Resource Management in Future Mobile Networks

From Millimetre-Wave Backhauls to Airborne Access Networks

Rui Li

Doctor of Philosophy

Institute of Computing Systems Architecture

School of Informatics

University of Edinburgh

2019

Abstract

The next generation of mobile networks will connect vast numbers of devices and support services with diverse requirements. Enabling technologies such as millimetre-wave (mm-wave) backhauling and network slicing allow for increased wireless capacities and logical partitioning of physical deployments, yet introduce a number of challenges. These include among others the precise and rapid allocation of network resources among applications, elucidating the interactions between new mobile networking technology and widely used protocols, and the agile control of mobile infrastructure, to provide users with reliable wireless connectivity in extreme scenarios.

This thesis presents several original contributions that address these challenges. In particular, I will first describe the design and evaluation of an airtime allocation and scheduling mechanism devised specifically for mm-wave backhauls, explicitly addressing inter-flow fairness and capturing the unique characteristics of mm-wave communications. Simulation results will demonstrate 5$\times$ throughput gains and a 5-fold improvement in fairness over recent mm-wave scheduling solutions. Second, I will introduce a utility optimisation framework targeting virtually sliced mm-wave backhauls that are shared by a number of applications with distinct requirements. Based on this framework, I will present a deep learning solution that can be trained within minutes, following which it computes rate allocations that match those obtained with state-of-the-art global optimisation algorithms. The proposed solution outperforms a baseline greedy approach by up to 62%, in terms of network utility, while running orders of magnitude faster. Third, the thesis investigates the behaviour of the Transport Control Protocol (TCP) in Long-Term Evolution (LTE) networks and discusses the implications of employing Radio Link Control (RLC) acknowledgements under different link qualities, on the performance of transport protocols. Fourth, I will introduce a reinforcement learning approach to optimising the performance of airborne cellular networks serving users in emergency settings, demonstrating rapid convergence (approx. 2.5 hours on a desktop machine) and a 5dB improvement of the median Signal-to-Noise-plus-Interference-Ratio (SINR) perceived by users, over a heuristic based benchmark solution. Finally, the thesis discusses promising future research directions that follow from the results obtained throughout this PhD project.

Lay Summary

High-speed wireless connectivity has become an essential resource for the human society. With growing demand for Internet access, mobile networks are evolving towards a fifth generation (5G), aiming to support a multitude of application scenarios with high quality of service guarantees, including greater capacity, seamless connectivity, and shorter communication delay. Advances in networking technologies such as millimetre-wave (mm-wave) communications, network slicing, and self-deployable emergency networks show great potential to make 5G a reality. However, these emerging technologies pose several challenging research questions concerning the optimisation of the allocation of resources. This includes how to determine how much rate should be allocated to traffic flows as they traverse backhaul networks and how much airtime should be assigned to each flow on different links along the paths taken, when links employ mm-wave technology for transmission, the data services accommodate have different performance objectives, or certain fairness guarantees must be met. In contrast, with extreme cases such as emergency networking where existing deployments are damaged or do not provide coverage, the essential resource is network connectivity, and the pressing question is how to optimally control the mobility of substitution infrastructure (e.g. airborne base stations) to guarantee user connectivity when these move as well. This thesis tackles these resource optimisation problems in the envisioned 5G settings, dealing with issues specific to the new technologies, in scenarios including wireless backhauling (i.e. connecting wireless points of access to the Internet without cabled infrastructure) and emergency airborne networks, with the goal of achieving the best performance.

In multi-hop mm-wave backhaul networks that mainly support data intensive applications, it is crucial to provision the network resources fairly among data traffic flows, so that all users experience similar performance, despite various dissimilarities in the radio channel quality. Towards this end, the thesis first presents an airtime allocation and scheduling mechanism that achieves fair flow rate allocations. Via simulation based evaluation, I will show that the proposed solution significantly outperforms the state-of-the-art in terms of total network throughput and fairness among data flows traversing the network.

On the other hand, multi-service networks are expected to support applications with different performance requirements. For example immersive experiences are bandwidth intensive, tele-operation of robots is delay sensitive, and Internet of Things ap-

plications can be satisfied by best-effort services. To meet these diverse requirements simultaneously, as well as to achieve the highest utility of the network, I will formulate a network utility optimisation problem that takes into account customisable utility functions for different 5G applications. I will then prove that such optimisation is difficult to solve, and argue that existing methods are either time consuming or unable to obtain the best, i.e. globally optimal, solution. To address this, I propose a novel approach based on a deep neural network, which can find accurate data rate allocations for different traffic flow mixes in a timely manner.

In addition, the thesis investigates the performance of the widely used Transport Control Protocol (TCP) in the latest cellular network standard, i.e. Long-Term Evolution (LTE). The study provides insights into the interactions between TCP and LTE, and I will show that, despite good link conditions, the acknowledgement mode (AM) in the radio link control layer contributes to noteworthy communication overheads. However, the AM scheme can help recover corrupt data packets when the channel quality is poor, which often happen when the user is located at the cell edge.

In emergency scenarios, wireless connectivity is an essential resource for rescue services, civilian operations, as well as for the citizens, yet substitution of unavailable network infrastructure is often challenging due to difficult terrain and environment conditions. In this case, flexible deployment of mobile connectivity can be enabled with aerial base stations. However, to provide the users with good connectivity, requires accurate and rapid mobility control of multiple base stations mounted on Unmanned Aerial Vehicles (UAVs). Towards this end, I will present a machine learning approach that tackles the challenges associated with the multi-UAV control task, including uncertainly in users' movement and the random nature of wireless channel quality. Importantly, the proposed approach does not rely on explicit knowledge of the wireless channel behaviour or user mobility models. I will demonstrate that this solution converges fast and provides significant improvements in the signal quality perceived by the users.

Based on the results obtained and insights uncovered, I will close the thesis by discussing several interesting research directions that are worth addressing in the future.

Acknowledgements

First and foremost, I want to express my deepest gratitude towards my principal supervisor, Dr Paul Patras, who spared no effort to provide prudent guidance on my research and career development throughout my journey of pursuing this PhD. I thank my examiners, Dr Mahesh Marina and Dr Francesco Gringoli for their valuable feedback, which help me prepare the final version of this thesis. I thank the amazing people I have had pleasure to work with, including my colleague Chaoyun Zhang at Informatics Edinburgh, Prof John S Thompson and Dr Pan Cao from IDCom, School of Engineering, Edinburgh, Dr Razvan Stanica and Prof Fabrice Valois of INSA Lyon/INRIA, France, Prof Maziar Nekovee and Dr Mehrdad Shariat of Samsung Research UK, Gek Hong Sim of TU Darmstadt, Germany, Cristina Cano of Universitat Oberta de Catalunya, Spain, Dr David Malone of Maynooth University, Ireland, and Prof Joerg Widmer of IMDEA Networks, Spain.

I want to thank the School of Informatics and those who contributed to making the Forum a particularly open, welcoming, and friendly environment for informatics researchers.

Pursuing a PhD is particularly challenging for mental health, a recent study revealing that graduate students report significantly higher rates of anxiety and depression.[1] I therefore find myself extremely lucky to be surrounded by many wonderful people whom I can call my friends, in Edinburgh and around the world, who have made the past four years an unique and joyful experience. Particularly, I would like to mention those I got to know from the MF1 lunch club: Yota Katsikouli, Andrew McLeod, Kate Haag, Janie Sinclair, Maria Astefanoaei, Akash Srivastava, Sam Ribeiro, Natalia Zoń, Valentin Radu, Alex Dawson, and Galini Tsoukaneri. I am grateful for those roboticists and friends for the lovely time we spent together outside of work: Henrique Ferrolho, Xinnuo Xu, Yiming Yang, Ruiqiu Wang, Zhibin Li, Kate Uzar, Vlad Ivan, João Moura, Qingbiao Li, Lei Yan and many more from IPAB. I thank those I shared my office and nerdy jokes with: Chris Vasiladiotis, Jörg Thalheim, Maurice Bailleu, Christof Schlaak, Praveen Tammana, and Andew McPherson. I thank Yo Li for being an almost-sister and for her company during my ups and downs, Lin Zhao, Yu Wang, Berlin Liu, Hongjian Dai, Yueying Lu, Linli Ma, and Yixin Shi for decade-old friendships that continue to grow. My appreciation travels to Duyao Zhang and Jie Shen for the good times we spent together.

[1] https://www.nature.com/articles/d41586-018-03803-3

Special thanks to Ludovica Vissat, Wolf Merkt, and Stan Manilov for their inspirations, support, friendship, and love during the most challenging year.

Finally, I thank myself for being determined, courageous and resilient, and these nice facts largely come from the family who gave birth to and raised me up. I thank my parents and close family members for doing their best in every possible way to support my education and development as a person, giving me the freedom to be myself, and trusting me to make my own decisions.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified. Part of the materials in this dissertation have been published in the following publications, where the contribution from fellow PhD student is limited to setting up simulation environment:

1. **Rui Li** and Paul Patras, "*WiHaul: Max-Min Fair Wireless Backhauling over Multi-Hop Millimetre-Wave Links*", the 3rd ACM HotWireless, collocated with MobiCom, New York City, NY, USA, Oct. 2016

2. **Rui Li** and Paul Patras, "*Max-Min Fair Resource Allocation in Millimetre-Wave Backhauls*", accepted for publication in IEEE Transactions on Mobile Computing, May 2019

3. **Rui Li**, Chaoyun Zhang, Pan Cao, Paul Patras, and John S. Thompson, "*Delmu: A Deep Learning Approach to Maximising the Utility of Virtualised Millimetre-Wave Backhauls*", in International Conference on Machine Learning for Networking, Paris, France, Nov. 2018

4. **Rui Li**, Mehrdad Shariat, and Maziar Nekovee, "*Transport Protocols Behaviour Study in Evolving Mobile Networks*", Recent Results, ISWCS, Poznan, Poland, Sept. 2016

5. **Rui Li**, Chaoyun Zhang, Razvan Stanica, Fabrice Valois, and Paul Patras, "*Learning Driven Mobility Control of Airborne Base Stations in Emergency Networks*", in Workshop on AI in Networks, Toulouse, France, Dec. 2018

(*Rui Li*)

To world peace.

Table of Contents

Chapter 1

Introduction

The next generation of mobile networks, commonly referred to as 5G, will embrace pervasive connectivity and accommodate diverse types of user devices and applications, which will pose distinct performance requirements. While key technological advances in network programmability enable network virtualisation, and millimetre-wave (mm-wave) backhauling has brought multi-Gbps data rates closer to practice, several important research questions remain to be rigorously addressed before the 5G enabling technologies can be deployed and fully exploited.

This thesis approaches network resource management in the context of 5G from multiple angles. I will start with mm-wave multi-hop backhauling scenarios, where I highlight the challenge posed by the employment of directional antenna patterns, which is required to overcome the severe signal attenuation at mm-wave frequencies and to allow for agile deployment of base stations with mm-wave backhauling functionality. I further identify fairness issue in throughput oriented backhauling scenarios, where rigorous design of airtime allocation is essential in order to achieve carrier-grade performance and exploit the full potential of the mm-wave airtime resources. Towards this end, I devise the first max-min fair rate allocation and airtime scheduling mechanism for mm-wave backhauls, which is compatible with scheduled access regimes specified by existing standards or on-going mm-wave standardisation efforts including IEEE 802.11ad, IEEE 802.11ay, and 5G new radio (NR). This study primarily concerns data intensive applications. As the discussion of 5G key performance indicators (KPI) continues with the development of this thesis, and thanks to the recent advance in network programmability and the emerging network slicing paradigm, 5G networks will be able to support diverse use cases that require distinct performance metrics. As such, I am the first to investigate utility fairness in rate allocation for individual flows

generated by applications with different requirements and present a utility framework for 5G networks, which is generally applicable but particularly useful for mm-wave backhauls. I will show that optimising an arbitrary combination of different types of utility functions is a hard problem, given the non-convexity and high-dimensionality of the objective; I will therefore present a supervised learning approach to optimise such mixed utilities. Taking a complete system level view of future mobile systems, I will then investigate in depth how legacy transport protocols interact with the current 4G Long-Term Evolution (LTE) networks and how different configurations impact on the overall throughput performance under various channel conditions. Specifically, I study the different settings of the radio link control (RLC) layer under different data and control channel conditions, particularly looking into the performance of transport protocols at the cell edge in comparison with those at the cell centre. I will complete this study by looking at providing wireless connectivity as a resource under extreme circumstances, i.e. the agile deployment and control aspects of emergency cellular networks carried on unmanned aerial vehicles (UAVs). Thereby, I will tackle with a deep reinforcement learning approach the challenging problem of mobility management of such airborne networks, with the goal of optimising user connectivity.

In the following sections of this chapter, I will present in details a number of important research challenges towards resource management in the future generation of mobile networks, and I will highlight contributions made by this thesis towards addressing some of the most pressing issues.

1.1 Key Research Questions

To fulfil the potential of 5G networks while employing new technologies such as mm-wave communications, several key technical challenges must be addressed. These span physical layer (PHY) optimisation, user multiplexing and scheduling, reliable end-to-end networking over wireless mobile networks, cross-layer design, and many more.

1) Beamforming Mechanisms and Codebook Design for Mm-Wave Systems. Multiple input multiple output (MIMO) signal processing tailored to the particularities of mm-wave frequencies is critical for enabling multi-Gbps link rates, yet differs substantially from techniques that target lower frequencies (Heath et al., 2016). Specifically, practical considerations such as power consumption and circuit technology bring new hardware constraints. High resolution analog-to-digital converters are expensive and

power-consuming (Alkhateeb et al., 2014b), hence hybrid beamforming architectures and the use of low-resolution analog-to-digital converters are explored. Codebook design, in particular, is critical in enabling rapid and accurate beamforming and subsequently allows for high data rate. This challenging topic has been so far well studied in e.g. (Wang et al., 2009; Lee and Ko, 2011; Li et al., 2013; Zhou and Ohashi, 2012), addressing the important issues of power allocation and signal-to-interference-and-noise-ratio (SINR) optimisation from multiple different approaches. Multi-user MIMO mm-wave beamforming in cellular and Wi-Fi networks opens another area of research (Alkhateeb et al., 2015; Choi, 2015; Chen et al., 2016; Li et al., 2017b).

2) Channel Measurement and Modelling Techniques for Mm-Wave MIMO Systems. Signal processing in mm-wave communication often requires accurate channel models; propagation in mm-wave spectrum is however complicated (Rappaport et al., 2014; Neil et al., 2017). For example, diffraction patterns can be smaller due to a smaller Fresnel zone, while penetration loss is much higher in these bands. Moreover, the channel environment can vary significantly across different use cases, e.g. indoor personal area networks, densified cellular (including access, backhaul, and even self-backhaul), or vehicular networks with high user mobility. On one hand, channel measurement based analysis in mm-wave bands in different environments is needed (Rappaport et al., 2013b). On the other hand, research effort is required towards mathematical modelling of mm-wave radio propagation based on knowledge of the propagation environment. These issues have been well pursued by previous research effort (Gustafson et al., 2014; Blumenstein et al., 2014; Qian et al., 2015; Samimi and Rappaport, 2016; Peter et al., 2016; Kim et al., 2017).

3) Resource Allocation in Mm-Wave Backhauls. In response to accelerating mobile traffic demands, cell densification is a first step towards substantially extended capabilities of current mobile network infrastructure (Bhushan et al., 2014). This, however, entails revisiting existing backhauling practices, in order to be able to transfer vast volumes of data between the access and core networks. In particular, the cost of deploying traditional, fibre-based backhauls surges with network density, whilst reconfiguration of such solutions is limited. Wireless alternatives have thus far been confined to microwave spectrum (0.3–30GHz) of restricted capacity and already overcrowded with numerous applications, e.g. Wi-Fi, cellular access, and RADAR. The mm-wave band (30–300GHz) is in contrast underutilised and exposes considerably wider spectral re-

sources that could support an order of magnitude higher data rates (Rappaport et al., 2013b). Previous research efforts provide sufficient evidence that in order to mitigate characteristic severe signal attenuation and to harness the potential of mm-wave for small cell backhauling, highly-directional beamforming using multiple antennas and phase arrays is required (Hur et al., 2013; Alexandropoulos, 2017). Directionality intrinsically eliminates interference and enables better spatial reuse, though introduces the risk of *link blockage* due to moving obstacles and *terminal deafness*, i.e. receivers can hardly be aware of transmitters, unless their beams are mutually aligned (Roh et al., 2014). The latter is particularly problematic in deployments with small form factor base stations that serve large numbers of end-users over Wi-Fi/cellular and communicate with gateways using single mm-wave transceivers over multiple hops. Fig. 1.1 illustrates an example deployment of mm-wave small cell backhaul at street lamppost level where base stations (STAs) form very narrow beams towards each other, as depicted between STAs 3 and 4, to achieve multi-Gbps communication.

Backhaul solutions designed with legacy multi-hop wireless technology operating in sub-6GHz bands are inappropriate, given the unique properties of mm-wave communications. As the infrastructure has commercial value, it is essential to ensure resources are not left underutilised, while customers remain satisfied with the level of service provided. Several 5G standards define carrier-grade mechanisms that allow for precise scheduling (e.g. 3GPP NR (3GPP, 2018) and IEEE 802.11ad[1] with Service Period operation (IEEE 802.11ad Std., 2014)), yet the *airtime allocation and scheduling tasks, which are crucial for backhauling, are left open to implementation*.

4) Utility Optimisation in Mm-wave Backhaul Networks. 5G networks will accommodate a new wave of applications with distinct requirements (NGMN, 2015). For example, ultra-high definition video streaming and immersive applications (e.g. augmented reality/virtual reality, or AR/VR) typically demand very high data throughput. Autonomous vehicles and remote medical care are stringently delay-sensitive, and belong to a new class of Ultra-Reliable Low-Latency Communications (URLCC) services (Schulz et al., 2017). In contrast, IoT applications, including smart metering and precision agriculture, can be satisfied with a best-effort service. In order to simultaneously meet such diverse performance requirements, while enabling new verticals, mobile network architectures are adopting a *virtually sliced* paradigm (3GPP, 2017b).

[1]Note that, although the IEEE 802.11ad is primarily intended for single-hop wireless local area networks, this protocol could also be used for multi-hop solutions in unlicensed bands, e.g. 60GHz, serving community networks.

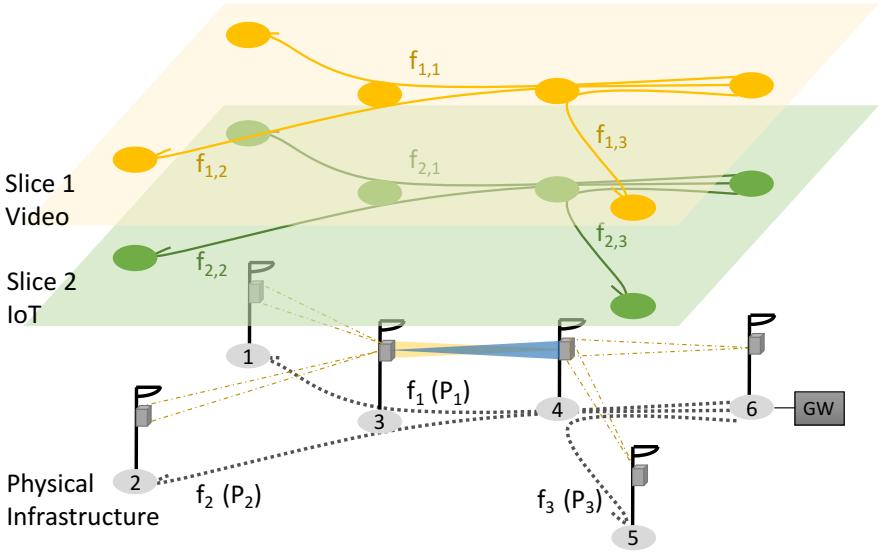


Figure 1.1: Example lamppost backhaul deployment of 6 base stations. STA 6 is connected to the gateway via wireline, while the other links operate in the mm-wave band. Aggregate flows traverse the network via different paths (P_1, P_2, and P_3). To achieve multi-Gbps communications, STAs form narrow beams, as depicted between STA 3 and STA 4 with yellow and blue shades. Terminal deafness occurs if STA 6 tries to talk to STA 4 at this point. Such a network can be logically sliced; here Slice 1 accommodates video streaming with sigmoid utility and Slice 2 carries Internet of Things (IoT) traffic with logarithmic utility. $f_{i,j}$, denotes a flow on slice i and path j.

The core idea of slicing is to partition physical network infrastructure into a number of logically isolated networks, i.e. slices. Each slice corresponds to a specific service type, which may potentially belong to a certain tenant operator.

Partitioning sliced mm-wave backhauls, and in general backhauls that employ any other communications technology, among traffic with differnt requirements, as in the example shown in Fig. 1.1, is essential for mobile infrastructure providers (MIPs). By and large, MIPs aim to extract as much value as possible from network resources due to the high capital expenditure and fierce market competition, yet achieving this in sliced backhauls is not straightforward. The notion of rate utility is widely used to quantify the worth of an allocation of resources to multiple flows. The question is: *what type of utility is suitable to such multi-service scenarios?* Logarithmic utility as proposed in (Kelly, 1997) has been adopted for elastic services and remains suitable for best-effort IoT traffic. On the other hand, applications such as video streaming typically throttle below a threshold, whilst an increase in service level is mostly imperceptible by users when the allocated rate grows beyond that threshold. Hence, the utility of such traffic can be modelled as a step-like sigmoid (Yin et al., 2015). Had there been real-time applications to accommodate, their utility is typically formulated

through polynomial functions (Fazel and Chiang, 2005; Wang et al., 2017). Further, in the case of traffic for which the MIP allocates resources solely based on financial considerations, a linear utility function can be employed. Hence, as the application scenarios diversify, a single type of utility cannot capture the distinct features of different services. However, combining all these utility functions may lead to non-concave expressions and *computing in a timely manner the optimal rate allocation that maximises their value becomes a challenging task.* Global search metaheuristics explore the feasible solution space intelligently to find global maxima (Ugray et al., 2007), yet often involve unacceptably long computational times. Thus they fail to meet 5G specific delay requirements in highly dynamic environments, where application demands change frequently. Greedy approaches can be used to overcome the runtime burden, though these will likely settle on sub-optimal solutions.

5) Routing in Mm-Wave Backhaul Networks. Data packet routing in mm-wave backhauls is challenging and traditional routing mechanisms may not be suitable. This is due to several factors. On one hand, the routing algorithm needs to satisfy the diverse and stringent 5G user requirements, including ultra high throughput and very low delay, which can even be conflicting with the flows produced by different applications that originate/terminate at the same base station. On the other hand, mm-wave particularities, such as vast link transmission rates that are however susceptibility to blockage, require routing solutions that are robust against frequent changes in link conditions. Moreover, a routing algorithm for mm-wave requires insights into per-hop queue-lengths, as the vast link rates can otherwise lead to excessive delays, buffer overflows and subsequently packet loss, if the transmission rate is overly aggressive from the perspective of buffer capacity at the receiving station. Addressing these problems may involve jointly taking into consideration scheduling (Seppänen et al., 2016; Seppnen and Kapanen, 2016; Yun et al., 2016; Niu et al., 2019) and end-to-end flow rate adaptation (Vu et al., 2018). As this has been a well explored topic in legacy settings as well as mm-wave, I instead pursue in this thesis a rate and airtime scheduling mechanism that is compatible with any stand-alone[2] routing protocols, and adapt quickly even when the routing algorithm changes.

[2]This excludes joint routing algorithms requiring MAC-layer modifications, e.g. joint routing and scheduling.

6) Transport Performance in Evolving Mobile Networks. The widely used transport control protocol (TCP) ensures reliable data exchange through error detection, packet reordering, and congestion avoidance mechanisms. As reported in (Huang et al., 2013), over 95% of the internet data traffic is based on TCP. However, when deployed in wireless networks, TCP's performance may vary considerably, depending on how the congestion control reacts to the unpredictable radio link environment, how much impact the protocol overhead has, and to what level the retransmission scheme recovers packets. *Understanding the throughput and delay experienced by users located at the cell edge, where the SINR perceived by the user equipment is often unsatisfactory, remains elusive.* Elucidating this is particularly important in order to achieve seamless cell edge performance, which is a KPI for 5G.

Moreover, in the control plane of current LTE systems, correct decoding of the Data Control Indicators (DCIs) depends on the correct interpretation of both Physical Control Format Indicator Channel (PCFICH) and Physical Downlink Control Channel (PDCCH) (3GPP, 2016a). However, PCFICH and PDCCH are not protected by any Automatic Repeat Request (ARQ) scheme. When an error occurs in these symbols, the data carried in a subframe will no longer be decoded. *It is therefore important to understand the impact of DCI errors on user experience and whether existing recovery mechanisms, such as the radio link control (RLC) layer acknowledged mode (AM), can mitigate this problem.*

7) Mobility Control of Emergency Airborne Networks. Wireless connectivity is particularly critical in emergency scenarios such as post-disaster rescue and recovery, where it conveys information relevant to people's life and property safety. For instance, communication failures led to an increase in the number of missing people in the California wild fires in November 2018 (BBC, 2018). Certain areas under extreme conditions, e.g. following earthquakes, floods, fire, and nuclear plant emergencies, are hardly accessible with legacy emergency cellular infrastructure carried on vans (i.e. cells-on-wheels). Meanwhile, following recent hardware and software advances, commercially available UAVs are being increasingly used for various applications, including aerial imaging and asset inspection. As a result, regulatory bodies, such as the Federal Aviation Administration (FAA), defined rules to enforce the safe operation of commercial UAVs (RCR Wireless, 2018). The telecom industry also shows growing interest in deploying UAV-mounted airborne base station (airSTA) for sporadic cellular services, with an emphasis on challenging use cases. For instance, following hurricane

Marie's devastation of Puerto Rico, AT&T obtained FAA approval to fly UAVs for temporary cellular coverage (FAA, 2017).

Providing wireless connectivity in a large area to a sizeable group of users, including citizens and rescue teams (e.g. police force, medical personnel, and firefighters), often requires more than one airborne vehicle with networking capabilities. In contrast to cellular networks, where the base station locations are fixed, airSTAs are mobile themselves. *Coordination among flying airSTAs and movement control is of paramount importance in order to provide the needed coverage and ensure sufficient and stable user data rates*, whilst any signal failure can be catastrophic to critical missions. Stochastic wireless channels and user movement uncertainty, however, render the airSTA mobility management task complex, involving an exponentially growing action space as the number of airSTAs increases. Traditional solutions, such as optimal control, require precise environment models, which are hardly obtainable in real-time and require strong assumptions that can compromise their usefulness. Heuristic alternatives only produce sub-optimal results.

1.2 Contributions

This thesis addresses research challenges facing resource management in future mobile networks as identified above. Figure 1.2 summarises the major contributions this thesis makes towards a subset of the challenges discussed above (namely 3, 4, 6, and 7), and illustrates how these contributions are related or complementary to each other, as I will further explain in this section.

1) Max-min Resource Allocation in Mm-wave Backhaul Networks. I will present a novel approach to jointly solve the airtime allocation and per-link scheduling of aggregate traffic flows, i.e. flow bundles that originate/terminate at the same base station, [3]which traverse multi-hop mm-wave backhauls. Chapter 3 of this thesis focuses on allocating resources at the medium access control layer (MAC) layer for general mm-wave systems. I do not make contributions in terms of PHY layer optimisation and argue that aspects including power allocation, codebook design, or beamform training can be largely decoupled from MAC operation; however, I explicitly take into account the distinct features of mm-wave technology, i.e. terminal deafness and susceptibility

[3]Hereafter, whenever there is no scope for confusion, I use the terms 'flow' and 'aggregate flow' interchangeably.

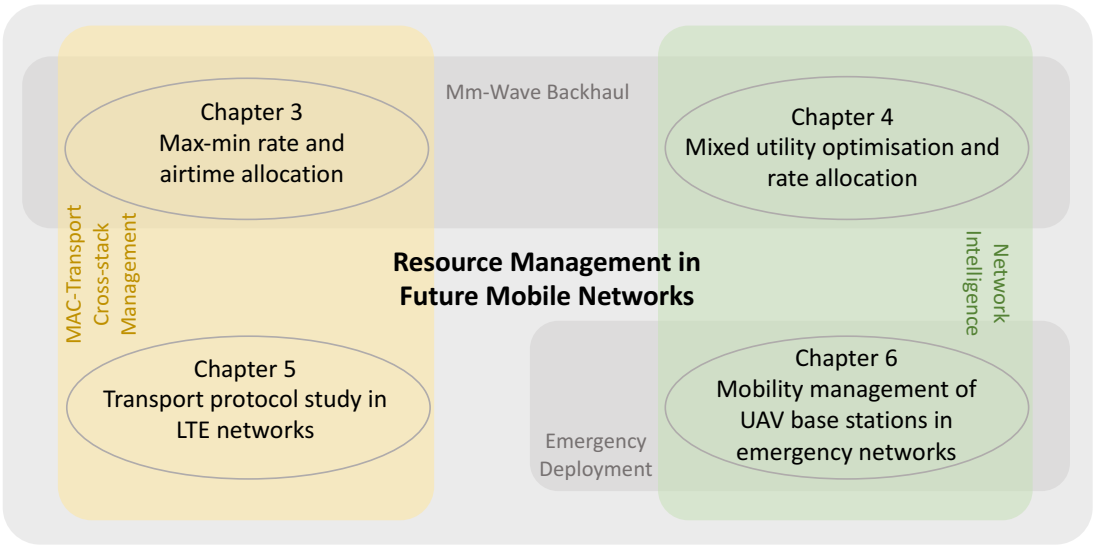


Figure 1.2: Diagramatic overview of the thesis contributions.

to link blockage, as well as realistic heterogeneous traffic demand regimes. The goal of this study is to *achieve a good balance between overall network throughput performance and inter-flow fairness*. That is, the revenue obtained from operating backhauls can be maximised, whilst aggregate flows encountering low capacity links and/or increased competition are not unnecessarily throttled (high quality of service). The focus is on providing small cell backhauling that can cater for real-time applications where latencies below tens of milliseconds are not perceivable by the user, instead of minimising latency, as required by ultra low-latency applications.

As such, Chapter 3 casts backhaul resource allocation as a max-min[4] optimisation problem with mm-wave specific terminal deafness and potential secondary interference, and traffic demand constraints. I demonstrate that a max-min fair solution exists and it is unique in scheduled-based multi-hop mm-wave networks. Subsequently, I propose WIHAUL, a backhauling scheme comprising a progressive filling algorithm that solves the max-min optimisation problem and computes per-hop airtime shares for each aggregate flow, and a light-weight scheduling procedure that works on top of any time-division multiplexing (TDM) protocol for mm-wave systems, enforces the computed airtimes, and coordinates multi-hop transmissions, enabling spatial reuse. I implement WIHAUL in the NS-3 simulator, building on preliminary mm-wave PHY

[4]I work with the max-min criterion instead of the popular Jain's fairness index, as I aim to avoid resource under-utilisation incurred when equalising throughputs. Instead, I seek to fulfil flow demands in increasing order, while sharing remaining network capacity among flows with higher demands. In the absence of an established quantitative measure of max-min fairness, we work with the notion of economic inequality (i.e. the Gini coefficient (Gini, 1921)) and extend a generic fairness model (Lan et al., 2010) to further quantify max-min fairness.

measurements and incorporating the IEEE 802.11ad specification, with extended functionality for multi-hop settings. Although this does not bear features specific to cellular systems, the MAC operation in the time domain is largely similar, thus the results obtained are relevant to such systems as well. The thesis evaluates the performance of the WIHAUL solution over different network topologies, link dynamics, routing paradigms, and traffic regimes, and demonstrates up to 5-fold throughput and fairness gains over existing access schemes.

2) Utility Framework and Optimisation in Sliced Mm-wave Backhauls. In order to maximise the utilisation of the mm-wave backhaul and to meet the divergent requirements of 5G applications, Chapter 4 puts forward a general utility framework for sliced backhaul networks, incorporating all known utility functions. Arbitrary combinations of such utility functions lead to highly non-convex sums and optimising these can be proven NP-hard. Henceforth, this thesis tackles the complexity of such maximisation (NUM) problem by proposing DELMU, a deep neural network model that learns the relations between traffic demands and optimal flow rate allocations. Augmented with a simple post-processing algorithm that ensures minimum service levels and admissibility within the network's capacity, the evaluation results show that DELMU makes close-to-optimal inferences, while requiring substantially shorter computation time, as compared to state-of-the-art global search and a baseline greedy algorithm, over which DELMU achieves 62% utility gains. The proposed approach can be trained within minutes and the millisecond scale inference times make the solution particularly suitable for highly dynamic traffic regimes in 5G networks.

In view of the current technological trends, I particularly focus on backhauls that operate in mm-wave bands. However, the proposed utility framework and deep learning approach are sufficiently general and can be applied to other systems employed in microwave or sub-gigahertz bands.

3) Transport Protocol Performance Evaluation in Future Cellular Networks. The evolution of current 4G system towards 5G calls for a more thorough understanding of the end-to-end performance of transport protocols, given that 5G NR will largely inherit the LTE structure. To this end, in Chapter 5, I investigate key metrics in the cellular network that affect directly the user experience, i.e. the end-to-end throughput, packet loss, under various round-trip-time (RTT) values and congestion control window (CWND) sizes for the case of TCP, different RLC protocol configurations,

and different channel qualities. The thesis identifies a number of performance issues in the interactions between the LTE stack and transport layer, especially in the case of poor channel quality. Particularly, inferior SINR values at the cell edge will cause significant throughput and delay degradation for both user datagram protocol (UDP) and TCP traffic. Although traffic running on top of UDP obtains marginally better throughput, it observes very high packet loss.

Further, this chapter reveals that commonly used transport protocols (i.e. UDP and TCP) are sensitive to control plane errors, which occur frequently at the cell edge in LTE networks. The RLC acknowledged mode can marginally overcome the protocol data unit (PDU) loss, yet introduces additional overhead and thus compromises throughput and delay performance under good link conditions.

4) Mobility Optimisation of Networks-on-Drones. Chapter 6 tackles the challenges of UAV mobility control in emergency networking deployed on drones, through a deep reinforcement learning approach. Particularly, a domain-specific reward function that encourages the UAV mobility control agent to provide high quality signal coverage to users was devised, and I employ an Asynchronous Advantage Actor-Critic (A3C) scheme to learn the optimal action policy via interaction with the wireless environment. This design is motivated by the rapid convergence requirements specific to emergency settings. Simulation results demonstrate that the proposed solution converges rapidly (within 4×10^5 steps) and, once trained, it makes accurate movement control decisions, outperforming a benchmark scheme that has perfect knowledge of the stochastic channel. More precisely, the proposed deep reinforcement learning solution obtains a 5dB median SINR improvement, while only requiring current location and association information.

The contributions of this thesis as detailed above complement each other, as depicted in Figure 1.2. Overall, I investigate resource optimisation and performance issues across different components of 5G mobile networks, including wireless backhauls (Chapter 3 and Chapter 4) and radio access (Chapter 6), the interactions between MAC and trasport protocols (Chapter 5), covering both commercial (Chapter 3 - Chapter 5) and emergency public service scenarios (Chapter 6). In terms of methodology, Chapters 3 and 4 tackle different fairness metrics, i.e. max-min fairness and respectively utility fairness, whilst Chapters 4 and 6 explore state-of-the art supervised learning and deep reinforcement learning methods for network resource management purposes.

1.3 Thesis Organisation

The rest of the thesis is organised as follows. In Chapter 2, I discuss relevant background and briefly survey previous work related to this PhD project. In Chapter 3, I delve into the details of the proposed max-min fair resource allocation solution for mm-wave backhauls. In Chapter 4, I introduce the utility framework for sliced backhauls and show the deep learning approach to optimising any combination of service utilities. In Chapter 5, I discuss thoroughly transport performance in cellular networks, and in Chapter 6, I detail the deep reinforcement learning approach to mobile base stations control. Finally, Chapter 7 gives concluding remarks and discusses possible future research directions.

Chapter 2

Background and Related Work

5G promises enhanced throughput performance, reduced latency and seamless connectivity to vast number of mobile devices in diverse use cases. Key enablers include mm-wave backhauling and network slicing, while employing machine learning algorithms is increasingly promising to achieve the 5G KPIs. This chapter reviews key technology advances and discusses previous work of direct relevance to this thesis, highlighting their limitations, which the thesis seeks to address.

