Low-Latency Communication over Heterogeneous Fiber-Wireless Networks for Human-to-Machine Applications

Lihua Ruan
ORCID: 0000-0002-9892-5823

Submitted in partial fulfilment of the requirements of the degree of

Doctor of Philosophy

September 2019
Department of Electrical and Electronic Engineering

THE UNIVERSITY OF MELBOURNE
Declaration

This is to certify that:

- the thesis comprises only my original work towards the Ph.D.,

- due acknowledgement has been made in the text to all material used,

- the thesis is less than 100,000 words in length, excluding tables, maps, bibliography and appendices.

Lihua Ruan
September 2019
Preface

This thesis comprises only the original work completed during the PhD period of the student at the University of Melbourne, conducted under the supervision of Professor Elaine Wong and Dr. Pubuduni Imali Dias. The supervisors contributed to insightful technical guidance, comments and discussions. The student independently completed the theoretical analysis, analytical models, simulations, and experiments. The published outcomes of the thesis are detailed in Chapter 1 Section 1.4. For each publication listed in Chapter 1 Section 1.4, the contributions of its associated authors are clarified as follows:

There have been 6 journal and 10 conference publications arising from Chapters 3 – 7 of this thesis. The student is the first and also the primary author of the first 3 publications of Chapter 3, the first 2 publications of Chapter 6, and the publications in Chapters 4, 5, 7. The student was responsible for the planning, execution, simulation, result analysis, and writing up the manuscripts. In this process, the student benefited greatly from the supervisors in meetings and discussions where they provided technical comments and assisted the student in revising the manuscripts. The suggestions and comments from the supervisors were invaluable. Other than the supervisors, the contributions from the co-authors in two of these publications are as follows:

In the 2nd publication of Chapter 6, co-author Sourav was in collaboration with technical discussions and poster presentation at the conference.

In the 1st publication of Chapter 7, co-author Prof. Martin Maier helped review, comment, and proofread the manuscript.

The thesis has not been submitted for other qualifications. All the work towards the thesis was carried out after the enrolment in the degree. No third-party editorial assistance was provided in the preparation of the thesis. The student is jointly sponsored by the University of Melbourne and China Scholarship Council.
Abstract

With the advent of the Tactile Internet, next-generation communication networks will see the emergence of a variety of low-latency human-to-machine (H2M) applications. In these applications, human beings will be able to remotely control and manipulate machines/robots in real-time, and concurrently experience haptic feedback such as tactile and kinetic sensations. Such applications demand stringent low latency in milliseconds in their transmission for effective H2M interactions. Current communication networks need to reduce its latency to support emerging H2M applications. Motivated by the necessity to address the latency challenge, this thesis discusses crucial building blocks to realise Tactile Internet and support H2M applications, and to improve their latency performance with novel solutions. Particularly, in this thesis, we emphasise the importance of wireless body area network (WBANs) and heterogeneous optical and wireless access networks for H2M applications. In these networks, the medium access control (MAC) layer impacts the latency significantly as it determines the way bandwidth resource is utilised. As such, we focus on the MAC layer designs in the above-mentioned networks, comprehensively investigate existing MAC layer solutions, and proposes novel solutions for low-latency H2M applications.

WBANs, comprising sensors and actuators on or around the human body, is essential for personal area H2M application delivery. For WBANs, we overview the evolution and standardisation of WBAN systems. Among existing standards, we pay special interest to the ETSI smart body area network (SmartBAN), which has been proposed to achieve low system complexity and power consumption for WBANs. Existing studies on SmartBANs mainly focus on the energy performance considering the limited energy capacity of miniaturised sensors, and on the uplink transmission for monitoring and reporting-based applications. In this thesis, we address the critical aspects in the: (a) assessment of SmartBAN performance in terms of both energy and latency via analytical models and simulations; (b) investigation on SmartBAN MAC channel access mechanism designs for both uplink and downlink transmission. Our studies yield the proposal of low-latency and high-energy-efficiency MAC frameworks for SmartBANs in supporting
In realising remote H2M communications, such as that between SmartBANs and distance clinicians in telemedicine, heterogeneous optical and wireless access network is considered as a promising underlying architecture. The converged application delivery can benefit from the high capacity and reliability of optical fiber communication and the mobility and wide coverage of wireless networks. In this thesis, we consider the integration of passive optical networks (PONs) and wireless local area networks (WLANs) since PONs are recognised as the most efficient technology to provide wired access and WLANs are cost-effective and flexible in their deployment compared to mobile networks. In reducing the end-to-end latency over heterogeneous PON and WLAN networks, we present a detailed analysis of MAC layer bandwidth allocation solutions in WLANs and in PONs. Particularly, we explore the benefit of using machine learning (ML) in analysing and improving existing bandwidth allocation solutions. In our study, a deep neural network (DNN) is utilised to critically characterise the dependency of network latency on multiple bandwidth allocation decisions parameters and network features in PONs and WLANs via supervised training. Then, with this dependency learnt, the optimal bandwidth decisions that reduce the end-to-end latency are derived by using the trained DNN.

State-of-the-art research on H2M communications reports a bursty traffic profile. When such traffic aggregates upstream into the integrated optical network units and wireless access points (ONU-APs), adaptive bandwidth allocation to ONU-APs based on their bandwidth demand is critical in reducing latency. To this end, we propose a machine learning-based predictive dynamic bandwidth allocation (DBA) scheme, termed MLP-DBA, to address the bandwidth contention among ONU-APs, and the latency bottleneck caused by the bursty arrivals. In MLP-DBA, we exploit an artificial neural network (ANN) at the central office (CO) to predict H2M packet bursts to each ONU-AP, thereby enabling the bandwidth demand of each ONU-AP to be estimated. As such, arrivals to ONU-APs can be allocated bandwidth for transmission by the CO without having to wait extra transmission cycles. Then, the MLP-DBA makes adaptive bandwidth allocation decisions by classifying each ONU-AP according to its estimated bandwidth, thereby reducing the latency and packet drop compared
to that in existing schemes.

Since the development of Tactile Interent and H2M applications is in its infancy, current understandings on H2M traffic characteristics are still limited. In this thesis, we develop experimental H2M applications to study the traffic characteristics and innovative bandwidth allocation schemes for H2M applications based on its traffic characteristics. We present our H2M applications developed in a haptic teleoperation system and analyse the human control and haptic feedback traffic traces in these applications. In an attempt to find suitable models that characterise H2M arrivals, we analyse the statistical distributions of packet inter-arrival times. We also analyse the time-domain correlations of control and feedback packets and report a high cross-correlation between control and feedback traffic. This characteristic is defined as the traffic causality in H2M applications in this thesis. Based on this finding, we propose an artificial intelligence-facilitated interactive bandwidth allocation (AIBA) scheme in supporting low-latency H2M applications over access networks. In the AIBA scheme, the CO estimates and pre-allocates bandwidth to subsequent haptic feedback when forwarding the human control, thereby expediting the feedback delivery. Moreover, since our future access networks will need to support both H2M and conventional content-centric applications, we discuss priority differentiation between H2M and content traffic. The capability of existing schemes and the proposed AIBA scheme in reducing and constraining latency for H2M application is comprehensively evaluated and compared.

Overall, the technical contributions presented in this thesis provide novel MAC layer solutions into key enabling networks, including WBANs and heterogeneous PON and WLAN networks, for low-latency H2M applications. We extend the discussions on how ML techniques can be explored to facilitate intelligent bandwidth allocation in access networks. Experimental study on H2M traffic and understandings on human control and haptic feedback traffic in H2M applications are reported. Future directions that can be extended from this thesis are also discussed.
Acknowledgement

I would like to express my sincerest gratitude to my principal supervisor, Prof. Elaine Wong. With her depth of knowledge, profound thinking, and timeless energy, Elaine guides and supports me in every aspect to help me shape the research work in completing a PhD. Throughout my candidature, Elaine encouraged me to keep going, assisted and inspired me with her expertise. Herself is an encouraging role model full of energy and bravery to overcome any difficulty and achieve the best. Her insights, unique angle and sharp sense into the cutting-edge technologies continuously motivate my learning. During my PhD, Elaine paid countless time and patience in helping me with my research, writing and presentation skills, which I deeply appreciate. More than her responsible and professional supervision, the understandings and considerations from Elaine supported me personally and spiritually. At my hard moments, Elaine offered her compassion and care, guided what is truly important to be valued in life. Without the enlightenment from Elaine, I could not have gone where I am today. Her wisdom and philosophy would have an everlasting impact on me. It is truly my great luck to have Prof. Elaine as my principal supervisor.

My special thanks go to my co-supervisor, Dr. Pubuduni Imali Dias, for her constant encouragement and help on my research. Over the four years, Imali spared her valuable time in reviewing my work and providing insightful comments. She offered me prompt technical support and also kind help on my personal matters. I will always cherish the time of us having free discussions as well as her signature cupcake as a perfect celebration for memorable moments. The help and support from Elaine and Imali are immense. Taking this opportunity, I would like to convey my heartfelt gratitude to both my supervisors for their dedications that make my PhD journey a meaningful and fruitful one.

My appreciation also goes to Prof. Dragan Nesic, for chairing my committee, and providing constructive feedback on my research progress.

Here, I would like to thank Elaine once more for her generous support and foresighted suggestions on my professional development. During my PhD, I had the privilege to learn from many renowned researchers and received valuable
feedback from them. I would like to thank Vincent Chan at the Research Laboratory of Electronics, Massachusetts Institute of Technology, for offering a precious research visiting opportunity. His inspirational lectures and discussions bring me new understandings into the theory in the optical communication fields. I would like to express my thanks to Prof. Martin Maier at the Institut National de la Recherche Scientifique for his valuable comments and suggestions that helped me improve my work in the last chapter of my thesis.

At the university, I received help from my friends, colleagues and seniors. I would like to thank Prof. Christina Lim and Dr. Alan Lee for their help in the lab and for providing tutoring opportunity under their lectured subjects. In my final-year study, I was awarded the John Collier scholarship donated by alumni, Mrs. Deirdre Collier, which partly supported my research travel. I am grateful for this award from the University, and would like to thank Mrs. Collier for this gift and for sharing the globe-trotting adventures of her family.

I would like to say thank you to everyone at the Photonics Research Lab (PRL) who helped me in the past: Dr. Yifei Wang, Dr. Tian Liang, Dr. Tian Yu, Tingting Song, Sourav Mondal, Ye Feng, Sampath Surendra, Dr. Chathurika Ranaweera, Dr. Bhagya Amarasekara, Dr. Di Che, Dr. Jian Fang, Chuanbowen Sun, Honglin Ji. At PRL, they helped solve questions and doubts of mine with their expertise. The lunch and coffee hours we had together were delightful and refreshing escapes from piles of papers and lines of coding in daily research.

The joint sponsor from the China Scholarship Council (CSC) and the University of Melbourne is gratefully acknowledged. I would like to thank the University of Melbourne for providing me opportunities and resources to pursue my study at such a world-leading institution, and for the outstanding research environment it has created.

My PhD study at the University of Melbourne is made possible also with the help and enlightenment from the teachers and supervisors in my earlier-year education back in Xinxiang, Harbin and Xi’an. My special appreciation goes to Prof. Yong Li and Mrs Yuan Liu for their knowledge, help and trust, and their encouragement on me to pursue higher studies and my dreams.

I owe my deepest thanks to my beloved family. I am forever indebted to my parents, my mother LiLi Sun and father Zhigang Ruan, and my grandparents,
Meizhen Tao and Guiyun Ruan, for their unconditional love and devotion. Words can never thank them enough for their nurturing. Their love, trust and support are the lights and warmth, and the source of strength to me. I take this chance to thank my dear younger brother, Xiangxiang Ruan, for taking care of our mother and sharing family responsibility as a strong adult. A special thank you sends to Zhibo Wu for always being there. Thank you for lending a sympathetic ear in the tearful time and sharing joyful moments in my PhD journey.

Lihua Ruan
September 2019
Dedicated to my parents, younger brother, and my grandparents
Content

Chapter 1 Introduction ................................................................. 1
  1.1 Tactile Internet and Human-to-Machine Applications .................. 1
  1.2 Motivations: Latency Challenge in the Underlying Networks ......... 4
    1.2.1 Optical and Wireless Access Networks ............................ 5
    1.2.2 Wireless Body Area Networks ....................................... 7
  1.3 Thesis Outlines and Original Contributions ............................. 8
  1.4 Publications ....................................................................... 19

Chapter 2 Introduction to Medium Access Control Layer Solutions .... 23
  2.1 Introduction ........................................................................ 23
  2.2 Wireless Body Area Network ................................................. 25
    2.2.1 IEEE 802.15.4 WPAN ................................................... 26
    2.2.2 IEEE 802.15.6 WBAN ................................................... 29
    2.2.3 ETSI Smart Body Area Network ....................................... 31
  2.2 Wireless Local Area Network ................................................ 33
    2.2.1 IEEE 802.11 Distributed Coordination Function .................. 35
    2.2.2 IEEE 802.11 Point Coordination Function .......................... 36
    2.2.3 IEEE 802.11 Hybrid Coordination Function ......................... 37
    2.2.4 A Reference Design for HCCA ........................................ 39
  2.3 Optical Access Networks: Dynamic Bandwidth Allocation .......... 40
    2.3.1 Classic DBA Schemes .................................................... 41
    2.3.2 Predictive DBA Schemes ............................................... 44
    2.3.3 Statistical Bandwidth Estimation Algorithms ....................... 45
  2.4 Machine Learning for Bandwidth Resource Allocation ............... 49
    2.4.1 Overview: Opportunities and Challenges ............................ 49
    2.4.2 Machine Learning Techniques Used in Bandwidth Resource
      Allocation ........................................................................ 50
  2.5 Summary ........................................................................... 61

Chapter 3 SmartBAN Uplink Latency and Energy Performance Study: A
  Time-Optimised MAC for Periodic Monitoring and Emergency Traffic 65
  3.1 Introduction ...................................................................... 65
  3.2 SmartBAN MAC Protocol and Uplink Transmission .................... 68
3.2.1 SmartBAN MAC Frame ................................................................. 68
3.2.2 Uplink Channel Access and Latency Models ................................. 71
3.3 Time-Optimised MAC for Low Latency and High Energy Efficiency ..... 75
  3.3.1 Periodic Uplink Latency and Optimal IBI ..................................... 75
  3.3.2 Adaptive-IBI Algorithm for Energy-Savings ................................. 78
  3.3.3 EM Uplink Latency and Optimal CMP ........................................ 82
  3.3.4 Proposed Time-Optimised MAC framework .................................. 86
3.4 Performance Evaluations ................................................................ 87
  3.4.1 Periodic Uplink Latency and Optimal IBI Validation ......................... 87
  3.4.2 Adaptive-IBI Performance Evaluations ......................................... 89
  3.4.3 EM Uplink Latency and Energy Consumption Evaluations .............. 95
3.5 Summary ....................................................................................... 98

Chapter 4 SmartBAN Downlink Performance Study: A Novel Downlink Transmission Framework for Reducing Latency and Energy Consumption ................................................................. 101
4.1 Introduction .................................................................................. 101
4.2 Downlink Transmission Mechanisms in SmartBAN ............................ 104
  4.2.1 SmartBAN Supplementary Downlink Mode .................................. 104
  4.2.2 Limitation of SDM Latency and Improved SDM (ISDM) ............... 105
  4.2.3 Conventional Downlink Transmission Mechanisms: Challenges in SmartBAN ................................................................. 107
4.3 Proposed Downlink Transmission Modes for SmartBANs .................. 108
  4.3.1 Limited-exhaustive and Fully-exhaustive Downlink Mode .............. 108
  4.3.2 Embedded Markov Chains for LEDM and FEDM Latency Performance ..................................................................................... 109
4.4 Proposed Downlink Transmission Framework for SmartBAN ......... 114
  4.4.1 Energy-saving Mechanisms for SDM, ISDM, LEDM and FEDM .... 115
  4.4.2 Numerical Analysis of Latency and Energy-savings ....................... 116
  4.4.3 A Novel Downlink Transmission Framework ................................ 118
4.5 Performance Evaluations ................................................................ 120
  4.5.1 Model Validations and Performance Comparisons ......................... 120
  4.5.2 Performance of the proposed Downlink Transmission Framework 123
4.6 A Discussion on SmartBAN Round-trip Latency ............................... 125
  4.6.1 Round-trip Latency Modelling .................................................... 125
  4.6.2 Performance Evaluations ............................................................. 127
4.7 Summary ....................................................................................... 128
Chapter 5 Deep Neural Network Supervised Bandwidth Allocation
Decisions for Low-Latency Heterogeneous E-health Network .......... 131

5.1 Introduction ........................................................................................................................................ 131
5.2 End-to-end Latency of Heterogenous Networks .............................................................................. 135
  5.2.1 Latency of DBA Schemes in PONs ......................................................................................... 135
  5.2.2 Latency of WLAN HCF ........................................................................................................... 137
  5.2.3 End-to-End Uplink Latency Analysis ...................................................................................... 138
5.3 DNN in Predicting End-to-End Latency ............................................................................................. 139
  5.3.1 DNN Architecture and Learning Model .................................................................................... 139
  5.3.2 Supervised Training ................................................................................................................... 141
5.4 Bandwidth Allocation Decisions Analysis ......................................................................................... 145
  5.4.1 Uplink Latency, \( D_{\text{HCF}} \), of WLAN ....................................................................................... 145
  5.4.2 Uplink Latency, \( D_{\text{DBA}} \), of PON .......................................................................................... 149
  5.4.3 A Discussion on the Execution of DNN-supervised Bandwidth Decisions ...................................... 151
5.5 Performance Evaluations .................................................................................................................... 152
  5.5.1 Network Simulation Parameters ............................................................................................... 152
  5.5.2 End-to-End Performance Improvements ................................................................................... 152
5.6 Summary ................................................................................................................................................ 155

Chapter 6 Machine Learning based Bursty Bandwidth Prediction for Low-Latency H2M Applications ................................................ 159

6.1 Introduction ........................................................................................................................................ 159
6.2 Bursty Traffic Characteristic and Generation ..................................................................................... 162
6.3 Proposed Artificial Neural Network for Bandwidth Prediction ......................................................... 165
  6.3.1 Traffic Features in Predictive DBA Operation ........................................................................... 165
  6.3.2 ANN-based Bandwidth Prediction Model .................................................................................... 166
  6.3.3 Supervised Training .................................................................................................................... 168
  6.3.4 Bandwidth Estimation for Bursty Traffic ................................................................................... 171
6.4 Proposed MLP-DBA for Bandwidth Allocation .................................................................................... 174
  6.4.1 Optical Network Units Classification ......................................................................................... 174
  6.4.2 MLP-DBA Scheme for Bandwidth Allocation ............................................................................. 176
6.5 Performance Evaluations .................................................................................................................... 178
  6.5.1 Uplink Latency and Packet Drop Ratio ......................................................................................... 179
  6.5.2 ONU-AP Classification Analysis .................................................................................................. 182
6.6 Summary ................................................................................................................................................ 184
Chapter 7  Understanding Human-to-Machine Traffic and Bandwidth Allocation Schemes for Low-Latency H2M Applications ............... 187
  7.1 Introduction ........................................................................................................... 187
  7.2 Related Work ......................................................................................................... 190
  7.3 Experimental H2M Applications and Traffic Analysis ....................................... 191
    7.3.1 Haptic Teleoperation System ......................................................................... 191
    7.3.2 Experiment H2M Applications ...................................................................... 193
    7.3.3 Control and Haptic Feedback Traffic Analysis .............................................. 195
  7.4 AI-facilitated Interactive Bandwidth Allocation Scheme ...................................... 201
    7.4.1 Proposed AIBA Scheme .............................................................................. 201
    7.4.2 Performance Evaluation .............................................................................. 204
  7.5 AIBA Scheme for Aggregated H2M and Content Applications ............................ 208
    7.5.1 AIBA Scheme with Traffic Priority Differentiation ...................................... 208
    7.5.2 Performance Evaluation .............................................................................. 210
  7.6 Summary ............................................................................................................... 217

Chapter 8  Conclusions and Future Directions .............................................................. 221
  8.1 Introduction ........................................................................................................... 221
  8.2 Summary of Key Contributions ............................................................................ 223
  8.3 Future Research Directions .................................................................................. 227
    8.3.1 Optimising WBAN MAC .............................................................................. 227
    8.3.2 Wavelength Allocation in Optical Access Networks .................................... 228
    8.3.3 Exploration of ML for Intelligent Resource Allocation ............................... 229
    8.3.4 H2M Traffic Modelling .............................................................................. 230
  8.4 Conclusions .......................................................................................................... 231

References .................................................................................................................. 233
List of Figures

1.1 An illustration of the Tactile Internet components 3
1.2 Converged human-to-machine application delivery over wireless local area networks and optical access networks 5
2.1 An illustration of WBAN and the ubiquitous e-health 25
2.2 Evolution and standardisation of WBANs 26
2.3 Beacon-enabled MAC of IEEE 802.15.4 WPAN, IEEE 802.15.6 WBAN, and ESTI SmartBAN. 27
2.4 IEEE 802.15.4-defined CSMA/CA access 28
2.5 IEEE 802.15.6-defined CSMA/CA access 31
2.6 An illustration of WLAN network topology 34
2.7 IEEE 802.11 DCF CSMA/CA access 35
2.8 IEEE 802.11 PCF implementation 37
2.9 IEEE 802.11 HCCA implementation 39
2.10 Converged uplink transmission over optical access network 41
2.11 DBA operation with REPORT and GATE 42
2.12 Predictive DBA with arrival prediction and bandwidth estimation 45
2.13 An illustration of sigmoid function and logistic regression 55
2.14 An illustration of support vector machine 56
2.15 An illustration of ANN layered network architecture 59
3.1 SmartBAN topology and uplink transmission 66
3.2 SmartBAN CCH and DCH framework 69
3.3 SmartBAN transmission time structure and data frame format 69
3.4 SmartBAN uplink transmission and latency models 74
3.5 Numerical analysis of periodic monitoring average uplink latency 76
3.6 SmartBAN transmission and operation modes 79
3.7 State transit diagram of EM in SmartBAN 82
3.8 Numerical analysis of EM traffic average uplink latency and energy consumption 85
3.9 Latency and energy consumption trade-off 86
3.10 Average latency $D_i$ with varying $T_{IBI}$ and $T_i$ 88
3.11 Average latency $D_i$ with varying $T_{IBI}$ and $m$ ($T_i = 750$ ms) 89
3.12 Simulations of SmartBAN periodic uplink latency 91
3.13 SmartBAN energy-savings 92
3.14 Average energy-savings simulation for both hub and sensors 93
3.15 Channel efficiency of both hub and sensors 93
3.16 Comparisons with the IEEE 802.15.6 MAC 94
3.17 EM latency validation 96
3.18 Simulations of SmartBAN EM uplink latency and energy consumption ($\lambda = 2$ packets/s) 96
3.19 Simulations of SmartBAN EM uplink latency and energy consumption ($\lambda = 20$ packets/s) 97

4.1 SmartBAN SDM transmission and embedded Markov chain time instants for LEDM and FEDM. 104
4.2 Embedded Markov model and the state transition under LEDM 110
4.3 Embedded Markov model and the state transition under FEDM 113
4.4 Numerical analysis of FEDM and LEDM delay and energy-savings ($N = 0, G = 0.1$) 117
4.5 Numerical analysis of FEDM and LEDM delay and energy-savings ($N = 0, G = 0.8$) 118
4.6 Steady-state probabilities of embedded states 120
4.7 Queue length distribution of LEDM and FEDM 121
4.8 Average downlink delay and energy savings ($T_{IH} = 10$ ms) 122
4.9 Performance of the proposed downlink transmission framework 124
4.10 SmartBAN MAC and round-trip latency modeling illustration 126
4.11 Average round-trip latency validation with traffic load 127
4.12 SmartBAN round-trip delay under varying $M$ and $G$ ($N = 4$) 128

5.1 Converged delivery of monitoring, emergency and response data over a heterogeneous WBAN + WLAN + PON for ubiquitous e-health 132
5.2 Dynamical bandwidth allocation and IEEE 802.11 HCF for end-to-end uplink transmission in heterogeneous networks 136
5.3 An illustration of the proposed DNN learning and bandwidth decision learning model 141
5.4 DNN architecture, and the input and output features 142
5.5 DNN supervised training 144
5.6 WLAN uplink latency under $T_{TXOP}$, $T_{CP}$ and traffic load $l_{WLAN}$ 147
5.7 New connection time vs uplink transmission latency in WLAN 148
5.8 $D_{DBA}$ analysis with trained DNN 150
5.9 $D_{HCF}$ comparisons with shortest TXOP/SI and reference TXOP/SI 154
5.10 $D_{DBA}$ comparisons with limited-service DBA 154
5.11 $D_{E2E}$ performance comparison ($d_{O2C} = 10$ km) 155
6.1 Network architecture that supports H2M communications 160
6.2 Bursty traffic patterns of master control and slave feedback in a teleoperation session 163
6.3 Predictive DBA operation and traffic features 165
6.4 Schematic of the proposed ANN architecture to predict the ON/OFF status (y) of the bursty traffic at ONU-AP j 167
6.5 ON/OFF status prediction performance 169
6.6 Evaluation of sensitivity to training set size and model variation 170
6.7 ON/OFF period prediction accuracy with ANN and k-NN 172
6.8 Prediction performance comparison with MR-DBA and LSTP-DBA 173
6.9 An illustration of MLP-DBA procedure with three ONU-APs 175
6.10 Average uplink latency with the proposed MLP-DBA 180
6.11 Packet drop ratio with the proposed MLP-DBA 181
6.12 Distribution of Class A, B and C ONU-APs 183

7.1 H2M applications over fiber-wireless access networks 188
7.2 Teleoperation system settings for haptic H2M applications 192
7.3 Force sensors and actuators 193
7.4 Experimental H2M (teleoperation) applications 194
7.5 Control and feedback packet inter-arrival time cumulative distribution functions of Applications A to C 196
7.6 An illustration of control and feedback traffic in Application C 198
7.7 Control and haptic feedback traffic correlation analysis and comparisons with Internet applications 200
7.8 Uplink bandwidth allocation over optical access networks 202
7.9 ANN for haptic feedback bandwidth estimation in AIBA 204
7.10 AIBA operation with M-ONUs (masters) and S-ONUs (slaves) 205
7.11 Haptic feedback bandwidth prediction comparisons 206
7.12 Haptic H2M traffic latency performance comparisons 207
7.13 Latency performance comparison 212
7.14 Average polling cycle duration comparison 213
7.15 Packet latency statistics 215
## List of Tables

3.1 Notations for the SmartBAN latency modelling 71
3.2 Simulation parameters 90

5.1 Bandwidth decision, protocol and network parameters 139
5.2 Standardised network parameters 146

6.1 Traffic features collected in the $i$-th polling cycle 167
6.2 Principle of ONU-AP classification and MLP-DBA bandwidth allocation 175
6.3 Network and protocol parameters 179

7.1 Experimental applications and control/haptic feedback traffic 193
7.2 Average CDF fitting error 197
7.3 Fitted parameters in Generalized Pareto distributions 197
List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACK</td>
<td>Acknowledgement</td>
</tr>
<tr>
<td>AP</td>
<td>Access point</td>
</tr>
<tr>
<td>ANN</td>
<td>Artificial neural network</td>
</tr>
<tr>
<td>BE</td>
<td>Backoff exponent</td>
</tr>
<tr>
<td>BI</td>
<td>Beacon interval</td>
</tr>
<tr>
<td>BS</td>
<td>Base station</td>
</tr>
<tr>
<td>CAP</td>
<td>Contention access phase</td>
</tr>
<tr>
<td>CBR</td>
<td>Constant bit rate</td>
</tr>
<tr>
<td>CDF</td>
<td>Cumulative distribution function</td>
</tr>
<tr>
<td>CCA</td>
<td>Clear channel assessment</td>
</tr>
<tr>
<td>CFP</td>
<td>Contention free period</td>
</tr>
<tr>
<td>CMP</td>
<td>Control and management period</td>
</tr>
<tr>
<td>CO</td>
<td>Central office</td>
</tr>
<tr>
<td>CP</td>
<td>Contention period</td>
</tr>
<tr>
<td>CSMA/CA</td>
<td>Carrier-sensing multiple access with collision avoidance</td>
</tr>
<tr>
<td>CW</td>
<td>Contention window</td>
</tr>
<tr>
<td>DBA</td>
<td>Dynamic bandwidth allocation</td>
</tr>
<tr>
<td>DCF</td>
<td>Distributed coordination function</td>
</tr>
<tr>
<td>DIFS</td>
<td>Distributed interframe space</td>
</tr>
<tr>
<td>DNN</td>
<td>Deep neural network</td>
</tr>
<tr>
<td>DT</td>
<td>Decision tree</td>
</tr>
<tr>
<td>DoF</td>
<td>Degree of Freedom</td>
</tr>
<tr>
<td>EAP</td>
<td>Exclusive access phase</td>
</tr>
<tr>
<td>EDCA</td>
<td>Enhanced distributed channel access</td>
</tr>
<tr>
<td>EPON</td>
<td>Ethernet passive optical network</td>
</tr>
<tr>
<td>ETSI</td>
<td>European Telecommunications Standards Institute</td>
</tr>
<tr>
<td>EM</td>
<td>Emergency</td>
</tr>
<tr>
<td>FiWi</td>
<td>Fiber-wireless</td>
</tr>
<tr>
<td>G-ACK</td>
<td>Group ACK</td>
</tr>
<tr>
<td>GPON</td>
<td>Gigabits passive optical network</td>
</tr>
<tr>
<td>HCCA</td>
<td>HCF controlled channel access</td>
</tr>
<tr>
<td>HCF</td>
<td>Hybrid coordination function</td>
</tr>
<tr>
<td>H2H</td>
<td>Human-to-human</td>
</tr>
<tr>
<td>H2M</td>
<td>Human-to-machine</td>
</tr>
<tr>
<td>H2R</td>
<td>Human-to-robot</td>
</tr>
<tr>
<td>IAP</td>
<td>Inactive period</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>IBI</td>
<td>Inter-beacon interval</td>
</tr>
<tr>
<td>IEEE</td>
<td>Institute of Electrical and Electronics Engineers</td>
</tr>
<tr>
<td>IoT</td>
<td>Internet-of-Things</td>
</tr>
<tr>
<td>ISM</td>
<td>Industrial, Scientific and Medical</td>
</tr>
<tr>
<td>ITU</td>
<td>International Telecommunication Union</td>
</tr>
<tr>
<td>kNN</td>
<td>$k$ nearest neighbourhood</td>
</tr>
<tr>
<td>LR</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>LRD</td>
<td>Long range dependence</td>
</tr>
<tr>
<td>MAC</td>
<td>Medium access control</td>
</tr>
<tr>
<td>MAP</td>
<td>Managed access phase</td>
</tr>
<tr>
<td>ML</td>
<td>Machine learning</td>
</tr>
<tr>
<td>MLE</td>
<td>Maximum likelihood estimation</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean square error</td>
</tr>
<tr>
<td>M2M</td>
<td>Machine-to-machine</td>
</tr>
<tr>
<td>MPDU</td>
<td>MAC protocol data unit</td>
</tr>
<tr>
<td>NB</td>
<td>Naïve Bayes</td>
</tr>
<tr>
<td>ONU</td>
<td>Optical network unit</td>
</tr>
<tr>
<td>OLT</td>
<td>Optical line terminal</td>
</tr>
<tr>
<td>PCF</td>
<td>Point coordination function</td>
</tr>
<tr>
<td>PON</td>
<td>Passive optical network</td>
</tr>
<tr>
<td>PHY</td>
<td>Physical</td>
</tr>
<tr>
<td>PPDU</td>
<td>Physical protocol data unit</td>
</tr>
<tr>
<td>RAP</td>
<td>Random access phase</td>
</tr>
<tr>
<td>RTT</td>
<td>Roundtrip transmission time</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver operating characteristic</td>
</tr>
<tr>
<td>SAP</td>
<td>Scheduled access period</td>
</tr>
<tr>
<td>SDM</td>
<td>Supplementary downlink mode</td>
</tr>
<tr>
<td>SI</td>
<td>Service interval</td>
</tr>
<tr>
<td>SIFS</td>
<td>Short interframe space</td>
</tr>
<tr>
<td>SRD</td>
<td>Short range dependence</td>
</tr>
<tr>
<td>STA</td>
<td>Station</td>
</tr>
<tr>
<td>SVM</td>
<td>Support vector machine</td>
</tr>
<tr>
<td>TCP</td>
<td>Transmission control protocol</td>
</tr>
<tr>
<td>TXOP</td>
<td>Transmission opportunity</td>
</tr>
<tr>
<td>UDP</td>
<td>User data diagram</td>
</tr>
<tr>
<td>UP</td>
<td>User priority</td>
</tr>
<tr>
<td>WBAN</td>
<td>Wireless body area network</td>
</tr>
<tr>
<td>WLAN</td>
<td>Wireless local area network</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Tactile Internet and Human-to-Machine Applications

The advancements of communication technologies continuously reshape people’s lifestyle and daily activity. Broadband Internet and mobile networks nowadays provide content-centric applications such as email, message, voice, and video worldwide. These advancements have overcome the barriers in connecting human beings, i.e., human-to-human (H2H), at different locations across the world at any given time. Taking a step further, the recent development in the Internet-of-Things (IoT) has enabled machine-to-machine (M2M) communications across the world today [1]. With billions of devices being connected and exchanging information with each other, M2M applications facilitate automation in transportation, manufacture, education, healthcare, public security, smart city and home, and many other aspects. The next evolution of this trend is envisioned to be the Tactile Internet [2], where human control and interaction are performed over various smart devices, machines and robots, i.e., human-to-machine (H2M) communications.

The term Tactile Internet was first coined in early 2014 by Professor Gerhard Fettweis from the Technical University of Dresden in Germany and is further defined by the International Telecommunication Union (ITU) in its Tactile Internet Technology Watch Report [3]. A Tactile Internet has been described as an ultra-responsive, highly-reliable and secure communication network that supports remote H2M communications [3]. By this definition, real-time and remotely-controlled H2M applications is a distinctive type of application of the Tactile Internet. Moreover, the Tactile Internet targets to realise haptic communication in H2M applications, e.g. touch, sensations, position and movement proprioception where human beings manipulate and interact with both real and virtual environment. An application user in a Tactile Internet
environment will be able to control remote machine or robotic systems, and concurrently experience vivid audio-visual, haptic, and kinesthetic feedback. The innovative H2M applications promised by the Tactile Internet benefit almost every aspect of our society, and the next-generation Internet is therefore known as the Internet of Skills [4,5]. Some of the major societal areas that will be positively affected by the Tactile Internet and the emerging H2M applications are listed as follows:

- **Healthcare**: Harnessing biomedical sensors, actuators, robotic exoskeletons, artificial limbs, the Tactile Internet facilitates ubiquitous and personalised e-health such as telediagnosis, telerehabilitation, and telesurgery. By remotely controlling actuators and edge robots, medical experts in a hospital can supervise medicine injection and assist therapeutic rehabilitation of patients at home. This frees patients from costly in-hospital treatment and relieves the burden of medical systems.

- **Education and entertainment**: A new chapter of e-learning and gaming will be opened. The immersive interactions between learners or game players and the augmented reality, virtual reality environment will improve the learning performance and gaming experience.

- **Industrial automation and intelligent systems**: The inherent ability of Tactile Internet to remotely control and steer machines enable unmanned driving and manufacturing. Not limited to the preset and repetitive tasks in current industrial automation, H2M communications will facilitate intelligent decisions and reactions to complex situations with timely human supervision and calibration.

- **Extreme and high-risk tasks**: Real-time H2M applications will play a vital role in working and research activities in an extreme environment, such as in earth core, deep ocean, and outer space. Moreover, remotely controlling robots for search and rescue operations in emergency situations such as earthquake and bushfires is also an important application that relies on H2M interactions.

Based on these areas of examples, the emerging H2M applications featured in the Tactile Internet, significantly differentiate itself from earlier generations of
Tactile Internet and Human-to-Machine Applications

Internet applications. Specifically, the Tactile Internet can be grouped into three domains as illustrated in Fig. 1.1: (a) a master domain where the human operator and control interface are located, (b) a slave domain that consists of slave machines and/or robots, and (c) a network domain that connects the master and slave domain, and supports the bilateral transmission of control and feedback traffic [4]. Importantly, the realisation of H2M applications hinges upon ultra-reliable and low-latency network domain communication, which at the moment is constrained by the limited performances of existing communication networks [6]. To upgrade current networks for the emerging H2M applications, the following key challenges must be addressed carefully [7]:

- **Low end-to-end latency**: Latency of transmitting control/steering/haptic packets from human operators to slave devices, and vice versa, need to be constrained within 1 – 10 ms for effective real-time interactions. For haptic data transmission, in particular, a latency above 1 ms can disrupt the communication, causing a scenario known as ‘cybersickness’.

- **High reliability**: Since Tactile Internet packets need to be delivered in real-time, an ultra-reliable network with an availability above 99.9%, i.e., the failure rate in $10^{-7}$ order, is required. This means that the underlying network outage needs to be less 1 ms per day and 3 s in total per year.

- **High capacity**: Given that there is already an exponential growth in Internet-based applications, the capacity of existing networks must be upgraded to support high-density usage, e.g., 100 devices/m². Increasing the network capacity is critical to ensure that the data exchange in H2M applications adheres to their latency and reliability requirements.
In light of the above challenges in the network domain, a complete rethink of the underlying network infrastructure and the enabling technologies, including both wired and wireless technologies, are necessary [4]. Particularly, it is critical to ensure that the performances of the networks that potentially underpin the Tactile Internet, such as the core, metro-access, local access, and personal area network, can meet the stringent latency, reliability and capacity requirements mentioned above. A bottleneck in a single network segment affects the performance of the entire network, therefore the success of H2M applications [8].

Among the above-discussed challenges, the latency requirement is the most stringent regarding that the capability of our current communication networks is far from being able to ensure an end-to-end latency in the order of milliseconds [9]. As such, motivated by the demand to address the network domain latency challenge, this thesis presents an overview of the building block networks for the Tactile Internet, attempts to critically assess their latency performance, and propose novel solutions to improve their latency in supporting emerging H2M applications. As will be explained in the following sections, this thesis mainly focus on the wireless body area networks, and the heterogeneous optical and wireless access networks for the Tactile Internet and H2M applications. Our discussion on potential network architecture, performance evaluation of each involved network segment, and the investigation on novel resource allocation solutions to reduce latency for H2M applications are comprehensively presented in this thesis. Our motivations and original contributions are detailed as follows.

1.2 Motivations: Latency Challenge in the Underlying Networks

As introduced above, the main challenge in H2M applications lies in maintaining a millisecond-low end-to-end latency across the network domain. The time lag between a human operator making operation and perceiving visual and audio feedback needs to be < 150 ms [9]. The latency constraint is more demanding when delivering haptic data. For example, pressure and force feedback need to be delivered within 10 ms [10]. The latency of delivering
1.2 Motivations: Latency Challenge in the Underlying Networks

vibration and material perception shall be less than 5 ms [10]. For real-time and high-precision applications such as telesurgery, less than 1 ms end-to-end latency is required [11,12]. In addressing the latency challenge, the architecture and performance of current access networks have been paid increasing attention. This is because the low end-to-end latency requirement constrains the geographical distance between a human operator and the slave machines/robots being controlled. Exchanging control and feedback packets through metro- or core networks may deteriorate end-to-end latency due to the long-distance transmission. As such, converged optical and wireless access networks with edge intelligence placed at the central office (CO) and at the interface of optical network units and wireless networks have been considered as a future solution to support low-latency H2M applications [2,4,8].

1.2.1 Optical and Wireless Access Networks

Currently, the integration of passive optical networks (PONs) and wireless access networks is a promising technological trend for the Tactile Internet. The optical fiber communication can ensure high capacity and reliability, and wireless networks enlarge mobility and coverage [13]. Fig. 1.2 illustrates an example of the integrated network architecture, in which the wireless access points are integrated with the optical network units (ONU-AP) in PONs. Then, by deploying proxy servers and edge cloudlets as shown in Fig. 1.2, access to relevant data and

**Figure 1.2** Converged human-to-machine application delivery over wireless local area networks and optical access networks.
processing of the packets can be done at the CO or at the wireless edge, i.e., ONU-APs, thereby effectively reducing communication latency [14].

Note that PON technology is superior in capacity, reliability, low cost, and upgrade capability compared to other cabled connections such as copper or digital subscriber line, and therefore becomes a popular wired access solution for the emerging applications [13]. A PON can be integrated with different types of wireless front-ends. For example, authors in [15] demonstrated an LTE-based system design and test towards 1-ms end-to-end transmission. In [16], a human-to-robot (H2R) interactive teleoperation system is supported by LTE access to deliver force/position feedback from remote robots. Research work in [17] investigated the potentiality of fiber-wireless enhanced LTE-A, which unifies low-cost Ethernet PON (EPON) and the pervasive mobile cellular networks. Further, exploitation of mobile and edge computing in the above-proposed H2R network architectures were studied in [18-20]. Another major branch of research has foreseen that 5G mobile networks will underpin wireless communication for the Tactile Internet and the emerging H2M applications [5,6,11,21]. At present, 5G enabling technologies are still being developed and many open issues need to be addressed in the on-going standardization before 5G can pervade to facilitate H2M applications.

Note that not limited to the studies mentioned above, much of the on-going research focuses on integrating the mobile cellular networks, such as LTE and 5G-enabled access networks, with PONs towards meeting H2M latency targets. The popular and proliferating wireless local area networks (WLANs), however, has somewhat been overlooked. Although cellular networks provide higher data rate and capacity compared to WLANs, WLANs are more flexible and cost-effective in terms of deployment, and can be easily dedicated to H2M applications without having to share bandwidth with other content-centric applications. As such, the capability and potential of WLANs for low-latency H2M applications also need to be explored. In this thesis, we pay special attention to the heterogeneous access networks that integrate PON and WLAN. Moreover, note that the latency bottleneck in converged H2M delivery as illustrated in Fig. 1.2 mainly lies in the uplink direction since users in WLANs contend uplink bandwidth for data transmission, and multiple ONU-APs also contend uplink
bandwidth to transmit to the CO. Uplink bandwidth allocation schemes in PON and WLAN are hence critical for the latency performance of the heterogeneous access network. In this regard, the existing bandwidth allocation schemes in both PON and WLAN segments have been comprehensively investigated in this thesis, and novel low-latency solutions for converged H2M application delivery have been proposed. Particularly, in studying the performance of existing bandwidth allocation schemes in both PON and WLAN, we explore the potential of machine learning (ML) techniques. In state-of-the-literature, ML has been considered in various aspects in network control and regulation, such as routing and network planning [22], protection and reconfiguration [23], resource allocation [24], etc. With the capability in processing and analysing big data, ML provides novel solutions that facilitate intelligent network control and resource allocation decisions. In this thesis, we investigate the capability of ML in improve resource allocation, and therefore the latency performance of heterogeneous access networks. Note that due to the distinct characteristics of different ML techniques and the flexibility in their applications, critical questions such as how to select suitable ML techniques, how to apply an ML technique to facilitate bandwidth allocation to achieve low latency, and the pros and cons of applying ML over the existing schemes, have remained unanswered. These open questions are comprehensively investigated and discussed in this thesis.

1.2.2 Wireless Body Area Networks

In addition to access networks, the wireless body area network (WBAN) is a vital technology that supports the H2M applications and needs to be critically studied. WBANs deploy various sensors and actuators near, on or implanted inside the human body, thereby realising personal area H2M application delivery [25]. To support body area haptic actuators and robotic devices, WBAN systems must have their latency reduced. However, current research on low-latency WBANs is limited.

Specifically, the advent of WBAN technology is most beneficial for public healthcare. The uses of WBANs is changing the conventional in-hospital treatment towards proactive, ubiquitous and personalized e-health [26]. At present, WBAN systems are mainly designed for health status monitoring and
reporting, emergency warning, etc [27]. As a result, research efforts for WBANs focus on reducing energy consumptions, expecting prolonged operation time and low-radiation effects on the human body [28]. These efforts compromise the latency performance of the existing WBAN systems. Note that haptic actuators and body-worn robotic devices such as the exoskeleton, robotic limbs, etc., are essential entities for H2M interactions, and will eventually complement the use of bio-medical sensors in WBANs. WBANs, however, are not ready for these low-latency H2M applications. As such, this thesis pays special attention to the latency performance in WBAN systems, aiming at addressing the current lack of understandings and explore solutions for WBANs in achieving both low latency and high energy efficiency.

Importantly, without needing to modify the physical (PHY) layer design, the channel access mechanisms governed by the medium access control (MAC) layer play a critical role in determining the WBAN latency and energy performances. The MAC schedules when and how a sensor or actuator node can access the channel and transmit the buffered packets, thereby significantly impacting WBAN latency performance. To this end, in this thesis, we overview the international WBAN standardization activities and present the technical details of the existing MAC layer protocols. The performance of current channel access mechanisms in WBANs has been critically evaluated, yielding to novel mechanisms that address the latency and energy challenges in WBANs.

In summary, in meeting the stringent low end-to-end communication latency requirement of emerging H2M applications, the heterogeneous access networks, integrating the PON and WLAN, together with the WBAN that supports body area H2M application delivery, are comprehensively studied in this thesis. In particular, the bandwidth allocation schemes in PON and WLAN, MAC layer channel access mechanisms in WBAN, are the primary focuses. The performances of the existing schemes in allocating bandwidth in each of these network segments are carefully re-evaluated and novel solutions are investigated. The outlines and the contributions of this thesis are detailed in the subsection below.

1.3 Thesis Outlines and Original Contributions

The thesis comprises of 8 chapters including this chapter. This chapter
provides an overview of the Tactile Internet revolution and highlights the low latency challenge associated with the emerging remotely-controlled H2M applications. The motivations to reduce end-to-end latency for H2M applications and the underlying issues that are met within the optical, wireless and wireless body access networks are discussed in detail. An overview of the content in the remaining chapters is organised as follows:

**Chapter 2 Introduction to Medium Access Control Layer Solutions**

This chapter presents an overview of the channel access mechanisms in WBAN, the bandwidth allocation schemes in PON and WLAN. Note that these mechanisms and schemes are stipulated by the MAC layer protocols corresponding to individual network segments. Therefore, the overview of this chapter provides the technical details and operation of prevalence MAC layer solutions in WBAN, WLAN, and PON, respectively. Accordingly, the technical challenges of the existing solutions in reducing the latency for H2M applications are analysed.

Specifically, for the WBAN, we provide an overview of the standardized WBAN systems, including IEEE 802.15.4 and IEEE 802.15.6-defined WBAN, and the ETSI Smart Body Area Network (SmartBAN). The MAC layer protocols of each of these standards are compared and the latency challenges in the current MAC design are summarised. Further, in the heterogeneous access networks, we introduce the IEEE 802.11 MAC that governs the channel access in WLAN and the existing dynamic bandwidth allocation (DBA) schemes in the PON segments. Regarding the open issues in reducing the latency, we present a survey of the recent research on using ML techniques in improving the bandwidth resource allocation over converged optical and wireless networks. Overall, open areas to be explored with respect to MAC layer solutions in WBAN, WLAN, and PON, along with the exploitation of ML techniques, to support the emerging H2M applications are highlighted in this chapter. The overview presented in this chapter provides insights into the motivations and contributions of this thesis.

**Chapter 3 SmartBAN Uplink Performance Study: A Time-Optimised MAC Framework for Period Monitoring and Emergency Traffic.**

Chapter 3 focuses on the performance studies, in term of latency and energy
consumption, of the recently-proposed SmartBAN system. Standardised by the European Telecommunications Standards Institute, the SmartBAN is designed to achieve low system complexity and ultra-low power consumption. As a newly proposed system, the performance of SmartBANs has not been fully examined in the existing literature. Particularly, since the latency is not the primary focus in existing WBAN designs, including SmartBANs, the capability of SmartBANs in supporting low-latency applications and how to improve the SmartBAN MAC design in reducing the latency remain unanswered. As current WBANs are mainly used for monitoring purposes, e.g., health status monitoring and reporting, this chapter focuses on the uplink transmission performance of SmartBANs. In particular, we consider two major types of traffic in SmartBAN applications, the monitoring traffic that is periodically sampled and generated by bio-medical sensors, and the emergency (EM) reports that are generated in critical situations [29]. The SmartBAN MAC layer transmission mechanisms that can accommodate the distinct demands of both the monitoring and EM traffic are investigated.

In light of the above, this chapter develops analytical uplink latency models according to the uplink transmission mechanisms specified in SmartBAN MAC protocol. Based on accurate modelling, the impact of timing parameters in a SmartBAN MAC frame on the uplink latency performances of transmitting periodic monitoring and EM traffic is critically analysed. Specifically, in this analysis, we provide insights into how to determine a hybrid MAC frame that meets the latency and energy demands of both periodic monitoring and EM traffic. We discuss the energy-saving mechanisms considering sleep-capable SmartBANs that deploy sensors with sleep mode. The latency and energy-saving trade-off when exploiting the hybrid MAC is analysed. In order to achieve low latency and high energy-savings, we propose a time-optimised MAC framework based on the above analytical study. The performance of the proposed framework is validated via extensive simulations, and compared with existing uplink transmission mechanisms defined in SmartBAN and IEEE 802.15.6-defined WBAN MAC.

**Original contributions of Chapter 3 are listed as follows:**

- Presented detailed discussions and performance evaluations in terms of latency and energy consumptions of SmartBAN MAC in transmitting periodic monitoring and EM reports for the first time.
Developed analytical models that accurately characterise uplink latency of periodic monitoring and EM traffic in SmartBANs.

Validated the developed mathematical models and provided both analytical and simulation-based analysis that enhance understandings on the dependency of uplink latency on MAC timing parameter selections.

Proposed the optimal inter-beacon interval (Optimal IBI) to reduce the latency of periodic monitoring traffic, and the optimal control and management period (Optimal CMP) to reduce the latency for emergency traffic based on the analytical latency modelling.

Proposed energy-saving mechanisms for sleep-capable SmartBANs that harness the sensor with sleep mode and analysed the energy-savings performances considering both periodic monitoring and EM traffic in SmartBANs.

Critically analysed the latency and energy-savings trade-off related to varying SmartBAN MAC timing parameter selections.

Proposed a time-optimised MAC framework that determines MAC layer timing parameters to yield a hybrid MAC frame that reduces uplink latency and energy consumptions for both monitoring and EM traffic.

Validated the proposed framework via extensive simulations and presented comprehensive performance comparisons with existing uplink transmission mechanisms in WBANs.

Chapter 4 SmartBAN Downlink Performance Study: A Novel Downlink Transmission Framework for Reducing Latency and Energy Consumption

With uplink performances of SmartBANs studied in the previous chapter, Chapter 4 presents the first study on SmartBAN downlink latency and energy performances. Note that the existing research on WBANs, including SmartBANs, lacks investigations into the downlink transmission as current WBAN systems are mainly used for monitoring purposes. However, with sensors complemented by haptic actuators and robotic devices, such as an exoskeleton, to support H2M applications, downlink transmission of control and actuation traffic in WBANs need to meet stringent latency constraints. To this end, this chapter focuses on
the downlink latency and energy performances of SmartBANs.

In this chapter, we first investigate the latency performance of the supplementary downlink mode (SDM) specified in the SmartBAN MAC protocol. An improved SDM (ISDM) is proposed to address the shortcoming of the SDM. Insights into downlink transmission in general WBANs are highlighted via studying the SDM in SmartBANs. Further, we show a common limitation in SDM and ISDM in terms of reducing latency, and propose limited-exhaustive downlink mode (LEDM) and fully-exhaustive downlink mode (FEDM) to overcome the latency bottleneck. Two embedded Markov chain models are developed to critically evaluate the latency performance of the proposed LEDM and FEDM. Further, plausible energy-saving mechanisms for SDM, ISDM, LEDM, FEDM are discussed. Insights into downlink latency and energy consumptions are then presented via numerical analysis. Finally, based on the above performance study, a novel downlink transmission framework that selects suitable downlink transmission mode and MAC frame timing parameters for latency-constraint SmartBAN applications is proposed. Extensive simulations show the effectiveness of our proposed downlink modes and the transmission framework. In addition, in conjunction with uplink latency studied in Chapter 3, round-trip transmission latency performance in SmartBAN is also evaluated via analytical modelling and simulation analysis at the end of this chapter.

