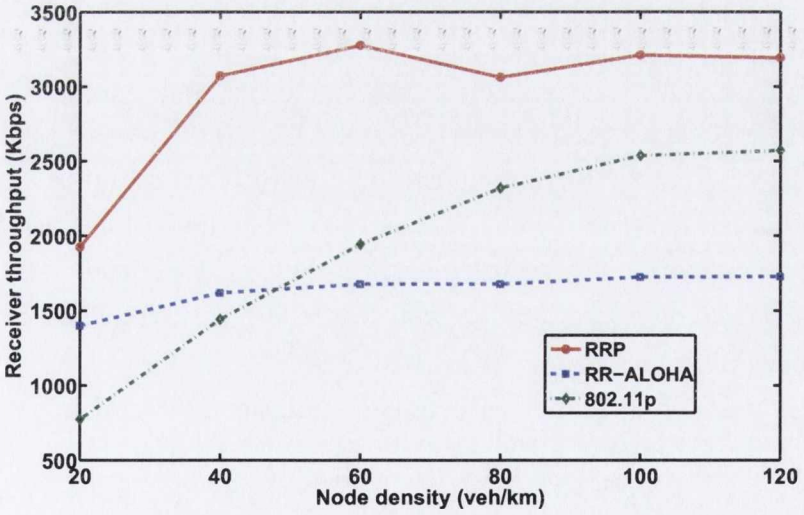

Figure 5.61: Receiver throughput vs. Node density

which is measured in Figure 5.61 with the same configuration as the sender throughput. For all protocols, as node density increases, the respective receiver

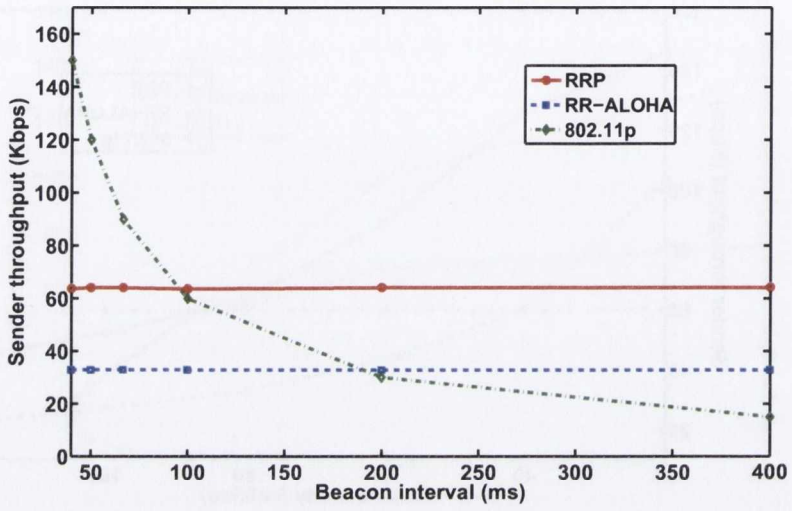

Figure 5.62: Sender throughput vs. Beacon interval

throughput also grows, until it reaches a saturation point with RRP achieving the highest receiver throughput. It is interesting to observe that the saturation is achieved under the condition that sender throughput is decreasing in RRP and RR-ALOHA, but remains constant in 802.11p. As shown in Figure 5.61, when more nodes exist in the communication range with constant sender throughput, 802.11p's receiver throughput does not grow continuously, which indicates that more packets are lost due to collision. On the other hand, as the node density increases, reservation-based protocols reduce their transmission rate to avoid channel saturation.

The sender throughput with respect to beacon interval is presented in Figure 5.62, with node density set at 80 vehicle/km. Due to the fact that node density is constant, RI for RRP and RR-ALOHA is constant because it is an environment-specific metric, as discussed in Section 5.7.2. Therefore the sender throughput is also constant for RRP and RR-ALOHA. For 802.11p, as beacon interval increases, the actual number of packets that arrived at the MAC layer decreases, which brings down the sender throughput.