2.1 Background

2.1.1 Mm-wave Standardisation and Industrial Efforts

Mm-wave frequency bands expose significantly wider spectral resources that enable up to multi-Gbps link rates. Regulatory bodies such as Ofcom in the UK have been encouraging nation-wide 5G trials in the mm-wave band (Ofcomm, 2018), and industry stakeholders have begun collaborating on building multi-Gbps mm-wave backhaul solutions in urban areas (e.g. Qualcomm's and Facebook's participation in the Terragraph project (Facebook Inc., 2018; Qualcomm Technologies Inc., 2018). Moreover, early customer premise equipments are developed to showcase mm-wave capacity (Qualcomm, 2019).

The 3rd Generation Partnership Project (3GPP) further promotes mm-wave technology through the specification of 5G NR in Release 15 (3GPP, 2018), with the first systems already being prototyped by Qualcomm (Qualcomm Technologies Inc., 2017). To give a bigger picture, Release 15 is the first 5G standardisation proposal by the

3GPP standards group, targeting non-standalone and standalone[1] operations of NR on frequency ranges below 6GHz and above 6GHz, supporting use cases including Enhanced Mobile Broadband (eMBB) and URLCC. Specific to MAC, 5G NR extends the LTE numerology to support diverse spectrum bands and deployment models, by allowing different types of sub-carrier spacing and slot lengths (3GPP, 2018). The 10ms frame structure of LTE with 1ms subframes is preserved. Moreover, NR specifies slot based scheduling that supports a fixed length of 14 orthogonal frequency-division multiplexing (OFDM) symbols, with the possibility of slot aggregation, and non-slot based scheduling (or mini-slot based) that allows for 7, 4, or 2 OFDM symbols. The airtime allocation and scheduling mechanism proposed in this thesis can work with both slot based and mini-slot based approach.

On the other hand, the IEEE 802.11ad standard specifies MAC and PHY protocols for directional multi-gigabit communications exploiting the vast spectral resources available in the 60GHz band (IEEE 802.11ad Std., 2014). The standard assumes mm-wave stations are equipped with phased arrays or a set of switched beam antennas, to form very narrow 'quasi-optical' beams that can mitigate the high signal attenuation associated with mm-wave frequencies. Building upon 802.11ad, the 802.11ay draft intends to multiply the mm-wave link capacity by employing 4-stream MIMO (Cerwall (ed), 2017). Medium access can be either contention based, whereby stations alternate between listening in quasi-omnidirectional fashion, and directional multi-gigabit (DMG) transmission mode that involves narrow beam forming; or scheduled, as the standard specifies a Service Period (SP) based mode of operation, by which SPs for different communicating pairs can be scheduled at the beginning of beacon intervals, which are followed by actual data transmissions, as illustrated in Fig. 2.1.

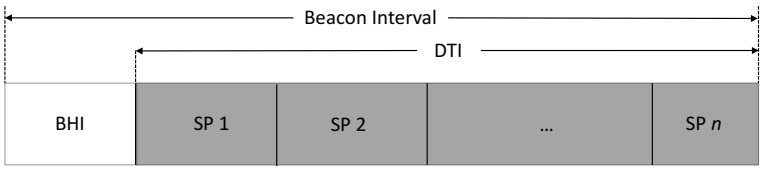


Figure 2.1: Typical beacon interval followed by IEEE 802.11ad communications with service periods (SPs). Transmissions are scheduled during the beacon interval header (BHI) and take place during the data transmission interval (DTI).

[1]Non-standalone NR uses LTE as control plane anchor, whereas standalone NR incorporates full control plane capability.

2.1.2 Network Slicing

Recent advances in network programmability enable to partition a physical network infrastructure into multiple logically isolated networks, which is also known as network slicing. Such virtual slicing paradigm promises improved network utilisation, flexibility, and cost-efficiency. Therefore, 5G standardisation efforts have also shed light upon the evolution towards sliced network architectures (3GPP, 2017b, 2016e), and industry has seen many early demonstrations of the slicing technology to support emerging 5G applications (ZDNet.co.kr, 2015; Huawei Co., Ltd, 2016; Telefonaktiebolaget LM Ericsson, 2018). Recent research suggests sharing mobile infrastructure among virtual operators can achieve considerable improvements in terms of quality of service (Samdanis et al., 2016; Sciancalepore et al., 2017, 2018), which is encouraging further research in this area. Moreover, prototype implementations of radio access network (RAN) slicing such as Orion (Foukas et al., 2017a) allow for the dynamic allocation of radio resources, providing flexibility for resource allocation mechanisms to fulfil diverse types of service level agreements (SLAs). This enables deploying different flow rate allocation schemes in real 5G networks and optimising network utility (Li et al., 2018), as I explain in Chapter 4. The utility framework proposed in this thesis and the deep learning approach to solve the utility optimisation make my contribution fundamentally different from these prior research efforts.

2.1.3 Emergency Networking

Intra- and inter-agency co-ordination is crucial in disaster response scenarios (Smith and Dowell, 2000), and multi-media data communications plays a vital role in helping first responders (Wang et al., 2016), e.g. police officers, medical teams, and firefighters. Legacy private mobile radio (PMR) network services, e.g. push-to-talk and off-network device-to-device communications are mainly voice-centric and allow only for limited-speed (hundreds of kilobits per second) data communications (ETSI, 2011). Satellite Internet, on the other hand, is subject to excessive latency – the average latency in satellite data services offered by U.S. providers is around 600ms (Hanson, 2016). Therefore, this connectivity solution can hardly meet the real-time constraints of mission critical applications. Previous work discusses the feasibility of adopting LTE for high-bandwidth low-latency communication in emergency and disaster scenarios (Doumi et al., 2013). More recently, ongoing standardisation specifies several technical aspects of isolated LTE E-UTRAN that incorporate local LTE evolved

packet core (LTE-EPC) for public safety operations (3GPP, 2015). These efforts are further reviewed and discussed in (Oueis et al., 2017). Oueis *et al.* identify a number of research perspectives including local EPC placement, radio resource control, and network dynamics. Among these, this thesis tackles the network dynamics problem, towards which I propose a solution for controlling the mobility of airborne base stations. Although current cell-on-wheels solutions are widely used to expand LTE service coverage at specific events, such as the Super Bowl (Cellsite Solutions LLC, 2017), and off-the-shelf wearable base stations in backpacks are available on the market (Air Lynx, 2018), there exist certain landscapes that are hard-to-reach and scenarios that are extremely challenging for terrain vehicles, e.g. following floods, earthquakes, or nuclear plant disasters.

Loon LLC initiated the idea of flying balloons carrying base stations into stratosphere and delivering LTE connectivity to unserved and underserved communities (Loon LLC, 2018). Despite good intentions, the experiment has caused several incidences (Wikipedia, 2018), including a balloon hitting power lines in a Washington town and disrupting the power supply (Welch, Chris, 2018), due to the difficulty of manoeuvring the balloons at the stratosphere level, especially when the airborne platforms are out of battery. On the other hand, UAVs are much more robust, flying at lower altitudes, and hence have received increasing interests from the telecom industry as potential bearers of aerial base stations in temporary cellular deployment (FAA, 2017). In this context, the research question of mobility management of a fleet of cellular base stations on UAVs is tackled in Chapter 6.

2.2 Brief Survey of Related Work

1) Mm-wave Characterisation: Recent efforts to characterise mm-wave spectral resources suggest that the 30–300GHz band enables the deployment of multi-Gbps link rates (Rappaport et al., 2013b) with small form factor equipment. Hence, mm-wave has been considered as a viable solution for wireless backhauling (MacCartney and Rappaport, 2014; Dehos et al., 2014). Channel measurement efforts also confirm that beamforming necessary to mitigate attenuation in mm-wave bands drastically reduces interference, and links can often be regarded as pseudo-wired (Singh et al., 2011). Mm-wave codebook based beamforming has so far been well studied (Wang et al., 2009; Lee and Ko, 2011; Li et al., 2013; Zhou and Ohashi, 2012). Specifically, Wang *et al.* propose a codebook based beamforming protocol to setup multi-Gbps mm-wave

communication links (Wang et al., 2009). Li *et al.* formulate a global optimisation problem to select the optimal beam pattern index at the transmitter and the receiver that provides the best SINR, following which the authors then devise a gradient-based approach to solve the optimisation problem (Li et al., 2013). Lee and Ko proposed a multi-level codebook-based beamforming scheme with the goal of reducing the beamforming setup time (Lee and Ko, 2011). At each beamforming level, the algorithm selects transmit and receive antennas by sending a number of training sequences with different weight vectors from a pre-defined codebook and taking the feedback of the best weight vectors options that optimise an effective signal-to-noise ratio (SNR) from the receiver, and then passing on these vectors to the next following level until the optimal selection of transmit and receive antennas are found.

Qiao *et al.* believe concurrent beamforming algorithm is desired for multi-hop indoor mm-wave networks to reduce the setup time and increase the system throughput (Qiao et al., 2015). These studies focus on beamforming techniques and target wireless personal area networks, which differ from what I try to solve in this thesis, i.e. scheduling for end-to-end data flows in small cell backhauling scenarios.

Further, Hur *et al.* design a beam alignment mechanism for mm-wave backhauling scenarios, tackling the effects of wind-induced beam misalignment (Hur et al., 2013). With mandatory use of beamforming, however, terminal deafness becomes a key challenge when scheduling transmissions/receptions (Nitsche et al., 2014).

Moreover, the throughput and energy consumption characteristics of different mm-wave bands are studied in (Mesodiakaki et al., 2016). While I do not explicitly address energy efficiency aspects in my work, I recognise that a certain degree of energy efficiency can be inherently achieved through optimal airtime allocation and scheduling, which are at the core of my contribution.

2) Medium Access Control in Ad-Hoc Networks with Directional Antennas: Previous MAC protocol designs for networks operating on legacy frequencies and employing directional transmission patterns are not feasible or insufficient to be employed in mm-wave networks. Choudhury *et al.* present one of the few MAC schemes that concern directional antenna patterns (Choudhury et al., 2006), namely MMAC. MMAC employs multi-hop RTS to build up single hop directional links, then uses the directional link for transmissions. The focus of MMAC is hence to establish links between two nodes, while multi-hop airtime scheduling and inter-flow fairness are not consid-

ered in this work. Korakis *et al.* exploit the increased coverage of directional antennas and propose a MAC protocol to tackle the deafness issue (Korakis et al., 2003). However, multi-hop scheduling and flow level fairness are not tackled by this protocol.

Bao and Garcia-Luna-Aceves propose a distributed receiver-oriented multiple access (ROMA) protocol to achieve scheduled access in ad-hoc networks with directional antennas (Bao and Garcia-Luna-Aceves, 2002). ROMA employs a type of multi-beam adaptive array (MBAA), which allows a node to activate multiple transmission beams at the same time. This MBAA scheme, however, focuses on link level scheduling, whereas in backhauling scenarios, which are at the core of Chapter 3.1 of this thesis, efficient resource allocation requires flow level scheduling. Zhang and Datta propose a directional antenna based MAC protocol for scheduling at each node in sensor networks, employing carrier sensing (Zhang and Datta, 2005). The incentive of employing directional antenna patterns there is mainly for energy conservation, and the directional beamwidth considered is as large as 60°. Beamforming in mm-wave makes carrier sensing difficult and hence this MAC approach is infeasible for mm-wave. Moreover, flow level scheduling and fairness are not considered in the design of this protocol.

3) Medium Access & Scheduling in Mm-Wave Networks: Medium Access Control protocols for mm-wave communications can be grouped into two main classes – contention-based and (pseudo-)scheduled. The IEEE 802.11ad standard (IEEE 802.11ad Std., 2014) specifies both contention-based and Service Period (SP) driven (scheduled) channel access mechanisms for communications in the unlicensed 60GHz band. Building upon 802.11ad, the 802.11ay draft aims to achieve link rates of up to 100Gbps, by employing a number of enhancements, including 4-stream MIMO (Cerwall (ed), 2017). On the other hand the 3GPP New Radio (NR) specification extends the LTE numerology by allowing different types of sub-carrier spacing and slot lengths (3GPP, 2018). The 10ms frame structure of LTE with 1ms subframes is preserved. It is worth noting that both IEEE and 3GPP standards leave open the airtime allocation and multi-hop transmission coordination tasks.

Based on the scheduled access specified in the IEEE 802.11ad standard (IEEE 802.11ad Std., 2014), Hemanth and Venkatesh analyse the performance of the SP mechanism in terms of frame delay (Hemanth and Venkatesh, 2016). A line of work builds upon the standard, specifying MAC protocol improvements for single-hop wireless local area networks (WLANs) (Chandra et al., 2014; Sim et al., 2016b; Chen et al., 2013). Chandra *et al.* employ adaptive beamwidth to achieve improved channel util-

isation (Chandra et al., 2014). Sim *et al.* exploit dual-band channel access to address terminal deafness and improve throughput (Sim et al., 2016b). Optimal client association and airtime allocation is pursued in (Facchi et al., 2017) to maximise the utility of enterprise mm-wave network deployments. However, all of the above mentioned works consider only single hop scenarios, whereas in this thesis, I tackle issues specific to mm-wave multi-hop scheduling and optimal resource management.

Coordinating TX-RX chains in dense small cell deployments, where data flows typically traverse multiple hops, is of great significance, since mis-aligned beams at a single hop will result in packet loss of an end-to-end data flow and hence wastage of mm-wave resources. Therefore, rethinking MAC protocols for multi-hop backhauling scenarios is required, taking into consideration mm-wave specifics. Towards multi-hop mm-wave scenarios, a directional cooperative MAC protocol is introduced in (Chen et al., 2013), where user devices select intermediate nodes to relay packets to access points, in order to establish multi-hop paths that exhibits higher signal-to-noise-ratio (SNR) than direct links. Further, Mandke and Nettles propose a dual-band architecture for multi-hop 60GHz networks where scheduling and routing decisions are communicated at 5.2GHz (Mandke and Nettles, 2010). These designs, however, do not solve the airtime allocation and transmission scheduling problem, as investigated in this thesis. Based on their feasibility study of in-band wireless backhauling, Taori *et al.* present a qualitative scheduling framework for inter-base station communications (Taori and Sridharan, 2015). This resembles closely the Type 2 TDD scheme of LTE, with the difference that the authors apply it to in-band backhauling scenarios, whereas in the LTE standard this is specified for cellular access only. Despite considering the implications of terminal deafness, these designs do not tackle the airtime allocation problem. Relay selection so as to overcome blockage and scheduling in mm-wave backhauls is tackled in (Hu and Blough, 2017), with the aim of maximising throughput. However, neither airtime allocation nor fairness are taken into account.

Several distributed opportunistic medium access schemes for mm-wave multi-hop scenarios have been proposed to date (Sim et al., 2016a; Singh et al., 2010). In these works, each base station chooses randomly a time to access an intended receiver, and marks the results of success or failure. Then based on this experience the base stations decide individually when to attempt to transmit next. Once a station finds a block of time suitable for transmission, it will keep using the same period of time in the following schedules. MDMAC as proposed in (Singh et al., 2010) operates with a slotted channel, whereby a station's transmission can occupy one or multiple slots, but the

slot duration remains fixed for all participants (20 μs by default), which may harm efficiency (Singh et al., 2010). Unslotted approaches (Bin)DLMAC are introduced in (Sim et al., 2016a) to improve protocol efficiency and 'learn' when to transmit in the presence of terminal deafness. Both schemes do not explicitly consider inter-flow fairness, as each node seeks to transmit as much as possible. Our results confirm this leads to poor performance for flows encountering lower capacity links. In contrast, the mechanism proposed in Chapter 3 not only improves throughput performance, but is also significantly more fair, as the proposed method takes into account all flow demands, link rates, and level of competition.

Su and Zhang solve optimal network throughput allocation heuristically in multi-channel settings, without fairness guarantees (Su and Zhang, 2009). Ford *et al.* target sum utility maximisation in self-backhauled mm-wave setting (Ford et al., 2015). Seminari *et al.* formulate the sharing of mm-wave backhauls as a one-to-many matching game, seeking to maximise the average sum rate (Semiari et al., 2017). Zhu *et al.* propose a maximum independent set (MIS) based scheduling algorithm to maximise QoS in mm-wave backhauls (Zhu et al., 2016). Similarly, Niu *et al.* propose MIS based scheduling that aims to minimise the energy consumption (Niu et al., 2017). A joint scheduling and power allocation problem is also solved with MIS in (Li et al., 2017a). In this body of work scheduling is performed with the explicit goal of achieving concurrent transmissions among non-interfering links. The WIHAUL mechanism I propose allows for concurrent transmissions by default. Moreover, my solution not only improves throughput performance, but also explicitly addresses fairness, while I take into account all flow demands, link rates, and the level of competition among them. In particular, I address airtime allocation and scheduling in multi-hop mm-wave networks using the max-min fairness criterion.

4) Fairness in Multi-hop Wireless Networks: Bertsekas and Gallager consider max-min fairness for flow control in wired networks (Bertsekas and Gallager, 1992) and subsequently Le Boudec and Radunovic demonstrate that max-min fairness is a geometric property of the set of feasible allocations (Radunović and Le Boudec, 2007). The 802.11 rate region is proven log-convex, and station attempt probabilities and burst sizes in 802.11 mesh networks are derived for max-min fair regimes in (Leith et al., 2012). This, however, only holds in multi-channel mesh topologies where stations employ multiple interfaces, which is impractical with small form factor mm-wave devices equipped with a single interface. Wang *et al.* argue that channel time rather

than flow rate should be used with the max-min allocation criterion in wireless multi-hop networks and accordingly propose a new definition of max-min fairness (Wang et al., 2008). Unfortunately, under this definition, flows traversing larger number of hops will, by design, obtain considerably smaller throughput than those close to gateways. This implies inferior service performance for distant users, hence the approach is ill-suited to carrier-grade backhauls.

Lan *et al.* propose a unified fairness measure that enables to explicitly quantify max-min fairness, which is largely perceived as qualitative (Lan et al., 2010). This general measure of fairness is employed in this thesis to derive a max-min fair metric and evaluate the gains achieved by our proposal. To add further perspective I also resort to economic notions of inequality, i.e. the Gini coefficient (Gini, 1921).

5) Network Slicing. Network slicing allows logical partitioning of shared physical network infrastructure and is envisioned as a key technology for 5G to meet the diverse service requirements of numerous emerging mobile applications (Alliance, 2016; Zhang et al., 2017b; Foukas et al., 2017b; Ordonez-Lucena et al., 2017).

Samdanis *et al.* and Sciancalepore *et al.* showed in their recent studies that significant quality of service improvements can be achieved by sharing mobile infrastructure among virtual operators (Samdanis et al., 2016; Sciancalepore et al., 2017, 2018). In (Samdanis et al., 2016), a slice broker concept which enables mobile infrastructure providers to dynamically manage the shared network resources is proposed. Based on this concept, a machine learning approach that addresses admission control in sliced networks is given in (Sciancalepore et al., 2017). An online slice brokering solution is studied in (Sciancalepore et al., 2018) with the goal of maximising the multiplexing gain in shared infrastructure. These efforts, however, do not tackle different service requirements of diverse application scenarios, which instead I address in Chapter 4.

3GPP standards has specified a simple form of slicing called dedicated core networks (DECOR) (3GPP, 2017a). There has been significant research effort implementing slicing functionalities in both core and radio access parts of the cellular networks (Kim et al., 2014; Nikaein et al., 2015; Nakao et al., 2017; Foukas et al., 2017a). Foukas *et al.* present a prototype implementation of radio access network (RAN) slicing system namely Orion (Foukas et al., 2017a). Building upon open-source LTE platform, Orion is the first RAN slicing system that enables functional and performance isolation of slices, while enabling dynamic allocation of radio resources, providing flexibility for resource allocation mechanisms to fulfil diverse types of service level

agreements (SLAs). These efforts are important steps towards the deployment of 5G network slicing, while they also open up new challenges of resource allocation for diverse application use cases, which will be addressed in Chapter 4.

6) Network Utility Maximisation (NUM). Optimising a mixture of concave and non-concave utilities for inelastic traffic has been studied in (Fazel and Chiang, 2005; Hande et al., 2007; Chen et al., 2011). Fazel *et al.* propose a sum-of-squares method to solve non-concave NUM problems that tackle primarily polynomial utility (Fazel and Chiang, 2005). Hande *et al.* study the sufficient conditions for the standard price-based (sub-gradient based dual) approach to converge to global optima with zero duality gap, which relies on capacity provisioning (Hande et al., 2007). Chen *et al.* consider NUM with mixed elastic and inelastic traffic, and develop a heuristic method to approximate the optimal solutions (Chen et al., 2011). Recent work investigates a convex relaxation of polynomial NUM and employs distributed heuristics to approximate the global optimal allocation (Wang et al., 2017). Udell and Boyd define a general class of non-convex problems as sigmoidal programming and propose an approximation algorithm to solve these (Udell and Boyd, 2013).

The heuristics methods mentioned above share the limitation of convergence times that are in the order of seconds, which can hardly meet the latency requirements of 5G networks. In contrast, the deep learning approach presented in Chapter 4 infers close to optimal rate allocations within milliseconds.

7) Transport Protocols in Wireless Networks. Originally designed for wired networks with low bit error rate, TCP struggles to adapt to wireless environment and often suffer from low throughput and frequent disconnection. Existing studies on TCP behaviour in LTE-EPC networks suggest that a sudden load increase in a cellular network will lead to significant bandwidth reduction and delay increase (Nguyen et al., 2014). Due to the uncertainty specific to wireless networks, especially varying link qualities, TCP retransmission leads to larger overhead and network inefficiency (Huang et al., 2013). Zhang *et al.* argue that a small handover offset leads to better throughput performance in spite of the increasing probability of ping-pong handover (Zhang et al., 2012). Challenges of optimising cell-edge SINR are presented in (Fujitsu, 2011), where it is suggested that inter-cell interference coordination schemes should be employed in PDCCH.

Over the years, there have been extensive research efforts on initiating new TCP

congestion mechanisms to mitigate the performance degradation of TCP in wireless networks (Balakrishnan et al., 1996; Brown and Singh, 1997).

A number of work take split-connection approach (Bakre and Badrinath, 1988; Brown and Singh, 1997). Indirect-TCP (I-TCP) (Bakre and Badrinath, 1988) divides each end-to-end TCP connection between a sender and receiver into two separate connections at the base station, i.e. one TCP connection between the sender and the base station, and the other between the base station and the receiver, each having separate acknowledgements. The authors show that I-TCP achieves improved throughput when compared with legacy TCP over wireless links. However, such separation does not address the fundamental issue of TCP performance degradation on wireless links and I-TCP in particular has been criticised for producing time-outs causing stalls of the sender (Balakrishnan et al., 1996). In addition, under the split-connection approach, every packet incurs the overhead of going through TCP protocol processing twice at the base station, which is not resource efficient.

The seminal work by Floyd *et al.* initiates the idea of explicit congestion notification (ECN) (Floyd, 1994), which, as the name suggests, involves taking congestion indication feedback from the network to decide whether to react to packet loss. Since then, several works have been build upon the ECN semantics. Katabi *et al.* propose a new congestion control algorithm namely XCN, by which the sender is informed about the degree of congestion at the bottleneck instead of receiving binary feedback about the congestion state, as in ECN (Katabi et al., 2002; Ramani and Karandikar, 2000; Athuraliya et al., 2001). XCN demonstrates fair bandwidth allocation, high utilisation, small standing queue size, and near-zero packet drop, with both steady and highly varying traffic. Moreover, Ramani and Karandikar propose a TCP protocol that tries to determine the cause of packet drop in wireless networks, based on ECN. This approach incorporates a number of modifications, mainly at the transmitting side, including queue monitoring. The proposed mechanism achieves throughput improvement over unmodified TCP (Ramani and Karandikar, 2000). Athuraliya *et al.* suggest a new active queue management scheme, namely random exponential marking (REM), to decouple the congestion measure based on ECN from performance metrics such as loss, queue length, or delay (Athuraliya et al., 2001). The proposed method achieves good utilisation of resources with slightly compromised loss and delay.

To the best of my knowledge, the work presented in Chapter 5 is the first to study transport protocols' performance in LTE networks with a specific focus on cell edge users and the issue introduced by control plane errors, and uncover potential through-

put gain by enabling the RLC AM mode in the LTE protocol stack .

In a recent study, Polese *et al.* look into the cell-edge performance in mm-wave frequency bands and observe similar TCP throughput degradation as I present in Chapter 5, revealing a reduction when MAC HARQ is disabled (Polese et al., 2017). This study further investigates the performance of multi-path TCP connections leveraging different radio access interfaces on mm-wave and legacy frequencies, and shows that multi-path TCP flows traversing both legacy frequency and mm-wave links achieve higher throughput and lower latency at mid-cell (100m) and cell-edge (150m) distances, when compared to single-path or multi-path TCP flows traversing mm-wave only links.

Furthermore, the IETF has put standardisation effort into Google's QUIC transport paradigm(Hamilton et al., 2016), which is essentially an UDP based protocol that exploits multiplexed connections between two UDP endpoints. QUIC demonstrates reduced latency and additional data loss avoidance, as compared to TCP in lossy channel conditions. However, QUIC has also been criticised for being less competitive when compared with TCP in small loss rate, large buffer, or large propagation delay scenarios (Yu et al., 2017). In 2016, Google developed a new congestion control algorithm named bottleneck bandwidth and round-trip propagation time (BBR) (Cardwell et al., 2016). BBR is a model-based congestion control mechanism that maintains an explicit model of the network using the recent maximum bandwidth and round-trip time experienced by the outbound packets. Despite achieving higher throughput and lower latency, BBR is nonetheless criticised for being unfair to other flows and causing considerate packet loss (Hock et al., 2017). The question of what the best transport protocol for the ever changing Internet ecosystem remains open (Huston, 2018).

8) Airborne and Emergency Networking. Oueis *et al.* give a technical overview of LTE operation for public safety (Oueis et al., 2017), which provides useful insights for the work on mobile base stations control, presented in Chapter 6. Fotouhi *et al.* propose a heuristic based mobility model to improve spectral efficiency in drone-based base station deployments. However, the authors consider only fixed user group coverage without handovers (Fotouhi et al., 2017). Orsino *et al.* highlight that drones carrying radio transceivers improve network coverage and bring higher data rates to challenging locations (Orsino et al., 2017). Grossglauser and Tse study the per-user throughput in mobile ad-hoc networks and conclude that performance can be improved dramatically when base stations are mobile (Grossglauser and Tse, 2002). Unlike the

findings presented in Chapter 6 of this thesis, these works fail to consider realistic deployment aspects of airborne base stations and the challenges associated with the mobility management task, including practical group mobility behaviour and potential handover effects.

The SWARMIX Project (SWARMIX.org, 2014) delivered a number of studies concerning synergistic interactions between heterogeneous agents including robots, humans, and animals. Part of this project investigates ad-hoc wireless networking aspects. Specifically, based on empirical studies of micro aerial vehicle networks, Asadpour *et al.* identified potential research directions including how controlled mobility could bring benefits at each network layer (Asadpour et al., 2014). This links to my study of mobility control to achieve connectivity performance gains for users. In (Asadpour et al., 2013), the authors present an implementation of a 2-hop UAV network that is capable of searching for a missing person via smartphone emitted beacon messages, and streaming video of the area of interest via the 2-hop network. This study demonstrates the feasibility of employing UAV mounted Wi-Fi access points in emergency settings, but neither addresses signal coverage for a handful of ground users with complex movement, nor the mobility management of UAVs, which are the main concerns of my study. Path planing of UAVs, as tackled in (Flushing et al., 2014; Di Caro et al., 2014) and other research papers in the field of autonomous robotics, are a school of study that relates to finding the optimal path to a destination subject to certain constrains, such as maximum travel distance. Although these works cover movement control of UAVs, they consider a different setting than emergency wireless networking, as this thesis does, where no prior knowledge of a target destination can be obtained.

9) Deep Learning in Networking and Network Resource Management. Recent improvements in computational power and the growing numbers of public data sets have led to a range of deep learning applications in the computer and communications networking domain, intelligent mobile networking is becoming an important research direction to enable service requirements promised by 5G (Kibria et al., 2017; Zhang et al., 2018; Agiwal et al., 2016; Gupta and Jha, 2015). For instance, Kato *et al.* devise a fully-connected neural network to find optimal routes in wired/wireless heterogeneous networks (Kato et al., 2017). Zhang *et al.* employ dedicated convolutional neural networks (CNNs) to infer fine-grained mobile traffic consumption from coarse traffic aggregates (Zhang et al., 2017a), improving measurement resolution by up to 100$\times$ while maintaining high accuracy. CNNs have also been employed in (Zhang and Pa-

tras, 2018), where the authors incorporate a 3D-CNN structure into a spatio-temporal neural network, to perform long-term mobile traffic forecasting. Moreover, following advances in deep reinforcement learning (DRL) such as the asynchronous advantage actor-critic method (A3C), which achieved remarkable performance in game-play applications (Mnih et al., 2016), Mao *et al.* employed the A3C technique for adaptive video streaming (Mao et al., 2017).

Several studies employ deep learning for network resource management in various contexts. Ying *et al.* formulate the wireless channel as a finite-state Markov channel and employ Q-leaning to obtain best user selection policy in interference alignment wireless networks (Ying et al., 2018). The proposed approach therein achieves higher sum rate and energy efficiency over benchmark solutions. Xu *et al.* present a deep Q-learning based DRL framework for power allocation in cloud RANs. Specifically, a DRL agent determines the ON/OFF mode of the remote radio heads based on the current mode and user demand. A simulation campaign demonstrates improved power efficiency as compared to benchmark solutions. Ferreira *et al.* propose a hybrid radio resource allocation management control algorithm based on a multi-objective reinforcement learning frame work, i.e. Action-Reward-State-Action (SARSA) in cognitive communications, which requires reduced computational resources (Ferreira et al., 2017). Sun *et al.* propose to utilise the input and output of a resource allocation algorithm as an unknown non-linear mapping and employ a deep neural network to approximate such mappings under interference-limited wireless network environments (Sun et al., 2017). The authors demonstrate that via imitation learning, the inferences made by a neural network resemble closely the teacher with significantly reduced runtime.

To the best of my knowledge, the work presented in Chapter 4 is the first that employs deep learning to solve utility optimisation problems in sliced backhauls. Further, the work presented in Chapter 6 is the first to employ DRL for airborne base station mobility control.

Chapter 3

Resource Allocation in Mm-wave Backhaul Networks

The mobile networking community is pursuing densification of small cell deployments to address the capacity crisis inherent to the projected exponential increase in mobile data traffic. Connecting to the Internet tens of thousands of base stations is non-trivial, especially in urban scenarios where optical fibre is difficult and costly to deploy.

Meanwhile, mm-wave communication technology has been advancing rapidly in recent years, demonstrating up-to multi-Gbps link rates and is promoted by both regularisation bodies (Ofcomm, 2018), and standardisation efforts such as IEEE 802.11ad (IEEE 802.11ad Std., 2014), 802.11ay (Cerwall (ed), 2017), and 3GPP 5G NR (3GPP, 2018). The mm-wave spectrum is a promising candidate for inexpensive multi-Gbps wireless backhauling, i.e. a solution favoured by industry stakeholders (Facebook Inc., 2018; Qualcomm Technologies Inc., 2018), but exploiting this band for effective multi-hop data communications is challenging. In particular, resource allocation and scheduling of very narrow transmission/reception beams requires to overcome terminal deafness and link blockage problems, while managing fairness issues that arise when flows encounter dissimilar competition and traverse different numbers of links with heterogeneous quality.

In this chapter, I propose WIHAUL, an airtime allocation and scheduling mechanism that overcomes these challenges specific to multi-hop mm-wave networks, guarantees max-min fairness among traffic flows, and ensures the overall available backhaul resources are fully utilised. An evaluation of the proposed WIHAUL scheme over a broad range of practical network conditions is presented, demonstrating up to $5\times$ individual throughput gains and a fivefold improvement in terms of measurable fairness,

over recent mm-wave scheduling solutions.

3.1 System Model

I focus on dense mobile broadband deployments where B fixed base stations provide wireless access to mobile users with different traffic demands. While serving a number of smart devices that consumes or generate data flows, base stations are connected via mm-wave links to wired Internet gateways, possibly over multiple hops.

PHY Layer Considerations: Although PHY layer optimisation is outside the scope of this work, this section briefly summarises the PHY aspects that are relevant to the design of WIHAUL. Theoretically, the proposal is compatible with either MIMO antenna arrays via beamforming, as well as fixed horn antennas (fixed or mechanically steered). I nevertheless envision beamforming to be a more practical solution for backhauling of densely deployed cellular base stations, especially in urban scenarios, for a number of reasons. First of all, beamforming allows for rapid beamswitching. As one may argue that a base station with multiple fixed horn antennas can achieve transmission towards multiple directions as well, this notwithstanding incurs larger form factors when compared with beamforming infrastructures such as steerable arrays (Singh et al., 2011). More importantly, beamforming copes better with dynamic routing by forming new links to tackle blockage at established links or sudden changes in link conditions, which are almost inevitable in mm-wave. Furthermore, horn antennas are usually subject to large beamwidth, whereas beamforming allows for narrow beamwidths that are less susceptible to inter-link interference and therefore offer better spatial reuse. In fact, a number of affordable off-the-shelf commercial devices supporting mm-wave, e.g. TP-Link Talon AD7200 Router (TP-LINK, 2017) and Mikrotik wAP60 (Mikrotik, 2017), employ phased array to perform beamforming.

Assume each backhaul node employs N TX/RX antennas and adopt the mm-wave MIMO channel model proposed in (Alkhateeb et al., 2014a), where hybrid analog/digital precoding is employed. By (Alkhateeb et al., 2014a), the channel is subject to limited scattering and geometric models are generally applicable (Rappaport et al., 2013a, 2012). The channel matrix can be expressed as

$$\mathbf{H} = \sqrt{\frac{N_{tx}N_{rx}}{\bar{P}_L}} \sum_{l=1}^{L} \alpha_l \mathbf{a}_{rx}(\theta_l^{AOA}) \mathbf{a}_{tx}^H(\theta_l^{AOD}) \tag{3.1}$$

where $\bar{P_L}$ denotes the average pathloss between the transmitter and receiver, L is the number of scatters, and α_l is the complex gain of the l-th channel, following Rayleigh distribution $\alpha_l \sim N(0, \bar{P_R}), \forall l \in \{1, 2, ..., L\}$, and $\bar{P_R}$ is the average power gain. Moreover, $\theta_l^{AOD} \in [0, 2\pi]$ and $\theta_l^{AOA} \in [0, 2\pi]$ denote the azimuth angles of the departure and arrival (AOD and AOA) respectively, and $\mathbf{a}_{tx}(\theta_l^{AOD})$ and $\mathbf{a}_{rx}(\theta_l^{AOA})$ are the antenna array response vectors at the transmitter and the receiver. While extensions to 3D beamforming is possible (Ayach et al., 2014), I focus here on horizontal 2D beamforming by neglecting the elevation angle. Assuming uniform linear arrays, the antenna response vector can be written as:

$$\mathbf{a}_{tx}(\theta_l^{AOD}) = \frac{1}{\sqrt{N_{tx}}}[1, e^{j(2\pi/\lambda)d\sin(\theta_l^{AOD})}, ..., e^{j(N_{tx}-1)(2\pi/\lambda)d\sin(\theta_l^{AOD})}]^T \tag{3.2}$$

where λ is the wavelength, and d is the distance between antenna elements. The response vector of the receiver antenna array, i.e. $\mathbf{a}_{rx}(\theta_l^{AOA})$ has a similar form.