**Original contributions of Chapter 4 are listed as follows:**

- Presented the first comprehensive downlink latency and energy performances study for SmartBANs.
- Evaluated the latency performance of SmartBAN SDM and highlighted the latency bottleneck caused by MAC frame design.
- Proposed ISDM that addresses the limitation in SDM and guide downlink transmission mechanism designs for current WBAN systems.
- Proposed novel LEDM and FEDM that effectively reduce downlink latency and discussed energy-saving mechanisms for each of the downlink modes studied.
- Developed analytical models that accurately evaluate the latency and energy consumptions of LEDM and FEDM.
• Critically analysed the dependency of downlink latency and energy consumptions on the MAC layer timing parameters under SDM, ISDM, LEDM, and FEDM.

• Proposed a novel downlink transmission framework that can flexibly accommodate downlink latency and energy consumptions for latency-constraint applications.

• Validated the developed models and the effectiveness of the proposed downlink transmission framework via extensive simulations.

• Analytically modelled and evaluated SmartBAN round-trip transmission latency considering both uplink and downlink latency.

Chapter 5 Deep Neural Network Supervised Bandwidth Allocation Decisions for Low-Latency Heterogeneous E-health Network

So far, we have investigated the latency and energy performances of SmartBANs and proposed MAC layer solutions to prepare SmartBANs towards low-latency H2M applications. In this chapter, we further study the end-to-end latency of delivering latency-sensitive packets generated in SmartBANs over heterogeneous access networks. In particular, we focus on an integrated PON and WLAN network to support the aggregated SmartBAN traffic. As discussed earlier, bandwidth allocation in both PON and WLAN segments are critical in addressing the uplink latency bottleneck caused by bandwidth contention among multiple ONU-APs in the PON and multiple SmartBAN users in WLANs. In this chapter, we first present an end-to-end latency analysis of existing bandwidth allocation solutions, the DBA schemes in PON and the channel access mechanisms defined by the IEEE 802.11 WLAN MAC. Important bandwidth allocation parameters that affect the uplink latency performance in PON and WLAN are analysed and discussed.

Specifically, in order to fully understand the performances of these existing schemes and optimizing bandwidth allocation decisions to reduce the latency, we propose to exploit a deep neural network (DNN) to characterise the dependency of end-to-end latency on key bandwidth allocation parameters over heterogeneous PON and WLAN network. A DNN, characterised as a neural network with multiple layers and neurons, is advantageous in recognising abstract patterns and complex associations among multiple features [30], and therefore is
considered as an ideal tool to learn and supervise bandwidth allocation decisions that involve multiple decision parameters. Moreover, considering that bandwidth in current PON and WLAN is mainly allocated based on the traffic load, we also take into multiple network features, including the traffic load, the number of SmartBANs, the number of ONU-APs, and the distance between ONU-APs and the CO, into account in this study. By training a DNN to map different bandwidth allocation parameters together with different network features to the network end-to-end latency, adaptive bandwidth allocation parameters in PON and WLAN that minimises the latency can be supervised by the DNN in return.

In this chapter, we introduce the proposed DNN learning and decision-making model in supervising bandwidth allocation over heterogeneous access networks. We present the details of implementing supervised training, including the determination of DNN architecture, selection of input features and target output, and the sensitivity in generating and selecting the training set. We show that a trained DNN can accurately predict the end-to-end latency corresponding to varying bandwidth allocation decisions, thereby critically analysing the impact of multiple bandwidth parameters and network features has on latency. Using the trained DNN, optimal bandwidth to be allocated can be derived by minimising the latency in PON and WLAN, respectively. Simulations are performed to validate the DNN-supervised bandwidth allocation parameters, and the latency improvements are compared to the existing bandwidth allocation schemes in heterogeneous PON and WLAN networks.

**Original contributions of Chapter 5 are listed as follows:**

- Presented the first exploitation of a DNN in supervising bandwidth allocation to reduce the end-to-end latency over heterogeneous PON and WLAN networks.
- Proposed the DNN learning and decision-making model to characterise the dependency of the network end-to-end latency on bandwidth allocation parameters and multiple network features, and supervise the optimal bandwidth allocation in PON and WLAN in turn.
- Presented the technical details of DNN supervised training and provided insights into the selection of training samples, input and output features and the layered architecture in harnessing a DNN for bandwidth
• Critically analysed how different bandwidth allocation parameters and network features affect end-to-end uplink latency using the trained DNN.
• Re-evaluated existing bandwidth allocation schemes in PON and WLAN and highlight the overlooked features in existing bandwidth allocation schemes.
• Validated that the trained DNN facilitates adaptive bandwidth allocation in PON and WLAN corresponding to varying network features via simulations.
• Validated that the DNN-supervised bandwidth allocation effectively reduces the latency in PON and WLAN segments, thereby reducing the overall end-to-end latency of the heterogeneous network compared to the existing schemes.

Chapter 6 Machine Learning based Bursty Bandwidth Prediction for Low-Latency H2M Applications

In the previous chapters, we investigate the latency performance of SmartBANs and the end-to-end latency experienced when delivering latency-sensitive SmartBAN traffic over heterogeneous PON and WLAN access networks. From this chapter onwards, we explore the latency performance and bandwidth allocation schemes in heterogeneous PON and WLAN access networks, considering the characteristics of aggregated H2M applications. Note that the Tactile Internet and H2M applications are still in their infancy, and have not been widely practised over our communication networks. Therefore, our current knowledge on H2M traffic characteristics is limited. Nevertheless, several pioneering studies have anticipated the bursty pattern of H2M traffic [31-33]. When such traffic is aggregated over access networks, accurate bandwidth demand estimation of ONU-APs is challenging.

Motivated by the necessity to accurately estimate the bursty bandwidth demand and facilitate adaptive bandwidth allocations that reduce end-to-end latency, in this chapter, we propose to predict the packet burst status at the ONU-APs and estimate bandwidth demand to facilitate DBA operation accordingly. Specifically, we propose the use of an artificial neural network (ANN)
at the CO to achieve the bursty status prediction, i.e., if there is a burst present at an ONU-AP. We present the selection of input features and output target, and the details of supervised training. The resultant training outcomes are analysed, which show accurate predictions on packet bursts at ONU-APs, and the accurate bandwidth demand estimations that are achieved.

As the trained ANN is capable of predicting packet bursts to the ONU-APs, we explore novel predictive DBA schemes that can efficiently distribute bandwidth to ONU-APs that are receiving bursty packets and preserve bandwidth during long-idle intervals to reduce uplink latency and improve bandwidth utilisation. The proposed predictive DBA scheme in this chapter is named as Machine Learning-based Predictive DBA (MLP-DBA). In contrast to the existing DBA schemes that grant bandwidth to all ONU-APs based on bandwidth requests, MLP-DBA makes adaptive bandwidth allocation decisions by classifying each ONU-AP according to its estimated bandwidth. Using simulations, we compare the effectiveness of MLP-DBA scheme in reducing uplink latency and packet drop ratio against the existing DBA schemes. The latency performance of MLP-DBA scheme for latency-constrained H2M applications is analysed in this chapter.

**Original contributions of Chapter 6 are listed as follows:**

- Proposed the first exploitation of an ANN at the CO to predict the packet burst status of ONU-APs, and thereby estimate bandwidth demand in executing DBA scheme in PON.
- Presented the technical details of supervised training, selection of input features and output targets, and analysed the sensitivity and robustness of the training performance corresponding to traffic pattern variations.
- Analysed the performance of the trained ANN in predicting the bursty status of ONU-APs, and compared the bandwidth demand estimation performance to that when using bandwidth estimation methods in existing DBA schemes.
- Proposed the classification of ONU-APs based on the predicted bursty status and estimated bandwidth demand, and showed the bottleneck of existing DBA schemes in allocating bandwidth to ONU-APs that belong to different classes.
1.3 Thesis Outlines and Original Contributions

- Proposed MLP-DBA that adaptively allocates bandwidth to classified ONU-APs, thereby efficiently allocating bandwidth to ONU-APs that are experiencing packet bursts.

- Used simulations to validate and analyse the importance of the proposed ONU-AP classification and adaptive bandwidth allocation decisions in MLP-DBA.

- Used simulations to prove the effectiveness of the proposed MLP-DBA in reducing uplink latency and packet drop compared to existing DBA schemes.

Chapter 7 Understanding Human-to-Machine Traffic and Bandwidth Allocation Schemes for Low-Latency H2M Applications

In Chapter 6, we discussed the burstiness of H2M traffic reported in pioneering studies and proposed novel bandwidth allocation scheme to achieve adaptive bandwidth allocation for bursty packets aggregated at the ONU-APs. As reported and highlighted in Chapter 6, our current understandings on the traffic in H2M applications, and the impact of its aggregation over optical access networks are limited since the advancement of H2M applications is still in its infancy. Investigation on the unique characteristics of human control and feedback traffic in H2M applications is imperative. Insights into the traffic characteristics are critical in innovating changes to present bandwidth allocations schemes designed for content-centric applications for the emerging H2M applications.

To this end, in this chapter, we first develop a series of H2M applications based on a teleoperation experiment system and collect human control and haptic feedback traffic traces during H2M interactions. In understanding the traffic characteristics, we analyse the statistical distributions of packet inter-arrival times, and time-domain self- and cross-correlation of control and feedback traces. In particular, we report a unique cross-correlation between human control and the associated haptic feedback traffic and define this cross-correlation as traffic causality in real-time H2M applications. Exploiting this high cross-correlation between control and feedback traffic, we propose an ANN-facilitated interactive bandwidth allocation (AIBA) scheme to expedite haptic feedback delivery. In the proposed AIBA scheme, the CO estimates haptic feedback bandwidth demand
using an ANN and pre-allocates the bandwidth for feedback traffic based on the control traffic forwarded. By injecting our experimental traffic traces into a network simulation platform, we show that the ANN achieves more accurate haptic feedback bandwidth estimation compared to the bandwidth estimation methods in existing DBA schemes. The proposed AIBA scheme effectively reduces the latency for feedback traffic. With the performance of AIBA scheme validated, we discuss the prioritisation of H2M traffic over content traffic. As future networks are expected to support the aggregation of both H2M and content traffic, we present our design on AIBA scheme with priority differentiation. Via extensive simulations using our experimental traces, we comprehensively evaluate the performances of existing DBA scheme and the AIBA scheme in reducing and constraining latency for H2M applications in the presence of content traffic. Simulation results again validate the effectiveness of the AIBA scheme in improving the latency performance for H2M applications over existing schemes.

**Original contributions of Chapter 7 are listed as follows:**

- Developed a series of haptic H2M applications to experimentally investigate the characteristics of human control and haptic feedback traffic.
- Presented our statistical analysis on H2M packet inter-arrival time distribution, and time-domain self-correlation and cross-correlation of the human control and haptic feedback traffic.
- Defined the traffic causality in real-time H2M applications and exploited this finding to improve bandwidth allocation schemes in supporting converged delivery of H2M applications over optical access networks.
- Proposed the AIBA scheme that uses an ANN to estimate bandwidth for haptic traffic, and facilitate predictive feedback bandwidth allocation when forwarding the control traffic.
- Using simulations, together with experimental H2M traffic traces, to validate the effectiveness of ANN-based feedback bandwidth estimation and AIBA scheme in expediting haptic feedback traffic delivery compared to existing DBA schemes.
- Discussed priority differentiation between H2M and conventional content
applications, and presented details schematic of the AIBA scheme that considers differentiated priorities.

- Using simulations to comprehensive evaluate and compared the latency performance of existing DBA schemes and the proposed AIBA scheme in supporting low-latency H2M applications.

Chapter 8 Conclusions

Chapter 8 summarises the research work, important insights and findings, and contributions reported in this thesis. Further, in this chapter, we discuss potential future directions raised from this thesis.

1.4 Publications

Chapter 3


- E. Wong, L. Ruan and M. P. I. Dias, "Minimizing Latency of Periodic Monitoring Traffic in Smart Body Area Networks (SmartBANs)," 2016 18th *International Conference on Transparent Optical Networks (ICTON)*, Trento, 2016, pp. 1-4.

Chapter 4

- L. Ruan and E. Wong, "SmartBAN Downlink Performance Study: A Novel Transmission Framework for Reducing Delay and Energy
Chapter 1 Introduction


Chapter 5


Chapter 6


Conference on Computer Communications Workshops (INFOCOM WKSHPS), Honolulu, 2018, pp. 1-2.


Chapter 7

Chapter 2

Introduction to Medium Access Control Layer Solutions

2.1 Introduction

In Chapter 1, we introduced the key network segments, the heterogeneous access network, including passive optical network (PON) and wireless local area network (WLAN), and the wireless body area network (WBAN) involved in the Tactile Internet and H2M applications. In particular, we stressed the need to reduce the latency in these network segments in meeting the stringent latency requirement of H2M applications. Based on existing research, the underlying medium access control (MAC) layer protocols govern how bandwidth resource is allocated in these networks, and therefore impact the latency performance significantly. As such, this chapter overviews the MAC protocols in WBAN, WLAN, and PON, respectively.

As introduced in Chapter 1, WBANs harness body-worn sensors and actuators to collect personal information and delivery control/actuation feedback. Before the standardisation of WBANs, a variety of MAC layer channel access mechanisms such as S-MAC [34], BodyMAC [35], CA-MAC [36], have been proposed to support WBAN applications. In order to better regulate and support WBAN operations, standardised specifications such as IEEE 802.15.4 wireless personal area network (WPAN) [37], IEEE 802.15.6 WBAN [38] and ETSI Smart Body Area Network (SmartBAN) [39,40], have been adopted in realising WBAN systems. These standards stipulate the physical (PHY) and MAC layer protocols that define the fundamental technology, and transmission framework, while enabling the freedom in selecting protocol parameters and channel access methods for WBANs. This chapter highlights the existing WBAN MAC layer channel access mechanisms based on the aforementioned standards and compares their
similarities and differences. For WLANs, the IEEE 802.11 family is the accepted standard that defines WLAN PHY and MAC layer functions [41]. After the overview of WBAN MAC, this chapter will present the technical details of the IEEE 802.11 MAC protocol and the existing channel access mechanisms in WLANs. Unlike the WBAN and WLAN that utilise a wireless medium for transmission, PONs rely on passive optical components for data transmission. Consequently, there are significant differences in allocating bandwidth in PONs compared to that in WBANs and WLANs. Considering the inherent difference of medium, in this thesis, we refer to the bandwidth allocation in PON as bandwidth allocation schemes and that in WBANs and WLANs as channel access mechanisms. Specifically, bandwidth allocation in PONs is mainly achieved by implementing dynamic bandwidth allocation (DBA) schemes. Although the PHY and MAC layers differ in different types of PONs, such as Ethernet PON (EPON), Gigabit PON (GPON), etc., the underlying concepts of DBA schemes are similar [42]. In recent studies, researchers have also incorporated traffic prediction to achieve bandwidth demand estimation in DBA schemes in order to improve latency performance [44-47]. These schemes are known as predictive DBA schemes. This chapter comprehensively examines the various channel access mechanisms and existing DBA schemes, and highlights the advantages and disadvantages of these existing solutions.

In addition to the extensive investigations on the WBAN, WLAN, and PON MAC, this chapter also discusses the uses of machine learning (ML) techniques in bandwidth allocation over heterogeneous access networks in the state-of-the-art literature. The technical details of popular ML techniques and how they have been utilised in improving bandwidth resource allocation are presented. In particular, we analyse and highlight the distinctive characteristics of individual techniques, aiming to provide guidance on the selection of suitable ML techniques in formulating and solving practical problems. Overall, the analysis of different MAC protocols in WBAN, WLAN and PON, and the investigation on the use of ML techniques will highlight the key characteristics and open issues in the existing literature, thereby explaining the motivation behind our novel bandwidth allocation solutions proposed in Chapters 3 to 7. In the next sections, we will review the different MAC protocols in detail.
2.2 Wireless Body Area Network

WBANs enable a variety of wearable healthcare applications by using short-range wearable or implantable bio-medical sensors and/or actuators. Fig 2.1 illustrates a typical topology of a WBAN and how the use of WBANs shapes ubiquitous and personalised healthcare. As shown in Fig. 2.1, the sensors and actuators are connected to a central hub, typically in a star topology. The hub acting as a gateway transmits/receives the WBAN data to/from wireless access networks, thereby enabling information exchange between individual users and remote clinicians. At present, WBAN systems are designed with a special focus on energy consumption as miniaturised sensors and actuators demand prolonged battery life for long-term operation and lower electro radiations to the human body. Consequently, existing MAC layer channel access mechanisms are designed to maximise energy-savings [48, 49]. Due to the lack of attention on the latency performance, WBANs are not ready to support the emerging low-latency H2M applications that will harness haptic actuators and assistive robotic devices, such as exoskeletons, in WBANs. As such, it is critical to reassess the WBAN MAC layer performance in terms of both latency and energy consumption.

As introduced in the previous section, standardised protocols used to support WBANs are represented by IEEE 802.15.4 WPAN, IEEE 802.15.6 WBAN and the ETSI SmartBAN. WBANs generally operate in the 2.4 GHz Industrial, Scientific and Medical (ISM) band, which is the reserved spectrum open for industrial, medical, and research practices other than telecommunication. Although the

![Figure 2.1 An illustration of WBAN and the ubiquitous e-health.](image-url)
three standards define distinct PHY layer techniques to support WBAN functions in ISM band, their MAC layers share some similarities, and therefore face similar challenges in reducing latency and energy consumption for WBANs. In the following subsections, we present a detailed comparison of IEEE 802.15.4 and IEEE 802.15.6-defined WBAN and the SmartBAN. In order to better understand the characteristics of these standards and the connections among them, the evolution of WBANs is presented in Fig. 2.2. Following the timeline shown in the figure, the technical details of the IEEE 802.15.4 WPAN MAC are first overviewed in the next subsection.

### 2.1.1 IEEE 802.15.4 WPAN

The IEEE 802.15.4 WPAN standard was ratified in 2003 and is the basis for industry-pervasive sensor networks such as Zigbee [50], WirelessHART [51], 6LoWPAN [52]. It partitions the ISM band into 16 channels, and covers a range of 10 m – 75 m. The WPAN supports a maximum data rate of 250 kbps. Although this data rate nowadays might be insufficient for some medical sensors and healthcare applications [53], the underlying MAC of WPAN is inspirational for realising WBAN systems. Specifically, the channel access in WPANs is governed by channel beacons that are broadcasted by the hub. The beacon message notifies

---

**Figure 2.2** Evolution and standardisation of WBANs.
the attached sensor/actuator nodes of the access durations, i.e., when the channel is accessible, and how they should transmit to avoid collisions on the shared channel. Note that such a beacon-enabled MAC is widely considered in WBAN MAC designs, including the IEEE 802.15.6 WBAN and ETSI SmartBAN. Fig. 2.3 presents a comparison of the IEEE 802.15.4, IEEE 802.15.6, and the SmartBAN beacon-enabled MAC.

A typical MAC frame of IEEE 802.15.4 WPAN is depicted in Fig. 2.3(a). As illustrated in Fig. 2.3(a), the time interval between two consecutive beacons, known as the beacon interval (BI), comprises three distinct access periods: a contention access period (CAP), a contention-free period (CFP), and an inactive period (IAP). During CAP, sensors in WPANs compete for bandwidth through slotted carrier-sensing multiple access with collision avoidance (CSMA/CA) [30]. By CSMA/CA, a sensor attempting to transmit packets first senses the channel availability. If the channel is in use, the sensor waits for a certain period of time.
and then perform channel sensing again. This process to avoid transmission conflict is known as backoff sensing. Sensors with higher priority levels are typically endowed with short backoff time. A sensor only transmits its packets when the channel is sensed as idle.

The details of the IEEE 802.15.4-defined CSMA/CA is illustrated in Fig. 2.4. Before performing backoff sensing, a sensor sets a backoff timer. The timer starts with a random number within \([0, 2^{BE} - 1]\), where \(BE\) is the abbreviation of the backoff exponent. The initial value of the timer determines the backoff time duration, and hence is also termed as contention window (CW). When the timer countdown to 0, the sensor performs sensing to check the status of the channel, termed as clear channel assessment (CCA). If the channel remains idle for two consecutive CCAs, i.e., CCA1 and CCA2 illustrated by Node 2 in Fig. 2.4, packets can be transmitted by the sensor. Otherwise, the sensor resets the timer and repeats backoff and CCAs such as performed by Node 1 in Fig. 2.4. A short interframe space (SIFS) is reserved between consecutive packet transmissions to avoid interference. Acknowledgement (ACK) is required to confirm a successful transmission. As the hub does not coordinate sensors to transmit in CAP, collisions will occur when two or more sensors transmit their packets at the same time. To deal with the collisions, collided sensors need to adjust BE to restart a longer timer and perform CCAs for retransmission.

To avoid collisions, scheduled transmissions are considered in CFP. Guaranteed time slots (GTSs) as shown in Fig. 2.4 are assigned to sensors by the hub. As such, sensors can only transmit using the designated GTSs. Finally,
2.2 Wireless Body Area Network

during IAP, no transmissions are permitted. Sensors hence can opt to switch off their transceivers to save energy [53]. With the BI frame structure defined, practical implementations can be flexible in different WPAN applications. For example, the time duration of a BI and individual access periods, i.e., CAP, CFP and IAP, are left open in the standard, and mechanisms to allocate GTSs also depend on the characteristics and requirements of applications [54]. A general mechanism for WPAN is the contention-dominated transmission using CAP. When a sensor requires extra bandwidth or prioritizes its traffic, it sends a request to the hub for GTSs [55].

It can be observed that IEEE 802.15.4 defined a hybrid MAC frame that exploits both contention-based, i.e., CSMA/CA, and scheduled access, i.e., via GTSs. In general, contention-based access methods yield lower latency since packets can be transmitted immediately when the channel is sensed available and centralised scheduling is not required. However, frequent collisions could occur when the traffic load increases. Channel sensing and packet retransmissions also consume extra energy. Scheduled access, on the other hand, requires packets to be transmitted in the designated time slots, thereby avoiding collisions and retransmissions, and achieving energy savings. However, packets could experience longer latency due to waiting for the scheduled time slots. The IEEE 802.15.4 MAC indicates the importance of the hybrid MAC frame design and its impact on network latency and energy performances.

In summary, the IEEE 802.15.4 standard is widely implemented to support low power, short-range, and low data rate wireless sensor network applications, including the earlier WBAN applications. However, as introduced at the beginning, it is challenging for WPANs with a ≤ 250 kbps to cater the low latency and high data rates demands of some emerging medical and healthcare applications [56]. Although the IEEE 802.15.4 standard is not ratified particularly for body area networks and medical applications, its MAC layer design manifests the key factors that impact the network latency and energy performance. These factors also need to be carefully considered in the subsequent IEEE 802.15.6 and ETSI SmartBAN MAC protocols.

2.1.2 IEEE 802.15.6 WBAN
The IEEE 802.15.6 WBAN is the first international standard for supporting communications in the vicinity of or inside the human body. The Task Group 6 for the standardisation of WBAN was initiated by IEEE 802 in 2007, and the completed version of IEEE 802.15.6 was released in 2012 [38]. The IEEE 802.15.6 standard offers a high transmission bit-rate of up to 10 Mbps and greater flexibility in channel access, but at the cost of a highly complex MAC superframe [57-59]. Operation bands in the IEEE 802.15.6 MAC are divided into 79 channels covering (a) narrowband, which is within the ISM band, (b) ultra-wideband (UWB), and (c) human body communication (HBC). The ISM band that is discussed in this thesis supports a data transmission rate of up to 971.4 kbps.

It is important to note that a WBAN can operate with and without beacons. When used without the beacons, sensors in a WBAN always access the channel by IEEE 802.15.6-specified CSMA/CA. A wide range of existing WBAN systems still uses beacon messages, i.e., beacon-enabled MAC, to govern channel access of the sensors [60-62]. The superframe between beacons is shown in Fig. 2.3 (b). A superframe is divided into two exclusive access phases (EAP1 and EAP2), two random access phases (RAP1 and RAP2), two managed access phases (MAP1 and MAP2), and a contention access phase (CAP). EAP, RAP, and CAP are contention-based where either slotted ALOHA or CSMA/CA can be adopted for transmission. In comparison, the MAP can be configured for either scheduled or contention-based access. Moreover, different access phases can be used for different traffic. For example, the IEEE 802.15.6 MAC protocol defined eight levels of user priority (UP). The EAP is specifically reserved for sensors that have the highest UP. RAP and MAP can be accessed by sensors with all traffic types. The CAP is an additional contention access phase whereby if enabled by the hub, an additional beacon (Beacon 2) as shown in Fig. 2.3(b) is sent to notify all sensors of this additional transmission phase.

The details of IEEE 802.15.6-defined CSMA/CA are illustrated in Fig. 2.5. The sensors first set the timer for backoff sensing. It should be noted that unlike IEEE 802.15.4 CSMA/CA, the sensors in WBANs keep sensing the channel during the backoff and only decrements the timer value if the channel is sensed idle. As illustrated by Node 1, if the channel becomes busy during its backoff sensing, the timer countdown suspends until the channel becomes idle again.
its timer reaches 0, a sensor transmits a packet to the hub and waits for the ACK. Group ACK (G-ACK) for sequential packet transmissions is allowed in IEEE 802.15.6 WBANs. This reduces the overhead and improves the channel utilisation compared to the IEEE 802.15.4 WPAN. However, as the sensors consistently sense the channel during the backoff, more energy is consumed than the IEEE 802.15.4 WPAN.

Similar to the IEEE 802.15.4 MAC frame introduced in Section 2.1.1, time durations of these access phases in IEEE 802.15.6 MAC superframe are not explicitly defined. In practice, the MAC superframe can be modified by enabling or disabling any of these access phases. At present, research work in evaluating and exploiting the IEEE 802.15.6 MAC favours contention access via CSMA/CA, using one or more selected phases [60-65]. For instance, latency studies reported in [60-65] took EAP1 and RAP1 into account and the lengths of other phases deliberately set to 0. In the e-health project known as CANet, only the RAPs of the IEEE 802.15.6 MAC superframe was adopted for data transmission for the simplicity of realisation [39]. Furthermore, although several access phases were jointly considered in modelling the IEEE 802.15.6 MAC in [66-68], the exact same contention access mechanism was adopted within all access phases. Overall, due to the complex superframe defined as well as the reliance on contention access, the performance of WBANs are not yet fully examined.

2.1.3 ETSI Smart Body Area Network

As discussed in the previous paragraphs, WPANs and WBANs suffer from

**Figure 2.5** IEEE 802.15.6-defined CSMA/CA access.
limited data rates and complex frame structure, respectively. In designing low-complexity and low-power consumption WBAN systems for medical and healthcare applications, the European Telecommunication Standards Institute initiated the technical committee (TC) Smart Body Area Network (SmartBAN) in 2013. In 2015, the SmartBAN ultralow-power PHY and MAC specifications were published [39,40]. The recent SmartBAN research outlined in [69-70] provided an overview of TC SmartBAN activities and the proposed standards. Research in [29] highlighted that compared to IEEE 802.15.6, SmartBAN exhibited a faster initial setup and connection times.

The SmartBAN MAC specifies two types of channels: control channel (C-Ch) and data channel (D-Ch). In total, there are 3 C-Ch and 37 D-Ch in ISM band and each data channel capable of transmitting up to 1 Mbps. Using C-Ch as an announcement channel, the hub shares network information with all its sensors as well as co-existing SmartBANs. When a sensor requests a new connection or to change to a different channel, this information is sent through the C-Ch. The D-Ch, on the other hand, is responsible for data packet transmission, and therefore, is the focus of this study. The SmartBAN D-Ch MAC frame is illustrated in Fig. 2.3(c). Each fixed-length inter-beacon interval (IBI) between two adjacent data channel beacons (D Beacons), contains a scheduled access period (SAP) with time division multiple access (TDMA), a control/management period (CMP) accessed by slotted ALOHA, and an inactive period (IAP). Unlike contention-dominated access adopted by WBANs, SmartBANs emphasize on scheduled channel access, whereby TDMA in SAP is the default access mechanism for data transmission. CMP is utilised for sending occasionally generated system management packets. It is clear that with scheduled access, transmission collisions can be eliminated, and power consumption can be reduced since no backoff sensing is required and sensors can switch to sleep mode outside its designed transmission time slots. However, the latency performance may be compromised.

Since the SmartBAN is a recently-proposed WBAN system, existing research on SmartBAN is relatively limited and the performance of SmartBANs is not fully understood. As such, investigation on the SmartBAN MAC and critical examinations of the latency and energy performances are imperative. Currently,
there are many areas of the SmartBAN that need careful studies such as the criteria for determining the IBI, SAP, CMP and IAP durations, as well as the impact of these timing parameters on the SmartBAN latency and energy consumptions. Taking these into consideration, this thesis focuses on the SmartBAN and comprehensively studies its MAC design. Comparisons are made with IEEE 802.15.6-defined WBANs. The evaluations and findings from our SmartBAN studies are detailed in Chapters 3 and 4 in this thesis.

In summary, the latency and energy performances of WBANs need to be critically evaluated and improved to support the emerging H2M applications in the Tactile Internet era. The beacon-enabled MAC that defines a channel beacon and a hybrid frame to govern channel access is commonly adopted by the IEEE 802.15.4, IEEE 802.15.6 and ETSI SmartBAN. Note that scheduled-based access methods such as TDMA and polling are advantageous in saving energy, while contention-based access methods such as CSMA/CA and ALOHA could be superior in low latency [71]. As a result, the implementation of either scheduled access, such as the default TDMA in the SmartBAN, or contention-dominated access, such as the existing use of CSMA/CA in WBANs, will not provide satisfactory energy and latency performances. Instead, a careful hybrid MAC design that best exploits these access methods based on the traffic and quality-of-service (QoS) characteristics of the sensor/actuator nodes is necessary. The study of the SmartBAN MAC in this thesis discusses such hybrid MAC designs towards achieving both low latency and high energy efficiency, which provides solutions for the aforementioned open areas in WBANs.

2.2 Wireless Local Area Network

The WLAN has become a prevalent wireless access technology that complements cellular technologies. WLANs are flexible and cost-effective in their deployments and can be easily dedicated to specific user groups or applications. The IEEE 802.11 family is the widely accepted standard that defines a series of PHY and MAC specifications for implementing WLANs. First ratified in 1997 (IEEE 802.11-1997), the IEEE 802.11 standard has gone through a series of amendments to keep up with the advancement of technology and increasing user demands [72]. IEEE 802.11b was the first to receive worldwide acceptance,
followed by IEEE 802.11g and IEEE 802.11n. In 2012, task group 802.11 documented and released IEEE 802.11-2012 which included all the previous amendments [41]. The latest updated release in this series is the IEEE 802.11-2016 [73]. The typical frequency spectrums used by WLANs are 2.4 GHz ISM band and 5 GHz band. The supported data rate ranges from 2 Mbps up to 600 Mbps depending on the modulation methods, with a gigabit data rate expected in the near future. A WLAN can either be in an infrastructure mode or an ad-hoc mode as illustrated in Fig. 2.6 [41]. The former is featured by the deployment of a central wireless access point (AP), which acts as the gateway or coordinator for its connected stations (STAs), i.e., user devices in a WLAN. In comparison, an ad-hoc WLAN can be initiated by any eligible STAs. This initiate STA then acts as the gateway and other STAs joined the WLAN communicate to this STA for accessing the outer networks. In this thesis, we focus on the infrastructure mode WLANs and their integrations with the PON to support converged application delivery. As introduced in Chapter 1, an infrastructure WLAN can avoid multi-hop relay of the packets over Ad-hoc STAs, and therefore reduce the end-to-end latency.

The IEEE 802.11 MAC defines two fundamental channel mechanisms: (a) the distributed coordination function (DCF), and (b) point coordination function (PCF). The DCF is a contention-based random channel access mechanism where IEEE 802.11-defined CSMA/CA is adopted by the STAs to transmit to an AP. The PCF, on the other hand, is a centralised mechanism in which an AP schedules the transmissions of STAs. The PCF is primarily designed to support QoS-sensitive traffic such as voice and video streams since centralised control

![Figure 2.6 An illustration of WLAN network topology.](image)
supports collision-free transmissions and the control of latency and jitter [74]. The basic mechanism for best-effort traffic transmission is actioned through DCF, where all STAs contend for access in CSMA/CA [75]. Alternatively, access to the channel can alternate between DCF and PCF, within the so-called contention period (CP) and contention free period (CFP), respectively, leading to the hybrid coordination function (HCF) in WLANs. Similar to the WBAN MAC as introduced in the previous section, in order to exploit both DCF and PCF, beacon-enabled MAC with a hybrid frame structure is a primary feature of HCF. In particular, IEEE 802.11e specifies the implementations of HCF particularly for supporting latency-sensitive applications [76,77]. The following subsections discuss the implementation of DCF, PCF, and HCF in detail.

### 2.2.1 IEEE 802.11 Distributed Coordination Function

Fig. 2.7 presents the DCF operation by the IEEE 802.11-defined CSMA/CA. A STA intending to send a packet needs to sense the channel status first, i.e., idle or busy. If the channel is sensed as busy, the STA continues sensing until the channel is detected to be idle for a time duration of distributed interframe space (DIFS) as shown in Fig. 2.7. After the DIFS, the STA performs backoff sensing. This backoff sensing procedure is similar to that stipulated in IEEE 802.15.6 as introduced in Section 2.1. In recent literature, the backoff sensing procedure and the resulting latency were critically analysed via Markov modelling in [78-80], and the throughput analysis was presented in the literature [81,82]. Multiple studies have also proposed means of improving the DCF performance through

![Figure 2.7 IEEE 802.11 DCF CSMA/CA access.](image-url)
strategically selecting timer and backoff algorithms, such as dynamic CW tuning, binary exponential backoff, collision rate based backoff, and sliding DCF [83-86]. However, the inherent drawback of CSMA/CA is its collisions. Degraded throughput with an increasing number of STAs and/or traffic load at STAs has been shown in above-mentioned studies. DCF does not ensure QoS such as latency and jitter within a WLAN, and therefore is considered to provide best-effort access [87].

2.2.2 IEEE 802.11 Point Coordination Function

With DCF as the fundamental access mechanism, PCF is an additional access mechanism for contention-free access as illustrated in Fig. 2.8. The channel is divided into CFP accessed by PCF and CP accessed by DCF. An AP initiates a CFP by broadcasting a beacon to the STAs attached to it and then sending a CF-Poll message to schedule the transmission of STAs. Similar to DCF, a SIFS is reserved between packets. The AP broadcasts a CF-End message to all STAs to terminate a CFP, after which the STAs are notified of the CP and implement the channel access occurs through CSMA/CA.

The PCF operation is characterised by the following key features. The STAs must be scheduled via CF-Poll messages for uplink transmission, and each time, only one packet can be sent to the AP as shown in Fig. 2.8. A successful transmission needs to be acknowledged via ACK. Note that packets generated by an application at an STA typically have equal length except for the last packet. The last packet contains the residual payload after fragmenting the application data into packets. This is attributed to the data fragmentation specified in IEEE 802.11 MAC. Finally, the DCF traffic during CP will affect the time interval between consecutive beacons, and therefore the time interval between two CFPs.

For downlink-dominated transmission such as audio/video download, PCF is favoured due to the QoS support and elimination of collisions in centralised scheduling [88]. For instance, when latency-sensitive video streams are buffered at an AP, the AP can initiate sequential CFPs to guarantee the latency and jitter in sending such packets. The studies reported in [88-91] theoretically analysed the PCF performances in terms of latency and throughput. Results show that the scheduling algorithms in PCF significantly impact the network performances.
2.2 Wireless Local Area Network

Moreover, a variety of PCF algorithms such as adaptive and dynamic PCF were proposed in existing literature for different applications [92,93]. Many research studies also focused on achieving power saving in PCF such as in [94,95]. Overall, regarding the above performance studies, it has been recognised that when STAs have uplink-dominated traffic, CF-Poll, ACK, SIFS tied to each packet incurs latency between consecutive transmissions. In the emerging IoT and Tactile Internet era, machine-to-machine and human-robot/machine applications will have small packet sizes [96]. The overhead caused by frequent CF-Poll, ACK and SIFS hence affects the channel utilisation efficiency, thereby increasing latency. In this regard, the HCF could provide a better access solution. HCF is defined to achieve better QoS provision. The details of its implementation are presented in the following subsection.

2.2.3 IEEE 802.11 Hybrid Coordination Function

The HCF defined by IEEE 802.11e implements enhanced distributed channel access (EDCA) for contention-based access and HCF controlled channel access (HCCA) for scheduled access [77]. Note that a wireless channel in a WLAN is partitioned into time slot units, which is the minimum unit to synchronize or regulate transmissions in DCF and PCF. The basic time unit in HCF to allocate transmissions is known as transmission opportunity (TXOP), which is a time duration that contains multiple numbers of time slot unit. EDCA improves the basic DCF by enabling four levels of service priority and designated different TXOP durations and contention window parameters accordingly. One of the primary topics in this area is to dynamically choose EDCA parameters to reduce

![Figure 2.8 IEEE 802.11 PCF implementation.](image)
latency such as the algorithms proposed in [96,97]. Research studies in [98-100] presented the latency performance analysis and simulation validations for traffic with different priority levels. Results show that compared to the basic DCF, EDCA reduced the latency of the highest-priority traffic, while compromising the performance of traffic flows with lower priorities. Although EDCA is an improvement from DCF and achieves latency differentiations among different service priorities, it cannot assure a specified latency for latency-sensitive applications [99]. In comparison, HCCA is able to provide QoS support and address the beacon interval variation in PCF. Currently, HCCA is widely studied in achieving low-latency uplink transmissions [100-103].

Fig. 2.9 illustrates the implementation of HCCA. As shown in Fig. 2.9, channel access by HCCA is governed by periodically-broadcasted beacons. A beacon interval is partitioned into multiple service intervals (SIs). Each SI comprises a controlled access period and a CP. During the contention period, STAs access the channel by EDCA. In the controlled access period, STAs transmit in corresponding TXOPs allocated by the AP, where an AP allocates TXOPs using QoS(+) CF-Poll message defined by the IEEE 802.11e. This allocation can be done in a flexible manner such as (a) sending QoS(+) CF-Poll to STAs following the beacon, (b) sending QoS(+) CF-Poll to STAs during CP, and (c) piggybacked QoS(+) CF-Poll on data packets to STAs. In a general HCCA implementation, when allocated by the AP in a beacon interval, TXOPs remain the same in each SI and STAs transmit in the assigned TXOPs as shown in Fig. 2.9. The unique characteristics and differences with PCF are as follows:
2.2 Wireless Local Area Network

- HCCA defines a hybrid MAC frame, whose timing structure is governed by the channel beacons, and different access methods need to fit the pre-configured access periods. In PCF, however, a CFP impromptu starts using a beacon and ends with a CF-End. The duration of a CFP hence is dynamic in nature and transmissions in DCF during CP affect the interval between CFPs. As such, HCCA is considered to better provision QoS support than PCF [101].

- Unlike the basic DCF and PCF that only allow one packet in each transmission attempt, HCCA allows sequential packet transmissions of STAs in the assigned TXOPs. This reduces the latency and overhead since CF-Poll and ACK are no longer needed for individual packet transmissions.

- In HCCA, both ACK and the QoS(+) CF-Poll can be piggybacked on data packets such that channel can be used more efficiently compared to the basic PCF.

- The latency performance of HCCA relies on the timing parameters in a MAC frame. The durations of TXOP, SI are critical and need to be appropriately set by an AP in order to improve the HCCA performance.

2.2.4 A Reference Design for HCCA

IEEE 802.11e provides a reference implementation of HCCA [102]. In the reference HCCA, an AP determines the duration of a TXOP by estimating the number of packets generated by the STAs in an SI. Let us denote the duration of a SI as $T_{SI}$, the data rate as $R_{WLAN}$ and packet length as $L$. Assuming the packet arrival rate at a STA is $\lambda$, TXOP allocated to this STA can be calculated as follows:

$$TXOP = \left\lceil \frac{T_{SI} \times \lambda}{L} \right\rceil \times \frac{L}{R_{WLAN}} + X$$

(2.1)

where $\lceil x \rceil$ is the ceiling function to accommodate an integer number of arrivals, and $X$ is the overhead in a TXOP including any ACK and SIFS. The duration of SI, $T_{SI}$, is a preset value by the AP that depends on the number of STAs and the duration of the CP. The reference HCCA setting defined in (2.1) is suitable for applications with constant arrivals [104]. Further, the studied in [105] focused on
improving TXOP allocations for application with variable bit rate and varying packet lengths. Comprehensive analysis and comparisons of DCF, PCF and HCF performances and limitations were presented in [75]. From these studies, it can be noted that the exploitation of HCCA can be flexible and the potential of HCCA has not been fully explored.

Considering the advantages of HCCA and the open issues mentioned above, in this thesis, we will present our investigations on the HCCA in WLANs in improving the end-to-end communication over heterogeneous access networks. Detailed performance evaluations and our proposed solutions for HCCA implementation are provided in Chapter 5.

2.3 Optical Access Networks: Dynamic Bandwidth Allocation

PON technologies are one of the most important deployment technologies that underpin next-generation broadband access networks [13]. The high capacity, data rate, reliability and robustness of optical fiber communication make PONs a superior solution to support the Tactile Internet and low-latency H2M applications [4]. As such, integrating PON and wireless access technologies to support converged application delivery has been widely discussed in the existing literature. To achieve seamless convergence in millisecond low end-to-end latency, the latency in PONs can be only up to a few hundreds of \( \mu s \), and the rest latency budget lies in wireless access segments such as in the WLANs and WBANs [8]. Fig. 2.10 illustrates the converged network architecture and uplink transmission in PONs. The downlink transmission in a PON is broadcasting by nature. In the uplink direction as shown in Fig. 2.10, multiple integrated optical network units and wireless access point (ONU-APs) share the same wavelength and contend uplink bandwidth for transmissions.

Uplink bandwidth allocation to ONU-APs is mainly realised by implementing DBA schemes. DBA schemes hence are critical to the latency performance in PONs. Although PONs underpin different types of access networks, such as Ethernet PON (EPON), Gigabit PON (GPON), the next-generation PON (NG-PON), etc., the underlying concepts of DBA operations are similar [42]. In a
2.3 Optical Access Networks: Dynamic Bandwidth Allocation

Typical DBA operation, REPORT and GATE messages are utilised to exchange information between ONU-APs and the CO. An ONU-AP can request bandwidth using a REPORT message that notifies the CO of its buffer occupancy (in bytes). Upon learning the bandwidth requirement of an ONU-AP, the CO makes bandwidth allocation decisions and grants bandwidth by sending a GATE message to the ONU-AP. A GATE mainly indicates an ONU-AP the assigned transmission timeslot using a pair of values \{startTime, duration\}. Then, an ONU-AP starts transmitting at startTime for the specified duration. It can be noted that the above framework underpins the message exchange between the CO and ONU-APs in DBA operation. The bandwidth to be allocated is determined by the CO. Based on how the CO allocate bandwidth when receiving bandwidth request from ONU-APs, different DBA schemes have been developed.

Existing DBA schemes can be categorised into classic DBA schemes and predictive DBA schemes. In classic DBA schemes, the CO mainly allocates bandwidth based on the requested bandwidth of ONU-APs. In predictive DBA schemes, the CO adopts traffic prediction to estimate and pre-allocate surplus bandwidth in addition to the requested amount [106]. In the following subsection, we discuss some typical classic and predictive DBA schemes that are adopted in the existing literature.

### 2.3.1 Classic DBA Schemes

Fig. 2.11 illustrates the DBA operation using REPORT and GATE messages. As presented in the figure, an ONU-AP sends a REPORT message piggybacked to
its uplink data, to request $BW_{req}$ amount of bandwidth for its next transmission. This $BW_{req}$ indicates the queue length buffered at the ONU-AP, i.e., how many bytes of packets are buffered at the ONU-AP. Upon receiving the REPORT, the CO calculates the bandwidth demand of the ONU-AP and generate a GATE message to allocate bandwidth to the ONU-AP. The amount of bandwidth allocated is represented as $BW_{grant}$ in Fig. 2.11. Repeating this above procedure, the CO allocates bandwidth to all ONU-APs in a round-robin manner. The time interval between consecutive transmissions from an ONU-AP is known as the transmission polling cycle.

It is important to note that different $BW_{grant}$ allocated by the CO will impact the latency performance. The fundamental classic DBA scheme is the fixed-cycle DBA scheme. In this scheme, the CO allocates the same fixed amount of bandwidth, i.e., a fixed $BW_{grant}$, to each ONU-APs in each polling cycle. It can be noted that this scheme operates in a TDMA manner and is easy to implement. However, due to a lack of flexibility in adjusting $BW_{grant}$ based on the dynamic queue length of ONU-APs, a DBA scheme which adheres to a fixed transmission polling cycle, results in high transmission latency that increases linearly with polling cycle duration. The exhaustive-service DBA scheme aims to cater to the bandwidth demand of all ONU-APs by allocating $BW_{grant}$ equal to $BW_{req}$. This effectively reduces the latency compared to the fixed-cycle DBA scheme, however, the uplink bandwidth will be monopolised by heavily-loaded ONU-APs [109]. To overcome this disadvantage, in the limited-service DBA scheme, the CO sets a
limit, $BW_{max}$, on $BW_{grant}$, which allocates $BW_{grant}$ as follows:

$$BW_{grant} = \min\{BW_{req}, BW_{max}\}$$ (2.2)

The bandwidth allocation as described in (2.2) ensures the amount of bandwidth allocated among ONU-APs is fair as $BW_{grant}$ cannot exceed $BW_{max}$. Taking a step further, the constant credit DBA scheme is proposed, by which the CO allocates an additional amount of bandwidth to the $BW_{req}$ as follows:

$$BW_{grant} = \min\{BW_{req} + C, BW_{max}\}$$ (2.3)

The rationale behind allocating an extra amount of bandwidth $C$ is that packets arrive to ONU-APs always need to be reported first before getting bandwidth for transmission. Allocating an additional bandwidth $C$ would allow the arrivals to be transmitted directly without needing to be reported to the CO, thereby reducing the latency. The linear credit DBA scheme achieves the same objective by multiplying $BW_{req}$ with a constant coefficient $c$ as follows:

$$BW_{grant} = \min\{BW_{req} \times c, BW_{max}\}$$ (2.4)

When $c$ is greater than 1, additional bandwidth is allocated such that unreported arrivals can be transmitted. In constant and linear credit DBA schemes, how the additional bandwidth $C$ and $c$ is determined, is crucial to reducing the latency in PONs. The latency performances of the aforementioned classic DBA schemes have been comprehensively analysed and compared in existing research [42,106-108]. Results show that the latency performances of exhaustive-service and limited-service DBA schemes are similar, and the latency improvement is marginal when using constant credit and linear credit DBA schemes. As such, the limited-service DBA scheme is the commonly-adopted baseline scheme in many research studies investigating DBA algorithms in PONs.

Note that both constant and linear credit DBA schemes aim to reduce the latency over the limited-service DBA scheme by allocating some additional bandwidth for the arriving packets. The performance, however, is limited as determining a suitable $C$ or $c$ values is challenging. In this regard, predictive DBA schemes have been proposed in state-of-the-art studies to take advantage of traffic prediction to estimate bandwidth demand. Details of the operation and
challenges in the predictive DBA schemes are presented next.

### 2.3.2 Predictive DBA Schemes

As discussed earlier, in classic DBA schemes, arriving packets cannot be transmitted directly in the current polling cycle. Instead, new arrivals need to be reported to the CO first and then transmitted in the next transmission timeslot allocated by the CO. As such, the average latency of classic DBA schemes is on average 1.5 times of the polling cycle duration, accounting for 0.5 polling cycle duration for new packets to be reported and 1 polling cycle duration for these packets to wait for its transmission timeslot [42]. To reduce the latency, predictive DBA schemes predict the incoming arrivals, estimate bandwidth demand of ONU-APs, and allocates the estimated bandwidth to ONU-APs such that the arrivals can be transmitted without needing to be reported. The operation of predictive DBA schemes is illustrated in Fig. 2.12. As shown in the figure, after sending a REPORT the CO, an ONU-AP continues to receive arriving packets. The CO or the ONU-APs will estimate bandwidth, $BW_{\text{pred}}$, for these new arrivals. The estimated $BW_{\text{pred}}$ is then allocated together with $BW_{\text{req}}$ to the ONU-APs such that the arriving packets can be transmitted without having to request bandwidth from the CO. Consequently, the latency is expected to be reduced by 0.5 polling cycle time compared to the classic DBA schemes.

It should be noted that unlike constant and linear credit DBA schemes that allocate a constant $BW_{\text{pred}}$ to ONU-APs in each polling cycle, predictive DBA schemes estimate $BW_{\text{pred}}$ based on incoming arrivals at individual ONU-APs as shown in Fig. 2.12. Overall, a CO that adopts predictive DBA schemes allocates $BW_{\text{grant}}$ in GATE as follows.

$$BW_{\text{grant}} = \min\{BW_{\text{req}} + BW_{\text{pred}}, BW_{\text{max}}\} \tag{2.5}$$

where $BW_{\text{max}}$ is the maximum available bandwidth that can be allocated to ONU-APs. Based on (2.5), the performance of predictive DBA schemes relies on $BW_{\text{pred}}$. An accurate $BW_{\text{pred}}$ that can accommodate the newly arrived packets will effectively help reduce the latency [110]. An inaccurate prediction of bandwidth, on the other hand, can lead to over- or under-provisioning of bandwidth and result in increased latency [106]. However, the underlying
algorithms used for accurate $\text{BW}_{\text{pred}}$ estimation are dependent on traffic characteristics and therefore, differ across different application scenarios. In the following subsections, we discuss some prevalent bandwidth estimation algorithms that have been used in predictive DBA schemes.

### 2.3.3 Statistical Bandwidth Estimation Algorithms

The estimation of $\text{BW}_{\text{pred}}$, in general, relies on the statistical characteristics of the traffic. Different bandwidth estimation algorithms have been proposed in existing studies to use different traffic feature to predict $\text{BW}_{\text{pred}}$. In general, existing algorithms estimate $\text{BW}_{\text{pred}}$ by using long-term or short-term traffic statistics such as bit rate, arrival rate, packet inter-arrival time, arrivals (in bytes) in past polling cycles at ONU-APs. The characteristics and application scenarios of these algorithms are analysed in detail in this section.

#### 2.3.3.1 Bandwidth Estimation for CBR Traffic

Content-centric applications such as voice, video, and voice-over-IP are widely supported in currently-deployed communication networks. A number of algorithms in existing literature such as [111-117] focused on traffic generated by these applications. These applications typically generate constant bit rate (CBR)
traffic, making the $\text{BW}_{\text{pred}}$ estimation straightforward. As discussed in the previous section, with accurate traffic predictions, the average latency can be effectively reduced from $\sim 1.5$ polling cycle duration towards $\sim 0.5$ polling cycle duration. Research in [113][114] estimated $\text{BW}_{\text{pred}}$ of ONU-APs in integrated fiber-wireless networks, considering packets received from the WLAN segment. The $\text{BW}_{\text{pred}}$ estimation is made as follows:

$$
\text{BW}_{\text{pred}} = \left[ \frac{T_{\text{poll}}}{T_{\text{CBR}}} \right] \times L_{\text{CBR}}
$$

in which $T_{\text{poll}}$ is the duration of a polling cycle, $T_{\text{CBR}}$ is the period of CBR arrivals (seconds/packet), $R_{\text{PON}}$ is the data rate of the link/channel (bytes/second) and $L_{\text{CBR}}$ is the packet length (bytes). Recall that $[x]$ is the ceiling function to accommodate an integer number of arrivals. Following the same concept, the authors in [115,116] proposed a M-DBA scheme. In the proposed scheme, the baseband unit pre-allocated bandwidth to remote radio head for CBR traffic from users. Research in [117] proposed a data mining forecasting DBA scheme, termed DAMA, for video streams. DAMA utilised machine learning technique, $k$-nearest neighborhood ($k$NN), to estimate $\text{BW}_{\text{pred}}$ by averaging the number of packets in $k$ number of past polling cycles that have similar durations to the current cycle. The average latency of DAMA was shown as $\approx 0.5$ polling cycle duration. However, the effectiveness of DAMA was only validated with constant-flow videos. From the above analysis, the predictive DBA schemes following (2.6) can improve the latency of CBR traffic, but the performance could be affected if the traffic pattern variates [42].