The receiver throughput with respect to beacon interval is presented in Fig-

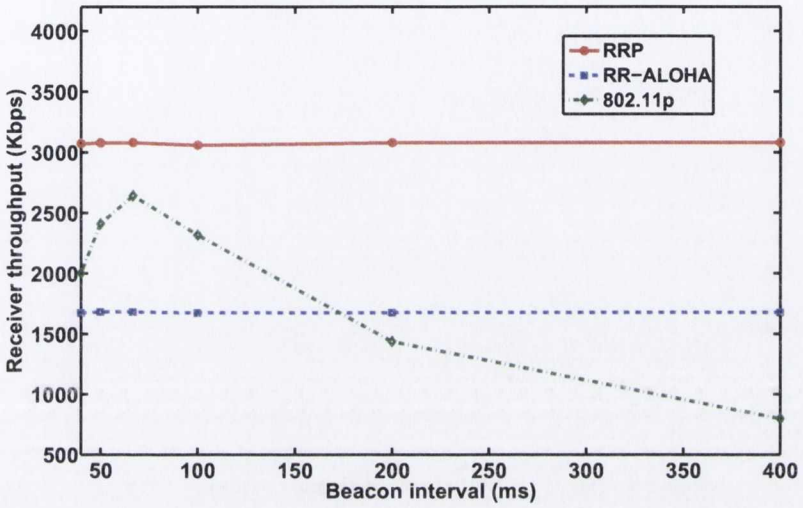

Figure 5.63: Receiver throughput vs. Beacon interval

ure 5.63, with the same configuration as sender throughput. For RRP and RR-ALOHA, the sender throughput as well as the node density are all constant, so that the receiver throughput is constant with various beacon intervals. For 802.11p, there exists a tipping point of the receiver throughput at approximately 70 ms of beacon interval. Prior to the tipping point, the increased number of packets begin to saturate the channel which reduces the receiver throughput, while beyond the tipping point, the decreased number of packets reduces the offered load which also results in a reduced receiver throughput.

5.7.4 Mean and Deviation of Staleness

The staleness observed in vehicles combines the experienced medium access delay and the packet delivery rate, and is the indicator of communication quality for safety applications in the study. A desirable communication protocol provides reliable transmission as well as timely medium access, which translates to a low staleness on average, and a small deviation as well.

In the following, the average and the deviation of staleness are measured and compared for RRP, RR-ALOHA and 802.11p. In Figure 5.64, the average stale-

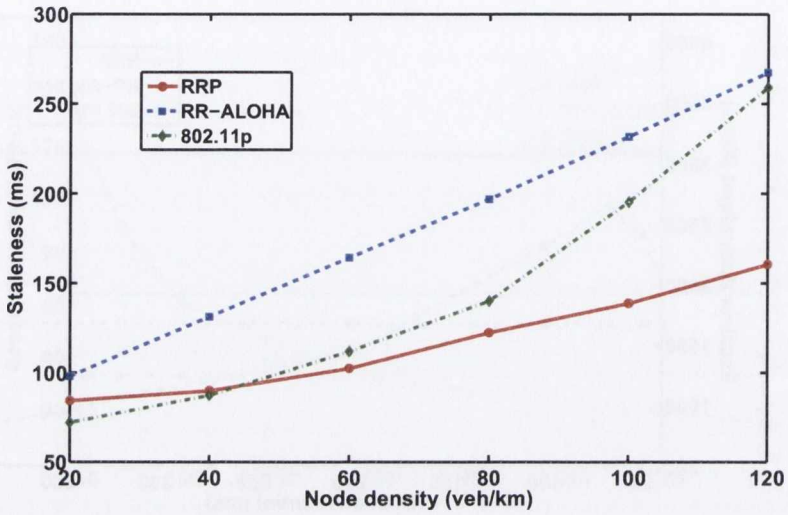

Figure 5.64: Mean staleness vs. Node density

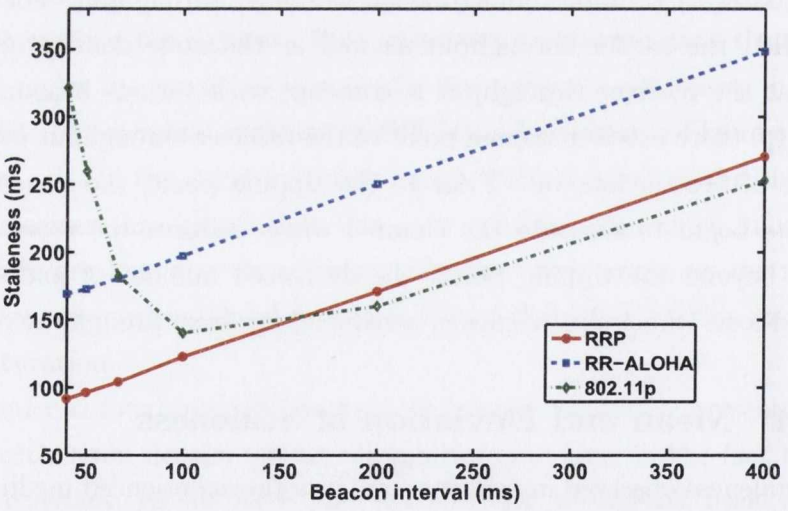

Figure 5.65: Mean staleness vs. Beacon interval

ness is evaluated with respect to various node densities, and the beacon interval is constant at 100ms. In general, for all protocols, the average staleness grows