According to (Alkhateeb et al., 2014a), with efficient design of the precoders ($\mathbf{F}_{BB}$ for baseband and $\mathbf{F}_{RF}$ for radio frequency (RF)) and combiners ($\mathbf{W}_{BB}$ for baseband and $\mathbf{W}_{RF}$ for RF), the achievable rate of the MIMO system is formulated as:

$$R = \log_2 |\mathbf{I}_{N_s} + \frac{P}{N_S}\mathbf{R}_n^{-1}\mathbf{W}_{BB}^H\mathbf{W}_{RF}^H\mathbf{H}\mathbf{F}_{RF}\mathbf{F}_{BB}\mathbf{F}_{BB}^H\mathbf{F}_{RF}^H\mathbf{H}^H\mathbf{W}_{RF}\mathbf{W}_{BB}|, \tag{3.3}$$

where the post-processing noise covariance matrix $\mathbf{R}_n$ is given by $\mathbf{W}_{BB}^H\mathbf{W}_{RF}^H\mathbf{W}_{RF}\mathbf{W}_{BB}$.

MAC Paradigm: My work targets mm-wave systems where channel multiplexing is performed following time division principles. As such, the proposed solution is applicable to both TDMA-based cellular backhaul scenarios and single-/multi-hop deployments based on the IEEE 802.11ad (IEEE 802.11ad Std., 2014) or 802.11ay (Cerwall (ed), 2017) working with SPs, e.g. in rural and community networks. With these in mind, this chapter addresses rigorously the airtime allocation and TX/RX beam scheduling in multi-hop backhaul networks. The proposed WIHAUL solution observes a periodic superframe/beacon interval structure where beamform training information is exchanged and TX/RX scheduling is performed at the start of a superframe, following which link transmissions take place as per computed schedules, as depicted in Fig. 3.1.

Centralised Control: I envision a centralised architecture, whereby a controller has full knowledge of the network topology, periodically collects link rate and flow demand information, and subsequently performs airtime allocation and beam scheduling. In practice, centralised control is achievable through software-defined network (SDN)

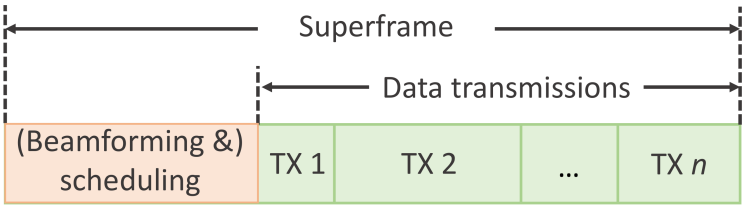


Figure 3.1: TDM superframe structure observed bv WIHAUL. Beamform training, scheduling and control message exchange take place periodically at the start. Data transmissions (possibly of different durations) follow. The structure can equally apply to 3GPP NR and IEEE 802.11ad with SPs.

primitives (Haleplidis et al., 2015), for instance running OpenFlow (Haque and Abu-Ghazaleh, 2016) over a dedicated narrow-band low frequency channel. Similar out-of-band control schemes have been previously used in wide spectrum networks (Giannoulis et al., 2013). Assume the controller is also responsible for computing paths p_k for all flows k traversing the backhaul, which is orthogonal to the problem this chapter attacks and thus not explicitly considered herein. This is aligned with previous work on mm-wave backhauling where path computation and link scheduling are dealt with separately (Seppänen et al., 2016). An alternative approach is to tackle the problem using decentralised approach to achieve scheduled access. This however, as will be discussed further in this Chapter, is subject to a convergence time which introduces additional overhead every time the system need to recompute, and as we will show in the simulation campaign, the state-of-the art solutions are outperformed by the proposed approach of this chapter. I give an overview of the overall envisioned system in Fig. 3.2.

The objective here is to allocate the airtime resources available on the mm-wave backhaul links to aggregate traffic flows and co-ordinate transmissions among base stations. Flows either enter the network via gateways, are relayed by intermediary hops, before reaching the end users (downlink), or originate at different base stations and are forwarded externally by the gateways (uplink). The problem pursued in this chapter is challenging and fundamentally different to previous efforts in multi-hop wireless networks (e.g. (Wang et al., 2008)), since the backhaul system is prone to terminal deafness and a receiver may experience secondary interference only when situated in the range and on the direction of another active beam. Since deployments with small form factor base stations equipped with a single mm-wave interface are considered, intra-flow competition occurs and fairness issues arise as flows are relayed by base stations, unlike in multi-radio mesh networks (Leith et al., 2012). Meanwhile,

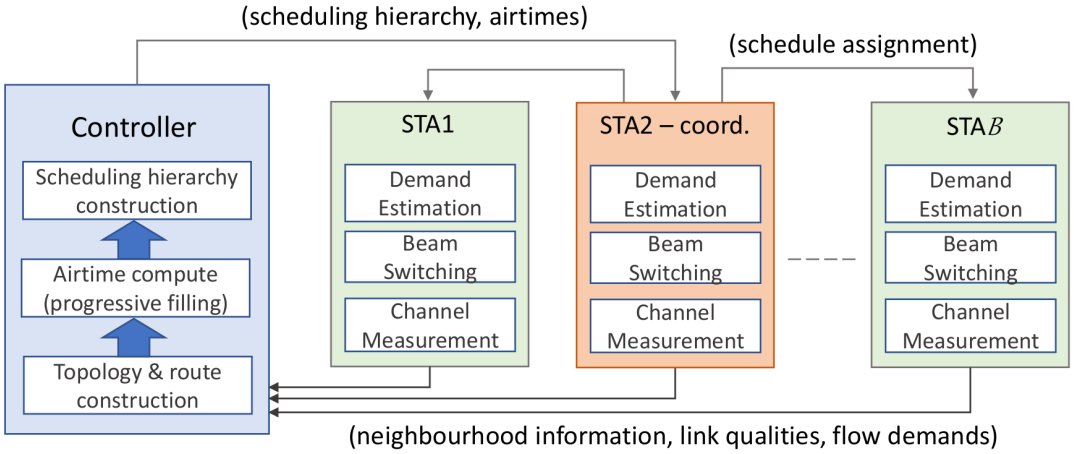


Figure 3.2: High-level overview of the envisioned system. WIHAUL **runs on the controller and computes flow airtime allocations and schedules, based on topology information and paths computed by routing logic. Scheduling hierarchy and airtimes sent to a scheduling coordinator, which dictates the TX/RX timing to backhaul nodes.**

concurrent transmissions on non-interfering links is feasible, which allows for spatial reuse and appropriate network utilisation at a lower cost.

3.2 Problem Formulation

The goal of this chapter is to devise a way of distributing the mm-wave backhaul resources among flows, such that network utilisation is maximised while flows with lower demand or originating/terminating further away from the gateway are not throttled. My focus is on the MAC layer and I assume PHY layer aspects such as power allocation, codebook design, or beamform training can be dealt with separately. This assumption is reasonable, because PHY optimisation will ultimately result in different capacity constraints imposed on the MAC layer. Mm-wave specific aspects such as link blockage are inherently captured in my formulation through constraints. I further take into account potential link-blockage conditions when allocating resources, circumventing these as we explain below. As such, I work with the *max-min fair* criterion (Bertsekas and Gallager, 1992), seeking to ensure flow demands are fulfilled in increasing order where possible, whilst any remaining network capacity is shared among flows with higher demands. That is, I aim to maximise the end-to-end throughput r_k of each aggregate flow k, subject to individual demands d_k, whilst any flow allocation increase would not harm others with already smaller or equal throughputs. To the best of my

knowledge, max-min fair resource allocation in mm-wave backhauls, which present unique terminal deafness challenges, has not been considered previously.

Notation	Description
$\mathcal{F}$	The set of flows traversing the backhaul
$\mathcal{F}_{\mathcal{A}}$	Active flows, i.e. the set of flows that have not been allocated resources
k	Index of an aggregate flow
r_k	End-to-end throughput of flow k
d_k	Demand of flow k
p_k	The path traversed by flow k
$l_{i,j}$	Link between stations i and j
$s_{k,i,j}$	The segment of flow k traversing link $l_{i,j}$
$t_{k,i,j}$	Airtime assigned to flow segment $s_{k,i,j}$
C_q	A clique with index q

Table 3.1: Notations and descriptions.

The mathematical notations used in this section is listed in Table 3.1. Formally, by denoting $\mathcal{F}$ the set of flows traversing the backhaul, p_k the path of flow k, i.e. the sequence of links this follows from source to destination (within the backhaul), and considering flow k is assigned airtime $t_{k,i,j}$ on link $l_{i,j}$, I want to find the vector

$$\mathbf{t} := \{t_{k,i,j} | k \in \mathcal{F}, l_{i,j} \subset p_k\}$$

that achieves max-min fair allocation of flow throughputs. This requires to iteratively solve the following optimisation problem:

$$\mathbf{t}^* = \bigcup_{k \in \mathcal{F}_{\mathcal{A}},\ \mathcal{F}_{\mathcal{A}} \subset \mathcal{F}} \arg\max_{\mathbf{t}} \min_{k} r_k \tag{3.4}$$

$$\text{s.t. } r_k \leq d_k, \forall k \in \mathcal{F}_{\mathcal{A}}, \tag{3.5}$$

$$\sum_{s_{k,i,j} \in C_q} \frac{r_k}{c_{i,j}} \leq 1, \forall k \in \mathcal{F}_{\mathcal{A}}, \forall C_q \in \mathcal{C}. \tag{3.6}$$

In the above, $\mathcal{F}_{\mathcal{A}} \subset \mathcal{F}$ denotes the set of flows that have not yet been allocated end-to-end resources (active flows) and (3.5) represents a *demand constraint* that ensures any allocated flow rate does not exceed the corresponding demand, so that no resources will be left unused. $s_{k,i,j}$ in (3.6) represents the segment of flow k traversing link $l_{i,j}$ for which $t_{k,i,j}$ airtime is allocated. $c_{i,j}$ denotes the maximum achievable data rate

between an (i, j) base station pair, and I work with aggregate data traffic flows between base stations and the gateway.

As single transceiver stations can only send to, or receive from one neighbour at a time, let us construct a conflict graph $G(V,E)$, where a flow segment corresponds to a vertex $v \in V$. An edge $e \in E$ exists between any two vertices, if the corresponding flow segments cannot be simultaneously active, either because they traverse the same node or because they may cause secondary interference onto one another, due to beam alignment and transmission range. C_q denotes a *clique*, which follows the definition given below.

Definition 1. *A 'clique' is the set of all flow segments that cannot be active simultaneously.*

Note that a flow segment can belong to multiple cliques and denote $\mathcal{C}$ the set of all cliques. I exemplify the conflict graph and clique notions with the simple topology depicted in Fig. 1.1, for which the equivalent conflict graph is shown in Fig. 3.3. Observe that two cliques exist in this example and the segments of flows 1 and 2 over link $l_{3,4}$, i.e. $s_{1,3,4}$ and $s_{2,3,4}$, simultaneously belong to both. [1] Returning to the problem to be solved in this chapter, by (3.6) I introduce a *clique constraint* that guarantees the total time consumed by all flow segments in a clique does not exceed 1, i.e. $\sum_{s_{k,i,j} \in C_q} t_{k,i,j} \leq 1, \forall k \in \mathcal{F}_{\mathcal{A}}, \forall C_q \in \mathcal{C}$.

In solving this problem, it will also prove useful to work with the notion of *conflict node*, defined on the actual network topology as below.

Definition 2. *In a backhaul network, a 'conflict node' is a base station that forwards traffic on behalf of others.*

For the example shown in Fig. 1.1, base stations 3 and 4 are conflict nodes.

Solution Existence

To verify whether a solution to the problem (3.4)–(3.6) exists, i.e. max-min fair allocation in a multi-hop mm-wave network is feasible, I first characterise the network's rate region.

Lemma 1. *The rate region of a multi-hop mm-wave backhaul network is convex.*

[1] In this example, cliques are only formed as a results of single-transceivers operating at each node and no secondary interference can be observed. Had node 1 been on the same direction as the (4,6) link, $s_{1,1,3}$ would have formed a third clique with $s_{1,4,5}, s_{2,4,6}$, and $s_{3,4,6}$.

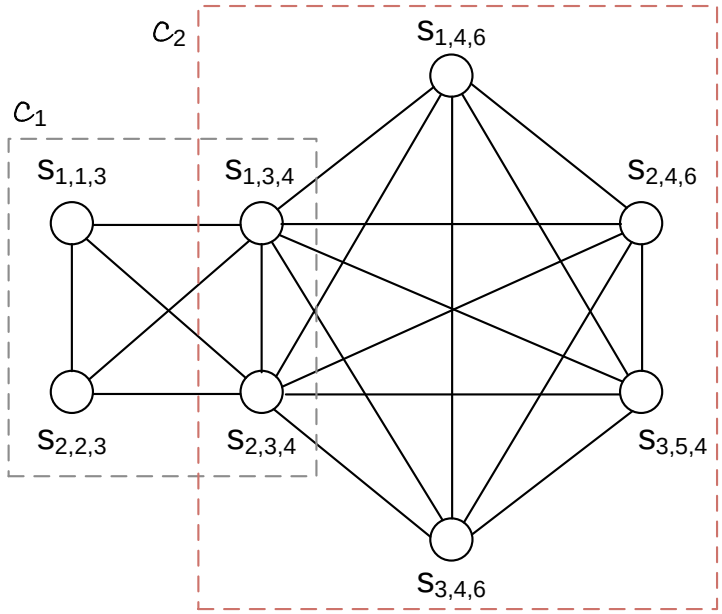


Figure 3.3: Conflict graph corresponding to the topology in Fig. 1.1. Each vertex corresponds to a segment of a flow k between base stations i and j. Cliques highlighted with dashed lines.

Proof. Since I consider transmissions between base stations are precisely scheduled, channel access in a clique can be seen as a single-hop time division multiplex (TDM) instance, which is known to have a convex capacity region (Cover and Thomas, 2006). The throughput of any sub-flow $s_{k,i,j}$ in a clique C_q is upper bounded by the minimum between the throughput allocated in the clique C_{q-1} traversed previously and the total flow demand d_k. The network rate region is obtained by the appropriate intersection of the rate regions of the component cliques. Thus it is convex. □

The following key result follows.

Corollary 1. *Max-min fair allocation in multi-hop mm-wave networks exists and it is unique.*

Proof. Following Theorem 2 in (Jaffe, 1981), Gafni and Bertsekas demonstrate that if a max-min fair rate allocation vector exists, then it is unique (Gafni and Bertsekas, 1984, p.1011). Radunovic and Le Boudec, prove by contradiction that compact convex sets are max-min achievable (Radunović and Le Boudec, 2007). As per Lemma 1 above, the rate region of a scheduled mm-wave backhaul is convex. Therefore a max-min fair rate allocation vector exists in mm-wave backhauls and it is unique. □

Finally, the rate region has the free disposal property (Radunović and Le Boudec, 2007) since each element of the rate vector $\mathbf{r} = \{r_k \mid k \in \mathcal{F}\}$ is lower bounded by

zero and any non-zero feasible allocation can always be decreased. It follows that a progressive filling algorithm can be employed to find the solution to the max-min fair allocation problem with mm-wave particularities.

3.3 WiHaul: Max-min Fair Backhauling

This section presents a max-min fair multi-hop mm-wave backhauling mechanism, namely WIHAUL. This consists of a progressive filling algorithm that solves the optimisation problem (3.4)–(3.6) in polynomial time, and a light-weight scheduling protocol that distributes airtime solutions among base stations, ensuring they communicate at the right time for the computed duration. As an alternative, a brute-force based approach could be employed to exhaustively search for a max-min allocation in the feasible space, but computational complexity would increase exponentially as the number of flows or the demand vary. The proposed solution handles mm-wave specific PHY impairments such as link blockage, as the progressive filling routine updates airtime allocation as a result of changes in the rate regions when such events occur. The scheduling procedure further handles terminal deafness (and secondary interference), as it builds on the notion of clique introduced above, which ensures appropriate spatial reuse while transceivers and receivers always have their beams aligned when intending to communicate.

3.3.1 Progressive Filling Algorithm

Algorithm 1 summarises the progressive filling procedure proposed to achieve max-min fair allocation of the backhaul resources under clique and demand constraints, and I detail its operation next. The progressive filling starts with all flow rates equal to zero and considering none of the aggregate flows have been allocated resources (lines 1–2). Flows for which an allocation was not performed are called active flows. The algorithm gradually increases flow rates simultaneously, in steps of size ε Kbps (line 4) until one or more flows either meet their demands (line 6) or activate a clique constraint (line 13). Note ε is a configurable parameter whose magnitude impacts on algorithm runtime. If a flow's demand d_k is satisfied, Algorithm 1 freezes the allocated rate r_k to the demand and remove that flow from the active set (line 8), thereafter considering it inactive and its resources frozen.

When a clique is fully utilised, Algorithm 1 will stop increasing the rates of the

Algorithm 1 Progressive Filling

$r_k = 0, \forall k$ ▷ Initialisation
$\mathcal{F}_\mathcal{A} := \mathcal{F}$ ▷ Set of active flows
while $\mathcal{F}_\mathcal{A} \neq \emptyset$ **do** ▷ Loop until all flows allocated
    $r_k += \varepsilon, \forall f_k \in \mathcal{F}_\mathcal{A}$ ▷ Increase rates of all active flows with same step
    for $\forall f_k \in \mathcal{F}_\mathcal{A}$ **do**
        if $r_k \geq d_k$ **then** ▷ Flow demand satisfied
            $r_k := d_k$;
            $\mathcal{F}_\mathcal{A} = \mathcal{F}_\mathcal{A} \setminus \{f_k\}$ ▷ Remove flow from active set
        end if
    end for
    for $q = 1 : |\mathcal{C}|$ **do** ▷ Loop over all cliques
        $t_{k,i,j} = r_k / c_{i,j}, \ \forall s_{k,i,j} \in C_q$ ▷ Time consumed by each flow segment in C_q
        if $\sum_{s_{k,i,j} \in C_q} t_{k,i,j} \geq 1$ **then** ▷ Clique constr. not met
            $t_{\text{left}} = 1$ ▷ Total airtime budget
            $S = 0$ ▷ Sum of inverse capacities of links traversed by active flows
            for $\forall s_{k,i,j} \in C_q$ **do** ▷ Loop over all sub-flows
                if $f_k \in \mathcal{F} \setminus \mathcal{F}_\mathcal{A}$ **then** ▷ Flow inactive
                    $t_{\text{left}} = t_{\text{left}} - t_{i,j,k}$ ▷ Subtract airtime already reserved
                else ▷ Flow active
                    $S = S + 1/c_{i,j}$ ▷ Update sum for subsequent airtime weighting
                end if
            end for
            $R = t_{\text{left}} / S$ ▷ Rate to allocate for all active flows
            for $\forall f_k \in \mathcal{F}_\mathcal{A}$ **do** ▷ Loop over all active flows
                $r_k = R; \ t_{k,i,j} = r_k / c_{i,j}$ ▷ Allocate rate and airtime on each link
                Freeze r_k; $\mathcal{F}_\mathcal{A} = \mathcal{F}_\mathcal{A} \setminus \{f_k\}$ ▷ Freeze rate remove flow from active set
            end for
        end if
    end for
end while

flows traversing it and proceed with computing from scratch the rates should be assigned according to the remaining airtime budget. To this end, Algorithm 1 subtracts from the total available airtime, i.e. 1 (line 14), the fractions already reserved for *inactive flows* (line 18) and sum up the inverse of the link capacities corresponding to *active flows* in that clique (line 20). The latter will allow us to provide all active flows with the same rate R (line 23), which under heterogeneous link rate conditions translates into allocating airtimes to each sub-flow that are inversely proportional to the traversed link's capacity (line 25), i.e.

$$t_{k,i,j} = \frac{t_{\text{left}}}{c_{i,j} \sum_{s_{k,l,m} \in \mathcal{F}_{\mathcal{A}} \cap C_q} \frac{1}{c_{l,m}}}.$$

It is straightforward to verify that airtimes $t_{k,i,j}$ above sum to t_{left}, as required. Subsequently, the algorithm freezes the rates r_k of flows in clique C_q and remove them from the active set (line 26).

Algorithm 1 repeats this procedure for the remaining active flows, until meeting their demand or activating other clique constrains. The progressive filling algorithm terminates when the set of active flows is empty (line 3). At that point I have obtained the airtimes to be allocated for each flow on each traversed backhaul link, in order to fulfil the max-min fair allocation of the rates.

Algorithm 1's runtime is a function of the highest flow rate divided by the step-length, which recall is configurable, and the total number of flows. Therefore the algorithm solves the max-min fairness optimisation problem posed in polynomial time. The results presented in Sec. 3.4.7 confirm this assessment.

3.3.2 Scheduling Procedure

Terminal deafness is a major challenge in mm-wave networks. Therefore, unless stations know to which neighbour to steer their beams, when, and for how long, they may be locked out, which would lead to frame loss and overall performance degradation. Such degradation may also occur when beams of different communicating pairs partially overlap, resulting in secondary interference. Algorithm 1 described previously addresses the computation of airtimes for each flow segment, in order to attain max-min fair rates. To convey the computed airtimes and overcome TX/RX issues, i.e. deafness or secondary interference, WIHAUL employs a network-wide co-ordination

procedure based on a scheduling hierarchy. This enables a centralised controller to dictate when nodes can transmit to others without conflict and in which order, so as to maximise spatial reuse.

Algorithm 2 gives the pseudocode of WIHAUL's scheduling operation, which I explain next with the example topology shown in Fig. 1.1. I assume a central controller (typically placed at the gateway; here node 6) has full knowledge of the network topology, including the hop distance to each base station, which of these are conflict nodes (i.e. have more than one neighbour), as well as their addresses, i.e.

(1) H_i: hop distance from node i to the gateway,

(2) S_i: node i's conflict state,

$$S_i = \begin{cases} 1, & \text{if } i \text{ is a conflict node,} \\ 0, & \text{if } i \text{ is a leaf node;} \end{cases}$$

(3) A_i: node i's unique ID (e.g. its IP address).

With this information and the airtime shares computed by Algorithm 1, the controller constructs a hierarchy to establish when a node should transmit/receive and when it should schedule its neighbours, respectively. Specifically, WIHAUL first considers all conflict nodes as eligible candidates for acting as scheduling coordinators (in our example nodes 4 and 3). Among these, the one with the lowest hop distance $H_c = \min_{\{i|S_i=1\}} H_i$ is designated as the root coordinator and placed at the top of the scheduling hierarchy, namely at Level 0. In this example it is node 4 that acts as coordinator, while 6 (the gateway) is not a conflict node. The remaining nodes with $S_i = 1$ will be placed at a level that depends on the difference between their H_i value and that of the main coordinator (H_c) i.e. Level $i = |H_i - H_c|$. Nodes with $S_i = 0$ will be placed at Level$_i$ below their neighbouring conflict node. As such, in our example nodes 5 and 6 reside at Level 1, while 1 and 2 at Level 2, as illustrated in Fig. 3.4.

At each level of the hierarchy, WIHAUL assigns airtime top–down, a node accepting the time allocated by its parent and assigning SPs to its children. In the considered example, the protocol first assigns SPs for 4 and then the nodes at Level 1, i.e. 3, 5 and 6. In turn, node 3 assigns SPs to 1 and 2, outside the interval when it is involved in communication with 4. This allows for spatial reuse, as links $l_{4,5}$ and $l_{3,1}$, and respectively $l_{6,4}$ and $l_{3,2}$ will be active simultaneously.

Algorithm 2 Max-min Fair Scheduling

Obtain air time shares $t_{k,i,j}, \forall k,i,j$ with Algorithm 1
$\mathcal{H}$ = BUILD SCHEDULING HIERARCHY(network topology)
Root coordinator of $\mathcal{H}$ assigns slots to its child nodes, i.e. Level 1 nodes, given total airtime available
while !bottom level of $\mathcal{H}$ **do**
    Order conflict nodes by A_i in increasing order
    for all conflict nodes **do**
        Accept airtime assigned by parent node
        for all child nodes of current parent **do**
            if node's priority lower than others in clique **then**
                Mark time slots used by other nodes as taken
            end if
            Assign airtime to child nodes
        end for
        Move to the next level
    end for
end while

function BUILD SCHEDULING HIERARCHY(topology)
    $\mathcal{L} \leftarrow 0$; ▷ Level 0
    Set node with $H_c = \min_{\{i|S_i=1\}} H_i$ as root coordinator
    Place the root coordinator on $\mathcal{L}$
    while !(all nodes assigned a level) **do**
        $\mathcal{L} \leftarrow \mathcal{L}+1$ ▷ Advance level
        Place on current level nodes i with $|H_i - H_c| = \mathcal{L}$
    end while
    return Scheduling hierarchy $\mathcal{H}$
end function

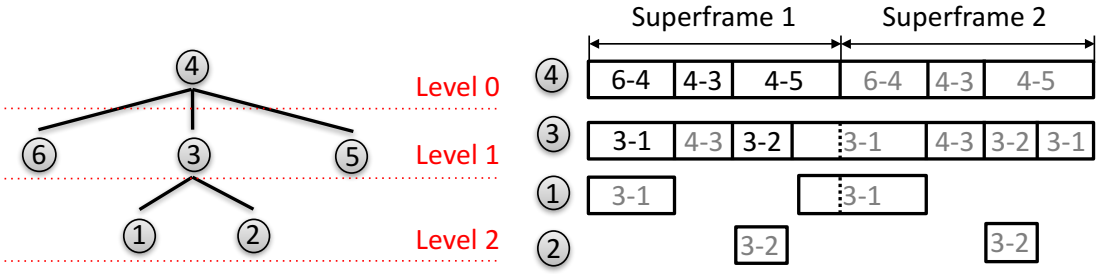


Figure 3.4: Hierarchical scheduling structure corresponding to the topology in Fig. 1.1 (left) and SPs allocated for each communicating pair transporting different flow segments (right). Links $l_{4,5}$ and $l_{3,1}$, and respectively $l_{6,4}$ and $l_{3,2}$ can be simultaneously active. SPs labelled in black represent those being scheduled by node circled; those in grey SPs by the parent of current node.

In case of multi-path routing, it may happen that two or more nodes on the same level share the same neighbouring node that they could schedule. In such cases, the node with the smallest identifier A_i takes priority and will be the one scheduling. In turn, the child informs the other candidate parents of the assigned time, to resolve the tie and avoid conflicts. This process is repeated until all computed SPs have been disseminated to all stations.

Subsequently, nodes will periodically switch their beams towards the corresponding neighbours for transmission/reception during the assigned times. To adapt to the dynamics of physical channel conditions (e.g. link blockage) and the changing flow demands, the controller will periodically (e.g. every beacon interval) collect link quality and flow demand information, run the progressive filling algorithm from scratch, and re-schedule flow segments as appropriate. The system check for network update during BIH as frequent as by the specific scenario (up-to one BI), and as we will show in Section 3.4.4, changes will take effect within one BI.

3.4 Performance Evaluation

To evaluate the performance of WiHaul, I implement this in NS-3 and conduct extensive simulations under different scenarios,[2] comparing with recent scheduling schemes for mm-wave networks, including DLMAC (Sim et al., 2016a), MDMAC (Singh et al., 2010), and variations of these. I examine achievable gains in terms of flow throughput distribution and overall network throughput, and the level of fairness each approach attains over realistic multi-hop topologies. I further analyse WiHaul's behaviour in

[2]The source code of the implementation is available at `https://git.io/wihaul`.

terms of allocated flow throughputs and airtimes, and give insight into the impact of link rates and flow demands on the partitioning of resources. Lastly, I evaluate the solution with real data traffic traces and examine end-to-end delay performance.

It is worth noting that making a definite comparison of the complexity entailed by my solution and the benchmarks considered is difficult. This is largely due to the different paradigms employed, i.e. centralised vs distributed, and random vs scheduled channel access. Unlike the proposed mechanism, the benchmarks are also subject to convergence times that depend on neighbourhood size and payload lengths, and may require restarting to cope with traffic dynamics. Slot alignment is also problematic in distributed settings, yet not explicitly discussed by the respective authors. By the proposed solution, the airtime allocation is a function of the highest flow rate and a configurable step length, while scheduling runtime depends on the number of nodes in a given topology.

3.4.1 Simulation Environment

While the proposed solution is applicable to any multi-hop mm-wave backhauls that operate in a scheduled mode, for evaluation purposes we employ the 802.11ad PHY with the SP based MAC, as this standard is already mature. To incorporate multi-hop frame relaying, the controller logic, and the progressive filling algorithm in NS-3, I extended the IEEE 802.11ad SP based MAC implementation of Facchi *et al.* (Facchi et al., 2017). The simulator incorporates the 802.11ad MAC frame structure and simple PHY functionality for directional multi-gigabit OFDM transmissions. The Beacon Interval (BI) header occupies a fixed fraction of the BI duration (100ms customisable), a configurable fraction of the BI duration. This proposal uses the default IEEE 802.11ad setting, i.e. 10% of the BI duration for overhead, which is also in agreement with the LTE and 5G NR frame structures.[3] Overall, this overhead interval is reserved for beamform training, control message exchange, and schedule dissemination. Specifically, should there be any changes in link capacity, routing, or flow demand, as will be discussed in details in Sec. 3.4.4, the BI overhead will cover the time required to propagate to backhaul nodes the flow rate allocations recomputed from scratch by the controller. Actual packets are exchanged during the data transmission interval (DTI), as scheduled by WIHAUL. I employ the Friis path loss model based on which the received

[3]LTE and subsequent 5G NR Type 2 frames for TDD access dedicate 10% of the frame duration for Downlink Pilot Time Slot (DwPTS), Guard Period (GP), and Uplink Pilot Time Slot (UpPTS), to handle TDD operation specifics (3GPP, 2018)

power and the SNR is computed, and then I use the SNR to map to specific modulation and coding scheme (MCS) which corresponds the link capacity, $c_{i,j}$, $\forall\{i,j\}$.

I implement a central controller that executes the proposed WIHAUL, including the progressive filling and scheduling algorithms, and incorporate measurement of MAC queue length to monitor events such as buffer bloat. If changes in flow demand or link capacity take place, the progressive filling and scheduling operations will be triggered to perform resource allocations in the next BHI. Further, the simulation tool incorporates MAC protocol data unit aggregation (A-MPDU) and MAC frame relaying to support efficient multi-hop backhauling scenarios. Assume that during DTI the TX/RX base stations are perfectly beam aligned as per the 'quasi-optical' vision in IEEE 802.11ad (IEEE 802.11ad Std., 2014). While PHY design remains outside the scope of this work, I also investigate the impact of secondary interference, and explains how WIHAUL tackles such interference in Sec. 3.4.5.

Given the switched operation of transmissions and receptions, and the high PHY bit rates employed on links, to avoid excessive delays and buffer overflows at relaying stations, WIHAUL divides the airtime allotted to each sub-flow into multiple SPs each of shorter duration. In the simulation evaluation, I work with 20 short SPs that sum up to the computed airtime allocations.

The NS-3 build-in module allows for full-stack simulation including application, transport and internet layers on top of the 802.11ad MAC and PHY. I work with applications that generate fixed packets of 1470 Bytes, except when experimenting with real traffic traces. The parameters used in simulation are summarised in Table 3.2.

3.4.2 Fairness Metrics

An allocation is max-min fair if increasing the rate of a flow is only possible by decreasing that of others (Bertsekas and Gallager, 1992). Note that max-min is a *qualitative* fairness criterion, and unlike e.g. Jain's fairness index, this typically does not have a directly measurable value. Therefore, to quantify fairness, I first resort to the concept of inequality distribution used in economics, and compute Gini coefficients (Gini, 1921), using the following formula:

$$\mathrm{G} = \frac{\sum_{k=1}^{n}\sum_{l=1}^{n}|r_k - r_l|}{2n\sum_{k=1}^{n} r_k},$$

where r_k is the rate allocated to flow k, and n is the total number of flows. The lower this coefficient is with a certain rate allocation vector, the more fair the distribution of resources is.

Parameter	Value
TX power	10dBm
TX/RX antenna gain	20dBi
BI duration	102400μs
BI overhead	10240μs
Progressive filling step length (ε)	10Kbps
UDP payload	1470B
TCP MSS	1460B
Time fraction allocated for TCP ACKs	0.06
TCP Initial Slow Start Threshold	64KB
TCP TX/RX Buffer Size	10MB

Table 3.2: Simulation settings.

To add further perspective and quantify to what extent the minimum flow rate in the network might be higher with WIHAUL than with other schemes, this chapter employs the generalised measure of fairness defined in (Lan et al., 2010), as follows

$$\mathcal{M}_\beta(\mathbf{r}) = \text{sign}(1-\beta)\cdot\left[\sum_{k=1}^{n}\left(\frac{r_k}{\sum_l r_l}\right)^{1-\beta}\right]^{\frac{1}{\beta}},$$

where β dictates different types of fairness measures. For max-min fairness $\beta\to\infty$, and $\mathcal{M}_\beta(\mathbf{r})$ becomes

$$\begin{aligned}\mathcal{M}_\beta(\mathbf{r}) &= \lim_{\beta\to\infty}\text{sign}(1-\beta)\left[\sum_{k=1}^{n}\left(\frac{r_k}{\sum_{l=1}^{n} r_l}\right)^{1-\beta}\right]^{\frac{1}{\beta}}\\ &= -e^{\lim_{\beta\to\infty}\log\left[\sum_{k=1}^{n}\left(\frac{r_k}{\sum_{l=1}^{n} r_l}\right)^{1-\beta}\right]^{\frac{1}{\beta}}}.\end{aligned}$$

Denote $y_k = (\sum_l r_l)/r_k$ and solve the limit above by applying l'Hôpital's rule, which leads to $\lim_{\beta\to\infty}(\sum_{k=1}^{n} y_k^{\beta-1}\log(y_k))/(\sum_{k=1}^{n} y_k^{\beta-1})$. As $\beta\to\infty$, the numerator is dominated by the highest y_k term, i.e. $\max_k\{y_k\log(y_k)\}$, hence the limit converges to $\max_k \sum_l r_l/r_k$ and max-min fairness can be measured with

$$\mathcal{M}_\beta(\mathbf{r}) = -\max_k\left\{\frac{\sum_l r_l}{r_k}\right\}. \tag{3.7}$$

3.4.3 Comparison with State-of-the-Art Solutions

I compare the performance of WiHaul against that of recent mm-wave scheduling schemes, namely DLMAC (Sim et al., 2016a) and MDMAC (Singh et al., 2010), in terms of mean and total network throughput, and inter-flow fairness. I conduct the evaluation over several topologies generated with the Cerdà-Alabern model that captures the characteristics of real-world multi-hop wireless deployments (Cerda-Alabern, 2012). The topologies considered comprise 10 to 15 stations (including the Internet gateway) and the number of aggregate flows traversing the network varies between 7 and 10. Fig. 3.5 illustrates four of these topologies[4], where the X and Y axes represent the base stations' coordinates, with base station 0 being the gateway. Link rates vary between 2.772–6.756Gbps, depending on distance between stations.

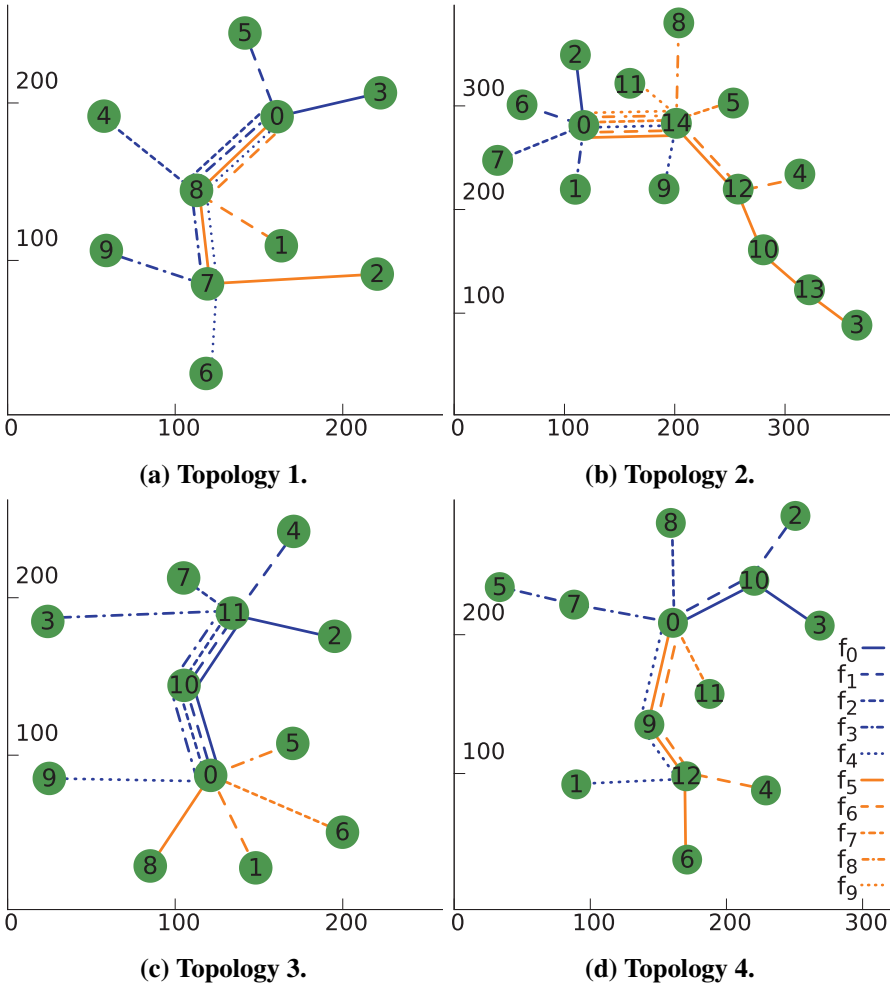
I also compare against optimised DLMAC and MDMAC versions that seek to reduce gaps between transmissions (BinDLMAC) (Sim et al., 2016a) and operate with slot sizes that maximise transmission efficiency respectively (OptMDMAC).[5] Note all these are decentralised and do not explicitly consider fairness in their design. Each approach transports backlogged aggregate flows (unlimited demand) transmitted over UDP.