### 2.3.3.2 Inter-Arrival Time based Bandwidth Estimation

Another type of predictive DBA schemes estimate $\text{BW}_{\text{pred}}$ using packet inter-arrival time information. The inter-arrival time is an important feature considered in network traffic prediction and classification, etc [118]. The inter-arrival time indicates the statistical characteristic of an arrival process, hence, can be used to predict packet arrivals within a certain period. Research in [119] proposed to estimate the inter-arrival time using Bayesian estimation, and then estimate $\text{BW}_{\text{pred}}$ accordingly. Results show a faster and more accurate
inter-arrival time estimation compared to the arithmetic average and exponential smoothing based estimation in [120] and [121]. In [122], the authors developed a linear predictor to predict the number of arrivals in user datagram protocol (UDP) flows. The proposed algorithm estimated packets accumulated within a polling cycle, thereby yielding $BW_{\text{pred}}$. In the studies mentioned above, $BW_{\text{pred}}$ estimation using inter-arrival time can be represented as follows:

$$BW_{\text{pred}} = \frac{1}{T_{\text{inter}}} \times T_{\text{poll}} \times L_{\text{avg}}$$

(2.7)

in which $\lambda$ is the packet arrival rate, i.e., the reciprocal of the average inter-arrival time $T_{\text{inter}}$ in between consecutive packets, and $L_{\text{avg}}$ is the average packet length in optical access networks. The average inter-arrival time used in predicting $BW_{\text{pred}}$ in (2.7) reflects the statistic feature of the traffic over the network. When the traffic is bursty, the effectiveness of the predictive DBA schemes using such prediction algorithms may deteriorate [123].

### 2.3.3.3 Moving Average based Bandwidth Prediction

Existing studies on Internet traffic reported that the aggregated traffic from multiple applications is in fact highly bursty [124-126]. This burstiness is featured by alternating ON and OFF intervals. In an ON interval, consecutive packets generated by end devices/robots are wirelessly transmitted through the WLANs and aggregated at ONU-APs, whereas no packet is generated in the OFF interval.

From Ethernet LAN data traces collected from 1989 to 1992, Taqqu et al. reported the phenomenon of traffic burstiness in the aggregated arrival process. Traffic burstiness is characterised by self-similarity and long-range dependence (LRD) [124] where the Hurst parameter, denoted as $H$, is utilised to parameterise the burstiness. The reason for the observed burstiness is explained in [125] as a result of superpositions of alternating ON and OFF processes. Numerical research papers reported consistent findings regarding aggregated wide-area and local area traffic. Research in [126] investigated and corroborated the LRD property based on Internet traffic traces in the past 17 years. Experimental investigations based on the traffic monitored in major cities in France again
proved Taqqu’s results and showed that burstiness and LRD can be well constructed by aggregation of ON/OFF processes [127]. More recently, on-going experiments of H2M applications over teleoperation systems clearly showed an ON/OFF pattern in interactive H2M traffic similar to that in conventional Internet applications [32,33].

For bursty traffic, the variance of ON/OFF intervals can be infinite, which means that long-lasting packets transmissions/idle intervals are likely to occur. To estimate $BW_{pred}$ for bursty traffic, subsequent research studies focused on making short-term bandwidth prediction. The predictive DBA scheme in [128], termed most-recently (MR)-DBA, used the most-recently received packets from the previous polling cycle to predict $BW_{pred}$. Uplink latency was shown to be reduced as compared to the limited-service DBA. Further, the limited sharing DBA scheme with traffic prediction (LSTP-DBA) [129] exploited a 4-order autoregressive model, i.e., counting on the received packets in the previous 4 polling cycles, to estimate $BW_{pred}$. The underlying principle of $BW_{pred}$ estimation in LSTP-DBA scheme is explained below.

Considering $BW_{pred}$ estimation for an ONU-AP, let us denote $BW_{pred}(i)$ as the predicted arrivals in the $i$-th polling cycle, $BW_{co,rec}(i)$ as the actually received packets by the CO. The estimated $BW_{pred}(i+1)$ for the $(i+1)$-th polling cycle in LSTP-DBA is formulated as follows:

$$BW_{pred}(i+1) = \sum_{k=0}^{3} w_{i,k} BW_{co,rec}(i-k)\quad (2.8)$$

in which $w_{i,k}$ is the coefficients. An iterative method was proposed to update $w_{i,k}$ based on the estimation error. Simulation results show that LSTP-DBA scheme outperforms classic DBA schemes regarding latency performance. The authors in [130] improved LSTP-DBA scheme by including additional terms of downlink bandwidth, while, the basic autoregressive estimation model was the same.

Summarising from the above discussions on estimating $BW_{pred}$ in the predictive DBA scheme, the following key characteristics and open issues in existing predictive DBA schemes have been identified:

- The selection of bandwidth estimation algorithms depends on traffic. In order to accurately estimate $BW_{pred}$ and improve the performance of
predictive DBA schemes, knowledge of traffic characteristics is demanded. Hence, to achieve accurate $BW_{\text{pred}}$ estimation and effective bandwidth allocation for H2M applications, the traffic characteristics of H2M applications need to be studied and understood.

- **Existing $BW_{\text{pred}}$ estimation algorithms highly rely on packet arrival statistics, e.g., inter-arrival time, the number of arrivals in the previous polling cycles. However, other network features, such as network configurations, protocol parameters, are not taken into account. As such, the impact of diverse network features on $BW_{\text{pred}}$ estimation and the latency performance of predictive DBA schemes is unclear.**

- **The latency bottleneck for bursty traffic.** Since predictive DBA algorithms allocate bandwidth in (2.5) following the limited-service discipline, $BW_{\text{max}}$ constrains the maximum bandwidth that can be allocated to individual ONU-APs. When long packet queues are buffered at ONU-APs especially when the traffic is bursty, $BW_{\text{max}}$ would be allocated to ONU-APs in each polling cycle regardless of $BW_{\text{pred}}$. As such, the performance of applying $BW_{\text{pred}}$ in the predictive DBA scheme will be limited.

In light of the above, we present a detailed latency performance evaluation and comparison between existing DBA schemes, including both classic and predictive schemes in Chapters 5 and 6. Our proposed solutions to address the above-mentioned open issues arising from the inadequacy of the existing scheme in support bursty H2M traffic, will be presented. Specifically, in Chapter 7, we will focus on investigating the H2M traffic characteristics in addressing the currently limited understandings on improving existing DBA schemes for future applications.

### 2.4 Machine Learning for Bandwidth Resource Allocation

#### 2.4.1 Overview: Opportunities and Challenges

In recent history, machine learning and artificial intelligence (AI) have gained increasing popularity in various areas in our society such as economy, marketing, transportation systems, and industrial automation. Researchers have explored
the potential of ML in communication networks, and the benefits brought by ML in different aspects have been witnessed [131-133]. For example, studies reported in [134] and [135] have applied deep learning to facilitate routing path selections at individual routers. Compared to benchmark routing algorithms, the deep learning solution significantly reduces routing latency and signalling overhead in communication networks [136]. In the areas of network reliability and security, ML techniques have been utilised to predict and detect network failures [137], identify intrusive user behaviours [138]. Network traffic classification is another important area that uses ML techniques to identify traffic from different applications [133,139], elephant and mice flows [140], useful or useless data [141] in order to allocate network resources efficiently. Based on recent research, ML techniques have distinctive characteristics. The selection of ML techniques and how to apply them depend on the underlying network problems.

As our focus is on reducing communication latency for emerging low-latency H2M applications, in this thesis, we explore the use of ML to examine the latency performance of access networks and to improve existing bandwidth allocation solutions for H2M applications. Note that in the previous sections, we have overviewed existing bandwidth allocation schemes in PONs and channel access mechanisms in WLANs. In the following overview, we will present the technical details of ML techniques, including kNN, decision tree (DT), naïve Bayes (NB), logistic regression (LR), support vector machine (SVM), and artificial neural network (ANN), and how each technique was used in existing literature for bandwidth allocation in heterogeneous PON and WLAN networks. Moreover, when using ML for such purposes, it is important to determine which ML technique will be suitable and should be used to address the underlying problem. For this purpose, we will compare the characteristics, pros and cons of different ML techniques in our overview. This discussion includes the time and space complexity to show the time cost and memory usage associated with each technique. In the next section, we will discuss the key concepts involved in ML.

### 2.4.2 Machine Learning Techniques Used in Bandwidth Resource Allocation

In general, ML solves two types of problems namely, classification and
2.4 Machine Learning for Bandwidth Resource Allocation

regression [142]. In a classification problem, ML techniques are used to identify the class or group that a given sample belongs. A sample is a data point that represents an entity, object, or observation to be classified. The selected key features of a sample, parameterised in a vector or matrix, are the input used by a ML technique to yield a classification. The classification result is known as the output. Note that the classification outputs are generally discrete values that represent different classes. In comparison, in a regression problem, a ML technique can be used to predict numeric values and fit curves. The outputs are generally continuous values. Supervised learning, unsupervised learning, and semi-supervised learning are the main methods adopted in ML to characterise the abstract associations between inputs and outputs, thereby solving classification or regression problems [143]. The principles of these three methods are introduced as follows:

- **Supervised learning:** Training sets are essential for supervised learning. A training set contains labelled samples whose input features and outputs are explicitly known. Based on the example input-output pairs provided in the training sets, the mapping between the input and output can be learned. This process is known as supervised training. When the training is complete, prediction or classification for any new inputs can be achieved.

- **Unsupervised learning:** Unsupervised learning does not require training sets and supervised training phases. Rather, it explores hidden patterns of any given samples [143]. Therefore, unsupervised learning can be implemented using unlabelled samples whose correct outputs are unknown. One common unsupervised learning example is clustering, where samples are grouped into different clusters based on the similarity of their input features.

- **Semi-supervised learning:** Semi-supervised learning falls between unsupervised and supervised learning, utilising both labelled and unlabelled samples. For example, a small set of labelled samples can be utilised to validate the unsupervised learning outcomes, thereby improving the learning algorithm. The outputs yielded in unsupervised learning can be used to label samples, which provide a reference in
examining the performance of a learning algorithm in turn.

A specific ML technique can be used for either classification or regression purpose, and learning can be achieved via either supervised, unsupervised and semi-supervised learning depending on the selection of ML techniques and the formulation of a problem. In the following subsections, we analyse commonly-used ML techniques, including kNN, DT, NB, LR, SVM, and ANN, in network control and resource allocation, and survey on their present use in bandwidth allocation over heterogeneous PON and WLAN networks. Except kNN, the use of these techniques primarily relies on supervised training in the literature as will be overviewed in each subsection.

2.4.2.1 \textit{k}-Nearest Neighborhood

The \textit{k}-Nearest Neighborhood (kNN) technique is a simple ML technique that achieves classification or regression of a sample referring to \textit{k} number of example samples that are closest to this sample, e.g., in Euclidean distance. As such, kNN does not require supervised training and can be used given either labelled or unlabelled samples. Since the rest techniques are supervised learning-based, for consistency, we explain the technical details of kNN when a training set is provided.

Let us denote a training set as $S = \{(x_i, y_i) | i = 1, \ldots, N\}$ in which, $(x_i, y_i)$ is a labelled training sample where the parameters $x_i, y_i,$ and $N$ represent input feature vector, labelled output and the total number of samples in a training set, respectively. For a given new sample with input vector $x$, kNN finds top $k$ inputs, $x_i$, in the training set $S$ that are most similar to the new input $x$. A common measure of this similarity is Euclidean distance and $k$ is an integer usually less than 20 [143]. Let us denote $D_k(x, S)$ as the subset containing the top $k$ similar samples found by kNN. At this point, a majority vote approach is adopted, where the majority class in $D_k(x, S)$ is selected as the output class of the new sample. Similarly, for a regression problem, kNN averages the output values in $D_k(x, S)$ as the fitted output value for the new sample as follows:

\begin{equation}
    y = \frac{1}{k} \sum_{i \in D_k(x, S)} y_i
\end{equation}

\textit{Related work:} The DAMA DBA scheme proposed in [117] utilises kNN to estimate $BW_{pred}$ by averaging the number of packets in $k$ number of past polling
cycles that have similar durations to the current cycle. In DAMA, the CO utilises past polling cycle durations of an ONU-AP as an input feature and the arrival data (in bytes), as the target output. In this case, \( S = \{(x_i, y_i) \mid x_i = \text{duration of the } i\text{-th polling cycle}; y_i = \text{arrival data in the } i\text{-th polling cycle}\} \). For a new cycle with duration \( x \), the predicted arrivals can be estimated using (2.9). As discussed in Section 2.3, using \( k\)NN, the CO adopting DAMA scheme pre-allocates bandwidth for arriving packets, thereby reducing the latency. The effectiveness of DAMA scheme is validated using video streams.

**Pros and cons:** The \( k\)NN is simple to implement and does not require any prior statistical knowledge and assumptions on the samples. However, \( k\)NN results in high time and computation costs. In [144], we investigated the capability of \( k\)NN in predicting the busy/idle status of ONUs when the aggregated traffic is bursty in nature. Our results highlight that using \( k\)NN in a DBA scheme yields long computation times. For instance, in estimating \( BW_{\text{pred}} \), \( k\)NN needs memory at the CO to store training sets (one for each ONU-AP) for the Euclidean distance comparison. Moreover, with \( k\)NN, any input feature, \( x \), needs to be compared with all \( x_i \) in \( S \). This procedure is computationally inefficient and slow in operation, especially when \( S \) is large.

**Time and space complexity:** Let us assume that an input vector \( x \) contains \( p \) number of features, and recall \( N \) is the number of samples in the training set \( S \). The time complexity of \( k\)NN can be estimated as \( O(Np) \) [145]. This indicates that the worst-case time cost in making a classification grows linearly with \( Np \). The space complexity of \( k\)NN can also be estimated as \( O(Np) \). This reflects the memory usage to store the reference samples when using \( k\)NN technique.

### 2.4.2.2 Decision Tree and Naïve Bayes Classifiers

Decision tree and naïve Bayes are mainly used for classification purposes [143]. With a training set \( S \), DT continuously splits the data samples into different branches based on the most informative features, thereby creating a tree hierarchy with nodes representing the features considered for splitting and associated branches directing to split groups. The splitting of samples terminates when all the samples are classified and the split groups, termed as leaf nodes in a DT, purely contain the same-class samples. As a result, supervised training of DT...
results in a tree structure with layers of nodes and branches. Then, the class of a new sample is determined by tracing the features through the trained DT until this sample falls into one of the leaf nodes, which indicates the classification result. In comparison, NB relies on Bayesian theorem [16], which make classifications based on statistical knowledge. Provided the training set \( S = \{(x_i, y_i) | i = 1, ..., M\} \), the conditional probability of output class, \( y = c \), of a given input vector \( x = \{x_1, x_2, ..., x_n\} \), can be calculated as follows:

\[
P(y = c | x) = \frac{P(y = c) \prod_{j=1}^{n} P(x_j | y = c)}{P(x)} \tag{2.10}
\]

where the statistics of \( P(y = c) \) and \( P(x_j | y = c) \) can be easily calculated based on \( S \).

Note that to classify \( x \) into \( y = \{c_1, c_2, ..., c_p\} \) different classes, the denominator \( P(x) \) remains the same for any \( P(y = c_k | x) \) to be computed by (2.10). Therefore, the exact value of \( P(x) \) is insignificant, and the numerator in (2.10) determines the likelihood of the class of \( x \).

**Related work:** DT and NB techniques are primarily utilised for traffic classification, and thereby for improving bandwidth allocation decisions in existing literature [140, 147-148]. In [149], authors applied DT and NB to classify ‘mice’ and ‘elephant’ traffic flows at a packet-switched data centre. Using this classification result, the authors proposed an intra-data centre bandwidth scheduling to prioritise long packet queues and delayed packets. Research in [150] proposed and compared \( k\)NN and DT classifier to justify the quality-of-transmission over an optical link, and then adaptively allocate bandwidth and wavelength accordingly. In particular, their results showed that DT is significantly faster in making classification decisions than \( k\)NN.

**Pros and cons:** The tree structure in DT provides easy and direct interpretations of the input features [143]. However, the possible risk of overfitting due to taking many trivial features into account is one of the drawbacks of DT. Moreover, both DT and NB are sensitive to the methods followed in preparing samples to some extent [144]. In general, DT and NB are most suitable for classification than regression problems and prefer discrete-valued input features.

**Time and space complexity:** The time complexity for training DT ranges
from $O(N \log N)$ to $O(N^2)$. Once a tree has been built, classifying an input is quick, with a worst-case complexity of $O(d)$, where $d$ is the depth of the tree [151]. If we further denote the total number of leaf nodes as $l$, the space complexity can be estimated by $O(ldp)$ [152]. For the NB technique, the time complexity in training is $O(Np)$ and the resulting space complexity is $O(pc)$, where $c$ is the number of output classes [153]. NB is also quicker when classifying a single input, with a time complexity of $O(pc)$.

### 2.4.2.3 Logistic Regression

Based on the maximum likelihood estimation, LR intends to determine a line that best split the inputs, i.e., binary classification [143]. LR associates input $x$ and output $y$ using sigmoid function $g(z) = 1/(1+e^{-z})$, where $z = \theta^T x + b$ and $y \in \{0, 1\}$. The sigmoid function $g(z)$ ranges from $[0, 1]$, and can be interpreted as mapping input $x$ to a probability $P(y = 1 | x) = g(\theta^T x + b)$. An illustration of the sigmoid function and logistic regression is presented in Fig. 2.13. The more $z = \theta^T x + b$ is away from 0, the more we trust the classification results. Therefore, by maximising the likelihood $P(y | x) = [P(y = 1 | x)]^y [1 - P(y = 1 | x)]^{1-y}$ with respect to the training set $S$, $\theta$ and $b$ can be derived. Then, given any new input, the output class can be determined using $g(z)$. Moreover, note that $g(z)$ realises a non-linear mapping from $x$ to $y$ as plotted in Fig. 2.13. The sigmoid function $g(z)$ is selected as one of the functions used in the neuron units in neural networks as will be introduced later.

**Related work:** LR is primarily used for classification and a common use is to

![Figure 2.13 An illustration of sigmoid function and logistic regression.](image)
classify user traffic. In [154], authors compared the NB and LR classifiers in predicting the active ($y = 1$) and idle ($y = 0$) status of mobile users and pre-allocated spectrum resource blocks to active users. The results in this study showed that LR outperforms NB with a more accurate prediction on user active/idle status and incurs less time cost in making a prediction. The proposed spectrum resource allocation based on user status prediction reduced call block ratio and increased network capacity. The study in [155] investigated Internet service classification using LR and studied the impact of different input features on the classification results.

**Pros and cons:** LR can be easily implemented using the sigmoid function. However, LR is typically used to classifications, and prone to underfitting due to its shallow structure [143].

**Time and space complexity:** The supervised training time mainly relies on optimising the parameters $\theta$ and $b$. The training time complexity is approximately $O(p^2N + p^3)$ in solving optimisation via gradient descent method [143]. Once the training completes, computation for classification is $O(p)$. The memory usage of LR is minor as it only needs to store the optimised parameter $\theta$ and $b$.

### 2.4.2.4 Support Vector Machine

Unlike DT, NB and LR classifiers, SVM formulates an optimisation problem with the objective to determine a hyperplane $\mathbf{w}^T\mathbf{x} + b$ that acts as the decision boundary to splits the inputs. Recall that LR splits samples via maximising the likelihood using sigmoid function, i.e., $z = \theta^T\mathbf{x} + b >> 0$ for $y = 1$ or $z = \theta^T\mathbf{x} + b << 0$.

![An illustration of support vector machine.](image)
2.4 Machine Learning for Bandwidth Resource Allocation

for \( y = 0 \) as shown in Fig. 2.13. In comparison, SVM maximises the distance \( \frac{(w^T x_i + b)}{\|w\|} \), termed as margin, of the inputs \( x_i \) to the decision boundary described by the hyperplane as shown in Fig. 2.14. SVM maps input \( x \) to the output value \( y \in \{ -1, 1 \} \) and a larger value of \( y \times \frac{(w^T x_i + b)}{\|w\|} \) indicates a greater margin, and therefore, more confidence in the classification decisions. The optimisation problem in convex form is finally derived as follows [142]:

\[
\min_{w, b} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{m} \xi_i \tag{2.11}
\]

subject to:

\[
y_i \left( w^T x_i + b \right) \geq 1 - \xi_i, \quad i = 1, 2, \ldots, m \tag{2.12}
\]

\[
\xi_i \geq 0 \tag{2.13}
\]

where \( C \) and \( \xi_i \) are parameters used when samples cannot be separated linearly. This optimisation problem can be further solved using the Lagrangian dual optimisation with the following objective and constraints [142]:

\[
\min_{\alpha} W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j < x_i, x_j > \tag{2.14}
\]

subject to:

\[
C \geq \alpha_i \geq 0, \quad i = 1, 2, \ldots, m \tag{2.15}
\]

\[
\sum_{i=1}^{m} \alpha_i y_i = 0 \tag{2.16}
\]

Note that when solving this dual optimisation problem, only samples with a margin equal to 1 have non-zero \( \alpha_i \) and the input vectors of such samples are termed support vectors. Using support vectors, the hyperplane used to classify samples can be determined by solving (2.11)-(2.13). Moreover, in many cases, inputs \( x_i \) are mapped to a higher-dimensional space, achieved by feature mapping in \( \phi(x_i) \), in order to better separate the samples. Kernel trick is hence played to define the value of inner product of \( < x_i , x_j > \) as \( K(x_i, x_j) = \phi^T(x_i) \phi(x_j) \), in which \( K(x_i, x_j) \) represents the inner product of \( x_i \) and \( x_j \). By selecting an appropriate kernel, \( K(\cdot, \cdot) \), the mapping, \( \phi(\cdot) \), the performance of SVM can be improved.

**Related work:** As a classifier, SVM has been studied for Internet traffic classifications and comparisons have been made with DT, NB, and neural
networks [154]. In [156], authors adopted SVM to predict the traffic flows, i.e., load and pattern variations, in a WLAN network and compared the prediction accuracy with the neural networks. For bandwidth allocation, research in [157] considered the data aggregated from cellular networks and exploited SVM to classify useful and useless data. When requesting bandwidth and transmitting to the CO, ONUs sifted useless data such that bandwidth can be allocated and utilised more efficiently.

**Pros and cons:** SVM is typically utilised for classification problems. Classification results are sensitive to the selection of kernels and parameters $C$ and $\xi$. The original form of SVM is designed for binary classification. For multi-class classification, individual SVMs may need to be trained to separate one class from the rest [158]. SVM can be well trained with a small-size training set. However, when the training set size increases, the training time may grow exponentially.

**Time and space complexity:** The time complexity for supervised training can be estimated by $O(N^3)$ [159]. When training is completed, classification time can be estimated by $O(N_{sv}p)$, where $N_{sv}$ is the number of support vectors. Note that compared to the previous techniques introduced, the training time complexity of SVM is proportional to $N^3$. As such, SVM is generally used when the training set size is relatively small, e.g., tens or a few hundreds of samples.

### 2.4.2.5 Artificial Neural Networks

Compared to the aforementioned ML techniques, ANNs have greater flexibility in addressing a variety of problems such as curve fitting, prediction, pattern recognition, classification, and etc. This is attributed to the ANNs’ capability to learn and characterise highly complex patterns and non-linear structures of data samples [146]. An ANN adopts layered learning architecture, i.e., input, output, and hidden layer, comprising neuron units, which are non-linear activation functions to resemble how the human brain learns. To illustrate the learning process of an ANN, one common type of ANNs, feedforward ANN, is presented in Fig. 2.15. In a feedforward ANN, input and intermediate results move from the input layer to the output layer. The neuron units in each layer are non-linear functions that map the associated inputs to an output. General activation
functions include sigmoid function, ReLU, and tanh function [158]. Recall that the LR technique uses a single sigmoid function to achieve classification. As such, LR can be viewed as a type of ANN with only one layer and a single sigmoid neuron unit. By selecting suitable architectures and training methods, ANNs can be used for solving various problems. A multiple layered ANN is also known as a deep neural network (DNN), which can better extract features and learn complex associations than a shallow single-layer ANN [160].

In supervised training, an ANN iteratively adjusts the parameters in the activation functions and weight matrix, $W$, associated with its neurons and layers as shown in Fig. 2.15 to minimise the error between its predicted outputs and target outputs, thereby accurately characterising the pattern between input features and output targets. Optimisation methods such as gradient descent and batch gradient descent can be applied to optimise the parameters in the activation functions and weight matrix between layers.

**Related Work:** Recently, numerous studies have exploited ANNs to classify network traffic, predict bandwidth resource demand, and supervise resource allocation decisions. In [161], the authors used a feed-forward ANN to estimate the bandwidth demand from mobile fronthaul and pre-allocate that bandwidth in the DBA schemes. The proposed ANN uses the most recent packet arrivals to predict the incoming packets, thereby pre-allocating bandwidth for future arrivals. Research in [162] utilised a recurrent neural network (RNN) to predict the traffic

![Figure 2.15 An illustration of ANN layered network architecture.](image-url)
volume hour by hour in an optical backbone and dynamically allocate resource based on traffic prediction. Similarly, in [163], the authors exploited a RNN to predict incoming traffic 30 min ahead in a cloud radio access network and reconfigure resource allocation among baseband units. In the above studies, the ANN is used to predict traffic, and thereby estimate bandwidth demand for specific networks such as mobile fronthaul and for time periods such as peak hours. ANNs have also been used to analyse network performances such as QoS metrics, to facilitate adaptive resource allocations. In [164], the authors proposed to use an ANN to predict the QoS performance and optimise resource allocation accordingly. In [165], an ANN was proposed to learn and tune a PID controller to achieve better QoS provision. Research studies [166, 167] also investigated the use of ANN for resource allocation in WLANs. In [166], an ANN was utilised to predict the idle and busy time slots in different WLAN channels. Research in [167] used an ANN to first learn the impact of historical traffic load in different time scales such as in a minute and an hour, on the selection of WLAN channels. Based on this knowledge, the future traffic load on each channel is predicted, enabling them to select the best channel for transmission.

Pros and cons: Based on our analysis of ML techniques, neural networks are more flexible compared to the other ML techniques we discussed in the previous paragraphs. Different types of classification and prediction problems in existing communication networks are addressed via ANNs by carefully designed architecture and training. ANNs are sophisticated in characterising complex nonlinear relationships between dependent and independent variables. However, ANNs face the risk of converging to local minimum/maximum values when determining its weights and parameters. This drawback can be overcome by testing, validating, and by appropriate initialization and selection of learning algorithms [136].

Time and space complexity: For ANNs, the primary time cost is attributed to the training. The training time and resultant space complexity could vary in different use cases. In general, for a trained ANN, the time complexity in making a classification or prediction can be estimated by $O(pn_1 + n_1n_2 + \ldots + n_in_{i+1} + \ldots)$, where $n_i$ is the number of neurons in the $i$-th layer. The memory usage is mainly attributed to storing the weights and bias that are optimised.
In summary, existing use of ML techniques has the following characteristics:

- The selection of ML techniques depends on the property of tasks, e.g., classification or regression, linear or non-linear. The pros and cons also need to be considered when selecting an appropriate ML technique for a given problem.

- Generally, $k$NN-based classifier spends the longest time in making a prediction with time complexity of $O(Np)$. The DT, NB, and LT techniques are more suitable for classification problems as they report a low time complexity in making a classification. In comparison, both SVM and ANN generally demand longer training time. The challenge in SVM lies in selecting the kernel and optimisation parameters for the specific problem. The use of ANNs, on the other hand, is more flexible in addressing different types of problems.

- In the aspect of network control and resource allocations, ML techniques have been used for different purposes such as traffic prediction, bandwidth demand estimation, classification of user activity, and channel status in existing literature discussed above. Using ML techniques in these areas benefits effective bandwidth resource allocation in optical and wireless networks.

It can be noted that among the aforementioned ML techniques, ANN is the most versatile technique that has been used in a wide variety of practical problems. As such, in this thesis, we consider neural networks and deep learning as promising techniques to improve bandwidth allocation solutions in supporting low-latency applications in heterogeneous PON and WLAN networks. Critical issues such as the selection of input features, output targets, design of architecture, and the implementation of supervised training will be discussed in the following chapters.

### 2.5 Summary

Existing WBANs and heterogeneous access networks that integrate WLANs and PONs are important networks that can support emerging H2M applications. The latency in these networks must be reduced in meeting the stringent
end-to-end latency demand for H2M applications. The MAC layer channel access mechanisms in WBANs and WLANs and bandwidth allocation schemes in PONs directly affects the network latency performance. As such, it is important to study existing MAC layer latency performance and explore novel MAC layer solutions in addressing the latency challenge.

In this chapter, we presented a comprehensive overview of current MAC protocols in the WBAN, WLAN, and PON. For WBANs, we overviewed the existing standards used for WBAN systems, including the IEEE 802.15.4 WPAN, IEEE 802.15.6 WBAN, and ETSI SmartBAN. The MAC layer protocols and channel access mechanism designs in these standards were presented and compared in detail. Specifically, we focused on the open issues in the recently-proposed SmartBAN and highlighted the necessity to investigate the hybrid MAC channel access, in conjunction with the traffic characteristics, to achieve both low latency and high energy efficiency.

Moreover, for converged application delivery over heterogeneous access networks, we investigated the channel access mechanisms defined by the IEEE 802.11 WLAN MAC protocol and the DBA schemes in PONs. In WLANs, technical details of the three fundamental channel access mechanisms, DCF, PCF, and HCF, were introduced. The advantages and existing issues in each mechanism were analysed. In our investigation, we reported the utilisation of HCF mechanism for low latency applications and introduced the reference HCCA design. We also discussed the open areas to be explored in WLAN MAC in reducing the latency. In PONs, we presented existing DBA schemes both classic and predictive, in facilitating uplink bandwidth allocation to ONU-APs. In classic DBA schemes, the operation principle and latency performance of fixed-cycle, exhaustive-service, limited-exhaustive, constant, and linear credit schemes were analysed and compared. Further, we introduced prevalent statistical bandwidth estimation algorithms used to predict resource allocation. We reported different statistic traffic features used in these algorithms to estimate bandwidth for ONU-APs and analysed challenges in these algorithms.

In exploring improved bandwidth allocation solutions that reduce the latency over heterogeneous access networks, we overviewed state-of-the-art research on using ML techniques in resource allocation in both optical and wireless access
networks. We summarised popular ML techniques including kNN, DT, NB, LR, SVM, and ANN, and reported their existing applications in resource allocations. The discussion included technical details, related studies, the pros and cons, and the time and space complexity of each technique. Based on our overview, we highlighted the potential of exploiting ML, particularly ANN and deep learning, to address challenges in bandwidth resource allocation over heterogeneous PON and WLAN access networks.

Overall, from Sections 2.1 to 2.3, state-of-the-art research on MAC layer channel access mechanisms in WBANs and WLANs, and DBA schemes in PONs are overviewed. Recent research in using ML techniques in improving the discussed MAC layer solutions in optical and wireless networks are presented in Section 2.4. We highlighted the opportunities and challenges in using ML techniques to reduce latency over heterogeneous access networks. Taking existing work and current challenges presented in this chapter into consideration, we propose multiple solutions for WBAN and heterogeneous PON and WLAN networks. In Chapters 3 to 7, we present our technical contributions in addressing these challenges in detail.
Chapter 3

SmartBAN Uplink Latency and Energy Performance Study: A Time-Optimised MAC for Periodic Monitoring and Emergency Traffic

3.1 Introduction

Wireless body area networks (WBANs) reshape healthcare services towards ubiquitous and personalised e-health by deploying biomedical sensors and actuators on or in the vicinity of the human body. Emerging real-time and remotely-controlled applications, such as telediagnosis and telerehabilitation, aim to assist individuals in emergency situations anytime and anywhere. However, as introduced in the previous chapters, WBAN systems at present are primarily designed to maintain low energy consumption since miniaturized sensors and actuators are limited in battery capacity and are expected to work sustainably. The latency performance of existing WBANs is comprised to some extent. In realising future remote e-health applications, both low latency and high energy efficiency need to be achieved in WBAN systems. As such, it is critical to accurately assess the latency performance of WBANs. Note that the medium access control (MAC) layer channel access mechanisms govern when and how the sensors shall access the channel, and therefore impact the latency and energy performances of WBANs. This focus of this chapter is hence on evaluating the performances of existing MAC channel access mechanisms and explore novel solutions to improve the latency and energy performances for WBANs.

As reviewed in Chapter 2, the Smart Body Area Network (SmartBAN) is a recently-proposed WBAN system. Compared to conventional WBANs, SmartBAN is specifically designed to achieve low system complexity and ultra-low power
Chapter 3 SmartBAN Uplink Latency and Energy Performance Study: A Time-Optimised MAC for Periodic Monitoring and Emergency Traffic

As a newly-standardised WBAN system, SmartBANs have not been fully investigated, in terms of its latency and energy consumption, in existing studies. In particular, data transmissions from SmartBAN sensors are governed by a hybrid beacon-enabled MAC, whereby different access periods, i.e., scheduled access period (SAP), control/management period (CMP), inactive period (IAP), are specified to allow flexible exploitation of both collision-free scheduled channel access and contention-based access mechanisms [168]. The time interval of sending the channel beacon, known as inter-beacon interval (IBI), as well as the setting of individual access periods, i.e., SAP, CMP, and IAP, impact the latency and energy performances of SmartBANs significantly. However, to date, such impact of these MAC timing parameters on the SmartBAN performances has not been critically studied, and consequently, there is a lack of criteria on how to determine suitable timing parameters for SmartBAN hybrid MAC frame in achieving low latency and high energy efficiency.

In light of the above, this chapter presents a comprehensive study of SmartBAN uplink transmission latency and energy performances. The focus is on the uplink direction since current WBANs, including SmartBANs, are mainly used for the purposes of health monitoring and supervision, and such applications are often uplink transmission dominated [169-171]. As illustrated in Fig. 3.1, the SmartBAN is organised in a star topology and the uplink latency studied in this chapter refers to the waiting time upon a data packet generated by a sensor before uplink transmission. To understand the impact of the timing parameters in SmartBAN hybrid MAC on uplink latency and energy performances, two common

![Figure 3.1 SmartBAN topology and uplink transmission.](image_url)
3.1 Introduction

types of traffic in health monitoring scenarios are considered in this study, i.e., periodic monitoring traffic [172] and sporadic emergency (EM) traffic [173]. Accurate analytical uplink latency models are then developed for the monitoring and EM traffic, respectively. Further, by exploiting the developed models, analytically evaluations of the impact of the timing parameters on latency and energy performances are presented. A time-optimised MAC framework that reduces the latency and energy consumption for both periodic monitoring and EM traffic in SmartBANs is proposed.

Specifically, under the proposed time-optimised MAC framework, we show that the latency and energy consumption of transmitting periodic monitoring traffic is highly dependent on the duration of IBI. An Optimal IBI frame is hence proposed for SmartBANs to reduce the uplink latency of periodic monitoring packets. Further, in achieving energy-savings, an adaptive inter-beacon interval (A-IBI) algorithm is proposed to determine the Optimal IBI during the network connection between SmartBAN hub and its attached sensors. Then, sleep mode in sensors and doze mode in the hub are introduced to reduce energy consumption under the proposed framework. For EM traffic, we present our analysis on the impact of the durations of SAP, CMP, and IAP on the latency and energy performance of transmitting EM packets. An Optimal CMP is proposed to support the immediate transmission of EM packets in meeting certain latency constraints. Extensive simulations validate the developed models and the effectiveness of the proposed MAC framework. With Optimal IBI, the percentage of energy-savings and channel efficiency for periodic monitoring are evaluated and significant performance improvement is shown compared to the IEEE 802.15.6-defined WBAN in simulations. With Optimal CMP, energy-savings for EM packet transmission are achieved and latency constraints are met. The contributions of this chapter are threefold: (a) development of closed-form analytical latency models for SmartBAN hybrid MAC considering both periodic monitoring and EM traffic; (b) provision of an understanding of the impact of MAC timing parameters on the latency and energy performances in SmartBANs, and (c) proposal of a time-optimised MAC framework that determines a hybrid MAC frame that achieves low latency and high energy-savings for SmartBANs.

The rest of this chapter is organised as follows. In Section 3.2, we introduce the
details of SmartBAN MAC protocol and the uplink transmission mechanisms for both periodic monitoring and EM traffic. In Section 3.3, we derive the closed-form expressions for periodic and EM uplink latency. A time-optimised MAC framework that determines Optimal IBI and Optimal CMP, and meanwhile schedules the sleep and active time for sensors is proposed. In Section 3.4, we verify the accuracy of our latency models through simulations. The effectiveness of the proposed framework in reducing uplink latency and energy consumption for periodic monitoring traffic is validated. We also evaluate the latency and energy consumption with varying EM traffic load in detail. Finally, our main findings are summarized in Section 3.5.

3.2 SmartBAN MAC Protocol and Uplink Transmission

3.2.1 SmartBAN MAC Frame

A SmartBAN is organised in a star topology with one central hub and a group of sensors. To enable communication between the hub and sensors, a SmartBAN utilises two types of channels, a control channel (CCH) and a data channel (DCH). The operation frequency of a SmartBAN falls in the Industrial Scientific Medical (ISM) band (2401-2481MHz). The channel bandwidth is 2MHz and central frequency thus is $f_c = 2402 + 2 \times n$ MHz, where $n = 0$ to 39. There are 3 CCHs and 37 DCHs in total and at any time, a SmartBAN uses a pair of CCH and DCH to facilitate on-body communication.

Fig. 3.2 illustrates the framework of the SmartBAN CCH and DCH. The hub periodically transmits a control channel beacon (C Beacon) on CCH to broadcast channel information and control frames (Fig. 3.2(a)). Data, control, and management packets transmission is carried out on DCH. As shown in Fig. 3.2(b), the working cycle of DCH is controlled by a data beacon (D Beacon), which is sent periodically. The time interval between consecutive D Beacons, termed IBI, is separated into distinct periods, namely: (a) SAP with time division multiplex access (TDMA) uplink transmission; (b) CMP with slotted ALOHA for both uplink and downlink; and (c) IAP where no transmission is allowed for both the hub and sensors. A SmartBAN MAC frame is time-slotted. The duration of the slot unit, SAP, CMP, IAP, and IBI are denoted as $T_{\text{min}}$, $T_{\text{SAP}}$, $T_{\text{CMP}}$, $T_{\text{IAP}}$ and $T_{\text{IBI}}$, respectively.
3.2 SmartBAN MAC Protocol and Uplink Transmission

Moreover, as shown in Fig. 3.2(b), a time slot, denoted as $T$, utilises multiple slot units for data frame transmission, while, D Beacon is sent in a single slot unit. The transmission time structure for both SAP and CMP is shown in Fig. 3.3.
Chapter 3 SmartBAN Uplink Latency and Energy Performance Study: A Time-Optimised MAC for Periodic Monitoring and Emergency Traffic

3.3(a) [40]. After data transmission, an acknowledgment frame (ACK) is sent to indicate successful transmission. At least one inter-frame space (IFS) is followed after data or ACK transmission. The duration of IFS is set to 150 μs [40]. The SmartBAN MAC protocol data unit (MPDU), is illustrated in Fig. 3.3(b). The length of a MAC frame depends on the data type. For ACK, the frame body \( L_d = 0 \). The length of an ACK frame is 64 bits. The D Beacon has a length of 144 bits [40].

To assist the selection of timing parameters in Section 3.3 and Section 3.4 and to analyse the SmartBAN MAC performance, we present an estimation on the duration of \( T_{IBI} \) based on the data generated by biomedical sensors in practical healthcare applications. First, let us assume the total number of sensors in the SmartBAN is \( N \). Further, consider that the designated TDMA slot in contains \( l \) unit slots for packet transmission. The duration of a TDMA slot is hence \( T = l \times T_{\text{min}} \), where \( T_{\text{min}} = 625 \, \mu\text{s} \) and \( l = 2^b \) (\( b = 0,1,2,3,4,5 \)) [40]. Within an IBI, the total number of unit slots is \( L \), where \( L = T_{\text{IBI}} / T_{\text{min}} \). We have also assumed that the durations of a TDMA slot and an ALOHA slot are equal. The number of time slots within SAP and CMP is represented as \( N_s \) and \( N_{CM} \), respectively. A summary of the notations for the SmartBAN MAC and the parameters of the latency model derived in the following section is presented in Table 3.1.

According to SmartBAN PHY protocol, the symbol rate is 1 M symbols/s and the maximum data rate is 1 Mbps. As such, the durations for D Beacon and ACK transmissions are 200 μs and 120 μs, respectively. For data frames, the length of a MAC frame body varies with the actual data content generated by the hub and sensors. Take ECG as an example, the ECG sensor samples at 250 Hz and each sample is 8 bits [174]. The ECG sensor sends one packet every 100 ms. Hence each packet contains 250 Hz × 100 ms = 25 ECG samples, which is 200 bits. At the rate of 1 Mbps, the transmission time is less than 1 ms. Considering the extreme case where the maximum length of the MAC body is 1600 bits [54]. The transmission time can then be calculated as \((56 + 48 + 1600 + 16 + 64) \, \text{bits}/1 \, \text{Mbps}) + (150 \times 2 \, \mu\text{s}) = 2.084 \, \text{ms} \). With \( l = 4 \), \( T = 4 \times 625 \mu\text{s} = 2.5 \, \text{ms} \). Since SAP is based on TDMA, the length of SAP can be estimated by the number of sensors, \( N \). For example, to support a SmartBAN with 6 and 8 sensors, the duration of SAP should be at least \( 6 \times 2.5 \, \text{ms} \) and \( 8 \times 2.5 \, \text{ms} \), which is 15 ms and 20 ms, respectively. We will revisit these parameters in Section 3.4 when the
3.2 SmartBAN MAC Protocol and Uplink Transmission

Table 3.1 Notations for the SmartBAN latency modelling.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_{\text{min}}$</td>
<td>The minimum time unit</td>
<td>$T$</td>
<td>The duration of a time slot</td>
</tr>
<tr>
<td>$l$</td>
<td>The number of slot units in a time slot</td>
<td>$L$</td>
<td>Total number of slot units in an IBI</td>
</tr>
<tr>
<td>$N$</td>
<td>The number of sensors</td>
<td>$T_{\text{IBI}}$</td>
<td>The duration of an IBI ($T_{\text{IBI}} = LT_{\text{min}}$)</td>
</tr>
<tr>
<td>$l_{\text{S}}$</td>
<td>The number of slot units in SAP</td>
<td>$l_{\text{CM}}$</td>
<td>The number of slot units in CMP</td>
</tr>
<tr>
<td>$N_{\text{S}}$</td>
<td>The number of time slots in the SAP</td>
<td>$T_{\text{SAP}}$</td>
<td>The duration of the SAP ($T_{\text{SAP}} = N_{\text{S}}T = NT$)</td>
</tr>
<tr>
<td>$\Psi$</td>
<td>The number of time slots in the CMP</td>
<td>$T_{\text{CMP}}$</td>
<td>The duration of the CMP ($T_{\text{CMP}} = \Psi T$)</td>
</tr>
<tr>
<td>$p_n$</td>
<td>The slot unit number of the $n$-th packet</td>
<td>$T_i$</td>
<td>The data generation period of Sensor $i$</td>
</tr>
<tr>
<td>$m$</td>
<td>The first packet generation slot unit number</td>
<td>$k$</td>
<td>The beginning slot unit number of the designated TDMA time slot</td>
</tr>
<tr>
<td>$D_i$</td>
<td>The average uplink latency of Sensor $i$ with periodic traffic</td>
<td>$D_{\text{EM}}$</td>
<td>The average uplink latency of EM traffic</td>
</tr>
</tbody>
</table>

proposed A-IBI algorithm adopted in a 6- and 8-sensor SmartBAN is evaluated with simulations and theory.

3.2.2 Uplink Channel Access and Latency Models

The SAP in SmartBAN MAC is accessed by TDMA, and is utilised by default for uplink transmission. Therefore, we considered SAP for transmitting the periodic monitoring packets. On the other hand, EM packets are typically latency-sensitive, demanding timely transmission in life-critical situations. SmartBAN experimental study presented in [54] has considered utilising both SAP and CMP for delivering EM packets. We adopt this hybrid TDMA + ALOHA transmission mechanism for EM packets in both SAP and CMP and critically investigate the impact of the timing parameters in SmartBAN hybrid MAC on latency and energy performances. Since the SmartBAN MAC provides a hybrid frame for flexible access, the position in an IBI when a sensor generates a packet,
determines how the sensor performs channel access and therefore, leads to different uplink transmission latency. Figs. 3.4 (a) and (b) demonstrate the general case of periodic and EM packet generation and corresponding uplink latency, respectively.

For periodic uplink traffic, the data generation period of Sensor $i$ is denoted as $T_i$ ($i=1,2,...N$), as in Table 3.1. A sensor may generate data packets at any time slot during the IBI. Therefore, the uplink latency is the duration between when a packet is being generated and when it is successfully transmitted in its designated TDMA slot, as depicted in Fig. 3.4(a). To analyse this latency, we assume the first packet of Sensor $i$ is generated at the $m$-th slot during an IBI. Since the uplink data is generated periodically, the slot number of a subsequent data packet generated from the same sensor is deterministic. Therefore, we use $p_n$ to represent the slot number where the $n$th packet of Sensor $i$ is generated. With the known start time ($m$-th slot) and periodicity ($T_i$), we can derive $p_n$ as:

$$p_n = \left\lfloor \frac{(m-1)T_{min} + nT_i - T_{IBI}}{T_{IBI}} \right\rfloor + \left\lfloor \frac{(m-1)T_{min} + nT_i}{T_{IBI}} \right\rfloor + 1$$

(3.1)

where $\lfloor x \rfloor$ equals the maximum integer that is less than $x$. Since both hub and sensors transmit data periodically, we can simplify $T_i/T_{IBI}$ to $M/K$ ($M$ and $K$ are integers) by removing common factor(s) from the numerator, $T_i$, and denominator, $T_{IBI}$. Then, (3.1) shows that $p_n$ is periodic with period $K$ as proven below:

$$p_{n+K} = \left\lfloor \frac{m+(n+K)L}{K} \right\rfloor + \left\lfloor \frac{m+(n+K)LM}{K} - L \times \left\lfloor \frac{m-1}{L+M} \right\rfloor \right\rfloor + 1$$

(3.2)

From (3.2), the periodic packets’ generation slot number can be concluded in Proposition 1 as follows.

**Proposition 1:** The slot number where a periodic uplink packet is generated is periodic with period $K$, where $K$ satisfies $T_i/T_{IBI} = M/K$. 

Note that a TDMA time slot with $T = lT_{\text{min}}$ is allocated to an individual sensor during SAP, and that data generated within the designated slot needs to wait until next IBI for transmission as shown in Fig. 3.4(a). As such, we adopt the discrete unit step function $u(\cdot)$ to differentiate the latency when a packet is generated before its designated TDMA slot, or the other way around within an IBI. The average uplink latency, $D$, of periodic monitoring packets is modeled as:

$$D = \frac{1}{K} \sum_{n=1}^{K} [L \times u(p_n - k) + k - p_n] \times T_{\text{min}}$$  \hspace{1cm} (3.3)

In (3.3), the assigned TDMA slot number, $k$, of Sensor $i$ is the threshold of the unit step function. Consequently, $u(p_n - k)$ differentiates the latency in two cases: (a) when $p_n < k$, latency of the $n$-th packet arises from waiting for the TDMA time slot in current IBI, which is $k - p_n$; (b) otherwise when $p_n \geq k$, latency is attributed to waiting for the TDMA time slots in the next IBI, which is $L + k - p_n$.

For EM packets, both SAP and CMP are utilised for transmission. Therefore, the latency depends on the region in which packet is generated and to be sent by a sensor as illustrated in Fig. 3.4(b). Let us demote the inter-arrival time of EM packets as $X$. Assuming Poisson arrivals, $X$ follows an exponential distribution $X \sim \text{Exp}(G)$, where $G$ is the offered traffic load on DCH. As proven in [175], the arriving time of Poisson packets in a network is random regarding a fixed working cycle, i.e., the IBI in SmartBANs. As such, the distribution of the EM packets’ generation slot number is described as follows.

**Proposition 2:** The time slot number where an EM packet is generated at a sensor follows a uniform distribution $U([1, T_{\text{IBI}} / T])$.

As shown in Fig. 3.4(b), the uplink latency of an EM packet is determined by the following cases: (a) when an uplink packet is generated before its designated TDMA slot (Region 1) or during the IAP (Region 4), latency arises from waiting for its TDMA slot within the current IBI or in the next IBI accordingly; (b) if an EM is generated at Sensor $i$ after its TDMA starting slot $k$ and before CMP (Region 2), latency is composed of waiting for the CMP and contention access attempts within the CMP; and (c) for EM packets generated within the CMP (Region 3), latency is determined by how many times a packet is retransmitted until it is successfully sent to the hub. Therefore, with Proposition 2, the average uplink latency for an EM packet can be derived by averaging the latency when EM packets arrive at
each of the four Regions as shown in Fig. 3.4(b). Let $d_i$ $(i=1,2,3,4)$ denote the latency of an EM packet generated in the four Regions. Correspondingly, $P_i$ denotes the probability of an EM packet being generated in Region $i$. Since the slot number where an EM is generated within an IBI is random, $P_i$ can be derived as follows:

$$P_i = \begin{cases} 
k T_{\text{min}} / T_{\text{IBI}} & i = 1 \\
(l_S - k)T_{\text{min}} / T_{\text{IBI}} & i = 2 \\
(T_{\text{CMP}} - T) / T_{\text{IBI}} & i = 3 \\
(T_{\text{AP}} + T) / T_{\text{IBI}} & i = 4 \end{cases}$$

(3.4)

in which $l_S$ is the total number of slot units in SAP. Then, the average latency of EM packets can be derived as follows:

$$D_{\text{EM}} = \sum_{i=1}^{4} P_i d_i$$

(3.5)
Using (3.3) and (3.5), we further derive closed-form analytical models, which result in the proposed time-optimised MAC framework in Section 3.3.

### 3.3 Time-Optimised MAC for Low Latency and High Energy Efficiency

#### 3.3.1 Periodic Uplink Latency and Optimal IBI

**Analytical Modelling**

According to our estimation in Section 3.2.1, $T_{\text{IBI}}$ is usually shorter than the transmission period of sensors. Also, due to the high data rate in a SmartBAN, a relatively short $T_{\text{IBI}}$ can support multi-sensor transmissions. Therefore, for a sensor with transmission period $T_i$, $T_i$ can be represented by $T_i = qT_{\text{IBI}} + x$, where $q$ is an integer and $x$ is the remainder, $x \in [0, T_{\text{IBI}})$. Substitute $T_i = qT_{\text{IBI}} + x$ into (3.3), then, the average uplink latency of the sensor can be derived as follows:

$$D_i = \frac{1}{K} \sum_{k=1}^{K} \left\{ L \times u \left( m + \frac{nLx}{T_{\text{IBI}}} - L \times \left\lfloor \frac{m-1}{L + \frac{nx}{T_{\text{IBI}}} \right\rfloor - k \right) + k - \left\lfloor m + \frac{nLx}{T_{\text{IBI}}} - L \times \left\lfloor \frac{m-1}{L + \frac{nx}{T_{\text{IBI}}} \right\rfloor \right\} \right\} \times \frac{T_{\text{IBI}}}{L}$$  \hspace{1cm} (3.6)

Further, let us simplify (3.6) by substituting $x/T_{\text{IBI}}$ with $y$, $y \in [0,1)$ as follows:

$$D_i = \frac{1}{K} \sum_{k=1}^{K} \left\{ L \times u \left( m + \frac{nLy}{L} - L \times \left\lfloor \frac{m-1}{L + ny} \right\rfloor - k \right) + k - \left\lfloor m + \frac{nLy}{L} - L \times \left\lfloor \frac{m-1}{L + ny} \right\rfloor \right\} \times \frac{T_{\text{IBI}}}{L}$$  \hspace{1cm} (3.7)

$$D_i = \begin{cases} \frac{T_{\text{IBI}}}{L} (L + k - m) & m > k \\ \frac{T_{\text{IBI}}}{L} (k - m) & m \leq k \end{cases}$$  \hspace{1cm} (3.8)

From (3.6) and (3.7), it can be noted that $D_i$ depends on the values of $T_{\text{IBI}}$, assigned TDMA slot number $k$, as well as the initial packet generation slot number $m$. Fig. 3.5 provides an illustrative overview of how these parameters impact $D_i$. The data generation period of Sensor $i$, $T_i$ is chosen to be 1000 ms in this case [41]. Specifically, a 3D relationship among $k$, $m$, and $D_i$ is shown in Fig. 3.5(a) with an
Figure 3.5 Numerical analysis of periodic monitoring average uplink latency ($T_i=1000\text{ms}$).
arbitrarily chosen $T_{\text{IBI}}$ of 130 ms. Note that when $T_{\text{IBI}}$ is set as a divisor of $T_i$, the $D_i$ in Fig. 3.5(a) can be further reduced to that presented in Fig. 3.5(b). Fig. 3.5(b) plots $D_i$ in (3.7) with variable $m$, $k$, and $T_{\text{IBI}}$. Here, the values of $T_{\text{IBI}}$ are divisors of $T_i = 1000$ ms. It can be noted that when the $T_{\text{IBI}}$ is selected as a divisor of the $T_i$, i.e., $y = 0$ in (3.7) and $D_i$ can be rewritten in (3.8).