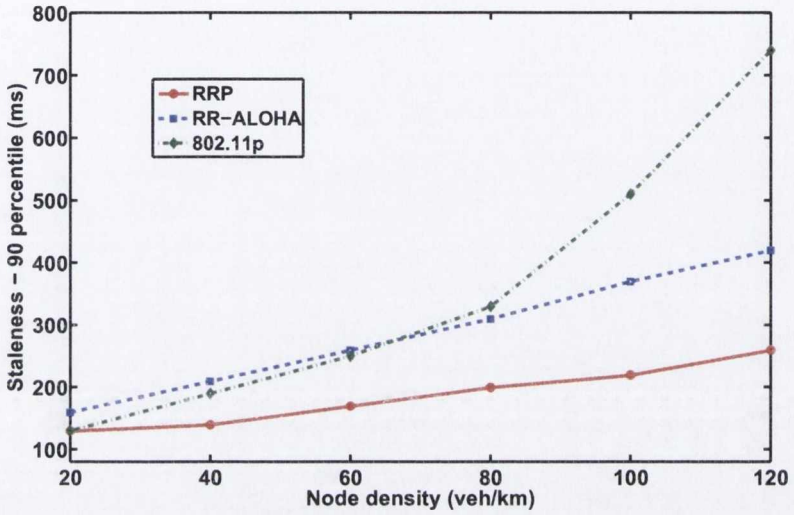

Figure 5.66: Staleness 90-percentile vs. Node density

as node density increases, since the contention for resources becomes more intensive. For RRP and RR-ALOHA, it grows almost linearly with node density, while 802.11p rises much faster in high-density scenarios, as it suffers from severe transmission collision. Except in the light density case (20 vehicle/km), RRP achieves the best performance in terms of the average staleness in all node densities.

In Figure 5.65, the average staleness is measured with respect to the beacon interval, and the node density is at fixed at 80 vehicle/km. 802.11p performs poorly when the beacon interval is small as more messages are congesting the channel. After the tipping point at 100 ms, for all protocols, it grows linearly as the beacon interval increases. For reservation-based protocols, given a specific node density, smaller staleness can be achieved by reducing the beacon interval. In fact, if the beacon message is not periodically generated, but polled whenever a slot is available, even smaller staleness can be achieve as the time wasted on waiting for a slot is eliminated. However, this improvement only applies to reservation-based protocol as the exact time to transmit is known a priori.

The deviation of staleness is studied in terms of the 90 and 99 percentile of staleness, meaning that 90 or 99 percent of the staleness values fall within such

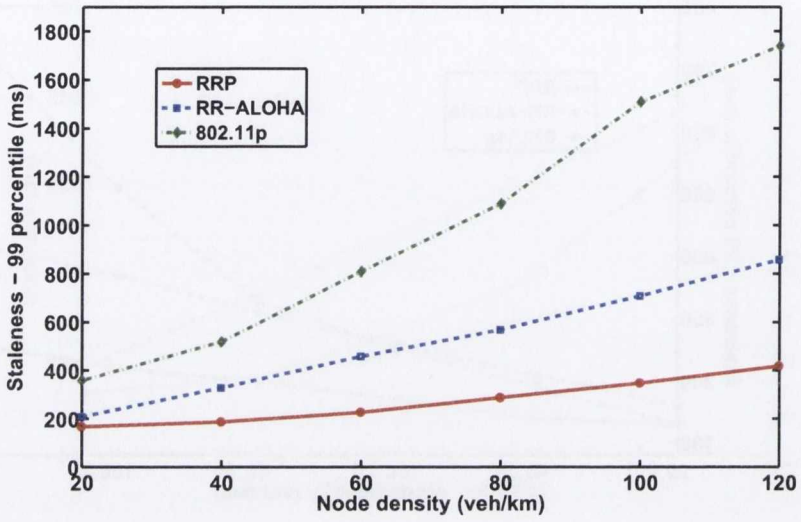

Figure 5.67: Staleness 99-percentile vs. Node density

a bound. In Figure 5.66 and Figure 5.67, the 90-percentile and 99-percentile of staleness are plotted against node density for RRP, RR-ALOHA, and 802.11p, with a fixed beacon interval of 100 ms. For the 90 percentile, the trend is similar as the average staleness in Figure 5.64 such that the measurement for all protocols grow as node density increases. However, the difference is that the deviation of staleness of the 802.11p soars in high-density scenarios. In the 99 percentile scenario, this phenomena becomes more obvious as the value of 802.11p reaches over 1000 ms in 80 vehicle/km node density and beyond. For a safety application, if 802.11p is used as the communication protocol, then in approximately 1% of the time, a vehicle's neighbor will be out of reach for more than 1000 ms, which is unacceptable for most safety applications.

On the other hand, the advantage of deterministic and time-bounded access scheme is clearly shown in staleness deviation figures. The 90-percentile and 99-percentile for RRP is below 300 and 400 ms in all node densities, which is considerably improved compared with 802.11p. The results have demonstrated the capability and suitability of RRP as a medium access protocol in a VANET.

Table 5.1: Comparison of 802.11p, RR-ALOHA and RRP

	802.11p	**RR-ALOHA**	**RRP**
Mobility support	High	Low	High
Channel congestion handling	Low	Medium	High
Reliability	Low	High	High
Predictability	Low	Medium	High
Timeliness	Low	Low	High
Operational cost	Low	Medium	High

5.8 Conclusion

In the following, the properties of 802.11p, RR-ALOHA and RRP are summarized and compared in terms of their suitability in the envisaged vehicular networks.

