

Supporting Latency-Critical Applications by Wireless and Mobile Networks: MAC Layer Approaches

Ye Feng

ORCID: 0000-0002-1120-9382

Submitted in total fulfilment of the requirements of the degree of

Doctor of Philosophy

Department of Electrical and Electronic Engineering
THE UNIVERSITY OF MELBOURNE

Jul 2020

Copyright © 2020 Ye Feng

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm or any other means without written permission from the author.

Dedicated to my parents, fiancée and parents-in-law.

Abstract

THE Internet has evolved a long way from transporting basic web data to transporting traffic instigated by new emerging services that are available today. One such service is wireless-based remote human-to-machine/human interaction, in which touch and actuation related information is delivered over the network. Such a service is referred to as the Tactile Internet, and it requires a network, integrating wired and wireless communication technologies, to provide high reliability and ultra-low end-to-end latency in the millisecond range. To date, some wired access networks can partly meet the requirement of the Tactile Internet, while currently deployed wireless access networks may not be able to fulfil these needs. Specifically, uplink latency resulting from the signalling process and queueing in the MAC layer becomes the bottleneck of applying wireless networks to the Tactile Internet. Today, wireless local area network (WLAN) and mobile cellular network have become the most dominant wireless solutions to connecting vast majority of users around the globe, due to their characteristics such as ease of deployment, lower cost and power consumption. Network designers, however, need a rethink of the underlying signalling mechanism such that latency of wireless transmissions can be significantly reduced.

The main objective of this thesis is to propose latency-reduction algorithms and transmission schemes for WLAN and mobile cellular network and study the feasibility of applying the proposed algorithms and schemes to the Tactile Internet. For this purpose, we have selected the HCF Controlled Channel Access (HCCA) of WLAN and Semi-Persistent Scheduling (SPS) scheme of LTE network as our candidates. Unlike other WLAN MAC protocols which are prone to packet collisions, HCCA operates a polling mechanism with guaranteed airtime for some certain admitted users. We studied the

performance of the HCCA by analytically deriving a closed-form expression for its average uplink latency, defined by various network and timing parameters. The derived expression was proved accurate, as compared with discrete-event simulations. We then conducted global sensitivity analyses to further understand the implications of these parameters on the latency performance. Lastly, we proposed a strategic parameter selection algorithm to fine-tune network parameters of WLAN such that the average latency of HCCA can be reduced effectively. Our analytical and simulative studies of the proposed algorithm considered small and constant payload size, in order to match the traffic characteristics of the Tactile Internet. Results showed that given proper network settings, a WLAN implementing HCCA is able to provide satisfying latency performance with a sub-millisecond latency.

On the other hand, the SPS was selected for mobile cellular network to meet stringent latency requirement because it not only eliminates the delay of signalling process by pre-allocating uplink radio periodically but also requires zero consumption in the control channel, which is especially beneficial for increasing the accessibility of a mobile cellular network. However, the mismatch between the pre-defined periodic resource allocation at the station side and the actual resource demand at the user side has resulted in over/under-scheduling problem, which limits the practical implementation of SPS for variable bit rate (VBR) data traffic. In this thesis, we addressed the over/under-scheduling problem by proposing a predictive SPS scheme, which enables dynamic resource allocation that closely matches the actual bandwidth requirement of Tactile users. In addition, a Feasible Resource Allocator (FRA) was proposed to introduce flexibility into resource scheduling and allow coexistence among Tactile and non-Tactile devices.

Finally, we further enhanced the scheduling accuracy by incorporating machine learning in the proposed predictive SPS scheme. In particular, we designed an Attention-based prediction model that implements dimension expansion at the pre-processing stage and introduces an initial input containing learnable parameters to the decoder, such that highest prediction accuracy was achieved during the training stage. It is important to note that simulations carried out for the proposed predictive SPS scheme and the Attention-based prediction model were based on traced data traffic, which was generated by haptic-

tactile experiments. Simulation results revealed that a sub-3 ms uplink latency in a mobile cellular network could be realised by our proposed SPS scheme with a simple autoregressive prediction model, and the uplink latency performance could be further improved by our Attention-based model.

Declaration

This is to certify that

1. the thesis comprises only my original work towards the PhD,
2. due acknowledgement has been made in the text to all other material used,
3. the thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Ye Feng, Jul 2020

Preface

This thesis contains publications of which I am the primary author. The publication status of each technical chapter and contributions of all persons involved are described as follows:

Chapter 3

- “IEEE 802.11 HCCA for Tactile Applications” was published by International Telecommunication Networks and Applications Conference (ITNAC) on Nov 2017 and it corresponds to Section 3.3.
- “Hybrid Coordination Function Controlled Channel Access for Latency-Sensitive Tactile Applications” was published by IEEE Global Communications Conference on Dec 2017 and it corresponds to Sections 3.3 and 3.4.
- “A Feasibility Study of IEEE 802.11 HCCA for Low-Latency Applications” was published by IEEE Transactions on Communications on July 2019 and it corresponds to Sections 3.3 and 3.4.

For the published works in Chapter 3, I conducted the original research, wrote the initial draft of the work and performed subsequent editing in response to co-authors. C. Jayasundara has helped verifying the mathematical derivations in Section 3.3, while A. Nirmalathas and E. Wong have given editorial advice and guidance towards research objectives, such as evaluation metrics for the performance of our proposed algorithm and sensitivity analyses of derived analytical equations.

Chapter 4

- “A Predictive Semi-Persistent Scheduling Scheme for Low-Latency Applications in LTE and NR Networks” was published by IEEE International Conference on Communications (ICC) on May 2019 and it corresponds to Sections 4.3.2.1 and 4.4.1.
- Sections 4.3.1, 4.3.2 and 4.4.2 contain unpublished materials that have not submitted for publication.

For the published and unpublished works in Chapter 4, I conducted the original research, wrote the initial draft of the work and performed subsequent editing in response to co-authors. A. Nirmalathas and E. Wong have given editorial advice and guidance towards research objectives, such as evaluation metrics for the performance of our proposed scheme.

Chapter 5

- Chapter 5 contains unpublished materials that have not submitted for publication.

For the unpublished works in Chapter 5, I conducted the original research, wrote the initial draft of the work and performed subsequent editing in response to co-authors. A. Nirmalathas and E. Wong have given editorial advice and guidance towards research objectives, such as the investigation into the necessity of applying machine learning to our research problem. In addition, I would like to thank L. Ruan, since she kindly provided the traced data traffic from her haptic-tactile experiments.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor Prof. Elaine Wong for the continuous support for my PhD study. As my principal supervisor, Elaine has encouraged me to explore new ideas and provided guidance in all the time of research and writing of this thesis. I could not have completed this thesis without Elaine's supervision. Her patience, passion, depth of knowledge have shown me what a dedicated researcher and friendly mentor is, and her hard-working attitude towards research goals has inspired me a lot. I would like to thank my co-supervisor Prof. Ampalavanapillai Nirmalathas, who introduced me to Prof. Elaine, and has always been supportive in our regular meetings with insightful suggestions. I would also like to thank my previous co-supervisor Dr Chamil Jayasundara, who has helped a lot with my first research objective, such as giving me editorial and technical advice for my first two conference papers. I would also like to thank my previous committee chair A/Prof. Leigh Johnston and current committee chair Prof. Jonathan Manton, for their presence in milestones of my candidature and their valuable feedback.

I would like to thank the University of Melbourne for providing me with great facilities and academic resources such that I can pursue my PhD degree. I am grateful for the financial support from the University of Melbourne in the form of Melbourne Research Scholarship and travel funding. I would also like to thank the administrative staff in the Department of Electrical and Electronic Engineering at the University of Melbourne, for their effort to keep our work environment safe and organised.

Throughout my PhD candidature, I have met many fantastic friends—Lihua, Xiaoxiao, Bin, Michael, Sourav, Imali, Chathurika, Tingting, from whom I have received support and advice. I wish them all the best for their future career.

I take this opportunity to thank my lecturers and demonstrators at the University of Melbourne and Monash University. Especially, I would like to thank Prof. Jean Armstrong, who had led me into the field of communications and had been my most supportive referee when I applied for postgraduate studies.

Most importantly, I would like to express my deepest gratitude to my loving parents Mr Xiaoming Feng and Mrs Jie Qian, who has been my utmost support for pursuing my goals. Their unconditional love, wisdom, patience and understanding has shaped me and made me who I am today.

Finally, I am extremely grateful to my fiancée Mengyan Guo, who has been accompanying me throughout my PhD program. I would not give up when I suffered setbacks due to simulation crash-down, paper rejection or weeks of tedious programming, because she would always be there to back me up with emotional support. Her considerate care, love, encouragement, understanding have been my most valuable incentive towards my PhD degree.

Contents

1	Introduction	1
1.1	Wireless and Mobile Cellular Communications	1
1.2	The Tactile Internet	3
1.3	Thesis Objectives and Approaches	6
1.4	Thesis Outline and Original Contributions	11
2	Introduction to Low-Latency Wireless and Mobile Cellular Networks	17
2.1	Wireless Networks Supported by IEEE 802.11	17
2.1.1	Technical Aspects of IEEE 802.11	18
2.1.2	Uplink Transmission Protocols of IEEE 802.11	20
2.1.3	IEEE 802.11 MAC Layer Solutions for Improved Latency Performance	29
2.2	Mobile Cellular Networks Supported by LTE	36
2.2.1	Technological Advancement of LTE	37
2.2.2	LTE Time-Frequency Division Transmissions	39
2.2.3	LTE Scheduling Algorithms for Ultra-Low Latency Performance .	44
2.3	Mobile Cellular Networks Enhancement by 5G	53
2.3.1	Technical Advancement of 5G	53
2.3.2	5G MAC Layer Solutions for Ultra-Low Latency Performance . . .	55
2.4	Haptic Codecs for the Tactile Internet	57
2.5	Conclusions	59
3	HCCA of IEEE 802.11 WLAN for Latency-Critical Haptic Applications	61
3.1	Introduction	61
3.2	Related Work	63
3.3	Initial Average Uplink Latency Analysis for HCCA	66
3.3.1	Derivation of the First Closed-Form Expression	66
3.3.2	Results and Discussion	75
3.4	Accurate Average Uplink Latency Analysis for HCCA	79
3.4.1	Derivation of the Accurate Closed-Form Expression	79
3.4.2	Results and Discussions	84
3.5	Conclusions	95
4	Predictive Semi-Persistent Scheduling Scheme of LTE for Latency-Critical Services	97
4.1	Introduction	97

4.2	Related Work	99
4.3	Predictive SPS Scheme	102
4.3.1	Feasible RBP Allocator	105
4.3.2	Predictive SPS Schedulers	107
4.4	Simulation Results and Discussion	111
4.4.1	Performance of LAR Algorithm for Latency-Critical Services	112
4.4.2	Performance of LAR and AR-5 Algorithms for Latency-Critical Ser- vices over Heterogeneous Networks	116
4.5	Conclusions	125
5	Predictive Semi-Persistent Scheduling Scheme Powered by Supervised Machine Learning	127
5.1	Introduction	127
5.2	Related Work	129
5.3	A Sequence-to-Sequence Model for Buffer Occupancy Prediction	131
5.4	Simulation Results and Discussion	137
5.4.1	Traced Data Traffic from H2M Experiment	137
5.4.2	Data for Training and Testing Proposed Attention-Based Recurrent Neural Network Model	138
5.4.3	Prediction Accuracy of Proposed Attention-Based Recurrent Neu- ral Network Model during Off-Line Training	138
5.4.4	Simulation Performance of Attention-Based Predictive SPS Scheme	140
5.5	Conclusions	145
6	Conclusions and Future Directions	147
6.1	Introduction	147
6.2	Key Contributions	148
6.3	Future Directions	152
6.3.1	Key Assumptions	152
6.3.2	Reinforcement Learning for Intelligent Resource Scheduling	153
6.3.3	Predictive SPS Scheme Implemented on 5G Networks	154
6.4	Conclusions	155

List of Figures

2.1	Different connections provided by IEEE 802.11.	18
2.2	Flow chart of transmitting a data packet in CSMA/CA operation.	20
2.3	Timeline of transmitting a data packet and an ACK in CSMA/CA.	22
2.4	Timeline of PCF protocol operation.	23
2.5	Flow chart of transmitting a data packet in EDCA operation.	26
2.6	Operation timeline of IEEE 802.11 HCCA protocol.	28
2.7	Mechanism of NPHCCA (adopted from[82]).	31
2.8	LTE radio-access network architecture.	36
2.9	LTE radio resource structure.	39
2.10	Timing diagram of Dynamic Scheme (DS) in LTE.	40
2.11	Timing diagram of Contention Based Scheme (CBS) in LTE.	42
2.12	Timing diagram of Semi-Persistent Scheduling (SPS) in LTE.	44
2.13	Illustration of 1-DoF deadband applied to a signal, force intensity I as a function of time t (adopted from [127]).	58
3.1	Network infrastructure (comprising optical and wireless domains) that enables the support of haptic-tactile applications.	62
3.2	HCCA queueing model.	67
3.3	Graphical argument approach for residual time R	68
3.4	Waiting time for reservation and contention phase intervals under different conditions.	70
3.5	Average uplink latency for T_{SI} equals 0.1 ms, 0.125 ms and 0.150 ms ($M=3$, $N=20$, $i=3$).	76
3.6	Sensitivity Analysis: Impact of TXOP on average uplink latency.	77
3.7	Sensitivity Analysis: Impact of SI on average uplink latency.	77
3.8	Achievable latency performance under different TXOP and SI values.	78
3.9	Achievable throughput under different TXOP and SI values.	78
3.10	Comparison between exact and approximated value of $E(\lfloor \frac{\lambda \overline{W}}{c} \rfloor)$	83
3.11	Average uplink latency performance and parameter variations when $T_{SI} = 80 \mu s$, $0 < \lambda < 25000$ packet/sec.	86
3.12	Average uplink latency performance and parameter variations when $T_{SI} = 80 \mu s$, $25000 < \lambda < 137000$ packet/sec.	87
3.13	Average uplink latency performance and parameter variations when $T_{SI} = 100 \mu s$, $0 < \lambda < 10000$ packet/sec.	88

3.14	Average uplink latency performance and parameter variations when $T_{SI} = 100 \mu s, 10000 < \lambda < 70000$ packet/sec.	89
3.15	Average uplink latency performance and parameter variations when $T_{SI} = 100 \mu s, 70000 < \lambda < 200000$ packet/sec.	90
3.16	Changes of HCCA settings for improved uplink latency by the proposed algorithm.	94
4.1	Timing diagram of the proposed predictive SPS scheme (3 ms SPS periodicity).	104
4.2	RBP allocation diagram of predictive SPS scheme.	105
4.3	Average uplink MAC layer latency under varying traffic loads.	114
4.4	Measured data plane utilisation under varying traffic loads.	114
4.5	Number of connected UEs under varying traffic loads.	115
4.6	Average NMSE of RBP allocation under varying traffic loads.	116
4.7	Average residual length of UEs' buffer under varying traffic loads.	116
4.8	Average uplink MAC layer latency of Tactile UEs.	118
4.9	Average used RBPs per Tactile UE per 10 ms.	119
4.10	Detailed comparison of latency performance between different SPS algorithms.	119
4.11	Average uplink MAC layer latency of non-Tactile UEs.	120
4.12	Overall data plane utilisation of the LTE cell.	121
4.13	Number of connected (admitted) Tactile UEs as a function of total offered traffic load.	121
4.14	Average NMSE of RBP allocation as a function of total offered traffic load.	122
4.15	Average NMSE of predicted buffer occupancy as a function of total offered traffic load.	122
4.16	Comparison of complexity of AR-5 and LAR algorithms.	124
4.17	Average packet drop rate for non-Tactile UEs.	124
5.1	Overall architecture of implemented encoder-decoder model.	131
5.2	Detailed structure of the LSTM unit at encoder.	133
5.3	Additive attention mechanism implemented by the proposed model.	134
5.4	Data pre-processing procedure.	136
5.5	Data arrivals of haptic interaction experiment.	137
5.6	Average uplink MAC layer latency of Tactile UEs.	141
5.7	Detailed comparison of latency performance between AR-5 model and Attention-based model.	141
5.8	NMSE of RBP allocation as a function of total offered traffic load.	143
5.9	NMSE of predicted buffer occupancy as a function of total offered traffic load.	143

List of Tables

1.1	Performance requirements of emerging services	4
2.1	Preliminary analysis of initial access latency	45
2.2	Preliminary analysis of transmission latency	46
3.1	Notations for queueing analysis	68
3.2	Seven scenarios of different waiting time for reservation and contention phase intervals.	71
3.3	Constants for numerical average latency analysis.	75
3.4	Constants for numerical average latency analysis	85
3.5	Total sobol indices of eight timing parameters.	91
3.6	Description of examples and resulting uplink latencies (by simulation) . .	93
4.1	Prediction target and available observations to SPS scheduler for n^{th} SPS transmission	108
4.2	Simulation parameters	113
4.3	Simulation parameters	117
5.1	Notations for Attention-based prediction model	132
5.2	Statistics of data inter-arrival time (ms).	137
5.3	NMSE of prediction for 2 ms periodicity.	139
5.4	NMSE of prediction for 3 ms periodicity.	139
5.5	MAE of prediction for 2 ms periodicity.	139
5.6	MAE of prediction for 3 ms periodicity.	140
5.7	Simulation parameters	140
5.8	Time complexity (3 ms) of executed AR-5 scheduler and proposed Attention-based scheduler.	144

CHAPTER 1

Introduction

1.1 Wireless and Mobile Cellular Communications

Communications network has become an everyday commodity of a majority of the world's population. During recent decades, our society has witnessed a dramatic growth of the Internet, resulting from deployment of multimedia applications, such as video streaming, Voice over IP (VoIP), teleconferencing and multi-player online gaming. Among diverse communications technologies, wireless communications such as IEEE 802.11 (commonly known as Wi-Fi) and mobile cellular communications such as Long-term evolution (LTE) and LTE-Advanced (LTE-A) have attracted a mass market due to their availability, mobility and low cost. Today, wireless and mobile cellular networks have become a significant solution to “last mile” connectivity for Internet access, while wired communications such as passive optical network serve as favourable technologies for high-capacity backhaul [1–3].

The IEEE 802.11 standards were firstly published in 1997, and they specified a set of media access control (MAC) and physical (PHY) layer protocols for deploying wireless local area network (WLAN) and have become a dominant wireless solution in the home, office, building and public areas [4]. Due to an explosive growth of data traffic in heterogeneous networks, the IEEE 802.11 network is also considered as a promising solution to offloading data from mobile cellular networks [5, 6]. The IEEE 802.11 standards have evolved from transmitting data at 2 MBps to competing with the speed of the fastest wired connections, due to the frequently upgraded technologies and standards. To date, the latest complete standard is IEEE 802.11-2016 while several latest amendments, including 802.11aj-2018, 802.11ak-2018 and 802.11aq-2018 are approved. A typical WLAN architecture implemented today comprises multiple **Stations (STA)** and an **Access Point**

(AP). A station is a mobile device that is equipped with access mechanisms to the wireless medium and is able to communicate with the AP via the wireless medium. A station can be used as a direct interface for the user to interact with the Internet or acts as a data collection/transmission device, depending on actual use cases. An AP, connected to a wireless router, handles connection requests and forwards data flows between a station and the Internet [7]. Another widely-deployed network architecture by IEEE 802.11 is called wireless mesh network (WMN). It is formed by multiple mesh STAs to create a backhaul for data exchange between external STAs [8]. In this thesis, however, we focus on the direct wireless connection between STAs and an AP since this offers the minimum network latency from STAs to the Internet.

Unlike WLAN, which is designed for small coverage areas, LTE technologies are proposed by 3rd Generation Partnership Project (3GPP) for wide-area coverage which usually reaches a distance of several kilometres. In recent years, smartphones have reshaped our society due to their capability to complete various tasks on a single digital device. This has significantly stimulated the deployment and research on mobile cellular communications. Since LTE networks are built upon a universally accepted single technology, which is critical to reducing the cost for both users and service providers, LTE has achieved global success in connecting users to the Internet since its first introduction in 2009. In particular, LTE has covered 86.48% of total areas in Australia by 2018 [9]. Similar to IEEE 802.11, LTE also underwent significant technical evolution such as advanced antenna technologies, utilisation of unlicensed spectrum and densified deployments, which have increased the peak data rate of LTE networks to multi-Gbps [10]. LTE standards define protocols comprising packet data convergence protocol (PDCP), radio-link control (RLC), MAC and PHY, to enable a mobile cellular network connection. In a radio access network that implements LTE, an **E-UTRAN NodeB (eNodeB or eNB)** is used in the base station to respond to all radio-related functions, and it communicates with **user equipment (UE)**. An eNodeB is further connected to an evolved packet core (EPC) which permits connected UEs the access to the core network.

1.2 The Tactile Internet

The vision of the future Internet is known as the Tactile Internet, which was first coined in early 2014 [11, 12]. Today's communications networks have evolved a long way from transmitting plain text data to transmitting network traffic generated by emerging applications such as teleconferencing and online gaming. Modern communications networks are not only expected to deliver data at a higher rate but also with less latency, and this is mainly driven by the acceptable delay of human senses. For example, the average reaction time for human auditory sense is around 150 ms, which translates to a one-way latency requirement of less than 150 ms for a smooth and natural verbal conversation [13], and this forms the fundamental performance requirement of VoIP. On the other hand, teleconferencing—an application with higher latency requirement has been realised and has become an everyday commodity for remote business, due to the advancement of communications networks. Such an application exerts a higher demand of round-trip latency—less than 20 ms for an uninterrupted session, sufficing the expected human reaction time when the visual sense is engaged. With the development of machines and mobile devices, the Tactile Internet suggests a natural leap over the communications network capability—from transmitting voice and video content to transmitting haptic content. Applications enabled by haptic communications include tele-rehabilitation and tele-surgery for health care service and controlled exoskeletons in the sports industry.

To provide users with imperceptible time-lag experience, different Tactile Internet services require different levels of latency and reliability performance. This is elaborated by considering specific use cases of the Tactile Internet in the following paragraphs. The corresponding requirements of latency, reliability and user experienced data rate of these services are extracted from 3GPP specifications [14] and shown in Table 1.1.

Process and discrete automation

Process and discrete automation refer to implementing controlled devices (e.g. industrial robots) for enhancing productivity and are the major driving source of the fourth industrial revolution. Traditionally, industrial control has been performed over wired net-

works, however, due to significant cost reduction of installation and maintenance, avoiding physical wear and tear of wired connections and introducing flexibility of deployment [15], wireless communications for process and discrete automation has attracted industrial and research attention in recent years.

A typical scenario of process and discrete automation could be a master station periodically contacting sensors and actuators to aggregate and/or process data, such that the parameters of machine tools can be updated immediately or the production of parts can be moved rapidly. Such a process of discrete automation requires an Ultra-Reliable Low-Latency Communication (URLLC), which makes the Tactile Internet a promising candidate for this application [16]. Feasible communications technologies include IEEE 802.11, power line communications [17] and satellite communications [18].

Intelligent Transportation

Intelligent transportation has been proposed to ensure road safety and driver comfortability due to today's complex traffic conditions and growing number of vehicles. Intelligent transportation refers to a system where a vehicle communicates with other vehicles (V2V communications) or roadside infrastructure (V2I communications), such that the vehicle is capable of making decisions based on information collected by itself or surrounding objects. With the intelligent transportation system, an automated driving vehicle is able to avoid collisions without human intervention. Such an application requires

Table 1.1: Performance requirements of emerging services

Scenarios	End-to-end latency	Reliability	User Data Rate	Experienced
Process automation-remote control	50 ms	99.9999%	1 Mbps ~ 100 Mbps	
Discrete automation	10 ms	99.99%	10 Mbps	
Intelligent transport systems	10 ms	99.9999%	10 Mbps	
Electricity distribution-high voltage	5 ms	99.9999%	10 Mbps	
Remote control	5 ms	99.999%	Low ~ 10 Mbps	
Tactile interaction	0.5 ms	99.999%	Low	

receiving information with a latency of 5~10 ms [15] due to the high-speed mobility of vehicles and significant damages imposed by unsuccessful service guarantee. Current candidates for enabling intelligent transportation are Dedicated Short Range Communications (DSRC) and Fifth Generation (5G) [19–21].

Smart grid

The increasing usage of renewable energy technologies (e.g. solar panels) has resulted in small, distributed energy suppliers, this has led to unsteady energy and out-of-phase power injection into the power grid, which imposes risks for secure energy supply. Smart grid is therefore proposed and developed to coordinate decentralised power suppliers to ensure reliable power supply. The smart grid uses a control grid to gain status information of distributed power grid and use the collected data to control power grid such that optimised utility is achieved [16].

To switch on or off distributed power suppliers properly, the smart grid must obtain all necessary data in time and forward the final control signals immediately, which results in a strict latency requirement. The Tactile Internet is therefore considered an enabler of the smart grid not only due to its guaranteed latency, but also its high security and reliability. Appropriate communications technologies for smart grid include wireless networks such as IEEE 802.15.4, IEEE 802.11, cellular networks [22], and wired networks such as power line communications and fibre-optic communications [23–25].

Remote Control and Tactile Interaction

Remote control and tactile interaction refer to services that allow a person to control robots and interact with real or virtual environments remotely. A typical example of remote control and tactile interaction is remote surgery, where a patient and a surgeon are at different locations, and the surgeon controls robotic arms to execute the treatment procedure. To provide smooth operation and true “immersive experience” for users, these services require communications networks to guarantee the most stringent latency performance (1~5 ms). Such a latency limit is set based on human reaction time to haptic

stimuli [11, 12, 16, 26]. Although some preliminary haptic-related applications are now available such as virtual reality (VR) and controlled robotics, they cannot be conducted over the Internet, due to unacceptable communications latencies. To the end, up-to-date communications technologies such as IEEE 802.15.4 [27], fiber-wireless (FiWi) enhanced LTE-A [11] and 5G [28] have been proposed to fulfil the stringent latency requirement of remote control and tactile interaction services.

We can see that with guaranteed communications latency, the Tactile Internet does not only support human-to-human (H2H) and human-to-machine (H2M) communications but also bring other exciting applications to reality. As such, the realisation of the Tactile Internet is expected to impact our society dramatically. A preliminary analysis has revealed that a potential market of US\$ 20 trillion worldwide could be activated by the Tactile Internet [11].

1.3 Thesis Objectives and Approaches

As explained in the preceding section, ultra-low latency is one of the most important characteristics of enabling wireless technologies of the Tactile Internet. Although current wireless communications such as IEEE 802.11 and LTE have succeeded in meeting the demand of conventional Internet applications, they need to be re-designed to cater to these emerging services [16]. On the one hand, existing wireless communications are not able to deliver packets with an ultra-low latency, which necessitates a rethink of their underlying frame structure and transmission mechanism such that latency of wireless transmission can be significantly reduced. On the other hand, conventional wireless communications were designed for content-centric networks, which mainly transmit medium to large data packets. In services of the Tactile Internet, however, content as well as control information need to be delivered on time, creating a big paradigm shift from content-centric networks to control-centric networks. In other words, the wireless technologies to support future networks need to be re-designed for data with small and constant payload size.

The overall latency in communication system consists of processing latency T_{proc} ,

queueing latency T_{queue} , transmission latency T_{trans} and propagation latency T_{prop} , where T_{trans} depends on processing speed of communications devices and frame structures of LTE standards and T_{prop} depends on transmission distance and speed of radio signal. In other words, T_{trans} and T_{prop} can only be reduced by redefining LTE standards and hardware advancement. On the other hand, T_{proc} is affected by the transmission protocols deployed by communication standards and T_{queue} is significantly influenced by scheduling functions. In this thesis, we study the feasibility of applying IEEE 802.11 and LTE technologies for latency-critical services by novel MAC layer designs for uplink transmissions, which is the key to reducing the T_{proc} and T_{queue} in communications networks. As explained in the preceding section, both IEEE 802.11 and LTE technologies have achieved competitive transmission speed, compared with the fastest wired networks. Therefore, attempts to further enhance the wireless bandwidth might not lead to significant latency reduction. Instead, a recent investigation [29] has revealed that the major bottleneck of supporting mobile VR in LTE networks lies in MAC protocol overhead. In particular, the delay introduced by queueing packets and waiting for grants for accessing wireless medium in a multi-user environment has a significant contribution to wireless transmission latency. Meanwhile, this thesis targets at shortening transmission latency in the uplink direction. An uplink transmission occurs when information is transmitted from a communications equipment towards the network core, which corresponds to from a station to an AP in an IEEE 802.11 network or from a UE to an eNodeB in an LTE network, respectively. Whereas a downlink transmission refers to cases where information is transmitted in the opposite direction. It is more difficult to achieve ultra-low latency for uplink transmissions than downlink transmissions [30]. This is because unlike downlink transmissions where an AP/eNodeB can directly send downlink packets to STAs/UEs, STAs/UEs awaiting uplink transmission need to obtain permission to access the radio resources before they can send uplink data, which introduces additional latency source. As a result, the uplink latency is often significantly larger than the downlink latency and becomes a bottleneck of applying wireless communications technologies for latency-critical services.

We first investigate the MAC layer protocols of IEEE 802.11 networks and determine

factors that have the most significant impact on uplink latency. Conventional IEEE 802.11 networks deploy a MAC protocol called distributed coordination function (DCF), which allows STAs to contend for accessing the wireless medium and attempts to avoid packet collision via random backoffs. The probability of collision instigated by this protocol is susceptible to the number of STAs and could reduce reliability and increase latency when a collision happened. Therefore, we focus on one of the latest IEEE 802.11 MAC protocols—HCF Controlled Channel Access (HCCA). HCCA was first defined in IEEE 802.11e amendment, and it essentially implements a polling system for connected STAs with a guaranteed amount of Transmission Opportunity (TXOP). Such protocol does not only avoid the packet collision problem of the traditional DCF but also introduce flexibility by variable TXOP allocation.

Next, in order to study the feasibility of applying HCCA for latency-critical services, the latency performance of HCCA needs to be evaluated. We develop a mathematical model that captures the operation procedure of HCCA and then derive the first closed-form expression for the average MAC layer uplink latency. The accuracy of the derived expression is evaluated with a discrete-event simulation in SimEvents® [31]. The analytic expression contains several timing and network parameters whose values need to be carefully chosen by network designers for ultra-low latency performance. We then conduct a global sensitivity analyses on the derived expression to gain understandings of implications of these timing and network parameters on the uplink latency.

Then, we develop an effective strategic parameter selection algorithm for the HCCA such that stringent latency requirement can be met. This strategic parameter selection algorithm is designed based on the knowledge gained from the global sensitivity analyses. Specifically, it tunes the timing and network parameters and evaluates the uplink latency performance of current network settings via the derived expression. It should be noted that our analysis considers data packets with small and constant payload size, which conforms to the data type in haptic communications [32]. Some existing literature has focused on improving the throughput and fairness of HCCA by adjustable TXOP [33–36] instead of reducing latency. Other literature attempts to introduce modifications on either IEEE 802.11 data structure or HCCA operation [37–42] to achieve reduced latency,

but the considered data (e.g. video frames) are content-centric instead of control-centric, therefore the achieved results may be inapplicable for the Tactile Internet.

The second objective of this thesis is to investigate the suitability of applying LTE to Tactile Internet services. We compare three major uplink MAC schemes supported by LTE—Dynamic Scheme (DS), Contention-Based Scheme (CBS) and Semi-Persistent Scheduling (SPS) scheme. As explained in the preceding section, the conventional DS scheme suffers from additional delay for transmission grant acquisition. Although some existing literature proposed advanced schedulers for improved quality-of-service (QoS) [43–46], the intrinsic extra delay of the DS scheme remains. The CBS, on the other hand, allows a UE to instantly transmit uplink data using shared radio resources [47, 48]. However, CBS may suffer from packet collision and thus leads to increased uplink latency. The SPS scheme was originally developed for quality service of VoIP, and it allows a UE to transmit data using a fixed amount of radio resources periodically. For variable bit rate (VBR) data traffic, however, SPS may suffer from extra delay introduced by a mismatch between allocated resources and actual resource demand. We select the SPS as our candidate for latency-critical applications based on two reasons. First and foremost, the contention-free nature of the SPS matches well with the high reliability requirement of the Tactile Internet. Second, the SPS does not use control channel during uplink transmissions, which benefits other services that consume control channel resources (e.g. massive machine-type communications).

To solve the resource mismatch problem, we propose a feasible resource allocator (FRA). Instead of using a fixed amount of radio resources, the FRA allows variable resource allocation and hence provides flexibility for the proposed SPS scheme. Next, a predictive scheduling function is developed to alleviate the resource mismatch issue of the original SPS scheme, by scheduling radio resources according to predicted resource demands.

Simulation results reveal that prediction accuracy has a significant impact on the scheduling efficiency and latency performance of the proposed SPS scheme, especially for a fast-changing data flow with bursty traffic. Therefore, the third objective of this thesis is to study and exploit an advanced technique—machine learning to enhance the over-

all performance of the proposed predictive SPS scheme. First, we interpret the associated problem into a Sequence-to-Sequence (seq2seq) task in the supervised machine learning paradigm [49, 50]. Then, we apply an attention mechanism to capture temporal dependencies within a given sequential input in order to generate accurate traffic prediction [51]. Our proposed method is different to the ones proposed in [52, 53], such that dimension expansion and an initial decoder input with learnable parameters are introduced for extra prediction flexibility. The prediction accuracy of our proposed machine learning model during off-line training is compared with that of existing attention-based prediction models, and results indicate that it achieves the best performance. Finally, simulation results show that our proposed predictive SPS scheme implementing the attention-based model is able to further enhance the latency performance as well as scheduling efficiency due to improved prediction accuracy during network simulation.

It is important to note that we wish to validate our proposed algorithms and protocols against real-life haptic-tactile applications. Therefore, for the first objective of this thesis, we implement Poisson traffic to model the random and irregular data traffic of the Tactile Internet for analysing and evaluating the average uplink latency of HCCA and effectiveness of the strategic parameter selection algorithm. While for the second and third objectives, we employ traced data traffic from actual haptic-tactile experiments and apply them to a system-level simulation framework based on SimuLTE [54].

The objectives of this thesis are summarised as follows:

- To characterise delay components contributing to overall network latency and identify the bottleneck of applying IEEE 802.11, LTE and 5G technologies for the Tactile Internet.
- To investigate the latency performance of IEEE 802.11 HCCA MAC protocol and propose a feasible solution to applying HCCA for ultra-low latency applications.
- To investigate the latency performance of SPS protocol of LTE/5G and propose a predictive SPS scheme for ultra-low latency applications.
- To predict the future radio resource requirements using Attention-based recurrent neural network and investigate its capability to further reduce the latency of the

proposed predictive SPS scheme.

1.4 Thesis Outline and Original Contributions

This thesis consists of six chapters, including the current chapter, Introduction (Chapter 1). In the current chapter, we discuss the background, motivation, thesis objectives, outline and original contributions of this thesis. The rest of the thesis is structured as follows:

Chapter 2: Introduction to Low-Latency Wireless and Mobile Cellular Networks

Chapter 2 presents a detailed discussion of operation details of MAC layer protocols of IEEE 802.11, LTE and 5G networks and how latency control is achieved in these existing protocols. In particular, to better understand the motivation of our proposed solutions, we first explain how IEEE 802.11, LTE and 5G are implemented in an example Tactile Internet scenario. We then discuss the limiting factors of latency performance of these protocols with a basic uplink latency analysis. We then discuss solutions proposed to reduce MAC layer latency of IEEE 802.11, LTE and 5G networks the existing literature.

Chapter 3: HCCA of IEEE 802.11 WLAN for Latency-Critical Haptic Applications

This chapter presents our first mathematical model of the uplink MAC layer latency for a Tactile Internet scenario which implements IEEE 802.11 Hybrid Coordination Function (HCF) Controlled Channel Access (HCCA) MAC protocol and a novel strategic parameter selection algorithm for effective uplink latency reduction. We show with MAC layer level simulations, that our proposed algorithm is able to bring the HCCA uplink MAC layer latency down to below 0.1 ms, with lower throughput being the trade-off. It is important to note that the simulations are performed using MATLAB [55] and the parameters used in the simulation comply to IEEE Standard 802.11-2016 [56].

Original contributions and publications from this chapter include:

- A closed-form formula for the average uplink latency of HCCA of IEEE 802.11,

which can be expressed using numerical values of various HCCA parameters. The formula is derived from an M/G/1 queueing model based on the operation of the HCCA protocol. In particular, it considers the fixed reservation periods created by Beacon, Poll and Acknowledge frames and variable reservation periods resulted from contention intervals. The expression also includes two approximation techniques for accurate results. A discrete event simulation that implements the MAC layer operation of HCCA protocol is developed using MATLAB [55]. It is used to verify the accuracy of the derived expression of the average uplink latency of HCCA.

- Implications of various timing and network parameters on the average uplink latency of HCCA, which is obtained by global sensitivity analyses. The analyses consider a total of 8 timing parameters and use a variance-based method and Monte Carlo method based on the closed-form expression of the average uplink latency of HCCA.
- A novel strategic parameter selection algorithm that uses the insights gained from the global sensitivity analyses and the proposed expression of the average uplink latency, to adaptively change the timing parameters of the HCCA protocol. The algorithm recommends new parameter values iteratively until the resulting average uplink latency is below a target value. It is important to note that we point out that an ultra-low average uplink latency could result in a decreased number of supported users.
- Y. Feng, C. Jayasundara, A. Nirmalathas, and E. Wong, "A Feasibility Study of IEEE 802.11 HCCA for Low-Latency Applications," *IEEE Transactions on Communications*, vol. 67, no. 7, pp. 4928-4938, July 2019.
- Y. Feng, C. Jayasundara, A. Nirmalathas, and E. Wong, "IEEE 802.11 HCCA for Tactile Applications," in *2017 27th International Telecommunication Networks and Applications Conference (ITNAC)*, Nov 2017, pp. 1-3.
- Y. Feng, C. Jayasundara, A. Nirmalathas, and E. Wong, "Hybrid Coordination Func-

tion Controlled Channel Access for Latency-Sensitive Tactile Applications,” in GLOBE-COM 2017 - 2017 IEEE Global Communications Conference, Dec 2017, pp. 1-6.

Chapter 4: Predictive Semi-Persistent Scheduling Scheme of LTE for Latency-Critical Services

In this chapter, we study the suitability of applying LTE for latency-critical haptic applications. We exploit the idea of SPS and design a novel predictive SPS protocol that does not only allocate radio resources periodically but also reduce uplink latency of an LTE network by mapping the predicted resource demand to variable resource allocation. We also show that our proposed SPS protocol allows coexistence of other MAC protocols, such that the preserved-but-unused radio resources of the SPS scheme can be allocated for other non-SPS network traffic. We utilise a well-known discrete event simulation platform OMNeT++ to prove the effectiveness of our proposed protocol. In particular, the predictive SPS protocol can achieve a ~ 3 ms uplink latency. Although this uplink latency performance satisfies latency-critical applications such as remote control, we point out that the sub-millisecond latency of a Tactile Internet could be achieved by deploying the predictive SPS protocol in a 5G system.

Original contributions and publication from this chapter include:

- A Feasible Resource Allocator (FRA) is proposed for variable SPS allocation functionality. The FRA determines the number as well as the location of a set of preserved radio resources for each SPS-connected UE. The location is determined such that resource preservation of SPS-connected UEs is safe from resource collisions in future uplink transmissions. During uplink transmissions, if preserved radio resources are not fully utilised, the FRA is able to schedule the unused resources for other non-SPS uplink transmissions.
- A predictive scheduler that takes recorded information including transmitted data, buffer status and granted resources as input and calculates a predicted resource demand for the immediate next uplink transmission using a simple auto-regressive model of order 5. The prediction is made in two steps. First, it extracts the informa-

tion of newly arrived data between adjacent uplink transmissions and forms a time series for a k -step ahead prediction. Second, it estimates a UE's buffer occupancy based on the predicted newly arrived data and previous buffer sizes.

- A discrete event program that simulates the operation of the proposed FRA and predictive scheduler is developed using SimuLTE and OMNeT++. For simulations, we assume there exist 20 non-SPS UEs and 10 SPS UEs, and we demonstrate the effectiveness of the proposed predictive SPS protocol by comparing its performance with other existing protocols with following metrics—average latency, resource utilisation, data plane utilisation, number of allowed SPS UEs, prediction accuracy, buffer occupancy, computation complexity and mean packet drop rate of non-SPS UEs. Results show that our proposed predictive SPS protocol achieves an average uplink latency of ~ 3 ms when SPS periodicity is 1 ms.
- Y. Feng, A. Nirmalathas and E. Wong, “A Predictive Semi-Persistent Scheduling Scheme for Low-Latency Applications in LTE and NR Networks,” in ICC 2019 - 2019 IEEE International Conference on Communications (ICC), May 2019, pp. 1-6.

Chapter 5: Predictive Semi-Persistent Scheduling Scheme Powered by Supervised Machine Learning

This chapter presents a machine learning model for further improving the performance of our proposed SPS scheme. We employ a recurrent encoder-decoder model, which achieved success in natural language translation, and apply an attention mechanism to selectively identify relevant input states for a more accurate target prediction. We use an off-line method to extract the recorded SPS information during simulation and use this to train our proposed model. Finally, we deploy the trained Attention-based model in system-level simulations with our proposed predictive SPS scheme to achieve improved latency performance and scheduling accuracy. Furthermore, we point out that the Attention-based model produces more accurate predictions for higher traffic loads.

Original contributions from this chapter include:

- An Attention-based recurrent neural network for predictive SPS scheme. In particular, the sequential-format input and output of the corresponding prediction task are firstly processed with dimension expansion and logarithmic normalisation to allow high-dimension feature detection and to remove data skewness. Then, data is fed into an encoder-decoder architecture such that temporal dependencies within input can be captured. An additive attention mechanism is implemented to let the proposed model selectively focus on the most relevant input states for improved prediction accuracy. Finally, at the prediction stage, an initial input with learnable parameters is introduced at the decoder, which enables the model to perform with extra flexibility.
- A preliminary Exploratory Data Analysis (EDA) of a traced traffic during a real-time haptic-tactile experiment that involves free space moving, touching the surface of a virtual object and grasping/moving multiple virtual objects. Specifically, we show that the data traffic generated by the haptic-tactile experiment exhibits randomness and burstiness.
- An off-line training framework for multiple alternative time series prediction techniques including statistical method (Auto-Regressive), simple feed-forward artificial neural network and other attention-based models. By comparing the Normalised Mean Square Error (NMSE) and Mean Absolute Error (MAE) of these models against test dataset, we show that our proposed model achieves the best prediction accuracy.
- A predictive SPS scheme that implements Attention-based recurrent neural network. The simulation is based on the traced data traffic of a real-time haptic-tactile experiment, and system-level simulation results reveal that this Attention-based SPS scheme can further improve the latency performance and scheduling efficiency, as compared with the SPS scheme with AR-5 model.

Chapter 6: Conclusions and Future Directions

This chapter firstly summarises the motivation and key contributions of this thesis. Then, future research directions that could be extended beyond this study are discussed, these include an investigation into the actual traffic flow (e.g. arrival pattern and data sizes) generated by the Tactile Internet, reinforcement learning for intelligent scheduling decision and performance evaluation of 5G network integrated with predictive SPS scheme. Finally, based on the simulation performance of our proposed algorithms and schemes, we argue that HCCA protocol and predictive SPS scheme are viable solutions to the wireless communications for the Tactile Internet deploying WLAN and mobile cellular networks, respectively.

Other related technical publications include:

- Y. Lv, M. P. I. Dias, L. Ruan, E. Wong, Y. Feng, N. Jiang, and K. Qiu, "Request-Based Polling Access: Investigation of Novel Wireless LAN MAC Scheme for Low-Latency E-Health Applications," *IEEE Communications Letters*, vol. 23, no. 5, pp. 896-899, May 2019.
- Y. Lv, L. Ruan, M. P. Imali Dias, E. Wong, Y. Feng, N. Jiang, and K. Qiu, "Dynamic Polling Sequence Arrangement for Low-Latency Wireless LAN," in *2018 Asia Communications and Photonics Conference (ACP)*, Oct 2018, pp. 1-3.
- L. Ruan, M. P. I. Dias, Y. Feng, and E. Wong, "Round-Trip Delay Modeling for Smart Body Area Networks," *IEEE Communications Letters*, vol. 21, no. 11, pp. 2528-2531, Nov 2017.

Introduction to Low-Latency Wireless and Mobile Cellular Networks

2.1 Wireless Networks Supported by IEEE 802.11

IEEE 802.11 WLAN is now widely deployed across the globe. Since its first release in 1997, wireless networks based on IEEE 802.11 have played a vital role in keeping the society connected. Currently, the IEEE 802.11 network is capable of providing fast broadband services at home and in public (e.g. campus or small buildings), in which case users could be stationary or moving at pedestrian speed. IEEE 802.11 is expected to have a fundamental influence on next-generation wireless communications networks due to its simplicity, flexibility and cost-effectiveness [57]. For instance, fibre-wireless (FiWi), which provides the concept of combining the advantages of fibre optical network (fast, reliable and high bandwidth) and IEEE 802.11 network (availability and wireless access), is becoming an important research topic [2, 58, 59]. In [11], a FiWi network consisting of a WLAN and an Ethernet Passive Optical Network (EPON) can significantly reduce the end-to-end latency and energy consumption by employing power saving mode (PSM) in the MAC layer, which makes IEEE 802.11 a potential candidate as the wireless access technology for the Tactile Internet.

As described in Section 1.1, within a WLAN constructed by IEEE 802.11 protocols, STAs can either directly communicate to other STAs within the service range, or transmit packets to other STAs via AP. Fig. 2.1 illustrated these two connection types offered by IEEE 802.11, where STA_1 and STA_2 are wirelessly connected to AP_1 and STA_3 is wirelessly connected to AP_2 . The APs are further connected to the distribution services which corresponds to wired Ethernet networks, by which the Internet access and inter-AP com-

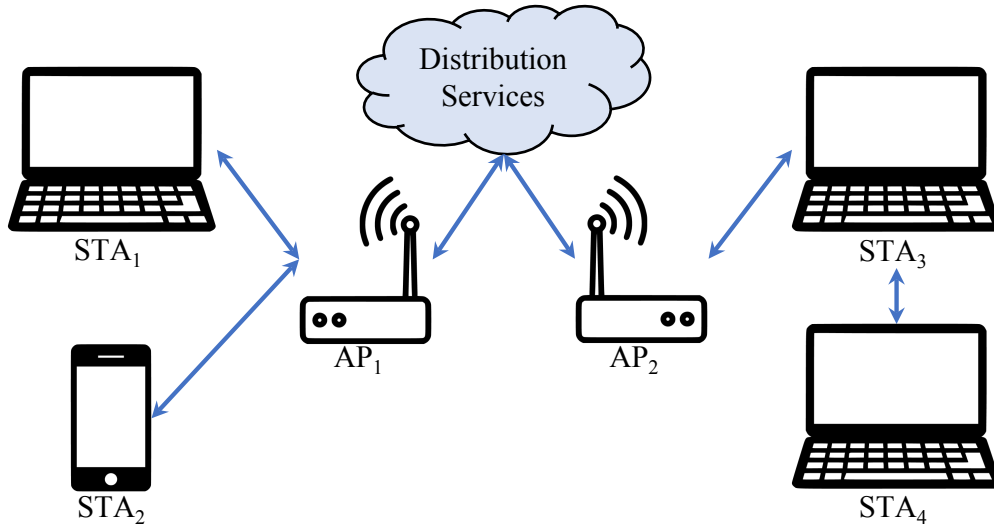


Figure 2.1: Different connections provided by IEEE 802.11.

munications are enabled. As such, the STA₁, STA₂ and STA₃ have wireless Internet access within the coverage of each service set. Furthermore, STA₃ is functioning as a hot spot, and it does not only bridge network traffic from STA₄ to AP₂, but also directly communicate to STA₄.

In the following sections, we will briefly introduce PHY and MAC layer technologies provided by IEEE 802.11. Then, we will discuss and elaborate operation principles of MAC uplink transmission schemes of IEEE 802.11 and their strengths and weaknesses in terms of QoS. Existing literature proposed for improving uplink latency performance with enhanced MAC layer protocols will also be discussed.