Let us examine first Figs. 3.5e–3.5h, where I show the average and 95% confidence intervals of individual flow throughputs attained with WiHaul, DLMAC, MDMAC, and their variations, in each topology considered. These figures also plot the average throughput performance over all flows as the last cluster of bars to the right of each plot. Observe in these clusters that the bars corresponding to WiHaul are indeed the highest and the total network throughput ranges between 2.25-2.5Gbps in all cases. Hence, WiHaul achieves the highest average flow throughput (and therefore total network throughput), irrespective of the number of hops flows traverse and with how many competing flows they share links.

Flows that encounter less competition attain superior performance with the proposed approach, without negatively impacting on the others. This can be observed in Figs. 3.5e and 3.5f, where with WiHaul flows f_0 and f_1, and respectively f_0–f_3 achieve approximately 450Mbps and 100Mbps more throughput than the other flows traversing the backhaul. At the same time, the proposed method reduce the gross per-

[4]Experiments conducted with more topologies show similar conclusions. Results obtain with four of these are included for conciseness.

[5]The default MDMAC design works with a slotted channel where slot size is fixed to 20μs. The optimised version considered works with slots that can accommodate exactly one transmission burst.



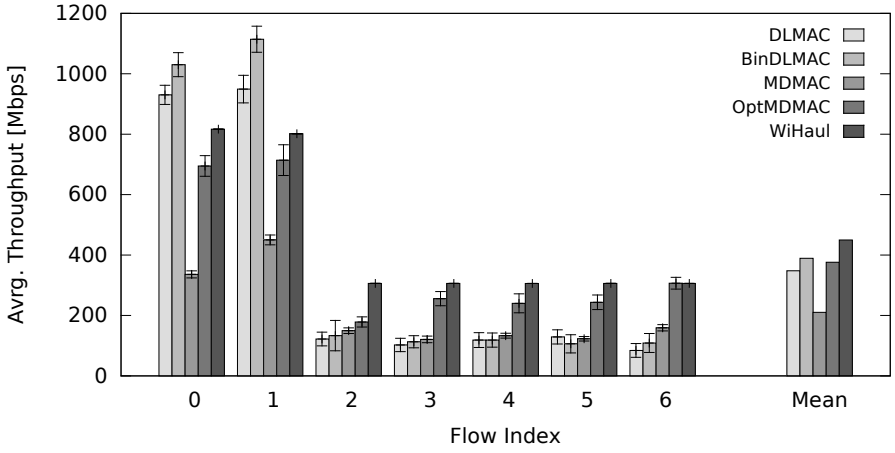
(a) Topology 1. (b) Topology 2.

(c) Topology 3. (d) Topology 4.

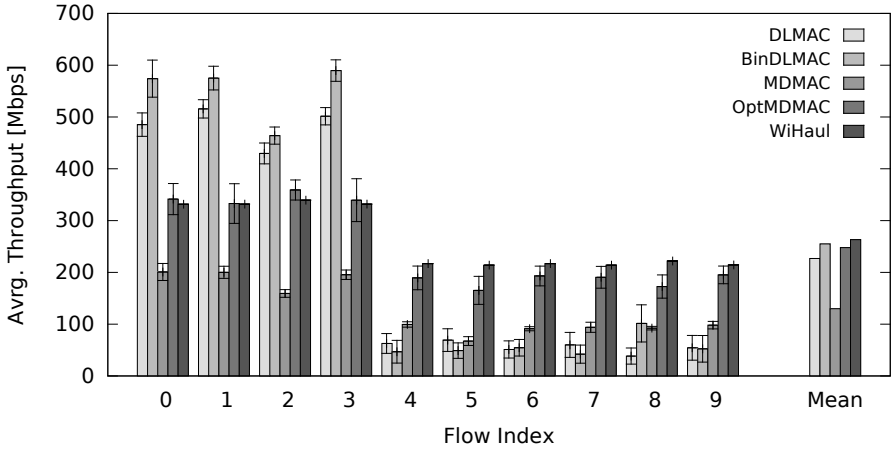
Figure 3.5: Multi-hop topologies used for performance evaluation, generated with the Cerdà-Alabern model (Cerda-Alabern, 2012); x-y axis in metres.

formance dissimilarity between flows (e.g. up to 1Gbps between flows f_1 and f_3 with BinDLMAC in topology 1). In addition, the flows penalised by earlier approaches attain up to 5× higher throughput with WIHAUL (observe flow f_4 in Fig. 3.5f with WIHAUL and BinDLMAC). Hence, with WIHAUL, flows attain similar throughput as long as they share the same cliques, while additional underutilised network resources are equally divided among unconstrained flows.

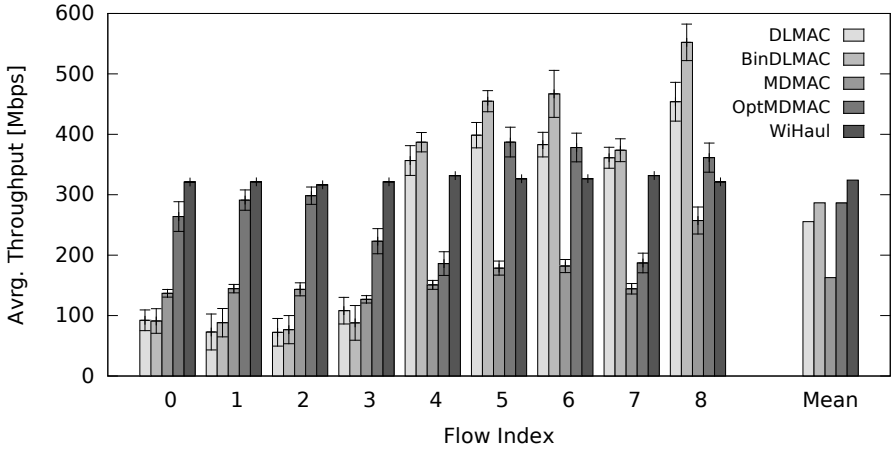
Note in Figs. 3.5g–3.5h that with WIHAUL all flows achieve the same throughput for topologies 3–4, unlike with DLMAC, MDMAC, and their variations, which largely favour flows terminating closer to the gateway and penalise those with endpoints multiple hops away. (Opt)MDMAC is less prone to such behaviour, though has the disadvantage of requiring appropriate configuration of the slot size, which is impractical. Nonetheless, although the 'optimised' MDMAC version performs rela-



(e) Throughput distribution in Topology 1.

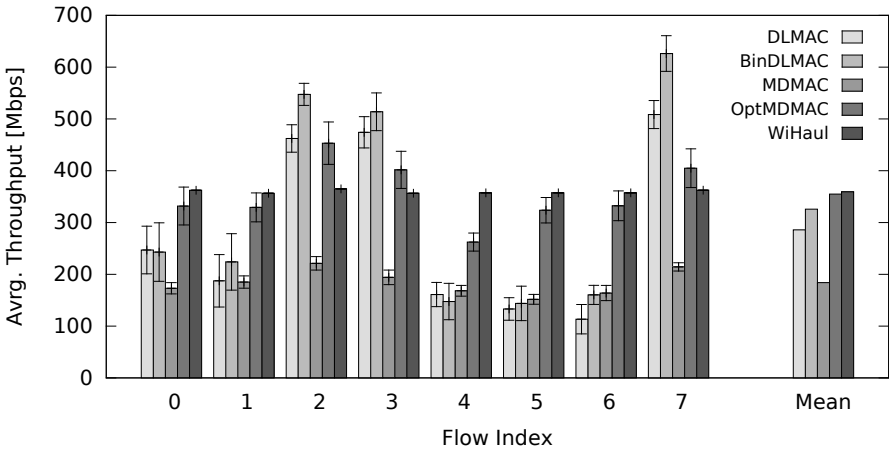


(f) Throughput distribution in Topology 2.



(g) Throughput distribution in Topology 3.

tively well overall, it still carries unfairness, as e.g. with this scheme flow f_7 in the third topology attains nearly half the throughput provided by WiHaul (Fig. 3.5g). I



(h) Throughput distribution in Topology 4.

Figure 3.5: Throughput comparison of WiHaul and existing schemes over the topologies shown in Fig. 3.5. Simulation results.

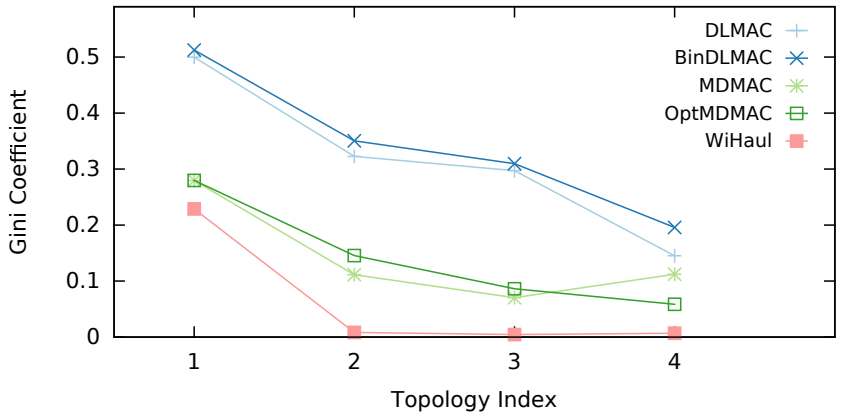


Figure 3.6: Gini coefficients corresponding to the throughput distribution attained by each scheme in topologies in Fig. 3.5. Simulation results.

conclude that WiHaul does not unnecessarily penalise flows that terminate/originate further away from gateways.

To examine closer the fairness properties of all schemes, Fig. 3.6 plots the Gini coefficients corresponding to the flow rate allocations each of these yields in the 4 topologies considered. Recall the Gini coefficient gives a numerical representation of inequality, with a lower value corresponding to a fairer allocation. Observe that although these values depend on the network topology, number of flows, and link rates, WiHaul outperforms the existing schemes, being in particular considerably more fair than the DLMAC variants. Precisely, the Gini coefficients when the network operates with BinDLMAC range between 0.2 and 0.5 and are the highest in all 4 topologies.

Topology \ Scheme	DLMAC	BinDLMAC	MDMAC	OptMDMAC	WIHAUL
1	-28.912	-25.658	-12.186	-14.75	**-10.289**
2	-31.783	-33.605	-11.542	-13.848	**-9.2215**
3	-20.174	-18.116	-9.7012	-10.826	**-8.0635**
4	-20.084	-28.705	-12.31	-11.517	**-9.3743**

Table 3.3: $\mathcal{M}_\beta$ measure of max-min fairness as derived in (3.7) following Lan's model (Lan et al., 2010). Simulation results.

DLMAC performs marginally better, while (Opt)MDMAC yields Gini coefficients in the 0.1–0.3 range. This proposal leads to the lowest Gini coefficients in all topologies (0.004–0.2), being substantially less unfair than the others. These properties are further confirmed by the results given in Table 3.3, which shows the fairness measure as derived in (3.7) for the proposed approach and the benchmarks considered. Indeed $\mathcal{M}_\beta$ is up to $5\times$ higher with the proposed approach, which also indicates WIHAUL ensures superior performance for the smallest flow, yet remains fair to the others.

I conclude that existing decentralised approaches bias against flows with longer hop-distance and/or inferior link rates; in contrast, the proposed WIHAUL not only achieves more fair partitioning of resources among all traffic flows, but also higher throughput for the smallest flow and overall higher mean throughput performance. This is because the decentralised approaches by design fail to take into account the end-to-end flow rate provisioning, henceforth the flows' throughputs are capped by the lowest throughput of the links traversed. This will naturally lead to wastage of network resources. In contrast, the proposed WIHAUL performs rate allocation based on knowledge of the conditions on the links traversed by each flow, and takes into account how much the users actually demand in order to best utilise channel resources. Although these comparisons are carried on UDP traffic, one should envision that TCP will only make the contrast in terms of throughput performance sharper, as without the end-to-end flow rate provisioning the decentralised approaches suffer from packet drop at intermediate base stations, which will in turn trigger TCP congestion control to slow down transmissions. All of these have important practical implication on cellular backhauls where WIHAUL could provide superior and more homogeneous service guarantees to users.

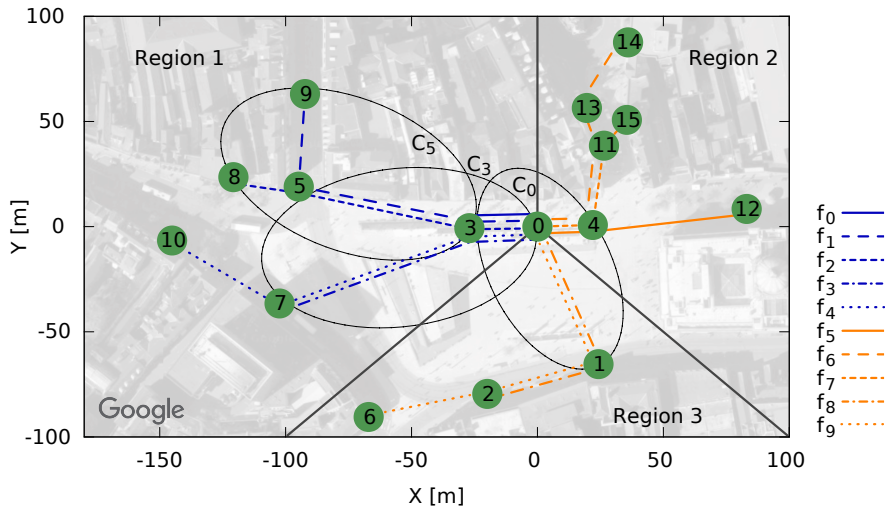


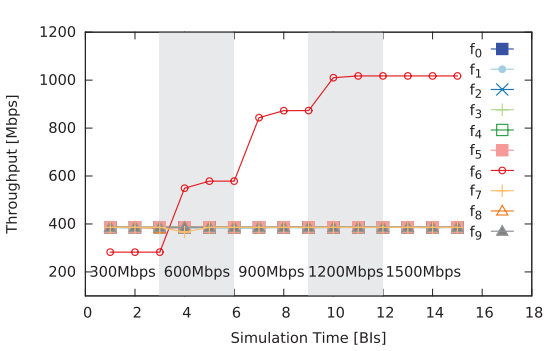
Figure 3.7: Lamppost small cell backhaul deployment in Nottingham city centre, operating on mm-wave links. Geographic information extracted from open data set (Council, 2012); backhaul carries 10 aggregate flows; cliques of interest circled; topology 'partitioned' into 3 regions.

3.4.4 Dynamic Conditions

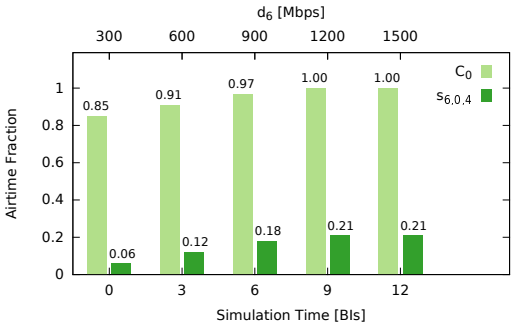
Next I undertake an in-depth analysis of WIHAUL's operation, investigating the impact of network dynamics in terms of link quality, flow demand, and routing strategy, on the airtime allocation and end-to-end performance. It is important to verify the performance of the proposed protocol under the above mentioned dynamic conditions, which occur often in mm-wave backhauls. For this I envision a lamppost based deployment in the Old Market Square of Nottingham as shown in Fig. 3.7, which I obtain from a publicly available data set (Council, 2012). This topology consists of 16 base stations (STAs) that communicate over mm-wave links and we envision 10 aggregate flows from the gateway (STA0). Also shown in the figure are three cliques of interest and, for ease of explanation, I consider the deployment as 'partitioned' into three regions. The results shown in what follows will demonstrate that max-min fair backhauling requires a non-trivial partitioning of the available airtime resources, which depends on the demand of each flow, the paths traversed, and the capacities of the links these comprise.

3.4.4.1 Demand Variation

Let us first examine a scenario where the demand of a single flow (i.e. f_6 originating at STA0 and terminating at STA14) grows from 300Mbps to 1.5Gbps, while that of the others remains fixed to 400Mbps. The goal here is to understand how this impacts



(a) Time evolution of average flow throughputs.



(b) Airtime allocation for f_6, i.e $t_{6,0,4}$, and total airtime in C_0.

Figure 3.8: Throughput performance and distribution of resources with WIHAUL in the topology in Fig. 3.7 as the demand of f_6 increases from 300Mbps to 1.5Gbps, in 300Mbps steps every 3 BIs (in plot/top labels), while the demand of others remains at 400Mbps. Simulation results.

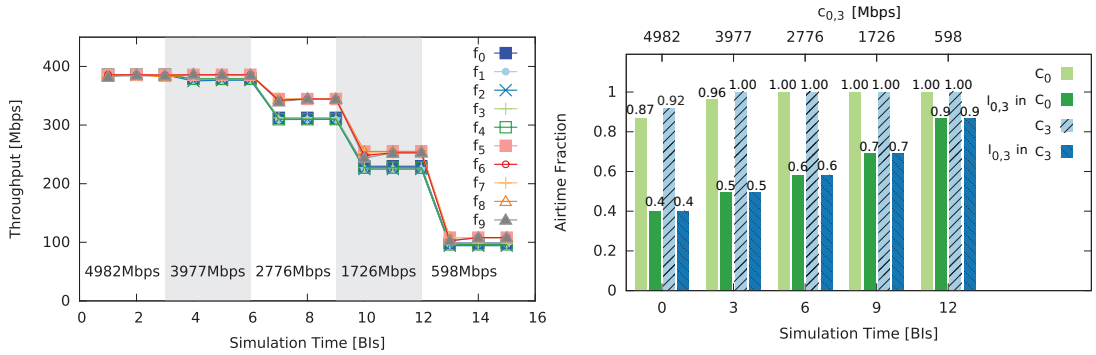
on airtime allocations and verify that the rates of the smallest flows are unaffected. Fig. 3.8 illustrates the results of this experiment, where I plot (a) the time evolution of the individual throughputs and (b) the fraction of airtime allocated to f_6 on link $l_{0,4}$, as well as the total airtime allocated in Clique C_0, which constrains f_6.

Observe that the throughput of f_6 increases with demand, up to 1Gbps, when the clique constraint is activated (total airtime in C_0 reaches 1) and the throughput is capped despite further growth in demand. As intended, the throughput of the remaining flows stays at 400Mbps, which indicates their demand is satisfied throughout. Note that the scheduling process is repeated every BI, link rate and demand updates are collected during BHIs, and takes one BI duration for the demand increase to propagate through the network.

To better understand the reasons behind these flow throughputs, we examine in Fig. 3.8b the airtime utilisation in the bottleneck clique C_0 and the time allocated to the demand-varying flow, $t_{6,0,4}$. Observe that initially there exist sufficient resources to accommodate the entire demand of flow f_6; this holds for a demand up to 900Mbps, when $t_{6,0,4}$ is tripled. Further increasing this demand does not result in a throughput increase above 1Gbps. This is because the proposed solution protects the remaining flows, which complies with the max-min fair allocation paradigm proposed.

3.4.4.2 Shared Link Degradation

Next let us examine the impact of link quality variation on the performance of all flows traversing such a link, when max-min fair allocation is performed. To this end, I



(a) Time evolution of flow throughputs.

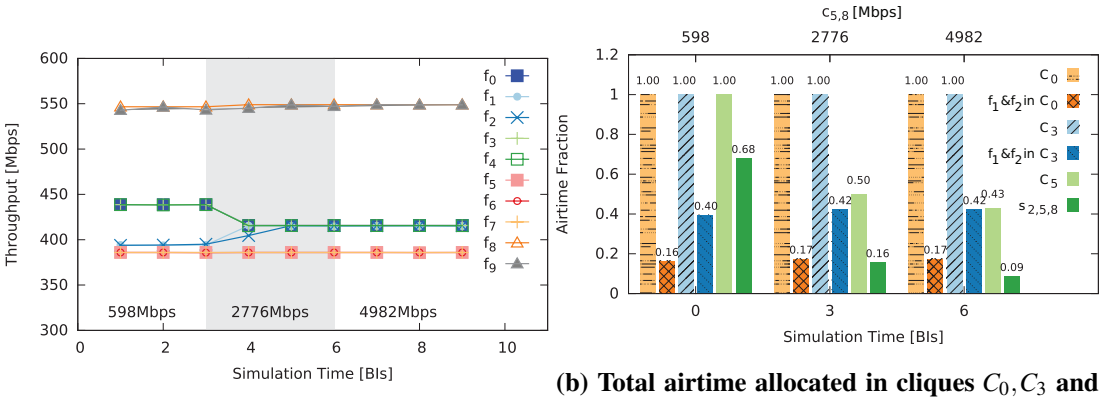
(b) Airtime on link $l_{0,3}$ in cliques C_0 and C_3; total airtime in these.

Figure 3.9: Flow throughputs and airtime fractions allocated to flow segments traversing $l_{0,3}$ in both cliques, as $l_{0,3}$ degrades. Flow demands remain at 400Mbps. MCS used with different link conditions labelled on white/shaded areas or show as top x-ticks. Simulation results.

simulate different degrees of link blockage between STA3 and STA0 (i.e. $l_{0,3}$), which results in signal attenuation between 5dB and 20dB. As a result, the MCS employed is reduced from 4.9Gbps to 598Mbps, to preserve link reliability. In this scenario, assume the bit rates of the other links remains constant and the demand of all flows is 400Mbps.

Fig. 3.9 illustrates the results of this experiment, where I measure (a) the individual flow throughputs and (b) the total time utilisation in cliques C_0 and C_3, as well as the sum of airtime fractions allocated to all flow segments traversing $l_{0,3}$, from the perspective of these cliques. Note that the airtime allocation on $l_{0,3}$ is effectively fixed under each link quality condition, but it may well represent different fractions from cliques' perspectives. When the link quality is high (i.e. $c_{0,3} = 4.982$Gbps), the total airtime consumption in C_0 and C_3 is below 1, hence all flows are satisfied. This is indeed confirmed by the flow throughputs shown in Fig. 3.9a. Subsequently, when a 5 dB attenuation is introduced at the third BI, the throughputs of flows f_0–f_4 drop slightly, while those of f_5–f_9 remain satisfied. That is because C_0 still has sufficient resources (airtime consumed sums to 0.96), while the C_3 clique constraint becomes active (airtime reaches 1). This can be observed indeed in Fig. 3.9b, where we also see that the total airtime allocated on link $l_{0,3}$ increases from 0.4 to 0.5 in both cliques, as a result of signal degradation.

Further attenuation on link $l_{0,3}$ (yielding 2.776Gbps bit rate), leads to the activation of the C_0 constraint (observe in Fig. 3.9b that the total airtime in clique C_0 reaches 1), and consequently to a decrease in the throughput of all flows. However, as C_3 becomes



(a) Time evolution of flow throughputs.

(b) Total airtime allocated in cliques C_0, C_3 and C_5; time assigned to f_1 and f_2 in C_0 and C_3; time allocated to f_2 on $l_{5,8}$, i.e. $t_{2,5,8}$.

Figure 3.10: Throughput performance and resource partitioning as $c_{5,8}$ increases from 598 to 4,982Mbps. Flow demands in region 1, 2 and 3 are 500, 400, and 600Mbps respectively. Shaded and label areas/x-ticks at the top correspond to bit rates on $l_{5,8}$ as link quality changes. Simulation results.

constrained before C_0, flows f_0–f_4 attain slightly lower (approx. 30Mbps) throughput than f_5–f_9. Lastly, this performance gap shrinks as link $l_{0,3}$ degrades further (BI 9 onward) and additional degradation would completely close the gap to meet the max-min fairness criterion. Meanwhile, the total time consumed by link $l_{0,3}$ to transport all flows is increasing to as much as 0.9 at the end of the simulation (see BI 12 in Fig. 3.9b).

I conclude that degradation of an intensively shared link (and clique) has a significant impact on the throughput performance of the entire network. Nevertheless, WIHAUL guarantees max-min fair allocation of the flow rates.

3.4.4.3 Heterogeneous Demands and Cascaded Cliques

Following we consider more complex circumstances where the demands of flows in regions 1–3 as shown in Fig. 3.7 are 500, 400, and 600Mbps respectively, while the quality of link $l_{5,8}$ varies. Signal attenuation decreases and capacity grows from 598 to 4,982Mbps on this link after every third BI. As $l_{5,8}$ only carries flow f_2, we investigate in Fig. 3.10b the changes in time allocation within all the cliques that f_2 traverses, i.e. C_5, C_3, and C_0, and show the time evolution of individual flow throughputs in Fig. 3.10a.

Note that as $c_{5,8}$ increases, more airtime is made available for both f_1 and f_2, as they share the same clique C_5. In effect, the constraint of this clique is removed (total

airtime consumption drops from 1 to 0.5) and this also impacts on the flows with which f_1 and f_2 share cliques C_3 and C_0, i.e. f_0, f_3 and f_4. Precisely, the throughput of these drops to 415Mbps after the third BI. As the quality of $l_{5,8}$ further increases, the total airtime allocated to f_2 on this link, i.e. segment $s_{2,5,8}$, decreases, though the flows in region 1 are together constraint by C_3. This confirms the proposed max-min fair allocation strategy ensures f_2 is not allocated more resources in cliques C_3 and C_0, as this would come at the cost of a decrease in f_0, f_3 and f_4's throughput. Lastly, observe that the throughput of the other flows remains unaffected, as the demand of $f_5, f_6,$ and f_7 is the smallest among all (i.e. 400Mbps) and changes in $c_{5,8}$ do not affect clique C_0, which is shared by all flows.

3.4.4.4 Dynamic Routing

Routing changes can happen in backhaul networks due to link blockage, buffer overflows, or other routing decisions made by a routing algorithm running at the networking layer. Next, we investigate the impact of route changes on the airtime allocation and end-to-end throughput performance, when the backhaul is managed with the proposed WIHAUL solution. To this end, we consider a situation where part of the traffic traversing links $l_{0,4}$ and $l_{0,1}$ in the topology depicted in Fig. 3.7, i.e. flows f_6, f_7, and f_9 are rerouted to STA2 and STA11 (i.e. no longer traverse STA1 and STA4), while the routes followed by f_5 and f_8 remain unchanged. After 6 beacon intervals, the initial routing topology is restored. We illustrate these changes in Fig. 3.11.

Fig. 3.12a illustrates the end-to-end throughput dynamics for all flows, as a result of these route changes, and in Fig. 3.12b the corresponding time allocation on links $l_{0,4}$, $l_{0,11}$, $l_{0,1}$, and $l_{0,2}$. Observed that WIHAUL reacts fast by re-allocating the airtime resources and the network throughput is only marginally affected. Flows f_9 and f_7 experience a 35Mbps drop at BI 4 due to the fact that packets buffered at STA4 and STA1 are partially dropped when the routes change, but the throughput recovers in the following BI. These results also confirm that WIHAUL will not unnecessarily penalise flows traversing more hops. In particular, when the routes change and the number of hops traversed by flows f_6, f_7, and f_9 decreases, after recomputing rates with the max-min criterion, their throughput is actually reduced, due to the fact that the clique c_0 consists of link segments, $l_{0,11}$ and $l_{0,2}$ that observe lower capacity as compared to links on the original paths, i.e. $l_{0,4}$ and $l_{0,1}$.

As expected, the time allocated on links $l_{0,4}$ and $l_{0,1}$ is reduced by approximately $2/3$ and $1/2$ when the routes of f_6, f_7 and f_9 change. Meanwhile, the time fractions

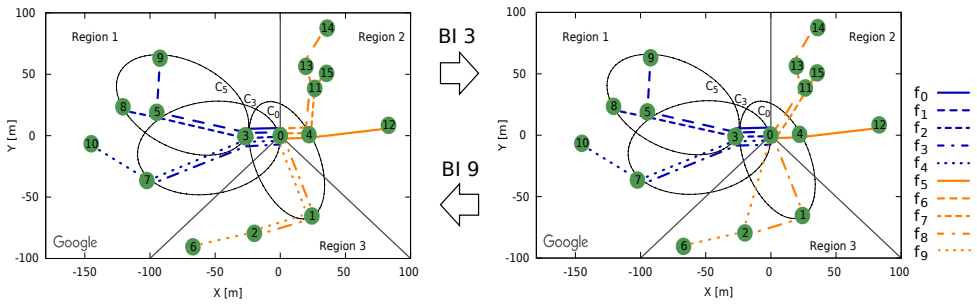
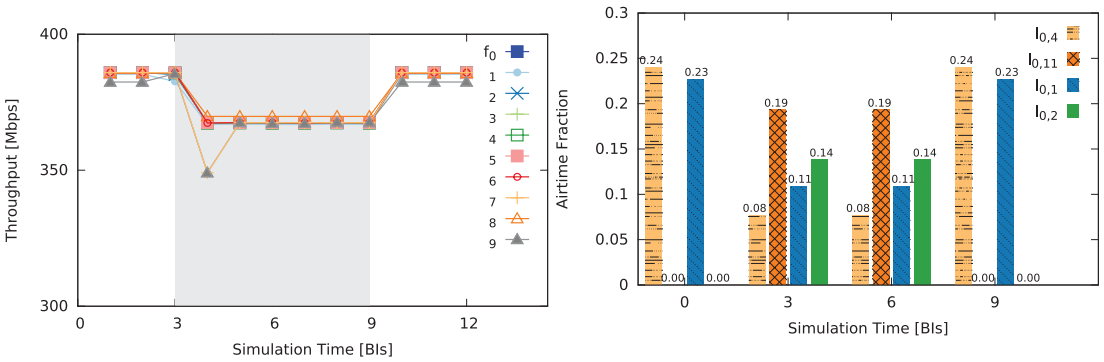


Figure 3.11: Changes in routing taking place at BI 3 and BI 9. At time BI 3, flows f_6, f_7 and f_9 are diverted from STA1 and STA4, while paths of f_5 and f_9 remain. At time BI 9, the routing changes back to original.



(a) Time evolution of flow throughputs.

(b) Airtime allocated for links $l_{0,4}$, $l_{0,11}$, $l_{0,1}$, and $l_{0,2}$.

Figure 3.12: Throughput performance and resource partitioning as routing of f_6, f_7 and f_9 changes. Flow demands for all are 400Mbps respectively. Shaded areas correspond to routing changes. Simulation results.

allocated to $l_{0,11}$ and $l_{0,2}$ increase from 0 to 0.19 and 0.14, which are both more than the amount reduced in $l_{0,4}$ and $l_{0,1}$. This is because the capacity of the new links employed by the new routes are lower than those on the initial routes.

3.4.5 Secondary Interference

In this subsection, we examine the potential impact of secondary interference, showing how WiHaul can overcome this by constructing cliques that capture such circumstances and avoiding their simultaneous activation during scheduling. We also discuss the complexity cost incurred when accounting for such secondary interference.

We simulate again the topology shown in Fig. 3.5b, where some links may interfere with each other when their TX-RX beam pairs are aligned. Specifically, when

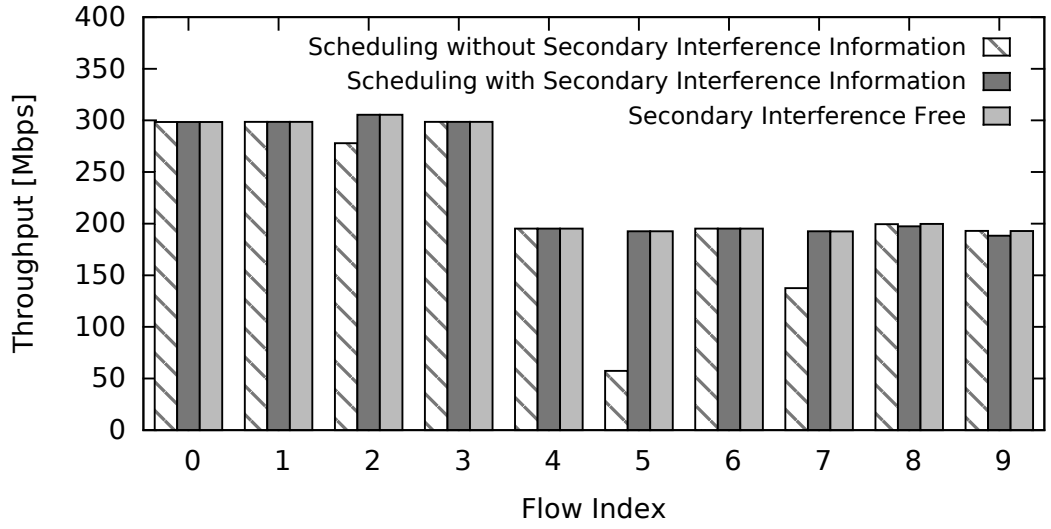


Figure 3.13: The existence of secondary interference can degrade the throughput of flows traversing interfering links (i.e. Flows 2, 5, and 7). Taking into account the potential secondary interference, WIHAUL**'s mitigates this effect.**

STA7 is receiving from STA0 and STA5 is transmitting to STA14, STA7 experiences secondary interference as the power of the signal it receives from STA5 has a level of -62.7 dBm. STA5 can suffer the same if receiving from STA14 and STA7's transmission to STA0 happens at the same time. Moreover, STA13's transmission to STA10 will interfere with STA14, if this is beam-switched to STA12 for reception, and vise versa. Fig. 3.13 illustrates the individual flow throughput averaged over 20 BIs when 1) secondary interference exists but the scheduling ignores this 2) secondary interference exists and WIHAUL incorporates this information when performing scheduling, and 3) the system is free of secondary interference. Observe that flows traversing the interfering links, i.e. f_2, f_5, and f_7, experience 50Mbps, 150Mbps, and respectively 70Mbps throughput degradation when schedules are assigned without accounting for such interference. When WIHAUL employs this knowledge for transmission coordination, the cliques are constructed such that none of the potentially interfering links are active simultaneously, regardless of whether this is due to secondary interference. As a result, the flow throughputs obtained when secondary interference is accounted for are virtually the same as those achieved in the idealistic case of the topology being free of secondary interference (given perfect beam shapes and pseudo-wired communication).

Given that secondary interference is in most cases marginal, and only 10% of the links in the simulated topologies shown in Fig. 3.5 experience secondary interference,

it is worth understanding the computational cost of scheduling with secondary interference in mind. Each group of interfering links forms a clique and a link with lower priority in the scheduling hierarchy (see Sec. 3.3.2) will have to store the time slots used by the links with higher priority in the hierarchy, which introduces n_{sche} iterations. Hence, accounting for secondary interference will increase the computation complexity of WIHAUL proportionally with the number of links that may interfere with each other if active simultaneously.