On the bright side, the 802.11p protocol is simple to implement and the communication overhead is minimal compared to RRP and RR-ALOHA, as nodes do not need to maintain their neighbors' status. In addition, 802.11p is generally not affected by node mobility, which is an important advantage in VANETs. The down side of 802.11p is two-fold. Firstly, the protocol performs poorly when offered load increases, due to its contention-based nature. Major performance indexes begin to deteriorate fast when the level of contention intensifies. Secondly, due to its probabilistic nature, the protocol cannot provide sufficient level of communication QoS in terms of reliability, predictability and timeliness which are critical for safety applications.

The RR-ALOHA protocol has moderate implementation complexity and communication overhead compared to 802.11p and RRP. In the protocol, the medium is accessed in a deterministic and predictable manner, which maintains high reliability in stressed load conditions. However, a major issue with RR-ALOHA is its inability to adapt in various node densities because of the fixed frame length. In addition, the performance of the protocol is susceptible to node mobility as schedules need to be constantly refreshed.

The RRP protocol is very complex compared to 802.11p and RR-ALOHA, and is a heavy-weight protocol in terms of communication overhead, since geographic and reservation information need to be disseminated in the neighborhood. However, the merit of this protocol is the adaptability to mobility, and the ca-

pability to gracefully degrade performance when the offered load increases. In addition, RRP provides a high level of communication QoS, i.e., high reliability, deterministic and time-bounded medium access, which are crucial for safety applications.

The comparison of 802.11p, RR-ALOHA and RRP are summarized in Table 5.1.

Chapter 6

Conclusions and Future Work

This thesis presents a new medium access control protocol (RRP) to address the problem of real-time medium access control in vehicular ad hoc networks. By leveraging the concept of pre-scheduling, the RRP protocol achieves reliable and time-bounded medium access, which are crucial for the envisaged safety-critical applications in vehicular networks.

This chapter summarizes the most significant contributions presented in this thesis, and discuss their relevance to the state of the art. The chapter concludes with a discussion on possible improvements and long-term evolution of the protocol that remain open for the future.

6.1 Contribution

The motivation of the work presented in this thesis stems from the observation of a gap between the communication requirements of safety-related applications and the quality of service provided by the state-of-the-art communication protocols. As discussed in Chapter 2, neither contention-based nor reservation-based MAC protocols proposed in the literature support real-time communication in vehicular environments, due to their intrinsic flaws and the challenging nature of the vehicular networks such as high mobility and large scale.

In order to achieve real-time communications in VANETs, reliable and time-bounded medium access are the two fundamental issues that cannot be circum-

vented. In this thesis, the intuition of solving the problem in an extremely dynamic vehicular environment is to allocate resources in a proactive way, i.e., via pre-scheduling. Based on this idea, the proposed RRP protocol predicts the dynamics of the network in the future, and allocate time slots among prospective neighbors. The pre-scheduling-style slot allocation algorithm in RRP guarantees exclusive slot access, which significantly reduces the packet collision probability, and bounds the maximum waiting time that a node experiences in order to obtain a slot. Consequently, in the RRP protocol, a node is able to access the medium with guaranteed delays and the transmissions are free from collisions with high probabilities.

It is worth mentioning that, in order to achieve theses properties above, assumptions are made and certain constraints need to be satisfied. For example, the medium access delay guarantee is achieved only if the communication channel provides enough bandwidth for potential communicating nodes in a local area. In addition, the reliability of a transmission is subject to the unpredictable fluctuation of the wireless signal during its propagation. A conservative estimation of the wireless channel increases the reliability of the communication protocol, but reduces its efficiency. As as result, a balance is always needed between guaranteed communication properties and the practicality of a protocol in the real world.

Another contribution of the thesis is proposing a new application-level metric - staleness, which characterizes the perceived communication QoS by an application. Via a simulation-based study, the proposed RRP protocol is evaluated in terms of staleness, packet delivery rate, medium access delay and other metrics, and compared with 802.11p and RR-ALOHA protocol. The results have demonstrated the feasibility of achieving reliable communication and time-bounded medium access in a vehicular ad hoc network.

6.2 Future Work

As is always the case in research, there are a number of issues that remain open for possible future work, both in terms of protocol improvement and long-term evolution.

In terms of protocol improvements, the RRP protocol can be strengthened in

the following aspects:

First of all, the length of the advance period for mobility prediction and the slot allocation is fixed in RRP. A larger value of prediction length may provide more preparation time for the slot allocation algorithm, which may reduce the number of undecided slots and improve the protocol's performance. However, longer prediction length also increases the possibility of mobility prediction error due to, e.g., unexpected driver behaviors. Consequently, a trade-off need to be made between performance and usability of the protocol in terms of the prediction length. In future version of RRP, the value of the parameter of prediction length need to be optimized considering the current vehicle speed as well as the local environment.