Considering a group of sensors in SmartBAN whose data generation period can be denoted as a sequence $\{T_i\}$, $T_{\text{IBI}}$ can, therefore, be set as the maximum common divisor of $\{T_i\}$. We refer to this $T_{\text{IBI}}$ value as the **Optimal IBI**. In this case, $D_i$ can be significantly minimised by manipulating $m$ and $k$ despite how vigorous the oscillation in $D_i$ is (refer to Fig. 3.5(b)). Hence, when an Optimal IBI is decided, the uplink latency can be further reduced by manipulating the initial packet position $m$. Fig. 3.5(c) (data period $T_i=1000\text{ms}$ and fixed $k = 2$) again shows that the Optimal IBI induces a larger latency deviation, which enables the latency minimisation through the synchronisation of $m$.

**A Discussion of $m$, $k$, and $T_{\text{IBI}}$**

It can be noticed in Fig. 3.5(c) that once Optimal IBI duration is determined, the manipulation of the $m$ is crucial for minimising the uplink latency. As such, we first present an analysis of the average uplink latency of a sensor when its initial packet generation slot number, $m$, is random. This average latency over random $m$ can be estimated as:

$$D_{\text{avg},j} = \frac{1}{LK} \sum_{m=1}^{L} \sum_{n=1}^{K} [L \times u(p_n - k) + k - p_n] \times T_{\text{min}}$$  \hspace{1cm} (3.9)

When the Optimal IBI is determined, (3.9) can be further simplified to:

$$D_{\text{avg},j} = \frac{T_{\text{IBI}}}{2} (1 - \frac{1}{L})$$  \hspace{1cm} (3.10)

From (3.10), it can be observed that when an Optimal IBI is chosen, without control of $m$ value, the average latency linearly increases with increasing $T_{\text{IBI}}$. As such, when a SmartBAN is established, the minimum uplink latency can be achieved if a sensor packet generation is synchronised to its TDMA slot which is a stringent synchronisation requirement for the SmartBAN especially in the situation that TDMA slots are dynamically allocated. As the periodic traffic, in general, is more tolerable to the latency than the EM packet, an alternative is to
synchronise $m$ to $L$ through the periodically-sent D Beacon. This manipulation of $m$ could be facilitated by the sensors with cognitive capability where physical layer parameters can be adjusted [176]. In this chapter, we focus on the impact of above-mentioned timing parameters on SmartBAN uplink latency and energy performances. Therefore, we do not further discuss the realisation of synchronisation, and we assume that the parameter, $m$ can be synchronised in the following analytical and simulation analysis.

To summarise, for periodic packets, we propose to select $T_{IBI}$ and the sensors’ transmission periods, denoted by sequence $\{T_i\}$, to minimise uplink latency. Based on the analytical model, the optimisation of $T_{IBI}$ is made to best suit the sensor group to minimise uplink latency. Then, the necessity to synchronise sensors’ packet generation in reducing latency has been discussed.

### 3.3.2 Adaptive-IBI Algorithm for Energy-Savings

A vital demand of BAN design, including the SmartBAN, is to reduce energy consumption. To maintain low power operation, sleep-capable sensors that can transit between active and sleep modes are typically utilised. The selection of $T_{IBI}$ of a SmartBAN hence impacts the energy performance significantly when such sensors are deployed. Moreover, the hub, in general, is in active, i.e., the transmitter and receiver are powered up, at all times, as is considered in previous BAN studies [64,65]. This can be energy consuming and inefficient especially in SmartBANs where the $T_{IBI}$ is usually short. In this work, we consider that the hub of the SmartBAN to have doze mode capability, whereby its transmitter can be powered down during the inactive period. We consider SmartBAN sensors to have sleep mode capability whereby both transmitter and receiver can be powered down during periods of inactivity. Harnessing sleep-capable sensors and hub with doze function, we propose an adaptive IBI algorithm (A-IBI) based on the proposed Optimal IBI in Section 3.3.1 to reduce uplink latency as well as enhance energy efficiency for daily health monitoring applications.

**Sleep Mode for SmartBAN**

During each IBI, there exist fixed idle time slots in IAP and SAP and some potential idle time slots in CMP for hub and sensors. In turn, during these idle time slots, the sensors can transition into sleep mode. Fig.3.6 illustrates a sensor’s
possible sleep durations within an IBI in accordance with uplink and downlink traffic. As shown in Fig.3.6, a sensor will wake up at the end of IAP, just in time before the start of the next IBI to receive the D Beacon from the hub. From the D Beacon, each sensor will learn the arrival time of its assigned TDMA slot and whether there will be downlink data to receive during CMP. Within SAP, each sensor remains active only during its designated TDMA slot for uplink transmission, and sleeps during TDMA slots allocated to other sensors. During CMP, sensors that do not have downlink data to receive from the hub will also sleep, waking up just in time for the D Beacon in next IBI. In IAP where no transmission is allowed, both the hub and sensors can sleep. Before the end of each IAP, the hub will wake up to send a D Beacon and will remain active during SAP and CMP. The processes for a sensor to switch from active to sleep mode (A2S) and sleep to active (S2A) are represented by the gray triangular slopes in Fig.3.6 and are considered in our evaluation below.

Here, we evaluate the worst energy consuming case whereby sensors are continuously kept in active mode during the designated TDMA slots for uplink transmission and during the CMP for possible downlink packets. We evaluate the energy-savings arising from using a sleep mode capable SmartBAN as compared to an always active SmartBAN where both hub and sensors are always on. The energy-savings of one sensor node is hence defined as follows:

$$\eta_{\text{sensor}} = (1 - \frac{T_{\text{active}} P_{\text{active, sensor}} + T_{\text{sleep}} P_{\text{sleep, sensor}} + T_{\text{switch}} P_{\text{switch}}}{T_{\text{IBI}} P_{\text{active, sensor}}}) \times 100\% \quad (3.11)$$

**Figure. 3.6** SmartBAN transmission and operation modes.
In (3.11), $T_{sleep}$ and $T_{active}$ represent the sleep and active intervals within an IBI, respectively. The time incurred for switching between A2S and S2A is denoted $T_{switch}$. In sleep mode, sensors operate in ultra-low power $P_{sleep, sensor}$. In active mode, the hub and sensors transmit data with maximum power $P_{active, hub}$, $P_{active, sensor}$ respectively. The power consumption for A2S and S2A of a sensor is denoted $P_{switch}$. In an always-on SmartBAN, $T_{IBI} = T_{active}$ as indicated in the denominator in (3.11). In comparison, in sleep mode capable SmartBAN, $T_{IBI} = T_{active} + T_{sleep} + T_{switch}$.

Moreover, sleep mode (where both transmitter and receiver are powered down) is not applicable to the hub as this leads to high re-synchronisation overhead. Hence, we consider only the doze mode for the hub, where only the transmitter is powered down. For performance evaluation, we estimate the energy-savings of the hub in doze mode as follows:

$$\eta_{hub} = (1 - \frac{T_{active, hub} P_{active, hub}}{T_{IBI} P_{active, hub}} + \frac{T_{doze, hub} P_{doze, hub}}{T_{IBI} P_{active, hub}}) \times 100\% \quad (3.12)$$

in which $T_{active, hub}$ and $T_{doze, hub}$ denote the hub active and doze intervals within the IBI, respectively. The power for the hub in doze mode is denoted $P_{doze, hub}$. For $N$ sensors in the Smart BAN, the average energy-savings of sensors in SmartBAN can be evaluated as follows:

$$\eta_{avg} = \frac{1}{N} \sum_{i=1}^{N} \eta_{sensor, i} \quad (3.13)$$

**Adaptive IBI Algorithm**

For a SmartBAN with $N$ sensors, we propose an adaptive IBI (A-IBI) algorithm to ensure that the SmartBAN will be established to operate with Optimal IBI. Sensors within the SmartBAN will be notified of the time slot duration and the total number of time slots in an IBI from the C Beacon and D Beacon, respectively [33]. Thus, upon receiving the Beacons, sensors will be made aware of the IBI duration. For a SmartBAN to determine its Optimal IBI, it is important for the hub to learn the parameters of all its sensors, especially the transmission period $T_i$. Therefore, we propose to utilise the reserved bits within the connection request (C-req) on DCH to relay the $T_i$ to the hub [40]. The C-req is a frame sent by new
sensors for connection to the hub. When the hub learns \( \{T_i\} \) from all sensors, the Optimal IBI can then be determined. Consequently, the hub will set \( T_{\text{SAP}} \) according to \( T_{\text{SAP}} = N \times T_s \). The shortest IBI is then obtained by setting \( T_{\text{CMP}} = T_{\text{SAP}} \). With A-IBI, the shortest IBI is extended (through extending the IAP) such that the total length of IBI is the maximum common divisor of \( \{T_i\} \), i.e. Optimal IBI. Therefore, an optimised MAC frame achieved by A-IBI can increase the energy efficiency of a SmartBAN, and at the same time, minimise uplink latency. The A-IBI algorithm is illustrated below:

Since the Optimal IBI is a divisor of \( T_i \), there is the possibility that during one or even several SAPs, a sensor does not have uplink data packet to send. Then, the assigned SAP slots to that sensor for uplink transmission will remain idle. Therefore, we define a channel efficiency function to evaluate the average uplink channel usage. Let us denote the number of idle slots during SAP in \( j \)-th IBI as \( N_{\text{idle}}(j) \) and the total number of slots during SAP period is \( N_s \). The channel efficiency function, \( \eta_{\text{channel}} \), is defined as follows:
\[ \eta_{\text{channel}} = \lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} \left(1 - \frac{N_{\text{idle}}(j)}{N_{\text{SAP}}} \right) \times 100\% \]  

(3.14) can be considered in evaluating the channel utilisation of schedule-based access mechanism, i.e. TDMA for all types of traffic not limited to periodic monitoring traffic.

### 3.3.3 EM Uplink Latency and Optimal CMP

**Analytical Modelling**

To analyse the uplink latency for EM traffic, we derive a closed-form solution for (3.7) by developing a state transition diagram as illustrated in Fig. 3.7. This is because EM packets exploit both SAP and CMP for uplink transmission, whereby a hybrid TDMA + ALOHA access shall be analysed.

For slotted ALOHA access, with an offered traffic load of \( G \), the successful access and transmission probability can be denoted as \( p_s = \exp\{-GT\} \) [165]. In Fig. 3.7, the state ‘Idle’ represents the state where no packet is waiting in the buffer of Sensor \( i \). When a packet is generated and is to be sent by Sensor \( i \), the sensor will transit from Idle to either waiting for the TDMA slot (\( W_{\text{TDMA}} \)) state, waiting for the ALOHA slot (\( W_r, r=1\ldots\Psi - \kappa \)) state, or waiting in IAP (\( W_{\text{IAP}} \)) state. When Sensor \( i \) in \( W_{\text{TDMA}} \) or \( W_{\text{IAP}} \) state obtains its designated TDMA slot, its current state (\( W_{\text{TDMA}} \) or \( W_{\text{IAP}} \)) changes to \( T_{\text{TDMA}} \) state, which represents TDMA transmission. \( W_r \) is the waiting state before the r-th transmission in the CMP. \( T_r \) represents the r-th transmission state.
transmission attempt. When a packet is to be transmitted at the \( \kappa \)-th ALOHA slot, the maximum number of transmission attempts allowed is \( \Psi \cdot \kappa \), where \( \Psi \) is the total number of ALOHA slots (recall the notation in Table 3.1). The probability that a packet within the CMP arrives at the \( \kappa \)-th ALOHA slot is \( 1/(\Psi-1) \). Consequently, the probability of a sensor transitioning from Idle state to CMP access is \( P_{\text{CMP}}(\kappa) \), and is derived as follows:

\[
P_{\text{CMP}}(\kappa) = \begin{cases} 
\frac{1}{\Psi - 1} P_3 & \kappa \in [1, \Psi - 1] \\
\frac{1}{P_3} & \kappa = 0
\end{cases} \quad (3.15)
\]

For an EM packet using slotted ALOHA, the probability of successfully transmitting at the \( r \)-th attempt is \( P[R = r] = (1-p_s)^{(r-1)} p_s \). For a packet sent at the \( \kappa \)-th ALOHA slot, the probability that it will fail to transmit \( \Psi \cdot \kappa \) times is represented by the probability \( (1-p_s)^{\Psi \cdot \kappa} \). Therefore, the average transmission attempt times, \( R_\kappa \), during the CMP can be obtained as follows:

\[
R_\kappa = \sum_{r=1}^{\Psi \cdot \kappa} rP[R = r] + (\Psi - \kappa)(1 - p_s)^{\Psi \cdot \kappa} = \frac{1}{P_3} [1 - (1-p_s)^{\Psi \cdot \kappa}] \quad (3.16)
\]

The average contention access latency, \( D_{\text{CMP}}(\kappa) \), for slotted ALOHA access within the CMP is determined by how many times a packet is retransmitted, before being successful. \( D_{\text{CMP}}(\kappa) \) can be mathematically derived as follows:

\[
D_{\text{CMP}}(\kappa) = \sum_{r=1}^{\Psi \cdot \kappa} rTP[R = r] + [\Psi - \kappa]T + T_{\text{IAP}} + kT_{\text{min}} \] \((1 - p_s)^{\Psi \cdot \kappa} \\
= \frac{T}{P_3} [1 - (1 - p_s)^{\Psi \cdot \kappa}] + (T_{\text{IAP}} + kT_{\text{min}})(1 - p_s)^{\Psi \cdot \kappa} \quad (3.17)
\]

According to Fig. 3.7, when a sensor generates an EM packet in Region 2, the sensor has to first wait for CMP and then transmit the packet in the ALOHA slots. As such, the maximum number of transmission attempts is \( \Psi \). Thus, \( \kappa = 0 \) in (3.17) and the average latency, \( d_2 \), in (3.7) can be derived as follows:

\[
d_2 = (l_s - k)T_{\text{min}} / 2 + \frac{T}{P_3} + (T_{\text{IAP}} + kT_{\text{min}} - \frac{T}{P_3})(1 - p_s)^\Psi \quad (3.18)
\]

Similarly, the average latency \( d_3 \) for Region 3 can be calculated by averaging
the waiting time for all possible $\kappa$ values as follows:

$$d_3 = T/2 + \sum_{\kappa=1}^{\psi-1} D_{CMP}(\kappa) = T/2 + \frac{T}{p_s} + \frac{1}{\psi-1}(T_{LAP} + k T_{min} - T)\left[\frac{(1 - p_s) - (1 - p_s)^{\psi-1}}{p_s}\right]$$ \hspace{1cm} (3.19)

Since an EM packet in Regions 1 and 4 is not sent using slotted ALOHA, the latency is caused by waiting for TDMA slot. Therefore, the average latency $d_1$ and $d_4$ can be represented as (3.20) and (3.21), respectively.

$$d_1 = k T_{min} / 2$$ \hspace{1cm} (3.20)

$$d_4 = (T_{LAP} + T) / 2 + k T_{min}$$ \hspace{1cm} (3.21)

Substituting $d_1, d_2, d_3, d_4$ into $D_{EM}$ in (3.7), $D_{EM}$ for EM arrivals can be derived as:

$$D_{EM} = \frac{1}{T_{lat}}\left[\frac{(k T_{min})^2}{2} + (T_{LAP} + T)(T_{LAP} + T/k T_{min}) + (I_S - k) T_{min}\left[\frac{(I_S - k) T_{min}}{2}\right]ight]

+ \frac{T}{p_s} + (T_{LAP} + k T_{min} - T)\left[\frac{T_{IAP} - T}{2} + \frac{T}{p_s}\right] + \frac{1}{\psi-1}(T_{LAP} + k T_{min} - T)\left[\frac{(1 - p_s) - (1 - p_s)^{\psi-1}}{p_s}\right]\right]$$ \hspace{1cm} (3.22)

Similarly, to reduce energy consumption, sensors can switch to sleep mode when it does not have EM packet to transmit. Specifically, a sensor sends an EM packet by either TDMA in SAP or contend to access the channel by slotted ALOHA. Once EM packet is successfully sent during CMP or IAP arrives, the sensor switches to sleep mode to save energy. Since EM packets are latency-sensitive, we estimate the energy consumed in transmitting EM packets in both SAP and CMP. Given that the power consumption of a sensor in active and in sleep modes are $P_{active,sensor}$ and $P_{sleep,sensor}$, respectively, the energy consumption of a single sensor is derived as follows:

$$\Phi = \frac{1}{T_{lat}}\left[\left(\sum_{\kappa=0}^{\psi-1} P_{CMP}(\kappa)R_s\right)T_{active,sensor} + P_{sleep,sensor}T(\psi - \left(\sum_{\kappa=0}^{\psi-1} P_{CMP}(\kappa)R_s\right)) + P_{sleep,sensor}T_{LAP}

= (P_{active,sensor} - P_{sleep,sensor})\left\{P_s \frac{1}{p_s}[1 - (1 - p_s)\psi] + P_s \frac{1}{p_s}[1 - (1 - p_s)^{\psi-1}]T\left[\frac{(1 - p_s) - (1 - p_s)^{\psi-1}}{\psi - 1}\right]\right\} + P_{sleep,sensor}(T_{IAP} + T_{LAP})$$ \hspace{1cm} (3.23)
Latency and Energy Consumption Trade-off

The average uplink latency in (3.22) and energy consumption in (3.23) as a function of $T_{IBI}$ and $T_{SAP} + T_{CMP}$ are depicted in Fig. 3.8. It should be noted that with TDMA, $T_{SAP}$ is generally a fixed duration depending on the number of sensors, therefore, a longer $T_{IBI}$ indicates a longer $T_{IAP}$ considering a given $T_{CMP}$ value. As shown in Fig. 3.8(a), increasing $T_{IBI}$ leads to more opportunity for sensors to sleep in IAP and thereby reducing energy consumption. From Fig.
Chapter 3 SmartBAN Uplink Latency and Energy Performance Study: A Time-Optimised MAC for Periodic Monitoring and Emergency Traffic

3.8(b), a long $T_{IBI}$ is shown to not always result in high latency if we increase the duration of $T_{CMP}$. Therefore, we propose to determine optimal CMP by investigating the trade-off between latency and energy when an IBI is preset, e.g., the Optimal IBI proposed in the previous section. When an IBI is determined, the impact of $T_{CMP}$ on the average EM uplink latency and energy consumption is presented in Fig. 3.9. The selection of $T_{CMP}$ is proposed as follows.

**Optimal CMP:** From Fig. 3.9, $T_{CMP}$ and hence $T_{IAP}$ can be optimised by minimising the energy consumption derived in (3.23), subject to the EM latency constraint of $D_{EM} < D_{ST}$ (constraint condition for EM traffic). If there is no $D_{ST}$, for example, the intersection point in Fig. 3.9 can be selected to determine $T_{CMP}$ so that both latency and energy consumption can be minimised.

### 3.3.4 Proposed Time-Optimised MAC framework

From the above analytical studies, the dependency of SmartBAN uplink latency and energy performances on the timing parameters, i.e., $T_{SAP}$, $T_{CMP}$, $T_{IAP}$, and $T_{IBI}$, of a hybrid MAC frame is comprehensively discussed. A time-optimised MAC framework can be derived to achieve low latency and high energy efficiency for both periodic monitoring and EM packets, which provides the first criterion for selecting suitable timing parameters in SmartBAN implementations.

The proposed time-optimised MAC framework is described as follows. For
3.4 Performance Evaluations

periodic traffic, the A-IBI algorithm can be adopted to initialise MAC layer functions of a SmartBAN, in which the Optimal IBI can be configured and energy-saving mechanism that enables periodic sleep of sensors and doze of the hub are stipulated. Moreover, within the Optimal IBI, the duration of SAP is set depending on the number of sensors within a SmartBAN and the duration of a TDMA slot. As such, in supporting latency-sensitive EM packets, the Optimal CMP is derived considering a trade-off between latency and energy consumption.

In the following section, we first validate the developed analytical models for periodic monitoring and EM traffic, respectively. The effectiveness of the proposed MAC framework with the Optimal IBI, Optimal CMP, and energy-savings achieved via A-IBI algorithm is verified and evaluated via extensive simulations.

3.4 Performance Evaluations

3.4.1 Periodic Uplink Latency and Optimal IBI Validation

We verify our periodic uplink latency analytical model with simulations and demonstrated the effectiveness of the proposed Optimal IBI for SmartBANs. Firstly, three different sensor uplink transmission periods, namely 500 ms, 900 ms, and 1300 ms, are arbitrarily chosen to analyse the average uplink latency with varying $T_{IBI}$. The average latency results are presented in Fig. 3.10. To evaluate the influence of $m$, $k$ and $T_{IBI}$ on $D_i$, we assume a fixed total slot number of $L = 30$ and a TDMA slot of $l = 2$. In our simulations, the initial packet generation slot $m$ is synchronised to the last slot ($m=L$) of the IBI, which is in compliance with our proposed optimised framework.

For a given designated TDMA slot number $k$ (arbitrarily chosen as $k = 2$ in Fig. 3.10(a) and $k = 4$ in Fig. 3.10(b)), results in Fig. 3.10 show that the average latency can be significantly reduced when $T_{IBI}$ is set as a divisor of each $T_i$, i.e. the Optimal IBI. Comparing Figs. 3.10(a) and (b), we can observe that a smaller $k$ gives rise to lower uplink latency. Simulation results in Fig. 3.10 verify the effectiveness in minimising periodic uplink latency with our proposed Optimal IBI for SmartBAN MAC.

The average latency results in Fig. 3.10 are obtained with $m$ synchronised to
the $L$-th slot. In Fig. 3.11, the average latency as a function of $T_{IBI}$ is obtained with different initial packet arrival time slot $m$. As shown in Fig. 3.11, when IBI duration is a divisor of $T_i$, ($T_i = 750$ ms for illustrative purpose), the average latency with different choices of $m$ varies largely and is in compliance with the theoretical latency given by (3.9). From Fig. 3.11, it is important to manipulate a sensor’s initial packet generation when the Optimal IBI has been set. Meanwhile, we can again observe that a minimum latency can only be achieved when $T_{IBI}$ is a divisor of $T_i$. 

Figure 3.10 Average latency $D_i$ with varying $T_{IBI}$ and $T_i$. 

(a) $m=L=30, k=2, l=2$

(b) $m=L=30, k=4, l=2$
Overall, the results in Fig. 3.10 to Fig. 3.11 have proven the effectiveness of the proposed Optimal IBI for SmartBAN. In the following sub-section, we evaluate the effectiveness of our proposed A-IBI algorithm in reducing uplink latency and energy consumption in two SmartBANs.

### 3.4.2 Adaptive-IBI Performance Evaluations

Using MATLAB, two SmartBANs, one with 6 and the other with 8 on-body sensors, were considered to test our proposed A-IBI algorithm. The number of sensors chosen in each SmartBAN is typical of the number of on-body medical sensors as reported in [177]. Each sensor has a different application and hence transmits a different type of periodic uplink data, with parameters highlighted in Tables 3.2(a) and (b), respectively. We evaluate the latency and energy-saving performances of both of these SmartBANs using simulations and theory.

As discussed in Section 3.2.1, the length of an Interframe Space, D Beacon, and ACK frame is 150 μs, 200 μs, and 120 μs, respectively. As specified in the SmartBAN MAC, we set the duration of a unit slot $T_{\text{min}}$ as 625μs. As evaluated in Section 3.3.1, we consider a TDMA slot to contain 4 slot units (4 × 625μs = 2.5 ms), to ensure that a maximum length uplink packet can be transmitted within a TDMA slot. Note that a D Beacon possesses a single time slot. We can then

![Figure 3.11 Average latency $D_i$ with varying $T_{\text{IBI}}$ and $m$ ($T_i = 750$ ms).](image-url)
estimate that the shortest IBI for the 6- and 8-sensor SmartBANs to be (6 + 1)× 2.5 ms × 2 = 35 ms and (8 + 1)× 2.5 ms × 2 = 45 ms, respectively.

Figs. 3.12(a) and (b) plot the average uplink latency, $D_i$, of each sensor in the 6- and 8-sensor SmartBANs, respectively with and without adopting A-IBI. Note that without A-IBI, the shortest IBI is implemented whereas with A-IBI, the IAP is extended such that $T_{\text{IBI}}$ is equivalent to the Optimal IBI. Results from Figs. 3.12(a) and (b) show that the manipulation (the two synchronisation cases discussed in Section 3.3.2) of $m$ under Optimal IBI can effectively reduce the latency. On the contrary, implementation of a SmartBAN with any other IBI values, e.g the shortest IBI and the Optimal IBI without the synchronisation of $m$, will lead to approximately $T_{\text{IBI}}/2$ average uplink latency as described in (3.10). For the 6-sensor SmartBAN, the Optimal IBI length is 100 ms, thus a higher average

### Table 3.2 Simulation parameters.

(a) 6-sensor SmartBAN (Optimal IBI 100 ms, $L = 160$)

<table>
<thead>
<tr>
<th>Sensor Num</th>
<th>Application</th>
<th>Slot Unit Num, $k$</th>
<th>Data Period, $T_i$</th>
<th>Frame Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ECG</td>
<td>4</td>
<td>100 ms</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>EEG</td>
<td>8</td>
<td>100 ms</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>EMG</td>
<td>12</td>
<td>100 ms</td>
<td>Maximum</td>
</tr>
<tr>
<td>4</td>
<td>Sleep monitoring</td>
<td>16</td>
<td>200 ms</td>
<td>Length</td>
</tr>
<tr>
<td>5</td>
<td>Capsule endoscope</td>
<td>20</td>
<td>500 ms</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Sports ECG/pulse</td>
<td>24</td>
<td>1 s</td>
<td></td>
</tr>
</tbody>
</table>

$P_{\text{active,sensor}} = 136$ mW  $P_{\text{sleep,sensor}} = 0.2$ mW  $P_{\text{switch}} = 1$ mW

(b) 8-sensor SmartBAN (Optimal IBI 50 ms, $L = 80$)

<table>
<thead>
<tr>
<th>Sensor Num</th>
<th>Application</th>
<th>Slot Unit Num, $k$</th>
<th>Data Period, $T_i$</th>
<th>Frame Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ECG</td>
<td>4</td>
<td>100 ms</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>EEG</td>
<td>8</td>
<td>100 ms</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>EMG</td>
<td>12</td>
<td>100 ms</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Sleep monitoring</td>
<td>16</td>
<td>200 ms</td>
<td>Maximum</td>
</tr>
<tr>
<td>5</td>
<td>Fall monitoring</td>
<td>20</td>
<td>250 ms</td>
<td>Length</td>
</tr>
<tr>
<td>6</td>
<td>Capsule endoscope</td>
<td>24</td>
<td>500 ms</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Sports Velocity</td>
<td>28</td>
<td>750 ms</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Sports ECG/pulse</td>
<td>32</td>
<td>1 s</td>
<td></td>
</tr>
</tbody>
</table>

$P_{\text{active,sensor}} = 136$ mW  $P_{\text{sleep,sensor}} = 0.2$ mW  $P_{\text{switch}} = 1$ mW
uplink latency of approximately 50 ms is shown if \( m \) is randomly generated compared to the 8-sensor SmartBAN, whose Optimal IBI length is 50 ms.

To evaluate the energy performance, we use the sum of the receiving and transmitting power given in [53]. In a simulation time of 500 seconds, we evaluate energy consumptions of a sleep-capable SmartBAN versus an always-active SmartBAN and compare the energy-savings with the theoretical value obtained from (3.11) and (3.12). The doze power, \( P_{\text{doze, hub}} \) and active power, \( P_{\text{active, hub}} \) for the
hub are assumed to be 100 mW and 200 mW, respectively. Here, we assume that power for the hub in doze mode is half of the active-mode power. Figs. 3.13(a) and (b) depict the simulation and theoretical energy-savings of sensors, $\eta_{\text{sensor},i}$ achieved under the 6- and 8-sensor SmartBANs, respectively. Results highlight that adopting A-IBI increases energy-savings of sensors. Comparing the results in Figs. 3.13(a) and (b), we can observe that the 6-sensor SmartBAN achieves higher energy-savings than the 8-sensor SmartBAN. This is due to the selection of $T_{\text{IBI}}$ by A-IBI. By adopting A-IBI, the Optimal IBI for the 6-sensor SmartBAN is 100 ms,
whereas that for the 8-sensor SmartBANis 50 ms. With A-IBI, the 6-sensor SmartBAN therefore necessitates a longer IAP to achieve an Optimal IBI of 100 ms, thus enabling the sensors to sleep for a longer duration, facilitating more energy savings. Fig. 3.14 provides an overall comparison between the hub’s energy savings, $\eta_{hub}$ and sensors’ average energy-savings, $\eta_{avg}$. The hub in the 6-sensor SmartBAN is able to save more energy due to a longer Optimal IBI. From the results in Figs. 3.13 to 3.14, we can see that by adopting A-IBI, the

**Figure. 3.14** Average energy-savings simulation for both hub and sensors.

**Figure. 3.15** Channel efficiency of both hub and sensors.
SmartBAN is able to achieve lower average uplink latency and higher energy-savings. We further evaluate the percentage of channel utilisation, represented by channel efficiency in (3.13). Simulations were carried out for a maximum of 10 sensors. As plotted in Fig. 3.15, the A-IBI improves uplink latency between the SmartBAN with Optimal IBI and the IEEE 802.15.6 with the same $T_{IBI}$.

(b) Energy consumption of the periodic uplink transmission among the SmartBAN with Optimal IBI, shortest IBI and the IEEE 802.15.6.

**Figure. 3.16** Comparisons with the IEEE 802.15.6 MAC.
channel efficiency since the Optimal IBI turns possible idle TDMA slots under shortest IBI into IAP without influencing uplink latency.

Finally, the average uplink latency and energy consumption comparisons between the SmartBAN and CSMA/CA-based IEEE 802.15.6 MAC are provided in Figs. 3.16(a) and (b), respectively. The beacon interval length of the IEEE 802.15.6 MAC is intentionally set to the Optimal IBI length under the 6-sensor scenario. From Fig. 3.16(a), a SmartBAN with \( m \) synchronised to \( k \) under the Optimal IBI achieves better latency performance. Furthermore, as the SmartBAN adopts scheduled-access mechanism during the SAP, it saves more energy compared with the IEEE 802.15.6 as shown in Fig. 3.16(b). It can be observed in Fig. 3.16(b) that sensors with shorter data periodicity spend more energy. Note that in the comparison with the IEEE 802.15.6 MAC, only the energy consumption of the uplink transmission is considered due to the different access mechanisms used in SmartBAN and IEEE 802.15.6. When considering only uplink transmission, the energy consumption for the SmartBAN with the Optimal IBI and the shortest IBI are the same due to the TDMA access.

Overall, simulation results verify that based on our proposed A-IBI in the SmartBAN, the latency and energy performance for delivering periodic monitoring traffic can be improved compared with the SmartBAN with the shortest IBI duration as well as the CSMA/CA-based IEEE 802.15.6 network.

3.4.3 EM Uplink Latency and Energy Consumption Evaluations

In order to evaluate the EM uplink latency performance, we have considered a SmartBAN with 6 sensors. The packet generation rate at each sensor is set as \( \lambda = 2 \) packets/s and the offered traffic load for the SmartBAN is \( G = \lambda N \). Fig. 3.17 plots the average latency, \( D_{EM} \) in (3.22) as a function of contention access duration, \( T_{CMP} \). Increasing \( T_{CMP} \) provides more ALOHA access opportunity and hence reduces \( D_{EM} \). When \( T_{CMP} = 0 \), only TDMA-based SAP is utilised, resulting in accumulated EM packets in sensors and therefore, increased latency. Based on Figs. 3.17(a) and (b), although a longer \( T_{IBI} \) generally results in increased average latency, by selecting a longer \( T_{CMP} \) within an IBI, the uplink latency for EM packets can be reduced. Simulation results in Fig. 3.17 validate the developed latency model in (3.22).
Chapter 3 SmartBAN Uplink Latency and Energy Performance Study: A Time-Optimised MAC for Periodic Monitoring and Emergency Traffic

Figure. 3.17 EM latency validation.

![Latency Validation](image1.png)

(a) $T_{IBI} = 100$ ms  
(b) $T_{IBI} = 200$ ms

Figure. 3.18 Simulations of SmartBAN EM uplink latency and energy consumption ($\lambda = 2$ packets/s).

![Simulations](image2.png)

(a) Latency ($T_{IBI} = 100$ ms)  
(b) Energy consumption ($T_{IBI} = 100$ ms)

![Latency Simulations](image3.png)

(a) Latency ($T_{IBI} = 200$ ms)  
(b) Energy consumption ($T_{IBI} = 200$ ms)
Now, we further analyse the latency and energy consumption performances of the SmartBAN under varying aggregated traffic loads. Figs. 3.18 and 3.19 compare the latency and energy consumption performances rising from different traffic loads of $\lambda = 2$ packets/s and $\lambda = 20$ packets/s. For packet size of 1600 bits, the aggregated traffic load is normalized as $1600 \text{ bits/packet} \times \lambda N / 1 \text{ Mbps}$. The normalized aggregated traffic loads at $\lambda = 2$ and $\lambda = 20$ are 0.02 and 0.2, respectively.

As shown in Fig. 3.18, increasing $T_{\text{CMP}}$ reduces the uplink latency, while the energy consumption is increased from 0.02 to 0.08 mJ. This is because a longer $T_{\text{CMP}}$ results in a larger probability that an EM packet is generated and sent during CMP. Note that at $\lambda = 2$, the network is lightly-loaded and the simulation results agree with that of the analytical EM latency model. For $\lambda = 20$, Fig. 3.19

**Figure.** 3.19 Simulations of SmartBAN EM uplink latency and energy consumption ($\lambda = 20$ packets/s).
shows that the queuing performance of a sensor depends on $T_{CMP}$. For $T_{CMP} < 60$ ms (Fig. 3.19(a)) and $T_{CMP} < 160$ ms (Fig. 3.19(c)), the uplink latency increases rapidly and severely deviates from our analytical model. On the contrary, when $T_{CMP} > 60$ ms (Fig. 3.19(b)) and $T_{CMP} > 160$ ms (Fig. 3.19(d)), the energy consumption significantly reduces from 15 mJ to less than 2 mJ. This is due to the innate drawback of slotted ALOHA access. Under slotted ALOHA, whenever multiple sensors attempt to transmit packets simultaneously, a collision occurs and all collided packets need to be retransmitted. Therefore, it was noted that when $T_{CMP}$ is not sufficiently long, packets accumulated during $T_{IAP}$ and $T_{SAP}$ undergo severe collisions. When all $N$ sensors have packets waiting in their buffers, a packet is unlikely to be transmitted successfully. This state, where each sensor in a $N$ sensor SmartBAN always has packets waiting in the buffer, is defined as the \textit{blocking state} and a SmartBAN in blocking state is referred to as ‘blocked’. In this case, the energy consumption increases linearly with increasing $T_{CMP}$ due to the constant retransmission (refer to $T_{CMP} < 60$ ms in Fig. 3.19(b) and $T_{CMP} < 160$ ms in Fig. 3.19(d)) and is significantly higher than the energy consumption reported in Figs. 3.18(b) and (d). Conversely, if $T_{CMP}$ is increased such that the sensors are not blocked, both latency and energy consumption are reduced significantly.

3.5 Summary

The SmartBAN has been proposed to address the need for low system complexity and ultra-low power consumption in wireless body area networks. This chapter has provided a critical analysis on how timing parameters, i.e., time durations of SAP, CMP, IAP, and IBI, of the SmartBAN hybrid MAC frame can impact on uplink latency and energy consumption. A time-optimised MAC has been proposed to determine an Optimal IBI and Optimal CMP in SmartBAN hybrid MAC frame to achieve low latency and high energy-savings for transmitting both periodic health monitoring traffic and latency-sensitive EM traffic.

For periodic monitoring packets generated by SmartBAN sensors, we have presented a closed-form uplink latency model and analysed the impact of sensors’ data period, designated TDMA slots, sensors’ initial packet slots and time duration of an IBA on the latency. From our model, uplink latency can be
minimised through the implementation of Optimal IBI (i.e. maximum common divisor of the sensor data periods). An adaptive IBI algorithm is further proposed to facilitate sleep and doze mode operations of sensors and the hub under the proposed Optimal IBI. Simulations of two illustrative SmartBANs with differing number of sensors and Optimal IBI highlight results which are in agreement with our theory. The proposed SmartBAN MAC framework has been shown to outperform latency and energy consumption performance of the IEEE 802.15.6 MAC. For EM packets relying on hybrid TDMA and slotted ALOHA access, we presented an insight into how the durations of different access periods of the SmartBAN MAC influence the uplink latency and energy consumption. Closed-form uplink latency and energy consumption models were derived and verified through simulations. Based on these analytical evaluations, the Optimal CMP can be determined by considering the latency and energy trade-off in practical applications. The impact of EM traffic load on the latency performance is also discussed, whereby a blocking state is reported and should be avoided when deploying SmartBANs.

Based on both theoretical and simulation analyses for periodic monitoring and EM traffic in SmartBANs, our findings conclude that: (a) periodic uplink latency can be minimized by selecting the duration of IBI as a common divisor of sensors’ data generation periods; (b) the time duration of CMP determines the latency and energy consumption of EM traffic, and can be selected based on the developed model considering latency constraint for certain applications; (c) for a SmartBAN with EM traffic, there exists a blocking state where each sensor always has packets waiting in the queue (the CMP needs to be determined with the aggregated traffic such that the SmartBAN is not blocked); and (d) our models are suited to unblocked SmartBANs and the minimum CMP can be solved considering the latency constraint in practical applications.

Overall, the analyses carried out in this chapter facilitated understandings on the uplink latency and energy performances of transmitting periodic monitoring and EM traffic in SmartBANs. In the next chapter, we investigate the latency and energy consumption in the downlink direction, where control and actuation traffic needs to be delivered from the SmartBAN hub to its attached sensor/actuator nodes. A novel downlink transmission framework is proposed in
achieving both low latency and high energy efficiency for SmartBAN downlink transmission.
Chapter 4

SmartBAN Downlink Performance Study: A Novel Downlink Transmission Framework for Reducing Latency and Energy Consumption

4.1 Introduction

In Chapter 3, we have overviewed that a majority of existing studies on wireless body area networks (WBANs) is focused on uplink transmission performance since current WBANs are mainly utilised for monitoring purposes, such as daily health status monitoring. As a result, little attention was paid to the downlink transmission in WBANs in state-of-the-art literature. However, with the advent of remotely-controlled human-to-machine (H2M) applications, future WBANs are expected to support interactive and therapeutic healthcare applications, such as tele-diagnosis and tele-rehabilitation, vision and touch enhancement, and movement assistance [178]. These applications will rely on miniaturised actuators and/or robotic devices worn on the human body to enable control and haptic feedback functionalities. For such applications, stringent low latency needs to be achieved in WBAN downlink, where control/actuation packets are transmitted from the WBAN hub to its attached actuator nodes. Due to the current lack of attention on the WBAN downlink latency and energy performance, expanding the capability of WBANs, including SmartBANs, into the above-mentioned applications requires a critical reassessment and novel medium access control (MAC) layer solutions for downlink transmission. Chapter 3 has comprehensively studied the uplink latency and energy performances of SmartBANs. In this chapter, we investigate the downlink transmission mechanisms, and the downlink latency and energy consumption in SmartBANs.
Note that the study on SmartBAN downlink transmission not only helps improve SmartBAN downlink performance, but would also provide insights into the downlink transmission mechanism design for existing WBAN systems.

In SmartBAN MAC protocol, a supplementary downlink transmission mode (SDM) is stipulated, which aims at transmitting occasionally generated system management packets to sensor nodes [40]. The downlink latency and energy performances of SDM have not been critically evaluated. The capability of the specified SDM in supporting latency-sensitive downlink applications remains unknown. In this regard, a comprehensive evaluation of SmartBAN SDM latency and energy performances is warranted. Importantly, novel downlink transmission mechanisms that can address the stringent low latency for emerging applications need to be explored.

Accurate evaluation of SmartBAN downlink latency and energy performance, however, is challenging due to the SmartBAN predefined MAC frame. As introduced in Chapter 3, in beacon-enabled SmartBANs, sensor/actuator nodes are notified of the MAC timing parameters, e.g., durations of inter-beacon interval (IBI), scheduled access period (SAP), control and management period (CMP), and inactive period (IAP), upon receiving D Beacons. The durations of an IBI and individual access periods are preset, and the values selected determine the latency and energy performances of the SmartBANs. As such, understanding how these timing parameters impact on the downlink performance of SmartBANs is critical to improving downlink transmission mechanism design. Models that can accurately characterise the dependency of downlink latency and energy performances on timing parameters setting in SmartBAN downlink transmission are required. Such analytical modelling for SmartBANs is challenging since existing analytical latency models, e.g., M/M/1, M/D/1 or M/G/1, are developed for conventional transmission mechanisms featured by alternating ‘vacation’ and ‘transmission’ periods [175]. Specifically, in the conventional mechanisms, a downlink transmission duration ends when the downlink buffer of the hub becomes empty and the hub starts a ‘vacation’ with a period irrelevant to the transmission time [175,179]. Radically different from the conventional mechanism, in beacon-enabled SmartBANs, transmission and vacation time are preconfigured in a MAC frame, which determines the queuing and latency
performances in turn. As such, novel models that suit the SmartBAN beacon-enabled MAC are demanded.

In light of the above, this chapter presents the first comprehensive downlink latency and energy performance study for SmartBANs. The downlink latency studied is defined as the queuing time between a downlink packet arriving at the SmartBAN hub to its downlink transmission. Recall that in Chapter 3 Section 3.3.1, we have defined the energy-savings metric as the percentage of the energy saved by a sleep-capable SmartBAN as compared to an always-active SmartBAN. In this chapter, we also evaluate the energy-saving capability in the downlink direction in SmartBANs. In understanding the performance of SDM defined by the SmartBAN MAC protocol, we first model and estimate the latency of SDM and analyse the factors that limit the latency performance of SDM. Then, an improved supplementary downlink mode (ISDM) is proposed to address the limitation in SDM. To further reduce the latency for latency-constraint applications, we propose two types of exhaustive mechanisms in SmartBANs, termed as limited-exhaustive (LEDM) and fully-exhaustive (FEDM) downlink transmission mode. Corresponding energy-saving mechanisms in each mode are discussed, respectively. Further, we develop embedded Markov chains to accurately evaluate the impact of SmartBAN MAC timing parameters on downlink latency and energy-savings of LEDM and FEDM. Based on the understandings on the SDM, ISDM, LEDM and FEDM latency and energy performances, a novel downlink transmission framework is finally proposed for the SmartBAN hub to flexibly decide transmission mode and access durations in achieving low latency and high energy-savings in SmartBAN downlink transmission. Overall, the contributions of the work in this Chapter are threefold: (a) provision of the first comprehensive downlink latency and energy-savings study of SmartBANs; (b) proposal of ISDM, LEDM, and FEDM, and development of models that accurately evaluate the impact of SmartBAN MAC timing parameters on downlink latency and energy-savings performances; and (c) design of a novel downlink transmission framework that can accommodate downlink latency and energy-savings for latency-constraint applications.

The rest of this chapter is organised as follows. In Section 4.2, the SmartBAN MAC protocol, estimation of SDM and the proposed ISDM are introduced.
Conventional exhaustive mechanisms for downlink transmission in a communication network are presented. Based on SmartBAN beacon-enable MAC, LEDM and FEDM, in conjunction with accurate analytical latency models for LEDM and FEDM, are proposed in Section 4.3. In Section 4.4, the energy-savings of SmartBAN downlink transmission is analysed. The proposed downlink transmission framework rising from the insights attained in SmartBAN downlink performance study is illustrated. In Section 4.5, simulations analyses are presented. Further, in Section 4.6, modelling and analyses of SmartBAN round-trip transmission including both uplink and downlink are presented. Finally, conclusions are summarised in Section 4.7.

4.2 Downlink Transmission Mechanisms in SmartBAN

4.2.1 SmartBAN Supplementary Downlink Mode

For downlink transmission, SmartBAN has defined SDM whereby nodes attached to the hub are informed of downlink transmission demands by a downlink list in the D Beacon. Following that, all listed nodes will listen for
downlink data during CMP. An illustration of this scheduling procedure is presented in Fig. 4.1. In SDM, a D Beacon can only notify nodes the number of downlink packets arrived during the previous IBI. Therefore, the latency of a downlink packet is the queuing time until this packet is polled by a D Beacon and all arrivals ahead of this packet are transmitted. We estimate the average downlink latency of SDM, denoted as $D_{SDM}$, using the residual time approach [175]. Let us assume that arriving packets at the hub will be polled by the next D Beacon and transmitted during CMP in the next coming IBI. Note that in practice, buffered packets may need to wait for several IBIs under heavy traffic load. However, as will be shown, using the assumption above and the residual time approach will help us understand the impact of the order of SAP and CMP on the SDM latency performance easily. The estimated $D_{SDM}$ in this case, provides the lower bound of the average SDM latency. Downlink latency models for LEDM and FEDM are developed using Markov chains as will be presented in Section 4.3, thereby are not limited by this assumption.

### 4.2.2 Limitation of SDM Latency and Improved SDM (ISDM)

In our model, the residual time, $R_i$, as seen by the $i$-th arrival refers to the remaining time until the end of current IBI. Let us consider that upon the arrival of the $i$-th packet, this packet will find $n_i$ packets in the downlink buffer, including $n_j$ packets to be transmitted in the current IBI and $n_k$ packets to be transmitted in the next IBI. The latency, $d_{queue,i}$, of the $i$-th packet waiting in the queue can be derived as follows:

$$d_{queue,i} = R_i + n_i T - n_j T + T_{poll} = R_i + n_k T + T_{poll}$$

(4.1)

where $T_{poll} = T + T_{SAP}$. $T_{poll}$ can be viewed as the time to poll a packet before the transmission period, i.e. CMP, begins. Further, since the arrival time is random within an IBI, the average residual time is $E[R_i] = T_{IBI}/2$. Let us denote the average number of $n_k$ as $\bar{n_k}$. Then, $D_{SDM}$ is formulated as follows:

$$D_{SDM} = T_{IBI} / 2 + \bar{n_k} T + T_{poll}$$

(4.2)

Note that $n_k$ is the number of arrivals within the time interval between the arriving time of the $i$-th packet and the beginning of the IBI that the $i$-th arrival
Chapter 4 SmartBAN Downlink Performance Study: A Novel Downlink Transmission Framework for Reducing Latency and Energy Consumption

falls in. We utilise the concept of backward recurrence time (BRT) following [179] to derive \( n_k \). BRT is defined as the elapsed time between an observation time, i.e. the \( i\)-th packet arriving, and a particularly selected time instant, i.e. beginning of the IBI that the \( i\)-th arrival falls in. In this case, BRT is a random variable uniformly distributed within \([0, T_{\text{IBI}}]\). As downlink arrivals follow a Poisson process, \( n_k \) can be computed by averaging all possible number of arrivals during BRT as follows:

\[
\overline{n}_k = \sum_{m=0}^{Q} m \int_{0}^{T_{\text{IBI}}} e^{-G \times \text{BRT}} \frac{(G \times \text{BRT})^m}{m!} \text{d}(\text{BRT}) = \frac{1}{GT_{\text{IBI}}} \sum_{m=0}^{Q} m (1 - \sum_{n=0}^{m} \overline{h}_n) \quad (4.3)
\]

where \( \overline{h}_n = e^{-GT_{\text{IBI}} \frac{(GT_{\text{IBI}})^n}{n!}} \). Substituting (4.3) into (4.2), \( D_{\text{SDM}} \) is derived as:

\[
D_{\text{SDM}} = \frac{T_{\text{IBI}}}{2} + \frac{1}{GT_{\text{IBI}}} \sum_{m=0}^{Q} m (1 - \sum_{n=0}^{m} \overline{h}_n) + T_{\text{poll}} \quad (4.4)
\]

The first two terms on the right-hand side in (4.4) only depends on the duration of an IBI, i.e. \( T_{\text{IBI}} \), whereas the third term, \( T_{\text{poll}} \), is determined by MAC frame configuration, i.e. the order of SAP and CMP in an IBI. To reduce \( T_{\text{poll}} \), we propose to consider changing the order of SAP and CMP to reduce downlink latency. This operation could be achieved using the reserved bits the SmartBAN MAC header. As our focus is on understanding the downlink performance associated with a beacon-enabled MAC, we do not further discuss the specific implementation here. We termed this mechanism as improved supplementary downlink mode, ISDM, and the latency, \( D_{\text{ISDM}} \), is given as:

\[
D_{\text{ISDM}} = \frac{T_{\text{IBI}}}{2} + \frac{1}{GT_{\text{IBI}}} \sum_{m=0}^{Q} m (1 - \sum_{n=0}^{m} \overline{h}_n) + T \quad (4.5)
\]

Note that ISDM can reduce the latency by an amount of \( T_{\text{SAP}} = NT \) compared to SDM. Moreover, since uplink transmission in SAP adopts TDMA by default, uplink latency performance will not be affected by this adjustment (referring to the models in [180]). Through the above analysis, a more general conclusion for scheduling downlink can be made: when downlink transmission is polled by beacons, downlink transmissions can begin immediately following the beacon for reducing latency.
4.2.3 Conventional Downlink Transmission Mechanisms: Challenges in SmartBAN

Both SDM and ISDM incur high latency since packets need to be polled before transmission. As shown in (4.4) and (4.5), $T_{IBI}/2$ is the average waiting time for a packet to be polled by a D Beacon. As such, SDM and ISDM may not be favourable choices for latency-sensitive applications. Specifically, compared to a polling-based transmission mechanism such as SDM and ISDM, the exhaustive transmission has been extensively studied and proven to be a latency-efficient method in a system with a single queue and one server, e.g., SmartBAN downlink broadcasting. Generally speaking, by exhaustive transmission, downlink packets can be broadcasted or unicasted without needing to notify nodes via polling message, e.g., channel beacons, thereby reducing packets’ waiting time to be polled by the hub. Therefore, we further investigate exhaustive transmission in supporting SmartBANs downlink, rather than using polling-based SDM and ISDM. The conventional exhaustive transmission mechanisms and the challenges of applying such mechanisms in SmartBANs are illustrated as follows.

Conventional exhaustive transmission mechanisms with a vacation can be described as that the server starts a busy period transmitting packets when a vacation ends and then goes on another vacation period whenever the packet buffer becomes empty. In this chapter, we refer these two types of periods as transmission period and vacation period. Such mechanisms have been widely adopted for downlink transmission in current communication networks. For example, the exhaustive transmission is considered from wireless access points to machine-type devices in machine-to-machine networks [181]. A basic $M/G/1$ queuing model in exhaustive transmission has been studied in [182]. A closed-form model of 1-limited transmission, which only allows 1 packet to be sent in the transmission period of a node, was developed. Further, in [175], a $k$-limited exhaustive transmission was introduced where a vacation begins either when the maximum $k$ number of packets are transmitted or when the packet buffer is empty. It should be noted that conventional exhaustive mechanisms lead to a dynamic downlink transmission period. For example, when applied to the SmartBAN, the duration of CMP, $T_{CMP}$, and hence the duration of an IBI, would be highly dependent on the packet arrivals and traffic load. As such,
downlink transmission in SmartBANs by conventional exhaustive transmission will result in a dynamic $T_{CMP}$ and $T_{IBI}$. This dynamic characteristic directly limits the energy-saving capability of sleep mode sensors/actuators, as will be shown in Section 4.4.

As introduced earlier in Chapter 3, SmartBAN beacon-enabled MAC is featured by the use of preset and fixed durations of MAC parameters, i.e., $T_{IBI}$, $T_{SAP}$, $T_{CMP}$ and $T_{IAP}$, during SmartBAN operation. One of the important benefits of this fixed MAC configuration is that sensors can be notified of inactive periods (no uplink/downlink transmissions are expected) through D Beacon and therefore, can turn off their transceivers (sleep mode) to save energy. As such, exploiting CMP for SmartBAN downlink transmission, we proposed novel exhaustive downlink transmission mechanisms where downlink channel access is governed by periodically broadcast D Beacons and fixed $T_{SAP}$ and $T_{CMP}$ in an IBI. Two novel downlink modes: limited-exhaustive (LEDM) and fully-exhaustive (FEDM) downlink mode, are defined and described in the following section.

### 4.3 Proposed Downlink Transmission Modes for SmartBANs

#### 4.3.1 Limited-exhaustive and Fully-exhaustive Downlink Mode

To address the challenges of applying conventional exhaustive transmission mechanisms for SmartBAN downlink transmission, the implementation of LEDM and FEDM are described as follow.

- **Limited-exhaustive downlink transmission, LEDM:** In an IBI, the SmartBAN hub broadcasts downlink packets to nodes during CMP. Downlink transmission ends when the hub finds the downlink packet buffer empty or when the CMP ends. In LEDM, once the downlink buffer becomes empty during CMP, the remaining CMP duration becomes part of a vacation period.

- **Fully-exhaustive downlink transmission, FEDM:** In FEDM, downlink transmission ends only when CMP ends in an IBI. That is the hub does not start a vacation period immediately, rather, it will continue to wait and
transmit any new arrivals until CMP ends. Therefore, in FEDM, CMP is the transmission period, whilst, D Beacon, SAP and IAP form the vacation. Since the entire CMP duration is fully utilised for transmission, the latency can be further reduced in FEDM as compared to LEDM.

### 4.3.2 Embedded Markov Chains for LEDM and FEDM Latency Performance

To accurately model downlink latency performance of the proposed LEDM and FEDM, we develop two types of embedded Markov chains by examining particular time instants in each IBI as shown in Fig. 4.1. For LEDM, we focus on the downlink queue length at the hub at the time instants of a downlink vacation completion as well as the termination of each CMP slot used for packet transmission. A transmission period ends when the downlink buffer becomes empty during CMP. This time instant is termed as *First Idle* as shown in Fig. 4.1. For FEDM, we select observation time instants at a downlink vacation period completion and each CMP time slot termination no matter it is idle or used for sending a packet.

Further, we denote \( \{L, \xi\} \) as the state that the downlink packet queue length at the hub is \( L \) at a specific time instant, \( \xi \). We have \( L = 0, 1, 2, \ldots Q \), where \( Q \) is the maximum number of downlink packets that can be buffered at the hub, and \( \xi = 0, 1, 2, \ldots M \), where \( M \) is the total number of time slots in CMP. The value \( \xi = 0 \) represents the time instant of a vacation period completion and \( \xi \neq 0 \) corresponds to the \( \xi \)-th CMP time slot termination. Then, we use \( p_{k,m} = \Pr\{L=k, \xi=m\} \) to represent the steady-state probability of the two Markov chains. Under steady state, \( p_{k,m} \) represents the proportion of time the system, described by the embedded Markov chain, spends at the embedded time instant \( m \) with queue length \( k \). The state transition following the definition of LEDM and FEDM are illustrated in Fig. 4.2 and Fig. 4.3, respectively. Explanations are provided as follows.

**Analytical Modelling of LEDM**

As shown in Fig. 4.2, let us denote the downlink vacation of LEDM as \( V_{d,m} \). Then, \( V_{d,m} = T_{IBI} - mT \) when First Idle occurs at the \( m \)-th CMP. Further, we denote \( g_i \) and \( h_{i,m} \) as the probability that \( i \) number of downlink packets arrive at
Chapter 4 SmartBAN Downlink Performance Study: A Novel Downlink Transmission Framework for Reducing Latency and Energy Consumption

The hub during a CMP time slot time, $T$, and the downlink vacation time, $V_{d,m}$, respectively. Then, we can derive $g_i = e^{-GT} \left(\frac{(GT)^i}{i!}\right)$ and $h_{t,m} = e^{-GV_{d,m}} \left(\frac{(GV_{d,m})^t}{t!}\right)$ considering Poisson arrival.

The state transition of the embedded Markov chain depends on the number of arrivals during the time interval between two consecutive embedded time instants. A state $\{k, m\} (k \neq 0, m \neq M)$, i.e. there are $k$ packets in the buffer at the time of the $m$-th CMP time slot termination, will transit to the state $\{k + i - 1, m + 1\}$ if $i$ ($i = 0, 1, 2, \ldots$) number of packets arrive during the $(m + 1)$-th CMP time slot. Hence, the state transition probability, in this case, is $g_i$. When $k = 0$ (First Idle) or $m = M$, a downlink vacation period begins. Hence, the next state to be observed will be $\{k + i, 0\}$ with $i$ arrivals during the vacation period. In this case, the state transition probability from $\{0, m\}$ or $\{k, M\}$ to $\{k + i, 0\}$ is $h_{i,m}$ with $V_{d,m} = T_{\text{IL}} - mT$. Under steady state, the state transition of the embedded Markov chain can be represented in (4.6), in which, $g_i^t = \sum_{k=i} g_k$ and $h_{t,m}^f = \sum_{j=1} h_{j,m}$. Then, with a normalised equation $\sum_{m=0}^M \sum_{k=0}^\infty p_{k,m} = 1$, the steady-state probability, $p_{k,m}$, of each state can be solved uniquely. Note that $p_{k,m} = \Pr\{L=k, \xi=m\}$ gives us the queue length distribution at each embedded time instant. With $p_{k,m}$, we are interested in the queue length distribution, $\Pr\{L=k\}$ ($k =

\textbf{Figure 4.2} Embedded Markov model and the state transition under LEDM.
0,1,2...Q, at any arbitrary observation time. This arbitrary observation time could fall into any one of the time intervals between two consecutive embedded time instants.

\[
p_{k,m} = \begin{cases} 
\sum_{i=1}^{k+1} g_{k+1-i} p_{i,m-1} & m = 1,2,...M, k = 0,1,2...Q-2 \\
\sum_{i=1}^{k} g_{k+1-i} p_{i,m-1} & m = 2,...M, k = Q-1 \\
\sum_{j=0}^{M-1} h_{k,j} p_{0,j} + \sum_{j=0}^{k} h_{k-i,m} p_{i,M} & m = 0, k = 0,1,2...Q-1 \\
\sum_{j=1}^{M-1} h_{Q,j} p_{0,j} + \sum_{j=1}^{Q-1} h_{Q-i,M} p_{i,M} & m = 0, k = Q 
\end{cases} \
\]

(4.6)

\[
\Pr\{A = j \mid \chi(T)\} = \int_0^T \exp(-Gt) \frac{(Gt)^j}{j!} \frac{1}{T} dt = \frac{1}{GT} \left(1 - \sum_{i=0}^{j} g_i\right) 
\]

(4.7)

\[
\Pr\{A = j \mid \chi(V_{d,m})\} = \int_0^{V_{d,m}} \exp(-Gt) \frac{(Gt)^j}{j!} \frac{1}{V_{d,m}} dt = \frac{1}{GV_{d,m}} \left(1 - \sum_{i=0}^{j} h_{i,m}\right) 
\]

(4.8)

Therefore, to derive \(\Pr\{L=k\}\), we define \(\chi(t)\) to represent the event that our observation falls into a certain time period, \(t\). \(\chi(V_{d,m})\) and \(\chi(T)\) hence indicate that an observation falls into a downlink vacation period, \(V_{d,m}\), or one of the CMP time slots with duration \(T\), respectively. Recall that the BRT introduced in Section 4.2 is the elapsed time between an observation time and its last embedded time instant. If the observation time falls into any one of the CMP time slots, BRT is a random variable uniformly distributed within \([0,T]\). If our observation time falls within a downlink vacation period, BRT obeys uniform distribution within \([0,V_{d,m}]\). Denoting \(A\) as the number of arrivals during a BRT, the probability of \(j\) arrivals during the BRT of an arbitrary observation in \(T\) and in \(V_{d,m}\), can be derived as (4.7) and (4.8), respectively.

Further, the probability that a particular embedded time instant, \(\xi = m\), is visited, is derived as \(\Pr\{\xi=m\} = \sum_{k=0}^{Q-1} p_{k,m}, (m = 1,2,...,M)\) and \(\Pr\{\xi=0\} = \sum_{k=0}^{Q} p_{k,0}\). In
each IBI, the completion of a vacation period, i.e. state with $\xi = 0$, will always be visited once, therefore, the probability that the rest embedded time instants will be visited in an IBI is derived as $p_{k,m}/\Pr\{\xi = 0\}$. Then, utilising (4.7) and (4.8), arbitrarily-observed queue length distribution, $\Pr\{L=k\}$, can be derived as follows:

$$
\Pr\{L = k\} = \begin{cases} 
\sum_{m=0}^{M} \frac{P_{0,m}}{\Pr\{\xi = 0\}} \Pr\{A = 0 \mid \chi(V_{d,m})\} \frac{T_{IBI} - mT}{T_{IBI}}, & k = 0 \\
\sum_{m=0}^{M} \frac{P_{0,m}}{\Pr\{\xi = 0\}} \Pr\{A = k \mid \chi(V_{d,m})\} \frac{T_{IBI} - mT}{T_{IBI}}, & k = 1, 2, \ldots, Q \\
+ \sum_{m=1}^{M-1} \sum_{l=1}^{k} \frac{P_{l,m-1}}{\Pr\{\xi = 0\}} \Pr\{A = k - i \mid \chi(T)\} \frac{T}{T_{IBI}} 
\end{cases} 
$$

(4.9)

in which $\frac{T_{IBI} - mT}{T_{IBI}}$ and $\frac{T}{T_{IBI}}$ is the probability that an observation falls into a vacation and one CMP slot in transmission, respectively. With (4.9), the average downlink queue length, $N_{LEDM}$, is then derived as:

$$
N_{LEDM} = \sum_{k=0}^{Q} k \Pr\{L = k\} 
$$

(4.10)

From (4.9), $\Pr\{L=Q\}$ is the blocking probability, $P_b$, at which there is no space in the downlink buffer and a new-arrival packet has to be dropped by the hub. According to Little’s theorem, the average downlink latency of LEDM, $D_{LEDM}$, can be derived as follows:

$$
N_{LEDM} = \sum_{k=0}^{Q} k \Pr\{L = k\} 
$$

(4.11)

**Analytical Modelling of FEDM**

In FEDM, the duration of a downlink vacation period, denoted as $V_d$, is fixed as $V_d = T + T_{SAP} + T_{IAP}$. Let us denote $h_i$ as the probability that $i$ number of downlink packets arrive at the hub during $V_d$, then $h_i = e^{-\alpha V_d (\frac{(G_d)}{i})}$. Similar to LEDM, the state transition probability from any $\{k, m\} (m \neq M)$ to $\{k + i - 1, m + 1\}$ is also equal to $g$. It should be noted the downlink vacation for FEDM only starts at states $\{k, M\}$. The state transition probability from $\{k, M\}$ to $\{k + i, 0\}$ hence is $h_i$. The state transition can be formulated in (4.12), in which, $h_i' = \sum_{j=1}^{M} h_j$. 

$$
\text{Chapter 4 SmartBAN Downlink Performance Study: A Novel Downlink Transmission Framework for Reducing Latency and Energy Consumption} 
$$

112
4.3 Proposed Downlink Transmission Modes for SmartBANs

Again, with a normalised equation \( \sum_{m=0}^{M} \sum_{k=0}^{Q} p_{k,m} = 1 \), the steady-state probability, \( p_{k,m} \), of each state can be solved. For FEDM, we look into the time instant of each CMP slot termination, therefore, the probability of selecting a particular embedded time instant is given by \( \Pr\{\xi = m\} = 1/(M+1) \). As such, the conditional queue length distribution at a given embedded time instant, \( \Pr\{L=k \mid \xi = m\} \), can be derived in (4.13).

\[
p_{k,m} = \begin{cases} 
g_k P_{0,m-1} + \sum_{i=1}^{k+1} g_{k+1-i} P_{i,m-1}, & m = 1,2,...M, \; k = 0,1,2...Q-2 \\
g_{Q-1} P_{0,m-1} + \sum_{i=1}^{Q-1} g_{Q-1-i} P_{i,m-1} + p_{Q,m-1}, & m = 1,2,...M, \; k = Q-1 \\
g_0 P_{0,m-1}, & m = 1,2,...M, \; k = Q \\
\sum_{i=0}^{k} h_{k-i} P_{i,M}, & m = 0, k = 0,1,2...Q-1 \\
p_{Q,M} + \sum_{i=1}^{k} h_{k-i} P_{i,M}, & m = 0, k = Q 
\end{cases}
\]

(4.12)

\[
\Pr\{L=k \mid \xi = m\} = \frac{p_{k,m}}{\Pr\{\xi = m\}} = (M+1)p_{k,m}
\]

(4.13)
To derive $\Pr\{L=k\}$ ($k = 0, 1, 2...Q$), at any arbitrary observation time, we again utilise BRT and $\chi(t)$ that have been introduced. $\Pr\{L=k\}$ ($k=0,1,2...Q$), dependent on the number of arrivals during a BRT as well as the queue length distribution of the corresponding embedded time instant, can be formulated in (4.14), in which

$$\Pr\{A = j | \chi(V_d)\} = \frac{1}{\sigma_d^V} (1 - \sum_{i=0}^{j} h_i).$$

With (4.12), the average downlink queue length can be derived as $N_{FEDM} = \sum_{k=0}^{Q} k \Pr\{L = k\}$. The average downlink latency, $D_{FEDM}$, then is given in (4.15).

$$D_{FEDM} = \frac{N_{FEDM} - T}{G(1 - P_b)}$$

We briefly discuss the existence of the steady state for the above two embedded Markov chain models. Let us denote $\eta = GT$ as the channel utilisation by downlink transmission. When the buffer capacity, $Q$, is large, e.g, 500 packets, the queuing performance is asymptotical to the infinite buffer case. Hence, it is should be noted that when $\eta \geq T_{CMP}/T_{IBI}$, there is no steady-state solution for the embedded Markov chain and latency performance significantly deteriorates. On the contrary, with $\eta < T_{CMP}/T_{IBI}$ and large $Q$, the SmartBAN will operate under steady state with $\Pr\{P_b = 0\} = 1$. Thus, downlink latency can be estimated by $D_{LEDM} \approx N_{LEDM}/G - T$ and $D_{FEDM} \approx N_{FEDM}/G - T$.

### 4.4 Proposed Downlink Transmission Framework for
4.4 Proposed Downlink Transmission Framework for SmartBAN

SmartBAN

4.4.1 Energy-saving Mechanisms for SDM, ISDM, LEDM and FEDM

Energy-saving Mechanisms for SDM, ISDM, LEDM and FEDM

In WBANs, including SmartBANs, energy-saving capability is achieved through sleep capability whereby nodes are capable of transitioning between active and sleep modes to save energy. In this section, we discuss the mechanism for sleep/active mode transition in downlink transmission and develop models for energy-savings that could be achieved in LEDM and FEDM. For sleep-capable SmartBANs, the receiver of a node remains active when receiving downlink packets. The power consumption of a transmitter and a receiver is denoted as $P_{TRX}$ and $P_{REC}$, respectively. When no transmission is expected, both transmitter and receiver are powered down with a baseline power denoted as $P_{SLP}$.

For SDM and ISDM, nodes are notified of the downlink demand through D Beacons. Therefore, nodes could transition into sleep mode after polled packets are received during CMP. While, for FEDM, nodes stay in active mode during the whole CMP since even the downlink buffer becomes empty, nodes have to wait for any new arrivals until CMP ends. Therefore, FEDM is more energy consuming than SDM and ISDM. The energy-savings of a node in FEDM, denoted as $E_{FEDM}$, is represented as follows:

$$E_{FEDM} = \left(1 - \frac{(T + T_{CMP})P_{REC} + (T_{SAP} + T_{IAP})P_{SLP}}{T_{IBI}(P_{TRX} + P_{REC})}\right) \times 100\% \tag{4.16}$$

in which $(T + T_{CMP})P_{REC}$ is the energy spend in receiving D Beacon and downlink packets.

For LEDM, the First Idle will be dynamic in each IBI and nodes by no means will know if a transmission period ends without informed by the hub. Thus, for nodes to transition to sleep mode when a downlink vacation starts, two types of mechanisms can be considered. Firstly, a hub can broadcast a SLEEP message following the First Idle. Upon receiving this message, nodes transition into sleep mode and wake up before next D Beacon. If First Idle does not occur during CMP, no SLEEP message is needed and nodes transition to sleep mode automatically when IAP starts. Secondly, a hub can utilise 1 reserved bit in the
MAC header [40] to indicate that current packet is the last packet in the downlink buffer. After receiving this packet, nodes shall transit to sleep mode to save energy. In this paper, we adopt the second mechanism since it does not incur extra overhead on DCH.

To evaluate the energy-savings of LEDM, let us use $T_{\text{IDLE}}$ to represent the average duration of the CMP that is in idle. Revisiting the embedded Markov Chain model for LEDM, the First Idle refers to the state $\{0, \xi\}$ and the rest CMP slots following the First Idle become part of a vacation. Therefore, $T_{\text{IDLE}}$ can be derived as follows:

$$T_{\text{IDLE}} = \frac{\sum_{m=0}^{M} p_{0,m} (M - m)T}{\sum_{k=0}^{Q} p_{k,0}}$$

(4.17)

With (4.17), the average energy-savings of LEDM, denoted as $E_{\text{LEDM}}$, can be formulated in (4.18). Moreover, it should be noted that the average duration of CMP in idle of SDM, ISDM and LEDM are the same since channel utilisation for these three mechanisms only depends on the traffic load. Therefore, the average energy-savings of SDM, ISDM and LEDM are the same as given in (4.18).

$$E_{\text{LEDM}} = \left(1 - \frac{(T + T_{\text{CMP}} - T_{\text{IDLE}}) P_{\text{REC}} + (T_{\text{IDLE}} + T_{\text{SAP}} + T_{\text{IAP}}) P_{\text{SLP}}}{T_{\text{IBI}} (P_{\text{TRX}} + P_{\text{REC}})} \right) \times 100\%$$

(4.18)

### 4.4.2 Numerical Analysis of Latency and Energy-savings

As can be observed from the above derivations, the latency and energy-savings of LEDM and FEDM are closely tied to the SmartBAN MAC timing parameters, i.e $T$, $T_{\text{IBI}}$, $T_{\text{SAP}}$, $T_{\text{CMP}}$ and $T_{\text{IAP}}$. In Fig. 4.4 and Fig. 4.5, we present a numerical view of how the selection of these timing parameter values can impact on latency and energy-savings. The latency, $D_{\text{FEDM}}$ and $D_{\text{LEDM}}$, and energy-savings, $E_{\text{FEDM}}$ and $E_{\text{LEDM}}$, in Fig. 4.4 and Fig. 4.5 are numerically obtained by using (4.11), (4.15), (4.16) and (4.18), respectively.

Note that $T_{\text{IBI}} = T + T_{\text{SAP}} + T_{\text{CMP}} + T_{\text{IAP}}$, in which $T_{\text{SAP}} = NT$ and $T_{\text{CMP}} = MT$. Since CMP is utilised for downlink, the pair of $T_{\text{CMP}}$ and $T_{\text{IBI}}$ is the focus in this paper. $T_{\text{SAP}}$ is considered as fixed by a preset $N$. For illustrative purposes in our numerical analysis, we consider $T_{\text{SAP}} = 0$ since under a fixed $T_{\text{IBI}}$, the
maximum downlink channel utilisation is determined by $T_{\text{CMP}}/T_{\text{IBI}}$. Further, we consider a total packet length of 64 bytes, which therefore requires $T = 0.625$ ms. To present the latency and energy performance variation with varying traffic loads, the aggregated downlink packets arriving rate, $G$, is set as 0.1 packets/ms and 0.8 packets/ms, corresponding to channel utilisation $\eta = 0.0625$ and $\eta = 0.5$, respectively. Note that plotted by MATLAB in a unit of 1 ms, the values in the black area as shown in Fig. 4.4 and Fig. 4.5 are not applicable since $T_{\text{CMP}} + T < T_{\text{IBI}}$ and $\eta = GT < T_{\text{CMP}}/T_{\text{IBI}}$ are required for steady-state solutions. For evaluating the energy performance, we use $P_{\text{REC}} = 116$ mW, $P_{\text{TX}} = 20$ mW and $P_{\text{SLP}} = 0.2$ mW following [53]. In Fig. 4.4, we evaluate $T_{\text{IBI}}$ from 3 ms to 13 ms, while in Fig. 3, a longer $T_{\text{IBI}}$ and $T_{\text{CMP}}$ is considered due to the larger $G$ selected.

From Figs. 4.4(a) and 4.5(a), it can be seen that in FEDM, increasing $T_{\text{CMP}}$ effectively reduces $D_{\text{FEDM}}$. Even when $G = 0.8$ in Fig. 3(a), $D_{\text{FEDM}}$ can be reduced to below 1 ms. However, a reduction in latency is at the cost of energy.
consumption. $E_{\text{FEDM}}$ decreases significantly with longer $T_{\text{CMP}}$ as shown in Figs. 4.4(c) and 4.5(c). In contrast, LEDM performance is less impacted by $T_{\text{CMP}}$ values, instead, $D_{\text{LEDM}}$ (in Figs. 4.4(b) and 4.5(b)) and $E_{\text{LEDM}}$ (in Figs. 4.4(d) and 4.5(d)) are highly dependent on $G$. This is because the time duration in CMP utilised for transmission depends on the traffic load. In summary, FEDM is advantageous in reducing downlink latency, while, LEDM significantly saves energy for SmartBANs. Therefore, the hub needs to properly determine its transmission mechanisms considering the latency and energy-saving requirements in practical applications. As such, we propose a novel downlink transmission framework as will be described in the following sub-section.

4.4.3 A Novel Downlink Transmission Framework

For a SmartBAN in which a downlink latency constraint, $D_{\text{ST}}$, needs to be
guaranteed for certain services, it is expected that energy-savings can be maximised with $D_{ST}$ satisfied. In general, for a relaxing $D_{ST}$ that LEDM can achieve, there is no necessity to adopt FEDM due to the high energy cost. For applications relying on stringent $D_{ST}$, FEDM needs to be considered in assuring $D_{ST}$. As such, assuming that $G$ is known and $D_{ST}$ is within the capability of the proposed LEDM or FEDM, we design a downlink transmission framework that includes three parts: (a) decision making; (b) parameter selection; and (c) sleep indication. The details of each part are specified below.

In order to determine the transmission mechanism (decision making), the hub first needs to compute the minimum $D_{LEDM}$, denoted as $D_{LEDM, MIN}$, that can be achieved by LEDM. To solve this $D_{LEDM, MIN}$ via the derived model in (4.15), a $T_{IBI, MAX}$ value is set to limit the range of $T_{IBI}$ within $[T_{IBI, MIN}, T_{IBI, MAX}]$ in which $T_{IBI, MIN} = 2T + T_{SAP}$ since at least one slot shall be assigned in CMP for downlink transmission. Then, if $D_{LEDM, MIN} < D_{ST}$, LEDM is adopted, otherwise, FEDM is implemented by the hub. Once the transmission mechanism is selected, suitable timing parameters will be set by the hub (parameter selection). The pair of $T_{CMP}$ and $T_{IBI}$ is determined by maximising energy-savings under $D_{ST}$. To do so, a dichotomy method to search and return suitable $T_{IBI}$ and $T_{CMP}$ values via solving the developed Markov chains are employed. Finally, with MAC timing parameters determined, the hub utilises the 1 reserved bit during SmartBAN operation (see Section 4.4.1 for energy-savings in LEDM). Since the information of $T_{IBI}$, $T_{CMP}$ and $T_{IAP}$ are encoded in D Beacons, the hub only needs to inform nodes to transition switch to sleep mode at an appropriate time (sleep indication). If LEDM is adopted, the hub will enable the reserved 1 bit, i.e. set to 1, when the packet to be sent is the last packet in the buffer. Otherwise, this bit is set to 0. It should be noted that in FEDM, this reserved bit always remains as 0. Therefore, it is not necessary for the nodes to know which transmission mechanism, i.e. LEDM or FEDM, the hub has adopted. The nodes only need to read the designated bit in the MAC header. If 1 is detected, the nodes are able to sleep until the arrival of the next D Beacon. Otherwise, the nodes will stay in active mode until CMP ends. With the above-mentioned framework, there is no extra computation requirement on the nodes. The resource-rich hub [183] is responsible for decision making and the control of sleep time.
4.5 Performance Evaluations

4.5.1 Model Validations and Performance Comparisons

To verify our embedded Markov chain model, we implement packet-driven simulations using MATLAB with MAC parameters set in accordance to the SmartBAN MAC specification. Considering the reasonable number of nodes (typically less than 10) deployed in wireless BANs [177], we set TDMA time slot number in SAP as \( N = 5 \) in our model validation and \( N = 0, 2, 4 \) in the following simulation performance discussions. To validate our model, the chosen values of MAC parameters are as follows: data rate = 1 Mbps; packet length = 64 bytes; \( T = 0.625 \) ms; \( T_{\text{CMP}} = 6.25 \) ms \((M = 10)\); \( T_{\text{SAP}} + T = 3.75 \) ms \((N = 5)\); and \( T_{\text{IAP}} = 0 \) ms, yielding \( T_{\text{IBI}} = 10 \) ms. Under the selected MAC parameter setting, channel

\[\text{Figure. 4.6} \text{ Steady-state probabilities of embedded states.}\]
utilisation, $\eta$, should satisfy $\eta < 0.625$, to guarantee steady state as analysed in Section 4.3. Therefore, in our simulations, we consider a packet arrival rate, $G$, to be within the range of $0.1 \sim 0.9$ packets/ms, corresponding to $\eta$ varying from $0.0625 \sim 0.56$. The simulation time is 500 s and was repeated 5 times.

The results in Fig. 4.6 present an understanding of the states in the developed two embedded Markov chains. The results also verify the steady-state probability of our theoretical results in (4.6) and (4.12). Figs. 4.6(a) and (b) show the states of the embedded Markov chain for LEDM under $G = 0.2$ and $G = 0.5$, corresponding to $\eta = 0.1$ and $\eta = 0.3$, respectively. Figs. 4.6(c) and (d) plot that of FEDM. In Figs. 4.6(a) and (b), the probability of visiting a state decreases with increasing $m$, while the opposite trend is shown in Figs. 4.6(c) and (d). This observation is in agreement with our analysis whereby for FEDM Markov chain, each embedded time instant is visited equally, while when the First Idle occurs in LEDM model, states with embedded time instant in the rest CMP are not visited. Moreover, the queuing performance at each embedded state varies in Fig. 4.6. A shorter queue length is presented at a larger $m$ in both LEDM and FEDM. This is due to the consecutive downlink packet broadcasting during CMP. Note that the queue length distribution in FEDM as shown in Figs. 4.6(c) and (d) becomes identical with increasing $m$. This is because when the downlink buffer is empty, the queue length, $k$, at an embedded state in FEDM only depends on the packet arrivals in the corresponding transmission time slot. Further, Fig. 4.7 validates the queue length distribution derived in (4.9) and (4.14) and shows that the average queue

\begin{figure}[h]
\centering
\begin{subfigure}{0.49\textwidth}
\centering
\includegraphics[width=\textwidth]{figure4.6a.png}
\caption{$G = 0.2$}
\end{subfigure}
\hfill
\begin{subfigure}{0.49\textwidth}
\centering
\includegraphics[width=\textwidth]{figure4.6b.png}
\caption{$G = 0.5$}
\end{subfigure}
\caption{Queue length distribution of LEDM and FEDM.}
\end{figure}
Chapter 4 SmartBAN Downlink Performance Study: A Novel Downlink Transmission Framework for Reducing Latency and Energy Consumption

length is longer in LEDM as compared to that of FEDM.

Figs. 4.8(a) and (b) compare the average downlink latency, $D_{SDM}$, $D_{ISDM}$, $D_{LEDM}$, and $D_{FEDM}$ under varying $G$. Theoretical latency values are solved by (4.11) and (4.15). Fig. 4.8(a) follows the previous MAC parameter setting, in which $T_{SAP} = 3.125$ ms ($N = 5$). In Fig. 4.8(b), we adjust the duration of CMP to $T_{CMP} = 8.125$ ms ($M = 13$) under $T_{IBI} = 10$ ms, which gives $T_{SAP} = 1.25$ ms ($N = 2$). As expected, in Figs. 4.8(a) and (b), ISDM reduces the latency by 3.125 ms and 1.25 ms compared to SDM, respectively. This validates the conclusion derived in Section 4.2. Moreover, FEDM exhibits the minimum downlink latency in the studied four mechanisms. Note that in Figs. 4.8(a) and (b), when $G$ increases, the latency of LEDM slightly decreases. This is due to the SmartBAN predefined MAC. In compliance with our previous analysis, since $T_{CMP}$ and $T_{IBI}$ are

Figure. 4.8 Average downlink delay and energy savings ($T_{IBI} = 10$ ms).
preselected as fixed, the utilisation, i.e. time duration utilised for downlink data transmission in CMP, increases with increasing $G$, leading to shorter vacation periods. Unlike LEDM, the downlink vacations in FEDM, SDM and ISDM are independent from $G$. As such, the latency increases with increasing $G$. Further, Figs. 4.8(c) and (d) compare the energy-savings performance under $P_{REC} = 116$ mW, $P_{TRX} = 20$ mW and $P_{SLP} = 0.2$ mW. Theoretical values of $E_{FEDM}$ and $E_{LEDM}$ in (4.16) and (4.18) are validated. As shown, FEDM is most energy consuming and $E_{FEDM}$ only depends on the predefined MAC frame. In contrast, energy-savings of SDM, ISDM and LEDM are the same and highly depend on $G$. This is in compliance with our previous analysis in Section 4.4. As such, the proposed LEDM exhibits better performance than the originally specified SDM since LEDM can effectively reduce latency without extra energy cost. FEDM can further reduce latency, but with high energy cost.

4.5.2 Performance of the proposed Downlink Transmission Framework

With the performance of LEDM, FEDM, SDM and ISDM understood, we implement the proposed downlink transmission framework at the hub. We consider two latency constraints in our simulations: $D_{ST} = 1$ ms for haptic feedback and $D_{ST} = 10$ ms for telesurgery [176]. With the proposed framework, the hub determines which mechanism to adopt, LEDM or FEDM, and set the pair of $T_{IBI}$ and $T_{CMP}$ that achieves the most energy-savings under $D_{ST}$ constraint. Two different $T_{SAP} = NT$ values are considered in simulation: $N = 0$ and $N = 4$. Further, $T_{IBI,MAX}$ is set as 100 ms. Fig. 4.9 shows latency and energy-savings of the proposed framework compared to individual LEDM, FEDM and SDM under the same $T_{IBI}$ and $T_{CMP}$ settings. As shown in Figs.4.9(a) and (c), the hub adopts FEDM and $D_{ST} = 1$ ms is assured. However, meeting this stringent latency is at the expense of a higher energy consumption as plotted in Figs. 4.9(b) and (d). For $D_{ST} = 10$ ms, it can be seen from Figs. 4.9 (e) and (f) that the hub implements LEDM and latency is guaranteed below 10 ms. Meanwhile, 50% up to 90% energy-savings can be achieved when varying $G$. Overall, the simulation results validate our theoretical models and show the effectiveness of the proposed framework. Both theoretical and simulation results show that LEDM can reduce the latency by around 5 ms, which can satisfy the latency demand of most existing
Figure 4.9 Performance of the proposed downlink transmission framework.

applications [183,184]. Thus, we highlight LEDM as a favourable solution as
compared to SDM for downlink data transmission in SmartBANs since it can effectively reduce latency, whilst achieve high energy-savings.

4.6 A Discussion on SmartBAN Round-trip Latency

4.6.1 Round-trip Latency Modelling

So far, we have comprehensively analysed and evaluated the latency and energy-savings performances of uplink-dominated SmartBANs for health monitoring and supervisions (Chapter 3) and downlink-dominated SmartBANs in anticipation of remotely-controlled therapeutic applications. In this section, we provide a discussion on the round-trip transmission in SmartBANs that comprise both sensors and actuators. In achieving ubiquitous healthcare, real-time EM reports from patients and timely feedback from remote clinicians need to be delivered by SmartBANs. In such scenarios, round-trip latency of SmartBANs, which is defined as the sum of uplink latency and downlink latency, need to be minimised for effective interactions between patients and clinicians. As such, with the developed models, we formulate a novel model for analysing round-trip transmission in SmartBANs and discuss the round-trip latency performance and its dependency on MAC layer parameters. This study also facilitates an understanding of the latency gap in existing SmartBAN systems towards future e-health applications that requires stringent low latency.

For a SmartBAN with mixed uplink EM packets and downlink feedback packets, we consider the transmission of EM in SAP, which is the default uplink period defined by SmartBAN MAC and downlink transmission in CMP by FEDM in minimising latency. Note that in this study, EM packets are only sent during the designated TDMA slots in SAP since access by slotted ALOHA in CMP is likely to incur high collisions with downlink packets broadcasted in the same time period. This channel access mechanism for round-trip transmission is illustrated in Fig. 4.10. Since uplink transmission is scheduled during SAP, the uplink latency of a packet is caused by waiting for its designated TDMA time slot. As such, we formulate uplink latency using residual time approach [175].

As shown in Fig.4.10, the residual time, $R_i$, seen by the $i$-th packet generated by a sensor node refers to the remaining time until either current TDMA time slot
or current vacation is complete. It is worthwhile to note that for any sensor intending to send EM packets, the downlink transmission duration, $T_{\text{CMP}}$, inactive period, $T_{\text{IAP}}$, D Beacon duration together with the duration of time slots assigned to other $N-1$ actuator nodes form an uplink vacation, $V_u$ as illustrated in Fig. 4.10. Hence, we have $V_u = T_{\text{IAP}} + T_{\text{CMP}} + NT$ and the average residual time, $R$, can be represented as follows:

$$R = \frac{1}{2} \lambda T^2 + \frac{(1 - \rho_{\text{up}})}{2} V_u$$

(4.19)

in which $\rho_{\text{up}} = \lambda T$ is the channel utilisation by uplink transmission of an individual sensor node. Assuming the average queue length at a sensor node is $N_{Q,u}$, the average uplink latency, $D_{\text{up}}$, is represented as:

$$D_{\text{up}} = R + (T + V_u) N_{Q,u}$$

(4.20)

According to Little’s theorem, $N_{Q,u} = \lambda D_{\text{up}}$, thus, $D_{\text{up}}$ can be solved as follows:

$$D_{\text{up}} = \frac{\lambda T^2 + (1 - \rho_{\text{up}}) V_u}{2(1 - \rho_{\text{up}} - \rho_{\text{up}} V_u)} = \frac{\lambda T^2 + (1 - \rho_{\text{up}})(NT + T_{\text{CMP}} + T_{\text{IAP}})}{2[1 - \rho_{\text{up}} - \lambda (NT + T_{\text{CMP}} + T_{\text{IAP}})]}$$

(4.21)

For $D_{\text{up}}$ to be bounded, $\rho_{\text{up}} + \lambda V_u < 1$ is required. Therefore, the packet
4.6 A Discussion on SmartBAN Round-trip Latency

The generation rate at a sensor node should satisfy $\lambda < 1/T_{IBI}$. Otherwise, there is no steady state of uplink latency model such that queue length and the waiting time of a packet tend to be infinite. For the convenience of specification, we denote downlink latency as $D_{down}$, which is modeled by (4.15). Moreover, $\rho_{down} = GT$, which is equal to the $\eta$ used in previous discussions for FEDM, represents the downlink channel utilisation considering aggregated downlink packets at the hub to actuators in a SmartBAN. Overall, based on (4.15) and (4.21), the round-trip latency, $D_{round}$, can be evaluated by:

$$D_{round} = D_{up} + D_{down}$$

(4.22)

Given (4.22), we analytically evaluate how varying durations of SAP, CMP, IAP will influence $D_{round}$ and how $D_{round}$ can be minimised.

4.6.2 Performance Evaluations

In this section, we first validate our model using simulations and show that the round-trip latency, $D_{round}$, can be minimised by selecting optimal downlink transmission duration. $D_{down}$ and $D_{up}$ under varying downlink/uplink traffics are verified in Figs. 4.11(a) and (b), respectively. Since $T_{CMP}/T_{IBI} = 0.67$, $G$ is within 0.05 ~ 0.25, i.e. $\rho_{down}$ from 0.125 ~ 0.625 in Fig. 4.11(a) to ensure the steady-state of downlink queue. Similarly, aggregated uplink utilisation, $\rho_{up} \times N$, is less than 0.25 in Fig. 4.11(b). Due to the TDMA scheme, uplink latency is
affected more by the increasing traffic load with higher average latency and larger deviation. Further, we present an understanding of how MAC timing parameters impact latency. In Fig. 4.12, we numerically solve round-trip latency, \( D_{\text{round}} \), in our model. \( T_{\text{SAP}} \) is determined by the number of nodes due to TDMA, i.e., \( T_{\text{SAP}} = NT \), thus \( D_{\text{round}} \) of a SmartBAN depends on the downlink access durations, i.e. \( T_{\text{CMP}} = MT \) as shown in Figs. 4.12. It should be noted that there is a trade-off between uplink and downlink latency since increasing \( M \) can reduce downlink, but in return incur high uplink latency. As such, \( M \) values that minimise \( D_{\text{round}} \) as shown by the red line in Figs. 4.12(a) and 4(b) can be numerically solved based on our round-trip latency model. Since \( D_{\text{up}} \) contributes more to \( D_{\text{round}} \), allocating more \( M \) to reduce downlink latency helps reduce \( D_{\text{round}} \) when uplink traffic is light as shown in Fig. 4.12(b), while less \( M \), meaning shorter \( T_{\text{CMP}} \) and \( T_{\text{IBI}} \), is demanded to reduce \( D_{\text{up}} \) as \( \rho_{\text{up}} \) increasing such that \( D_{\text{round}} \) can be effectively reduced.

4.7 Summary

This chapter presented the first downlink latency and energy performance study for SmartBANs. We first evaluated the latency of the SDM specified by the SmartBAN MAC protocol, and proposed ISDM to address the limitation in SDM design. In order to achieve low latency and high energy efficiency in SmartBANs, we proposed LEDM and FEDM, in conjunction with energy-saving mechanisms, for transmitting downlink packets in SmartBANs. Embedded Markov chain
models were developed for LEDM and FEDM, which accurately evaluated how $T_{SAP}$, $T_{CMP}$ and $T_{IBI}$ impact on the downlink latency and energy-savings performances. Based on the comprehensive knowledge on the latency and energy-savings of SDM, ISDM, LEDM and FEDM, a novel downlink transmission framework was proposed to achieve flexible downlink transmission mode selection and transmission time duration determination in SmartBANs.

The numerical and simulation analyses arising from this chapter highlight the following insights: (a) ISDM reduces latency by an amount of $T_{SAP}$ compared to SDM, suggesting that beacon-enabled downlink transmission period shall begin following the beacon scheduling; (b) compared to SDM and ISDM, LEDM can effectively reduce latency without consuming extra energy; and (c) FEDM can further reduce latency in meeting stringent latency constraint, but incurs more energy consumption. As such, for latency-constrained applications, downlink transmission mode, and the values of $T_{IBI}$ and $T_{CMP}$ can be selected to meet the latency constraint. For general applications without specific latency requirement, LEDM would be a favourable solution since it effectively reduces latency compared to the SDM, and retains the high energy-saving performance.

Finally, the round-trip transmission that includes both uplink and downlink of SmartBANs was investigated. An accurate model was developed to characterise the round-trip latency performance of SmartBANs. With our proposed model, the downlink, uplink and round-trip latency were precisely analysed under different selections of access durations, i.e. $T_{SAP}$, $T_{CMP}$, $T_{IAP}$ in SmartBAN beacon-enabled MAC frame. A trade-off between uplink and downlink latency was highlighted and discussed. Analytical and simulation results showed that compared to uplink transmission duration, i.e., $T_{SAP}$, the downlink transmission duration, i.e., $T_{CMP}$, plays a more important role in reducing the round-trip latency.

Overall, in this chapter, the downlink latency and energy consumption in SmartBANs were critically studied, and the round-trip transmission latency was discussed. In the next chapter, we investigate the latency performance of heterogeneous optical and wireless access networks that realise converged transmission of the traffic aggregated from multiple SmartBAN users in local area networks.
Chapter 4 SmartBAN Downlink Performance Study: A Novel Downlink Transmission Framework for Reducing Latency and Energy Consumption
Chapter 5
Deep Neural Network Supervised Bandwidth Allocation Decisions for Low-Latency Heterogeneous E-health Network

5.1 Introduction

Healthcare is one of the major areas to be revolutionised by the rapid advancement of Internet-of-Things (IoT) and Tactile Internet paradigms [1-3]. The emerging remotely-controlled human-machine/robot (H2M) applications as well as haptic communications in IoT and Tactile Internet are envisioned to enable remote and real-time interactions between clinicians and patients. The innovations in e-health applications are expected to drive ubiquitous healthcare services, such as telediagnosis, telerehabilitation, and ultimately telesurgery, anytime and anywhere. In Chapters 3 and 4, latency and energy performances of the smart body area network (SmartBAN) have been comprehensively investigated. With miniaturised sensors and actuators placed on or near the human body, SmartBANs are essential in delivering sensing/control and actuation data from/to the individuals. To realise real-time interaction between SmartBAN users and remote clinicians, an ultra-responsive and ultra-reliable communication system is required.

In this regard, the integration of passive optical network (PON) with wireless local area network (WLAN) access technology has been considered a promising solution for next-generation e-health systems [8,185,186]. The convergence of the PON and WLAN combines the advantages of high data rate, reliability, and capacity of PON and the mobility of wireless communication [13]. Specifically, the Ethernet PON (EPON) or Gigabit PON (GPON) technologies provide reliable backhaul support for broadband access, and the existing PONs can be sustainably upgraded to cater to increasing capacity demand [13]. Compared to LTE and
WiMAX technologies used in the wireless front-end, the deployment of WLANs is flexible and cost-effective, and complements cellular networks by offloading mobile data via WiFi access points. Importantly, WLANs can be easily dedicated to specific user groups or applications with certain quality-of-service (QoS) demands such as emerging H2M applications, e-health applications. Taking the aforementioned advantages of PON and WLAN, we investigate the capability of integrated PON and WLAN in supporting low-latency SmartBAN applications. Critically understanding the impacts of different bandwidth allocation solutions in such heterogeneous access networks on network end-to-end latency performance is a primary goal in this chapter.

Fig. 5.1 illustrates a heterogeneous network architecture that includes a SmartBAN, WLAN, and PON. A WLAN provides access for SmartBAN users within its coverage and data from multiple SmartBANs are aggregated into the integrated optical network unit and wireless access point (ONU-AP). In order to expedite the data processing time and reduce the end-to-end latency, the local health server/database is placed at the central office (CO) of the PON and edge cloudlets are used to push data and services close to users. The latency bottleneck over such a heterogeneous network, however, mainly depends on the uplink direction since uplink bandwidth is shared among multiple ONU-APs in the PON, and multiple SmartBAN users located in the same WLAN contend bandwidth for uplink transmission. As such, in this chapter, we focus on the uplink end-to-end
5.1 Introduction

latency, termed $D_{E2E}$. This $D_{E2E}$ comprises the transmission, propagation, processing, and queuing latency experienced by a packet from the SmartBAN hub until it reaches the CO. In this regard, medium access control (MAC) and bandwidth allocation schemes in both WLAN and PON are critical in improving the uplink latency.

As overviewed in Chapter 2, the uplink bandwidth in PONs is generally allocated via implementing dynamic bandwidth allocation (DBA) schemes at the CO. REPORT and GATE messages are exchanged between the CO and ONU-APs to facilitate bandwidth allocation. The REPORT message is used by ONU-APs to notify CO of the requested bandwidth, while the GATE is used by the CO to inform the allocated bandwidth to the ONU-APs. Note that based on the bandwidth requested through REPORT, the CO decides how much of bandwidth is to be allocated to ONU-APs, yielding to different DBA schemes. The amount of the bandwidth allocated to ONU-APs by the CO impacts the network latency performance. As explained in Chapter 2, existing DBA schemes can be categorized into classic and predictive schemes. In classic DBA schemes, the CO allocates the requested amount of bandwidth to the ONU-APs. Thus, arriving packets need to be reported by ONU-APs, and then transmitted in the next-granted timeslot. Fixed-cycle, limited-service, exhaustive-service DBA schemes are common classic DBA schemes. Among them, the limited-service DBA scheme has been widely adopted as a baseline scheme [44-46]. In predictive DBA schemes, the CO allocates some additional bandwidth to the requested bandwidth. This is to reduce packets' waiting time for reporting and bandwidth granting. Arrivals can be transmitted in the current polling cycle using the additional bandwidth without needing to request bandwidth for them. This additional bandwidth in predictive DBA schemes can be a static amount of bandwidth, or statistically estimated based on ONU traffic load as detailed in Chapter 2. It can be noted that bandwidth allocation in both classic and predictive DBA schemes relies on the requested bandwidth from ONU-APs. On the other hand, existing studies have shown that multiple network features, such as the traffic aggregated in WLAN and round-trip transmission time (RTT) in fiber links, affect the bandwidth allocated among ONU-APs [187-190]. Based on existing research, the impact of these features on the allocated bandwidth has not been fully evaluated yet.
Unlike the bandwidth allocation schemes in PONs, the channel access in WLANs is governed by periodically broadcasted channel beacons as specified in IEEE 802.11 MAC [41]. Contention-based distributed coordination function (DCF) via carrier sensing multiple access with collision avoidance (CSMA/CA) and contention-free point coordination function (PCF) via centralized scheduling are two fundamental mechanisms. Hybrid coordination function (HCF) further specifies a hybrid MAC frame that incorporates both PCF and DCF to ensure QoS. Using HCF, latency-sensitive traffic can be prioritized by scheduling transmission opportunities (TXOPs) in PCF. In comparison to bandwidth allocation that relies on REPORT in PONs, the timing parameters that define a MAC frame such as access durations for PCF and DCF within a frame and duration of a scheduled TXOP, impact the latency in a WLAN. A reference HCF frame design provided by IEEE 802.11 [77] heuristically determines the TXOP considering the traffic load. Moreover, studies reported in [191-193] have evaluated the end-to-end uplink latency, $D_{E2E}$, over fiber-wireless networks with time division multiple access (TDMA) deployed in WLANs and PONs. In [194-196], the authors have studied the cooperation between an ONU and its AP. Although the proposed methods in these studies effectively reduce the uplink latency in PON, they do not consider the bandwidth allocation in the WLAN. As we are considering an integrated network with PON and WLAN, it is important to study the impact of different bandwidth allocation decisions in WLAN and PON have on $D_{E2E}$.

To address these important aspects, in this chapter, we propose to exploit machine learning, particularly deep neural networks (DNNs), to learn the dependency of $D_{E2E}$ on bandwidth allocation decisions made in (a) DBA in the PON and (b) HCF in the WLAN segment. With this dependency learnt, the DNN can provide comprehensive evaluations on the bandwidth values to be allocated in reducing $D_{E2E}$ under different network scenarios in turn. A DNN, characterised as a neural network with multiple layers and neurons, is advantageous in recognizing abstract patterns and complex associations among multiple features [160]. Therefore, we consider the DNN as a useful technique to evaluate bandwidth allocation decisions that involve multiple decision parameters and network protocol/configuration features. In this chapter, we introduce the proposed DNN architecture, including the selection of input features and output targets and present the supervised training process and the resultant learning
outcome. Further, we critically analyse how the bandwidth allocation decisions, i.e., selection of timing parameters in WLAN HCF MAC frame and bandwidth allocation in DBA, impact the uplink latency with the trained DNN. Finally, adopting bandwidth allocation decisions supervised by the trained DNN, we show that the uplink latency in each network segment and $D_{E2E}$ of the entire network can be effectively reduced compared to existing schemes. For this purpose, we will first analyse the latency of DBA schemes and WLAN HCF and highlight the key bandwidth decision parameters, and network features that impact the latency.

### 5.2 End-to-end Latency of Heterogenous Networks

In this section, we present the existing bandwidth allocation schemes and their latency performance in a heterogeneous network. We first discuss the latency in standalone PONs and WLANs followed by the latency of a heterogeneous network where both PON and WLAN coexists.

#### 5.2.1 Latency of DBA Schemes in PONs

Fig. 5.2 illustrates the end-to-end uplink transmission in a heterogeneous SmartBAN + WLAN + PON e-health network. As shown, the CO coordinates uplink transmissions from ONU-APs using REPORT and GATE messages. The time interval between two consecutive transmissions of an ONU-AP is known as a polling cycle. In each polling cycle, an ONU-AP sends a REPORT to notify the CO about its queue length (in bytes). The CO then calculates and grants bandwidth to the ONU-AP using the GATE message, which indicates the next transmission start time and the duration. Let us denote the requested bandwidth in the REPORT as $BW_{req}$ and the bandwidth granted by the CO as $BW_{grant}$. Note that the value of $BW_{grant}$ depends on the DBA algorithm adopted. Considering Fig. 5.2, the uplink latency experienced by a packet in the PON, $D_{DBA}$, can be evaluated as follows:

$$D_{DBA} = D_{poll} + D_{grant} + D_{queue_{onu}}$$  \hspace{1cm} (5.1)

For a given packet, $D_{poll}$ is the time interval between this packet arriving at an ONU-AP and the next REPORT being sent by that ONU-AP; $D_{grant}$ is the time
period from sending the REPORT until the start of the granted transmission time slot during which this packet is transmitted. $D_{\text{queue_onu}}$ is the queuing time from the start of the transmission time slot until the packet is transmitted in that time slot. On average, $D_{\text{queue_onu}}$ is approximately half of a transmission slot time, which means $D_{\text{queue_onu}}$ is insignificant compared to $D_{\text{poll}}$ and $D_{\text{grant}}$.

As explained in Chapter 2 Section 2.3, in the baseline limited-service DBA, the CO allocates $BW_{\text{grant}} = \min\{BW_{\text{req}}, BW_{\text{max}}\}$ to an ONU-AP in each polling cycle. $BW_{\text{max}}$ is the maximum available bandwidth for ONU-APs. As such, only the requested bandwidth, $BW_{\text{req}}$, is allocated to ONU-APs in limited-service DBA scheme when $BW_{\text{req}} < BW_{\text{max}}$. In comparison, in predictive DBA schemes, the CO predicts incoming arrivals and estimate bandwidth $BW_{\text{pred}}$ for these arrivals. When allocating bandwidth to ONU-APs, the CO pre-allocates this surplus $BW_{\text{pred}}$ in additional to ONU-APs’ request $BW_{\text{req}}$, that is, $BW_{\text{grant}} = \min\{BW_{\text{req}} + BW_{\text{pred}}, BW_{\text{max}}\}$. Note that in the limited-service DBA scheme, packets are deferred at least 1 polling cycle in their transmission as shown in Fig. 5.2 due to the report-then-grant sequence. In predictive DBA schemes, pre-allocating the $BW_{\text{pred}}$ estimated allows incoming arrivals to be transmitted in the current polling cycle without reporting, thereby reducing the $D_{\text{grant}}$. On the other hand, adding $BW_{\text{pred}}$ to $BW_{\text{req}}$ increases $BW_{\text{grant}}$ compared to that in the limited-service DBA scheme. This may increase the overall polling cycle duration, which increases $D_{\text{poll}}$ in turn.

**Figure. 5.2** Dynamical bandwidth allocation and IEEE 802.11 HCF for end-to-end uplink transmission in heterogeneous networks.
As such, $BW_{\text{pred}}$ is a critical decision parameter that needs to be carefully determined. It is important to determine $BW_{\text{pred}}$ that ensures the summation of $D_{\text{poli}}$ and $D_{\text{grant}}$ is minimized to achieve a low $D_{\text{DBA}}$.

Conventionally, this $BW_{\text{pred}}$ in predictive DBA schemes is estimated using statistical algorithms overviewed in (2.6) and (2.7) in Chapter 2. These algorithms estimate $BW_{\text{pred}}$ by predicting packet arrivals (in bytes) within a polling cycle time. Such $BW_{\text{pred}}$ values, however, may not be the optimal solutions in reducing $D_{\text{DBA}}$. Comprehensively evaluating the dependency of $D_{\text{DBA}}$ on $BW_{\text{pred}}$ is demanded. Note that by investigating different $BW_{\text{pred}}$ used, we are able to understand and compare the performance of existing DBA schemes. For example, when $BW_{\text{pred}} = 0$, the latency performance corresponds to that of the baseline limited-service DBA scheme, whereas if $BW_{\text{pred}} = BW_{\text{max}}$ is allocated, the latency performance converges to that of a fixed-cycle DBA scheme. As such, in this chapter, $BW_{\text{pred}}$ is the focus in investigating the DBA performance in PONs. We address critical questions of: (a) the impacts of $BW_{\text{pred}}$ decisions in $BW_{\text{grant}} = \min\{BW_{\text{req}} + BW_{\text{pred}}, BW_{\text{max}}\}$ on $D_{\text{DBA}}$; and (b) the amount of $BW_{\text{pred}}$ that shall be allocated in reducing $D_{\text{DBA}}$ for diverse network scenarios, e.g., with different network loads, and network configurations.

### 5.2.2 Latency of WLAN HCF

As introduced in Chapter 2 Section 2.2, the HCF defined in WLAN MAC is considered a favourable channel access mechanism to support latency-sensitive applications with QoS demand. The HCF hybrid MAC frame, comprising service interval (SI), controlled access period (CAP), contention period (CP), and uplink latency under HCF are presented in Fig. 5.2. Let us consider a scenario where $N$ number of SmartBAN users are connected to an individual WLAN. An SI contains transmission opportunities, TXOP: ($i=1, 2, ..., N$), a CP, and short inter-frame spaces (SIFSs) between allocated TXOPs as shown in Fig. 5.2. Let us denote the duration of SI, TXOP, SIFS, and CAP as $T_{\text{SI}}$, $T_{\text{TXOP}}$, $T_{\text{SIFS}}$, and $T_{\text{CP}}$, respectively, where $T_{\text{SI}} = N(T_{\text{TXOP}} + T_{\text{SIFS}}) + T_{\text{CP}}$. Noticeably, the selection of $T_{\text{TXOP}}$, $T_{\text{SI}}$, and $T_{\text{CP}}$ impacts the latency in WLANs. The uplink latency by HCF, $D_{\text{HCF}}$, can be estimated as follows:

$$D_{\text{HCF}} = D_{\text{res}} + nT_{\text{SI}} + D_{\text{queue \_ user}}$$ (5.2)
where $D_{res}$ is the residual time for completing an on-going packet transmission or waiting for the next TXOP, and $n$ is the number of SIs that a packet has to wait until the start of a TXOP, in which this packet can be transmitted. The parameter $D_{queue\_user}$ is the queuing time at the SmartBAN until this packet is transmitted in the given TXOP. For a given $T_{CP}$, increasing $T_{TXOP}$ increases the transmission opportunity within an SI increase in $T_{SI}$ and $D_{res}$. When the number of SmartBAN users, $N$, or traffic load, $l_{WLAN}$, at an individual user increases, a longer $T_{TXOP}$ is expected, which also increases the $T_{SI}$. As such, suitable $T_{TXOP}$, $T_{CP}$, and $T_{SI}$ values should be chosen based on $N$ and $l_{WLAN}$ in order to reduce $D_{HCF}$ in (5.2).

Bandwidth allocation in WLAN via assigning values of $T_{TXOP}$, $T_{CP}$, and $T_{SI}$ is therefore critical for reducing $D_{E2E}$. Note that $T_{TXOP}$, $T_{CP}$ and $T_{SI}$ represent the duration of the corresponding time slot. Therefore, we discuss these parameters in the unit of μs.

### 5.2.3 End-to-End Uplink Latency Analysis

With above analysis, the end-to-end latency, $D_{E2E}$, of delivering a packet from a SmartBAN user to the CO can be represented as follows:

$$D_{E2E} = D_{HCF} + D_{DBA} + D_{pro} + D_{trans}$$  \hspace{1cm} (5.3)

where $D_{pro}$ is the processing time for protocol and packet format translation at the ONU-AP and $D_{trans}$ is the transmission time of that packet in WLAN and PON in conjunction with the propagation time in both network segments. As per (5.3), $D_{DBA}$ and $D_{HCF}$ is the key to reduce $D_{E2E}$.

The selection of $T_{TXOP}$, $T_{CP}$, $T_{SI}$ and $BW_{pred}$ however, is challenging since multiple protocol parameters, network configurations, and decision variables are involved in reducing the latency. For example, for WLAN, the physical layer data rate, $R_{WLAN}$, number of connected SmartBANs, $N$, and associated packet statistics, e.g. packet length and arrival rate, will impact the choice of $T_{TXOP}$ and $T_{SI}$. Further, aggregated traffic load from a WLAN, $l_{WLAN}$, ONU-to-CO distance, $d_{O2C}$, the number of ONU-APs, $M$, data rate of the optical link, $R_{PON}$, and packet statistics in the optical access network need to be considered to determine the $BW_{pred}$. Based on the above analysis, critical bandwidth allocation decision, protocol, and network parameters that impact $D_{E2E}$ are listed in Table 5.1. To
5.3 DNN in Predicting End-to-End Latency

5.3.1 DNN Architecture and Learning Model

<table>
<thead>
<tr>
<th>Notations</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$BW_{grant}$ (in bytes)</td>
<td>Bandwidth granted to ONU-APs by DBA operation; $BW_{grant} = BW_{req} + BW_{pred}$</td>
</tr>
<tr>
<td>$T_{poll, max}$</td>
<td>Maximum polling cycle duration</td>
</tr>
<tr>
<td>$M$</td>
<td>Number of ONU-APs, i.e., WLANs</td>
</tr>
<tr>
<td>$R_{PON}$</td>
<td>Optical link data rate</td>
</tr>
<tr>
<td>$d_{O2C}$</td>
<td>ONU-to-CO distance</td>
</tr>
<tr>
<td>$S_{min}/S_{max}/S_{avg}/S_{var}$</td>
<td>Optical network packet statistics: minimum/maximum/average/variance length</td>
</tr>
<tr>
<td>$l_{WLAN}$</td>
<td>Aggregated traffic load in a WLAN</td>
</tr>
<tr>
<td>$N$</td>
<td>Number of SmartBAN users</td>
</tr>
<tr>
<td>$T_{TXOP} (\mu s)$</td>
<td>Time duration of TXOP</td>
</tr>
<tr>
<td>$T_{CP} (\mu s)$</td>
<td>Time duration of CP</td>
</tr>
<tr>
<td>$T_{SI} (\mu s)$</td>
<td>Time duration of SI</td>
</tr>
<tr>
<td>$R_{WLAN}$</td>
<td>Physical layer data rate in WLAN</td>
</tr>
<tr>
<td>$P_{min}/P_{max}/P_{avg}/P_{var}$</td>
<td>WLAN packet statistics: minimum/maximum/average/variance length</td>
</tr>
</tbody>
</table>

fully understand the association between $D_{DBA}$, $D_{HCF}$, $D_{E2E}$ and bandwidth allocation decision parameters, $T_{TXOP}$, $T_{CP}$, $T_{SI}$ and $BW_{pred}$ in a heterogeneous network, we apply a deep neural network to learn their dependency since DNN is capable of recognising complex associations among multiple parameters. Moreover, compared to other machine learning techniques, the application of DNN is flexible depending on the selection of input features and output targets without being limited by specific problems. With the above-mentioned association learnt, the DNN can then be used to comprehensively evaluate bandwidth allocation decisions in WLAN HCF and DBA, and improve existing solutions to reduce $D_{HCF}$, $D_{DBA}$ and $D_{E2E}$ in turn. The details of the proposed DNN-supervised bandwidth allocation decision-making is illustrated in the following section.
As introduced in Chapter 2 Section 5, neural networks, particularly deep neural networks that comprise multiple layers of non-linear computational neuron units, are superior in recognising abstract patterns and complex associations among multiple features. Exploiting a DNN, insights into the dependency of uplink latency on multiple network features and bandwidth allocation decisions can be attained. Specifically, the proposed DNN learning model is presented in Fig. 5.3. With supervised training, we first train a DNN at the CO to learn the association between $D_{DBA}$, $D_{HCF}$, $D_{E2E}$ and bandwidth allocation decisions and multiple network features. Then, when supervised training is completed, the trained DNN is able to predict $D_{DBA}$, $D_{HCF}$, $D_{E2E}$ corresponding to different bandwidth allocations decisions, i.e., $T_{TXOP}$, $T_{CP}$, $T_{SI}$, and $BW_{pred}$, thereby supervising the optimal values of these decision parameters that reduce the $D_{E2E}$ as illustrated in Fig 5.3.

The detailed architecture and the input and output of the proposed DNN are illustrated in Fig. 5.4. As introduced in Chapter 2, a multiple-layered DNN can identify abstract patterns and relationships between given inputs and outputs. In Fig. 5.4, a DNN comprises an input layer, an output layer, and multiple hidden layers in between. Each layer is composed of neuron units, which are non-linear functions that map the associated inputs to an output. With supervised training, a DNN iteratively adjusts the weights associated with its neurons and layers to minimise the error between its predicted outputs and target outputs, thereby accurately characterizing the pattern between input features and output targets. Here, the objective of the proposed DNN is to learn $D_{HCF}$, $D_{DBA}$ and $D_{E2E}$ under various bandwidth allocation, network and traffic conditions, which parameters are listed in Table 5.1. In characterising this tension among bandwidth allocation decisions and network latency performances, we formulate $x = \{N, T_{TXOP}, T_{CP}, l_{WLAN}, M, d_{O2C}, BW_{pred}\}$ as input features to the DNN since these parameters generally vary in different heterogeneous networks and impact latency significantly as analysed above. The target output is formulated as $y = \{D_{HCF}, D_{DBA}, D_{E2E}\}$. With input features and target output characterised as discussed above, the DNN architecture is finalised via supervised training. The objective of supervised training is to minimise the mean square error (MSE) between the predicted outputs from the DNN and target outputs provided by the training set. Specifically, the number of layers and the number of neuron units within each
layer are determined by greedy layer-wise training [158]. The weight matrix is iteratively tuned by classic gradient-descent-based backward propagation algorithm [158].

5.3.2 Supervised Training

To implement supervised training, a training set \( S = \{ (x_i, y_i) \mid i = 1, 2, \ldots, k \} \) is first collected using an event-driven packet-level simulation environment (MATLAB). In collecting each sample \((x_i, y_i)\) in \(S\), we use a randomly-selected \(x_i\) and record the corresponding \(y_i\). Specifically, for a given \(x_i\), the simulated network is configured by \(N(i)\) number of SmartBANs in a WLAN, \(M(i)\) number of ONU-APs with \(d_{O2C}(i)\) distance, packets generated by each SmartBAN follows a Poisson process with the aggregated load as \(l_{WLAN}(i)\). By allocating \(T_{TXOP}(i), T_{CP}(i)\) in WLAN HCF and \(BW_{pred}(i)\) in DBA, the \(y_i\) is obtained by measuring the average

**Figure. 5.3** An illustration of the proposed DNN learning and bandwidth decision learning model.
Chapter 5 Deep Neural Network Supervised Bandwidth Allocation Decisions for Low-Latency Heterogeneous E-health Network

$D_{HCF}$, $D_{DBA}$, and $D_{E2E}$ over a period of 500s of the network running time. This procedure is repeated until sufficient $(x_i, y_i)$ samples, e.g. $> 10^4$ in this paper, are collected. Note that test samples, independent from the training samples, can also be generated following the same procedure. Once training is complete, the performance of the trained DNN can be evaluated using a test set as follows:

$$\text{MSE} = \frac{1}{l} \sum_{i=1}^{l} |D_{E2E}(i) - \hat{D}_{E2E}(i)|^2 \quad (5.4)$$

where $l$ is the number of test samples, $D_{E2E}(i)$ is end-to-end latency value of the $i$-th test sample, and $\hat{D}_{E2E}(i)$ is the predicted value by the trained DNN.

The architecture of the DNN and the selection of training samples impact the training performance. We adopt the layer-wise training method and sensitivity of the training sample selection is analysed in Fig. 5.5. As can be viewed in Fig. 5.5(a), the DNN has an initial 1 hidden layer with 3 neuron units, thus showing a large MSE. As the number of neuron units is increased, its training performance improves dramatically and then slows down to show minimal improvement beyond 8 neurons. This is mainly due to the limitation of the single-layer architecture. Hence, a second hidden layer is added to the DNN, and the resultant MSE is shown to be effectively reduced. Again, more neurons are added until the training performance improvement stalls. A new layer is then added. This process is repeated until an MSE in the order of 1 is achieved.

In Fig. 5.5(b), we investigate the impact of training set size on the training performance. Training sets that contain 0.2 to $2 \times 10^4$ samples as illustrated in Fig.

**Figure 5.4** DNN architecture, and the input and output features.
5.5(b) are used to train the DNN, respectively, and the MSE results after individual trainings are plotted. As shown in Fig. 5.5(b), increasing the number of training samples, in general, improves training performance. Note that when the traffic load increases, either the WLAN or PON, or both, will saturate, resulting in high $D_{E2E}$ compared to those reported in the non-saturated situation. Here, saturation in the WLAN refers to network overload that the packet buffer at each SmartBAN is always non-empty, which incurs high $D_{HCF}$. Similarly, saturation in the PON refers to when the packet buffer at each ONU-AP is always non-empty and $D_{DBA}$ increases dramatically. By selecting the training samples collected in non-saturated network scenarios, significantly fewer samples are required to reduce the MSE of the trained DNN. This is because either saturated WLAN or PON, due to overload, incurs higher $D_{E2E}$ compared to those reported in the non-saturated situation. The DNN tends to reduce the prediction error for these saturated $D_{HCF}$, $D_{DBA}$ or $D_{E2E}$ values as erroneously predicting saturated $D_{HCF}$, $D_{DBA}$ or $D_{E2E}$ affects MSE severely. However, to achieve low $D_{E2E}$, accurate learning of non-saturated $D_{E2E}$ is required. As such, we propose to classify training samples into two classes following (5.5):

$$I_{WLAN} < \min \left( \frac{N_{TXOP}}{N(T_{TXOP} + T_{SIFS}) + T_{CP}}, \frac{R_{PON \cdot P_{avg}}}{MR_{WLAN} \cdot S_{avg}} \right)$$

(5.5)

where $\frac{N_{TXOP}}{N(T_{TXOP} + T_{SIFS}) + T_{CP}}$ and $\frac{R_{PON \cdot P_{avg}}}{MR_{WLAN} \cdot S_{avg}}$ are the traffic load thresholds that lead to a saturated WLAN and PON, respectively. By selecting samples satisfying (5.5), the training performance is improved (orange line in Fig. 5.5(b)) compared to that using a raw training set (blue line in Fig. 5.5(b)). With the proposed sample selection using (5.5), the DNN quickly converges to optimal weights and requires significantly fewer samples for training. Based on the analysis in Fig. 5.5(b), we use a training set with $1.5 \times 10^4$ samples to train the DNN via layer-wise training (Fig. 5.5(a)). In Fig. 5.5(c), we present the outcome of the trained DNN in predicting $D_{HCF}$, $D_{DBA}$ and $D_{E2E}$ of 300 test samples for illustrative purpose. Results in Fig. 5.5(c) show that providing the trained DNN with bandwidth allocation decision and network feature parameters, i.e., input feature $x$, accurate predictions of $D_{HCF}$, $D_{DBA}$ and $D_{E2E}$ are attained.

The above results in supervised training indicate the importance of selecting
Chapter 5 Deep Neural Network Supervised Bandwidth Allocation Decisions for Low-Latency Heterogeneous E-health Network

(a) Layer-wise learning

(b) Training set size and sample selection

(c) Training outcome

**Figure 5.5** DNN supervised training.
training samples, designing DNN architecture in improving the training performance. Once training is complete, the impact of plausible bandwidth allocation decisions, along with multiple network features, on uplink latency can be analytically studied using the trained DNN. Detailed analyses are provided in the following section.

5.4 Bandwidth Allocation Decisions Analysis

In this section, we present the dependency of $D_{HCF}$, $D_{DBA}$, and $D_{E2E}$ on varying bandwidth allocations, i.e., $T_{TXOP}$, $T_{CP}$, $T_{SI}$, and $BW_{pred}$, learnt through the trained DNN. Note that different types of PONs, such as EPON, GPON, and WLANs in IEEE 802.11 series can be used to underpin a converged heterogeneous network. The PON DBA and WLAN HCF operations are similar. For illustrative purpose, we present our analysis based on an EPON and IEEE 802.11e WLAN. The parameters of the input feature $x$ along with their numerical values considered in this analysis are listed in Table 5.2. In this work, we consider fixed length (128 bytes) WLAN packets referring to [102]. Upon arriving at an ONU-AP for transmission to the CO, these packets are re-encapsulated into 64 bytes. This is because IEEE 802.11 packets need to have their irrelevant frames removed and be re-formatted for transmission in PON.

5.4.1 Uplink Latency, $D_{HCF}$, of WLAN

To study the dependency of $D_{HCF}$ on WLAN parameters, we vary $N$, $T_{TXOP}$, $T_{CP}$, $l_{WLAN}$ and input them into the trained DNN. The output $D_{HCF}$ values are plotted in Fig. 5.6. In this analysis, we consider $N = 4$ SmartBAN users and $T_{TXOP} = 10 \mu s$ allocated to each user. Fig. 5.6(a) shows the $T_{TXOP}$ allocations under varying $l_{WLAN}$. As shown, when the aggregated traffic load in a WLAN increases, $T_{TXOP}$ allocated to each user increases accordingly, resulting in a decrease in latency. For instance, to manage aggregated arrivals with $l_{WLAN} = 0.8$, a $T_{TXOP} > 60 \mu s$ is required. This complies with our analysis in (2). Although a small $T_{TXOP}$ leads to a shorter $T_{SI}$, a packet may need to wait multiple SIs before the packets buffered ahead of this packet are transmitted. On the other hand, allocating a large $T_{TXOP}$ to SmartBAN users cannot always guarantee a reduction in $D_{HCF}$ since increasing $T_{TXOP}$ increases $T_{SI}$, hence the waiting time for a TXOP. In comparison, when
$l_{\text{WLAN}} = 0.2$ as shown in Fig. 5.6(a), a $T_{\text{TXOP}} = 10 \mu s$ is considered as sufficient and $D_{\text{HCF}}$ is shown to increase with $T_{\text{TXOP}}$. Fig. 5.6(b) shows the impact of the number of SmartBAN users $N$ on $T_{\text{TXOP}}$ decision in a WLAN. For this analysis, we have considered a fixed $l_{\text{WLAN}} = 0.2$ and as a result, a larger $N$ implies less load generated at individual SmartBAN users. However, as can be observed, a longer $T_{\text{TXOP}}$ is still required since more packets are accumulated at the buffer due to an increased number of SIFS and increased $T_{\text{SIFS}}$. Fig. 5.6(c) illustrates $D_{\text{HCF}}$ as a function of $T_{\text{TXOP}}$, $T_{\text{CP}}$ and $l_{\text{WLAN}}$. For a given $T_{\text{CP}}$, there is an optimal $T_{\text{TXOP}}$ that minimises $D_{\text{HCF}}$. An inappropriate $T_{\text{TXOP}}$, either too short or long, will increase latency. Based on the results in Fig. 5.6, bandwidth allocation via $T_{\text{TXOP}}$ in WLAN can be formulated to reduce uplink latency $D_{\text{HCF}}$ as described in (5.2). Suitable $T_{\text{TXOP}}$ values can be determined considering varying $T_{\text{CP}}$, $N$, and $l_{\text{WLAN}}$ conditions. Specifically, a short $T_{\text{CP}}$ duration is generally expected in order to reduce latency. As such, in the simulation in Section 5.5, we mainly validate the selection of $T_{\text{TXOP}}$ and set the $T_{\text{CP}} = 0$. Since the CP is mainly used for delivering connection request from new SmartBAN users, a brief discussion of $T_{\text{CP}}$ and the connection time is provided below.

From the above analytical study, a long $T_{\text{CP}}$ within an SI is likely to increase the uplink latency $D_{\text{HCF}}$, while, the connection establishment time of a new SmartBAN joining the WLAN will be reduced with an increasing $T_{\text{CP}}$. Recall that upon receiving a WLAN channel beacon, SmartBAN hubs are notified of $T_{\text{SI}}$ and $T_{\text{CP}}$. Readily-connected SmartBANs are assigned TXOPs for uplink data

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Specifications and values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_{\text{PON}}$</td>
<td>1 Gbps</td>
</tr>
<tr>
<td>$T_{\text{guard}}/T_{\text{process}}$</td>
<td>1 μs guard time/ processing time</td>
</tr>
<tr>
<td>$T_{\text{REPORT}}/T_{\text{GATE}}$</td>
<td>64 bytes</td>
</tr>
<tr>
<td>$T_{\text{poll, max}}$</td>
<td>1 ms</td>
</tr>
<tr>
<td>Optical network packet</td>
<td>64 bytes Ethernet format</td>
</tr>
<tr>
<td>$T_{\text{pro}}$</td>
<td>1 μs packet processing time at ONU-AP</td>
</tr>
<tr>
<td>$R_{\text{WLAN}}$</td>
<td>540 Mbps</td>
</tr>
<tr>
<td>$T_{\text{SIFS}}$</td>
<td>10 μs</td>
</tr>
<tr>
<td>WLAN packet</td>
<td>128 bytes</td>
</tr>
</tbody>
</table>
5.4 Bandwidth Allocation Decisions Analysis

Figure 5.6 WLAN uplink latency under $T_{\text{TXOP}}$, $T_{\text{CP}}$ and traffic load $l_{\text{WLAN}}$. 

- (a) $T_{\text{TXOP}}$ vs $l_{\text{WLAN}}$ ($N = 4$, $T_{\text{CP}} = 10 \mu s$)
- (b) $T_{\text{TXOP}}$ vs $N$ ($l_{\text{WLAN}} = 0.2$, $T_{\text{CP}} = 10 \mu s$)
- (c) $T_{\text{TXOP}}$ vs $T_{\text{CP}}$ ($N = 4$, $l_{\text{WLAN}} = 0.2$)
transmission in each SI. A connection request to join the WLAN shall be sent during CP via CSMA/CA. As such, a newly-present SmartBAN needs to wait for CP, completes backoff sensing, and then sends its connection request. Thus, the connection establishment time, $D_{\text{new\_connect}}$, here defined as the waiting time when a SmartBAN hub receives a beacon until a connection request is successfully sent, can be estimated as follows:

$$D_{\text{new\_connect}} = \left(1 + \left\lfloor \frac{T_{\text{backoff}}}{T_{\text{CP}}} \right\rfloor \right) \times (T_{\text{SI}} - T_{\text{CP}}) + T_{\text{backoff}}$$  \hspace{1cm} (5.6)$$

where $T_{\text{backoff}}$ is the CSMA/CA backoff time and $\lfloor u \rfloor$ indicates the maximum integer that is less than $u$. $\lfloor T_{\text{backoff}}/T_{\text{CP}} \rfloor$ is the number of SIs that a SmartBAN hub has to experience before backoff sensing is completed. Note that here, we assume that only one new SmartBAN presents and requests to join the network at a time to estimate $D_{\text{new\_connect}}$. Based on (5.6), the trade-off between the time to establish the connection for a new SmartBAN and $D_{\text{HCF}}$ of existing SmartBANs in the WLAN is presented in Fig. 5.7. Again, we consider $N = 4$ existing SmartBANs and $T_{\text{TXOP}} = 10 \mu$s for illustrative purposes. As shown in Fig. 5.7, a shorter CP duration, i.e., $T_{\text{CP}} = 10 \mu$s and $20 \mu$s, incur higher $D_{\text{new\_connect}}$ since the new SmartBAN hub need to utilise CPs in consecutive SIs to complete backoff sensing. When increasing $T_{\text{CP}}$ to $30 \mu$s, $D_{\text{new\_connect}}$ is effectively reduced in Fig. 5.7. This is because backoff sensing and connection request sending can be

![Figure 5.7](image-url)  

**Figure. 5.7** New connection time vs uplink transmission latency in WLAN.
completed within one CP. Therefore, further increasing $T_{CF}$ does not see $D_{new\_connect}$ improvement, but yields increasing $D_{HCF}$ as shown in Fig. 5.7.

5.4.2 Uplink Latency, $D_{DBA}$, of PON

In this section, we use the DNN to analyse the impact of $BW_{pred}$ on $D_{DBA}$. In Fig. 5.8, we present the latency analysis as a function of $BW_{pred}$ that incorporate different network features and normalised network load. Note that the aggregated traffic load traversing the heterogeneous network is normalised by $l_{network} = \lambda NS_{\text{avg}}/R_{PON}$. As shown in Fig. 5.8(a), the trained DNN provides $D_{DBA}$ corresponding to all possible $BW_{pred}$ decisions in a network with 16 WLANs and 10 km ONU-to-CO fiber links. As expected, when $l_{network}$ increases, $BW_{pred}$ will be adjusted in order to reduce $D_{DBA}$. Optimal $BW_{pred}$, in this case, is highlighted in the blue curve in Fig. 5.8(a). As such, in each polling cycle, granting these $BW_{pred}$ values in addition to the reported $BW_{req}$ to individual ONU-APs minimises $D_{DBA}$ as shown in Fig. 5.8(a). Moreover, it can be noted that when $BW_{pred} = 0$, the resultant $D_{DBA}$ corresponds to the limited-service DBA scheme since only $BW_{req}$ is allocated to ONU-APs. Conversely, when $BW_{pred}$ is sufficient to transmit all arrivals accumulated during a polling cycle, the reported buffer length will be 0, i.e., $BW_{req} = 0$, in each polling cycle. As a result, the fixed $BW_{pred}$ is allocated to ONU-APs in each polling cycle. Therefore, as pointed out in Fig. 5.8(a), $D_{DBA}$ converges to a fixed-cycle DBA scheme, in which, $D_{DBA}$ linearly increases with increasing $BW_{pred}$. In comparison, $D_{DBA}$ of the statistical predictive DBA scheme using (2.7) is presented in the red curve. When ONU-APs are lightly-loaded, e.g., $l_{network} < 0.4$, insufficient $BW_{pred}$ is estimated in the predictive DBA scheme due to small $\lambda$. Conversely, when $l_{network}$ increases, $BW_{pred}$ will be over-estimated due to the increased $T_{poll}$. Consequently, as shown in Fig. 5.8, the predictive DBA scheme does not achieve optimal bandwidth allocation decisions corresponding to network feature variations. Finally, given any fixed $BW_{pred}$, $D_{DBA}$ corresponds to the performance of constant credit DBA scheme, by which a constant $BW_{pred}$ is allocated on top of the $BW_{req}$. It can be noted that $D_{DBA}$ is susceptible to the selection of $BW_{pred}$ values.

Fig. 5.8(b) shows $BW_{pred}$ for $l_{network} = 0.5$ when the ONU-to-CO distances vary from 1 ~ 20 km. In this instance again, optimizing the $BW_{pred}$ values effectively
adjusts $D_{\text{poll}}$ and $D_{\text{grant}}$, thus resulting in a low $D_{\text{DBA}}$. It should be noted that a larger $\text{BW}_{\text{pred}}$ is generally required when $d_{O2C}$ increases, as polling cycle duration of DBA is lower bounded by RTT time in an optical access network. Allocating suitable $\text{BW}_{\text{pred}}$ values as highlighted in Fig. 5.8(b) enables the packets to be transmitted without needing to be reported, thereby improving $D_{\text{DBA}}$. In Fig. 5.8(b), $D_{\text{DBA}}$ ranges from 50 ~ 100 $\mu$s corresponding to the DNN-supervised $\text{BW}_{\text{pred}}$ when network load varies from 0.1 ~ 0.8. These $D_{\text{DBA}}$ values are significantly lower than those of the classic and predictive DBA schemes as indicated in Fig.
In summary, DNN-selected BW\textsubscript{pred} values as shown in Fig. 5.8 are adaptive to multiple network features. Compared to the classic and predictive DBA schemes discussed, allocating these BW\textsubscript{pred} in BW\textsubscript{grant} = min\{BW\textsubscript{req}+BW\textsubscript{pred}, BW\textsubscript{max}\} effectively prevents bandwidth under- or over-granting, thereby reducing D\textsubscript{DBA}. Results in Figs. 5.6 and 5.8 highlight the capability of the trained DNN in characterising the dependency of D\textsubscript{HCF} and D\textsubscript{DBA} with multiple bandwidth allocation decision and network feature parameter, and in turn, achieving optimal bandwidth allocation decisions corresponding to diverse networks.

**5.4.3 A Discussion on the Execution of DNN-supervised Bandwidth Decisions**

Note that using the trained DNN, the optimal bandwidth decision values highlighted by the blue curves in Fig. 5.8 can be derived using common optimisation methods such as greedy search and Newton’s method. As our primary objective is to use the trained DNN to critically address how bandwidth allocation decisions impact the latency, we adopt the greedy search method to obtain the optimal BW\textsubscript{pred} values that minimise D\textsubscript{DBA}. Specifically, we feed the trained DNN using all plausible BW\textsubscript{pred}, with a granularity of 64 bytes, in the inputs. The corresponding D\textsubscript{DBA} outputs are plotted in Fig. 5.8, and the optimal BW\textsubscript{pred} values are hence obtained. In this way, D\textsubscript{DBA} associated with different BW\textsubscript{pred} can be fully understood. To reduce the D\textsubscript{DBA} and thereby the D\textsubscript{E2E}, the optimal BW\textsubscript{pred} values supervised by the trained DNN should be allocated.

From the above analyses, the optimal T\textsubscript{TXOP} in WLANs and BW\textsubscript{pred} in the PON can be obtained by minimising D\textsubscript{HCF} and D\textsubscript{DBA} using the trained DNN, respectively. In this study, we consider that supervised training and optimising T\textsubscript{TXOP} and BW\textsubscript{pred} are performed initially before network operation commences. For a given heterogeneous network configured by N number of SmartBANs, traffic load l\textsubscript{WLAN} in a WLAN, and M number of ONU-APs and distance d\textsubscript{O2C} in a PON, we perform greedy search using the trained DNN for the optimal T\textsubscript{TXOP} and BW\textsubscript{pred} value that minimises D\textsubscript{HCF} and D\textsubscript{DBA}, respectively. Specifically, note that analytical results show that BW\textsubscript{pred} does not impact the D\textsubscript{HCF} in WLAN. Using the trained DNN, we first fix BW\textsubscript{pred} = 0 in the inputs to search the optimal T\textsubscript{TXOP}, and then the optimal BW\textsubscript{pred} is searched as discussed above. Finally, in network
Chapter 5 Deep Neural Network Supervised Bandwidth Allocation Decisions for Low-Latency Heterogeneous E-health Network

operation, the optimal $T_{TXOP}$ is set in WLAN HCF and the optimal $BW_{pred}$ is adopted in DBA, in which $BW_{grant} = \min\{BW_{req} + BW_{pred}, BW_{max}\}$ is allocated to each ONU-AP in each polling cycle.

In the following section, we implement simulation to validate the effectiveness of allocating such DNN-supervised bandwidth values as discussed above in reducing the latency over the heterogeneous network. The latency performances corresponding to allocating these DNN-supervised bandwidth values, i.e., $T_{TXOP}$ and $BW_{pred}$, are evaluated and validated using packet-level simulations.

5.5 Performance Evaluations

5.5.1 Network Simulation Parameters

The end-to-end latency $D_{E2E}$ resultant from allocating the DNN-supervised bandwidth allocation decisions are evaluated using packet-level simulations in MATLAB. We apply the knowledge of bandwidth allocation decisions, i.e., $T_{TXOP}$ and $BW_{pred}$ that reduce uplink latency to a 16-ONU-AP heterogeneous network. The number of SmartBAN users $N$ in a WLAN varies from 2 ~ 8. Uplink latency performances, $D_{HCF}$, $D_{DBA}$, and $D_{E2E}$, are evaluated when $l_{network}$ varies from 0.1 ~ 0.8 (corresponding to $l_{WLAN}$ from 0.04 ~ 0.2 in individual WLANs). Figs. 5.9 to 5.11 presents $D_{HCF}$, $D_{DBA}$, and $D_{E2E}$ under varying network loads. Specifically, the $D_{HCF}$ arising from DNN-supervised bandwidth allocation decisions is compared against existing bandwidth allocations with the shortest TXOP/SI setting and the TXOP/SI setting in IEEE 802.11 reference HCF design [86]. Likewise, the $D_{DBA}$ arising from our trained DNN is compared against existing baseline limited-service DBA and predictive DBA schemes using statistical prediction algorithm in (2.6). Finally, the overall $D_{E2E}$ experienced by a packet generated by a SmartBAN user across a heterogeneous network is evaluated.

5.5.2 End-to-End Performance Improvements

In Fig. 5.9, in the shortest TXOP/SI, an ONU-AP allocates $T_{TXOP}$ duration for one packet transmission, thereby providing the shortest SI for each SmartBAN. However, when network load increases, more than one packet is expected to be generated and buffered within an SI. The allocated $T_{TXOP}$ is therefore insufficient
to transmit all arrivals, leading to high queuing latency and a saturated WLAN as shown in Fig. 5.9. Compared to the shortest TXOP/SI, the reference TXOP/SI proposed in the standard linearly increases $T_{TXOP}$ with increasing traffic load, which prevents the saturation in the shortest TXOP/SI. Note that in Fig. 5.9, the $D_{HCF}$ curves in this case exhibit inflexion points. This is because, when the increased $T_{TXOP}$ is less than the transmission time of one packet, $D_{HCF}$ will not be improved and remains the same as that of the shortest TXOP/SI. Compared to bandwidth allocations by shortest and reference TXOP/SI, assigning the DNN-supervised $T_{TXOP}$ values reduces $D_{HCF}$ corresponding to different $N$ in network. As shown in Fig. 5.9, $D_{HCF}$ is effectively reduced and constrained below 100 μs across all the network loads under consideration.

Moreover, $D_{HCF}$ is lower when fewer SmartBANs share the bandwidth in a WLAN, as illustrated by different $N$ in Fig. 5.9. However, since under a fixed $l_{network}$, a smaller $N$ indicates a higher traffic load contributed by individual SmartBANs, a larger $T_{TXOP}$ needs to be allocated in a WLAN. As such, it is shown in Fig. 5.9 at the saturation in shortest TXOP/SI and inflexion in reference TXOP/SI occur as $l_{network} = 0.4$ when $N = 2$, whereas, similar situations are observed until $l_{network}$ reaches 0.5 when $N = 4$ and 6. In comparison, allocating the $T_{TXOP}$ values optimised from the trained DNN in WLAN effectively prevents $D_{HCF}$ increasing with increasing traffic load. This is because the DNN is trained with multiple network features taken into account. Regarding different $N$, adaptive $T_{TXOP}$ values ensure sufficient bandwidth is allocated to SmartBANs in the WLAN. This avoids the latency deviation observed in the shortest and reference TXOP/SI in Fig. 5.9.

Fig. 5.10 compares the $D_{DBA}$ with DNN-supervised BW$_{pred}$ and the limited-service DBA. For this purpose, we have considered ONU-to-CO optical fiber lengths of 5 km, 10 km, and 15 km. As shown in Fig. 5.10, compared to limited-service DBA that is affected by RTT severely, allocating DNN-supervised BW$_{pred}$ values in DBA operation effectively limits $D_{DBA}$ deterioration with increasing $d_{O2C}$ and achieves a $D_{DBA} < 100$ μs. This is again because DNN adaptively adjusts and selects bandwidth values that suit varying $d_{O2C}$, thereby overcoming the RTT limitation and improving channel utilisation. Fig. 5.11 compares $D_{E2E}$ between allocating DNN-supervised $T_{TXOP}$ and BW$_{pred}$, and existing
bandwidth allocation over heterogeneous WLAN + PON networks. In existing bandwidth allocation, i.e., the reference TXOP/SI and limited-service DBA are adopted in WLAN and optical access network, respectively. As shown, using the reference TXOP/SI and limited-service DBA scheme in WLAN and PON leads to a $D_{E2E} > 200$ μs. In contrast, allocating the DNN-supervised bandwidth decisions achieves $D_{E2E}$ of 100 ~ 200 μs across 0.1 ~ 0.8 network load. This improvement can be attributed to the optimised bandwidth allocation decisions, i.e., $T_{TXOP}$, and

---

**Figure. 5.9** $D_{HCF}$ comparisons with shortest TXOP/SI and reference TXOP/SI.

**Figure. 5.10** $D_{DBA}$ comparisons with limited-service DBA.
5.6 Summary

This chapter has considered integrated PON and WLAN to support converged application delivery from SmartBAN users to remote e-health servers/clinicians. To address the uplink latency bottleneck caused by the uplink bandwidth contention, bandwidth resource allocation solutions in both WLAN and PON were comprehensively investigated and improved. In this chapter, we have critically analysed the end-to-end latency, including $D_{HCF}$, $D_{DBA}$ and $D_{E2E}$, over heterogeneous PON and WLAN, and highlighted crucial bandwidth allocation decision parameters, i.e., $T_{TXOP}$, $T_{CP}$, $T_{SI}$ in WLAN HCF, and $BW_{pred}$ in DBA schemes in the PON. In order to fully understand how these bandwidth allocation decision parameters, together with multiple network feature parameters, impact $BW_{pred}$, for both the WLAN HCF and the DBA in the PON. Moreover, it can be observed that using existing schemes lead to a $D_{E2E}$ deviation of more than 200 $\mu$s, this due to the adoption of the reference TXOP/SI in the WLAN segment. In comparison, executing the DNN-supervised bandwidth allocation decisions effectively alleviates this deviation in $D_{E2E}$, achieving a 40% – 70% $D_{E2E}$ reduction for a heterogeneous network.

**Figure. 5.11** $D_{E2E}$ performance comparison ($d_{O2C} = 10$ km).
on the latency performances, we have presented our exploitation of a DNN to characterise this complex dependency.

Specifically, based on the analysis of $D_{HCF}$, $D_{DBA}$ and $D_{E2E}$, we illustrated the selection of DNN input features and output targets, and presented the details of the supervised training. In supervised training, the DNN was trained to learn the association between key bandwidth allocation decisions and $D_{HCF}$, $D_{DBA}$, $D_{E2E}$. When supervised training was completed, we showed that the DNN can accurately predict $D_{HCF}$, $D_{DBA}$, $D_{E2E}$ corresponding to various bandwidth allocation decisions considered. Then, using the trained DNN, impacts of multiple bandwidth decision parameters and network features have on latency were critically analysed in this chapter. The selections of $T_{TXOP}$, $T_{CP}$ in WLAN HCF and $BW_{pred}$ in PON DBA to reduce $D_{HCF}$, $D_{DBA}$, and $D_{E2E}$ were comprehensively evaluated. Based on these analyses, insights into the latency performance of existing bandwidth allocation schemes, i.e., the shortest TXOP/SI, referent TXOP/SI in WLAN HCF and classic and predictive DBA schemes were provided. Analytical results have shown that existing schemes have not made the best bandwidth allocation decisions in terms of reducing latency, since existing schemes determined the bandwidth values to be allocated mainly relying on the traffic load. To improve latency performance of a network, multiple network features, such as traffic load, the number of SmartBANs in a WLAN, the number of ONU-APs and the ONU-to-CO distance, need to be considered in determining key bandwidth decisions parameters for WLAN HCF and PON DBA operation. In this regard, the trained DNN explicitly indicated the optimal values of the above-discussed decision parameters in minimising latency.

Specifically, with the trained DNN, we applied the heuristic greedy search method to derive the optimal $T_{TXOP}$ and $BW_{pred}$ values corresponding to the network feature $N$, $l_{WLAN}$, $M$ and $d_{O2C}$. Both the supervised training and search for the optimal $T_{TXOP}$ and $BW_{pred}$ were performed offline before network operation commencing. Then, we implemented extensive network simulations to validate the effectiveness of using these DNN-supervised $T_{TXOP}$ and $BW_{pred}$ decisions in WLAN HCF and PON DBA operation to reduce $D_{HCF}$, $D_{DBA}$, and ultimately $D_{E2E}$. Simulation results showed that compared to the use of reference TXOP/SI in WLAN and limited-service DBA scheme in PON, $D_{HCF}$, $D_{DBA}$, and $D_{E2E}$ were
effectively reduced under different network scenarios by using the DNN-supervised $T_{\text{TXOP}}$ and $BW_{\text{pred}}$ in allocating bandwidth in WLAN and PON. Less than 100 $\mu$s latency was achieved in each network segment, and an overall end-to-end latency of less than 200 $\mu$s was ensured under a network load of 0.1 – 0.8. Moreover, the latency deviation as seen under existing schemes was alleviated by allocating the DNN-supervised $T_{\text{TXOP}}$ and $BW_{\text{pred}}$.

In summary, with Chapters 3 and 4 focusing on SmartBANs for the next-generation e-health, this chapter comprehensively investigated the end-to-end communication between individual SmartBANs and remote servers/clinicians supported by a heterogeneous PON and WLAN network. Note that the SmartBAN applications is only a subset of emerging IoT and Tactile Internet applications. Our future access networks will need to support diverse new H2M and tactile-haptic applications. As such, from the next chapter onwards, we discuss the potential traffic pattern of aggregated H2M applications and delve into the latency performance and bandwidth resource allocation strategy in heterogeneous optical and wireless access networks in supporting aggregated low-latency H2M applications.
Chapter 6

Machine Learning based Bursty Bandwidth Prediction for Low-Latency H2M Applications

6.1 Introduction

In the previous chapters, we have studied wireless body area networks, and heterogeneous optical and wireless local access networks in support of latency-sensitive healthcare applications. This chapter aims to examine and improve the latency performance of heterogeneous access networks in delivering aggregated human-to-machine/robot (H2M) applications. Similar to the heterogeneous network architecture studied in the previous chapter, the underlying access network architecture for H2M communications is illustrated in Fig. 6.1. By strategically placing H2M servers, database, cloudlets at optical and wireless access networks, data processing and the delivery of H2M applications can be expedited to meet the millisecond latency requirement [1-3]. As shown in Fig. 6.1, bidirectional H2M packets traverse different segments of the Internet including passive optical access networks, the integrated optical network unit and wireless access points (ONU-AP), and wireless local area networks (WLANs). To reduce the latency, data processing and control/actuation feedback delivery can be done at the H2M server and/or the edge cloudlets [4]. State-of-the-art literature has reported that an end-to-end latency within 1 – 10 ms over such a heterogeneous network must be adhered to for effective interaction in H2M applications [11,12]. In particular, a stringent 1-ms latency is required in transmitting tactile-haptic data in an immersive communication environment. Subtracting the latency budgets in wireless local and personal area networks, the latency in the optical access network segment can be only up to a few hundreds of microseconds [8].
Chapter 6 Machine Learning based Bursty Bandwidth Prediction for Low-Latency H2M Applications

Currently, the development of H2M application is still in infancy, and diverse H2M applications have not been widely practised in daily life. Several pioneering studies on H2M applications and haptic communications have demonstrated that H2M traffic generated in H2M applications exhibits bursty pattern with alternate ON and OFF intervals [31-33], which is very similar the burstiness of Internet traffic [124-126]. For example, in an ON interval, consecutive packets generated by end devices/robots are wirelessly transmitted through the WLANs and aggregated at ONU-APs, whereas no packet is generated in the OFF interval. While downlink transmissions in Fig. 6.1 are typically broadcast, the latency bottleneck mainly lies in the uplink transmission due to bandwidth contention amongst multiple ONU-APs. To reduce the latency in supporting converged H2M applications delivery over heterogeneous access networks, uplink bandwidth resource allocation decision made by the central office (CO) considering the bursty pattern of H2M traffic aggregated at the ONU-APs is critical.

As introduced in the previous chapters, uplink bandwidth allocation in optical access networks is mainly realised by implementing dynamic bandwidth allocation (DBA) schemes at the central office (CO). Compared to the classic DBA schemes that allocate bandwidth based on the bandwidth request from ONU-APs, predictive DBA schemes incorporate traffic prediction to estimate and pre-allocate bandwidth. As such, predictive DBA schemes could help reduce uplink latency since packets can be transmitted without needing to be reported by an ONU-AP. Recall that in predictive DBA schemes, the CO allocates $BW_{grant} = \min\{BW_{req} + BW_{pred}, BW_{max}\}$ to each ONU-AP based on its requested bandwidth $BW_{req}$, and the

Figure 6.1 Network architecture that supports H2M communications.
addition bandwidth $BW_{\text{pred}}$, which is a bandwidth estimation based on short-term packet arrival prediction. For bursty traffic, this estimation of $BW_{\text{pred}}$ is critical and remains an open challenge in improving the performance of a predictive DBA scheme. First of all, the effectiveness and performance of a predictive DBA is dependent on its packet arrival prediction accuracy [15]. Accurately predicting the arrival of bursty traffic is necessary to reduce uplink latency [14]. Therefore, it is imperative that traffic prediction methods and the $BW_{\text{pred}}$ estimation of bursty traffic is investigated in order to reduce the uplink latency of networks that support H2M traffic. Further, considering the ON/OFF alternations, it is anticipated that the bandwidth allocated by the CO should be in accordance to uplink traffic characteristics and load [16]. For highly bursty H2M traffic, the ability of DBA schemes in allocating bandwidth during bursty intervals and preserving bandwidth during long-idle intervals is key to reducing uplink latency.

As overviewed in Chapter 2 Section 2.3.2, the most-recently (MR) DBA and limited sharing DBA with traffic prediction (LSTP-DBA), are two typical predictive DBA schemes for bursty traffic [128,129]. MR-DBA scheme uses the most-recently received packets from the previous polling cycle to estimate $BW_{\text{pred}}$. LSTP-DBA scheme uses a 4-order autoregressive model, i.e., counting on the arrival packets in the previous 4 polling cycles, to estimate $BW_{\text{pred}}$. Simulation results show that MR-DBA and LSTP-DBA outperform classic DBA schemes regarding the uplink latency performance. Recently, a machine learning technique (ML), $k$-nearest neighbourhod ($k$NN), was used to evaluate $BW_{\text{pred}}$ for video streams. The proposed scheme, named DAMA (data mining forecasting DBA), estimates $BW_{\text{pred}}$ by averaging the received packets in $k$ number of past polling cycles that have similar durations to the current cycle [117]. MR-DBA, LSTP-DBA and DAMA predict $BW_{\text{pred}}$ based on the packet arrivals in the near past. How accurate they predict the packet arrivals and estimate bandwidth demand has yet to be investigated. Moreover, note that even bursty arrivals can be accurately predicted by the CO, allocating additional bandwidth $BW_{\text{pred}}$ to ONU-APs will not always be effective in reducing uplink latency. This is because overwhelmingly long packet queues may be accumulated at ONU-APs due to packet bursts. However, the available bandwidth that can be allocated to an ONU-AP is restricted by $BW_{\text{max}}$. This $BW_{\text{max}}$ is adopted in both classic and predictive DBA schemes to prevent bandwidth monopolisation by heavily-loaded
ONU-AP. With BW_{max}, the fairness among ONU-APs is ensured, however, the CO will not be able to cater ONU-APs that demand more bandwidth than BW_{max} due to packet bursts. To address this bottleneck, identifying idle ONU-APs that are in OFF intervals and reserve bandwidth from these ONU-APs is critical.

In light of the above, we exploit the use of an artificial neural network (ANN) at the CO to predict the ON and OFF status of bursty H2M traffic arriving at each ONU-AP. Prediction of the ON/OFF status consequently enables the evaluation of the uplink bandwidth demand of each ONU-AP. This is different from the above-mentioned predictive DBA schemes that directly estimate BW_{pred} based on the number of received packets in the past polling cycles. In justifying prediction performance, we present a comparative study of ANN versus existing bursty traffic prediction algorithms used in predictive DBA schemes. Through simulations, we first show that our trained ANN can achieve superior accuracy (> 90%) in predicting the ON and OFF status of ONU-APs, and therefore effectively improving the accuracy of the bandwidth demand estimation. The proposed ANN prediction is then implemented at the CO to facilitate bandwidth allocation decisions through the proposed machine learning based predictive DBA (MLP-DBA) scheme. In contrast to the baseline limited-service DBA scheme and predictive DBA schemes that grant requested/estimated bandwidth to all ONU-APs, the proposed MLP-DBA scheme classifies each ONU-AP based on (a) the predicted ON and OFF status of ONU-APs and (b) the estimated bandwidth demand. Then, the CO adaptively allocates bandwidth to each ONU-AP. In this chapter, we show that with ANN prediction, the proposed MLP-DBA scheme successfully improves uplink latency performance and reduces the packet drop ratio as compared to the limited-service DBA, MR- and LSTP-DBA schemes.

The rest of this chapter is organised as follows. In Section 6.2, self-similarity and long-range dependence of bursty traffic and the synthetic generation for H2M traffic are introduced. Our ANN-based ON/OFF status prediction is proposed in Section 6.3. Section 6.4 details the proposed MLP-DBA scheme. Analytical and simulation results and discussions are presented in Section 6.5. A summary of the main contributions of this paper is presented in Section 6.6.

6.2 Bursty Traffic Characteristic and Generation
Studies on network traffic have shown that bursty user traffic can be characterised by self-similarity and long-range dependence (LRD) [124-126]. Self-similarity is a natural phenomenon that an object is similar to a part of itself. In statistics, a self-similar process, \( \{X(t)\} \), is described as follows:

\[
\{X(at)\} = \{a^H X(t)\}
\]

(6.1)

where \( a \) is a scale factor and \( H \) is known as the Hurst parameter that measures the dependence over time. (6.1) indicates that for a self-similar process the same pattern can be observed at different time scale. The aggregated user traffic over the Internet exhibits self-similarity over time and is modelled by long-trail distributions. The Pareto distribution has been proven to fit various practical traffic traces, and is a widely-accepted model to generate synthetic network traffic [42]. In particular, a Pareto distribution with Hurst parameter within 0.5 – 1 possesses the LRD property [126]. A large \( H \) in traffic model means stronger LRD and hence the burstiness of the traffic [127].

For H2M applications, there is a lack of traffic profiles of aggregated H2M applications over heterogeneous networks since a majority of current H2M research is still confined to dedicated systems for specific tasks/experiments. Nevertheless, on-going experiments have clearly shown an ON/OFF pattern in

![Figure. 6.2 Bursty traffic patterns of master control and slave feedback in a teleoperation session (adopted from [32]).](image-url)
individual H2M tasks similar to that in content-centric end-user applications as presented in Fig. 6.2. Given such ON/OFF sources, research in [204] has confirmed that aggregation of packets, protocols, bitrate mechanisms and packet/flow level controls do not impact the LRD characteristic. A primary consideration remains to estimate $H$ parameter regarding actual user applications. As such, we adopt the Pareto-traffic model in this chapter to synthesize the bursty traffic ON/OFF intervals of H2M applications. Using synthetic traffic sources, we look into the performance of existing DBA schemes in supporting aggregated bursty traffic. Traffic with different levels of burstiness, characterised by different Hurst parameter values, is generated and investigated in this chapter. Details on the synthetic traffic generation are presented below.