3.4.6 Real-Time Traffic

I complete the evaluation of WIHAUL by conducting experiments with real-time traffic potentially subject to latency constraints. We are particularly interested in the delay packets experience while traversing multi-hop mm-wave backhauls, where cascaded queues could have a negative impact on user experience. To this end, we emulate dynamic adaptive streaming over HTTP (DASH) by extracting meta-data from mobile traffic traces collected in New York City (Fund et al., 2014). We replay 100 such video sessions in parallel towards different base stations (download) in the topology shown in Fig. 3.7. The distribution of the session bit rates is shown in Fig. 3.14, where observe that individual bit rates vary between 100 Kbps and 3.4Mbps.

Under these circumstances, we measure the packet round-trip-time (RTT) for each aggregate flow over 30 seconds, as well as the average throughputs. We plot the RTT experienced by TCP segments in Fig. 3.15a, where observe this is below 30ms, with median values for all aggregates falling between 8 and 15ms. This complies with the NGMN Alliance specifications for end-to-end delay (20ms) in small cell backhauls

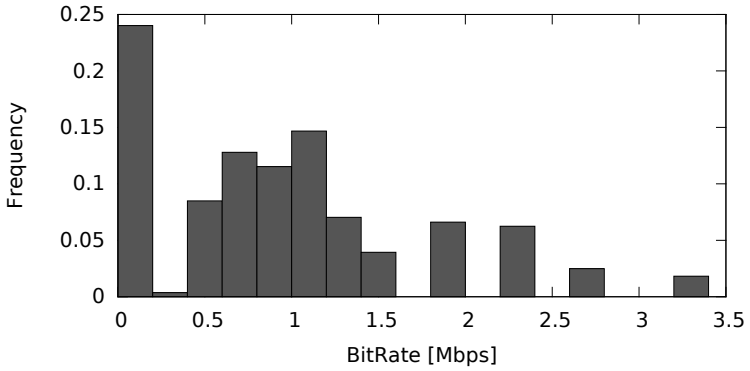
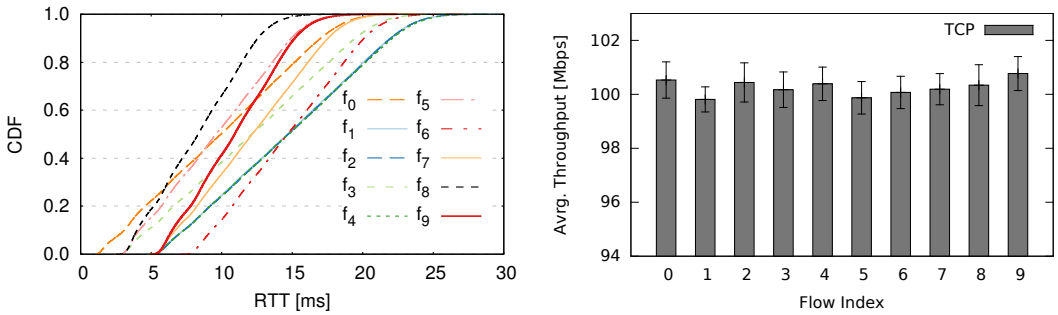


Figure 3.14: Distribution of DASH flow bit rates measured in New York City as reported in (Fund et al., 2014) and used here for the evaluation of WIHAUL.



(a) Round trip time distribution. **(b) Average aggregate flow throughput.**

Figure 3.15: CDF of packet RTTs for the aggregate flows (each consisting of 100 HTTP sessions following the bit rate distribution in Fig. 3.14) in the Nottingham topology shown in Fig. 3.7, and their average throughputs. Simulation results.

(Alliance, 2012). As expected, RTTs are proportional to the number of hops traversed, however, their distribution also depends on how frequently they are served. Precisely, note that the slope of the CDFs decreases with the number of aggregates traversing the first hop from the gateway and thus the latency in different regions is only scaled up by the number of hops each aggregate traverses. For instance, flows f_1, f_2, and f_4 are 3 hops away from the gateway (STA0) and share $l_{0,3}$ with f_0. As such, the RTTs they experience are identical (overlapping curves). Flow f_9 also traverse 3 hops, but only shares $l_{0,1}$ with f_8, hence their RTT distributions start at $\sim$5ms, but quickly diverge (medians 11 and respectively 15ms).

Turning attention to aggregate flow throughputs, we show the average and 95% confidence intervals of this metric in Fig. 3.15b. We see that overall performance is homogeneous (despite flows traversing different number of hops and experiencing different link rates), fluctuating around 100Mbps for each aggregate. Note that in this scenario all flows are satisfied and cliques are not constrained.

3.4.7 Runtime Performance

Lastly, we examine the runtime convergence of WIHAUL's progressive filling routine, to understand the practical feasibility of executing this algorithm periodically in order to perform airtime allocation. To this end, we take again the Nottingham topology depicted in Fig. 3.7, as this has a reasonably large number of nodes (i.e. 14) and aggregate flows traversing it (i.e. 9), which directly impact on the complexity. We measure the total time required by an off-the-shelf workstation, equipped with an Intel Core

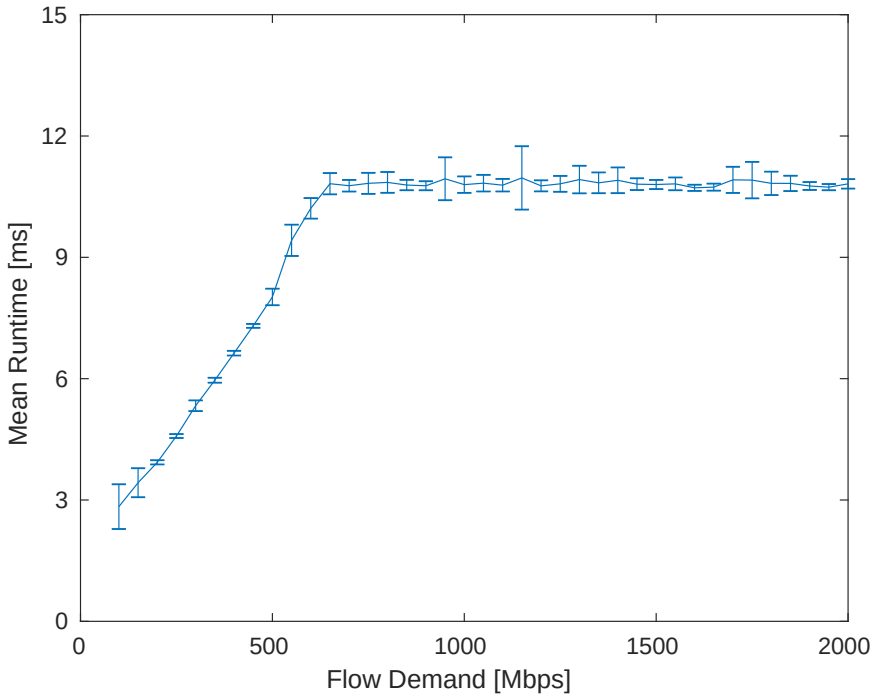


Figure 3.16: Progressive filling runtime performance as flow demands increase, in the topology shown in Fig. 3.7. Simulation results.

i5-4570 CPU clocked at 3.20GHz, to complete the execution of the progressive filling. For these measurements, we consider all flows have equal demands that range between 100Mbps and 2Gbps, with 50Mbps increments. For each case, we set the step length of the progressive filling algorithm to 10Mbps, execute this 100 times, and compute the mean runtime with 95% confidence intervals.

Observe that the proposed solution converges within a number of steps that, as long as clique constraints are not met (which will eventually happen, given limited channel capacities), strictly depends on the demand and the step size. For this topology, the runtime will not increase beyond approximately 11ms as the demand exceeds 650Mbps. We argue that this cost is negligible if the algorithm is run e.g. every second, while the granularity can be increased if the airtime allocation routine is fed with the output of a traffic forecasting mechanism (Zhang and Patras, 2018), and executed in anticipation of the expected traffic. Moreover, the step size used by the progressive filling algorithm will also have an impact on the runtime performance, i.e. the longer the step size, the faster the algorithm convergence, subject to potential reduction on accuracy.

3.5 Discussion

While this chapter of thesis addresses max-min fair rate and airtime allocation for aggregated flows traversing backhauls with a general focus on data traffic, ongoing 5G research efforts are increasingly concerned with accommodating diverse types of application scenarios. In order to address the different needs of various applications in shared network infrastructures, the 'network slicing' paradigm has emerged, thanks to advances in network programmability. Therefore, the next chapter changes the focus of my research towards optimal resource in backhauls that accommodate services with potentially conflicting requirements, including rate-intensive, delay-sensitive, best-effort, and revenue-driven.

As a possible future extension of this Chapter, the proposed max-min fair resource allocation scheme could be prototyped on commercial hardware, given that several platforms are emerging (TP-LINK, 2017; Mikrotik, 2017). The centralised control protocol is associated with certain communication overhead, while the decentralised benchmark solutions do not explicitly address fairness. It would therefore be interesting to investigate potential decentralised algorithms to achieve max-min fairness in mm-wave backhaul settings.

Although I provide a rate allocation and scheduling approach with focus on max-min fairness and achieving scheduled access, so that each base station knows when to transmit to whom, one could argue that certain schedules might be better than others. A possible extension is to devise a quantitative measure of different schedules' performance, by taking into account e.g. beam switching angles (which may also be associated with energy and/or airtime consumption).

3.6 Summary

By supporting multi-Gbps link rates, mm-wave technology is becoming a promising enabler of wireless backhauling solutions in ultra-dense cellular deployments. Highly directional beam-forming is mandatory to combat severe signal attenuation specific to these frequencies, though gives rise to cumbersome terminal deafness issue that must be tackled to fully exploit vast bandwidth resources. This chapter built upon the scheduled access paradigm as per IEEE 802.11ad (IEEE 802.11ad Std., 2014), 802.11ay (Cerwall (ed), 2017), and 3GPP 5G NR (3GPP, 2018) and proposed WI-HAUL, a network-wide airtime resource allocation and scheduling mechanism, which

explicitly guarantees inter-flow max-min fairness in mm-wave backhauls. The solution was validated over a broad range of network conditions and demonstrated via extensive simulations that WIHAUL achieves up to 5× higher measurable fairness as compared to existing mm-wave MAC proposals, improving up to fivefold the throughput of otherwise limited flows, while attaining superior overall network throughput. Further, I demonstrated that the progressive filling routine devised for airtime allocation completes within milliseconds and its complexity strictly depends on the highest flow demand. Lastly, the proposed approach is able to meet the typical delay constrains of real-time applications.

Chapter 4

Utility Framework and Optimisation in Mm-wave Backhauls

Advances in network programmability enable operators to 'slice' the physical infrastructure into independent logical networks. By this approach, each network slice aims to accommodate the demands of increasingly diverse services, which is particularly important for 5G use cases. However, precise allocation of resources to slices across future 5G millimetre-wave backhaul networks, in order to optimise the total network utility, is challenging. This is because the performance of different services often depends on conflicting requirements, including bandwidth, sensitivity to delay, or the monetary value of the traffic incurred. While the previous chapter tackled max-min fairness among aggregated flows of all data types, resource allocation must become increasingly aware of application-specific demands. Towards this end, this chapter will present a general rate utility framework for slicing mm-wave backhaul links, encompassing all known types of service utilities, i.e. logarithmic, sigmoid, polynomial, and linear. Inter-flow fairness is thus regarded as *utility fairness*, where the notion of *utility* quantifies value of rate allocations as perceived by the applications and MIP.

A deep learning solution, namely DELMU (Li et al., 2018), will be proposed to tackle the complexity of optimising non-convex objective functions built upon arbitrary combinations of such utilities. Specifically, by employing a stack of convolutional blocks, DELMU can learn correlations between traffic demands and achievable optimal rate assignments. The inferences made by the neural network will be further regulated by a simple 'sanity check' routine, which guarantees both flow rate admissibility within the network's capacity region and minimum service levels. The proposed method can be trained within minutes, following which it computes rate allocations

that match those obtained with state-of-the-art global optimisation algorithms, yet orders of magnitude faster. This confirms the applicability of DELMU to highly dynamic traffic regimes and demonstrates up to 62% network utility gains over a baseline greedy approach.

4.1 System Model

Consider a backhaul network deployment with $\mathcal{B}$ STAs inter-connected via mm-wave links.[1] Each STA is equipped with a pair of transceivers, hence is able to transmit and receive simultaneously, while keeping the footprint small to suit dense deployment. STAs employ MIMO antenna arrays and attain a total capacity as formulated in Sec. 3.1. To meet carrier-grade requirements and ensure precise TX/RX beam coordination, the network operates with a time division multiple access (TDMA) scheme. Assume carefully planned deployments where STAs have a certain elevation, e.g. on lampposts, hence interference is minimal and blockage events occur rarely.

I focus on settings where the backhaul network is managed by a single MIP and is partitioned into I logical slices to decouple different services (e.g. as specified in (3GPP, 2017b)). $\mathcal{F}$ user flows traverse the network and are grouped by traffic type i corresponding to a specific slice, i.e. $\mathcal{F} = \cup_{i\in\{1,\dots,I\}}\mathcal{F}_i$. $f_{i,j}$ denotes a flow on slice i that traverses path P_j. We consider an application specific flow demand $d_{i,j}$, and a minimum service rate of $\delta_{i,j}$, where the assigned flow rate $r_{1,j}$ to $f_{i,j}$ should satisfy $\delta_{i,j} \leq r_{i,j} \leq d_{i,j}$. Same as proposed in Chapter 3, the incentives to employ $\delta_{i,j}$ is to guarantee the availability of the service and allocating rate no more than requested, i.e. $d_{i,j}$, is to avoid wastage on network resources. The MIP's goal is to adjust the flow rates according to corresponding demands, in order to maximise the overall utility of the backhaul network. Flow demands are defined by upper and lower bounds. Lower bounds guarantee minimum flow rates, so as to ensure service availability, whilst upper bounds eliminate network resources wastage. I assume a controller (e.g. 'network slice broker' (Samdanis et al., 2016)) has complete network knowledge, periodically collects measurements of flow demands from STAs, solves NUM instances, and distributes the flow rate configurations corresponding to the solutions obtained.

[1] Although this chapter primarily focus on mm-wave backhauls, due to their potential to support high-speed and low latency communications, the optimisation framework and deep learning solution this chapter presents are generally applicable to other technology.

4.2 Problem Formulation

The objective is to find the optimal end-to-end flow rates that maximise the utility of sliced multi-service mm-wave backhaul networks. This chapter first introduces a general network utility framework, based on which I formulate the NUM problem, showing that in general settings this is NP-hard.

4.2.1 Utility Framework

Recall that network utility refers to the value obtained from exploiting the network, which can be monetary, resource utilisation, or level of user satisfaction. For any flow f, consider four possible types of utility functions of flow rate r, depending on which slice $\mathcal{F}_i$ that flow belongs to. The utilities considered are parameterised by α_i and β_i, whose values have practical implications, such as the amount billed by the MIP for a service. Given an allocated rate r, I distinguish the following types of services that can be mapped onto slices, whose utilities I incorporate in the proposed framework:

1. Services for which the MIP aims to maximise solely the attainable **revenue**. Denoting $\mathcal{F}_1$ the set of flows in this class, their utility is formulated as a linear function (Ahuja et al., 1993):

$$U_{\text{lnr}}(r) = \alpha_1 r + \beta_1, \quad \forall f \in \mathcal{F}_1. \tag{4.1}$$

   We note that $U_{\text{lnr}}(r)$ is both concave and convex.

2. Flows $f \in \mathcal{F}_2$ generated by applications that require certain level of **quality of service**, e.g. video streaming, and whose corresponding utility is thus formulated as a sigmoid function (Yin et al., 2015):

$$U_{\text{sig}}(r) = \frac{1}{1 + e^{-\alpha_2 (r - \beta_2)}}, \quad \forall f \in \mathcal{F}_2. \tag{4.2}$$

   Observe that $U_{\text{sig}}(r)$ is convex in $[0, \beta_2)$ and concave in (β_2, ∞), therefore non-concave over the entire domain.

3. **Delay sensitive** flows, $f \in \mathcal{F}_3$, whose utility is modelled as a polynomial function (Fazel and Chiang, 2005):

$$U_{\text{ply}}(r) = \alpha_3 (r^{\beta_3}), \quad \forall f \in \mathcal{F}_3, \tag{4.3}$$

   where β_3 is in the range $(0, 1]$, for which the above expression is concave.

4. **Best-effort** traffic, $f \in \mathcal{F}_4$, that does not belong in any of the previous classes, and whose utility is commonly expressed through a logarithmic function (Kelly, 1997):

$$U_{\log}(r) = \log(\alpha_4 r + \beta_4), \quad \forall f \in \mathcal{F}_4. \tag{4.4}$$

It is easy to verify that $U_{\log}(r)$ is also concave.

The proposed general utility framework encompasses all the four types of traffic discussed above (which may be parametrised differently for distinct tenants), therefore the overall utility of the sliced backhaul network can be expressed as

$$\begin{aligned}\mathcal{U} := \sum_{f\in\mathcal{F}} U(r) = \sum_{f_1\in\mathcal{F}_1} U_{\text{lnr}}(r_1) + \sum_{f_2\in\mathcal{F}_2} U_{\text{sig}}(r_2) \\ + \sum_{f_3\in\mathcal{F}_3} U_{\text{ply}}(r_3) + \sum_{f_4\in\mathcal{F}_4} U_{\log}(r_4).\end{aligned} \tag{4.5}$$

Arbitrary combinations of both concave and non-concave utility functions may result in non-concave expressions $\mathcal{U}$, as exemplified in Fig. 4.1. This figure shows the total utility when combining 4 flows with different utility functions, two of them sigmoidal and two polynomial, each with different parameters. Let us assume the rates of each type of flow increase in tandem. Observe that even in a simple setting like this one, the network utility is highly non-concave and finding the optimal allocation that maximises it is non-trivial. I will next formalise this problem with practical mm-wave capacity constraints, following which I will discuss its complexity.

4.2.2 Network Utility Maximisation

Consider a set of flows that follow predefined paths, $P_j, j \in \{1,2,...,J\}$, to/from the local gateway, where the number of possible routes in the network is J. $f_{i,j}$ denotes a flow on slice i that traverses path P_j, which is allocated a rate $r_{i,j}$. By contract, $r_{i,j}$ shall satisfy $\delta_{i,j} \leq r_{i,j} \leq d_{i,j}$, where $\delta_{i,j}$ is the minimum rate that guarantees service availability, and $d_{i,j}$ is the upper bound beyond which the service quality cannot be improved. $d_{i,j}$ is no less than $\delta_{i,j}$ by default. Furthermore, each path P_j consists of a number of mm-wave links, and the link between STAs m and n is subject to a link capacity $c_{m,n}$. I use $\tau^{s}_{j,m} \in \{0,1\}, s \in \{Tx, Rx\}$, to indicate whether node m transmits or receives data of flows traversing path P_j. The total network utility in (4.5) can be rewritten as:

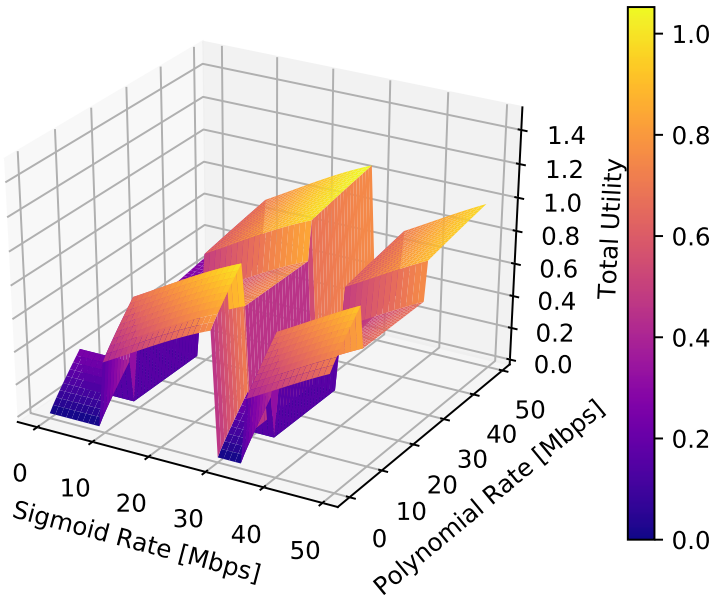


Figure 4.1: Total utility when combining four flows with different utility functions; namely, two have sigmoid utility of rate r, i.e. $U_{\mathbf{sig}}(r) = \frac{1}{1+e^{-0.08(r-15)}}$ and $U'_{\mathbf{sig}}(r) = \frac{1}{1+e^{-0.08(r-40)}}$, in the ranges of $[10-30]$ Mbps and respectively $[35-50]$ Mbps; the other two flows have polynomial utility functions, i.e. $U_{\mathbf{ply}}(r) = 0.04(r^{0.9})$ between $[0-10]$ Mbps, and $U'_{\mathbf{ply}}(r) = 0.03(r^{0.6})$ in $[30-50]$ Mbps. Rates increased in tandem for each type of flow.

$$\sum_{f \in \mathcal{F}} U(r) = \sum_{i=0}^{I} \sum_{j=0}^{J} U_i(r_{i,j}). \tag{4.6}$$

Finding the flow rate allocation vector $\mathbf{r_{i,j}}, \forall i, j$, that maximises this utility requires to periodically solve the following optimisation problem:

$$\max \sum_{i=0}^{I} \sum_{j=0}^{J} U_i(r_{i,j}) \tag{4.7}$$

$$\text{s.t. } \delta_{i,j} \leq r_{i,j} \leq d_{i,j}, \forall i, j; \tag{4.8}$$

$$\sum_{i=0}^{I} \sum_{j=0}^{J} \tau^{s}_{j,m} \frac{r_{i,j}}{c_{m,n}} \leq 1, \{m,n\} \in P_j, s \in \{Tx, Rx\}. \tag{4.9}$$

In the formulation above, (4.7) is the overall objective function and (4.8) specifies the demand constraints. Each STA can transmit and receive to/from one and only one STA simultaneously, and the total time allocated at a single node for all flow TX/RX should not exceed 1, which is captured by (4.9). Here $r_{i,j}/c_{m,n}$ denotes the time fraction allocated to flow $f_{i,j}$ on link $l_{m,n}$.

4.2.3 Complexity

In what follows I briefly show that the network utility optimisation problem formulated above, where the objective function is a linear combination of linear, sigmoid, polynomial, and logarithmic functions, is NP-hard. By Udell and Boyd (Udell and Boyd, 2013) any continuous function can be approximated arbitrarily well by a suitably large linear combination of sigmoidal functions (Udell and Boyd, 2013). Thus $\sum U(r)$ can be regarded as a sum of sigmoids and a larger number of other sigmoidal functions. Following the approach in (Udell and Boyd, 2013), one can reduce an integer program

$$\begin{aligned} &\text{find } \mathbf{r} \\ &\text{s.t. } A\mathbf{r} \leq Z;\ \mathbf{r} \in \{0,1\}^n, \end{aligned}$$

to an instance of a sigmoidal program

$$\begin{aligned} &\max \sum_i g(r_i) = \sum_i r_i(r_i - 1) \\ &\text{s.t. } A\mathbf{r} \leq Z;\ 0 \leq r_i \leq 1. \end{aligned}$$

Here $g(r_i)$ enforces a penalty on non-integral solutions, i.e. the solution to the sigmoidal program is 0 if and only if there exists an integral solution to $A\mathbf{r} = Z$. Since the integer program above is known to be NP-hard (Papadimitriou and Steiglitz, 1998), the reduced sigmoid program is also NP-hard, and therefore the NUM problem cast in (4.7)–(4.9) is also NP-hard.

4.3 The Deep Learning Approach

To tackle the complexity of the optimisation problem formulated in the previous section and compute solutions in a timely manner, I propose DELMU, a deep learning approach specifically designed for sliced mm-wave backhauls and also applicable to other technologies. In essence, this proposal learns correlations between traffic demands and allocated flow rates, to make inferences about optimal rate assignments. The results show that, with sufficient training data, the proposed deep neural network finds solutions close to those obtained by global search, while requiring substantially less runtime.

4.3.1 Convolutional Neural Network

I propose to use a Convolutional Neural Network (CNN) to imitate the behaviour of the optimal rate assignments. I employ a metaheuristic method named global search (GS), the optimality of which is proven in (Ugray et al., 2007), to produce ground-truth data to be used for training and testing of the proposed deep neural network. The GS method works by starting from multiple points within the feasible space and searching for local optima in their vicinity, then concluding on the global optimum from the set of local optima obtained (Ugray et al., 2007). Specifically, the GS algorithm generates a number of starting points using the scatter search algorithm (Glover, 1998), then eliminates those starting points that are not promising, judging by the corresponding value of the objective function and constraints. It then repeatedly executes a constrained nonlinear optimisation solver, i.e. `fmincon` by MATLAB®, to search for local maxima around the remaining start points. Eventually the largest of all local maxima is taken as the global maximum, if one exists. Note that simpler approximations such as semidefinite programming are constrained to convex optimisation problems, thus inappropriate for this task.

The CNN is trained by minimising the difference between ground-truth flow rates allocations (obtained with global search) and those inferred by the neural network. In general CNNs preform weight sharing across different feature channels (Goodfellow et al., 2016). This significantly reduces the number of model parameters as compared to traditional neural networks, while preserving remarkable performance. At the same time, the proposed approach aims to work well with a limited amount of training data, which makes CNNs particularly suitable for this problem. Therefore, I design a 12-layer CNN to infer the optimal flow rate and illustrate its structure in Fig. 4.2. The choice is motivated by recent results that confirm neural network architectures with 10 hidden layers, like the one employed in this work, can be trained relatively fast and perform excellent hierarchical feature extraction (Srivastava et al., 2015). Alternative neural network architectures including multi-layer perceptron (MLP) and CNN with various number of layers were considered. After empirically comparing the performance of these different network architectures, I found that the particular structure chosen achieves a good balance between inference accuracy and computation complexity.

The minimum and maximum traffic demand, and topology information are concatenated into a single vector, which will be subsequently fed to a sequence of con-

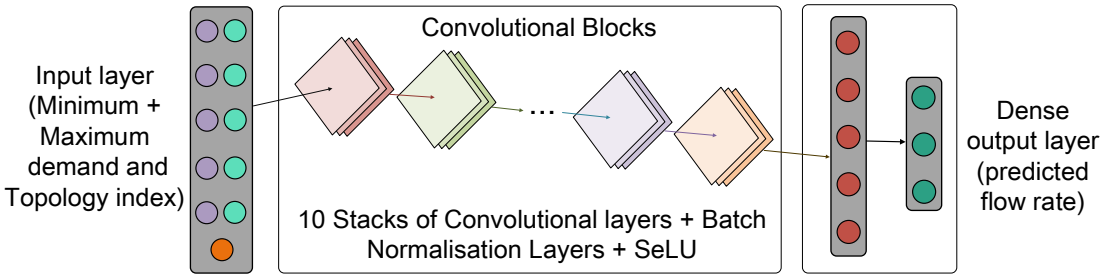


Figure 4.2: Proposed Convolutional Neural Network with 10 hidden layers, which takes traffic demand and topology index as input, and infers the optimal flow rate allocations.

volutional blocks. Each block consists of a one-dimensional convolutional layer and a Scaled Exponential Linear Unit (SELU) (Klambauer et al., 2017), which takes the following form:

$$\mathrm{SELU}(x) = \omega \begin{cases} x & x > 0 \\ \eta e^x - \eta & x \leqslant 0. \end{cases} \tag{4.10}$$

Here $\omega = 1.0507$ and $\eta = 1.6733$ by default. Employing SELU functions aims at improving the model representability, while enabling self-normalisation without requiring external techniques (e.g. batch normalisation). This enhances the robustness of the model and eventually yields faster convergence. Features of traffic demands are hierarchically extracted by convolutional blocks, and they are sent to fully-connected layers for inference. I train the CNN using a stochastic gradient descent (SGD) based method named Adam (Kingma and Ba, 2015), by minimising the following mean square error:

$$L_e = \frac{1}{Q \times I \times J} \sum_{q=0}^{Q} \sum_{i=0}^{I} \sum_{j=0}^{J} (r_{q,i,j} - r'_{q,i,j})^2. \tag{4.11}$$

Q denotes the number of training data points, $r_{q,i,j}$ denotes the allocated rate allocated to flow j on slice i, with demand instance q, as suggested by global search. $r'_{q,i,j}$ is the corresponding rate inferred by the neural network. The CNN is trained with 500 epochs, with an initial learning rate of 0.0001.

4.3.2 Post-Processing Algorithm

The output of the CNN on its own occasionally violates the constraints (4.8) and (4.9), because the model is only fed with traffic demands without embedding of constraints. I address this issue by designing a post-processing algorithm that adjusts the CNN solutions to fall within feasible domains, while maintaining minimum utility degradation

and very short computation times. The idea is to first decrease recursively with a large step-length the rate of flows that breach the constraints, then increase repeatedly with a smaller step-length the rate of flows that can achieve the largest utility gains.

Algorithm 3 CNN Post-Processing Algorithm

1: Compute the time between each pair of nodes $t^s_{m,n}$
2: Compute the utility of each flow $u_{i,j} = U_i(r_{i,j})$
3: **while** Any $t^s_{m,n} > 1$ **do**
4: Find the link $l_{m,n}$ with the maximum $t^s_{m,n}$
5: deStepLen $= \min\{10, r_{i,j} - \delta_{i,j}\}$
6: **for** Flows satisfying $\tau^s_{j,m} == 1$ or $\tau^s_{j,n} == 1$ for $l_{m,n}$ **do**
7: Potential utility decrease $u'_{i,j} = U_i(r_{i,j} - \text{deStepLen})$
8: **end for**
9: Find the $f_{i,j}$ with the minimum non-zero $\Delta u_{i,j} = u_{i,j} - u'_{i,j}$
10: Decrease rate of $f_{i,j}$, i.e. $r_{i,j} = r_{i,j} - \text{deStepLen}$
11: Update $t^s_{m,n}$ and $u_{i,j}$
12: **end while**
13: **while** Any flow rate can be increased **do**
14: inStepLen $= \min\{1, d_{i,j} - r_{i,j}\}$
15: Potential utility increase $u''_{i,j} = U_i(r_{i,j} + \text{inStepLen}), \forall f_{i,j}$
16: Find the $f_{i,j}$ with the maximum $\Delta u_{i,j} = u''_{i,j} - u_{i,j}$
17: Increase rate of $f_{i,j}$, i.e. $r_{i,j} = r_{i,j} + \text{inStepLen}$
18: Update $t^s_{m,n}$ and $u_{i,j}$
19: **end while**

Algorithm 3 shows the pseudo-code of this procedure. The routine starts by computing the total time on each link for all traversing flows, i.e. $t^s_{m,n} = \sum_i \sum_j \tau^s_{j,m} r_{i,j}/c_{m,n}$ (line 1) and the utility of each individual flow based on the rate allocation returned by CNN (line 2). Then it searches recursively for a flow to decrease (lines 3–12). At each step, Algorithm 3 selects the link with the highest total time (line 4) and reduces the rate of the flow traversing the link with minimum possible utility loss (lines 5–10). Then the total link time and the flow utilities are updated (line 11). The process (lines 4–11) is repeated until the time for all links comply with the time constraints. Next, I increase iteratively a flow that yields the maximum potential utility gain, while ensuring that all constraints are satisfied (lines 13–19). This is done by tentatively increasing each flow, with a step-length that complies with the demand constraint (line 14), computing the corresponding utility increment (line 15), then finding the flow with

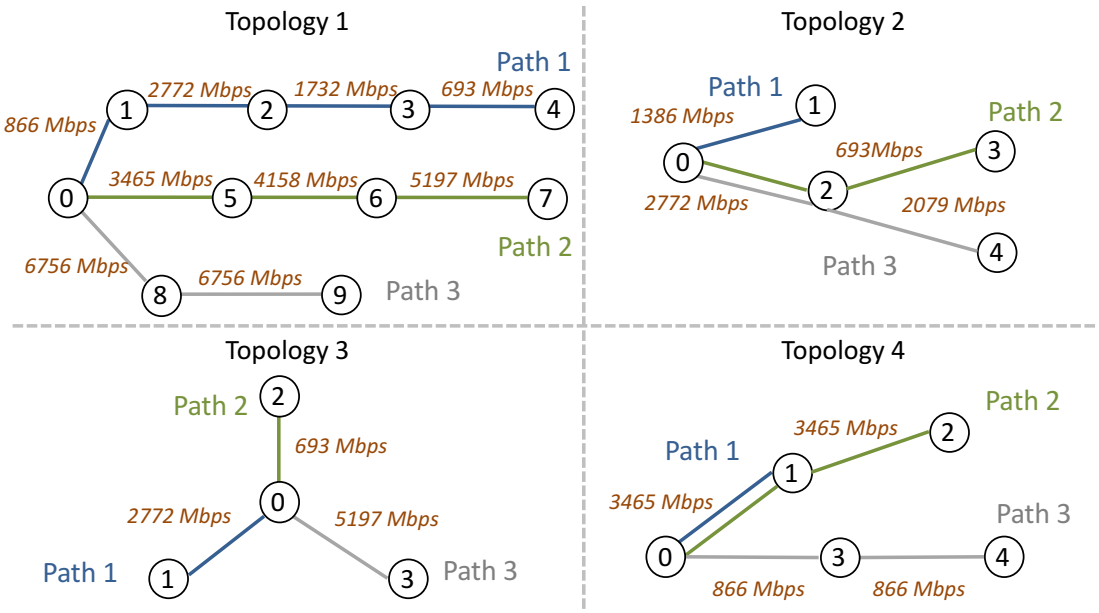


Figure 4.3: The four network topologies used for evaluation. Circles represent the STAs, flow paths are shown with lines of different colour, and link capacities are labelled.

maximum possible utility increase (line 16), and confirming the rate increment for that flow (line 17). Before the next round of increasing the rates, Algorithm 3 recomputes the total time on all links to verify that further rate increases are possible, and updates the utility of each flow (line 18).

4.4 Performance Evaluation

I evaluate the proposed DELMU solution, which encompasses the CNN structure and the post processing algorithm, on different backhaul topologies under a range of conditions. Specifically, I use four different topologies as illustrated in Fig. 4.3, where the number of STAs varies from 4 to 10, and link rates range from 693 Mbps to 6.8 Gbps. Each path carries up to four types of flows, i.e. with individual sigmoid, linear, polynomial, and logarithmic utilities. For each topology, 10,000 combinations of flow demands in the range $[0-750]$ Mbps in increments of 50 Mbps are randomly generated. The corresponding minimum service rates are generated uniformly at random in the range $[0-100]$ Mbps as integer values, and are capped by the maximum flow demand. The parameters shown in Table 4.1 are used to model utility.

To train and subsequently test the neural network, I run the GS algorithm on each of the 10,000 network settings described above. 80% of the results obtained are used to

Utility Type	Linear	Sigmoid	Polynomial	Logarithmic
α_i	0.00133	0.08000	0.03651	0.00229
β_i	0	350	0.5	1

Table 4.1: α_i and β_i parameters for the utility functions used in the evaluation.

construct a synthetic dataset that is used in the training process, which effectively seeks to minimise the mean square error expression defined in (4.11), by means of SGD. I use the remaining 20% of cases for as ground truth for testing the accuracy of the optimal rate allocation inferences that DELMU makes. More precisely, the performance of DELMU in terms of total network utility and computational time is compared against the solutions obtained with GS and those computed with a baseline greedy approach that is devised in this thesis. In this chapter, I will discuss both benchmarks in more detail in the following subsection.

To compute solutions with the GS and greedy algorithms, and make inferences with the proposed CNN, I use a workstation with an Intel Xeon E3-1271 CPU @ 3.60GHz and 16GB of RAM. The CNN is trained on a NVIDIA TITAN X GPU using the open-source Python libraries TensorFlow (Abadi et al., 2016) and TensorLayer (Dong et al., 2017). I implement the greedy solution in Python and employ the GS solver of MATLAB®.

4.4.1 Benchmark Solution

I engineer a baseline greedy algorithm for the purpose of evaluation, with the goal of finding reasonably good solutions *fast*. The greedy approach starts by setting all flow rates to the minimum demand and then recursively chooses a flow to increase its rate, with the aim of achieving maximum utility gain at the current step, as long as the constraints (4.8)–(4.9) are respected. A solution is found when there are no remaining flows whose rates can be further increased. For fair comparison, the greedy approach takes exactly the same flow demands and the corresponding minimum service rates as used by GS and DELMU. A step size of 1 Mbps is employed.