In addition, in the current design of the neighbor identification algorithm, once a node is determined as a neighbor when the neighbor identification algorithm is invoked, the results cannot be changed even when some new information has been received. A dynamic approach regarding the neighbor identification process may be beneficial, as erroneous neighbor identification results due to prediction error or incomplete information can be corrected once new knowledge becomes available.

The RRP protocol has the ability to support heterogeneous nodes with different transmission powers. However, in the current design, all transmission powers are assumed to be equal. A scenario with nodes tuning their transmission powers at different levels may create a much more complex environment, which is interesting to investigate in terms of the performance of RRP and other benchmark protocols.

In terms of long-term evolution of the RRP protocol, the following aspects may be of interest:

In theory, RRP is able to support slot reservation with various priorities by changing the priority numbers used in slot contentions. Consequently, a node may be allocated with slots belonging to different priority categories. Messages that belong to different priority categories are delivered in their respective prioritized slots. Supporting prioritized message delivery improves the flexibility of the RRP in supporting applications with various communication requirements.

In addition, the slot allocation algorithm in current RRP protocol treats nodes

in a binary manner, i.e., neighbor, or non-neighbor. However, if the geographic property of a neighbor can be exploited, the spacial reuse and the overall efficiency of the protocol can be improved. For example, the slot allocation algorithm may avoid simultaneous transmissions between a neighbor that is close to the current node, but may schedule a shared slot with a neighbor that is far away from the current node which is therefore less likely to cause interferences. To improve the performance of slot allocation algorithm by analyzing the geographic position of neighbors and utilizing results from graph theory would be of interest.

References

1609.1 (2006). Trial-Use Standard for Wireless Access in Vehicular Environments (WAVE) - Resource Manager. *IEEE Std 1609.1-2006*, 1–71. 25

1609.2 (2006). IEEE Trial-Use Standard for Wireless Access in Vehicular Environments - Security Services for Applications and Management Messages. *IEEE Std 1609.2-2006*. 26

1609.3 (2010). IEEE Standard for Wireless Access in Vehicular Environments (WAVE) - Networking Services. *IEEE Std 1609.3-2010 (Revision of IEEE Std 1609.3-2007)*, 1–144. 26

1609.4 (2011). IEEE Standard for Wireless Access in Vehicular Environments (WAVE)–Multi-channel Operation. *IEEE Std 1609.4-2010 (Revision of IEEE Std 1609.4-2006)*, 1–89. 26

802.11 (1997). IEEE Standard for Information Technology-Telecommunications and Information Exchange Between Systems-Local and Metropolitan Area Networks-Specific Requirements-Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. *IEEE Std 802.11-1997*. 8, 17, 21

802.11E (2005). IEEE Standard for Information Technology - Telecommunications and Information Exchange Between Systems - Local and Metropolitan Area Networks - Specific Requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements. *IEEE Std 802.11e-2005 (Amendment to IEEE Std 802.11, 1999 Edition (Reaff 2003)*. 23

802.11p (2010). IEEE Standard for Information technology– Local and metropolitan area networks– Specific requirements– Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 6: Wireless Access in Vehicular Environments. *IEEE Std 802.11p-2010 (Amendment to IEEE Std 802.11-2007 as amended by IEEE Std 802.11k-2008, IEEE Std 802.11r-2008, IEEE Std 802.11y-2008, IEEE Std 802.11n-2009, and IEEE Std 802.11w-2009)*, 1–51. 26, 119

Abhyankar, S. & Agrawal, D.P. (2002). Distributed mobility-aware route selection for wireless ad hoc networks. In *Performance, Computing, and Communications Conference, 2002. 21st IEEE International*, 241–247, IEEE. 84

Abramson, N. (1970). The aloha system: another alternative for computer communications. In *Proceedings of the November 17-19, 1970, fall joint computer conference*, AFIPS '70 (Fall), 281–285, ACM, New York, NY, USA. 14

Bahl, P., Chandra, R. & Dunagan, J. (2004). Ssch: slotted seeded channel hopping for capacity improvement in ieee 802.11 ad-hoc wireless networks. In *Proceedings of the 10th annual international conference on Mobile computing and networking*, MobiCom '04, 216–230, ACM, New York, NY, USA. 33

Bai, F. & Krishnan, H. (2006). Reliability Analysis of DSRC Wireless Communication for Vehicle Safety Applications. In *Intelligent Transportation Systems Conference, 2006. ITSC '06. IEEE*, 355–362. 28, 122

Bao, L. & Garcia-Luna-Aceves, J.J. (2001). A new approach to channel access scheduling for ad hoc networks. In *Proceedings of the 7th annual international conference on Mobile computing and networking*, MobiCom '01, 210–221, ACM, New York, NY, USA. 52

Basu, P., Khan, N. & Little, T.D. (2001). A mobility based metric for clustering in mobile ad hoc networks. In *Distributed Computing Systems Workshop, 2001 International Conference on*, 413–418, IEEE. 85