The probability density function (pdf) of the Pareto distribution is given as follows:

$$f(x) = \frac{ab^a}{x^{a+1}}, \quad x \geq b$$

(6.2)

where $a$ is a shape parameter and $b$ is a location parameter. The Hurst parameter of a Pareto distribution can be calculated by $H = (3-a)/2$. A Pareto-distributed interval can be generated by:

$$X = \frac{b}{U^{1/a}}$$

(6.3)

where $U$ is a uniform random variable within the range (0,1]. By setting different $b$ and $a$ values, traffic source with different levels of burstiness can be generated. For example, in generating an ON interval, the default $b$ value is 1, and the measured $a$ value based on Internet traffic traces is 1.4, corresponding to $H = 0.8$. Note that the relationship of ON and OFF interval, termed $X_{ON}$ and $X_{OFF}$, with the traffic load, $l$, is:

$$l = \frac{E(X_{ON})}{E(X_{ON}) + E(X_{OFF})}$$

(6.4)

Using (6.4), the $b$ and $a$ values used to generate a Pareto-distributed OFF interval can be derived. Finally, by synthesizing alternating ON and OFF intervals, a single traffic stream can be generated.
6.3 Proposed Artificial Neural Network for Bandwidth Prediction

6.3.1 Traffic Features in Predictive DBA Operation

Since bursty traffic is featured by alternating ON and OFF intervals, we propose to exploit an ANN to predict the ON and OFF status of ONU-APs, that is whether an ONU-AP is experiencing packet bursts in individual polling cycles. Then based on this ON and OFF status prediction, bandwidth demand of an ONU-AP can be estimated. In order to realise the ON and OFF status prediction, we first look into relevant traffic features that could be used by the ANN during DBA operation.

Fig. 6.3 illustrates the traffic features seen by the CO during predictive DBA operation. Let us denote \( T_{\text{poll}}(i, j) \) as the duration of the \( i \)-th polling cycle of ONU-AP \( j \). As illustrated, in the \( i \)-th polling cycle, ONU-AP \( j \) transmits \( k(i, j) \) bytes, or equivalently \( n(i, j) \) number of packets, in the granted transmission duration and reports the remaining queue length, denoted as \( BW_{\text{req}}(i, j) \), in the REPORT message to the CO. Hence, \( BW_{\text{req}}(i, j) \) is the requested bandwidth by ONU-AP \( j \). The CO then allocates bandwidth, \( BW_{\text{grant}}(i+1, j) \), by sending a GATE message that indicates the next transmission start time and duration of ONU-AP \( j \). The duration of the \((i+1)\)-th polling cycle, \( T_{\text{poll}}(i+1, j) \), is updated at the CO.

Figure. 6.3 Predictive DBA operation and traffic features.
Upon receiving the GATE message from the CO, ONU-AP \( j \) waits for the next transmission start time which it then transmits packets for a duration equivalent to \( BW_{\text{grant}}(i+1, j) \).

As shown in Fig. 6.3, while waiting for the transmission start time in the \((i+1)\)-th polling cycle, an amount of \( a(i+1, j) \) bytes is received by ONU-AP \( j \) for uplink transmission. To reduce uplink latency, the CO predicts \( a(i+1, j) \), or the equivalent bandwidth \( BW_{\text{pred}}(i+1, j) \). The total bandwidth, \( BW_{\text{grant}}(i+1, j) \), that is allocated by the CO to ONU-AP \( j \), should therefore be:

\[
BW_{\text{grant}}(i+1, j) = \min\{BW_{\text{req}}(i, j) + BW_{\text{pred}}(i+1, j), BW_{\text{max}}\} \tag{6.5}
\]

With \( BW_{\text{pred}}(i+1, j) \) allocated, \( a(i+1, j) \) bytes arriving during the \((i+1)\)-th polling cycle can be transmitted within the polling cycle without needing to be reported to the CO, thereby reducing latency. However, the exact \( a(i+1, j) \) is not known apriori and so accurately predicting \( a(i+1, j) \), i.e., \( BW_{\text{pred}} \), is critical to improving uplink latency performance. Let us now denote the prediction error as:

\[
e(i+1, j) = |BW_{\text{pred}}(i+1, j) - a(i+1, j)| \tag{6.6}
\]

### 6.3.2 ANN-based Bandwidth Prediction Model

As introduced earlier, when considering bursty traffic, it is critical for the CO to identify if an ONU-AP is currently in an ON interval, i.e., the ONU-AP is receiving packets aggregated from its WLAN, or otherwise in an OFF interval, i.e., an idle interval with no arriving packet bursts. Erroneously allocating additional bandwidth to ONU-APs that are in their OFF intervals increases polling cycle time, and thereby increases the uplink latency of the packets at those ONU-APs.

Hence, we use an ANN to predict the ON/OFF status of ONU-AP \( j \) and estimate \( BW_{\text{pred}}(i+1, j) \) from this prediction. Fig. 6.4 illustrates a schematic of our proposed ANN comprising an input layer, an output layer and some hidden layers in between. Each layer is composed of neuron units, which are non-linear functions that map the associated inputs to an output. When receiving the REPORT from ONU-AP \( j \) in the \( i \)-th polling cycle, the CO uses available traffic features, as listed in Table 6.1 to predict the ON/OFF burst status of ONU-AP \( j \) in the \((i+1)\)-th polling cycle. With \( x(i, j) = \{k(i, j), n(i, j), a(i, j), BW_{\text{req}}(i, j), T_{\text{poll}}(i, j), \} \),
6.3 Proposed Artificial Neural Network for Bandwidth Prediction

Figure 6.4 Schematic of the proposed ANN architecture to predict the ON/OFF status \( y \) of the bursty traffic at ONU-AP \( j \).

Table 6.1 Traffic features collected in the \( i \)-th polling cycle.

<table>
<thead>
<tr>
<th>Notations</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k(i, j) )</td>
<td>The number of received bytes from ONU-AP ( j )</td>
</tr>
<tr>
<td>( n(i, j) )</td>
<td>The number of received packets from ONU-AP ( j )</td>
</tr>
<tr>
<td>( a(i, j) )</td>
<td>The number of arrival bytes at an ONU-AP ( j )</td>
</tr>
<tr>
<td>( BW_{\text{req}}(i, j) )</td>
<td>The remaining buffer length of ONU-AP ( j )</td>
</tr>
<tr>
<td>( T_{\text{poll}}(i, j) )</td>
<td>The duration of the ( i )-th polling cycle</td>
</tr>
<tr>
<td>( T_{\text{poll}}(i+1, j) )</td>
<td>The duration of the ((i+1))-th polling cycle</td>
</tr>
</tbody>
</table>

\( T_{\text{poll}}(i+1, j) \) as the input traffic features to the ANN in the \( i \)-th polling cycle, the target output \( y \) is denoted in (6.7) as follows:

\[
y(i+1, j) = \begin{cases} 
1, & \text{ONU-AP } j \text{ is ON in } (i+1)\text{-th polling cycle} \\
0, & \text{ONU-AP } j \text{ is OFF in } (i+1)\text{-th polling cycle} 
\end{cases} 
\]  
(6.7)

Recall that \( BW_{\text{pred}}(i+1, j) \) is the estimated bandwidth of \( a(i+1, j) \) at ONU-AP \( j \). Considering lastmile compatibility \[13\] and the ON/OFF status \( y(i+1, j) \) predicted by the ANN, \( BW_{\text{pred}}(i+1, j) \) can be estimated using:

\[
BW_{\text{pred}}(i+1, j) \approx y(i+1, j)T_{\text{poll}}(i+1, j)R_{\text{WLAN}} 
\]  
(6.8)
\( R_{WLAN} \) is the data rate of the WLAN associated with ONU-AP \( j \). Once the ON/OFF status \( y(i+1, j) \) is predicted, and the \( BW_{\text{pred}}(i, j) \) is estimated using (6.8), the CO evaluates the total bandwidth demand and allocates \( BW_{\text{grant}}(i+1, j) = \min\{BW_{\text{req}}(i, j) + BW_{\text{pred}}(i+1, j), BW_{\max}\} \), of ONU-AP \( j \) in \((i+1)\)-th polling cycle.

### 6.3.3 Supervised Training

Prior to using the ANN to predict the ON/OFF status of ONU-APs, we must first train the ANN using training sets collected during the training phase of network operation. A training set, \( S \), is represented as \{\((x(i, j), y(i+1, j))\) | for \( i = 1, 2, \ldots, N, j = 1, 2, \ldots, M\)\} with \( x(i, j) \) as the traffic features collected by the CO in the \( i \)-th polling cycle and \( y(i+1, j) \) as the labelled output. The parameter \( N \) is the total number of polling cycles, and \( M \) is the number of ONU-APs. Here, we generate training sets using event-driven packet-level simulations of 16- and 32-ONU-AP optical access networks. The data rate of the optical access network and WLANs are 1 Gbps and 100 Mbps, respectively. Traffic features from 250,000 polling cycles under various network loads at each ONU-AP are collected and then is utilized for supervised training.

After supervised training with the collected training set, the final architecture of the ANN is shown in Fig. 6.5 whereby in addition to the input and output layers, there is one hidden layer of ten neurons. The optimal weight matrix for each layer is determined using the gradient descent method. Since prediction accuracy does not improve with further layers and neurons, this architecture is used as a final architecture in subsequent investigations. Note that once the training phase is complete, the CO only needs to store the weight matrix to map the input traffic features to output the predicted ON/OFF status (6.7), estimate \( BW_{\text{pred}} \) in (6.8) and evaluate \( BW_{\text{grant}} \) in (6.5).

To verify the predict performance of the trained ANN, we use a test set, \( T \), with another 250,000-cycle traffic features, to investigate the accuracy of its ON/OFF status prediction. Fig. 6.5(a) plots the ON/OFF prediction accuracy for the 16 and 32 ONU-AP networks as a function of aggregated network load, highlighting that the trained ANN achieves high prediction accuracy for network loads from 0.1 to 1. The prediction accuracy is > 96% under light and high network loads due to the ONU-APs having longer idle (OFF) and bursty (ON) intervals, respectively. By
comparison, the prediction accuracy decreases under moderate traffic load, i.e., 0.4 – 0.7. This behaviour can be attributed to the frequent alternation of ON and OFF intervals. To investigate further, we plot the receiver operating characteristic (ROC) curves at a network load of 0.6 (Fig. 6.5(b)). Here, the true positive rate is the percentage of correct ON predictions in all ON samples, and the false positive rate is the percentage of mis-predicting OFF over all OFF samples. Note that the curves approach (0,1), indicating that the trained ANN achieves effective prediction for both ON and OFF status. Also, the trained ANN achieves better prediction performance for the 32-ONU-AP network, as also
indicated in Fig. 6.5(a). Considering the same network load, each individual ONU-AP in the 32-ONU-AP network contributes to a lower network load than an individual ONU-AP in a 16-ONU-AP network.

In Fig. 6.6, we investigate the dependency of the ON/OFF prediction accuracy on the training set size, Hurst parameter variation in the Pareto traffic model, and an exponential traffic model for a 32-ONU-AP. In Fig. 6.6(a), results for various network loads at 0.3, 0.6 and 0.9 are shown for illustrative purposes. As

(a) Training set size variation

(b) Traffic model variation

**Figure. 6.6** Evaluation of sensitivity to training set size and model variation.
can be observed, increasing the training set size improves prediction accuracy. Further, the prediction accuracy varies under different network traffic loads. As shown in Fig. 6.6(a), network loads of 0.3 and 0.9 require a smaller training set to reach > 90% accuracy whereas a larger training set is required for the network load of 0.6 to achieve the same prediction accuracy. For the network loads considered, prediction accuracy is low when the training set is small due to an inadequate number and imbalance of ON/OFF samples. Specifically, as each time of packet burst at each ONU-AP differentiates significantly, a small training set only containing features of a few ON/OFF intervals cannot be well generalized to predict a test set. When samples are enough to include features from enough ON/OFF alternations at all 32 ONU-APs, a clear rise in prediction accuracy can be observed. Results show that overall, across the three network loads, a training set size with around $10^4$ samples is sufficient to train the ANN whereby any further increase in training set size would not yield further improvement in prediction accuracy.

Fig. 6.6(b) shows the sensitivity of the trained ANN. The trained ANN is first applied to predict the ON/OFF status of ONU-APs when the Hurst parameter varies from 0.6 – 0.9. A larger $H$ represents stronger LRD and traffic burstiness, indicating that consecutive long ON intervals are more likely to occur. Results in Fig. 6.6(b) show that the trained ANN achieves > 95% prediction accuracy regardless of $H$ variations in traffic. In comparison with LRD traffic, the trained ANN is also applied to traffic characterised by short-range dependence (SRD) with exponential ON/OFF intervals. As can be viewed in Fig. 6.6(b), the prediction accuracy (green bar) reduces, but > 90% prediction accuracy can still be achieved. Further, when re-training the ANN corresponding to the SRD traffic, the prediction accuracy (yellow bar) is effectively improved to 98%. Results in Fig. 6.6(b) validate the robustness of the trained ANN.

6.3.4 Bandwidth Estimation for Bursty Traffic

ON/OFF Status Prediction: ANN vs $k$-NN

In the DAMA scheme, ML technique, $k$-NN, has been considered for making bandwidth prediction, while the performance was only validated with video streams [117]. To evaluate the capability of the proposed ANN prediction for
Chapter 6 Machine Learning based Bursty Bandwidth Prediction for Low-Latency H2M Applications

bursty traffic, we first compare the use of the trained ANN and $k$-NN in [117] in predicting the ON/OFF status of ONU-APs during the DBA operation. The prediction accuracy of the proposed ANN-based prediction is plotted and compared against $k$-NN technique with $k = 1, 5, \text{and } 10$. Results in Figs. 6.7(a) and (b) highlight that the trained ANN achieves $> 90\%$ accuracy in predicting ON/OFF periods for both 16- and 32-ONU-AP networks. In contrast, the prediction accuracy of $k$-NN varies depending on the network load and is less effective than ANN. Particularly, the low prediction accuracy of $k$-NN is observed in the 0.6 to 0.8 network load region. This can be attributed to the
short and frequent change between ON and OFF periods. On the other hand, during low and high network loads, moderate prediction accuracy (> 60% in k-NN) is observed. This observed increase in prediction accuracy is due to the occurrence of long-OFF and -ON periods in low and high network loads, respectively.

The kNN-based prediction is significantly time-consuming, in tens to hundreds of seconds depending on the size of a training set, compared to using a trained ANN. This is attributed to the fact that in kNN any new input feature vector $x$ needs to be compared with all the existing samples in a training set to attain a
predicted output \( y \). It should be noted that a DBA polling cycle is in general in millisecond order. As such, we consider \( k \)-NN not a favourable technique to apply at the CO in making bandwidth allocation decisions. If \( k \)-NN needs to be used, the computation time and learning outcome must be critically evaluated and improved. In contrast, although an ANN needs to be trained first with training time varying in the order of seconds, the prediction can be done immediately within a few microseconds by a trained ANN. In the following, we utilize the proposed ANN-based prediction to estimate bandwidth demand and facilitate bandwidth allocation in predictive DBA operation.

**Bandwidth Demand Estimation**

With the trained ANN predicting the ON/OFF status of all ONU-APs using (6.7), and estimating \( BW_{\text{pred}} \) using (6.8), we evaluate the average prediction error:

\[
E_{\text{avg}} = \frac{1}{NM} \sum_{j=1}^{M} \sum_{i=1}^{N} e(i, j) = \frac{1}{NM} \sum_{j=1}^{M} \sum_{i=1}^{N} [BW_{\text{pred}}(i, j) - a(i, j)]
\] (6.7)

Figs. 6.8 (a) and (b) plots \( E_{\text{avg}} \) for a 16- and 32-ONU-AP network, respectively and compares the results obtained using our proposed ANN-based prediction method to that obtained using the MR-DBA and LSTP-DBA. Fig. 6.8 shows that \( E_{\text{avg}} \) increases with traffic load and that our proposed ANN-based prediction has the best performance, especially under heavy traffic loads. \( E_{\text{avg}} \) of the 32-ONU-AP network is about half of that of the 16-ONU-AP network. Again, this is attributed to the lower traffic load per individual ONU-APs in 32-ONU-AP network when considering the same aggregated network load.

**6.4 Proposed MLP-DBA for Bandwidth Allocation**

**6.4.1 Optical Network Units Classification**

The CO in an optical access network will typically allocate a bandwidth equivalent to \( \min\{BW_{\text{req}} + BW_{\text{pred}}, BW_{\text{max}}\} \) to ONU-APs during each polling cycle. When ONU-APs are heavily-loaded with long accumulated packet queues, \( BW_{\text{req}} \) exceeds \( BW_{\text{max}} \). For heavily-loaded ONU-APs, reducing the time interval between consecutive transmissions and increasing the allocated bandwidth would reduce the uplink latency of the arriving packets, but this must be done without
6.4 Proposed MLP-DBA for Bandwidth Allocation

Impacting the less heavily-loaded ONU-APs. In order to expedite the delivery of arriving packets, MLP-DBA enables the CO to classify ONU-APs based to their BW_{req} + BW_{pred} and allocate bandwidth accordingly, as described below.

Upon receiving a REPORT message from an ONU-AP, the CO predicts the ON/OFF status of the ONU-AP, i.e. y using the trained ANN, and estimates BW_{pred}. Then, the CO classifies each ONU-AP into one of 3 different classes following the principle listed in Table 6.2.

- If according to the train ANN, the status of an ONU-AP is ON in the next polling cycle, i.e., y = 1, and meanwhile BW_{req} + BW_{pred} < BW_{max}, the ONU-AP will be classified as **Class A**. For such an ONU-AP, the CO will

<table>
<thead>
<tr>
<th>Class</th>
<th>Classification Principle</th>
<th>Estimated BW_{pred}</th>
<th>Allocated Bandwidth, BW_{grant}</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>y = 1, and BW_{req} + BW_{pred} &lt; BW_{max}</td>
<td>T_{poll}R_{WLAN}</td>
<td>BW_{req} + T_{poll}R_{WLAN}</td>
</tr>
<tr>
<td>B</td>
<td>y = 0, and BW_{req} + BW_{pred} &lt; BW_{max}</td>
<td>0</td>
<td>BW_{req}</td>
</tr>
<tr>
<td>C</td>
<td>BW_{req} + BW_{pred} ≥ BW_{max} for both y = 0 and 1</td>
<td>yT_{poll}R_{WLAN}</td>
<td>BW_{max} and generating an additional GATE</td>
</tr>
</tbody>
</table>

**Figure. 6.9** An illustration of MLP-DBA procedure with three ONU-APs.
allocate through a GATE message a bandwidth equivalent to $BW_{grant} = BW_{pred} + BW_{req}$ for transmission in the next polling cycle, allowing arriving packets to be transmitted without having to be reported to the CO and without having to wait for the next transmission time.

- If the status of the ONU-AP is predicted to be OFF, i.e., $y = 0$, with $BW_{req} + BW_{pred} < BW_{max}$, it will be classified as Class B. The CO will allocate a Class-B ONU-AP a bandwidth that is equivalent to $BW_{req}$ since no packet is expected to arrive during its OFF interval.

- Heavily-loaded ONU-AP with a $BW_{req} + BW_{pred} > BW_{max}$ is classified as Class C. For such an ONU-AP, the CO will first allocate $BW_{max}$ through a GATE message. Further, to alleviate the number of queued packets, an additional GATE message is generated for the ONU-AP by the CO such that an additional transmission opportunity is given to the ONU-AP before scheduling the next polling cycle.

### 6.4.2 MLP-DBA Scheme for Bandwidth Allocation

To explain the operation of MLP-DBA for Class-A, B, and C ONU-APs more clearly, we present the MLP-DBA operation in Fig. 6.9 with 3 ONU-APs for illustrative purposes. As shown in Fig. 6.9, in the $(i-1)$-th polling cycle, ONU-AP 1, 2, and 3 are classified as Class A, C, and B, respectively. As such, in responding to their REPORTs, the CO first grants $BW_{grant}$ to ONU-AP 1, 2, and 3 in the $i$-th polling cycle, respectively. Moreover, since ONU-AP 2 is classified as Class C, at the beginning of the $i$-th polling cycle, an additional GATE message is generated for ONU-AP 2 by the CO such that an additional transmission opportunity is given to ONU-AP 2 before scheduling the $(i+1)$-th polling cycle. As such, the transmission interval for ONU-AP 2 is reduced as presented in the figure. As ONU-AP 3 is OFF with no packet arrivals expected, the additional transmission opportunity interval does not affect ONU-AP 3. Note that additional GATEs are generated and send at the beginning of each polling cycle based on the classification results in the previous cycle. As such, the CO does not respond to the REPORTs received during additional transmission opportunity intervals. The CO repeats this procedure during each polling cycle for all ONU-APs. The pseudocode of the MLP-DBA scheme is presented as follow:
MLP-DBA: Pseudocode of bandwidth allocation at the CO

$t_{\text{scheduled}}$ — time up to which the uplink channel has been scheduled
$RRT$ — round-trip transmission time
$T_g$ — guard time
$T_{\text{process}}$ — time for processing a GATE message at an ONU-AP
$T_{\text{REPORT}}$ — duration of a REPORT message
$\delta(i, j) \in \{A, B, C\}$ — class of ONU-AP $j$ in $i^{\text{th}}$ polling cycle
$t_{\text{start}}(i, j)$/$t_{\text{add\_start}}(i, j)$ — the start time of granted timeslot for ONU-AP $j$ in the $i^{\text{th}}$ polling cycle
$t_{\text{end}}(i, j)$/$t_{\text{add\_end}}(i, j)$ — the end time of granted timeslot for ONU-AP $j$ in the $i^{\text{th}}$ polling cycle
$R_{\text{PON}}$ — data rate in the optical network
$R_{\text{WLAN}}$ — data rate in the WLAN
$BW_{\text{add\_grant}}(i, j)$ — the granted bandwidth for ONU-AP $j$ in the additional opportunity if an additional GATE is generated for ONU-AP $j$ in the $i^{\text{th}}$ polling cycle
$\text{ANN}(\mathbf{x})$ — trained neural network with input feature vector $\mathbf{x}$

The operation for ONU-AP $j$ at the CO at the beginning of $i^{\text{th}}$ polling cycle (repeat for all $i$ and $j$):

```plaintext
1 { 
2 if $\delta(i, j) = C$ 
3 // schedule the start time in the $i^{\text{th}}$ additional cycle 
4 $t_{\text{add\_start}}(i, j) = \min\{\text{LocalTime} + RRT/2 + T_{\text{process}},$
5 $t_{\text{scheduled}} - RRT/2 - T_{\text{process}}\};$
6 $BW_{\text{add\_grant}}(i, j) = BW_{\max};$
7 $t_{\text{add\_end}}(i+1, j) = t_{\text{add\_start}}(i, j) + BW_{\max}/R_{\text{PON}} + T_{\text{REPORT}} + T_{\text{guard}};$
8 generate and send a GATE for $t_{\text{start}}(i + 1, j)$ and $BW_{\text{add\_grant}}(i, j);$ 
9 $t_{\text{scheduled}} = t_{\text{add\_end}}(i, j);$ 
10 else 
11 $BW_{\text{add\_grant}}(i, j) = 0;$ 
12 end 
13 }
```