4.4.2 Total Utility

Let us first examine the overall utility performance of the proposed DELMU, in comparison with that of the greedy and the GS solutions. Fig. 4.4 illustrates the distributions

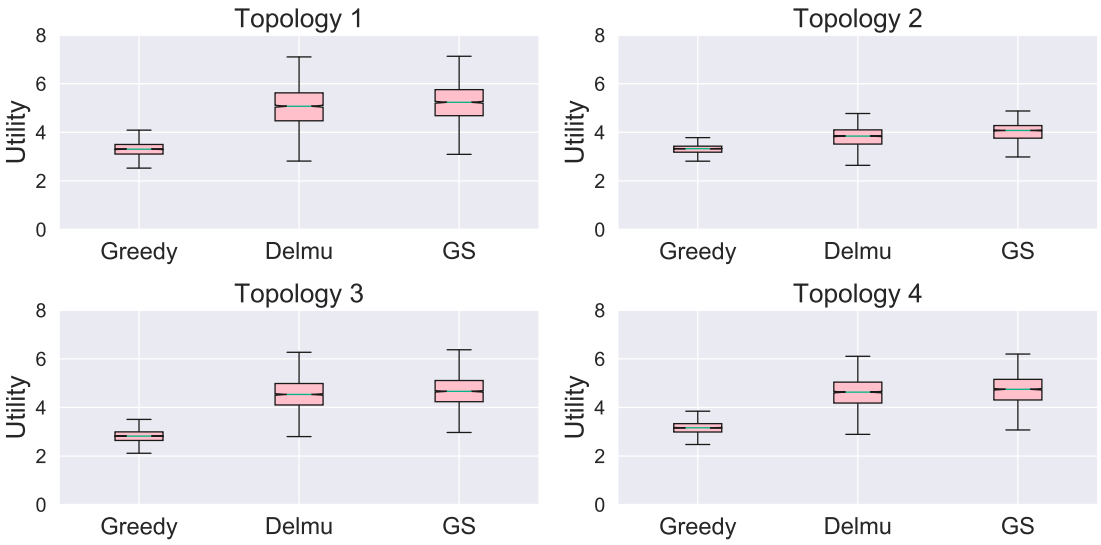


Figure 4.4: Distribution of the total utility attained by the proposed DELMU, and the benchmark GS and greedy algorithms, for the four topologies shown in Fig. 4.3. Numerical results.

of the total network utility for the 12 flows traversing the network, over the 2,000 instances tested. Observe that, among the 4 topologies used, the distribution of the total utility obtained by DELMU is almost the same as that of the optimal solution obtained with GS, as confirmed by the similar median values, the distance between the first and third quartiles, as well as the whiskers (minima and maxima). Specifically, the median values of the total utility attained by GS in Topologies 1–4 are 5.23, 4.07, 4.66, and 4.75, while those achieved by the proposed DELMU are 5.09, 3.88, 4.56, and 4.64. In sharp contrast to the DELMU's close-to-optimal performance, the greedy solution attains the medians of 3.30, 3.32, 2.81, and 3.16 utility units in the 4 topologies considered. Among these, for the case of Topology 3, DELMU obtains a 62% total utility gain over the greedy approach. It is also worth remarking that, although a greedy approach can perform within well-defined bounds from the optimum when working on submodular objective functions (Son et al., 2015), this is clearly suboptimal in the case of general utility functions as addressed herein.

4.4.3 Decomposing Performance Gains

To understand how DELMU achieves close-to-optimal utility, and why the benchmark greedy solution performs more poorly, this subsection examines one single instance for each topology, and dissects the utility values into the components corresponding to each type of traffic (i.e. slice). Fig. 4.5 illustrates the sum of utilities for each type of

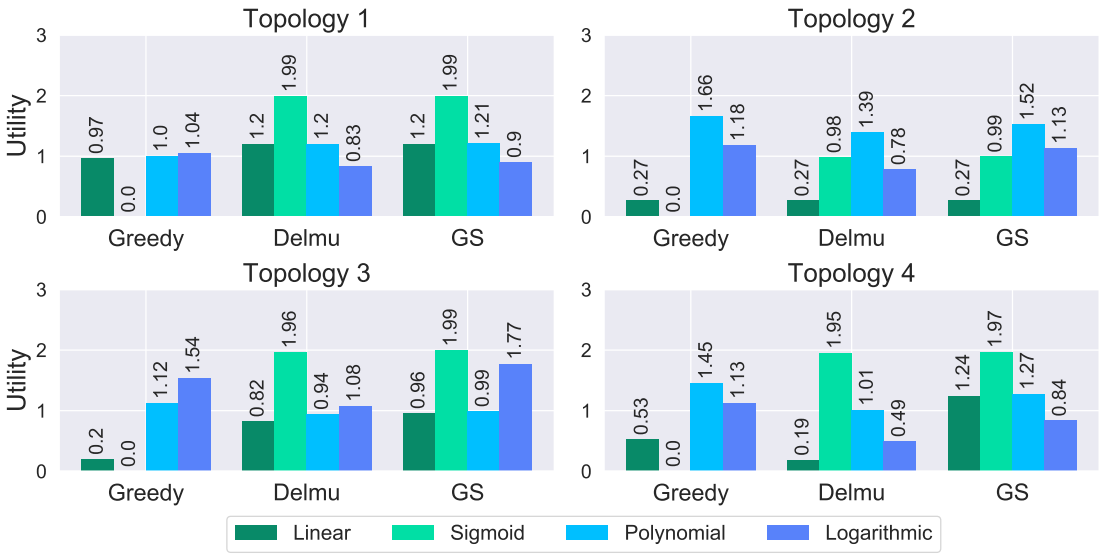


Figure 4.5: An example instance of the utility corresponding to each traffic type in each topology. Bars represents the sum utility of flows in the same slice. Numerical results.

traffic, attained with the greedy, CNN, and GS approaches. Note that the greedy solution tends to allocate more resources to traffic with logarithmic utility (in all topologies) and respectively polynomial utility (in Topologies 2, 3, and 4). In contrast, the CNN allocates higher rates to traffic subject to sigmoid utility in all the scenarios studied, which results in higher overall utility. This is because the greedy approach gives more resources to the flows that yield utility gains in the first steps of the algorithm's execution and fails to capture the inflection point of the traffic with sigmoid utility, which can contribute to a higher overall utility, under limited resource constraints. Furthermore, the allocations of rates to different traffic types by DELMU show close resemblance to the GS behaviour, which confirms the fact that DELMU achieves overall close to optimal utility allocations, at a lower computational cost, as will be shown next.

Let us delve deeper into the utility attained by each flow on each slice, along different paths, and in Fig. 4.6 compare the performance of the proposed approach and the benchmarks considered in the case of Topology 1. Flows corresponding to slices that have linear, sigmoid, polynomial, and respectively logarithmic utility are indexed from 1 to 4. Again, observe that the greedy approach assigns zero utility to traffic subject to sigmoid utility, in stark contrast with the GS method. While DELMU obtains the highest gains from traffic with linear and sigmoid utility on paths 2 and 3, greedy dedicates most of the network resources to traffic with logarithmic and exponential utility, without obtaining significantly more utility from these types of flows. DELMU

achieves accurate inference, as the performance is nearly the same with that of GS for all flows.

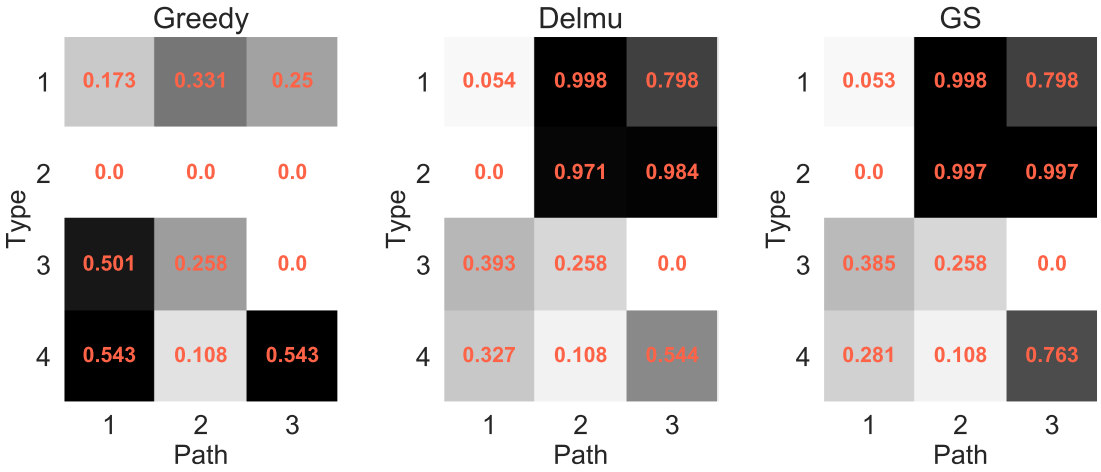


Figure 4.6: Utility of all data flows (on different slices and over different paths) attained by greedy, DELMU, and GS in one demand instance in Topology 1. In each subfigure, darker shades represent higher utility and the actual values are labelled. Numerical results.

4.4.4 Real-time Inference

To shed light on the runtime performance of the proposed DELMU solution, I first examine the average time required for inferring a single solution throughout the performance analysis presented in Section 4.4.2. I compare these computation times with those of the greedy and GS approaches over 2,000 instances and list the obtained results in Table 4.2. Note that the values for DELMU include the post-processing time.

Topology Index	1	2	3	4
GS	8.4339s	4.6075s	3.4492s	4.8311s
Greedy	0.1500s	0.1590s	0.1178s	0.1345s
DELMU	**0.0036s**	**0.0035s**	**0.0025s**	**0.0026s**

Table 4.2: Average computation time required to obtain a single solution to the NUM problem in Topologies 1–4 using GS, greedy, and the proposed CNN mechanism.

Observe that GS takes seconds to find a solution, while the greedy approach, although inferior in terms of utility performance, has runtimes in the order of hundreds of milliseconds for a single instance. In contrast, the CNN makes and adjusts inferences

within a few milliseconds. That is, as compared to the greedy algorithm, CNN generally requires two orders of magnitude smaller computation time. On the other hand, the GS algorithm, although working optimality, has three orders of magnitude higher runtimes as compared to DELMU. Lastly, note that the CNN inference itself requires $\sim$1.5ms per instance, and hence the post-processing dominates the overall execution time in the first two topologies. In terms of training time, the 4 topologies were trained together and the time taken was 16mins, making it 4mins per topology. This seems to be high as compared to the benchmarks does not require training, we argue that once the neural network is trained, it will be able to make rapid inference for a range of scenarios that observe similar input-output properties. This chapter concludes that the proposed DELMU is suitable for highly dynamic backhauls.

I complete this analysis by investigating the ability of the proposed DELMU solution to handle network dynamics in sliced mm-wave backhaul settings, including changes in traffic demand due to e.g. on/off behaviour of user applications and variations in capacity triggered e.g. by occasional blockage on the mm-wave links. Consider Topology 3 in Fig. 4.3, transporting a mix of flows with linear, polynomial, and logarithmic utility and different lifetimes, considering a 10 Mbps minimum level of service, in all cases when a flow is active. Precisely, in Fig. 4.7 I examine the time evolution of the throughput DELMU allocates to flows on each slice, according to a sequence of events. In particular, flows subject to sigmoid utility start with 0 Mbps demands, whilst all flows of the other types on all path have each an initial demand of 200 Mbps. After 100 ms, a flow with sigmoid utility on path 2 (i.e. $f_{2,2}$) becomes active, adding a 400 Mbps demand to the network. At time 200 ms, partial link blockage occurs on the link between STA 0 and STA 1, causing the corresponding capacity $c_{0,1}$ to drop from 2,772 Mbps to 693 Mbps. $f_{2,2}$ finishes 100 ms later.

Observe in the figure that the DELMU performs a correct allocation as soon as a change occurs and, given the millisecond scale inference times, the transition is almost instantaneous even at the 100 ms granularity. For instance, when $f_{2,2}$ joins, the allocation of network resources is immediately rearranged, so that the request of $f_{2,2}$ is mostly satisfied, whereas the rest of flows receive reduced rates. In this case, all the flows with linear utility are reduced to close to the minimum level of service, i.e. each to 11 Mbps rate. The drop in $c_{0,1}$ capacity at 200 ms leads to a significant degradation of the rates assigned to flows with polynomial and logarithmic utility, while the linear and sigmoid flows remain unaffected. Eventually, at 300 ms, when flow $f_{2,2}$ finishes, the rate of the flows with polynomial and logarithmic utility are increased, yet remain

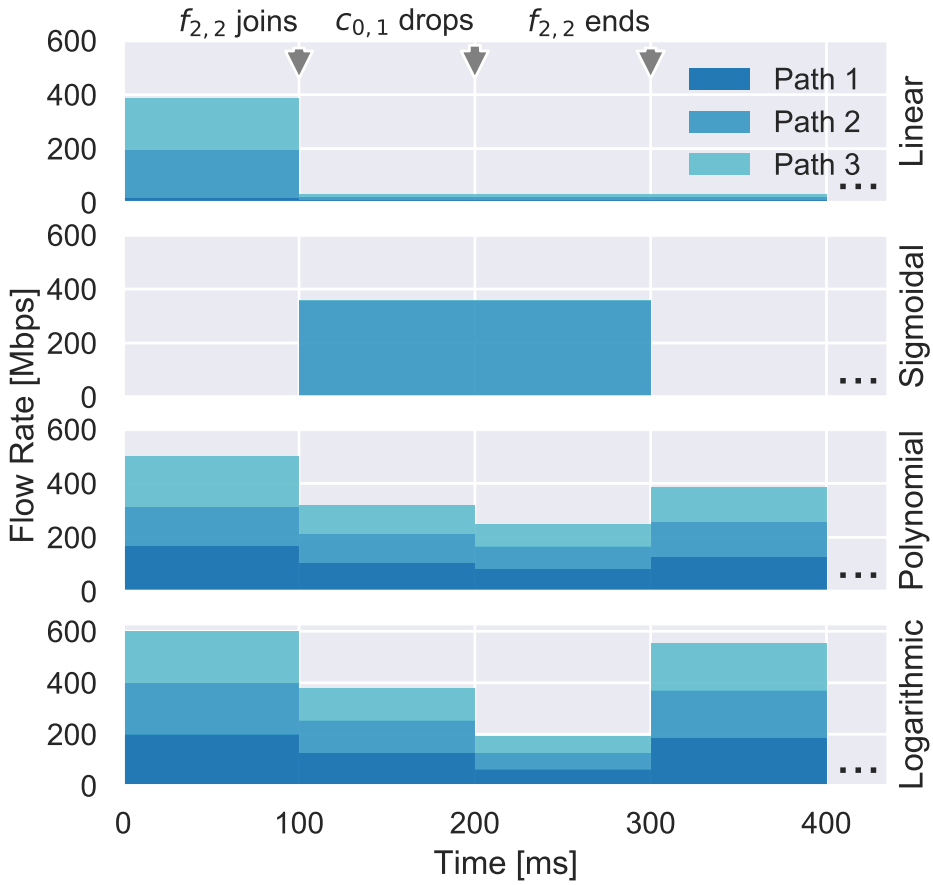


Figure 4.7: Rate allocations performed by DELMU for flows of different slices and paths over time in Topology 3 (see Fig. 4.3), as a sequence of demand and capacity changes occur as labeled at the top of the figure. Numerical results.

below the values assigned initially, due to the inferior $c_{0,1}$ capacity. Hence, the proposed DELMU is suitable for highly dynamic backhaul environments, as it makes close to optimal inferences fast and is able to adapt to sudden changes.

4.5 Discussion

The proposed deep learning based method, i.e. DELMU, achieves close-to-optimal flow rate allocations in virtually sliced mm-wave backhaul scenarios with achievable three orders of magnitude smaller computation time as compared to the optimal GS solution. It is worth noting that the proposed utility optimisation framework is generally applicable to different type of networks, but is particularly useful in mm-wave back-

haul networks due to the requirement of precise access scheduling and rate allocation as discussed in Chapter 3, in order to achieve carrier grade performance.

From a methodological perspective, one limitation of this study is the usage of simulation generated numerical data. This is, however, intrinsic to the school of supervised learning methods. The practical implications of such dependency on data are two-fold. On one hand, when significant changes occur in terms of topology, range of user demands and the dimension of the input (and output) of the neural network, the model should be retrained. Training of a new neural network, as discussed before, takes a couple of minutes in the topologies shown in the previous section. On the other hand, the performance of the neural network based approach is highly dependant on the quality of the ground-truth data obtained by the global optimisation solver. In cases where the global optimisation solver fails to find the global optimal, the neural network may settle on non-optimal solutions.

As a centralised approach, DELMU is subject to communication overhead including collection of flow demands and dissemination of allocated flow rates.

4.6 Summary

In this chapter of my PhD thesis, I tackled utility optimisation in sliced mm-wave networks by proposing a deep learning approach that learns correlations between traffic demands and optimal rate allocations. The proposed method specifically deals with scenarios where traffic is subject to conflicting requirements and maximise non-concave utility functions that reconcile all services, while overcoming the inherent complexity of the problems posed. Evaluation results demonstrated that the proposed convolutional neural network attains up to 62% utility gains over a greedy approach, infers close to optimal allocation solutions within orders of magnitude shorter runtimes as compared to global search, and responds quickly to network dynamics.

Chapter 5

Transport Protocols Performance in Next Generation Cellular Networks

TCP ensures reliable data exchange through error detection, packet reordering, and congestion control mechanisms, and hence today's internet data traffic is dominated by TCP (Huang et al., 2013). However, TCP's performance is known to be liable to variations in wireless networks, depending on how the congestion control reacts to the unpredictable radio link environment, how much impact the protocol overhead has, and to what level the retransmission scheme recovers packets. On the other hand, UDP experience less transmission overhead, but provides no delivery guarantees. It is therefore important to understand the throughput and delay experienced by users different channel conditions.

This chapter presents simulation based performance evaluation of the two widely used transport protocols, i.e. TCP and UDP, operating on top of LTE networks. I investigate key metrics that influence directly the user experience, such as the end-to-end throughput, under various channel conditions and LTE protocol settings. I also identify a number of performance issues that emerge when the current LTE channel is exposed to inferior channel quality. Specifically, for the user located at the cell edge, the SINR decreases significantly, hence the throughput and delay performance degrade for both UDP and TCP traffic. Although traffic running on top of UDP obtains marginally better throughput, it observes very high packet loss. Further, I will reveal that the transport protocols investigated are sensitive to LTE control plane errors. Enabling RLC acknowledged mode (AM) can mitigate partially the protocol data unit (PDU) loss, thereby improving TCP throughput remarkably at the cell edge. However, the acknowledgement scheme introduces additional overhead, thus affecting the throughput

and delay performance under good link conditions. Finally, this chapter concludes that in the presence of high error rates on control channels, robust modulation and coding schemes are needed. Alternatively, the RLC acknowledged mode can be employed to combat the packet loss, when TCP is used as transport protocol.

5.1 LTE-EPC Simulation Setup

In order to perform a practical analysis of the interaction between transport protocols and lower layers of the LTE stack, as well as an end-to-end full-stack evaluation, this chapter utilises the build-in LTE module of NS-3 simulator, namely LENA (CTTC, 2016). This section describes the simulation settings and reviews some of the design aspects in LENA.[1]

I examine the performance of UDP and TCP downlink data traffic from a remote server to a single UE. During each 50s simulation, the UE is assigned a 20m square box as an area of activity, and it moves in a random direction with 3kmph velocity, within this box. The simulations are run 14 times for each configuration, with the UE's area of activity placed at different distances to the associated eNodeB. The distance between the centre of UE's active area to the eNodeB ranges from 40m to 300m, and the distance between the centres of two adjacent boxes for two simulation runs is 20m. 4 groups of simulations are performed under the following different types of settings: RLC operates in unacknowledged mode (UM) with and without the existence of control frame errors, and respectively using AM with a control frame error model switched on and off. For each simulation scenario, both TCP and UDP are examined. Default simulation seeds are used for all simulations, hence when horizontally comparing the simulation runs at the same location of UE but different simulation settings, the channel environment, e.g. SINR, is the same.

The LTE-EPC network topology used is shown in Fig. 5.1. Specifically, the UE is connected to a single eNodeB, which has wired connections with the Service Gateway (SGW) and other eNodeBs. The SGW links to the remote server based on a high-speed point-to-point (P2P) connection of 10Gbps and this link introduces a delay of 10ms. In the LTE network, the eNodeBs are grouped in three-sector sites laid out on a hexagonal grid, as depicted in the Radio Environment Map (REM) in Fig. 5.2. In order to evaluate realistic interference scenarios, the cell site of interest is surrounded by 2 layers of three-sector sites, which generates interference on both data and control

[1]The simulation scripts used are available at https://github.com/ruihuili/TCP_LTE_NS3.git

channels. All nodes are assumed to be placed outdoor. Throughput and packet loss measurements are collected by the NS-3 Flow Monitor module at the IP layer. As for TCP congestion control protocol, by default NS-3 employs the New Reno version and disables Selective Acknowledgement (SACK). I record the RTT and CWND trace for all simulation runs with TCP traffic. Table 5.1 lists the overall configuration of the simulations.

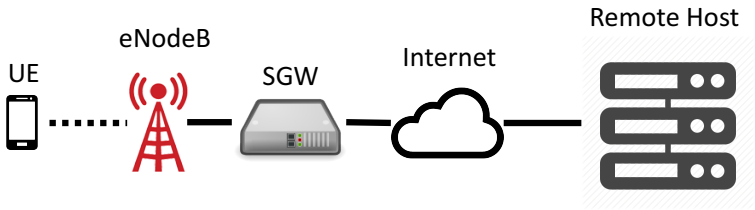


Figure 5.1: LTE-EPC network topology.

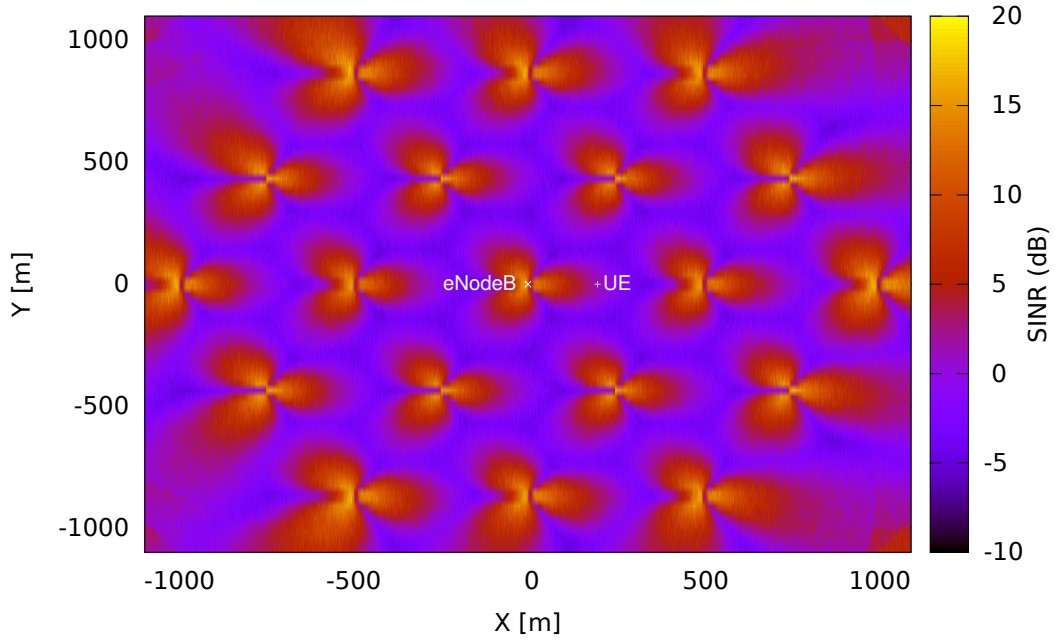


Figure 5.2: REM for LTE RAN control channel.

Regarding the propagation model, the ITU-R P1411 path loss model (ITU, 2015) is used throughout the experiments, together with a log-normal shadowing model. LENA takes fading trace computed offline using MATLAB. In the simulations, multi-path fading conditions follow the Extended Pedestrian A profile specified in Annex B.2 of the 3GPP standard TS 36.104 (3GPP, 2016d). The fading amplitude is a random process and fading values are computed based on the commonly used Rayleigh model, which is a function of both time and frequency.

Parameter Name	Value
Antenna type	Parabolic
Beamwidth	70°
Transmission power	46dBm
Site height	30m
Sector offset at each 3-sector site	0.5m
Inter-site distance	500m
UE height	1.5m
Carrier Frequency for downlink	2.1GHz
Carrier Frequency for uplink	1.9GHz
Bandwidth	50 RBs (10MHz)
Standard deviation of shadowing	$\sigma = 1$
Traffic Pattern	Backlogged
Packet size	1024bytes
TCP EPS bearer	QNGBR_VIDEO_TCP_DEFAULT
RLC transmission buffer size	1024Kbytes

Table 5.1: Simulation settings.

Frequency-division duplexing is implemented in LENA and the Transmission Time Interval (TTI) is 1ms. The reference signal power received every TTI is used to calculate the SINR, and Channel Quality Information (CQI) feedback is generated using the SINR obtained. Interference is modelled using a Gaussian model, according to which the overall interference power is calculated by summing up the power of all interfering signals. The adopted error model for both control and data planes is based on link-to-system mapping. Furthermore, Hybrid ARQ (HARQ) is utilised on data channels, employing a soft combining hybrid incremental redundancy scheme with multiple stop-and-wait, which means that the retransmissions contain only new information with respect to the previous transmissions. The HARQ model is integrated with the error model, and the retransmissions are arranged by the scheduler.

As per the TS 36.211 specification (3GPP, 2016a), downlink control frames, i.e. PCFICH and PDCCH, start at the beginning of each subframe and in total last no more than three symbols. The subframe structure in LENA is implemented accordingly (NS-3 Project, 2016), as shown in Fig. 5.3. PCFICH indicates the actual length of the control frame, and PDCCH mainly carries the DCI assigned by the MAC layer,

including information about the resources allocated for the UEs. Errors in the control channel thus result in the loss of corresponding Transport Blocks (TBs) transmitted in the TTI.

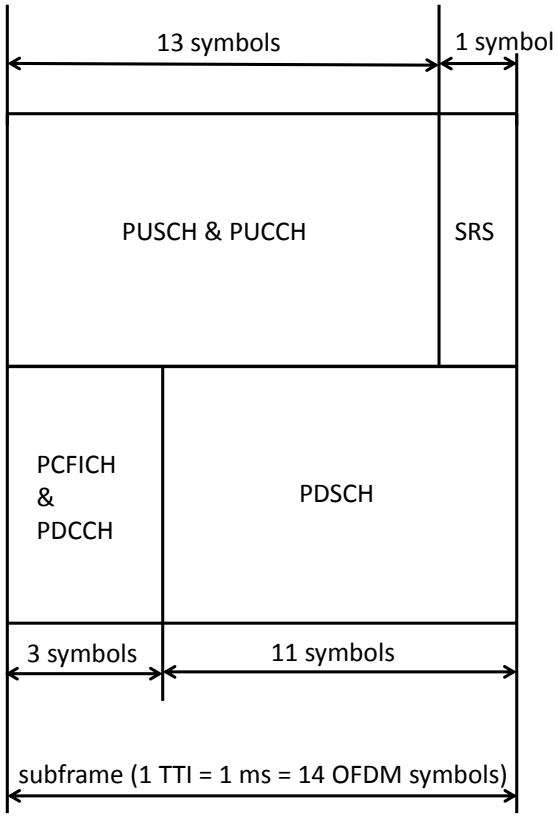


Figure 5.3: LTE subframe structure (NS-3 Project, 2016).

5.2 Simulation Results

I begin the simulation study by analysing the throughput and packet loss in the absence of errors on the control channel, when RLC runs in AM and UM, with UDP and respectively TCP traffic. Then I will investigate the throughput performance of TCP when exposed to control frame errors, contrasting with UDP throughput. Further, I compare the TCP throughput when LTE operates with RLC AM and UM. Finally, I will examine the Congestion Window (CWND) and Run Trip Tim (RTT) traces for TCP traffic in a number of representative scenarios.

5.2.1 Error Free Control Channels

Fig. 5.4 depicts the average throughput and packet loss at each location, for TCP and UDP when RLC operates with AM and UM settings and the control channels are error free. Observe that as the distance between the UE's active area and the eNodeB increases from 40m to 300m, both UDP and TCP traffic will see a significant reduction in throughput in both RLC modes. Specifically, UDP achieves at most approximately 0.5Mbps higher throughput than TCP in UM, and around 1Mbps higher throughput in AM. However, TCP manages to eliminate all packet losses through its retransmission scheme, whereas UDP suffers from severe packet loss with RLC UM. With the help of RLC ARQ, UDP packet loss is reduced approximately by half compared to the equivalent cases where RLC operates without ARQ.

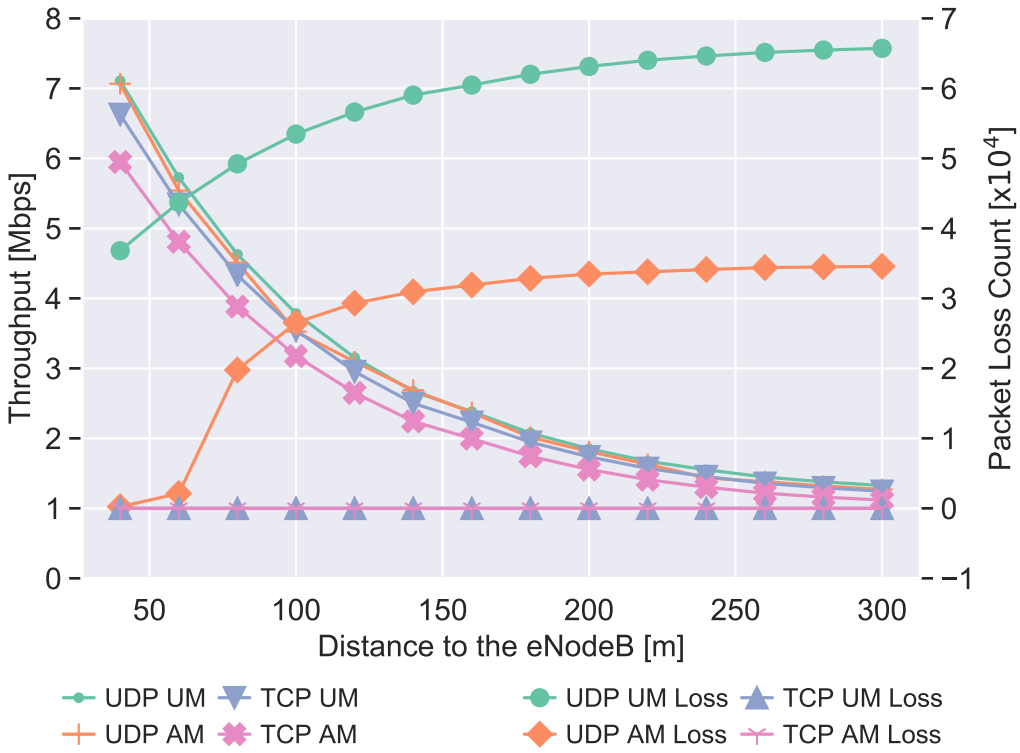


Figure 5.4: Average TCP and UDP throughput and packet loss versus UE–eNodeB distance, when RLC operates with AM and UM. Control channel is error free.

It is also worth noting that the scheduler takes into account only the amount of data in RLC PDUs and discards RLC headers when making scheduling decisions. Consequently, an RLC PDU may have been allocated a TB according to the amount of data it carries, but after adding RLC and MAC headers the information may no longer fit the TB size. In such cases RLC will perform segmentation and the PDU will require

an additional transmit opportunity. This results in a slight UDP throughput reduction, but has a greater impact on TCP.

From this set of results, I conclude that in the absence of control channel errors, despite achieving lower throughput than UDP, TCP is able to recover all packets lost in all cases. In contrast, UDP experiences significant packet loss, even though it shows a marginally higher throughput. In real life, such behaviour can hardly satisfy data-intensive applications such as video streaming. However, when RLC employs ARQ for retransmission, lost or out-of-order packets can be partially recovered for UDP. Hence the user experience can potentially be improved to some extent, without considerably compromising the throughput.

Nevertheless, there exists a maximum number of RLC PDU retransmission attempts. As per TS 36.322 and TS 36.331 (3GPP, 2016b,c), upon such events, RLC shall inform the upper layers to trigger a Radio Link Failure (RLF). Note that RLF is not currently implemented in NS-3, and RLC will simply stop forwarding any PDUs when the maximum retransmission threshold is reached. However, this means that in practice RLC may fail to recover all packets lost, and in extreme cases, it will rely on upper layers to recover these. In such cases, UDP will take packet loss for granted, whereas TCP will endeavour to recover as many lost packets as possible while performing congestion control.

5.2.2 Control Channels Prone to Errors

In the above studied cases, note that the congestion control of TCP is hardly constraining the transmission behaviour, thanks to the HARQ that provides timely correction of byte-wise errors, and TCP itself that detects and corrects packet losses efficiently enough without triggering slow-start frequently. In what follows, I investigate the case of error-prone control plane, where the control channel is assumed to be exposed to full interference from the surrounding cells, and consequently, PCFICH and PDCCH symbols become corrupt. Each occurrence of such errors in the control plane, as mentioned previously, will result in the loss of all the TBs carried in the TTI. It can be therefore anticipated that when UEs observe high packet loss rates, TCP may experience severe performance degradation. Switching to RLC AM, however, can presumably recover the packet loss to some extent. This is indeed confirmed by the simulation results shown in Fig. 5.5.

Similar to the previous observation of throughput-distance behaviour in error-free

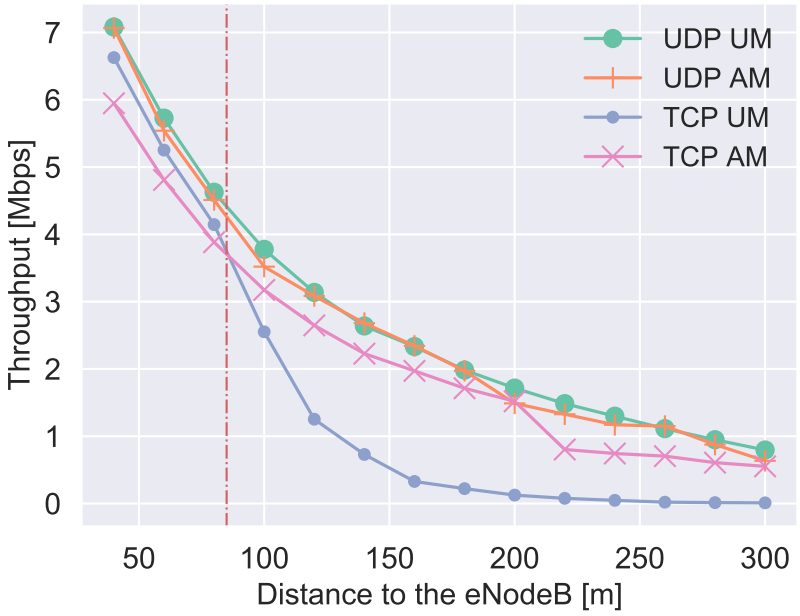


Figure 5.5: Average TCP throughput with RLC AM and UM as compared to UDP throughput with RLC UM, versus the distance to an eNodeB. Control channel error model switched on.

control channel cases, the throughput of an UE decreases dramatically when this moves away from the cell centre and control symbols encounter errors. Fig. 5.5 also suggests that the degradation is more significant for TCP. This is because that when the link quality is poor, the TCP is aware of the high packet loss, therefore it reduces the CWND and carries out retransmissions to recover the lost packets. Such retransmissions will degrade the throughput in UM, particularly at the cell edge scenarios.

Further, when the RLC AM is enabled, a number of packet lost due to control symbol errors can be actively recovered. Then fewer duplicate acknowledgements (DupAcks) will be received by the TCP state machine at the transmitter's side, hence congestion as seen by TCP is perceived as less serious and larger CWNDs allow more packet to be transmitted. Consequently, the TCP throughput is improved, up to a level that is just marginally lower than the typical UDP throughput.