Bharghavan, V., Demers, A., Shenker, S. & Zhang, L. (1994). Macaw: a media access protocol for wireless lan's. In *Proceedings of the conference*

on Communications architectures, protocols and applications, SIGCOMM '94, 212–225, ACM, New York, NY, USA. 16

Bilstrup, K., Uhlemann, E., Ström, E.G. & Bilstrup, U. (2009). On the ability of the 802.11p mac method and stdma to support real-time vehicle-to-vehicle communication. *EURASIP J. Wirel. Commun. Netw.*, **2009**, 5:1–5:13. 28, 54

Borgonovo, F., Capone, A., Cesana, M. & Fratta, L. (2002). RR-ALOHA, a Reliable R-ALOHA broadcast channel for ad-hoc inter-vehicle communication networks. *proceedings of MedHocNet*. 41

Borgonovo, F., Capone, A., Cesana, M. & Fratta, L. (2004). ADHOC MAC: new MAC architecture for ad hoc networks providing efficient and reliable point-to-point and broadcast services. *Wireless Networks*, **10**, 359–366. 41, 119

Briesemeister, L., Schafers, L. & Hommel, G. (2000). Disseminating messages among highly mobile hosts based on inter-vehicle communication. In *Intelligent Vehicles Symposium, 2000. IV 2000. Proceedings of the IEEE*, 522–527, IEEE. 75

Cai, Z., Lu, M. & Georghiades, C. (2003). Topology-transparent time division multiple access broadcast scheduling in multihop packet radio networks. *Vehicular Technology, IEEE Transactions on*, **52**, 970 – 984. 37

Camp, T., Boleng, J. & Davies, V. (2002). A survey of mobility models for ad hoc network research. *Wireless communications and mobile computing*, **2**, 483–502. 81, 124

Chen, J., Sheu, S.T. & Yang, C.A. (2003). A new multichannel access protocol for ieee 802.11 ad hoc wireless lans. In *Personal, Indoor and Mobile Radio Communications, 2003. PIMRC 2003. 14th IEEE Proceedings on*, vol. 3, 2291–2296, IEEE. 32

CHLAMTAC, I. & FARAGO, A. (1994). Making transmission schedules immune to topology changes in multi-hop packet radio networks. *Networking, IEEE/ACM Transactions on*, **2**, 23 –29. 35

CHLAMTAC, I., MYERS, A., SYROTIUK, V. & ZARUBA, G. (2000). An adaptive medium access control (mac) protocol for reliable broadcast in wireless networks. In *Communications, 2000. ICC 2000. 2000 IEEE International Conference on*, vol. 3, 1692 –1696 vol.3. 47

CHOI, N., SEOK, Y. & CHOI, Y. (2003). Multi-channel mac protocol for mobile ad hoc networks. In *Vehicular Technology Conference, 2003. VTC 2003-Fall. 2003 IEEE 58th*, vol. 2, 1379 – 1382 Vol.2. 32

COZZETTI, H. & SCOPIGNO, R. (2009). Rr-aloha+: A slotted and distributed mac protocol for vehicular communications. In *Vehicular Networking Conference (VNC), 2009 IEEE*, 1–8. 56

CRICHIGNO, J., WU, M.Y. & SHU, W. (2008). Protocols and architectures for channel assignment in wireless mesh networks. *Ad Hoc Networks*, **6**, 1051 – 1077. 29

EICHLER, S. (2007). Performance evaluation of the ieee 802.11 p wave communication standard. In *Vehicular Technology Conference, 2007. VTC-2007 Fall. 2007 IEEE 66th*, 2199–2203, IEEE. 28

GARCIA-LUNA-ACEVES, J.J. & FULLMER, C.L. (1999). Floor acquisition multiple access (fama) in single-channel wireless networks. *Mob. Netw. Appl.*, **4**, 157–174. 15

GERLA, M., BAJAJ, L., TAKAI, M., AHUJA, R. & BAGRODIA, R. (1999). Glomosim: a scalable network simulation environment. *University of California, Los Angeles, Computer Science Department, Technical Report*, **990027**. 134

HAAS, Z. & DENG, J. (2002). Dual busy tone multiple access (dbtma)-a multiple access control scheme for ad hoc networks. *Communications, IEEE Transactions on*, **50**, 975–985. 8, 15, 18

Härri, J., Bonnet, C. & Filali, F. (2008). Kinetic mobility management applied to vehicular ad hoc network protocols. *Computer Communications*, **31**, 2907–2924. 84

Harri, J., Filali, F. & Bonnet, C. (2009). Mobility models for vehicular ad hoc networks: a survey and taxonomy. *Communications Surveys & Tutorials, IEEE*, **11**, 19–41. 81, 82

Hartenstein, H. & Laberteaux, K.P. (2008). A tutorial survey on vehicular ad hoc networks. *Communications Magazine, IEEE*, **46**, 164–171. 2

Ho, C., Obraczka, K., Tsudik, G. & Viswanath, K. (1999). Flooding for reliable multicast in multi-hop ad hoc networks. In *Proceedings of the 3rd international workshop on Discrete algorithms and methods for mobile computing and communications*, 64–71, ACM. 75

Hoi-Sheung, So, W., Walrand, J. & Mo, J. (2007). Mcmac: A parallel rendezvous multi-channel mac protocol. In *Wireless Communications and Networking Conference, 2007.WCNC 2007. IEEE*, 334 –339. 33

Hong, X., Gerla, M., Pei, G. & Chiang, C.C. (1999). A group mobility model for ad hoc wireless networks. In *Proceedings of the 2nd ACM international workshop on Modeling, analysis and simulation of wireless and mobile systems*, 53–60, ACM. 124

Jain, N., Das, S.R. & Nasipuri, A. (2001). A multichannel csma mac protocol with receiver-based channel selection for multihop wireless networks. In *Computer Communications and Networks, 2001. Proceedings. Tenth International Conference on*, 432–439, IEEE. 30

Johnson, D.B. & Maltz, D.A. (1996). Dynamic source routing in ad hoc wireless networks. *Kluwer International Series in Engineering and Computer Science*, 153–179. 74

Ju, J.H. & Li, V. (1998). An optimal topology-transparent scheduling method in multihop packet radio networks. *Networking, IEEE/ACM Transactions on*, **6**, 298 –306. 39

KARN, P. (1990). Maca-a new channel access method for packet radio. In *ARRL/CRRL Amateur radio 9th computer networking conference*, vol. 140, 134–140. 15

KLEINROCK, L. & TOBAGI, F. (1975). Packet switching in radio channels: Part i–carrier sense multiple-access modes and their throughput-delay characteristics. *Communications, IEEE Transactions on*, **23**, 1400–1416. 14

LEE, S.J., SU, W. & GERLA, M. (1999). Ad hoc wireless multicast with mobility prediction. In *Computer Communications and Networks, 1999. Proceedings. Eight International Conference on*, 4–9, IEEE. 85

LI, J., HAAS, Z.J., SHENG, M. & CHEN, Y. (2003). Performance evaluation of modified ieee 802.11 mac for multi-channel multi-hop ad hoc network. In *Advanced Information Networking and Applications, 2003. AINA 2003. 17th International Conference on*, 312–317, IEEE. 30

MARINA, M., KONDYLIS, G. & KOZAT, U. (2001). Rbrp: a robust broadcast reservation protocol for mobile ad hoc networks. In *Communications, 2001. ICC 2001. IEEE International Conference on*, vol. 3, 878–885 vol.3. 47

MCDONALD, A.B. & ZNATI, T.F. (1999). A mobility-based framework for adaptive clustering in wireless ad hoc networks. *Selected Areas in Communications, IEEE Journal on*, **17**, 1466–1487. 85

MODELER, O. (2009). OPNET Technologies Inc. *Bethesda, MD. URL: http://www. opnet. com*. 128

MORENO, M.T. (2007). *Inter-vehicle communications: achieving safety in a distributed wireless environment*. Ph.D. thesis, Dissertation, Shaker. 27

MOUSTAFA, H. & ZHANG, Y. (2009). *Vehicular networks: techniques, standards, and applications*. Auerbach publications. 2

MURRAY, T., COJOCARI, M. & FU, H. (2008). Measuring the performance of ieee 802.11 p using ns-2 simulator for vehicular networks. In *Electro/Information Technology, 2008. EIT 2008. IEEE International Conference on*, 498–503, IEEE. 28

Perkins, C. & Royer, E. (1999). Ad-hoc on-demand distance vector routing. In *Mobile Computing Systems and Applications, 1999. Proceedings. WMCSA '99. Second IEEE Workshop on*, 90–100. 74

PTV, A. (2004). Vissim. *Web page, URL http://www. english. ptv. de/cgi-bin/traffic/traf_vissim. pl*. 125

Rozovsky, R. & Kumar, P.R. (2001). Seedex: a mac protocol for ad hoc networks. In *Proceedings of the 2nd ACM international symposium on Mobile ad hoc networking & computing*, MobiHoc '01, 67–75, ACM, New York, NY, USA. 54

Rubio, L., Reig, J. & Cardona, N. (2007). Evaluation of nakagami fading behaviour based on measurements in urban scenarios. *AEU-International Journal of Electronics and Communications*, **61**, 135–138. 134

Sánchez, M. & Manzoni, P. (2001). Anejos: a java based simulator for ad hoc networks. *Future generation computer systems*, **17**, 573–583. 124

Scopigno, R. & Cozzetti, H. (2009). Mobile slotted aloha for vanets. In *Vehicular Technology Conference Fall (VTC 2009-Fall), 2009 IEEE 70th*, 1–5. 56

Shacham, N. & King, P. (1987). Architectures and performance of multichannel multihop packet radio networks. *Selected Areas in Communications, IEEE Journal on*, **5**, 1013–1025. 29

Sharma, S., Alatzeth, V., Grewal, G., Pradhan, S. & Helmy, A. (2004). A comparative study of mobility prediction schemes for gls location service. In *Vehicular Technology Conference, 2004. VTC2004-Fall. 2004 IEEE 60th*, vol. 5, 3125–3129, IEEE. 83