2.1.1 Technical Aspects of IEEE 802.11

IEEE 802.11 standards define the PHY and MAC layer protocols deployed in WLAN [56], each with their own functionalities and hence future PHY layer upgrades can be developed and deployed without the need of altering MAC layer protocols. The legacy IEEE 802.11-1997 defined three alternative PHY layers to support wireless transmission—infrared (1 Mbps), Frequency Hopping Spread Spectrum (FHSS) (1 or 2 Mbps) and Direct Sequence Spread Spectrum (DSSS) (1 or 2 Mbps). However, these transmission speeds could not fulfil the demands by the unprecedented growth of versatile wireless appli-

cations. Since the development of IEEE 802.11a, WLAN adopted Orthogonal Frequency Division Multiplexing (OFDM) as its PHY layer transmission protocol, in which case transmission speed of more than 54 Mbps could be supported.

The conventional MAC layer protocol of IEEE 802.11 deploys Carrier-Sense Multiple Access with Collision Avoidance (CSMA/CA). As in a wireless medium, a STA cannot directly detect the existing transmissions within a service set, which causes transmission collisions when multiple STAs transmitting data packet at the same time. Therefore, the conventional IEEE 802.11 MAC layer protocol can only attempt data transmission with a collision avoidance mechanism given no coordination was maintained between individual STAs. CSMA/CA allows multiple STAs to contend for the wireless medium, such that there is at most one STA transmitting a frame via the medium at a time. The general idea behind CSMA/CA is to listen to the channel for any ongoing transmission before sending.

First, it should be noted that the operation of CSMA/CA is based on slotted time. Specifically, there are three types of inter-frame spacing (IFS)—short IFS (SIFS), point coordination function IFS (PIFS) and distributed coordination function IFS (DIFS) and they are defined in the form of time (μs). In particular, SIFS is the smallest IFS to separate transmissions of different frames (e.g. data frame and acknowledgement frame) from the same sender, and its value is $10 \mu\text{s}$ [56]. SIFS is used as a guard time during which the channel must be idle, such that inter symbol interference could be minimised [60]. On the other hand, DIFS is the waiting time before a STA is allowed to send a new transmission. All STAs with the same service range are required to listen to the wireless medium to determine whether there is an ongoing transmission to avoid collisions. Each STA maintains a network allocation vector (NAV) timer, which instructs individual STA about the delay time of queued packets. In other words, STAs can only start transmission after their NAVs reach zero.

The detailed operation procedure of CSMA/CA is illustrated in the flow chart in Fig. 2.2. If a source STA wanted to transmit a data packet to a destination STA or AP, it first needed to listen to the wireless channel. If the medium was idle for DIFS, the source STA then sent its data immediately. On the other hand, if the channel was sensed

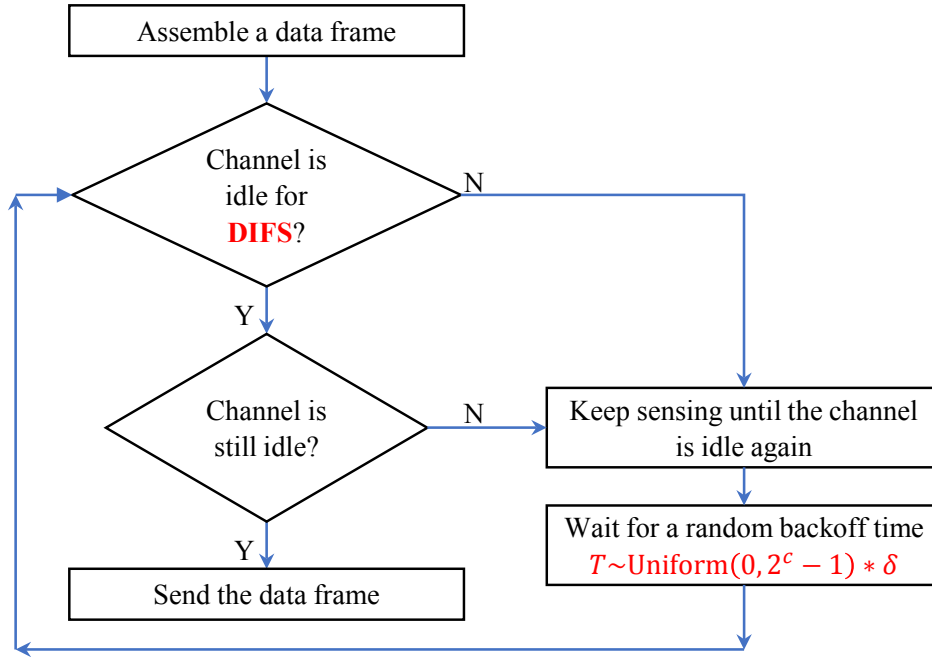


Figure 2.2: Flow chart of transmitting a data packet in CSMA/CA operation.

busy within one DIFS, the source STA shall update its NAV and keep sensing the channel until it was idle (i.e. when ongoing the transmission was finished). When the channel became available again, the source STA would go through the previous procedure after a random backoff time. It should be noted that a timer is used to count down the random backoff time and it decrements only when the channel was sensed idle.

If the destination STA received the data correctly, it should respond by sending back an acknowledgement frame (ACK), at SIFS after data reception. Fig. 2.3 represents the corresponding timeline of the discussed procedure. However, if the ACK was not received by the source STA, a collision was assumed to occur, and the collided data packet would be retransmitted.

2.1.2 Uplink Transmission Protocols of IEEE 802.11

2.1.2.1 Distributed Coordination Function (DCF)

CSMA/CA is a fundamental multiple access mechanism for wireless communications, and IEEE 802.11 deploys a MAC layer protocol called Distributed Coordination Func-

tion (DCF) to determine the random backoff time to enable better collision avoidance. In particular, DCF used a binary exponential backoff algorithm to determine the amount of deferral applied to each sensing period in the CSMA/CA mechanism, which attempts to disperse the transmitting instant of awaiting STAs statistically. The reason for implementing such an approach is two-fold. First, STAs benefit from the reduced collision probability by retransmitting at random backoff time in the wireless medium. Second, the collision probability increases with the number of STAs in an IEEE 802.11 network, which is unknown by an AP. The maximum possible backoff time is exponentially increased after a detected signal collision, such that STAs have a significant chance to successfully transmit in the next attempts regardless of the number of contending STAs. In DCF, after c times of collisions, the starting value of the random backoff timer in a CSMA/CA mechanism is set as $N_c\delta$, where $N_c \sim \text{Uniform}(0, 2^c - 1)$ and δ is the granularity of a backoff slot.

There exist extensive studies on the performance of DCF for WLAN networks, where M/G/1 or G/G/1 queueing models and Markov chain were frequently used for these analyses [61–65]. In particular, Bianchi [66,67] has developed a two-dimensional Markov chain to model the behaviour of DCF, which enables calculating the probability of saturation throughput and transmission failure due to collision by closed-form expressions. The latency performance and collision probability of DCF are largely affected by the number of contending STAs [65]. On the one hand, more STAs competing for channel access leads to a higher probability of multiple STAs attempting to transmit at the same time, which causes a higher collision rate. On the other hand, more STAs trying to send their data packets results in less channel idle time of transmission opportunity for all STAs within the service range, which significantly deteriorates latency performance and in turn, leads to higher packet blocking rate.

A drawback of this simple DCF protocol is that it does not contain service differentiation for different types of data traffic. As pointed out in [57], the throughput and latency performance of all active traffic within a WLAN would decrease when the number of STAs exceeded a certain limit. This can be explained in a queueing model aspect, that a DCF-assisted WLAN has only one unified queue that is shared by all traffic flows, and

thus they shall experience the same amount of latency performance. Therefore, the traditional IEEE 802.11 DCF does not provide QoS guarantee for high-priority traffic flows such as audio and video traffic, unless additional admission control is applied. In addition, numerical analysis has proven that the binary exponential backoff mechanism results in a heavy-tailed latency distribution when the number of collision-caused retransmission is unlimited [68]. This suggests that the contention-based nature of DCF is prone to long latency and is not suitable for real-time applications [69]. The disadvantage of DCF necessitated the development of the Enhanced Distributed Channel Access (EDCA), which will be discussed in Section 2.1.2.3.

2.1.2.2 Point Coordination Function (PCF)

Point Coordination Function (PCF) was introduced in IEEE 802.11 as an alternative MAC protocol, but it is not widely implemented in IEEE 802.11 hardware devices due to its complexity. Different to the DCF, PCF provides a contention-free solution for the transmissions between STAs and a corresponding access point (AP). Within a WLAN employing PCF, an AP acts as a point coordinator (PC) to actively poll each STA. First, the AP gains control of the wireless channel with the highest priority by sensing the medium to be idle for one PIFS period, which is much smaller than DIFS. Then, the AP divides the channel airtime into super-frames containing a contention-free period (CFP) and a contention period (CP), where STAs are allowed to transmit data once it is polled by the

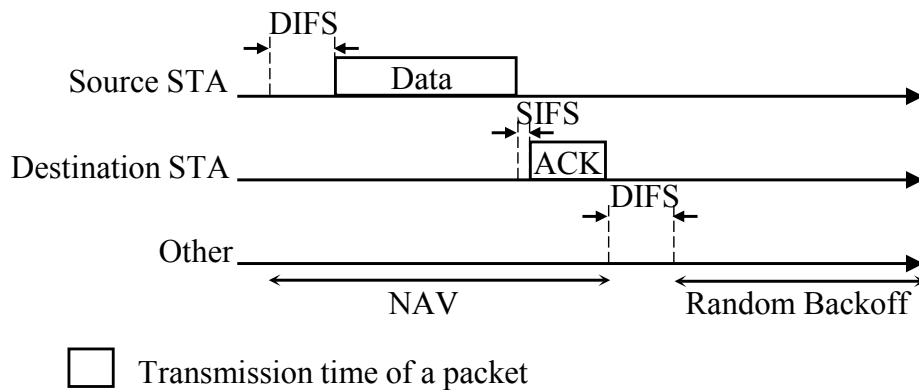


Figure 2.3: Timeline of transmitting a data packet and an ACK in CSMA/CA.

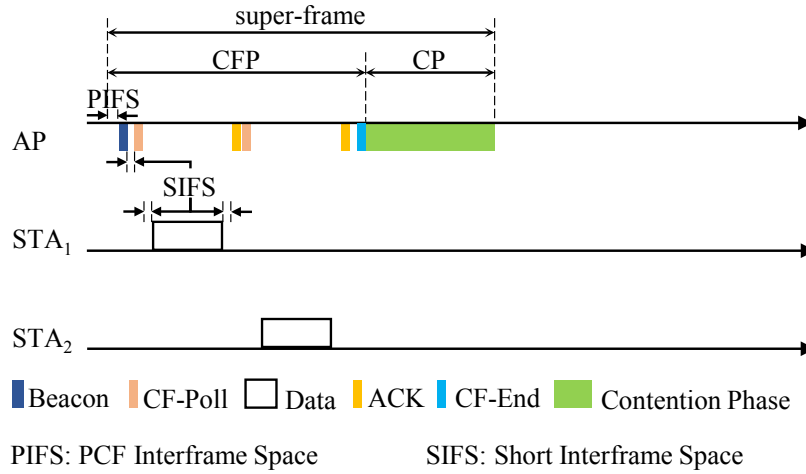


Figure 2.4: Timeline of PCF protocol operation.

AP during CFP and STAs only use DCF for contention-based transmissions during CP. At the start of a super-frame, the AP broadcasts a beacon which is used to denote the next beacon transmission time. In this way, all active STAs will remain still by setting corresponding values of their NAV, unless a CF-End or CF-END+ACK frame is received.

During operation, the AP will maintain a list of registered STAs and polls each STA according to this list. In particular, the AP sends a CF-Poll frame in the downlink direction to the first STA in the polling list, at SIFS after broadcasting the beacon. At SIFS after the reception of CF-Poll, the first STA will send its data packet/s in the uplink direction to the AP. If the uplink data is successfully received, the AP then sends an ACK frame to the first STA and polls the second by sending a second CF-Poll frame. It should be noted that a STA will remain silent if it did not have pending data, and a STA being silent for a long time shall be removed from the polling list. The operation of an AP polling 2 STAs using PCF protocol is illustrated in Fig. 2.4.

It is shown in [70] that when the number of STAs increases, PCF offers higher saturation throughput compared with DCF. This is because PCF can support a larger number of successfully transmitted packets due to its contention-free nature. However, PCF could potentially waste channel bandwidth if a STA was polled during which it does not have uplink data to transmit.

Although the PCF was designed for latency-bounded services, it still may lead to poor

QoS performance due to several reasons. First, PCF suffers from the same drawback as DCF—they cannot provide differentiated QoS for traffic flow with distinct characteristics and service requirements. This is because the AP polls each registered STAs in a simple round-robin fashion without any admission control mechanism, which could potentially lead to an unbounded waiting time before a STA gets its opportunity for transmission. As such, the PCF does not guarantee QoS, especially when traffic load is high [71]. Second, the AP-centric polling scheme brings additional latency when two STAs within the service range wants to communicate with each other. Third, the introduction of CP within the super-frame may lead to unpredictable latency for STAs in the next polling cycle. Specifically, an AP releases the control of the wireless medium during CP, and all STAs are allowed to contend for channel access via DCF. Therefore, the AP is not guaranteed to obtain the control of wireless medium again by sensing the channel for the shortest time period (PIFS), if the channel was found busy at the starting time of the next super-frame. In this case, the beacon frame, as well as registered STAs, are delayed [72,73].

An adaptive coordination function (ACF) was introduced in [69], where instead of separating a super-frame into a CFP and a CP, the ACF divides a super-frame into an adaptive period and a TDMA period. Within the adaptive period, STAs are allowed to transmit uplink resource requests by Request For Slot (RFS) frame—a new control frame explicitly designed for ACF. Additionally, STAs are also allowed to transmit packets using either DCF or PCF during the adaptive period. After receiving RFS frames, the AP then exploits an admission control and slot reservation algorithm to allocate channel resources according to RFS frames. When TDMA period starts, the AP firstly broadcasts a beacon frame which contains AP's decision about admitted uplink transmissions. It should be noted that the proposed ACF only allows transmissions of real-time traffic in the TDMA period, while other latency-tolerant traffic flows shall use the adaptive period for uplink transmission. By analytically analysing the performance of PCF and DCF protocols under various network conditions (e.g. number of STAs, packet arrival rate, payload size etc.), a network operator can choose the best protocol for adaptive period operation. However, the overall performance gain of the ACF was not discussed in [69], and the author did not suggest a reference algorithm for admission control and slot reser-

vation.

2.1.2.3 Hybrid Coordination Function (HCF)

IEEE 802.11e has proposed a Hybrid Coordination Function (HCF) in order to support both integrated service (IntServ) and differentiated services (DiffServ) of QoS on the Internet. The HCF contains two media access controls, namely Enhanced Distributed Channel Access (EDCA) and HCF Controlled Channel Access (HCCA). To improve the performance of the conventional DCF and PCF protocols in terms of providing differentiated QoS, HCF has defined access category (AC) queues and traffic stream (TS) queues at the MAC layer. When a data packet of a traffic flow arrives at the MAC layer of a STA or an AP, it is assigned with a traffic priority identifier (TID), and it determines if EDCA or HCCA shall be used for the transmission of this traffic flow. Another new feature introduced by HCF is Transmission Opportunity (TXOP), and it is a contention-free transmission time assigned for a specific STA. Different to the PCF, a TXOP can be allocated to either a CFP or CP in the HCF protocol. A detailed explanation of EDCA and HCCA is presented in the following sections.

Enhanced Distributed Channel Access (EDCA)

In essence, EDCA is DCF powered by traffic flow prioritisation. In other words, high-priority traffic flows should have a higher chance to start a transmission. The prioritisation enabled by EDCA is achieved by two mechanisms. First, a STA deploying EDCA uses an arbitration IFS (AIFS) instead of DIFS. The AIFS of each AC is determined as follows:

$$\text{AIFS}_i = N_i \delta + \text{SIFS} , \quad (2.1)$$

where i denotes the index of each AC and δ represents one slot time. Specifically, the value of N is set to 1 and 2 for higher and lower priority ACs, respectively. When EDCA is deployed, a STA wishing to transmit data needs to sensing the channel to be idle for AIFS, in which case higher priority ACs with shorter AIFS will have greater chance to start transmissions compared with lower priority ACs.

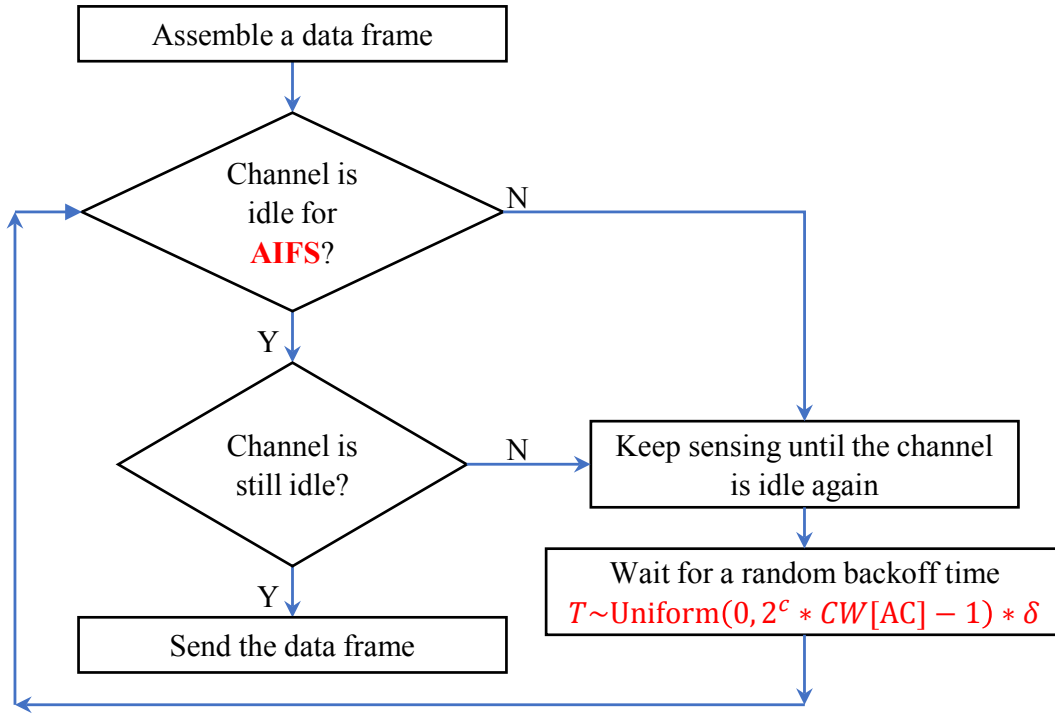


Figure 2.5: Flow chart of transmitting a data packet in EDCA operation.

Second, different random backoff times are allocated to different ACs within a STA. As mentioned in Section 2.1.2.1, if STAs sensed the channel busy, they need to start backoff timers before the second transmission attempt. With EDCA, the range of the random number generator for the backoff timer of a higher priority AC is smaller than a lower priority AC. Therefore, prioritisation is also provided in the case of channel sensed busy [74].

Furthermore, EDCA exploits service differentiation with the idea of AC and uses an “internal collision” mechanism to provide better performance than the DCF. With EDCA, an AC queue needs to sense the existence of other AC queues within a STA, and it has to start a backoff timer whenever a higher priority AC queue is present. This effectively creates an “internal collision” avoidance mechanism, in which case the scheduler within a STA will grant the EDCA-TXOP to the highest priority AC when an internal collision happened. As an analogy to the DCF, the flow chart representing the operation of EDCA is presented in Fig. 2.5, where CW denoted the contention window which defines the default backoff timer for each AC.

Although EDCA provides a solution to prioritising different traffic flows due to their unique QoS requirements and reduce the “internal collision” within a STA, the external collisions between the same priority ACs from multiple STAs are still high when network is heavily loaded [75]. It is shown by analytical and stimulative evaluation [76,77], that the average latency experienced by a data packet grows with the number of traffic flows when EDCA is employed.

HCF Controlled Channel Access (HCCA)

The traditional DCF and its modified version-EDCA allows opportunistic uplink transmission from a source STA to the destination AP, in which case collision is still possible. Due to this nature, only a soft QoS can be provided by the DCF and EDCA. On the other hand, the latest HCCA scheme deploys a centralised QoS-aware access point (QAP) whereby a hybrid coordinator (HC) assigns resources to associated STAs based on traffic characteristics reported in a traffic specification frame (TSPEC) [56]. The TSPEC contains information such as maximum service interval (MSI), mean data rate, and nominal MAC service data unit (MSDU) size. The operation of HCCA is illustrated in Fig. 2.6. Upon network initialisation, the HC gains control of the wireless medium. It then broadcasts a Beacon frame to all STAs and announces the next beacon transmission time. The time between one Beacon and the next is known as Beacon Interval. Meanwhile, the HC partitions each Beacon Interval into multiple service intervals (SIs) which is calculated as follows: let m be the minimum of all SIs for all admitted streams, the SI is then determined by a number lower than m that is also a submultiple of the Beacon Interval.

The HC then divides SI into two periods, namely a contention free period (CFP) and a contention period (CP). During CFP, the HC polls STAs according to its polling list. For example, consider STA_i . It receives a CF-Poll frame sent by the QAP and then responds with data frames. Upon receiving the data frames from STA_i , the QAP sends an Acknowledgement (ACK) frame to terminate the transmission from STA_i within current SI. It is noted that the HCCA allows multiple data frame transmissions as long as the sum of data, CF-Poll, and ACK frame transmission time is within the designated Transmission Opportunity for STA_i ($TXOP_i$). Once the last frame from the final STA is received, the

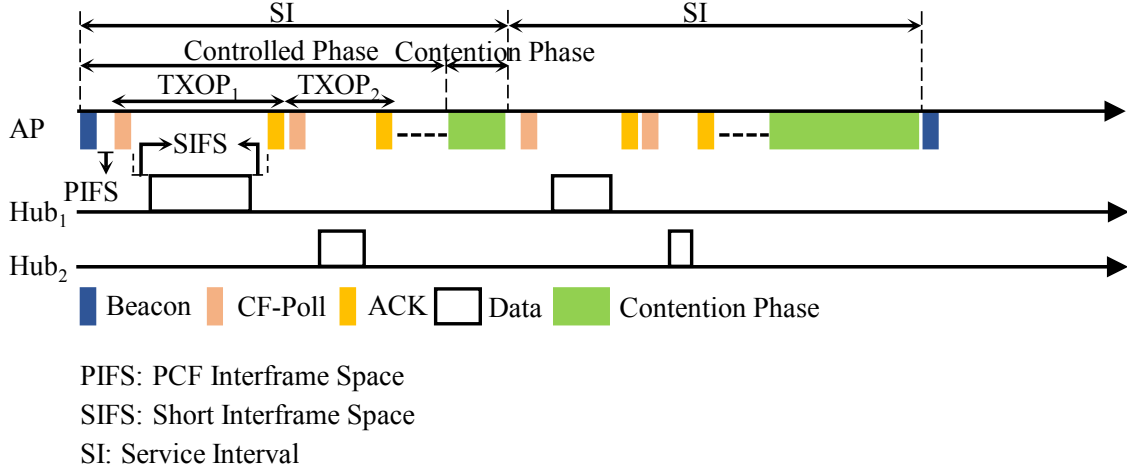


Figure 2.6: Operation timeline of IEEE 802.11 HCCA protocol.

HC ends the CFP with a CF-End frame, and transitions into CP. During CP, all wireless devices compete for channel access using a contention-based transmission mechanism. This CP phase lasts until the start of the next SI.

It should be noted that HCCA allows enhancement of polling efficiency by permitting earlier TXOP termination if a STA finished its transmission earlier. In addition, when a STA has no pending data to transmit, it responds to the AP with a Null-Poll frame. This can potentially advance the transmission of subsequent STAs and potentially allows for a longer CP. Therefore, the average latency of connected STAs in a polling cycle can be reduced, and devices utilising the CP can have more opportunity for data transmission. A simple realisation of HCCA called the reference design that takes into consideration the TXOPs and the number of admitted STAs was proposed in [56]. In this reference design, two major functions are performed by an HC, namely the scheduler function and the admission function. The scheduler function determines the duration of TXOP assigned to each STA based on:

$$\text{TXOP}_i = \max\left(\frac{N_i \times L_i}{R_i} + O, \frac{M}{R_i} + O\right). \quad (2.2)$$

where N is the expected number of packet arrivals during one SI, L is the nominal MSDU size, M is the maximum MSDU size, R is the physical transmission rate and O is the

overhead in time units. On the other hand, the admission function accepts/rejects new connection requests from STAs. For instance, if k STAs have already connected to a QAP using HCCA, a $(k + 1)^{\text{th}}$ STA can be admitted only if:

$$\frac{\text{TXOP}_{k+1}}{SI} + \sum_{i=1}^k \frac{\text{TXOP}_i}{SI} \leq \frac{T - T_{\text{CP}}}{T} \quad (2.3)$$

where T represents the Beacon Interval, and T_{CP} is the duration of CP.

It was reported that the performance of EDCA and an HCCA-assisted HCF could both meet the defined QoS requirements in the testing scenarios [74]. However, the reduced contention overhead by HCCA significantly enhances channel utilisation by about 20%, which in turn improves latency by approximately 15%. Therefore, HCCA can provide better services for higher priority traffic flows compared with EDCA.

2.1.3 IEEE 802.11 MAC Layer Solutions for Improved Latency Performance

The WLAN has long past the expectation of merely delivering data packets successfully. In order to fulfil the needs of exponentially increasing data rate by multimedia applications, collective efforts have been put into re-designing of all layers of WLAN networks. Specifically, the MAC layer has attracted much attention since it controls the media-sharing mechanism among multiple coexisting users, and it is responsible for reliable data transmission with certain QoS. As such, the functionality of the WLAN MAC layer includes providing service differentiation among multimedia applications, latency bound for real-time applications and fairness among associated users. The capability of IEEE 802.11 networks to support low-latency applications has been pursued in extensive studies. An early work of Costa *et al.* has compared the performance of different IEEE 802.11 MAC protocols on real-time network traffic [78]. Results show that DCF/EDCA suffers from high collision rate when the network is exposed to external non-real-time traffic, while HCCA preserves the real-time traffic requirements with a limited number of connected STAs. A more comprehensive study has been conducted in [79], such that configuration parameters of EDCA such as contention window can be determined for given performance criteria of throughput and latency of each STA. On the

other hand, [80] compared five HCCA algorithms in terms of suitability for soft real-time guarantees and results show a clear trade-off between network latency and the number of supported STAs.

In this section, MAC layer enhancements targeting at reducing latency are discussed, and they can be classified into three categories. In addition, their individual strengths and weaknesses with respect to the service performance will also be discussed.

QoS Enhancement by Advanced Medium Access Mechanism

Many studies paid attention to modifying the standardised uplink transmission protocols themselves rather than restricting to the already defined MAC protocol. These studies inherent the basic form of existing MAC protocols and improve their QoS performance by novel approaches, such as deploying additional handshake procedure and modifying control frame field.

Lo *et al.* have developed a novel efficient multipolling mechanism incorporating contention-based transmission for HCF [81]. Instead of polling STAs by a Poll frame for each STA, the proposed multipolling mechanism employs a CP-Multipoll frame which aggregates multiple polling information and is broadcast to all participating STAs. This greatly reduces the polling overhead of HCCA and thus offers better channel utilisation for uplink transmissions. Furthermore, this mechanism outperforms conventional HCCA with higher polling accuracy. In the conventional HCCA, an AP is required to determine the amount of TXOP allocated to a particular STA, which might mismatch the actual bandwidth demand and cause either bandwidth waste due to Null-Poll issue or extra latency due to insufficient TXOP allocation. However, the proposed multipolling mechanism enables each STA decides when to finish the uplink transmission and allows an awaiting STA to detect the channel busyness immediately. Therefore, it can eliminate the demand mismatch issue faced by conventional HCCA. Simulation results demonstrate that the proposed multipolling mechanism can serve real-time applications with VBR traffic flows such as audio and video communications. However, this mechanism does not consider the priority of traffic flows and thus cannot offer service differentiation. Another drawback of this mechanism is that it is prone to hidden STA issue, where mul-

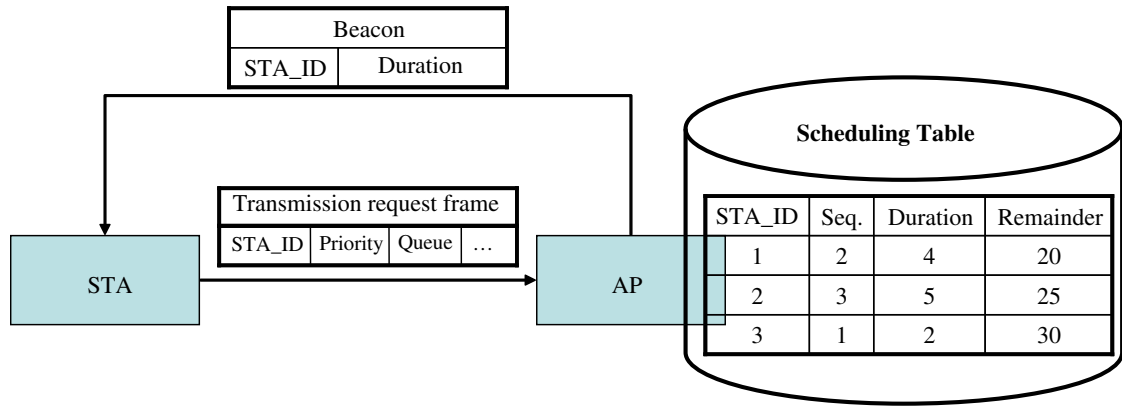


Figure 2.7: Mechanism of NPHCCA (adopted from[82]).

tiple STA could transmit data simultaneously due to the existence of undetected STAs.

A non-polling based HCCA (NPHCCA) was proposed in [82], which allows STAs to inform about their buffer occupancy by specifically sending transmission request to the associated AP. The NPHCCA divides the transmission time into three parts. The first part, called Beacon_Time indicates the time when all STAs listen to a beacon frame from an AP for synchronisation and notification of scheduled transmission time. Then, STAs could content for channel access using DCF during Req_Time and send transmission request to the AP, in which case uplink demand including priority and queue status are acknowledged and stored in a scheduling table in the AP. Finally, in the TX_Time the AP can use the stored information in the scheduling table to poll STAs, which can effectively eliminate null-polling issue of the original HCCA protocol. The overall process employed by NPHCCA is illustrated in Fig. 2.7. Simulation results show that improved latency performance can be achieved by NPHCCA when the network is lightly loaded, and this is mainly due to higher transmission efficiency by the reduced null-polling phenomenon. However, NPHCCA involves additional control frame transmission, which inevitably introduces extra signal overhead. Further, the existing beacon frame needs to be modified for new fields, which could also increase signalling overhead.

In [83], the authors focused on the impact of aggregation mechanisms of the existing IEEE 802.11 MAC protocol. Although the original MAC layer aggregation mechanism

was introduced to reduce MAC frame overhead by combining multiple frames into a single long frame, it also instigates longer headers which may impair latency performance of latency-critical applications. Targeting at alleviating the negative impact of aggregation on small-payload data frames (e.g. VoIP) due to large repeated header, a new aggregation scheme called mA-MSDU was therefore employed. The proposed aggregation mechanism results in small headers and subframe integrity check by introducing a control field into the subframe header. Simulation results reveal that the proposed mechanism can effectively improve throughput and reduce latency experienced by VoIP streams. However, the authors did not discuss the impact of proposed mechanism on backoff process and ACK frame calculation.

QoS Enhancement by Advanced Admission Control

When traffic load is heavy, networks tend to saturate, and traffic flows of latency-sensitive services are likely to be impaired. Therefore, admission control is important as it controls the amount of attempting traffic flows such that QoS of higher priority applications is not deteriorated. For an infrastructure-based WLAN network, an admission control protocol can be implemented for either distributed MAC protocols (DCF and EDCA) and coordinated MAC protocol (PCF and HCCA). Representative approaches of both types will be discussed here.

A measurement-based admission control protocol called AC-VVD was proposed for EDCA in [84]. In this work, real-time communications are protected from potential QoS degradation due to data communications, which fulfils the demand of service differentiation. Specifically, STAs carrying voice and video sessions shall listen to the available budgets from an AP about admission acceptance or rejection, this helps to guarantee the QoS performance of existing real-time sessions and fairness among them. Besides, the AP can dynamically control the global parameters associated with EDCA protocol for STAs carrying data communications according to current traffic condition. Simulation results prove that the proposed approach provides service differentiation among different ACs and good fairness within the same AC. Moreover, the QoS performance of real-time communications (voice and video) is greatly improved when AC-VVD is employed.

Chen *et al.* have proposed a model-based call admission control (CAC) and a rate control scheme to regulate attempting traffic [76]. Different to AC-VVD where admission decisions are made based on measurement, the proposed scheme is based on estimated channel condition. First, the theoretical average latency of employing EDCA under unsaturated channel is derived using G/G/1 queueing model and Markov chain. This is utilised by the CAC to evaluate the impact of admitting a new traffic flow on the current network utilisation and latency performance, and this helps to determine whether to accept or reject this new flow. Second, a channel busyness ratio is derived to represent the network condition accurately, and it is used to instruct both the CAC and rate control scheme. With the busyness ratio, latency-tolerant flows can use leftover bandwidth for best-effort transmission. Simulation results show that with the proposed admission control protocol, EDCA can achieve stringent QoS requirements with high channel utilisation. However, the theatrical framework only considers contention window differentiation, which may lead to inaccuracy in latency evaluation.

In [85], an enhanced distributed control algorithm was developed for EDCA. This algorithm requires the cooperation of both AP and STAs. At an AP, transmission budget indicating allowable transmission time per AC is encapsulated in beacon and broadcast to each STA. Then, associated STAs determine the amount of transmission limit of active AC. As such, local voice/video transmission time is restricted and therefore prevent the network from overloading. Although simulation results demonstrate that the proposed admission control protects latency performance of existing video streams, heavily loaded traffic condition was not considered. Similarly, Wu *et al.* presented another CAC function to secure resources for high priority connections [86]. Specifically, a new AC is set to active only when (a) the average bandwidth for existing higher or equal priority AC call could be guaranteed and (b) the minimum bandwidth requirement of existing lower priority AC call could be met.

Indeed, the idea of admission control intrinsically requires a certain level of central coordination and information sharing, which is contrary to the distributed format of EDCA. For example, STAs need to listen to the available budget before any transmission attempts in [84], and a beacon frame must be in place for the enhanced distributed control algo-

rithm to work [85], which introduce extra delay component or complexity compared with the conventional EDCA. On the other hand, the admission control naturally fits HCCA as central coordination is already provided.

PHCCA was proposed for simple priority-based admission control for HCCA in order to provide QoS provisioning [87]. First, connected traffic flows are divided into three classes, which correspond to the importance of all associated users. For the admission control algorithm, a higher priority class traffic finding the channel full is allowed to borrow bandwidth of lower priority classes. In particular, the bandwidth allocated for each class is further divided into a borrowable (B_B) and a non-borrowable portion (B_{NB}), where the B_{NB} is designed for starvation protection. Any new incoming traffic flow can request to use available B_B if the bandwidth of its class is already full, creating an admission control mechanisms catering for high priority traffic flows. Simulation results reveal that the number of accepted requests from the highest priority class increased by 70% compared with the original HCCA. On the other hand, the performance of the middle-class traffic was similar to regular HCCA.

Mansoor *et al.* proposed a feedback-based admission control Unit (FACU) targeting at fully utilising the surplus bandwidth within current SI. Specifically, surplus bandwidth exists when one or more STAs do not exhaust their allocated TXOPs. In such a case, the surplus channel bandwidth will be reallocated to allow extra STAs to transmit without impairing the performance of existing STAs. Analytical results prove the effectiveness of FACU, such that considerable amount of unutilised bandwidth can be beneficial for admitting extra QoS flows, which in turn enhances the channel utilisation and provides better QoS provisioning compared to conventional HCCA.

QoS Enhancement by Advanced Bandwidth Allocation Mechanism

Numerous studies have focused on dynamic and accurate bandwidth (TXOP) allocation for HCCA since it could significantly impact the latency performance of IEEE 802.11. In the reference design defined in [56], TXOP is determined based on the average bandwidth demand, which can be calculated from the information contained in the TSPEC frame. This mechanism works well for CBR data traffic but is not effective for VBR traffic

[71, 88, 89]. The issues associated with the latency performance of HCCA include null-polling and under-scheduling. As described in [56], all STAs registered in the polling list of an AP will be polled during an SI regardless of STAs' buffer occupancy. This could result in bandwidth waste when a STA has no pending data to send. On the other hand, if an allocated TXOP is not sufficient to transmit data stored in the STA buffer, the remaining data could only be transmitted in the next TXOP, which leads to an extra delay of SI. Advanced scheduling mechanisms proposed for HCCA include feedback-based, estimation-based, optimisation-based and cross-layering approaches, and they will be discussed in detail in Section 3.2. Here we introduce two example HCCA schedulers which employ urgency estimation and cross-layer concept, respectively.

Rashid *et al.* proposed a PRO-HCCA scheduler to alleviate the deficiencies of reference HCCA [90]. The novelty of PRO-HCCA is a derived urgency index that is used to differentiate services with individual latency requirements. A prediction mechanism was also proposed so that the urgency index could be estimated for uplink transmissions. The PRO-HCCA collocated at the AP can then conduct optimisation based on the collected urgency indices to determine the allocated TXOP for each STA. The PRO-HCCA scheduler was tested on interactive video, non-interactive video and VoIP streams and simulation results reveal that PRO-HCCA scheduler enhances the robustness and stability of HCCA in terms of providing lower latency than the reference HCCA.

Adaptive feedback-based HCCA (AF-HCCA) was introduced in [91]. Similar to [42], which leverages the accurate arrival time of the next video frame via cross-layering, AF-HCCA exploits the availability of information of pre-recorded video sources for accurate scheduling. In essence, an AP can determine the amount of TXOP that accurately matches the actual demand, in which case it eliminates the over-polling issue and enhances latency performance and channel utilisation. Simulation results demonstrated that AF-HCCA could achieve the least latency as compared with existing HCCA protocols such as reference HCCA design [56], E-EDD [39] and Feasible HCCA [42]. Despite AF-HCCA's success in scheduling within HCCA, advanced EDCA algorithms were also suggested as a solution to efficiently utilising the amount of surplus channel bandwidth caused by a tight match between TXOP and bandwidth demand [89].

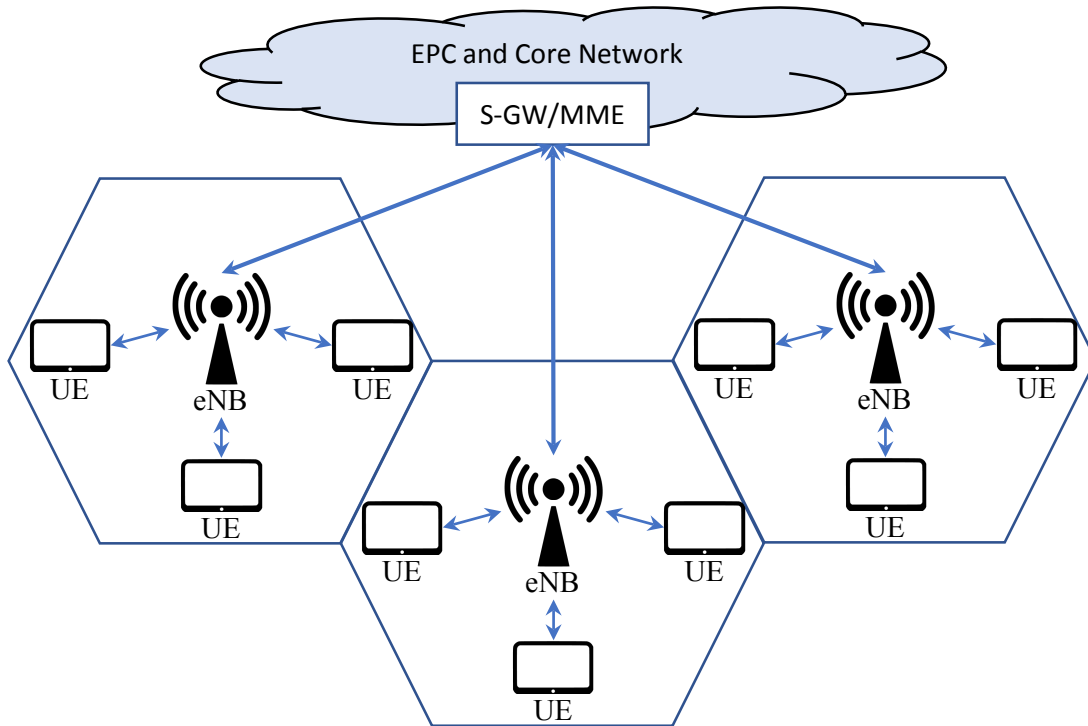


Figure 2.8: LTE radio-access network architecture.

2.2 Mobile Cellular Networks Supported by LTE

The rapid growth of Internet usage and increasing popularity of mobile devices have thrust the development of mobile cellular networks. Different from IEEE 802.11 networks as described in the preceding section, mobile cellular networks have a much wider coverage of several kilometres, where users could be moving at a moderate speed. In particular, LTE systems have been designed to support versatile applications on mobile devices such as video streaming, teleconferencing and web browsing. Today, LTE systems are deployed widely and gained significant success in the market [9].

Being developed by 3GPP as the Fourth Generation (4G) of mobile cellular technologies, LTE outperforms earlier generations and suits the demand of high data rates, low latency and high capacity. The first LTE standards were introduced in 3GPP release 8 [92], and it was designed as a full-Internet Protocol (IP) technology for packet-data support. Unlike 3G systems where voice was transmitted in a circuit-switch approach, voice in LTE is delivered through a VoIP technology [10,60]. Later in 3GPP release 10 [93], LTE

has evolved in terms of spectral efficiency and flexibility, capacity and data rates and the term LTE-Advanced (LTE-A) was assigned to this improved LTE system. It should be noted that as our study on the MAC layer protocols is applicable for both LTE and LTE-A systems, the term **LTE** is used to refer both LTE systems. Today we observe a wave of development of the Fifth Generation (5G) mobile cellular technologies but LTE infrastructure and technology will remain as an important complementary part of the future 5G world for a long time, as described in the 5G Non-Standalone (NSA) deployment by 3GPP NR [94]. A typical LTE network architecture is depicted in Fig. 2.8, where UEs are connected to an eNB within a mobile cell, and the eNBs are further communicating with the core network through serving gateway (S-GW) and mobility management entity (MME), by which the user-plane node and control-plane node of EPC are configured.

In the following sections, we will briefly introduce the technological features provided by LTE standards. Then, we will discuss and elaborate operation principles of LTE uplink transmission schemes and their effects on the latency performance of LTE systems in detail. Existing literature proposed for performance improvement will also be discussed.

2.2.1 Technological Advancement of LTE

Comparing with earlier generations of cellular technologies, LTE offers a rich set of characteristics. First, Orthogonal Frequency Division Multiplexing (OFDM) is used as the basic transmission scheme in PHY layer of LTE. OFDM scheme enables LTE to transmit radio signals with a high degree of robustness against multipath interference and allows the deployment of frequency-dependent scheduling. It should be noted that although conventional OFDM is used in downlink transmissions, uplink packets are transmitted using Single-Carrier Frequency Division Multiplexing (SC-FDMA) in order to gain orthogonality among uplink users with low peak-to-average ratio, which is essential for power-constraint user devices. It is worth to mention that as SC-FDMA can be interpreted as an OFDM modulator with an additional Discrete Fourier Transform (DFT) precoder, SC-FDMA is also referred to as DFT-spread OFDM (DFTS-OFDM). Second, LTE standards allow an eNodeB or UE to be configured in time division duplexing (TDD) and

frequency division duplexing (FDD) mode, which significantly enhances system flexibility in choosing uplink-to-downlink radio resource ratio.

Third, Adaptive Modulation and Coding (AMC) is supported by LTE. It is known that Bit Error Rate (BER) of wireless transmission is susceptible to wireless channel conditions. Although a coding scheme with a higher modulation order can transmit data with a higher rate, the resulting BER might be worse compared with transmitting using a coding scheme with a lower modulation order. To achieve a balance between data rate and BER for LTE systems, AMC is therefore deployed to actively select a proper modulation and coding scheme according to Channel Quality Indicator (CQI), which is determined based on sensed channel quality. AMC is essential for LTE systems to reach the maximum throughput in a time-varying channel.

Another equally important feature of LTE is time-frequency-division resource scheduling for both uplink and downlink transmissions. In a multiple access environment, UEs in LTE systems can transmit data in the same time slot, given that the radio resources in the frequency domain are different. This feature is important to achieve low latency performance compared with earlier mobile technologies. In an LTE network, a data packet in a UE's buffer can be transmitted once the packets ahead are cleared, without an extra latency of waiting for transmission completion of other UEs.

LTE-A, embedded with several enhancements compared with LTE, were introduced in release 10. One of the most important upgrades is carrier aggregation. With carrier aggregation, multiple LTE carriers can be used to transmit/receive data to/from the same device, in which case an overall wider bandwidth is used, and hence higher link data rate is achieved. In the most extreme case in release 10, five component carriers with 20 MHz individual bandwidths can be aggregated to form a 100 MHz bandwidth link channel. The maximum capacity of LTE-A is further increased to 640 MHz in release 13 [95].

Another improvement introduced by LTE-A compared with LTE is the multi-antenna transmission technique—Multiple Input Multiple Output technology (MIMO). With MIMO, data are transmitted with multiple antenna pair, which significantly enhances the quality of received signals and reduces co-channel interference [60]. As a result, enhanced spectral efficiency can be realised by MIMO.

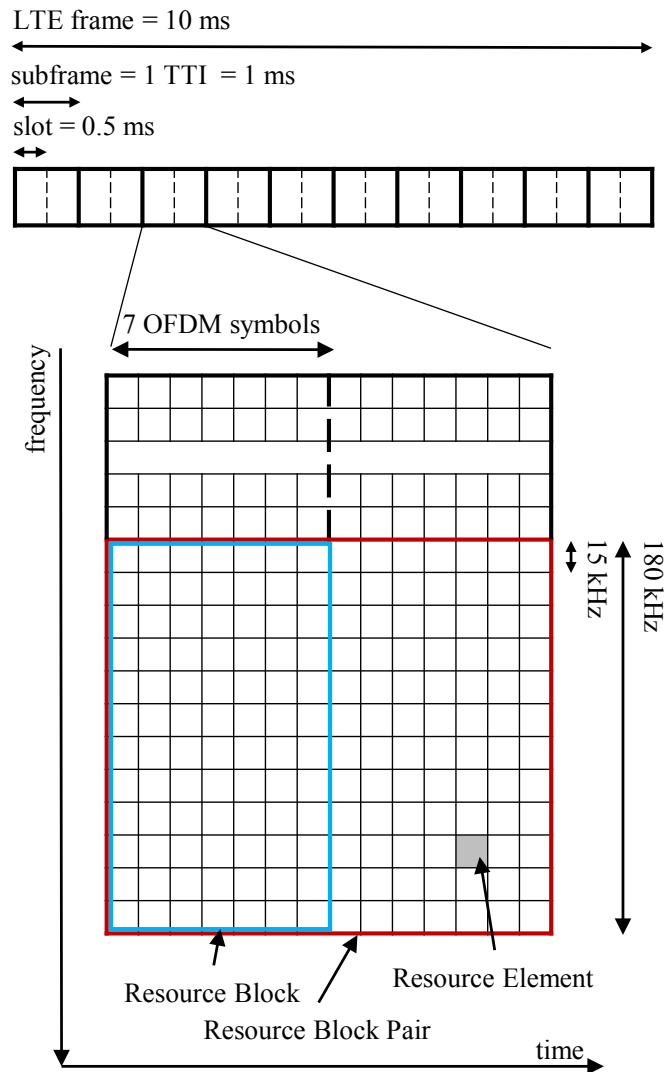


Figure 2.9: LTE radio resource structure.