On the other hand, when the UE is located near the eNodeB, e.g. at the distance of 40 to 80m, both the data and control planes observe less interference and path-loss (i.e. better SINR). Thus the impact on packet loss is much less significant than at the cell edge, and subsequently RLC AM commits fewer PDU retransmissions. The PDU segmentation induced by AM, however, is the dominate cause of throughput decrease, as can be seen in Fig. 5.5. Specifically, when the UE-eNodeB distance is below approximately 80m, TCP with AM observes a moderate throughput reduction of 0.5Mbps

as compared with TCP with UM, which is in line with the observation made from Fig. 5.4. A potential approach to leverage RLC AM in TCP performance is to switch between the AM and UM modes based on channel conditions. One simple way to implement this is to automatically made such switch based on user location. Focusing to this scenario, as shown by the vertical dotted line at around 80m from the cell centre, if the user is located below this distance, RLC UM performs well enough. When the UE-eNodeB distance is greater than 80m, employing the AM mode would benefit throughput performance. The generalisation of this rule remains to be investigated further.

5.2.3 TCP Congestion Window Behaviour

Next, I study the CWND size in detail, and compare CWND values under different scenarios as before, i.e. with both RLC AM and UM in the presence of control channel error and not, at different locations. Fig. 5.6 shows the median value of the CWND size over the simulation run time. As can be observed in the figure, in the absence of control channel errors, the median CWNDs is relatively higher as compared to when the control channel is susceptible to errors. This is because in the absence of frequent and consecutive packet loss, the CWND will be increased every RTT. When moving away from the cell centre, the UE may observe an increase in RTT and therefore the CWND increase is slower. The median CWND value is thus lower.

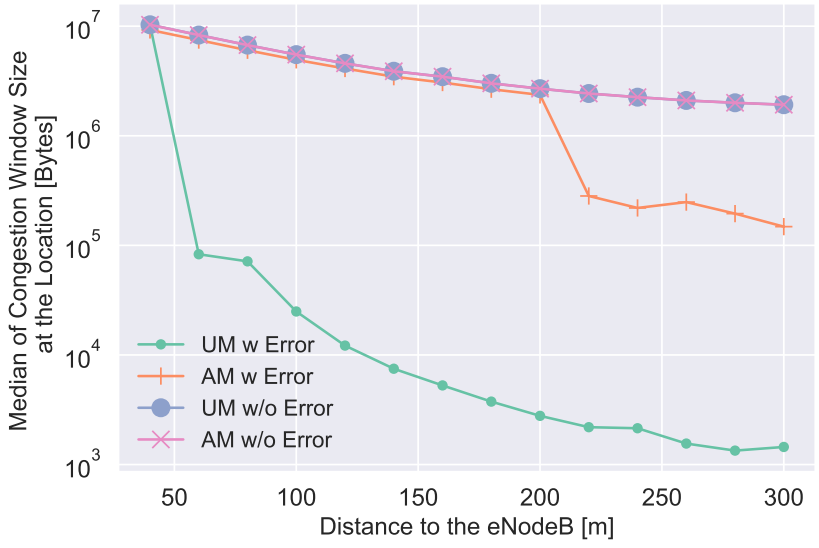


Figure 5.6: Median CWND when the UE located at different distances from the eNodeB.

When control channel errors exist but the distance between UE and eNodeB is

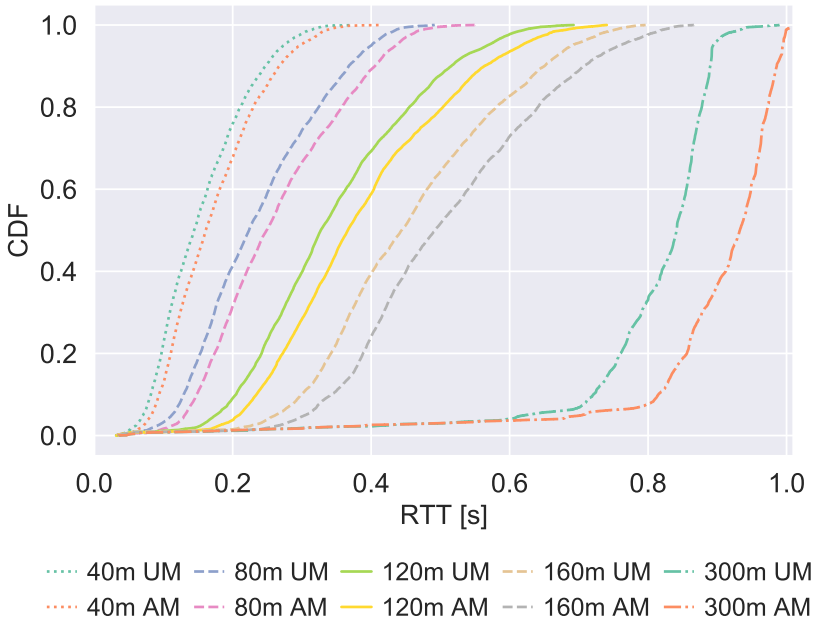


Figure 5.7: CDF of the RTT at different locations without control plane errors under RLC AM and UM.

small, and consequently the packet loss rate is also lower, the CWND is maintained relatively high. This suggests that under good link quality, congestion control is hardly constraining the transmission of packets. Conversely, when control symbols become corrupt and RLC does not employ AM, the CWND stays low, which will result in a very limited number of transmissions allowed by the TCP congestion control algorithm. This likely contributes to the TCP throughput degradation observed when UM is used, and explains why the UDP is much higher. Activating ARQ at the RLC layer, the CWND is consistently larger and more TCP segments are allowed to be transmitted.

The time required for a complete TCP segment transmission, i.e. from the point when the segment leaves the transmitter until the successful reception of an ACK, is defined as RTT. Fig. 5.7 shows the CDF of the RTT when either AM or UM employed and the UE–eNodeB distance is respectively $\{40, 80, 120, 160, 300\}$m. Unsurprisingly, for both AM and UM cases, the larger the UE–eNodeB distance, the poorer the link quality becomes and the longer the RTT tends to be. This behaviour is likely because 1) more retransmissions take place, and hence a packet occupies more TTIs; and 2) queues build up at lower layers and thus PDUs wait longer to be transmitted. When compared with UM at the same locations, AM observes marginally higher RTT overall. This can be explained by the issue of unwanted segmentation taking place at the RLC layer, as already mentioned in the above paragraphs. Moreover, the gap between

median values attained with the UM and AM schemes at the same location increases as the UE–eNodeB distance grows. Concretely, these gaps are {0.02, 0.03, 0.04, 0.05, 0.09} respectively for the distance set sampled, i.e. {40, 80, 120, 160, 300}m. This is presumably because RLC has to retransmit more frequently the PDUs as the channel quality worsens, therefore delaying the successful delivery of a TCP segment.

Considering the throughput performance discussed previously, I argue that cell edge users will experience both increased delay and lower throughput, which could potentially lead to degraded user experience.

5.3 Discussion

This study aimed to provide insights into the interaction between classic transport protocols and the LTE stack, which is currently widely employed in mobile networks. The main limitations of my work are that the analysis is purely based on simulations and the study confined to only one version of TCP congestion control paradigm. As potential future directions, the findings presented in this chapter could be enhanced with an experimental study could unravel further practical implications. On the other hand, it is also important to gain better understanding of any reciprocal influences between newer transport protocol versions (e.g. BBR, QUIC, and newer TCP congestion control algorithms), and emerging wireless technologies, e.g. mm-wave access and backhauling. Note however that, different flavours of TCP may bring marginal improvements in TCP performance, but the overall findings should still hold. In terms of TCP throughput performance, this should be upper-bounded by that of UDP, while the packet loss rate incurred by TCP should remain small, as shown in Fig. 5.4. Moreover, the conclusion drawn from the comparison between RLC AM/UM mode should intuitively hold in better performing TCP versions even though the significance may be subject to change.

Following my work, Polese *et al.* have provide early results of cell-edge performance in mm-wave frequency bands, and the authors observe similar TCP throughput degradation, revealing a reduction when MAC HARQ is disabled (Polese et al., 2017). This study further leverages different radio access interfaces, including LTE deployed in legacy frequencies and mm-wave bands, to support multi-path TCP connections. Experiments show that multi-path TCP flows traversing both legacy frequency and mm-wave links, when compared to single-path or multi-path TCP flows traversing

mm-wave only links, attain higher throughput and lower latency at mid-cell (100m) and cell-edge (150m) distances. The authors therefore conclude that the presence of a secondary reliable LTE path improves the overall throughput of the connection.

In late 2016, Google developed a new congestion control algorithm named BBR (Cardwell et al., 2016). BBR is a model-based congestion control mechanism that maintains an explicit model of the network using the recent maximum bandwidth and round-trip time experienced by the outbound packets. Unlike loss-based congestion control methods, BBR suits better the contemporary network interface controllers with Gbps links. Although BBR achieves higher throughput and lower latency, it is nonetheless criticised for being unfair to other flows and causing considerable packet loss (Hock et al., 2017). Moreover, Google QUIC (Hamilton et al., 2016) exploits multiplexed connections between two UDP endpoints, in order to reduce the latency and avoid additional data loss experienced by TCP in lossy channel conditions. However, as working on top of UDP, it does not provide delivery guarantees, as TCP does. It is therefore less competitive when compared with TCP, in terms of loss rate and buffering, especially in large propagation delay scenarios (Yu et al., 2017).

Therefore, for future research, the question of what the best transport protocol for the ever changing Internet ecosystem remains open (Huston, 2018).

5.4 Summary

This chapter investigated the impact of RLC AM/UM configurations and control channel error induced in LTE networks that influence the performance of TCP and UDP. One can conclude that as the distance between UE and eNodeB increases, the performance of both transport protocols decreases significantly. When the LTE control plane is susceptible to errors at low SINR, i.e. cell edge, further degradation in the performance of TCP and UDP was observed. On the other hand, data packet loss is currently relying on data plane ARQ schemes for recovery. Consequently, the RLC AM overcomes the performance bottleneck introduced by control symbol errors up to a certain level, this can be beneficial for TCP at the cell edge. Therefore, robust modulation and coding schemes are needed to mitigate errors in the control channels of LTE networks, in order to address the performance issue of the transport protocols studied at the cell edge. Future work can be carried out to improve the transport protocol design and enhance user experience, for instance by enabling cross-layer cooperation – one possible approach is to switch the RLC AM/UM mode by taking into account the channel

conditions, the transport protocol employed, and the transport protocol attributes.

Chapter 6

Mobility Optimisation of Network-on-Drones

In emergency services post floods, earthquakes, or nuclear plant disasters, mobile base stations mounted on UAVs provide viable wireless coverage solutions in challenging landscapes and conditions, where cellular/WiFi infrastructure is unavailable and access is too difficult for human drivers of current cell-on-wheels solutions (Cellsite Solutions LLC, 2017), or carriers of wearable base stations (Air Lynx, 2018). Flying at lower altitudes and supported by advanced hardware and software platforms, UAVs have received increasing interest from the telecom industry, as potential bearers of aerial base stations in temporary cellular deployments (FAA, 2017).

Operating multiple such airborne base stations, to ensure reliable user connectivity, demands intelligent control of UAV movements, as poor signal strength and user outage can be catastrophic to mission critical scenarios. Recent advances in deep reinforcement learning (DRL) methods have led to a number of successful applications, e.g. AlphaGo (Silver et al., 2017), and successful research outcomes of applying DRL to various domains that are subject to large action and state spaces, which otherwise would be difficult to solve (Schmidhuber, 2015). Towards this end, this chapter investigates a DRL based solution to tackle the challenges of base stations mobility control. Asynchronous Advantage Actor-Critic (A3C) algorithm is employed, and I design a custom reward function, which incorporates SINR and outage events information, and seeks to provide mobile user coverage with the highest possible signal quality. Preliminary results reveal that the DRL solution converges after 4×10^5 steps of training, which corresponds to approximately 2.5 hour in real life on a desktop machine, after which it outperforms a benchmark gradient-based alternative, as the proposed solution

attains 5dB higher median SINR during an entire test mission of 10,000 steps.

6.1 System Model

We consider a deployment of $\mathcal{B}$ airSTAs providing broadband wireless access to a handful of users in emergency network settings. The system complies with the LTE isolated E-UTRAN specified in (3GPP, 2015) with simplified LTE functionality, e.g. disabled MME authentication, to simplify the overall architecture and prolong network lifetime. Each airSTA serves a number of UEs and is connected wirelessly (via satellite or μ-/mm-wave links) to a central controller. The controller hosts a DRL agent that learns to make optimal decisions about the airSTAs mobility control. We assume the backhaul communication between the airSTAs is operated on an orthogonal channel from the access links, and airSTAs do not communicate directly with each other.

Wireless Channel: Consider airSTAs share the same frequency band, i.e. reuse factor 1. I focus on the downlink communication, assuming the SINR is directly related to the quality of service received by UEs. The transmit power employed by airSTA b to a user is P_b and is denoted by $G_{b,u}$ the channel gain between airSTA b and user u, which is a linear combination of the free-space path-loss $l_{b,u}$, shadow fading, and antenna gain G_a. The log-distance path-loss $l_{b,u}$ can be computed following the 3GPP model for urban cellular scenarios with standard coefficients α and β, i.e. $l_{b,u} = \alpha + \beta \log(D_{b,u})$, where $D_{b,u}$ is the Euclidean distance between b and u. Given $I_b \subset \mathrm{B} \backslash b$ the set of airSTAs that interfere with b and N_0, the power of per-channel additive white noise, the SINR observed by UE u is:

$$\mathrm{SINR}_{b,u} = \frac{P_b G_{b,u}}{N_0 + \sum_{b' \in I_b} P_{b'} G_{b',u}} \tag{6.1}$$

UE Mobility and LTE Handovers: Assume a reference group mobility model, by which users cluster around group centres that move along random way points (Hong et al., 1999). The motivation for employing this mobility model is that, in the envisioned emergency scenario, rescue and medical teams, or fire fighters, rush towards a target scene while the population may be moving away from that location. Further, as both UEs and airSTAs continuously change their position, standard LTE S1 based handover policy is employed including hysteresis and time-to-trigger to avoid ping-pong effects. Specifically, when the received signal strength is below a certain threshold $SINR_{th}$ for a duration of $t_{trigger}$, the user can be handed off to adjacent cells, if a new airSTA provides an SINR higher by σ than the value currently measured.

6.2 Problem Formulation and the Deep Reinforcement Learning Method

This chapter address mobility control of airborne airSTAs in emergency settings, considering stochastic wireless channels and user mobility, as modelled in Sec. 6.1. While this is a challenging problem subject to large state dimension including users location and user and airSTA association as well as exponentially increasing action space, deep reinforcement learning (DRL) has achieved promising results in similarly complex tasks, such as Atari game play (Mnih et al., 2015) and adaptive video streaming (Mao et al., 2017). This motivates the application of a DRL method, formulating the mobility control task as Markov Decision Process (MDP), and designing an A3C based solution tailored to the target networking scenario. Alternative to the proposed DRL approach, optimal control or exhaustive search seem to be plausible solutions. However, optimal control methods require precise models of the environment, which are not always easy to obtain in the emergency scenarios considered. An exhaustive search method, as the results presented will show, gain marginal performance improvements at the cost of significantly longer computation times.

Markov Decision Process (MDP): The airSTAs mobility control can be modelled as a 5-tuple MDP, $< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma >$, where $\mathcal{S}$ is the state of the environment including airSTAs locations, $\mathcal{L}_{\mathcal{B}}(t)$, UEs locations, $\mathcal{L}_{\mathcal{U}}(t)$, and their associations, $\alpha_{\mathcal{B},\mathcal{U}}(t)$. $\mathcal{A}$ is the action taken by each agent, i.e. the movement direction of each airSTA, $\mathcal{M}_b(t)$, and $\mathcal{P}$ is the state transition matrix. Precisely, the system moves from state s to s' following action a according to $\mathcal{P}^a_{ss'} = \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = a]$. $\mathcal{R}$ is the reward function, which quantifies the system performance following an action, i.e. $\mathcal{R}^a_s = \mathbb{E}[R_{t+1} |$
$S_t = s, A_t = a]$. In the problem, this will depend on the SINR experienced by the UEs, as I detail next. $\gamma \in [0,1]$ is the discount factor, which dictates the importance of future rewards.

A policy π is the probability distribution of taking an action a in a given state s, i.e. $\pi(a|s) = \mathbb{P}[A_t = a | S_t = s]$. To attain an optimal policy $\pi*$, I employ a DRL method and give an overview of the learning procedure in Fig. 6.1. As shown in the figure, the agent updates a policy π. By this, at time step t, given state s_t including UEs and airSTAs locations and their associations, the agent takes an action a_t, according to which each drone $b \in \mathcal{B}$ moves from (x_b, y_b, z_b) to (x'_b, y'_b, z_b). The agent thereby receives a reward r_t, and the system enters a new state s_{t+1}. This process is repeated until the episodic return, i.e. the sum of future discounted rewards, converges. This

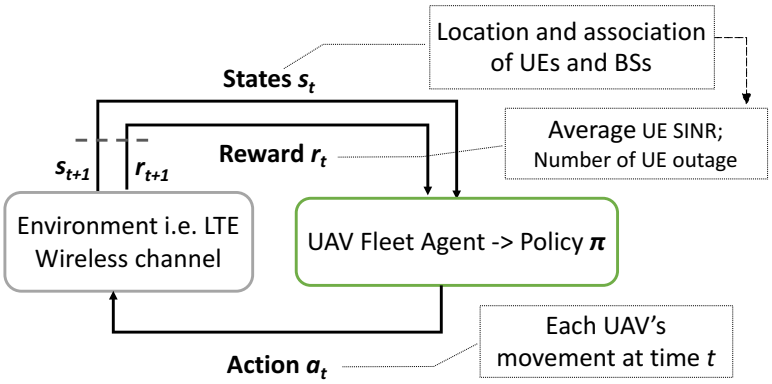


Figure 6.1: Overview of the learning loop for the proposed intelligent UAV mobility management agent.

indicates that an optimal policy was found.

There exist two main approaches to solving control problems through reinforcement learning (RL): policy-gradient RL methods learn by 'trial-and-error' a policy, i.e. the probability distribution of actions to take given the state of the environment; value-based methods learn to estimate the value for each action. Combining these two approaches, an actor-critic RL agent employs an 'actor' that performs actions to improve the policy, while a 'critic' makes judgements on the actor's performance and learns to estimate the action values.

To tackle the UAV mobility control problem, this thesis proposes a custom asynchronous advantage actor-critic (A3C) algorithm. A3C is a state-of-the-art actor-critic method that exploits multi-threading to create several learning agents, each exploring the state space in their own environment and updating periodically a global neural network with learned knowledge (Mnih et al., 2016). The key advantages of this approach are that it can be trained on a CPU, it de-correlates past experiences gained by each learning agent, and it converges rapidly. Such paralleled learning is a viable alternative to experience replay (i.e. Deep Q-Learning), as it removes large memory requirements.

The proposed A3C-based solution employs two neural networks that have the most simple deep learning architecture, i.e. the multi-layer perceptron (MLP). MLPs have fully connected layers and the neurons in the hidden layers implement non-linear activation functions of the output of the previous layer. The parameters (weights and biases) of these functions are obtained by training through back-propagation (LeCun et al., 2015). In this case, one of the MLPs acts as the actor and the other as the critic. Both of them employ 2 hidden layers, each consisting of 200 neurons.

I design a reward function that captures the specifics of mobile users served by

airborne airSTAs, aiming to ensure best connectivity, i.e. highest SINR and lowest outage likelihood. Hence, the proposed reward function is

$$R = \theta * \overline{\text{SINR}} - \frac{N_{out}}{N_{UE}}, \tag{6.2}$$

where $\overline{\text{SINR}}$ is the mean SINR computed across all the N_{UE} users, θ denotes a normalising factor, and N_{out} is the number of UEs whose SINR is below a minimum service requirement.

6.3 Evaluation and Results

6.3.1 Simulation Setup

Consider a 100×100 grid area, with a grid cell width of 5m. Within this area, 40 users move in groups of 10 UEs, following the group reference model (Hong et al., 1999), and are provided with connectivity by airSTAs mounted on UAVs. For each user the simulator computes the SINR of the link to the serving airSTA, using the model described in Sec. 6.1. A sample SINR heatmap and UE locations are illustrated on the left in Fig. 6.2. The histogram on the right shows the distribution of the SINR experienced by users in this instance. The UAVs move at a fixed altitude, [1] i.e. 10m from the ground level. The movement of a airSTA at each time step is chosen between 4 candidate directions (i.e. N, S, W, E) towards adjacent points on the grid, or idling (no move).

The DRL model is trained with 10 random seeds, each time with 1,000 training episodes. A single episode lasts 2,000 time steps and the locations the UAVs are reset to the same coordinates at beginning of each episode. At each discrete time step, the mobility control agent performs actions based on the current learning policy, and chooses from $5^4 = 625$ possible actions to take (4 directions or movement plus idling, 4 UAVs). The simulation parameters used are summarised in Table 6.1.

The model is trained and tested on a 8-core desktop with Intel Xeon W-2125 CPU clocked at 4.00GHz, and Python library TensorFlow is employed to implement the neural networks (Abadi et al., 2016).

[1] It is expected that even with 3D mobility management, adjustments in transmit power and gain will lead to the similar ground signal coverage.

	Parameter	Value
Wireless Channel	BS Transmit power	20 dBm
	Antenna gain	2 dB
	Log normal shadowing	$\mathcal{N}(0, 2)$
	Gaussian noise	-121 dBm
	Handover time-to-trigger	3
	Handover threshold	1 dB
	Minimum SINR	-5 dBm
Learning	Learning rate	0.0001
	Discount factor	0.9
	Number of A3C workers	4
	Global update step	10
	Normalising factor θ	0.05

Table 6.1: Simulation parameters.

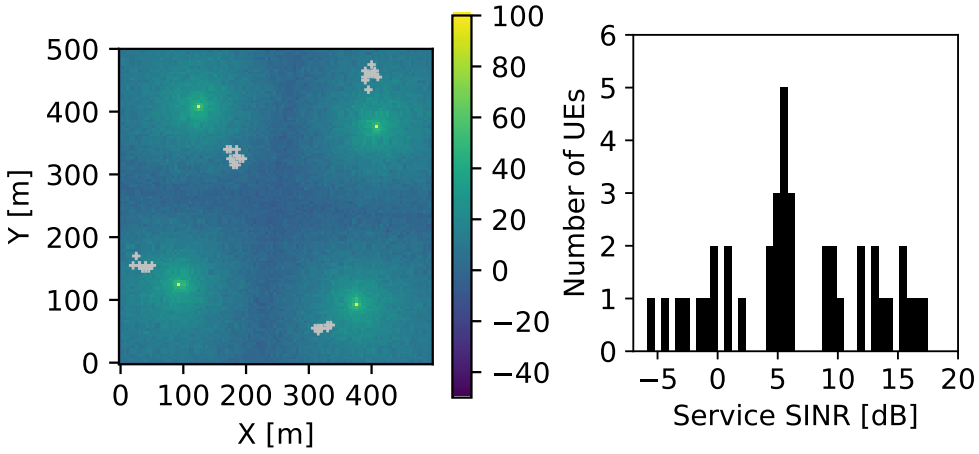


Figure 6.2: Left: SINR heatmap and UEs location (grey dots). Right: The corresponding distribution of UEs' experienced SINR.

Benchmark: To assess the performance of the proposed DRL solution, a benchmark SINR gradient based method is devised, and tested with the same network settings as with the proposed DRL approach. At each time step, this benchmark computes the average SINR at the associated UEs along each of the possible directions of movement. It then moves that UAV in the direction of the lowest average SINR. By this approach, the aim is to avoid outage while maintaining good signal quality for all the UEs served.

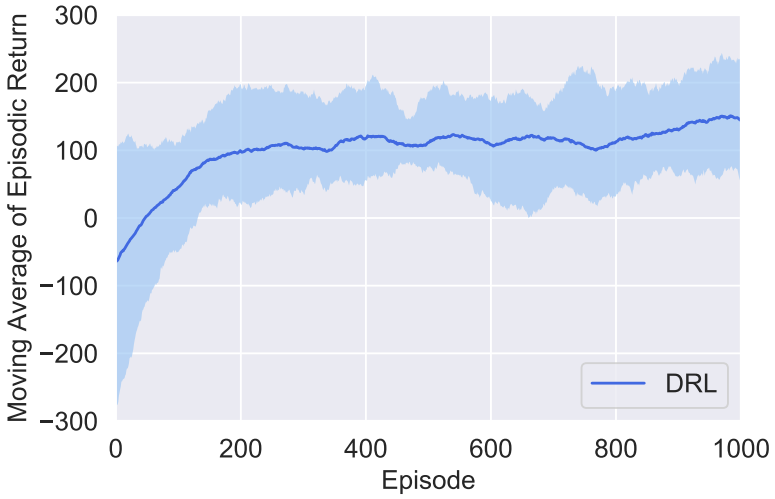


Figure 6.3: Moving average of the episodic return during 1,000 training episodes with different random seeds. Solid line represents the mean; shaded area represents the region between maximum and minimum return.

6.3.2 Simulation Results

Training Convergence: At the end of each episode, the moving average of the episodic return $\mathcal{R}$ is computed using the following formula:

$$\mathcal{R} = 0.01 * r_{ep} + 0.99\mathcal{R}, \tag{6.3}$$

where r_{ep} is the total reward of the episode. To examine the learning convergence, Fig. 6.3 plots the evolution of $\mathcal{R}$. The training converges within 200 episodes (i.e. $4 * 10^5$ steps), corresponding to approximately 2.5 hrs of training in real world. Such training can be performed once during pre-deployment stage, after which the agent can be used for multiple missions, given the wireless channel characteristics remain largely similar. Observe that during this phase the proposed A3C solution improves the average episodic return from around -50 to 100, with the minimum value being improved from -280 to 80. Once trained, the agent can be used directly to make decisions about airSTAs movement, according to the current conditions.

Performance: Ultimately, this chapter concerns quantifying the performance gains the DRL approach can attain over other solutions, such as the SINR gradient-based benchmark considered. To this end, the trained neural network model was trained over 10,000 steps, resetting the environment every 2,000 steps, and examine the signal quality provided by both approaches. Specifically, in Fig. 6.4 I plot the cumulative distribution (CDF) of the SINR experienced by all users with the benchmark scheme and the proposed DRL algorithm, studying also the impact of the number of training

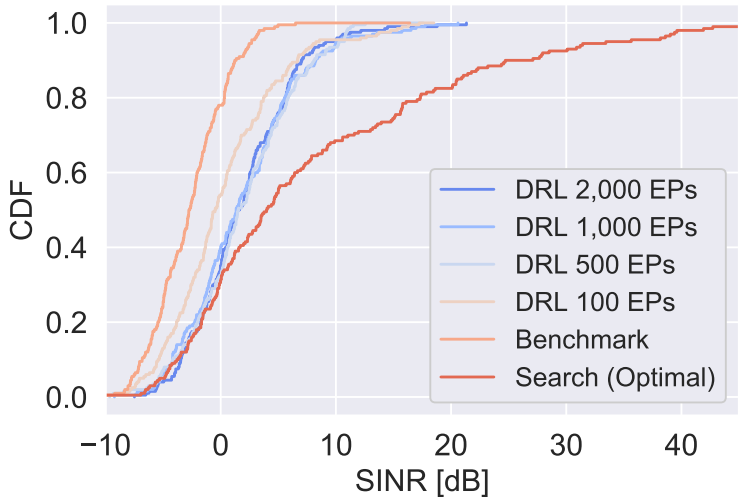


Figure 6.4: CDF of the SINR attained by all users with the proposed DRL method after 100, 500, 1,000, and 2,000 training episodes (EPs), the benchmark gradient method, and the optimal solution. Each scheme is tested over 10,000 testing steps.

episodes (i.e. 100, 500, 1,000, and 2,000 episodes) on the performance achieved. Moreover, in order to understand the performance difference between the proposed method and the best performing policy, Figure 6.4 includes the SINR distribution of such *optimal* solution. This is obtained by searching through all the possible moves and choosing the one leading to the highest SINR value.

The results confirm that the proposed DRL solution achieves a 5dB improvement of the median SINR, over the gradient-based benchmark. If considering -5dB as the signal outage level, the proposed DRL scheme only experiences approximately 5% user outage, which is one forth of that experienced by the benchmark gradient method, i.e. 20%. When compared with the optimal solution, the lower 40% of the SINR distribution achieved with the DRL approach overlaps with that of the optimal solution. The median SINR of the DRL solution is approximately 2.5dB lower than the optimal, but the computation time of the DRL approach during inference is on average only 0.02 of that required to find the optimum. The results further confirm that after convergence, the learning algorithm performs stably. Specifically, the distribution of the SINR after 2,000 training episodes achieves a median value that is only marginally better than that of an algorithm trained over 500 episodes. Let us conclude that the proposed DRL scheme for mobility control of airborne airSTAs attains more than 3$\times$ higher median SINR and 4$\times$ lower outage rate compared to the gradient-based benchmark. Furthermore, the proposed solution performs stably after training.

6.4 Discussion

The envisioned deployment of the proposed emergency networking solution relies on pre-deployment training of the mobility management agent, which is a lightweight MLP network that takes mobility control decisions for the airSTAs fleet at each time step, in a simulated environment. Specifically, the simulation environment provides a life-like signal propagation environment with channel metrics that can be gathered from similar emergency settings. The simulation environment should also observe close resemblance to the actual deployment of the airSTAs system, e.g. number of airSTAs, landscape of the deployment environment, which guarantees mission critical requirements to a certain degree. Once trained, the mobility management agent should be able to perform control of the fleet movement in practical deployments with a range of similar wireless channel propagation conditions and system settings. The model should undertake extensive and thorough testing before being deployed to guarantee security of operation.

In this Chapter, I focused on the wireless access part of the communications and the solution proposed is based on the assumptions of centralised control and the existence of efficient backhauling for the airSTAs fleet. In simplified scenarios airSTAs can be connected to a central controller via single-hop wireless channels that are orthogonal to the data channel. It is important to notice that the communication overhead of such centralised protocol would increase linearly as the size of the single-hop network grows. Although I envision such overhead can be accommodated in the PDCCH or PBCH (Physical Broadcast Channel), this remains a limitation of the current work.

As the network size grows, multi-hop backhauling should also be enabled to support inter-airSTA communications, so that airSTAs located closer to the gateway will relay packets for those are far away from the gateway. This, however, will further complicates the system model, as the backhaul topology and inter-airSTA link rates will introduce constraints on the capacity of individual airSTAs. One may consider to integrate the resource allocation mechanism proposed in Chapter 3, to achieve end-to-end flow throughput fairness, particularly when the link budget or hop-count encountered by data flows traversing the backhaul are heterogeneous. Another potential research direction is to take into consideration of power consumption hence limited life time of each airSTAs, by incorporating models of energy expenses in both transmission power and UAV movements.

6.5 Summary

This chapter introduced a deep reinforcement learning solution for the mobility control of multiple airborne base stations, with the goal of providing reliable connections to the users in the scenarios where wireless infrastructure is unavailable. This chapter took a A3C approach and designed a reward function that captures the specifics of such scenarios. By means of simulation experiments of this study reveals that the proposed DRL algorithm converges fast, and achieves 5dB higher median SINR and $4\times$ lower outage range, as compared to a gradient-based benchmark solution, achieving close-to-optimal performance while requiring only 0.02 of the total computation time required for searching the optimum.

Chapter 7

Conclusions and Future Directions

This thesis made several key contributions towards important resource optimisation problems in the domain of 5G mobile networks. Starting with mm-wave multi-hop backhauling, the thesis presented an airtime allocation and scheduling mechanism that can achieve max-min fair rate allocation among aggregate data traffic flows. Then, I presented an utility framework to describe the diverse performance requirements encountered by 5G, and detailed a rate allocation mechanism that optimise this mixed utilities via a deep learning approach. Furthermore, this thesis studied end-to-end network performance at the transport layer, specifically on the interactions between legacy transport protocols, i.e. UDP and TCP, and the acknowledgement mode (AM) of RLC in the current 4G LTE networks under different data and control plane settings. Finally, the thesis presented a study on mobility management of base stations mounted on UAVs via a deep reinforcement learning approach.

Based on the results obtained throughout this PhD project, there are a number of potential research directions identified, which are detailed below.

1) Model compression and acceleration of machine learning models. Towards more agile deployment of machine learning models in mobile networks, especially when involving users' end terminals, model compression is necessary to extend the battery life and preserve the memory space available in mobile devices or on the mobile edge. Popular methods in image processing such as parameter pruning and sharing (Gong et al., 2014), tensor decomposition (Kim et al., 2015) and more recently, knowledge distillation (Hinton et al., 2015), can be exploited for both convolutional and fully connected neural networks (Cheng et al., 2018). When applying these techniques to specific problems, either for the tasks tackled in this thesis i.e. utility op-

timisation and mobility management or other networking related topics, it is to be investigated which compression algorithm suits the problem best and what the right trade off between the level of compression and the loss in accuracy should be sought.

2) Backhaul traffic routing, power consumption, and user association considerations with airborne base stations. Chapter 6 provides a feasibility study of employing a deep reinforcement learning framework, namely A3C, in emergency cellular network deployments. A couple of future directions have been identified through this study. One is to take into consideration multi-hop backhauling to support inter-airSTA communications, in order to cope with larger network sizes. This will lead to a more complicated system model, as the backhaul topology and link rates will introduce constraints on the capacity of subsets of flows traversing certain paths. Routing naturally becomes an important task that requires to be solved with objectives such as optimal network throughput. Moreover, as UAVs are battery-powered, power consumption models should introduce further constrains on the lifetime of each UAV, when taking into account the transmission power and movement of UAVs. Another potential research direction is to consider user association in such scenarios, to achieve high user throughput and airSTA load-balancing.

3) Decentralised decision making in airborne base stations. Based on the preliminary study on mobility management of airborne base stations introduced in Chapter 6, a decentralised multi-agent learning approach can be employed to relax the assumption of centralised control of the multi-UAV system. This will presumably reduce the communication overhead and backhauling traffic load among UAVs, at the potential cost of performance accuracy, due to the lack of global knowledge. The problem can be formulated into a partially observable Markov decision process, and can be treated using multi-agent reinforcement learning methods, for example where each UAV learns their own policy while treating the other UAVs as part of the environment.

4) Evolving economic models for mobile networks. Along the line of the utility framework proposed in this thesis for 5G, a more complex economic situation for multi-service networks can be considered. As the future generation of mobile networks is providing service to more than just smart devices, but a diversified collection of businesses and service portfolio including digital advertising, gaming, and content delivery/caching (Intel, 2018), more sophisticated economic models can be devised.

Utility optimisation concerning requirements from this range of clients can potentially optimise revenue. On the other hand, smart pricing schemes for network resources as commodities may also be applied.

5) Traffic and network capacity forecast. More accurate provisioning of network resources requires incorporating forecasting schemes to estimate users demand and available network capacity. Recent work investigates long-tem forecasting (Zhang and Patras, 2018), yet this needs to be tailored to network specific scenarios, i.e. dense urban deployments or emergency self-deployable network settings, and more importantly, short-term forecasting for more precise and rapid resource partitioning.

Bibliography

3GPP (2015). Technical Specification Group Services and System Aspects; Isolated Evolved Universal Terrestrial Radio Access Network (E-UTRAN) operation for public safety; Stage 1 (Release 13). *3GPP TS 22.346*.

3GPP (2016a). Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation. Standard, 3rd Generation Partnership Project Technical Specification Group.

3GPP (2016b). Evolved Universal Terrestrial Radio Access (E-UTRA); Radio link control (RLC); Protocol specification. Standard, 3rd Generation Partnership Project Technical Specification Group.

3GPP (2016c). Evolved Universal Terrestrial Radio Access (E-UTRA);Radio resource control (RRC); Protocol specification. Standard, 3rd Generation Partnership Project Technical Specification Group.

3GPP (2016d). Radio Access Network Evolved Universal Terrestrial Radio Access (E-UTRA) Base Station (BS) radio transmission and reception. Standard, 3rd Generation Partnership Project Technical Specification Group.