Simulator, N. (1989). ns-2. 125

So, J. & Vaidya, N.H. (2004). Multi-channel mac for ad hoc networks: handling multi-channel hidden terminals using a single transceiver. In *Proceedings of the 5th ACM international symposium on Mobile ad hoc networking and computing*, MobiHoc '04, 222–233, ACM, New York, NY, USA. 31

SOMMER, C. & DRESSLER, F. (2008). Progressing toward realistic mobility models in vanet simulations. *Communications Magazine, IEEE*, **46**, 132–137. 125

STIBOR, L., ZANG, Y. & REUMERMAN, H.J. (2007). Evaluation of communication distance of broadcast messages in a vehicular ad-hoc network using ieee 802.11 p. In *Wireless Communications and Networking Conference, 2007. WCNC 2007. IEEE*, 254–257, IEEE. 28

SU, W., LEE, S.J. & GERLA, M. (2000). Mobility prediction in wireless networks. In *MILCOM 2000. 21st Century Military Communications Conference Proceedings*, vol. 1, 491–495, IEEE. 81

SU, W.W.L. (2000). *Motion prediction in mobile/wireless networks*. Ph.D. thesis, University of California Los Angeles. 85

TANENBAUM, A. (2002). *Computer Networks*. Prentice Hall Professional Technical Reference, 4th edn. 14

TANG, K. & GERLA, M. (2001). Mac reliable broadcast in ad hoc networks. In *Military Communications Conference, 2001. MILCOM 2001. Communications for Network-Centric Operations: Creating the Information Force. IEEE*, vol. 2, 1008–1013 vol.2. 19

TANG, Z. & GARCIA-LUNA-ACEVES, J. (1999). A protocol for topology-dependent transmission scheduling in wireless networks. In *Wireless Communications and Networking Conference, 1999. WCNC. 1999 IEEE*, 1333–1337 vol.3. 47

TONGUZ, O., WISITPONGPHAN, N., BAI, F., MUDALIGE, P. & SADEKAR, V. (2007). Broadcasting in vanet. In *2007 mobile networking for vehicular environments*, 7–12, IEEE. 75

TSENG, Y.C., NI, S.Y., CHEN, Y.S. & SHEU, J.P. (2002). The broadcast storm problem in a mobile ad hoc network. *Wireless networks*, **8**, 153–167. 75

Tzamaloukas, A. & Garcia-Luna-Aceves, J. (2000). Channel-hopping multiple access. In *Communications, 2000. ICC 2000. 2000 IEEE International Conference on*, vol. 1, 415 –419 vol.1. 33

VCS (2006). Vehicle Safety Communications Project. Final Report HS 810 591, DOT. 2, 6, 28, 120

Venkateswaran, A., Sarangan, V., Gautam, N. & Acharya, R. (2005a). Impact of mobility prediction on the temporal stability of manet clustering algorithms. In *Proceedings of the 2nd ACM international workshop on Performance evaluation of wireless ad hoc, sensor, and ubiquitous networks*, 144–151, ACM. 83

Venkateswaran, A., Sarangan, V., Gautam, N. & Acharya, R. (2005b). Impact of mobility prediction on the temporal stability of manet clustering algorithms. In *Proceedings of the 2nd ACM international workshop on Performance evaluation of wireless ad hoc, sensor, and ubiquitous networks*, 144–151, ACM. 85

Wang, K.H. & Li, B. (2002). Group mobility and partition prediction in wireless ad-hoc networks. In *Communications, 2002. ICC 2002. IEEE International Conference on*, vol. 2, 1017–1021, IEEE. 85

Wegener, A., Hellbruck, H., Fischer, S., Schmidt, C. & Fekete, S. (2007). Autocast: An adaptive data dissemination protocol for traffic information systems. In *Vehicular Technology Conference, 2007. VTC-2007 Fall. 2007 IEEE 66th*, 1947–1951, IEEE. 75

Wu, C. & Li, V. (1987). Receiver-initiated busy-tone multiple access in packet radio networks. *SIGCOMM Comput. Commun. Rev.*, **17**, 336–342. 18

Xu, K., Gerla, M. & Bae, S. (2002). How effective is the ieee 802.11 rts/cts handshake in ad hoc networks. In *Global Telecommunications Conference, 2002. GLOBECOM '02. IEEE*, vol. 1, 72–76 vol.1. 21

Zhu, C. & Corson, M. (1998). A five-phase reservation protocol (FPRP) for mobile ad hoc networks. In *INFOCOM'98. Seventeenth Annual Joint Confer-*

ence of the IEEE Computer and Communications Societies. Proceedings. IEEE, vol. 1, 322–331, IEEE. 44

ZIMMERMANN, H. (1980). Osi reference model–the iso model of architecture for open systems interconnection. *Communications, IEEE Transactions on*, **28**, 425–432. 7