2.2.2 LTE Time-Frequency Division Transmissions

As mentioned in the preceding section, LTE systems deploying OFDM signalling are able to divide radio resources according to time and frequency. As specified in LTE standards, the shortest radio transmission time—Transmission Time Interval (TTI) is defined as 1 ms. An LTE frame is defined as 10 ms long and consists of 10 subframes, each corresponding to 1 TTI. A subframe can be further divided into 2 slots, where each slot transmits 6 or 7 OFDM symbols depending on the type of chosen OFDM cyclic prefix. Furthermore, radio resources in an LTE network are divided into many time-frequency units called

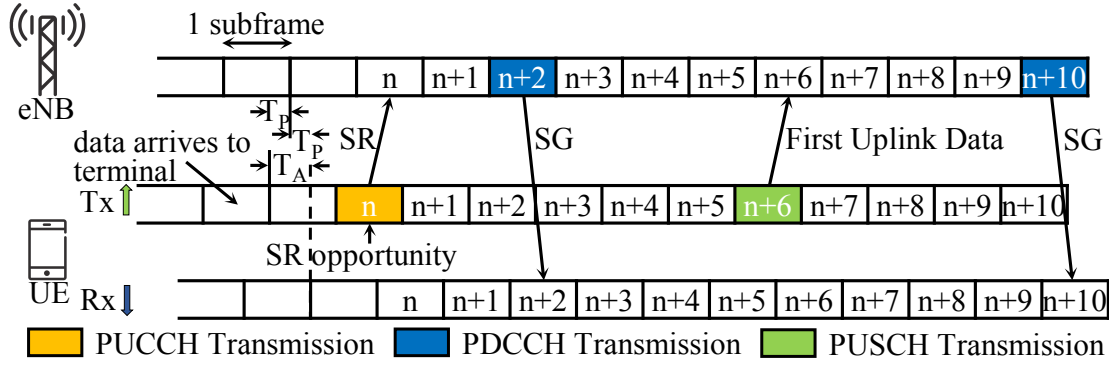


Figure 2.10: Timing diagram of Dynamic Scheme (DS) in LTE.

Resource Elements (RE), and they correspond to one OFDM symbol duration times 15 kHz. The smallest resource unit that can be scheduled to a connected device is called a Resource Block Pair (RBP), and it is formed by an aggregation of adjacent REs, which occupies a total time-frequency unit of 1 ms times 180 kHz. An overview of the LTE radio resource grid where each slot contains 7 OFDM symbols is illustrated in Fig. 2.9. In addition, the allocated RBPs must be contiguous for uplink transmission, because of SC-FDMA operation.

In LTE, the MAC layer defines both transmission scheme and transmission scheduler which are deployed for uplink transmissions. They guide scheduling decisions which consist of both position and size of allocated radio resources and this has a significant influence on the processing and queueing latency components among overall latency. Details of different uplink transmission schemes and their corresponding schedulers are discussed in the following sections.

2.2.2.1 Dynamic Scheme for Uplink Transmissions in LTE

A conventional LTE network uses the DS scheme for uplink transmission. An overview of this operation diagram is illustrated in Fig. 2.10. Once a UE is connected and synchronised to an eNB, the UE's uplink and downlink transmissions are then controlled by a mechanism called transmit timing advance [95]. The transmit timing advance T_A sets a negative offset between the start of a received downlink subframe and a transmitted uplink subframe, such that time alignment between the eNB and the UE can be achieved. In

practice, T_A is determined by $2T_P$, where T_P denotes the propagation time latency from the UE to the eNB. Once the UE detects the arrival of new uplink data, it transmits a Scheduling Request (SR) to the eNB on Physical Uplink Control Channel (PUCCH) at the next SR opportunity—subframe n . Upon successful reception of the SR, the eNB transmits a downlink control packet on PDCCH called Scheduling Grant (SG) at subframe $n + 2$, informing the radio resource scheduling decision for the UE. The scheduling decision includes both the number and position (frequency range) of assigned RBPs. The first uplink data frame is then transmitted by the UE on Physical Uplink Shared Channel (PUSCH), at 4 subframes after the UE received the SG. An important feature of LTE uplink data frame is that it contains not only the uplink data but also a Buffer Status Report (BSR), indicating the size of leftover data in UE's buffer. By obtaining the buffer occupancy information, the eNB then dynamically schedules further RBPs and sends the second SG to the UE at subframe $n + 10$. The subsequent uplink data packets are therefore transmitted every 8 subframes until the UE's data buffer is cleared.

As radio resources of an LTE cell are finite, one of the most significant roles of the LTE MAC layer protocol is to allocate resources to connected UEs in order to meet different Quality of Service (QoS) requirements. Three conventional scheduling algorithms are commonly deployed in today's LTE networks—round robin, max CQI and proportional fair (PF). Round robin scheduler selects and schedules UEs based on a round robin order, hence each connected UE will have almost equal resource allocation for uplink transmissions. The drawback of this approach is that a certain amount of radio resources might be allocated to UEs experiencing poor channel condition, which deteriorates the overall achievable throughput within an LTE cell. On the other hand, max CQI scheduler focuses on maximising cell throughput and prioritises UEs according to channel condition. As a result, UEs who have better higher CQI values will be assigned with more radio resources. The disadvantage of max CQI scheduler is that UEs with the worst channel condition might never have access to radio resources, which leads to degraded user experience. Finally, PF scheduler [96] ranks UEs according to long-term transmission rate, which reflects both the channel condition and amount of service starvation. Therefore, PF is expected to achieve a balance between channel quality and resource allocation fairness.

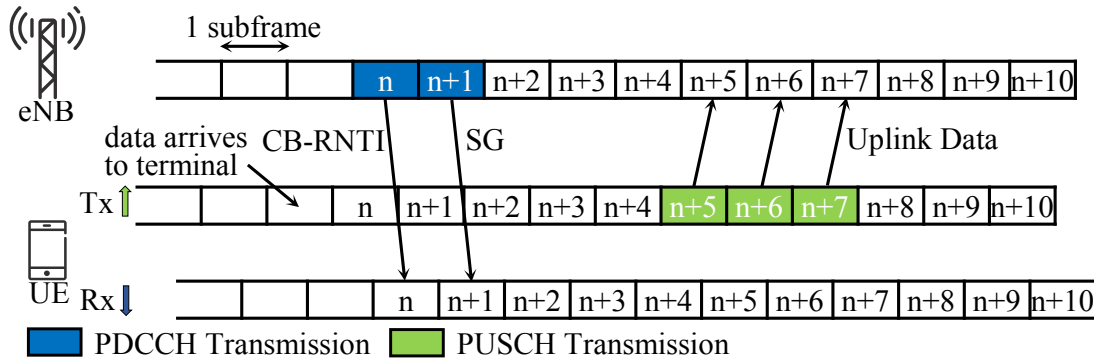


Figure 2.11: Timing diagram of Contention Based Scheme (CBS) in LTE.

2.2.2.2 Contention-Based Scheme for Uplink Transmissions in LTE

Contention Based Scheme (CBS) [47, 48] was proposed to address the latency caused by the DS scheme. In general, CBS achieves statistical multiplexing by allowing multiple UEs to contend for the same PUSCH resources for uplink transmission. First, a group of UEs are assigned with CB-Radio Network Temporary Identifiers (CB-RNTIs), and these UEs listen to PDCCH for available CB grants. After receiving the CB grants, the UEs are allowed to send their uplink data directly using the shared PUSCH resources without waiting for an SR and an SG [97]. The procedure of CBS uplink transmission is illustrated in Fig. 2.11. Without the need of exchanging SR and SG between UEs and eNB, CBS is able to conduct uplink transmissions with less latency. However, CBS suffers from data packet collision when multiple UEs are transmitting data using the same RBPs at the same subframe, in which case uplink packets cannot be correctly received and decoded. Andreev *et al.* implemented CBS for Machine-Type Communications (MTC), which is characterised by small data transmissions [98]. Their solution to reducing data collision is a random backoff procedure followed by a packet retransmission mechanism. In [99], CBS is tested and evaluated with the implementation of consecutive Hybrid Automatic Repeat reQuest (HARQ) retransmission. The proposed HARQ retransmission approach is used to increase the reliability of CBS, and it is shown that CBS is more efficient in terms of resource scheduling than traditional SPS scheme and the proposed CBS is able to meet reliability requirement, given data collision is properly handled.

2.2.2.3 Semi-Persistent Scheduling Scheme for Uplink Transmissions in LTE

The conventional DS scheme imposes a significant amount of control-signal overhead when transmitting small-sized packets. Additionally, instead of dynamic resource allocation, the regularly occurring pattern of constant bitrate (CBR) traffic motivates an uplink scheme to schedule resources periodically and deterministically. As such, the Semi-Persistent Scheduling (SPS) scheme was proposed to address these two issues, and it is most notably implemented for VoIP [100]. In essence, an eNB periodically assigns predefined RBPs to an SPS-configured UE without consuming resource on PDCCH or PUCCH, except for the initial connection setup stage. The principle of operation of the SPS scheme with 10 ms SPS periodicity is illustrated in Fig. 2.12. The SPS scheme follows a similar process as in the DS scheme until subframe $n + 6$. In particular, all SPS-related parameters including SPS periodicity, scheduling decision, selected Modulation and Coding Scheme (MCS) and a maximum number of SPS transmission are encapsulated in an SPS-SG. At subframes $n + 16$ and $n + 26$, a UE initiates uplink transmission using the granted periodic RBPs, and this continues until the maximum number of SPS transmission is reached or other termination mechanisms are triggered. An apparent disadvantage of implementing the ordinary SPS scheme for the Tactile Internet is the incapability of adjusting scheduled RBPs according to the network traffic fluctuation. This leads to a trade-off between low/high latency and high/low resource waste in an over/under-scheduling problem. In particular, SPS might fail to transmit all the buffered data immediately when data traffic is bursty [101] or induce noticeable resource waste when data traffic reduces.

Following the same procedure in [102], the minimum initial access and transmission latencies of DS, CBS and SPS schemes can be derived, and they are shown in Table 2.1 and Table 2.2, respectively. It should be noted that the transmission latency is valid when sufficient radio resources are accessible. It can be seen that the DS scheme has a maximum initial access latency of 7.5 ms and average transmission latency of 5 ms, which also forms the lower bound of achievable latency by the DS. Although the contention-free feature of DS can provide latency and jitter with certain upper bound, the time wasted in sending SR and receiving SG may not be tolerable for ultra-low latency services. Such latency performance is clearly unacceptable for latency-critical applications, such as tel-

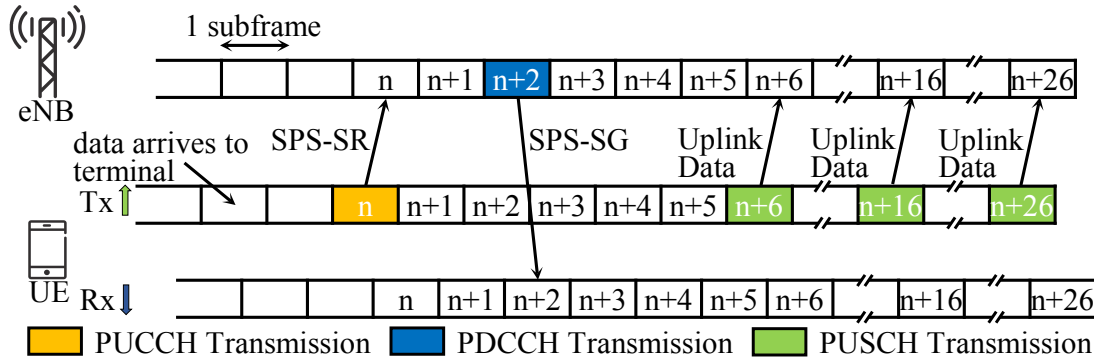


Figure 2.12: Timing diagram of Semi-Persistent Scheduling (SPS) in LTE.

erehabilitation for health care and smart transportation for road safety.

On the other hand, we can see that without exchanging SR and SG between UEs and eNB, CBS offers 3.5 ms less transmission latency as compared with the DS. When CBS is implemented, however, packet collision is likely to happen, which necessitates the implementation of collision reduction mechanisms and efficient retransmission techniques [99, 103].

It is observed that because the SPS scheme and DS share a similar process of sending SR and receiving SG, the latency in the initial access is the same for these two uplink transmission schemes. More importantly, the SPS scheme offers a minimum transmission latency as CBS does, given UEs are scheduled with sufficient RBPs. This motivates our interest in designing an SPS scheme that allows flexible scheduling and ensures a contention-free transmission with low-latency performance.

2.2.3 LTE Scheduling Algorithms for Ultra-Low Latency Performance

As explained in Section 1.3, the communications community strives for designing LTE such that emerging new services can be supported with one of the most stringent requirements—ultra-low latency. For instance, it is reported that current LTE technology is not able to support mobile VR with high quality [29]. In particular, some VR frames might experience excessive latency and therefore leads to poor quality of experience. Therefore, it is critical to analyse the composition of overall latency in LTE networks and design feasible solutions for effective latency reduction. The study in [29] shows that instead of

Table 2.1: Preliminary analysis of initial access latency

Component	DS	CBS	SPS
Mean waiting time (ms) for PUCCH (1 ms SR period)	0.5	--	0.5
Mean waiting time (ms) for CB-RNTI signalling	--	0.5	--
SR transmission (ms)	1	--	1
CB-RNTI transmission (ms)	--	1	--
eNB decodes SR and generates an SG (ms)	1	--	1
SG transmission (ms)	1	1	1
UE Processing Delay (ms)	3	3	3
UL data transmission (ms)	1	1	1
Total latency (ms)	7.5	6.5	7.5

being limited by LTE bandwidth, latency incurred by the signalling process of LTE networks forms a bulk portion of network latency. In this thesis, our research is focused on reducing the T_{proc} and T_{queue} , which stem from MAC layer protocols in LTE networks.

Designs of latency-constrained schedulers are discussed in detail in this section because it is important to provide acceptable latency performance for latency-critical communications services in terms of determining a proper number of radio resources for uplink traffic. It should be noted that all scheduling algorithms discussed here are designed for the DS protocol, and therefore they only deal with queueing latency. Techniques proposed for CBS and SPS schemes will be discussed in Section 2.3 and 4.2, respectively.

Existing literature has devoted largely into designing feasible resource scheduler algorithms that can enhance performance metrics of LTE networks, such as latency, throughput, power efficiency and fairness among users. For instance, one of the most widely-deployed uplink schedulers is PF, as described in Section 2.2.2.1, where throughput and fairness among connected devices within a cell are addressed. Numerous studies have

Table 2.2: Preliminary analysis of transmission latency

Component	DS	CBS	SPS
Mean waiting time (ms) for transmission opportunity	4	0.5	0.5
PUSCH transmission (ms)	1	1	1
Total latency (ms)	5	1.5	1.5

been conducted for designing schedulers targeting at one or more QoS metrics, and they can be classified into two categories—prioritisation-based schedulers and optimisation-based schedulers.

Prioritisation-based Scheduling for LTE

Scheduling algorithms based on prioritisation are implemented by sorting connected devices according to one or more criteria and then allocating available RBPs according to the sorted order. The considered criteria usually reflect QoS requirements, but they do not provide a guarantee about meeting those requirements. Therefore, these algorithms are regarded as heuristic or best-effort.

The first uplink scheduler algorithm that considers both channel conditions and latency requirement is called Channel Dependent Earliest Deadline Due (CD-EDD) [104]. In here, a novel performance metric of each device is calculated as $metric_i = a_i \frac{\mu_i(t)}{\bar{\mu}_i} \frac{W_i(t)}{d_i(t)}$, where a_i , $\mu_i(t)$, $W_i(t)$ and $d_i(t)$ represent the weighting parameter reflecting the statistical QoS requirements, current transmission rate, delay experienced by the head-of-line packet and remaining time to expiry of the head-of-line packet of the i^{th} user at time t , respectively. It is clear that the second term $\frac{\mu_i(t)}{\bar{\mu}_i}$ reflects the average transmission rate experienced while the third term $\frac{W_i(t)}{d_i(t)}$ reflects how urgent a user is to transmit its buffered data. The proposed scheduler then determines the number of allocated RBPs to devices according to $metric_i$. It is reported that gains in throughput and latency can be achieved by this scheduler. In addition, a set of scheduling policies were also proposed to achieve fairness of latency, throughput and packet loss rate among connected devices. However,

how shall a scheduler in an eNB access the information of remaining time to expire was not provided in [104].

An Adaptive Transmission Bandwidth (ATB) scheduling algorithm is introduced in [105]. It firstly needs to compute the utility function which represents the performance metric:

$$M_{\text{sum}} = \sum M_{i,j}, i \in \Omega_i, j \in \Omega_j,$$

where $M_{i,j}$, Ω_i and Ω_j are the metric for the i^{th} UE and j^{th} RBP, set of remaining UEs and set of remaining RBPs, respectively. In particular, $A_{i,j} = \{0, 1\}$, where 1 indicates that j^{th} RBP is scheduled to i^{th} UE. It is noted that this is just a generic expression such that metrics like CQI, SNR, latency or fairness can be a candidate of performance metric. During operation, the proposed scheduler needs to go through a time-domain scheduling process to determine which UEs are available at time slot t . The algorithm then performs a frequency-domain scheduling process such that it allocates resources to UEs in the order of descending $M_{i,j}$. In particular, for the i^{th} UE, the algorithm firstly allocates the j^{th} RBP where $j' = \text{argmax}_j M_{i,j}$ and then expands RBP allocation as long as $M_{i',j'}$ has the highest priority. The algorithm continues to schedule resources for other UEs in the same procedure. For performance verification, the proposed algorithm adopted the PF metric, while other aspects of uplink scheduling performance were not addressed. Simulation results clearly show a 20% improvement on average cell throughput is achieved. Moreover, the scheduling procedure of the proposed algorithm provides low computational complexity and QoS provisioning flexibility. However, the proposed algorithm does not consider varying loading conditions of different users and individual hard QoS constraints.

Authors in [106] proposed two uplink scheduling algorithms for latency-critical MTC. The first algorithm schedules traditional LTE users first, followed by sorting MTC devices with respect to $\gamma_{k,l}$, which represents the detected SNR over the l^{th} RBP from the k^{th} MTC user. Secondly, the first scheduler allocates the l^{th} RBP to an MTC user with the best $\gamma_{k,l}$ if latency tolerance of this user is smaller than a predefined threshold. Remaining RBPs which are adjacent to the l^{th} RBP are also allocated to the k^{th} MTC user if this user has the best SNR. The second step is repeated for the remaining RBPs. Finally, unscheduled RBPs are scheduled according to the second step without considering latency constraints.

Simulation results show that this scheduling algorithm outperforms the conventional channel-aware scheduler and latency-based class scheduler in terms of latency budget violation ratio. The second algorithm proposed in [106] follows the same procedure as the previous one, except that MTC devices are sorted solely by their latency tolerance, and the most urgent device is thus scheduled which has the best $\gamma_{k,l}$. Via simulation, the second scheduling algorithm exhibits the best latency performance, whereas it suffers from worse BER among considered algorithms. It is noted that the first algorithm puts more weight on channel conditions while the second on latency requirements, which explains their performance difference in terms of BER and latency. An obvious drawback of these proposed algorithms is that they do not consider fairness among users, which means if there existed an MTC device that always required the least latency tolerance, other MTC devices could easily be starved. A better design would be to sort devices according to the remaining latency budget instead of latency tolerance. Furthermore, this study reveals the commonly known trade-off between throughput and latency in wireless communications [107].

A class-based, hybrid scheduler was developed in [108]. First, the authors proposed a new set of QoS Class Identifier (QCI) for Machine-to-Machine (M2M) devices, which corresponds to latency-sensitive, real-time and latency-tolerant M2M. Then, the scheduler divides available resources into two disjoint sets, dedicated for H2H (Human-to-Human) and M2M users, respectively. The division ratio depends on the network operator, which could either be determined by dynamic network traffic conditions or be configured statistically. Upon scheduling, this algorithm allocates available resources to M2M users in the order of descending QCI, which equivalently forms a priority-based scheduler. Although segmentation of RBPs introduces flexibility in uplink resource allocation, the proposed scheduler does not consider any actual service constraints, leaving scheduling fairness, latency, allocation efficiency unguaranteed. Moreover, no fairness mechanism is deployed in this algorithm, leading to starvation issue for low priority M2M devices similar to [106]. On the other hand, this algorithm provides scheduling flexibility as separate scheduling policies can be applied for H2H and M2M devices.

Afrin *et al.* have designed an uplink scheduler for M2M traffic catering to latency bud-

get and buffer occupancy [109]. Firstly, the proposed algorithms require slight alterations of existing standards—1 byte in addition to standard BSR field, such that the information of packet age (PA) can be stored and transmitted to an eNB. The proposed algorithm is still a priority-based algorithm scheduler, by which it sorts awaiting devices according to an urgency metric determined as:

$$U_i = \begin{cases} \frac{B_i}{\max\{B\}} \times \frac{T_{SF}}{D_i - t}, & D_i - t > 1 \\ 1, & \text{otherwise,} \end{cases}$$

where B_i , T_{SF} , $D_i - t$ represent BSR index of the i^{th} user, LTE subframe duration (1 ms) and remaining latency budget of the i^{th} user, respectively. It is clear that such a proposed urgency metric considers both latency budget and buffer occupancy of a connected device and hence data packet from latency-critical traffic can be scheduled with high priorities. After the urgency metric is computed for every MTC device, the algorithm then schedules RBPs to devices with the order of descending U_i . Simulation results show that this scheduling algorithm outperforms existing reference configurations in terms of probability of meeting latency budget, in the expense of higher average resource usage. The proposal does not discuss how the contention between MTC devices and traditional LTE user can be resolved.

A relatively complex scheduler which carefully treats H2H and M2M separately according to traffic congestion levels was introduced in [110]. In the time-domain scheduling stage similar to [105], the proposed scheduler firstly determines which devices shall be scheduled within the current TTI. Meanwhile, this scheduler prioritises H2H devices over M2M devices. Then, a metric representing the current resource demand is computed for all selected H2H devices:

$$\widehat{RB}(u, t) = \min(r_H^{\min}, \frac{BS(u, t)RB^{avg}(u, t-1)}{BS^{avg}(u, t-1)}),$$

where $\widehat{RB}(u, t)$, BS , RB^{avg} and $r_H^{\min}$ denote the computed demand for device u at TTI t , the reported buffer size, the average number of allocated RBPs and minimum threshold for each H2H device, respectively. Then, the maximum number of allowable resource

allocation for M2M devices is determined by:

$$U_M^{\max}(t) = r_M^{\min} \times \max \left(\left\lfloor \frac{\rho \mathcal{R}}{r_M^{\min}} \right\rfloor, \left\lfloor \frac{\mathcal{R} - \widehat{RD}_H(t)}{r_M^{\min}} \right\rfloor \right),$$

where $r_M^{\min}$, ρ ($0 \leq \rho \leq 1$), $\mathcal{R}$ and $\widehat{RD}_H$ represent a predefined small amount of resources for each M2M device, the minimum ratio of available resources for M2M transmissions, sum of available RBPs and sum of H2H resource demand, respectively. It is important to note that authors of [110] further proposed a dynamic monitoring mechanism such that the actual value of ρ is adjusted according to the extent of traffic congestion. So far, the available uplink radio resources are distributed among H2H and M2M groups based on a compound factor of H2H demand and the minimum ensured M2M resource allocation. Finally, the metrics for M2M devices is calculated based on the remaining latency budget and the exponential moving average of the throughput. In sum, this proposed scheduling algorithm addresses latency and fairness among M2M users. Furthermore, this algorithm helps prevent M2M devices from starvation by admission control over traditional H2H traffic.

Salman *et al.* tackled the QoS provisioning problem of LTE networks by a prioritised M/G/1 queueing protocol [111]. A queueing model is constructed at the eNB side, where each uplink SR can be regarded as buffered data with individual traffic identification and latency requirement. In particular, with the proposed H2H relaxation scheme, SR can be classified into H2H video (type I, latency-sensitive), H2H data and audio (type II, latency-insensitive) and M2M traffic, which can then be queued into three separate buffers. The server of the queueing model then starts processing these buffers, where type I H2H requests have the highest priority while M2M requests have the lowest. In order to reduce the latency of M2M devices, however, the type II H2H requests are delayed by a predefined number R , which means these type of requests will be temporarily assigned the lowest priority. This queueing policy does not only lead to lower latency for M2M traffic but also bring a minimum negative impact on traditional H2H traffic. Simulation results reveal that average waiting time of M2M data packets decreased with larger R , while the waiting time of type I H2H data packets were kept at a minimum level for considered

packet arrival rates. It should be noted that the simulation only considered packet arrival rate up to 1 packet/slot, which may not be sufficient to model the real-world scenario. One disadvantage of the proposed schemes is that they do not provide absolute guarantees to the latency of M2M traffic, especially when an LTE network experiences high H2H traffic loads.

Optimisation-based Scheduling for LTE

Contrary to the prioritisation-based scheduling algorithms discussed in the preceding section, optimisation-based scheduling algorithms firstly formulate a utility function targeting at different QoS metrics and then form optimisation problems with system-level constraints. Solving these problems is usually difficult because the number of considered parameters and constraint is huge, and the scheduler needs to make decisions within a hard time limit (every TTI, i.e. 0.1 ms). In addition, as mentioned in Section 2.2.2, the allocation of scheduled RBPs must be contiguous in the frequency domain, which makes the optimisation problems of LTE uplink scheduling NP-Hard [112].

Delgado *et al.* have proposed the first optimisation-based scheduling algorithm [113] for LTE uplink transmissions that considers latency requirement. It was also the first algorithm that complies to contiguous RBP scheduling. In addition to assuring average latency under maximum allowable end-to-end delay, the objective equation in the optimisation problem was formulated concerning fairness among uplink users. After acknowledging the high complexity of solving the optimisation problem, two greedy heuristic algorithms were proposed to enable practical implementation of the proposed scheduler. The first algorithm, called LC-Delay assigns an RBP to a user only if both latency and throughput requirements are satisfied for all users. Otherwise it assigns this RBP to a user such that either one of the requirement is met. On the other hand, the second algorithm called PF-Delay gives priority to users with high proportional throughput while meeting latency requirement. Simulation results reveal that the LC-Delay outperforms conventional low complexity algorithm in terms of fairness, while both proposed algorithms achieved the least maximum latency compared with existing schedulers.

She *et al.* formulated a joint optimisation problem by minimising the total scheduled

bandwidth while constrained with latency requirements of uplink and downlink transmissions [45]. In the proposed model, DS was deployed for uplink transmissions while a statistical multiplex queueing model was utilised for downlink transmissions. It was proved that less service rate is required when the statistical multiplex queueing model was adopted, compared with the sum of individual required service rates. The proposed optimisation problem was solved by a two-step method, where the bandwidth assignment was firstly optimised with respect to latency and packet loss criteria and then uplink/downlink bandwidth assignment is optimised given the end-to-end latency requirements. The proposed scheduling algorithm was verified with simulations, and it could effectively save almost 50 % of the total allocated bandwidth required by an unoptimised scheduler. However, the proposed algorithm prohibited Channel State Information (CSI) for overhead reduction, such that only approximated error probability can be calculated and this could result in inaccuracy in formulating the optimisation problem.

In [43], the authors considered the problem as minimising the difference between total resource allocation of uplink and downlink transmissions, forming a symmetric joint resource allocation task for uplink/downlink design. The reason behind was that in a typical tactile interaction session, user-generated uplink command and actuator-responded downlink feedback are likely to generate symmetric traffic flow in the uplink and downlink directions. In addition, uplink and downlink latency requirements and power allocation constraints are considered in the optimisation formulation. To solve this challenging joint optimisation problem, an optimal power control policy was firstly derived, and it relates allocated RBPs, QoS exponent due to experienced latency and allocated power. Then, a binary integer programming (BIP) formulation was carried out for RBP allocation. Finally, the authors proposed a low-complexity greedy algorithm and a canonical dual solution algorithm to solve the BIP. In the simulation stage, a 5G LTE-A network architecture was deployed, where PHY layer OFDM numerology was modified such that the TTI was $100 \sim 200 \mu s$ and the maximum considered haptic data rate was 2.0 Mbps. Simulation results demonstrate that the proposed algorithm outperforms existing non-optimisation scheduling algorithms in terms of uplink/downlink rate difference. Further, the proposed algorithm permits more successful users to transmit their tactile data

by meeting the latency requirements.

A QoE-based optimisation formulation was proposed for 5G-enabled Tactile Internet in [44]. The objective function was based on BIP and a QoE equation expressed in terms of packet latency for considered stabilisation approaches (i.e. Time-Domain Passivity Approach (TDPA) and Model-Mediated Teleoperation (MMT)). In order to solve this exponentially complex optimisation problem, authors derived a heuristic solution that transforms the original problem into a max-min one, such that the proposed scheduling algorithm can finish within hard time constraint (0.1 ms). Simulation results demonstrate that the proposed scheduler generated improved goodput compared with the traditional PF scheduler. On the other hand, simulation results proved that the proposed algorithm could effectively reduce the latency level for both TDPA and MMT schemes. It is worth mentioning that although the aforementioned scheduling algorithms use either average latency or experienced with satisfying probability as a constraint of the optimisation formulation, they did not indicate how this information can be achieved or estimated by the eNB.

2.3 Mobile Cellular Networks Enhancement by 5G

The growing number of mobile devices and mobile applications has instigated investment of research and development of the next generation mobile cellular network—5G. Although it was reported that the main driving force of such rapid growth is video streaming [114, 115], demands from emerging services such as enhanced mobile broadband (eMBB), ultra-reliable and low latency communications (URLLC) and massive machine type communications (mMTC) have also stimulated the development of 5G. In this section, we discuss the technical advancement proposed by 5G, especially those promise to provide ultra-low end-to-end latency.

2.3.1 Technical Advancement of 5G

First, to cope with the exponentially growing demand for data rate by users, 5G provides increased bandwidth and massive MIMO antenna technology. Specifically, 5G networks

could provide larger bandwidth than LTE by utilising the available spectrum ranging from 28-300 GHz. On the other hand, the increased frequency in 5G enables advanced massive MIMO transmission at the base station, which significantly improves spatial multiplexing gain and hence reduces the impact of path loss on wireless communications.

Second, the idea of placing a moderate service centre at the edge of a mobile cellular network is known as mobile edge cloud (MEC), which is an effective approach by 5G networks to bring server close to the base station [26, 116]. Namely, highly adaptive and energy-efficient computing (HAEC) units can be employed such that application servers can run as close to users as possible, which could significantly reduce propagation delay compared with deploying a heavy dedicated infrastructure remotely. However, there exist open challenges such as load and task scheduling among edge servers and choosing locations of edge server deployment for maximised performance and energy efficiency [117–119].

Third, non-orthogonal transmission scheme is proposed as a promising PHY layer technology to offer multiplexing gain by overloading the spectrum[120]. Candidate schemes include Non-Orthogonal Multiple Access (NOMA), Sparse Code Multiple Access (SCMA) and Interleave Division Multiple Access (IDMA). By offering users extra dimensions of radio resources, the number of allowable active mobile connections is expected to increase by two to three times, which could in turn boost network throughput by 50%. In particular, non-orthogonal transmission scheme works by using interference cancellation or joint detection technologies to successfully receive and decode radio signals from multiple mobile users which uses the same radio resource in frequency and time. This is especially helpful for CBS uplink protocol as it tolerates a certain extent of transmission collision, which greatly reduces the chance of failed data transmission for a contention-based uplink protocol.

Finally, due to the vast variability of target use cases, 5G networks must be designed to be flexible in order to meet the connection requirement of current and emerging services. In LTE, OFDM is considered adequate for good system capacity, yet flexibility needs to be introduced in the numerology of frame structure so that both bandwidth-limited applications(e.g. video streaming) and latency-limited applications (e.g. haptic

interaction) can be supported. To support PHY layer flexibility, 5G allows to have up to 2^M OFDM symbols per TTI. In other words, the time period of a TTI can be as short as one OFDM symbol, catering for URLLC scenario.

2.3.2 5G MAC Layer Solutions for Ultra-Low Latency Performance

In this section, we discuss how the latency performance of CBS and SPS protocols can be improved with 5G advancement. As discussed in the preceding section, Non-Orthogonal Multiple Access (NOMA) has been proposed as a promising PHY layer technology for better resource utilisation efficiency and enhanced reliability. By using the power difference of received signals or codebook signature and applying Successive Interference Cancellation (SIC) at the receiver side, overlapped/superposed radio signals from different users can be received and decoded correctly. This benefit can naturally be applied to CBS to reduce collision probability among uplink users. Although SIC for uplink NOMA is conducted by LTE-A/5G eNodeB, there is a strong motivation towards offloading SIC computation to dedicated devices for faster processing speed. For instance, a GPU-based approach has been introduced to alleviate heavy computation loads of SIC receivers, such that significant speedup can be achieved due to the parallel processing capability of GPU [121].

Li *et al.* studied the feasibility of applying NOMA to CBS and proposed the first UE pairing procedure [122]. The proposed NOMA-CBS protocol allows NOMA for scheduled/unscheduled uplink devices and for granted/ungranted radio resources. For the pairing algorithm employing Contention Based Radio Network Temporary Identifiers (CB-RNTIs), each unscheduled device is assigned with one type of CB-RNTIs, by which the SNR of each device can be identified as high/middle/low. After then, the eNB can assign radio resources that are already occupied by low/vacant/high SNR, to unscheduled devices with high/middle/low SNR. As such, the uplink signals using the same radio resource at the same time can be distinguished well by NOMA, which efficiently enhances utilisation efficiency and reduces transmission failure probability. Results from simulation of NOMA-assisted LTE-A reveal that NOMA-assisted CBS achieved the least latency as compared with DS and conventional CBS protocols. However, the mean arrival rate of

each uplink user was 2 packets/second with Poisson distribution, and this assumption needs further justification. On the other hand, the correct detection ratio in the NOMA process was a pre-assumed value, lacking theoretical explanation, and therefore it may lead to inaccurate performance estimation.

Contrary to the preceding work where only NOMA was applied to CBS, an overall evaluation of a 5G system with CBS was conducted in [99]. In particular, the considered 5G network consisted of hopping-aided semi-static resource configuration, blind UE detection empowered by referencing signal and consecutive HARQ retransmissions mechanism. It was shown that an increased number of retransmissions could improve the block error rate (BLER) significantly, which could potentially meet the reliability requirement of URLLC. Simulation results also prove that reliability could be further enhanced by NOMA algorithms (e.g. Message Passing Algorithm (MPA) or Expectation Propagation Algorithm).

Instead of simulating the 5G architecture in simplified scenarios (e.g. single cell), the work presented in [103] studied the effect of different HARQ schemes in an outdoor 3GPP urban macro setup with 21 cells. In particular, three HARQ schemes were proposed to enhance the reliability of CBS uplink protocol—conventional, K-rep and proactive HARQ. The K-rep HARQ allows a device to transmit associated uplink data for a predefined K consecutive times without waiting for HARQ feedback. Such an HARQ scheme improves the reliability of CBS uplink transmission while potentially degrading resource efficiency. On the other hand, proactive HARQ allows a device to dynamically determine the number of replicas of uplink data by early termination informed by HARQ feedback. Therefore, the waste of resources can be reduced while the transmission reliability is unharmed. Simulation results show that conventional HARQ is able to support the largest traffic load as compared with K-rep and proactive HARQ schemes, as no resource was wasted for duplicate data transmission. On the other hand, proactive HARQ was strongly recommended for the best latency performance, when the network was exposed to light traffic.

Different to the proactive HARQ, by which the dynamic data repetition is enabled, Singh *et al.* attempted to achieve the same reliability performance by determining the op-

timal number of repetitions [123]. By deriving the expression for collision rate for a given load and number of available resources, a network operator can choose an optimal number of repetitions for required reliability and latency performance. Further, by acknowledging that CBS could under-perform scheduled allocation under heavy traffic load, the authors proposed a flexible access scheme such that network operator can choose to use either K-rep CBS or scheduled contention-free allocation, such that prominent reliability gains and lower implementation complexity could be achieved.

As for SPS, the latency improvement brought by short TTI (sTTI) in 5G standards was evaluated in [124]. Simulation results show that when a single UE uploaded data to an eNB, the average experienced latency is 2.04 ms when sTTI was set as 0.14 ms. However, the tested traffic load/data arrival rate was not listed. Nevertheless, simulation results reveal that SPS combined with sTTI offers the feasibility for LTE technologies to provide ultra-low latency for uplink UEs.

In summary, CBS offers a promising protocol for instant uplink transmissions. However, mechanisms that prevent signal reception failure from multiple uplink devices sending signals using the same resources at the same time must be developed and carefully deployed, in order to achieve satisfying latency and reliability performance. Advanced PHY layer technologies such as NOMA can be applied to allow successful data detection and decoding for superposed data, which effectively reduce the data collision rate. On the other hand, techniques such as K-rep HARQ helps to improve reliability performance, if collisions happened. As for SPS, it is expected to outperform the conventional DS scheme in terms of uplink latency and signalling overhead.

2.4 Haptic Codecs for the Tactile Internet

The well-known compression technique for kinesthetic perception called deadband principle is introduced in this section. The Tactile Internet aims at providing a complete remote immersion by physical interaction with the virtual/remote environment over the Internet [28]. As such, a new type of data called haptic data is transmitted in the Tactile Internet, which differ significantly to conventional data types such as video and audio

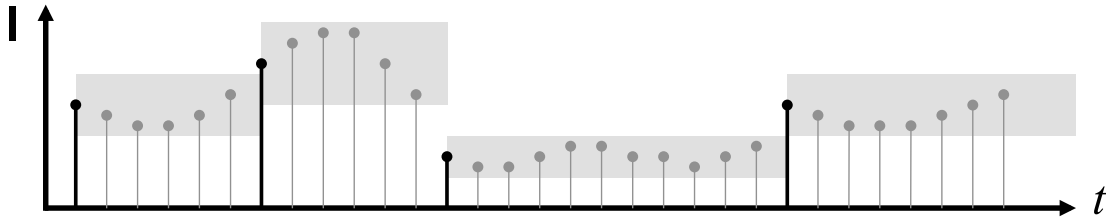


Figure 2.13: Illustration of 1-DoF deadband applied to a signal, force intensity I as a function of time t (adopted from [127]).

contents. Specifically, haptic data can be classified into two categories—kinesthetic and tactile information. The kinesthetic perception delivers information of limb position and movement in space and perceived force/torques acting by the human body. On the other hand, the tactile information is used by human to perceive the texture of an object, and it usually contains several dimensions of information, namely friction, hardness, warmth conductivity, macroscopic and microscopic roughness [28]. Due to the intrinsic characteristics of haptic data, modern codecs which control the acquisition, compression and transmission of conventional data types are not applicable for the Tactile Internet [125].

The deadband principle was introduced as a promising haptic data reduction technique for a TelePresence and TeleAction (TPTA) system [28, 126, 127]. A major conclusion from vast human perception studies is that human kinesthetic perception can be approximated by Weber's Law:

$$\Delta I = kI \quad (2.4)$$

which states that the smallest still perceivable change of human stimulus intensity ΔI is determined by the current stimulus intensity I and a constant k —just noticeable difference (JND). This conclusion provides an opportunity for kinesthetic data compression called the deadband principle for the Tactile Internet. This is because some sampled kinesthetic data can be treated as “unnecessary for transmission” and removed from the application's data buffer, given that these kinesthetic data is assumed unperceivable by the operator.

Fig. 2.13 illustrates the process of using the deadband principle for 1-D kinesthetic data compression. The first black dot denotes the initial generated kinesthetic data, and

the first group of grey dots represents the subsequent kinesthetic data that are within the JND range of the initial kinesthetic data, which shall be ignored because they do not affect the perception of an operator. On the other hand, the value of the second black dot is beyond the JND range of the first data, and hence it is stored into the application buffer and defines a new perceptual deadband for ongoing haptic operation. The authors in [126] reported a packet reduction rate of 8.7% for velocity and 7.4% for force with barely noticeable influence on the extent of immersion when the deadband principle is applied.

2.5 Conclusions

In this chapter, we show that the latency performance of wireless communications is limited by signalling process. Particularly, the signalling process in the uplink direction forms a bulk proportion of the overall latency of today's wireless communications network. In addition, we discuss and compare MAC layer protocols and corresponding latency improvement techniques of two dominant wireless technologies—WLAN and mobile cellular network. As for WLAN, HCCA does not only provide admission control, which is important for QoS provisioning, but also schedules guaranteed channel airtime for admitted devices. Therefore, HCCA could be a viable MAC layer solution for WLAN to support latency-critical services. Complex mechanisms have been designed for HCCA to achieve efficient dynamic TXOP allocation. However, current literature lacks an analytical evaluation of the implications of different timing and network parameters on the latency performance of HCCA. On the other hand, most research about QoS provisioning for mobile cellular network focus on resource scheduling for the DS scheme, which leaves signalling delay unaffected and hence may not meet the ultra-low latency requirement. The SPS scheme, which shortens the signalling delay in uplink transmissions, is hence considered a potential MAC solution for mobile cellular network to support latency-critical services. It should be noted that most existing works about these two wireless technologies are not designed for data packets with small and constant payload size—a major paradigm shift envisioned for the Tactile Internet.

Based on the studies reported in this chapter, there is an urgent need to further in-

investigate the feasibility of applying WLAN and mobile cellular network for the Tactile Internet. As such, we design a strategic parameter selection algorithm for HCCA based on an analytical study of latency performance and propose a predictive SPS scheme for LTE networks to address the ultra-low latency requirement. A detailed description of our technical contributions will be presented in Chapter 3, 4 and 5.

HCCA of IEEE 802.11 WLAN for Latency-Critical Haptic Applications

3.1 Introduction

Unlike conventional applications such as video streaming, teleconferencing and data transmission, a communications network that supports future machine-to-machine (M2M) and haptic-tactile applications may need to provide stringent end-to-end latency guarantees (~ 1 ms) [11, 16]. These applications promise users an immersive experience with unnoticeable latencies. Otherwise, due to extremely low human reaction time to haptic stimuli, users will suffer from a symptom known as “cyber sickness”, similar to motion sickness caused by conflicting visual and vestibular inputs. A detailed survey of “cyber sickness” can be found in [128].

To support these future applications and services, the wireless domain of a considered network is expected to contribute no more than ~ 0.2 ms of the 1 ms end-to-end latency [16]. To date, research in wireless technologies is focused on cellular 5G network [26]. Due to ease of deployment and cost-effectiveness of WLAN technologies, investigating the feasibility of these technologies in supporting haptic-tactile communications is crucial and has yet to be addressed.

An example of a communications network that supports haptic-tactile applications is shown in Fig. 3.1. Here control/steering data which is sent to/from individual actuator and sensor (A/S) is aggregated at a user device called Hub. Communications between several A/Ss and a Hub can be enabled by using, for example, a Smart Body Area Network (SmartBAN) [129]. In turn, each Hub connects to an access point using an HCCA-enabled WLAN, which then connects to the Internet through fibre optical com-

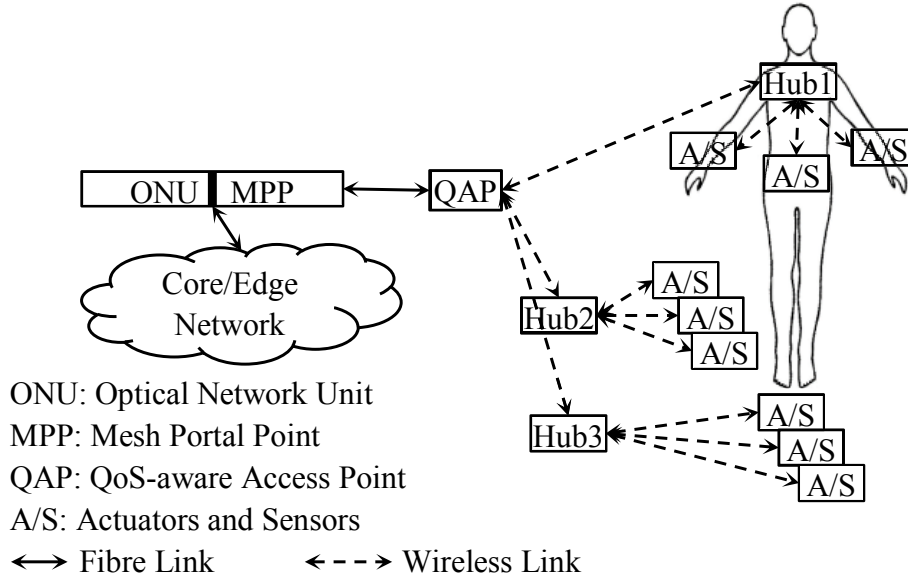


Figure 3.1: Network infrastructure (comprising optical and wireless domains) that enables the support of haptic-tactile applications.

munications.

As mentioned in Section 2.1.2.3, DCF has been widely deployed by conventional IEEE 802.11 networks but wireless packet collision cannot be completely avoided, which leads to unbounded latency. On the other hand, HCF controlled Channel Access (HCCA) was first introduced in IEEE 802.11e amendment. It is worth mentioning that while the predecessor of the HCCA such as the point coordinate function (PCF) can be analysed by classical M/G/1 queueing theory, the performance of the HCCA is usually studied simulatively. In this chapter, we analyse and derive the first closed-form expressions for the average uplink latency introduced by HCCA. We then investigate the suitability of HCCA in meeting stringent latency requirements. In essence, the HCCA implements a polling mechanism that guarantees a certain amount of service time for connected users. Due to this nature, HCCA is expected to outperform other existing MAC protocols specified in [56], in terms of providing quality of service (QoS). In [78], Costa *et al.* have compared the performance of DCF, PCF, EDCA and HCCA for an IEEE 802.11 simulation network. Results show that HCCA had the best performance for average latency and deadline missing ratio among considered MAC protocols under high network load and it was the only protocol that preserved real-time traffic requirements under the impact of

external traffic sources. However, DCF and EDCA demonstrated capability to support more STAs than HCCA, which stemmed from HCCA's admission control mechanism.

The main contrition of this chapter includes the first closed-form expression for the average MAC layer uplink latency for HCCA, which fills the gap of theoretical analysis of the impact on HCCA timing and network parameters. In particular, our work considers haptic-tactile communications, whereby data frames are expected to be constant and short-length (64 bytes) [32] with a latency target at a sub-millisecond range. Moreover, we conduct global sensitivity analyses to achieve a comprehensive understanding on how uplink latency varies with each timing and network parameter. Further, using the insights gained from our theoretical model, we propose a strategic parameter selection algorithm which ensures that the stringent uplink requirements are met. As an initial step towards designing an ultra-low latency MAC layer protocol for the IEEE 802.11 to support emerging services and applications, our work answers the question of what latency level can the IEEE 802.11 technology achieve and which system parameters have the most significant impact on the latency performance of the HCCA protocol.

The remaining of this chapter is structured as follows. In Section 3.2, we discuss advanced HCCA designs in literature, especially those having a significant impact. Section 3.3 presents the detailed derivation of the first closed-form expression of the average uplink latency of HCCA, followed by simulation results. In Section 3.4, we elaborate significant improvement for our analytic model and explain the working principle of our proposed algorithm. Finally, we conclude this chapter with a summary of our findings in Section 3.5.

3.2 Related Work

As mentioned in Section 2.1.2.3, [56] defines a reference design for the HCCA scheduler, where Service Interval (SI) and Transmission Opportunity (TXOP) are pre-determined based on data traffic statistics. Extensive studies on this reference design showed that this reference design works well for constant bitrate (CBR) traffic, Whereas the fixed SI setting and TXOP assignment of HCCA is not able to adapt bursty data for variable bi-

trate (VBR) traffic [36, 130, 131]. Therefore, many advanced scheduling algorithms have been proposed to assist HCCA with dynamic TXOP assignment, such that traffic fluctuation can be accommodated for improved system performance. In this section, we discuss and compare some significant algorithms in the existing literature.

The very first dynamic TXOP algorithm—FHCF was proposed in [33]. Firstly, it uses a moving window to calculate the expected deviation of ideal and estimated buffer sizes for each STA. It then uses this deviation to dynamically determine the needed length of a TXOP for $STA_i - t_i^{est}$, where i represents the index of each STA. To further assure fairness among connected STAs, the actual allocated TXOPs are proportionally shrunk, in case the length of Controlled Access Phase (CAP) cannot host all calculated t^{est} . In essence, FHCF is a scheduling algorithm that considers history statistics of buffer sizes. One drawback of the FHCF mechanism is that it only considers the absolute value of buffer length difference, making the adjustment of TXOP purely positive. Via simulation in heterogeneous traffic flow environment—Audio, VBR video and CBR MPEG4 video, FHCF outperforms the reference design in terms of packet drop rates of VBR video and CBR MPEG4 video traffic flows and therefore improves the overall fairness among transmitted data. Simulation results further show that by deploying FHCF, the latency performance of VBR video flow can be significantly improved at the expense of latency increase of other two traffic flows. Whether such performance improvement results from the dynamic TXOP mechanism or the proportional shrunk mechanism is unanswered.

Another well-known scheduling algorithm proposed for HCCA is Adaptive Resource Reservation Over WLANs (ARROW) [34], which performs adaptive TXOP allocation by exploiting the queue size (QS) field of QoS data frames. ARROW determines the amount of assigned TXOP solely based on the information of buffer size transmitted in previous SI. Although such a mechanism can reflect the data fluctuation for VBR traffic, it is not sufficient for estimating current queued data size. Therefore, [34] states that with ARROW, data arriving at STA_i between SI_t and SI_{t+1} can be transmitted no earlier than $TXOP_{i,t+2}$, where i represents the index of each STA, and t is a time index for each transmission. Simulation results show that increased throughput was achieved by ARROW, especially when the number of connected STAs exceeds a certain value. While the la-

tency performance of ARROW-assisted HCCA deteriorates when the number of connected STAs is either very small or very large. These results indicate that ARROW does not conquer the throughput-delay trade-off commonly observed in communications networks [107].

Another interesting scheduling algorithm called Dynamic TXOP HCCA (DTH) was proposed in [36, 132]. This algorithm consists of a reclaiming mechanism and an estimation mechanism. The reclaiming mechanism simply allows changeable starting time for each TXOP, in case the earlier polled STAs does not fully utilise their scheduled TXOPs. The estimation mechanism calculates the moving average of effectively utilised transmission resource by each STA and uses this value as the TXOP allocation for subsequent transmissions. Moreover, the DTH incorporates an anti-saturation mechanism, such that the TXOP allocation will not stop growing at a certain circumstance. Simulation results show that DTH was effective to reduce the CF-Null sending rate of all considered traffic flows, indicating that DTH can effectively reduce the amount of assigned, but unused transmission time. On the other hand, DTH offered reduced packet drop rate for some VBR video traffic flows, while other traffic flows (i.e. VoIP and teleconference) suffered from higher packet drop rate. In terms of latency performance, VoIP traffic flow had the lowest latency (~ 10 ms) when DTH was deployed, only a limited amount of improvement compared with the reference design.

Authors [42] took a different approach and designed a feasible HCCA polling scheme (F-Poll). Via cross-layering, the exact next frame arrival time of STA_{*i*}'s traffic flow-arrival_{*i*} can be obtained by STAs. To enable F-Poll, protocol changes are required such that the obtained information is encapsulated in the QS field of QoS data frames. Upon receiving arrival_{*i*}, a QAP then selectively polls STAs that are ready to transmit in order to reduce the poll overhead and hence reduces the HCCA latency. Simulation results reveal that F-Poll can significantly reduce the end-to-end latency compared with the reference design. However, as the number of transmitting STAs ahead of a target STA can vary, the latency of F-Poll-assisted HCCA can largely fluctuate. Furthermore, the number of transmitted CF-Null frames is close to zero, meaning that TXOP allocation by F-Poll is efficient such that no transmission resource was wasted. One major constraint of F-Poll is that it

strongly depends on the cross-layer concept for information extraction of prerecorded VBR video streams, whereas such method is inapplicable to other interactive real-time communications.

Nonetheless, existing studies are based on computer simulations without closed-form analysis of resulting uplink latency, limiting the theoretical analysis of the impact of HCCA timing and network parameters. Moreover, existing HCCA scheduling schemes have not considered haptic-tactile communications, whereby data frames are expected to be constant and short-length (64 bytes) [32] with a latency target at a sub-millisecond range. Therefore, our work aims at filling this gap.

3.3 Initial Average Uplink Latency Analysis for HCCA

In this section, we develop the first analytical uplink latency model for the HCCA MAC protocol based on an M -user $M/G/1$ queueing system, in order to evaluate its latency performance. The memoryless Poisson distribution assumes that the distribution of arrival time for a particular network packet is not affected by the arrivals of previous data, which is an appropriate model if arrived data are generated from a large number of independent sources. Therefore, Poisson distribution and $M/G/1$ queueing system have been widely implemented to model the arrivals of network data packets in existing literature [59,64–67,111,133]. More importantly, we exploit this analytical model to understand the implications of diverse timing parameters on latency performance. To the best of our knowledge, this is the first service interval selection method for the HCCA MAC protocol that can be used as a guideline for network and service operators to meet latency requirements.

3.3.1 Derivation of the First Closed-Form Expression

First, we use $\bar{D}_{\text{HCCA}}$ to represent the average uplink latency of HCCA. We also model the arriving process of aggregated uplink data from various A/Ss to a particular Hub as a Poisson process with rate λ and hence the overall uplink arrival rate is given by $M\lambda$. Additionally, we consider that the QAP in this network polls Hubs according to its

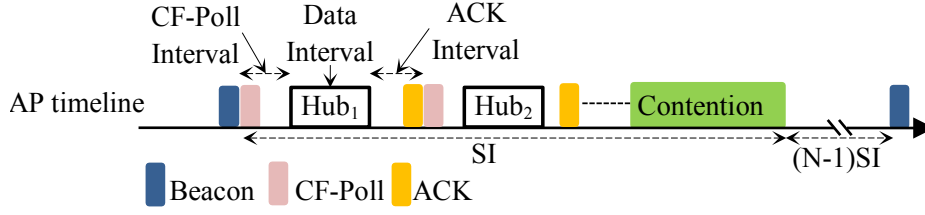


Figure 3.2: HCCA queueing model.

polling list in a fixed order. The queueing model comprises reservation intervals, which includes CP and poll-signalling frames (i.e. CF-Poll and ACK frames). The queueing model of HCCA is illustrated in Fig. 3.2. We assume that Hubs with latency-sensitive (LS) traffic transmit data to QAP via HCCA only, while CP is used for transmitting non-latency sensitive (NLS) traffic. Let $\bar{Z}$ and $\bar{Z}^2$ represent the first and second moments of a random variable Z , respectively.

First, $\bar{D}_{\text{HCCA}}$ can be expressed by $\bar{W} + X$ where $\bar{W}$ is the average waiting time in the queue and X is the transmission time of an LS data frame. Here, we assume that haptic data frames are short and have fixed length [32], and hence the term X is a constant. $\bar{W}$ can be further decomposed into three terms [134]:

$$\bar{W} = \bar{R} + \bar{Y} + \bar{Q}, \quad (3.1)$$

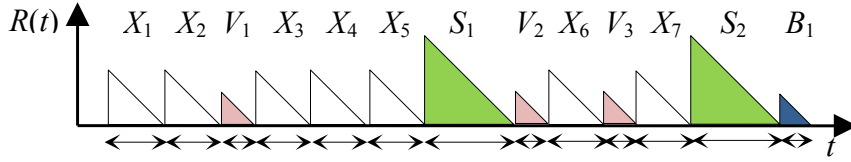
where $\bar{R}$ represents the average residual time to finish an ongoing transmission. $\bar{Y}$ denotes the average waiting time incurred by reservation intervals, while $\bar{Q}$ represents the waiting time due to transmissions of earlier LS frames, respectively. In particular, $\bar{Y}$ can be further split into two parts: $\bar{Y}_1$, the average waiting time caused by reservation intervals when the assigned TXOP is sufficient and $\bar{Y}_2$, the extra waiting time caused by reservation intervals when TXOP is insufficient. The notations used for the queueing model is presented in Table 3.1.

Lemma 3.1. *The mean residual time $\bar{R}$ is given by:*

$$\bar{R} = \frac{\rho_{\text{total}}X}{2} + \frac{MN(\mu^2 + \nu^2)}{2(NT_{\text{SI}} + B)} + \frac{N\bar{S}^2}{2(NT_{\text{SI}} + B)} + \frac{B^2}{2(NT_{\text{SI}} + B)}. \quad (3.2)$$

Table 3.1: Notations for queueing analysis

Notation	Description	Notation	Description
ρ_{total}	Total traffic load (equals $M\rho$)	μ	Time for a CF-Poll+SIFS frame
ρ	Traffic load of each Hub	ν	Time for an ACK+SIFS frame
X	Time for transmitting one haptic data frame	λ	Hub data frame arrival rate of each Hub
M	Number of currently supported Hubs in the system	T_{max}	The assigned TXOP for a Hub reduced by V
N	Number of service intervals between two successive beacon frames	D_{HCCA}	Uplink MAC layer latency
V	Time for reservation intervals (CF-Poll+ACK+2SIFS)	B	Time to transmit a beacon frame
S	Time for a contention phase	S'	Time for a contention phase when Hub _{i} is using full allocated TXOP
T_{SI}	Time of a service interval	i	Index of the Hub of interest

Figure 3.3: Graphical argument approach for residual time R .

Proof. We use a graphical argument method to prove (3.2) [134]. Fig. 3.3 shows the residual time as a function of time t . Note that when a new reservation, data or contention phase interval begins, $R(t)$ decays linearly for X , V or S time units. Therefore, $\bar{R}$ can be written as:

$$\begin{aligned}
\bar{R} &= \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t R(\tau) d\tau = \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{i=1}^{X(t)} \frac{X_i^2}{2} + \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{i=1}^{\mu(t)} \frac{\mu_i^2}{2} + \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{i=1}^{\nu(t)} \frac{\nu_i^2}{2} + \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{i=1}^{S(t)} \frac{S_i^2}{2} \\
&\quad + \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{i=1}^{B(t)} \frac{B_i^2}{2}, \tag{3.3}
\end{aligned}$$

where $X(t)$, $\mu(t)$, $\nu(t)$, $S(t)$ and $B(t)$ are the number of data frame intervals, CF-Poll inter-

vals, ACK intervals, contention phase interval and beacon frame intervals, respectively up to time t . Likewise, X_i , μ_i , ν_i , S_i and B_i represent the durations of these intervals.

When $t \rightarrow \infty$, $\frac{X(t)}{t}$ will be equivalent to the data frame generation rate, which is $\frac{\rho_{\text{total}}}{X}$ by Little's Law. Thus, the first term on the right-hand side (RHS) of (3.3) is:

$$\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{i=1}^{X(t)} \frac{X_i^2}{2} = \frac{1}{2} \left(\lim_{t \rightarrow \infty} \frac{X(t)}{t} \right) \left(\lim_{t \rightarrow \infty} \frac{\sum_{i=1}^{X(t)} X_i^2}{X(t)} \right) = \frac{\rho_{\text{total}} \overline{X^2}}{2X}. \quad (3.4)$$

As the size of the control/actuation data frame is assumed to be constant [32], we can write (3.4) as:

$$\frac{\rho_{\text{total}} \overline{X^2}}{2X} = \frac{\rho_{\text{total}} X}{2}. \quad (3.5)$$

When $t \rightarrow \infty$, $\frac{V(t)}{t}$ will be equivalent to the number of reservation intervals within two successive beacon frames, which is given by $\frac{MN}{NT_{\text{SI}} + B}$. As a result, the second and third terms of the RHS of (3.3) can be simplified to:

$$\begin{aligned} \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{i=1}^{\mu(t)} \frac{\mu_i^2}{2} + \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{i=1}^{\nu(t)} \frac{\nu_i^2}{2} &= \frac{1}{2} \left(\lim_{t \rightarrow \infty} \frac{\mu(t)}{t} \right) \left(\lim_{t \rightarrow \infty} \frac{\sum_{i=1}^{\mu(t)} \mu_i^2}{\mu(t)} \right) \\ &+ \frac{1}{2} \left(\lim_{t \rightarrow \infty} \frac{\nu(t)}{t} \right) \left(\lim_{t \rightarrow \infty} \frac{\sum_{i=1}^{\nu(t)} \nu_i^2}{\nu(t)} \right) = \frac{MN(\mu^2 + \nu^2)}{2(NT_{\text{SI}} + B)}. \end{aligned}$$

Similarly, the rest of (3.3) can be written as:

$$\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{i=1}^{S(t)} \frac{S_i^2}{2} + \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{i=1}^{B(t)} \frac{B_i^2}{2} = \frac{N\overline{S^2}}{2(NT_{\text{SI}} + B)} + \frac{B^2}{2(NT_{\text{SI}} + B)},$$

which concludes the proof. $\square$

Now in order to determine the mean waiting times during reservation and contention phase intervals, we decompose $\overline{Y}$ into two terms, $\overline{Y}_1$ and $\overline{Y}_2$. The term $\overline{Y}_1$ represents the mean waiting time during reservation and contention phase intervals if the assigned

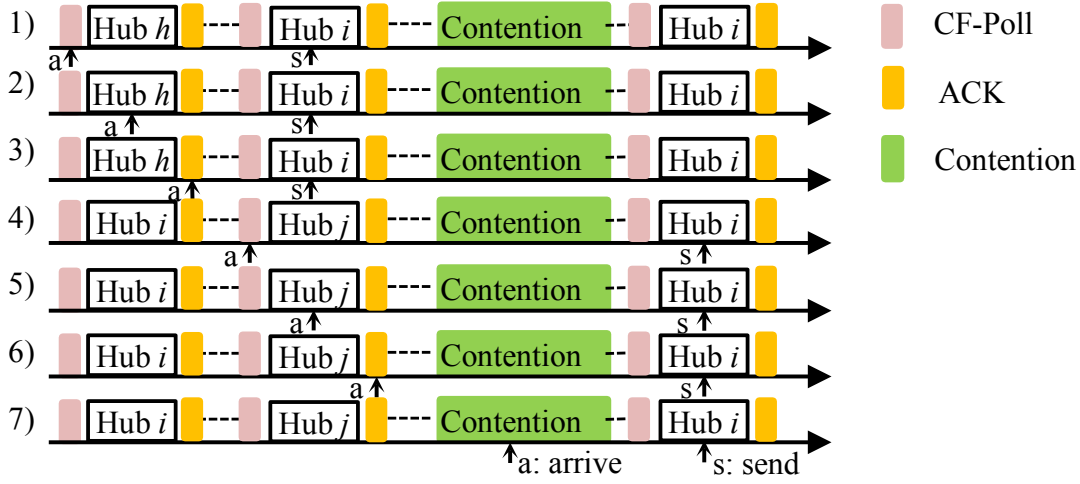


Figure 3.4: Waiting time for reservation and contention phase intervals under different conditions.

TXOP is sufficient to transmit all the queued data frames up to the data frame of interest. On the other hand, $\bar{Y}_2$ is the extra mean waiting time due to limited TXOP.

Lemma 3.2. *The term $\bar{Y}_1$ can be written as:*

$$\begin{aligned} \bar{Y}_1 = & \frac{(iV - v)\bar{S}}{T_{SI}} + \frac{\mu(M^2V + 2M\bar{S} - MV - 2i\bar{S})}{2T_{SI}} \\ & + \frac{\rho(M^2V + MV + 2(M - i + 1)\bar{S})}{2} \\ & + \frac{v(M^2V + MV - 2Mv + 2(M - i + 1)\bar{S})}{2T_{SI}}. \end{aligned} \quad (3.6)$$

Proof. The mean waiting time during reservation and contention phase intervals depends on the arrival time of the data frame. We assume that the data frame of interest is generated by Hub_i, which has a traffic load of ρ . Let I_n denotes the order of a contending STA–Hub_n in the AP’s polling list. Assume Hub_h, Hub_i and Hub_j are scheduled such that $I_h < I_i < I_j$, Table 3.2 and Fig. 3.4 summarise possible scenarios that can lead to different waiting times for Hub_i during reservation and contention phase intervals. For instance, in scenario (1), a data packet arrives at Hub_i when the AP is transmitting a CF-Poll to Hub_j. The data packet can be transmitted by Hub_i given Hub_i has sufficient TXOP. While in scenario (4), a data packet arriving at Hub_i when the AP is transmitting

Table 3.2: Seven scenarios of different waiting time for reservation and contention phase intervals.

Frame arrives during:	Probability	Value of Y_1
Hub j 's CF-Poll ($j \leq i$)	$\frac{\mu}{T_{SI}}$	$(i - j)V$
Hub j 's data interval ($j < i$)	ρ	$(i - j)V$
Hub j 's ACK ($j < i$)	$\frac{\nu}{T_{SI}}$	$(i - j)V - \nu$
Hub j 's CF-Poll ($j > i$)	$\frac{\mu}{T_{SI}}$	$(M + i - j)V + \bar{S}$
Hub j 's data interval ($j \geq i$)	ρ	$(M + i - j)V + \bar{S}$
Hub j 's ACK ($j \geq i$)	$\frac{\nu}{T_{SI}}$	$(M + i - j)V - \nu$
Contention phase	$\frac{\bar{S}}{T_{SI}}$	$iV - \nu$

a CF-Poll to Hub $_j$ can only be transmitted in the next polling cycle. It is worth noting that V is considered the summation of CF-Poll, ACK intervals and two SIFS intervals. It is also important to note that we do not consider the effect of the beacon frame in our analysis as its impact on $\bar{Y}_1$ is negligible, especially when N is large. The expected value of $\bar{Y}_1$ for a data frame that is generated by Hub $_i$ is then derived from the Table :

$$\begin{aligned}
\bar{Y}_1 = & \sum_{j=1}^i \frac{\mu(i-j)V}{T_{SI}} + \sum_{j=1}^{i-1} \rho(i-j)V + \sum_{j=1}^{i-1} \frac{\nu[(i-j)V - \nu]}{T_{SI}} \\
& + \sum_{j=i+1}^M \frac{\mu[(M+i-j)V + \bar{S}]}{T_{SI}} + \sum_{j=i}^M \rho[(M+i-j)V + \bar{S}] \\
& + \sum_{j=i}^M \frac{\nu[(M+i-j)V - \nu + \bar{S}]}{T_{SI}} + \frac{\bar{S}(iV - \nu)}{T_{SI}},
\end{aligned}$$

yielding the expression in Lemma 3.2.

□

Lemma 3.3. *The term $\bar{Y}_2$ can be written as:*

$$\bar{Y}_2 = \frac{\rho \bar{W}(MV + \bar{S}')}{T_{max}}. \quad (3.7)$$

Proof. When a data frame arrives at the queue of a Hub, the average number of data frames that are already waiting in the queue is $\lambda \bar{W}$ by Little's Theorem. During one TXOP, a maximum number of $\frac{T_{max}}{X}$ data frames will be transmitted. Thus, on average, the

number of TXOPs required to transmit all the previously arrived data frames is $\frac{\lambda \bar{W} X}{T_{\max}} = \frac{\rho \bar{W}}{T_{\max}}$.

Let S' denote the contention phase interval in such a scenario. Extra waiting time during reservation and contention phase intervals of $MV + \bar{S}'$ will be introduced for each TXOP group, thus $\bar{Y}_2$ can be expressed in (3.7).

□

Let T' denote the average actual data frame transmission time if a Hub does not use the full assigned TXOP, i.e. when $T' < T_{\max}$. Also, let q denote the probability that a Hub uses full TXOP.

Lemma 3.4. $\bar{S}$, $\bar{S}'$ and $\bar{S}^2$ can be approximated as:

$$\bar{S} \approx T_{SI} - MV - \rho_{total} T_{SI} , \quad (3.8)$$

$$\bar{S}' \approx T_{SI} - MV - (M-1)qT_{\max} - T_{\max} , \quad (3.9)$$

$$\bar{S}^2 \approx (T_{SI} - MV)^2 + T_{\max} \rho_{total} T_{SI} \left(1 + \frac{\rho_{total} T_{SI}}{T_{\max}} - \frac{\rho_{total} T_{SI}}{MT_{\max}} \right) - 2\rho T_{SI}(T_{SI} - MV) . \quad (3.10)$$

Proof. By definition, the contention phase is the time remaining within a service interval, after HCCA has completed the controlled phase. As Hubs utilise their corresponding TXOP independently, the random variable S follows the binomial distribution:

$$Pr\{S = T_{SI} - MV - jT_{\max} - (M-j)T'\} = \binom{M}{j} q^j (1-q)^{M-j} ,$$

where j represents the number of Hubs using full TXOP. The value of $\bar{S}$ is therefore:

$$\begin{aligned} \bar{S} &= \sum_{j=0}^M [T_{SI} - MV - jT_{\max} - (M-j)T'] \binom{M}{j} q^j (1-q)^{M-j} \\ &= T_{SI} - MV - MT' - Mq(T_{\max} - T') . \end{aligned} \quad (3.11)$$

By the same approach, it is easy to prove that the expected contention phase interval when one Hub is using its full TXOP is:

$$\bar{S}' = T_{SI} - MV - (M - 1)T' - (M - 1)q(T_{\max} - T') - T_{\max} . \quad (3.12)$$

Next, we approximate the value of q with an approach similar to [135]. The ratio between time consumed by data frame transmission and reservation intervals can be expressed as:

$$\frac{\rho_{\text{total}}}{1 - \rho_{\text{total}}} = \frac{M[qT_{\max} + (1 - q)T']}{MV + \bar{S}} .$$

By rearranging the above expression, we can get:

$$q = \frac{\rho_{\text{total}}(MV + \bar{S})}{M(T_{\max} - T')(1 - \rho_{\text{total}})} - \frac{T'}{T_{\max} - T'} . \quad (3.13)$$

Taking the value of $\bar{S}$ in (3.11) into (3.13) leads to:

$$q = \frac{\rho_{\text{total}}}{1 - \rho_{\text{total}}} \left[1 - \frac{MT_{\max}}{M(T_{\max} - T')} + \frac{T_{SI} - Mq(T_{\max} - T')}{M(T_{\max} - T')} \right] + 1 - \frac{T_{\max}}{T_{\max} - T'} .$$

When $T_{\max}$ is sufficiently large, q can be approximated by assuming that $T_{\max} - T' \approx T_{\max}$, which lead to:

$$q \approx \frac{\rho_{\text{total}}(T_{SI} - MqT_{\max})}{(1 - \rho_{\text{total}})MT_{\max}} . \quad (3.14)$$

By rearranging (3.14) and bringing q to the left-hand side, we can have:

$$q \approx \frac{\rho_{\text{total}}T_{SI}}{MT_{\max}} . \quad (3.15)$$

Taking (3.15) and $T_{\max} - T' \approx T_{\max}$ in (3.11) and (3.12), the terms $\bar{S}$ and $\bar{S}'$ can be approximated:

$$\bar{S} \approx T_{SI} - MV - \rho_{\text{total}}T_{SI} ,$$

$$\bar{S}' = T_{SI} - MV - (M - 1)qT_{\max} - T_{\max} .$$

Following the same assumption, an approximation for $\overline{S^2}$ can be derived:

$$\begin{aligned}\overline{S^2} &= \sum_{j=0}^M [T_{SI} - MV - jT_{\max} - (M-j)T']^2 \binom{M}{j} q^j (1-q)^{M-j} \\ &\approx \sum_{j=0}^M [T_{SI} - MV - jT_{\max}]^2 \binom{M}{j} q^j (1-q)^{M-j} .\end{aligned}\quad (3.16)$$

By further simplifying, (3.16) yields (3.10), which completes the proof of Lemma 3.4. $\square$

Finally, by using Little's Law, we can calculate the waiting time during processing data frames which arrived before the data frame of interest as:

$$\overline{Q} = \rho_{\text{total}} \overline{W} . \quad (3.17)$$

Theorem 3.1. *The average latency of HCCA MAC protocol is given by:*

$$\overline{D}_{HCCA} = \frac{M^2 T_{\max} (\overline{R} + \overline{Y}_1)}{(MT_{\max} - \rho_{\text{total}} T_{SI})(M + \rho_{\text{total}} - M\rho_{\text{total}})} + X . \quad (3.18)$$

Proof. Taking (3.7), (3.9) and (3.17) into (3.1), we have:

$$\overline{W} = \frac{M^2 T_{\max} (\overline{R} + \overline{Y}_1)}{(MT_{\max} - \rho_{\text{total}} T_{SI})(M + \rho_{\text{total}} - M\rho_{\text{total}})} ,$$

where $\overline{R}$ and $\overline{Y}_1$ are given in (3.2) and (3.6), respectively.

As the wireless signals propagate through air at the speed of light and the distance between each Hub and an AP is limited to only a few meters, the propagation latency can be considered negligible [133]. Therefore, the HCCA MAC layer latency is the sum of queueing latency and transmission latency. Hence, the average latency of the HCCA MAC protocol is given by:

$$\overline{D}_{HCCA} = \overline{W} + X ,$$

which proves Theorem 3.1. $\square$

Table 3.3: Constants for numerical average latency analysis.

Parameter	Value
ACK size (MAC header + FCS)	112 bits
Poll size (MAC header + FCS)	160 bits
Beacon size	320 bits
Payload size	512 bits
SIFS ($j \geq i$)	10 μ sec
FCS	32 bits
MAC header size	288 bits
PHY header size	192 bits
PHY data rate	540 Mbps

3.3.2 Results and Discussion

In this section, we numerically analyse the average uplink latency of the HCCA MAC protocol using the expressions derived in Section 3.3.1 and compare them with discrete-event simulations carried out using MATLAB. It is worth mentioning that the closed-form expression derived in Section 3.3.1 is denoted as Model 1. Furthermore, we comparatively analyse the latency performance under different MAC protocol parameters in order to understand their implications on the overall latency. We then point out a general guideline for network designers such that uplink latency requirement can be met by fine-tuning system parameters. Table 3.3 summarises the parameter values used in our evaluation, complying with numerology specified in [56].

3.3.2.1 Impact of Traffic Load and Service Interval (SI) on Average Uplink Latency

Fig. 3.5 shows the variation of average uplink latency with traffic load and service interval. Here, we choose to vary SI from 0.100 ms to 0.150 ms and see if these settings could meet the latency requirement. M and N are assigned with an arbitrary value of 3 and 20. The corresponding traffic load is varied until the HCCA can no longer support M Hubs. As can be seen from Fig. 3.5, the latency increases with traffic load. This is expected because the increase in traffic load results in more frequent arrivals of data

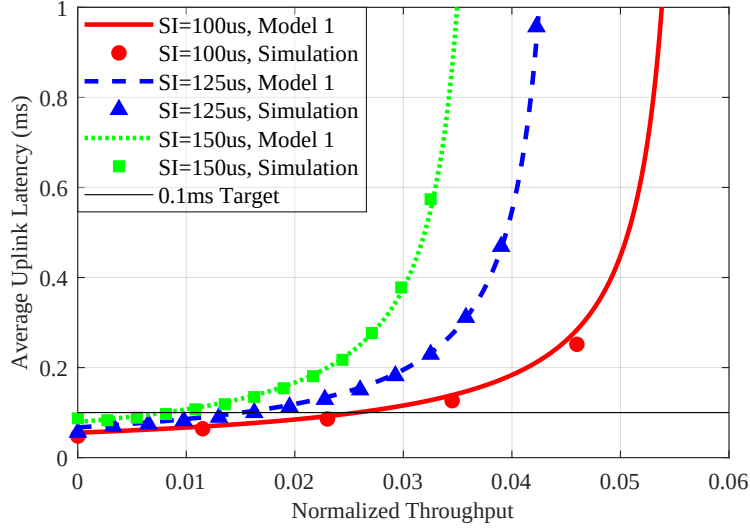


Figure 3.5: Average uplink latency for T_{SI} equals 0.1 ms, 0.125 ms and 0.150 ms ($M=3$, $N=20$, $i=3$).

frames, which thereby increases the queue length in a Hub and leads to increased waiting time. In Fig. 3.5, the numerical results from (3.18) in Section 3.3.1 are compared with simulation results, showing good agreement that validates the queueing model and our derived average latency model.

On the other hand, results in Fig. 3.5 highlight the significance of SI on the average uplink latency. The average latency increases more rapidly with traffic load when the SI value is larger. In the HCCA reference design, SI is fixed and thereby results in fixed cycle time for the polling system. However, in general, the longer the SI, the larger the waiting time before a Hub can transmit the data frames in its queue. Therefore, SI must be carefully selected in order to satisfy the latency requirement.

3.3.2.2 Impact of TXOP and Service Interval (SI) on Average Uplink Latency

In Fig. 3.6 and 3.7, we emphasise the importance of two timing parameters of HCCA—TXOP and service interval (SI), via localised sensitivity analysis when λ is set to 13000 packets/second. From Fig. 3.6, we clearly see that the increase of TXOP reduces the average latency. However, when TXOP is above a certain level, its impact on reducing latency diminishes as the sensitivity approaches zero. In Fig. 3.7, the increase of latency with SI

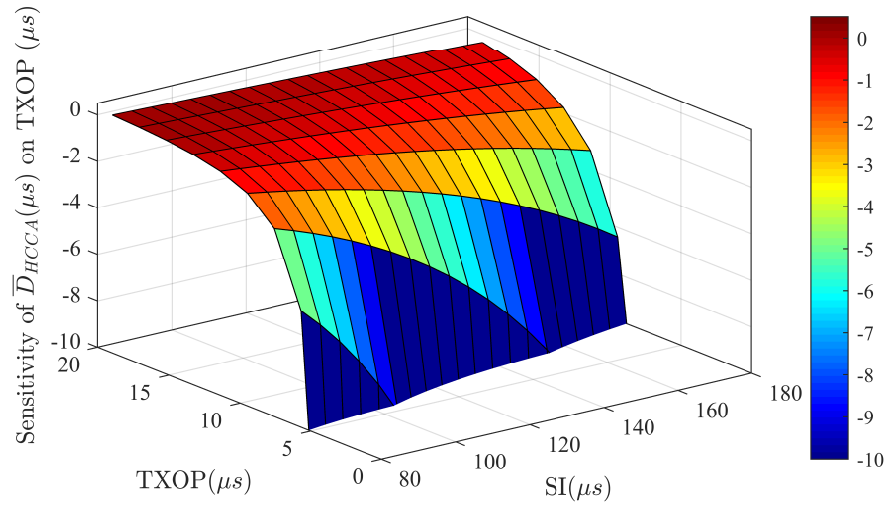


Figure 3.6: Sensitivity Analysis: Impact of TXOP on average uplink latency.

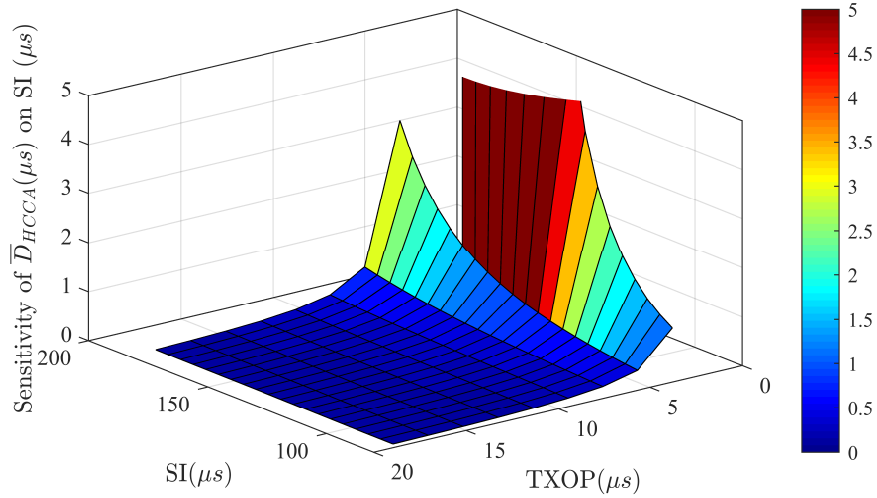


Figure 3.7: Sensitivity Analysis: Impact of SI on average uplink latency.

is greater when smaller TXOP value is selected. Therefore, if the reference design cannot provide required latency performance due to inappropriate SI and TXOP settings, the network designer may firstly increase the TXOP value, then reduce the SI value if further latency reduction is necessary.

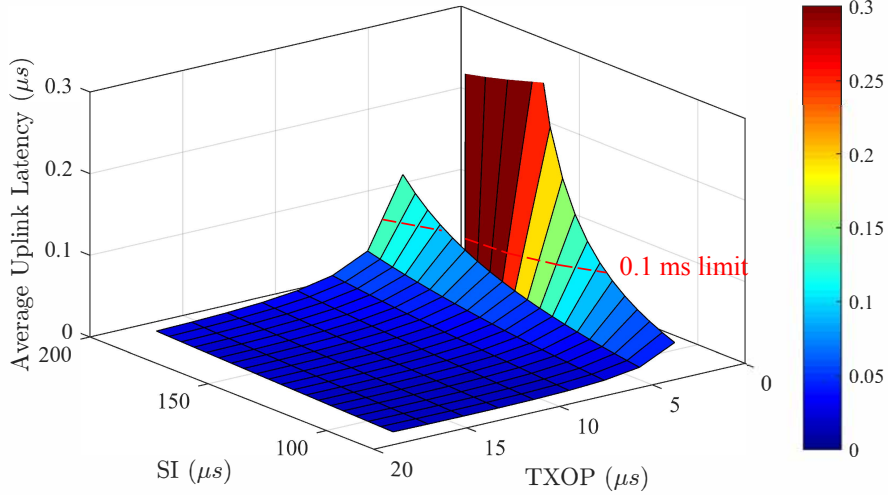


Figure 3.8: Achievable latency performance under different TXOP and SI values.

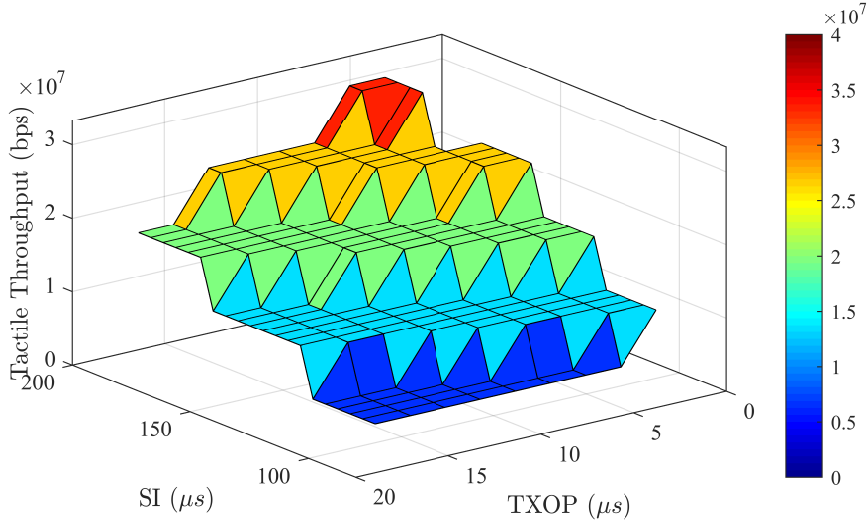


Figure 3.9: Achievable throughput under different TXOP and SI values.

3.3.2.3 Trade-off between Achievable Latency and Tactile Throughput

Further, we highlight a general guideline for selecting TXOP and SI values for the best achievable tactile throughput. Considering a $\lambda = 13000$ packets/second, Fig. 3.8 and 3.9 demonstrate a trade-off between achievable latency and tactile throughput. Although higher TXOP and lower SI setting guarantees lower latency, the maximum throughput will be impaired due to the reduced number of admitted Hubs. Therefore, one can use

our derive closed-form expression for TXOP and SI selection in order to achieve the highest achievable throughput while maintaining 0.1 ms latency.

3.4 Accurate Average Uplink Latency Analysis for HCCA

In the preceding sections, we derive the first closed-form expression (Model 1) for the uplink latency of an HCCA-enabled WLAN and critically studied two of the HCCA timing and network parameters, namely the Service Interval (SI) and the Transmission Opportunity (TXOP). The limitation of Model 1 is that the accuracy of analytical results depends on the network load. To ensure a more versatile theoretical model that accurately captures the uplink latency behaviour of HCCA at all practical network loads, in this section, we incorporate (a) a flooring function associated with the reservation interval in our model and (b) use of a Markov Chain model to accurately evaluate the probability of a user device using its full TXOP. Hence, we denote expression derived in this section as Model 2. Moreover, we conduct global sensitivity analyses to achieve a comprehensive understanding of how uplink latency varies with each timing and network parameter. Further, using the insights gained from our theoretical model, we propose a strategic parameter selection algorithm which ensures that the stringent uplink requirements are met. As an initial step towards designing an ultra-low latency MAC layer protocol for the IEEE 802.11 to support emerging services and applications, our work answers the question of what latency level can the IEEE 802.11 technology achieve and which system parameters have the most significant impact on the latency performance of the HCCA protocol.

3.4.1 Derivation of the Accurate Closed-Form Expression

The notations used in this section are listed in Table 3.1. First, we adopt the exact derivation from Lemma 3.1, Lemma 3.2 and (3.17) in Section 3.3.1. Therefore, we have:

$$\bar{R} = \frac{\rho_{\text{total}}X}{2} + \frac{MN(\mu^2 + \nu^2)}{2(NT_{\text{SI}} + B)} + \frac{N\bar{S}^2}{2(NT_{\text{SI}} + B)} + \frac{B^2}{2(NT_{\text{SI}} + B)}, \quad (3.19)$$

$$\begin{aligned} \bar{Y}_1 = & \frac{(iV - \nu)\bar{S}}{T_{SI}} + \frac{\mu(M^2V + 2M\bar{S} - MV - 2i\bar{S})}{2T_{SI}} \\ & + \frac{\rho(M^2V + MV + 2(M - i + 1)\bar{S})}{2} \\ & + \frac{\nu(M^2V + MV - 2M\nu + 2(M - i + 1)\bar{S})}{2T_{SI}}, \end{aligned} \quad (3.20)$$

$$\bar{Q} = \rho_{\text{total}}\bar{W}. \quad (3.21)$$

For an LS data frame that has just arrived at Hub_{*i*}, let N_Q and $T_{\max}$ denote the number of LS data frames waiting ahead and the scheduled transmission time for LS data frames within one TXOP, receptively. The latency term $\bar{Y}_2$ is induced when the data frame of interest cannot be transmitted within the next available TXOP, for example if $N_Q > T_{\max}/X$. In contrast to Section 3.3.1, here we use a different approach to derive $\bar{S}$ in (3.20), $\bar{S}^2$ in (3.19) and $\bar{Y}_2$. We show in Section 3.4.2.1 that this new approach yields a more accurate result of the average HCCA uplink latency. Without loss of generality, we assume that the data frame of interest belong to Hub₁. Let $\bar{S}'$ be the average contention period when Hub₁ uses its full TXOP and let q be the probability of this case. Then, let T' denote the average transmission time when Hub₁ does not use full TXOP, i.e. $T' < T_{\max}$.

Lemma 1. T' can be expressed as:

$$T' = \frac{\rho T_{SI} - q T_{\max}}{1 - q}. \quad (3.22)$$

Proof. From queueing theory, the utility ρ is the percentage of time a server spends on transmitting Hub_{*i*}'s LS data frames, Therefore

$$\rho_{\text{total}} T_{SI} = q T_{\max} + (1 - q) T', \quad (3.23)$$

with further rearrangement, (3.23) yields (3.22). $\square$

By taking (3.22) into (3.11), we have:

$$\bar{S} = T_{SI}(1 - \rho_{\text{total}}) - MV. \quad (3.24)$$

Lemma 3.5. *We can use Markov Chain method to find the accurate value of q .*

Proof. We use a Markov Chain model to describe the number of LS data frames in Hub₁'s queue at the start of Hub₁'s TXOP. When HCCA is implemented, at most $\lfloor T_{\max}/X \rfloor$ LS data frames are allowed to be transmitted. Let i and i' denote the number of data frames at Hub₁'s queue at the start of current and following TXOP, respectively. $U \sim \text{Poisson}(\lambda, T_{\text{SI}})$ represents the number of new data frames that arrive during current service interval. We shall have

$$i' = \begin{cases} U & \text{if } i \leq \lfloor \frac{T_{\max}}{X} \rfloor, \\ U + i - \lfloor \frac{T_{\max}}{X} \rfloor & \text{if } i > \lfloor \frac{T_{\max}}{X} \rfloor. \end{cases} \quad (3.25)$$

Define $F(u)$ as the probability of u new data frames arriving during current service interval, where $F(u) = e^{-\lambda} \frac{\lambda^u}{u!}$. Let $\pi = [\pi_0, \pi_1, \pi_2, \dots, \pi_n, \dots]$, $n = 0, 1, 2, \dots$, represents the steady state probability of n LS data frames waiting in Hub₁'s queue. The corresponding state transition matrix can be formatted as:

$$\mathbf{P} = \left[\begin{array}{cccc} F(0) & F(1) & F(2) & \dots \\ F(0) & F(1) & F(2) & \dots \\ \vdots & \vdots & \vdots & \\ 0 & F(0) & F(1) & F(2) & \dots \\ 0 & 0 & F(0) & F(1) & F(2) & \dots \\ & & & \ddots & \ddots & \ddots \end{array} \right] \left. \vphantom{\begin{array}{cccc} F(0) & F(1) & F(2) & \dots \\ F(0) & F(1) & F(2) & \dots \\ \vdots & \vdots & \vdots & \\ 0 & F(0) & F(1) & F(2) & \dots \\ 0 & 0 & F(0) & F(1) & F(2) & \dots \\ & & & \ddots & \ddots & \ddots \end{array}} \right\} \lfloor \frac{T_{\max}}{X} + 1 \rfloor.$$

π can then be solved by:

$$\begin{cases} \pi \mathbf{P} = \pi, \\ \pi \mathbf{1} = 1. \end{cases}$$

and finally, q can be determined by $q = 1 - \sum_{i=0}^{\lfloor \frac{T_{\max}}{X} \rfloor - 1} \pi_i$. Note that for computational efficiency, we consider the first $m \times m$ entries of $\mathbf{P}$ and the first m terms of π , as long as the dropped terms have insignificant effects on q . Empirically, $m = 3 \lfloor \frac{T_{\max}}{X} \rfloor$ yields a good

result. □

Therefore, by following (3.12) and (3.16), $\bar{S}^2$ and $\bar{S}'$ can be directly expressed by q and T' :

$$\bar{S}^2 = \alpha^2 - 2\alpha(T_{\max} - T')Mq + (T_{\max} - T')^2(Mq(1 - q + Mq)) , \quad (3.26)$$

$$\bar{S}' = T_{\text{SI}} - MV - (M - 1)T' - (M - 1)q(T_{\max} - T') - T_{\max} , \quad (3.27)$$

where $\alpha = T_{\text{SI}} - MV - MT'$.

Now, we derive the expression of $\bar{Y}_2$. By definition, the extra waiting time caused by reservation intervals with an insufficient TXOP allocation is $\lfloor \frac{N_Q}{T_{\max}/X} \rfloor (MV + \bar{S}')$, except when the LS data frame of interest arrives during Hub₁'s data frame transmission. In such a case, before the next available TXOP, the number of data frames ahead of the newly-arrived frame is reduced by $\frac{T_{\max}}{X} - x$, where x is the number of data frames that have already been transmitted during current TXOP. The corresponding waiting time in this case is then determined by $\lfloor \frac{N_Q - (\frac{T_{\max}}{X} - x)}{T_{\max}/X} \rfloor (MV + \bar{S}')$ and the probability of this situation is ρ . The Poisson arrival process ensures that the position of a packet arrival is uniformly distributed within a considered time interval. Therefore, we have the following expression for $\bar{Y}_2$:

$$\begin{aligned} \bar{Y}_2 &= (1 - \rho) \lfloor \frac{N_Q}{T_{\max}/X} \rfloor (MV + \bar{S}') + \rho \sum_{x=1}^{\frac{T_{\max}}{X}} \frac{1}{\frac{T_{\max}}{X}} \lfloor \frac{N_Q - (\frac{T_{\max}}{X} - x)}{T_{\max}/X} \rfloor (MV + \bar{S}') \\ &= (1 - \rho) \lfloor \frac{N_Q}{T_{\max}/X} \rfloor (MV + \bar{S}') + \rho \left(\frac{N_Q + 1}{\frac{T_{\max}}{X}} - 1 \right) (MV + \bar{S}') . \end{aligned} \quad (3.28)$$

Lemma 3.6. *The first term on the right-hand side of (3.28) can be approximated by $(1 - \rho) \lfloor \frac{\lambda \bar{W}}{T_{\max}/X} - (0.5 - \frac{1}{2T_{\max}/X}) \rfloor (MV + \bar{S}')$.*

Proof. Since the arrival of LS data frames at each Hub follows a Poisson process, it is natural to denote $N_Q \sim \text{Poisson}(\lambda \bar{W})$, with $\bar{N}_Q = \lambda \bar{W}$. Here we derive an expression for approximating the value of $\lfloor \frac{N_Q}{T_{\max}/X} \rfloor$.

Define $c = \lfloor \frac{T_{\max}}{X} \rfloor$. Due to the HCCA scheduler, c is a positive integer. Then let $p(x)$ and $q(x)$ be the probability generating function (pgf) of $c \lfloor \frac{N_Q}{c} \rfloor$ and N_Q respectively. We shall

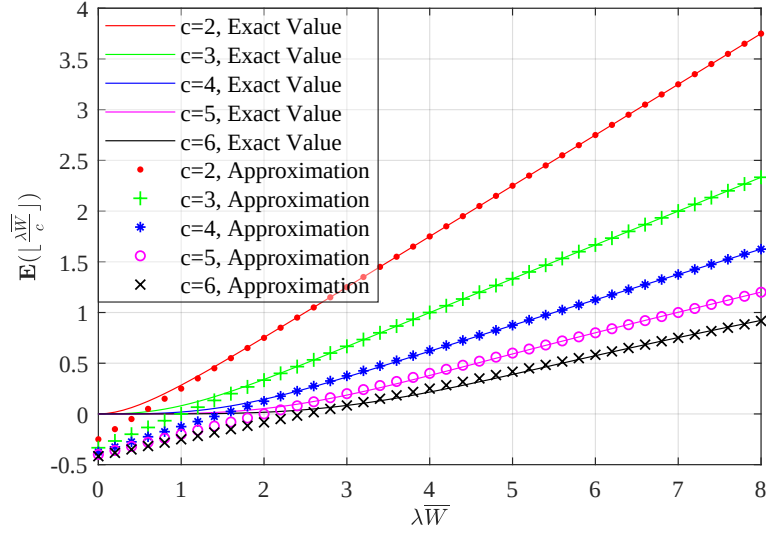


Figure 3.10: Comparison between exact and approximated value of $E(\lfloor \frac{\lambda \bar{W}}{c} \rfloor)$.

have $p(x) = e^{\lambda \bar{W}(x-1)}$ for Poisson distribution of parameter $\lambda \bar{W}$. It can be shown that

$$q(x) = \frac{x(1-x^{-c})}{c} \sum_{j=0}^{c-1} \frac{p(\omega^{-j}x)}{x - \omega^j}, \quad (3.29)$$

where ω is any primitive c^{th} root of unity. To find the mean value of $c \lfloor \frac{N_Q}{c} \rfloor$, we need to find the first derivative of $q(x)$ valued at $x = 1$. It follows:

$$q'(x)|_{x=1} = \lambda \bar{W} - \frac{1}{2}(c-1) + e^{-\lambda \bar{W}} \sum_{j=1}^{c-1} \frac{e^{\lambda \bar{W} \omega^{-j}}}{1 - \omega^j}. \quad (3.30)$$

With the conditions given, the mean value of $\lfloor \frac{N_Q}{c} \rfloor$ can be approximated as:

$$E(\lfloor \frac{\lambda \bar{W}}{c} \rfloor) \approx \frac{\lambda \bar{W}}{c} - \frac{1}{2} + \frac{1}{2c}. \quad (3.31)$$

Fig. 3.10 shows the comparison between exact and approximated values of $E(\lfloor \frac{\lambda \bar{W}}{c} \rfloor)$ with varying c . We can see that (3.31) approximates $E(\lfloor \frac{\lambda \bar{W}}{c} \rfloor)$ well. $\square$

With further rearrangement, $\bar{Y}_2$ can be expressed as:

$$\bar{Y}_2 = [0.5(\rho + 1)(\frac{X}{T_{\max}} - 1) + \frac{\rho \bar{W}}{T_{\max}}](MV + \bar{S}'). \quad (3.32)$$

Finally, $\bar{D}_{\text{HCCA}}$ of Model 2 can be determined by combining (3.1), (3.21) and (3.32):

$$\bar{D}_{\text{HCCA}} = \frac{\bar{R} + \bar{Y}_1 - 0.5(\rho + 1)(1 - \frac{X}{T_{\max}})(MV + \bar{S}')}{1 - \rho_{\text{total}} - \frac{\rho}{T_{\max}}(MV + \bar{S}')} + X, \quad (3.33)$$

where $\bar{R}$, $\bar{Y}_1$ and $\bar{S}'$ are determined using (3.19), (3.20), (3.24), (3.26), (3.27) and the approach described in Lemma 3.5.

3.4.2 Results and Discussions

In this section, we consider a simple IEEE 802.11 HCCA network, and compare the analytical results obtained through $\bar{D}_{\text{HCCA}}$ with results from discrete event simulations. The MAC level simulations are constructed using SimEvents[®] [31], and they model the polling operation of HCCA for 10 seconds and then calculate the average waiting time of data packets. Moreover, we compare the $\bar{D}_{\text{HCCA}}$ values from Model 1 and Model 2. We then conduct sensitivity analyses to better understand the impact of different MAC timing and network parameters on the uplink latency. Finally, based on the sensitivity analyses, we propose a novel strategic parameter selection algorithm to ensure that the average uplink latency constraint is met.

For numerical results, we consider a simple WLAN similar to the one in Fig. 3.1, whereby Hubs and a QAP are connected via HCCA reference design. Specifically, the haptic data transmitted from Hubs to the QAP forms the uplink traffic of our study. The parametric constants used in the numerical analysis conform to [56], and they are listed in Table. 3.4.

3.4.2.1 Numerical Results and Comparison between Models

In the analysis, we assume that the QAP attempts to admit as many Hubs as possible, subject to the constraint given in (2.3). Further, we consider a network with SI set to 80 μs and 100 μs respectively. The LS data frame arrival rate λ is increased from a very low value to a point where the QAP can admit only 1 Hub. The normalised effective throughput representing the transmission rate of LS data is calculated by $\rho_{\text{total}} = M\lambda X$.

Table 3.4: Constants for numerical average latency analysis

Constant	Value
ACK size (MAC header + FCS)	112 bits
Poll size (MAC header + FCS)	160 bits
Beacon size	320 bits
Payload size of LS data	512 bits
PIFS	25 μs
SIFS	10 μs
FCS	32 bits
MAC header	288 bits
PHY header size	192 bits
PHY transmission rate	540 Mb/s
CP time	2X + PIFS
T_{SI}	80 μs , 100 μs

The results of these comparisons are shown in Fig. 3.11(a)-3.15(a). It should be noted that ρ_{total} represents the network throughput driven by data packets, excluding the throughput from control packets and guard intervals such as ACK, CF-Poll, Beacon, contention period, SIFS and PIFS. As such, the considered network is close to saturation when ρ_{total} is approaching 28%. To better explain and understand the performance of HCCA with varying ρ_{total} , two other latency parameters that impact $\bar{D}_{\text{HCCA}}$ —the number of connected Hubs and $\frac{T_{\text{max}}}{X}$ are shown in Fig. 3.11(b)-3.15(b).

It can be observed that for both SI values, the results of $\bar{D}_{\text{HCCA}}$ from Model 2 are in close agreement with simulations. Although Model 1 produces similar results at small ρ_{total} values, their deviations from the simulation results grow as ρ_{total} increases. The deviations from Model 1 result from the fact that it does not consider the flooring function in $\bar{Y}_2$. For instance, if N_Q and $\frac{T_{\text{max}}}{X}$ are 13 and 3 respectively, the number of extra polling cycles a data frame needs to wait is $\lfloor \frac{13}{3} \rfloor = 2$ instead of 2.6. It is observed that all the two models track the simulation results closely when $\frac{T_{\text{max}}}{X} = 1$. This is because the flooring function in $\bar{Y}_2$ does not affect in this case.

As shown in Fig. 3.11(a)-3.15(a), the HCCA reference design exhibits discontinuous latency performance with increasing ρ_{total} . Comparing (a) and (b) of Fig. 3.11-3.15, it is

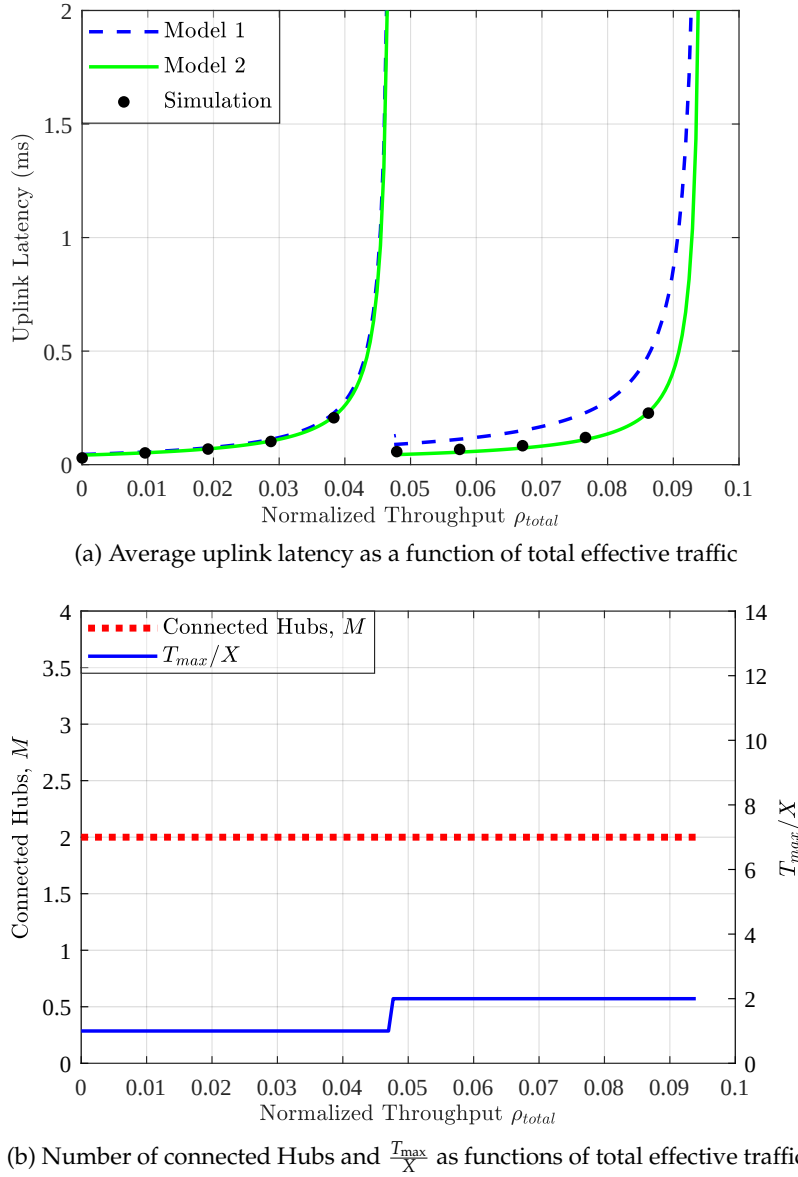
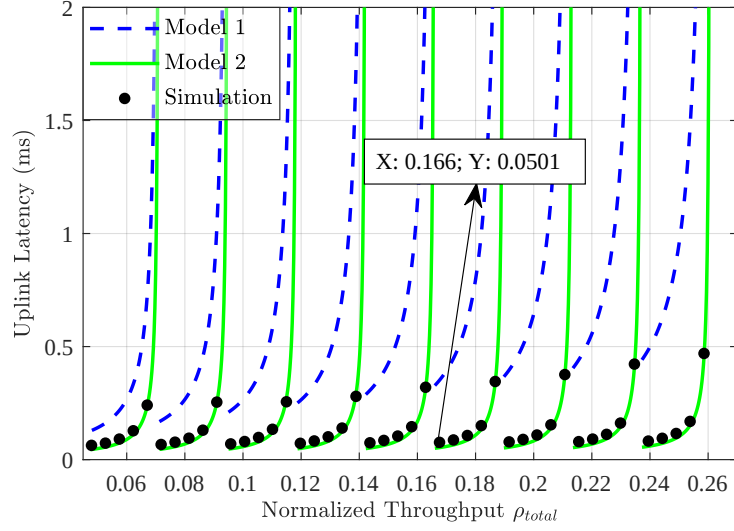


Figure 3.11: Average uplink latency performance and parameter variations when $T_{SI} = 80 \mu s$, $0 < \lambda < 25000$ packet/sec.

observed that the latency discontinuity aligns with the discontinuity of $\frac{T_{max}}{X}$, which is determined by $\lceil T_{SI}\lambda \rceil$. This can be explained by interpreting $T_{SI}\lambda$ and $\frac{T_{max}}{X}$ as the average number of frame arrivals during an SI and the allowed number of frames to be transmitted during an SI, respectively. When $T_{SI}\lambda$ approaches $\frac{T_{max}}{X}$, it reduces the ability of HCCA to clear data frame queue and thus introduces a higher waiting time. In particular, when $T_{SI}\lambda \rightarrow \frac{T_{max}}{X}$, the data frame queue of a particular Hub tends to saturate, and this results



(a) Average uplink latency as a function of total effective traffic

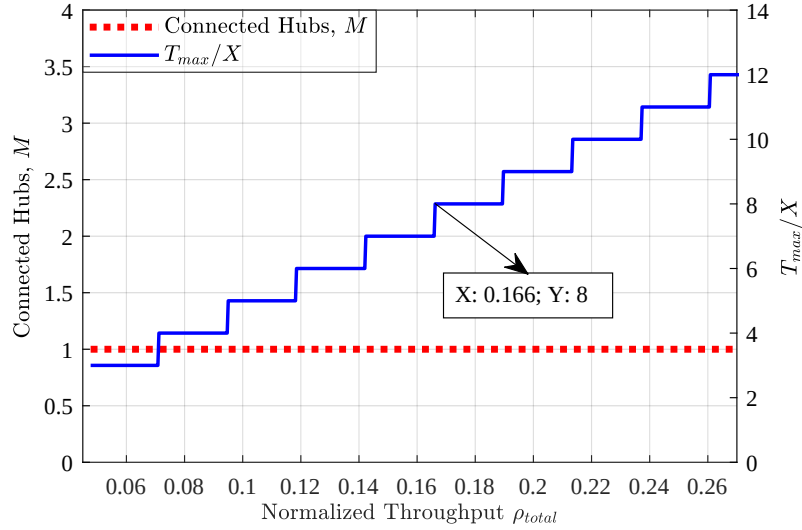
(b) Number of connected Hubs and $\frac{T_{max}}{X}$ as functions of total effective traffic

Figure 3.12: Average uplink latency performance and parameter variations when $T_{SI} = 80 \mu s$, $25000 < \lambda < 137000$ packet/sec.

in drastically increased average latency. Although the rapid increase of latency due to the effect of $T_{SI}\lambda$ and $\frac{T_{max}}{X}$ can easily cause violation of latency restriction, it can be seen that the HCCA reference design still provides acceptable latency value under certain traffic loads. For example, as shown in Fig. 3.12, when $\frac{T_{max}}{X} = 8$, the lowest average latency is 0.0501 ms when ρ_{total} is 0.166. Similarly, in Fig. 3.14 and 3.15, low uplink latencies of 0.0927 ms and 0.0824 ms are achieved by HCCA when $\frac{T_{max}}{X} = 3$ at 0.091 normalised

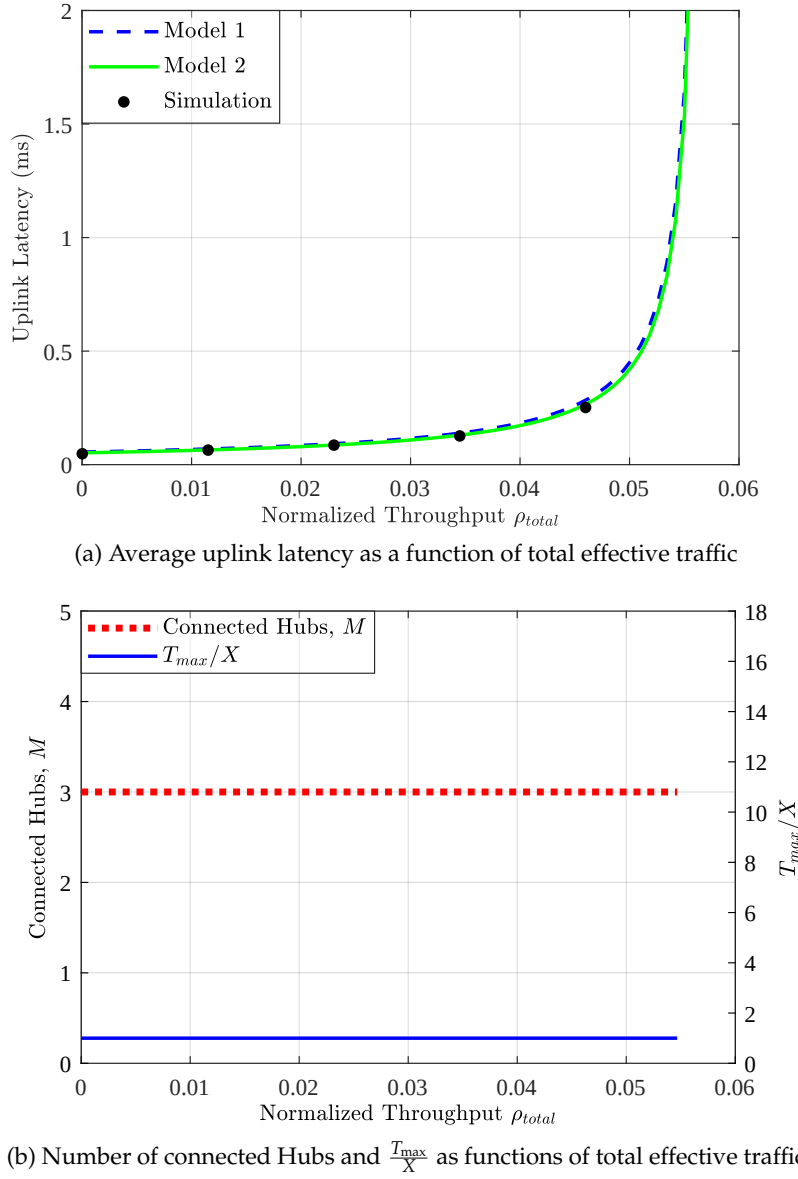
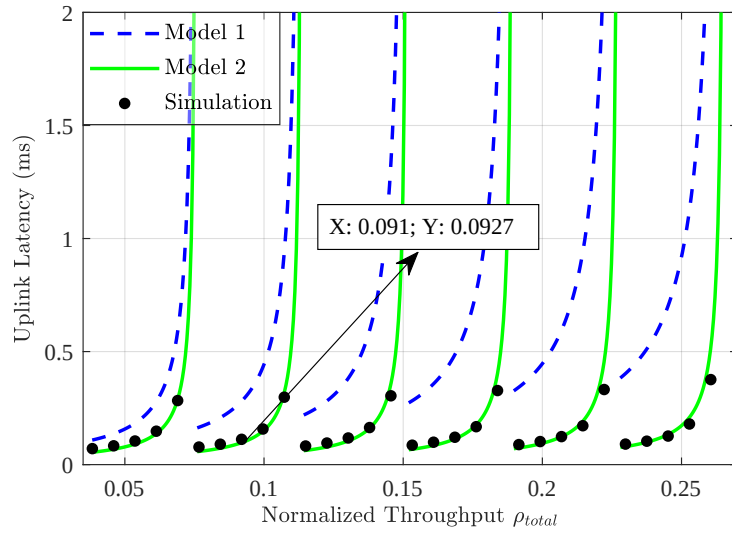


Figure 3.13: Average uplink latency performance and parameter variations when $T_{SI} = 100 \mu s$, $0 < \lambda < 10000$ packet/sec.

throughput and when $\frac{T_{max}}{X} = 11$ at 0.195 normalised throughput, respectively. Therefore, it is possible to reduce the latency of LS data frames to acceptable value by fine-tuning the timing and network parameters of HCCA. It is also observed that the achievable throughput under low latency requirement is relatively low, especially for a low SI value. This is because to maintain an acceptable average latency, the SI value needs to be sufficiently small to ensure a short polling cycle and hence the number of supported Hubs is limited.



(a) Average uplink latency as a function of total effective traffic

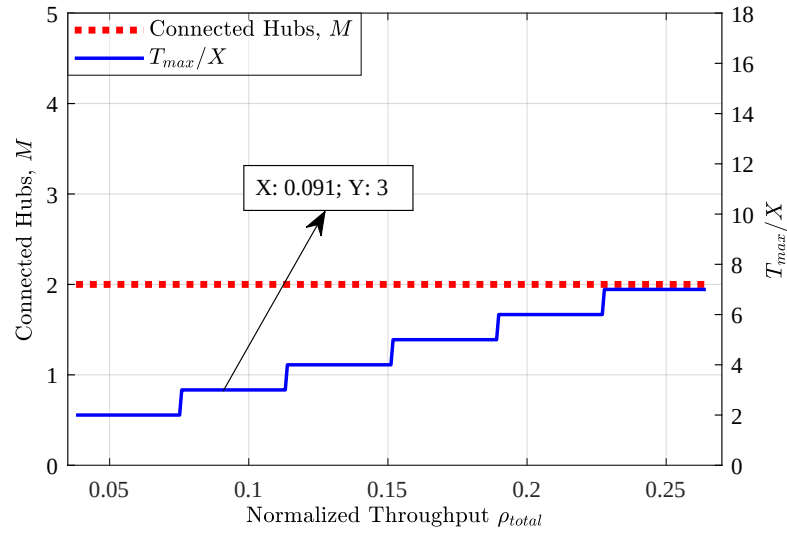
(b) Number of connected Hubs and $\frac{T_{max}}{X}$ as functions of total effective traffic

Figure 3.14: Average uplink latency performance and parameter variations when $T_{SI} = 100 \mu s$, $10000 < \lambda < 70000$ packet/sec.

This trade-off between latency performance and throughput requires a rethink of the existing network structure. To enhance system scalability, vertical network slicing could be used to provide differentiated QoS.

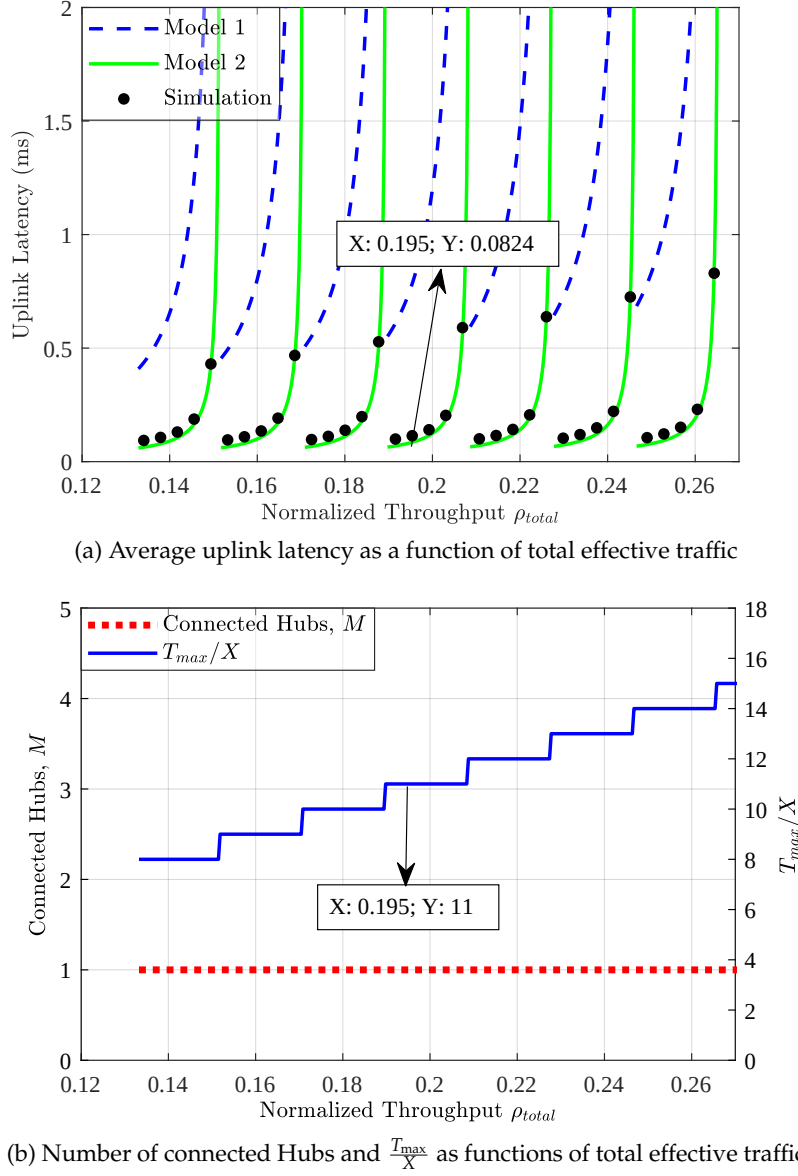


Figure 3.15: Average uplink latency performance and parameter variations when $T_{SI} = 100 \mu s$, $70000 < \lambda < 200000$ packet/sec.

3.4.2.2 Sensitivity Analyses

Motivated by the need to adjust HCCA parameters for meeting the uplink latency constraint, we study the importance of different parameters on the latency. This can only be achieved by quantitatively understanding the parameters' impact. Here we use a variance-based method and Monte Carlo simulations to conduct sensitivity analyses of

Table 3.5: Total sobol indices of eight timing parameters.

Timing and network parameters	Total sobol index		
	240 Mbps	540 Mbps	1040 Mbps
ACK-frame size	1.104×10^{-3}	1.879×10^{-4}	3.243×10^{-5}
Poll-frame size	9.401×10^{-4}	1.589×10^{-4}	2.784×10^{-5}
Beacon-frame size	1.482×10^{-7}	2.936×10^{-8}	8.363×10^{-9}
SIFS period	2.383×10^{-2}	1.761×10^{-2}	9.754×10^{-3}
T_{SI}	8.830×10^{-1}	9.286×10^{-1}	9.306×10^{-1}
$\frac{T_{max}}{X}$	6.319×10^{-1}	5.960×10^{-1}	6.062×10^{-1}
λ	8.374×10^{-1}	8.377×10^{-1}	8.327×10^{-1}
LS data payload	2.226×10^{-2}	4.442×10^{-3}	3.236×10^{-3}

our analytical model with eight different timing and network parameters. The parameters under consideration are ACK, Poll and Beacon frame sizes, short interframe space (SIFS) period, T_{SI} , $\frac{T_{max}}{X}$, LS data arrival rate and LS data payload size. It should be noted that the first four parameters relate to IEEE 802.11 MAC Standards, while the fifth and sixth relate to actual HCCA implementation and the last two relate to LS traffic characteristics. We use the values in Table. 3.4 with $T_{SI} = 100 \mu s$ and $\frac{T_{max}}{X} = 6$ as the base case. The range of each parameter is from 50% to 150% of the corresponding base value. These timing and network parameters are correlated because of the constraints set by (2.3) and $T_{SI}\lambda < \frac{T_{max}}{X}$. Hence, we need to apply the general equation for total sobol indices, as specified by 5.4 in [136]. It is explained in [137] that this index represents the importance of each parameter on the variation of subject value, excluding the effect of parameter correlation. Here, these sensitivity indices are calculated against three different physical transmission rates: 240 Mbps, 540 Mbps and 1040 Mbps, respectively. The first and second physical transmission rates correspond to high throughput (HT) PHY layer settings, and the third physical transmission rate corresponds to very high throughput (VHT) settings [56]. The results are shown in Table. 3.5.

From Table. 3.5, we can clearly see that T_{SI} , $\frac{T_{max}}{X}$ and λ have greater impact on the uplink latency performance than the others, with their total sobol indices being around 8.8×10^{-1} , 6.3×10^{-1} and 8.3×10^{-1} , respectively. This implies that fine-tuning T_{SI} and

Algorithm 1 Strategic Parameter Selection algorithm

```

1: start HCCA session, use default value  $T_{SI,start}$  for  $T_{SI}$ 
2: while iteration  $\leq$  maximum iteration value and number of admitted Hubs is sufficient
   do
3:   calculate  $\bar{D}_{HCCA}$  under current  $T_{SI}$  and TXOP setting
4:   if  $\bar{D}_{HCCA} \leq$  latency limit then
5:     break
6:   else
7:     update  $T_{SI}$  as  $\sum TXOP + T_{CP}$ 
8:     if current  $T_{SI} \geq$  previous  $T_{SI}$  then
9:       increase TXOP by 1 LS data frame transmission time
10:      if number of admitted Hubs is sufficient then
11:        apply (2.3) to terminate the HCCA connection of low-priority Hubs
12:      else
13:        update  $T_{SI}$  as  $\sum TXOP + T_{CP}$ 
14:      end if
15:    end if
16:  end if
17: end while
18: use the final setting for HCCA connection

```

$\frac{T_{max}}{X}$, and controlling the frame arrival rate through haptic data compression techniques are the most effective ways of improving the HCCA latency performance. Moreover, it is observed that the total sobol index of T_{SI} increases with the physical transmission rate, revealing that the strategic selection of T_{SI} remains one of the most effective ways of meeting latency constraint, even with the possibly increased physical transmission rates in future IEEE 802.11 Standards.

Comparing the results in Table. 3.5, we see a trend of increasing importance of SIFS period and LS data payload size on latency with the reduced data rate. In particular, their total sobol indices become 2.383×10^{-2} and 2.226×10^{-2} respectively, when the data rate is 240 Mbps. This implies that designing new values for SIFS in future standards and developing effective haptic data encoding techniques are crucial for uplink latency reduction, especially when a high physical transmission rate is unavailable.

3.4.2.3 Strategic Parameter Selection Algorithm

In this section, we propose a novel strategic parameter selection algorithm which can

Table 3.6: Description of examples and resulting uplink latencies (by simulation)

Cases	Description	Resulting uplink latency (ms)
I	$T_{SI,start} = 110 \mu s, \lambda = 40000$ packet/sec	0.0643
II	$T_{SI,start} = 110 \mu s, \lambda = 60000$ packet/sec	0.0811
III	$T_{SI,start} = 110 \mu s, \lambda = 135000$ packet/sec	0.0526
IV	$T_{SI,start} = 95 \mu s, \lambda = 40000$ packet/sec	0.0590
V	$T_{SI,start} = 95 \mu s, \lambda = 60000$ packet/sec	0.0718
VI	$T_{SI,start} = 95 \mu s, \lambda = 135000$ packet/sec	0.0527
VII	$T_{SI,start} = 80 \mu s, \lambda = 40000$ packet/sec	0.0751
VIII	$T_{SI,start} = 80 \mu s, \lambda = 60000$ packet/sec	0.0434
IX	$T_{SI,start} = 80 \mu s, \lambda = 135000$ packet/sec	0.0607

effectively reduce the uplink latency of HCCA protocol by incorporating the analytical model of HCCA and the insights gained from the sensitivity analyses. Our proposed algorithm, which is presented in Algorithm 1, adjusts the value of TXOP or T_{SI} in each iteration to achieve a lower uplink latency. The basic operation of our algorithm is explained as follows. When a latency reduction is necessary, the proposed algorithm first reduces T_{SI} by shortening the CP to a minimal value (Step 7). If such an approach does not provide a satisfying uplink latency, it then increases the term $\frac{T_{max}}{X}$ by 1 (Step 9). The strategic parameter selection algorithm adjusts HCCA settings until it finds a suitable pair of TXOP and T_{SI} such that a desired theoretical average uplink latency is obtained. Examples illustrating the behaviour of the proposed algorithm are presented in Fig. 3.16. We choose 110, 95 and 80 μs as three random starting values of T_{SI} , and 40000, 60000 and 135000 packet/sec as three example packet arrival rates of a single Hub. It is noted that these arrival rates are selected such that at least two example cases require latency reduction. A detailed description of these testing cases is shown in Table. 3.6.

For Case VII in Fig. 3.16(c), the proposed algorithm uses the closed-form uplink latency expression (3.33) and detects a satisfying theoretical uplink latency (≈ 0.06 ms) by the starting HCCA setting ($T_{SI} = 80 \mu s$), therefore an improvement of HCCA setting is not required. On the other hand, the strategic parameter selection algorithm reduces the theoretical uplink latencies in Case I, II, III, VI, VIII, IX to desired values by decreas-

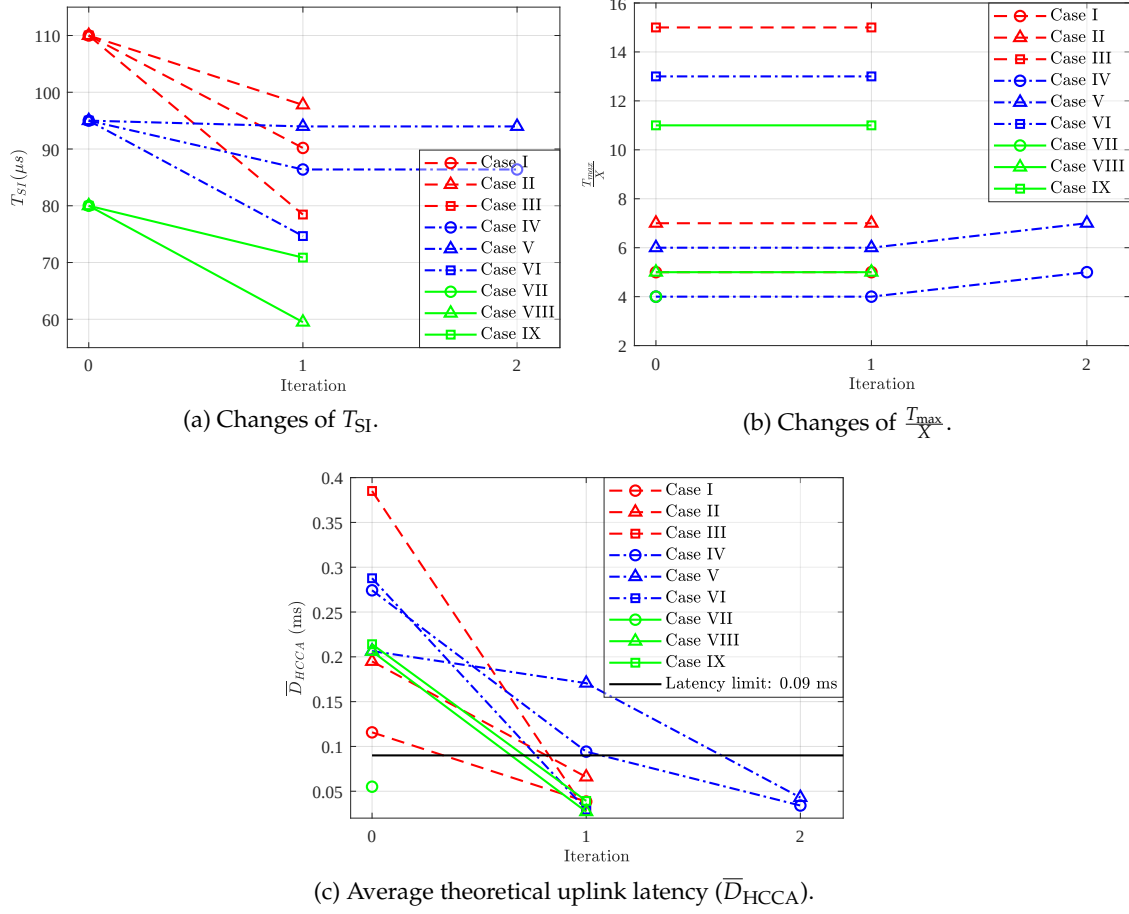


Figure 3.16: Changes of HCCA settings for improved uplink latency by the proposed algorithm.

ing the T_{SI} during the first iteration, as shown in Fig. 3.16(c) and Fig. 3.16(a). For Case IV and V, however, the proposed algorithm takes more steps to achieve latencies of less than 0.09 ms. This is because according to the theoretical calculation by (3.33), effective latency reduction shall be achieved by both reducing T_{SI} (iteration 1 in Fig. 3.16(a)) and increasing $\frac{T_{max}}{X}$ (iteration 2 in Fig. 3.16(b)). To demonstrate the effectiveness of the proposed algorithm, we simulate and calculate the MAC level average uplink latency by using the final HCCA settings, and the results are presented in Table. 3.6. It can be seen that the proposed algorithm achieves successful latency reduction by tuning the HCCA timing and network parameters, such that the shortened uplink latencies are below 0.1 ms.

3.5 Conclusions

In this chapter, we have investigated the feasibility of applying the HCCA MAC protocol defined in IEEE 802.11 for latency-sensitive (LS) applications. Specifically, we considered the uplink transmission model in MAC level where multiple Hubs transmit data packets wirelessly to an access point. By leveraging the M/G/1 queueing model, we analysed the uplink latency of HCCA and showed our proposed analytical model to be more accurate than previous works in terms of evaluating average latency under high network load conditions. We used this analytical model to motivate the pertinence of network and HCCA parameter adjustment of the current protocol in order to meet stringent latency requirements. Further, we presented the first global sensitivity analyses to study the impact of eight timing and network parameters on the latency performance. Numerical results demonstrate that under high physical transmission rate, the most effective ways to achieve latency reduction are tuning T_{SI} and $\frac{T_{max}}{X}$ values, and controlling data frame arrival rate in application's perspective. Additionally, when the physical transmission rate is low, latency reduction can also be achieved by shortening the SIFS period and the payload of LS data frames. Furthermore, with the insights obtained from our sensitivity analyses, we developed an efficient iterative algorithm called strategic parameter selection that adjusts the values of T_{SI} and $\frac{T_{max}}{X}$ according to the latency performance of HCCA. Numerical results have shown that HCCA can meet the 0.1 ms latency requirement with proper parameters selected by the proposed algorithm.

Indeed, there exists a trade-off between latency performance and overall achievable throughput of HCCA. One possible solution is to utilise the unused CP in HCCA protocol to serve the most urgent applications such as monitoring of vital health signals. This approach can improve both uplink latency and network throughput, while an investigation of the impact of CP length on HCCA performance could be pursued in future work by extending the results of this chapter. Another approach to enhanced network throughput is to apply network slicing technology to achieve QoS differentiation for different network traffics and applications. The integrated design of network slicing and MAC protocol could therefore form another interesting research direction.

Another equally important wireless communication technology—mobile cellular net-

work has attracted users' attention due to unprecedented mobile applications, and this suggests another solution to the wireless access network for the Tactile Internet and other emerging services due to its vast availability and connectivity. However, the current 4G network implementing Dynamic Scheduling (DS) induces unacceptable uplink latency, which necessitates research efforts into a feasible MAC layer protocol for ultra-low latency uplink transmissions. In the next chapter, we show that Semi-Persistent Scheduling (SPS) scheme of LTE, which was originally developed for VoIP application, could be the potential candidate of uplink transmission protocol for 4G and beyond, in order to meet latency requirement of the Tactile Internet.

Predictive Semi-Persistent Scheduling Scheme of LTE for Latency-Critical Services

4.1 Introduction

Cellular networks have successfully connected people around the globe by providing quality services for data traffic such as video streaming and voice data. The next natural leap of communications technologies is to enable human haptic data transmission, and this forms the fundamental concept of the Tactile Internet [12]. A typical application of the Tactile Internet is a TelePresence and TeleAction (TPTA) system comprising an operator (OP) and a teleoperator (TOP), where the OP transmits haptic data to the TOP and the TOP sends feedback by interacting with the remote environment. One of the most challenging system requirements of the Tactile Internet and other latency-critical services such as automation motion control and high voltage distribution, is to transmit information with unnoticeable latencies ranging from 0.5 to 5 ms [14].

The content-oriented LTE (4G) cellular networks have been playing a vital role in providing downlink wireless broadband services with high data rates and high capacity. However, LTE may fail to provide services for emerging haptic-tactile applications due to its non-negligible uplink latency. A typical LTE network consists of an evolved NodeB (eNB) and user equipment (UE). Given sufficient channel capacity, the downlink latency of an LTE network is unnoticeable because the eNB gains instantaneous knowledge about the amount of downlink data to be delivered to individual UEs. While for the conventional uplink transmission scheme—Dynamic Scheduling (DS) scheme, each UE needs to send a scheduling request to the eNB and wait for a scheduling grant before transmitting

its uplink data packets. Hence, the uplink latency is significantly larger than the downlink latency and becomes a bottleneck for applying cellular communications networks for latency-critical services. Succeeding LTE, the fifth generation of cellular mobile communications (5G) has attracted significant attention. One of the use scenarios of 5G is Ultra-Reliable Low-Latency Communications (URLLC), and this makes 5G a potential technology enabler of latency-critical services, including the Tactile Internet [26]. However, the first deployment stage of 5G–Non-Standalone (NSA) from 3GPP states that the first version of 5G system is constructed upon existing LTE infrastructure and this means LTE will remain as an important complementary technology in the future 5G world for a very long time. Therefore, an efficient uplink transmission scheme which alleviates the uplink latency bottleneck of LTE also benefits 5G communications networks in providing services for latency-critical applications. In other words, our work is not only applicable to current LTE networks but also goes in the direction of current advances in 5G network design.

The main contribution of this chapter is a novel predictive Semi-Persistent Scheduling (SPS) scheme that reduces the uplink latency of LTE and 5G networks to satisfy the latency requirements of the emerging haptic communications. The proposed SPS scheme leverages its periodic transmission feature and schedules radio resources in a proactive approach, without requiring a grant for every transmission. This transmission scheme does not consume resources in Physical Downlink Control Channel (PDCCH) of an LTE cell, and this effectively increases the maximum number of allowable UE connections. Specifically, two prediction models have been proposed for the SPS scheme, which determine the number of allocated radio resources based on past observations of newly arrived data and estimation of current UE's buffer occupancy. Moreover, we introduce the first Feasible Resource Allocator (FRA) which allows the practical deployment of the predictive SPS scheduler, such that UEs governed by the predictive SPS scheme can coexist with UEs governed by other uplink transmission schemes. Our proposed SPS scheme is verified by extensive system-level simulations with experimental haptic-tactile traffic traces. Results show that with the help of prediction models, the proposed SPS scheme yields significant improvement on uplink MAC layer latencies, as compared with the DS

scheme and existing SPS schemes. In particular, the proposed predictive SPS scheme reduces the uplink latency of an LTE network to sub-3 ms for all considered traffic loads, when 1 ms is selected as the SPS periodicity. It is worth mentioning that our proposed predictive SPS scheme does not introduce any new frames or signalling process in the PUCCH or PUSCH. This means our proposed scheme is 3GPP-compliant and modification purely resides in the scheduling process in the MAC layer.

The rest of this chapter is organised as follows. In Section 4.2, we discuss existing literature that is closely relevant to our work in detail. We also point out the key differences between our proposed SPS scheme and existing ones. Section 4.3 elaborates the detailed process of FRA and the proposed predictive schedulers. Simulation results are presented in Section 4.4. Finally, we conclude this chapter with a summary of our findings in Section 4.5.

4.2 Related Work

As discussed in section 2.2.2, SPS scheme was firstly designed for quality services of VoIP [100], which exhibits constant bitrate (CBR) feature. While for variable bitrate (VBR) data traffic, the fixed amount of scheduled radio resources in traditional SPS scheme could result in over/under scheduling issue. Therefore, various attempts have been proposed to introduce flexibility into the SPS scheme, such that it can adapt to data traffic fluctuations and made efficient scheduling decisions.

Jin *et al.* have proposed a PHY layer method that improves SPS performance by selecting the optimal Modulation and Coding Scheme (MCS) for uplink transmission in a time-correlated Nakagami-m fading channel [138]. According to specifications of the SPS scheme, selection of deployed MCS is fixed during an SPS session. A proper MCS selection is hence essential for long term resource scheduling efficiency, especially when packets are corrupted and HARQ is deployed to restore corrupted packets. The optimal MCS transmission rate is defined as:

$$R_{optimal} = \min_{R \in MCS_{set}} \{B(R) \cdot N_{mean}(R) + B_{head}\}$$

$$\cdot [N_{mean}(R) - 1] P_{outage}(R) \leq Q_{loss}, \lambda, m, \rho, N_{max}\}, \quad (4.1)$$

where R , B , N_{mean} and B_{head} denote transmission rate, allocated resource size for packet transmission, average number of transmission attempts and allocated resource size for signalling overhead, respectively. While for SPS configuration settings, P_{outage} , Q_{loss} , λ , m , ρ and N_{max} denote actual and required packet loss probability, inverse of SNR, diversity and time-correlation factor of Nakagami-m channel, respectively. In essence, the optimal MCS transmission rate is selected such that minimum resource is consumed while packet loss requirement is met. The difficulty of this method lies in calculating the exact value of $N_{mean}(R)$ and $P_{outage}(R)$. Authors of [138] approximated these two parameters and then applied the proposed MCS selection method for a practical VoIP service in an IEEE 802.16e network. Numerical results show that this method effectively reduces radio resources used for HARQ retransmissions while theoretical packet loss rate requirement is met.

Avocanh *et al.* improve the SPS scheme with a MAC Layer method—an SPS scheme with Provisioning (SPS-P) that proactively allocates resources to a number of next frames by prediction [139]. The prediction model used by the SPS-P is based on GBAR (Gamma-Beta Auto Regressive) process, which is specialised for estimating short-term fluctuation of Videotelephony data traffic. Simulation results show that the percentages of packets experiencing a delay of less than 50 ms for SPS-P and DS schemes are 90% and 50%, respectively. Meanwhile, the achievable packet loss rate of the SPS-P is almost 7 times smaller than that of the DS scheme. Although the SPS-P scheme outperforms the conventional DS scheme in serving Videotelephony traffic, it might not be applicable for new emerging applications. This is because Videotelephony traffic consists of large, variable-sized packets with constant inter-arrival time, while emerging services such as M2M communications and the Tactile Internet mainly contain short and constant sized packets with variable inter-arrival time. Another drawback of the SPS-P is that the number of future resource requirement predictions is limited— N_f making it unsuitable for long-term traffic prediction. This, in turn, would require the SPS-P to initiate an entirely new SPS-P session after N_f times of uplink transmissions, which will inevitably lead to

increased latency for the overall Videotelephony session.

For Machine-to-Machine (M2M) communications, authors in [140, 141] designed an adaptive SPS scheme which utilises the latest buffer occupancy information contained in the BSR filed to adjust the resource allocations for the next uplink SPS transmissions. Moreover, the proposed adaptive SPS scheme incorporates a channel condition measuring window, such that MCS selections that match current channel measurement can be updated, which provides certain diversity gains similar to the DS scheme. Simulation results of the proposed adaptive SPS scheme demonstrate that 85% packets have met delay budget when 50 gateways are deployed, roughly 3 times higher than the DS scheme. Moreover, the adaptive SPS scheme is shown to effectively reduce the uplink packet drop rate by almost 50% compared with the DS scheme. The proposed adaptive SPS scheme, however, does not elaborate how positions of scheduled resources are determined, making it infeasible for practical implementation. Moreover, the calculation of resource allocation does not consider new incoming uplink data after the BSR transmission. Lastly, the periodicity of adaptive SPS is over 10 ms, and hence it cannot meet the ultra-low latency requirement.

Li *et al.* [142] have improved the latency performance of the SPS scheme for industrial wireless networks in an interesting approach. It is known that in an industrial automation process, the triggering of a Machine-Type Communications (MTC) device is likely to increase its adjacent devices, which in turn generate data in quick succession. A DPre (predictive pre-allocation) framework is therefore developed such that resources would be allocated to candidate devices correctly. A highlight of the DPre framework is that scheduling entity in the base station is trained using supervised and sequential learning and a successful prediction ratio of 70% is achieved. However, the design principle of the DPre framework does not apply to other emerging services, as there is no correlation of data generation among massive neighbouring devices in applications like VR and telerehabilitation.

In [143], a soft resource reservation SPS scheme was proposed for teleoperation over mobile networks. The reservation strategy allows a UE to selectively send an SR to an eNB about whether new incoming data packets have been received at the UE buffer. Via

this procedure, if the UE did not have any uplink data to send, the eNB will reschedule the reserved resources for other communications within the cell. Although this soft resource reservation SPS scheme does not introduce any drawback as compared to the conventional DS scheme, it may cause worse latency performance compared to the traditional SPS scheme. First, unlike the traditional SPS scheme, where no Physical Uplink Control Channel (PUCCH) resource is consumed after SPS establishment, the soft resource reservation SPS scheme must send uplink SR messages each time when transmission opportunity is needed. This may have a significant impact on accessible mobile connections of other communications within the cell. Second, as explained in Section 2.2.2, an eNB needs to schedule resources for uplink transmissions 4 ms in advance and the processing time of PUCCH message is 3 ms, hence there will be at least 7 ms gap after the opportunity of an SR transmission, which introduces the possibility of a UE being unable to transmit even though it has new incoming uplink data. Results have shown that the resource reservation SPS scheme achieves an average uplink latency of 10 ms, outperforming the conventional DS scheme by 5 ms. However, no comparison was made between the soft resource reservation SPS scheme and its preceding counterpart—the traditional SPS.

4.3 Predictive SPS Scheme

In our proposed predictive SPS scheme, identical predictive schedulers that take in input information and generate resource allocation decisions are deployed in both MAC layers of eNB and corresponding UEs. As such, the proposed SPS scheme preserves the advantage of zero control-signalling overhead because of identical resource allocation decisions. More importantly, as the number of assigned RBPs is expected to vary during each SPS transmission, a Feasible RBP Allocator (FRA) that is more complex than the allocator of ordinary SPS scheme needs to be designed for supporting the predictive SPS functionality.

The key contribution of our work can be summarised as follows:

- A novel predictive SPS scheme is proposed to reduce the uplink latency of LTE and

5G networks. Specifically, the proposed SPS scheduler determines the number of assigned radio resources by extrapolating past observations of newly arrived data and estimating current UEs buffer occupancy.

- We introduce the first Feasible Resource Allocator (FRA) which allows practical deployment of the predictive SPS scheduler, such that UEs governed by the predictive SPS scheme can coexist with UEs governed by other transmission schemes.
- Different to [139], our work focuses on emerging services that are in control-centric networks, where the data packets are expected to be short, constant sized with variable inter-arrival time.
- Different to [143], the proposed scheme leverages the periodic transmission feature and schedules radio resources in a proactive approach, without requiring a grant for every transmission. Therefore, this scheme does not consume resources in PDCCH of an LTE cell, and this effectively increases the maximum accessible connection within the mobile network cell.
- Different to [143] where the soft reservation scheme schedules resources in an all-or-none approach, our predictive SPS scheme allows variable resource allocation, which further enhances the flexibility of uplink SPS scheme.
- We verify the performance of the predictive SPS scheme in a more realistic ultra-low latency environment by applying the Deadband Principle (described in Section 2.4) to raw haptic data traffic. Extensive simulation results show that the proposed SPS scheme achieves significant improvement in terms of uplink latency and resource allocation efficiency.

To better understand the procedure of the predictive SPS scheme, we illustrate the scheduling process of the proposed predictive SPS scheme in Fig. 4.1, assuming there is a UE requiring uplink transmission to an eNB via predictive SPS scheme with a 3 ms periodicity. The major operations of each denoted steps in Fig. 4.1 is further explained as follows:

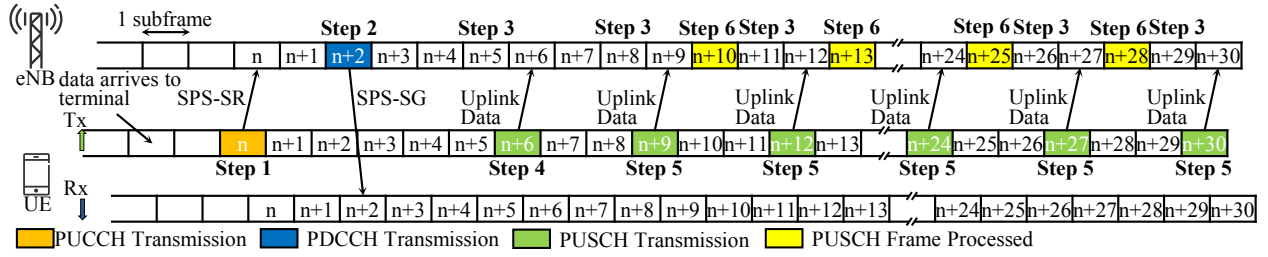


Figure 4.1: Timing diagram of the proposed predictive SPS scheme (3 ms SPS periodicity).

- Step 1:** The UE sends an SPS-SR to the eNB, asking for an SPS connection for uplink transmission.
- Step 2:** The eNB attempts to determine the initial number of scheduled RBPs (N_1) for current SPS session, the maximum number of RBPs ($N_{\max}$) that could be used by this UE, maximum number of SPS transmission (T_{SPS}) and selected modulation and coding scheme (MCS) for the UE. The FRA at the eNB side decides whether the connection request is granted and determines actual allocations of assigned RBPs. An SPS-SG containing all necessary information will be transmitted to the UE if an SPS connection is established.
- Step 3:** By deploying the predictive SPS scheduler, the eNB reserves a certain number of RBPs (N_i) for i^{th} SPS transmission.
- Step 4:** The UE receives the SPS-SG and extracts the encapsulated SPS connection information. The UE then transmits an initial uplink data frame using the scheduled N_1 RBPs.
- Step 5:** By deploying the predictive SPS scheduler, the UE determines the number of RBPs (N_i) for i^{th} SPS transmission. The transmitted LTE frame always contains haptic data packets and a BSR.
- Step 6:** For each uplink transmission, the SPS scheduler keeps a history record of N_i , size of the transmitted data (D_i) and UE's buffer length (B_i).

When the number of SPS transmission reaches T_{SPS} , the UE deletes the configured SPS grant, and the eNB frees the reserved RBPs. It is worth noticing that similar to traditional

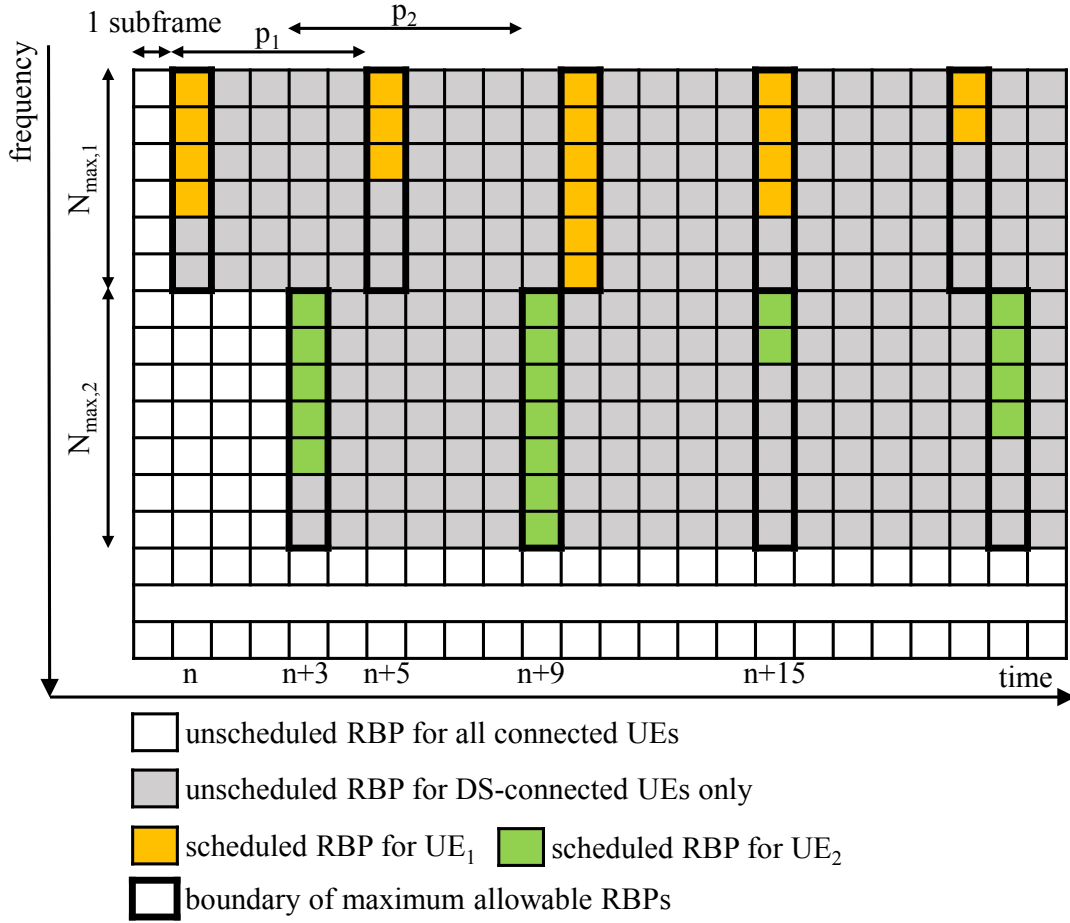


Figure 4.2: RBP allocation diagram of predictive SPS scheme.

SPS, the predictive SPS scheme allows a reset of SPS configuration in case there is a need to adjust the MCS selection when channel quality changes. It is important to note that depending on selected SPS periodicity, **Step 3** can happen either before or after **Step 4**.

4.3.1 Feasible RBP Allocator

The FRA is designed such that dynamic RBP allocations do not cause resource collision among connected UEs. Let $N_{\max,j}$ and $N_{n,j}$ denote the maximum number of scheduled RBPs and the number of scheduled RBPs of UE_j at subframe n , respectively. The functionality of FRA within eNB's MAC layer is shown in the algorithm described in Algorithm 2. For a configured SPS connection between an eNB and UE_j, let BR_j , p_j and c_j represent the

Algorithm 2 Scheduling process at eNB for subframe n

```
1: for all UEs that are connected through SPS do
2:   if UE $j$ 's periodic counter reaches 0 then
3:     determine  $N_{n,j}$  using predictive SPS scheduler
4:     determine the position of scheduled RBPs using FRA
5:   end if
6: end for
7: while detection of new SR do
8:   for all new SPS-SR do
9:     determine  $N_{\max,j}$  based on traffic characteristics
10:    if  $N_{\max,j} \leq$  maximum available contiguous RBPs then
11:      determine the position of  $N_{\max,j}$  RBPs using FRA, allocate them to UE $j$  and
      mark these RBPs unavailable to new SPS-SR
12:      determine  $N_{n,j}$  using predictive SPS scheduler
13:      determine the position of scheduled RBPs using FRA and mark these RBPs
      unavailable to other UEs
14:      set a periodic counter that corresponds to  $p_j$  and record configuration of this
      successful SPS connection
15:    else
16:      SPS request is unsuccessful
17:    end if
18:    delete current SPS-SR
19:  end for
20:  for all new DS-SR do
21:    go through conventional DS procedure
22:  end for
23: end while
24: for all UEs that are connected through DS do
25:   go through conventional DS procedure using available RBPs
26: end for
```

bitrate, SPS periodicity and bytes per RBP, respectively. Consider a scenario where UE₁ requests an SPS connection with a p_1 periodicity and the operation of FRA is explained with Algorithm 2 and RBP allocation diagram in Fig. 4.2. At subframe n , the eNB has determined $N_{\max,1}$ based on:

$$N_{\max,j} = \lceil a_j \frac{BR_j \times p_j}{c_j} \rceil \quad (4.2)$$

where a_j is a scaling factor that is greater than 1 (line 9). In the case that the amount of remaining contiguous RBPs is less than $N_{\max,1}$, the SPS request is denied by the eNB (line 16). This situation is denoted as “Limitation of UE Accessibility”. Otherwise, the eNB selects and assigns $N_{\max,1}$ contiguous RBPs to UE₁ and marks these RBPs unavailable to other SPS-connected UEs, as described in line 11. Meanwhile, the FRA determines the position of $N_{n,1}$ RBPs for the initial uplink transmission of UE₁ and the remaining $(N_{\max,1} - N_{n,1})$ RBPs become available to DS-connected UEs at subframe n (line 13). At subframe $n + 3$ in Fig. 4.2, UE₂ has successfully requested an SPS connection with a p_2 periodicity, and it goes through a similar SPS-establishment process. At subframe $n + 5$ and $n + 9$, the FRA determines the positions of $N_{n+5,1}$ and $N_{n+9,2}$ RBPs within the range of assigned $N_{\max,1}$ and $N_{\max,2}$ RBPs for the second uplink SPS transmission of UE₁ and UE₂, respectively (line 2–5). Let us pay attention to subframe $n + 15$, it can be seen that with FRA, the varying RBP allocations of our proposed predictive SPS scheme do not cause resource collision, even when multiple SPS transmissions occur at the same time. Furthermore, the unused $(N_{\max,j} - N_{n,j})$ RBPs are safe to be scheduled for DS-connected UEs. Therefore, the proposed SPS scheme can flexibly allocate radio resources to SPS-connected UEs and DS-connected UEs and achieve improved RBP utilisation efficiency, as compared with the ordinary SPS scheme.

4.3.2 Predictive SPS Schedulers

The predictive SPS scheduler is designed such that it predicts the UE's MAC buffer length and determines an appropriate number of RBPs for the next SPS transmission. In essence, an allocation decision $\hat{G}_n$ for n^{th} SPS transmission is determined based on available history observations, including transmitted data size $\{D_{n-k-M}, \dots, D_{n-k}\}$, re-

Table 4.1: Prediction target and available observations to SPS scheduler for n^{th} SPS transmission

Sequence	New	Transmitted Data	BSR	Granted Data Size
$n - k - M$		D_{n-k-M}		
$n - k - M + 1$	$T_{n-k-M+1}$	$D_{n-k-M+1}$	$B_{n-k-M+1}$	
$\dots$	$\dots$	$\dots$	$\dots$	
$n - k - 2$	T_{n-k-2}	D_{n-k-2}	B_{n-k-2}	
$n - k - 1$	T_{n-k-1}	D_{n-k-1}	B_{n-k-1}	$\dots$
$n - k$	T_{n-k}	D_{n-k}	B_{n-k}	G_{n-k}
$\dots$	--	--	--	$\dots$
$n - 1$	--	--	$\hat{B}_{n-1}$	G_{n-1}
n	$\hat{T}_n$	--	--	$\hat{G}_n$

maining buffer length $\{B_{n-k-M+1}, \dots, B_{n-k}\}$ reported in BSRs and previously scheduled RBPs size $\{G_{n-k}, \dots, G_{n-1}\}$. The prediction target and available history observations for n^{th} SPS transmission are shown in Table 4.1, where M represents the maximum observation window. It is important to note that there exists a 4 ms leading time between RBP scheduling and uplink data transmission and another 4 ms processing time for decoding an uplink data packet. This results in an “Information Lagging” issue such that for n^{th} SPS transmission, an eNB only have observed transmitted data and BSR up to $(n - k)^{\text{th}}$ order. The value of k can be determined by:

$$k = \left\lfloor \frac{7 \text{ ms}}{p_j} \right\rfloor + 1 \quad (4.3)$$

For instance, consider an eNB that has reached subframe $n + 26$ in Fig. 4.1. It can be observed that the eNB has to schedule for the uplink transmission at subframe $n + 30$ (the 9th transmission), but it just finished processing the uplink data transmitted at subframe $n + 21$ (the 6th transmission). Therefore, the eNB only has observed up to D_6 , B_6 and G_8 and the order of “Information Lagging” is 3.

In essence, the proposed predictive scheduler determines the number of assigned RBPs based on the estimated length of the previous data buffer ($\hat{B}_{n-1}$) and predicted size

of new data that arrived between previous and next SPS transmission ($\hat{T}_n$) based on:

$$N_i = \begin{cases} N_{\max}, & \text{if } \lceil \frac{\hat{G}_n}{c} \rceil > N_{\max} \\ \lceil \frac{\hat{G}_n}{c} \rceil, & \text{otherwise.} \end{cases} \quad (4.4)$$

where $\hat{G}_n = \hat{B}_{n-1} + \hat{T}_n$. The process of determining $\hat{B}_{n-1}$ and $\hat{T}_n$ is elaborated in the following sections.

4.3.2.1 Predicting the Newly Arrived Data $\hat{T}_n$

The previous values of newly arrived data bytes can be derived by:

$$T_i = D_i + B_i - B_{i-1} \quad (4.5)$$

where $n - k - M + 1 \leq i \leq n - k$. We then propose two prediction models to extrapolate and derive the value of $\hat{T}_n$ —Logical AutoRegressive model (LAR) and Order-5 AutoRegressive model (AR-5). It is important that due to “Information Lagging”, it is necessary to use the prediction models recursively to conduct a k -step ahead prediction, in order to determine $\hat{T}_n$.

Logical AutoRegressive Model

In [144], an autoregressive (AR) model is proposed to predict the content of next haptic data, and it was proved viable with a prediction accuracy of $> 96\%$. This motivates our derivation of a data traffic prediction model that is specifically developed for haptic data when Deadband Principle is deployed. Following a similar procedure as in [144], we assume that acceleration of the current haptic data can be derived from previous ones with a first-order AR model:

$$v'(n) = \beta v'(n-1) + \epsilon(n), \quad (4.6)$$

where $v'(n)$ denotes the acceleration of current haptic data content and $\epsilon(n)$ represents a

noise term with zero mean. It is noted that the acceleration can also be expressed as:

$$v'(n) = \frac{v(n) - v(n-1)}{\Delta T}, \quad (4.7)$$

where $v(n)$ is the change in haptic data and ΔT is the sampling interval of a haptic transmitter. As explained in Section 2.4, the number of transmitted haptic data increases with the change in haptic data. Therefore, it is reasonable to assume a linear relationship between the number of transmitted data and $v(n)$:

$$T(n) = \alpha v(n), \quad (4.8)$$

where $T(n)$ is the number of transmitted haptic data during an SPS interval and α is a constant that is greater than 0. From (4.6), (4.7) and (4.8), we form a second-order AR model for $T(n)$:

$$T(n) = (1 + \beta)T(n-1) - \beta T(n-2) + \alpha \Delta T \epsilon(n). \quad (4.9)$$

The coefficient β can be solved by differentiating the logarithmic likelihood function of (4.9)[145], and it can be estimated by using previously observed values:

$$\hat{\beta} = \frac{\sum_{j=3}^N [T(j) - T(j-1)][T(j-1) - T(j-2)]}{\sum_{j=3}^N [T(j)T(j-1)]^2}, \quad (4.10)$$

where N is the number of previous observations. For the actual implementation of the proposed predictive scheduler, an observation window of size 70 is used. As in the aforementioned example illustrated in Table 4.1, the second-order AR model can be used recursively for the calculation of $\{\hat{T}_i | i = n - k + 1, n - k + 2, \dots, n\}$.

Order-5 AutoRegressive Model

Here we use an AR-5 model to determine $\hat{T}_n$. The algorithm of k -step ahead prediction using Yule-Walker equations is shown in Algorithm 3, where k is determined by (4.3).

Algorithm 3 k -step ahead prediction of AR-5 using Yule-Walker equations

Input: $\{T_{n-k-M+1}, \dots, T_{n-k}\}$
Output: $\hat{T}_n$

- 1: $\mu \leftarrow \frac{1}{M} \sum_{i=n-k-M+1}^{n-k} T_i$
- 2: $\mathbf{T}_0 \leftarrow \{T_{n-k-M+1} - \mu, \dots, T_{n-k} - \mu\}$
- 3: compute 5×5 correlation matrix $\mathbf{R}$ from $\mathbf{T}_0$
- 4: compute 5×1 correlation vector ϕ from $\mathbf{T}_0$
- 5: AR-5 coefficients: $\rho \leftarrow \mathbf{R}^{-1}\phi'$
- 6: **for** $i = 1$ to k **do**
- 7: i^{th} -step prediction:
 $T_{\text{new},i} \leftarrow \rho' \cdot [\mathbf{T}_0(\text{end}), \dots, \mathbf{T}_0(\text{end}-4)]'$
- 8: append $T_{\text{new},i}$ to $\mathbf{T}_0$
- 9: **end for**
- 10: $\hat{T}_n \leftarrow \max\{\mathbf{T}_0(\text{end}) + \mu, 0\}$

4.3.2.2 Estimating the UE's Previous Buffer Occupancy $\hat{B}_{n-1}$

From Table 4.1, it can be seen that the predictive scheduler may not have access to the latest BSR observations. Therefore, the approximation $\hat{B}_{n-1}$ can be determined by recursive calculation of the following equation:

$$\hat{B}_j = \begin{cases} \max\{B_{j-1} + \hat{T}_j - G_j, 0\}, & j = n - k + 1 \\ \max\{\hat{B}_{j-1} + \hat{T}_j - G_j, 0\}, & n - k + 1 < j \leq n - 1. \end{cases} \quad (4.11)$$

4.4 Simulation Results and Discussion

The proposed predictive SPS schemes are implemented on an LTE simulation program based on SimuLTE [54]. SimuLTE is a system-level simulator developed on the well-known discrete event simulation platform OMNeT++ and it is widely used to simulate LTE, and LTE-Advanced systems. To validate the predictive SPS design and evaluate its performance against actual haptic communications systems, we utilise raw haptic data that were captured during a 25-second haptic interaction experiment [146]. We process the raw uplink data using the Deadband Principle with 5% JND and obtain a vector of inter-arrival time, which forms the original uplink data traffic. We choose 5% JND because it corresponds to the least level of haptic data compression, which is essential to

preserving motion/force accuracy in a haptic-tactile application such as remote surgery. Data traffic with different loads is then generated by multiplying the inter-arrival time vector with varying common weights, such that each derived traffic corresponds to a certain traffic load. Simulations of uplink transmission using an LTE system are conducted by feeding the resulting traffic to UEs' application layer and implementing the proposed SPS scheme and FRA in the LTE MAC layer. Since the Tactile Internet is expected to support and deliver short-length data packets [32], we assume that each haptic data stream carries two 3-Dof data types (master velocity and force) each with a 16-bit resolution, which results in one haptic packet having a 12-byte payload [126].

For the conducted simulations, the number of RBPs for a UE that is configured with the normal SPS is determined by $N_{fix} = \lceil 1.05 \times \frac{CLt}{8M} \times \frac{1}{c} \rceil$, where $\frac{CLt}{8M}$ represents the expected number of arriving bytes at a UE during one SPS period. While for the adaptive SPS and the proposed predictive SPS, the maximum number of RBPs for a UE is determined by $N_{max} = \lceil 1.75 \times 1.05 \times \frac{CLt}{8M} \times \frac{1}{c} \rceil$.

4.4.1 Performance of LAR Algorithm for Latency-Critical Services

In order to compare the performance of different LTE scheduling schemes, we repeated the simulations using the conventional DS scheme, the normal SPS scheme, the adaptive SPS scheme [140, 141] and the proposed predictive SPS scheme with different SPS periodicities. It is assumed that all UEs attempt to send uplink haptic data to an eNB and the LTE parameters used in the simulations are summarised in Table 4.2.

4.4.1.1 Latency Performance

Fig. 4.3 compares the latency performance of different uplink transmission schemes. On the one hand, the normal SPS scheme does not provide improved latency performance for the haptic traffic under low-to-moderate traffic loads, in comparison with the conventional DS scheme. This verifies that the normal SPS scheme is not suitable for the VBR haptic traffic. When the traffic load reaches above 80%, however, the performance of the DS scheme degrades as the UEs cannot receive enough RBPs for data transmission. On

Table 4.2: Simulation parameters

Parameter	Value	Notation
Simulation time	100 s	
Number of Cell	1	
Number of UEs	10	M
Bandwidth	20 MHz	
Uplink RBPs	50	
Theoretical uplink capacity	35.6 Mbps	C
Haptic traffic start time	Uniform(0 s, 1 s)	
Haptic packet payload	12 B	
Total haptic traffic load	10%–100%	L
Scheduler algorithm for DS	Proportional fair [96]	
SR periodicity	1 ms	
Tested SPS period	3, 8 ms	t
UE MAC buffer size	0.75 MB	
Bytes per RBP	89	c
LTE multiplexing mode	FDD	
Tx mode	Single antenna	
Maximum UE Tx power	398 mW	
Pathloss model	ITU-R, Rural Macro	

the other hand, the predictive SPS scheme produces the best latency performance among all tested uplink transmission schemes. In particular, when the SPS periodicity is 3 ms and 8 ms, the predictive SPS scheme outperforms the adaptive SPS scheme by roughly 2.5 ms and 3.5 ms respectively, under all traffic loads.

4.4.1.2 Uplink Data Plane Utilisation and Number of Connected UEs

Fig. 4.4 and 4.5 show the total data plane utilisation and the number of successfully connected UEs for different uplink schemes with various traffic loads, respectively. From these figures, the DS scheme is able to support all UEs under all traffic loads, and the corresponding data plane utilisation increases linearly with the traffic load. The only exception is when the traffic load is at 100%, a proportion of the uplink data packets

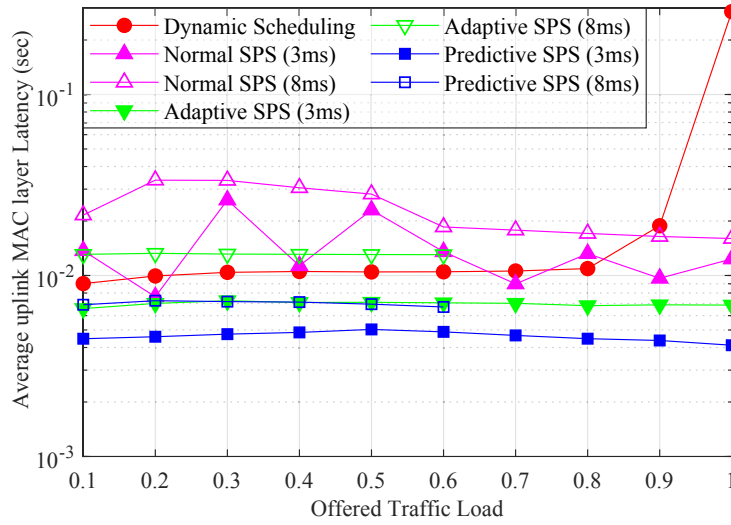


Figure 4.3: Average uplink MAC layer latency under varying traffic loads.

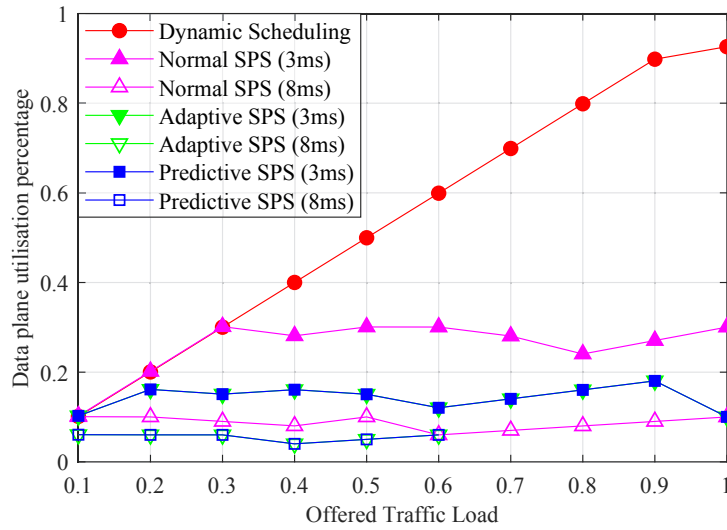


Figure 4.4: Measured data plane utilisation under varying traffic loads.

are dropped at UEs' buffers due to the limited buffer size, and hence leads to a data plane utilisation lower than 100%. The comparatively low data plane utilisation of all SPS schemes results from a decreased number of connected UEs, as demonstrated in Fig. 4.5. Such limitation imposes a restricted number of UEs for the deployment of SPS. It also implies that the SPS scheme allows coexistence of other traffic that uses non-SPS scheme (such as the DS and CBS schemes) to consume the unoccupied radio resources.

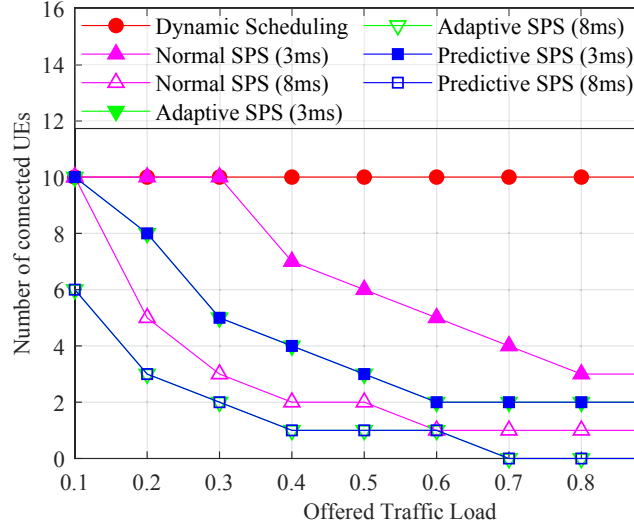


Figure 4.5: Number of connected UEs under varying traffic loads.

4.4.1.3 NMSE of Scheduled RBPs and Residual Buffer Lengths

Let $N_i^{correct}$ denotes the minimum number of RBPs which can successfully clear a UE's uplink data buffer, and it can be calculated by $N_i^{correct} = \lceil \frac{l_i}{c} \rceil$, where l_i is the buffer length before i^{th} SPS transmission. The normalised mean square error (NMSE) of scheduled RBPs is an effective measure of the RBP allocation accuracy, and it can be determined by $NMSE_{RBP} = \frac{1}{\sigma^2 n} \sum_{j=1}^n (N_j^{correct} - N_j)^2$, where σ is the sample standard deviation of $N_j^{correct}$ and n is the number of observed N_j [147]. Fig. 4.6 compares the calculated $NMSE_{RBP}$, and it shows that among all SPS schemes, the proposed SPS scheme is capable of allocating RBPs with the highest accuracy.

Fig. 4.7 plots the average length of UEs' buffers after each uplink transmission. It can be observed that when the SPS periodicity is 8 ms, the mean residual buffer length of the adaptive SPS scheme is similar to that of the DS scheme. This is because both schemes schedule RBPs based on the latest BSR only, in which case the newly arrived data are not considered. Together with Fig. 4.6, Fig. 4.7 shows that the proposed predictive SPS scheme is able to reduce the buffer to a minimal level, by both predicting the newly arrived data and estimating the latest buffer length with maximal accuracy.

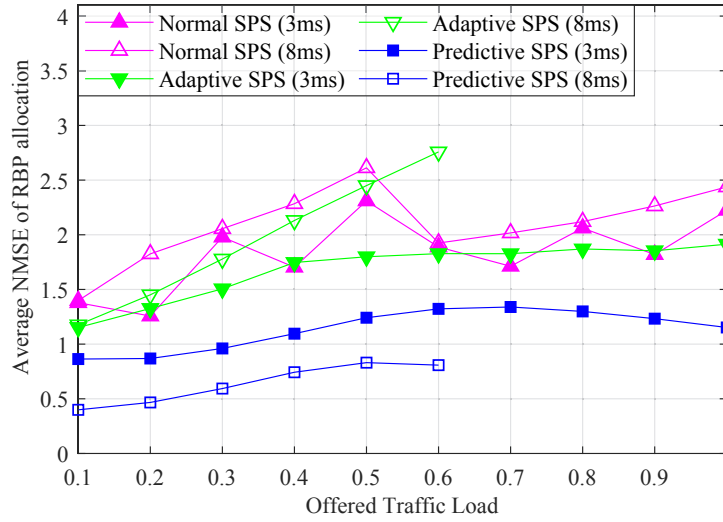


Figure 4.6: Average NMSE of RBP allocation under varying traffic loads.

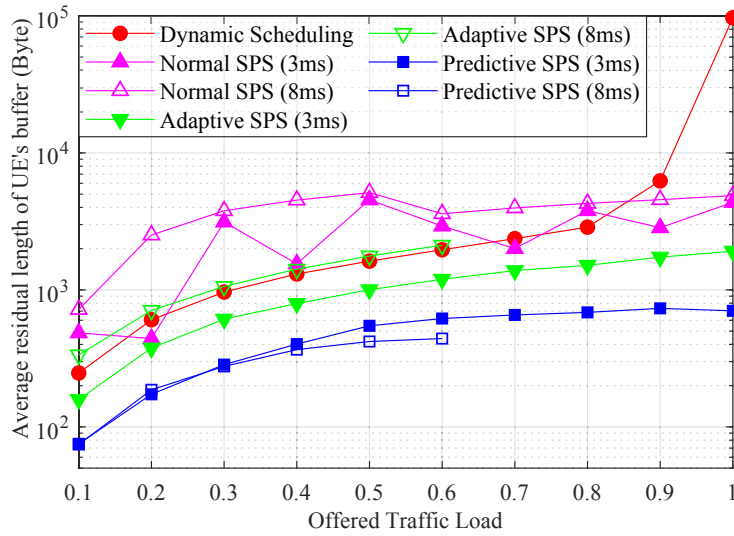


Figure 4.7: Average residual length of UEs' buffer under varying traffic loads.

4.4.2 Performance of LAR and AR-5 Algorithms for Latency-Critical Services over Heterogeneous Networks

To demonstrate the effectiveness of latency reduction for supporting Tactile Internet applications and the compatibility with the conventional DS scheme, simulations are conducted in a heterogeneous network traffic environment. In this situation, 10 UEs with Tactile Internet traffic and 20 non-Tactile UEs with heterogeneous traffic are sharing the

Table 4.3: Simulation parameters

Parameter	Value	Notation
Simulation time	100 s	
Number of Cell	1	
Number of tactile UEs	10	M
Number of non-Tactile UEs	20	
Bandwidth	20 MHz	
Uplink RBPs	50	
Theoretical uplink capacity	35.6 Mbps	C
Total traffic load	10%–100%	L
Non-Tactile traffic load ratio	60%	α
Non-Tactile traffic	Burst model	
Non-Tactile packet payload	Uniform(20,1500) B	
Tactile traffic load ratio	40%	β
Haptic packet payload	12 B	
Scheduling algorithm for DS scheme	Proportional fair	
SR periodicity	1 ms	
Tested SPS period	1, 2 ms	t
UE MAC buffer size	1 MB	
Bytes per RBP	89	c
LTE multiplexing mode	FDD	
Tx mode	Single antenna	
Maximum UE Tx power	398 mW	
Pathloss model	ITU-R, Rural Macro	

PUSCH of an LTE cellular network. We compare results arising from using different uplink transmission schemes, including the DS scheme, normal SPS scheme, adaptive SPS scheme [141] and the proposed AR-5 SPS scheme with different SPS periodicities (1 and 2 ms).

The LTE parameters and selected traffic characteristics are summarised in Table 4.3. Moreover, we repeat the simulation for 10 times, where each UE attempts to start uplink transmission at a different time. Simulation results are presented with 95% confidence interval.

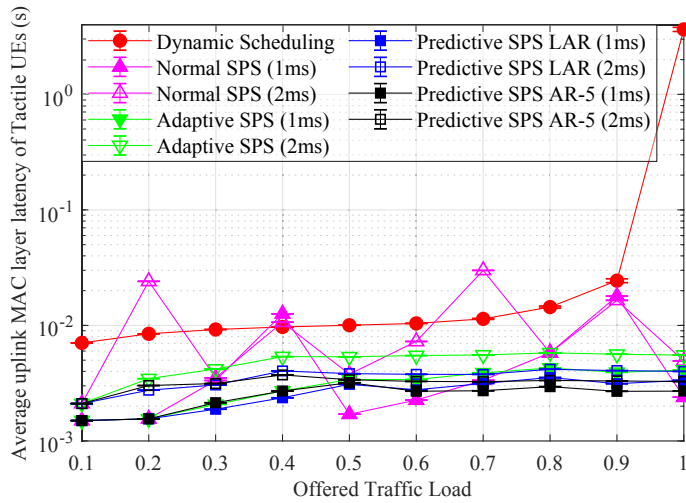


Figure 4.8: Average uplink MAC layer latency of Tactile UEs.

4.4.2.1 Latency Performance

Fig. 4.8 shows the latency performance of Tactile UEs with different uplink scheduling schemes. It can be seen that all advanced SPS schemes (Adaptive, LAR and AR-5) produce steady latency performance against varying traffic loads and outperform the conventional DS scheme. Fig. 4.9 compares the average radio resource consumption by each Tactile UE. It is observed that the normal SPS scheme generates consumption curves that do not follow closely to the traffic loads, indicating that its fixed scheduling policy (scheduling N_{fixed} RBPs for every uplink transmission) cannot adapt to traffic variation and hence explains its unsteady latency performance presented in Fig. 4.8. Moreover, the conventional DS scheme has the lowest resource consumption. This is expected because the DS scheme allocates resources reactively and does not over-schedule radio resources.

A detailed comparison of the Adaptive SPS and proposed predictive SPS schemes (LAR and AR-5) is presented in Fig. 4.10. It can be observed that the predictive SPS schemes produce lower latencies than the Adaptive SPS scheme when SPS periodicity is 2 ms for all considered conditions and 1 ms for most conditions. Furthermore, the proposed AR-5 SPS scheme outperforms the LAR SPS scheme because it produces lower latency values under all conditions except for 0.2–0.5 traffic loads. Simulation results

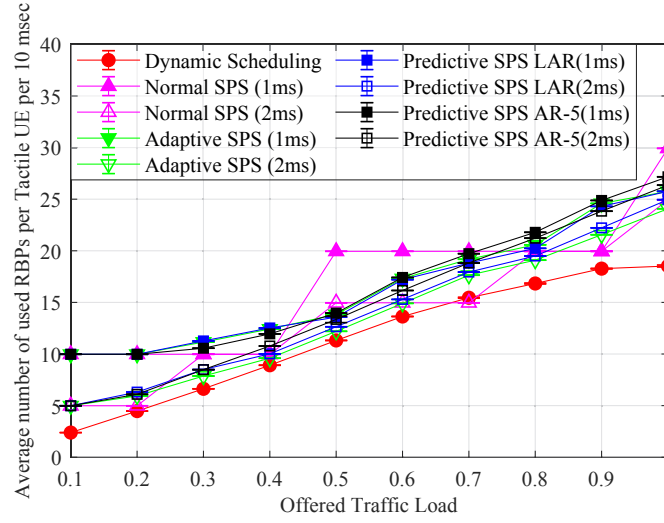


Figure 4.9: Average used RBPs per Tactile UE per 10 ms.

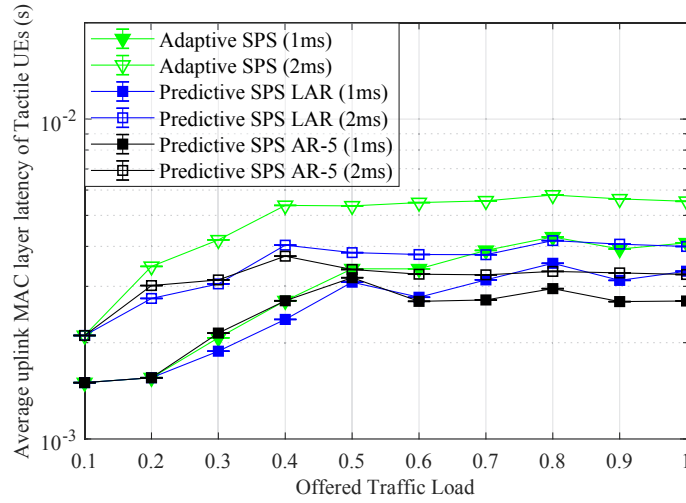


Figure 4.10: Detailed comparison of latency performance between different SPS algorithms.

show that the proposed AR-5 SPS scheme achieves a sub-3 ms uplink latency under any given traffic load when SPS periodicity is chosen to be 1 ms.

Fig. 4.11 shows the latency performance of non-Tactile UEs affected by different SPS schemes. Results clearly show that by implementing the FRA, all SPS schemes have a negligible adverse impact on the latency performance of non-Tactile UEs up to a 0.8 traffic load, such that average uplink latencies of these UEs remain at a sub-30 ms level.

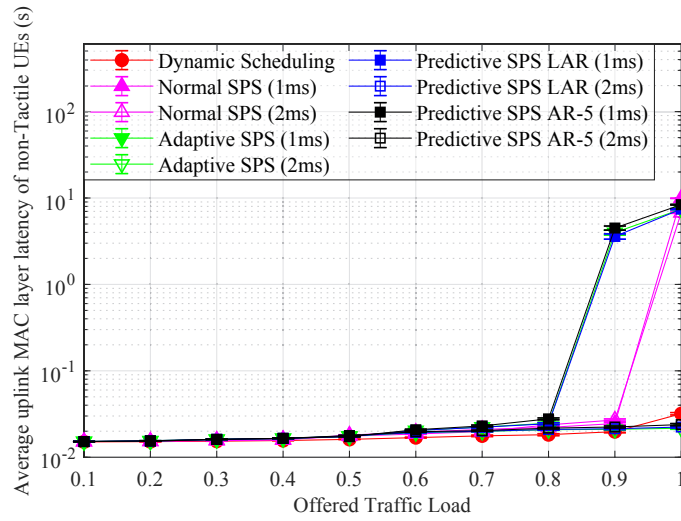


Figure 4.11: Average uplink MAC layer latency of non-Tactile UEs.

When the traffic load is above 0.9, however, the Adaptive SPS (1 ms), LAR SPS (1 ms) and AR-5 (1 ms) schemes affect the latency performance of non-Tactile UEs significantly. This highlights the clear trade-off of latency performance between Tactile and non-Tactile UEs when the traffic load approaches 100%. Improvements can be achieved by implementing latency-aware scheduling algorithms described in [106, 148, 149] for non-Tactile UEs. On the other hand, the Normal SPS (1 and 2 ms) scheme has much less impact on the latency of non-Tactile UEs because it under-schedules radio resources for the Tactile UEs (as shown in Fig. 4.9) and therefore allows more resources to be allocated to non-Tactile UEs.

4.4.2.2 Data Plane Utilisation and Number of Admitted Tactile UEs

Fig. 4.12 and 4.13 compare the data plane utilisation and the corresponding number of connected (admitted) Tactile UEs for different uplink schemes. It can be seen from Fig. 4.12 that the utilisation curves of all schemes increase linearly with the traffic load up to 0.6. The Adaptive SPS (2 ms), LAR SPS (2 ms) and AR-5 SPS (2 ms) schemes produce less utilisation when the traffic load is above 0.7. This conforms to the decreased number of connected Tactile UEs, as shown in Fig. 4.13. On the other hand, the Normal SPS (1 and 2 ms), Adaptive SPS (1 ms), LAR SPS (1 ms) and AR-5 SPS (1 ms) schemes present

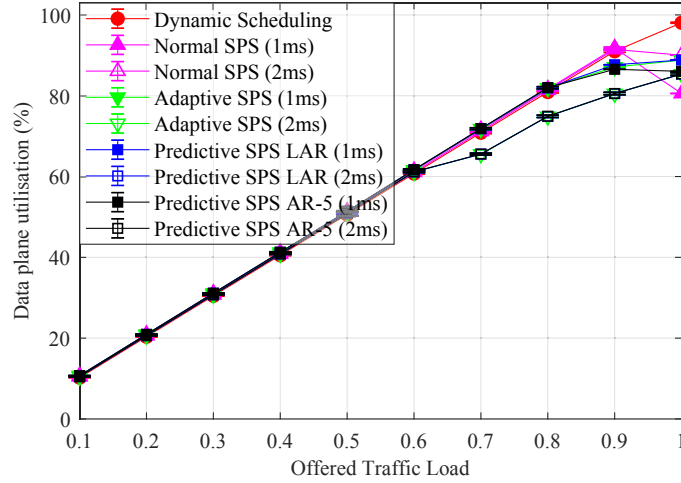


Figure 4.12: Overall data plane utilisation of the LTE cell.

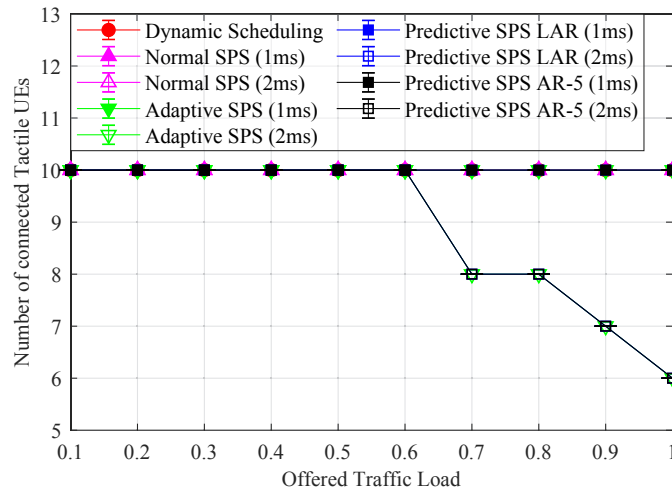


Figure 4.13: Number of connected (admitted) Tactile UEs as a function of total offered traffic load.

similar levels of utilisation to that of the DS scheme up to 0.9 and then deviate with different extents at 100% traffic load, which will be further explained in Section 4.4.2.4. In general, the major cause of the reduced utilisation of our proposed SPS scheme is the decreased number of connected Tactile UEs due to the “Limitation of UE Admission” issue, as explained in Section 4.3.1.

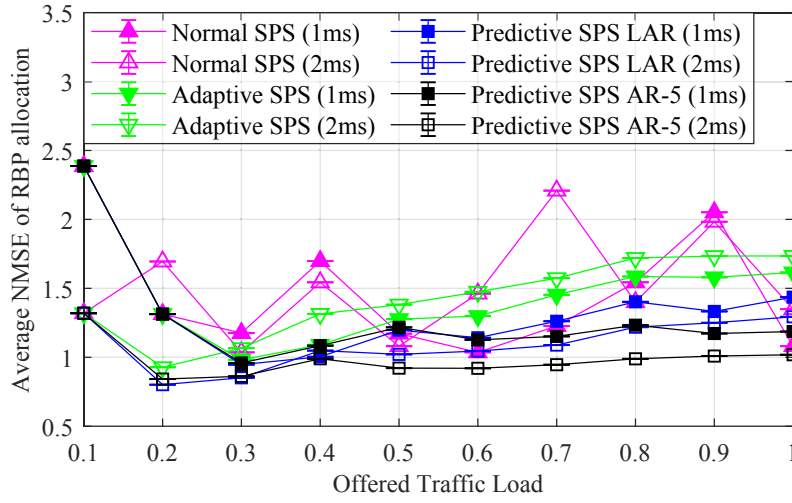


Figure 4.14: Average NMSE of RBP allocation as a function of total offered traffic load.

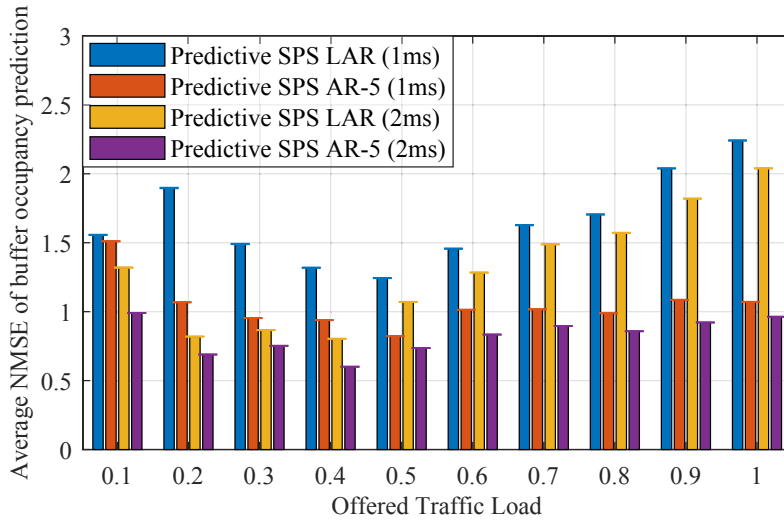


Figure 4.15: Average NMSE of predicted buffer occupancy as a function of total offered traffic load.

4.4.2.3 Scheduling and Prediction Performance

In Fig. 4.14 and 4.15, the efficiency of different SPS scheduling schemes are measured using a unitless quantity termed Normalised Mean Square Error (NMSE). Let y_n and $\hat{y}_n$ represent the n^{th} correct value and the n^{th} actually assigned/predicted value, respectively. The NMSE is defined as $\text{NMSE}_y = \frac{1}{\sigma^2 t} \sum_{n=1}^t (y_n - \hat{y}_n)^2$, where σ is the sample standard deviation of y_n and t is the total number of recorded values [147]. The NMSE provides an

overall idea of how assigned/predicted values differ from correct ones. For the NMSE of RBP allocations, y_n is the minimum required number of RBPs to transmit all buffered data at n^{th} SPS transmission, and $\hat{y}_n$ is the number of actually assigned RBPs. As shown in Fig. 4.14, the predictive SPS schemes have relatively steady and low NMSE curves, as compared with the normal SPS and Adaptive SPS schemes. This indicates that resource allocations performed by the predictive schedulers are closer to true resource demands and hence less over/under-scheduling. It is observed that higher NMSE values occur when the traffic load is at 0.1 for the predictive SPS schemes, this is because the data buffer when under very low traffic load conditions is most likely to be empty and thus y_n equals 0. However, due to the scheduling policy described in (4.4), at least 1 RBP is scheduled for every SPS transmission, in which case the probability of allocation mismatch is increased, and this leads to higher NMSEs.

Fig. 4.15 compares scheduling efficiency between the LAR and AR-5 scheduling algorithms. The efficiency is determined as the NMSE of buffer occupancy, where y_n is true buffer occupancy measured in bytes at n^{th} SPS transmission and $\hat{y}_n$ is predicted buffer occupancy. It is observed that the proposed AR-5 scheduler outperforms the LAR scheduler for all considered conditions. In particular, the AR-5 scheduler achieves significant enhancement of buffer occupancy prediction when SPS periodicity is 1 ms, except for 0.1 traffic load where data arrival is sparse, and prediction of buffer occupancy is generally less accurate.

4.4.2.4 Algorithm Complexity and Packet Drop Rate

Fig. 4.16 shows the complexity of the LAR and AR-5 scheduling algorithms in terms of average processing time. It should be noted that the simulations are conducted using an AMD Opteron 62xx class CPU. We can see that the processing time of both algorithms is steady under varying traffic loads. In particular, the LAR scheduling algorithm takes an average time of 0.025 ms, while the AR-5 scheduling algorithm needs less than 0.15 ms to complete. Although the AR-5 algorithm requires more processing time to achieve a better prediction accuracy than the LAR algorithm, such time expense can be tolerated as it is much less than one TTI of an LTE system. It can be observed that 1 ms and 2 ms SPS

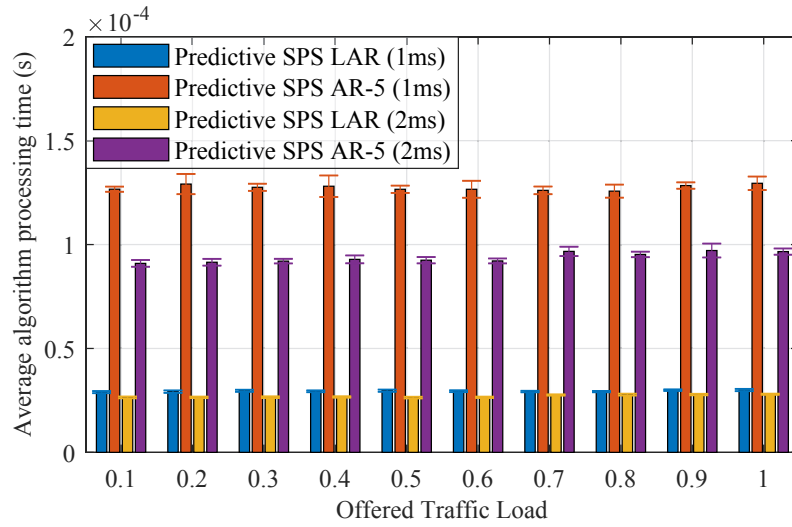


Figure 4.16: Comparison of complexity of AR-5 and LAR algorithms.

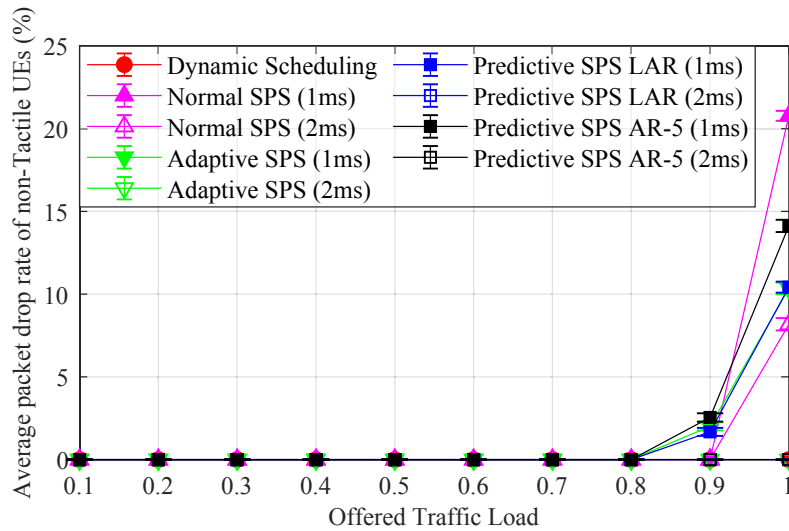


Figure 4.17: Average packet drop rate for non-Tactile UEs.

periodicities require an 8-step ahead and a 4-step ahead prediction, respectively, which leads to the longer average processing time for the former case.

Fig. 4.17 shows the average packet drop rate of non-Tactile UEs under varying traffic loads. It is observed that packet drop only occurs when the Normal SPS (1 and 2), Adaptive SPS (1 ms), LAR SPS (1 ms) and AR-5 (1 ms) SPS schemes are implemented at 100% traffic load. The latency and packet drop rate of non-Tactile UEs at 100% traffic shown

in Fig. 4.11 and Fig. 4.17 explain the utilisation deviation presented in Fig. 4.12. This is caused by the excessive waiting time of packets in non-Tactile UEs' buffer, which in turn results in a higher packet drop rate and decreased transmitted packets and hence less data plane utilisation. This system deterioration can be alleviated by prioritising non-Tactile UEs according to their QoS requirements using QoS-aware scheduling algorithms and increased buffer capacity.

4.5 Conclusions

In this chapter, we presented a novel predictive SPS scheme for alleviating the uplink latency bottleneck of mobile networks, such that latency-critical services could be realised by the current 4G and future 5G networks. To unlock the full potential of the SPS scheme and to solve the associated over-/under-scheduling problem, we introduced the first FRA, which allows variable resource allocations under SPS configuration. Next, we proposed two predictive SPS schedulers that not only predict the amount of newly arrived data between adjacent SPS transmissions but also estimate UE's previous buffer occupancy in a recursive approach. Incorporating the proposed FRA and predictive schedulers, our SPS scheme determines the sizes and positions of scheduled radio resources such that extremely low uplink latency can be achieved. Additionally, the proposed SPS scheme does not require a new signalling process to be introduced to the traditional SPS scheme. Therefore, no PDCCH resource will be consumed after an SPS session is established, which significantly increases the number of accessible mobile connections of a mobile cellular network. This is especially beneficial for future 5G cellular networks supporting massive connectivity. We verified the proposed SPS scheme by evaluating latency performance, resource allocation efficiency, prediction accuracy and remaining buffer length with other existing uplink schemes when Tactile UEs coexist with other UEs transmitting heterogeneous network traffic. From the numerical results, we showed that the proposed SPS scheme achieved significant uplink latency improvement. Results have also demonstrated that more accurate prediction results are achieved by the AR-5 SPS scheduler compared with the LAR scheduler. Finally, the FRA supports variable resource alloca-

tions with only marginal impact on non-Tactile traffic for most tested conditions.

It can be seen from Fig. 4.10 that the latency of Tactile UEs increases with traffic load, this could be caused by increasing prediction error of the applied LAR and AR-5 algorithms, as is reflected in Fig. 4.15. This motivates our research into applying more advanced techniques for higher prediction accuracy and thus further improve the latency performance and resource efficiency of the proposed predictive SPS scheme. Our study of such an approach is presented in the next chapter.

Predictive Semi-Persistent Scheduling Scheme Powered by Supervised Machine Learning

5.1 Introduction

With the goal of ultra-low end-to-end latency, the Tactile Internet cannot be realised without shortening delay components of haptic data transmission, including application domain (data generation/collection and processing delay etc.), transmission domain (queueing delay and signalling delay etc.) and network domain (data generation/responding delay etc.). Therefore, powerful techniques that improve latency performance during data transmission need to be studied and exploited. Our work presented in Chapter 4 provides a solution to ultra-low latency uplink transmission protocol of the wireless access network for the Tactile Internet by exploiting an existing uplink MAC protocol called SPS defined by LTE and 5G standards. However, the prediction accuracy deteriorates with the increase of traffic load, which suggests that the prediction algorithm based on a simple autoregressive model might be insufficient for fast-changing VBR data traffic generated by future emerging applications.

In this chapter, we focus on using machine learning to overcome the weakness of our previous work. With the advancement of the processing power of contemporary computer and the availability of large dataset, tools from the machine learning paradigm have been proved effective for many applications such as regression, classification and decision making process [50]. Contrary to traditional computational algorithms which need to be explicitly programmed, machine learning algorithms can learn and improve themselves from the input data and are expected to understand the structure/feature

of a given dataset. Three basic machine learning paradigms exist, namely supervised and unsupervised machine learning and reinforcement learning. The supervised machine learning takes labelled data as its input and applies statistical analysis, Bayesian models etc. to perform regression or classification on unlabelled data. When learning from the data, machine learning algorithms improve themselves by evaluating the deviation between predicted and given labels. Examples of regression-type machine learning are weather forecasting, future stock prediction and natural language processing (NLP), while classification-type machine learning involves facial recognition, optical character recognition and future event prediction. On the other hand, unsupervised machine learning algorithms are provided with unlabelled or partially labelled data, and they need to learn the hidden feature, pattern and commonalities among the data so that categorisation and labelling can be performed [150]. Finally, reinforcement learning works by interacting with the known/unknown environment. It deploys an *agent* to take *action* on every encountered situation within the *environment*, and this agent receives *rewards*/penalties for performing correctly/incorrectly. By deploying reinforcement learning techniques, the agent is expected to learn the dynamics of the environment and causality between an action and its resulting reward/penalty, in which case the performance of the agent can be improved by choosing the correct action. Different to supervised learning, where the ground-true label of input is provided, the agent of a reinforcement learning algorithm only receives an immediate reward for its action, which does not necessarily represent the overall worthiness of that action. In other words, an action resulting in the best immediate reward might not be the correct choice to achieve the highest long-term gain.

In this chapter, we exploit supervised machine learning to make accurate predictions on the unobserved network traffic, and we present an Attention-based recurrent neural network for a Sequence-to-Sequence (seq2seq) model. The rest of this chapter is structured as follows. In Section 5.2, we cover related machine learning techniques for our predictive SPS scheme. Section 5.3 discusses the detailed structure and working principle of our seq2seq model for predicting future arrivals of a traffic flow in haptic-tactile applications. Then, the prediction performance during off-line training and latency performance during the simulation of the proposed model are evaluated in Section 5.4. Fi-

nally, Section 5.5 concludes this chapter with a summary of our findings.

5.2 Related Work

We introduce research on time series prediction using attention mechanism in this section. As pointed out in Chapter 4, precise prediction on the future network traffic is vital for allocating radio resources to uplink devices accurately and efficiently. By equation (4.3), we can see that for SPS periodicity equals p ms, and there are in total $\lfloor \frac{Z}{p} \rfloor + 1$ future predictions need to be performed in order to determine the current buffer occupancy. Such a requirement fits the concept of Sequence-to-Sequence model, whose input and output are in a sequential format. Among seq2seq tasks, natural language translation in the NLP topic has become a popular research topic. It attempts to read a sentence in one language and outputs translation in another.

To this end, encoder-decoder models are frequently used in natural language translation as it targets at extracting a fixed-length feature vector from the input language by an encoder and then interpreting the vector-form feature into variable-length outputs by a decoder. Both encoder and decoder consist of a recurrent neural network (RNN) unit such as Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) [49]. As for the encoder, each element in the source sentences is fed into the RNN one by one to generate a context vector. The decoder behaves differently at training and translation stages. At the training stage, the RNN is fed with the target sentence while it is initialised with the context vector from the encoder. While at the translation stage, the decoder is firstly fed with a start-of-sentence (sos) token and then generates the first translated output. The following outputs are then produced by feeding the sos token as well as the already translated output(s). Results show that this encoder-decoder structure outperforms other machine learning model such as statistical machine translation system in terms of accuracy [49, 151], yet machine translation is still facing unsolved problems such as rapid accuracy deterioration for long input sentences [152, 153].

In 2015, Bahdanau *et al.* proposed the first Attention-based encoder-decoder model for machine translation [51]. For an output at a given position in the output sequence,

the proposed model can selectively put more weights (attention) on more relevant input elements instead of the entire input sequence. Furthermore, contrary to traditional encoder-decoder models where the encoded feature needs to capture all the necessary information of source sentences, the proposed Attention-based model relaxed this issue by allowing the decoder to look at the entire source sentence. By visualising the weights of trained attention matrices, one can see that when translating an output word Out_i , the decoder indeed pays more attention to relevant input words (e.g. In_{i-1} , In_i and In_{i+1}). Tested results show that the proposed model significantly outperforms the conventional encoder-decoder model (RNNencdec) for both short and long sentences in an English-French translation task.

To this end, the Attention-based encoder-decoder model was proposed as a promising machine learning method for seq2seq tasks because it is able to capture the long-term temporal correlation of time series data. This also fits the demand of k -step ahead prediction for the proposed predictive SPS scheme as described in Section 4.3.2.1, as both the input and output are in a sequential format. Currently, only a little work has focused on time series prediction using Attention-based encoder-decoder model. Qin *et al.* proposed a dual-stage Attention-based neural network (DA-RNN) for predicting the target time series based on history observation of multiple driving series [52]. Apart from the attention operation introduced in [51], their proposed model performs an additional attention operation among different driving series. In other words, the proposed model performs spatial attention at the encoder to select relevant driving series, then applies temporal attention at the decoder to select the relevant element of the output of the encoder. This design was verified by using room temperature as target series and 16 other monitoring data for about 40 days as driving series. Results show that DA-RNN outperforms not only traditional statistical time series prediction method such as ARIMA and NARX, but also the conventional machine learning model such as RNNencdec and Attention-based RNN network.

Similarly, a multi-level attention network for geo-sensory data (GeoMAN) was proposed in [53]. This network also employs spatial and temporal attention mechanism to select relevant time series data from relevant sensors. Furthermore, external factors

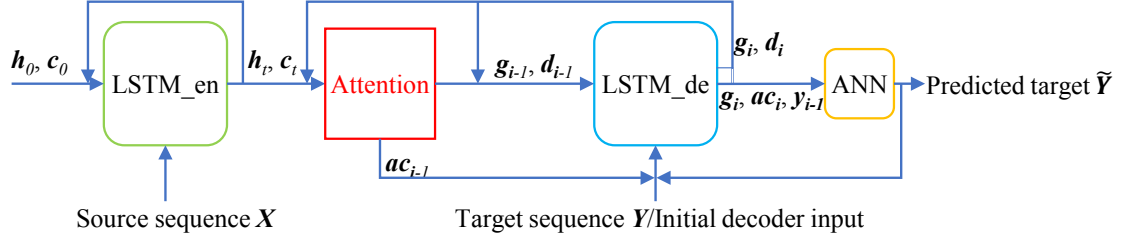


Figure 5.1: Overall architecture of implemented encoder-decoder model.

arising from the change of the surrounding environment are used together with the results from the spatial-temporal attention at the last layer, to predict future readings of a geo-sensor. This network was tested against water quality data with 14 types of sensor readings and air quality data with 35 types of sensor readings. Results show that the proposed GeoMAN achieved the best performance among considered alternatives such as ARIMA, LSTM, RNNencdec and DA-RNN, in terms of root mean square error (RMSE) and mean absolute error (MAE).

5.3 A Sequence-to-Sequence Model for Buffer Occupancy Prediction

The proposed Attention-based recurrent neural network for time series prediction in this work employs an RNN encoder-decoder architecture which is illustrated in Fig. 5.1. Specifically, an LSTM unit is selected as the RNN unit. The notation used for explaining the procedure of our proposed Attention-based prediction model is listed in Table 5.1.

Encoder using LSTM

The encoder is used to extract compressed information from the source sequences to help the decoder to generate accurate predictions, and it is achieved by feeding elements of source sequences into an LSTM unit at the encoder and extracting generated hidden state (h_t) and cell state (c_t). LSTM unit was first introduced by Hochreiter *et al.* [154], and it has a recurrent structure such that the generated information can be fed back as an input for the next computation, in which case long-term temporal dependencies in the input

Table 5.1: Notations for Attention-based prediction model

Notation	Description	Notation	Description
X	Source sequence	T	Length of source sequence
n	Dimension of source sequence	m	Dimension of hidden state of LSTM.en
h_t	t^{th} hidden state of LSTM.en	c_t	t^{th} cell state of LSTM.en
Y	Target sequence	I	Length of target sequence
p	Dimension of target sequence	q	Dimension of hidden state of LSTM.de
g_i	i^{th} hidden state of LSTM.de	d_i	i^{th} cell state of LSTM.de
$\tilde{Y}$	Predicted target sequence	ac_i	i^{th} weighted compressed information from output of LSTM.en

sequence could be captured. Moreover, the LSTM unit offers a non-linear transformation for the input signal without suffering from vanishing gradient problem. Conventional non-linear activation unit such as sigmoid, tanh or rectified linear unit (ReLU) functions may encounter vanishing gradient in the backward propagation process, which in turn leads to non-optimised results or slow convergence speed. This issue has less effect on LSTM because it accumulates information flows over time, and this allows gradients to be unchanged. A modern version of LSTM contains a forget gate (f), an input gate (i), a cell gate and an output gate (o), by which hidden and cell states can be computed.

For a given source sequence $X = (x_1, x_2, \dots, x_t, \dots, x_T) \in \mathbb{R}^{n \times T}$ and target sequence $Y = (y_1, y_2, \dots, y_i, \dots, y_I) \in \mathbb{R}^{p \times I}$, the flow chart describing the computation process of a LSTM at encoder is presented in Fig. 5.2 and its calculation is shown as follows:

$$i_t = \sigma(W_i[h_{t-1}; x_t] + b_i) , \quad (5.1)$$

$$f_t = \sigma(W_f[h_{t-1}; x_t] + b_f) , \quad (5.2)$$

$$g_t = \tanh(W_g[h_{t-1}; x_t] + b_g) , \quad (5.3)$$

$$o_t = \sigma(W_o[h_{t-1}; x_t] + b_o) , \quad (5.4)$$

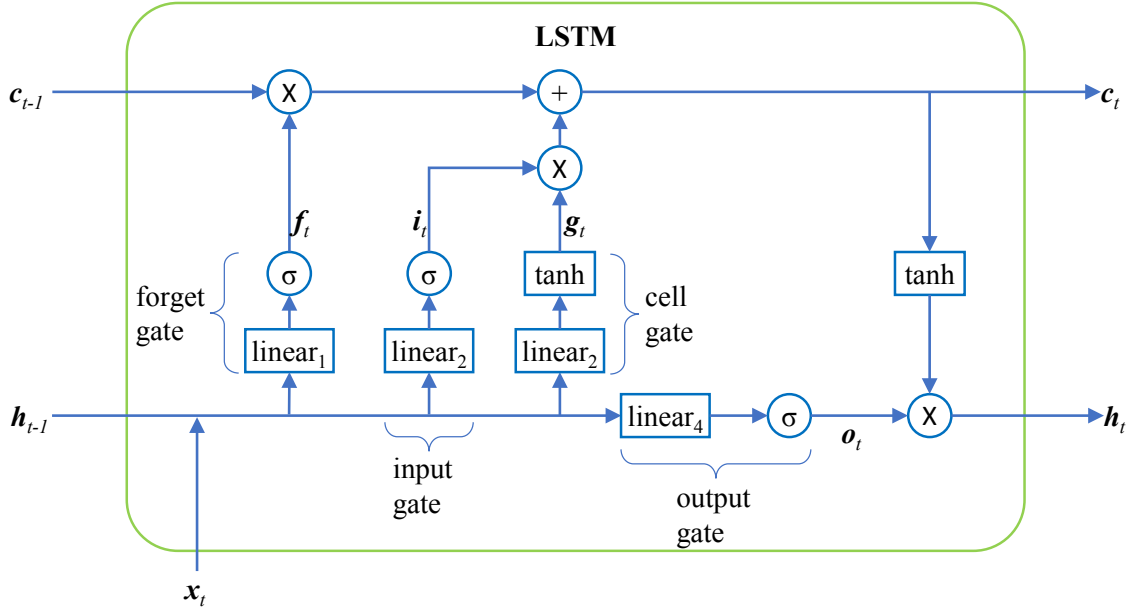


Figure 5.2: Detailed structure of the LSTM unit at encoder.

$$c_t = f_t \otimes c_{t-1} + i_t \otimes g_t, \quad (5.5)$$

$$h_t = o_t \otimes \tanh(c_t). \quad (5.6)$$

where $x_t \in \mathbb{R}^n$ is the t^{th} element in the input sequence, $h_{t-1} \in \mathbb{R}^m$ is the hidden state from the previous LSTM calculation and $[h_{t-1}; x_t]$ represents the concatenation of these two vectors. $W_i, W_f, W_g, W_o \in \mathbb{R}^{m \times (n+m)}$ and $b_i, b_f, b_g, b_o \in \mathbb{R}^m$ are multiplication matrices and bias terms of input, forget, cell and output gates. The symbol σ and $\otimes$ denote sigmoid function and element-wise multiplication, respectively. Specifically, f_t determines the extent of forgetting the previous cell information and o_t determines how much information of new cell state shall contribute to the calculation of new hidden state h_t . It should be noted that when feeding the first element of source sequence into the LSTM_en, the hidden and cell states denoted as h_0 and c_0 need to be initialised and provided.

Additive Attention Mechanism

A collection of hidden and cell states are available when encoding finishes and they are

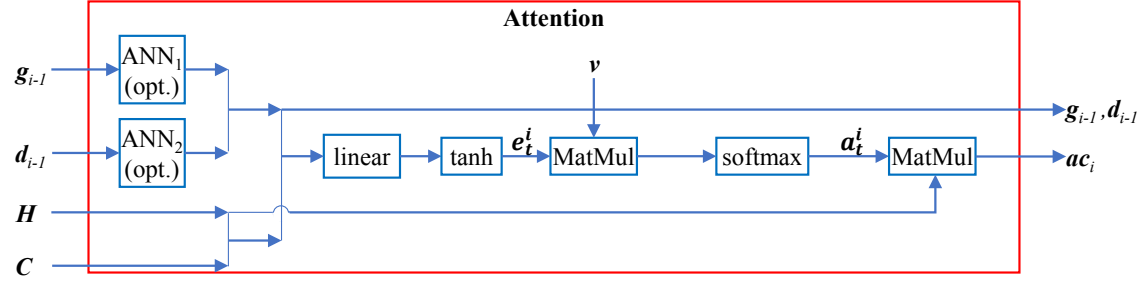


Figure 5.3: Additive attention mechanism implemented by the proposed model.

then processed by the attention mechanism. In our work, we employ an additive attention mechanism [51] between the accumulated hidden and cell states from LSTM.en and the generated hidden and cell states from LSTM.de. The detailed operation is illustrated in Fig. 5.3, where $\mathbf{H} = (h_1, h_2, \dots, h_T)^\top \in \mathbb{R}^{T \times m}$ and $\mathbf{C} = (c_1, c_2, \dots, c_T)^\top \in \mathbb{R}^{T \times m}$ denote the accumulated hidden and cell state, respectively. The detailed mathematical derivation of the additive attention can be expressed as follows:

$$e_t^i = \tanh(W_a[h_t; c_t; g_{i-1}; d_{i-1}] + b_a), \quad (5.7)$$

$$a_t^i = \frac{\exp(v^\top e_t^i)}{\sum_{t=1}^T \exp(v^\top e_t^i)}, \quad (5.8)$$

$$ac_i = \sum_{t=1}^T a_t^i [h_t; c_t], \quad (5.9)$$

where $g_i \in \mathbb{R}^q$ and $d_i \in \mathbb{R}^q$ are the hidden and cell states calculated by LSTM.de. In particular, the energy score e_t^i measures the importance of t^{th} hidden and cell states from the encoder to the i^{th} element of decoder, in order to make accurate target sequence prediction $\hat{y}_i$. The energy scores are then normalised by a softmax function to ensure that the sum of learnt attention weights a_t^i equals 1. Then, the weighted sum of the accumulated compressed information from LSTM.en can be calculated using a simple matrix multiplication operation. Such a procedure is expected to select the most relevant outputs from LSTM.en for a more accurate target prediction. $W_a \in \mathbb{R}^{q \times (2q+2m)}$, $b_a \in \mathbb{R}^q$ and $v \in \mathbb{R}^q$ are trained parameters of the attention mechanism. It should be noted that at $i = 1$, the hidden and cell states of LSTM.de are initialised with the last hidden and cell state

of LSTM_{en} processed with feed-forward networks, which means $g_0 = \text{ANN}_1(h_T)$ and $d_0 = \text{ANN}_2(c_T)$.

Decoder using LSTM

Finally, the target sequence prediction is calculated by the LSTM_{de} followed by a simple feed-forward network. In particular, the LSTM_{de} generates hidden and cell states by taking the previous states, weighted compressed information from LSTM_{en} and previous target prediction as inputs. The detailed calculation can be expressed as follows:

$$i'_i = \sigma(W'_i[g_{i-1}; z_i] + b'_i) , \quad (5.10)$$

$$f'_i = \sigma(W'_f[g_{i-1}; z_i] + b'_f) , \quad (5.11)$$

$$g'_i = \tanh(W'_g[g_{i-1}; z_i] + b'_g) , \quad (5.12)$$

$$o'_i = \sigma(W'_o[g_{i-1}; z_i] + b'_o) , \quad (5.13)$$

$$d_i = f'_i \otimes d_{i-1} + i'_i \otimes g'_i , \quad (5.14)$$

$$g_i = o_i \otimes \tanh(d_i) . \quad (5.15)$$

where $z_i \in \mathbb{R}^{m+p}$ is the concatenation of the i^{th} element in the target sequence and weighted compressed information. In particular, $z_i = [ac_i; y_i]$ in training stage and $z_i = [ac_i; \tilde{y}_i]$ in evaluation stage, respectively. $W'_i, W'_f, W'_g, W'_o \in \mathbb{R}^{q \times (m+p)}$ and $b'_i, b'_f, b'_g, b'_o \in \mathbb{R}^q$ are multiplication matrices and bias terms of input, forget, cell and output gates of LSTM_{de}. The i^{th} predicted target sequence is then derived by feeding the concatenation of y_{i-1} , ac_i and g_i into a simple feed-forward artificial neural network denoted as ANN in Fig. 5.1.

Data Pre-Processing and Learnable Initial Decoder Input

Different to [52, 53] where multi-variant data were fed directly into a prediction model, our proposed model pre-processes raw input and target data to form proper source and

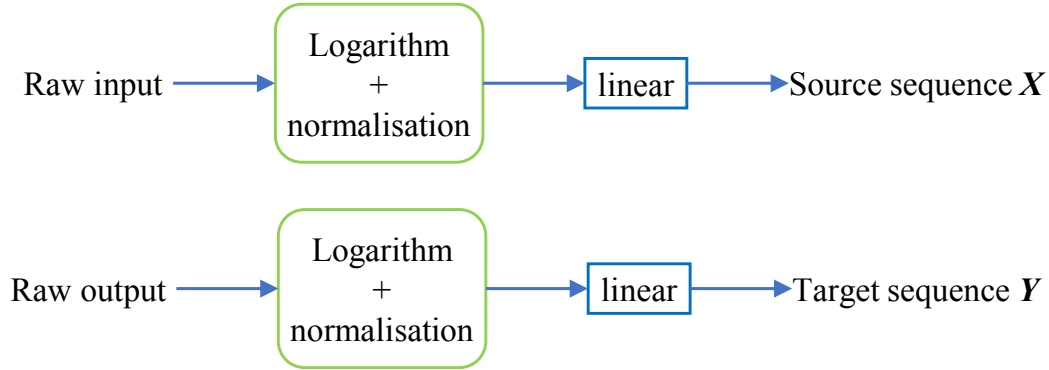


Figure 5.4: Data pre-processing procedure.

target sequences. First, the logarithm values of raw data are computed, followed by normalisation. This can remove potential skewness in the data samples. Dimension expansion, often quoted as “embedding” in neural language translation, is critical because it converts one-hot encoded words into manageable vectors, which helps remove sparsity of the raw input data. Similarly, both normalised input and target data in our proposed model undergo a dimension expansion process implemented by matrix multiplication, such that potential high-dimension pattern of the input data could be captured. The pre-processing procedure of our proposed model is illustrated in Fig. 5.4.

An Attention-based recurrent neural network designed for natural language translation requires feeding a fixed start-of-sentence (sos) signal into the decoder to generate the first target prediction. Such action is inapplicable for our time series task, hence our proposed model utilises an input vector for the decoder with learnable parameters to initiate generation of the target sequence, which introduces extra flexibility into target prediction.

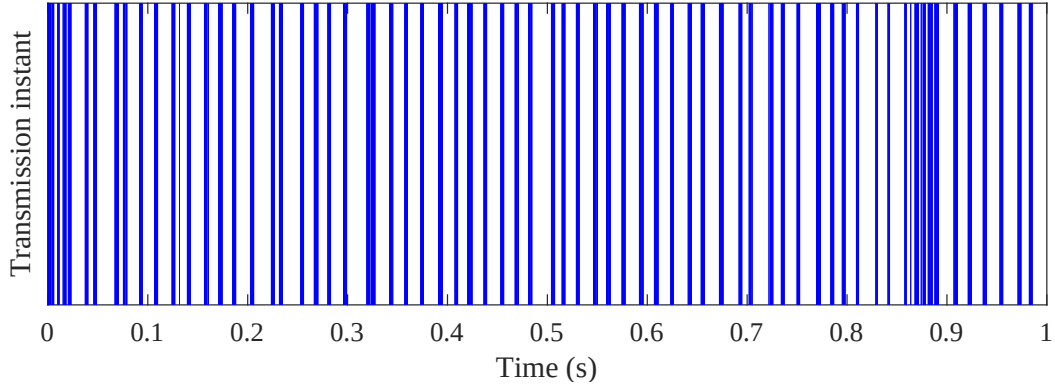


Figure 5.5: Data arrivals of haptic interaction experiment.

Table 5.2: Statistics of data inter-arrival time (ms).

Experiment indices	Statistics			
	Min	Mean	Max	Std
1	0.0	1.151	2.530×10^1	3.337
2	0.0	1.450	1.957×10^2	5.533
3	0.0	1.117	7.639×10^1	3.215
4	0.0	1.146	3.318×10^1	3.273
5	6.000×10^{-3}	4.882	1.979×10^2	6.606

5.4 Simulation Results and Discussion

5.4.1 Traced Data Traffic from H2M Experiment

To verify the proposed prediction model and to evaluate its performance with predictive SPS scheme, we utilise a traced data traffic of H2M experiment ¹ to form the data traffic generated by the application layer in our simulations. Specifically, a 22-sensor glove with position tracker and hand exoskeleton was used as Master domain haptic control device, and a virtual environment containing virtual object was implemented at Slave domain. Different actions were performed during H2M experiments, including free space movement, touching the surface of a virtual object and grasping/moving multiple virtual objects. An example of generated data flow with a time window of 1 second

¹I would like to thank Lihua Ruan for sharing her traced data traffic in haptic-tactile experiments.

is illustrated in Fig. 5.5². In addition, the statistics of inter-arrival time of generated data flows from 5 separate H2M experiments are shown in Table 5.2³. These results reveal that the data flow generated by the haptic interaction application exhibits randomness and bursty characteristics. In particular, the inter-arrival time between two adjacent haptic data could range from 0 ms (multiple data packets were generated simultaneously) to 200 ms. Therefore, regular SPS with fixed resource allocation might fail to deliver high QoS for such an H2M interaction application. We apply the same procedure as explained in Section 4.4, such that data traffic with different loads are generated by multiplying the inter-arrival time vector from an H2M experiment with varying common weights.

5.4.2 Data for Training and Testing Proposed Attention-Based Recurrent Neural Network Model

To obtain the training and testing data for our supervised prediction model, we firstly apply traced H2M data to regular SPS scheme with a periodicity of 2 and 3 ms then extract relevant samples including the amount of transmitted data, BSR and scheduled data for each uplink transmission. Hence, the newly arrived data N during an SPS period can be calculated by eq. (4.5).

5.4.3 Prediction Accuracy of Proposed Attention-Based Recurrent Neural Network Model during Off-Line Training

In this section, the extracted time series of newly arrived data N during an SPS period is used for off-line supervised training for our proposed prediction model. In this work, we use an observation window of 30, which means 30 history observations of N form a raw input and k immediate future N form a raw output. In particular, we verify our proposed prediction model by comparing the accuracy of the k -step ahead prediction task. In addition, we compare the accuracy of the proposed model with other alternative models include AR-5, simple feed-forward Artificial Neural Network (ANN), DA-RNN [52], Dot-Product Attention-based network (DP-Attn) [155] and Self-Attention-based network

²Fig. 5.5 is generated from Lihua Ruan's traced data traffic.

³The numbers in Table 5.2 are calculated from Lihua Ruan's traced data traffic.

Table 5.3: NMSE of prediction for 2 ms periodicity.

	NMSE									
Traffic load	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
AR-5	0.9158	0.9144	1.0060	1.0245	0.9461	0.9503	0.9882	0.9243	0.9084	0.9230
ANN	0.6467	0.9262	0.9876	1.0522	0.9450	0.8776	0.9809	0.9538	0.8592	0.9565
DA-RNN	1.1370	1.2423	1.1205	1.0258	1.3155	1.1065	1.1765	1.0096	1.1469	0.9615
DP-Attn	1.5144	1.7567	1.1584	0.9976	1.2948	0.9188	1.0215	1.2374	1.1531	1.1522
S-Attn	0.9433	0.9983	0.9986	0.9949	1.0015	0.9910	1.0015	0.9968	0.9894	1.0037
Proposed model	0.5794	0.5508	0.7352	0.7638	0.5982	0.6110	0.8508	0.6645	0.5626	0.7391

Table 5.4: NMSE of prediction for 3 ms periodicity.

	NMSE									
Traffic load	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
AR-5	0.8900	0.9996	0.9410	0.9161	0.9558	0.8400	0.9081	0.8555	0.8947	0.8241
ANN	0.5611	1.0134	0.9875	0.8909	0.9701	0.8443	0.9638	0.8243	0.9382	0.8914
DA-RNN	0.7247	1.1291	1.0516	1.0363	1.0134	0.9581	1.0701	0.9681	0.9682	0.9954
DP-Attn	1.2310	1.4359	0.8925	1.0854	1.0526	1.0456	2.5316	1.6783	1.0717	0.9483
S-Attn	0.9506	1.0029	0.9904	0.9870	0.9928	0.9814	0.9849	0.9848	0.9844	0.9789
Proposed model	0.5211	0.7468	0.5992	0.6023	0.7516	0.5949	0.7201	0.5528	0.6547	0.6792

Table 5.5: MAE of prediction for 2 ms periodicity.

	MAE									
Traffic load	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
AR-5	0.7824	0.8250	0.7087	0.7488	0.8277	0.7683	0.6485	0.7704	0.7922	0.6842
ANN	0.6027	0.8770	0.7188	0.6474	0.8363	0.7599	0.5945	0.8168	0.7445	0.6664
DA-RNN	0.8125	0.9466	0.8342	0.6872	0.9349	0.8210	0.7139	0.8382	0.8932	0.6578
DP-Attn	0.9890	1.0472	0.8223	0.7426	0.9518	0.7372	0.6198	0.9041	0.8952	0.7649
S-Attn	0.8447	0.9260	0.7106	0.7561	0.9118	0.8238	0.5977	0.8483	0.8585	0.6737
Proposed model	0.5428	0.5518	0.5416	0.5790	0.5869	0.5630	0.5548	0.6043	0.5623	0.5786

(S-Attn) [156]. It should be noted that our specific prediction task contains only one driving sequence, and therefore the spatial attention mechanism of DA-RNN is removed for actual implementation. The prediction accuracy measured by Normalised Mean Square Error (NMSE) and Mean Absolute Error (MAE) for different periodicities and varying loading conditions are listed in Table 5.3 to 5.6, respectively. From the testing results, we can clearly see that our proposed model generates the least prediction error among all considered existing methods, which indicates that our proposed model shall enhance the latency performance and allocation efficiency of predictive SPS scheme.

Table 5.6: MAE of prediction for 3 ms periodicity.

	MAE									
Traffic load	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
AR-5	0.8108	0.7083	0.7910	0.7447	0.7253	0.7479	0.6266	0.7341	0.6824	0.6896
ANN	0.5845	0.7904	0.8154	0.7802	0.7133	0.7695	0.5762	0.7547	0.6353	0.7502
DA-RNN	0.6619	0.8042	0.8332	0.7746	0.7041	0.8423	0.6910	0.8304	0.6649	0.7710
DP-Attn	0.8629	0.9728	0.7563	0.7299	0.6704	0.8349	1.1894	0.9992	0.7341	0.7940
S-Attn	0.9063	0.7198	0.8541	0.8178	0.7671	0.8467	0.6348	0.8555	0.6535	0.8215
Proposed model	0.5302	0.5463	0.5588	0.5503	0.5766	0.6044	0.5039	0.5496	0.5639	0.6055

Table 5.7: Simulation parameters

Parameter	Value	Notation
Simulation time	100 s	
Number of Cell	1	
Number of tactile UEs	10	M
Number of non-Tactile UEs	20	
Bandwidth	200 MHz	
Uplink RBPs	500	
Theoretical uplink capacity	36.0 Mbps	C
Total traffic load	10%–100%	L
Non-Tactile traffic load ratio	60%	α
Non-Tactile traffic	Burst model	
Non-Tactile packet payload	Uniform(20,1500) B	
Tactile traffic load ratio	40%	β
Haptic packet payload	12 B	
Scheduling algorithm for DS scheme	Proportional fair	
SR periodicity	1 ms	
Tested SPS period	2, 3 ms	t
UE MAC buffer size	1 MB	
Bytes per RBP	9	c
LTE multiplexing mode	FDD	
Tx mode	Single antenna	
Maximum UE Tx power	398 mW	
Pathloss model	ITU-R, Rural Macro	

5.4.4 Simulation Performance of Attention-Based Predictive SPS Scheme

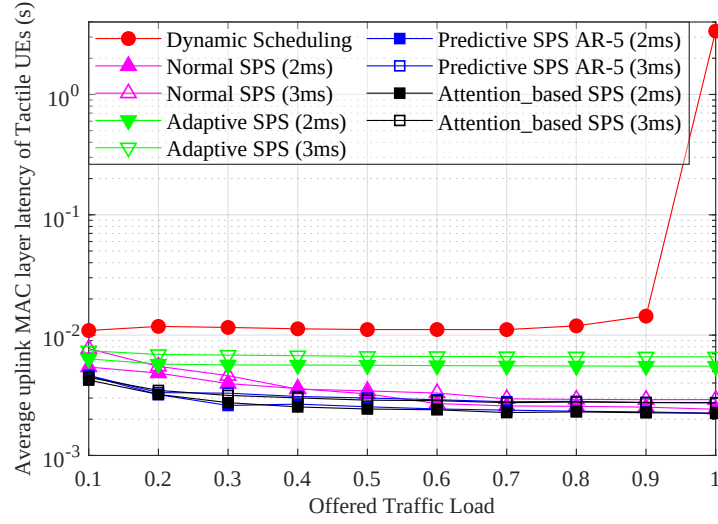


Figure 5.6: Average uplink MAC layer latency of Tactile UEs.

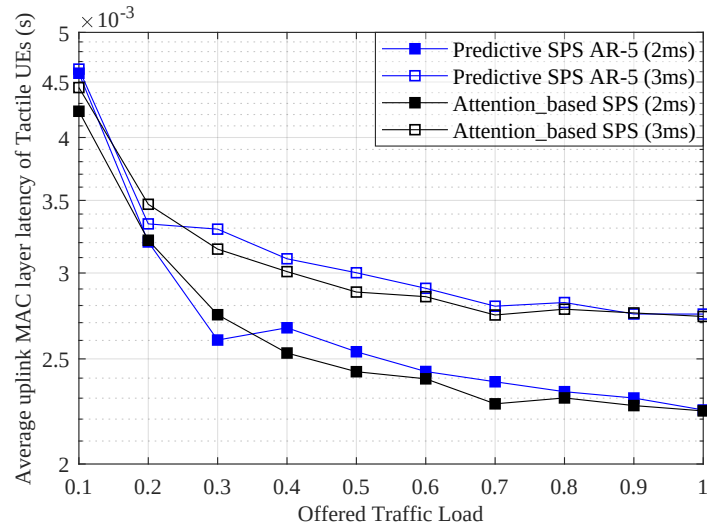


Figure 5.7: Detailed comparison of latency performance between AR-5 model and Attention-based model.

To verify the performance enhancement of predictive SPS scheme, we apply the trained Attention-based model and simulate for 100 seconds under varying traffic loads. The LTE parameters and applied traffic characteristics are summarised in Table 5.7.

5.4.4.1 Latency Performance

The average MAC layer latency performance of Tactile UEs by various uplink transmission protocols is presented in Fig. 5.6. It shows that the achievable uplink latency of conventional DS protocol is around 10 ms latency when traffic loads are under 90%, and it rises drastically when the network is near saturation, leading to almost 3 seconds latency when traffic load is 100%. On the other hand, SPS schemes provide relatively stable latency performance, which is always under 7 ms. It is worth noting that the resulting average uplink latency of Adaptive SPS [140, 141] is higher than that of other SPS schemes. This could result from the “Information Lagging” issue (Section 4.3.2), in which case the latest captured information from BSR does not represent the actual buffer status of connected UEs.

A detailed comparison of uplink latencies between AR-5 model and our proposed Attention-based model is presented in Fig. 5.7. It is observed that the Attention-based mode outperforms AR-5 model in almost all considered traffic loads, except for 0.3 and 0.2 loadings under 2 ms and 3 ms periodicity, respectively. Therefore, our proposed Attention-based model is able to enhance the latency performance of predictive SPS scheme. The average MAC layer uplink latency of both AR-5 and Attention-based model increases for lower traffic loads, this is due to higher prediction error, and a more detailed explanation is presented in the next section.

5.4.4.2 Scheduling and Prediction Performance

The ability of different SPS schemes to allocate sufficient RBPs for uplink transmission is measured by the Normalised Means Square Error (NMSE) of RBP allocation, where lower NMSE indicates less deviation between actual and ideal allocations and hence higher allocation accuracy and efficiency. The simulation results presented in Fig. 5.8 reveal that our proposed Attention-based predictive SPS scheme provides the best allocation accuracy, which explains the improvement of latency performance presented in Fig. 5.6.

The prediction accuracy of buffer occupancy during simulation is shown in Fig. 5.9. In general, the prediction performed by the proposed Attention-based predictive SPS

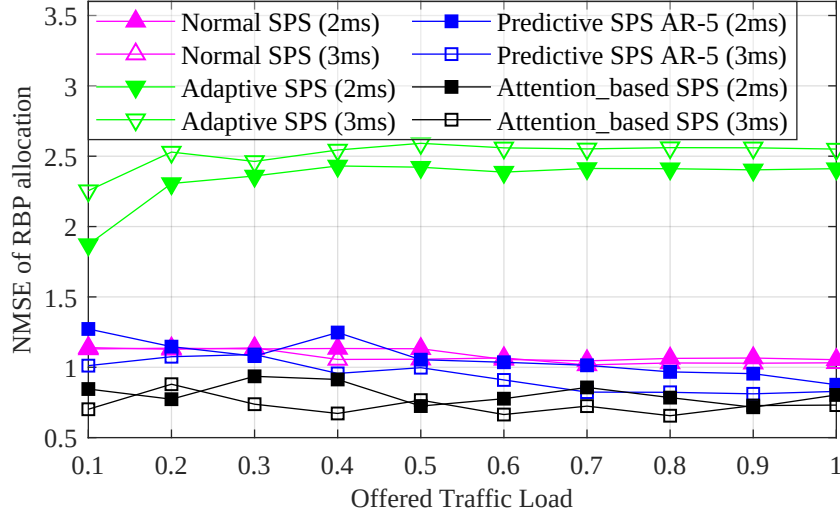


Figure 5.8: NMSE of RBP allocation as a function of total offered traffic load.

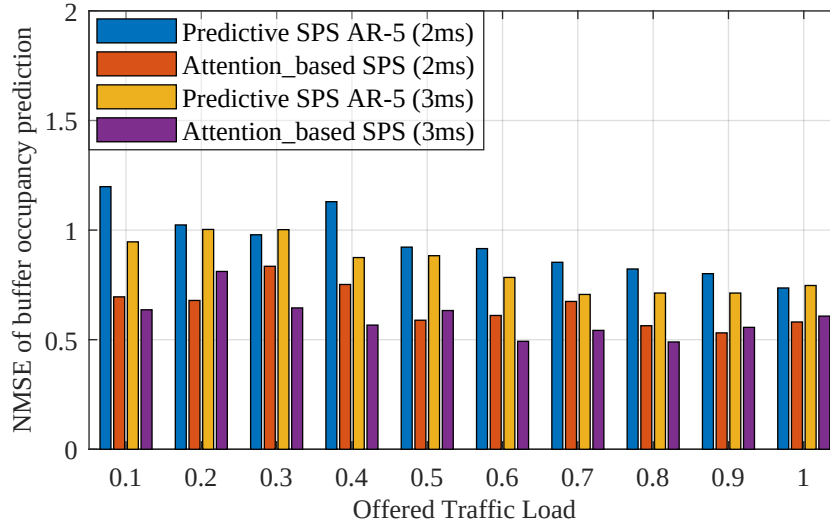


Figure 5.9: NMSE of predicted buffer occupancy as a function of total offered traffic load.

scheme is more accurate than the one by AR-5 model. However, when the operation condition denoted as (load, periodicity (ms)) is (0.2, 3), (0.3, 2), (1.0, 2) and (1.0, 3), the prediction accuracies of these two model are close, and this conforms with the similar latency performance as depicted in Fig. 5.7. Furthermore, the results from Fig. 5.9 and 5.7 indicate an increasing trend of prediction error as traffic load decreases. A possible reason for this trend could be the growing sparsity of data arrivals when traffic is lightly

Table 5.8: Time complexity (3 ms) of executed AR-5 scheduler and proposed Attention-based scheduler.

		Time complexity (ms)									
Traffic load		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
2 ms	AR-5 model	0.0592	0.0595	0.0595	0.0588	0.0592	0.0595	0.0598	0.0608	0.0619	0.0611
	Proposed model	4.8761	5.0803	4.9832	5.1005	4.9589	4.9405	4.9416	5.1503	5.1293	5.0812
3 ms	AR-5 model	0.0552	0.0562	0.0551	0.0560	0.0578	0.0584	0.0584	0.0604	0.0603	0.0579
	Proposed model	4.6766	4.6703	4.7480	4.7607	4.8347	4.8203	4.7131	4.8061	4.9066	4.9403

loaded, in which case temporal dependencies is harder to obtain due to the limited size of input sequences.

5.4.4.3 Time Complexity and Power Consumption

The time complexity measured by calculation time of prediction is listed in Table 5.8, where an AMD Opteron 62xx class CPU is used. It is observed that the computation time of the proposed Attention-based prediction model is about 100 times longer than the AR-5 model. This is expected as the number of parameters and the amount of required computation of Attention-based model is much larger than that of the AR-5 model. Specifically, the largest average time complexity of the proposed Attention-based prediction model is 5.1503 ms. Although such a computation time is unsuitable for SPS periodicity of less than 5 ms, we would expect a reduced time complexity when the proposed model is implemented by devices of 5G and beyond, because these devices offer advanced computation power such that edge intelligence implementing AI can be successfully enabled [26].

Another essential requirement to conduct machine learning for network devices is high power consumption at the user side. One feasible solution to this issue is establishing a default, pre-trained model on individual UEs, while eNodeB is obliged to conduct the full computation of our proposed Attention-based model and provide an updated model for UEs. Since eNodeB already has the full set of required training data by the nature of our predictive SPS scheme, no additional data need to be communicated between eNodeB and UEs. As such, the power consumption and computation requirement at UEs are drastically reduced.

5.5 Conclusions

In this chapter, we present a novel scheduler for predictive SPS scheme using Attention-based recurrent neural network, in order to achieve enhanced latency performance for LTE and 5G networks. Specifically, we interpret the “Information Lagging” issue arising from predictive SPS scheme into a sequence-to-sequence task and design a machine learning model for this task. Different to existing research, our Attention-based model implements dimension expansion at the pre-processing stage and introduces an initial input containing learnable parameters to the decoder. In addition, we study the characteristics of traced data traffic from real-time haptic interaction experiments, and results from the off-line training reveal that our proposed model outperforms other alternative ones. Finally, compared with the AR-5 model described in Section 4.3.2, simulation results prove that our proposed Attention-based predictive SPS scheme is able to further enhance the MAC layer uplink latency for almost all considered traffic loads. Meanwhile, we point out that the prediction accuracy degrades if data is sparse and temporal dependencies are weak.

Conclusions and Future Directions

6.1 Introduction

With the capability to deliver haptic information through the network, the Tactile Internet is envisioned as the future Internet. This is because by wirelessly transmitting haptic, video and audio data in real time, not only will users be provided a true immersive experience, but also emerging services can be realised and successfully supported. These include intelligent transportation system, smart grid and process automation [12]. Nowadays, wireless local area network (WLAN) and mobile cellular network have been widely deployed as solutions to wireless access technology. In particular, they have achieved unprecedented success in serving users with a broadband connection for short and long range scenarios, respectively. Although current WLAN and mobile cellular technologies are sufficient to meet service requirements of today's applications, they may not be viable for the Tactile Internet due to ultra-low end-to-end latency requirement—from 0.5 to 10 ms for various use cases.

The focus of this thesis has therefore been to improve the latency performance of these two wireless access networks by advanced MAC layer designs. This is because among different delay sources contributing to the overall latency in the wireless domain, signal processing and queueing delay are significantly affected by the MAC layer protocol, which has been viewed as the latency bottleneck of a wireless communications network [29]. To this end, we have compared different uplink transmission protocols in the MAC layer of WLAN and mobile cellular network, and we argue that the HCCA protocol of WLAN and SPS scheme of LTE/5G network are viable solutions to both reliability and latency requirements for the Tactile Internet. On the one hand, HCCA does not only perform a polling operation that promises a certain amount of channel time called TXOP

for uplink transmission during CFP, but also allows all stations to contend for channel access during CP. On the other hand, the SPS scheme was proposed to periodically allocate a fixed number of radio resource blocks to uplink UEs, which functions well for CBR traffic but suffers from resource mismatch issue for VBR traffic.

For enhancing latency performance of the HCCA protocol, we proposed a strategic parameter selection algorithm that exploits the derived average uplink latency formula and provides a guideline for the network operator to fine-tune system parameters— $T_{\max}$ and T_{SI} , such that resulting average uplink latency can meet the requirement and this was proved by our numerical evaluations.

In order to alleviate the over/under scheduling issue of the SPS scheme, we proposed a predictive SPS scheme to proactively determine the scheduled radio resources based on the prediction of future network traffic. Meanwhile, we designed a feasible resource allocator (FRA), such that non-SPS UEs and SPS UEs could coexist and preserved-but-unscheduled resources can be reused efficiently. Simulation results reveal that the average uplink latencies under 3.5 ms can be achieved for all considered traffic loads. In addition, we argued that sub-millisecond latency could be achieved if predictive SPS scheme and sTTI were implemented in a 5G network.

Furthermore, we deployed an Attention-based recurrent neural network model as the scheduling entity for the predictive SPS scheme. As a result, we showed that the proposed model achieved the best prediction accuracy and hence further improved the latency performance and scheduling efficiency of the SPS scheme. The following sections summarise the key contributions of this thesis and discuss the future directions of our research.

6.2 Key Contributions

In Chapter 1, we briefly introduced WLAN and mobile cellular networks. We then explained the concept of the Tactile Internet, highlighting the stringent requirements of reliability and latency for different use cases. The chapter then summarised our objectives in reducing uplink latency produced by the MAC layer protocols of WLAN and mobile

cellular networks, due to the technical bottleneck imposed by uplink transmission. Finally, the chapter outlined how objectives were achieved via our technical contributions, followed by our published research work.

The motivation behind our technical contributions lies in improving the latency of existing MAC layer protocols of WLAN and mobile cellular networks. Therefore, in Chapter 2, we critically analysed and compared existing work proposed for latency enhancement. In particular, the chapter first provided a detailed description of the MAC protocols proposed for uplink transmission in a WLAN supported by IEEE 802.11 technologies, and these include DCF, EDCA and HCCA protocols. Then, a comprehensive literature review on the improvement of throughput, fairness and latency of WLAN was presented. The existing solutions to latency reduction for a WLAN can be classified into three categories—medium access, admission control and bandwidth allocation mechanism. By advanced medium access mechanism, information of exact uplink bandwidth demand is acquired by additional handshake procedure or modified control frame field. As for admission control methods, service differentiation could be achieved by determining the amount of permitted traffic flows such that QoS of high priority flows are protected. The rest of the existing solutions designed advanced allocation mechanisms to dynamically predict or estimate the bandwidth demand of a specific user, such that pending data can be transmitted immediately. It is important to note that most of the existing work has targeted on throughput, while whether WLAN is suitable for ultra-low latency was not addressed thoroughly. Furthermore, we argued that existing research had studied the HCCA simulatively, lacking an understanding of the implications of various network parameters on the latency performance.

Further in Chapter 2, general technical aspects of LTE networks were introduced, with a focus of describing the MAC layer uplink transmission protocols—DS, CBS and SPS. Existing work for enhanced MAC layer schedulers designed for the DS was then presented, and it can be classified into two categories—prioritisation-based and optimisation-based. The prioritisation-based schedulers attempt to sort connected users according to measured/estimated requirement/urgency metrics in order to provide a best-effort service for latency-critical applications. On the other hand, optimisation-based schedulers de-

termine the position and number of allocated radio resources by maximising a utility function with energy, latency, fairness being constraints. It should be mentioned that due to the characteristics of SC-FDMA for uplink transmissions, the optimisation problem is usually NP-Hard [112]. Therefore, these formulated problems were often addressed by heuristic algorithms. Chapter 2 further discussed the innovation brought by 5G technologies regarding latency performance. Specifically, we highlighted that K-rep HARQ and NOMA are feasible techniques to enhance reliability and latency performance of CBS, whereas sTTI is a feasible solution to bring latency of SPS to a sub-millisecond level.

In Chapter 3, we presented our initial technical contributions, the first closed-form expression for the average uplink latency of HCCA and a parameter selection algorithm for controlling the uplink latency of HCCA. The HCCA protocol is equivalent to a polling system with resource reservation dedicated for admitted latency-critical traffic flows. In Chapter 2, we identified that existing research lacked analytical studies of the impact of network parameters on the uplink latency of HCCA. The expression was derived by modelling the operation of a reference design of HCCA [56] with a multi-user M/G/1 queueing system. Moreover, we deployed an estimation framework and a Markov Chain method such that numerical values of average latency obtained by the expression matched closely with the results of discrete-event simulations. Then, global sensitivity analyses were conducted to acknowledge the importance of various network parameters on the uplink latency, which filled the gap of existing literature. Specifically, we analysed eight parameters and argued that $T_{\max}$ and T_{SI} have the most significant impact on latency, whereas the payload size of transmitted packets and SIFS period exhibited increasing importance for low-speed WLAN networks. Results also revealed that data compression techniques could serve as an alternative solution to latency reduction by controlling the generated data arrival rate at the application layer.

Further in Chapter 3, we proposed a strategic parameter selection algorithm that utilises the gained knowledge from the global sensitivity analyses, and it can iteratively fine-tune $T_{\max}$ and T_{SI} such that a predefined latency budget could be met. We numerically verified the algorithm on an HCCA network against nine different starting conditions. Results showed that our proposed algorithm could successfully help HCCA to

meet the 0.1 ms latency requirement. To this point, we argue that a WLAN implementing IEEE 802.11 HCCA MAC protocol is able to support the Tactile Internet using reference design, provided that network parameters were chosen properly.

Starting from Chapter 4, we turned our attention into another widely deployed wireless communications technology—mobile cellular network, for supporting the Tactile Internet. As analysed in Chapter 2, the original SPS scheme was introduced in LTE networks to provide quality service for VoIP by periodically allocating a fixed amount of radio resources to connected users. However, it could suffer from over/under scheduling issue, which impairs either scheduling efficiency or latency performance of an LTE network. Therefore, we proposed a predictive SPS scheme for latency-critical services of the Tactile Internet. First, we proposed a feasible resource allocator (FRA) that does not only support variable allocation for users connected using SPS scheme but also allows re-allocating unused resource for non-SPS users, by which improved scheduling efficiency can be achieved. Second, we deployed a simple prediction model that utilises recorded history observations to estimate the buffer occupancy of a particular uplink device. Such a predictive approach raised the probability of transmitting pending data immediately. We verified our predictive SPS scheme by system-level simulations, and results showed that a sub-3.5 ms uplink latency could be achieved for considered traffic loads when 1 ms periodicity was implemented.

Another equally important finding of the evaluation for predictive SPS scheme in Chapter 4 was that latency performance and prediction accuracy deteriorated with experienced traffic load, which indicated that the simple auto-regressive model might fail to capture the fast-changing characteristics of data traffic when the network was heavily-loaded. This has become the motivation behind our novel Attention-based recurrent neural network proposed for enhancing the performance of predictive SPS scheme. First, we interpreted the prediction task into a sequence-to-sequence problem and deployed a recurrent neural network architecture to record and represent the long-term dependencies within the input sequences. Then, inspired by the recent success in natural language translation [51], we proposed to use an additive attention mechanism to perform accurate prediction for our specific sequence-to-sequence problem. Most importantly, we imple-

mented dimension expansion at data pre-processing stage and introduced an initial input containing learnable parameters to the decoder, in order to improve prediction accuracy with extra flexibility. Results from off-line training showed that our proposed Attention-based model achieved the highest prediction accuracy among all considered models and simulation results reveal that both scheduling efficiency and latency performance were enhanced when the Attention-based model was utilised.

It is important to note that simulations conducted in Chapter 4 and 5 were based on a well-known discrete-event simulation platform OMNeT++ and SimuLTE [54]. The data traffic from the application layer was generated by processing traced data of haptic-tactile experiments and matching it to various traffic loads of an LTE network. This represented our best effort to model the data traffic of actual haptic-tactile applications and therefore, to increase the accountability of evaluation of our proposed SPS scheme.

6.3 Future Directions

As highlighted in previous sections, technical contributions of this thesis consist of a comprehensive study of IEEE 802.11 HCCA protocol and LTE/5G SPS protocol, as solutions to achieving ultra-low latency for the Tactile Internet and emerging latency-critical services. During this study, the characteristics of actual network traffic generated by haptic-tactile applications have been discussed but could not be pursued due to the limitation of our research scope. This section presents related areas that need further investigation.

6.3.1 Key Assumptions

For the technical contributions presented in Chapter 3, we have assumed that the packet arrival of a haptic-tactile application of the Tactile Internet follows a Poisson process. The Poisson process is one of the most dominant and oldest models used for analysing traffic in traditional telephony networks, and it is an appropriate model for packets generated from many independent sources [157]. Further, the Poisson process is characterised as a memoryless renewal process and the inter-arrival time of a Poisson process has an exponential distribution, which means the arrival of an event in a Poisson process is inde-

pendent of the others. This has made the Poisson process a reasonable traffic model for the network illustrated in Fig. 3.1, where a Hub receives data from independent actuators/sensors and forwards the received data to an access point in the uplink direction. However, the traffic model from an actual haptic-tactile application may be different (e.g. Normal distribution or Pareto distribution), which will complicate the mathematical analysis of latency derivation. In such cases, phase-type distribution combined with a Markov chain model shall be used, and latency distribution could be obtained instead of a closed-form expression of average latency [158, 159]. Nevertheless, it will be worth investigating HCCA against different traffic models with respect to various performance metrics, such as packet drop rate, throughput, average latency, probability of latency violation.

With the availability of generated haptic data from haptic-tactile experiments [146], we have enhanced the validity of the numerical evaluation of our proposed predictive SPS scheme in Chapter 4. To closely simulate the actual data traffic of haptic-tactile applications, we have assumed that the deadband principal with a 5% JND was employed as the default haptic codecs. The reason for such an assumption was two-fold. First, the deadband principal was introduced as a promising data compression technique for haptic data [125, 127]. Second, we selected a 5% JND because it ensures a maximum sensitivity for force/position difference and can provide the highest accuracy for actuators in a haptic-tactile application such as TelePresence and TeleAction (TPTA). However, alternative haptic codecs or different JND values could be implemented during an actual haptic-tactile application, which could affect the latency performance of the predictive SPS. To this end, a collaboration between haptic codecs developers and haptic-tactile application providers will be needed in order to determine the final version of haptic codecs and hence the resulting form of generated traffic model.

6.3.2 Reinforcement Learning for Intelligent Resource Scheduling

In recent years, reinforcement learning has attracted network designers' attention as an approach to make intelligent decisions based on measurement/observation of the external environment. For example, reinforcement learning has been used to solve the combi-

natorial optimisation problem in order to obtain a scheduling policy for the best performance [160, 161]. In particular, by observing the immediate reward produced by taking a certain action, a reinforcement learning-based scheduler is able to learn the long-term effect of such an action progressively. Therefore, by using the gained knowledge from previous interactions, the scheduler can achieve the best gain when it encounters a similar environment. The concept of reinforcement learning was also applied in [162], where a reinforcement learning-based scheduler was trained to reach the minimum sum cost of offloading resources onto mobile edge computing (MEC) units. Authors have expressed the immediate reward as a weighted sum of computation time and energy consumption, such that these two metrics could be catered simultaneously.

To this end, a reinforcement learning-based scheduler could be designed for predictive SPS, such that the widely known trade-off between throughput and latency in wireless communications [107] could be properly addressed. Specifically, for our proposed SPS scheme, the number of allocated radio resources was determined as the minimum number that is able to transmit all pending data. Such an approach assures the best achievable latency with possible resource and energy waste. However, by allowing the scheduler to consider the weighted sum of latency, scheduling efficiency and energy consumption in a reinforcement learning algorithm, better allocation decisions could be performed.

6.3.3 Predictive SPS Scheme Implemented on 5G Networks

In Chapters 4 and 5, we have proposed and evaluated the performance of predictive SPS scheme under traced traffic from real-time haptic-tactile experiments. We have highlighted that the uplink latency of LTE networks could be brought down to sub-3.5 ms level due to the restriction of 1 ms TTI, while further latency reduction could be achieved by deploying 5G networks, where sTTI could be as short as 0.14 ms. Therefore, developing a universally-recognised and open-source framework for simulating the 5G network at a system level [163] will be vital for verifying innovative concepts and techniques proposed for enhanced performance. Such an effort will have significant contributions to both academics and industries, considering that 5G technologies cover numerous techni-

cal contributions and are undergoing further research and development.

6.4 Conclusions

The vision of the Tactile Internet and emerging services such as tele-rehabilitation and smart grid requires the communications network to transmit data with stringent requirements. Currently deployed wireless communication networks, however, are not able to support services with ultra-low latency and ultra-high reliability. This chapter summarises our motivation and technical contributions to the MAC layer protocols of WLAN and mobile cellular networks. In this thesis, uplink latency reduction solutions were proposed and evaluated. As for WLAN, motivated by the QoS capability provided by the polling mechanism of HCCA, we have derived the first closed-form expression for the uplink latency of HCCA. A strategic parameter selection algorithm was then proposed to fine-tune network parameters of HCCA in order to control the uplink latency within a predefined limit. Instead of studying the performance of existing MAC protocol, our proposed solution to the mobile cellular network incorporates variable resource allocation and prediction method for the SPS scheme. Such a predictive scheme alleviates the shortcomings induced by original SPS protocol, such that radio resources could be proactively allocated to SPS users, and the unused sources can be shared between non-SPS users. We have further enhanced the proposed SPS scheme by implementing an Attention-based recurrent neural network as the prediction mechanism, such that resource allocation could be performed with improved accuracy. Finally, in this chapter, we discussed the future directions of our research work and they include extending the investigation of data traffic generated by haptic-tactile applications, implementation of reinforcement learning for intelligent decision-making and evaluating the performance of the predictive SPS scheme for 5G networks.

Bibliography

- [1] G. Shen, R. S. Tucker, and C. Chae, "Fixed Mobile Convergence Architectures for Broadband Access: Integration of EPON and WiMAX [Topics in Optical Communications]," *IEEE Communications Magazine*, vol. 45, no. 8, pp. 44–50, August 2007.
- [2] N. Ghazisaidi, M. Maier, and C. M. Assi, "Fiber-Wireless (FiWi) Access Networks: A Survey," *IEEE Communications Magazine*, vol. 47, no. 2, pp. 160–167, February 2009.
- [3] S. Sarkar, S. Dixit, and B. Mukherjee, "Hybrid Wireless-Optical Broadband-Access Network (WOBAN): A Review of Relevant Challenges," *Journal of Lightwave Technology*, vol. 25, no. 11, pp. 3329–3340, Nov 2007.
- [4] G. R. Hiertz, D. Denteneer, L. Stibor, Y. Zang, X. P. Costa, and B. Walke, "The IEEE 802.11 Universe," *IEEE Communications Magazine*, vol. 48, no. 1, pp. 62–70, January 2010.
- [5] Y. He, M. Chen, B. Ge, and M. Guizani, "On WiFi Offloading in Heterogeneous Networks: Various Incentives and Trade-Off Strategies," *IEEE Communications Surveys Tutorials*, vol. 18, no. 4, pp. 2345–2385, Fourthquarter 2016.
- [6] H. Zhou, H. Wang, X. Li, and V. C. M. Leung, "A Survey on Mobile Data Offloading Technologies," *IEEE Access*, vol. 6, pp. 5101–5111, 2018.
- [7] P. E. Clint Smith and C. Daniel, *Wireless Networks: Design and Integration for LTE, EVDO, HSPA, and WiMAX*, 3rd ed. McGraw-Hill Education, 2014, ch. 10, pp. 201–213.

- [8] S. Sampaio, P. Souto, and F. Vasques, "A review of scalability and topological stability issues in IEEE 802.11s wireless mesh networks deployments," *International Journal of Communication Systems*, vol. 29, no. 4, pp. 671–693, 2016.
- [9] OPENSIGNAL, "The State of LTE(February 2018)," <https://www.opensignal.com/reports/2018/02/state-of-lte>, accessed: 2019-08-02.
- [10] E. Dahlman, S. Parkvall, and J. Sköld, "Chapter 1 - Introduction," in *4G LTE-Advanced Pro and The Road to 5G*, third edition ed., E. Dahlman, S. Parkvall, and J. Sköld, Eds. Academic Press, 2016, pp. 1 – 5.
- [11] M. Maier, M. Chowdhury, B. P. Rimal, and D. P. Van, "The Tactile Internet: Vision, Recent Progress, and Open Challenges," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 138–145, May 2016.
- [12] G. P. Fettweis, "The Tactile Internet: Applications and Challenges," *IEEE Vehicular Technology Magazine*, vol. 9, no. 1, pp. 64–70, March 2014.
- [13] *One-way Transmission Time*, TELECOMMUNICATION STANDARDIZATION SECTOR OF ITU (ITU-T) ITU-T Recommendation G.114, May 2==3.
- [14] 3GPP, "Service Requirements for Next Generation New Services and Markets," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 22.261, 09 2018, version 16.5.0. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3107>
- [15] H. Chen, R. Abbas, P. Cheng, M. Shirvanimoghaddam, W. Hardjawana, W. Bao, Y. Li, and B. Vucetic, "Ultra-Reliable Low Latency Cellular Networks: Use Cases, Challenges and Approaches," *IEEE Communications Magazine*, vol. 56, no. 12, pp. 119–125, December 2018.
- [16] G. Fettweis, H. Boche, T. Wiegand, E. Zielinski *et al.*, "The Tactile Internet," International Telecommunication Union, Tech. Rep., August 2014. [Online]. Available: <https://www.itu.int/oth/T2301000023/en>

- [17] V. Gazis, "A Survey of Standards for Machine-to-Machine and the Internet of Things," *IEEE Communications Surveys Tutorials*, vol. 19, no. 1, pp. 482–511, Firstquarter 2017.
- [18] V. Gungor and F. Lambert, "A survey on communication networks for electric system automation," *Computer Networks*, vol. 50, no. 7, pp. 877 – 897, 2006. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1389128606000193>
- [19] S. Zeadally, J. Guerrero, and J. Contreras, "A tutorial survey on vehicle-to-vehicle communications," *Telecommunication Systems*, vol. 73, no. 3, pp. 469–489, 2020.
- [20] E. Benalia, S. Bitam, and A. Mellouk, "Data dissemination for Internet of vehicle based on 5G communications: A survey," *Transactions on Emerging Telecommunications Technologies*, vol. 31, no. 5, 2020.
- [21] B. Ji, X. Zhang, S. Mumtaz, C. Han, C. Li, H. Wen, and D. Wang, "Survey on the Internet of Vehicles: Network Architectures and Applications," *IEEE Communications Standards Magazine*, vol. 4, no. 1, pp. 34–41, 2020.
- [22] Y. Kabalci, "A survey on smart metering and smart grid communication," *Renewable and Sustainable Energy Reviews*, vol. 57, pp. 302 – 318, 2016. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1364032115014975>
- [23] P. Siano, "Demand response and smart gridsA survey," *Renewable and Sustainable Energy Reviews*, vol. 30, pp. 461 – 478, 2014. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1364032113007211>
- [24] M. Erol-Kantarci and H. T. Mouftah, "Energy-Efficient Information and Communication Infrastructures in the Smart Grid: A Survey on Interactions and Open Issues," *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 179–197, 2015.
- [25] X. Fang, S. Misra, G. Xue, and D. Yang, "Smart Grid The New and Improved Power Grid: A Survey," *IEEE Communications Surveys Tutorials*, vol. 14, no. 4, pp. 944–980, Fourth 2012.

- [26] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5G-Enabled Tactile Internet," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 460–473, March 2016.
- [27] S. Ullah, H. Higgins, B. Braem, B. Latre, C. Blondia, I. Moerman, S. Saleem, Z. Rahman, and K. Kwak, "A comprehensive survey of wireless body area networks on PHY, MAC, and network layers solutions," *Journal of Medical Systems*, vol. 36, no. 3, pp. 1065–1094, 2012.
- [28] K. Antonakoglou, X. Xu, E. Steinbach *et al.*, "Toward haptic communications over the 5g tactile internet," *IEEE Communications Surveys Tutorials*, vol. 20, no. 4, pp. 3034–3059, Fourthquarter 2018.
- [29] Z. Tan, Y. Li, Q. Li, Z. Zhang, Z. Li, and S. Lu, "Supporting Mobile VR in LTE Networks: How Close Are We?" *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 2, no. 1, pp. 8:1–8:31, Apr. 2018.
- [30] Chih-Ping Li, Jing Jiang, W. Chen, Tingfang Ji, and J. Smee, "5G Ultra-Reliable and Low-Latency Systems Design," in *2017 European Conference on Networks and Communications (EuCNC)*, June 2017, pp. 1–5.
- [31] M. A. Gray, "Discrete Event Simulation: A Review of SimEvents," *Computing in Science Engineering*, vol. 9, no. 6, pp. 62–66, Nov 2007.
- [32] E. Wong, M. P. I. Dias, and L. Ruan, "Predictive Resource Allocation for Tactile Internet Capable Passive Optical LANs," *Journal of Lightwave Technology*, vol. 35, no. 13, pp. 2629–2641, 2017.
- [33] P. Ansel, Q. Ni, and T. Turletti, "FHCF: A Simple and Efficient Scheduling Scheme for IEEE 802.11e Wireless LAN," *Mobile Networks and Applications*, vol. 11, no. 3, pp. 391–403, 2006.
- [34] D. Skyrianoglou, N. Passas *et al.*, "ARROW: An Efficient Traffic Scheduling Algorithm for IEEE 802.11e HCCA," *IEEE Transactions on Wireless Communications*, vol. 5, no. 12, pp. 3558–3567, 2006.

- [35] W. . Kuo, "Traffic Scheduling for Multimedia Transmission over IEEE 802.11e Wireless LAN," *IET Communications*, vol. 2, no. 1, pp. 92–97, January 2008.
- [36] G. Cecchetti, A. L. Ruscelli, A. Mastropaolo, and G. Lipari, "Dynamic TXOP HCCA Reclaiming Scheduler with Transmission Time Estimation for IEEE 802.11e Real-time Networks," in *Proceedings of the 15th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, ser. MSWiM '12. New York, NY, USA: ACM, 2012, pp. 239–246.
- [37] G. Boggia, P. Camarda, L. A. Grieco *et al.*, "Feedback-Based Control for Providing Real-Time Services With the 802.11e MAC," *IEEE/ACM Transactions on Networking*, vol. 15, no. 2, pp. 323–333, April 2007.
- [38] A. Grilo, M. Macedo, and M. Nunes, "A Scheduling Algorithm for QoS Support in IEEE802.11 Networks," *IEEE Wireless Communications*, vol. 10, no. 3, pp. 36–43, June 2003.
- [39] D.-Y. Lee, S.-R. Kim, and C.-W. Lee, "An Enhanced EDD QoS Scheduler for IEEE 802.11e WLAN," in *Advances in Computational Science and Engineering*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 45–59.
- [40] J. Huang, Y. Chen, and D. Shiung, "A Four-Way-Polling QoS Scheduler for IEEE 802.11e HCCA," in *TENCON 2010 - 2010 IEEE Region 10 Conference*, Nov 2010, pp. 1986–1991.
- [41] Y.-S. Chen, Y.-W. Lee, and J. H. Park, "Enhanced HCCA Mechanism for Multimedia Traffics with QoS Support in IEEE 802.11e Networks," *Journal of Network and Computer Applications*, vol. 34, no. 5, pp. 1566 – 1571, 2011, dependable Multimedia Communications: Systems, Services, and Applications.
- [42] M. Al-Maqri, M. Othman, B. Ali, and Z. Hanapi, "Feasible HCCA Polling Mechanism for Video Transmission in IEEE 802.11e WLANs," *Wireless Personal Communications*, vol. 85, no. 3, pp. 899–924, 2015.

- [43] A. Aijaz, "Toward Human-in-the-Loop Mobile Networks: A Radio Resource Allocation Perspective on Haptic Communications," *IEEE Transactions on Wireless Communications*, vol. 17, no. 7, pp. 4493–4508, July 2018.
- [44] S. Liu, M. Li, X. Xu, E. Steinbach, and Q. Liu, "QoE-Driven Uplink Scheduling for Haptic Communications Over 5G Enabled Tactile Internet," in *2018 IEEE International Symposium on Haptic, Audio and Visual Environments and Games (HAVE)*, Sep. 2018, pp. 1–5.
- [45] C. She, C. Yang, and T. Q. S. Quek, "Joint Uplink and Downlink Resource Configuration for Ultra-Reliable and Low-Latency Communications," *IEEE Transactions on Communications*, vol. 66, no. 5, pp. 2266–2280, May 2018.
- [46] S. Oulaourf, A. Haidine, A. Aqqal, and H. Ouahmane, "Review on Radio Resource Allocation Optimization in LTE/LTE-Advanced Using Game Theory," *International Journal of Communication Networks and Information Security*, vol. 9, no. 1, pp. 117–156, 2017.
- [47] "R2-154191, Contention based uplink transmission," Huawei, HiSilicon, Technical Report, October 2015.
- [48] "R2-100125, Impacts of contention based uplink in RAN2," Ericsson, ST-Ericsson, Technical Report, January 2010.
- [49] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," *CoRR*, vol. abs/1409.3215, 2014. [Online]. Available: <http://arxiv.org/abs/1409.3215>
- [50] C. M. Bishop, *Pattern Recognition and Machine Learning*, ser. Information science and statistics. Springer, 2006.
- [51] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," 2015.
- [52] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. W. Cottrell, "A Dual-Stage Attention-Based Recurrent Neural Network for Time Series

- Prediction,” in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 2627–2633. [Online]. Available: <https://doi.org/10.24963/ijcai.2017/366>
- [53] Y. Liang, S. Ke, J. Zhang, X. Yi, and Y. Zheng, “GeoMAN: Multi-level Attention Networks for Geo-sensory Time Series Prediction,” in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, 7 2018, pp. 3428–3434. [Online]. Available: <https://doi.org/10.24963/ijcai.2018/476>
- [54] A. Virdis *et al.*, “Simulating LTE/LTE-advanced networks with simuLTE,” *Advances in Intelligent Systems and Computing*, vol. 402, pp. 83–105, 2015.
- [55] MATLAB, *version 9.3 (R2017b)*. Natick, Massachusetts: The MathWorks Inc., 2017.
- [56] IEEE, “IEEE Standard for Information technology Telecommunications and information exchange between systems Local and metropolitan area networks Specific requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications,” *IEEE Std 802.11-2016 (Revision of IEEE Std 802.11-2012)*, pp. 1–3534, Dec 2016.
- [57] Q. Ni, L. Romdhani, and T. Turletti, “A survey of QoS Enhancements for IEEE 802.11 Wireless LAN,” *Wireless Communications and Mobile Computing*, vol. 4, no. 5, pp. 547–566, 2004.
- [58] N. Ghazisaidi and M. Maier, “Fiber-Wireless (FiWi) Access Networks: Challenges and Opportunities,” *IEEE Network*, vol. 25, no. 1, pp. 36–42, January 2011.
- [59] D. Pham Van, B. P. Rimal, M. Maier, and L. Valcarenghi, “Design, Analysis, and Hardware Emulation of a Novel Energy Conservation Scheme for Sensor Enhanced FiWi Networks (ECO-SFiWi),” *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1645–1662, May 2016.
- [60] K. Al Agha, G. Pujolle, and T. Ali-Yahiya, *Mobile and Wireless Networks*, 2016, vol. 2.

- [61] O. Tickoo and B. Sikdar, "Queueing Analysis and Delay Mitigation in IEEE 802.11 Random Access MAC Based Wireless Networks," in *IEEE INFOCOM 2004*, vol. 2, March 2004, pp. 1404–1413 vol.2.
- [62] L. Hai and S. Taka, "Collision Probability in Saturated IEEE 802.11 Networks," *Australian Telecommunication Networks & Applications Conference (ATNAC)*, 2012.
- [63] O. Tickoo and B. Sikdar, "Modeling Queueing and Channel Access Delay in Unsaturated IEEE 802.11 Random Access MAC Based Wireless Networks," *IEEE/ACM Transactions on Networking*, vol. 16, no. 4, pp. 878–891, Aug 2008.
- [64] V. Gupta, M. Gong, S. Dharmaraja, and C. Williamson, "Analytical Modeling of Bidirectional Multi-Channel IEEE 802.11 MAC Protocols," *International Journal of Communication Systems*, vol. 24, no. 5, pp. 647–665, 2011.
- [65] A. Bozkurt, "Optimal Delay Analysis for Real-Time Traffics over IEEE 802.11 Wireless LANs," *EURASIP Journal on Wireless Communications and Networking*, vol. 2016, no. 1, p. 52, 2016. [Online]. Available: <https://doi.org/10.1186/s13638-016-0545-0>
- [66] G. Bianchi, "IEEE 802.11-saturation throughput analysis," *IEEE Communications Letters*, vol. 2, no. 12, pp. 318–320, 1998.
- [67] ———, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535–547, 2000.
- [68] T. Sakurai and H. L. Vu, "MAC Access Delay of IEEE 802.11 DCF," *IEEE Transactions on Wireless Communications*, vol. 6, no. 5, pp. 1702–1710, May 2007.
- [69] S.-T. Cheng and M. Wu, "Adaptive coordination function for IEEE 802.11 wireless LANs," *Wireless Personal Communications*, vol. 43, no. 2, pp. 743–757, Oct 2007.
- [70] S. Choi, "PCF vs. DCF: Limitations and Trends," *IEEE 802.11-01/154*, 2001.
- [71] X. Yu, P. Navaratnam, and K. Moessner, "Resource Reservation Schemes for IEEE 802.11-Based Wireless Networks: A Survey," *IEEE Communications Surveys Tutorials*, vol. 15, no. 3, pp. 1042–1061, Third 2013.

- [72] S. Mangold, S. Choi, P. May, O. Klein, G. Hiertz, and L. Stibor, "IEEE 802.11e Wireless LAN for Quality of Service," *IEEE 802.11e Wireless LAN for Quality of Service*, 2018.
- [73] M. A. Visser and M. El Zarki, "Voice and Data Transmission over an 802.11 Wireless Network," in *Proceedings of 6th International Symposium on Personal, Indoor and Mobile Radio Communications*, vol. 2, Sep. 1995, pp. 648–652 vol.2.
- [74] P. Garg, R. Doshi, R. Greene, M. Baker, M. Malek, and Xiaoyan Cheng, "Using IEEE 802.11e MAC for QoS over Wireless," in *Conference Proceedings of the 2003 IEEE International Performance, Computing, and Communications Conference, 2003.*, April 2003, pp. 537–542.
- [75] L. Romdhani, Qiang Ni, and T. Turletti, "Adaptive EDCF: Enhanced Service Differentiation for IEEE 802.11 Wireless Ad-Hoc Networks," in *2003 IEEE Wireless Communications and Networking, 2003. WCNC 2003.*, vol. 2, March 2003, pp. 1373–1378 vol.2.
- [76] X. Chen, H. Zhai, X. Tian, and Y. Fang, "Supporting QoS in IEEE 802.11e Wireless LANs," *IEEE Transactions on Wireless Communications*, vol. 5, no. 8, pp. 2217–2227, Aug 2006.
- [77] I. Inan, F. Keceli, and E. Ayanoglu, "Analysis of the 802.11e Enhanced Distributed Channel Access Function," *IEEE Transactions on Communications*, vol. 57, no. 6, pp. 1753–1764, June 2009.
- [78] R. Costa, P. Portugal, F. Vasques, C. Montez, and R. Moraes, "Limitations of the IEEE 802.11 DCF, PCF, EDCA and HCCA to handle real-time traffic," in *2015 IEEE 13th International Conference on Industrial Informatics (INDIN)*, 2015, pp. 931–936.
- [79] P. Serrano, A. Banchs, P. Patras, and A. Azcorra, "Optimal Configuration of 802.11e EDCA for Real-Time and Data Traffic," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 5, pp. 2511–2528, 2010.

- [80] G. Cecchetti and A. Ruscelli, "Real-time support for HCCA function in IEEE 802.11e networks: A performance evaluation," *Security and Communication Networks*, vol. 4, no. 3, pp. 299–315, 2011.
- [81] Shou-Chih Lo, Guanling Lee, and Wen-Tsuen Chen, "An Efficient Multipolling Mechanism for IEEE 802.11 Wireless LANs," *IEEE Transactions on Computers*, vol. 52, no. 6, pp. 764–778, June 2003.
- [82] Y.-S. Chen, Y.-W. Lee, and J. H. Park, "Enhanced HCCA Mechanism for Multimedia Traffics with QoS Support in IEEE 802.11e Networks," *Journal of Network and Computer Applications*, vol. 34, no. 5, pp. 1566 – 1571, 2011, dependable Multimedia Communications: Systems, Services, and Applications.
- [83] A. Saif, M. Othman, S. Subramaniam, and N. AbdulHamid, "Impact of Aggregation Headers on Aggregating Small MSDUs in 802.11n WLANs," in *2010 International Conference on Computer Applications and Industrial Electronics*, Dec 2010, pp. 630–635.
- [84] Yang Xiao and H. Li, "Voice and Video Transmissions with Global Data Parameter Control for the IEEE 802.11e Enhance Distributed Channel Access," *IEEE Transactions on Parallel and Distributed Systems*, vol. 15, no. 11, pp. 1041–1053, Nov 2004.
- [85] Yang Xiao and Haizhon Li, "Evaluation of Distributed Admission Control for the IEEE 802.11e EDCA," *IEEE Communications Magazine*, vol. 42, no. 9, pp. S20–S24, Sep. 2004.
- [86] Ho-Ting Wu, Min-Hua Yang, and Kai-Wei Ke, "The Design of QoS Provisioning Mechanisms for Wireless Networks," in *2010 8th IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, March 2010, pp. 756–759.
- [87] S. Hantrakoon and A. Phonphoem, "Priority Based HCCA for IEEE 802.11e," in *2010 International Conference on Communications and Mobile Computing*, vol. 3, April 2010, pp. 485–489.

- [88] E. Charfi, L. Chaari, and L. Kamoun, "PHY/MAC Enhancements and QoS Mechanisms for Very High Throughput WLANs: A Survey," *IEEE Communications Surveys Tutorials*, vol. 15, no. 4, pp. 1714–1735, Fourth 2013.
- [89] M. A. Al-Maqri, M. A. Alrshah, and M. Othman, "Review on QoS Provisioning Approaches for Supporting Video Traffic in IEEE802.11e: Challenges and Issues," *IEEE Access*, vol. 6, pp. 55 202–55 219, 2018.
- [90] M. M. Rashid, E. Hossain, and V. K. Bhargava, "HCCA Scheduler Design for Guaranteed QoS in IEEE 802.11e Based WLANs," in *2007 IEEE Wireless Communications and Networking Conference*, March 2007, pp. 1538–1543.
- [91] M. A. Al-Maqri, M. Othman, B. M. Ali, and Z. M. Hanapi, "An Efficient HCCA Scheduler for Video Streaming with QoS Support," *Procedia Computer Science*, vol. 109, pp. 116 – 123, 2017, 8th International Conference on Ambient Systems, Networks and Technologies, ANT-2017 and the 7th International Conference on Sustainable Energy Information Technology, SEIT 2017, 16-19 May 2017, Madeira, Portugal.
- [92] "Release 8," <https://www.3gpp.org/specifications/releases/72-release-8>, 3GPP, 2019.
- [93] "Release 10," <https://www.3gpp.org/specifications/releases/70-release-10>, 3GPP, 2019.
- [94] "Release 15," <https://www.3gpp.org/release-15>, 3GPP, 2019.
- [95] E. Dahlman, S. Parkvall, and J. Sköld, "Chapter 5 - Physical Transmission Resources," in *4G LTE-Advanced Pro and The Road to 5G*, third edition ed., E. Dahlman, S. Parkvall, and J. Sköld, Eds. Academic Press, 2016, pp. 1 – 5.
- [96] R. Kwan *et al.*, "Proportional Fair Multiuser Scheduling in LTE," *IEEE Signal Processing Letters*, vol. 16, no. 6, pp. 461–464, June 2009.
- [97] L. Zhang and J. Ma, *Grant-Free Multiple Access Scheme*. Cham: Springer International Publishing, 2019, pp. 515–533.

- [98] S. Andreev, A. Larmo *et al.*, "Efficient Small Data Access for Machine-Type Communications in LTE," in *2013 IEEE International Conference on Communications (ICC)*, June 2013, pp. 3569–3574.
- [99] C. Wang, Y. Chen, Y. Wu, and L. Zhang, "Performance Evaluation of Grant-Free Transmission for Uplink URLLC Services," in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, June 2017, pp. 1–6.
- [100] D. Jiang, H. Wang, E. Malkamaki, and E. Tuomaala, "Principle and Performance of Semi-Persistent Scheduling for VoIP in LTE System," in *2007 International Conference on Wireless Communications, Networking and Mobile Computing*, Sept 2007, pp. 2861–2864.
- [101] J. Seo and V. C. M. Leung, "Performance Modeling and Stability of Semi-Persistent Scheduling with Initial Random Access in LTE," *IEEE Transactions on Wireless Communications*, vol. 11, no. 12, pp. 4446–4456, December 2012.
- [102] "3GPP TR 36.881, Study on latency reduction techniques for LTE," 3GPP, Technical Report, June 2016.
- [103] T. Jacobsen, R. Abreu, G. Berardinelli, K. Pedersen, P. Mogensen, I. Z. Kovacs, and T. K. Madsen, "System Level Analysis of Uplink Grant-Free Transmission for URLLC," in *2017 IEEE Globecom Workshops (GC Wkshps)*, Dec 2017, pp. 1–6.
- [104] A. K. F. Khattab and K. M. F. Elsayed, "Channel-quality Dependent Earliest Deadline Due Fair Scheduling Schemes for Wireless Multimedia Networks," in *Proceedings of the 7th ACM International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, ser. MSWiM '04. New York, NY, USA: ACM, 2004, pp. 31–38.
- [105] F. D. Calabrese, C. Rosa, M. Anas, P. H. Michaelsen, K. I. Pedersen, and P. E. Mogensen, "Adaptive Transmission Bandwidth Based Packet Scheduling for LTE Uplink," in *2008 IEEE 68th Vehicular Technology Conference*, Sep. 2008, pp. 1–5.

- [106] A. S. Lioumpas and A. Alexiou, "Uplink Scheduling for Machine-to-Machine Communications in LTE-Based Cellular Systems," in *2011 IEEE GLOBECOM Workshops (GC Wkshps)*, Dec 2011, pp. 353–357.
- [107] A. E. Gamal, J. Mammen, B. Prabhakar, and D. Shah, "Throughput-Delay Trade-off in Wireless Networks," in *IEEE INFOCOM 2004*, vol. 1, March 2004, p. 475.
- [108] I. Abdalla and S. Venkatesan, "A QoE Preserving M2M-Aware Hybrid Scheduler for LTE Uplink," in *2013 International Conference on Selected Topics in Mobile and Wireless Networking (MoWNeT)*, Aug 2013, pp. 127–132.
- [109] N. Afrin, J. Brown, and J. Y. Khan, "Performance Analysis of an Enhanced Delay Sensitive LTE Uplink Scheduler for M2M Traffic," in *2013 Australasian Telecommunication Networks and Applications Conference (ATNAC)*, Nov 2013, pp. 154–159.
- [110] A. M. Maia, D. Vieira, M. F. de Castro, and Y. Ghamri-Doudane, "A Fair QoS-Aware Dynamic LTE Scheduler for Machine-to-Machine Communication," *Computer Communications*, vol. 89-90, pp. 75 – 86, 2016, internet of Things Research challenges and Solutions.
- [111] S. A. AlQahtani, "Delay-Aware Resource Allocation for M2M Communications Over LTE-A Networks," *Arabian Journal for Science and Engineering*, vol. 44, no. 4, pp. 3639–3653, Apr 2019.
- [112] S. . Lee, I. Pefkianakis, A. Meyerson, S. Xu, and S. Lu, "Proportional Fair Frequency-Domain Packet Scheduling for 3GPP LTE Uplink," in *IEEE INFOCOM 2009*, April 2009, pp. 2611–2615.
- [113] O. Delgado and B. Jaumard, "Scheduling and Resource Allocation in LTE Uplink with a Delay Requirement," in *2010 8th Annual Communication Networks and Services Research Conference*, May 2010, pp. 268–275.
- [114] M. Agiwal, A. Roy, and N. Saxena, "Next Generation 5G Wireless Networks: A Comprehensive Survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 1617–1655, thirdquarter 2016.

- [115] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, "5G: A Tutorial Overview of Standards, Trials, Challenges, Deployment, and Practice," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1201–1221, June 2017.
- [116] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A Survey on Mobile Edge Computing: The Communication Perspective," *IEEE Communications Surveys Tutorials*, vol. 19, no. 4, pp. 2322–2358, Fourthquarter 2017.
- [117] M. Ryden, K. Oh, A. Chandra, and J. Weissman, "Nebula: Distributed Edge Cloud for Data Intensive Computing," in *2014 IEEE International Conference on Cloud Engineering*, March 2014, pp. 57–66.
- [118] S. Mondal, G. Das, and E. Wong, "Efficient Cost-Optimization Frameworks for Hybrid Cloudlet Placement over Fiber-Wireless Networks," *J. Opt. Commun. Netw.*, vol. 11, no. 8, pp. 437–451, Aug 2019.
- [119] S. Mondal, G. Das, and E. Wong, "A Novel Cost Optimization Framework for Multi-Cloudlet Environment over Optical Access Networks," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Dec 2017, pp. 1–7.
- [120] M. Schellmann, P. Weitkemper, E. Lhetkangas, E. Strm, C. Bockelmann, S. B. Slimane, M. Dohler, and T. Nakamura, *The 5G Radio-Access Technologies*. Cambridge University Press, 2016, p. 158207.
- [121] T. Manglayev, R. Kizilirmak, Y. Kho, and N. Hamid, "GPU Accelerated Successive Interference Cancellation for NOMA Uplink with User Clustering," *Wireless Personal Communications*, vol. 103, no. 3, pp. 2391–2400, 2018.
- [122] A. Li, X. Chen, and H. Jiang, "Contention Based Uplink Transmission with NOMA for Latency Reduction," in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, June 2017, pp. 1–6.

- [123] B. Singh, O. Tirkkonen, Z. Li, and M. A. Uusitalo, "Contention-Based Access for Ultra-Reliable Low Latency Uplink Transmissions," *IEEE Wireless Communications Letters*, vol. 7, no. 2, pp. 182–185, April 2018.
- [124] Z. Amjad, A. Sikora, J. Lauffenburger, and B. Hilt, "Latency Reduction in Narrow-band 4G LTE Networks," in *2018 15th International Symposium on Wireless Communication Systems (ISWCS)*, Aug 2018, pp. 1–5.
- [125] E. Steinbach, M. Strese, M. Eid, X. Liu, A. Bhardwaj, Q. Liu, M. Al-Jaafreh, T. Mahmoodi, R. Hassen, A. El Saddik, and O. Holland, "Haptic Codecs for the Tactile Internet," *Proceedings of the IEEE*, vol. 107, no. 2, pp. 447–470, Feb 2019.
- [126] P. Hinterseer *et al.*, "Perception-Based Data Reduction and Transmission of Haptic Data in Telepresence and Teleaction Systems," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 588–597, Feb 2008.
- [127] E. Steinbach, S. Hirche, J. Kammerl, I. Vittorias, and R. Chaudhari, "Haptic Data Compression and Communication," *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 87–96, Jan 2011.
- [128] S. Martirosov and P. Kopecek, "Cyber Sickness in Virtual Reality - Literature Review," 2017, pp. 718–726.
- [129] L. Ruan, M. P. I. Dias, Y. Feng, and E. Wong, "Round-trip Delay Modeling for Smart Body Area Networks (SmartBANs)," *IEEE Communications Letters*, vol. PP, no. 99, pp. 1–1, 2017.
- [130] A. Grilo and M. Nunes, "Performance evaluation of IEEE 802.11e," in *The 13th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, vol. 1, Sep. 2002, pp. 511–517 vol.1.
- [131] S.-L. Tsao, "Extending earliest-due-date scheduling algorithms for wireless networks with location-dependent errors," vol. 1, no. 52 ND, 2000, pp. 223–228.

- [132] G. Cecchetti, A. L. Ruscelli, A. Mastropaolo, and G. Lipari, "Providing Variable TXOP for IEEE 802.11e HCCA Real-Time Networks," in *2012 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2012, pp. 1508–1513.
- [133] D. P. Van, B. P. Rimal, M. Maier, and L. Valcarenghi, "Eco-fiwi: An energy conservation scheme for integrated fiber-wireless access networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 6, pp. 3979–3994, June 2016.
- [134] D. Bertsekas and R. Gallager, *Data Networks (2Nd Ed.)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1992.
- [135] S. Bharati and P. Saengudomlert, "Analysis of mean packet delay for dynamic bandwidth allocation algorithms in epons," *Journal of Lightwave Technology*, vol. 28, no. 23, pp. 3454–3462, Dec 2010.
- [136] S. Kucherenko, S. Tarantola, and P. Annoni, "Estimation of global sensitivity indices for models with dependent variables," *Computer Physics Communications*, vol. 183, no. 4, pp. 937–946, 2012.
- [137] S. Kucherenko, O. V. Klymenko, and N. Shah, "Sobol' Indices for Problems Defined in Non-rectangular Domains," *Reliability Engineering and System Safety*, vol. 167, pp. 218–231, 2017.
- [138] H. Jin, C. Cho, N. Song, and D. K. Sung, "Optimal Rate Selection for Persistent Scheduling with HARQ in Time-Correlated Nakagami-m Fading Channels," *IEEE Transactions on Wireless Communications*, vol. 10, no. 2, pp. 637–647, February 2011.
- [139] J. T. S. Avocanh, M. Abdennebi *et al.*, "A Semi-persistent Scheduling Scheme for Videotelephony Traffics in the Uplink of LTE Networks," in *Proceedings of the 17th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, ser. MSWiM '14. New York, NY, USA: ACM, 2014, pp. 321–325.
- [140] N. Afrin, J. Brown, and J. Y. Khan, "An Adaptive Buffer Based Semi-persistent Scheduling Scheme for Machine-to-Machine Communications over LTE," in *2014*

- Eighth International Conference on Next Generation Mobile Apps, Services and Technologies*, Sept 2014, pp. 260–265.
- [141] N. Afrin *et al.*, “Design of a buffer and channel adaptive LTE semi-persistent scheduler for M2M communications,” in *2015 IEEE International Conference on Communications (ICC)*, June 2015, pp. 5821–5826.
- [142] M. Li, X. Guan, C. Hua, C. Chen, and L. Lyu, “Predictive Pre-allocation for Low-latency Uplink Access in Industrial Wireless Networks,” in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, April 2018, pp. 306–314.
- [143] M. Condoluci, T. Mahmoodi, E. Steinbach, and M. Dohler, “Soft Resource Reservation for Low-Delayed Teleoperation Over Mobile Networks,” *IEEE Access*, vol. 5, pp. 10 445–10 455, 2017.
- [144] N. Sakr, N. D. Georganas, J. Zhao, and X. Shen, “Motion and Force Prediction in Haptic Media,” in *2007 IEEE International Conference on Multimedia and Expo*, July 2007, pp. 2242–2245.
- [145] A. B. K.S. Shanmugan, *Random signals: detection, estimation, and data analysis*. New York:Wiley, 1988.
- [146] I. E. Steinbach, *IEEE P1918.1.1 - Haptic Codecs for the Tactile Internet*, 2016, available: <http://www.lmt.ei.tum.de/en/research/selected-projects/ieee-p191811-haptic-codecs.html>.
- [147] R. Adhikari and R. K. Agrawal, “An introductory study on time series modeling and forecasting,” *CoRR*, vol. abs/1302.6613, 2013.
- [148] S. Zhenqi, Y. Haifeng, C. Xuefen, and L. Hongxia, “Research on uplink scheduling algorithm of massive M2M and H2H services in LTE,” in *IET International Conference on Information and Communications Technologies (IETICT 2013)*, April 2013, pp. 365–369.

- [149] H. Safa, W. El-Hajj, and K. Tohme, "A QoS-Aware Uplink Scheduling Paradigm for LTE Networks," in *2013 IEEE 27th International Conference on Advanced Information Networking and Applications (AINA)*, March 2013, pp. 1097–1104.
- [150] S. J. Russell and P. Norvig, *Artificial Intelligence : A Modern Approach*, ser. Pearson custom library. Pearson, 2014.
- [151] K. Cho, B. van Merriënboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," *CoRR*, vol. abs/1406.1078, 2014. [Online]. Available: <http://arxiv.org/abs/1406.1078>
- [152] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches," *CoRR*, vol. abs/1409.1259, 2014. [Online]. Available: <http://arxiv.org/abs/1409.1259>
- [153] J. Pouget-Abadie, D. Bahdanau, B. van Merriënboer, K. Cho, and Y. Bengio, "Overcoming the Curse of Sentence Length for Neural Machine Translation using Automatic Segmentation," *CoRR*, vol. abs/1409.1257, 2014. [Online]. Available: <http://arxiv.org/abs/1409.1257>
- [154] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: <https://doi.org/10.1162/neco.1997.9.8.1735>
- [155] M. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *CoRR*, vol. abs/1508.04025, 2015.
- [156] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.
- [157] A. Mohammed and A. Agamy, "A survey on the common network traffic sources models," *International Journal of Computer Networks*, vol. 3, no. 2, pp. 103–115, 2011.
- [158] R. Ghazizadeh and P. Fan, "Queuing Analysis of HCCA for Multi-Rate Wireless LANs with Truncated ARQ Protocol," *Wireless Personal Communications*, vol. 55,

- no. 4, pp. 607–630, Dec 2010. [Online]. Available: <https://doi.org/10.1007/s11277-009-9822-0>
- [159] R. Ghazizadeh and P. Fan, “Queuing Analysis for HCCA with Adaptive Modulation Coding over Wireless LANs,” in *2008 11th IEEE International Conference on Communication Technology*, Nov 2008, pp. 139–143.
- [160] I. Coma, S. Zhang, M. E. Aydin, P. Kuonen, Y. Lu, R. Trestian, and G. Ghinea, “Towards 5G: A Reinforcement Learning-Based Scheduling Solution for Data Traffic Management,” *IEEE Transactions on Network and Service Management*, vol. 15, no. 4, pp. 1661–1675, Dec 2018.
- [161] I. Comsa, A. De-Domenico, and D. Ktenas, “QoS-Driven Scheduling in 5G Radio Access Networks - A Reinforcement Learning Approach,” in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Dec 2017, pp. 1–7.
- [162] J. Li, H. Gao, T. Lv, and Y. Lu, “Deep Reinforcement Learning Based Computation Offloading and Resource Allocation for MEC,” in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2018, pp. 1–6.
- [163] M. Mezzavilla, S. Dutta, M. Zhang, M. R. Akdeniz, and S. Rangan, “5G mmWave Module for ns-3 Network Simulator,” *CoRR*, vol. abs/1506.08801, 2015. [Online]. Available: <http://arxiv.org/abs/1506.08801>