The operation for ONU-AP $j$ at the CO upon receiving the REPORT in the $i^{\text{th}}$ polling cycle (repeat for all $i$ and $j$):

```plaintext
14 { 
15 // schedule the start time in the $(i + 1)^{\text{th}}$ polling cycle 
16 $t_{\text{start}}(i + 1, j) = \min\{\text{LocalTime} + RRT/2 + T_{\text{process}}, t_{\text{scheduled}} - RRT/2 - T_{\text{process}}\};$
17```
//ANN-based prediction and classification
\[ x_{i,j} = \{ k(i,j), n(i,j), a(i,j), BW_{req}(i,j), T_{POLL}(i,j), T_{POLL}(i+1,j) \}; \]
\[ y = \text{ANN}(x_{i,j}); \]
\[ BW_{pred}(i+1,j) = yR_{WLAN \text{start}(i+1,j)} - yR_{WLAN \text{start}(i,j)}T_{\text{start}(i,j)}; \]

// following the limited-service principle, BW_{grant} is bounded by BW_{max}
\[ BW_{grant}(i+1,j) = \min\{BW_{req}(i,j) + BW_{pred}(i+1,j) - BW_{\text{add}_{\text{grant}}}(i,j), BW_{max}\}; \]

// update time schedule and generate a GATE message
\[ t_{\text{end}}(i+1,j) = t_{\text{start}}(i+1,j) + BW_{\text{grant}}(i+1,j)/R_{\text{PON}} + T_{\text{REPORT}} + T_{\text{guard}}; \]
\[ t_{\text{scheduled}} = t_{\text{end}}(i+1,j); \]

The operation for ONU-AP j at the CO upon receiving the REPORT in the i-th additional opportunity (repeat for all i and j):

\[ \{ \]
\[ x_{i,j} = \{ k(i,j), n(i,j), a(i,j), BW_{req}(i,j), T_{POLL}(i,j), \}
\[ T_{POLL}(i+1,j); \]

\[ y = \text{ANN}(x_{i,j}); \]
\[ \text{update } \delta(i+1,j); \]
\[ } \]

### 6.5 Performance Evaluations

To verify the effectiveness of the proposed MLP-DBA scheme, we implemented packet-level simulations (MATLAB) of an optical access network which parameters are listed in Table 6.3. The bursty ON and OFF intervals follows a Pareto distribution with a Hurst parameter of 0.8, which best fits existing Internet traffic trace [42]. We analyse the ONU-AP classifications during MLP-DBA operation and compare the uplink latency and packet drop ratio performances of MLP-DBA with/without ONU-AP classification, MR-DBA, LSTP-DBA and the baseline limited-service DBA schemes. In the case of MLP-DBA without classification, BW_{pred} is estimated using the proposed ANN and (6.7) and (6.8) but bandwidth allocation follows the typical DBA operation whereby \( \min\{BW_{req} + BW_{pred}, BW_{max}\} \) is allocated to all ONU-APs. In contrast, MLP-DBA with classification utilizes the ANN to predict the ON/OFF status and estimate BW_{pred} of each ONU-AP and based on these results, classify and allocate
bandwidth to each ONU-AP accordingly.

### 6.5.1 Uplink Latency and Packet Drop Ratio

Figs. 6.10 (a) and (b) verify the effectiveness of MLP-DBA in predicting bandwidth demand and improving uplink latency performance. The uplink latency is defined as the duration between the time a packet arrives at a buffer of an ONU-AP until its uplink transmission. Latency constraints, $D_n = 1 \text{ ms}$ and $10 \text{ ms}$, are considered in supporting H2M interactions. Several observations can be highlighted. First of all, compared to the limited-service DBA, MR-DBA, LSTP-DBA and the proposed MLP-DBA without classification, the uplink latency performance of the MLP-DBA with classification is improved under light-to-moderate network loads, i.e., $0.1 – 0.5$ in Fig. 6.10(a) and $0.1 – 0.4$ in Fig. 6.10(b). This is attributed to the high accuracy of ON/OFF status prediction, $\text{BW}_{\text{pred}}$ estimation and the subsequent allocation of $\text{BW}_{\text{grant}} = \text{BW}_{\text{pred}} + \text{BW}_{\text{req}}$ to ensure that arriving packets can be transmitted without having to wait for the next transmission cycle. MLP-DBA without classification achieves better performance compared to MR-DBA and LSTP-DBA due to its superior ON/OFF status prediction performance of the ANN as compared to the most-recent and the 4-order autoregressive models, respectively. However, MR-DBA, LSTP-DBA and MLP-DBA without classification result in high uplink latency similar to that of the limited-service DBA scheme when network load is beyond $0.5$ in Fig. 6.10(a) and (b). This is because, under heavy network loads, packet bursts at ONU-APs

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_{\text{PON}}$</td>
<td>1 Gbps</td>
</tr>
<tr>
<td>$R_{\text{WLAN}}$</td>
<td>100 Mbps</td>
</tr>
<tr>
<td>Number of ONU's</td>
<td>16 and 32</td>
</tr>
<tr>
<td>CO to ONU-AP distance</td>
<td>10 km</td>
</tr>
<tr>
<td>$T_{\text{REPORT}}$</td>
<td>0.6 $\mu$s</td>
</tr>
<tr>
<td>$T_g/T_{\text{process}}$</td>
<td>1 $\mu$s</td>
</tr>
<tr>
<td>Packet length</td>
<td>$64 \sim 1518$ bytes</td>
</tr>
<tr>
<td>Buffer capacity</td>
<td>1000 packets within 1 M bytes</td>
</tr>
<tr>
<td>Maximum polling cycle</td>
<td>1 ms</td>
</tr>
</tbody>
</table>
tend to have shorter OFF intervals, leading to accumulated packet queues at ONU-APs. Consequently, existing predictive DBAs that allocates min\{BW_{req} + BW_{pred}, BW_{max}\} to ONU-APs are less effective when BW_{max} is allocated in each polling cycle and uplink latency is mainly attributed to multiple-cycle waiting time before a packet is allocated bandwidth for transmission. Adaptively allocating bandwidth to ONU-APs based on the proposed classification addresses this latency bottleneck.

**Figure 6.10** Average uplink latency with the proposed MLP-DBA.
As shown in Figs. 6.10(a) and (b), implementing MLP-DBA with ONU-AP classification effectively reduces the uplink latency even under heavy network loads. This is because bandwidth is allocated to heavily loaded Class C ONU-APs, by adding additional transmission opportunities in each polling cycle. In comparison, for network loads below 0.5 in the 16-ONU-AP network and 0.3 in the 32-ONU-AP network, MLP-DBA with and without classification exhibit similar latency performance since Class-A ONU-APs form the majority type of
ONU-APs. Overall, considering H2M applications, MLP-DBA is shown to support approximately 10% more network traffic than existing DBA schemes under $D_{st} = 1$ ms and 10 ms. As presented in Fig. 6.10, with MLP-DBA, a 16-ONU-AP network can support a network traffic load of up to 0.6 and 0.8 under $D_{st} = 1$ ms and $D_{st} = 10$ ms, respectively, and a 32-ONU-AP network can support a network load of up to 0.4 for $D_{st} = 1$ ms and 0.6 for $D_{st} = 10$ ms.

Fig. 6.11 plots the packet drop ratio under limited-service DBA, MR-DBA, LSTP-DBA and MLP-DBA with/without classification. As shown in the figures, implementations of MR-DBA, LSTP-DBA and MLP-DBA without classification help alleviate packet drops under heavy network loads. However, the improvement shown is marginal with $< 5\%$ compared to limited-service DBA in both networks. For a 32-ONU-AP network, the packet drop ratio is about twice that in a 16-ONU-AP network. This is attributed to the fact that under the same maximum polling cycle duration of 1 ms, the $BW_{max}$ of a 32-ONU-AP network is half of that for a 16-ONU-AP network. In comparison, the proposed MLP-DBA based on ONU-AP classification shows significant improvement in reducing packet drops. Overall, as illustrated in the figures, the packet drop ratio for MLP-DBA with classification is reduced by 10% as compared to the baseline limited-service DBA at high traffic loads. This is due to the reduced polling cycle durations for heavily-loaded Class C ONU-APs by introducing additional transmission opportunities. Since MLP-DBA can predict and classify ONU-APs that have long packet queues, GATE messages are generated prior to receiving REPORT messages from these ONU-APs. As a result, the polling cycle duration of these ONU-APs is reduced, and additional bandwidth is allocated, thereby improving both latency and packet drop performances. The performance of the proposed MLP-DBA can be further explained by viewing the varying distribution of Class-A, B and C ONU-APs in the two networks in Fig. 6.11.

### 6.5.2 ONU-AP Classification Analysis

In each polling cycle, the CO classifies ONU-APs upon receiving the REPORTs. Figs. 6.12(a) and (b) show the distribution of the classification results of ONU-APs, i.e. Class A, B, and C, during MLP-DBA operation over 500s simulation time in 16-ONU-AP and 32-ONU-AP network, respectively. In the lightly-loaded region,
i.e., for normalized network loads under 0.4, Class-B ONU-APs dominate since the ONU-APs tend to experience long OFF intervals. In this case, when an ONU-AP occasionally encounters an ON interval, arriving packets can be transmitted within $BW_{\text{max}}$. When the network traffic load increases, the percentage of Class A increases and followed by Class C, which rapidly increases and dominates the network due to the heavy packet bursts at these ONU-APs. In this case, existing predictive DBAs become less effective as $BW_{\text{max}}$ is allocated in majority of the polling cycles due to the dominated Class-C ONU-APs.
Therefore, a marginal improvement in reducing latency and packet drop ratio is shown in Figs. 6.10 and 6.11 when network loads are beyond 0.5 with MR-DBA, LSTP-DBA and MLP-DBA without classification. In contrast, MLP-DBA with the proposed ONU-AP classification adaptively allocates $BW_{grant}$ to Class-A ONU-APs and enables additional transmission opportunities to heavily-loaded Class-C ONU-APs, thereby effectively reducing latency and packet drop ratio as presented in Figs. 6.10 and 6.11. Moreover, since bandwidth contention amongst ONU-APs is more competitive in the 32-ONU-AP network, when network load increases, more ONU-APs are classified as Class C in the 32-ONU-AP network. Overall, results in Fig. 6.12 verify the capability of the proposed MLP-DBA in making adaptive bandwidth allocation decisions that reduce uplink latency and packet drops. Concluding, simulations presented in Figs. 6.10 to 6.12 in this section highlight the importance of classification and adaptive bandwidth allocation decisions.

6.6 Summary

The emerging remotely-controlled H2M applications demand stringent low-latency in the order of $1 - 10$ ms for effective interactions between human operators and remote slave robots/devices. To improve uplink latency performance for H2M traffic over access networks, we proposed the MLP-DBA scheme to adaptively allocate bandwidth to classified ONU-APs based on ANN-enabled bursty ON/OFF status prediction and bandwidth estimation. State-of-the-art research has anticipated the aggregated H2M traffic to have a bursty pattern featured by alternating ON and OFF intervals. Regarding the bursty traffic, the challenges in (a) accurately predicting arrivals and estimating bandwidth demand of ONU-APs, and (b) adaptively allocating bandwidth to ONU-APs with accumulated packet queues, need to be addressed in existing predictive DBA schemes.

In light of the above, in the proposed MLP-DBA scheme, we first exploited an ANN at the CO to first predict the ON/OFF packet burst status at individual ONU-APs, and then bandwidth demand was estimated accordingly. Specifically, we showed that with supervised training, the proposed ANN achieved $> 90\%$ accuracy in identifying the ON/OFF status of ONU-APs, thereby
yielding superior ONU-AP bandwidth estimation as compared to that in MR- and LSTP-DBA schemes. Comparisons have also been made with $k$-NN based ON/OFF status prediction. Results showed that the use of ANN technique was more flexible and achieved more accurate ON/OFF status prediction than the use of $k$-NN technique. Moreover, the sensitivity of the proposed ANN to different levels of traffic burstiness was investigated. The trained ANN achieved > 95% prediction accuracy regardless of $H$ variations in traffic.

Based on both ON/OFF prediction and the estimated bandwidth, we further proposed the classification of ONU-APs. Bandwidth was then allocated adaptively to classified ONU-APs using MLP-DBA scheme. Based on the classification of ONU-APs, the proposed MLP-DBA was able to (a) allocate the estimated bandwidth to ONU-APs that are in ON intervals (Class A), (b) reserve bandwidth from ONU-APs that are in OFF intervals (Class B), and (c) concurrently providing additional transmission opportunities to heavily-loaded ONU-APs (Class C). Consequently, MLP-DBA scheme effectively reduced the uplink latency and alleviated packet drops as compared to the baseline limited-service DBA and predictive MR-DBA and LSTP-DBA schemes.

Overall, this chapter is dedicated to addressing the latency bottleneck in heterogeneous networks in supporting low-latency H2M applications. The novel bandwidth resource allocation solution, MLP-DBA scheme, was proposed and evaluated based on the bursty traffic pattern reported in state-of-the-art literature. However, since Tactile Internet and H2M applications are still in their infancy, the unique traffic characteristics of H2M applications, and that of their aggregation over the access networks, are not fully understood to date. As such, in the next chapter, we develop several different H2M application based on a haptic teleoperation system and experimentally study the characteristics of H2M traffic. We analyse human control and feedback traffic traces generated during H2M interaction, and explore how existing bandwidth resource allocation schemes should be optimised for future Tactile Internet and H2M applications.
Chapter 7
Understanding Human-to-Machine Traffic and Bandwidth Allocation Schemes for Low-Latency H2M Applications

7.1 Introduction

In Chapter 6, we have discussed the bursty traffic pattern of human-to-machine (H2M) applications reported in recent research and exploited machine learning (ML) to improve bandwidth prediction and allocation for such traffic over heterogeneous optical and wireless access networks. However, H2M applications and haptic communications are still in their infancy and H2M applications have not been widely utilised over existing access networks. As such knowledge on H2M traffic characteristics and their aggregations over access networks is limited. As communication networks are rapidly evolving to support aggregation of conventional content-centric applications and emerging H2M applications, understanding the nature of H2M traffic and innovating bandwidth allocation solutions for future applications are imperative.

H2M applications typically comprise (a) a master domain with human operators and control interfaces, (b) a distant slave domain with execution machines and/or robotic devices, and (c) a network domain that communicates human control and slave machine feedback between the master and slave domains. As presented in the previous chapters, the heterogeneous optical and wireless access networks are promising in realising low-latency H2M applications [11-14]. The optical network units (ONUs) are integrated with wireless front-ends such as cellular base stations (BSs) and wireless local area network access points (APs) to support converged application delivery as illustrated in Fig. 7.1. Such a converged architecture benefits from the high capacity and reliability of optical networks, and the flexibility and mobility of wireless networks. As shown in Fig.
Chapter 7 Understanding Human-to-Machine Traffic and Bandwidth Allocation Schemes for Low-Latency H2M Applications

7.1, depending on the locations of masters and slaves in the heterogeneous network, H2M applications can be realised locally, i.e., intra-ONU communication, or non-locally, i.e., inter-ONU communication [18-20]. For local H2M connections between human operators and slave machines in the same wireless access network, data processing and exchange can be expedited by harnessing edge cloudlets deployed at ONU. In comparison, non-local H2M applications demand inter-ONU connections supported by the central office (CO), and their success hinges on stringent low latency in the passive optical network (PON). In this chapter, we focus on the inter-ONU H2M applications as these applications are more susceptible to network domain latency.

As analysed in previous chapters, effective H2M interactions require the time lag between human control and the subsequent feedback to be within 1 – 10 ms [1-3]. This demands the latency in a PON to be reduced to a few hundreds of $\mu$s for inter-ONU H2M communication. Note that the latency performance of PONs is affected by bandwidth allocation schemes, which at present are primarily designed and optimised for content applications over the network. The performance of existing schemes in supporting aggregated low-latency H2M applications has not been thoroughly examined in current literature. Further, the capability of existing schemes in ensuring latency for H2M application in the presence of conventional content applications is yet to be investigated. To address these key questions and change the existing bandwidth allocation schemes for future H2M and conventional traffic applications, understanding the unique characteristics of H2M traffic is vital.

Figure. 7.1 H2M applications over fiber-wireless access networks.
To address these key areas of interests, in this chapter, we develop interactive H2M applications based on a haptic teleoperation system to experimentally investigate the characteristics of human control and haptic feedback traffic. Understandings on the unique characteristics, including statistical distributions of control and feedback arrivals, and time-domain correlation of the collected traffic traces, are reported and is compared with existing studies on H2M traffic. In particular, we highlight a high correlation, defined as traffic causality, between human-controlled and haptic feedback traffic during H2M interactions. Exploiting this traffic causality, we propose an artificial intelligence-facilitated interactive bandwidth allocation (AIBA) scheme to reduce the latency for H2M applications over heterogeneous access networks.

In the proposed AIBA scheme, the central office (CO) exploits an artificial neural network (ANN) to estimate feedback bandwidth demand and pre-allocates this estimated bandwidth for feedback when forwarding the control traffic. The traffic prediction expedites feedback delivery by eliminating the bandwidth report-then-grant process in existing bandwidth allocation schemes. To validate the effectiveness of the proposed ANN in estimating feedback bandwidth, and the AIBA scheme in reducing latency for feedback traffic, we simulate a basic network scenario, considering only H2M applications over the simulated network where experimental traffic traces are injected to the network simulation platform to emulate H2M traffic for performance evaluations. The simulation results validate that the proposed ANN in AIBA scheme achieves more accurate feedback bandwidth estimation and the AIBA scheme reduces latency for haptic feedback by up to 60% compared to the baseline scheme for content traffic, and also outperforms existing predictive schemes by 40%.

As the simulation validates the effectiveness of the AIBA scheme in reducing feedback latency by harnessing the traffic causality, we further improve the AIBA scheme by incorporating (a) control arrival prediction and bandwidth estimation and (b) priority differentiation between H2M and conventional content applications. Then, we focus on the capability of the AIBA scheme and existing schemes, including both baseline and predictive schemes, in supporting aggregated low-latency H2M applications and conventional content applications. More specifically, we discuss the prioritisation of H2M traffic over content traffic.
in existing schemes. Via extensive simulations using our experimental traces, we evaluate the performances existing and our proposed AIBA scheme in reducing and constraining latency for H2M applications in the presence of content traffic. The simulation results again validate the effectiveness of the AIBA scheme in improving the latency performance for H2M applications over existing schemes.

The rest of this chapter is organised as follows. Section 7.2 presents an overview of state-of-the-art research on H2M traffic. Our experimental setup, developed H2M applications, and traffic analysis are presented in Section 7.3. Section 7.4 details the proposed AIBA scheme and validates the performance of AIBA scheme in reducing feedback latency by exploiting traffic causality. The improved AIBA scheme that includes control bandwidth estimation and H2M priority, together with comprehensive simulation evaluations on AIBA and existing schemes are presented in Section 7.5. Finally, Section 7.6 summarises the main contributions of this chapter.

### 7.2 Related Work

Real-time human control and haptic traffic in H2M applications are the latest traffic that needs to be supported by our communication networks. In meeting the stringent latency demand of such traffic, studying the H2M traffic characteristics is vital and therefore continues to drawing research interest. As detailed in [197], control and haptic feedback in H2M applications are mainly described by the degree of freedom (DoF), i.e. the number of different directions that create forces, e.g., 3-DoF force along x, y, z axis in a touch spot. The control/feedback packets may comprise 1 DoF to over hundreds of DoFs, which are sampled/generated by master control interface/slave machines [198]. Studies in [32,199,200] investigate haptic data compression and processing to reduce the number of packets transmitted over the network and reports the bursty arrival pattern. In [199], the authors have proposed dead-band coding based on the just-noticeable difference in force perception and discuss the bursty pattern of control and force feedback packets in their teleoperation. This knowledge of burstiness was used to design a resource reservation scheme for low-latency wireless access [32]. Recent research reported in [200] analyses the statistical nature of control and haptic feedback packet inter-arrival time using experimental traffic traces in two teleoperation
7.3 Experimental H2M Applications and Traffic Analysis

7.3.1 Haptic Teleoperation System

Our experiments comprise a series of interactive H2M applications based on a commercial haptic teleoperation system [201]. As illustrated in Fig. 7.2, the system is an innovative touch and force feedback teleoperation platform for fingers and hands that allows human operators to remotely touch and grasp computer-generated virtual objects. The master domain consists of a human...
operation interface and a control unit as shown in Fig. 7.2. The interface includes a 22-sensor data glove and a 6-DoF position tracker that capture human hand movements and actions. In addition, a hand exoskeleton is worn to react to the haptic feedback. During H2M interaction, the control traffic generated by hand actions is processed by the control unit and transmitted to slave-domain applications, where a virtual hand can be controlled to touch and/or grasp virtual objects as shown in Fig. 7.2. To explain the generation of human control and haptic feedback packets in the experimental system, details of haptic sensors and actuators are presented in Fig. 7.3.

As shown in Fig. 7.3(a), data glove sensors are distributed at the joints of human hands and the position track is placed at the wrist. The force variations caused by hand actions are sensed by the sensors and then mapped to different hand gestures shown in the slave-domain virtual space. The position tracker records 6-DoF measurements of the master-domain hand location, i.e., X, Y, Z position in inches and Azimuth Elevation and Roll attitude in degrees. These measurements are sent to slave-domain applications in synchronised master- and slave-side hand movements. The sensor- and tracker-sensed data are the main content of human control packets in the experimental system. In comparison, the haptic feedback packets primarily indicate different forms and values of reaction.
forces. As shown in Figs. 7.3(b) and (c), the exoskeleton is strapped to the glove with rings and finger loops placed at the fingertips and front joints. Bowden cables of the rings and loops are driven by motors in response to the feedback, thereby creating different resistive forces to each finger and the palm. In our study, we develop three types of H2M applications that create different haptic feedback as discussed in the following subsections.

7.3.2 Experiment H2M Applications

The master domain devices, i.e., data glove, hand exoskeleton, position tracker and the FCU are installed as illustrated in Fig. 7.2. The slave domain contains a PC, on which H2M applications are developed based on Visual Studio 2008 platform in C++. The operation system of the slave domain PC is Window 8 and the CPU is 3.2 GHz Intel Core i7. Fig. 7.4 shows three types of H2M applications
developed for this study as detailed below. The control and haptic feedback traffic in each application are specified in Table 7.1.

- **Application A**: moving the hand in free space. The virtual hand in virtual space moves with the human hand in real-time as shown in Fig. 7.4(a). Without interactions with virtual objects, the Bowden cables are pulled with a constant default force.

- **Application B**: touching a virtual ball surface. A virtual ball is placed in the virtual space. When the hand touches the ball as shown in Fig. 7.4(b), reaction forces are put onto fingers to imitate stiffness of the object texture.

- **Application C**: grasping and moving objects, i.e., a ball, a box, and a cone in a virtual space. When human operators pick an object and move it in virtual space as shown in Fig. 7.4(c), vibrated forces are put onto fingers to create the feeling of friction when moving the object.

Note that the control traffic in Applications A-C contains similar content, i.e., hand posture and position. The processing of the control traffic and the corresponding haptic feedback traffic, however, depend on the type of application. We collect human control and feedback traffic traces using Applications A-C. Specifically, in each experiment, human participants can start an application, perform operations such as grasping either a ball or a box, and move it to a random position such as in Application C, and pause/restart an application at any time. Each experiment lasts 20 mins. Then, we analyse the statistical characteristics of the experimental control and feedback traffic, and compare our...
results with existing studies on H2M traffic and the content-centric traffic over
the Internet.

7.3.3 Control and Haptic Feedback Traffic Analysis

7.3.3.1 Statistical Analysis: Packet Inter-Arrival Time

Packet inter-arrival time is one of the key features in studying network traffic. The statistics of packet inter-arrival time can be utilised in traffic classification and fitting arrival models [204]. Research in [200] has fitted the packet inter-arrival time distributions of human control and haptic feedback traces from two teleoperation systems. The candidate distributions including deterministic, Gamma, Generalised Pareto and Poisson, have been reported. In our study, we fit control and feedback packet inter-arrival times into a variety of distributions in existing network traffic studies [205,206]. The candidate distributions are listed as Exponential, Gamma, Generalized Pareto, Inverse Gaussian, Logistic, Lognormal, Nakagami, Normal, t-location, Weibull. The maximum likelihood estimation is utilized to fit the parameters in each candidate distribution. The goodness of fitting (GoF) is measured by comparing the average fitting error between the fitted cumulative distribution function (CDF) and the experimental CDF. Note that a smaller fitting error indicates a better fitting to a candidate distribution, i.e., higher goodness of fitting.

In Fig. 7.5, we plot the CDFs and fitting errors of the top-four bested fitted distributions, i.e., Generalised Pareto, t-location, Logistic, Exponential, to our experimental traces. Notably, in both control and feedback traces, the Generalised Pareto (blue dash line) best fits empirical distributions (black solid line). This observation is consistent in Applications A-C, and in both control and feedback traces. In comparison, the rest distributions incur significant higher fitting errors. The average CDF fitting error of each distribution, and the fitted parameters of Generalised Pareto distributions for Applications A-C are listed in Table 7.2 and Table 7.3, respectively. It can be noted that the parameter pairs in Table 7.3 only slightly vary in different applications, but are distinct in control and feedback directions. In Fig. 7.5(d), we show the fitted distributions of aggregated packet inter-arrival times in Applications A-C. The Generalised Pareto distribution remains the best fit. Notice that Generalised Pareto
Chapter 7 Understanding Human-to-Machine Traffic and Bandwidth Allocation Schemes for Low-Latency H2M Applications

196

Figure. 7.5 Control and feedback packet inter-arrival time cumulative distribution functions of Applications A to C.
distribution is also reported in the H2M traffic study in [200]. In comparison, existing content traffic over the Internet has been widely studied and modelled by Pareto distribution as discussed in Chapter 2. The Pareto distribution is a particular case of its generalised counterpart by substituting \( \sigma_{GP} = \sigma_p / \xi_p \), and \( \xi_{GP} = 1 / \xi_p \) in the probability density function given in Table 7.3.

Based on both existing studies and our analysis on H2M traffic, we highlight the Generalized Pareto distribution as a potential model for H2M arrivals. However, it should be noted that due to current limited knowledge on H2M traffic, more rigorous experimental research using diverse H2M applications are still warranted to validate the preliminary observations. From Fig. 7.5, the CDFs of control and feedback packet inter-arrival time indicate intensive control and feedback packets transmission during H2M interactions. Up to 80% of the packets are sent within < 1ms as shown in the figure due to the demand of real-time haptic interaction in H2M applications. As such, we further investigate the time-domain correlation in the collected traffic traces.

### 7.3.3.2 Time Domain Analysis: Control/Feedback Correlation
From the above distribution fitting, it can be noted that in our experiments, control and feedback packets are exchanged within milliseconds and the time intervals between packets follow Generalised Pareto distributions. In Fig. 7.6, we present the human control and haptic feedback traffic pattern by plotting the dynamic traffic volume (Mbps) in Application C on different time scales, i.e., 1 ms, 1s and 10s. Fig. 7.6 shows an example experiment in which human operator first moves hand in virtual space, then grasps the three virtual objects, i.e., the ball, cone, and cube as depicted in Fig. 7.4(c). Note that after the movement and grasping, the human operator pauses the action for a while. This is to distinguish traffic from different actions for illustrative purpose. In addition, the observations of the traffic in Applications A and B are similar to that in Fig. 7.6.

As can be observed in Fig. 7.6, control and feedback packets are closely correlated, and packet arrivals fluctuate on the 1-ms granularity compared to the 1-s and 10-s timescale. In optical access networks, the CO allocates uplink bandwidth to ONUs in a round-ribbon manner and the interval between consecutive transmissions of an ONU, known as a polling cycle, is in millisecond order. As such, the short-term (within ms) traffic characteristics is critical to improving bandwidth allocation for H2M applications. Based on the above understandings, we pay special attention to the correlation of human control and haptic feedback traffic on the timescale of polling cycles (~ms).

![Figure 7.6 An illustration of control and feedback traffic in Application C.](image-url)
Considering the duration of the polling cycle in 1 ms, we compute the self-correlation coefficients of individual control and haptic feedback traffic traces, respectively, as well as the cross-correlation coefficient between control traffic and the associated haptic feedback traffic. The correlation coefficient, ranging from 0 to 1, indicates the correlation between two stochastic processes. A larger coefficient value, i.e., approaching 1, manifests a stronger correlation. Let us denote the human control packets transmitted in sequential polling cycles as \( \{ ctr(t_1), ctr(t_2), ctr(t_3), \ldots, ctr(t_i), \ldots \} \), in which \( t_i \) refers to the \( i \)-th polling cycle and \( ctr(t) \) is the amount of control packets (in bytes) transmitted. Likewise, the haptic feedback packet sequence is denoted as \( \{ fbk(t_1), fbk(t_2), fbk(t_3), \ldots, fbk(t_i), \ldots \} \).

Considering the sequence with length \( N \), i.e., \( N \) number of polling cycles, the self-correlation coefficients of human control and haptic feedback traffic, denoted as \( \rho_{ctr}(t) \) and \( \rho_{fbk}(t) \) respectively, are calculated as follows:

\[
\rho_{ctr}(t) = \frac{1}{\sigma_{ctr}^2} \sum_{i=1}^{N} E\{ \text{ctr}(t_i) \cdot ctr(t_{i-t}) \}
\]

(7.1)

\[
\rho_{fbk}(t) = \frac{1}{\sigma_{fbk}^2} \sum_{i=1}^{N} E\{ \text{fbk}(t_i) \cdot fbk(t_{i-t}) \}
\]

(7.2)

Note that \( E\{ \text{ctr}(t_i) \cdot ctr(t_{i-t}) \} \) and \( E\{ \text{fbk}(t_i) \cdot fbk(t_{i-t}) \} \) represent the expectation of the control and haptic feedback packet sequences, respectively. \( \sigma_{ctr}^2 \) and \( \sigma_{fbk}^2 \) represent the variances, respectively. The variable \( t \) (\( t = 0,1,2,\ldots \)) in \( \rho_{ctr}(t) \) and \( \rho_{fbk}(t) \) indicates the time lag, in the unit of polling cycle. Further, the cross-correlation coefficient, termed \( \rho_{cross}(t) \), between control and haptic feedback traffic is given as follows:

\[
\rho_{cross}(t) = \frac{1}{\sigma_{ctr} \sigma_{fbk}} \sum_{i=1}^{N} E\{ \text{fbk}(t_i) \cdot ctr(t_{i-t}) \}
\]

(7.3)

Figs. 7.7(a), (b), and (c) plot \( \rho_{ctr}(k) \), \( \rho_{fbk}(k) \), and \( \rho_{cross}(k) \) of Applications A-C, respectively. The client and server traffic trace analysis of video, email, and web browsing are also plotted (in red line) in Figs. 7.7(a), (b), and (c), respectively. Note that the analyses of the three content applications are randomly plotted with Applications A-C, aiming to highlight the difference between the traffic in current applications and H2M applications. As shown in Fig. 7.7(a), human control and
haptic feedback traffic of Application A are most correlated at $k = 0$, with $\rho_{\text{cross}}(0) = 0.6$, while, both $\rho_{\text{ctr}}(k)$ and $\rho_{\text{fbk}}(k)$ are < 0.4. Likewise, in Figs. 4(b) and (c), a high $\rho_{\text{cross}}(k)$ of up to 0.8 is observed at $k = 1$, indicating that feedback traffic is highly correlated to the control traffic in one previous polling cycle. The values of $\rho_{\text{ctr}}(k)$ and $\rho_{\text{fbk}}(k)$ of Applications A-C are again smaller than $\rho_{\text{cross}}(k)$. Note that the
highest \( \rho_{\text{cross}}(k) \) occurs at \( k = 0 \) in Application A and \( k = 1 \) in Applications A-C. This is partly because constant control (feedback) packets are generated by (sent to) the glove in Application A whereas haptic feedback in Applications A-C is also related to how human hand interacts with the objects (Fig. 7.4). In comparison, the traffic cross- and self-correlation of the three content applications in Fig. 7.7 are significantly weak.

Overall, Fig. 7.7 indicates a strong cross-correlation, i.e., \( \rho_{\text{cross}} = 0.6 \sim 0.8 \), between haptic feedback traffic and control traffic in either current or the most recent polling cycle, and a weaker self-correlation, i.e., \( \rho_{\text{ctr}} < 0.6 \) and \( \rho_{\text{fbk}} < 0.4 \), for the individual traces. Moreover, compared to H2M traffic, the conventional content-centric traffic does not show the correlations on the timescale of polling cycles (\(~\text{ms}\)). As such, we define this unique strong cross-correlation between the real-time human control and haptic feedback traffic as traffic causality of H2M applications. Exploiting the reported traffic causality, we propose the AIBA scheme for H2M applications to expedite haptic feedback traffic delivery based on the control traffic to be forwarded, which is detailed as follows.

### 7.4 AI-facilitated Interactive Bandwidth Allocation Scheme

#### 7.4.1 Proposed AIBA Scheme

Fig. 7.8 compares the bandwidth allocation timing diagrams of existing schemes and our proposed AIBA scheme. As illustrated in Fig. 7.8, the CO broadcasts downlink packets to ONUs using downlink wavelength \( \lambda_{\text{down}} \). In the uplink, ONUs share uplink wavelength \( \lambda_{\text{up}} \) for uplink transmission. Recall from the overview in Chapter 2 that in existing bandwidth allocation schemes, the uplink bandwidth is allocated in a report-then-grant fashion using REPORT and GATE. As shown in Fig. 7.8(a), an ONU requests bandwidth using a REPORT piggybacked to the uplink data. The REPORT notifies the CO of its buffer occupancy \( \text{BW}_{\text{req}} \), i.e., buffered packets (in bytes). Upon receiving the REPORT, the CO grants bandwidth, \( \text{BW}_{\text{grant}} \), by sending a GATE to the ONU. The GATE indicates the assigned transmission start time and duration. This process
repeats for each ONU in each polling cycle. For inter-ONU communication, the CO forwards the received data from ONU $i$ to its destination ONU $j$ in broadcast as shown in Fig. 7.8(a).

Recall that the baseline limited-service and predictive schemes are commonly-adopted existing schemes for bandwidth allocation at the CO. In baseline scheme, the CO always allocates bandwidth equal to the request, i.e., $BW_{grant} = \min\{BW_{req}, BW_{max}\}$. The $BW_{max}$ is the maximum bandwidth can be assigned to an ONU in each polling cycle. As such, in baseline scheme, packets arriving in current polling need to be reported, and then transmitted in the next polling cycles. When $BW_{req} > BW_{max}$, packets may defer multiple polling cycles before transmission. Predictive schemes predict arrivals and allocate surplus
bandwidth $BW_{\text{pred}}$ such that $BW_{\text{grant}} = \min\{BW_{\text{req}} + BW_{\text{pred}}, BW_{\text{max}}\}$. This allows some arriving packets to be transmitted in the current polling cycle without needing to be reported, thereby reducing latency. Note that static predictive schemes, e.g., constant and linear credit schemes, allocate a constant $BW_{\text{pred}}$. Statistical predictive schemes, on the other hand, typically estimate $BW_{\text{pred}}$ based on short-term arrivals to ONUs in the past polling cycles. The details of existing bandwidth allocation schemes, including both the baseline and predictive schemes, have been presented in Chapter 2.

Note that in existing schemes, the CO allocates bandwidth to ONUs independently based on individual reports. The inter-ONU traffic correlation is not considered. For H2M traffic, we report the traffic causality, which is a strong cross-correlation between human control and haptic feedback, in above traffic analysis. As such, exploiting this cross-correlation, we propose the AIBA scheme to achieve interactive bandwidth allocation in reducing latency for H2M applications. The operation diagram of AIBA scheme is illustrated in Fig. 7.8(b).

In the AIBA scheme, the CO grants bandwidth to subsequent feedback traffic at the same time when forwarding control packets to the ONUs, thereby eliminating the bandwidth request process altogether and effectively reducing the latency experienced by the feedback packets.

Specifically, considering H2M applications comprise master and slave domains, we categorise ONUs associated with master human operators as a M-ONU, and the counterpart ONUs associated with slave machines/robots as a S-ONU as shown in Fig. 7.8(b). Note that an ONU can be both a M-ONU and S-ONU as it can support control and feedback traffic from multiple H2M applications. Then, in the AIBA scheme, upon receiving control traffic from an M-ONU in each polling cycle, the CO uses the ANN to estimate feedback bandwidth, and then grants this estimated bandwidth to S-ONUs when forwarding the control traffic to them. The architecture of the ANN is presented in Fig. 7.9. The inputs to the ANN are bilateral packets (in bytes), i.e., control and feedback bytes received from the M-ONU and S-ONU respectively, in the most-recent 4 polling cycles. The input layer hence has 8 neurons. The output is the feedback bandwidth granted to S-ONU. The ANN comprises 2 hidden layers with 5 and 3 neurons, which are
determined by layer-wise training. Note that increasing the number of layers and neurons does not see improvement in terms of bandwidth estimation accuracy.

Overall, the baseline scheme incurs waiting time for reporting and bandwidth granting. The predictive scheme reduces latency by estimating and allocating surplus bandwidth to individual ONU.
s. In comparison, the AIBA scheme reduces the latency of H2M applications by harnessing the traffic causality, i.e., the cross-correlation between human control and haptic feedback traffic. In the following, we validate the effectiveness of ANN-based feedback bandwidth estimation, and the AIBA scheme in expediting feedback delivery from S-ONUs.

### 7.4.2 Performance Evaluation

The performance of AIBA is verified using a packet-driven MATLAB simulation platform as shown in Fig. 7.10 with 16, 32, and 64 ONUs. As illustrated in Fig. 7.10, half of the ONUs are injected with experimental control traffic traces and the other half with experimental haptic feedback traffic traces collected from the three H2M applications. Note that in this section, we simulate the basic scenario, where an ONU is associated with either a control or feedback trace, as we aim to validate the effectiveness of exploiting the reported traffic causality in the AIBA scheme. The comprehensive evaluations for more complex simulation scenarios that include both H2M and content traffic will be discussed in the next section.
First, Fig. 7.11 plots the haptic feedback bandwidth estimated by the AIBA for S-ONUs injected with Applications A-C haptic traces. The figure also shows the bandwidth estimation error as compared to the static and statistical predictive schemes. Here, feedback bandwidth estimation in static and statistical predictive scheme relies on arithmetic average and moving average algorithm as overviewed in Chapter 2 Section 2.3.3, respectively. Note that both the two predictive schemes estimate feedback bandwidth only based on the historical feedback arrivals for S-ONUs. Results in Fig. 7.11 show that the AIBA yields the best performance as knowledge of control traffic is utilised when estimating bandwidth for the subsequent feedback traffic. Exploiting the traffic causality by using both human control traffic and haptic feedback arrivals, the ANN achieves more accurate bandwidth estimation for haptic feedback, thereby enabling more effective bandwidth allocation in the AIBA scheme.

Fig. 7.12 compares the average uplink latencies of the control and haptic feedback traffic, i.e., duration from control and feedback packet arrival at M-ONUs and S-ONUs until transmitted to the CO. The baseline scheme results in similarly high latency between control and feedback traffic since bandwidth is allocated to ONUs based on independent bandwidth requests. The latency improvement achieved by using predictive schemes is limited only to haptic feedback traffic and only to networks with a small number of ONUs. Importantly, the latency of control traffic is impacted by the inaccurate
Chapter 7 Understanding Human-to-Machine Traffic and Bandwidth Allocation Schemes for Low-Latency H2M Applications

bandwidth prediction of the haptic feedback traffic, leading to latencies higher than the baseline case (Figs. 7.12(b) and (c)). The AIBA exploits the reported traffic causality to achieve accurate haptic feedback prediction and bandwidth pre-allocation, thereby expediting haptic feedback delivery by up to 60% compared to the baseline scheme. Compared to the predictive schemes, the AIBA reduces feedback latency by 40% without affecting the control traffic. Moreover, the effectiveness of the AIBA scheme in reducing latency for haptic feedback is less affected by the number of ONUs. Around 40% latency reduction from the baseline scheme is attained in the 32- and 64-ONU networks. This is again because exploiting the time-domain correlation between human control and haptic feedback, feedback bandwidth is adaptively allocated to S-ONUs, enabling timely feedback packets delivery in response to control. In comparison, in
predictive schemes, the CO always allocate the estimated bandwidth for feedback at the time when receiving reports from S-ONUs. As such, the latency of predictive schemes is more susceptible to bandwidth estimation error as well as the number of ONUs.
The above simulation evaluations validate effective expedition of feedback delivery by harnessing the traffic causality to adaptively allocate bandwidth for feedback traffic. Basic scenarios where simulated networks only have H2M applications are investigated. Note that as H2M applications demand stringent low latency, constraining the latency of H2M traffic in the presence of content traffic is critical. The performances of existing schemes in differentiating H2M and content traffic priority, and in ensuring latency constraint for H2M traffic have yet to be investigated. As such, in the following section, we consider the scenarios when both H2M and content applications aggregate over access networks. Taking these aspects into account, we present the implementation of the AIBA scheme that estimates both control and feedback bandwidth, and prioritises H2M traffic over conventional content traffic. Then, a comprehensive evaluation of existing schemes, and the AIBA scheme in achieving low latency for H2M applications is reported.

7.5 AIBA Scheme for Aggregated H2M and Content Applications

7.5.1 AIBA Scheme with Traffic Priority Differentiation

Note that in the above study, feedback bandwidth is estimated and pre-allocated to S-ONUs to reduce latency for haptic feedback packets as illustrated in Fig. 7.12. In reducing latency for control packets, conventional bandwidth estimation algorithms adopted in predictive schemes can be used at the same time to estimate control bandwidth for M-ONUs. In this section, we incorporate the statistical algorithm for control bandwidth estimation into the AIBA scheme. This ensures that latency of both control and feedback packets can be reduced by pre-allocating the estimated bandwidth to them. It should be noted that our focus in this section is on differentiating H2M and content traffic, and investigating the capability of different schemes in reducing latency for H2M applications. Since H2M packets demand stringent low latency in their transmission, we consider prioritising H2M packet transmission by ONUs in the presence of content traffic. The ONUs buffer H2M traffic and content traffic in separate queues. In the AIBA scheme, the CO estimates and pre-allocates
feedback bandwidth when forwarding control packets. Content packets, together with any remaining H2M packets in the buffer, are reported to CO using REPORTs. Then, in the designated transmission timeslots, transmission of H2M packets is prioritised by ONUs as shown in Fig. 7.8(b). The implementation of AIBA scheme with traffic priority differentiation is detailed as follows:

- In each polling cycle, ONUs transmit uplink data and report buffer occupancy, i.e. $BW_{req}$, including both H2M and content packets in bytes, in a round-ribbon manner. In transmission timeslots, the transmission of H2M traffic is prioritised by ONUs.
- Based on the H2M packets received from an ONU $i$, the CO estimates bandwidth, $BW_{pred(ctr, i)}$ and $BW_{pred(fbk, j)}$, for ONU $i$’s control traffic and the feedback traffic for the destination S-ONU $j$, respectively.
- Upon forwarding the control packets to S-ONU $j$, the CO sends a GATE to grant $BW_{pred(fbk, j)}$ for feedback transmission.
- Then, the CO sends a GATE to grant $BW_{grant} = \min\{BW_{req} + BW_{pred(ctr, i)}, BW_{max} - BW_{pred(fbk, i)}\}$ for ONU $i$. Note that $BW_{pred(fbk, i)}$ is the pre-allocated bandwidth to ONU $i$ based on the control traffic to it. This ensures fairness among ONUs.

Note that when no H2M traffic is present at an ONU, the above bandwidth allocation in AIBA is the same as the existing schemes for content traffic. The pseudocode of the AIBA scheme is presented as follows:

---

**AIBA Scheme: Pseudocode of bandwidth allocation at the CO**

$t_{scheduled}$ — time up to which the uplink channel has been scheduled

$RTT$ — round-trip transmission time

$T_g$ — guard time

$T_{process}$ — time for processing a GATE message at an ONU-AP

$T_{REPORT}$ — duration of a REPORT message

$t_{start(k, i)}$ and $t_{end(k, i)}$ — start and end time of granted timeslot for ONU $i$ in the $k$th polling cycle

$t_{start_fbk(j)}$ and $t_{end_fbk(j)}$ — start time of granted timeslot for ONU $j$, which is the destination S-ONU of ONU $i$.

$R_{PON}$ — data rate in the optical access network

**The operation at the CO upon receiving the REPORT in the $k$th polling cycle from ONU $i$ (repeat for all $k$ and $i$):**

---
7.5.2 Performance Evaluation

To evaluate the latency performance of different bandwidth allocation schemes, we implement packet-level simulation of a 16-ONU network in MATLAB. Experimental control and feedback traces in Applications A-C are injected to ONUs to emulate H2M arrivals. Content arrivals to ONUs are generated using the common synthetic bursty traffic source [42]. Normalised traffic loads from 0.1 to 1 in the network are simulated and analysed. In our simulation, we pair two ONUs, e.g., ONU $i$ and $j$, and emulate inter-ONU H2M applications between ONU $i$ and $j$. For this purpose, we randomly select an application for ONU $i$ and $j$, respectively. The experimental control packets are injected to ONU $i$ and $j$. Concurrently, the associated feedback packets are injected to ONU $j$ and $i$ in response to the control packets. In this way, an ONU in the simulated network acts as both a M- and S-ONU. Meanwhile, content arrivals to ONUs are buffered
7.5 AIBA Scheme for Aggregated H2M and Content Applications

in a different queue at an ONU. In the following, we compare the latency performance of: (a) baseline scheme without priority, i.e., first-in-first-out for all packets; (b) baseline scheme with priority, i.e., prioritising H2M traffic; (c) predictive scheme that estimates total bandwidth for both H2M and content traffic and prioritizes H2M traffic transmission; (d) predictive scheme that estimates H2M bandwidth only and prioritizes H2M traffic; and (e) proposed AIBA scheme.

7.5.2.1 Average Latency Comparisons

Fig. 7.13 compares the average uplink latency of H2M and content packets under these different schemes. From Figs. 7.13(a) and (b), in the baseline scheme (without priority), the latency of H2M and content traffic is equally high, and increases with increasing network loads. When network load exceeds 0.5, the average latency of both H2M and content traffic exceeds 1 ms. Prioritising H2M traffic effectively prevents the latency increase with increasing network loads. As shown in Fig. 7.13(a), the latency of H2M traffic is below 500 μs under these schemes that prioritise H2M packets. With priority, H2M packets can be transmitted without to be reported. Therefore, the average latency, in general, approximates 0.5 polling cycle time. With a 1-ms polling cycle duration in maximum, the average latency of H2M traffic approaches 500 μs in the high load region, i.e., network loads 0.7 to 1 as shown in Fig. 7.13(a). The polling cycle durations of the discussed schemes are plotted in Fig. 7.14. Note that neither baseline nor predictive schemes can reduce latency from this 500 μs in the high load region. In comparison, the AIBA scheme effectively reduces this latency since it facilitates adaptively feedback bandwidth allocation based on the control traffic in a polling cycle. More details are explained as follows.

As shown in Fig. 7.13(a), in the light load region, i.e., network loads 0.1 to 0.3, adopting priority in baseline scheme does not improve the latency performance compared to baseline scheme (without priority). Due to the low content traffic loads, most content packets are transmitted within 1 – 2 cycles time. When the content buffer is empty, H2M packets still experience the report-grant latency in baseline scheme (with priority). As such, prioritising H2M traffic in the baseline scheme sees benefits only under network loads above 0.3 in Fig. 7.13(a).
Moreover, implementing predictive schemes can reduce latency, depending on the amount of surplus bandwidth estimated and allocated. This is illustrated by comparing two use cases of predictive scheme in Fig. 7.13, i.e., statistical bandwidth estimation for H2M and content traffic in total, and the estimation for pure H2M traffic. Fig. 7.13(a) shows that under the predictive scheme (total bandwidth), the latency of H2M traffic is below the baseline scheme (with priority) only under network loads less than 0.3, and increases dramatically with increasing network load. This is because the CO attempts to grant surplus bandwidth for all arriving packets to ONUs, expecting that both H2M and content arrivals can be transmitted without reporting. This causes the latency to increase.
7.5 AIBA Scheme for Aggregated H2M and Content Applications

with prolonged polling cycle duration as shown in Fig. 7.14. Moreover, since H2M packets are prioritised, the effectiveness in reducing content packet latency via predictive scheme is affected as will be discussed in the following. Compared to the rest schemes, only marginal latency reduction in the light load region is shown in Fig. 7.13(b) via the predictive scheme (total bandwidth).

On the other hand, implementing the predictive scheme (H2M bandwidth) reduces latency under network loads < 0.5 in Fig. 7.13(a). Since the CO grants surplus bandwidth only for H2M arrivals, the report-grant process can be eliminated for H2M packets without affecting the polling cycle duration as shown in Fig. 7.14. However, compared to the baseline scheme (with priority), the predictive scheme (H2M bandwidth) meets its bottleneck in the moderate to high load region, e.g., network loads > 0.5 in Fig. 7.13(a). As shown in Fig. 7.14, the polling cycle duration increases steeply when network load exceeds 0.5. As such, in Fig. 7.13(a), the average latency in predictive scheme (H2M bandwidth) quickly converges to baseline scheme (with priority). Overall, the latency performance of a predictive scheme is susceptible to its estimated bandwidth value, and the ability of any predictive scheme in reducing latency for H2M traffic is weakened by a high network load.

In comparison, the AIBA scheme shown in Fig. 7.13(a) reduces latency for H2M traffic in all network loads 0.1 to 1 over baseline schemes and the predictive scheme (total bandwidth). Note that compared to the predictive scheme (H2M

Figure. 7.14 Average polling cycle duration comparison.
bandwidth), the AIBA scheme incurs a slightly higher latency at load 0.5. This observation can be explained as follows. As shown in Fig. 7.14, polling cycle durations of the studied schemes primarily grow in moderate load region, i.e., network loads 0.4 to 0.7. The polling cycle growth starts earlier in the AIBA scheme i.e., at load 0.4, than the predictive scheme (H2M bandwidth). As such, the AIBA scheme and predictive scheme (H2M bandwidth) see comparable latency in a narrow load region 0.4 to 0.5 in Fig. 7.13(a). In other network loads, particularly the high load region 0.7 to 1, AIBA scheme effectively reduces latency for H2M traffic.

The impact of prioritising H2M traffic on the latency of content traffic is presented in Fig. 7.13(b). Compared to the baseline scheme (without priority), the rest schemes lead to higher latency since content packets are deferred in the presence of H2M packets. The predictive scheme (total bandwidth) is an exception in network loads 0.1 to 0.4. As explained earlier, this is because surplus bandwidth is allocated for both H2M and content traffic. This latency reduction for content traffic is at the cost of increasing polling cycle and the latency of H2M traffic. Moreover, among the schemes that prioritise H2M traffic in Fig. 7.13(b), the baseline scheme (with priority) shows a lower latency when the network load exceeds 0.5, since the baseline scheme intends to prevent bandwidth over-granting via the report-grant process. On the other hand, when the network load is less than 0.5, differences in latency performance of these schemes are minor. In summary, from Figs. 7.13(a) and (b), compared to the predictive schemes evaluated, the proposed AIBA scheme effectively reduces latency for H2M traffic in all network loads, and retains similar latency for content traffic.

7.5.2.2 Latency Statistics of H2M Packets

As the success of H2M applications and haptic interactions relies on stringent millisecond-low latency, we investigate the capability of different schemes in constraining latency for H2M packets. Based on the above latency evaluations, we narrow our focus to the baseline scheme (with priority), predictive scheme (H2M bandwidth), and AIBA scheme. In Figs. 7.15(a) to (c), we present and analyse the latency distribution and CDF of H2M packets under network load 0.2 (light load), 0.5 (moderate load), and 0.8 (high load), respectively.
Figure 7.15 Packet latency statistics.
Distinct centers and shapes of the latency distributions under the studied schemes are shown in Fig. 7.15. Note that compared to the baseline and predictive schemes, the AIBA scheme constrains a higher proportion of packets with latency less than 100 μs, and the probability of yielding longer latency values decreases in Figs. 7.15(a) to (c). This is because in the AIBA scheme, the CO exploits the correlation between control and feedback traffic to grant bandwidth, allowing timely feedback delivery in response to control. In addition, in Fig. 7.15(a), the peak probability of baseline and predictive scheme occurs at about 1.5 and 0.5 polling-cycle-time latency, respectively. This is consistent with our previous understandings. When the content packet buffer is empty, H2M packets under the baseline scheme need to wait 0.5 polling cycle to be reported and another 1 polling cycle for the granted timeslot. In the predictive scheme, the latency is reduced towards 0.5 polling cycle time via allocating surplus estimated bandwidth for H2M packets. AIBA scheme further reduces this latency as it adaptively allocates bandwidth in a polling cycle based on control and feedback correlation. The CDF curves in Fig. 7.15 indicate the percentage/probability of packets below different latency values under each scheme. At the network load 0.2, Fig. 7.15(a) shows that more than 90% of H2M packets are transmitted within 150 μs in the AIBA scheme, which is 10% and 15% higher than that in the baseline and predictive scheme, respectively. Note that in Fig. 7.15(b), the peak probability at the 1.5-polling-cycle latency of baseline scheme significantly drops. This is because the accumulation of content packets allows more H2M packets to be transmitted within 1 polling cycle time. Overall, at the network load 0.5, similar CDFs are shown in Fig. 7.15(b). The predictive scheme is slightly better, constraining 3% – 5% more H2M packets in 200 – 500 μs latency range compared to AIBA scheme. In both predictive and AIBA schemes, more than 80% of H2M packets wait less than 500 μs in the buffer, and 98% of packets are transmitted within 1 ms. At a high network load, i.e., 0.8 in Fig. 7.15(c), the latency distribution is nearly uniform within 1 ms under predictive scheme. In comparison, the distribution in baseline scheme exhibits a decreasing trend. This is because, although the average polling cycle duration approaches 1 ms in high load region, the variation of the duration is smaller in predictive scheme due to the surplus bandwidth allocated. Nevertheless, the CDFs of the baseline and predictive scheme are similar as shown in Fig. 7.15(c).
In this case, the AIBA scheme clearly shows a higher percentage of packets with latency < 500 μs. This is again because the correlation between control and feedback is effectively exploited in allocating bandwidth for feedback traffic. The packet latency statistics in Fig. 7.15 support our previous latency analysis and understandings on each scheme.

7.6 Summary

Human-controlled and haptic feedback data in emerging Tactile Internet human-to-machine (H2M) applications require stringent low latency in their transmission. Understanding the traffic characteristics of these low-latency applications is vital in innovating network control and resource allocation strategies to meet their latency demand. In this chapter, we have presented our experimental study on the characteristics of human control and haptic feedback traffic in H2M applications, and investigated novel bandwidth allocation schemes in supporting converged H2M application delivery over access networks. We introduced our haptic experiment system and the developed three types of H2M applications, including hand movement, touching, and object grasping. Then, we presented our analyses and understandings on control and feedback traffic in our experiments. In particular, we fitted different statistical distributions to analyse packet inter-arrival times, and reported that Generalised Pareto distribution was shown to be the best fit. Then, we paid special interests to the time-domain correlation in control and feedback traces, and reported a high cross-correlation between control and feedback traffic. This cross-correlation was defined as traffic causality in this chapter. Based on these understandings on H2M traffic, we further explored how to innovate bandwidth allocation schemes for emerging low-latency H2M applications converged over access networks.

We proposed the AIBA scheme to reduce latency for H2M application by harnessing the reported traffic causality between control and feedback. In AIBA scheme, the CO uses an ANN to estimate feedback bandwidth, and then pre-allocates it for feedback when forwarding the control traffic. Simulation validated the accurate bandwidth estimation achieved by ANN, and the effectiveness of the AIBA scheme in reducing latency for feedback traffic. Results showed that compared to the baseline scheme, AIBA scheme achieved 40%
– 60% latency improvement as it expedited feedback delivery by eliminating the bandwidth report-then-grant process baseline scheme. Compared to the predictive schemes, up to 40% latency improvement was achieved in the AIBA scheme. Importantly, the latency performance of the AIBA scheme was less affected by bandwidth estimation error and the increasing number of ONUs in a network. The latency of control traffic was shown to be not affected by the feedback bandwidth estimation and allocation in the AIBA scheme. These results validated the effectiveness of allocating feedback bandwidth exploiting the traffic causality reported.

With a proof of concept attained for the AIBA scheme, we investigated the performance of the existing schemes and AIBA scheme in supporting aggregated low-latency H2M applications and conventional content applications. We presented the detailed schematic of implementing AIBA scheme in the presence of both H2M and content traffic. Prioritising H2M traffic over content traffic was discussed for both existing and the proposed AIBA schemes. Via extensive simulations using experimental traffic traces, the capability of each scheme in reducing and constraining latency for low-latency H2M applications was comprehensively evaluated. Simulation results validated that compared to the existing schemes, the AIBA scheme effectively improved latency for H2M applications under all network load situations. Several understandings were highlighted: (a) the report-grant process is the main latency cause in the baseline scheme in light network loads; (b) the predictive schemes reduces the latency for H2M applications, depending on its estimated bandwidth and network loads; (c) the AIBA scheme adaptively allocates bandwidth for feedback based on the control traffic, thereby reducing latency for H2M applications over light to high network loads.

In this chapter, we have delved into critical open areas on understanding H2M traffic and rethinking bandwidth allocation schemes for emerging low-latency H2M applications. The reported experimental findings in this chapter consolidated the theoretical ground, results and conclusions in the previous chapters. Overall, this thesis has attempted to address the latency challenge in body area networks, and heterogeneous access networks for emerging H2M applications by improving bandwidth resource allocation in these network
segments. In the Tactile Internet evolution, continuous research efforts and innovative solutions in diverse research directions are demanded. In the next chapter, we will summarise the key contributions in Chapters 3 – 7 and discuss potential future directions of our work reported in this thesis.
Chapter 8
Conclusions and Future Directions

8.1 Introduction

Current communication networks are evolving from connecting only human beings to supporting machine-to-machine communications through Internet-of-Things (IoT) applications. The next stage of IoT applications is driven by the emerging Tactile Internet, which sees a plethora of real-time and remotely-controlled human-to-machine (H2M) applications. The Tactile Internet defines an ultra-reliable and low-latency network for manipulating and/or perceiving both virtual and real objects in a remote environment. In particular, with millisecond-low end-to-end latency specified as its target, the Tactile Internet envisions H2M applications with haptic capability, whereby human operators can ‘feel’ tactile and kinetic sensations when controlling remote objects and can immersively interact with the environment. Seemingly, human and machine/robot interaction mutually benefits IoT functionalities and human communication experiences. However, the latency in current communication systems is the main bottleneck in realising H2M applications. Stringent low end-to-end latency between master human operators and slave machines/robots within 1 – 10 ms is demanded for effective tactile-haptic interactions in H2M applications. In addressing this latency challenge, advanced technology, innovative network architecture solutions, and resource allocation strategies need to be investigated thoroughly.

The main focus of this thesis has been to improve the latency performance of the fundamental building block networks in the Tactile Internet, more specifically, the wireless body area networks (WBANs) and heterogeneous optical and wireless access networks. WBANs, comprising miniaturised sensors and actuators on or around human body, are vital for sensing human action and control, as well as enabling human perception of tactile and kinetic sensations. As sensors and
actuators has limited battery life, WBAN medium access control (MAC) layer channel access mechanisms are mainly designed to maximise their energy-savings. As such, lowering the latency remains an open challenge in existing WBAN systems. To this end, this thesis paid special attention to a recently-proposed WBAN system, the smart body area network (SmartBAN), and critically re-examined both its latency and energy performances. The thesis proposed novel transmission frameworks to achieve low-latency and high energy-savings for SmartBANs. Since the existing WBAN MAC layer designs share many commonalities with SmartBAN, the conclusions arising from our SmartBAN studies also helps in improving MAC layer solutions of existing WBANs.

In supporting remote service/data access for WBAN users, we considered converged WBAN applications delivery over heterogeneous optical and wireless access networks. For integrated passive optical networks (PONs) and wireless local area networks (WLANs), we examined the latency performance of existing MAC layer channel access mechanisms in WLANs and dynamic bandwidth allocation (DBA) schemes in PONs. In particular, machine learning (ML) techniques were exploited to comprehensively investigate the impact of different bandwidth allocation decision parameters and key network features on latency. By training a deep neural network (DNN) to learn the dependency between latency and bandwidth allocation decisions, we analysed in this thesis the limitations of existing bandwidth allocation solutions in WLAN and PON, and presented the latency improvement by optimising bandwidth allocation decisions corresponding to different network features in practical scenarios.

As current communication networks are expected to support aggregation of both emerging H2M applications and conventional content-centric applications, it is important to understand the traffic characteristics of H2M applications, their aggregations over access networks, and explore novel bandwidth allocation schemes to ensure low latency for H2M applications. For this purpose, we focused on bursty nature of H2M traffic reported in existing studies on H2M traffic. Machine learning-based predictive DBA (MLP-DBA) proposed in this thesis estimates bandwidth demand of integrated optical network units and wireless access point (ONU-APs) and adaptively allocates bandwidth among the ONU-APs.
8.2 Summary of Key Contributions

Specifically, we compared the performance of different statistical methods and ML techniques in estimating bandwidth demand. This justified the rationale behind exploiting an artificial neural network (ANN) in predicting ONU-APs bursty status and estimating bandwidth demand in our proposed MLP-DBA. Since the ANN can accurately predict the bursty status of ONU-APs, the MLP-DBA effectively reduces the latency by adaptively allocating bandwidth to ONU-APs that are receiving packets or having accumulated packets in the queue.

As the Tactile Internet and H2M applications remain in their early stages of development, the unique characteristics of human control and haptic feedback traffic in H2M applications have yet to be fully investigated. To address the lack of research in H2M traffic, this thesis experimentally investigated on the nature of human-controlled and haptic feedback traffic in a haptic teleoperation system. Using our experimental traffic traces, we performed: (a) statistical analysis on control and feedback packet inter-arrival times; and (b) time-domain analysis on the correlations exhibited by control and feedback traffic traces. The high cross-correlation between control and feedback traffic highlighted by the simulation results is defined as traffic causality in this thesis. Exploiting this characteristic, we proposed an artificial intelligence-facilitated interactive bandwidth allocation (AIBA) scheme to reduce the latency for inter-ONU H2M applications in PONs. The proposed AIBA scheme uses an ANN to estimate feedback bandwidth and pre-allocates this bandwidth when forwarding the control traffic. This approach effectively expedites feedback delivery since feedback packets can be timely transmitted in responding to control packets. Furthermore, this thesis discussed the priority differentiation between H2M and conventional content applications in the proposed AIBA scheme and existing schemes. Via extensive simulations, this thesis comprehensively analysed the capability of the investigated schemes in reducing and constraining latency for H2M applications. Overall, the studies reported in this thesis enhance our understandings on H2M traffic characteristics and potential changes to be made to existing bandwidth allocation solutions for future applications. The next section presents the contributions achieved in each chapter of this thesis.

8.2 Summary of Key Contributions
In Chapter 1, we introduced the evolution of the Tactile Internet and emerging H2M applications. We highlighted the enabling technologies for the Tactile Internet and the importance of reducing the latency in current communication networks to support the H2M applications. Chapter 1 summarised our objectives, original contributions, and the outline of this thesis. In Chapter 2, we overviewed existing MAC layer solutions of key underlying networks, namely, WBANs, heterogeneous PON and WLAN access networks. This chapter explained in detail the evolution of WBANs and compared existing IEEE 802.15.4, IEEE 802.15.6, and SmartBAN MAC designs. This was followed by a detailed account of heterogeneous PON and WLAN in supporting low-latency H2M applications. More specifically, this chapter reviewed the existing WLAN MAC mechanisms and DBA schemes in PON. We also overviewed existing literature that exploited ML techniques for bandwidth resource allocation over heterogeneous PON and WLAN networks. The characteristics of popular ML techniques and their applications in bandwidth resource allocations were summarised.

After an overview of our motivations and existing literature which highlights the key questions that are to be addressed in this thesis, in Chapters 3, 4, 5, 6, and 7, we reported on our performance studies and proposed solutions for WBANs, WLANs, and PONs. In Chapters 3 and 4, we focused on the SmartBAN and presented a comprehensive study on its latency and energy performance. We modelled the latency and energy-savings in both uplink and downlink directions. In Chapter 3, we discussed two types of common traffic, periodic monitoring and emergency (EM) traffic, in SmartBAN uplink and analysed the impact of SmartBAN MAC timing parameters on their latency. Based on our analysis, we proposed a time-optimised MAC, which provided a criterion in determining a hybrid MAC frame that reduces the latency and energy consumption in SmartBANs. In Chapter 4, we comprehensively investigated SmartBAN downlink latency and energy performances. In our investigation, we developed an analytical model to evaluate the default supplementary downlink transmission mode (SDM) defined by the SmartBAN MAC, and reported the limitation in achieving low latency due to the MAC frame design. Based on this important finding, we proposed the improved SDM (ISDM) to address the limitation in SDM, and two novel downlink transmission modes, namely, limited-exhaustive downlink transmission mode (LEDM) and fully-exhaustive downlink transmission mode (FEDM)
transmission mode (FEDM). Then, we formulated analytical models that accurately evaluate the latency and energy consumptions of LEDM and FEDM schemes. Finally, followed by better understandings on SDM, ISDM, LEDM, and FEDM, we proposed a novel downlink transmission framework that can flexibly accommodate downlink latency and energy consumptions for latency-constrained applications.

Overall, in Chapters 3 and 4, we validated the effectiveness of our proposed analytical models and the proposed uplink time-optimised MAC framework and downlink transmission framework for SmartBANs via extensive simulations. Concluding this work, we also presented extended discussions and modelling of the SmartBAN round-trip latency at the end of Chapter 4. This discussion details how MAC timing parameters can be adjusted to minimise the round-trip latency to supporting emerging H2M applications. In summary, the work presented in Chapters 3 and 4 consists of: (a) the first study that accurately models and evaluates SmartBAN latency and energy performances in meeting the latency demand for H2M applications; (b) proposing a novel criteria on how to determine a hybrid MAC frame in achieving low-latency and high energy consumption considering both uplink and downlink transmissions; and (c) proposing a novel time-optimised MAC framework for uplink periodic monitoring and emergency traffic in SmartBANs; and (d) proposing novel downlink transmission modes and downlink transmission framework for SmartBANs.

In Chapter 5, we investigated the latency performance of heterogeneous PON and WLAN access networks in supporting WBAN to deliver e-health applications. For this purpose, we analysed the key parameters in the hybrid coordination function (HCF) in WLAN MAC and DBA schemes in PON that impact the end-to-end latency. In order to characterise the association between these parameters and the end-to-end latency, we presented our exploitation of a DNN to learn this dependency. The study in Chapter 5 presents the details of selecting DNN input and output features and the supervised training. Multiple network features are taken into account when training the DNN, and analysing the bandwidth allocation decisions using the trained DNN. We showed that with supervised training, the DNN is able to predict the end-to-end latency given any bandwidth value to be allocated in the WLAN HCF and PON DBA operation.
Therefore, the optimal bandwidth values that reduce the end-to-end latency of a heterogeneous network can be derived using the trained DNN in turn. Via extensive simulations, we compared the latency experienced by the packets when bandwidth is allocated using DNN-supervised bandwidth values and that of the conventional bandwidth allocation schemes. The simulation results show 40% – 70% reduction in end-to-end latency and effective elimination of latency deviation by allocating DNN-supervised bandwidth values.

In Chapter 6, we considered the bursty nature of traffic generated by aggregated H2M applications, and analysed the latency bottleneck caused by: (a) bandwidth contentions among multiple ONU-APs; and (b) the difficulty in estimating and adaptively allocating bandwidth for bursty arrivals to individual ONU-APs in existing DBA schemes. To address these challenges, in Chapter 6, we exploited an ANN at the CO to predict the bursty status of ONU-APs and estimate the bandwidth demand accordingly. In contrast to the conventional DBA schemes whereby the CO grants requested bandwidth to all ONU-APs directly, in our proposed MLP-DBA scheme, the CO adaptively allocates bandwidth to classified ONU-APs based on their bursty statues. The effectiveness of MLP-DBA schemes is validated via extensive simulations, which show that the proposed ANN achieves > 90% accuracy in predicting the bursty status of ONU-APs. Compared to existing DBA schemes, MLP-DBA scheme achieves latency improvement under all light to high network loads 0.1 to 1.

In Chapter 7, we reported our experimental study on H2M traffic that is performed to better understand the unique traffic characteristics of H2M applications. For this purpose, we developed three H2M applications based on a haptic teleoperation system. Based on our analysis on experimental traffic traces, we defined the traffic causality, which is a high cross-correlation between real-time human control and haptic feedback during H2M interactions. Motivated by this observation, we proposed the AIBA scheme that uses an ANN to estimate haptic feedback bandwidth based on bilateral control and feedback traffic and pre-allocates bandwidth for feedback when forwarding control traffic. By injecting the experimental H2M traffic traces into a simulation platform, we showed that the proposed ANN achieves more accuracy in estimating haptic feedback bandwidth compared to conventional statistical algorithms adopted in
predictive DBA schemes. The proposed AIBA scheme effectively expedites haptic feedback delivery by up to 60% compared to the baseline limited-service DBA scheme and up to 40% compared to predictive DBA schemes. Using these results, we further investigated priority differentiation of H2M and conventional content applications in existing schemes and the proposed AIBA scheme via extensive simulations. This study resulted in the following conclusions: (a) the report-grant process is the main latency cause in baseline DBA scheme in light network loads; (b) predictive DBA schemes reduces the latency for H2M applications, depending on its estimated bandwidth and network load; and (c) AIBA scheme adaptively allocates bandwidth for feedback based on the control traffic, thereby reducing latency for H2M applications over light to high network loads. In summary, our technical contributions in Chapter 7 resulted in: (a) the first study of the traffic causality in H2M applications; (b) proposal of the AIBA scheme in supporting low-latency H2M applications; and (c) comprehensive analysis of priority differentiation between H2M and content traffic in existing DBA schemes.

8.3 Future Research Directions

This thesis comprised comprehensive investigations of MAC layer channel access and bandwidth allocation solutions in WBANs, and heterogeneous WLAN and PON access networks to support low-latency H2M applications. Yet, many research questions related to underlying networks, machine intelligence and H2M traffic, still warrant investigation, as to be discussed in the following subsections.

8.3.1 Optimising WBAN MAC

In our SmartBAN performance study, we analytically investigated the dependency of SmartBAN latency and energy consumption on MAC layer timing parameters. Our focus was on mathematically modelling how the duration of different access periods impact the latency and energy consumption. For analytical simplicity and tractability, the dynamics of WBAN communication channels and uplink/downlink traffic associated with human activities were not considered in our modelling. However, since WBANs, including SmartBANs, are deployed on human body, their generated traffic, latency, and energy performance are impacted by human movements, activities, lifestyle, and habits. As such,
SmartBAN MAC layer solutions can be further improved by a context-aware MAC, where access mechanisms take contextual cues such as user’s activity, mental and physical status into account. Towards such context-aware MAC designs, our developed models in Chapters 3 and 4 could potentially be utilised to determine MAC timing parameters accordingly to the sensor/actuator traffic generated in different human movements and activity scenarios.

Further in our proposed frameworks, we mainly considered two common types of traffic, periodic monitoring and EM traffic, generated by bio-medical sensors. Since EM traffic generally requires lower latency, its transmission is prioritised over the monitoring traffic. With the development of e-health and H2M applications, WBANs will need to support various types of traffic, such as the monitoring and EM traffic, remote-controlled traffic and haptic traffic. It will be worthwhile investigating the diversity and prioritisation of traffic in WBANs. Specifically, in provisioning differentiated quality-of-service (QoS), existing WBAN standards have indicated different levels of user priorities (UPs). For example, the IEEE 802.15.6 WBAN MAC has defined eight UP levels and the SmartBAN MAC has specified four levels. However, the assignment of UP levels regarding the types of sensors/actuators and QoS demand of the traffic in different applications has yet to be investigated. In our study, we showed that flexibly exploiting a hybrid MAC frame, i.e., EM traffic transmitted using both contention and scheduled access periods, and periodic traffic transmitted in designated time slots, achieves low latency and high energy-savings in SmartBANs. In our future studies, we hope to incorporate more types of traffic and appropriate UP assignments into our modelling. Doing so would improve the criterion in determining a hybrid SmartBAN MAC frame to cater to the latency demand of different traffic whilst retaining high energy efficiency in SmartBANs.

8.3.2 Wavelength Allocation in Optical Access Networks

In this thesis, we investigated the latency performance of existing DBA schemes and proposed novel bandwidth allocation solutions to reduce latency in supporting low-latency applications over access networks. With the exponential growth of data traffic from diverse applications in our communication networks, strategic allocation of both wavelength and bandwidth needs to be investigated.
As proven in Chapter 5, bandwidth allocation decisions in PONs are closely related to multiple network features and traffic characteristics of the network. Similar analysis should be done on how to dynamically distribute both wavelength and bandwidth among ONU{s according to their arrivals and configurations such as the distance to the CO. In particular, our proposed DNN learning and decision-making model in Chapter 5 could be modified to investigate wavelength assignment to ONU{s or grouped ONU{s distinguished by critical networks features such as traffic loads. Moreover, future access networks are anticipated to cater to different types of applications such as content-centric applications, IoT applications, and emerging H2M applications, which have different QoS demands. Studies on wavelength and bandwidth resource allocation should take this crucial fact into consideration. In Chapter 7, we discussed priority differentiation between H2M and conventional content applications and compared the latency performance of several existing DBA schemes that prioritise H2M arrivals. Results showed an increased latency of H2M applications when the content traffic load grows. This latency increase can be addressed by appropriately grouping ONU{s for wavelength sharing based on their H2M and content arrivals, and optimising bandwidth allocation to ONU{s in each group. Therefore, adapting our DNN model for wavelength and bandwidth allocation in PONs is critical to investigate the impact of priority-differentiated traffic loads at ONU{s on wavelength and bandwidth allocation decisions. This investigation can potentially yield priority-differentiated resource allocation schemes that efficiently utilise network resource, in terms of wavelength and bandwidth, in meeting the QoS demands such as latency, throughput and reliability of diverse applications in future access networks.

8.3.3 Exploration of ML for Intelligent Resource Allocation

In Chapter 2, we summarised commonly-used supervised training-based ML techniques and how these techniques are used in existing studies to improve resource allocation decisions in access networks. We also presented our work in Chapter 6 that exploits ML techniques to predict bursty status and estimates bandwidth demand. Based on our simulation results, using ML was proved to improve existing bandwidth allocation solutions, and thereby the latency performance of the heterogeneous access networks. Although state-of-the-art
research has highlighted the benefits of using different ML techniques in network control and resource allocation, comprehensive analysis of which technique to select and the costs associated with different techniques such as computation time and memory usage have not been critically addressed in existing studies. As such, we hope to extend our studies on both costs and gains of using ML techniques for resource allocation in access networks.

In this thesis, our exploitation of ML mainly focused on neural networks that rely on supervised training. This requires training set collection and incurs time cost in training phase as presented in Chapters 5 and 6. When the supervised training is complete, prior knowledge of the network such as traffic load is required in yielding the optimal bandwidth decisions using the trained neural network. In our future studies, we hope to explore the use of more advanced ML techniques that do not rely on supervised training to facilitate intelligent network resource allocation. Here, we will consider reinforcement learning (RL) as a promising solution to overcome the above-mentioned shortfalls in supervised training-based techniques. RL focuses on the interactions between an agent and the environment. The agent reinforces its actions for more rewarding outcomes by strategically exploring and exploiting actions. A potential RL-based bandwidth allocation solution is to let the CO (as the agent) adaptively adjust bandwidth allocation decisions (actions) based on the latency experienced by the packets (rewards). In this way, supervised training and prior knowledge of the network environment are no longer needed. However, in RL, the exploration of different actions is necessary in determining an optimal action. As such, to apply RL to improve bandwidth allocation in reducing the latency over access networks, we consider it is worthwhile investigating the trade-off between exploration and exploitation of bandwidth allocation decisions by the CO.

8.3.4 H2M Traffic Modelling

The preliminary findings arising from our experimental studies on human control and haptic feedback traffic in our developed H2M applications, was reported in Chapter 7. In that study, we fitted the distributions of packet inter-arrival times observed in our experiment to suitable statistical models that can characterise H2M traffic. However, as existing research and our current
findings on H2M traffic are still limited, rigorous experimental studies using diverse H2M applications in different teleoperation systems are still required to validate current preliminary observations. As such, we hope to continue our H2M traffic analysis and modelling with more H2M experiments. Some of these diverse applications include teleoperations, virtual reality (VR) and augmented reality (AR) applications with tactile-haptic capability. An expected outcome from analysing the traffic characteristics in these diverse H2M applications lies in building synthetic H2M traffic sources that can be widely utilised in assessing and optimising existing networks for emerging H2M applications. Moreover, as highlighted in this thesis, H2M applications have not been widely deployed over communication networks compared to content-centric applications. Using synthetic H2M traffic sources, we are able to extend our analysis on the characteristics of aggregated H2M traffic over access networks. Based on the above-mentioned H2M experiments, it would also be worthwhile studying the similarities and differences between the traffic in H2M applications and that in content and machine-centric applications. Such understandings would be helpful in improving network resource allocation schemes to cater to different QoS demands of diverse applications as indicated in Section 8.3.2.

8.4 Conclusions

The next-generation communication and Internet era will see the emergence of diverse remotely-controlled H2M and tactile-haptic applications that require millisecond-low latency in their transmission. Current communication networks need to reduce their latency in order to realise these low-latency H2M applications. To this end, this thesis discussed crucial underlying networks, including WBANs and heterogeneous access networks, for future applications and presented our contributions in improving existing MAC layer solutions to reduce latency in these networks.

For WBANs, we focused on the SmartBAN system and developed analytical models in critically studying its latency performance towards supporting low-latency applications. Based on our modelling, in Chapter 3, we proposed time-optimised MAC to improve uplink latency and energy-savings in SmartBANs. In Chapter 4, we proposed novel downlink transmission modes for
downlink transmission in SmartBANs. The investigation on different downlink transmission modes yields a low-latency and high-energy-efficiency downlink transmission framework for SmartBANs. In reducing latency for heterogeneous WLAN and PON access networks, we presented in Chapter 5 our exploitation of a DNN in analysing and optimising bandwidth allocation decisions in existing bandwidth allocation schemes. We showed that by allocating the optimal bandwidth values derived using the trained DNN, the latency in WLAN and PON, and accordingly the end-to-end latency, can be effectively reduced compared to that in existing schemes. Existing bandwidth allocation solutions in access networks can be further improved for H2M applications by considering the unique characteristics of H2M traffic. In Chapter 6, we discussed the bursty nature of H2M traffic and proposed the use of an ANN to estimate bursty bandwidth demand of ONU-APs. Then, using the ANN, we proposed the MLP-DBA scheme to make adaptive bandwidth allocation by classifying ONUs based on the estimated bandwidth. Further in Chapter 7, we experimentally studied the human control and haptic feedback traffic in H2M applications and highlighted the traffic causality characteristic. Based on this important understanding, we proposed the AIBA scheme to pre-allocate bandwidth for haptic feedback responding to the control traffic to be forwarded, which is proved to effectively reduce latency for H2M applications over existing schemes. Finally, in this chapter, we conclude the technical contributions arising from our work and outlined the potential areas of research that are extended from the thesis.
References


2015, pp. 501-506.


[39] Enhanced Ultra-Low Power PHY, Document TS DTS/SmartBAN-007 V1.1.1,
[40] Low Complexity Medium Access Control (MAC) for SmartBAN, Document TS DTS/SmartBAN-005 V1.1.1, ETSI TC SmartBAN, Sophia Antipolis, France, Apr. 2015.


Networks: In the Era of Big Data and Beyond," *IEEE Reviews in Biomedical Engineering*, vol. 8, pp. 4–16, Apr. 2015.


2013, pp. 2443-2448.

[94] C. Coutras, "Real-Time Medical Data over WLANs," 2015 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, 2015, pp. 668-673.


[143] A.C. Müller, S. Guido, Introduction to Machine Learning with Python: A


[194] K. Saito, H. Nishiyama, N. Kato, H. Ujikawa, K. Suzuki and N. Yoshimoto,
"Improving the Performance of FiWi Networks through Collaboration between ONU and APs," *2013 IEEE Global Communications Conference (GLOBECOM)*, Atlanta, 2013, pp. 4830-4834.


[202] CyberGlove System User Guide [Online]. [http://www.cyberglovesystems.com/user-guides/trn03w14ukdonygqrhuc0hh6i0dume](http://www.cyberglovesystems.com/user-guides/trn03w14ukdonygqrhuc0hh6i0dume)


[204] P. Assimakopoulos et al., “Statistical Distribution of Packet Inter-Arrival


Minerva Access is the Institutional Repository of The University of Melbourne

Author/s: Ruan, Lihua

Title: Low-Latency Communication over Heterogeneous Fiber-Wireless Networks for Human-to-Machine Applications

Date: 2019

Persistent Link: http://hdl.handle.net/11343/233705

File Description: Final thesis file

Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.