3GPP (2016e). Technical Specification Group Services and System Aspects; Study on Architecture for Next Generation System. *3GPP TS 23.799*.

3GPP (2017a). Architecture enhancements for dedicated core networks. *3GPP TS 23.707*.

3GPP (2017b). Technical Specification Group Services and System Aspects; System Architecture for the 5G System. *3GPP TS 23.501*.

3GPP (2018). 5G; NR; Physical channels and modulation. *3GPP TS 38.211*.

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). TensorFlow: A system for large-scale machine learning. In *Proc. OSDI*, volume 16, pages 265–283.

Agiwal, M., Roy, A., and Saxena, N. (2016). Next generation 5g wireless networks: A comprehensive survey. *IEEE Communications Surveys Tutorials*, 18(3):1617–1655.

Ahuja, R. K., Magnanti, T. L., and Orlin, J. B. (1993). *Network Flows: Theory, Algorithms, and Applications*. Prentice-Hall, Inc.

Air Lynx (2018). Air Lynx. https://www.air-lynx.com/.

Alexandropoulos, G. C. (2017). Position aided beam alignment for millimeter wave backhaul systems with large phased arrays. *arXiv preprint arXiv:1701.03291*.

Alkhateeb, A., El Ayach, O., Leus, G., and Heath, R. W. (2014a). Channel estimation and hybrid precoding for millimeter wave cellular systems. *IEEE Journal of Selected Topics in Signal Processing*, 8(5):831–846.

Alkhateeb, A., Leus, G., and Heath, R. W. (2015). Limited feedback hybrid precoding for multi-user millimeter wave systems. *IEEE transactions on wireless communications*, 14(11):6481–6494.

Alkhateeb, A., Mo, J., Gonzalez-Prelcic, N., and Heath, R. W. (2014b). MIMO precoding and combining solutions for millimeter-wave systems. *IEEE Communications Magazine*, 52(12):122–131.

Alliance, N. (2012). Small cell backhaul requirements. *White Paper*.

Alliance, N. (2016). Description of network slicing concept. *NGMN 5G P1*, 1.

Asadpour, M., Giustiniano, D., Hummel, K. A., and Egli, S. (2013). UAV networks in rescue missions. In *Proceedings of the 8th ACM international workshop on Wireless network testbeds, experimental evaluation & characterization*, pages 91–92. ACM.

Asadpour, M., Van den Bergh, B., Giustiniano, D., Hummel, K. A., Pollin, S., and Plattner, B. (2014). Micro aerial vehicle networks: An experimental analysis of challenges and opportunities. *IEEE Communications Magazine*, 52(7):141–149.

Athuraliya, S., Li, V. H., Low, S. H., and Yin, Q. (2001). Rem: Active queue management *. *IEEE Network the Magazine of Global Internetworking*, 15(3):48–53.

Ayach, O. E., Rajagopal, S., Abu-Surra, S., Pi, Z., and Heath, R. W. (2014). Spatially Sparse Precoding in Millimeter Wave MIMO Systems. *IEEE Trans. Wireless Communications*, 13(3):1499–1513.

Bakre, A. and Badrinath, B. R. (1988). I-tcp: indirect tcp for mobile hosts. *Proc.international Conf. on Distributed Computing Systems*, (5):136–143.

Balakrishnan, H., Padmanabhan, V. N., Seshan, S., and Katz, R. H. (1996). Comparison of mechanisms for improving tcp performance over wireless links. In *Conference on Applications*, pages 256–269.

Bao, L. and Garcia-Luna-Aceves, J. J. (2002). Transmission scheduling in ad hoc networks with directional antennas. *Proceedings of the 8th annual international conference on Mobile computing and networking*, pages 48–58.

BBC (2018). California wildfires: Why are so many listed as missing? https://www.bbc.co.uk/news/world-us-canada-46253575. Accessed on 20/11/2018.

Bertsekas, D. and Gallager, R. (1992). Data Networks. *PrenticeHall, Englewood Cliffs, NJ*.

Bhushan, N., Li, J., Malladi, D., Gilmore, R., Brenner, D., Damnjanovic, A., Sukhavasi, R., Patel, C., and Geirhofer, S. (2014). Network densification: The dominant theme for wireless evolution into 5G. *IEEE Communications Magazine*, 52(2):82–89.

Blumenstein, J., Mikulasek, T., Marsalek, R., and Prokes, A. (2014). In-Vehicle mm-Wave Channel Model and Measurement. In *Vehicular Technology Conference*.

Brown, K. and Singh, S. (1997). M-TCP: TCP for mobile cellular networks. *ACM SIGCOMM Computer Communication Review*, 27(5):19–43.

Cardwell, N., Cheng, Y., Gunn, C. S., Yeganeh, S. H., and Jacobson, V. (2016). BBR: Congestion-Based Congestion Control. *ACM Queue*, 14, September-October:20 – 53.

Cellsite Solutions LLC (2017). EXPANDING HIGH-DEMAND NETWORK COVERAGE WITH CELL ON WHEELS. https://cellsitesolutions.com/cell-on-wheels-expand-network-coverage/.

Cerda-Alabern, L. (2012). On the topology characterization of Guifi.net. In *Proc. IEEE WiMob*, pages 389–396.

Cerwall (ed), P. (2017). Status of Project IEEE 802.11ay. http://www.ieee802.org/11/Reports/tgay_update.htm.

Chandra, K., Prasad, R. V., Niemegeers, I. G., and Biswas, A. R. (2014). Adaptive beamwidth selection for contention based access periods in millimeter wave WLANs. In *Proc. IEEE CCNC*, pages 458–464.

Chen, J., Zhao, P., Wang, Z., and Quan, J. (2016). Enhanced beam selection for multi-user mm-wave massive MIMO systems. *Electronics Letters*, 52(14):1268–1270.

Chen, L., Wang, B., Chen, L., Zhang, X., and Dacheng, Y. (2011). Utility-based resource allocation for mixed traffic in wireless networks. *Proc. IEEE INFOCOM Workshops)*, pages 91–96.

Chen, Q., Tang, J., Wong, D. T. C., Peng, X., and Zhang, Y. (2013). Directional cooperative MAC protocol design and performance analysis for IEEE 802.11ad WLANs. *IEEE Transactions on Vehicular Technology*, 62(6):2667–2677.

Cheng, Y., Wang, D., Zhou, P., and Zhang, T. (2018). Model compression and acceleration for deep neural networks: The principles, progress, and challenges. *IEEE Signal Processing Magazine*, 35(1):126–136.

Choi, J. (2015). Beam Selection in mm-Wave Multiuser MIMO Systems Using Compressive Sensing. *IEEE Transactions on Communications*, 63(8):2936–2947.

Choudhury, R. R., Yang, X., Ramanathan, R., and Vaidya, N. H. (2006). On designing MAC protocols for wireless networks using directional antennas. *IEEE Transactions on Mobile Computing*, 5(5):477–491.

Council, N. C. (2012). Open Data Nottingham: Streetlights. http://www.opendatanottingham.org.uk/dataset.aspx?id=35. Accessed: 2018-12-20.

Cover, T. and Thomas, J. (2006). *Elements of Information Theory (2nd Edition)*. John Wiley & Sons, Inc.

CTTC (2016). LENA: LTE-EPC network simulator. "http://networks.cttc.es/mobile-networks/software-tools/lena/".

Dehos, C., González, J. L., De Domenico, A., Ktenas, D., and Dussopt, L. (2014). Millimeter-wave access and backhauling: The solution to the exponential data traffic increase in 5G mobile communications systems? *IEEE Communications Magazine*, 52(9):88–95.

Di Caro, G., Feo, E., and Gambardella, L. (2014). Use of time-dependent spatial maps of communication quality for network-aware multi-robot path planning. In *Proc. of the 8th Intl. Workshop on Wireless Sensor, Actuator and Robot Networks (WiSARN)*.

Dong, H., Supratak, A., Mai, L., Liu, F., Oehmichen, A., Yu, S., and Guo, Y. (2017). TensorLayer: A versatile library for efficient deep learning development. In *Proc. ACM Multimedia Conference*.

Doumi, T., Dolan, M. F., Tatesh, S., Casati, A., Tsirtsis, G., Anchan, K., and Flore, D. (2013). Lte for public safety networks. *IEEE Communications Magazine*, 51(2):106–112.

ETSI, T. (2011). 100 812:" terrestrial trunked radio (tetra). *Security aspects*.

FAA (2017). FAA Approves Drone to Restore Puerto Rico Cell Service. https://www.faa.gov/news/updates/?newsId=89185.

Facchi, N., Gringoli, F., and Patras, P. (2017). Maximising the utility of enterprise millimetre-wave networks. *arXiv preprint arXiv:1706.04278*.

Facebook Inc. (2018). Terragraph: Solving the Urban Bandwidth Challenge. https://terragraph.com/#terragraph.

Fazel, M. and Chiang, M. (2005). Network utility maximization with nonconcave utilities using sum-of-squares method. *Proc. IEEE CDC-ECC*, 2005(1):1867–1874.

Ferreira, P. V. R., Paffenroth, R., Wyglinski, A. M., Hackett, T. M., and Mortensen, D. J. (2017). Multi-objective reinforcement learning-based deep neural networks for cognitive space communications.

Floyd, S. (1994). Tcp and explicit congestion notification. *Acm Computer Communication Review*, 24(5):8–23.

Flushing, E. F., Kudelski, M., Gambardella, L. M., and Di Caro, G. A. (2014). Spatial prediction of wireless links and its application to the path control of mobile robots. In *Proceedings of the 9th IEEE International Symposium on Industrial Embedded Systems (SIES 2014)*, pages 218–227. IEEE.

Ford, R., Gmez-Cuba, F., Mezzavilla, M., and Rangan, S. (2015). Dynamic time-domain duplexing for self-backhauled millimeter wave cellular networks. In *Proc. IEEE ICC Workshops*, pages 13–18.

Fotouhi, A., Ding, M., and Hassan, M. (2017). DroneCells: Improving 5G Spectral Efficiency using Drone-mounted Flying Base Stations. *CoRR*, abs/1707.02041.

Foukas, X., Marina, M. K., and Kontovasilis, K. (2017a). Orion: RAN Slicing for a Flexible and Cost-Effective Multi-Service Mobile Network Architecture. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, pages 127–140. ACM.

Foukas, X., Patounas, G., Elmokashfi, A., and Marina, M. K. (2017b). Network slicing in 5g: Survey and challenges. *IEEE Communications Magazine*, 55(5):94–100.

Fujitsu (2011). Enhancing LTE cell-edge performance via PDCCH ICIC. Whitepaper, Fujitsu Network Communications Inc.

Fund, F., Wang, C., Liu, Y., Korakis, T., Zink, M., and Panwar, S. (2014). CRAWDAD dataset nyupoly/video (v. 2014-05-09). Downloaded from http://crawdad.org/nyupoly/video/20140509.

Gafni, E. and Bertsekas, D. (1984). Dynamic control of session input rates in communication networks. *IEEE Transactions on Automatic Control*, 29(11):1009–1016.

Giannoulis, A., Patras, P., and Knightly, E. (2013). Mobile Access of Wide-Spectrum Networks: Design, Deployment and Experimental Evaluation. In *Proc. IEEE INFOCOM*, Turin, Italy.

Gini, C. (1921). Measurement of inequality of incomes. *The Economic Journal*, 31(121):124–126.

Glover, F. (1998). A template for scatter search and path relinking. *Lecture Notes in Computer Science*, 1363:13–54.

Gong, Y., Liu, L., Yang, M., and Bourdev, L. (2014). Compressing deep convolutional networks using vector quantization. *arXiv preprint arXiv:1412.6115*.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Grossglauser, M. and Tse, D. N. (2002). Mobility increases the capacity of ad hoc wireless networks. *IEEE/ACM Trans. Networking*, 10(4):477–486.

Gupta, A. and Jha, R. K. (2015). A survey of 5g network: Architecture and emerging technologies. *IEEE Access*, 3:1206–1232.

Gustafson, C., Haneda, K., Wyne, S., and Tufvesson, F. (2014). On mm-Wave Multipath Clustering and Channel Modeling. *IEEE Transactions on Antennas & Propagation*, 62(3):1445–1455.

Haleplidis, E., Pentikousis, K., Denazis, S., Salim, J. H., Meyer, D., and Koufopavlou, O. (2015). Software-defined networking (SDN): Layers and architecture terminology. RFC 7426.

Hamilton, R., Iyengar, J., Swett, I., Wilk, A., et al. (2016). QUIC: A UDP-based secure and reliable transport for HTTP/2. *IETF, draft-tsvwg-quic-protocol-02*.

Hande, P., Zhang, S., and Chiang, M. (2007). Distributed rate allocation for inelastic flows. *IEEE/ACM Trans. Networking*, 15(6):1240–1253.

Hanson, W. A. (2016). Satellite internet in the mobile age. *New Space*, 4(3):138–152.

Haque, I. T. and Abu-Ghazaleh, N. (2016). Wireless software defined networking: A survey and taxonomy. *IEEE Communications Surveys & Tutorials*, 18(4):2713–2737.

Heath, R. W., Gonzalez-Prelcic, N., Rangan, S., Roh, W., and Sayeed, A. M. (2016). An overview of signal processing techniques for millimeter wave MIMO systems. *IEEE journal of selected topics in signal processing*, 10(3):436–453.

Hemanth, C. and Venkatesh, T. (2016). Performance analysis of service periods (SP) of the IEEE 802.11ad hybrid MAC protocol. *IEEE Transactions on Mobile Computing*, 15(5):1224–1236.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hock, M., Bless, R., and Zitterbart, M. (2017). Experimental evaluation of bbr congestion control. In *2017 IEEE 25th International Conference on Network Protocols (ICNP)*, pages 1–10. IEEE.

Hong, X., Gerla, M., Pei, G., and Chiang, C.-C. (1999). A group mobility model for ad hoc wireless networks. In *ACM MSWiM*, pages 53–60.

Hu, Q. and Blough, D. M. (2017). Relay Selection and Scheduling for Millimeter Wave Backhaul in Urban Environments. In *Proc. IEEE MASS*, pages 206–214.

Huang, J., Qian, F., Guo, Y., Zhou, Y., Xu, Q., Mao, Z. M., Sen, S., and Spatscheck, O. (2013). An in-depth study of LTE: effect of network protocol and application behavior on performance. In *ACM SIGCOMM Computer Communication Review*, volume 43, pages 363–374. ACM.

Huawei Co., Ltd (2016). Huawei and Deutsche Telekom Demonstrate 5G E2E Network Slicing Technology. https://www.huawei.com/en/press-events/news/2016/2/demonstrate-5g-e2e-network-slicing-technology.

Hur, S., Kim, T., Love, D. J., Krogmeier, J. V., Thomas, T. A., and Ghosh, A. (2013). Millimeter wave beamforming for wireless backhaul and access in small cell networks. *IEEE Transactions on Communications*, 61(10):4391–4403.

Huston, G. (2018). TCP and BBR. "https://ripe76.ripe.net/presentations/10-2018-05-15-bbr.pdf".

IEEE 802.11ad Std. (2014). Amendment 3: Enhancements for Very High Throughput in the 60GHz Band. *ISO/IEC/IEEE 8802-11:2012/Amd.3:2014(E)*.

Intel (2018). How 5G Will Transform the Business of Media and Entertainment. https://newsroom.intel.com/wp-content/uploads/sites/11/2018/10/ovumintel5gebook.pdf.

ITU (2015). Propagation data and prediction methods for the planning of short-range outdoor radiocommunication systems and radio local area networks in the frequency range 300 MHz to 100 GHz. Recommendation, International Telecommunication Union.

Jaffe, J. (1981). Bottleneck flow control. *IEEE Transactions on Communications*, 29(7):954–962.

Katabi, D., Handley, M., and Rohrs, C. E. (2002). Congestion control for high bandwidth-delay product networks. In *Conference on Applications*.

Kato, N., Fadlullah, Z. M., Mao, B., Tang, F., Akashi, O., Inoue, T., and Mizutani, K. (2017). The Deep Learning Vision for Heterogeneous Network Traffic Control: Proposal, Challenges, and Future Perspective. *IEEE Wireless Communications*, 24(3):146–153.

Kelly, F. (1997). Charging and rate control for elastic traffic. *Trans. Emerging Telecommunications Technologies*, 8(1):33–37.

Kibria, M. G., Nguyen, K., Villardi, G. P., Ishizu, K., and Kojima, F. (2017). Big Data Analytics and Artificial Intelligence in Next-Generation Wireless Networks. *IEEE Access*, PP(99):1–1.

Kim, M., Iwata, T., Umeki, K., Wangchuk, K., Takada, J. I., and Sasaki, S. (2017). Mm-wave outdoor-to-indoor channel measurement in an open square small cell scenario. In *International Symposium on Antennas & Propagation*.

Kim, Y.-D., Park, E., Yoo, S., Choi, T., Yang, L., and Shin, D. (2015). Compression of deep convolutional neural networks for fast and low power mobile applications. *arXiv preprint arXiv:1511.06530*.

Kim, Y. H. et al. (2014). Slicing the next mobile packet core network. In *2014 11th International Symposium on Wireless Communications Systems (ISWCS)*, pages 901–904. IEEE.

Kingma, D. and Ba, J. (2015). Adam: A method for stochastic optimization. *Proc. ICLR*.

Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S. (2017). Self-normalizing neural networks. In *Proc. NIPS*.

Korakis, T., Jakllari, G., Jakllari, G., Tassiulas, L., and Tassiulas, L. (2003). A MAC Protocol for Full Exploitation of Directional Antennas in Ad-hoc Wireless Networks. *Proc. of MOBIHOC*, pages 98–107.

Lan, T., Kao, D., Chiang, M., and Sabharwal, A. (2010). An axiomatic theory of fairness in network resource allocation. In *Proc. IEEE INFOCOM*.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436.

Lee, H.-H. and Ko, Y.-C. (2011). Low complexity codebook-based beamforming for MIMO-OFDM systems in millimeter-wave WPAN. *IEEE Transactions on Wireless Communications*, 10(11):3607–3612.

Leith, D. J., Cao, Q., and Subramanian, V. G. (2012). Max-min fairness in 802.11 mesh networks. *IEEE/ACM Transactions on Networking*, 20(3):756–769.

Li, B., Zhou, Z., Zou, W., Sun, X., and Du, G. (2013). On the efficient beam-forming training for 60GHz wireless personal area networks. *IEEE Transactions on Wireless Communications*, 12(2):504–515.

Li, R., Zhang, C., Patras, P., Cao, P., and Thompson, J. S. (2018). DELMU: A Deep Learning Approach to Maximising the Utility of Virtualised Millimetre-Wave Backhauls. *arXiv preprint arXiv:1810.00356*.

Li, Y., Luo, J., Xu, W., Vucic, N., Pateromichelakis, E., and Caire, G. (2017a). A joint scheduling and resource allocation scheme for millimeter wave heterogeneous networks. In *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6. IEEE.

Li, Z., Han, S., and Molisch, A. F. (2017b). Channel-statistics-based analog downlink beamforming for millimeter-wave multi-user massive MIMO. In *IEEE International Conference on Communications*.

Loon LLC (2018). Loon. https://loon.co/.

MacCartney, G. R. and Rappaport, T. S. (2014). 73GHz millimeter wave propagation measurements for outdoor urban mobile and backhaul communications in New York City. In *Proc. IEEE ICC*, pages 4862–4867.

Mandke, K. and Nettles, S. M. (2010). A dual-band architecture for multi-Gbps communication in 60GHz multi-hop networks. In *Proc. ACM International Workshop on MmWave Communications: From Circuits to Networks*, pages 9–14.

Mao, H., Netravali, R., and Alizadeh, M. (2017). Neural adaptive video streaming with pensieve. In *SIGCOMM*, pages 197–210. ACM.

Mesodiakaki, A., Kassler, A., Zola, E., Ferndahl, M., and Cai, T. (2016). Energy efficient line-of-sight millimeter wave small cell backhaul: 60, 70, 80 or 140 GHz? In *Proc. IEEE WoWMoM*, pages 1–9.

Mikrotik (2017). Mikrotik WAP60. https://mikrotik.com/product/wap_60g/.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *ICML*, pages 1928–1937.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529.

Nakao, A., Du, P., Kiriha, Y., Granelli, F., Gebremariam, A. A., Taleb, T., and Bagaa, M. (2017). End-to-end network slicing for 5g mobile networks. *Journal of Information Processing*, 25:153–163.

Neil, C. T., Shafi, M., Smith, P. J., Dmochowski, P. A., and Zhang, J. (2017). Impact of Microwave and mmWave Channel Models on 5G Systems Performance. *IEEE Transactions on Antennas & Propagation*, 65(12):6505–6520.

NGMN (2015). 5G White Paper. *Next generation mobile networks*.

Nguyen, B., Banerjee, A., Gopalakrishnan, V., Kasera, S., Lee, S., Shaikh, A., and Van der Merwe, J. (2014). Towards understanding TCP performance on LTE/EPC mobile networks. In *Proceedings of the 4th workshop on All things cellular: operations, applications, & challenges*, pages 41–46. ACM.

Nikaein, N., Schiller, E., Favraud, R., Katsalis, K., Stavropoulos, D., Alyafawi, I., Zhao, Z., Braun, T., and Korakis, T. (2015). Network store: Exploring slicing in future 5g networks. In *Proceedings of the 10th International Workshop on Mobility in the Evolving Internet Architecture*, pages 8–13. ACM.

Nitsche, T., Cordeiro, C., Flores, A. B., Knightly, E. W., Perahia, E., and Widmer, J. C. (2014). IEEE 802.11ad: Directional 60 GHz communication for multi-Gigabit-per-second Wi-Fi. *IEEE Communications Magazine*, 52(12):132–141.

Niu, Y., Ding, W., Wu, H., Li, Y., Chen, X., Ai, B., and Zhong, Z. (2019). Relay-Assisted and QoS Aware Scheduling to Overcome Blockage in mmWave Backhaul Networks. *IEEE Transactions on Vehicular Technology*, 68(2):1733–1744.

Niu, Y., Gao, C., Li, Y., Su, L., Jin, D., Zhu, Y., and Wu, D. O. (2017). Energy-efficient scheduling for mmwave backhauling of small cells in heterogeneous cellular networks. *IEEE Transactions on Vehicular Technology*, 66(3):2674–2687.

NS-3 Project (2016). LENA Design Documentation. "https://www.nsnam.org/docs/models/html/lte-design.html". Accessed: 2016-07-1.

Ofcomm (2018). Enabling 5G UK. https://www.ofcom.org.uk/__data/assets/pdf_file/0022/111883/enabling-5g-uk.pdf.

Ordonez-Lucena, J., Ameigeiras, P., Lopez, D., Ramos-Munoz, J. J., Lorca, J., and Folgueira, J. (2017). Network slicing for 5g with sdn/nfv: Concepts, architectures, and challenges. *IEEE Communications Magazine*, 55(5):80–87.

Orsino, A., Ometov, A., Fodor, G., Moltchanov, D., Militano, L., Andreev, S., Yilmaz, O. N., Tirronen, T., Torsner, J., Araniti, G., et al. (2017). Effects of Heterogeneous Mobility on D2D-and Drone-Assisted Mission-Critical MTC in 5G. *IEEE Communications Magazine*, 55(2):79–87.

Oueis, J., Conan, V., Lavaux, D., Stanica, R., and Valois, F. (2017). Overview of LTE Isolated E-UTRAN Operation for Public Safety. *IEEE Communications Standards Magazine*, 1(2).

Papadimitriou, C. H. and Steiglitz, K. (1998). *Combinatorial optimization: algorithms and complexity*. Courier Corporation.

Peter, M., Weiler, R. J., Keusgen, W., Eichler, T., Kottkamp, M., and Nhring, A. (2016). Characterization of mm-wave channel sounders up to W-Band and validation of measurement results. In *European Conference on Antennas & Propagation*.

Polese, M., Jana, R., and Zorzi, M. (2017). TCP and MP-TCP in 5G mmWave networks. *IEEE Internet Computing*, 21(5):12–19.

Qian, L., Shirani-Mehr, H., Balercia, T., and Papathanassiou, A. (2015). Validation of a Geometry-Based Statistical mmWave Channel Model Using Ray-Tracing Simulation. In *Vehicular Technology Conference*.

Qiao, J., Shen, X., Mark, J. W., and He, Y. (2015). MAC-layer concurrent beamforming protocol for indoor millimeter-wave networks. *IEEE Transactions on Vehicular Technology*, 64(1):327–338.

Qualcomm (2019). Reference Design sub 6Ghz and mmwave Fixed Wireless. https://www.qualcomm.com/news/releases/2019/02/25/qualcomm-announces-5g-reference-design-sub-6-ghz-and-mmwave-fixed-wireless.

Qualcomm Technologies Inc. (2017). Making 5G mmWave a commercial reality...in your smartphone. https://www.qualcomm.com/media/documents/files/making-5g-mmwave-a-commercial-reality-in-your-smartphone.pdf.

Qualcomm Technologies Inc. (2018). Qualcomm and Facebook to Bring High-Speed Internet Connectivity Over 60GHz to Urban Areas. https://www.qualcomm.com/news/releases/2018/05/21/qualcomm-and-facebook-bring-high-speed-internet-connectivity-over-60ghz.

Radunović, B. and Le Boudec, J.-Y. (2007). A unified framework for max-min and min-max fairness with applications. *IEEE/ACM Transactions on Networking*, 15(5):1073–1083.

Ramani, R. and Karandikar, A. (2000). Explicit congestion notification (ecn) in tcp over wireless network. In *IEEE International Conference on Personal Wireless Communications*.

Rappaport, T. S., Gutierrez, F., Ben-Dor, E., Murdock, J. N., Qiao, Y., and Tamir, J. I. (2013a). Broadband millimeter-wave propagation measurements and models

using adaptive-beam antennas for outdoor urban cellular communications. *IEEE transactions on antennas and propagation*, 61(4):1850–1859.

Rappaport, T. S., Heath Jr, R. W., Daniels, R. C., and Murdock, J. N. (2014). *Millimeter wave wireless communications*. Pearson Education.

Rappaport, T. S., Qiao, Y., Tamir, J. I., Murdock, J. N., and Ben-Dor, E. (2012). Cellular broadband millimeter wave propagation and angle of arrival for adaptive beam steering systems. In *Radio and Wireless Symposium (RWS), 2012 IEEE*, pages 151–154. IEEE.

Rappaport, T. S., Sun, S., Mayzus, R., Zhao, H., Azar, Y., Wang, K., Wong, G. N., Schulz, J. K., Samimi, M., and Gutierrez, F. (2013b). Millimeter wave mobile communications for 5G cellular: It will work! *IEEE Access*, 1:335–349.

RCR Wireless (2018). Drone test trials to shape FAA rules. https://www.rcrwireless.com/20180424/wireless/176063/drone_test_trials_to_shape_FAA_rules_tag41.

Roh, W., Seol, J.-Y., Park, J., Lee, B., Lee, J., Kim, Y., Cho, J., Cheun, K., and Aryanfar, F. (2014). Millimeter-wave beamforming as an enabling technology for 5G cellular communications: Theoretical feasibility and prototype results. *IEEE Communications Magazine*, 52(2):106–113.

Samdanis, K., Costa-Perez, X., and Sciancalepore, V. (2016). From network sharing to multi-tenancy: The 5G network slice broker. *IEEE Communications Magazine*, 54(7):32–39.

Samimi, M. K. and Rappaport, T. S. (2016). 3-D Millimeter-Wave Statistical Channel Model for 5G Wireless System Design. *IEEE Transactions on Microwave Theory & Techniques*, 64(7):2207–2225.

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Netw*, 61:85–117.

Schulz, P., Matthe, M., Klessig, H., Simsek, M., Fettweis, G., Ansari, J., Ashraf, S. A., Almeroth, B., Voigt, J., Riedel, I., Puschmann, A., Mitschele-Thiel, A., Muller, M., Elste, T., and Windisch, M. (2017). Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture. *IEEE Communications Magazine*, 55(2):70–78.

Sciancalepore, V., Samdanis, K., Costa-Perez, X., Bega, D., Gramaglia, M., and Banchs, A. (2017). Mobile Traffic Forecasting for Maximizing 5G Network Slicing Resource Utilization. *Proc. IEEE INFOCOM*.

Sciancalepore, V., Zanzi, L., Costa-Perez, X., and Capone, A. (2018). ONETS: Online Network Slice Broker From Theory to Practice. *arXiv preprint arXiv:1801.03484*.

Semiari, O., Saad, W., Bennis, M., and Dawy, Z. (2017). Inter-operator resource management for millimeter wave multi-hop backhaul networks. *IEEE Transactions on Wireless Communications*, 16(8):5258–5272.

Seppänen, K., Kilpi, J., Paananen, J., Suihko, T., Wainio, P., and Kapanen, J. (2016). Multipath routing for mmWave WMN backhaul. In *2016 IEEE International Conference on Communications Workshops (ICC)*, pages 246–253.

Seppnen, K. and Kapanen, J. (2016). Fair queueing for mmWave WMN backhaul. In *IEEE International Symposium on Personal*.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., and Bolton, A. (2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359.

Sim, G. H., Li, R., Cano, C., Malone, D., Patras, P., and Widmer, J. (2016a). Learning from experience: Efficient decentralized scheduling for 60GHz mesh networks. In *Proc. IEEE WoWMoM*.

Sim, G. H., Nitsche, T., and Widmer, J. C. (2016b). Addressing MAC layer inefficiency and deafness of IEEE 802.11ad millimeter wave networks using a multi-band approach. In *Proc. IEEE PIMRC*, pages 1–6.

Singh, S., Mudumbai, R., and Madhow, U. (2010). Distributed coordination with deaf neighbors: Efficient medium access for 60 GHz mesh networks. In *Proc. IEEE INFOCOM*, pages 1–9.

Singh, S., Mudumbai, R., and Madhow, U. (2011). Interference analysis for highly directional 60-GHz mesh networks: The case for rethinking medium access control. *IEEE/ACM Transactions on Networking*, 19(5).

Smith, W. and Dowell, J. (2000). A case study of co-ordinative decision-making in disaster management. *Ergonomics*, 43(8):1153–1166.

Son, K., Oh, E., and Krishnamachari, B. (2015). Energy-efficient design of heterogeneous cellular networks from deployment to operation. *Computer Networks*, 78:95–106.

Srivastava, R. K., Greff, K., and Schmidhuber, J. (2015). Training very deep networks. In *Advances in Neural Information Processing Systems*, pages 2377–2385.

Su, H. and Zhang, X. (2009). Joint Link Scheduling and Routing for Directional-Antenna Based 60 GHz Wireless Mesh Networks. In *Proc. IEEE GLOBECOM*, pages 1–6.

Sun, H., Chen, X., Shi, Q., Hong, M., and Sidiropoulos, N. D. (2017). Learning to optimize: Training deep neural networks for wireless resource management. In *IEEE International Workshop on Signal Processing Advances in Wireless Communications*.

SWARMIX.org (2014). SWARMIX: Synergistic Interactions in Swarms of Heterogeneous Agents. "http://www.swarmix.org/".

Taori, R. and Sridharan, A. (2015). Point-to-multipoint in-band mmwave backhaul for 5g networks. *IEEE Communications Magazine*, 53(1):195–201.

Telefonaktiebolaget LM Ericsson (2018). Autonomous control and management of drones using 5G network slicing. https://www.ericsson.com/en/videos/2018/2/autonomous-control-and-management-of-drones.

TP-LINK (2017). TP-LINK ad7200. https://www.tp-link.com/en/home-networking/wifi-router/ad7200/.

Udell, M. and Boyd, S. (2013). Maximizing a Sum of Sigmoids. *Optimization and Engineering*, pages 1–25.

Ugray, Z., Lasdon, L., Plummer, J., Glover, F., Kelly, J., and Martí, R. (2007). Scatter search and local NLP solvers: A multistart framework for global optimization. *INFORMS Journal on Computing*, 19(3):328–340.

Vu, T. K., Liu, C. F., Bennis, M., Debbah, M., and Latvaaho, M. (2018). Path Selection and Rate Allocation in Self-Backhauled mmWave Networks.

Wang, J., Ashour, M., Lagoa, C., Aybat, N., Che, H., and Duan, Z. (2017). Non-Concave Network Utility Maximization in Connectionless Networks: A Fully Distributed Traffic Allocation Algorithm. In *Proc. IEEE ACC*, pages 3980–3985.

Wang, J., Lan, Z., Pyo, C.-W., Baykas, T., Sum, C.-S., Rahman, M. A., Gao, J., Funada, R., Kojima, F., Harada, H., and Kato, S. (2009). Beam codebook based beamforming protocol for multi-Gbps millimeter-wave WPAN systems. *IEEE Journal on Selected Areas in Communications*, 27(8):1390–1399.

Wang, J., Wu, Y., Yen, N., Guo, S., and Cheng, Z. (2016). Big data analytics for emergency communication networks: A survey. *IEEE Communications Surveys & Tutorials*, 18(3):1758–1778.

Wang, P., Jiang, H., Zhuang, W., and Poor, H. (2008). Redefinition of max-min fairness in multi-hop wireless networks. *IEEE Transactions on Wireless Communications*, 7(12):4786–4791.

Welch, Chris (2018). Google's Project Loon suffers accident as balloon takes out power lines. https://www.theverge.com/2014/6/3/5777182/google-project-loon-balloon-takes-out-power-lines.

Wikipedia (2018). Loon (company) Incidents. https://en.wikipedia.org/wiki/Loon_(company)#Incidents.

Yin, X., Jindal, A., Sekar, V., and Sinopoli, B. (2015). A Control-Theoretic Approach for Dynamic Adaptive Video Streaming over HTTP. *Proc. ACM SIGCOMM*, pages 325–338.

Ying, H., Zheng, Z., Yu, F. R., Nan, Z., and Zhang, Y. (2018). Deep reinforcement learning-based optimization for cache-enabled opportunistic interference alignment wireless networks. *IEEE Transactions on Vehicular Technology*, PP(99):1–1.

Yu, Y., Xu, M., and Yang, Y. (2017). When quic meets tcp: An experimental study. In *2017 IEEE 36th International Performance Computing and Communications Conference (IPCCC)*, pages 1–8.

Yun, Z., Yong, N., Li, J., Wu, D. O., Yong, L., and Jin, D. (2016). QoS-aware scheduling for small cell millimeter wave mesh backhaul. In *IEEE International Conference on Communications*.

ZDNet.co.kr (2015). SK, Ericsson demo 5G network slicing technology. https://www.zdnet.com/article/sk-ericsson-demo-5g-network-slicing-technology/.

Zhang, C., Ouyang, X., and Patras, P. (2017a). ZipNet-GAN: Inferring Fine-grained Mobile Traffic Patterns via a Generative Adversarial Neural Network. In *Proc. ACM CoNEXT*, pages 363–375.

Zhang, C. and Patras, P. (2018). Long-term mobile traffic forecasting using deep spatio-temporal neural networks. In *Proc. ACM MobiHoc*, pages 231–240, Los Angeles, CA, USA.

Zhang, C., Patras, P., and Haddadi, H. (2018). Deep Learning in Mobile and Wireless Networking: A Survey. *arXiv preprint arXiv:1803.04311*.

Zhang, H., Liu, N., Chu, X., Long, K., Aghvami, A.-H., and Leung, V. C. (2017b). Network slicing based 5g and future mobile networks: mobility, resource management, and challenges. *IEEE Communications Magazine*, 55(8):138–145.

Zhang, L., Okamawari, T., and Fujii, T. (2012). Performance evaluation of tcp and udp during lte handover. In *Wireless Communications and Networking Conference (WCNC), 2012 IEEE*, pages 1993–1997. IEEE.

Zhang, S. and Datta, A. (2005). A Directional-Antenna Based MAC Protocol for Wireless Sensor Networks. In *Computational Science and Its Applications ICCSA 2005*, volume 3481 of *Lecture Notes in Computer Science*, pages 686–695. Springer Berlin Heidelberg.

Zhou, L. and Ohashi, Y. (2012). Efficient codebook-based MIMO beamforming for millimeter-wave WLANs. In *Proc. IEEE PIMRC*, pages 1885–1889.

Zhu, Y., Niu, Y., Li, J., Wu, D. O., Li, Y., and Jin, D. (2016). Qos-aware scheduling for small cell millimeter wave mesh backhaul. In *2016 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE.