Adaptive Medium Access Control for Internet-of-Things Enabled Mobile Ad Hoc Networks

by

Qiang Ye

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2016

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

An Internet-of-Things (IoT) enabled mobile ad hoc network (MANET) is a self organized distributed wireless network, in which nodes can randomly move making the network traffic load vary with time. A medium access control (MAC) protocol, as a most important mechanism of radio resource management, is required in MANETs to coordinate nodes' access to the wireless channel in a distributed way to satisfy their quality of service (QoS) requirements. However, the distinctive characteristics of IoT-enabled MANETs, i.e., distributed network operation, varying network traffic load, heterogeneous QoS demands, and increased interference level with a large number of nodes and extended communication distances, pose technical challenges on MAC. An efficient MAC solution should achieve consistently maximal QoS performance by adapting to the network traffic load variations, and be scalable to an increasing number of nodes in a multi-hop communication environment. In this thesis, we develop comprehensive adaptive MAC solutions for an IoT-enabled MANET with the consideration of different network characteristics.

First, an adaptive MAC solution is proposed for a fully-connected network, supporting homogeneous best-effort data traffic. Based on the detection of current network traffic load condition, nodes can make a switching decision between IEEE 802.11 distributed coordination function (DCF) and dynamic time division multiple access (D-TDMA), when the network traffic load reaches a threshold, referred to as MAC switching point. The adaptive MAC solution determines the MAC switching point in an analytically tractable way to achieve consistently high network performance by adapting to the varying network traffic load.

Second, when heterogeneous services are supported in the network, we propose an adaptive hybrid MAC scheme, in which a hybrid superframe structure is designed to accommodate the channel access from delay-sensitive voice traffic using time division multiple access (TDMA) and from best-effort data traffic using truncated carrier sense multiple access with collision avoidance (T-CSMA/CA). According to instantaneous voice and data traffic load conditions, the MAC exploits voice traffic multiplexing to increase the voice capacity by adaptively allocating TDMA time slots to active voice nodes, and maximizes the aggregate data throughput by adjusting the optimal contention window size for each data node.

Lastly, we develop a scalable token-based adaptive MAC scheme for a two-hop MANET with an increasing number of nodes. In the network, nodes are partitioned into different one-hop node groups, and a TDMA-based superframe structure is proposed to allocate different TDMA time durations to different node groups to overcome the hidden terminal problem. A probabilistic token passing scheme is adopted for packet transmissions

within different node groups, forming different token rings. An average end-to-end delay optimization framework is established to derive the set of optimal MAC parameters for a varying network load condition. With the optimal MAC design, the proposed adaptive MAC scheme achieves consistently minimal average end-to-end delay in an IoT-based two-hop environment with a high network traffic load.

This research on adaptive MAC provides some insights in MAC design for performance improvement in different IoT-based network environments with different QoS requirements.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor, Professor Weihua Zhuang, for her continuous help, guidance and support during my Ph.D. study and my personal life in University of Waterloo. Her insightful thoughts and valuable instructions always inspire me to do in-depth thinking on my research. Moreover, her rigorous attitude towards the way of doing research has also influenced me as the research spirit. During the four years' Ph.D. study, I deeply feel not only the inspirational research talents but also the personal charisma from my professor. I am so grateful and honored for being one of her students.

Secondly, I would like to thank Professor Mohamed Oussama Damen, Professor Oleg Michailovich, Professor Peisong Han, and the external examiner, Professor Dongmei Zhao, from McMaster University, for serving my thesis defense committee members. Their suggestions, comments and valuable questions have helped me significantly to improve the quality of my thesis. I would also like to express my special thanks to Professor Xuemin (Sherman) Shen for his generous help and continuous support during my study.

In the past four years, I have also gained lots of help from my current and former colleagues in our BBCR group where we have established precious friendships. I would like to express my sincere gratitude to all the BBCR members and my friends in University of Waterloo for their selfless help, especially Chong Lou, Kede Ma, Dr. Ju Ren, Deyu Zhang, Wen Wu, Dr. Hassan Omar, Dr. Kan Yang, Haixia Peng, Dr. Khadige Abboud, Dr. Nan Cheng, Dr. Hao Liang, Hesham Moussa, Dr. Yong Li, Kun Zhuge, Dr. Yong Zhou, Amit Kumar Tamang, Dr. Qinghua Shen, Dr. Lei Xu, Dr. Ning Zhang, Dr. Kuan Zhang, Dr. Haibo Zhou, Dr. Chengzhe Lai, Wenchao Xu, Dr. Miao Wang, Dr. Ran Zhang, Dr. Yuanguo Bi, Dr. Ning Lu, Dr. Tom H. Luan, Dr. Ruilong Deng, Dr. Rong Jiang, Baozhu Li, Dr. Kamal Malekshan, Dr. Wenbo Su, Miao He, Dr. Zhou Su, Dr. Ying Wang, Dr. Shan Zhang, Ying Chen, Nan Chen, Jianbing Ni, Dr. Jiping Li, Dr. Hui Zhang, Dr. Md. Shamsul Alam, and many others.

Finally, I would like to greatly thank my family members, my beloved father, mother, and my dearest wife, for their deep love and considerate care during my Ph.D. study.

Qiang Ye

Sept. 26, 2016

Waterloo, Ontario, Canada

This PhD thesis is dedicated to my beloved parents, Dazhen Ye and Rui Wang, and my dearest wife, Ye Qiu.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ACKs	Acknowledgments
AIFS	Arbitration Interframe Space
AP	Access Point
CFP	Contention-free Period
CP	Contention Period
CTP	Control Period
CSMA/CA	Carrier Sense Multiple Access with Collision Avoidance
D	Destination
DCF	Distributed Coordination Function
DIFS	Distributed Interframe Space
D-PRMA	Distributed Packet Reservation Multiple Access
D-TDMA	Dynamic Time Division Multiple Access
DTSA	Dynamic Time Slot Assignment
D2D	Device-to-Device
EDCA	Enhanced Distributed Channel Access
GPS	Global Positioning System
HC	High Contention
HOL	Head-of-Line
IoT	Internet-of-Things
MAC	Medium Access Control
MANET	Mobile Ad hoc Network
M2M	Machine-to-Machine
QoS	Quality-of-Service
R	Relay

RTS/CTS	Request-to-Send/Clear-to-Send
S	Source
T-CSMA/CA	Truncated Carrier Sense Multiple Access with Collision Avoidance
TDMA	Time Division Multiple Access
WLAN	Wireless Local Area Network
WSN	Wireless Sensor Network
VANET	Vehicular Ad hoc Network

List of Symbols

$\frac{1}{\alpha}$ / $\frac{1}{\beta}$	Average duration in *on/off* state of a voice node
λ	Average packet arrival rate at each node
$\lambda_{ac}(\lambda_{bc})$	Average external traffic arrival rate heading to area C from nodes in area A (B)
$\lambda_{ca}(\lambda_{cb})$	Average compound traffic arrival rate at a relay node for a destination in area A (B)
$\lambda_a(\lambda_b)$	Average external traffic arrival rate at nodes in area A (B) for local transmissions
λ_v/λ_d	Average packet arrival rate at each voice/data node for DAH-MAC
μ_j	Average packet service rate for nodes in R_j $(j = ac, bc, a, b)$
μ_d	Average packet service rate of each node with IEEE 802.11 DCF
μ_t	Average packet service rate of each node with D-TDMA
ρ	Individual node queue utilization ratio
τ	Packet transmission probability in a back-off slot
τ_{opt}	Optimal data packet transmission probability
φ	Maximum time fraction for voice traffic in each superframe
B	Average voice burst size
CW	Minimum contention window size
CW_{opt}	Optimal contention window size
$\overline{CW_2}$	Average back-off contention window time
$D_{ac}(D_{bc})$	Average delay for packet transmissions from area A (B) to C
$D_{ca}(D_{cb})$	Average delay for relay packet transmissions from area C to A (B)
$D_a(D_b)$	Average delay for local packet transmissions within area A (B)
$D_{ab}(D_{ba})$	Average end-to-end delay for the transmission direction from area A (B)

	to B (A)
D^{opt}	Minimal average end-to-end delay
D_T	Average packet delay for IEEE 802.11 DCF
D_{th}	A delay bound for local packet transmissions within area A and area B
D^*	Minimized average end-to-end delay under a certain superframe length
k_j	Number of token rotation cycles scheduled for token ring R_j $(j = ac, bc, a, b)$
k_j^*	Optimal number of token rotation cycles scheduled for token ring R_j $(j = ac, bc, a, b)$ under a certain superframe length
k_j^{opt}	Optimal number of token rotation cycles scheduled for token ring R_j $(j = ac, bc, a, b)$
L_j	Number of node members in token ring R_j $(j = ac, bc, a, b)$
$\mathcal{L}(j)$	Probabilistic token passing list for token ring R_j $(j = ac, bc, a, b)$
M_b	Maximum back-off stage
M_v	Maximum number of voice packets generated in a superframe
M	Total number of time slots for each superframe in TA-MAC
M_L	Retransmission limit
M_m	Number of minislots
M^{opt}	Optimal total number of time slots for each superframe
N	Total number of nodes in the network
N_a,N_b,N_c	Number of nodes in network areas A, B, C
N_{av}	Number of active voice nodes in each superframe
N_d/N_v	Number of data/voice nodes
$\overline{N_s}$	Average number of voice bursts scheduled for transmission
N_{sm}	Maximum number of voice bursts scheduled for transmission
N_{vm}	Number of minislots in each control period (voice capacity)
$\mathcal{N}(x)$	One-hop neighbor node IDs of node x
p	Conditional collision probability
P_L	Voice packet loss rate bound
p_v	Probability of a generic time slot inside the vulnerable period
R_{ac},R_{bc},$R_a(R_b)$	Four token rings formed among different node groups
S	Aggregate network throughput
S_d	Normalized saturation data throughput for T-CSMA/CA
T_{ac},T_{bc},T_{ab}	Three TDMA time durations in each superframe

T_c	Collision time packets experience during a collision
$\overline{T_c}$	Average collision time encountered before a packet is transmitted
$\overline{T_{cfp}}$	Average duration of contention-free period
T_{cfpm}	Maximum duration of contention-free period
T_o	Duration of a conflict period
$\overline{T_{cp}}$	Average duration of contention period
T_{ctrl}	Control period duration
T_f	Superframe length for TA-MAC
T_f^{opt}	Optimal superframe length for TA-MAC
T_m	Minislot duration
T_p	Packet transmission time for D-TDMA
T_{pd}/T_{pv}	Data/voice packet duration
T_{pl}	Duration of packet payload
T_s	Successful packet transmission time (slot) duration
T_{SF}	Superframe duration for DAH-MAC
T_1/T_2	Idle duration before data (token) packet / REQUEST packet transmissions
$\overline{T_v}$	Average duration of vulnerable period
$\overline{T_{vd}}$	Average virtual transmission time
W_{qt}	Packet queueing delay for D-TDMA
W_{st}	Packet access delay for D-TDMA
y_m	Maximum number of transmitted voice packets in each superframe
$W_{s,j}$	Packet service time for a node in token ring R_j ($j = ac, bc, a, b$)

Chapter 1

Introduction

Mobile ad hoc networks (MANETs) within the Internet-of-Things (IoT) framework is one of the most important wireless networks, in which mobile nodes are interconnected and transmit packets in a distributed way to provide various IoT-oriented services. To improve the performance for IoT-enabled MANETs, proper medium access control (MAC) is required to distributedly coordinate communications and interactions among mobile nodes. However, the unique characteristics of IoT pose technical challenges on MAC, i.e., the increasing network traffic load, heterogeneous quality-of-service (QoS) demands, and the increased interference level in a multi-hop environment with a continuous injection of nodes and longer communication distances. Therefore, MAC for an IoT-enabled MANET is required to achieve consistently maximal performance by adapting to network traffic load variations, providing the heterogeneous QoS guarantee, and eliminating the interference in a multi-hop environment. In this chapter, we first give an overview of the characteristics of IoT-enabled MANETs, upon which the MAC issues for MANETs are discussed. Then, we give our research objectives and thesis outline.

1.1 IoT-Enabled MANETs

The IoT is one of the most promising network infrastructures towards the next generation wireless network evolution. The IoT framework will interconnect a growing number of heterogeneous objects, i.e., smartphones, sensors and actuators, autonomous devices, via suitable wireless technologies for ubiquitous Internet access and pervasive spectrum sharing [3] [4]. Within this framework, various IoT-oriented intelligent applications can be realized, e.g., environment monitoring [3], intelligent control for smart homing [5], and industrial automation [6]. To support the increasing number of nodes and user demands, an IoT-enabled mobile ad hoc network (MANET) emerges as a promising wireless network to provide seamless Internet access for end users[1].

A MANET consists of a group of self-organized nodes, interconnected for communication in a peer-to-peer manner, without any centralized control. Due to low cost and simplified implementation, MANETs are widely deployed for applications such as smart home networking [3], prompt response in postdisaster areas [7], and tactical networks for the purpose of command interactions [8]. Some typical realizations for IoT-enabled MANETs have been invented and recently popularized, such as device-to-device (D2D) communications, which rely on the ad hoc networking of spatially-distributed smart devices for information relaying and sharing [9] [10], and machine-to-machine (M2M) communications, where an increasing number of heterogeneous objects are connected through wireless interfaces and exchange information without human intervention [6]. To achieve consistently high QoS performance for a MANET, proper MAC is imperative. A MAC protocol is a mechanism to coordinate nodes' access to the wireless medium to transmit their packets. However, the distinctive features of IoT-enabled MANETs make the MAC protocol design challenging:

- **Distributed network operation:** Since MANETs do not depend on any infrastructure or centralized control, each node interacts with others in a peer-to-peer distributed manner. Therefore, a distributed MAC protocol is required, upon which each node makes its transmission decision based on its local information. The distributed network operation makes the MAC more challenging.

- **Node mobility:** For a MANET, nodes can randomly move within the network coverage area, and also come into or depart from the network, making the network traffic load vary with time. The traffic load variations can lead to QoS performance degradation. Therefore, a MAC protocol should be adaptive to the varying number of nodes in the network, to maintain consistently high network performance.

- **Heterogeneous QoS demands:** With an increasing demand for supporting heterogeneous services for an IoT-based MANET, QoS support for different types of applications becomes an important task. For example, packet delay is important for delay-sensitive voice traffic, while throughput is more concerned for best-data traffic. Therefore, the network is expected to not only provide as high as possible throughput for best-effort data traffic, but also ensure a bounded packet loss rate for delay-sensitive voice communications or even multimedia streaming. Therefore, QoS-aware MAC is required to coordinate the channel access for heterogeneous traffic in a differentiated way to satisfy QoS requirements of all individual users [7] [11] [12].

[1] An end user can also mean an end device in this thesis.

- **Increased interference and scalability issue:** The IoT infrastructure has been envisioned to accommodate an increasing number of users, which may degrade the network performance due to the increased contention level. Therefore, the MAC protocol should be scalable to the number of nodes to achieve high throughput and low delay, especially under a high network load condition [13]; The increased number of nodes can enlarge the network coverage area, making the communication distance between a pair of end users beyond the one-hop transmission (communication) range. For multi-hop communications, transmission collisions are accumulated due to increased interference level from the hidden terminal problem [12] [14] and/or receiver blocking problem [15] [16], which become worse with an increasing number of nodes. On the other hand, in a multi-hop network, some nodes staying in the transmission ranges of both source and destination nodes (that are far apart) may relay traffic for the end nodes. Thus, the compound traffic arrival rate (superposition of the external traffic arrival rate and the relay traffic arrival rate) for each relay node can become high, resulting in a large overall delay for relay transmissions and thus for end-to-end transmissions. Therefore, to maintain consistently satisfactory end-to-end packet delay in a multi-hop environment with increased number of nodes, first, efficient multi-hop MAC is critical to coordinate the packet transmissions from nodes in each transmission hop, which should achieve high performance by avoiding the packet collisions due to hidden nodes and improving the spatial reuse of the network resources, and be scalable to the increased contention level; Second, to further improve the end-to-end performance, packet routing can be designed for nodes to select best relays along a multi-hop transmission path, which helps to balance the traffic load among relay nodes to increase the network capacity. Therefore, a joint MAC and routing cross-layer design for a multi-hop network is a more comprehensive solution for optimizing the end-to-end performance, which can be considered in further research. In this thesis, we focus on MAC for both single-hop and multi-hop networks.

1.2 MAC for MANETs

Based on specific characteristics of an IoT-enabled MANET, a comprehensive MAC solution is required to coordinate packet transmissions from mobile nodes by taking into considerations all the network features to achieve consistently satisfactory network performance, which is in general a very challenging task especially in a multi-hop communication environment with an increasing network traffic load. Therefore, in this section, we study our proposed MAC solutions upon investigating and incorporating three main aspects of the network characteristics step by step towards a more practical networking scenario.

1.2.1 Adaptive MAC

As discussed in Section 1.1, the distributed network operation makes the MAC challenging, as it requires additional control overhead to exchange neighboring information for synchronization and transmission opportunity allocation, and for network traffic load variations due to node mobility, resulting in the network performance degradation. Therefore, it is desired that a MAC protocol should 1) be efficient in distributed resource allocation and, at the same time, 2) achieve consistently high performance by adapting to network traffic load dynamics.

Existing MAC Solutions

With no reliance on topology and synchronization information, the carrier sense multiple access with collision avoidance (CSMA/CA) based contention MAC scheme, i.e., IEEE 802.11 distributed coordination function (DCF) [17], is widely used in current MANET implementations. However, as the network traffic load increases, the performance of IEEE 802.11 DCF experiences an inevitable degradation, due to an increased amount of control overhead for packet collision resolution. On the other hand, by avoiding packet transmission collisions among nodes, the channelization-based (reservation-based) time division multiple access (TDMA) schemes [18] [19] achieve higher resource utilization than the IEEE 802.11 DCF when the network traffic load is high. However, the distributed time slot acquisition of TDMA consumes a considerable amount of channel time for local information exchange among neighboring nodes, which makes the channel utilization of TDMA inferior to IEEE 802.11 DCF in a low network traffic load condition [20].

Because of the performance tradeoff between the contention-based MAC and contention-free channelization-based MAC, in order to make use of the network resources more efficiently, adaptive MAC schemes which are proposed in literature, combine CSMA/CA (or slotted-Aloha) with TDMA in a hybrid MAC frame pattern, by switching between the two MAC frame structures either periodically [21] [22] or via an adaptability to a changing network traffic load [23] [24] [25]. Generally, the adaptive MAC schemes make the switching between different MAC frame structures based on the estimation of network traffic load through measurements of some microscopic MAC operation parameters (e.g., the number of unused TDMA time slots [23], number of consecutively lost acknowledgments (ACK-s) [25], queue lengths [26]). These microscopic parameters can be effective to reflect the realtime network traffic load condition. However, how to determine the optimal value of the microscopic MAC switching point, with which the adaptive MAC solutions can achieve maximum performance, is a challenging issue. It is difficult to model the relationship between the microscopic network traffic load indicator and the MAC performance, which is mostly captured by either simulations [23] [25] or experiments [26]. Without an explicit analytical relationship between the MAC performance and the network traffic load indicator, the MAC switching points can only be set empirically, and the corresponding switching strategy does not necessarily achieve a maximal performance gain.

Proposed MAC Solution

We develop an adaptive MAC solution for a fully-connected MANET, in which the MAC switching point between IEEE 802.11 DCF and TDMA is determined based on a theoretical performance comparison of the MAC protocols. Our contribution lies in three aspects: First, for a homogeneous network traffic scenario, where all nodes have identical traffic generation statistics, we establish a mathematical relationship between the MAC performance metrics (i.e., throughput and delay) and the macroscopic network traffic load indicator (i.e., the total number of nodes); Second, most existing performance evaluations of either IEEE 802.11 DCF or TDMA rely on numerical methods (Markov chain modeling [17] [27], nonlinear system modeling based on mean value analysis [28]), which do not provide a closed-form expression for performance metrics and are thus computationally complex to conduct a performance comparison between the MAC protocols. To overcome the limitation, we establish a simplified and unified framework, considering both traffic saturation and non-saturation cases, to make the performance comparison tractable. Approximate and closed-form analytical relations are established between the MAC performance metrics and the total number of nodes in the network for both IEEE 802.11 DCF and TDMA, by using the *least-squares curve-fitting method* and *M/G/1 queueing analysis*, respectively; Third, according to the unified performance analysis framework, an adaptive MAC solution is developed to determine the MAC selection between IEEE 802.11 DCF and TDMA based on the MAC switching point calculation. The MAC switching point is adaptive to traffic load statistics of each node. It is demonstrated that the MAC solution maximizes the network performance in the presence of data traffic load dynamics.

1.2.2 QoS-Aware MAC

Some IoT-based wireless networks, e.g., wireless sensor networks (WSNs) and machine-to-machine (M2M) networks, are designed to support a large number of power-constrained sensor nodes or autonomous devices generating low data rate traffic. For the networks, energy-efficient MAC is required to coordinate packet transmissions among wireless nodes to prolong the whole network lifetime. Such studies include the CSMA/CA-based IEEE 802.15.4 ZigBee supporting energy efficient and low data rate communications in WSNs [4], a data gathering protocol in an IoT-based WSN with TDMA employed for intra-cluster data transmissions [29], and the IEEE 802.11ah operating at lower frequency bands to cover an increasing number of devices for an outdoor M2M network environment [30]. However, these MAC protocols, developed to support low data rate applications on power-limited devices, cannot guarantee the differentiated QoS requirements for heterogeneous services [3].

For a typical IoT-enabled MANET with power-rechargeable mobile nodes [31] (i.e., smartphones, laptops) generating a high volume of heterogeneous traffic, supporting heterogeneous services with differentiated QoS guarantee becomes an important but challenging task. The network is expected to not only provide as high as possible throughput for best-effort data traffic, but also ensure a bounded packet loss rate for delay-sensitive voice communications or even multimedia streaming. Therefore, QoS-aware MAC is required to coordinate the channel access for heterogeneous traffic in a differentiated way to satisfy QoS requirements of all individual users [7]. However, the characteristics of MANETs pose technical challenges in the QoS-aware MAC design: 1) Since MANETs do not depend on any central control, distributed MAC is required to coordinate the transmissions of neighboring nodes based on their local information exchanges; 2) Nodes are mobile, making the heterogeneous network traffic load change with time. The traffic load variations can lead to QoS performance degradation. Thus, MAC is expected to be context-aware, which adapts to the changing network traffic load to achieve consistently satisfactory service performance.

Existing MAC Schemes for Heterogeneous QoS Support

Contention-based MAC schemes with service differentiation are commonly used for supporting heterogeneous traffic [32–35]. The enhanced distributed channel access (EDCA), standardized in IEEE 802.11e [36], is one typical example, in which delay-sensitive realtime traffic is granted smaller arbitration interframe space (AIFS) and contention window size to access the channel with a higher probability than non-realtime traffic [33] [34]. It is demonstrated in [37] that the contention window size differentiation among realtime and non-realtime traffic is superior over the AIFS differentiation in achieving a smaller access delay for the realtime service in a traffic saturation condition. To grant voice traffic deterministic channel access priority for further improving the delay performance, busy-tone based contention protocols are proposed [2] [38], in which each voice node broadcasts a busy-tone signal, instead of decrementing a backoff counter, after an idle AIFS duration to prevent the contention intervention from data nodes. Even if the contention separation is achieved between voice traffic and data traffic in busy-tone based protocols, contention collisions still exist and accumulate among voice (data) nodes themselves after the voice (data) traffic load becomes high, making the delay (throughput) performance degraded to an unacceptable level.

By avoiding contention collisions, distributed TDMA schemes [39] [40] allocate time slots to each node in a distributed way for exclusive use. They are more effective than contention-based MAC schemes in guaranteeing the delay of realtime traffic especially in a high traffic load condition, where channel time is accumulated and wasted for packet collisions resolution in contention-based schemes. To maximize resource utilization, the distributed TDMA time slot allocation should be adaptive to the instantaneous voice traffic load [39]. In [41], a TDMA-based distributed packet reservation multiple access (D-PRMA) protocol is proposed, in which voice nodes are granted a higher probability than data nodes to contend for the channel based on slotted-Aloha. Once a contention is successful, the same time slot in each subsequent frame is reserved for the successful voice node until the slot is detected idle. Each D-PRMA frame consists of a fixed number of time slots to support transmissions from a certain number of voice and data nodes, which is not flexible when the number of nodes varies over a wide range. In addition, TDMA-based schemes can be underperformed in supporting best-effort data traffic. Since the data traffic generation is bursty, some of the time slots are wasted when the traffic load is low.

To guarantee the voice delay bound and achieve high resource utilization with multiplexing for best-effort data, hybrid MAC schemes are better options which combine a TDMA period for voice transmissions and a contention period for data transmissions using CSMA/CA-based mechanisms within a superframe [21] [22]. In [2], a hybrid MAC scheme is developed for wireless local area networks (WLANs), in which voice nodes in talk spurts are polled by an access point (AP) to transmit packets in a contention-free period, whereas the remaining idle voice nodes, once having packets to transmit, contend with data nodes to access the channel according to the busy-tone contention protocol in a contention period. Since contention-based MAC schemes experience throughput degradation with an increase of network traffic load, some existing methods adapt the contention window size to node density [42] or node relative velocity [43] to achieve consistently high network throughput. However, within a hybrid MAC superframe structure, how to achieve a consistently maximal data throughput over heterogeneous traffic load variations and how to adaptively allocate time slots based on instantaneous voice traffic load in a distributed way to maximize voice traffic multiplexing gain still remain unsolved.

Proposed MAC Solution

We propose a distributed and adaptive hybrid MAC scheme (DAH-MAC), in which distributed TDMA is employed for voice packet transmissions to guarantee a voice packet loss rate bound and truncated CSMA/CA (T-CSMA/CA) is used for data nodes to access the channel. Most of the existing contention-based MAC schemes evaluate the average access delay for voice traffic in a saturation condition or with a constant arrival rate [28], which is not the case in reality. We use a more accurate *on/off* model [2] for voice traffic generation, and exploit voice traffic multiplexing to improve the voice capacity [2]. Since voice service is realtime, a packet not transmitted after a delay bound should be dropped at the source, and the voice packet delay has to be evaluated in a stochastic manner [44] for calculating the packet loss probability. In this way, the delay requirement for voice traffic can be satisfied probabilistically by guaranteeing the voice packet loss rate below a given bound. The contributions of this work are three-folded:

1. To guarantee the voice packet loss rate bound, we present a distributed and traffic-adaptive TDMA time slot allocation scheme to allocate one time slot for each active voice node according to its transmission buffer state. Also, we establish an analytical model so that the MAC scheme can determine the voice capacity region by adjusting a MAC parameter, i.e., the maximum time fraction allocation requirement for voice traffic in each superframe, which facilitates voice session admission control for QoS guarantee. By exploiting the voice traffic multiplexing, the resource utilization for voice traffic is improved significantly;

2. The T-CSMA/CA based contention scheme is employed for data traffic access. We establish an analytical model of data saturation throughput for the DAH-MAC. The saturation throughput is a function of the number of voice and data nodes as well as the packet transmission probability of each data node;

3. For the saturation throughput of the DAH-MAC, we derive an approximate closed-form expression of the optimal data packet transmission probability as a function of the heterogeneous network traffic load. Further, we obtain a closed-form expression of the optimal contention window size, which establishes a mathematical relationship between the MAC layer parameter and the heterogeneous network traffic load. Based on the analysis, the maximum best-effort data saturation throughput can be achieved by adjusting the contention window size according to variations of the number of voice and data nodes.

1.2.3 Interference-Aware and Scalable MAC

For an IoT-enabled MANET, the increased number of nodes can enlarge the network coverage area and make the communication distance between a pair of source (S) and destination (D) nodes beyond the one-hop transmission (communication) range. Therefore, some intermediate nodes, located within both communication ranges of a pair of S-D nodes that are far apart, not only transmit data packets generated at their own application layer, but may also relay packets between the source node and destination node. Consequently, the total traffic arrival rate at a node depends on its location within the MANET, which indicates whether or not the node can act as a relay. The traffic load on each relay node can become high with a continuous injection of nodes into the network, resulting in a large overall delay for relay transmissions and thus for end-to-end transmissions. Therefore, an efficient MAC solution for a multi-hop MANET should not only avoid the hidden and exposed terminal problems [12] [15], by eliminating the interference among nodes and improving the spatial reuse of the network resources to achieve a high channel utilization, but also be scalable to an increasing number of nodes to maintain high throughput and low end-to-end delay, especially under high network load conditions.

Existing MAC Schemes for a Multi-hop Network

In literature, contention-based CSMA/CA with request-to-send/clear-to-send (RTS/CTS) handshaking schemes, e.g., IEEE 802.11 MAC [1], has been demonstrated not scalable in high network load conditions in a multi-hop environment, due to increased transmission collisions caused by the hidden terminal problem [14] and the receiver blocking problem [16], which become worse with an increasing number of nodes. The above problems can be solved by a dual-channel busy-tone based MAC solution [15] at the price of increased protocol complexity and additional circuitry [16]. TDMA protocols [45] [46] perform better for multi-hop transmissions, achieving high channel utilization by eliminating unintentional packet collisions due to the hidden terminal problem. In [45], a joint TDMA-based MAC and routing protocol is proposed for packet transmissions in a multi-hop vehicular ad hoc network (VANET), in which every vehicle can acquire a time slot that is not occupied by any of its two-hop neighbors upon listening to the neighboring information exchange within each frame. Dynamic TDMA time slot assignment (DTSA) is presented in [18] to support a varying number of users in a multi-hop MANET, where the frame length is doubled each time when no time slots are available for newly arriving nodes in current frame. Recently, hybrid MAC protocols, combining CSMA/CA with TDMA, are re-visited for a multi-hop environment, for example, the unused TDMA time slots contention based on C-SMA/CA [25] and the CSMA/CA-based time slots scheduling [6], to achieve a performance tradeoff between the two MAC approaches, which can be effective in a low load condition. But the network scalability is still throttled due to contention collision accumulation in high load conditions. Token-based MAC protocols, as a subset of contention-free protocols, have also generated many research interests for MANETs, due to its quality-of-service (QoS) provisioning capability [47] [48] [49] and the flexibility in supporting network topology changes [50]. A multi-channel token ring-based MAC protocol is proposed in [51] for supporting both safety and non-safety packet transmissions in a multi-hop VANET, where inter-ring communications are based on the CSMA/CA and token passing is employed for intra-ring data communications. In [52], a dual-channel token-based MAC protocol is proposed for multi-hop MANETs, in which a control channel is used for token passing and channel reservation, and data transmissions use a data channel. The performance analysis is carried out for a single-hop scenario.

End-to-end packet delay is an important performance metric to reflect the effectiveness of a MAC protocol in a multi-hop environment. However, most of the existing TDMA and token-based protocols [51] [52] allocate time slots and schedule the token passing without considering the end-to-end delay satisfaction due to the intractability of analytical modeling for the end-to-end delay and its optimization in a multi-hop network. Thus, the end-to-end packet delay can increase to an unacceptable level with an increasing node number if transmission opportunities are not adaptively allocated. Therefore, adapting TDMA time slot allocation and scheduling of token rotation cycles to the network traffic load variations is of paramount importance to ensure the protocol scalability, with a low end-to-end delay and a high aggregate network throughput.

Proposed MAC Solution

We consider a two-hop network as the first step towards a more general multi-hop environment, and propose a token-based adaptive MAC (TA-MAC) scheme. In the TA-MAC, both the number of token rotation cycles and the superframe duration are optimized and adapted to the instantaneous network traffic load to achieve consistently minimal average end-to-end packet delay. Our contributions are three-folded:

1. First, to eliminate the hidden terminal problem, a distributed TDMA-based superframe structure is considered for the TA-MAC, in which different one-hop node groups are allocated different TDMA durations. Inspired by [48] [49] for a single-hop network, each node group forms a token ring, and adopts a probabilistic token passing scheme among its group members for packet transmissions. Each token ring

maintains and updates its node members in a distributed way by adapting to the instantaneous number of nodes in the network;

2. Second, to determine the MAC parameters for performance optimization, we evaluate the average delay for end-to-end packet transmissions, the average delay for local packet transmissions, and the aggregate network throughput for the TA-MAC in closed-form functions of the MAC protocol parameters and the network traffic load;

3. Third, with a predefined superframe length, an optimization framework is established for minimizing the average end-to-end delay under the constraints of guaranteeing the bounded average delay for local transmissions and maintaining stable transmission queues of each node. The original non-convex minimization problem is then decoupled into a convex subproblem and a biconvex subproblem, which can be solved sequentially to obtain the minimized number of token rotation cycles for each token ring. Then, a distributed calculation algorithm is proposed to determine the optimal superframe length and the associated optimal numbers of token rotation cycles for each token ring, with which the minimal average end-to-end delay can be achieved.

1.3 Thesis Objective and Outline

The objective of this PhD research is to develop comprehensive distributed MAC solutions for IoT-enabled MANETs, which can achieve consistently maximal performance by adapting to the network traffic load variations, satisfy heterogeneous QoS requirements from differentiated services in presence of heterogeneous traffic load dynamics, and improve the protocol scalability with an increasing number of users and the enlarged communication distance. To attain the overall objective, the following challenges are tackled sequentially:

1. To develop a distributed and adaptive MAC solution, we study how to determine the optimal switching threshold between the contention-based MAC and the reservation-based MAC, with which the adaptive MAC can achieve consistently maximal network performance by switching between both MAC schemes in the presence of data traffic load dynamics;

2. To develop a context-aware hybrid MAC protocol for heterogeneous QoS support, we investigate how to adaptively allocate time slots based on instantaneous voice traffic load in a distributed way to maximize voice traffic multiplexing gain and how to achieve consistently maximal data throughput over heterogeneous traffic load variations, within a hybrid MAC superframe structure;

3. To develop an interference-aware and scalable MAC protocol for a two-hop network, we consider how to eliminate the hidden terminal problem based on spatial reservation of time resources from different node groups, and how to adaptively allocate time slots to nodes in each group in a varying network load condition. We also study how to derive the set of optimal MAC parameters to achieve minimal average end-to-end delay for ensuring the protocol scalability in an IoT environment.

The rest of the thesis is organized as follows. In Chapter 2, an adaptive MAC solution is presented for a fully-connected MANET supporting best-effort data traffic. The proposed MAC solution maximizes the network performance over traffic load variations by switching between IEEE 802.11 and D-TDMA. In Chapter 3, we present an adaptive hybrid MAC scheme for supporting heterogeneous traffic, which guarantees the voice packet loss rate bound and achieves maximum best-effort data saturation throughput in presence of heterogeneous traffic load dynamics. Chapter 4 introduces a token-based adaptive MAC scheme for a two-hop MANET, which demonstrates much better scalability over a wide range of the number of nodes in the network, especially in high network traffic load conditions. We conclude the thesis and give future research directions in Chapter 5.

Chapter 2

Traffic Load Adaptive MAC for Fully-Connected MANETs

In this chapter, we propose an adaptive MAC solution for a fully-connected MANET, supporting homogeneous best-effort data traffic. Based on the detection of current network load condition, nodes can make a switching decision between IEEE 802.11 DCF and dynamic TDMA (D-TDMA), when the network traffic load reaches a threshold, referred to as MAC switching point. The adaptive MAC solution determines the MAC switching point to maximize network performance. Approximate and closed-form performance analytical models for both MAC protocols are established, which facilitate the computation of MAC switching point in a tractable way. Extensive analytical and simulation results demonstrate that the adaptive MAC solution provides consistently maximal network performance in the presence of traffic load dynamics.

2.1 System Model

2.1.1 Network model

Consider a fully-connected MANET [53–55] with a single and error-free channel [44] [56]. There is no central controller in the network, and nodes coordinate their transmissions in a distributed way. The destination node for each source node is randomly selected from the rest nodes. Each mobile node generates best-effort data traffic. The data traffic arrivals at each node are modeled as a Poisson process with an average arrival rate λ

packet/s [25] [41] [57]. Packet loss among any pair of source-destination (S-D) nodes results from packet transmission collisions. The total number of nodes in the network is denoted by N, which changes slowly with time (with respect to a packet transmission time), due to user mobility.

2.2 Adaptive MAC framework

Consider two candidate MAC protocols maintained at each node in the adaptive MAC framework [26], in which a separate mediating MAC entity working on top of the MAC candidates can reconfigure the current MAC layer by switching from one MAC protocol to the other, based on the current network condition (e.g., interference level, and the total number of nodes). This adaptation of MAC to the networking environment has a potential to improve the network performance. The contention-based IEEE 802.11 DCF is considered as one candidate MAC protocol, which is a standardized and widely adopted MAC scheme based on CSMA/CA. It has a better channel utilization than slotted-Aloha [58], and has high performance at a low contention level. Since we consider a fully-connected MANET scenario where no hidden terminal problem exists, the basic access mechanism in IEEE 802.11 DCF is considered.

The channelization-based dynamic TDMA (D-TDMA) scheme [19] is chosen as the other MAC candidate, which is originally used in cellular networks. Time is partitioned to frames of constant duration. Each D-TDMA frame consists of a control period and data packet transmission period. The control period has a number of constant-duration minislots, and data transmission period is composed of a number of equal-length data slots. The duration of each data slot is the time used to transmit one data packet. The number of minislots indicates the maximum number of users the network can support, and the number of data slots equals current total number of nodes, N, in the network. The D-TDMA frame structure is shown in Fig. 2.1. In order to fit the distributed MANET

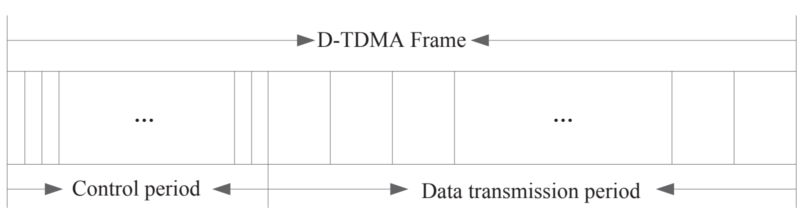


Figure 2.1: Frame structure of D-TDMA.

scenario, the minislots in the control period of each D-TDMA frame is used for local information exchange and distributed data slot acquisition of each node. The D-TDMA can support a varying number of nodes in the network and achieve high channel utilization in a high data traffic load.

2.3 Closed-form Performance Models for IEEE 802.11 DCF and D-TDMA

In this section, a unified performance analysis framework is established for both candidate MAC protocols. We present approximate and closed-form expressions for the relation between performance metrics (i.e., network throughput and packet delay) and the total number of nodes in the network. Both traffic saturation and non-saturation cases are considered. All the time durations of IEEE 802.11 DCF are normalized to the unit of a back-off time slot in the IEEE 802.11b standard.

2.3.1 Closed-form performance models for IEEE 802.11 DCF

In a homogeneous traffic case, because of the throughput fairness property of IEEE 802.11 DCF [59–61], the network throughput[1] definition can be made over one renewal cycle[2] of the transmission process. It is defined as the ratio of average payload transmission duration during one renewal cycle over the average length of the cycle [57], given by

$$S = \frac{NT_{pl}}{N\left(T_s + \frac{\overline{T_c}}{2}\right) + \overline{CW_2} + (1-\rho_r)[1-(N-1)\rho_r]\left(\frac{1}{\lambda} - D_T\right)}. \tag{2.1}$$

In (2.1), T_{pl} is the duration of each packet payload; T_s is the successful transmission time of one packet; $\overline{T_c} = \frac{p}{1-p}T_c$ is the average collision time encountered by each packet before it is successfully transmitted [44], assuming a large retransmission limit, T_c is the collision time that each packet experiences when a collision occurs, p is the packet collision probability conditioned on that the node attempts a transmission, which is assumed to be constant and independent of the number of retransmissions; $\overline{CW_2} = \frac{W_0}{2} + p\frac{W_1}{2} + p^2\frac{W_2}{2} + ... + p^{M_b}\frac{W_{M_b}}{2} +$

[1]The throughput in this chapter is normalized by the channel rate.

[2]The transmission attempt process of each node can be regarded as a regenerative process with the renewal cycle being the time between two successfully transmitted packets of the node [57].

$p^{M_b+1}\frac{W_{M_b}}{2} + ... + p^{M_L}\frac{W_{M_b}}{2}$ denotes the average back-off contention window time spent by the tagged node i during the cycle, where $W_j = 2^j CW (j = 0, 1, ..., M_b)$ is the back-off contention window size in the j-th back-off stage (CW is the minimum contention window size), M_b is the maximum back-off stage, M_L is the retransmission limit; $\rho_r = \frac{\lambda}{\mu_s} = \frac{\lambda}{N\mu_d}$ is the probability with which an incoming packet sees a non-empty queue [57], where μ_s denotes the average service rate of the IEEE 802.11 DCF, μ_d is the average packet service rate seen by an individual node; and D_T is the average packet delay, defined as the duration from the instant that a packet arrives at the transmission queue to the instant that the packet is successfully transmitted, averaged over all transmitted packets of each node.

Performance analysis in a traffic saturation case: In a traffic saturation case, (2.1) can be simplified to represent the saturation throughput S_1, given by

$$S_1 = \frac{NT_{pl}}{N\left(T_s + \frac{\overline{T_c}}{2}\right) + \overline{CW_2}} \tag{2.2}$$

which is a function of the number of nodes, N, and conditional collision probability p [62]. The collision probability p is correlated with N [28] [63],

$$p = 1 - (1-\tau)^{N-1} \tag{2.3}$$

where τ is the packet transmission probability of each node in any back-off time slot given a nonempty queue, and can be also expressed as a function of p. Eq. (2.3) captures the collision probability that each packet transmission of the tagged node encounters if at least one of the other $N-1$ nodes transmits in the same back-off time slot. In literature, there are mainly two ways to approximate τ: 1) $\tau = \frac{E[M_0]}{\overline{CW_2}}$ [28], where $E[M_0] = \frac{1-p^{M_L+1}}{1-p}$ is the average number of transmission attempts each node made before the packet is successfully transmitted or discarded due to the retransmission limit M_L; and 2) $\tau = \frac{1}{\overline{CW_1}}$ [63], where $\overline{CW_1} = \frac{1-p-p(2p)^{M_b}}{1-2p}\frac{CW}{2}$ is the average back-off contention window size between two consecutive packet transmission attempts of the tagged node. Both approximations of τ can be substituted into (2.3) for solving p with certain N.

Since variables p and N are correlated in (2.3), a high-degree nonlinear equation whose computational complexity gets higher with an increase of degree N, the saturation throughput S_1, as a function of both variables, can only be evaluated by solving (2.3) numerically. Thus, the throughput model of (2.2) and (2.3) is a nonlinear system that does not provide a closed-form expression for S_1. Based on this numerical performance model, it is computational complex, by referring to numerical techniques, e.g., Newton's method [57] and fixed-point theorem [64], to conduct a performance comparison between IEEE 802.11 DCF and the other MAC candidate. Therefore, we aim to make some approximation on (2.3) to get an explicit closed-form relation between p and N, which can be directly substituted into (2.2) to simplify S_1 as a closed-form function of only N.

Some approximations are available in literature to simplify (2.3) (e.g., first-order approximation [63], asymptotic analysis [64]). However, as shown in Fig. 2.2, the accuracy of these approximations drops when N becomes larger. In [57], the exact relationship between p and N is depicted by solving (2.3) for p over a wide range of N using numerical techniques. It is stated that p increases both monotonically and logarithmically with N, provided that M_b, M_L, and CW are specified based on the IEEE 802.11b standard. Thus, to get a more accurate closed-form function between p and N, we use a *nonlinear least-squares curve-fitting method* to fit the relation between both variables:

$$\begin{aligned}&\min_{\mathbf{a}=(a_1,a_2)} ||a_1 + a_2 \ln(\mathbf{N}) - \mathbf{P}||_2^2 \\ &\qquad \textit{subject to } a_2 \geq 0\end{aligned} \tag{2.4}$$

where vectors $\mathbf{N} = \{n \mid n \in \mathbb{Z}^+\}$ and $\mathbf{P} = \{p_n \mid n \in \mathbb{Z}^+\}$ are data sets of N and p, respectively, satisfying (2.3), and $\mathbf{a} = (a_1, a_2)$ is the vector of the fitting coefficients. In (2.4), the bounded constraint makes the optimization problem converge fast to an optimal solution [65].

Proposition 1 *Global optimal fitting coefficients in (2.4) exist since the logarithmic nonlinear least-squares curve-fitting is a convex optimization problem.*

The proof of Proposition 1 is given in Appendix 2.7.1. The logarithmic curve-fitting relation obtained between p and N is

$$p \approx a_1 + a_2 \ln(N) = -0.0596 + 0.1534 \ln(N). \tag{2.5}$$

Fig. 2.2 shows that the closed-form logarithmic fitting function is much more accurate to approximate p, over a wide range of N, than the existing approximations in [63] and [64]. Since the fitting function explicitly expresses p as a function of N, S_1 can be simplified as a closed-form function of only N, by substituting (2.5) into (2.2). However, since the average back-off contention window $\overline{CW_2}$ is a high-degree function of p, the approximate expression of S_1 is still complicated. The expression of $\overline{CW_2}$ can be approximated by an

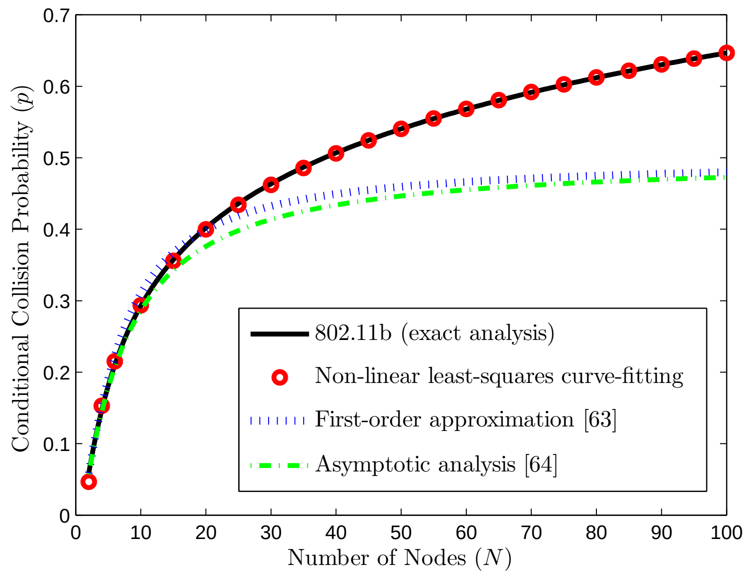


Figure 2.2: Least-squares curve-fitting between p and N.

exponential function of p, because the Taylor expansion of an exponential function has a mathematical form similar to the expression of $\overline{CW_2}$. That is,

$$\begin{aligned}
\overline{CW_2} &= \frac{W_0}{2} + \frac{W_1}{2}p + \frac{W_2}{2}p^2 + ... + \frac{W_{M_b}}{2}p^{M_b} + \frac{W_{M_b}}{2}p^{M_b+1} + ... + \frac{W_{M_b}}{2}p^{M_L} \\
&\approx b_1 + b_2 \exp(b_3 p) \\
&= (b_1 + b_2) + b_2 b_3 p + b_2 \frac{b_3^2}{2!}p^2 + b_2 \frac{b_3^3}{3!}p^3 + ...
\end{aligned} \tag{2.6}$$

where $(b_1, b_2, b_3) = (12.9590, 3.5405, 6.5834)$ are the coefficients of the exponential function obtained through the nonlinear least-squares curve-fitting method. Then, by substituting (2.5) into (2.6), $\overline{CW_2}$ can be further approximated by a closed-form function of N,

$$\begin{aligned}
\overline{CW_2} &\approx b_1 + b_2 \exp\left[b_3(a_1 + a_2 \ln(N))\right] \\
&= b_1 + b_2 \exp(b_3 a_1) \exp\left[b_3 a_2 \ln(N)\right].
\end{aligned} \tag{2.7}$$

Fig. 2.3 shows that $\overline{CW_2}$ has a nearly linear relation with N since $b_3 a_2 \approx 1$ in (2.7).

By substituting (2.5) and (2.7) into (2.2), we obtain a simplified and closed-form ex-

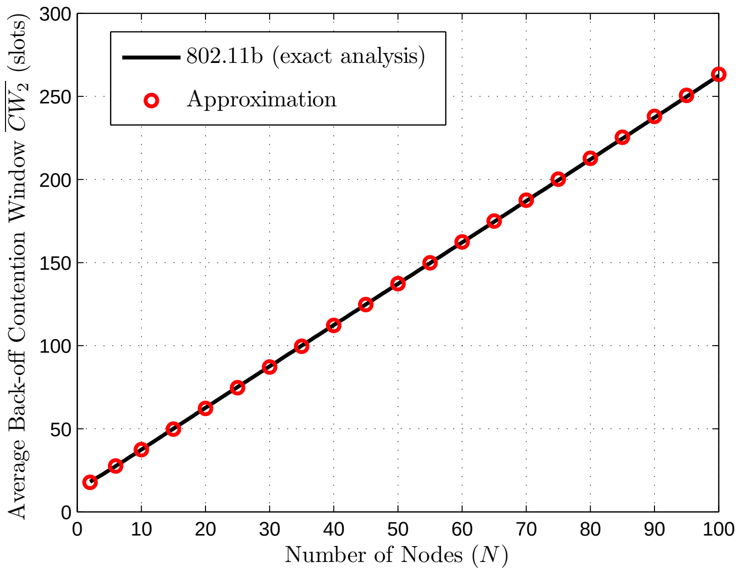


Figure 2.3: An approximation of average back-off contention window $\overline{CW_2}$.

pression of S_1, as a function of N, given by

$$S_1(N) = \frac{NT_{pl}}{NT_s + \frac{N}{2}\frac{a_1+a_2\ln(N)}{1-[a_1+a_2\ln(N)]}T_c + b_1 + b_2\exp(b_3a_1)\exp[b_3a_2\ln(N)]}. \tag{2.8}$$

where T_s, T_{pl}, and T_c are known parameters specified in the IEEE 802.11b standard. Fig. 2.4 demonstrates that the simplified analytical function $S_1(N)$ is an accurate approximation of the numerical performance model [57] [62] represented by the nonlinear system of (2.2) and (2.3).

In the traffic saturation case, the average packet access delay (average packet service time), D_1, is defined as the duration from the instant that the packet becomes the head of the transmission queue to the instant that the packet is successfully transmitted, averaged over all transmitted packets of each node. Since D_1 is the denominator of S_1 [57], it can also be approximated by a closed-form analytical function of N, given by

$$D_1(N) = NT_s + \frac{N}{2}\frac{a_1+a_2\ln(N)}{1-[a_1+a_2\ln(N)]}T_c + b_1 + b_2\exp(b_3a_1)\exp[b_3a_2\ln(N)]. \tag{2.9}$$

Performance analysis in a traffic non-saturation case: When the network is non-saturated, the average packet arrival rate λ of each node should not exceed its service

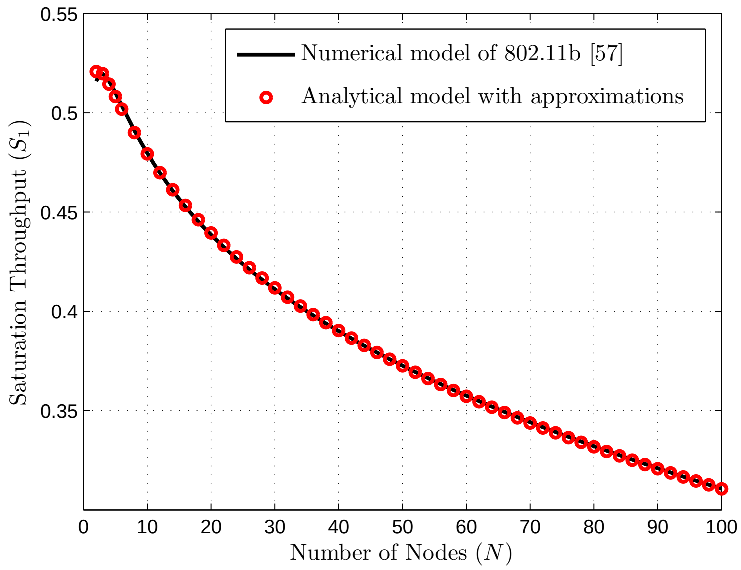


Figure 2.4: Saturation throughput S_1 and its approximation.

capacity share μ_d. The packet queue at each node possibly becomes empty upon successful transmission of the previous packet. The derivation of the packet transmission probability should account for the fact that a node attempting a transmission only when it has packets to transmit. Thus, Eq. (2.3) should be revised to

$$p = 1 - (1 - \rho \cdot \tau)^{N-1} \tag{2.10}$$

where $\rho = \frac{\lambda}{\mu_d}$ is the queue utilization ratio of an individual node, and $\rho\tau$ is the packet transmission probability of each node.

Due to its fairness property, the IEEE 802.11 system can be viewed as a server that schedules its resources to the contending nodes in a round robin manner [57]. In each scheduling cycle, every node (out of N nodes) can occupy an average fraction of $\frac{1}{N}$ system bandwidth to successfully transmit one packet. This service system is called processor sharing (PS) system. Thus, the IEEE 802.11 DCF can be modeled as an M/G/1/PS system with cumulative arrival rate $\lambda_s = N\lambda$ and service rate $\mu_s = N\mu_d$. According to [57], this M/G/1/PS system has the same access delay and queueing delay as the M/M/1 queueing system with equivalent average arrival rate λ_s and service rate μ_s. Thus, the average packet delay in the M/G/1/PS system is a summation of average packet access delay and average packet queueing delay (i.e., the duration from the instant that the packet arrives at the transmission queue to the instant that the packet becomes the queue head averaged over all transmitted packets of each node), given by

$$D_T = \frac{1}{\mu_s - \lambda_s} \tag{2.11}$$

where $\mu_s = \left[T_s + \frac{\overline{T_c}}{2} + \frac{\overline{CW_2}}{N}\right]^{-1}$ [57].

Similar to the traffic saturation case, since p and N are correlated as in the high-degree nonlinear relation, (2.10) and (2.11) form a nonlinear system with two variables p and N that can be solved using numerical techniques [28] [57]. To get a simplified and closed-form performance expression as a function of N, one approach is to obtain an explicit relation between p and N. A first-order approximation of (2.10) and linearizing the transmission probability as $\tau \approx \frac{2CW}{(CW+1)^2}(1-p)$ [66] can be applied to simplify (2.10) to a quadratic equation of p, given by

$$\begin{aligned} p &\approx (N-1)\lambda\left[N\left(T_s + \frac{\overline{T_c}}{2}\right) + \overline{CW_2}\right]\tau \\ &\approx (N-1)\lambda\left[\frac{2CWNT_s}{(CW+1)^2}(1-p) + \frac{NT_cCW}{(CW+1)^2}p + \frac{1}{1-p}\right], \quad (\tau, p \ll 1). \end{aligned} \tag{2.12}$$

Then, with packet arrival rate λ, a closed-form relation between p and N can be established by solving the quadratic equation of p. However, this first-order approximation is accurate only when p and τ are much less than one for a small value of N, as shown in Fig. 2.5(a) - 2.5(b). To have a more accurate approximation, we solve (2.10) for p over a wide range of N under the condition that all nodes are traffic non-saturated. It is found out that, with different values of λ, linearizing p as a function of N better characterizes the relation between p and N. Thus, a *linear least-squares curve-fitting method* is used to find a closed-form linear function between p and N, denoted by $p(N, \lambda)$ as an approximation of (2.10), shown in Fig. 2.5(a) - 2.5(b). Substituting $p(N, \lambda)$ into $\overline{T_c}$ in (2.1) yields a closed-form function $\overline{T_c(N, \lambda)} = \frac{p(N,\lambda)}{1-p(N,\lambda)}T_c$.

Since p shows a near linear relation with N for different values of λ and $\overline{CW_2}$ is approximately an exponential function of p in (2.6), $\overline{CW_2}$ can be approximately represented as an exponential function of N, denoted by $\overline{CW_2(N, \lambda)}$. Therefore, a closed-form expression for average packet delay D_T in terms of N is obtained as

$$D_2(N, \lambda) = \frac{1}{\mu_s(N, \lambda) - N\lambda} \tag{2.13}$$

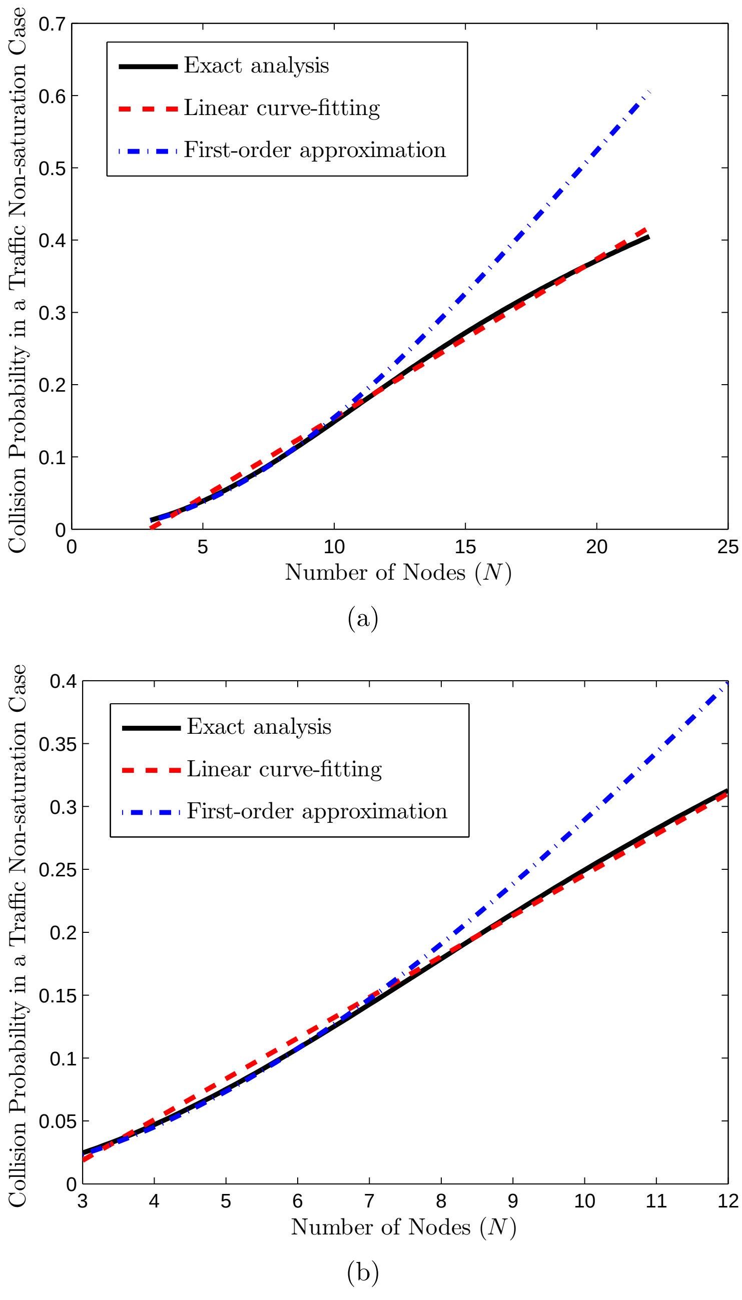


Figure 2.5: Approximations for collision probability in a traffic non-saturation case. (a) $\lambda = 25$ packet/s. (b) $\lambda = 50$ packet/s.

where $\mu_s(N,\lambda) = \left[T_s + \frac{\overline{T_c(N,\lambda)}}{2} + \frac{\overline{CW_2(N,\lambda)}}{N}\right]^{-1}$ is a closed-form expression for μ_s.

Similarly, the non-saturated network throughput, with the general form in (2.1), has the approximate and closed-form expression given as

$$S_2(N,\lambda) = \frac{NT_{pl}}{N\left(T_s + \frac{\overline{T_c(N,\lambda)}}{2}\right) + \overline{CW_2(N,\lambda)} + \left[1 - \frac{\lambda}{\mu_s(N,\lambda)}\right]\left[1 - (N-1)\frac{\lambda}{\mu_s(N,\lambda)}\right]\left[\frac{1}{\lambda} - D_2(N,\lambda)\right]}. \tag{2.14}$$

Fig. 2.6 and Fig. 2.7 show the exact values of average packet delay and non-saturation throughput as well as their accurate approximations over a wide range of N. It can be seen that, for λ equal to 25 and 50 packet/s, the traffic of each node enters the saturation state when N increases to the values greater than 22 and 12 respectively.

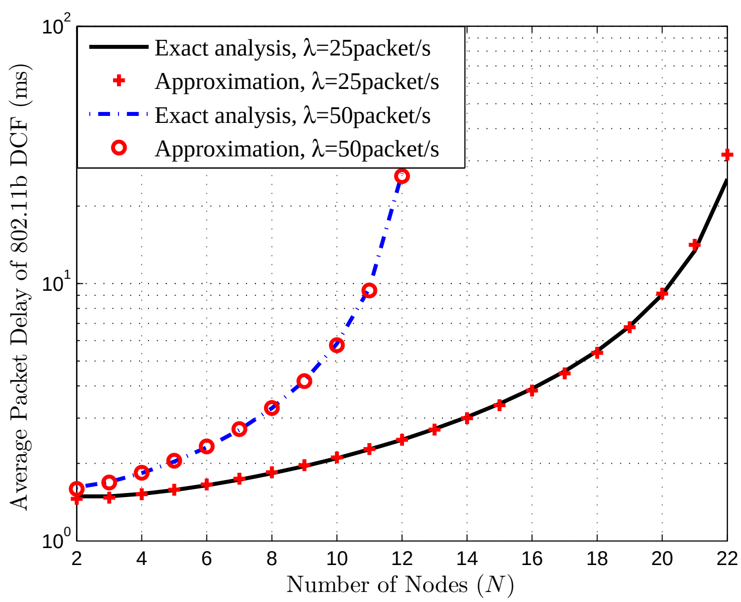


Figure 2.6: Average packet delay of IEEE 802.11 DCF and its approximation for $\lambda = 25$ and 50 packet/s respectively.

2.3.2 Closed-form performance models for D-TDMA

Performance analysis in a traffic saturation case: When the network is saturated, we can obtain closed-form expressions of throughput and delay as a function of N. Since N in

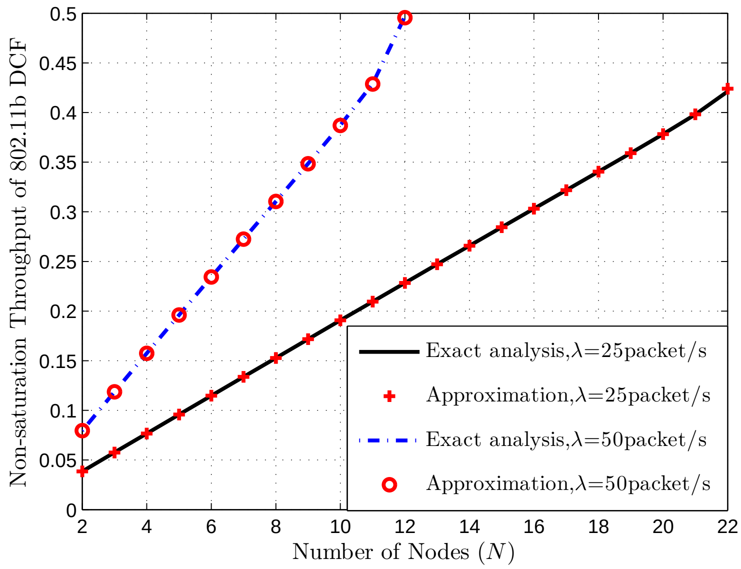


Figure 2.7: Non-saturation throughput of IEEE 802.11 DCF and its approximation for $\lambda = 25$ and 50 packet/s respectively.

general varies slowly with respect to the frame duration, the network saturation throughput S_3 is approximately given by

$$S_3(N) = \frac{NT_{pl}}{NT_p + M_m T_m} \tag{2.15}$$

where T_{pl} is the duration of payload information of each data packet, T_p is the data packet duration including headers, M_m denotes the number of minislots in the control period of each D-TDMA frame, and T_m is the duration of each minislot.

Also, the average packet access delay of D-TDMA, denoted by D_3, can be expressed as

$$D_3(N) = NT_p + M_m T_m. \tag{2.16}$$

Performance analysis in a traffic non-saturation case: In order to simplify the analysis of packet access delay and queueing delay, denoted by W_{st} and W_{qt} respectively, we assume that nodes release their data slots and randomly acquire new ones in the next frame, after transmitting a packet in the data transmission period of current frame [67]. This assumption guarantees that the service times of successive packets are i.i.d. random variables. Based on this assumption, the queue of each node in the traffic non-saturation case can be modeled as an M/G/1 queueing system [67], with an average service rate denoted by

μ_t packet/s. We derive the distribution of packet service time W_{st} to calculate the average packet access delay, $E[W_{st}]$, in the M/G/1 system. Then, the P-K formula [68] can be used to calculate the average packet queueing delay, $E[W_{qt}]$, for each M/G/1 queue, based on the second moment of W_{st}, denoted by $E[W_{st}^2]$. As a result, the average packet delay D_4, which is the summation of $E[W_{st}]$ and $E[W_{qt}]$ (see Appendix 2.7.2 for the derivation of $E[W_{st}]$ and $E[W_{qt}]$), is given by

$$D_4 = E[W_{st}] + \frac{\lambda E[W_{st}^2]}{2[1-\lambda E[W_{st}]]}. \tag{2.17}$$

Since $E[W_{st}]$ and $E[W_{st}^2]$ are both functions of N, D_4 is also a closed-form function of N, denoted by $D_4(N, \lambda)$.

As to the non-saturation throughput analysis, the probability that the queue of a tagged node is non-empty at the start of its designated time slot, denoted by P_{qn}, is given by

$$P_{qn} = \frac{\lambda}{\mu_t} \tag{2.18}$$

where $\mu_t = \frac{1}{E[W_{st}]} = \frac{2-\lambda(M_c+N-1)T_p}{(M_c+N+1)T_p}$, M_c denotes the duration of each control period normalized to the unit of one D-TDMA data slot duration, according to the delay analysis in Appendix 2.7.2.

Thus, we use random variable X to denote the number of nodes with non-empty queues at the start of their designated time slots during the time of one frame. The probability mass function (PMF) and the average of random variable X are given by [27]

$$P\{X = k\} = \binom{N}{k}\left(\frac{\lambda}{\mu_t}\right)^k\left(1-\frac{\lambda}{\mu_t}\right)^{N-k},\ k = 0, 1, ..., N; \tag{2.19}$$

$$E\left[X\right] = N \cdot \frac{\lambda}{\mu_t}. \tag{2.20}$$

Hence, the network non-saturation throughput S_4 can be approximated as a function of N,

$$S_4(N, \lambda) = \frac{N\lambda T_{pl}}{\mu_t(NT_p + M_m T_m)}. \tag{2.21}$$

In summary, we derive simplified and closed-form throughput and delay expressions $S_1(N)$, $S_2(N, \lambda)$, $D_1(N)$, $D_2(N, \lambda)$ for the IEEE 802.11 DCF, and $S_3(N)$, $S_4(N, \lambda)$, $D_3(N)$,

$D_4(N, \lambda)$ for D-TDMA, respectively, for both traffic saturation and non-saturation cases. The expressions can greatly simplify the MAC switching point calculation.

2.4 Adaptive MAC Solution

In this section, we present a MAC protocol which adapts to the changing traffic load in the MANET. The key element is to determine the MAC switching point, with which an appropriate candidate MAC protocol is selected to achieve better performance in terms of throughput and delay at each specific network traffic load condition. Based on the closed-form expressions derived in Section 2.3, we establish a unified performance analysis framework to evaluate the throughput and delay over a wide range of N for both non-saturated and saturated network traffic conditions. Taking throughput evaluation as an example, in this framework, when N is small, the network is non-saturated and the throughput is represented analytically by $S_2(N, \lambda)$ and $S_4(N, \lambda)$ for IEEE 802.11 DCF and D-TDMA, respectively. With an increase of N, packet service rates μ_d and μ_t of each node with both MAC protocols decrease consistently, making the queue utilization ratio of each node approach to one. After a specific network load saturation point in terms of N, say N_1 (N_2), where the arrival rate λ equals the service rate μ_d (μ_t), the network operating in IEEE 802.11 DCF (D-TDMA) enters the traffic saturation state. Thus, $S_1(N)$ and $S_3(N)$ are used to represent the network saturation throughput for each MAC candidate, respectively.

With this unified framework, performance comparison between the MAC candidates, with respect to N, can be conducted to calculate the MAC switching point. However, since IEEE 802.11 DCF and D-TDMA have different service capacity, the saturation points N_1 and N_2 are in general different, depending on λ. Therefore, the MAC switching point can be a specific network traffic load point, where the network with either IEEE 802.11 DCF or D-TDMA has four possible traffic state combinations: 1) the network is in the traffic saturation state with both MAC candidates; 2) the network is in the traffic non-saturation state with both MAC candidates; 3) the network is traffic saturated with IEEE 802.11 DCF and traffic non-saturated with D-TDMA; 4) the network is traffic non-saturated with IEEE 802.11 DCF and traffic saturated with D-TDMA. The established unified closed-form expressions facilitate performance comparison and the calculation of MAC switching point denoted by N_s (in terms of the number of nodes), for the four possible cases. The MAC switching point may vary, due to variations of λ at each node, in the homogeneous network traffic scenario. Algorithm 1 presents the detail steps of determining N_s. As an example, we illustrate step by step the switching point calculation for $\lambda = 25$ and

50 packet/s, respectively, based on network throughput comparison. Then, the complete MAC switching point calculation algorithm is provided considering all the possible cases.

1) $\lambda = 25$ packet/s:

Step 1. Compare the saturation points, N_1 and N_2, for both MAC candidates, $N_1 < N_2$;

Step 2. Compare the throughput of both MAC candidates at N_1, $S_1(N_1) > S_4(N_1, \lambda)$;

Step 3. Compare the throughput of both MAC candidates at N_2, $S_3(N_2) > S_1(N_2)$;

Step 4. The MAC switching point is calculated by solving equation $S_1(N) = S_4(N, \lambda)$, where the network has saturated traffic operating in IEEE 802.11 DCF and non-saturated traffic operating in D-TDMA.

2) $\lambda = 50$ packet/s:

Step 1. Compare the saturation points N_1 and N_2 for both MAC candidates, $N_1 = N_2 = N^*$;

Step 2. Compare the throughput of both MAC candidates at N^*, $S_1(N^*) = S_3(N^*)$;

Step 3. The MAC switching point is obtained as $N_s = N^*$, where the network has saturated traffic operating in either candidate MAC protocol.

The MAC switching point can also be determined based on comparison of average packet delay between the MAC candidates, which is expected to generate similar results since a higher throughput corresponds to a lower packet delay. In theory, the average packet delay can be evaluated only when the packet arrival rate is less than the service rate, where the network traffic is in the non-saturation state. Otherwise, the packet delay will theoretically approach infinity. Therefore, when the MAC switching point locates at an N value where the network is in a traffic saturation state with either candidate MAC protocol, the average packet access delay, $D_1(N)$ and $D_3(N)$, can be used to calculate the switching point.

Due to node mobility, the number of nodes, N, may fluctuate around the switching point when nodes move relatively fast, resulting in undesired frequent MAC switching (taking account of switching cost). In order to benefit from the MAC switching, the performance gain should be higher than the switching cost. Therefore, the MAC switching point can be replaced by a switching interval. The MAC switching is triggered only if the number of nodes, N, varies beyond the switching interval. The length of the switching interval depends on the performance gain and switching cost. The switching cost can be calculated as the communication overhead consumed for periodic control information exchange among nodes to acquire the updated network traffic load information for distributed

Algorithm 1: MAC switching point calculation algorithm

Input : The saturation points, N_1 and N_2, for IEEE 802.11 DCF and D-TDMA.
Output: The MAC switching point N_s.

if $N_1 < N_2$ **then**
    if $S_1(N_1) > S_4(N_1, \lambda)$ **then**
        if $S_1(N_2) < S_3(N_2)$ **then**
            $N_s \leftarrow$ *solving* $S_1(N) = S_4(N, \lambda)$;
        else
            $N_s \leftarrow$ *solving* $S_1(N) = S_3(N)$;
        end
    else if $S_1(N_1) < S_4(N_1, \lambda)$ **then**
        $N_s \leftarrow$ *solving* $S_2(N, \lambda) = S_4(N, \lambda)$;
    else
        $N_s \leftarrow N_1$;
    end
else if $N_1 > N_2$ **then**
    if $S_3(N_2) > S_2(N_2, \lambda)$ **then**
        $N_s \leftarrow$ *solving* $S_2(N, \lambda) = S_4(N, \lambda)$;
    else if $S_3(N_2) < S_2(N_2, \lambda)$ **then**
        if $S_3(N_1) > S_1(N_1)$ **then**
            $N_s \leftarrow$ *solving* $S_2(N, \lambda) = S_3(N)$;
        else
            $N_s \leftarrow$ *solving* $S_1(N) = S_3(N)$;
        end
    else
        $N_s \leftarrow N_2$;
    end
else
    if $S_1(N_1) \geq S_3(N_1)$ **then**
        $N_s \leftarrow$ *solving* $S_1(N) = S_3(N)$;
    else
        $N_s \leftarrow$ *solving* $S_2(N, \lambda) = S_4(N, \lambda)$;
    end
end

MAC switching. For a network with higher node mobility, nodes require more frequent control information exchange, resulting in an increased switching cost and a longer switching interval. Therefore, how to determine the optimal length of the switching interval to maximize the performance gain with the consideration of the switching cost and different node mobility patterns can be investigated in future research.

2.5 Numerical Results

In this section, we present analytical and simulation results for performance evaluation of both MAC candidates and the MAC switching point. The simulation results are used to demonstrate the accuracy of the MAC switching point calculation based on the closed-form expressions in Section 2.3. With an error-free wireless channel in the system, we use the network simulator, OMNeT++ [69] [70], to simulate the IEEE 802.11b DCF and the D-TDMA. In the simulation, a fully-connected network over a 50m $\times$ 50m square coverage area is deployed, where nodes are randomly scattered. Traffic arrivals for each node are realized as a Poisson process with λ being 25 and 50 packet/s, respectively, for the non-saturated traffic case, and with λ set as 500 packet/s for the saturated traffic case. The reason of using the same traffic model in computer simulations is to verify the accuracy of the analytical models proposed for the IEEE 802.11 DCF and the D-TDMA, since several assumptions and simplifications are made in the mathematical modeling and analysis, especially for traffic non-saturation conditions for both MAC schemes. For the IEEE 802.11 DCF, with homogeneous Poisson traffic arrivals, the 802.11 service system can be approximately modeled as an M/G/1/processor sharing (PS) system, for which the average packet delay analysis is greatly simplified; For the D-TDMA, to model the queue of each node in a traffic non-saturation condition as an M/G/1 queue, it is assumed that nodes release their data slots and randomly acquire new ones in the subsequent frame once the packet transmission is completed in current frame. Also, to simplify the derivation of the packet service time distribution, we normalize the control period of each D-TDMA frame to an integer multiple of one D-TDMA data slot duration and discretize the packet service time in the unit of one data slot, while neglecting the possibility of head-of-line (HOL) packets arriving within the duration of each data slot. These assumptions are necessary for a tractable analysis, and their effects on analysis accuracy should be evaluated by the simulations without the assumptions under the same traffic model. Each simulation point provides the average value of the corresponding performance metrics (i.e., throughput and delay). We also plot the 95% confidence intevals for each simulation result. Note that some confidence intervals are very small in the figures. Other main simulation parameters are summarized in Table 2.1.

Table 2.1: Simulation parameters used in IEEE 802.11b [1] and D-TDMA

Parameters \ MAC schemes	**IEEE 802.11b**	**D-TDMA**
Channel capacity	11Mbps	11Mbps
Basic rate	1Mbps	1Mbps
Back-off slot time	20 μs	—
Minimum contention window size (CW)	32	—
Maximum contention window size (W_m)	1024	—
Retransmission limit (M_L)	7	—
Guard time (GT) [19]	—	1 μs
PLCP & Preamble	192 μs	192 μs
MAC header duration	24.7 μs	24.7 μs
Packet payload duration (T_{pl})	$\frac{8184}{11}$ μs	$\frac{8184}{11}$ μs
Short interframe space (SIFS)	10 μs	—
ACK	10.2 μs	—
Distributed interframe space (DIFS)	50 μs	—
Minislot duration (T_m)	—	219.4 μs
Network size upper limit (M_m)	15/25/35 nodes	15/25/35 nodes
Queue length	10000 packets	10000 packets

2.5.1 Traffic saturation case

First, the saturation throughputs of both MAC candidates are plotted in Fig. 2.8(a) - 2.8(c), for $M_m = 15, 25, 35$, respectively. It is observed that the analytical and simulation results closely agree with each other. As M_m increases, the saturation throughput of D-TDMA decreases since the length of control period in each D-TDMA frame increases, which reduces the channel utilization. The two MAC candidates have near opposite throughput variation trends as the network traffic load increases. For IEEE 802.11 DCF, the saturation throughput decreases with an increase of the traffic load. On the other hand, the saturation throughput of D-TDMA experiences a consistent rise when the number of nodes increases. Therefore, the two throughput curves intersect at a specific network traffic load value, for example $N = 12.5$ when $M_m = 35$. Before this value, IEEE 802.11 DCF outperforms D-TDMA and, after this value, the D-TDMA performs better. Thus, the MAC switching point is the first integer number of nodes after the intersection, i.e., $N_s = 13$.

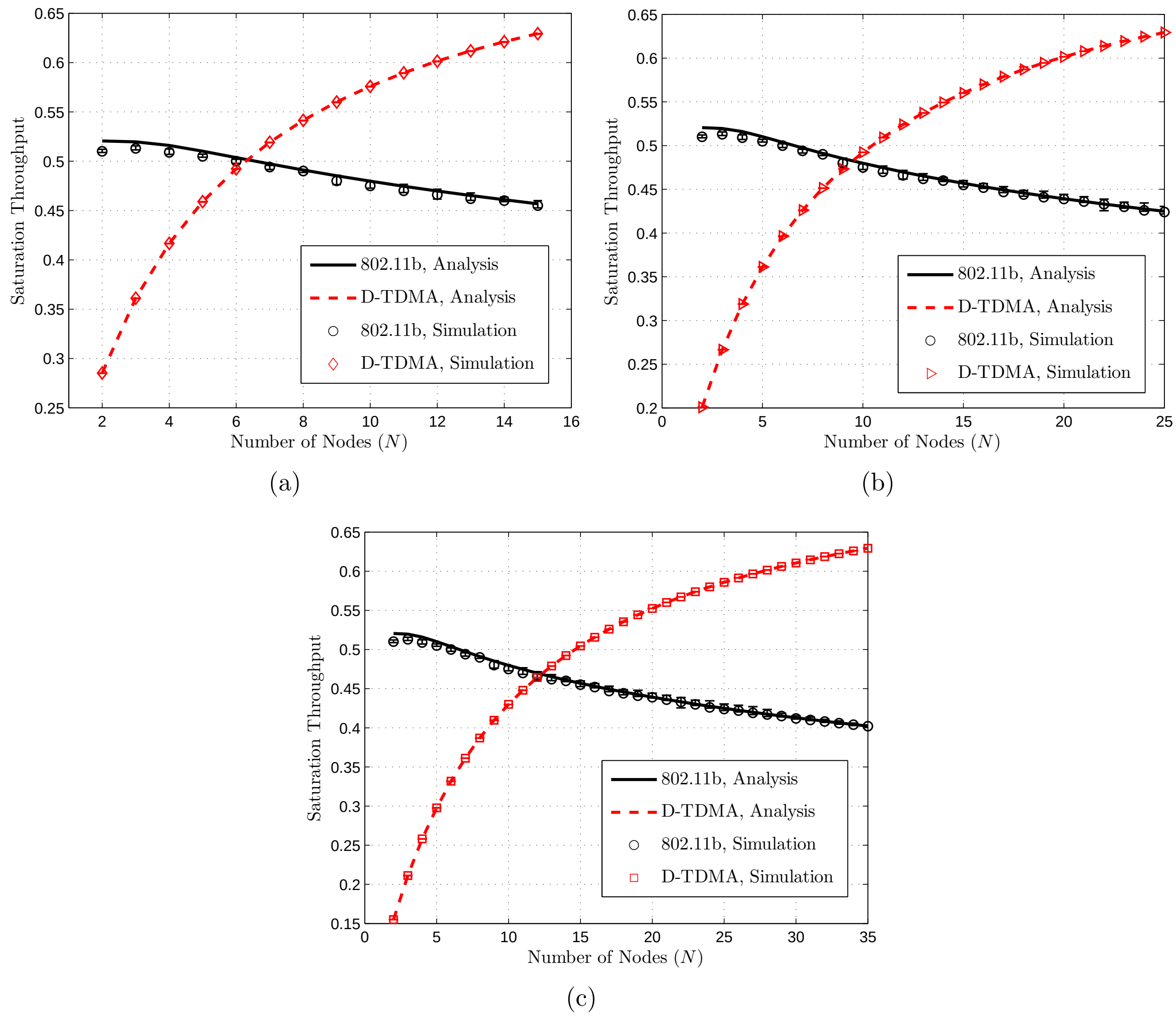


Figure 2.8: Saturation throughput of both MAC schemes. (a) $M_m = 15$. (b) $M_m = 25$. (c) $M_m = 35$.

The average packet access delay of both MAC candidates in a traffic saturation case are plotted in Fig. 2.9(a) - 2.9(c), for $M_m = 15, 25, 35$, respectively. It is observed that the MAC switching point is almost the same as that based on the saturation throughput.

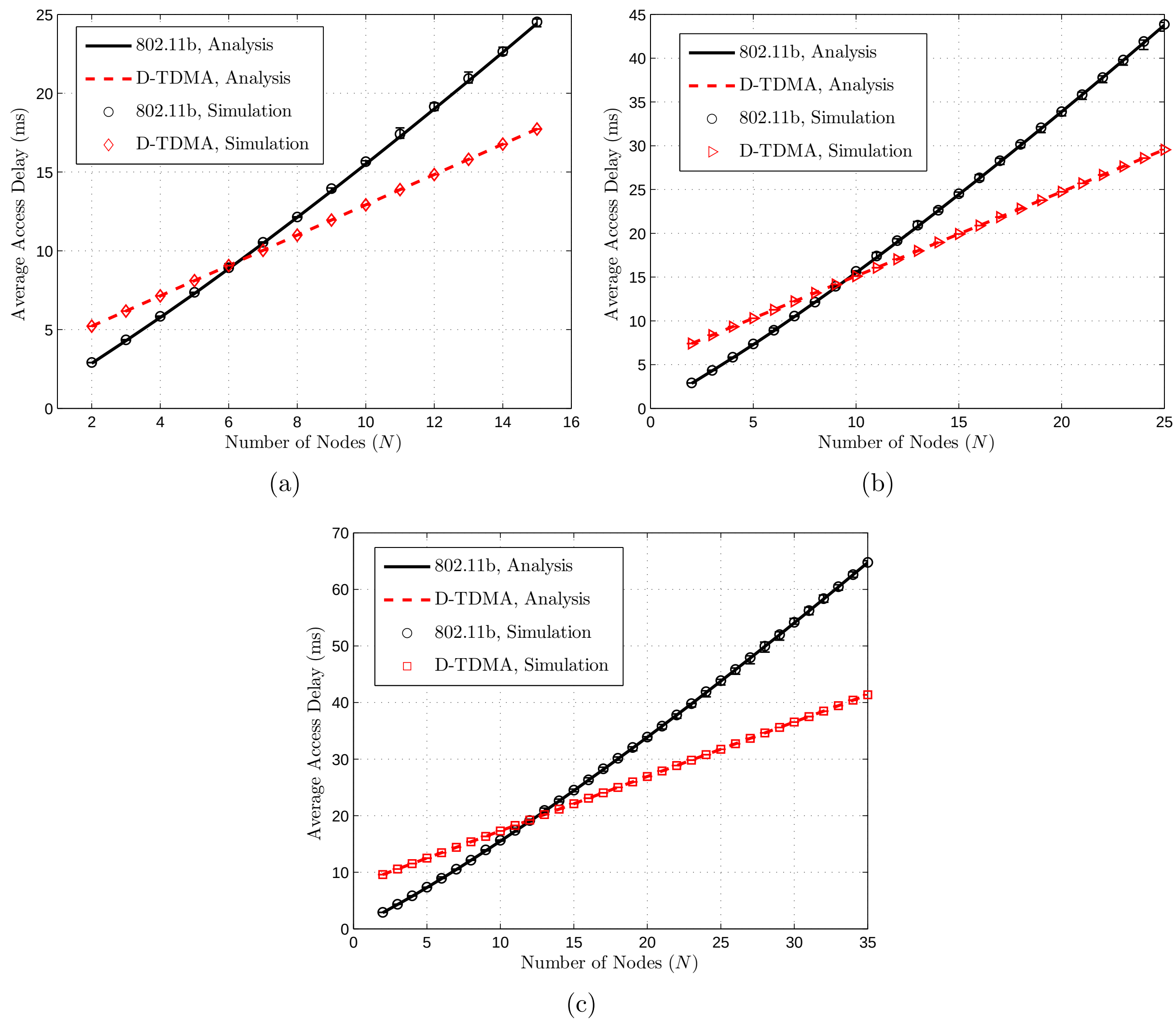


Figure 2.9: Average packet access delay of both MAC schemes. (a) $M_m = 15$. (b) $M_m = 25$. (c) $M_m = 35$.

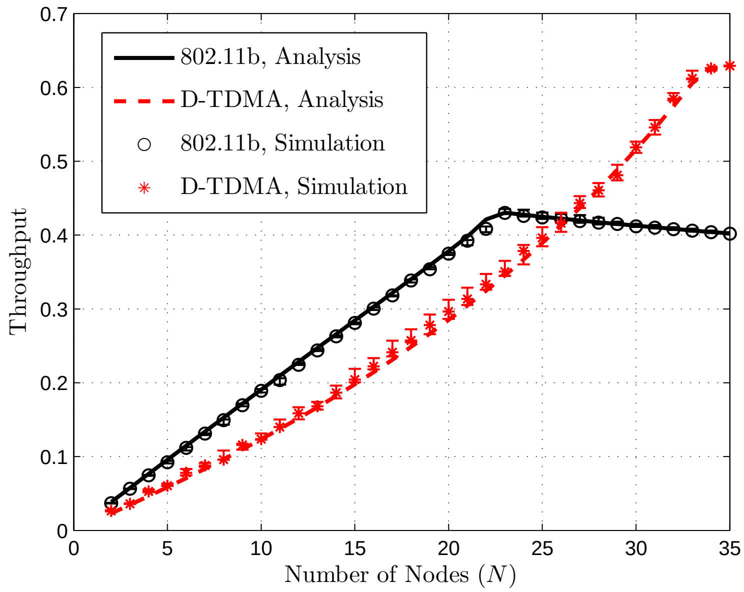
2.5.2 Traffic non-saturation case

Fig. 2.10(a) - 2.10(b) show how the network throughput changes with the number of nodes for both MAC schemes at $\lambda = 25$ and 50 packet/s, respectively. Again, the analytical results closely match the simulation results. In the simulation, we start from $N = 2$ where each node has a non-saturated traffic for both MAC candidates, and gradually increase the N value to $N = 35$. As N increases, the service rate for each node decreases, and the traffic at each node becomes saturated after N increases to a certain value. For IEEE 802.11 DCF, the saturation point locates at $N_1 = 23$ and 13 for λ equal to 25 packet/s and 50 packet/s, respectively. On the other hand, the corresponding saturation point of D-TDMA is $N_2 = 33$ and 13, respectively. When the traffic load is low, the non-saturation network throughput of IEEE 802.11 DCF is greater than that of D-TDMA and, therefore, nodes should choose IEEE 802.11 DCF as the initial MAC scheme. For $\lambda = 25$ packet/s, the MAC switching point is $N_s = 26$, where nodes with IEEE 802.11 DCF are traffic saturated. For $\lambda = 50$ packet/s, the switching point appears at $N_s = 13$, from which nodes with either MAC scheme have saturated traffic.

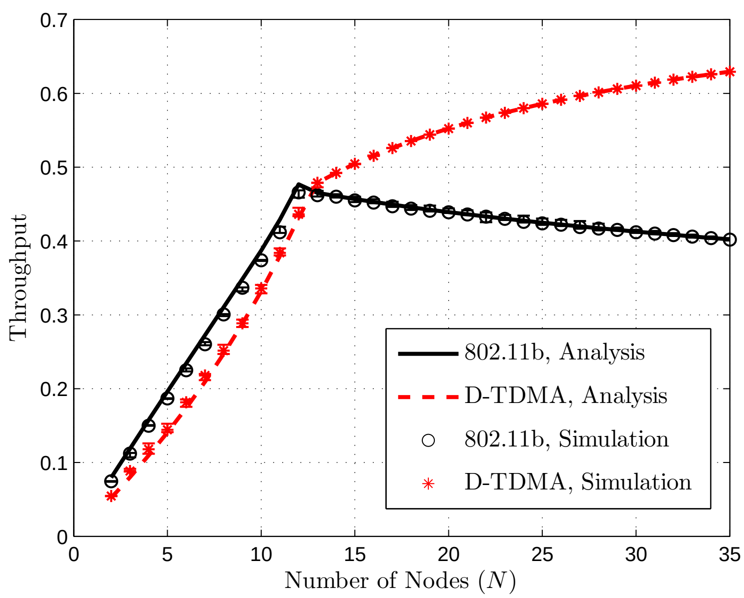
Fig. 2.11(a) - 2.11(b) show the average packet delay for both MAC candidates in a traffic non-saturation case with $\lambda = 25$ and 50 packet/s respectively, for N varying from 2 to the largest integer within traffic non-saturation load region. We can see that the analytical results closely match the simulation results and the confidence intervals are very small. For $\lambda = 25$ packet/s, the two delay curves are expected to intersect at the network load point where IEEE 802.11 DCF becomes traffic saturated and D-TDMA is still traffic non-saturated. Thus, the MAC switching point exists as the saturation point of IEEE 802.11 DCF, denoted as $N_s = 23$. For $\lambda = 50$ packet/s, the two delay curves do not intersect in the traffic non-saturation state. Thus, the MAC switching point is expected to exist at a traffic saturated load point greater than $N = 12$, which can be obtained analytically as $N_s = 13$ based on comparison of average packet access delay for both MAC candidates, shown in Fig. 2.9(c). The MAC switching point based on packet delay comparison is almost the same as that based on throughput comparison.

2.6 Summary

In this chapter, an adaptive MAC solution is proposed based on a MAC switching point calculation between the IEEE 802.11 DCF and the D-TDMA. Novel analytical models for throughput and delay in saturated and unsaturated traffic load conditions are developed for both MAC schemes in closed-form expressions of the total number of nodes in the network, which facilitate the distributed calculation of the MAC switching point. The adaptive MAC solution switches between the contention-based MAC and the reservation-based MAC based on the MAC switching point, which provides a way to improve the network performance over traffic load variations. In Chapter 3, we extend the adaptive MAC solution to support heterogeneous services (i.e., both realtime voice and non-realtime best-effort data applications) in a fully-connected MANET.

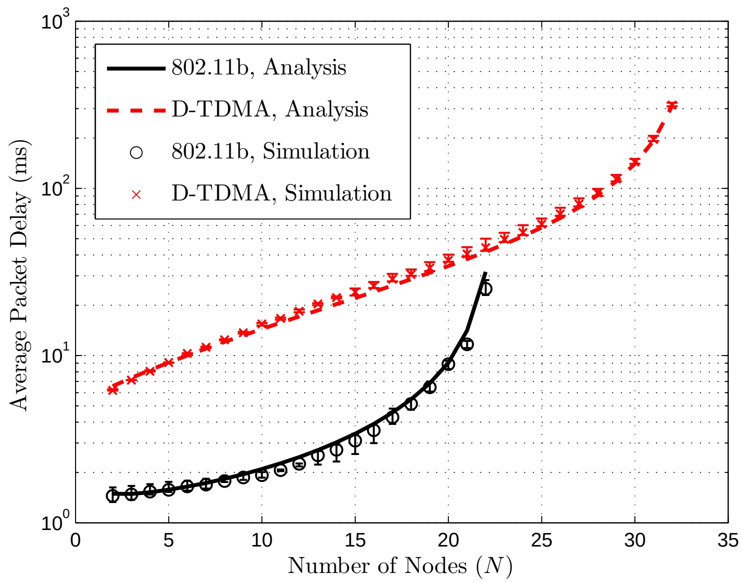


(a)

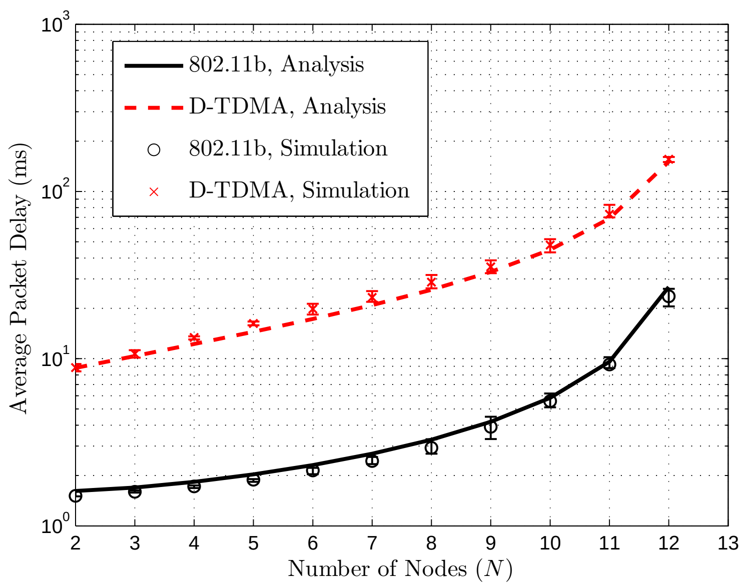


(b)

Figure 2.10: Network throughput versus the number of nodes. (a) $\lambda = 25$ packet/s. (b) $\lambda = 50$ packet/s.



(a)



(b)

Figure 2.11: Average packet delay versus the number of nodes. (a) $\lambda = 25$ packet/s. (b) $\lambda = 50$ packet/s.

2.7 Appendix

2.7.1 Proof of Proposition 1

The objective function of the logarithmic nonlinear least-squares curve-fitting problem can be written as

$$||a_1 + a_2 \ln(\mathbf{N}) - \mathbf{P}||_2^2 = \sum_{n=2}^{N} (a_1 + a_2 \ln(n) - p_n)^2 = \sum_{n=2}^{N} f_n^2(\mathbf{a}). \tag{2.22}$$

Then, $\forall\, \mathbf{a} \in \mathbf{dom}\, f_n$, we calculate the Hessian matrix of $f_n^2(\mathbf{a})$ as follows:

$$\mathbf{H}(f_n^2(\mathbf{a})) = \begin{bmatrix} \frac{\partial f_n^2(\mathbf{a})}{\partial a_1^2} & \frac{\partial f_n^2(\mathbf{a})}{\partial a_1 \partial a_2} \\ \frac{\partial f_n^2(\mathbf{a})}{\partial a_2 \partial a_1} & \frac{\partial f_n^2(\mathbf{a})}{\partial a_2^2} \end{bmatrix} = \begin{bmatrix} 2 & 2\ln(n) \\ 2\ln(n) & 2\ln^2(n) \end{bmatrix}. \tag{2.23}$$

The eigenvalues of the Hessian matrix can be derived by solving the eigenfunction of $\mathbf{H}(f_n^2)$

$$\begin{aligned} &\mathbf{det}(\lambda \mathbf{I} - \mathbf{H}(f_n^2)) = \begin{vmatrix} \lambda - 2 & -2\ln(n) \\ -2\ln(n) & \lambda - 2\ln^2(n) \end{vmatrix} = 0 \\ &\Longrightarrow \lambda_1 = 0, \lambda_2 = 2 + 2\ln^2(n). \end{aligned} \tag{2.24}$$

Because both eigenvalues of $\mathbf{H}(f_n^2)$ are nonnegative, the Hessian matrix $\mathbf{H}(f_n^2)$ is semidefinite. On the other hand, since $\mathbf{dom}\, f_n = \{(a_1, a_2) \,|\, a_2 \geq 0\}$ is a convex set, $f_n^2(\mathbf{a})$ is a convex function for all $\mathbf{a} \in \mathbf{dom}\, f_n$.

Hence, the objective function $\sum_{n=2}^{N} f_n^2(\mathbf{a})$ is a nonnegative sum of convex functions $f_n^2(\mathbf{a})\, (n = 2, 3, ..., N)$, which is also convex [65]. That is, the curve-fitting is a convex optimization problem.

2.7.2 Derivation of $E[W_{st}]$ and $E[W_{qt}]$

In order to calculate the average packet delay of the M/G/1 queue of a tagged node, we first derive the probability distribution of packet service time W_{st}. For analysis simplicity, we normalize the control period of each frame as an integer multiple of one D-TDMA data slot duration T_p, i.e., $M_c = \lceil \frac{M_m T_m}{T_p} \rceil$, where $\lceil \rceil$ is the ceiling function. The end instant of each slot along one D-TDMA frame is numbered from 1 to $M_c + N$, as shown in Fig. 2.12. Let random variable J denote the arriving instant of each head-of-line (HOL) packet. It

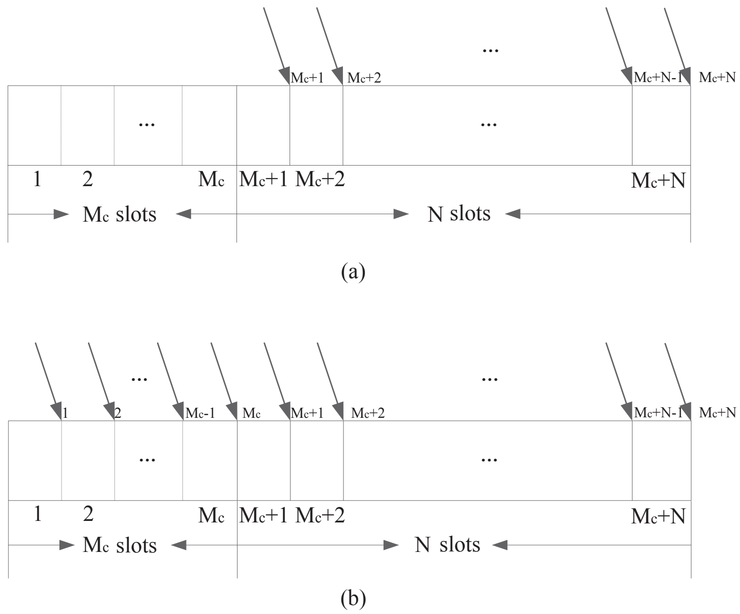


Figure 2.12: The HOL packets arrival patterns within one frame. (a) The node's queue is non-empty. (b) The node's queue is empty.

is assumed that HOL packets of the tagged node only appear at the end of each time slot, neglecting the possibility that HOL packets can arrive within the duration of each time slot [67], which means J takes discrete values from set $\mathbf{A} = \{1, 2, ..., M_c + N\}$. From Fig. 2.12, it can be seen that HOL packets have two different arriving patterns according to current status of the queue: (a) when the node's queue is non-empty (i.e., at least one packet staying in the queueing system), HOL packets can only appear at the end of its designated data slot in the data transmission period, which means J takes values from set $\mathbf{A}' = \{M_c + 1, M_c + 2, ..., M_c + N\}$; (b) when the node's queue is empty (i.e., no packets are in service.), HOL packets can arrive at any time instant in set $\mathbf{A}$. Next, we derive the distribution of W_{st} under these two cases.

When the node's queue is non-empty, based on the assumption that data slot is randomly selected for each node in the next frame upon the successful packet transmission in the current frame, we use random variable I, which takes values from set $\mathbf{B} = \{1, 2, ..., N\}$, to denote the data slot number that the node selects in the next frame. Thus, the probability distribution of the packet service time W_{st} in the unit of one data slot duration, denoted by W_s, is derived as

$$\begin{aligned}
&P\{W_s = M_c + k\} = \sum_{j\in\mathbf{A}',i\in\mathbf{B}} P\{W_s = M_c + k, J = j, I = i\} \\
&= \sum_{j\in\mathbf{A}',i\in\mathbf{B}} P\{W_s = M_c + k|J = j, I = i\}P\{J = j\}P\{I = i\} \\
&= \frac{k}{N^2} \quad (1 \le k \le N); \\
&P\{W_s = M_c + N + k\} = \frac{N-k}{N^2} \quad (1 \le k \le N-1).
\end{aligned} \tag{2.25}$$

When the node's queue is empty, HOL packets can arrive at any time instant in set $\mathbf{A}$. Thus, the probability distribution of W_s is derived in the following two cases:

(i) If $M_c \ge N$,

$$\begin{aligned}
P\{W_s = k\} &= \sum_{j=M_c-k+1}^{M_c-k+N} P\{W_s = k|J = j\}P\{J = j\} = \frac{1}{M_c+N} \quad (1 \le k \le N-1) \\
P\{W_s = M_c - k\} &= \sum_{j=k+1}^{k+N} P\{W_s = M_c - k|J = j\}P\{J = j\} = \frac{1}{M_c+N} \quad (0 \le k \le M_c - N) \\
P\{W_s = M_c + k\} &= \sum_{j=1}^{N-k} \sum_{j=M_c+N-k+1}^{M_c+N} P\{W_s = M_c + k|J = j\}P\{J = j\} = \frac{1}{M_c+N} \quad (1 \le k \le N);
\end{aligned} \tag{2.26}$$

(ii) If $M < N$,

$$\begin{aligned}
P\{W_s = k\} &= \sum_{j=M_c-k+1}^{M_c-k+N} P\{W_s = k|J = j\}P\{J = j\} = \frac{1}{M_c+N} \quad (1 \le k \le M_c) \\
P\{W_s = M_c + k\} &= \sum_{j=1}^{N-k} \sum_{j=M_c+N-k+1}^{M_c+N} P\{W_s = M_c + k|J = j\}P\{J = j\} = \frac{1}{M_c+N} \quad (1 \le k \le N).
\end{aligned} \tag{2.27}$$

Hence, the average service time, $E[W_{st}]$, and the second moment of service time, $E[W_{st}^2]$, are derived as follows:

$$
\begin{aligned}
E[W_{st}] &= P_{qn} \cdot \sum_{k_1 \in C} k_1 T_p P\{W_s = k_1\} + P_{qe} \cdot \sum_{k_2 \in D} k_2 T_p P\{W_s = k_2\} \\
&= \lambda E[W_{st}] \cdot \sum_{k_1 \in C} k_1 T_p P\{W_s = k_1\} + (1 - \lambda E[W_{st}]) \cdot \sum_{k_2 \in D} k_2 T_p P\{W_s = k_2\} \\
&\Longrightarrow E[W_{st}] = \frac{(M_c + N + 1)T_p}{2 - \lambda(M_c + N - 1)T_p};
\end{aligned} \tag{2.28}
$$

$$
\begin{aligned}
E[W_{st}^2] = \frac{(2M_c + 2N + 1)(M_c + N + 1)T_p^2}{6} + T_p^2 \lambda E[W_{st}] \cdot \\
\left[(M_c + N)^2 + \frac{N^2 - 1}{6} - \frac{(2M_c + 2N + 1)(M_c + N + 1)}{6} \right]
\end{aligned} \tag{2.29}
$$

where P_{qe} is the queue empty probability; **C** and **D** are two sets of possible values of W_s for the queue non-empty and queue empty cases, respectively.

Chapter 3

Distributed and Adaptive Hybrid MAC for IoT-Enabled MANETs

In this chapter, we propose a distributed and adaptive hybrid MAC (DAH-MAC) scheme for a single-hop IoT-enabled MANET supporting voice and data services. A hybrid superframe structure is designed to accommodate packet transmissions from a varying number of mobile nodes generating either delay-sensitive voice traffic or best-effort data traffic. Within each superframe, voice nodes with packets to transmit access the channel in a contention-free period using distributed TDMA, while data nodes contend for channel access in a contention period using truncated CSMA/CA (T-CSMA/CA). In the contention-free period, by adaptively allocating time slots according to instantaneous voice traffic load, the MAC exploits voice traffic multiplexing to increase the voice capacity. In the contention period, a throughput optimization framework is proposed for the DAH-MAC, which maximizes the aggregate data throughput by adjusting the optimal contention window size according to voice and data traffic load variations. Numerical results show that the proposed MAC scheme outperforms existing QoS-aware MAC schemes for voice and data traffic in the presence of heterogeneous traffic load dynamics.

3.1 System Model

Consider a single-channel fully connected MANET [41] [54] [56], where each node can receive packet transmissions from any other node. The fully connected network scenario can be found in various MANET applications, including office networking in a building or in a university library where users are restricted to move in certain geographical areas [54], users within close proximity are networked with ad hoc mode in a conference site [7], M2M communications in a residential network for a typical IoT-based smart home application where home appliances are normally within the communication range of each other [3]. The channel is assumed error-free, and packet collisions occur when more than one node simultaneously initiate packet transmission attempts. Without any network infrastructure or centralized controller, nodes exchange local information with each other and make their transmission decisions in a distributed manner. The network has two types of nodes, voice nodes and data nodes, generating delay-sensitive voice traffic and best-effort data traffic, respectively. Each node is identified by its MAC address and a unique node identifier (ID) that can be randomly selected and included in each transmitted packet [40]. We use N_v and N_d to denote the total numbers of voice and data nodes in the network coverage area, respectively. Nodes are mobile with a low speed, making N_v and N_d change with time.

For delay-sensitive voice traffic, each packet should be successfully transmitted within a bounded delay to achieve an acceptable voice communications quality; otherwise, the packet will be dropped. Therefore, as a main QoS metric for voice traffic, packet loss rate should be carefully controlled under a given threshold, denoted by P_L (e.g., 10^{-2}). The generic *on/off* characteristic of voice traffic allows traffic multiplexing in transmission. Each voice source node is represented by an *on/off* model, which is a two-state Markov process with the *on* and *off* states being the talk spurt and silent periods, respectively. Both periods are independent and exponentially distributed with respective mean $\frac{1}{\alpha}$ and $\frac{1}{\beta}$. During a talk spurt, voice packets are generated at a constant rate, λ_v packet/s. As for best-effort data traffic, data nodes are expected to exploit limited wireless resources to achieve as high as possible aggregate throughput. It is assumed that each data node always has packets to transmit. Nodes in the network are synchronized in time, which can be achieved such as by using the 1PPS signal with a global positioning system (GPS) receiver [40] [41].

In the network, time is partitioned into superframes of constant duration, denoted by T_{SF}, which is set to have the same duration as the delay bound of voice traffic. Each superframe is further divided into three periods: control period (CTP), contention-free period (CFP) and contention period (CP), the durations of which are denoted by T_{ctrl}, T_{cfp} and T_{cp} respectively, as shown in Fig. 3.1. The control period consists of N_{vm} fixed-duration (T_m) minislots, each with a unique minislot sequence number. It is to support a varying number of voice nodes in the network. Each voice node selects a unique minislot and broadcasts local information in its selected minislot, for distributed TDMA time slot allocation in the following contention-free period [19]. In the context of higher service priority to voice traffic, to avoid a complete deprivation of data service, there is a maximum

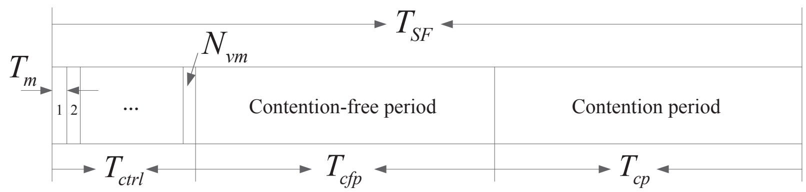


Figure 3.1: Superframe structure.

fraction of time, φ (< 1), for voice traffic in each superframe. The value of φ is assumed known to all nodes when the network operation starts, and can be broadcast by the existing nodes in each control period. The voice capacity is the maximum number of voice nodes allowed in the network, denoted by N_{vm} (same as the number of minislots in each CTP), under the QoS constraint, which depends on φ. The period following the control period is the CFP, which is further divided into multiple equal-duration TDMA time slots, each slot having a unique sequence number. Each voice node with packets to transmit (referred to as active voice node) occupies one time slot to transmit a number of voice packets, called a *voice burst*[1]. Thus, the number of TDMA slots in the CFP is determined by the number of voice burst transmissions scheduled for the superframe, denoted by N_s ($N_s \leq N_v$).

The last period CP is dedicated to best-effort data nodes for transmission according to T-CSMA/CA. Data packet transmissions are based on CSMA/CA and are periodically interrupted by the presence of CTP and CFP.

3.2 The DAH-MAC Scheme

In the following, we illustrate how voice nodes access their TDMA slots in each CFP without a central controller. For data nodes accessing the channel using T-CSMA/CA, we highlight differences between the T-CSMA/CA within the proposed hybrid superframe structure and the traditional CSMA/CA.

3.2.1 Accessing Minislots

In the distributed MAC, each voice node needs to exchange information with neighboring voice nodes by broadcasting control packets in the minislots in the control period of each

[1]A voice burst is the packets generated by an active voice node within one superframe that can be transmitted over a time slot.

superframe. When the network operation starts, the number of minislots (voice capacity N_{vm}) in the control period should be determined in a distributed way under the constraint that voice packet loss rate is bounded by P_L and the summation of T_{ctrl} and T_{cfp} does not exceed $\varphi \cdot T_{SF}$ in each superframe. After N_{vm} is determined, each voice node randomly chooses one minislot in the CTP of a superframe, and broadcasts a control packet in its selected minislot [19]. Each node broadcasts its control packet in the same occupied minislot of each subsequent superframe[2], until it is powered off or departs from the network. A control packet, shown in Fig. 3.2, includes five fields: a header, a set of IDs of the node's neighbors including the node itself, the node's occupied minislot sequence number (MSN, chosen from 1 to N_{vm}), buffer occupancy indication bit (BIB), the node's scheduled TDMA slot sequence number (SSN, a number within 0 to N_s) in the previous superframe.

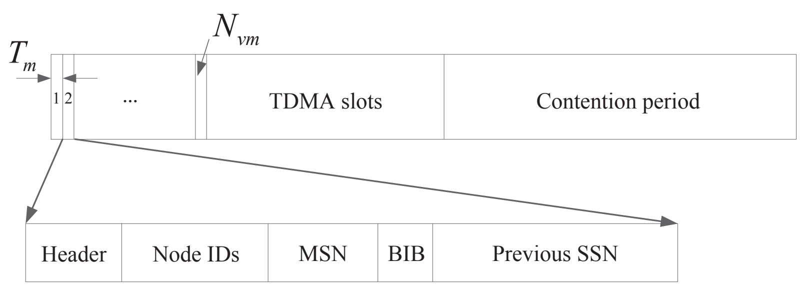


Figure 3.2: Format of control packet broadcast in each minislot.

Accessing a minislot from a tagged voice node is considered successful if the control packets received from other nodes in subsequent minislots contain the tagged node's ID [67]. Otherwise, an *access collision* happens due to simultaneous control packet transmissions in the same minislot by more than one node. All nodes involved in the collision wait until the next superframe to re-access one of the vacant minislots. The minislot accessing process is completed when all existing nodes successfully acquire their respective minislots. When a new node is powered on or entering the network coverage area, it first synchronizes in time with the start of a superframe, determines the number of minislots (based on φ), N_{vm}, and listens to all control packets in the CTP. Then, it randomly selects an unoccupied minislot and broadcasts a control packet in the minislot in the next superframe. If all N_{vm} minislots are occupied, which means the whole network reaches its voice capacity, the node defers its channel access and keeps sensing the CTPs of subsequent superframes until some existing minislots are released due to node departures. After the minislot accessing is successful, the node keeps using the same minislot of subsequent superframes to broadcast its control

[2]To ensure fair minislot access, voice nodes re-select minislots after using the previous ones for a predefined number of successive superframes [71].

packet.

3.2.2 Adaptive TDMA Time Slot Allocation

For efficient resource utilization, time slot allocation to voice nodes should adapt to traffic load variations. Taking account of the voice traffic *on/off* characteristic, only active nodes should be allocated one time slot each, in a superframe. We divide active voice nodes into two categories: Type I and Type II nodes. Type I nodes in the current superframe were not allocated a time slot in the previous superframe[3], and are named "current-activated" nodes; Type II nodes remain active in both previous and current superframes, and are called "already-activated" nodes.

For each Type I node, voice traffic transits from the *off* state to the *on* state during previous superframe and generates voice packets before the node broadcasts a control packet in current superframe. Because of the randomness of state transition time from *off* to *on* in the previous superframe, packet transmissions from Type I nodes should have higher priority to be scheduled as early as possible according to their minislot accessing sequence, as long as Type II nodes can transmit within the delay bound, in order to minimize the possibility of Type I packet loss due to delay bound violation. Each Type II node has a time slot in the previous superframe and remains active in the current superframe. It should transmit packets no later than in the same time slot in the current superframe to meet the delay bound requirement. As an example, Fig. 3.3 illustrates how TDMA time slots are allocated in one superframe, with $N_v = 9$ and $N_{vm} = 10$, i.e., how to obtain current SSN based on the information in control packets. Each node has a unique minislot. Three types of important information in a broadcast control packet are shown: 1) MSN, the unique sequence number of a specific minislot; 2) BIB, which is 1 if the node has packets to transmit and 0 otherwise; 3) previous SSN, showing the sequence number of a TDMA time slot allocated to a voice node in the previous superframe, with SSN = 0 if the node was not allocated a TDMA time slot. Each entry in the left part of Fig. 3.3 discloses the information broadcast by the nodes in their minislots. The information broadcast by Type I and Type II active nodes is distinguished by the dashed-line and solid-line bounding rectangles, respectively. It can be seen that Nodes 8 and 5 are Type I with BIB = 1 and with previous SSN = 0, whereas Nodes 7, 1 and 4 are Type II with BIB = 1 and previous SSN $\neq$ 0.

[3] In most cases, the reason of not having a time slot is that the voice node has no packets to transmit. In some occasions when the instantaneous voice traffic load becomes heavy, an active node may not be able to get a time slot, resulting in packet dropping at the transmitter.

Node ID	MSN	BIB	Previous SSN	Current SSN
Node 7	1	1	3	3
Node 8	2	1	0	1
Node 5	3	1	0	4
Node 3	4	0	1	0
	5			
Node 1	6	1	2	2
Node 2	7	0	5	0
Node 6	8	0	4	0
Node 4	9	1	6	5
Node 9	10	0	0	0

Figure 3.3: An example of TDMA time slot allocation.

Based on the information provided by the control packets, the right part of Fig. 3.3 shows the time slot allocation in the current superframe. Node 8 accessing the first minislot among all the Type I nodes will transmit in the first time slot (with current SSN = 1) in the CFP. Node 5 is allocated a time slot after Nodes 1 and 7 because the latter two nodes are Type II nodes and should transmit packets no later than in their previously allocated time slots (with previous SSN = 2 and 3 respectively). Packet transmissions in current superframe from Node 4 (a Type II node with the largest previous SSN) are scheduled in a time slot with the index less than its previously allocated time slot.

3.2.3 T-CSMA/CA based Contention Access

In DAH-MAC, best-effort data nodes access the channel within a CP of each superframe according to the T-CSMA/CA contention protocol, in which data nodes attempt packet transmissions according to CSMA/CA with exponential backoff [17] and the transmissions are periodically interrupted by the presence of a CTP and a CFP. Thus, the performance of T-CSMA/CA is different from the traditional CSMA/CA contention protocol without interruptions. First, the packet waiting time for transmission is increased by the interrupted periods; Second, before each source node initiates a packet transmission attempt at the end of its backoff counter decrementing process, it is required to check whether the remaining time in the CP is enough to support at least one packet transmission. To have an acceptable transmission attempt, the remaining time should be not less than the summation of a data packet duration (T_{pd}, including acknowledgment) and a gurad time (T_{gt}). This summation is called *conflict period*. If the remaining time in current CP is not long enough, a *virtual conflict* occurs with the imminent CTP of next superframe. A *hold-on* strategy can be used to resolve the conflict [21] [22], in which the packet attempts are suspended until the start of next CP. Other nodes that are not involved in the conflict can still decrement their backoff counters within the conflict period until the end of current CP. When the next CP arrives, the transmission process resumes and the suspended packets are transmitted immediately after the channel is sensed idle for a distributed interframe space (DIFS).

By referring to some methods in [21], we give a detailed illustration inside the CP of each superframe, shown in Fig. 3.4, to highlight the differences between the T-CSMA/CA and the traditional CSMA/CA protocol. Suppose that the CP starts at time instant 0 and ends

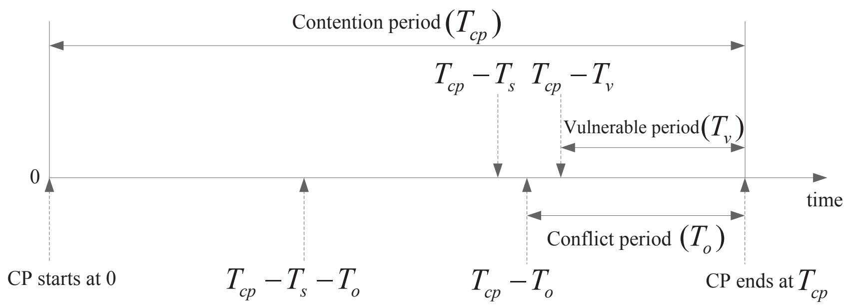


Figure 3.4: An illustration of the CP.

at T_{cp}. If a packet transmission attempt is initiated within the interval $[0, T_{cp} - T_s]$, the packet can be transmitted according to the CSMA/CA, either successfully or in collision, with a complete transmission duration T_s (T_{pd} plus a DIFS interval). The time instant $T_{cp} - T_o$ denotes the last time instant at which a packet transmission attempt can be initiated, and the conflict period is the following interval with duration T_o (T_{pd} plus T_{gt}), which is smaller than T_s. Thus, if a packet transmission starts in the interval $[T_{cp} - T_s, T_{cp} - T_o]$, the packet transmission time is on average $\frac{T_o+T_s}{2}$, assuming the transmission initiation instant is uniformly distributed within the interval. On the other hand, if the last packet transmission within the CP starts in the interval $[T_{cp} - T_s - T_o, T_{cp} - T_o]$, the transmission finishing point, denoted by $T_{cp} - T_v$, lies in the conflict period, where T_v is called *vulnerable period* [21] indicating the residual idle interval between the last transmission finishing point and the end of the CP; If no transmissions initiate during $[T_{cp} - T_s - T_o, T_{cp} - T_o]$, the starting point of the vulnerable period is $T_{cp} - T_o$, and the vulnerable period is the same as the conflict period. Thus, it can be seen that the vulnerable period is always not longer than the conflict period. The time interval $[0, T_{cp} - T_v]$ before the vulnerable period is called *non-vulnerable period*.

3.3 Performance Analysis

In this section, firstly, for a given maximum fraction of time (φ) for voice traffic in each superframe, the voice capacity, N_{vm}, under the packet loss rate bound is derived, which can facilitate voice session admission control. Secondly, for a specific N_v, the average number of voice burst transmissions scheduled in each superframe, $\overline{N_s}$, is derived, with which the average time duration of each CFP and CP, denoted by $\overline{T_{cfp}}$ and $\overline{T_{cp}}$, can be determined. Then, the aggregate throughput of the DAH-MAC for N_d data nodes is evaluated for each superframe, and maximized by adjusting the contention window size to the optimal value according to variations of N_v and N_d.

3.3.1 Voice capacity

As mentioned in Section 3.2, when nodes come into the network coverage area, they distributedly calculate the number of minislots, N_{vm}, in each CTP, indicating the maximum number of voice nodes supported in the network. Thus, within the voice capacity region, the following inequality needs to be satisfied to guarantee that the time duration for voice traffic not exceed the maximum fraction (φ) of each superframe time,

$$T_{ctrl} + T_{cfpm} = N_{vm}T_m + N_{sm}\lceil B \rceil T_{pv} \leq \varphi T_{SF} \tag{3.1}$$

where T_{cfpm} denotes the maximum value of T_{cfp}, T_m is the duration of each minislot, N_{sm} is the maximum value of N_s, indicating the maximum number of scheduled voice burst transmissions in a CFP to maintain the packet loss rate bound, B is the average size (number of voice packets) of a voice burst, $\lceil\rceil$ is the ceiling function, T_{pv} is the voice packet duration including header, and $\lceil B \rceil T_{pv}$ indicates the duration of each TDMA time slot (the duration of one time slot should allow an integer number of packet transmissions).

To determine N_{sm}, we first estimate that, with N_v voice nodes, how many generated voice packets in a superframe are required to be transmitted in the CFP to guarantee the packet loss rate bounded by P_L. Let X_i denote the number of packets generated by voice node i ($i = 1, 2, ..., N_v$) within a superframe, and y_m denote the maximum number of transmitted voice packets in the CFP to guarantee P_L. Since the length of a superframe is to be the same as the voice packet delay bound, lost packets are estimated as those generated but not transmitted within one superframe. Thus, y_m can be calculated by solving the following equation

$$\frac{E\left[X - y_m | X > y_m\right]}{E\left[X\right]} = P_L \tag{3.2}$$

where $X = \sum\limits_{i=1}^{N_v} X_i$.

Since $\{X_i, i = 1, 2, ..., N_v\}$ are independent and identically distributed (i.i.d) random variables, X can be approximated as a Gaussian random variable when N_v becomes relatively large (based on central limit theorem) [2], with mean $E[X]$ and variance $D[X]$ being $N_v E[X_i]$ and $N_v D[X_i]$ respectively. Thus, we estimate the distribution of X as a normal distribution, with which (3.2) is approximated as

$$\frac{\int_{y_m}^{N_v \cdot M_v} \frac{x - y_m}{\sqrt{2\pi N_v D[X_i]}} \cdot e^{-\frac{(x - N_v E[X_i])^2}{2 N_v D[X_i]}} dx}{N_v E[X_i]} = P_L \tag{3.3}$$

where $M_v = \lambda_v \cdot T_{SF}$ denotes the maximum number of packets generated by a voice source node within one superframe.

In (3.3), to derive $E[X_i]$ and $D[X_i]$, we calculate the distribution of X_i, which is the probability of generating k packets by voice node i within a superframe ready for transmission in the CFP, denoted by $P(k)$[4]. According to the *on/off* source model, the probability of a voice node staying at *on* (*off*) state, denoted by P_{on} (P_{off}), at any time instant, is $\frac{\beta}{\alpha+\beta}$ $\left(\frac{\alpha}{\alpha+\beta}\right)$. Let T_{on} (T_{off}) denote the time duration a voice node stays at *on* (*off*) state. We have

$$\begin{aligned} P(k) &= P_{on} \cdot P\left\{\frac{k-1}{\lambda_v} < T_{on} \le \frac{k}{\lambda_v}\right\} + P_{off} \cdot P\left\{T_{SF} - \frac{k}{\lambda_v} < T_{off} \le T_{SF} - \frac{k-1}{\lambda_v}\right\} \\ &= \frac{\beta}{\alpha+\beta}\left[e^{-\frac{\alpha(k-1)}{\lambda_v}} - e^{-\frac{\alpha k}{\lambda_v}}\right] + \frac{\alpha}{\alpha+\beta}\left[e^{-\beta\left(T_{SF} - \frac{k}{\lambda_v}\right)} - e^{-\beta\left(T_{SF} - \frac{k-1}{\lambda_v}\right)}\right] \quad (1 \le k \le M_v - 1) \end{aligned} \tag{3.4}$$

[4] Since $\{X_i, i = 1, 2, ..., N_v\}$ have identical probability distribution, we simply drop the voice node index i.

$$P(M_v) = P_{on} \cdot P\left\{T_{on} > \frac{M_v - 1}{\lambda_v}\right\} + P_{off} \cdot P\left\{T_{off} \leq \frac{1}{\lambda_v}\right\}$$
$$= \frac{\beta}{\alpha+\beta} e^{-\frac{\alpha(M_v-1)}{\lambda_v}} + \frac{\alpha}{\alpha+\beta}\left(1 - e^{-\frac{\beta}{\lambda_v}}\right) \tag{3.5}$$

and

$$P(0) = 1 - \sum_{k=1}^{M} P(k). \tag{3.6}$$

With the probability distribution of X_i, the average voice burst size B can be obtained. Based on y_m and B, the maximum number of scheduled voice bursts, N_{sm}, in each CFP is derived. Then, with a specific φ, Algorithm 2 can be used to determine the maximum number of voice nodes (voice capacity), N_{vm}, that can be supported in the network.

Algorithm 2: Voice capacity

Input : The maximum fraction of time, φ, for voice traffic in each superframe.
Output: Voice capacity N_{vm}, the CTP duration T_{ctrl}.

Initialization: $N_v \leftarrow 1$;
do
    $y_m \leftarrow$ solving (3.3);
    $N_{sm} \leftarrow \frac{y_m}{B}$;
    if $N_v T_m + N_{sm} \lceil B \rceil T_{pv} \leq \varphi T_{SF}$ **then**
        $N_v \leftarrow N_v + 1$;
    else
        $N_{vm} \leftarrow N_v - 1$;
        $T_{ctrl} \leftarrow N_{vm} T_m$;
        break;
    end
while $N_v > 0$;
return N_{vm} and T_{ctrl}.

3.3.2 Average number of scheduled voice bursts in a CFP

The actual number of generated voice bursts is likely less than N_{sm} and varies depending on the buffer occupancy states broadcast at the beginning of each superframe. In the following, for a specific N_v, we calculate the average number of scheduled voice bursts, $\overline{N_s}$, with which $\overline{T_{cfp}}$ and $\overline{T_{cp}}$ can be obtained.

We first determine the probability distribution of the number of active voice nodes, denoted by N_{av}, which broadcast control packets with BIB = 1 in their respective minislots of each superframe. Due to the voice source *on/off* characteristic, N_{av} is composed of two portions: 1) the number of nodes with a nonempty buffer staying at the *on* state, denoted by N_{av}^{on}; and 2) the number of nodes with a nonempty buffer staying at the *off* state, denoted by N_{av}^{off}.

Since packets periodically arrives at the transmission buffer every $\frac{1}{\lambda_v}$ second for each voice node at the *on* state, the probability of a voice node being active in its occupied minislot conditioned on that the node is at the *on* state can be derived by calculating a *posterior probability* as

$$P_{a|on} = P\left\{\text{on at } \left(t_i - \frac{1}{\lambda_v}\right) \middle| \text{on at } t_i\right\} = e^{-\frac{\alpha}{\lambda_v}} \tag{3.7}$$

where t_i is the time instant that voice node i broadcasts in its selected minislot in the current superframe. Eq. (3.7) indicates that the time duration of voice node i staying at the *on* state should last for at least the duration of $\frac{1}{\lambda_v}$ before it broadcasts at t_i to ensure a nonempty transmission buffer.

Similarly, the probability of voice node i being active at t_i, conditioned on that the node stays at the *off* state, is calculated as

$$\begin{aligned} P_{a|off} &= P\left\{\text{on at}\left(t_i - T\right), \text{on at } \left(t_i - T + \frac{1}{\lambda_v}\right) \middle| \text{off at } t_i\right\} \\ &= \frac{\beta}{\alpha}\left(e^{-\frac{\alpha}{\lambda_v}} - e^{-\alpha T}\right) \end{aligned} \tag{3.8}$$

where $T = T_{SF} - T_{cfpm}$. Thus, $t_i - T$ is a calculation for the time instant of the end of the CFP in previous superframe[5]. Eq. (3.8) indicates that voice node i should stay at the *on* state for at least the time interval of $\frac{1}{\lambda_v}$ after the end of previous CFP to ensure a nonempty transmission buffer before time instant t_i. Then, the probability distribution of

[5]For calculation simplicity, we assume that the duration spent in current CTP before t_i is T_{ctrl}, and the duration of previous CFP is T_{cfpm}.

N_{av}, denoted by $P_{N_{av}}(k)$, can be derived as

$$
\begin{aligned}
P_{N_{av}}(k) &= P\{N_{av}^{on} + N_{av}^{off} = k\} \\
&= \sum_{i=0}^{N_v} \sum_{j \in \mathbf{A}} P\{N_{av}^{off} = k - N_{av}^{on} \big| N_{av}^{on} = j, N^{on} = i\} P\{N_{av}^{on} = j | N^{on} = i\} P\{N^{on} = i\} \\
&= \sum_{i=0}^{N_v} \sum_{j \in \mathbf{A}} P(i,j,k) \qquad\qquad (0 \le k \le N_v)
\end{aligned}
\tag{3.9}
$$

where N^{on} denotes the number of nodes in the *on* state, set $\mathbf{A}$ denotes the value range of j depending on k and i, and

$$P(i,j,k) = \binom{N_v - i}{k - j} P_{a|off}^{k-j} \left(1 - P_{a|off}\right)^{N_v - i - k + j} \cdot \binom{i}{j} P_{a|on}^{j} \left(1 - P_{a|on}\right)^{i-j} \cdot \binom{N_v}{i} P_{on}^{i} P_{off}^{N_v - i}.$$

Then, the complete expression of $P_{N_{av}}(k)$ is obtained by delimiting j in (3.9) considering the following three cases:

(i) $N_v - k > k$,

$$P_{N_{av}}(k) = \sum_{i=0}^{k} \sum_{j=0}^{i} P(i,j,k) + \sum_{i=k+1}^{N_v - k} \sum_{j=0}^{k} P(i,j,k) + \sum_{i=N_v-k+1}^{N_v} \sum_{j=i-N_v+k}^{k} P(i,j,k); \tag{3.10}$$

(ii) $N_v - k < k$,

$$P_{N_{av}}(k) = \sum_{i=0}^{N_v-k} \sum_{j=0}^{i} P(i,j,k) + \sum_{i=N_v-k+1}^{k} \sum_{j=i-N_v+k}^{i} P(i,j,k) + \sum_{i=k+1}^{N_v} \sum_{j=i-N_v+k}^{k} P(i,j,k); \tag{3.11}$$

(iii) $N_v - k = k$,

$$P_{N_{av}}(k) = \sum_{i=0}^{k} \sum_{j=0}^{i} P(i,j,k) + \sum_{i=k+1}^{N_v} \sum_{j=i-N_v+k}^{k} P(i,j,k). \tag{3.12}$$

Thus, with $P_{N_{av}}(k)$, the probability mass function (pmf) of the number of scheduled voice bursts, N_s, is given by

$$P_{N_s}(k) = \begin{cases} P_{N_{av}}(k) & (0 \leq k \leq N_{sm} - 1) \\ \displaystyle\sum_{r=N_{sm}}^{N_v} P_{N_{av}}(r) & (k = N_{sm})\,. \end{cases} \tag{3.13}$$

Finally, $\overline{N_s}$, $\overline{T_{cfp}}$, and $\overline{T_{cp}}$ can be obtained accordingly, based on $P_{N_s}(k)$.

3.3.3 A data throughput optimization framework for the DAH-MAC

In a CP, data nodes access the channel according to the T-CSMA/CA contention protocol. Since we evaluate the average aggregate throughput for data nodes in each superframe, $\overline{T_{cp}}$ is used to denote the average duration of a CP[6]. For the T-CSMA/CA, nodes are restricted to transmit packets within each CP. Before any transmission attempts, nodes are required to ensure that the remaining time in current CP is long enough for at least one packet transmission; otherwise, all transmission attempts are suspended until the next CP starts.

According to the illustration inside the CP of each superframe in Fig. 3.4, the average value of T_v can be calculated by (3.14), assuming the transmission starting point is uniformly distributed inside $\left[\overline{T_{cp}} - T_s - T_o, \overline{T_{cp}} - T_o\right]$.

$$\overline{T_v} = (1-\tau)^{N_d T_s} \cdot T_o + \left[1 - (1-\tau)^{N_d T_s}\right] \cdot \frac{T_o}{2} = \frac{\left[1 + (1-\tau)^{N_d T_s}\right] T_o}{2} \tag{3.14}$$

where τ is the packet transmission probability of a data node with a nonempty transmission buffer at any backoff slot.

A generic time slot in the non-vulnerable period of the CP can be divided into three categories: 1) an idle backoff slot; 2) a complete transmission slot with the duration T_s (a successful transmission duration is assumed the same as a collision duration [72]); and 3) a restricted transmission slot with the duration $\frac{T_o+T_s}{2}$. Thus, the average duration of a generic time slot in the non-vulnerable period is calculated as

$$\sigma = (1-\tau)^{N_d} + \left[1 - (1-\tau)^{N_d}\right] T_a \tag{3.15}$$

[6]All time intervals in a CP are normalized to the unit of an idle backoff time slot duration as commonly seen in CSMA/CA based systems.

where $T_a = \left(\frac{T_s - T_o}{\overline{T_{cp}} - T_o} \cdot \frac{T_s + T_o}{2} + \frac{\overline{T_{cp}} - T_s}{\overline{T_{cp}} - T_o} \cdot T_s\right)$ denotes the average duration of a transmission slot in the non-vulnerable period.

Then, the probability that a generic slot is inside the vulnerable period is given by

$$p_v = \frac{\overline{T_v}}{\frac{\overline{T_{cp}} - \overline{T_v}}{\sigma} + \overline{T_v}}. \tag{3.16}$$

Thus, the duration of a generic slot including the vulnerable period is derived as

$$\begin{aligned} \sigma_d &= p_v + (1 - p_v)\sigma \\ &= p_v + (1 - p_v)\left[(1-\tau)^{N_d} + \left[1 - (1-\tau)^{N_d}\right] T_a\right]. \end{aligned} \tag{3.17}$$

Since packet transmissions or collisions cannot happen in T_v, the packet collision probability for each node at any backoff slot in a traffic saturation case is expressed as

$$p = 1 - (1 - p_v)(1 - \tau)^{N_d - 1}. \tag{3.18}$$

In (3.18), the transmission probability τ can be approximated, based on renewal reward theory, as a ratio of the average reward received during a renewal cycle over the average length of the renewal cycle [21] [72]. That is,

$$\tau = \frac{E[A]}{E[A] + E[W]} = \frac{\sum_{j=0}^{R_l - 1} p^j}{\sum_{j=0}^{R_l - 1} p^j + \sum_{j=0}^{R_l - 1} \left(\frac{CW_j}{2} \cdot p^j\right)} \tag{3.19}$$

where $E[A]$ and $E[W]$ denote the average number of transmission attempts and backoff slots experienced, respectively, before a successful packet transmission; R_l is the retransmission limit; and $CW_j = 2^j CW (j = 0, 1, ..., M_b)$ is the contention window size in backoff stage j (CW is the minimum contention window size and M_b is the maximum backoff stage).

Therefore, the aggregate data saturation throughput[7] for the DAH-MAC is expressed

[7]The throughput in this chapter is normalized by the channel capacity.

as

$$
\begin{aligned}
S_d &= \frac{N_d T_{pd} \tau (1-p)}{\sigma_d} \cdot \frac{\overline{T_{cp}}}{T_{SF}} \\
&= \frac{N_d T_{pd} \tau (1-p_v)(1-\tau)^{N_d-1}}{p_v + (1-p_v)[(1-\tau)^{N_d} + [1-(1-\tau)^{N_d}]\, T_a]} \cdot \frac{\overline{T_{cp}}}{T_{SF}}
\end{aligned}
\tag{3.20}
$$

where T_{pd} is the data packet duration, and $\overline{T_{cp}}$ is a function of N_v.

From (3.20), we can see that when N_v and N_d are given and other system parameters are set, e.g., according to IEEE 802.11b standard [1], the saturation throughput S_d is a function of τ, and can be evaluated by solving (3.18) for τ numerically. We rewrite (3.20) as

$$
S_d = \frac{T_{pd} \cdot \frac{\overline{T_{cp}}}{T_{SF}}}{\overline{T_{vd}}} \tag{3.21}
$$

where $\overline{T_{vd}} = \frac{p_v + (1-p_v)\left[(1-\tau)^{N_d} + \left[1-(1-\tau)^{N_d}\right] T_a\right]}{N_d \tau (1-p_v)(1-\tau)^{N_d-1}}$. $\overline{T_{vd}}$ is called *average virtual transmission time* [73], which indicates the total average time experienced (including backoff waiting time, collision time and packet transmission time) in each CP to successfully transmit one packet.

Therefore, with different values of N_d, we evaluate the relationship between $\overline{T_{vd}}$ and τ, shown in Fig. 3.5. It can be seen that there exists an optimal transmission probability τ_{opt} that achieves a minimum of $\overline{T_{vd}}$. The reason of existing such an optimal transmission probability can be explained as follows: When $\tau < \tau_{opt}$, an increasing amount of channel time remains idle before an transmission initiates, which consistently enlarges $\overline{T_{vd}}$ even if transmission collisions rarely happen in this scenario; However, when τ continues to increase beyond τ_{opt}, due to more frequent transmission attempts, the number of packet collisions rises, which consume an increasing fraction of channel time before packets are successfully transmitted. Therefore, the existence of τ_{opt} can be regarded as a compromise of the preceding two effects and achieves a minimum virtual transmission time and a maximum throughput. Since overheads consumed in one transmission collision are much greater than in an idle backoff slot, the optimal transmission probability is obtained as a relatively small value, as shown in Fig. 3.5, to lower the collision probability at the expense of consuming more idle slots. Therefore, our objective is to first derive τ_{opt} as a function of N_v and N_d. Then, by substituting τ_{opt} into (3.14)-(3.19), a closed-form mathematical relationship can be established between the optimal value of contention window size CW, denoted by CW_{opt}, and the heterogeneous network traffic load.

To do so, the expression of $\overline{T_{vd}}$ in (3.21) can be further derived as the summation of

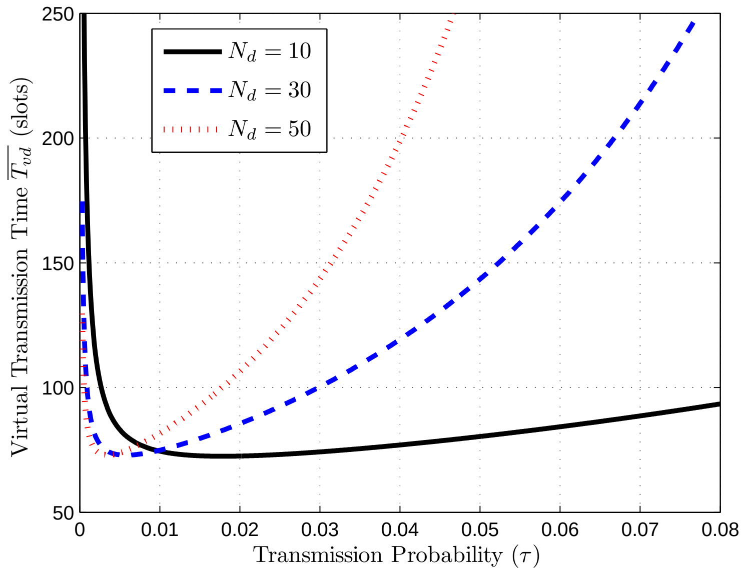


Figure 3.5: The evaluation of $\overline{T_{vd}}$ in a function of τ ($\varphi = 0.5$, $N_v = 20$).

the following three terms:

$$\overline{T_{vd}} = \frac{p_v}{N_d(1-p_v)\tau(1-\tau)^{N_d-1}} + \frac{1-\tau}{N_d\tau} + \frac{\left[1-(1-\tau)^{N_d}\right]T_a}{N_d\tau(1-\tau)^{N_d-1}}. \tag{3.22}$$

From (3.22), it is computational complex to obtain the first order derivative function of $\overline{T_{vd}}$ with respect to τ. The complexity mainly results from p_v which is a complex function of τ in (3.16). Thus, to make the derivation of $\overline{T_{vd}}$ tractable, an approximation of p_v can be obtained by simplifying $\overline{T_v}$, considering the following two cases:

1) For $\tau \geq 0.005$, since all time durations in a CP are normalized to the unit of an idle backoff slot duration, we have $T_s \gg 1$ (according to the IEEE 802.11b specification) and $N_dT_s \gg 1$. Thus,

$$\overline{T_v} = \frac{\left[1+(1-\tau)^{N_dT_s}\right]T_o}{2} \approx \frac{T_o}{2} \quad (\tau \geq 0.005); \tag{3.23}$$

2) For $\tau < 0.005$, the average duration of a generic time slot in the non-vulnerable period, σ, approaches 1. Moreover, since $\overline{T_v} \in \left[\frac{T_o}{2}, T_o\right]$, we have $\overline{T_v} \ll \overline{T_{cp}}$. Thus, Eq. (3.16)

can be approximated as

$$p_v = \frac{\sigma \overline{T_v}}{\overline{T_{cp}} + (\sigma - 1)\overline{T_v}} \approx \frac{\sigma \overline{T_v}}{\overline{T_{cp}}} \approx \frac{\sigma T_o}{2\overline{T_{cp}}} \quad (\tau < 0.005). \tag{3.24}$$

As a result, by using $\frac{T_o}{2}$ (the lower bound of $\overline{T_v}$) to approximate $\overline{T_v}$, the approximation of p_v is

$$\widetilde{p_v} = \frac{\frac{T_o}{2}}{\frac{\overline{T_{cp}} - \frac{T_o}{2}}{\sigma} + \frac{T_o}{2}}. \tag{3.25}$$

Therefore, by substituting (3.25) into (3.22) and after some algebraic manipulation, the approximation of $\overline{T_{vd}}$ is obtained as

$$\widetilde{\overline{T_{vd}}} = \frac{\overline{T_{cp}}}{\overline{T_{cp}} - \frac{T_o}{2}} \cdot \frac{T_a - (T_a - 1)(1-\tau)^{N_d}}{N_d \tau (1-\tau)^{N_d - 1}}. \tag{3.26}$$

Then, by taking the first order derivative of $\widetilde{\overline{T_{vd}}}$ with respect to τ and letting the derivative function equal to 0, we solve for an approximation of τ_{opt} (under the condition of $\tau \ll 1$) as a closed-form function of N_v and N_d, given by

$$\widetilde{\tau_{opt}} = \frac{\sqrt{1 + \frac{2(T_a - 1)(N_d - 1)}{N_d}} - 1}{(T_a - 1)(N_d - 1)}. \tag{3.27}$$

Fig. 3.6 shows the accuracy of the approximation by plotting τ_{opt} and $\widetilde{\tau_{opt}}$ over a wide range of N_d. Note that although the optimal transmission probability, τ_{opt}, for each data node in a backoff slot decreases to a relatively small value with the increase of N_d, the probability of a successful packet transmission in a backoff slot is much higher than τ_{opt} when N_d becomes large, to achieve the maximized throughput.

Then, by substituting $\widetilde{\tau_{opt}}$ and $\widetilde{p_v}$ into (3.15)-(3.19), we derive an approximate expression for the optimal contention window CW_{opt}, as a closed-form function of N_v and N_d, given by

$$\widetilde{CW_{opt}} = \frac{\left(1 - \widetilde{\tau_{opt}}\right)\left(1 - \widetilde{p}^{R_l}\right)}{\widetilde{\tau_{opt}}\left(1 - \widetilde{p}\right)\left(\sum_{j=0}^{M_b} 2^{j-1}\widetilde{p}^j + \sum_{j=M_b+1}^{R_l - 1} 2^{M_b - 1}\widetilde{p}^j\right)} \tag{3.28}$$

where $\widetilde{p} = 1 - (1 - \widetilde{p_v})(1 - \widetilde{\tau_{opt}})^{N_d - 1}$.

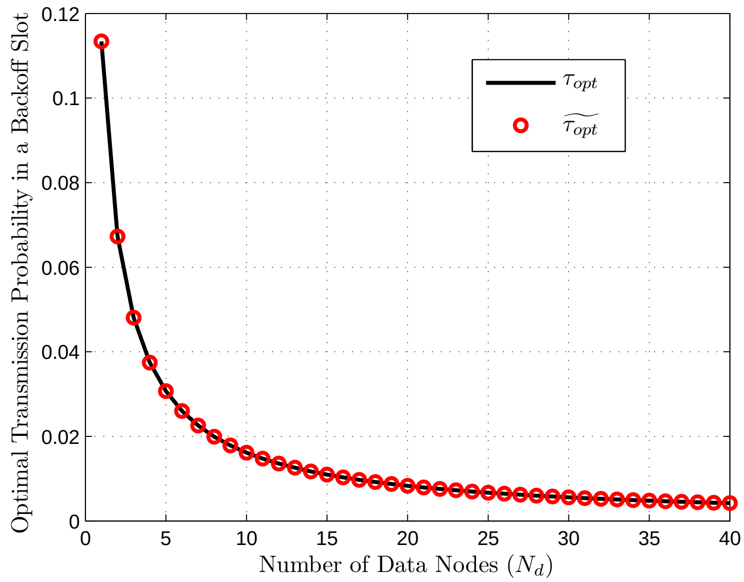


Figure 3.6: Optimal transmission probability in each backoff slot for data nodes ($N_v = 20$, $\varphi = 0.5$).

The proposed analytical framework not only provides an effective way to evaluate the performance of the DAH-MAC in supporting both voice and data traffic, but also provides some insights in MAC design in practical engineering for performance improvement: First, the voice capacity, N_{vm}, and the maximum number of voice bursts, N_{sm}, scheduled for each superframe are derived based on the analytical model and used as a reference for engineers in the protocol design to guarantee the voice delay bound in presence of the voice traffic load dynamics; Second, with the closed-form mathematical relationship provided in (3.28), data nodes operating T-CSMA/CA in each CP can adaptively adjust the minimum contention window size CW to the optimal value $\widetilde{CW_{opt}}$, based on the updated heterogeneous network traffic load information acquired in each superframe, to achieve consistently maximum aggregate data throughput.

3.4 Numerical Results

In this section, simulation results are provided to validate the accuracy of the analytical results. All simulations are carried out using OMNeT++ [69] [70] [74]. Nodes are interconnected and each source node randomly selects one of the rest nodes as its destination node. We run each simulation for 10000 superframe intervals to generate one simulation point. The main simulation settings for different MAC schemes with voice and data traffic are listed in Table 3.1. For a voice source, the GSM 6.10 codec is chosen for encoding the voice stream, with which voice packet payload size is 33 bytes and packets interarrival interval is 20 ms when the voice source node is at the *on* state [2]. Packet arrivals for each best-effort data node follow a Poisson process with the average arrival rate λ_d. We set λ_d to 500 packet/s to ensure each data transmission buffer is always saturated.

Table 3.1: Simulation parameter settings [1] [2]

MAC schemes / **Parameters**	**DAH-MAC**	**Busy-tone contention protocol [2]**	**D-PRMA [41]**
Channel capacity	11Mbps	11Mbps	11Mbps
Backoff slot time	20 μs	20 μs	n.a.
Minimum contention window size (voice/data)	n.a.	8/32	n.a.
Maximum contention window size (voice/data)	n.a.	16/1024	n.a.
Backoff Stage limit (voice/data)	n.a./5	1/5	n.a.
Retransmission limit (voice/data)	n.a./7	2/7	n.a.
PLCP & Preamble	192 μs	192 μs	192 μs
MAC header	24.7 μs	24.7 μs	24.7 μs
RTP/UDP/IP headers (voice)	$\frac{4\cdot 8}{11}$ μs	$\frac{4\cdot 8}{11}$ μs	$\frac{4\cdot 8}{11}$ μs
Packet payload length (voice/data)	$\frac{33\cdot 8}{11}$ / $\frac{8184}{11}$ μs	$\frac{33\cdot 8}{11}$ / $\frac{8184}{11}$ μs	$\frac{33\cdot 8}{11}$ / $\frac{8184}{11}$ μs
AIFS/DIFS (voice/data)	n.a./50 μs	30/50 μs	n.a.
Minislot duration	0.25 ms	n.a.	0.41 ms
Time slot duration	1.22 ms	n.a.	1.64 ms
Transmission time (voice/data)	0.244/1.18 ms	0.244/1.18 ms	0.244/1.18 ms
Gurad time (T_{gt})	20 μs	n.a.	n.a.
Average *on/off* time $\left(\frac{1}{\alpha} / \frac{1}{\beta}\right)$	352/650 ms	352/650 ms	352/650 ms
Minislot contention probability (voice/data)	n.a.	n.a.	0.6/0.2
Transmission queue length	10000 packets	10000 packets	10000 packets
Superframe time (delay bound)	100 ms	100 ms	100 ms

We first study the voice capacity, determined by Algorithm 2 in Section 4.3, with a variation of the maximum fraction of time (φ) for voice traffic in each superframe. Then, with a specific φ, the maximum number of voice burst transmissions supported in a CFP to guarantee the packet loss rate bound and the average number of scheduled voice bursts are both evaluated. For performance metrics, the voice packet loss rate and aggregate data throughput are considered. We also compare the performance of the proposed DAH-MAC scheme with two well-known MAC protocols.

3.4.1 Voice capacity

The voice capacity in the network, with a variation of φ, for the DAH-MAC is plotted in Fig. 3.7. The analytical results are obtained according to Algorithm 2 in Section 3.3.1. It can be seen that the analytical results closely match the simulation results especially when the voice capacity region is relatively large, since using the central limit theorem to approximate the distribution of X in (3.2) becomes more accurate when N_v gets larger.

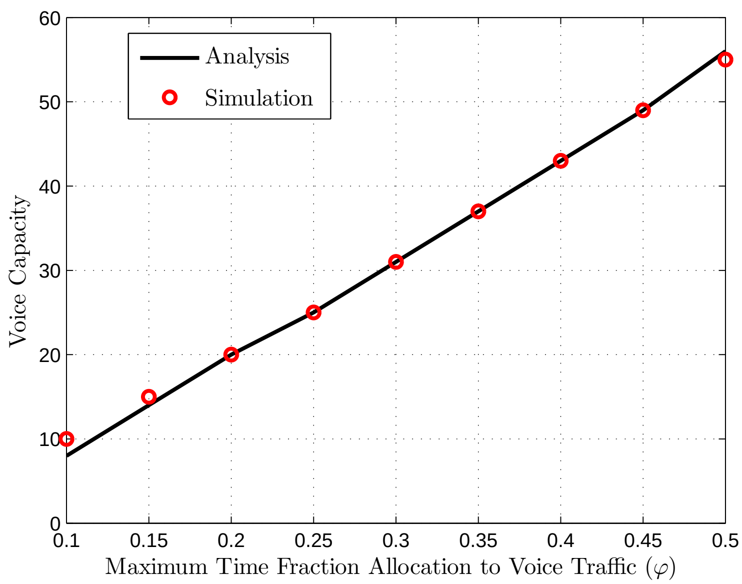


Figure 3.7: Voice capacity region with different φ.

3.4.2 Number of scheduled voice bursts (time slots) in a CFP

We also evaluate the number of voice burst transmissions (time slots) in a CFP with different N_v in the voice capacity region. Fig. 3.8 shows the average number of scheduled voice bursts ($\overline{N_s}$) in each superframe. It can be seen that the analytical and simulation results closely match, which verifies the accuracy of our analysis. In Fig. 3.8, we also plot the maximum supported voice bursts (N_{sm}) in each CFP. Due to the randomness of voice

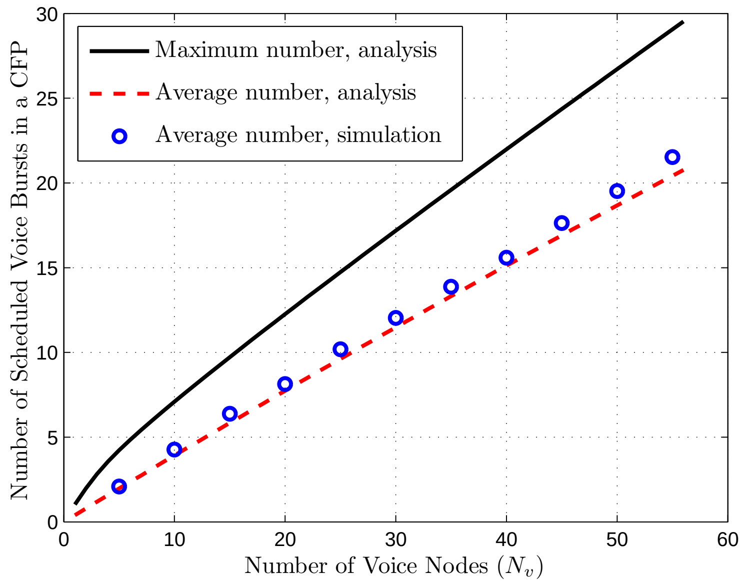


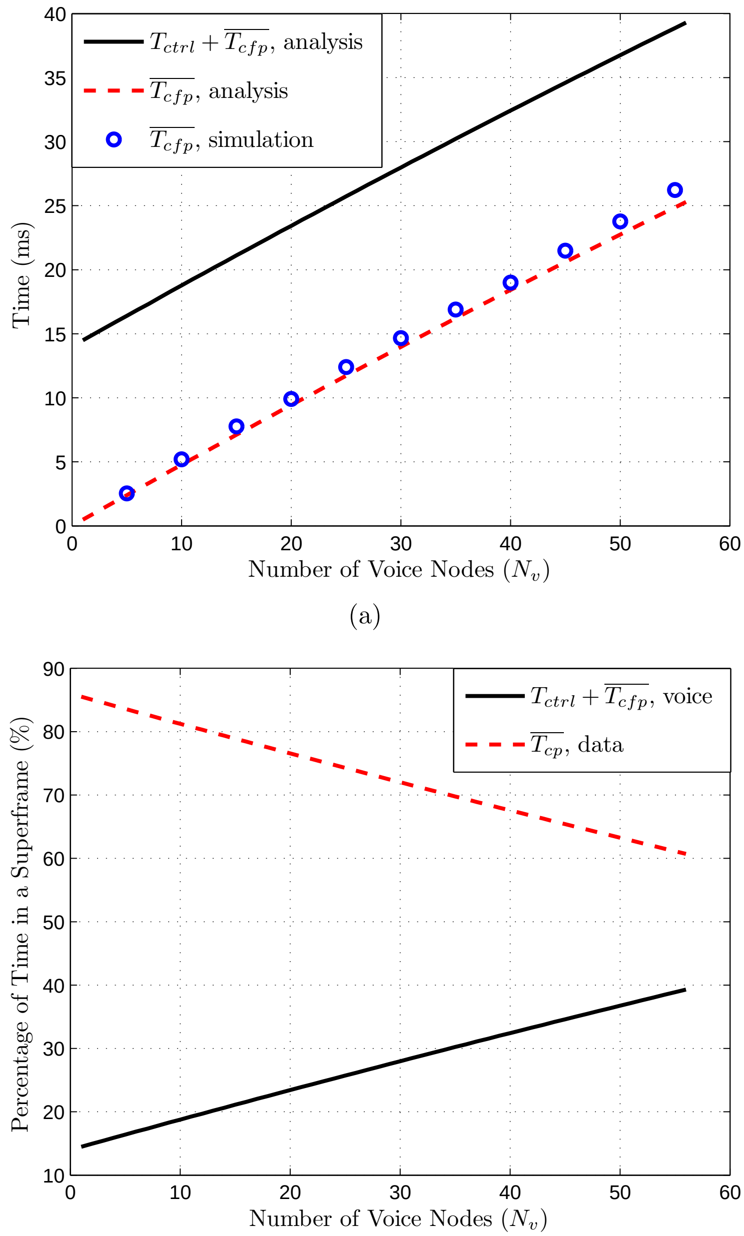
Figure 3.8: Number of scheduled time slots for voice traffic ($\varphi = 0.5$).

packet arrivals, the instantaneous voice traffic load fluctuates on a per-superframe basis. The gap between N_{sm} and $\overline{N_s}$ indicates the number of time slots allocated to voice bursts in each CFP is commonly below the maximum allowable value. Therefore, by adapting to the realtime voice traffic load in each superframe, the proposed distributed TDMA time slot allocation achieves a high resource utilization.

Fig. 3.9 shows the average time allocated to voice and data traffic in each superframe with a specific φ. In Fig. 3.9(a), we can see that the average time of each CFP ($\overline{T_{cfp}}$) for voice burst transmissions increases with N_v, with a fixed duration of each CTP (T_{ctrl}) for accommodating a certain number of voice nodes, which is determined according to the voice capacity with $\varphi = 0.5$; Fig. 3.9(b) shows the average percentage of time allocated to voice and data traffic in each superframe. It can be seen that within the capacity region, the average time allocated to voice traffic is always bounded by φT_{SF} under the packet loss rate constraint, and the residual average superframe time are occupied by data traffic.

3.4.3 Voice packet loss rate

Packet loss rate for voice traffic in a CFP is evaluated with different φ in Fig. 3.10. It is observed that the simulation results are close to the analytical results. Although some performance fluctuations appear when N_v is relatively small due to the central limit theorem



(a)

(b)

Figure 3.9: Average time allocation in a superframe ($\varphi = 0.5$). (a) Durations of CTP and CFP for voice traffic. (b) Percentage of time for voice and data traffic.

approximation and the rounding-off effect in deriving N_{sm} (set as a simulation parameter), the packet loss rate is always below the performance bound within the voice capacity region, which verifies the effectiveness of our proposed MAC in supporting voice service. If the number of minislots in the control period of each superframe is set beyond the voice capacity N_{vm}, the packet loss rate rises dramatically as shown in Fig. 3.10. Therefore, Algorithm 2 in Section 3.3.1 is employed in the DAH-MAC to calculate N_{vm} with different requirement of φ, which controls the number of voice nodes N_v within the capacity region to guarantee a bounded packet loss rate.

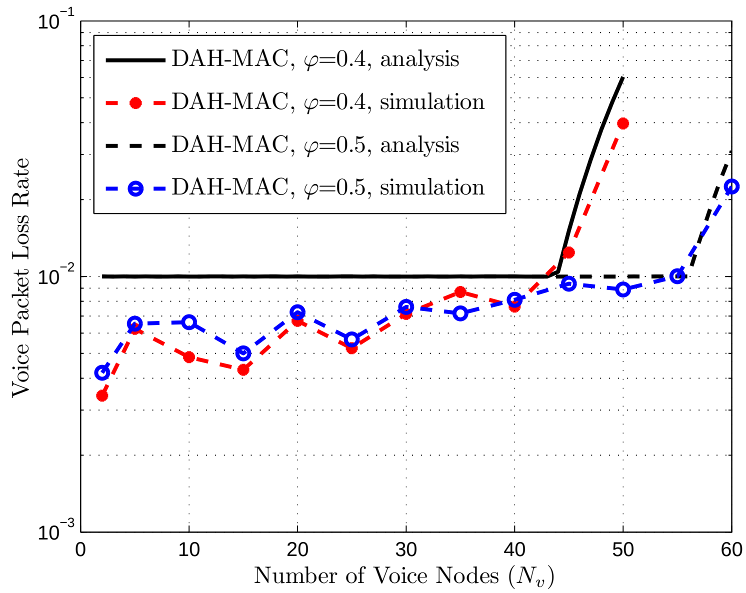


Figure 3.10: Voice packet loss rate in a CFP with different φ.

Fig. 3.11 displays a comparison of voice packet loss rates between the proposed DAH-MAC and two well-known MAC protocols: D-PRMA protocol [41] and busy-tone contention protocol [2], with a variation of N_v. The latter two MAC protocols are both effective in supporting voice packet transmissions. We can see that the D-PRMA can guarantee a bounded packet loss rate when N_v is relatively small. However, the packet loss rate increases dramatically since contention collisions rise when an increasing number of voice nodes start to contend for the transmission opportunity in each available time slot. Thus, the voice capacity region for D-PRMA is limited. Different from the D-PRMA, the busy-tone contention protocol grants a deterministic channel access priority for voice traffic. Thus, it can be seen that the voice packet loss rate is guaranteed over a wide range of N_v. Nevertheless, due to the contention nature, a consistent increase of voice packet colli-

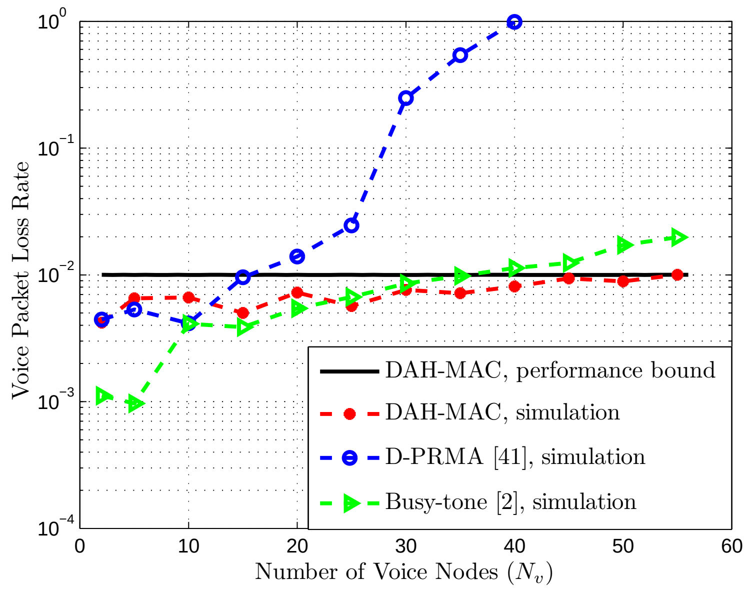


Figure 3.11: A comparison of voice packet loss rates ($N_d = 10$, $\varphi = 0.5$).

sions lead to accumulated channel access time, and the packet loss rate eventually exceeds the bound after around 35 voice nodes are admitted. In the DAH-MAC, the proposed distributed TDMA can admit more voice sessions by setting a higher value of φ, at the expense of local information exchanges in each enlarged control period. As can be seen in Fig. 3.11, the voice capacity region of the proposed MAC can be larger than the other two MAC protocols with a bounded packet loss rate.

3.4.4 Aggregate best-effort data throughput

We first evaluate the average channel utilization of T-CSMA/CA in each CP, defined as the ratio of average time used for successful data packet transmissions in a CP to average duration of the CP. It can be seen in Fig. 3.12 that the T-CSMA/CA achieves consistently high channel utilization with variations of data traffic load, since the proposed throughput analytical framework maximizes the T-CSMA/CA channel utilization within the DAH-MAC superframe structure.

Then, we make a comparison of aggregate data throughput between the proposed DAH-MAC and the busy-tone contention protocol with a variation of N_d and under different voice traffic load conditions. To ensure a fair comparison, we set φ as 0.33 for DAH-MAC to achieve the same voice capacity ($N_{vm} = 35$) with the busy-tone contention protocol. First,

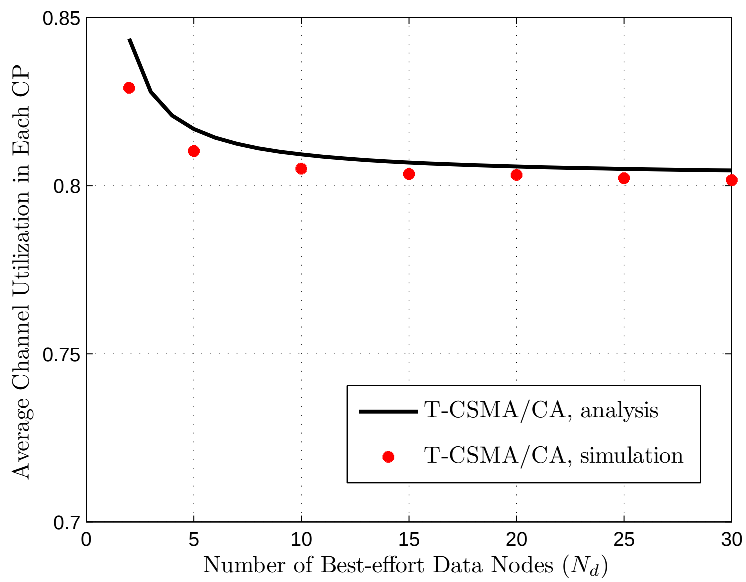


Figure 3.12: Channel utilization for data traffic in each CP ($N_v = 20$, $\varphi = 0.5$).

when $N_v = 35$ representing a high voice traffic load condition, it can seen from Fig. 3.13(a) that the DAH-MAC can achieve a consistently higher data throughput than the busy-tone contention protocol. This is because the distributed TDMA can achieve better resource utilization in high voice traffic load conditions, and the aggregate data throughput is maximized over a wide range of N_d; for the busy-tone contention protocol, with a high voice traffic load, an increasing fraction of channel time is consumed for voice collisions resolution, and a large number of voice nodes also limit the channel access opportunity for data traffic, since voice traffic has absolute priority of accessing the channel. The data throughput comparison is also conducted when the voice traffic load is moderate. It can be seen in Fig. 3.13(b) that throughputs of the DAH-MAC and the busy-tone contention protocol experience a similar trend with the increase of N_d as in Fig. 3.13(a), except for the cases where the busy-tone contention protocol achieves a higher throughput when N_d becomes relatively small. The advantage of busy-tone contention becomes more notable when the voice traffic load gets lower, as can be seen in Fig. 3.13(c) where $N_v = 5$. Therefore, the results from Fig. 3.13(b) - 3.13(c) demonstrate some effectiveness of busy-tone based contention when N_v and N_d are relatively small, since in a low heterogeneous network traffic load condition, contention collisions among voice or data nodes are largely reduced, thus improving the channel utilization of the contention-based MAC protocol. Overall, over a wide range of N_d, our proposed MAC can achieve a consistently higher throughput especially in a high voice traffic load condition.

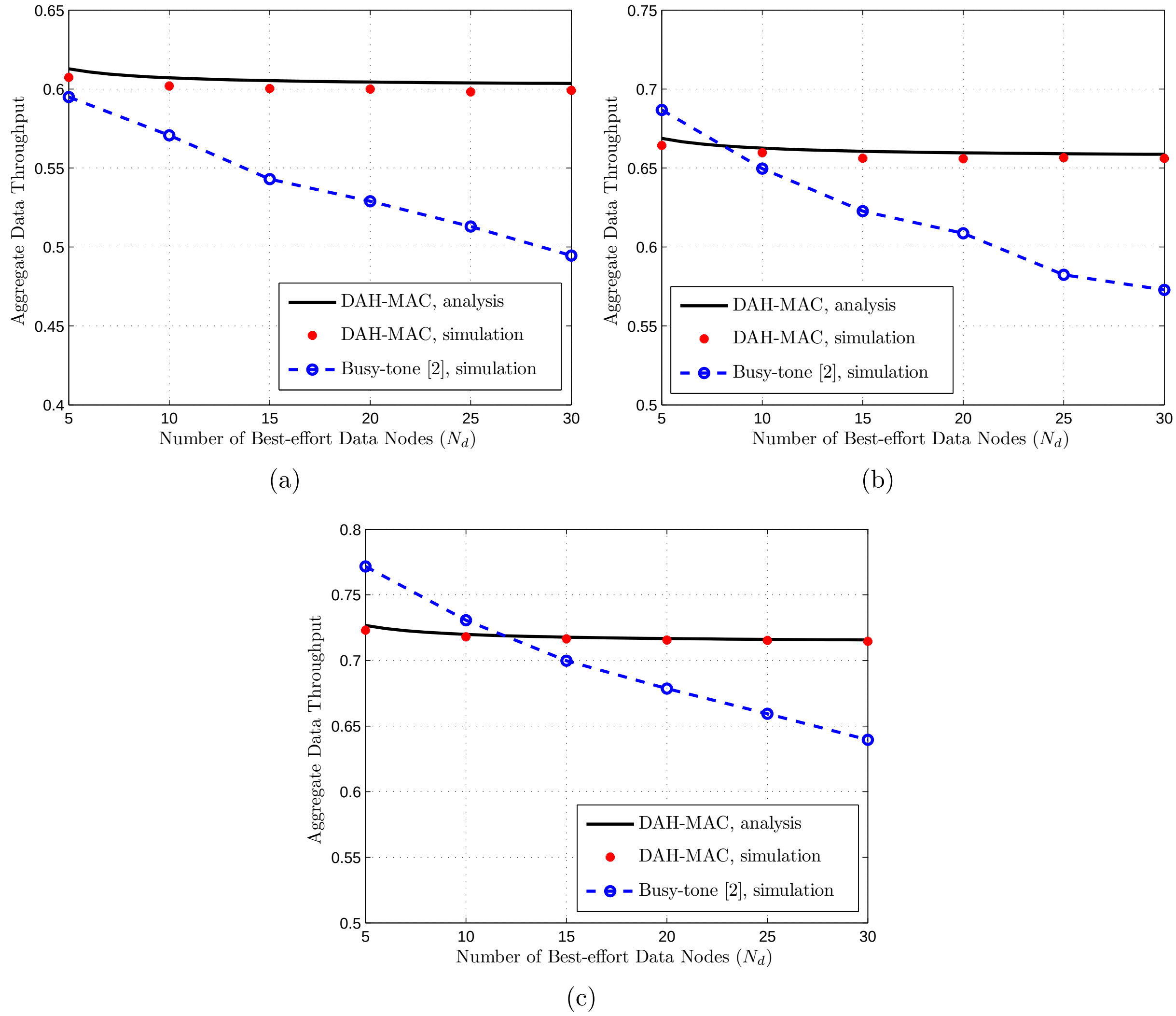


Figure 3.13: A comparison of the DAH-MAC maximum data throughput ($\varphi = 0.33$) with the busy-tone contention protocol. (a) $N_v = 35$. (b) $N_v = 20$. (c) $N_v = 5$.

We further conduct the throughput comparison between the DAH-MAC and the D-PRMA. It is shown in Fig. 3.14 that the DAH-MAC can achieve both a larger voice capacity region ($N_{vm} = 35$ with $\varphi = 0.33$) and a higher data throughput than the D-PRMA. Since the D-PRMA is designed to support the QoS of voice traffic, the throughput for best-effort data traffic is suppressed because data nodes contend the channel with a lower probability and they can only transmit once in a slot upon the successful contention.

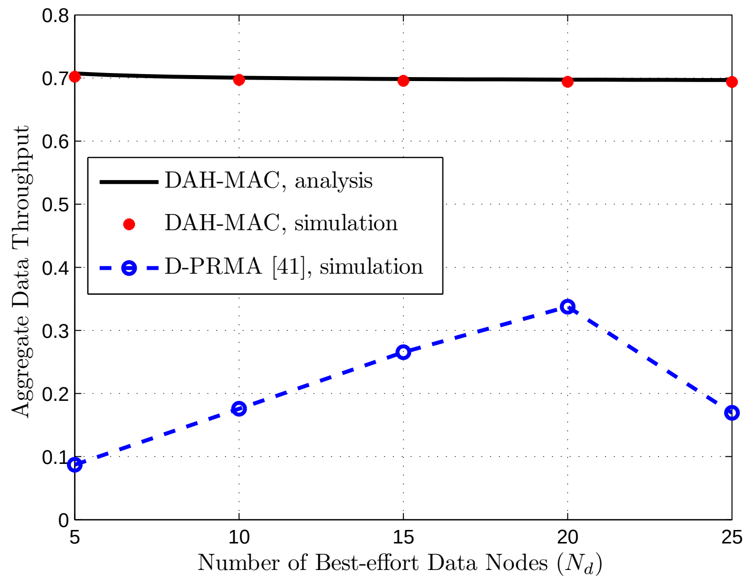


Figure 3.14: A comparison of the DAH-MAC maximum data throughput ($\varphi = 0.33$, $N_v = 10$) with the D-PRMA protocol.

Also, the D-PRMA uses slotted-Aloha based mechanism to contend for the transmission opportunity, which is inferior to the CSMA/CA based mechanism in terms of collisions resolution. We can see that the data throughput starts to decrease when N_d becomes large due to accumulated contention collisions.

3.5 Summary

In this chapter, a distributed and traffic-adaptive hybrid MAC scheme is proposed to support both voice and data traffic in a single-hop MANET. The proposed hybrid MAC adaptively allocates TDMA time slots to active voice nodes for guaranteeing a voice packet loss rate bound, and employs the T-CSMA/CA based contention scheme for data nodes with the contention window size adjusted to the optimal value upon the instantaneous numbers of voice and data nodes in the network for achieving maximum aggregate data throughput. The proposed adaptive hybrid MAC scheme provides differentiated QoS guarantee in the presence of heterogeneous traffic load dynamics.

Chapter 4

Token-Based Adaptive MAC for a Two-Hop IoT-Enabled MANET

In this chapter, a distributed token-based adaptive MAC (TA-MAC) scheme is proposed for a two-hop IoT-enabled MANET. In the TA-MAC, nodes are partitioned into different one-hop node groups, and a TDMA-based superframe structure is proposed to allocate different TDMA time durations to different node groups to overcome the hidden terminal problem. A probabilistic token passing scheme is devised for packet transmissions within different node groups, forming different token rings, which adapts to network load variations with node movement by updating the memberships of each token ring in a distributed way. To optimize the MAC design, performance analytical models are developed in closed-form functions of both the MAC parameters (i.e., the number of token rotation cycles for each token ring and the superframe length) and the network traffic load. Then, an average end-to-end delay optimization framework is established to derive the set of optimal MAC parameters under a certain network load condition. Numerical and simulation results demonstrate that, by adapting the MAC parameters to the varying network condition, the TA-MAC achieves consistently minimal average end-to-end delay, bounded delay for local transmissions, and high aggregate throughput. Further, the performance comparison between the TA-MAC and other MAC schemes show the scalability of the proposed MAC in an IoT-based two-hop environment with an increasing network traffic load.

4.1 System Model

For a multi-hop MANET, the communication distance between a pair of source-destination (S-D) nodes can be larger than the one-hop transmission range. Therefore, some intermediate relay (R) nodes, residing within both transmission ranges of the S-D nodes that are not reachable to each other directly, not only transmit data packets from their own application layers, but may also relay packets between the S-D node pair. Fig. 4.1(a) illustrates a

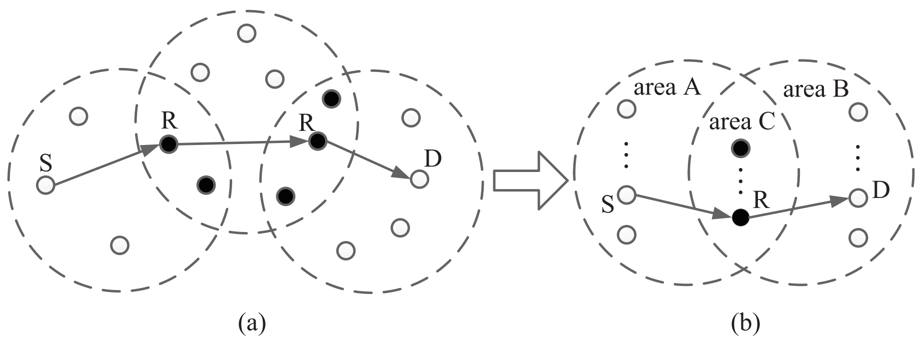


Figure 4.1: (a) A general multi-hop MANET. (b) A simplified two-hop network.

general MANET, where each dashed circle represents a fully-connected network (a one-hop cluster), i.e., nodes in the network are within the one-hop communication range of others. Some nodes, denoted by black dots, staying in the overlapping areas of different one-hop clusters can act as relays to forward packets for other S-D node pairs (denoted by white dots) beyond the direct communication range. Therefore, the basic communication unit in a multi-hop MANET is a two-hop network, and some nodes can be R nodes in addition to S-D nodes. With mobility, nodes can leave one two-hop network and become members of another one. In this chapter, we consider a basic two-hop network model as the first step towards a general multi-hop environment, shown in Fig. 4.1(b). There are three logical areas (A, B and C); Nodes enter or depart from the network coverage region, or move around in the three areas. For a given packet transmission direction, such as from left to right, a node can be an S (D) node or an R node depending on its area in A (B) or C.

Let N denote the total number of nodes in the network which can slowly vary with time due to node mobility, and N_a, N_b and N_c denote the numbers of nodes in areas A, B, and C, respectively. There is a single type of data traffic in the network. The compound traffic arrivals for each R node are different from those of an S (D) node, which consists of not only the traffic arrivals generated from its own application layer but also the relay traffic coming from nodes in both area A and area B [45]. Packet arrivals at each node are split into two traffic streams according to different transmission directions: an arriving packet at each S node in area A (B) is transmitted either to a local D node in the same area or to an end D node two-hop away in area B (A). Similarly, each R node transmits its self-generated packets to a D node in either area A or area B, and also relays packets from area A (B) to a D node in area B (A).

There is a single radio channel in the network, without transmission errors. Nodes access the channel in a distributed manner. We assume that each node is equipped with a global positioning system (GPS) receiver, and the time synchronization among nodes in the network can be achieved by using the 1PPS signal provided by any GPS receiver [40]. Transmission failures can happen due to packet collisions, i.e., more than one transmission attempts are initiated simultaneously by different nodes. Each node has an exclusive node identifier (ID) that can be selected at random and included in each transmitted packet [40]. For a tagged node x, we define the set of node IDs of all one-hop neighbors of node x (including x itself) as $\mathcal{N}(x)$.

In the network, time is divided into a sequence of superframes, and the length of each superframe, denoted by T_f, is determined upon the numbers of nodes in each network area. As shown in Fig. 4.2, we partition each superframe into durations, T_{ac}, T_{bc} and T_{ab}, which consist of, M_{ac}, M_{bc} and M_{ab}, numbers of time slots of equal duration T_s, respectively. Therefore, the duration of each superframe, T_f, is equal to $M \cdot T_s$, where M is the total number of time slots within a superframe. Each time slot can accommodate one data packet transmission, and nodes only transmit packets at the start of each time slot. To resolve the hidden terminal problem [46] [75] for the two-hop network, durations T_{ac} and T_{bc} are reserved for communications between nodes in areas A and C and between nodes in areas B and C, respectively, where the transmitting and receiving nodes of a communication pair are in different areas, as shown in Fig. 4.3 (a)-(b); The last duration, T_{ab}, is reserved for simultaneous communications among nodes in area A and among nodes in area B, where both the transmitting and receiving nodes are in the same area, as shown in Fig. 4.3(c). With this TDMA-type transmission duration reservation for the four node groups forming four one-hop subnetworks, packet collisions caused by hidden terminals can be completely eliminated.

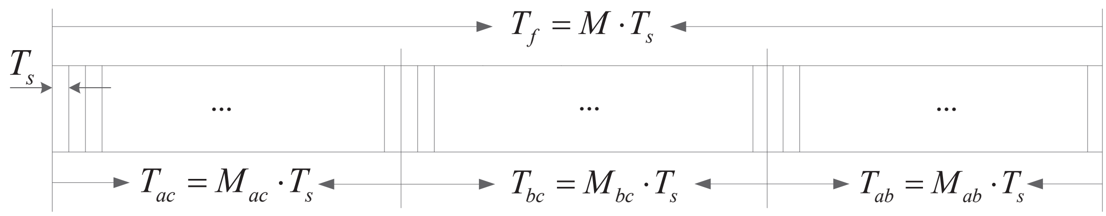


Figure 4.2: Superframe structure.

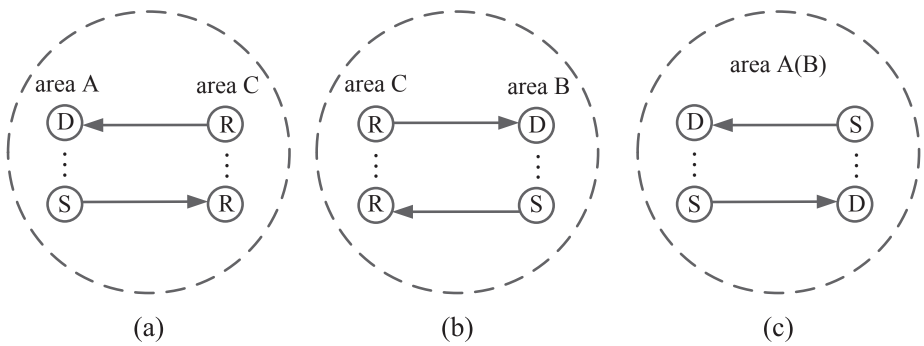


Figure 4.3: Packet transmissions for four one-hop subnetworks during (a) T_{ac}. (b) T_{bc}. (c) T_{ab}.

4.2 The TA-MAC Scheme

4.2.1 Probabilistic Token Passing within Each Node Group

In the TA-MAC, there are four tokens circulated separately among nodes in each group for packet transmissions, forming four token rings, R_{ac}, R_{bc} and R_a (R_b). For each token ring, when a node holds a token, it is assigned a time slot with duration T_s for transmission of either a data packet[1] or a token packet [51] [52], and the current token holder decides which node is the next token holder. We define *one token rotation cycle* as the time duration a token has visited all node members once in token ring R_j ($j = ac, bc, a, b$), which is calculated as $L_j \cdot T_s$, where L_j is the number of node members in token ring R_j. We also define *probabilistic token passing list*, $\mathcal{L}(j)$, as the set of node IDs of all node members in token ring R_j. Each node records a sequence of node IDs that the token has already visited for current token rotation cycle, and the current token holder selects the next token holder with equal probability from those nodes that have not been visited to maintain the

[1]To smooth the delay jitter, we assume that each backlogged node transmits one data packet each time the node holds the token.

fairness of channel access among all node members.

At the beginning of a superframe, a token starts to circulate among nodes in areas A and C during T_{ac}, forming token ring R_{ac}. Once a designated node in area A (C) gets a token, it first waits for the channel to be idle for the duration of T_1 [49], and then piggybacks the token on the head-of-line (HOL) packet (if any) waiting in its queue and transmits the packet to its destination node in area C (A). Note that the destination node (or the next-hop relay node) and the next token holder are not necessarily the same node. If the token holder does not have packets in its queue, it simply passes the token to the next token holder after T_1. When the current token rotation cycle finishes, a new token rotation cycle of token ring R_{ac} starts, conforming to the same token passing rule until the end of T_{ac}. Once the duration T_{ac} elapses, the current token circulation for R_{ac} ceases, and another token starts to circulate among nodes in areas C and B during T_{bc}, forming token ring R_{bc}, which proceeds the same way as in T_{ac}. The token rings R_a and R_b are formed when two tokens are circulated among nodes in area A and among nodes in area B, respectively, during T_{ab}. These two token rings operate simultaneously and independently for the two disjoint one-hop subnetworks in both areas. Therefore, the durations, T_{ac}, T_{bc} and T_{ab}, can be denoted by k_{ac}, k_{bc} and k_a (k_b) token rotation cycles for token rings, R_{ac}, R_{bc} and R_a (R_b), respectively, as shown in Fig. 4.4, indicating the number of times a token is held by each node in each token ring for packet transmissions. Note that the duration T_{ab} can be denoted by k_a token rotation cycles of R_a, or k_b token rotation cycles of R_b (not depicted in Fig. 4.4 for simplicity). It is possible that the numbers of time slots, M_{ac}, M_{bc} and M_{ab}, in each duration are not an integer multiple of the numbers of nodes, L_{ac}, L_{bc} and L_a (L_b), in respective token rings, making the numbers of token rotation cycles, k_j ($j = ac, bc, a, b$), a non-integer. In this case, the number of time slots in the last token rotation cycle, denoted by $M_j - (\lceil k_j \rceil - 1)L_j$ ($\lceil \cdot \rceil$ is the ceiling function), is less than L_j. Since nodes are granted a random time slot in each token rotation cycle based on the probabilistic token passing, each node in token ring R_j is statistically guaranteed to hold the token for k_j times in each superframe if k_j is not an integer. To ensure fair channel access among nodes, the node members in each token ring at least hold the token once in each duration (i.e., $k_j \geq 1$). Both k_j and M are important MAC parameters that affect the performance of the TA-MAC scheme (to be discussed in Subsection 4.2.5).

For any node, say node x, in the network, two types of (data/token) packets are transmitted: Type I and Type II packets, shown in Fig. 4.5. A Type I packet contains a header, a set of IDs of the one-hop neighbors of node x, $\mathcal{N}(x)$, including the probabilistic token passing list, $\mathcal{L}(j)$, for the current token ring R_j, and a payload for either a data packet or a token packet; A Type II packet is composed of a header and a payload. Each node in token ring R_j ($j = ac, bc, a, b$) transmits exactly one Type I packet in the first token

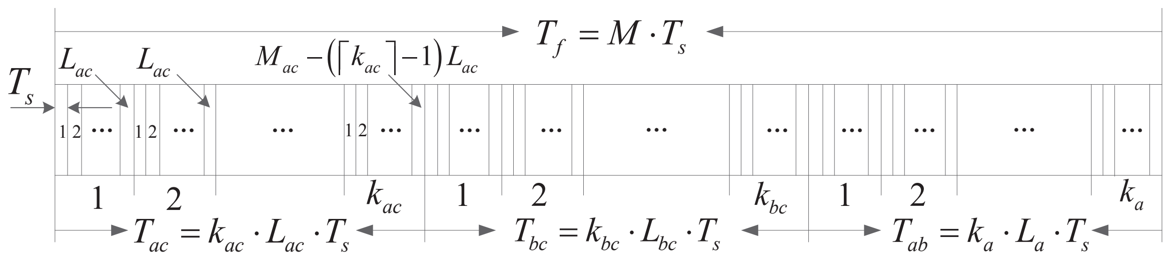


Figure 4.4: Token rotation cycles within T_{ac}, T_{bc} and T_{ab}.

rotation cycle to exchange local information with its two-hop neighbors for detecting (updating) the node location, and for distributedly calculating the durations, T_{ac}, T_{bc} and T_{ab}, in each superframe. If more than one token rotation cycles are scheduled for token ring R_j, Type II packets with a smaller overhead by removing the control information, $\mathcal{N}(x)$, are transmitted within other token rotation cycles for improving the MAC efficiency.

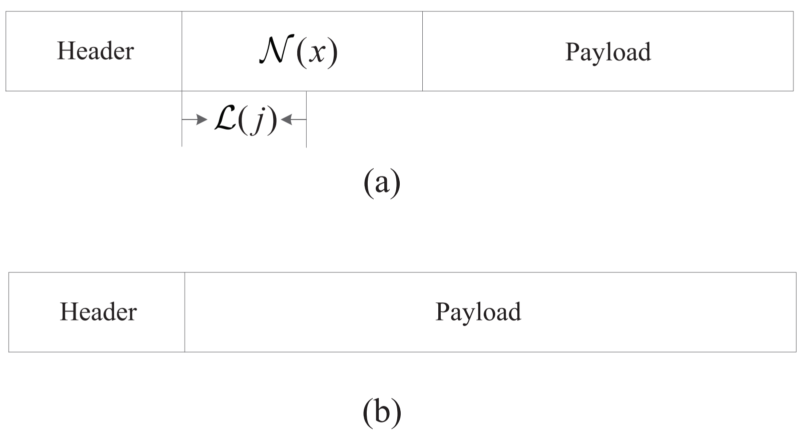


Figure 4.5: Packet types: (a) A Type I packet. (b) A Type II packet.

4.2.2 Nodes Joining/Leaving the Network

A node needs to join corresponding token rings for packet transmissions when entering the network. To do so, it first specifies its location in the two-hop network. Suppose a new node, x, is powered on, and synchronizes in time with its one-hop neighbors. After that, node x listens to packet transmissions on the channel for one superframe duration, from which it obtains $\mathcal{N}(x)$. Then, the node determines that

1) It is an S (D) node in area A or B, if $\exists\ ID_y \in \mathcal{N}(x)\backslash ID_x$, such that $\mathcal{N}(x) \subset \mathcal{N}(y)$, where ID_x and ID_y denote the IDs of node x and node y;

2) It is an R node in area C, if $\forall\ ID_y \in \mathcal{N}(x)\backslash ID_x$, we have $\mathcal{N}(y) \subseteq \mathcal{N}(x)$, and $\exists\ ID_z \in \mathcal{N}(x)\backslash ID_x$, such that $\mathcal{N}(z) \subset \mathcal{N}(x)$.

Furthermore, if node x is an S-D node and can only receive packets from R nodes in area C during T_{ac} (T_{bc}), it is located in area B (A).

After determining its location, node x broadcasts a REQUEST packet (REQUEST packets have a higher priority than data (token) packets) after waiting for the channel to be idle for a duration of T_2 ($< T_1$), to join corresponding token rings. Each REQUEST packet has two important information fields: JOINING and LEAVING, indicating the current network area that the node stays in and the previous area that it departed from. If the node is newly powered on, the LEAVING field is left blank. For instance, node x is powered on in area A, it broadcasts a REQUEST packet within T_{ac}, after sensing an idle channel for T_2, to join the token rings R_{ac} and R_a, respectively. Upon receiving the REQUEST packet, each one-hop neighbor y of x adds ID_x in the set $\mathcal{N}(y)$, and the token holders in R_{ac} and R_a also add ID_x in the probabilistic token passing lists, $\mathcal{L}(ac)$ and $\mathcal{L}(a)$, respectively. Consequently, if subsequent packet transmissions from any node z in R_{ac} and R_a, have $ID_x \in \mathcal{N}(z)$ and $ID_x \in \mathcal{L}(ac)$ and $\mathcal{L}(a)$, the admissions to corresponding token rings become successful.

On the other hand, when node x is expected to drain its power, it sends a REQUEST packet within T_{ac} before leaving area A, with the LEAVING field specifying area A (the JOINING field is left blank). Then, each one-hop neighbor y of x removes ID_x from $\mathcal{N}(y)$, and ID_x is also removed from $\mathcal{L}(ac)$ and $\mathcal{L}(a)$ by current token holders in R_{ac} and R_a, respectively. If node x is the current token holder in R_{ac} (R_a), it also passes the token to the next token holder before departure.

4.2.3 Existing Nodes Moving Across Network Areas

When an existing node, say node x, moves across network areas, its location change can be detected[2]:

1) When moving from area A to C, node x detects its location change within T_{bc} after receiving packets from nodes in area B; Then, it broadcasts a REQUEST packet, with

[2]Because of the geographical symmetry of areas A and B, we only consider S (D) nodes moving between areas A and C. Similar results can be obtained when nodes move between areas B and C.

JOINING and LEAVING fields specifying area C and area A, to join R_{bc} and leave R_a (ID_x is added in $\mathcal{L}(bc)$ and removed from $\mathcal{L}(a)$). ID_x is also added in $\mathcal{N}(y)$ for any node y in area B;

2) When moving from area C to A, node x detects its location change within T_{bc} if no packet transmission activity can be detected from any node in area B. Similarly, a REQUEST packet is broadcast from node x after the location change detection, and ID_x is then added in $\mathcal{L}(a)$ and removed from $\mathcal{L}(bc)$ by the token holders in R_a and R_{bc}, respectively. For token ring R_{bc}, if the current token holder is a node from area C, it removes ID_x from $\mathcal{L}(bc)$ directly upon receiving the REQUEST packet (Any node y in area B also removes ID_x from $\mathcal{N}(y)$ when receiving the updated $\mathcal{L}(bc)$ from the token holder); If the current token holder is a node from area B, it removes ID_x from $\mathcal{L}(bc)$ when it selects node x as the next token holder and no transmission activity is detected within a retransmission timeout (see details in Subsection 4.2.4).

Access collisions happen when more than one nodes, either newly arriving nodes or existing nodes, within the same communication range broadcast REQUEST packets at the same time, which can be detected by the nodes involved when their node IDs are not updated in corresponding token passing lists $\mathcal{L}(i)$ received from subsequent packet transmissions. Some random backoff based collision resolution schemes can be used for the nodes involved before re-broadcasting a REQUEST packet [49]. Since nodes are assumed to move with a low speed (e.g., a walking speed), access collisions rarely happen.

4.2.4 Lost Token Recovery

Occasionally, existing nodes are not aware of a node departure in the following three situations[3]: 1) The REQUEST packet broadcast by a node being powered off is in collision, and new REQUEST packet cannot be re-initiated due to the power depletion; 2) The current token holder cannot correctly receive the broadcast REQUEST packet from a moving node since the communication range exceeds the one-hop distance (e.g., the REQUEST packet broadcast by a node moving from area C to A is not received by the current token holder (if it is a node from area B) in R_{bc}.); 3) The next token holder departs from the network due to node movement.

When one of the preceding situations happens, the token can be lost, which is detected by the previous token holder as there is no packet transmission from the current token

[3]Since nodes move with a low speed, we assume that a token holder can pass the token to the next token holder before moving to a new network area.

holder. Then, the previous token holder enters into a token recovery process, in which it regenerates and passes a new token to the same current token holder for a maximum of N_{re} times [48] [51]. If there is still no transmission activity discovered before N_{re} is reached, a retransmission timeout is triggered and the previous token holder resends the token to a new node, with the old one removed from the probabilistic token passing list.

4.2.5 Important MAC Parameters

We consider the average end-to-end delay as the main performance metric to evaluate the effectiveness of the proposed MAC scheme for the two-hop network, which is defined as the summation of the delay from the time a packet arrives at an S node in area A (B) to the time it is received by an R node, averaged over all transmitted packets for the same transmission direction from area A (B) to C, and the delay from the moment a packet reaches a selected R node to the moment it is received by its D node in area B (A), averaged over all transmitted packets for the same transmission direction from area C to B (A). To achieve minimum average end-to-end packet delay[4], each node member in token ring R_j ($j = ac, bc, a, b$) is expected to determine the number of times it should get the token for packet transmissions, i.e., the number of token rotation cycles k_j ($j = ac, bc, a, b$), during T_{ac}, T_{bc} and T_{ab} in each superframe. Therefore, the design objective of the TA-MAC lies in determining k_j for respective token ring R_j in each superframe, which also indicates the number of time slots allocated to each node member of token ring R_j.

A small k_j gives better time slot utilization in a token rotation cycle, i.e., the percentage of nonempty time slots in a token rotation cycle, due to the increased node queue utilization ratios, but can result in a longer transmission queue length for each node. Thus, with an increase of k_j, the packet delay is expected to be reduced due to more resource reservation for high service capability. However, excessive resource reservation for one token ring lowers the delay for single-hop packet transmissions within the token ring, but shrinks the time slot resources for other token rings, thus increasing the packet delay for other transmission hops. Moreover, an excessive k_j prolongs the length of each superframe, which may cause the increase of average packet service time for each node, which is defined as the duration from the instant that a packet arrives at the head of a node queue to the instant it is successfully transmitted, averaged over all transmitted packets from the node. Therefore, to minimize the average end-to-end delay, first, time slot allocation for one individual token ring should be balanced with the others; Second, for the purpose of using a minimum total

[4]The reason of considering the average end-to-end delay minimization instead of the worst-case delay guarantee is to utilize the network resources more efficiently to achieve a larger network capacity.

amount of time resources to achieve the lowest average end-to-end delay, the total amount of time slots for each superframe, M, should also be optimized and adaptive to the varying number of nodes. We aim at finding the optimal number of token rotation cycles, k_j^{opt}, scheduled for token ring R_j ($j = ac, bc, a, b$), along with the optimal total number of time slots, M^{opt}, for each superframe, with a varying network load, to achieve the minimal average end-to-end delay, D^{opt}, under the constraints that each node queue is stable and the delay for local packet transmissions is bounded by an acceptable threshold. With an increase of the numbers of nodes in each network area, M^{opt} is expected to consistently increase and then remain stable in high traffic load conditions, making k_j^{opt} decrease with the network load, since the superframe length should be controlled to keep a low packet service time and a low overall packet delay. The set of optimal MAC parameters, k_j^{opt} and M^{opt}, are distributedly calculated by each node based on the detection of current traffic load conditions for all the three network areas, which are also dynamically updated upon variations of the number of nodes in each area. According to $\left[k_j^{opt}, M^{opt}\right]$, nodes can also determine the optimal durations for T_{ac}, T_{bc} and T_{ab} in each superframe.

4.3 Performance Analysis

In this section, we develop performance analytical models for the TA-MAC scheme in closed-form functions of k_j and M for each superframe.

4.3.1 Compound Packet Arrival Rate

Traffic arrivals at each node are modeled as a Poisson process with an average arrival rate λ packet/slot [49]. Each node in area A (B) transmits a packet either to a randomly-selected local destination node in the same area during T_{ab}, or to a random relay node in area C during T_{ac} (T_{bc}). Thus, the traffic arrivals at each node in area A (B) are split into two streams with the average arrival rates denoted by λ_a (λ_b) and λ_{ac} (λ_{bc}) for each transmission direction. For analysis simplicity, we assume that every packet generated from an S node's own application layer is transmitted equally likely for both directions. Thus, λ_a (λ_b) and λ_{ac} (λ_{bc}) are equal to $\frac{\lambda}{2}$ packet/slot. On the other hand, traffic arrivals at each relay node consist of two portions: 1) traffic generated at the node's own application layer and destined for a node either in area A or in area B equally as assumed for an S node, with the average arrival rate $\frac{\lambda}{2}$ packet/slot for both transmission directions; 2) the relay traffic received from nodes in area A (B) during T_{ac} (T_{bc}), which will be forwarded to a destination node chosen at random in area B (A) during T_{bc} (T_{ac}).

To analyze the average end-to-end delay from a source node in area A (B) to a destination node in area B (A) as well as the aggregate throughput for the two-hop network, a network of queues should be established. However, the exact relay traffic arrival process at each node in area C, consisting of the superposition of the departure processes from the nodes in area A (B), is difficult to be precisely modeled [68]. Therefore, inspired by [45], we approximate the relay traffic arrival process at a relay node as a single Poisson process with rate parameter, λ_{ar} (λ_{br}), being the sum of the traffic arrival rates heading to the common relay node from nodes in area A (B), as shown in Fig. 4.6. Since each source

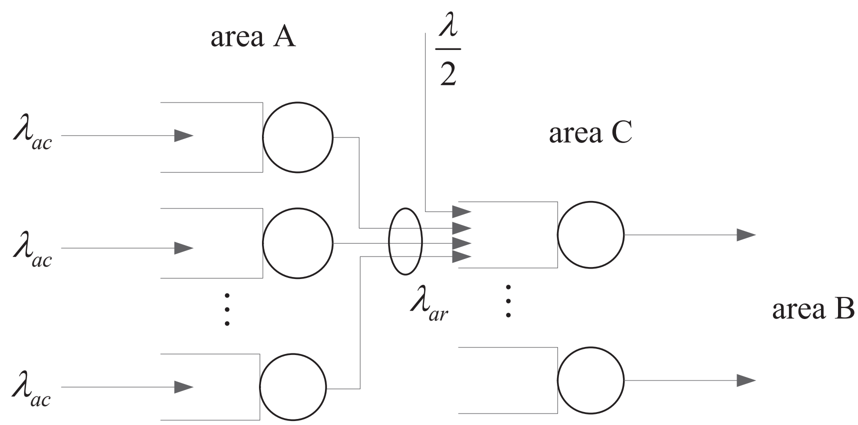


Figure 4.6: Poisson approximation for relay traffic arrival rate for transmission direction from area C to B.

node selects its relay from N_c nodes with equal probability, the compound traffic arrival rate at each relay node with the combination of the external traffic and the relay traffic for the transmission direction from area C to B or from area C to A, denoted by λ_{cb} or λ_{ca}, respectively, is approximated as

$$\lambda_{cb} \approx \frac{\lambda}{2} + \lambda_{ar} = \frac{\lambda}{2} + \frac{\lambda N_a}{2N_c} \qquad (C \to B) \tag{4.1}$$

or

$$\lambda_{ca} \approx \frac{\lambda}{2} + \lambda_{br} = \frac{\lambda}{2} + \frac{\lambda N_b}{2N_c} \qquad (C \to A)\,. \tag{4.2}$$

As an extension to *Kleinrock independence approximation* [68], this Poisson traffic approximation on each relay node is effective in analytically modeling the two-hop network as a network of M/G/1 queues for evaluating the average end-to-end delay. To justify the accuracy of this approximation, the analytical results are further compared with the simulation results in Subsection 4.5.2. The approximation becomes more accurate when the network traffic load increases [68] (i.e., a large number of nodes in each area with increased queue utilization ratios for each node), under which the TA-MAC scheme can also achieve high channel utilization.

4.3.2 Average Packet Service Time

Next, we calculate average packet service time for head-of-line (HOL) packet transmissions. Packet service time (in the unit of one time slot), $W_{s,j}$, for a node in token ring R_j ($j = ac, bc, a, b$), is defined as the duration from the instant that a packet arrives at the head of the node queue to the instant it is successfully transmitted. We use index q ranging from 1 to L_j to indicate the end instant of each time slot in a token rotation cycle of R_j. Taking nodes in token ring R_{ac} as an example, since there are k_{ac} token rotation cycles scheduled within each T_{ac}, we assume that HOL packets from a tagged node x in area A and area C can appear at the end of each allocated time slot in any of k_{ac} token rotation cycles along each T_{ac}, as shown in Fig. 4.7. For analysis tractability, we neglect the possibility that an HOL packet arrives within the duration of a time slot, and the possibility that a packet arrives at a node in R_{ac} and finds the node queue empty during T_{bc} and T_{ab} [67], which is more accurate for a higher traffic load condition. We make the same assumption for HOL packet arrivals at nodes in other token rings. Define a random

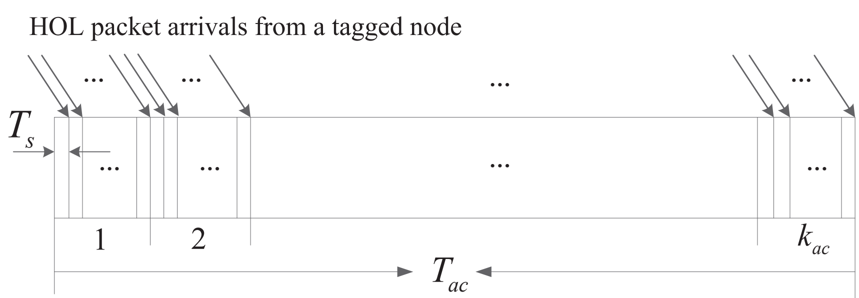


Figure 4.7: HOL packet arrivals along T_{ac} from a tag node in R_{ac}.

variable H_j, such that $H_j = 0$ denotes an HOL packet from a tagged node in R_j appears during the k_jth token rotation cycle, and $H_j = 1$ denotes an HOL packet arrives within other token rotation cycles. Since each node member randomly occupies a transmission slot in corresponding token passing sequence of each token rotation cycle, the service times for consecutive packet transmissions are independent and identically distributed (i.i.d) random variables [45], which is a necessary condition for the M/G/1 queue modeling for each node. Thus, the HOL packet can be transmitted by accessing one of the time slots with equal probability in the next token rotation cycle following the packet arriving instant. Therefore, we derive the average packet service time, denoted by $E[W_{s,j}]$, by considering the following two cases:

(1) When $H_j = 1$, the probability mass function (pmf) of $W_{s,j}$ conditioned on the arriving time instant Q of the HOL packet is expressed as

$$P\{W_{s,j} = i|Q = q, H_j = 1\} = \frac{1}{L_j} \quad (L_j - q + 1 \leq i \leq 2L_j - q,\ 1 \leq q \leq L_j), \tag{4.3}$$

where $L_{ac} = N_a + N_c$; $L_{bc} = N_b + N_c$; $L_a = N_a$; $L_b = N_b$, and the expectation of $W_{s,j}$ conditioned on $H_j = 1$ can be derived as

$$\begin{aligned} E[W_{s,j}|H_j = 1] &= \sum_{i=L_j-q+1}^{2L_j-q} \sum_{q=1}^{L_j} iP\{W_{s,j} = i|Q = q, H_j = 1\}P\{Q = q|H_j = 1\} \\ &= L_j; \end{aligned} \tag{4.4}$$

(2) When $H_j = 0$, we similarly have the conditional pmf and the conditional expectation of $W_{s,j}$ as

$$P\{W_{s,j} = i|Q = q, H_j = 0\} = \frac{1}{L_j} \quad (M'_j - q + 1 \leq i \leq M'_j - q + L_j,\ 1 \leq q \leq L_j) \tag{4.5}$$

where $M'_j = M - (k_j - 1)L_j$, and

$$\begin{aligned} E[W_{s,j}|H_j = 0] &= \sum_{i=M'_j-q+1}^{M'_j-q+L_j} \sum_{q=1}^{L_j} iP\{W_{s,j} = i|Q = q, H_j = 0\}P\{Q = q|H_j = 0\} \\ &= M'_j. \end{aligned} \tag{4.6}$$

Therefore, the average service time for each HOL packet from nodes in R_j can be derived as

$$\begin{aligned} E[W_{s,j}] &= E[W_{s,j}|H_j = 1]P\{H_j = 1\} + E[W_{s,j}|H_j = 0]P\{H_j = 0\} \\ &= L_j \cdot \left(1 - \frac{1}{k_j}\right) + M'_j \cdot \frac{1}{k_j} \\ &= \frac{M}{k_j}, \quad j = ac, bc, a, b. \end{aligned} \tag{4.7}$$

We also define the average packet service rate, μ_j, as the number of packets transmitted per slot time from a node in R_j. Thus, we have

$$\mu_j = \frac{1}{E[W_{s,j}]} = \frac{k_j}{M}, \quad j = ac, bc, a, b. \tag{4.8}$$

4.3.3 Aggregate Network Throughput

Since each time slot can accommodate one data packet transmission, we define the aggregate network throughput as the ratio of average number of transmitted packets over total number of time slots in each superframe, which is also the aggregate time slot utilization within each superframe,

$$S = \frac{1}{M}\left(k_{ac}\frac{N_a\lambda_{ac} + N_c\lambda_{ca}}{\mu_{ac}} + k_{bc}\frac{N_b\lambda_{bc} + N_c\lambda_{cb}}{\mu_{bc}} + \sum_{j\in\{a,b\}} k_j\frac{N_j\lambda_j}{\mu_j}\right). \tag{4.9}$$

From (4.7) and (4.9), the aggregate network throughput, S, is a function of the number of nodes in each area and the traffic arrival rate at each node, which does not vary with k_j and M. Actually, the variations of k_j and M affect the average packet service rate μ_j for each node in R_j and the time slot utilization within each token rotation cycle, resulting in a different packet delay. But the aggregate channel utilization within each superframe remains unchanged with variations of k_j and M. Therefore, higher aggregate network throughput is expected to be achieved with more nodes in each individual network area.

4.3.4 Average End-to-End Delay

The average end-to-end delay in the two-hop network for each transmission direction, either from area A to B or from area B to A, denoted by D_{ab} or D_{ba}, consists of the summation of two portions: the average delay a packet experiences from the instant it arrives at a source node in area A (B) to the instant it is received by a relay node, denoted by D_{ac} (D_{bc}), and the average delay when a packet arrives at the transmission queue of a selected relay node to the moment it is received by a destination node in area B (A), denoted by D_{cb} (D_{ca}). Each portion for a packet is further composed of the average queueing delay, i.e., the average duration the packet stays in the transmission queue after its arrival, and the average service time.

We derive the second moment of the packet service time $W_{s,j}$ for each node member in token ring R_j ($j = ac, bc, a, b$) as

$$E[W_{s,j}^2] = E[W_{s,j}^2 | H_j = 1]P\{H_j = 1\} + E[W_{s,j}^2 | H_j = 0]P\{H_j = 0\}. \tag{4.10}$$

Then, based on *P-K formula* [68], the average end-to-end delay, D_{ab} (D_{ba}), for either transmission direction, and the average delay for local transmissions within area A (B), denoted by D_a (D_b), can be derived as

$$\begin{aligned} D_{ab} &= D_{ac} + D_{cb} \\ &= \sum_{(n,j)\in\{(ac,ac),(cb,bc)\}} \left(E[W_{s,j}] + \frac{\lambda_n E[W_{s,j}^2]}{2\left(1 - \lambda_n E[W_{s,j}]\right)} \right) \\ &= \sum_{(n,j)\in\{(ac,ac),(cb,bc)\}} \left(\frac{\varepsilon_j}{k_j} + \frac{\lambda_n[\alpha_j k_j^2 + \beta_j k_j + \gamma_j]}{2(k_j - \lambda_n \varepsilon_j)} \right), \end{aligned} \tag{4.11}$$

$$D_{ba} = D_{bc} + D_{ca} = \sum_{(n,j)\in\{(bc,bc),(ca,ac)\}} \left(\frac{\varepsilon_j}{k_j} + \frac{\lambda_n[\alpha_j k_j^2 + \beta_j k_j + \gamma_j]}{2(k_j - \lambda_n \varepsilon_j)} \right), \tag{4.12}$$

and

$$D_j = \frac{\varepsilon_j}{k_j} + \frac{\lambda_j[\alpha_j k_j^2 + \beta_j k_j + \gamma_j]}{2(k_j - \lambda_j \varepsilon_j)} \quad (j = a, b) \tag{4.13}$$

where $\alpha_j = L_j^2$; $\beta_j = -\frac{5L_j^2 + 12ML_j + 1}{6}$; $\gamma_j = M^2 + 2ML_j$; $\varepsilon_j = M$.

From (4.11) to (4.13), it is observed that with a certain number of nodes in each area, both the average end-to-end delay and average delay for local transmissions are functions of k_j and M. As stated in Subsection 4.2.5, with a certain M value, an increased k_j for one token ring reduces the one-hop average packet delay among its node members, and also the time resources for other token rings, which can increase the average delay for other transmission hops. Thus, the numbers of token rotation cycles scheduled for each token ring should be balanced to minimize the average end-to-end delay. At the same time, the total number of time slots, M, for each superframe should be properly chosen to further improve the delay performance. Therefore, our design objective lies in how to determine k_j and M to achieve the minimal average end-to-end delay.

4.4 Optimal MAC Parameters

In this section, we propose an optimization framework to find the set of optimal MAC parameters, $\left[k_j^{opt}, M^{opt}\right]$, for each superframe, with which the minimal average end-to-end delay, D^{opt}, is achieved, and a bounded average delay, D_{th}, for local transmissions is guaranteed.

4.4.1 Average End-to-End Delay Minimization

An average end-to-end delay minimization problem is formulated as (**P1**), to derive the set of optimal numbers of token rotation cycles, $\mathbf{k}^* = [k_{ac}^*, k_{bc}^*, k_a^*, k_b^*]$, for each token ring, for a given M.

$$(\mathbf{P1}): \min_{\mathbf{k}=[k_{ac},k_{bc},k_a,k_b]} \{\max\{D_{ab}(k_{ac}, k_{bc}), D_{ba}(k_{ac}, k_{bc})\}\}$$

$$\text{s.t.} \begin{cases} k_{ac}L_{ac} + k_{bc}L_{bc} + k_a L_a = M & \text{(4.14a)} \\ k_a L_a = k_b L_b & \text{(4.14b)} \\ \rho_n = \dfrac{\lambda_n}{\mu_j} < 1 \quad (n, j) \in \{(ca, ac), (cb, bc)\} & \text{(4.14c)} \\ \rho_j = \dfrac{\lambda_j}{\mu_j} < 1 \quad (j = a, b) & \text{(4.14d)} \\ D_j(k_j) \leq D_{th} \quad (j = a, b) & \text{(4.14e)} \\ k_j \geq 1 \quad (j = ac, bc, a, b). & \text{(4.14f)} \end{cases}$$

In (**P1**), the objective is to minimize the average end-to-end delay for both transmission directions from area A to B and from area B to A, by finding the set of optimal numbers of token rotation cycles for each token ring in each superframe. Constraints (4.14a) - (4.14b) indicate the total number of time slots for each superframe is M, and the time slots allocated to nodes in area A and area B are balanced based on the numbers of nodes in both areas. Constraint (4.14c) guarantees each relay node queue in token rings R_{ac} and R_{bc} be stable, where $\rho_n = \frac{\lambda_n}{\mu_j}$ denotes queue utilization ratios for relay nodes in area C[5]. Similarly, ρ_j in Constraint (4.14d) denotes queue utilization ratios for nodes in token rings R_a and R_b. Constraint (4.14e) states that the average delays for local packet

[5]Since traffic arrival rates at each relay node are greater than at each source node, queue utilization ratios of source nodes in R_{ac} and R_{bc} are also guaranteed stable by constraint (4.14c).

transmissions within both area A and area B are guaranteed below a bound. Constraint (4.14f) guarantees that the node members in each token ring at least hold the token once in each superframe to ensure the fair channel access.

The main difficulty to solve (**P1**) is the non-convexity of the objective function in terms of the decision variables vector $\mathbf{k} = [k_{ac}, k_{bc}, k_a, k_b]$. Therefore, to make the problem tractable, we decouple (**P1**) into a convex subproblem and a biconvex subproblem [76] [77] with two separate sets of decision variables. By solving these two subproblems sequentially, the original problem can be tackled. First, we introduce an important proposition and its corollary.

Proposition 2 *In (***P1***), the one-hop average delays, D_{ac} (D_{ca}), D_{bc} (D_{cb}), and D_a (D_b), for packet transmissions from area A (C) to area C (A), from area B (C) to area C (B), and within the local area A (B), are all strictly convex functions in terms of k_{ac}, k_{bc}, and k_a (k_b), respectively.*

The proof of Proposition 2 is given in Appendix 4.7.1.

Corollary 1 *In (***P1***), the one-hop average delays, D_{ac} (D_{ca}), D_{bc} (D_{cb}), and D_a (D_b), are all strictly decreasing functions of k_{ac}, k_{bc}, and k_a (k_b), respectively.*

The proof of Corollary 1 is given in Appendix 4.7.2.

According to **Corollary 1**, each one-hop average delay is a decreasing function of k_j for corresponding token ring. Thus, to minimize the average end-to-end delay, the maximum amount of time slots are expected to be allocated to node members in token rings R_{ac} and R_{bc} among all feasible solutions for (**P1**). In other words, the minimum number of time slots should be reserved for the token rings R_a and R_b under the constraints (4.14b) - (4.14f). Therefore, to tackle (**P1**) efficiently, we first solve the following subproblem (**SP1**) to obtain the set of optimal numbers of token rotation cycles, $\mathbf{k_1^*} = [k_a^*, k_b^*]$, scheduled for token rings R_a and R_b, respectively.

$$(\mathbf{SP1}): \min_{\mathbf{k_1}=[k_a,k_b]} k_a L_a$$

$$\text{s.t.} \begin{cases} k_a L_a = k_b L_b & \text{(4.15a)} \\ \rho_j < 1 \quad (j = a, b) & \text{(4.15b)} \\ D_j(k_j) \le D_{th} \quad (j = a, b) & \text{(4.15c)} \\ k_j \ge 1 \quad (j = a, b). & \text{(4.15d)} \end{cases}$$

We further simplify (**SP1**) as (**SP1a**) with a single decision variable k_a, by substituting the constraint (4.15a) into the constraints (4.15b) - (4.15d).

$$(\mathbf{SP1a}) : \min_{k_a} k_a L_a$$

$$\text{s.t.} \begin{cases} \rho_a < 1 & (4.16a) \\ \rho_b \left(\dfrac{k_a L_a}{L_b}\right) < 1 & (4.16b) \\ D_a(k_a) \leq D_{th} & (4.16c) \\ D_b \left(\dfrac{k_a L_a}{L_b}\right) \leq D_{th} & (4.16d) \\ \dfrac{\max\left\{1, \frac{L_b}{L_a}\right\}}{k_a} \leq 1. & (4.16e) \end{cases}$$

In (**SP1a**), the objective function and the left-hand sides of the inequality constraints (4.16a) - (4.16c) and (4.16e) are all convex functions of k_a. For constraint (4.16d), the function $D_b\left(\frac{k_a L_a}{L_b}\right)$ is a composite function of k_a, where $D_b(\cdot)$ is a convex and decreasing function of k_b and $k_b = \frac{k_a L_a}{L_b}$ is a linear function of k_a. Therefore, according to the *scalar composition rules* [65], $D_b\left(\frac{k_a L_a}{L_b}\right)$ is also a convex function of k_a. Hence, (**SP1a**) is proved to be a convex optimization problem, which can be efficiently tackled to get the optimal solution k_a'. Since the total number of time slots reserved for R_a or R_b is required to be integer, we obtain the optimal numbers of token rotation cycles for each of the token rings as

$$k_a^* = \frac{\lceil k_a' L_a \rceil}{L_a} \tag{4.17}$$

and

$$k_b^* = \frac{\lceil k_a' L_a \rceil}{L_b}. \tag{4.18}$$

Note that k_a^* and k_b^* are guaranteed in the feasible set of (**SP1a**), since all the inequality constraint functions of (**SP1a**) are decreasing functions of the decision variable k_a.

By substituting the optimal set of values $[k_a^*, k_b^*]$ into (**P1**), the original optimization problem is reduced to the second subproblem (**SP2**).

$$(\mathbf{SP2}) : \min_{\mathbf{k_2}=[k_{ac}, k_{bc}]} \{\max\{D_{ab}(k_{ac}, k_{bc}), D_{ba}(k_{ac}, k_{bc})\}\}$$

$$\text{s.t.} \begin{cases} k_{ac}L_{ac} + k_{bc}L_{bc} = M^* & \text{(4.19a)} \\ \rho_n < 1 \quad (n = ca, cb) & \text{(4.19b)} \\ k_j \geq 1 \quad (j = ac, bc) & \text{(4.19c)} \end{cases}$$

where $M^* = M - k_a^* L_a$.

Theorem 1 *In (***SP2***), the two-dimensional decision variable vector* $\mathbf{k_2}$ *represents a biconvex set, if for any fixed* k_{ac} *(*k_{bc}*) from the feasible solutions,* $\mathbf{k_2}$ *is a convex set with respect to* k_{bc} *(*k_{ac}*).*

Proof: If we fix k_{ac} (k_{bc}) in $\mathbf{k_2}$ and rewrite the set of constraints of (**SP2**) in a standard form, the equality constraint function is an affine function of k_{bc} (k_{ac}), and both inequality constraint functions are convex functions of k_{bc} (k_{ac}). Therefore, the set of feasible solutions of k_{bc} (k_{ac}) satisfying all these constraints form a convex set [65].

Theorem 2 *In (***SP2***), the objective function defined on the biconvex set* $\mathbf{k_2}$ *represents a biconvex function, if for any fixed* k_{ac} *(*k_{bc}*) from the feasible solutions, the objective function is a convex function in terms of* k_{bc} *(*k_{ac}*).*

Proof: If we fix k_{ac} (k_{bc}) in $\mathbf{k_2}$, both functions D_{ab} and D_{ba} are regarded as a linear combination of a convex function in terms of k_{bc} (k_{ac}) and a constant, which are also convex. Moreover, the *max* function, $\max\{x, y\}$, which is proved to be convex on $\mathbf{R}^2$ in [65], is also nondecreasing in each of its two arguments. Therefore, according to the *vector composition rules* [65], the objective function $\max\{D_{ab}, D_{ba}\}$ is a convex function with respect to k_{bc} (k_{ac}).

Based on **Theorem 1** and **Theorem 2**, (**SP2**) is a biconvex optimization problem, since we have a biconvex objective function minimized over a biconvex set, which often has multiple local optima and is difficult to determine the global optimal solution [78]. Therefore, to solve (**SP2**) efficiently, we further simplify (**SP2**) into a single variable optimization problem (**SP2a**), by substituting the equality constraint (4.19a) into the objective function and other constraints.

$$(\mathbf{SP2a}) : \min_{k_{ac}}\{\max\{D_{ab}(k_{ac}, h(k_{ac})), D_{ba}(k_{ac}, h(k_{ac}))\}\}$$

$$\text{s.t.} \begin{cases} \rho_{ca} < 1 & \text{(4.20a)} \\ \rho_{cb}(h(k_{ac})) < 1 & \text{(4.20b)} \\ 1 \leq k_{ac} \leq \dfrac{M^* - L_{bc}}{L_{ac}} & \text{(4.20c)} \end{cases}$$

where $h(k_{ac}) = k_{bc} = \frac{M^* - k_{ac}L_{ac}}{L_{bc}}$.

Proposition 3 *(***SP2a***) is a convex optimization problem, with respect to the decision variable k_{ac}.*

The proof of Proposition 3 is given in Appendix 4.7.3.

Based on **Proposition 3**, the convex optimization problem (**SP2a**) can also be efficiently solved to get the optimal solutions k'_{ac} and k'_{bc}. Similar to (4.17) and (4.18), we further obtain the optimal numbers of token rotation cycles, k^*_{ac} and k^*_{bc}, for R_{ac} and R_{bc} respectively, by rounding $k'_{ac}L_{ac}$ to the integer number (within the feasible region) that achieves the minimum value of the objective function in (**SP2a**).

At this point, by tackling a sequential of tractable subproblems, we solve the original optimization problem (**P1**) to get the optimal numbers of token rotation cycles, $\mathbf{k}^* = [k^*_{ac}, k^*_{bc}, k^*_a, k^*_b]$, scheduled for corresponding token rings in each superframe under a predefined M, upon which the average end-to-end packet delay is minimized as D^*.

4.4.2 Optimal Total Number of Time Slots for Each Superframe

In Subsection 4.4.1, the average end-to-end delay is minimized as D^* for a given M. As stated in Subsection 4.2.5, D^* also varies with respect to M, which is expected to be reduced with an increase of M, due to more time slots reserved for each token ring. However, if M is set too large, excessive resource reservation prolongs the superframe length, causing an increase of packet delay. Therefore, we aim at determining the optimal total number of time slots, M^{opt}, for each superframe, with the optimal number of token rotation cycles, k_j^{opt}, for token ring R_j ($j = ac, bc, a, b$), upon which the minimal average end-to-end delay, D^{opt}, can be achieved. To this end, we propose an optimal superframe length calculation algorithm, as shown in Algorithm 3, for each node to distributedly determine and update the set of optimal MAC parameters, $\left[k_j^{opt}, M^{opt}\right]$, for the TA-MAC upon variations of the number of nodes in the network. The procedure of the algorithm is stated as follows:

Step 1. M is initialized as the minimum value to satisfy constraint (4.14f) in (**P1**). Both D^{opt} and M_s (the minimum value of M to make (**P1**) feasible) are initialized, and the maximum iteration limit is set to a large number;

Step 2. The sequential subproblems (**SP1a**) and (**SP2a**) are repeatedly solved by increasing M with the increment of one time slot in each iteration until a set of feasible solutions, $[k^*_j, D^*]$, for (**P1**) are found at $M = M_s$ (We assume that an efficient admission

Algorithm 3: The optimal superframe length calculation algorithm

Input : L_{ac}, L_{bc}, L_a, L_b, λ, D_{th}.
Output: k_j^{opt}, D^{opt}, M^{opt}, and T_f^{opt}.

Initialization: $M \leftarrow L_{ac} + L_{bc} + \max\{L_a, L_b\}$, $D^{opt} \leftarrow +\infty$, $M_s \leftarrow 0$, set the maximum iteration limit;
do
    $[k_j^*, D^*] \leftarrow$ solving (**SP1a**) and (**SP2a**);
    if *No feasible solutions are found* **then**
        if *The maximum iteration limit is reached* **then**
            break;
        end
        $M \leftarrow M + 1$;
    else
        $M_s \leftarrow M$;
        break;
    end
while *(***P1***) is not feasible*;
if $M_s > 0$ **then**
    while *The maximum iteration limit is not reached* **do**
        if $D^* < D^{opt}$ **then**
            $k_j^{opt} \leftarrow k_j^*$;
            $D^{opt} \leftarrow D^*$;
            $M^{opt} \leftarrow M$;
        end
        $M \leftarrow M + 1$;
        $[k_j^*, D^*] \leftarrow$ solving (**SP1a**) and (**SP2a**);
    end
    $T_f^{opt} \leftarrow M^{opt} \cdot T_s$;
    return k_j^{opt}, M^{opt}, T_f^{opt}, and D^{opt}.
end

control scheme for controlling the numbers of nodes in each network area is implemented so that a feasible M_s is always found within the network capacity.);

Step 3. Starting from a feasible M_s, we iteratively search for M^{opt} and k_j^{opt} to achieve the minimal average end-to-end delay D^{opt}, by continuously increasing M and solving (**SP1a**) and (**SP2a**) at each updated M until the maximum iteration limit is reached.

4.5 Numerical Results

In this section, numerical and simulation results are presented to verify the accuracy of the performance analytical results. All the simulations are carried out using OMNeT++ [69] [79]. In the simulation, nodes are scattered over a 150m $\times$ 150m square region, forming a two-hop network with three network areas similar to that shown in Fig. 4.1(b), where nodes within the transmission ranges (set to 50m) of all other nodes can relay traffic from the S-D node pairs that are not reachable to each other directly. For each transmitted packet, the source node randomly selects a next-hop node or a destination node, according to the packet's destination area, among a specific group of nodes. External traffic arrivals for each node are modeled as a Poisson process with the average rate of 0.01 packet/slot (10 packet/s). The delay bound for local packet transmissions within area A and area B is set as 400 ms. Each simulation point is generated by running the simulation for 10000 superframe rounds. The main simulation parameters are summarized in Table 4.1.

Table 4.1: Simulation parameters

MAC schemes / Parameters	**TA-MAC**	**LA-MAC [25]**	**DTSA [18] [25]**
Channel capacity	11Mbps	11Mbps	11Mbps
Time slot duration	1 ms	1 ms	1 ms
Idle duration (T_1/T_2)	20/10 μs	n.a.	n.a.
Node ID	7 bit	7 bit	7 bit
Data/token packet duration	0.96/0.24 ms	0.96 ms/n.a.	0.96 ms/n.a.
Packet arrival rate (λ)	0.01 packet/slot	0.01 packet/slot	0.01 packet/slot
Backoff slot duration	n.a.	20 μs	n.a.
Minimum contention window (slot owner)	n.a.	1	n.a.
Maximum contention window (slot owner)	n.a.	8	n.a.
Minimum contention window	n.a.	9	n.a.
Maximum contention window	n.a.	16	n.a.
State switch threshold	n.a.	5	n.a.
Retransmission limit	n.a.	7	n.a.
High contention (HC) state duration	n.a.	100 superframe rounds	n.a.
Transmission queue length	10000 packets	10000 packets	10000 packets
Delay bound for local transmissions (D_{th})	400 ms	n.a.	n.a.

By solving the sequential subproblems (**SP1a**) and (**SP2a**) using Algorithm 3, we determine the optimal total number of time slots, M^{opt}, for each superframe, and the optimal number of token rotation cycles, k_j^{opt}, scheduled for token ring R_j ($j = ac, bc, a, b$), upon which the minimal average end-to-end delay D^{opt}, the bounded average delays for local transmissions, D_a and D_b, in area A and area B, can be achieved by the TA-MAC scheme. Then, we analyze D^{opt}, D_a and D_b, and the aggregate network throughput with respect to variations of the network traffic load. Lastly, the TA-MAC scheme is compared with a hybrid MAC scheme and a dynamic TDMA scheme in terms of delay and throughput over a wide range of network traffic load.

4.5.1 Optimal MAC Parameters

In Fig. 4.8, with certain numbers of nodes in each network area, the optimal number of token rotation cycles, k_j^*, scheduled for token ring R_j ($j = ac, bc, a, b$) under different M is demonstrated by solving (**SP1a**) and (**SP2a**) sequentially. We can see that k_j^* increases with M, and more token rotation cycles are scheduled for token rings R_{ac} and R_{bc} to minimize the average end-to-end delay. The set of optimal MAC parameters, $\left[M^{opt}, k_j^{opt}\right]$, is also obtained, based on Algorithm 3, to achieve the minimal average end-to-end delay.

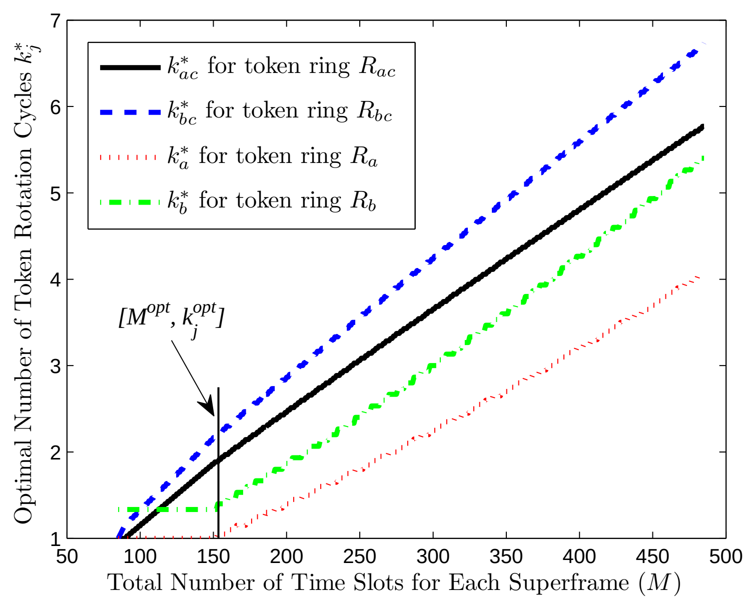


Figure 4.8: Optimal number of token rotation cycles k_j^* for token ring R_j ($j = ac, bc, a, b$) under different M ($N_a = 20$, $N_b = 15$, $N_c = 15$).

Fig. 4.9 demonstrates the minimized average end-to-end delay, D^*, and the average delays for local packet transmissions, D_a and D_b, versus M. It can be seen that D^* decreases with M at the beginning when M is small, which indicates that more token rotation cycles are required to achieve a smaller end-to-end delay; When M becomes large, D^* starts to increase since excessive time slot reservation for each superframe enlarges the packet service time. Therefore, the optimal total number of time slots for each superframe, M^{opt}, can be determined based on Algorithm 3 to achieve the minimal average end-to-end delay, D^{opt}. We can also see that the average delays for local transmissions, D_a and D_b, are below certain threshold. Nodes in token rings R_a and R_b are always guaranteed the minimum amount of time slots to maintain bounded average delays for local packet transmissions.

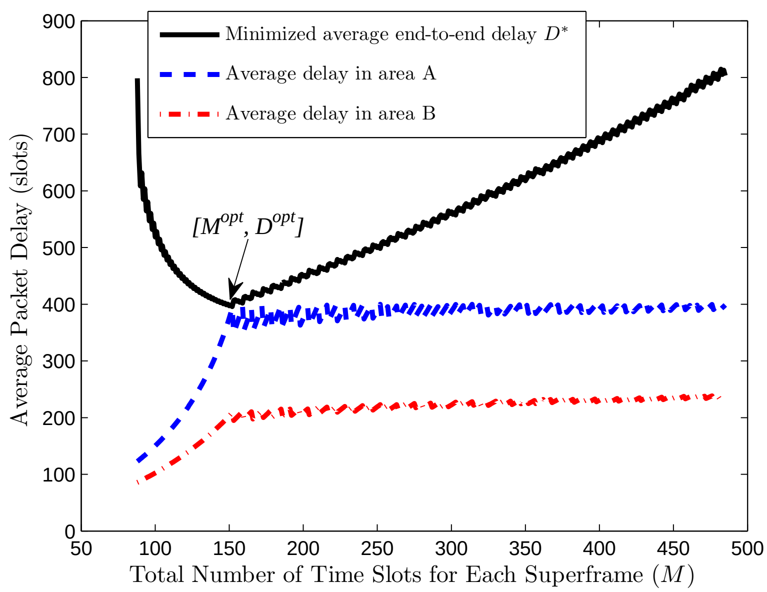


Figure 4.9: Average packet delay under different M ($N_a = 20$, $N_b = 15$, $N_c = 15$).

In Fig. 4.10, we also evaluate M^{opt} and k_j^{opt} in terms of an increasing total number of nodes, N, in the network, with the same numbers of nodes in each area ($N_a = N_b = N_c$). It can be seen that M^{opt} increases consistently with N and remains around a steady value when the network load becomes high, which indicates that the optimal superframe length for the TA-MAC is adaptive to the network traffic load variations and is controlled stable at high traffic load conditions. Within each superframe, the optimal numbers of token rotation cycles, k_{ac}^{opt} and k_{bc}^{opt}, keep decreasing with the increase of N, and k_a^{opt} and k_b^{opt} maintain at the minimum value in order to provide the maximum amount of resources for nodes in R_{ac} and R_{bc} to achieve the minimal average end-to-end delay.

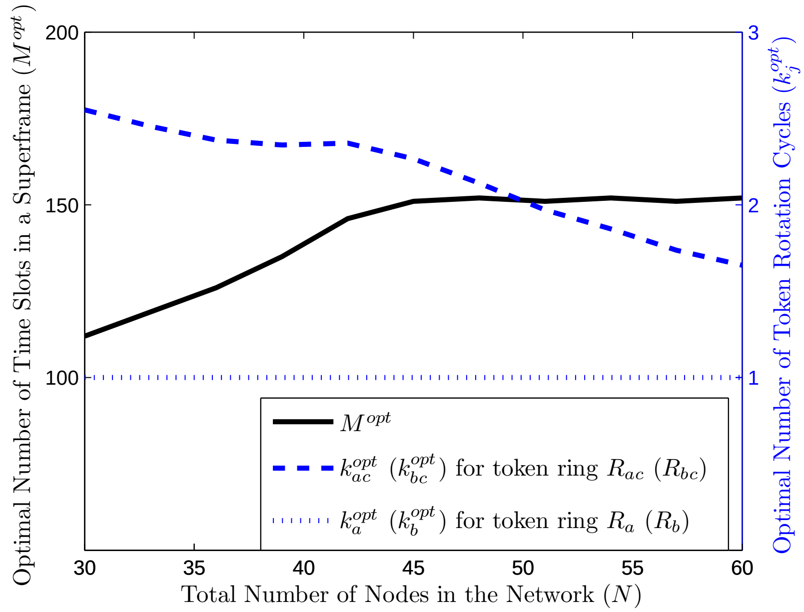
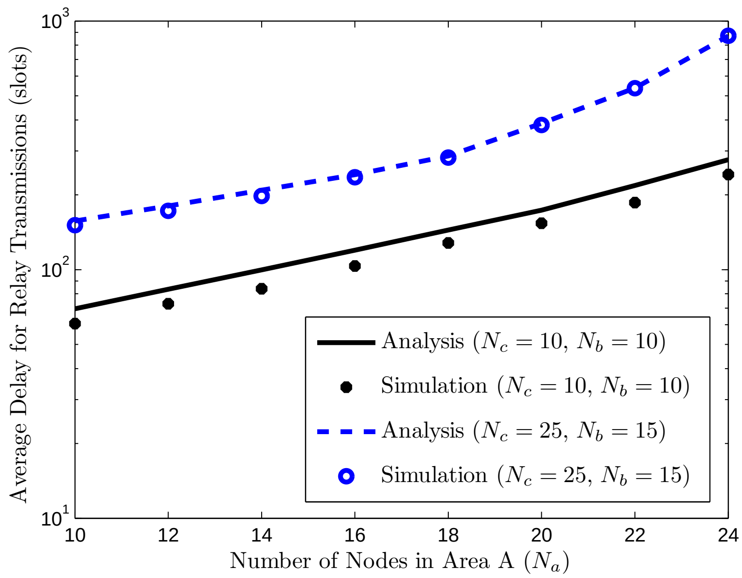


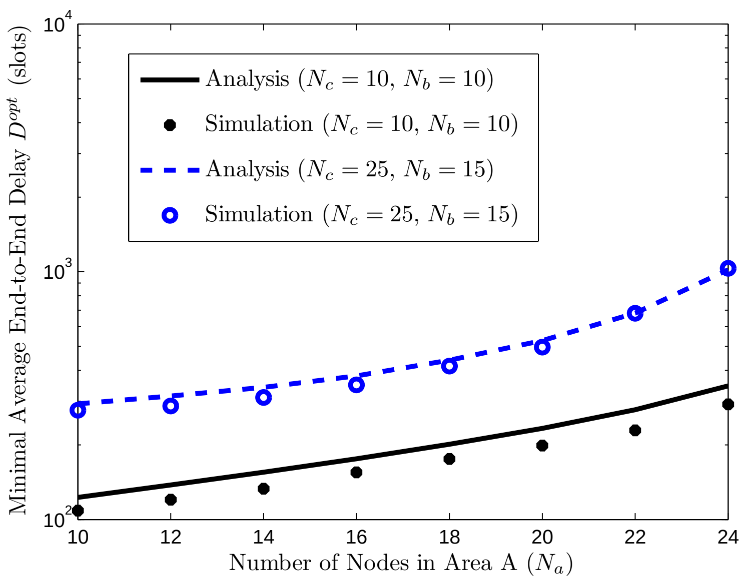
Figure 4.10: The optimal total number of time slots, M^{opt}, for each superframe and the optimal number of token rotation cycles, k_j^{opt}, for token ring R_j with respect to the total number of nodes, N ($N_a = N_b = N_c$).

4.5.2 Performance Metrics for the TA-MAC

We further evaluate the average delay for relay packet transmissions and the average end-to-end delay, with a varying number of nodes, N_a, in area A, for both low and high network traffic load conditions. In Fig. 4.11(a), it is shown that the average delay for relay transmissions increases consistently with N_a, and the simulation results match the analytical results more closely when the network traffic load becomes high ($N_c = 25$, $N_b = 15$), which verifies the effectiveness of the Poisson compound traffic arrival rate approximation on each relay node used in the analysis. Basically, the approximation becomes more accurate with an increasing number of nodes and node queue utilization ratios. Similar trends are observed in Fig. 4.11(b). We can see that, with the optimal MAC parameters, the minimal average end-to-end delay achieved by the TA-MAC is always controlled within an acceptable level when the network load varies in a wide range. The simulation results are also shown to be close to the analytical results.

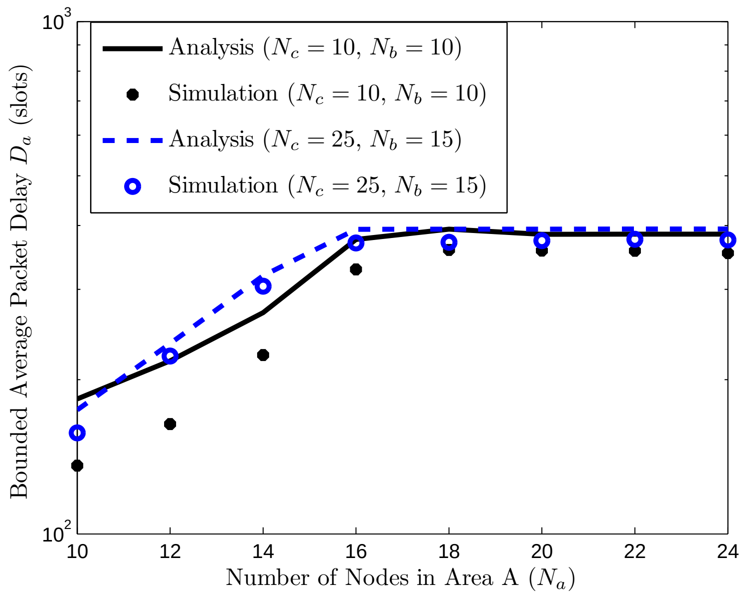


(a)

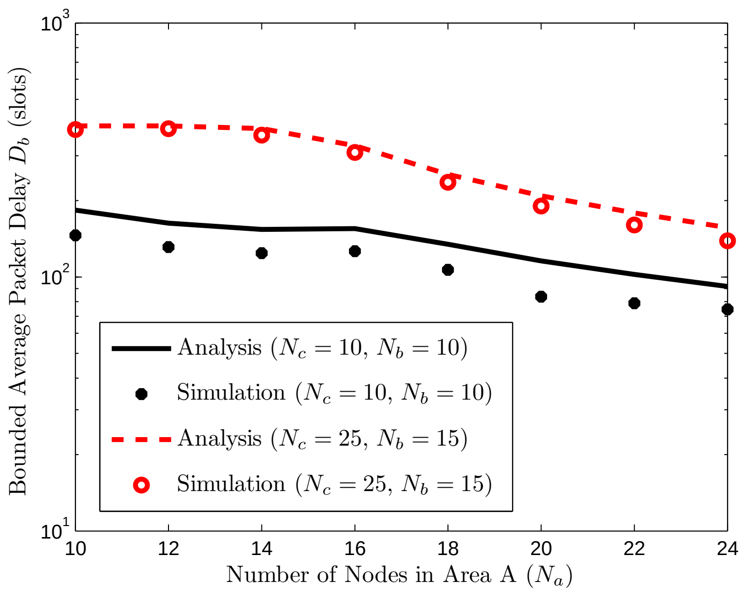


(b)

Figure 4.11: An evaluation of average delay for relay transmissions and average end-to-end delay under different network load conditions.



(a)



(b)

Figure 4.12: An evaluation of average delay for local transmissions in different network load conditions.

Fig. 4.12(a) and 4.12(b) demonstrate average delays for local transmissions in area A and B, D_a and D_b, which are guaranteed to be bounded under certain threshold with a varying network load, since the TA-MAC reserves the minimum amount of time slots for both areas A and B to achieve the minimal average end-to-end delay for pairs of end users. Similarly, the analytical results match the simulation results.

Fig. 4.13 shows the aggregate network throughput for the TA-MAC with varying traffic load conditions. We can see that the throughput continuously increases with N_a, and the simulation results match well with the analytical results. We also observe that higher network throughput can be achieved in a high network condition ($N_c = 25, N_b = 15$), with more packets transmitted in each superframe.

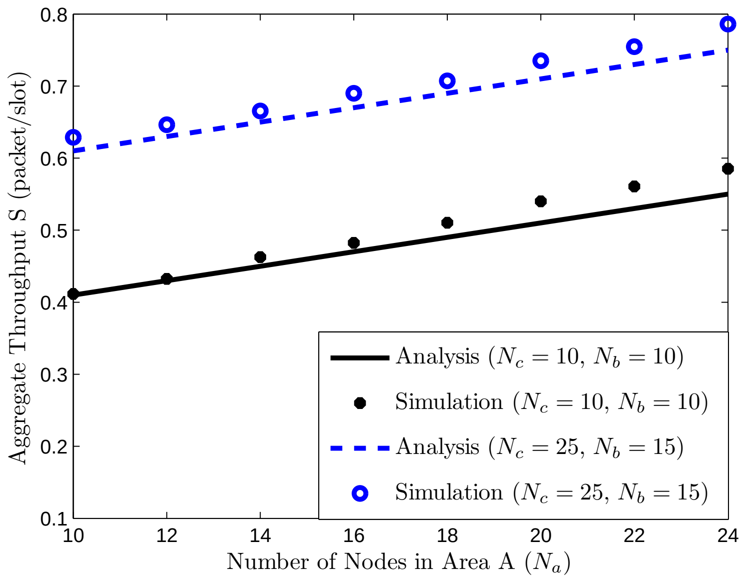


Figure 4.13: Aggregate network throughput under different network load conditions

4.5.3 Performance Comparison

Lastly, we compare the performance of the TA-MAC scheme with that of two existing MAC schemes proposed for multi-hop MANETs: a load adaptive MAC (LA-MAC) scheme [25] and a dynamic TDMA time slot assignment (DTSA) [18] [25]. LA-MAC is a hybrid MAC scheme, in which each node is allocated one time slot for exclusive use based on the DTSA. If current slot owner does not have packets to transmit when its designated time slot comes, all its two-hop neighbors can contend for the transmission opportunity in the slot based on a mechanism similar to the CSMA/CA. If a node's transmission attempts fail for a consecutive number of times, referred to as the state switch threshold, the node switches to the High Contention (HC) state and broadcasts a notification message, informing that only its one-hop neighbors can contend in the node's designated time slot. This state switch is used for reducing packet collisions caused by hidden nodes in high traffic load conditions; The DTSA is a dynamic TDMA scheme, where each node in a two-hop network is allocated one exclusive time slot within a time frame. The first slot in each frame is reserved for new nodes broadcasting request messages to join the network. If the current frame does not have available time slots for newly arriving nodes, the whole frame length is doubled to generate new available time slots. Thus, the scheme at most guarantees each node occupying two time slots in every time frame.

Fig. 4.14 demonstrates the comparison of the average end-to-end delay in terms of an increasing network load, with the same numbers of nodes in each area, between the TA-MAC scheme and the other two schemes. We can see that the LA-MAC achieves a smaller end-to-end delay in a low traffic load condition since nodes can contend to exploit the transmission opportunities in those time slots that are not used by the TDMA slot owners, making the MAC scheme behaving like CSMA/CA. However, the growth of the numbers of nodes in areas A, C and B results in a shrinking number of empty slots and accumulated contention collisions for the LA-MAC, making it behave close to the DTSA. Since the time slot allocation for nodes with DTSA are not optimized, the traffic of relay nodes become quickly saturated, making the average end-to-end delay for both LA-MAC and DTSA increase dramatically to very large values in high network load conditions, whereas the TA-MAC achieves consistently minimal average end-to-end delay performance within a wide range of network traffic load. Its advantage becomes more obvious with a high number of nodes in the network by maintaining the end-to-end delay within the acceptable level.

We further conduct the aggregate network throughput comparison in Fig. 4.15, with network traffic load variations. It can be seen that the throughput of all three schemes consistently increases with the network load. In a low network load condition, all the schemes achieve similar channel utilizations. However, when the network load increases, the proposed TA-MAC scheme achieves consistently higher throughput than the other two schemes, by optimizing the scheduling of token rotation cycles for each token ring and controlling the queue of each node in an unsaturated condition, whereas the throughputs for both LA-MAC and DTSA start to saturate from a moderate network load condition.

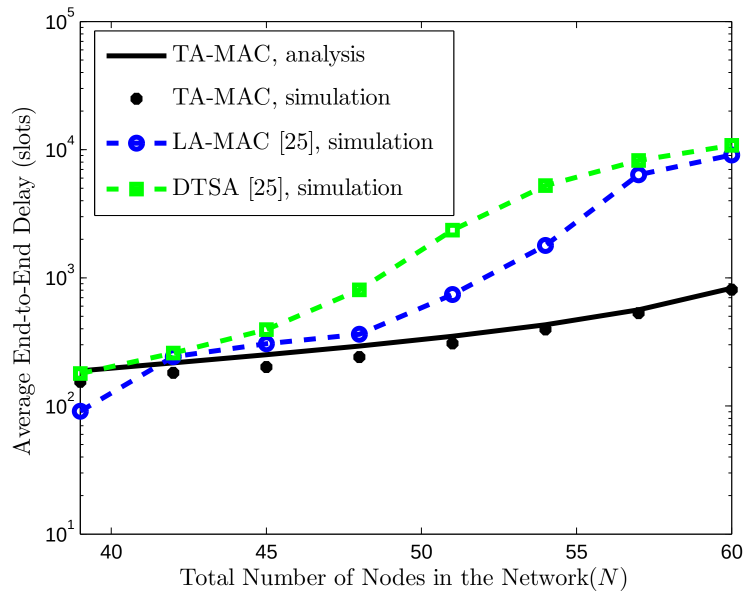


Figure 4.14: Average end-to-end packet delay comparison between the TA-MAC scheme and other MAC schemes

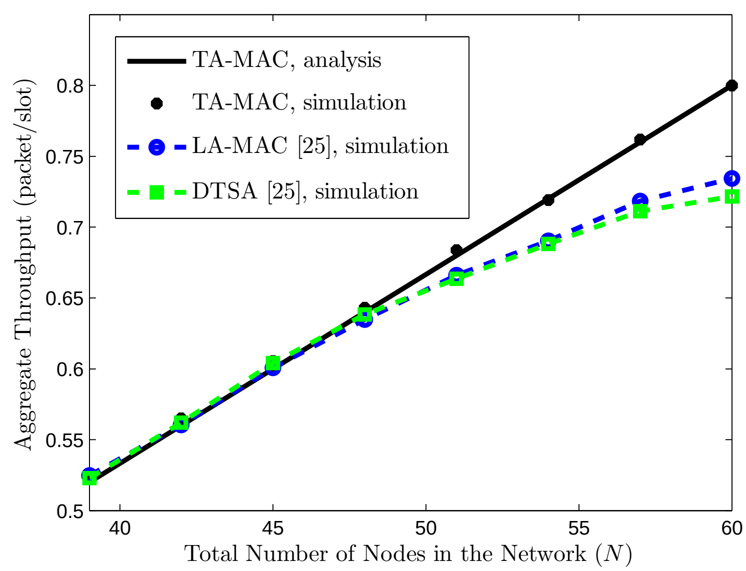


Figure 4.15: Aggregate throughput comparison between the TA-MAC scheme and other MAC schemes

4.6 Summary

In this chapter, a distributed token-based adaptive MAC (TA-MAC) scheme is proposed for a two-hop MANET. The TA-MAC eliminates the hidden terminal problem and adjust the set of MAC parameters, i.e., the numbers of token rotation cycles scheduled for each token ring and the superframe length, according to the numbers of nodes in each network area for performance optimization. An average end-to-end delay minimization framework is developed to find the set of optimal MAC parameters for each superframe. The proposed TA-MAC scheme achieves consistently minimal average delay for end-to-end transmissions, bounded delays for local transmissions and high aggregate throughput with variations of the number of nodes in the network. Based on a comparison with other two MAC schemes, the TA-MAC demonstrates much better scalability for the IoT-based two-hop environment in presence of network load dynamics, especially in a high traffic load condition.

4.7 Appendix

4.7.1 Proof of Proposition 2

For brevity, we only provide the proof for D_a. The proofs for other average delay functions can be carried out in a similar way. From (4.13), D_a is the combination of average queueing delay, D_q, and average service delay D_t. That is,

$$D_a = D_q + D_t = \frac{\lambda_a}{2} \cdot \frac{x_1 k_a^2 + x_2 k_a + x_3}{k_a + x_4} + \frac{\varepsilon_a}{k_a} = \frac{\lambda_a}{2} f_1(k_a) + f_2(k_a) \tag{4.21}$$

where x_1, x_2, x_3 and x_4 equal to the corresponding values in (4.13) ($x_1 = \alpha_a$, $x_2 = \beta_a$, $x_3 = \gamma_a$, $x_4 = -\lambda_a \varepsilon_a$).

In (4.21), $f_2(k_a)$ is a strictly convex function of k_a, since $f_2''(k_a) > 0$, $\forall k_a \geq 1$. On the other hand, the second-order derivative of $f_1(k_a)$ can be derived as

$$f_1''(k_a) = \frac{2x_1 x_4^2 - 2x_2 x_4 + 2x_3}{(k_a + x_4)^3} = \frac{g_1(x_4)}{(k_a + x_4)^3}. \tag{4.22}$$

Theoretically, x_1, x_2, and x_3 are fixed with a certain number of nodes, L_a, in area A, and k_a can take values from the interval $\left[1, \frac{M-L_{ac}-L_{bc}}{L_a}\right]$, due to constraints (4.14a) and (4.14f). Thus, by conforming to constraint (4.14d), we have $x_4 \in \left(-\frac{M-L_{ac}-L_{bc}}{L_a}, 0\right)$ with the variation of λ_a, to guarantee $(k_a + x_4)^3 > 0$ in (4.22). Therefore, the numerator of (4.22) can be regarded as a quadratic function of x_4, denoted by $g_1(x_4)$. We define $g_1(x_4)$ as an extension of $g_1(x_4)$ with $\mathbf{dom}\,\widetilde{g_1} \in (-\infty, \infty)$. Since $x_1 > 0$ and $x_2 < 0$, $g_1(x_4)$ represents a parabola, opening upward with the horizontal axis coordinate of its vertex $x_4^* = -\frac{5L_a^2+12ML_a+1}{12L_a^2}$. Because $x_4^* < -\frac{M-L_{ac}-L_{bc}}{L_a}$, it is concluded that $g_1(x_4)$ with $\mathbf{dom}\,g_1 \in \left(-\frac{M-L_{ac}-L_{bc}}{L_a}, 0\right)$ is a strictly increasing function of x_4. Furthermore, since $g_1\left(-\frac{M^{-}-L_{ac}-L_{bc}}{L_a}\right) > 0$, it is proved that $g_1(x_4) > 0$, $\forall x_4 \in \mathbf{dom}\,g_1$, and thus we have $f_1''(k_a) > 0$, $\forall k_a \in \left[1, \frac{M-L_{ac}-L_{bc}}{L_a}\right]$. Hence, D_a is a linear combination of two strictly convex functions of k_a, which is also known to be strictly convex [65].

4.7.2 Proof of Corollary 1

According to **Proposition 2**, D_a is a convex function of k_a. Thus, we have $D_a''(k_a) > 0$, indicating that $D_a'(k_a)$ is a strictly increasing function which is derived as

$$\begin{aligned} D_a'(k_a) &= \frac{\lambda_a}{2} \cdot \frac{x_1 k_a^2 + 2x_1 x_4 k_a + x_2 x_4 - x_3}{(k_a + x_4)^2} - \frac{\varepsilon_a}{k_a^2} \\ &= \frac{\lambda_a}{2} f_3(k_a) + f_4(k_a), \qquad k_a \in \left[1, \frac{M - L_{ac} - L_{bc}}{L_a}\right]. \end{aligned} \tag{4.23}$$

Therefore, the maximum value of $D_a'(k_a)$ is obtained when $k_a = \frac{M-L_{ac}-L_{bc}}{L_a}$. That is,

$$D_a'\left(\frac{M - L_{ac} - L_{bc}}{L_a}\right) = \frac{\lambda_a}{2} f_3\left(\frac{M - L_{ac} - L_{bc}}{L_a}\right) + f_4\left(\frac{M - L_{ac} - L_{bc}}{L_a}\right). \tag{4.24}$$

In (4.24), $f_3\left(\frac{M-L_{ac}-L_{bc}}{L_a}\right)$ is a function of x_4, with $x_4 \in \left(-\frac{M-L_{ac}-L_{bc}}{L_a}, 0\right)$, which is expressed as

$$\begin{aligned} f_3\left(\frac{M - L_{ac} - L_{bc}}{L_a}\right) &= \frac{\left[\frac{2x_1(M-L_{ac}-L_{bc})}{L_a} + x_2\right] x_4 + x_1\left(\frac{M-L_{ac}-L_{bc}}{L_a}\right)^2 - x_3}{\left(\frac{M-L_{ac}-L_{bc}}{L_a} + x_4\right)^2} \\ &= \frac{g_2(x_4)}{\left(\frac{M-L_{ac}-L_{bc}}{L_a} + x_4\right)^2}. \end{aligned} \tag{4.25}$$

Since $\frac{2x_1(M-L_{ac}-L_{bc})}{L_a} + x_2 < 0$, the linear function $g_2(x_4)$ is a strictly decreasing function with its maximum value being $g_2\left(-\frac{M^{-}-L_{ac}-L_{bc}}{L_a}\right) < 0$. Thus, we have $f_3\left(\frac{M-L_{ac}-L_{bc}}{L_a}\right) < 0$, and $D'_a\left(\frac{M-L_{ac}-L_{bc}}{L_a}\right) < 0$, $\forall x_4 \in \left(-\frac{M-L_{ac}-L_{bc}}{L_a}, 0\right)$. Hence, it is proved that $D'_a(k_a) < 0$, $\forall k_a \in \left[1, \frac{M-L_{ac}-L_{bc}}{L_a}\right]$. Similar proofs for the same property of other one-hop average delay functions can be made, which are omitted here.

4.7.3 Proof of Proposition 3

If we rewrite (**SP2a**) in a standard form, it is easy to verify that all the inequality constraint functions are convex with respect to k_{ac}. In the objective function, D_{ab} is a summation of D_{ac} and D_{cb} according to (4.11), which is further expressed as

$$D_{ab}(k_{ac}, h(k_{ac})) = D_{ac}(k_{ac}) + D_{cb}(h(k_{ac})). \tag{4.26}$$

Based on **Proposition 2**, we know that D_{ac} is a convex function of k_{ac} and D_{cb} is convex function of $h(k_{ac})$. It is also found that $h(k_{ac})$ is a linear function of k_{ac}. Thus, according to the scalar composition rules, $D_{cb}(h(k_{ac}))$ is also a convex function of k_{ac}. Hence, D_{ab}, a linear combination of two convex functions, is also convex with respect to k_{ac}. The same property also holds for D_{ba} with a similar proof. Moreover, as stated before, the two dimensional max function, $\max\{x, y\}$, is convex on $\mathbf{R}^2$ and nondecreasing in each of its two arguments. Therefore, according to the vector composition rules, the objective function of (**SP2a**) is a convex function with respect to the decision variable k_{ac}, which ends the proof.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

The objective of this PhD thesis is to develop comprehensive adaptive MAC solutions for an IoT-enabled MANET, with the consideration of different network features, to achieve consistently maximal QoS performance by adapting to the network traffic load variations. Specifically, we first consider a simplified fully-connected network with homogeneous best-effort data traffic support. Adaptive MAC is studied to switch between IEEE 802.11 DCF and D-TDMA to maximize network performance over traffic load variations. To make the performance comparison tractable between the two candidate MAC protocols, approximate and closed-form analytical relations are established between the MAC performance metrics (i.e., throughput and delay) and the total number of nodes in the network for both the IEEE 802.11 DCF and the D-TDMA, considering traffic saturation and non-saturation cases. According to the unified performance analysis framework, an adaptive MAC solution is developed to determine the MAC selection between the two candidate MAC protocols based on the MAC switching point calculation. With the switching point, nodes can make a switching decision between the IEEE 802.11 DCF and the D-TDMA in a distributed manner when the network traffic load varies. Analytical and simulation results demonstrate the high accuracy of the proposed analytical model in determining the MAC switching point, in both saturated and unsaturated traffic load conditions.

Then, a fully-connected network supporting delay-sensitive voice traffic in addition to the best-effort data traffic is considered. Due to different QoS requirements for voice and data traffic, an adaptive hybrid MAC scheme is proposed, in which distributed TDMA is employed for voice packet transmissions and truncated CSMA/CA (T-CSMA/CA) is used for data nodes to access the channel. To guarantee the voice packet loss rate bound, a traffic-adaptive TDMA time slot allocation scheme is presented to allocate one time slot for each active voice node according to its transmission buffer state. We establish an analytical model to determine the voice capacity region and the average number of scheduled voice bursts. The resource utilization for voice traffic is improved significantly by exploiting the voice traffic multiplexing. To maximize the channel utilization for data traffic, a data throughput analytical and optimization framework is developed, in which a closed-form mathematical relationship is established between the MAC layer parameter (i.e., the optimal contention window size) and the number of voice and data nodes in the network. Based on the analysis, data nodes operating T-CSMA/CA can adaptively adjust the contention window size to the optimal value based on the updated heterogeneous network traffic load, to achieve consistently maximum aggregate data throughput.

We further consider a two-hop network with an increasing number of nodes, as the first step towards a more general multi-hop environment. A token-based adaptive MAC scheme is proposed, in which different one-hop node groups are allocated different TDMA durations within a superframe structure, and each node group forms a token ring and adopts a probabilistic token passing scheme among its group members for packet transmissions. Each token ring maintains and updates its node members in a distributed way by adapting to the instantaneous number of nodes in the network. To determine the MAC parameters for performance optimization, we evaluate the average delay for end-to-end packet transmissions in closed-form functions of the MAC protocol parameters and the network traffic load. Then, we establish an optimization framework to determine the optimal superframe length and the associated optimal numbers of token rotation cycles for each token ring, with which the average end-to-end delay is minimized. Numerical and simulation results show that the proposed MAC scheme demonstrates much better scalability than the other two schemes for the two-hop network in presence of the network traffic load dynamics, especially in high traffic load conditions.

5.2 Future Research Directions

This PhD research can be extended in several directions by removing some of the assumptions made in the thesis:

1. If the wireless channel introduces transmission errors, a transmission without collision may fail to be correctly received by the receiver. Therefore, the transmission failure of a packet can result from not only packet collisions but also transmission errors

due to a poor channel condition. Hence, in the adaptive MAC framework, the performance analytical models for the MAC candidate protocols should be re-evaluated by calculating the packet transmission failure probability with the consideration of both packet collisions and packet transmission errors. Therefore, the optimal switching point derivation needs to be extended accordingly. For the traffic model, traffic arrival patterns among different nodes may not be homogeneous. For instance, each node can have a unique average traffic arrival rate, denoted by λ_i for node i. In this case, the network traffic load can be described by a combination of the total number of nodes and the different traffic arrival statistics of the nodes, $(N, \lambda_1, ..., \lambda_N)$. It is important to extend the proposed adaptive MAC solution by calculating the switching point with a more general traffic generation model. In addition to the MAC switching point calculation, how each node can coordinate in a distributed way to switch between the MAC candidate protocols for implementation and how the implementation overhead will affect the performance gain of the adaptive MAC solution are also important for future research discussions.

2. For supporting heterogeneous traffic, how to extend or improve the proposed adaptive and hybrid MAC scheme when the traffic types become more diverse (more than voice and data traffic types) can be investigated in an IoT-based environment. For example, if video traffic is also supported in the network which may have the delay jitter requirement for each video frame in addition to the delay bound requirement, how to design a more comprehensive adaptive and QoS-aware MAC scheme to satisfy more than two types of QoS demands in a varying heterogeneous network traffic load environment will be challenging.

3. For an IoT-enabled network, the number of nodes supported can be increased to hundreds or thousands in a more enlarged network region beyond the two-hop communication range. Therefore, to improve the scalability in a general multi-hop network, the proposed token-based adaptive MAC scheme can be extended to exploit the spatial reuse of the network resources for each transmission hop to maintain a high channel utilization. Also, the relay selection for each source node can also be considered to achieve the traffic load balancing and to further improve the end-to-end performance.

References

[1] "Supplement to 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: Higher-Speed Physical Layer Extension in the 2.4 GHz Band," *IEEE Std 802.11b-1999*, pp. i–90, 2000.

[2] P. Wang, H. Jiang, and W. Zhuang, "Capacity improvement and analysis for voice/data traffic over WLANs," *IEEE Trans. Wireless Commun.*, vol. 6, no. 4, pp. 1530–1541, 2007.

[3] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A survey on enabling technologies, protocols, and applications," *IEEE Commun. Surv. Tutor.*, vol. 17, no. 4, pp. 2347–2376, 2015.

[4] M. Palattella, N. Accettura, X. Vilajosana, T. Watteyne, L. Grieco, G. Boggia, and M. Dohler, "Standardized protocol stack for the Internet of (important) Things," *IEEE Commun. Surv. Tutor.*, vol. 15, no. 3, pp. 1389–1406, 2013.

[5] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," *Future Gener. Comput. Syst*, vol. 29, no. 7, pp. 1645–1660, 2013.

[6] Y. Liu, C. Yuen, X. Cao, N. U. Hassan, and J. Chen, "Design of a scalable hybrid MAC protocol for heterogeneous M2M networks," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 99–111, 2014.

[7] M. Natkaniec, K. Kosek-Szott, S. Szott, and G. Bianchi, "A survey of medium access mechanisms for providing QoS in ad-hoc networks," *IEEE Commun. Surv. Tutor.*, vol. 15, no. 2, pp. 592–620, 2013.

[8] J.-R. Cha, K.-C. Go, J.-H. Kim, and W.-C. Park, "TDMA-based multi-hop resource reservation protocol for real-time applications in tactical mobile adhoc network," in *Proc. IEEE MILCOM'10*, 2010, pp. 1936–1941.

[9] H. Nishiyama, T. Ngo, S. Oiyama, and N. Kato, "Relay by smart device: Innovative communications for efficient information sharing among vehicles and pedestrians," *IEEE Veh. Technol. Mag.*, vol. 10, no. 4, pp. 54–62, 2015.

[10] J. Liu and N. Kato, "A markovian analysis for explicit probabilistic stopping-based information propagation in postdisaster ad hoc mobile networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 81–90, 2016.

[11] M. Natkaniec, K. Kosek-Szott, S. Szott, J. Gozdecki, A. Głowacz, and S. Sargento, "Supporting QoS in integrated ad-hoc networks," *Wireless Pers. Commun.*, vol. 56, no. 2, pp. 183–206, 2011.

[12] P. Wang, H. Jiang, and W. Zhuang, "A new MAC scheme supporting voice/data traffic in wireless ad hoc networks," *IEEE Trans. Mobile Comput.*, vol. 7, no. 12, pp. 1491–1503, 2008.

[13] A. Rajandekar and B. Sikdar, "A survey of MAC layer issues and protocols for machine-to-machine communications," *IEEE Internet Things J.*, vol. 2, no. 2, pp. 175–186, 2015.

[14] R. Jurdak, C. V. Lopes, and P. Baldi, "A survey, classification and comparative analysis of medium access control protocols for ad hoc networks," *IEEE Commun. Surv. Tutor.*, vol. 6, no. 1, pp. 2–16, 2004.

[15] H. Zhai, J. Wang, and Y. Fang, "DUCHA: A new dual-channel MAC protocol for multihop ad hoc networks," *IEEE Trans. Wireless Commun.*, vol. 5, no. 11, pp. 3224–3233, 2006.

[16] K. Ghaboosi, M. Latva-aho, Y. Xiao, and Q. Zhang, "eMAC-A medium-access control protocol for the next-generation ad hoc networks," *IEEE Trans. Veh. Technol.*, vol. 58, no. 8, pp. 4476–4490, 2009.

[17] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE J. Select. Areas Commun.*, vol. 18, no. 3, pp. 535–547, 2000.

[18] A. Kanzaki, T. Uemukai, T. Hara, and S. Nishio, "Dynamic TDMA slot assignment in ad hoc networks," in *Proc. IEEE AINA'03.*, 2003, pp. 330–335.

[19] N. Wilson, R. Ganesh, K. Joseph, and D. Raychaudhuri, “Packet CDMA versus dynamic TDMA for multiple access in an integrated voice/data PCN,” *IEEE J. Select. Areas Commun.*, vol. 11, no. 6, pp. 870–884, 1993.

[20] W. Hu, X. Li, and H. Yousefi’zadeh, “LA-MAC: A load adaptive MAC protocol for MANETs,” in *Proc. IEEE GLOBECOM’09*, 2009, pp. 1–6.

[21] R. Zhang, L. Cai, and J. Pan, “Performance analysis of reservation and contention-based hybrid MAC for wireless networks,” in *Proc. IEEE ICC’10*, 2010, pp. 1–5.

[22] ——, “Performance study of hybrid MAC using soft reservation for wireless networks,” in *Proc. IEEE ICC’11*, 2011, pp. 1–5.

[23] R. Ahmed, “An adaptive multiple access protocol for broadcast channels,” in *Proc. IEEE IPCCC’97*, 1997, pp. 371–377.

[24] I. Chlamtac, A. Farago, A. Myers, V. Syrotiuk, and G. Zaruba, “ADAPT: a dynamically self-adjusting media access control protocol for ad hoc-networks,” in *Proc. IEEE GLOBECOM ’99*, 1999, pp. 11–15.

[25] W. Hu, H. Yousefi’zadeh, and X. Li, “Load adaptive MAC: A hybrid MAC protocol for MIMO SDR MANETs,” *IEEE Trans. Wireless Commun.*, vol. 10, no. 11, pp. 3924–3933, 2011.

[26] C. Doerr, M. Neufeld, J. Fifield, T. Weingart, D. C. Sicker, and D. Grunwald, “MultiMAC-an adaptive MAC framework for dynamic radio networking,” in *Proc. IEEE DySPAN’05*, 2005, pp. 548–555.

[27] D. Zheng and Y.-D. Yao, “Throughput performance evaluation of two-tier TDMA for sensor networks,” in *Proc. IEEE SARNOFF ’09*, 2009, pp. 1–5.

[28] L. Cai, X. Shen, J. W. Mark, L. Cai, and Y. Xiao, “Voice capacity analysis of WLAN with unbalanced traffic,” *IEEE Trans. Veh. Technol.*, vol. 55, no. 3, pp. 752–761, 2006.

[29] M. Dong, K. Ota, A. Liu, and M. Guo, “Joint optimization of lifetime and transport delay under reliability constraint wireless sensor networks,” *IEEE Trans. Parallel Distrib. Syst.*, vol. 27, no. 1, pp. 225–236, 2016.

[30] M. Park, “IEEE 802.11ah: Sub-1-GHz license-exempt operation for the Internet of Things,” *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 145–151, 2015.

[31] S. Tozlu, M. Senel, W. Mao, and A. Keshavarzian, "Wi-Fi enabled sensors for Internet of Things: A practical approach," *IEEE Commun. Mag.*, vol. 50, no. 6, pp. 134–143, 2012.

[32] A. Veres, A. T. Campbell, M. Barry, and L.-H. Sun, "Supporting service differentiation in wireless packet networks using distributed control," *IEEE J. Select. Areas Commun.*, vol. 19, no. 10, pp. 2081–2093, 2001.

[33] D. Xu, T. Sakurai, H. Vu, and T. Sakurai, "An access delay model for IEEE 802.11e EDCA," *IEEE Trans. Mobile Comput.*, vol. 8, no. 2, pp. 261–275, 2009.

[34] P. Serrano, A. Banchs, P. Patras, and A. Azcorra, "Optimal configuration of 802.11e EDCA for real-time and data traffic," *IEEE Trans. Veh. Technol.*, vol. 59, no. 5, pp. 2511–2528, 2010.

[35] Y. Xiao and Y. Pan, "Differentiation, QoS guarantee, and optimization for real-time traffic over one-hop ad hoc networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 16, no. 6, pp. 538–549, 2005.

[36] "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 8: Medium Access Control (MAC) Quality of Service Enhancements," *IEEE Std 802.11e-2005 (Amendment to IEEE Std 802.11, 1999 Edition (Reaff 2003)*, pp. 1–212, 2005.

[37] Y. Xiao, "Enhanced DCF of IEEE 802.11e to support QoS," in *Proc. IEEE WCNC '03*, vol. 2, 2003, pp. 1291–1296.

[38] J. Sobrinho and A. Krishnakumar, "Quality-of-service in ad hoc carrier sense multiple access wireless networks," *IEEE J. Select. Areas Commun.*, vol. 17, no. 8, pp. 1353–1368, 1999.

[39] P. Wang and W. Zhuang, "A collision-free MAC scheme for multimedia wireless mesh backbone," *IEEE Trans. Wireless Commun.*, vol. 8, no. 7, pp. 3577–3589, 2009.

[40] H. Omar, W. Zhuang, and L. Li, "VeMAC: A TDMA-based MAC protocol for reliable broadcast in VANETs," *IEEE Trans. Mobile Comput.*, vol. 12, no. 9, pp. 1724–1736, 2013.

[41] S. Jiang, J. Rao, D. He, X. Ling, and C. C. Ko, "A simple distributed PRMA for MANETs," *IEEE Trans. Veh. Technol.*, vol. 51, no. 2, pp. 293–305, 2002.

[42] M. Wang, Q. Shen, R. Zhang, H. Liang, and X. Shen, "Vehicle-density-based adaptive MAC for high throughput in drive-thru networks," *IEEE Internet Things J.*, vol. 1, no. 6, pp. 533–543, 2014.

[43] W. Alasmary and W. Zhuang, "Mobility impact in IEEE 802.11p infrastructureless vehicular networks," *Ad Hoc Netw.*, vol. 10, no. 2, pp. 222–230, 2012.

[44] A. Abdrabou and W. Zhuang, "Stochastic delay guarantees and statistical call admission control for IEEE 802.11 single-hop ad hoc networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 10, pp. 3972–3981, 2008.

[45] H. A. Omar, W. Zhuang, and L. Li, "Gateway placement and packet routing for multihop in-vehicle internet access," *IEEE Trans. Emerging Topics in Comput.*, vol. 3, no. 3, pp. 335–351, 2015.

[46] I. Rhee, A. Warrier, J. Min, and L. Xu, "DRAND: Distributed randomized TDMA scheduling for wireless ad hoc networks," *IEEE Trans. Mobile Comput.*, vol. 8, no. 10, pp. 1384–1396, 2009.

[47] P. Koutsakis, "Token- and self-policing-based scheduling for multimedia traffic transmission over WLANs," *IEEE Trans. Veh. Technol.*, vol. 60, no. 9, pp. 4520–4527, 2011.

[48] H. Liang and W. Zhuang, "DFMAC: DTN-friendly medium access control for wireless local area networks supporting voice/data services," *ACM Mobile Netw. Appl. (MONET)*, vol. 16, no. 5, pp. 531–543, 2011.

[49] P. Wang and W. Zhuang, "A token-based scheduling scheme for WLANs supporting voice/data traffic and its performance analysis," *IEEE Trans. Wireless Commun.*, vol. 7, no. 5, pp. 1708–1718, 2008.

[50] I. S. Liu, F. Takawira, and H. J. Xu, "A hybrid token-CDMA MAC protocol for wireless ad hoc networks," *IEEE Trans. Mobile Comput.*, vol. 7, no. 5, pp. 557–569, 2008.

[51] Y. Bi, K. H. Liu, L. X. Cai, X. Shen, and H. Zhao, "A multi-channel token ring protocol for QoS provisioning in inter-vehicle communications," *IEEE Trans. Wireless Commun.*, vol. 8, no. 11, pp. 5621–5631, 2009.

[52] P. Teymoori, N. Yazdani, and A. Khonsari, "DT-MAC: An efficient and scalable medium access control protocol for wireless networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, pp. 1268–1278, 2013.

[53] Z. Haas and J. Deng, "Dual busy tone multiple access (DBTMA)-a multiple access control scheme for ad hoc networks," *IEEE Trans. Commun.*, vol. 50, no. 6, pp. 975–985, 2002.

[54] L. Lei, S. Cai, C. Luo, W. Cai, and J. Zhou, "A dynamic TDMA-based MAC protocol with QoS guarantees for fully connected ad hoc networks," *Kluwer J. Telecommun. Syst.*, pp. 1–11, 2014.

[55] I. Chlamtac, M. Conti, and J. J.-N. Liu, "Mobile ad hoc networking: imperatives and challenges," *Ad Hoc Netw.*, vol. 1, no. 1, pp. 13–64, 2003.

[56] H. Jiang, P. Wang, and W. Zhuang, "A distributed channel access scheme with guaranteed priority and enhanced fairness," *IEEE Trans. Wireless Commun.*, vol. 6, no. 6, pp. 2114–2125, 2007.

[57] K. Medepalli and F. Tobagi, "System centric and user centric queueing models for IEEE 802.11 based wireless LANs," in *Proc. IEEE BroadNets'05*, 2005, pp. 612–621.

[58] L. Kleinrock and F. Tobagi, "Packet switching in radio channels: Part I–carrier sense multiple-access modes and their throughput-delay characteristics," *IEEE Trans. Commun.*, vol. 23, no. 12, pp. 1400–1416, 1975.

[59] A. Abdrabou and W. Zhuang, "Service time approximation in IEEE 802.11 single-hop ad hoc networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 1, pp. 305–313, 2008.

[60] G. Berger-Sabbatel, A. Duda, M. Heusse, and F. Rousseau, "Short-term fairness of 802.11 networks with several hosts," in *Mobile and Wireless Commun. Netw.*, 2005, pp. 263–274.

[61] G. Berger-Sabbatel, A. Duda, O. Gaudoin, M. Heusse, and F. Rousseau, "Fairness and its impact on delay in 802.11 networks," in *Proc. IEEE GLOBECOM'04*, vol. 5, 2004, pp. 2967–2973.

[62] K. Medepalli and F. Tobagi, "Throughput analysis of IEEE 802.11 wireless LANs using an average cycle time approach," in *Proc. IEEE GLOBECOM'05*, vol. 5, 2005, pp. 3007–3011.

[63] Y. Tay and K. C. Chua, "A capacity analysis for the IEEE 802.11 MAC protocol," *Wireless Netw.*, vol. 7, no. 2, pp. 159–171, 2001.

[64] A. Kumar, E. Altman, D. Miorandi, and M. Goyal, "New insights from a fixed-point analysis of single cell IEEE 802.11 WLANs," *IEEE/ACM Trans. on Netw.*, vol. 15, no. 3, pp. 588–601, 2007.

[65] S. P. Boyd and L. Vandenberghe, *Convex optimization.* Cambridge, U.K.: Cambridge Univ. Press, 2004.

[66] M. Carvalho and J. Garcia-Luna-Aceves, "Delay analysis of IEEE 802.11 in single-hop networks," in *Proc. IEEE ICNP'03*, 2003, pp. 146–155.

[67] H. Omar, W. Zhuang, A. Abdrabou, and L. Li, "Performance evaluation of VeMAC supporting safety applications in vehicular networks," *IEEE Trans. Emerging Topics in Comput.*, vol. 1, no. 1, pp. 69–83, 2013.

[68] D. P. Bertsekas, R. G. Gallager, and P. Humblet, *Data networks.* Englewood Cliffs, NJ, USA: Prentice-hall, 1987, vol. 2.

[69] "OMNeT++ 5.0," [Online]. Available: http://www.omnetpp.org/omnetpp.

[70] J. Ren, Y. Zhang, K. Zhang, A. Liu, J. Chen, and X. Shen, "Lifetime and energy hole evolution analysis in data-gathering wireless sensor networks," *IEEE Trans. Industr. Informatics*, 2015, to appear.

[71] A. Aijaz and A. Aghvami, "Cognitive machine-to-machine communications for Internet-of-Things: A protocol stack perspective," *IEEE Internet Things J.*, vol. 2, no. 2, pp. 103–112, 2015.

[72] X. Ling, K.-H. Liu, Y. Cheng, X. Shen, and J. W. Mark, "A novel performance model for distributed prioritized MAC protocols," in *Proc. IEEE GLOBECOM'07*, 2007, pp. 4692–4696.

[73] F. Cali, M. Conti, and E. Gregori, "Dynamic tuning of the IEEE 802.11 protocol to achieve a theoretical throughput limit," *IEEE/ACM Trans. on Netw.*, vol. 8, no. 6, pp. 785–799, 2000.

[74] Q. Ye, W. Zhuang, L. Li, and P. Vigneron, "Traffic load adaptive medium access control for fully-connected mobile ad hoc networks," *IEEE Trans. Veh. Technol.*, to appear. DOI: 10.1109/TVT.2016.2516910.

[75] P. Wang and W. Zhuang, "A collision-free MAC scheme for multimedia wireless mesh backbone," *IEEE Trans. Wireless Commun.*, vol. 8, no. 7, pp. 3577–3589, 2009.

[76] J. Ren, Y. Zhang, N. Zhang, D. Zhang, and X. Shen, “Dynamic channel access to improve energy efficiency in cognitive radio sensor networks,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3143–3156, 2016.

[77] D. Zhang, Z. Chen, J. Ren, N. Zhang, M. Awad, H. Zhou, and X. Shen, “Energy harvesting-aided spectrum sensing and data transmission in heterogeneous cognitive radio sensor network,” *IEEE Trans. Veh. Technol.*, to appear. DOI: 10.1109/TVT.2016.2551721.

[78] J. Gorski, F. Pfeuffer, and K. Klamroth, “Biconvex sets and optimization with biconvex functions: a survey and extensions,” *Math. Method. Oper. Res.*, vol. 66, no. 3, pp. 373–407, 2007.

[79] Q. Ye and W. Zhuang, “Distributed and adaptive medium access control for Internet-of-Things-enabled mobile networks,” *IEEE Internet Things J.*, to appear. DOI: 10.1109/JIOT.2016.2566659.