Algorithms and Experimentation for Future Wireless Networks:
From Internet-of-Things to Full-Duplex

Tingjun Chen

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

Abstract

Algorithms and Experimentation for Future Wireless Networks:

From Internet-of-Things to Full-Duplex

Tingjun Chen

Future and next-generation wireless networks are driven by the rapidly growing wireless traffic stemming from diverse services and applications, such as the Internet-of-Things (IoT), virtual reality, autonomous vehicles, and smart intersections. Many of these applications require massive connectivity between IoT devices as well as wireless access links with ultra-high bandwidth (Gbps or above) and ultra-low latency (10 ms or less). Therefore, realizing the vision of future wireless networks requires significant research efforts across all layers of the network stack. In this thesis, *we use a cross-layer approach and focus on several critical components of future wireless networks including IoT systems and full-duplex (FD) wireless, and on experimentation with advanced wireless technologies in the NSF PAWR COSMOS testbed.*

First, we study tracking and monitoring applications in the IoT and focus on ultra-low-power energy harvesting networks. Based on realistic hardware characteristics, we design and optimize Panda, a centralized probabilistic protocol for maximizing the neighbor discovery rate between energy harvesting nodes under a power budget. Via testbed evaluation using commercial off-the-shelf energy harvesting nodes, we show that Panda outperforms existing protocols by up to $3\times$ in terms of the neighbor discovery rate. We further explore this problem and consider a general throughput maximization problem among a set of heterogeneous energy-constrained ultra-low-power nodes. We analytically identify the theoretical fundamental limits of the rate at which data can be exchanged between these nodes, and design the distributed probabilistic protocol, EconCast, which approaches the maximum throughput in the limiting sense. Performance evaluations of EconCast using both simulations and real-world experiments show that it achieves up to an order of magnitude higher throughput than Panda and other known protocols.

We then study FD wireless – simultaneous transmission and reception at the same frequency – a key technology that can significantly improve the data rate and reduce communication la-

tency by employing self-interference cancellation (SIC). In particular, we focus on enabling FD on small-form-factor devices leveraging the technique of frequency-domain equalization (FDE). We design, model, and optimize the FDE-based RF canceller, which can achieve >50 dB RF SIC across 20 MHz bandwidth, and experimentally show that our prototyped FD radios can achieve a link-level throughput gain of 1.85–$1.91\times$. We also focus on combining FD with phased arrays, employing optimized transmit and receive beamforming, where the spatial degrees of freedom in multi-antenna systems are repurposed to achieve wideband RF SIC. Moving up in the network stack, we study heterogeneous networks with half-duplex and FD users, and develop the novel Hybrid-Greedy Maximum Scheduling (H-GMS) algorithm, which achieves throughput optimality in a distributed manner. Analytical and simulation results show that H-GMS achieves 5–$10\times$ better delay performance and improved fairness compared with state-of-the-art approaches.

Finally, we described experimentation and measurements in the city-scale COSMOS testbed being deployed in West Harlem, New York City. COSMOS' key building blocks include software-defined radios, millimeter-wave radios, a programmable optical network, and edge cloud, and their convergence will enable researchers to remotely explore emerging technologies in a real world environment. We provide a brief overview of the testbed and focus on experimentation with advanced technologies, including the integrating of open-access FD radios in the testbed and a pilot study on converged optical-wireless x-haul networking for cloud radio access networks (C-RANs). We also present an extensive 28 GHz channel measurements in the testbed area, which is a representative dense urban canyon environment, and study the corresponding signal-to-noise ratio (SNR) coverage and achievable data rates. The results of this part helped drive and validate the design of the COSMOS testbed, and can inform further deployment and experimentation in the testbed.

In this thesis, we make several theoretical and experimental contributions to ultra-low-power energy harvesting networks and the IoT, and FD wireless. We also contribute to the experimentation and measurements in the COSMOS advanced wireless testbed. We believe that these contributions are essential to connect fundamental theory to practical systems, and ultimately to real-world applications, in future wireless networks.

Table of Contents

List of Tables

List of Figures

x

xiii

Acknowledgments

I am grateful and fortunate to have many amazing people along my Ph.D journey. My first and deepest thanks go to my Ph.D. advisor, Professor Gil Zussman. I started working with Gil when I was a visiting undergraduate student, and I have been extremely fortunate to have him as my Ph.D. advisor. I am always inspired by his broad vision, profound knowledge, insightful thoughts, and curiosity and passion for research. I am also grateful for his friendship and genuine support, and for teaching me so many things, including and beyond research, that have gradually shaped me as a researcher and scientist. I wish that I will have as good advisor–advisee relationship and friendship with my students as that Gil has with me.

I would like to extend my special thanks to Professors Harish Krishnaswamy and Javad Ghaderi for being invaluable collaborators and for their support on my research. I have learned a great deal about hardware and circuits for wireless systems from Harish, and I am always inspired by his creative ideas as well as depth and breath of knowledge. Javad has always been inspiring me through his solid foundations in mathematics, algorithm design, and network optimization. Working with both of them has been a unique experience for me to bridge theory and systems.

During my Ph.D. studies, I have been fortunate to collaborate with many distinguished researchers. I am grateful to have worked with Professor Dan Rubenstein on the EnHANTs project, and with Professors Dan Kilper and Dipankar Raychaudhuri as well as Ivan Seskar on the COSMOS project. I would also like to express my appreciation to Professor Debasis Mitra for agreeing to be on my thesis committee. I spent a wonderful summer as a research intern at Nokia Bell Labs working with Dr. Dmitry Chizhik, Dr. Jinfeng Du, and Dr. Reinaldo Valenzuela, and I have benefited a lot from our long-term collaboration.

During my undergraduate studies at Tsinghua University, I was fortunate to be advised by Professors Zhisheng Niu and Sheng Zhou, who led me into the research world which later motivated me to pursue a doctoral degree. I am also grateful for having Professor Longbo Huang as a true friend and for his advice and guidance in many occasions.

I am extremely grateful to have Craig Gutterman as a true friend, and we have been discussing

research as well as sharing thoughts and feelings throughout our Ph.D. journey. My thesis would not have been possible without successful collaborations. I am thankful for working with Guy Grebla and Robert Margolies within the EnHANTs project, who guided me through the starting of my Ph.D. study. I am also thankful for having worked with Ph.D. students Mahmood Baraani Dastjerdi, Jelena Diakonikolas, Manav Kohli, Aravind Nagulu, and Jin Zhou within the FlexI-CoN project, and our collaboration has led to exciting results. Working with Jakub Kolodziejski, Artur Minakhmetov, Michael Sherman, Jiakai Yu, Shengxiang Zhu in various aspects within the COSMOS project has also been a unique and fruitful experience.

I deeply appreciate the guidance from my other senior mentors and colleagues: Tianwei Deng, Maria Gorlatova, Varun Gupta, Dr. Jonathan Ostrometzky, Negar Reikaramian, and Saleh Soltan. I have also been fortunate to work with many talented students at Columbia who have contributed to different projects, especially Gregory Chen, Tianyi Dai, Angel Daniel Estigarribia, Guy Farkash, Nicole Grimwood, Saahil Jain, and Jackson Welles.

I have made many amazing friends while at Columbia: Todd Arnold, Niloofar Bayat, Sasha Gazman, Xinyi Ge, Mahshid Ghasemi, Ege Gurmericliler, Jason Holley, Shixin Lan, Yuanbo Li, Kunal Mahajan, Vivek Mangal, Jingping Nie, Mehrnoosh Shafiee, and Zhe Sun. I thank all of them and many others for their sincere friendship and support. I am very fortunate to have Janet Kayfetz as a dear friend, who has provided immense help and support both in and outside classrooms. I also thank Jane Nisselson, with whom I really enjoy chatting about sneakers and fashion, for teaching me video shooting and filmmaking.

Finally, and above all, I thank my parents for their unconditional love and support.

To my amazing parents, Mei and Feng.

Chapter 1: Introduction

1.1 Future Wireless Networks and Systems

5G and next-generation wireless networks are driven by the rapidly growing wireless traffic stemming from diverse services and applications, such as the Internet-of-Things (IoT), virtual reality, autonomous vehicles, and smart intersections, as illustrated in Fig. 1.1. Many of these applications require massive connectivity between IoT devices [1, 2] and wireless access links with ultra-high bandwidth (Gbps or above) and ultra-low latency (10 ms or less) [3–5]. Therefore, realizing the vision of future wireless networks requires significant research efforts across all layers of the network stack.
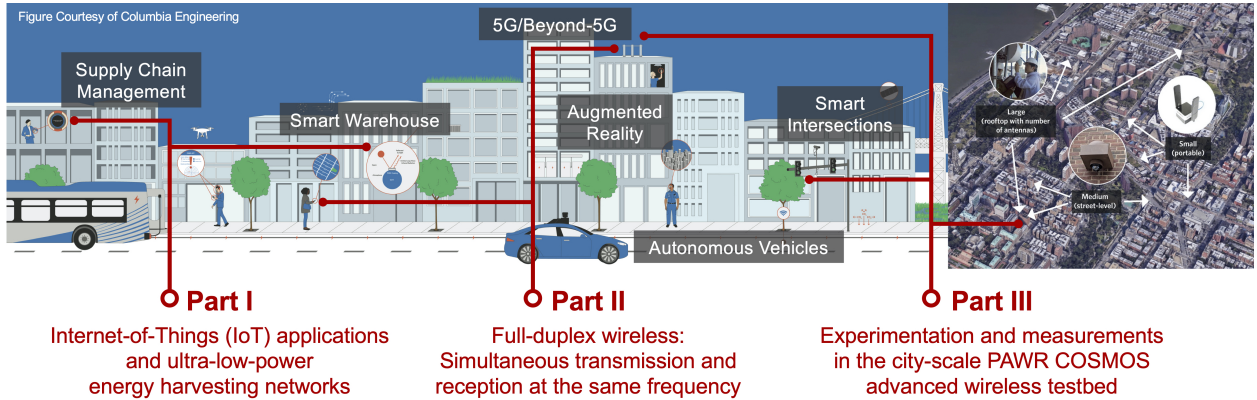
Figure 1.1: Future wireless networks will feature ultra-high bandwidth and ultra-low latency, capabilities that can support various real-time applications including augmented reality, autonomous vehicles, and smart intersections. The focus of this thesis is on algorithms design and experimental evaluations for Internet-of-Things (IoT) systems, and full-duplex (FD) wireless networks, as well as experimentation and measurements in the city-scale COSMOS advanced wireless testbed.

This thesis focuses on several critical components of next-generation wireless networks including IoT applications and ultra-low-power energy harvesting networks, full-duplex (FD) wireless, as well as on experimentation with advanced wireless technologies and millimeter-wave (mmWave)

1

channel measurements in a city-scale testbed (see Fig. 1.1). Specifically, we focus on emerging Physical layer technologies and their interactions with the higher layers in the network stack. By using a cross-layer approach, we design practical, efficient, and scalable algorithms and systems, and develop customized prototypes and testbeds to evaluate their performance.
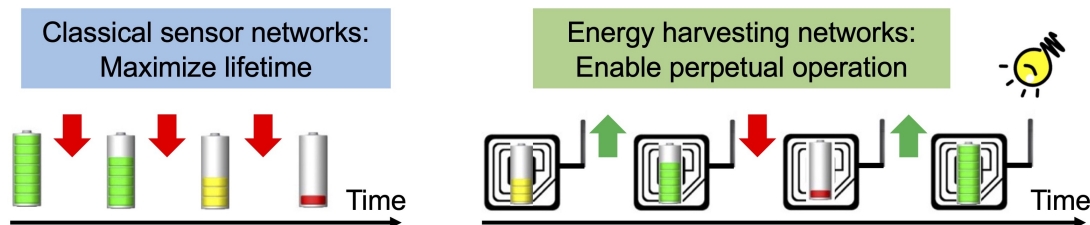
Figure 1.2: (Left) Sensor nodes in classical sensor networks are powered by batteries and the objective is usually to maximize their lifetime. (Right) Energy harvesting enables perpetual operation of ultra-low-power tags, where the energy level of each tag depends on both the energy harvesting profile as well as its own activities (e.g., sleep, listen/receive, and transmit).

In IoT systems that connect a massive number of objects equipped with sensors, actuators, and tags [6, 7], *energy harvesting and ultra-low-power communications* are two emerging technologies that can be used to support various tracking and monitoring applications. Compared to traditional sensor networks, where nodes are powered by batteries and the objective is usually to maximize their lifetime, energy harvesting and ultra-low-power communications can enable battery-less and perpetual operation of wireless nodes. An example is active tags that can be attached to physical objects, harvest energy from ambient sources (e.g., indoor light), and then communicate tag-to-tag toward gateways. Because this reliance on node-to-node communication requires less infrastructure than traditional (RFID/reader-based) systems, these ultra-low-power nodes can enable ubiquitous connectivity and, in particular, facilitate tracking applications spanning healthcare and smart building and agriculture.

In future wireless networks, *full-duplex (FD) communications* [8, 9] – simultaneous transmission and reception on the same frequency, as illustrated in Fig. 1.3 – is a key technique that can significantly improve spectrum efficiency and reduce communication latency at the Physical layer as well as provide many other benefits at the higher network layers. The fundamental challenge
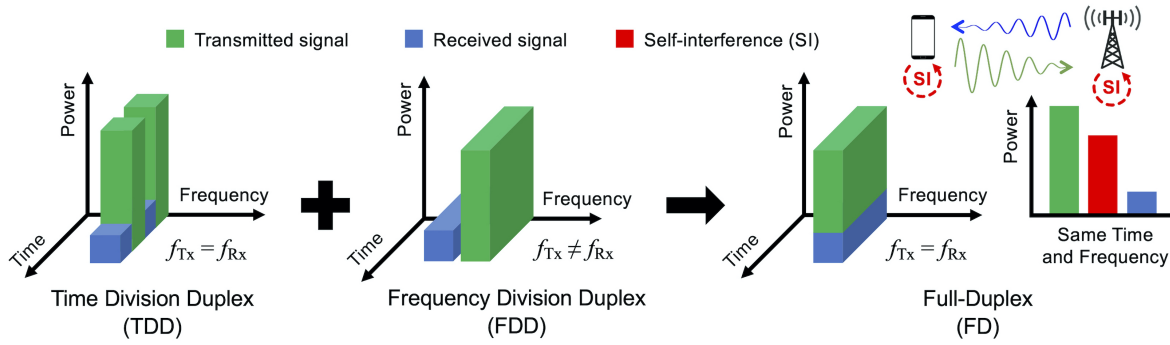
Figure 1.3: Concept of full-duplex (FD) wireless: a radio can simultaneously transmit and receive in the same frequency band by employing self-interference (SI) cancellation.

associated with enabling FD wireless is the tremendous transmitter self-interference (SI) that leaks into the receiver, which is usually at least a billion times ($>90\,\mathrm{dB}$) stronger than the desired signal and requires powerful SI cancellation. Another important technology is multi-antenna systems, such as phased arrays multiple-input-multiple-output (MIMO) systems, which can substantially enhance the communication range through the beamforming, a technique for directional signal transmission or reception utilizing spatial selectivity. Therefore, to achieve further improved data rates and enhanced link distance, it is beneficial to combine FD with multi-antenna systems.
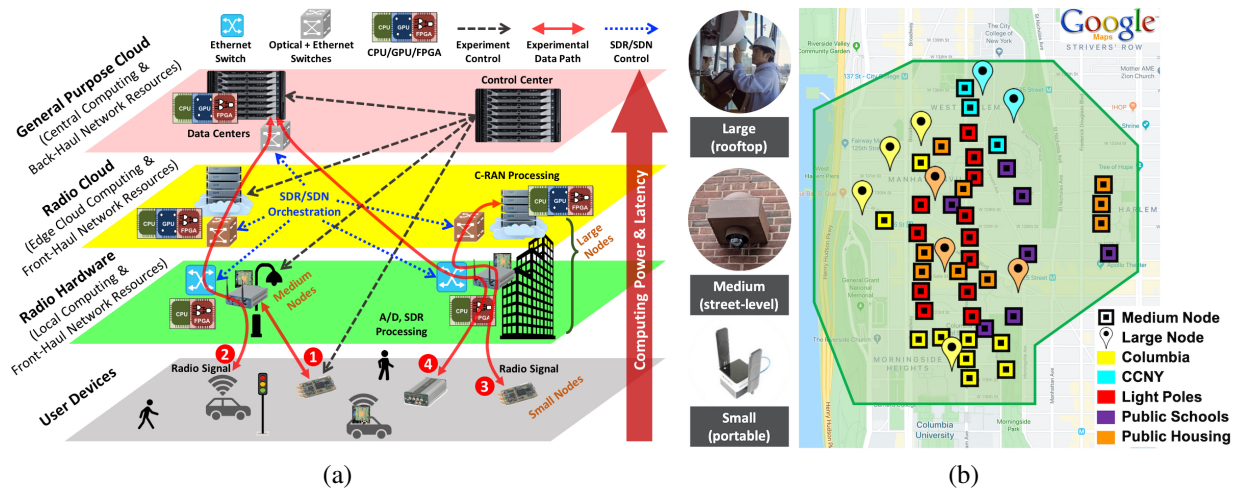
Figure 1.4: (a) COSMOS' multi-layered computing architecture, where paths 1–4 demonstrate the data paths of example experiments with local/remote computing. (b) COSMOS' envisioned deployment area of about 1 sq. mile covering about 15 city blocks and about 5 city avenues, with tentative deployment sites marked on the map.

Evaluation and realization of various emerging wireless technologies and the supported applications would benefit from a large-scale experimental platform. Therefore, we focus on the **C**loud-enhanced **O**pen **S**oftware-defined **MO**bile wireless testbed for city-**S**cale deployment (COSMOS) platform (see Fig. 1.4), which is being deployed in West Harlem, New York City, as part of the NSF Platforms for Advanced Wireless Research (PAWR) program. Open-access and programmability across all technology components (across wireless, optical, and edge cloud domains) and protocol layers (from Physical to applications) in COSMOS enable researchers to remotely explore beyond-5G technologies in a real world environment. In particular, we perform *extensive channel measurements for millimeter-wave (mmWave) communications*, which uses the widely available spectrum at these frequencies to achieve significantly increased data rates. We also focus on *optical-wireless x-haul networking*, where the convergence of wireless technologies with the underlying optical front-/mid-/back-haul (x-haul) networks can benefit various cloud radio access network (C-RAN) applications. These pilot experiments and measurements helped drive and validate the design of the COSMOS testbed, and can inform further deployment and experimentation in the testbed.

In the rest of the chapter, we discuss the contributions made in these three areas and present an overview of the contributions to literature. In Chapter 2, we discuss related work in each area and our unique contributions.

1.2 Contributions

In this section, we describe the contributions of this thesis to the three main research areas: (i) algorithm design and system optimization for ultra-low-power energy harvesting networks that support a wide range of IoT applications (Part I), (ii) the design, implementation, and optimization of compact FD radios and the corresponding gains obtained in FD networks (Part II), and (iii) the experimentation on advanced wireless technologies coupled with an optical x-haul network and edge cloud, as well as 28 GHz channel measurements in the city-scale COSMOS testbed (Part III).
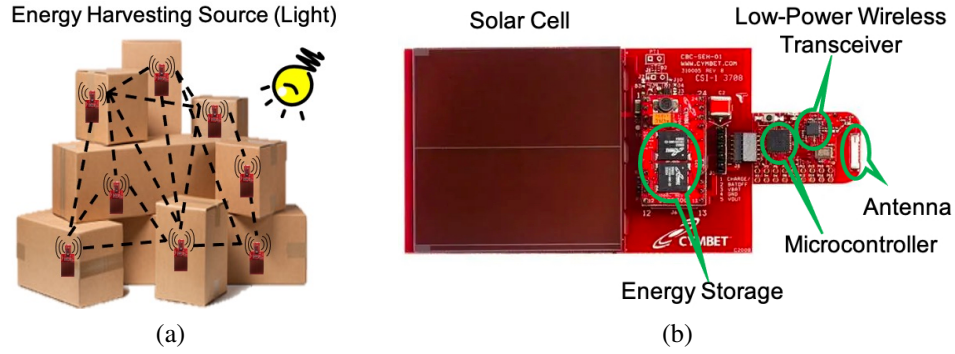
Figure 1.5: (a) Energy harvesting nodes can be attached to everyday Internet-of-Things (IoT) objects (e.g., boxes) and utilize neighbor discovery (ND) protocols for inventory tracking and monitoring applications. (b) The prototype energy harvesting node stores harvested energy in a capacitor and powers the microcontroller and transceiver.

1.2.1 IoT Applications and Ultra-Low-Power Energy Harvesting Networks

In Part I we study algorithm design and system optimization for ultra-low-power energy harvesting networks that support a wide range of IoT applications including object tracking and monitoring, as illustrated in Fig. 1.5(a). This research was conducted within the cross-disciplinary Energy Harvesting Active Networked Tags (EnHANTs) project [10] at Columbia University.

Neighbor Discovery Rate Maximization (Chapter 3): We investigate the specific application of neighbor discovery from a practical perspective where nodes are constrained by limited power budgets. We present the design and analysis of the Power Aware Neighbor Discovery - Asynchronously (Panda) protocol, which maximizes the rate at which nodes discover each other. Nodes running the Panda protocol follow an optimized pattern to transition between different radio states (sleep, listen/receive, and transmit) while incorporating real hardware features and characteristics. We evaluate the Panda protocol using both simulations and a customized testbed composed of ultra-low-power energy harvesting nodes (see Fig. 1.5(b)) and show that Panda outperforms related protocols by up to $3\times$ in terms of the neighbor discovery rate.

Distributed Scheduling and Network Throughput Maximization (Chapter 4): Although Panda outperforms other well-known protocols, it is not clear how efficient it is when compared to the optimal case. In Chapter 4, we further explore this problem and consider a general throughput
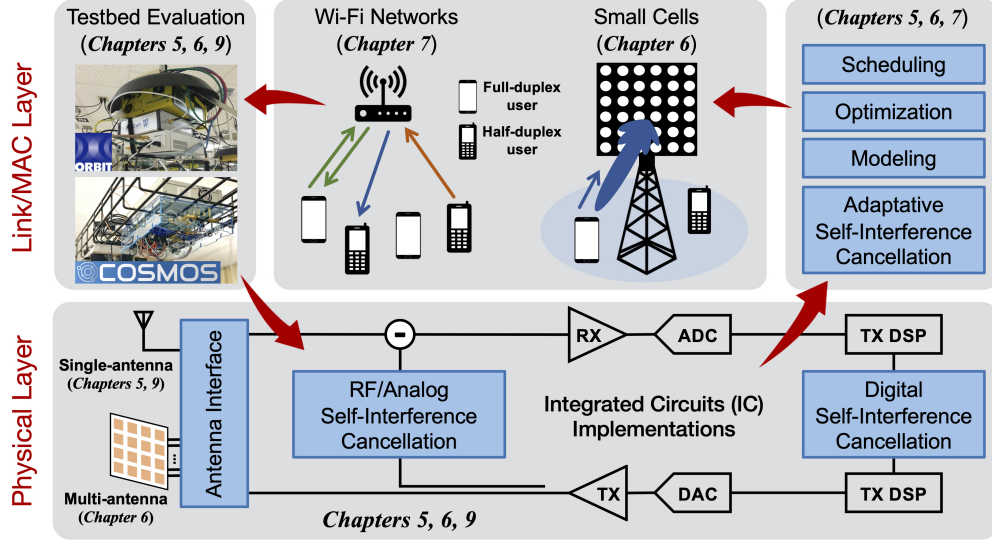
Figure 1.6: Overview of the interdisciplinary Columbia FlexICoN project and the contributions made in Part II of this thesis.

maximization problem among a set of heterogeneous energy-constrained ultra-low-power nodes. Our analysis identifies the theoretical fundamental limits of the rate at which data can be exchanged between these nodes. We then design and present the Energy-constrained Broadcast (EconCast) protocol, which is an efficient, practical, and fully distributed protocol that approaches maximum throughput in the limiting sense. We extensively evaluate the performance of the EconCast protocol in both simulations and real-world experiments, and show that EconCast outperforms Panda and other known protocols by up to an order of magnitude in realistic environments.

1.2.2 Full-Duplex Wireless

In Part II, we focus on the design, implementation, and optimization of compact FD radios integrating SI cancellation techniques grounded in integrated circuit (IC) implementations, and on the interactions between the Physical and the higher layers in FD networks (see Fig. 1.6). This line of research is conducted within the interdisciplinary Full-duplex wireless: From Integrated Circuits to Networks (FlexICoN) project [11] at Columbia University.

Single-Antenna Compact Full-Duplex Radios (Chapter 5): Enabling wideband FD wireless in small-form-factor devices requires novel SI cancellation techniques that are compatible to IC de-

6

signs. To evaluate the performance of IC-based FD radios, developed within the Columbia Flex-ICoN project, we designed and implemented two generations of RF SI cancellers using discrete components on printed circuit boards (PCBs). These cancellers emulate their RFIC counterparts and allow for easier integration and system-level experimentation using software-defined radios. We constructed realistic canceller hardware models and devised efficient optimized canceller configuration schemes for real-time adaptation to environmental changes. In Chapter 5, we present the 2nd-generation FD radios, which leverage the technique of frequency-domain equalization (FDE) to achieve significantly enhanced cancellation performance ($>50\,$dB RF SI cancellation across $20\,$MHz bandwidth) and bandwidth ($>5\times$ improvement over the 1st-generation FD radios). We also integrated this canceller with a CMOS circulator to implement FD radios with high power (e.g., multi-watt) handling.

Multi-Antenna Full-Duplex Phased Arrays (Chapter 6): Another important technology that can significantly increase signal-to-noise ratio (SNR) and communication range is phased arrays employing analog beamforming – directional signal transmission and reception utilizing spatial selectivity. Although combining FD with phased arrays can provide increased data rates, it is challenging due to the large number of SI channels between each pair of antennas, which result in $O(N^2)$ hardware complexity, energy consumption, and powerful SI. In Chapter 6, we develop efficient algorithms for obtaining the optimal transmit and receive analog beamforming weights (a.k.a., beamformers) that can achieve wideband RF SI cancellation with minimal beamforming performance degradation, while improving FD rate gains in small cell networks. The key insight is that in multi-antenna systems, RF SI cancellation can be achieved in the spatial domain without any explicit analog/RF circuits by repurposing beamforming degrees of freedom jointly across transmit and receive elements.

Distributed Scheduling in Heterogeneous Half- and Full-Duplex Networks (Chapter 7): Moving up in the network stack, random access (e.g., WiFi-like) networks with FD capabilities call for a novel design of the medium access control (MAC) layer. This is due to the fact that introducing FD users into legacy half-duplex (HD) networks will result in unfairness to HD users, given

their different abilities in channel contention. To address this challenge, in Chapter 7, we propose a novel low-complexity Hybrid Greedy Maximal Scheduling (H-GMS) algorithm that achieves throughput-optimality in a distributed manner. H-GMS is hybrid in the sense that it combines distributed queue length-based scheduling and centralized greedy scheduling with different levels of centralization. The analytical and simulation results show that H-GMS achieves 5–10× better delay performance and improved fairness between HD and FD users by up to 50% compared with state-of-the-art approaches.

1.2.3   Experimentation and Measurements in the COSMOS Testbed

In Part III, we present the overview of the design and deployment of the city-scale COSMOS testbed, as shown in Fig. 1.4, and focus on the supported experimentation on advanced wireless technologies coupled with an optical x-haul network and edge cloud.

Overview of the PAWR COSMOS Testbed (Chapter 8): Deployed in an area of about 1 sq. mile in West Harlem, NYC, the COSMOS testbed will allow the academic research community as well as the wireless industry to explore in the most demanding real-world environments the technology "sweet spot" of ultra-high bandwidth and ultra-low latency. In Chapter 8, we present an overview of the COSMOS testbed and provide context to the pilot COSMOS experiments on open-access full-duplex wireless converged optical-wireless networking as well as extensive $28\,\mathrm{GHz}$ channel measurements described in Chapters 9–11.

Open-Access Full-Duplex Wireless in the COSMOS Testbed (Chapter 9): In order to allow the broader community to experiment with full-duplex (FD) wireless, in Chapter 9, we present the integration of two generations of FD radios in the open-access ORBIT [12] and COSMOS [13, 14] wireless testbeds. Since interfacing an RFIC canceller with an SDR presents numerous technical challenges, we implemented the RF cancellers on printed circuit boards (PCBs) to facilitate cross-layered experiments with an SDR platform. We first present the integration of a Gen-1 *narrowband* RF canceller with a USRP N210 SDR in the ORBIT testbed [12]. We then present the integration of two Gen-2 *wideband* RF cancellers (an improved version of that presented in Chapter 5) with

USRP2 SDRs in the COSMOS testbed [13, 14]. We also develop open-source programs and experiments, as one of the COSMOS pilot experiments, which can be further extended to more complicated communication networking scenarios. These are the *first open-access FD radios* that allow remote experimental research and researchers from around the world have already started using these resources.

28 GHz Channel Measurements in the COSMOS Testbed Area (Chapter 10): A key technological block of COSMOS is the programmable mmWave radios. Since mmWave radio signals experience higher path loss and are more vulnerable to environmental changes (e.g., blockage and movement) compared to signals at sub-6 GHz, the operation of mmWave networks requires accurate channel models for specific deployment sites and understanding of the system-level performance. In Chapter 10, we present extensive 28 GHz channel measurements (over 24 million power measurements collected from over 1,500 links in 3 different sites) within the COSMOS testbed area, a representative urban street canyon environment. The results show that even in the non-line-of-sight (NLOS) scenario, the measured path gain values are on average 5–10 dB higher than that provided by the 3GPP 38.901 urban canyon NLOS model [15]. Moreover, link SNR values of at least 15 dB can be achieved on all sidewalks up to 194 m link distance. These results are used to make decisions about the COSMOS testbed deployment (i.e., the IBM 28 GHz programmable phased array antenna modules) and can also be used to study resource allocation and scheduling problems in mmWave networks at the higher layers.

Converged Optical-Wireless x-Haul Networking in the COSMOS Testbed (Chapter 11): Another key building block of COSMOS is the programmable optical transport network and its unique capability to emulate different optical front/mid/back-haul (x-haul) networks that can support a broad range of converged optical-wireless/wired and edge cloud applications. To demonstrate this capability, in Chapter 11, we present the design and prototyping of cloud radio access network (C-RAN) and mobile edge cloud functionalities using SDRs and dark fiber infrastructure deployed in New York City as part of the COSMOS optical network. We implement remote processing of digitized radio-over-fiber signals from the wideband FD radio described in Chapter 5, and dynamic

9

optical switching for C-RAN wireless handover through a customized software-defined networking (SDN) controller.

1.3 Contribution to Literature

The research about neighbor discovery rate maximization in energy harvesting networks in Chapter 3 appeared the proceedings of *IEEE INFOCOM'16* [16] and an extended journal version appeared in *IEEE Journal on Selected Areas in Communications* [17]. A demo of the developed Panda protocol was presented at *ACM SenSys'16* [18]. The research results about the design, analysis, and experimental evaluation of EconCast for distributed throughput maximization in ultra-low-power networks appeared in the proceedings of *ACM CoNEXT'16* [19], and the paper received the **Best Paper Award**. An extended journal version appeared in *IEEE/ACM Transactions on Networking* [20] and a U.S. patent was granted [21].

The research results about the design and experimentation of compact wideband FD radios based on the technique of frequency-domain equalization in Chapter 5 appeared in the proceedings of *ACM MobiCom'19* [22] and a *Springer book chapter* [23]. A corresponding poster on this topic was presented at *ACM MobiCom'19* [24] and won the **First Place in the ACM MobiCom 2019 Student Research Competition (SRC) Winner**. The developed RF canceller presented in this chapter was also integrated with a CMOS circulator to implement FD radios with high power (e.g., multi-watt) handling, and the results appeared in *IEEE ISSCC'20* [25] and *IEEE Journal on Solid-State Circuits* [26]. The research related to enabling wideband FD phased array through joint transmit and receive beamforming in Chapter 6 appeared in the proceedings of *ACM Mobi-Hoc'19* [27], and the paper was selected as a **Best Paper Finalist**. The design and implementation of an 8-element FD phased array RFIC appeared in the proceedings of *IEEE RFIC'18* [28]. The design and evaluation of the H-GMS algorithm for distributed scheduling in hybrid HD and FD networks appeared in the proceedings of *IEEE INFOCOM'18* [29] and an extended version appeared in *IEEE/ACM Transactions on Networking* [30]. In addition, papers summarizing the collaborative Columbia FlexICoN project were invited to *IEEE Communications Magazine* [31], *IEEE*

Microwave Magazine [32], *Asilomar'16* [33], and *ACM HotWireless'16* [34]. Demos describing two generations our prototyped FD radios were presented at *IEEE INFOCOM'17* [35] and *ACM MobiHoc'16* [36].

The overview of the design and deployment of the city-scale COSMOS advanced wireless testbed appeared in the proceedings of *ACM MobiCom'20* [13]. The integration of open-access FD radios in the ORBIT and COSMOS testbeds appeared in *ACM WiNTECH'20 Workshop* [37] and in [38], and the demonstrations were presented in [39, 40]. Some of the results on the 28 GHz channel measurements are summarized in *ACM mmNets'19 Workshop* [41] and in [42]. The detailed design of COSMOS' programmable optical network and the pilot experimentation on converged optical-wireless networking appeared in *IEEE/OSA OFC'20a* [43], *IEEE/OSA OFC'20b* [44], and *IEEE/OSA OFC'19* [45], as well as presented in [46].

Chapter 2: Related Work

In this chapter, we provide a brief overview of the related work. We first present an overview of the work related to ultra-low-power energy harvesting networks and neighbor discovery (ND) for IoT applications (Section 2.1). Then, we provide the background and review related work in the area of full-duplex (FD) wireless (Section 2.2). Finally, we describe the work related to the COSMOS testbed and the supported experimentation (Chapter 2.3).

2.1 IoT Applications and Ultra-Low-Power Energy Harvesting Networks

In Part I, we focus on ultra-low-power energy harvesting networks where wireless nodes are constrained by a limited power budget (e.g., indoor ambient light source), and study the specific application of neighbor discovery (ND) and the underlying throughput maximization problem.

ND for low-power wireless networks is a well studied topic (see [6, 47, 48] for a summary), where the ND protocols can be categorized into deterministic (e.g., [49–53]) and probabilistic (e.g., [54, 55]) based on the mode of transition between different radio states (sleep, listen/receive, and transmit). In particular, deterministic protocols select node states based on the "quorum" techniques or using prime numbers. Thus, they are able to guarantee an upper bound on *discovery latency*, while the choice of parameters (e.g., prime numbers) is often limited. On the other hand, nodes running probabilistic protocols transition between different radio states based on a probabilistic approach. It has been shown that the probabilistic protocols can achieve a higher average ND rate, but also suffer from an unbounded discovery latency [48].

There is vast amount of related literature on wireless sensor networks and ND that aims to limit energy consumption and improve energy efficiency. Most of the (deterministic and probabilistic) protocols do not explicitly account for different listen and transmit power consumption levels of

the nodes, or different power budgets [48–52, 55–59]. They mostly use a duty cycle during which nodes sleep to conserve energy and when nodes are simultaneously awake, a pre-determined listen-transmit sequence (e.g., prime number- or quorum-based approaches) with an unalterable power consumption level is used. However, for ultra-low-power nodes constrained by severe power budgets, the appropriate amount of time a node sleeps should explicitly depend on the relative listen and transmit power consumption levels. These prior approaches achieve throughput levels which are much below optimal. Additionally, these protocols often require some explicit coordination (e.g., slotting [52, 55] or exchange of parameters [50]) between the nodes, which are not suitable for emerging ultra-low-power nodes.

The *centralized* probabilistic protocol, Panda, developed in Chapter 3 is fundamentally different: other protocols (i) are constrained by a duty cycle, instead of a power budget, (ii) do not account for channel collisions when multiple nodes transmit at the same time, (iii) rely on nodes maintaining synchronization with each other, or (iv) do not consider practical hardware energy consumption costs (i.e., the power consumed by the node during the transition between different radio states). To the best of our knowledge, *Panda is the first ND protocol tailored for energy harvesting nodes* and the *first attempt to maximize the discovery rate, given a power budget*. As such , nodes running the Panda protocol have duty cycles between 0.1% – 0.6%, which is an order of magnitude lower than those typically considered in prior works [48].

The *distributed* probabilistic protocol, EconCast, developed in Chapter 4 is the *first distributed protocol with provable performance guarantees for achieving the maximum throughput in ultra-low-power networks*. The design of EconCast is inspired by prior work on network utility maximization [60] and queue-based CSMA [61–65]. The problem we consider is unique and novel in two aspects. First, in the past work on CSMA and network utility maximization, nodes or links make decisions based on the relative sizes of queues. Often, a queue is a backlog of data packets to send or the available energy. Prior work that considers the latter (e.g., [66]) uses the energy only for transmission, while listening is "free", which is a very different paradigm than the scenario considered in this chapter. Second, in our setting, the queue "backlogs" energy but there

13

is no clear mapping assumed in previous work between successful packet transmissions and the nodes' energy consumptions. A node's listen or transmit events will relieve the backlog, but *do not* increase utility (throughput) unless other nodes are appropriately configured (i.e., transmitting when no listening nodes exist or listening when no transmitting nodes exist does not increase the throughput). This coordination of state among nodes to utilize their energy makes the considered problem more challenging.

In recent years, hardware platforms have been develop for energy harvesting nodes [67–69], computational RFIDs [70], and mm^3-scale wireless devices [71]. Additionally, there are other radio features that achieve low energy consumption in wireless sensor networks such as preamble-sampling and wake up radios [72, 73]. However, the added power consumption of these features makes them impractical for the energy harvesting nodes we study. Furthermore, numerous options for low-power wireless communication exist (e.g., Bluetooth Low Energy [74]). In the experimental evaluations of Panda and EconCast, we use the TI eZ430-RF2500-SEH energy harvesting nodes [75], which is an emerging low-power energy harvesting platform with seamless support of the development of wireless protocols.

2.2 Full-Duplex (FD) Wireless

In Part II, we focus on cross-layering in FD wireless networks from the design of single- and multi-antenna FD radios in the Physical layer to distributed scheduling in the medium access control (MAC) layer (see Fig. 1.6). Extensive research related to FD wireless is summarized in [8, 76], including implementations of FD radios and systems, analysis of rate gains, and resource allocation in the higher layers of the networking stack.

RF Canceller and Single-/Multi-antenna FD Radios. RF self-interference cancellation (SIC) typically involves two stages: (i) isolation at the antenna interface, and (ii) SIC in the RF domain using cancellation circuitry. While a separate Tx/Rx antenna pair can provide good isolation and can be used to achieve cancellation [77–80], a shared antenna interface such as a circulator is more appropriate for single-antenna implementations [81, 82] and is compatible with FD multi-

antenna systems. Existing designs of analog/RF SIC circuitry are mostly based on a time-domain interpolation approach [81, 83], where real delay lines with different lengths and amplitude/phase controls [81, 83] are used emulate the SI channel. This essentially represents an RF implementation of a finite impulse response (FIR) filter. Based on the same RF SIC approach, several FD MIMO radio designs are presented [80, 84–87]. FD relays have also been successfully demonstrated in [88, 89]. Moreover, recent work has demonstrated the feasibility of achieving SIC via digital/analog beamforming in FD massive-antenna systems [90, 91].

Rate Gain and Scheduling in FD Networks. Related literature has focused on characterizing the capacity region and rate gains in FD network, as well as developing resource allocation algorithms under both perfect [92, 93] and imperfect SIC [94–96]. Similar problems have also been considered in FD multi-antenna/MIMO systems [90, 97, 98]. Network-level FD gain is analyzed in [77, 99–101] and experimentally evaluated in [79, 102] in *homogeneous*, where all the users are half-duplex (HD) or FD. Recent studies in the MAC layer has also focused on *homogeneous* networks [98, 99, 103–109], where signal-to-noise ratio (SNR) or a specific standard (e.g., IEEE 802.11 where an FD topic interest group was recently formed [110]) is considered. FD can also facilitate different applications including improved Physical layer security [111–113], localization [114], and neighbor discovery [115].

The techniques utilized in prior FD radio designs are incompatible with IC implementations, which are required for hand-held and form-factor-constrained devices. The wideband FD radio presented in Chapter 5 utilizes the technique of frequency-domain equalization (FDE) based on a previous work on the design of a wideband RFIC canceller in the FlexICoN project. To the best of our knowledge, this is the *first thorough study of wideband RF SIC achieved via a frequency-domain-based approach (which is suitable for compact implementations) that is grounded in real-world implementation and includes extensive system- and network-level experimentation.*

To enable FD operation in phased arrays, in Chapter 6, we propose to achieve wideband RF SIC in FD phased arrays via joint Tx and Rx beamforming, which results in lower hardware complexity and power consumption. Essentially, the idea is to repurpose the spatial degrees of freedom that

inherently exist in multi-antenna systems. The proposed approach is evaluated using both measurements and a customized 8-element FD phased array RFIC. This is the *first fundamental study of an FD phased array employing joint transmit and receive beamforming to achieve wideband RF SIC and FD rate gains*. We believe that the proposed approach is beneficial to other problems related to multi-antenna systems applying different beamforming techniques.

The H-GMS algorithm presented in Chapter 7 is *a novel distributed scheduling algorithm developed for heterogeneous network consisting of both HD and FD users*. H-GMS is proven to be throughput-optimal in heterogeneous networks and achieves 16–30× better delay performance compared with existing algorithms. We also, for the first time, consider and demonstrate the fairness between users that have different HD/FD capabilities, and show that H-GMS improves the fairness by up to 50% compared with the fully distributed Q-CSMA algorithm.

2.3   Experimentation and Measurements in the COSMOS Testbed

In Part III, we present an overview of the city-scale COSMOS advanced wireless testbed, and focus on the experimentation and measurements across the wireless and optical domains.

Field-Deployed Testbeds for Wireless and Optical Research. Several testbeds with multi-technology capabilities aimed to address the new challenges posed by 5G and future wireless networks have been recently developed and deployed. For example, Bristol Is Open [116] is a programmable testbed deployed in Bristol, UK, that provides a smart city development and research platform for wireless network connectivity, IoT hosting, and cloud computing. The ADRENALINE [117] testbed deployed in the Catalania region, Spain, is a circuit-switched optical testbed designed for experimental research on large-scale optical transport networks and distributed edge computing. POWDER-RENEW [118] is another advanced wireless testbed that shares some design/operation considerations with COSMOS, and is currently under development. It focuses on massive MIMO wireless and will support wireless experiments in Salt Lake City by using the cloud computing resources provided by CloudLab [119].

In comparison, as described in Chapter 8, *COSMOS is deployed in a densely populated urban*

environment in West Harlem (NYC), with unique technological emphasis on mmWave wireless, converged optical-wireless x-haul networking, and edge cloud, as well as full programmability across the optical, wireless, and application domains. In addition, the availability of dark fiber in NYC and recent advances in disaggregated optical components have enabled COSMOS to unique capability in the real-world environment with similar research capable radio and edge cloud computing at scale.

Open-Access FD Wireless. Related work in the area of FD wireless has been discussed in Section 2.2. In Chapter 9, we present the hardware and software development of two generations of FD radios and their integration in the ORBIT and COSMOS testbeds. In particular, the Gen-2 wideband FD radios integrated in the COSMOS testbed are based on the work described in Chapter 5. These are *the first open-access, remotely-accessible FD radios* that can facilitate the research in this area.

mmWave Channel Measurements. Various channel measurement campaigns have been conducted for different mmWave frequencies (e.g., 28/38/73 GHz) in urban [120–123] and suburban [120, 124–126] environments. Both directional horn antennas on mechanical steering platforms [122, 124, 127] and phased arrays applying beam steering [127–130] were considered. With the support of wideband channel sounders, these campaigns can provide measurements of not only the path gain, but also other metrics such as the power delay profile. However, most of these campaigns consider limited number of measurement links (usually at the order of 10s of links) with different Tx and Rx locations. Recent work also demonstrates link-level channel measurements using commercial 802.11ad devices [131, 132].

In our measurement campaign presented in Chapter 10, we use a narrowband 28 GHz channel sounder that can uniquely support (i) fast recording of received signal power on the entire 360° azimuth plane (and thereby the effective azimuth beamforming gain), and (ii) easy and continuous measurements on relatively long streets with fine-grained measurement step sizes. Hence, our measurements in the COSMOS testbed area focuses on accurately characterizing the 28 GHz mmWave channel *in a dense urban canyon environment along individual streets* with small link

17

distance step sizes. This results in extensive amount of collected measurement data.

Converged Optical-Wireless x-Haul Networking. Using optical networks to facilitate wireless HO has been investigated before. For example, a fixed PON ring network was proposed in [133] to reduce the HO latency in 4G LTE networks. A fast inter-base station ring architecture with millimeter-wave cellular networks was proposed in [134] to facilitate frequent HOs in 5G networks while satisfying the stringent requirements of high reliability and low latency. PONs, that are not based on a ring topology, can also help tackle frequent HOs. A concept of dynamic wavelength-specific virtual PONs was proposed in [135], where RRHs using the same radio frequency are merged into groups, and if user equipment (UE) moves between cells at the same radio frequency, HO will not be triggered and instead, PON time-division multiplexing features are used. A key element for each of [133–135] is the ability to coordinate optical switching operations in response to wireless HOs. Each of these methods could benefit from city-scale experimental evaluation in COSMOS. In Chapter 11, we describe the architecture design and deployment of COSMOS' programmable optical network, and present *a unique pilot experiment on wireless HO via dynamic optical switching, which integrates COSMOS' optical and radio resources (i.e., SDRs).*

Part I

IoT Applications and Ultra-Low-Power Energy Harvesting Networks

Chapter 3: Panda: Neighbor Discovery on a Power Harvesting Budget

Energy harvesting and ultra-low-power communications are both emerging techniques that can enable various tracking and monitoring applications in IoT. As described in Chapter 1, the application of neighbor discovery (ND) has always been an important part of many network scenarios [47, 48]. However, to consume power at the rate of power harvested, energy harvesting nodes require *extremely limited power budgets*: for example, even with optimized power spending, the duty cycles are between $0.1\% - 0.6\%$. Therefore, numerous assumptions from related work (see Chapter 2) no longer hold, including that switching times (between the sleep, listen/receive, and transmit states) draw negligible power and that the power costs to send and receive are identical.

In this chapter, we develop an ND protocol tailored for commercial off-the-shelf (COTS) energy harvesting nodes, based on the TI eZ430-RF2500-SEH prototype [75] (shown in Fig. 1.5(b)). The nodes harvest ambient light to supply energy to a low-power microcontroller and wireless transceiver. To maintain *perpetual* tracking of the (potentially) mobile objects, ND must be run continuously with the node operating in an ultra-low-power mode that consumes power at the rate of power harvested [136]. Hence, our objective is *to maximize the rate in which nodes discover their neighbors, given a constrained power budget at each energy harvesting node*. In particular, we design, analyze, and experiment with Panda – Power Aware Neighbor Discovery Asynchronously,[1] an ND protocol that *maximizes the average discovery rate under a given power budget*.

We first model a generic ultra-low-power energy harvesting node that captures the capabilities of our prototype and study, for the first time, important radio characterizations in the context of ND (Section 3.1). We then present the design and analysis of the Panda protocol in which an

---

[1]The protocol name, Panda, relates to the animal as both energy harvesting nodes and Pandas spend the majority of their time sleeping to conserve energy.

20

Component	Problem/Modification
Energy Storage	On-board battery cannot be monitored; disable on-board battery and replace with an external capacitor.
Solar Panel	On-board solar cell cannot be monitored; disable on-board cell, measure power harvested by connecting a ammeter in series with solar cell from [137].
Power Consumption	Unable to track power consumed; measure consumed power with an oscilloscope across a 10 ohm sense resistor, placed in series with the transceiver and the microcontroller.
12 kHz Clock Source	Clock frequency varies by up to 20% for each node; rectified by manually measuring/calibrating the number of clock ticks in one second for each device.

Table 3.1: Modifications to the TI eZ430-RF2500.

energy harvesting node discovers its neighbors by transitioning between the sleep, listen/receive, and transmit states at optimized rates while maintaining within its power budget (Sections 3.2 and 3.3). Furthermore, we present Panda-Dynamic (Panda-D), which extends Panda's applicability to non-homogeneous power harvesting and multihop topologies (Section 3.4).

Finally, we experimentally evaluate the performance of Panda using the eZ430-RF2500-SEH prototype and show that the real-life discovery rates are within 2% of the analytically predicted values (Section 3.5). Moreover, we show that Panda's experimental discovery rate is up to $3\times$ higher than the discovery rates from simulations of two best known low-power ND protocols [50, 55]. We also demonstrate that Panda-D adjusts the rate of ND for scenarios with non-homogenous power harvesting and multihop topologies.

This chapter's contributions were previously presented in [16, 17] and the work is in collaboration with Robert Margolies and Dr. Guy Grebla. Undergraduate students Gregory Chen and Saahil Jain contributed to the development of the Panda demonstration [18].

3.1 Ultra-Low-Power Energy Harvesting Node Prototype and System Model

3.1.1 Prototype Description

The prototype is shown in Fig. 1.5(b) and is based on the commercially available TI eZ430-RF2500-SEH [75]. We made some modifications to the hardware as summarized in Table 3.1. The prototype include the following main components:

Energy Harvesting Power Source: The prototype harvests light from a Sanyo AM 1815 amorphous solar cell [137]. The solar cell is set to a fixed harvesting voltage of 1.02 V (no power point tracking techniques are used). To measure the power harvested, we place an ammeter in series with the solar cell.

Energy Storage: The energy harvested by the solar cells is stored in a capacitor and the voltage is denoted by V_{cap}. The voltage is regulated to 3.5 V to power the node. We modified the board design to enable experimentation with varying capacitor sizes.

In practice, a node can withstand variability in the energy harvesting and consumption, so long as the storage is sufficiently sized. There are numerous related works discussing the impact of finite energy storage sizes (e.g., [138]) that are out of the scope of this work. Thus, unless stated otherwise, we use a 30 mF capacitor. To ensure stable voltage regulation, a software cutoff is imposed; if $V_{cap} \leq 3.6$ V, the node enters and remains in a low-power sleep state until enough power is harvested such that V_{cap} exceeds the cutoff.

Low-Power Microcontroller: A TI-MSP430 microcontroller [139] is used to provide computational capabilities. These include (i) sampling the capacitor voltage using an analog to digital converter (ADC), (ii) operating a low-power 12 kHz clock with an idle power draw of 1.6 uW to instruct the node to enter and exit an ultra-low-power sleep state, and (iii) receiving and sending messages to the radio layer.

Low-Power Transceiver: The prototype utilizes a CC2500 wireless transceiver (a 2.4 GHz transceiver designed to provide low-power wireless communication) [140] to send and receive messages. The transceiver operates at 250 kbps and consumes 64.85 mW while in receive state. The transmission power can be set in software and we utilize levels between -16 dBm and $+1$ dBm, with a resulting power consumption between 53.25 mW and 86.82 mW. At these levels, nodes within the same room typically have little or no packet loss.
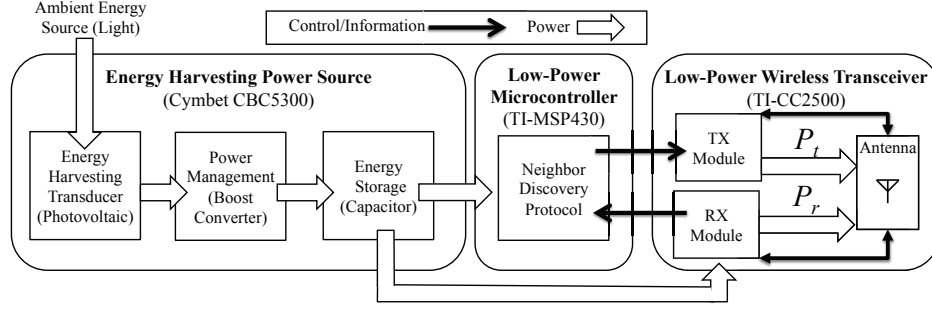
Figure 3.1: The hardware model of the energy harvesting node based on the TI eZ430-RF2500-SEH prototype [75].

V_{cap}	The voltage of the capacitor (V)
N	Number of nodes
P_b	Average power spending budget (mW)
P_t	Transmitting power consumption (mW)
P_r	Listening/Receiving power consumption (mW)
C_{ij}	Energy cost to transition from state i to j (uJ)
M	Discovery-packet duration (ms)
λ	Rate of exponential distribution (ms^{-1})
l	The duration of the listening period (ms)
ρ	Expected renewal duration (ms)
Y	Denotes role of node in the renewal
$\eta(\cdot)$	Expected energy spending (uJ)
$\Phi(\cdot)$	Expected power spending in a renewal (mW)
χ	Expected duration of idle listening (ms)
U	Discovery rate (s^{-1})

Table 3.2: Nomenclature.

3.1.2 System Model

The model is based on the prototype and is shown pictorially in Fig. 3.1. Yet, it is generalizable to a class of other prototypes (e.g., [69]). Table 3.2 shows a summary of the nomenclature.

A node can be in one out of three states, denoted by the set $\mathcal{S} = \{s, r, t\}$ for sleep (s), receive[2] (r), and transmit (t). A node in state $i \in \mathcal{S}$ consumes power of P_i. Since the power consumption in sleep state is negligible, we assume $P_s = 0$ throughout the chapter and remark that all results can be easily applied for $P_s > 0$, as described in Appendix 3.7.1. For the power budgets we consider,

---

[2]We refer to the *receive* and the *listen* states synonymously as the power consumption of the prototype in both states is similar.

the energy consumed by the radio to transition between different states is non-negligible. Hence, we denote by C_{ij} the energy (uJ) consumed to switch from state i to state j ($i, j \in \mathcal{S}$).

Unfortunately, the prototype does not have explicit power awareness (unlike, e.g., [69]). Therefore, we impose a power budget, P_b (mW) on each energy harvesting node. The power budget is set such that *energy neutrality* is achieved: nodes consume power (on average) at the power harvesting rate [141]. Hence, for an energy harvesting node harvesting more power (e.g., brighter light source), P_b is higher.

We denote by N the number of nodes in the network and present two important definitions:

Definition 3.1 *The discovery message is a broadcast packet containing the ID of the transmitter.[3] A discovery occurs when a node receives a discovery message from a neighbor. Multiple discoveries can occur per discovery message transmission.*

Definition 3.2 *The discovery rate, denoted by U, is the expected number of discoveries in the network per second.*

The objective of the ND protocol is to *maximize the discovery rate, subject to a given power budget*. This is in contrast to other works which seek to minimize the worst case discovery latency [50, 52], subject to a duty cycle. The discovery latency, or time between successive discoveries, is very much related to the discovery rate. In fact, the inverse of the average discovery rate is indeed the average discovery latency. Thereby, maximizing the average discovery rate is quite similar to minimizing the worst case latency.

In practice, both the average discovery rate and the worst-case latency are important to applications [48]. Reduced worst-case latency is important in cases where nodes are only collocated for short periods of time. However, in general, applications (such as the one described in Section 1.1) must be designed to handle the occasional missed-discovery and would receive greater benefit from a higher average discovery rate. As such, we focus on maximizing the average discovery rate and in Section 3.5, we also consider the discovery latency as a secondary performance metric.

---

[3]In practice, the discovery message may include information on already discovered neighbors, thus enabling *indirect* discoveries. However, we do not consider these indirect discoveries.
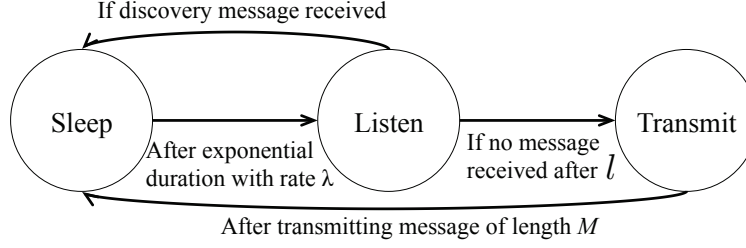
Figure 3.2: Panda protocol outline: energy harvesting nodes transition between different radio states (sleep, listen, and transmit) to maintain within a power budget.

3.2 Design and Analysis of Panda

3.2.1 Panda Protocol Description

Fig. 3.2 depicts the state transition diagram for the Panda protocol, from sleep to listen to transmit and then back to sleep. To ensure perpetual operation under the power budget P_b (mW), nodes initialize in a low-power sleep state to conserve energy. To maximize the discovery rate, Panda follows a probabilistic approach in which nodes sleep for an exponential duration with rate λ (ms^{-1}), which prevents unwanted synchronization among subsets of nodes.

Following sleep, nodes awaken and listen to the channel for discovery messages from their neighbors for a fixed duration of l (ms). If a message is received, the node remains in the listen state until it completes reception of this message. If no transmission is heard while in the listen state, the node transmits its discovery message of fixed duration M (ms).[4]

Note that in Panda, similar to CSMA, nodes always listen before they transmit. Therefore, a node's transmission will never collide with an ongoing transmission from a node that is within wireless communication range. Additionally, after a message is transmitted, the node returns to the sleep state. Hence, there is no acknowledgment of the discovery. This is because coordinating acknowledgment messages among multiple potential receivers can be costly, requiring additional listening by the transmitter and possibly collision resolution (e.g., [142]).

---

[4]The discovery message duration, M, is fixed, stemming from the fixed size of the node ID contained in the message.
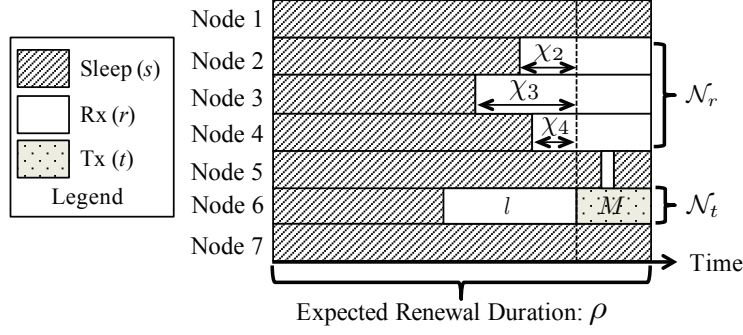
Figure 3.3: Renewal process representing one renewal cycle for $N = 7$ nodes: all nodes begin in a memoryless sleep state and the renewal restarts after the first node to wake up completes its transmission (node 6 in this example).

3.2.2 Analysis

While Panda can operate in general scenarios, for analytical tractability, we assume the following: **(A1)** All nodes are homogeneous, namely, have the same power budget P_b and the same hardware; **(A2)** Every pair of nodes can exchange packets (clique topology) with no packet errors due to noise; and **(A3)** The number of nodes, N, is known a priori.

These assumptions are applicable to some systems and envisioned applications. For example, when tracking boxes in a room (Fig. 1.5(a)), these assumptions are close to reality as nodes in close proximity harvest similar amounts of energy, have few packets lost, and the number of nodes can be estimated a priori. However, for scenarios in which these assumptions do not hold, in Section 3.4, we present Panda-D which is based on relaxed assumptions and discuss the implications.

We note that, as this is the first attempt to develop an ND protocol explicitly for energy harvesting nodes, it is natural to consider the homogeneity assumption (A1). Additionally, several other works also make assumptions similar to (A2) and (A3) (e.g., [50, 52] only consider ND for a link, $N = 2$, with no collisions or packet loss).

Using these assumptions, we now use techniques from renewal theory [143] to analyze Panda for a network of N nodes. The renewal process is shown pictorially in Fig. 3.3. The renewal initiates with all nodes in the sleep state and ends after one node completes its transmission, whether the message is heard or not. The sleep duration for each node follows a memoryless exponential

distribution. Therefore, for all analytical purposes, all nodes effectively initiate their sleep state at the start of the renewal.

In each renewal, the *first* node to wake up begins its listen state, and after a duration l, it transmits its discovery message. This is exemplified by node 6 in Fig. 3.3; we denote by $\mathcal{N}_t$ the set containing a single transmitting node in a renewal.

Nodes that are in the receive state (r) when a message transmission begins, will stay in this state until the transmission is completed and then switch to the sleep state (s). We denote by $\mathcal{N}_r$ the set of such nodes and $|\mathcal{N}_r|$ the size of the set, exemplified by nodes 2–4 in Fig. 3.3. The expected idle listening time of a node in $\mathcal{N}_r$ is denoted by χ. Fig. 3.3 shows examples of idle listening durations for nodes 2–4, denoted as χ_i. Any node which wakes up in the middle of the message transmission immediately senses the busy channel and returns to the sleep state. An example is node 5 in Fig. 3.3.

When the transmission is completed, all nodes are in sleep state and the renewal restarts. The average renewal duration is the time it takes for the first node to wake up (occurring with rate $N\lambda$), listen for a duration l, and transmit a message for a duration of M. Hence, the expected renewal duration ρ is:

$$\rho = 1/(\lambda N) + l + M. \tag{3.1}$$

3.2.3 Discovery Rate

Recall that the objective of Panda is to maximize the discovery rate, U (see Definition 3.2). Considering U as the reward function and applying the elementary renewal theorem for renewal-reward processes [143], we obtain:

$$U := \lim_{t \to \infty} \frac{u(t)}{t} = \frac{\mathbb{E}[|\mathcal{N}_r|]}{\rho}, \tag{3.2}$$

where $u(t)$ represents the number of discoveries (as defined by Definition 3.1) by time t and ρ is computed by (3.1).

There are $(N-1)$ nodes who are not the transmitter in the renewal, each of which is equally

and independently likely to discover the transmitter. A discovery occurs if the node wakes up from sleep within a period of time l after the transmitting node ($\mathcal{N}_t$) wakes up, an event with probability $(1 - e^{-\lambda l})$. Hence,

$$\mathbb{E}[|\mathcal{N}_r|] = (N - 1)(1 - e^{-\lambda l}). \tag{3.3}$$

3.2.4 Energy Consumption

Since all nodes are homogenous (A1), we let n denote an arbitrary node and define a random variable Y that indicates the set ($\mathcal{N}_t$, $\mathcal{N}_r$) in which the node resides in the renewal:

$$Y = 0 \text{ if } n \in \mathcal{N}_t; \ Y = 1 \text{ if } n \in \mathcal{N}_r; \ Y = 2 \text{ otherwise.} \tag{3.4}$$

The expected power consumed (mW) for a node in which $Y = y$ is denoted as $\Phi(y)$. It is computed as, $\Phi(y) = \Pr\{Y = y\}\eta(y)/\rho$, where $\eta(y)$ represents the expected amount of energy (uJ) consumed by a node in a renewal in which $Y = y$ and ρ is the expected renewal duration. The expected power consumed in a renewal must meet the power budget, and thus, $\Phi(0) + \Phi(1) + \Phi(2) \leq P_b$. The remainder of this section is used to derive $\Phi(0)$, $\Phi(1)$, and $\Phi(2)$. We will often refer to $\Phi(1)$ as the *discovery power*.

$$\eta(0) = C_{sr} + P_r l + P_t \cdot M + C_{ts}, \tag{3.5}$$

$$\Pr\{Y = 0\} = 1/N. \tag{3.6}$$

(3.5) defines the energy consumption of the transmitting node, which consumes energy to wake up from sleep (C_{sr}), listen for a period of l, transmit a message of length M, and then return to sleep (C_{ts}). By definition of the renewal, there will be exactly one transmitter in a renewal and due to assumption (A1), $\Pr\{Y = 0\} = 1/N$.

$$\eta(1) = C_{sr} + P_r(\chi + M) + C_{rs}, \tag{3.7}$$

$$\Pr\{Y = 1\} = \frac{N - 1}{N}(1 - e^{-\lambda l}). \tag{3.8}$$

For a receiving node, the expected energy consumption is defined in (3.7). A receiving node

28

consumes energy to wake up from sleep (C_{sr}), idle listen before the message transmission (for a duration of χ), and receive for the duration of the message M, and then return to sleep (C_{rs}). As shown in Fig. 3.3, χ denotes the expected duration of idle listening before receiving a message. We derive it without loss of generality, by assuming the transmitter in a renewal ($\mathcal{N}_t$, e.g., node 6 from Fig. 3.3) enters the listen state at $t = 0$, and at $t = l$, it transmits the discovery message. Let x denote the idle listening time for a given node where x is exponentially distributed with $0 < x < l$. We look to find $\chi = \mathbb{E}\left[x|x < l\right]$, i.e.,

$$\chi = \int_{t=0}^{+\infty} \Pr\{x > t | x < l\}\,\mathrm{d}t = \int_{t=0}^{l} \frac{(1 - e^{-\lambda l}) - (1 - e^{-\lambda t})}{1 - e^{-\lambda l}}\,\mathrm{d}t = \frac{1}{\lambda} - \frac{l e^{-\lambda l}}{1 - e^{-\lambda l}}.$$

(3.8) defines the likelihood of $n \in \mathcal{N}_r$. Conditioned on $n \notin \mathcal{N}_t$ w.p. $(\frac{N-1}{N})$, a node successfully receives the message, if it starts listening in a period of length l preceding the transmission. Ssince the sleep duration is exponentially distributed, this is an event occurring with probability, $(1 - e^{-\lambda l})$.

Throughout this chapter, we assume that nodes which sleep for the entire renewal (e.g., nodes 1 and 7 in Fig. 3.3), and those which wake up briefly and sense a busy channel (e.g., node 5), do not consume power, and thus $\eta(2) = 0$. In Appendix 3.7.1, we show how it can be relaxed.

3.3 Optimization of Panda

Clearly, the choice of the sleep rate (λ) and the listen duration (l) determines the power consumption of the node as well as the discovery rate U. First, we demonstrate that an analytical solution is difficult to obtain. Next, we describe the Panda Configuration Algorithm (PCA) which obtains the *configuration parameters* (λ, l) for Panda. Finally, we demonstrate that the PCA obtains a nearly-optimal discovery rate.

Algorithm 3.1 Panda Configuration Algorithm (PCA)
_____
1: **for** $K = [0, \epsilon, 2\epsilon, \ldots, \lfloor \frac{\rho_{\max}}{\epsilon} \rfloor \epsilon]$ **do**
2:      Find $(\lambda,\ l)$ that maximize (3.12) subject to (3.14)
3:      **if** (λ, l) satisfies (3.10) **then**
4:          Compute the discovery rate U
5:      **end if**
6: **end for**
7: **return** (λ, l) that maximize U, denoted as λ_A, l_A, and U_A.
_____

3.3.1 Problem Formulation and Preliminaries

Finding $(\lambda^\star,\ l^\star)$ that maximizes U is formulated as follows:

$$\max_{\lambda, l} : \quad U = (N-1)(1 - e^{-\lambda l})/\rho \tag{3.9}$$

$$\text{subject to} : \quad \Phi(0) + \Phi(1) \le P_b, \tag{3.10}$$

where (3.9) is derived using (3.2) and (3.3). Recall that ρ is computed from (3.1) and $\Phi(y)$ is computed using the results from Section 3.2.4. The problem as formulated above is non-convex[5] and non-linear, and is thereby challenging to solve.

In the following subsections, we will attempt to find nearly-optimal Panda configuration parameters (λ, l). We now provide several observations on the specific structure of the problem which are used throughout this section. First, the following Taylor-series approximation is useful:

$$e^{-x} \ge 1 - x \text{ for } x \ge 0, \text{ and } e^{-x} \approx 1 - x \text{ for } x \approx 0. \tag{3.11}$$

We substitute x with λl in (3.11),[6]

$$U \le (N-1)\lambda l/\rho := \overline{U}. \tag{3.12}$$

_____
[5]The non-convexity of the optimization problem is easy to prove by taking second order partial derivatives.
[6]Limited power budgets cause energy harvesting nodes to be in the sleep state much longer than in the listen state. Thus, $\lambda l \approx 0$ and (3.11) is a good approximation.

3.3.2 Panda Configuration Algorithm (PCA)

The Panda Configuration Algorithm (PCA) returns a configuration of λ and l that satisfy (3.10). To find a configuration with the highest discovery rate, the PCA utilizes a relaxed problem formulation as follows. An upper bound on the discovery power, $\overline{\Phi}(1)$, is computed by using (3.11) to obtain $(1 - e^{-\lambda l}) \leq \lambda l$, which leads to,

$$\Phi(1) \leq \overline{\Phi}(1) := \frac{N-1}{N\rho} \cdot \lambda l \cdot \left(P_r(\chi + M) + C_{sr} + C_{rs}\right). \tag{3.13}$$

The relaxed power budget constraint is then,

$$\text{subject to}: \quad \Phi(0) + \overline{\Phi}(1) \leq P_b. \tag{3.14}$$

The PCA analytically computes the values of (λ, l) that maximize $\overline{U}$ by solving for λ in terms of l in (3.13), and then finding the critical points where $\mathrm{d}\overline{U}/\mathrm{d}l = 0$. For computation tractability, the PCA replaces χ with a constant K in $\overline{\Phi}(1)$. The PCA uses the fact that, in practice, a node's sleep time is upper bounded, introducing an upper bound on the renewal duration $\rho_{\max}$. Thereby, the PCA sweeps values between $0 \leq \chi \leq \rho_{\max}$, and returns the best solution (i.e., the one that maximizes U). We denote the discovery rate that the PCA obtains by U_A and the configuration parameters by (λ_A, l_A).

Fig. 3.4 demonstrates the discovery rate of Panda for various power budgets and number of nodes. The performance of PCA is compared to the discovery rate provided by Monte Carlo solution to (3.9), denoted as $U^\star$.[7] In all test cases considered in Section 3.5 (see Table 3.5), the PCA was within 0.25% of the discovery rate of the Monte Carlo simulation demonstrating the near-optimality of the PCA for the parameters we consider.

---

[7]The Monte Carlo solution generates over 10^8 random configuration parameters and returns the (λ, l) that satisfies (3.10) with the largest ND rate.
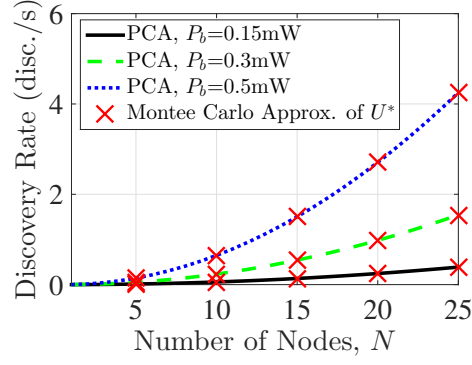
Figure 3.4: The performance of the PCA for varying number of nodes, N: the discovery rates resulting from the PCA (U_A) compared to a Monte Carlo approximation of $U^\star$.

3.4 Panda-Dynamic (Panda-D)

As described in Section 3.1, the objective of the ND protocol is to maximize the discovery rate, subject to a power budget. Panda is analyzed and optimized assuming that nodes are homogenous (A1), are arranged in a clique (A2), and the number of nodes N is known a priori (A3). However, when these assumptions do not hold, the expected power consumption of a node operating with Panda (see Section 3.2.4) will vary and the power budget is no longer satisfied. Therefore, in this section, we present Panda-Dynamic (Panda-D).

Panda-D attempts to maximize the discovery rate by operating with the same behavior as Panda, transitioning between the sleep, receive, and transmit states. However, to achieve energy neutrality in the general setting with the relaxed assumptions, the rate of the exponential sleep duration is dynamic, and is adapted based on the voltage of the capacitor,[8] V_{cap}. Thereby, if a node consumes too much power, its voltage will decrease and it will adapt by staying in the sleep state for longer durations.

Formally, the configuration parameters for Panda-D are computed as follows. In this case, P_b represents an *estimated* power budget for each node, yet we allow for each node to harvest power at varying rates around P_b. The sleep duration is scaled such that the nodes' anticipated power consumption is 0.01 mW when $V_{\text{cap}} = 3.6$ V, and is P_b when $V_{\text{cap}} = 3.8$ V. From the two points,

---

[8]A similar adaptation mechanism was also proposed in [141].

the desired power consumption, P_{des}, is computed as a linear function of V_{cap},

$$P_{\text{des}}(V_{\text{cap}}) = \frac{P_b - 0.01}{3.8 - 3.6} \cdot (V_{\text{cap}} - 3.6) + 0.01,\ 3.6 \le V_{\text{cap}} \le 4.$$

Based on the desired power consumption P_{des}, a node adjusts its sleep duration. As mentioned above, we cannot explicitly relate the sleep duration to the power consumption for a node. Instead, we estimate the power consumption by ignoring the *discovery power*. That is, we assume that a node always follows the sleep, receive, transmit cycle and is spending on average at rate,

$$P_{\text{est}} = \frac{\eta(0)}{1/\lambda + l + M} = \frac{P_r l + P_t M + C_{sr} + C_{ts}}{1/\lambda + l + M}.$$

The average sleep duration, $1/\lambda$, is computed as a function of V_{cap} by solving $P_{\text{est}} = P_{\text{des}}$,

$$1/\lambda = \frac{P_r l + P_t M + C_{sr} + C_{ts}}{P_{\text{des}}(V_{\text{cap}})} - l - M. \tag{3.15}$$

We remark that the listen time l is obtained using the PCA with $N = 2$ (i.e., we try to maximize the discovery rate for each *directional link*).

We claim that the robustness of Panda-D is two-fold. First, it is power aware and nodes can operate under different and varying power harvesting rates, relaxing (A1). Additionally, it does not require any a priori knowledge of the size or topology of the network, relaxing (A2), (A3).

3.5 Experimental Evaluation

We now evaluate Panda using a testbed, pictured in Fig. 3.5, composed of the TI eZ430-RF2500-SEH [75] prototypes (described in Section 3.1.1). First, we evaluate Panda in the context of the model presented in Section 3.1.2. We compare Panda's experimental discovery rate, denoted by U_E, to related work. Additionally, we present Panda's performance with varying parameters (e.g., transmission power, message length). Then, we evaluate Panda-D in scenarios with non-homogeneous power harvesting and multihop topologies.
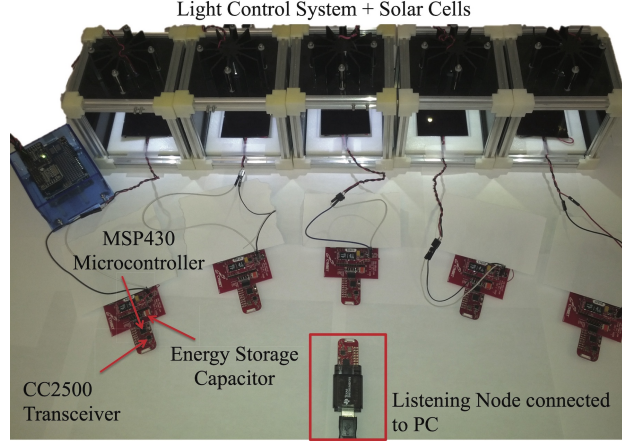
Figure 3.5: Panda experimental setup: 5 energy harvesting nodes harvest energy through the solar cells with neighbor discovery rates monitored by a listening node.

Byte	Data
0	Packet length (18 bytes)
1	Type
2-11	Neighbor table
12-13	Capacitor voltage
14-15	Debugging information
16	Transmissions counter
17	Originating node ID

Table 3.3: Discovery message structure.

Parameter	P_t	P_r	M	C_{sr}	C_{rs}	C_{tr}
Value	59.23 mW	64.85 mW	0.92 ms	74.36 uJ	13.48 uJ	4.83 uJ

Table 3.4: Measured prototype parameters.

3.5.1 Protocol Implementation

In accordance with Panda, the microcontroller builds the discovery message and sends it to the low-power transceiver. Table 3.3 illustrates the structure of the discovery message. The message contains debugging information, the source ID of the transmitting node, and the node's capacitor voltage (which is sampled from the ADC). Additionally, the message includes the number of discoveries from each neighbor since the initialization of the experiment, referred to as the node's *neighbor table*. The total length of a discovery message is 18 bytes and the resulting transmission

34
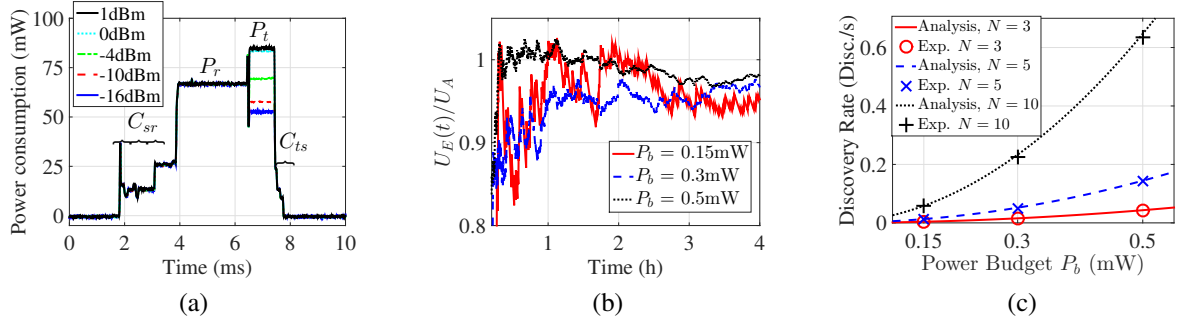
Figure 3.6: (a) Power consumption and transition costs for different transmission power levels for a node transitioning between the sleep, receive, and transmit states, and back to sleep. (b) Panda experimental evaluation with varying power budgets, P_b: Convergence of the experimental discovery rate (U_E) to the analytical discovery rate (U_A) for $N = 5$. (c) Panda's discovery rate with varying power budgets, P_b, and number of nodes, N

duration of the discovery message is 0.92 ms.

In order to characterize the energy costs, we measure the power consumption of the micro-controller and transceiver using an oscilloscope. Fig. 3.6(a) shows the power levels for a node transitioning between the sleep, receive, and transmit states. We compute the average power consumption and transition energy for each state, with values summarized in Table 3.4.

We note that the transition times to and from the sleep state are non-negligible (in some cases a few ms). To account for this, these transition times are considered as part of the sleep state and, are therefore, subtracted from the actual sleep duration. We elaborate further on the importance of incorporating these switching costs in Appendix 3.7.2.

The parameters in Table 3.4 compose the inputs to the PCA, which computes the rate of the exponential sleep λ_A and the duration of the listen state l_A as well as an expected discovery rate U_A. These *configuration parameters* are loaded into the nodes for experimental evaluation in which we observe the discovery rate as well as the power consumption.

3.5.2   Testbed and Experimental Setup

We consider networks of 3, 5, and 10 nodes ($N = 3, 5, 10$). We consider power budgets of $P_b = 0.15, 0.3, 0.5$ mW; these are aligned with other solar harvesting budgets [136]. Initially,

35

N	P_b (mW)	λ_A^{-1} (ms)	l_A(ms)	Duty Cycle (%)	U_A (Disc./s)	U_E (Disc./s)	Error (%)	Run Time (h)
3	0.15	1778.68	2.066	0.168	.0039	.0038	-1.35	36
	0.3	887.39	2.070	0.336	.0156	.0154	-1.23	36
	0.5	530.88	2.075	0.561	.0434	.0438	1.07	48
5	0.15	1777.18	2.068	0.168	.0130	.0132	1.43	96
	0.3	885.91	2.075	0.337	.0519	.0518	-0.33	60
	0.5	529.43	2.084	0.564	.1443	.1427	-1.15	18
10	0.15	1773.49	2.075	0.169	.0584	.0589	0.89	18
	0.3	882.32	2.089	0.340	.2332	.2341	0.38	18
	0.5	525.97	2.107	0.572	.6470	.6510	0.62	18

Table 3.5: Panda experimental parameters: (λ_A, l_A) generated using the PCA for every input (N, P_b) pair and the resulting analytical (U_A) and experimental (U_E) discovery rate.

to confirm the practicality of Panda when assumptions (A1), (A2), and (A3) hold, we place the nodes in close proximity with a homogenous power budget. In Section 3.5.7, we will evaluate Panda-Dynamic (Panda-D) and relax these assumptions by considering a multihop topology and non-homogenous power harvesting.

To facilitate experimental evaluation with up to $N = 10$ nodes, in addition to an energy harvesting node shown in Fig. 1.5(b), we also incorporate nodes powered by AAA batteries into the experiments. Both the energy harvesting node and the nodes powered by AAA batteries operate using the same configuration parameters and have identical behaviors (i.e., the source of power does not affect the behavior of Panda). However, we carefully logged the power consumption of the energy harvesting node by including control information in the discovery message (see Table 3.3).

We utilize a listening node consisting of a microcontroller and transceiver set to a promiscuous sniffing mode to log experimental results. Powered by a USB port on a monitoring PC, the listening node reports all received messages to the PC for storage and post processing. The experimental discovery rate, U_E, is computed by dividing the total number of discoveries since the initialization of the experiment by the experiment duration. Clearly, the time until which the experimental discovery rate converges depends on the rate of discovery. In Fig. 3.6(b), we observe the experimental discovery rate, U_E over time for $N = 5$ and $P_b = 0.15, 0.3, 0.5$ mW. Based on the results shown in Fig. 3.6(b), all experiments were conducted for up to 96 hours.

ND Table	1	2	3	4	5	Total RX
1	0	35	42	32	42	152
2	24	0	23	45	38	130
3	39	36	0	21	33	129
4	36	35	46	0	32	149
5	38	42	42	42	0	164

Table 3.6: Neighbor table for $N = 5$, $P_b = 0.3\,\mathrm{mW}$ after 4 hours. Entry (i, j) shows the number of discoveries of node j by node i.

The light levels are set to correspond to each of the power budgets, P_b. However, the performance of the solar cells vary significantly due to external effects such as aging, orientation, and temperature [69]. To mitigate these affects and facilitate repeatable and controllable experiments, we designed a software controlled light system which we describe in Appendix 3.7.3.

Additionally, as mentioned in Section 3.1, the prototype is not power aware. That is, although we can accurately measure the power harvested by the solar cell, it is difficult to control the energy actually stored in the capacitor, due to numerous inefficiencies of the harvesting circuitry, which are further described in Appendix 3.7.3. As such, we empirically estimated the harvesting inefficiency to be 50% and adjust the light levels to provide each node energy according to the value of P_b chosen.

3.5.3 Discovery Rate

For each (N, P_b) pair, we evaluate Panda, with the experimental parameters summarized in Table 3.5. First, we note that Panda's duty cycle is typically between 0.1% − 0.6%, which is significantly lower than the duty cycles considered in related protocols [48]. Additionally, note the accuracy of the analytical discovery rate, U_A, computed from (3.2), compared to the experimental discovery rate, U_E. On average, the error between them is around 1%. This confirms the practicality of Panda and the model described in Section 3.1.

In Fig. 3.6(c), we plot the experimental and analytical discovery rate for each value of (N, P_b) shown in Table 3.5 and observe the effect of varying N and P_b. As expected, the discovery rate increases as P_b increases. The number of nodes N is directly correlated with the discovery rate, as
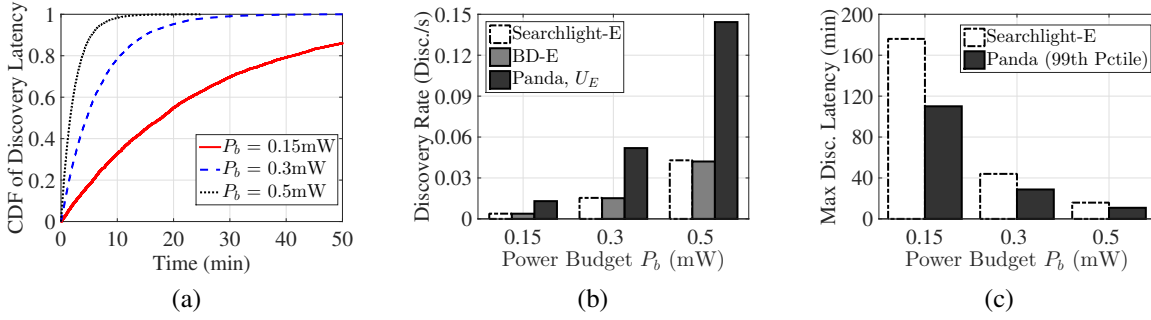
Figure 3.7: Panda experimental evaluation for $N = 5$: (a) CDF of per link discovery latency; comparison to SearchLight-E [50] and BD-E [55] of (b) the discovery rate and (c) the worst case latency.

indicated in (3.2) and (3.3). As such, the discovery rate increases as N increases.

Additionally, by tracking each nodes' neighbor table in Table 3.6, we confirm that all nodes discover one another and exhibit similar per link discovery rates.

3.5.4   Discovery Latency and Comparison to Related Work

The discovery latency is the time between consecutive discoveries for a *directional link*. It can be an important parameter for numerous applications where nodes are only within communication range for short periods of time. Although the objective of Panda is to maximize the discovery rate, in Fig. 3.7a, we show the CDF of the discovery latency for each *directional link* in an experiment with $N = 5$ and varying power budgets. Clearly, the average discovery latency decreases as the average discovery rate increases. Thus, for a higher power budget, the discovery latency decreases.

Previous work [49, 50, 52] focused on minimizing the worst case discovery latency for a link. We compare the discovery latency of Panda, shown in Fig. 3.7a, to previous work. However, as mentioned in Section 2.1, previous work considers a duty cycle constraint instead of a power budget (P_b). To provide a means of comparison, we use the following equation to relate the power constraint to a duty cycle.

$$P_b = \text{Duty Cycle } (\%) \cdot \text{Average Active Power (mW)} \tag{3.16}$$

We compare to the deterministic Searchlight protocol [50], which minimizes the worst case discov-

ery latency [52]. We also compare to the the well-known probabilistic Birthday (BD) protocol [55]. To account for the power budget, we modify these protocols based on (3.16) (with details explained in Appendix 3.7.4) and denote them as Searchlight-E and BD-E. Based on previous work [52], we set the slot size for Searchlight-E and BD-E to 50 ms and add an overflow guard time of 1 ms.

In Fig. 3.7b, we compare the average discovery rate for Panda vs. simulations of the Searchlight-E and BD-E protocols. We found that Panda typically outperforms the Searchlight-E and BD-E protocols by over $3\times$ in terms of the average discovery rate.[9] Furthermore, in Fig. 3.7c, we consider the *worst case* discovery latency and show that although Panda has a non-zero probability of having any discovery latency, for the experiments we considered, the 99th percentile of discovery latency outperformed the Searchlight-E protocol worst case bound by up to 40%.

Note that the Searchlight protocol was proven to minimize the worst case discovery latency. However, as shown through our evaluation, Panda outperforms Searchlight-E by a factor of $3\times$ in terms of average discovery rate. Moreover, in most cases (over 99%), the discovery latency is below the worst case bound from Searchlight-E. This emphasizes the importance of incorporating a detailed power budget, as is done in Panda, as opposed to a duty cycle constraint.

3.5.5 Power Consumption

Using Panda, a node consumes power at a rate of up to P_b (mW), *on average*. However, the power consumption is stochastic, and therefore, it is expected that the energy stored will vary over time. In Fig. 3.8(a), we show the capacitor voltage over time for a node with $N = 5$ and $P_b = 0.5$ mW. Energy neutrality is demonstrated by the oscillation in the energy level within the limits of the capacitor storage. Recall from Section 3.1 that if the energy drains below a software induced threshold of 3.6 V, the node temporarily sleeps for 10 s to regain energy. These periods of additional sleep affect the discovery rate and, as indicated by the accuracy of the experiments, these occurrences are rare.

Furthermore, in Fig. 3.8(b), we experiment with varying capacitor sizes ranging from 10–

---

[9]As described in Appendix 3.7.4, the simulations of Searchlight-E and BD-E do not account for packet errors or collisions. As such, the discovery rates for these protocols is likely to be lower in practice.
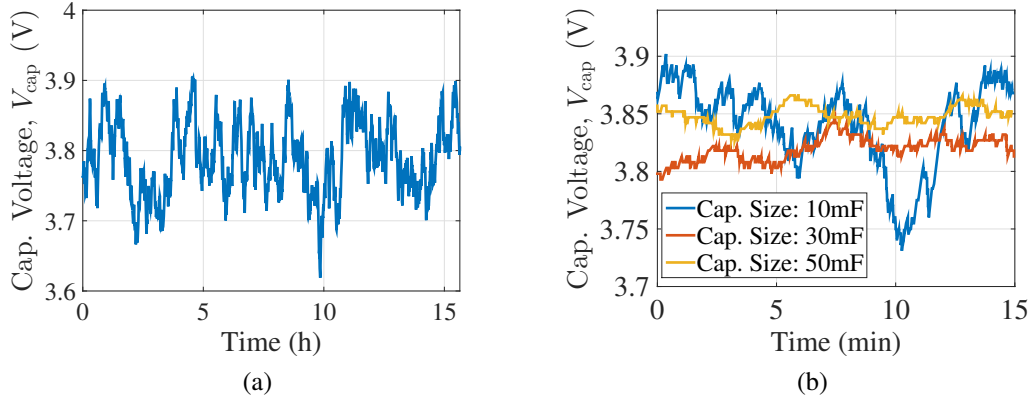
Figure 3.8: Panda experimental evaluation for $N = 5$: (a) capacitor voltage level (V_{cap}) for a node with a 30 mF capacitor and $P_b = 0.5$ mW, and (b) capacitor voltage level of nodes with varying capacitor sizes over 15 minutes with $P_b = 0.15$ mW.

50 mF. As expected, smaller capacitors have added variation in the voltage level. Therefore, smaller capacitors can reach the upper (fully charged) or lower (empty) voltage limits more frequently than larger capacitors. In practice, the capacitor should be sized with respect to the variation in the power consumption and power harvested.

3.5.6 Panda Design Considerations

We now consider Panda's performance with varying transmit power levels, P_t, and discovery message length, M.

Transmit Power, P_t. The transmission power can be set in software. A larger transmission power can result in more geographical coverage, but also consumes more energy. In Fig. 3.9(a), we consider $N = 5$ and $P_b = 0.5$ mW and observe how the discovery rate changes with varying transmission powers. A larger transmission power requires nodes to sleep longer before transmitting, resulting in less discoveries. Note that for this experiment the energy costs from Table 3.4 no longer hold and we remeasured them to compute the configuration parameters.

Discovery Message Duration, M. The discovery message requires M (ms) to be transmitted and contains the node ID and neighbor table information. By adjusting the modulation/coding of the radio or the data content, the packet length can be shortened. A shorter packet length results in
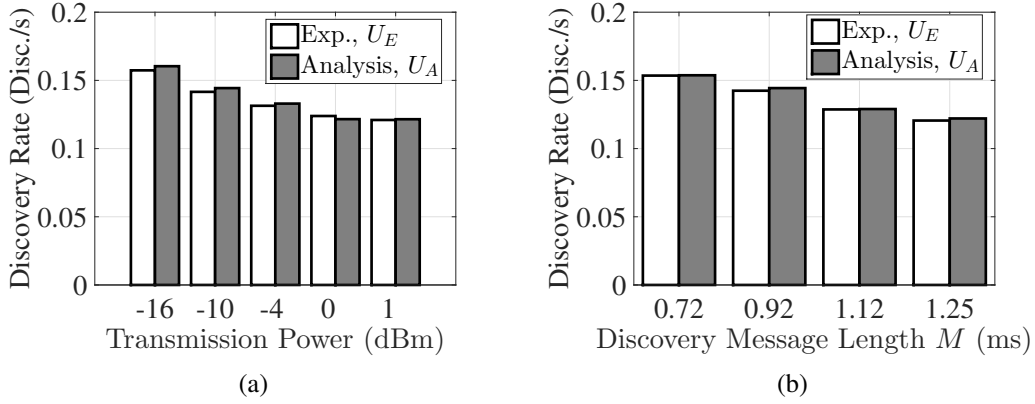
Figure 3.9: Performance evaluation of Panda with $N = 5$ and $P_b = 0.5\,\mathrm{mW}$: Experimental and analytical discovery rates under varying (a) transmission power level (P_t), and (b) discovery message length (M).

less time transmitting as well as less time listening for messages. As shown in Fig. 3.9(b), smaller packet sizes result in an increase in the discovery rate. This presents an application design decision if the contents of the packet can be adjusted to obtain a desired discovery rate.

3.5.7 Panda-D

We now evaluate Panda-D (described in Section 3.4). The only input to Panda-D is the estimated power harvesting rate, $P_b = 0.15\,\mathrm{mW}$, and the capacitor voltage V_{cap}. From (3.15), the average duration of the exponential sleep is then computed as,

$$\frac{1}{\lambda} = \frac{382.2238}{V_{\mathrm{cap}} - 3.5857} - 2.9843\,\mathrm{ms}. \tag{3.17}$$

Thus, the node scales its power consumption based on V_{cap}. For example, at $V_{\mathrm{cap}} = 3.6\,\mathrm{V}$ and $4\,\mathrm{V}$, the node will sleep on average for $26.75\,\mathrm{s}$ and $0.92\,\mathrm{s}$, respectively.

To estimate the average sleep duration for a given node in Panda-D, we compute the average value of V_{cap} over the course of an experiment. Based on the this value, the average sleep duration is estimated from (3.17).

Panda-D does not require a priori information of the number of neighbors, N. Therefore, throughout this section, (A3) is relaxed. Below, we observe the performance of Panda-D first
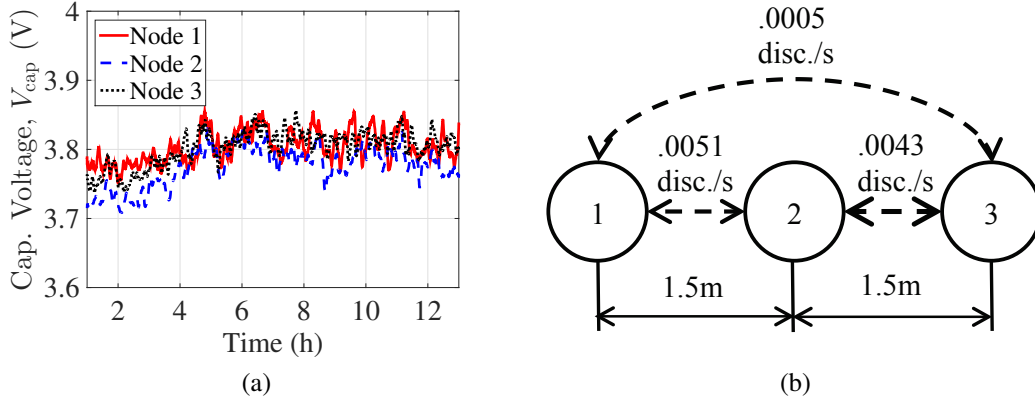
Figure 3.10: Panda-D experimental evaluation for $N = 3$ and homogeneous power harvesting $P_b = 0.15\,\text{mW}$: (a) Capacitor voltage, V_{cap}, in a clique topology. (b) Per link experimental discovery rates for a line topology after 50 hours.

when (i) nodes remain in a clique topology with homogenous power budgets. Then we consider Panda-D (ii) in a multihop topology (relaxing (A2)), and finally (iii) in non-homogenous power harvesting scenario (relaxing (A1)). Relaxing all assumptions together requires running a live real-world experiment and is a subject of future work.

Comparison to Panda. We first evaluate Panda-D with an experimental setup similar to the one shown in Fig. 3.5. Specifically, we consider a network of $N = 3$ nodes in close proximity with a power harvesting rate of $P_b = 0.15\,\text{mW}$.

Fig. 3.10(a) shows that the capacitor voltage for all 3 nodes stays approximately near 3.8 V. As described in Section 3.4, the average power consumption at 3.8 V is approximately P_b. Therefore, in this scenario, Panda-D and Panda have similar power consumption and discovery rates. As such, the experimental discovery rate of Panda-D is within 1% of the analytical estimate of Panda.

Multihop Topologies. Previously, we assumed that all nodes form a clique topology with no packet losses (A2) and the number of nodes N known (A3). Indeed, for the experiments conducted above with a transmission power of $-10\,\text{dBm}$, we found that nodes within about 20 m could be treated as a clique topology with over 99% packet success rates.

However, to evaluate a non-clique topology and relax (A2) and (A3), we manually reconfigured the transmission power to $-26\,\text{dBm}$ and set 3 nodes in a line topology with distance between nodes

1–2 and 2–3 of 1.5 m, as shown in Fig. 3.10(b). In this configuration, nodes rarely receive messages from their two-hop neighbors. Nodes run Panda-D and are given light levels corresponding to the power harvesting rate of $P_b = 0.15\,\mathrm{mW}$ (as described in Section 3.5.2). After 50 hours, the resulting discovery rate is shown on each link in Fig. 3.10(b).

The two extreme nodes (nodes 1 and 3) have very few discoveries from one other, due to the distance between them. However, the node in the middle (node 2) forms an effective clique of size 2 with each of its neighbors. We therefore can analyze the discovery rate per link. For example, the discovery rate of the link between nodes 1 and 2 is 0.0051 disc./s, which is within 1% of the analytical discovery rate for a clique with $N = 2$ and $P_b = 0.15\,\mathrm{mW}$. Therefore, even with non-clique topologies, each link that is within communication range can be analyzed as a network with $N = 2$. This implies that issues such as the hidden-node problem do not significantly affect the performance of Panda.

Non-Homogeneous Power Harvesting. We now consider nodes 2–5 using Panda-D with light levels corresponding to power harvesting of 0.075, 0.15, 0.225, 0.3 mW, respectively. Node 1 is a *control* node running Panda with $P_b = 0.15\,\mathrm{mW}$ and $N = 5$.

For each of the 4 Panda-D nodes, the capacitor voltage, V_{cap}, is shown in Fig. 3.11a and settles based on the power harvesting. Variations in the settling voltage stem from the dynamic average sleep duration at different power harvesting levels. For example, node 5 is given a light level of 0.3 mW, and therefore, has a shorter sleep duration than node 2 (light level of 0.075 mW). Correspondingly, Fig. 3.11b shows the neighbor table: entry (i,j) represents the number of discoveries of node j by node i over the experiment duration. Due to non-homogeneity, the discovery rate for each link depends on the power harvested; nodes with larger power budgets discover their neighbors, and are discovered, more frequently.

In Appendix 3.7.5, we treat each link with non-homogenous power harvesting as a clique ($N = 2$), and estimate its discovery rate; the approximation is within 20% of the experimental value.

Figure 3.11: Panda-D experimental evaluation for non-homogeneous power harvesting with $N = 5$ over 24 hours: (a) Capacitor voltage, V_{cap}, and (b) resulting neighbor table.
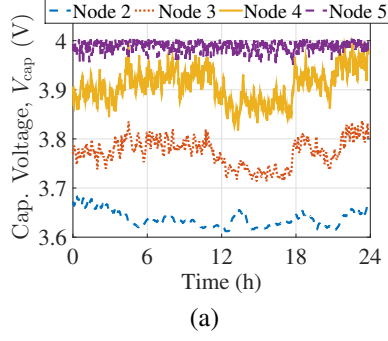
ND Table	1	2	3	4	5
1	0	11	53	72	96
2	9	0	9	15	20
3	55	13	0	71	113
4	70	16	61	0	177
5	93	28	106	175	0

3.6 Conclusion

In this chapter, we designed, analyzed, and evaluated Panda, an ND protocol for energy harvesting nodes. By accounting for specific hardware constraints (e.g., transceiver power consumption for transmission, reception, and state switching), Panda adheres to a *power budget*. Using renewal theory, we developed the Panda Configuration Algorithm (PCA) to determine the nodes' sleep and listen durations which maximize the discovery rate; the PCA achieves a nearly-optimal discovery rate (over 94%). We evaluated Panda using TI eZ430-RF2500-SEH energy harvesting nodes. The real-life accuracy was consistently within 2%, demonstrating the practicality of our model. Furthermore, Panda outperformed the closest related protocols Searchlight-E [50] and BD-E [55] by achieving a discovery rate that was up to $3\times$ higher. Finally, we showed that a version of the protocol, Panda-D, was able to adapt to scenarios with non-homogeneous power harvesting and multihop topologies.

Panda can be readily applied to nodes with a non-rechargeable battery, where the power budget is set based on the desired lifetime. In Chapter 4, we relax the assumptions in Section 3.2 and develop the EconCast protocol which considers: (i) nodes with heterogenous transmit and listen/receive power consumption levels as well as power budgets, (ii) the network size (i.e., number of nodes) is *not* known a priori and the nodes are *not synchronized* with each other, and (iii) an objective of maximizing the aggregate throughput (i.e., communication rate) in the MAC layer.

N	P_b (mW)	Pr{A node wakes up in the middle of a packet}	Expected Energy per Renewal (uJ)	Pct of P_b (%)
3	0.15	0.34e-3	0.0302	0.034
	0.3	0.69e-3	0.0605	0.068
	0.5	1.15e-3	0.1010	0.112
5	0.15	0.41e-3	0.0363	0.068
	0.3	0.83e-3	0.0728	0.135
	0.5	1.38e-3	0.1215	0.223
10	0.15	0.47e-3	0.0410	0.151
	0.3	0.94e-3	0.0822	0.300
	0.5	1.57e-3	0.1376	0.495

Table 3.7: Panda idle power consumption for every input (N, P_b) pair from Table 3.5: the probability of a node waking in the middle of a transmitted packet and the expected portion of the power budget (P_b) consumed.

3.7 Appendices

3.7.1 Appendix – Incorporating Idle Power Consumption

In this work, we disregard the power cost of nodes in the sleep state. In this section, we explain how these costs can be incorporated. As described in Section 3.5, the idle cost of the microcontroller is normally 1.6 uW. This draw is constant for all states (sleep, listen, and idle). As such, to incorporate it into our model, it is simply subtracted from the power budget P_b.

In Section 3.2.4, we ignore the expected amount of energy (uJ) consumed by a node when it begins to listen while a packet is currently being transmitted (exemplified by Node 5 in Fig. 3.3). In this case, the node spends energy to transition to and from the sleep state, as well as listen for a short fixed Clear Channel Assessment (CCA) period, denoted as t_{CCA}. The energy consumption is then $C_{sl} + P_r \cdot t_{\text{CCA}} + C_{ls}$. This event occurs in a renewal with probability given by $\frac{N-1}{N}e^{-\lambda l}(1 - e^{-\lambda M})$. Firstly, the node must not be the transmitter w.p. $(\frac{N-1}{N})$. As the node is asleep when the transmitter begins to listen, it must then sleep for at least l duration. Finally, given that it is in the sleep state when the transmitter begins to transmit, it must then wakeup before the message is transmitted (duration M).

For the evaluations in Section 3.5, the idle power costs are summarized in Table 3.7 for the
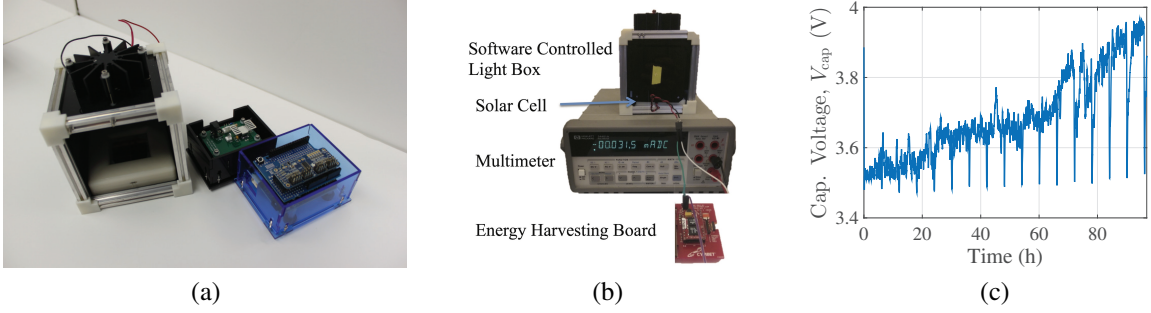
Figure 3.12: (a) Software controlled light system including a dark box enclosure, high-power LED driver, and an Arduino-based light controller. (b) Energy harvesting characterization measurement setup. (c) Capacitor voltage with adjusted light levels over 4 days to find the energy neutral point.

N	P_b (mW)	U_A, w/ Switching Costs (disc./s)	U_A, w/o Switching Costs (disc./s)	Power Consumed (mW)
3	0.15	0.0039	0.010	0.26
	0.3	0.0156	0.038	0.52
	0.5	0.0434	0.107	0.86
5	0.15	0.0130	0.032	0.26
	0.3	0.0519	0.128	0.52
	0.5	0.1440	0.359	0.87
10	0.15	0.0584	0.144	0.26
	0.3	0.2330	0.581	0.52
	0.5	0.6470	1.630	0.87

Table 3.8: Panda performance evaluation: the discovery rate U_A resulting from the PCA, and the actual power consumed when ignoring the switching costs.

experimental parameters originally presented in Table 3.5. As can be seen, this probability is quite small (under 0.2%). Thereby, the percentage of the power budget consumed, on average, is always less than 0.5% of the power budget, and therefore can be ignored. We note, however, that the PCA can easily be modified to incorporate this idle power consumption.

3.7.2 Appendix – Importance of Switching Costs

In this work, we incorporate the costs to switch to and from different radio states (sleep, receive, transmit). In Table 3.8, we demonstrate the importance of accounting for these costs (which are commonly overlooked in related work). In the table, the PCA is used to compute the parameters under the assumption that $C_{ij} = 0 \, (\forall i, j \in \{s, r, t\})$. As indicated in the table, the discovery

rate improves by $2\times$–$3\times$ compared to the discovery rate when including the switching costs from Table 3.4. However, the power consumed by transition causes the power budget to be exceeded by up to 80%. Therefore, ignoring the switching costs may improve the discovery rate, but also results in significantly higher power consumption.

3.7.3 Appendix – Software Controlled Light System and Harvesting Inefficiencies

We develop an advanced software controlled light system (shown in Fig. 3.12(a)) that uses a Java-based script and Arduino-based light control modules to precisely control the irradiance (light energy intensity) generated by LEDs. The system produces 1,024 irradiance levels from 0–$14\,\text{mW/cm}^2$ and the level can be changed every $100\,\text{ms}$. Dark box enclosures and 3D printed mounting fixtures ensure full control over the light conditions at the solar cells. This guarantees that our experimental evaluations *are based on the same energy inputs*.

Furthermore, we conduct extensive experiments utilizing a UV818 photodetector to carefully calibrate the irradiance of the light control system. We characterize the power harvested under both the ambient light and the software controlled light setup. The measurement setup is shown in Fig. 3.12(b). The solar cell is connected to an ammeter, and together with the cell voltage, the harvested power is easily computed. Using the software controlled light system, we record the mapping of irradiance to power harvested and can control the power harvested by the solar cell.

However, the actual power that is stored depends on numerous inefficiencies in the power harvesting circuitry. Specifically, the Cymbet CBC5300 up-converts the power harvested from the solar cell voltage to the capacitor charging voltage of $4\,\text{V}$, which consumes some overhead energy. In addition, there are inefficiencies in the regulation circuit which regulates an output voltage of $3.5\,\text{V}$ to power the load. These inefficiencies are difficult to characterize as they vary based on uncontrollable external factors such as the temperature and component variations.

In our evaluations (see Section 3.5), nodes were given light levels which corresponded to their power budget P_b. To accomplish this, given the inefficiencies described above, we conduct a 4-day experiment in which nodes operated using Panda, yet we varied the light levels every 6 hours. An

example of the capacitor voltage for one node in this experiment is shown in Fig. 3.12(c). Each valley represents a 10-minute "dark" period where the light is completely off before changing to the next light levels.

With limited light levels (i.e., hours 0–20 in Fig. 3.12(c)), the capacitor voltage operates near the minimum implying that the node is consuming more energy than it harvests. With larger light levels (i.e., hours 80–100), the node is harvesting more energy than it consumes and thus the capacitor voltage reaches its upper limit. However for the range of lights corresponding to 20–80 hours in Fig. 3.12(c), the node has a relatively stable voltage, implying that it is consuming power (on average) as the same rate it harvests; *energy neutrality* is obtained.

By performing this experiment for all nodes, we found the light levels at which each node is energy neutral. The neutral light levels varied significantly. Furthermore, by comparing the power harvested by the solar cell to the power budget (P_b), we found that the efficiency of the storage process to be between $40\% - 60\%$. This emphasizes the need to incorporate energy storage feedback into the ND protocol, as is done by Panda-D.

3.7.4   Appendix – Searchlight-E and BD-E

In this section, we describe the Searchlight [50] and Birthday [55] protocols, as well as how they are adapted to obey an power budget, termed Searchlight-E and BD-E. Both protocols divide time into slots, and in each time slot, a node must decide whether to be *active* or *idle*.[10] The percentage of active slots is termed the *duty cycle*. During active slots, nodes send a packet transmission, then listen to the channel for messages from neighbors, and end the slot with another packet transmission [49]. Thus, when the active slots of any two nodes are un-aligned and only partially overlap, a mutual discovery occurs. Within this model, the ND protocol designates each slot as active or idle such that they the node maintains the duty cycle constraint.

In Birthday, each node randomly decides whether to be active in a time slot with probability p.

---

[10]In its original form, the Birthday protocol further divided active slots into a listen slot and a transmit slot. However, since Birthday was published, Disco [49] demonstrated the *beacon slot* which combines the listen and transmit slots. Thus, for simplicity, we adapt Birthday to use beacons.

Clearly, p is the effective duty cycle. In Searchlight, two slots are active per cycle of t slots. The node is always active in the first slot of cycle, which is referred to as the anchor slot. Clearly if the anchor slot of two nodes overlap, then they will continue to discover each other in each cycle. Since nodes are unsynchronized, it is likely that the anchor slots of the two nodes will not overlap and, by construction, the offset of the two nodes must be less than $(t/2)$ slots. Thus, each node activates a second time in a cycle in what is called a *probe* slot. The probe slot begins in slot 2 of the cycle is successively incremented until slot $(t/2)$ at which point it is guaranteed to have discovered the neighboring node.

We now analyze the power consumption for each protocol. We will denote the time slot duration as d_s and the energy consumed per active slot is,

$$E_{\text{slot}} = 2P_t M + P_r \cdot (d_s - 2M) + C_{st} + C_{ts}.$$

The power budget for Searchlight results in two active slots per cycle of length t and is written as, $2E_{\text{slot}}/(td_s) \leq P_b$. In Birthday, each node transmits a beacon in a slot w.p. p and the power budget is simply, $pE_{\text{slot}}/d_s \leq P_b$. In our evaluation, we select t and p such that the power budget is fully consumed and term these protocols, Searchlight-E and BD-E.

We note that there are numerous aspects of related protocols [49–52, 55] which have not been considered. Specifically, existing works do not consider collisions occurring due to not listening before transmitting. Furthermore, numerous practical parameters are not considered such as the setting of the slot size. As can be seen above, the slot size impacts the average power consumption. In our simulation of Searchlight-E and BD-E, we ignore collisions and set the slot size to $d_s = 50$ ms with a guard time of 1 ms, as was done in [52].

3.7.5 Appendix – Approximate Analysis of Panda-D

We now approximate the directional discovery rate U_{ij} (i.e., the rate at which node i is discovered by node j) for a link between node i and j, under non-homogeneous power budgets λ_i and

	1	2	3	4	5
1	-	-17	8	-15	-9
2	-32	-	-22	-23	-20
3	12	12	-	-4	23
4	-17	-20	-17	-	12
5	-11	13	16	11	-

Table 3.9: Panda-D Discovery Rate Approximation for the $N = 5$ experiment from Fig. 3.11: Error rate (%) of the experimental per-link discovery rate compared to (3.18).

λ_j, respectively. Using similar analysis as in Section 3.2, we obtain

$$U_{ij} = \frac{\frac{\lambda_i(1-e^{-\lambda_j l})}{\lambda_i+\lambda_j}}{\frac{1}{\lambda_i+\lambda_j} + l + M}. \tag{3.18}$$

To evaluate this approximation, we apply it to the non-homogenous power harvesting experiment described in Section 3.5.7, with the total number of discoveries on each directional link presented in Fig. 3.11b.

Recall that the sleep rate for a node i, λ_i, is dynamically changing in Panda-D. We estimate the sleep rate based on the experimental average capacitor voltage using (3.15). In Table 3.9, we compute the error rate between the experiment per-link discovery rate and (3.18). The approximation is quite crude (typically within 25%). Yet it can still be used as a rough approximation of the per-link discovery rate. We remark here that the relatively high errors come from: (1) the small number of discoveries, (2) each node is operating independently without knowledge of N a-priori, and (3) errors in the ADC capacitor voltage sampling.

Chapter 4: Maximizing Broadcast Throughput Under Ultra-Low-Power Constraints

While the Panda protocol described in Chapter 3 outperforms other well-known protocols in terms of the achievable neighbor discover (ND) rate, it still has limitations and its efficiency compared to the optimal case has not been addressed. In this chapter, we generalize the ND rate maximization problem, described in Section 3.3, and formulate the problem of *maximizing broadcast throughput among energy-constrained nodes*. We design, analyze, and evaluate EconCast: **E**nergy-**con**strained Broad**Cast**. EconCast is an asynchronous distributed protocol in which nodes transition between sleep, listen/receive, and transmit states, while maintaining a power budget.

First, we derive oracle groupput and anyput (i.e., maximum throughput achieved by an oracle) and provide methods to efficiently compute their values (Sections 4.1 and 4.2). Then, we use Lagrangian methods and a Q-CSMA (Queue-based Carrier Sense Multiple Access) approach to design EconCast (Section 4.3). To respectively maximize the two alternative throughput measures, EconCast can operate in groupput or anyput mode. Nodes running EconCast dynamically adapt their transition rates between sleep, listen, and transmit states based on (i) the energy available at the node and (ii) the number (or existence) of other active listeners. To support the latter, a listening node emits a low-cost informationless "ping" which can be picked up by other listening nodes, allowing them to estimate the number (or existence) of active listeners. We briefly discuss how this method helps increasing the throughput and the implementation aspects. We analyze the performance of EconCast and prove that, in theory, it converges to the oracle throughput (Section 4.4).

We evaluate the throughput performance of EconCast numerically and via extensive simulations under a wide range of power budgets and listen/transmit consumption levels, and for various heterogeneous and homogeneous networks (Sections 4.5 and 4.6). Specifically, numerical results

51

$\mathcal{N}, N$	Set of nodes, number of nodes
$L_i, \mathbf{L}$	Node i's listen power consumption (W), $\mathbf{L} = [L_i]$
$X_i, \mathbf{X}$	Node i's transmit power consumption (W), $\mathbf{X} = [X_i]$
$\rho_i, \boldsymbol{\rho}$	Node i's power budget (W), $\boldsymbol{\rho} = [\rho_i]$
b_i	Energy storage level of node i (J)
$\mathbf{w}, \mathcal{W}$	Network state, the set of collision-free states
$\alpha_i, \boldsymbol{\alpha}$	Fraction of time node i listens, $\boldsymbol{\alpha} = [\alpha_i]$
$\beta_i, \boldsymbol{\beta}$	Fraction of time node i transmits, $\boldsymbol{\beta} = [\beta_i]$
$\gamma, \hat{\gamma}$	Indicator if existing some nodes listening, its estimated value
$c, \hat{c}$	Number of nodes listening, its estimated value
ν	Indicator if there is exactly one node transmitting
$\pi_{\mathbf{w}}, \boldsymbol{\pi}$	Fraction of time the network is in $\mathbf{w} \in \mathcal{W}$, $\boldsymbol{\pi} = [\pi_{\mathbf{w}}]$
$\mathcal{T}_{\mathbf{w}}$	Throughput of state $\mathbf{w} \in \mathcal{W}$
$\mathcal{T}, \mathcal{T}^{\star}$	Throughput and oracle throughput
$\mathcal{T}_g, \mathcal{T}_g^{\star}$	Groupput and oracle groupput
$\mathcal{T}_a, \mathcal{T}_a^{\star}$	Anyput and oracle anyput
$\eta_i, \boldsymbol{\eta}$	Lagrange multiplier of node i, $\boldsymbol{\eta} = [\eta_i]$

Table 4.1: Nomenclature.

show that EconCast outperforms Panda and prior art (Birthday [55] and and Searchlight [50]) by a factor of $6\times$–$17\times$ under realistic assumptions. In addition, we consider the performance of Econ-Cast in terms of burstiness and latency. We also consider the design tradeoffs of EconCast when oprating in groupput and anyput modes. We implement EconCast using the TI eZ430-RF2500-SEH energy harvesting nodes and experimentally show that in practice it obtains 57% – 77% of the achievable throughput. Moreover, we compare the experimental throughput to analytical throughput of Panda [17] (where the analytical throughput is usually better than the experimental performance) and show that, for example, EconCast outperforms Panda by $8\times$–$11\times$.

This chapter's contributions were previously presented in [19, 20].

4.1 Network Model and Problem Formulation

We consider a network of N energy-constrained nodes whose objective is to distributedly maximize the broadcast throughput among them. The set of nodes is denoted by $\mathcal{N}$. Table 4.1 summarizes the notations.

4.1.1 Basic Node Model

Power Consumption: A node $i \in \mathcal{N}$ can be in one of three states: *sleep* (s), *listen/receive*[1] (l), and *transmit* (x), and the respective power consumption levels are 0, L_i (W), and X_i (W).[2] These values are based on hardware characteristics.

Power Budget: Each node i has a *power budget* of ρ_i (W). This budget can be the rate at which energy is harvested by an energy harvesting node or a limit on the energy spending rate such that the node can maintain a certain lifetime. In practice, the power budget may vary with time [136, 144] and the distributed protocol should be able to adapt. For simplicity, we assume that the power budget is constant with respect to time, but our analysis in [145] can be easily extended to scenarios with time-varying power budget with the same constant mean. Each node i also has an energy storage (e.g., a battery or a capacitor) whose level at time t is denoted by $b_i(t)$.

Severe Power Constraints: Intermittently connected energy-constrained nodes cannot rely on complicated synchronization or structured routing approaches.

Unacquainted: Low bandwidth implies that each node i must operate with very limited (i.e., no) knowledge regarding its neighbors, and hence, does not know or use the information (ρ_j, L_j, X_j) of the other nodes $j \neq i$.

4.1.2 Architecture Assumptions

We assume that there is only *one frequency channel* and *a single transmission rate* is used by all nodes in the transmit state. Similar to CSMA, nodes perform *carrier sensing* prior to attempting transmission to check the availability of the medium. Energy-constrained nodes can only be awake for very short periods, and therefore, the likelihood of overlapping transmissions is negligible. We also assume that *a node in the listen state can send out low-cost, informationless "pings"* which can be picked up by other listening nodes, allowing them to estimate the number (or existence) of active listeners. We explain in Section 4.3 how this property will help us develop a distributed

---

[1] We refer the *listen* and *receive* states synonymously as the power consumption in both states is similar.

[2] The actual power consumption in *sleep* state, which may be non-zero, can be incorporated by reducing ρ_i, or increasing both L_i and X_i, by the sleep power consumption level.

protocol. In practice, a short *pinging interval* is introduced after each packet transmission, during which the transmitter collects pings from the recipients of the previous packet and performs the estimation (see Section 4.6).

4.1.3 Model Simplifications

At any time t, the network state can be described as a vector $\mathbf{w}(t) = [w_i(t)]$, where $w_i(t) \in \{s, l, x\}$ represents the state of node i. While the distributed protocol EconCast (described in Section 4.3) can operate in general scenarios, for analytical tractability, we make the following assumptions:

- The network is a *clique*.[3]

- Nodes can perform *perfect carrier sensing* in which the propagation delay is assumed to be zero.

These assumptions are suitable in the envisioned applications where the distances between nodes are small. Under these assumptions, the network states can be restricted to the set of *collision-free* states, denoted by $\mathcal{W}$ (i.e., states in which there is *at most* one node in transmit state). This reduces the size of the state space from 3^N to $(N + 2) \cdot 2^{N-1}$.

Let $\gamma_{\mathbf{w}} \in \{0, 1\}$ indicate whether there *exists* some nodes listening in state $\mathbf{w}$ and let $c_{\mathbf{w}}$ be the number of listeners in state $\mathbf{w}$. We use $\nu_{\mathbf{w}} \in \{0, 1\}$ as an indicator which is equal to 1 if there is *exactly* one transmitter in state $\mathbf{w}$ and is 0 otherwise. Based on these indicator functions, two measures of broadcast throughput, *groupput* and *anyput*, and the throughput of a given network state $\mathbf{w}$ are defined below.

Definition 4.1 (Groupput) *The groupput, denoted by $\mathcal{T}_g$, is the average throughput of the trans-missions received by* all *the receivers, where each transmitted bit is counted once per receiver to*

---

[3]We note that EconCast is not designed based on specific assumption, regarding the topology and that nodes do not need to know the properties of other nodes. Yet, in this chapter, we mainly focus on a clique topology (i.e., all nodes are within the communication range of each other), since it lends itself to analysis. We briefly extend the analytical results to non-clique topologies in Section 4.2.3, and also evaluate the performance of EconCast in such networks.

which it is delivered, i.e.,

$$\mathcal{T}_g = \lim_{T \to \infty} \frac{1}{T} \int_{t=0}^{T} \nu_{\mathbf{w}(t)} c_{\mathbf{w}(t)} dt. \tag{4.1}$$

Definition 4.2 (Anyput) *The anyput, denoted by $\mathcal{T}_a$, is the average throughput of the transmissions that are received by* at least *one receiver, i.e.,*

$$\mathcal{T}_a = \lim_{T \to \infty} \frac{1}{T} \int_{t=0}^{T} \nu_{\mathbf{w}(t)} \gamma_{\mathbf{w}(t)} dt. \tag{4.2}$$

Definition 4.3 (Network State Throughput) *The throughput associated with a given network state* $\mathbf{w} \in \mathcal{W}$, *denoted by* $\mathcal{T}_{\mathbf{w}}$, *is defined as*

$$\mathcal{T}_{\mathbf{w}} = \begin{cases} \nu_{\mathbf{w}} c_{\mathbf{w}}, & \text{for Groupput,} \\ \nu_{\mathbf{w}} \gamma_{\mathbf{w}}, & \text{for Anyput.} \end{cases} \tag{4.3}$$

Note that without energy constraints, the oracle (maximum) groupput is $(N-1)$ and is achieved when some node always transmits and the remaining $(N-1)$ nodes always listen and receive the transmission. Similarly, the oracle (maximum) anyput without energy constraints is 1 and is achieved when some node always transmits and some other node always listens and receives the transmission.

4.1.4 Problem Formulation

Define $\pi_{\mathbf{z}}$ as the fraction of time the network spends in a given state $\mathbf{z} \in \mathcal{W}$, i.e.,

$$\pi_{\mathbf{z}} = \lim_{T \to \infty} \frac{1}{T} \int_{t=0}^{T} \mathbf{1}_{\{\mathbf{w}(t)=\mathbf{z}\}} \, dt, \tag{4.4}$$

where $\mathbf{1}_{\{\mathbf{w}(t)=\mathbf{z}\}}$ is the indicator function which is 1, if the network is with state $\mathbf{z}$ at time t, and is 0 otherwise. Correspondingly, denote $\boldsymbol{\pi} = [\pi_{\mathbf{w}}]$.

Below, we define the *energy-constrained throughput maximization problem* (P1) where the fractions of time each node spends in sleep, listen, and transmit states are assigned while the node maintains the power budget. Define variables $\alpha_i, \beta_i \in [0, 1]$ as the fraction of time node i spends

in listen and transmit states, respectively. The fraction of time it spends in sleep state is simply $(1 - \alpha_i - \beta_i)$. In view of (4.1)–(4.4), (P1) is given by

$$(\mathsf{P1}) \ \max_{\boldsymbol{\pi}} \ \sum_{\mathbf{w} \in \mathcal{W}} \pi_{\mathbf{w}} \mathcal{T}_{\mathbf{w}} \tag{4.5}$$

$$\text{subject to} \quad \alpha_i L_i + \beta_i X_i \leq \rho_i, \ \forall i \in \mathcal{N}, \tag{4.6}$$

$$\alpha_i = \sum_{\mathbf{w} \in \mathcal{W}_i^l} \pi_{\mathbf{w}}, \ \beta_i = \sum_{\mathbf{w} \in \mathcal{W}_i^x} \pi_{\mathbf{w}}, \tag{4.7}$$

$$\sum_{\mathbf{w} \in \mathcal{W}} \pi_{\mathbf{w}} = 1, \ \pi_{\mathbf{w}} \geq 0, \ \forall \mathbf{w} \in \mathcal{W}, \tag{4.8}$$

where $\mathcal{W}_i^l$ and $\mathcal{W}_i^x$ are the sets of states $\mathbf{w} \in \mathcal{W}$ in which $w_i = l$ and $w_i = x$, respectively. Each node is constrained by a power budget, as described in (4.6), and (4.8) represents the fact that at any time, the network operates in one of the collision-free states $\mathbf{w} \in \mathcal{W}$.

Based on the solution to (P1), the maximum throughput is achievable by an *oracle* that can schedule nodes' sleep, listen, and transmit periods, in a centralized manner. Therefore, we define the maximum value obtained by solving (P1) as the *oracle throughput*, denoted by $\mathcal{T}^\star$. Respectively, we define the *oracle groupput* and *oracle anyput* as $\mathcal{T}_g^\star$ and $\mathcal{T}_a^\star$.

To evaluate EconCast, it is essential to compare its performance to the oracle throughput. However, (P1) is a Linear Program (LP) over an exponentially large number of variables (i.e., $|\mathcal{W}|$ is exponential in N) and is computationally expensive to solve. In Section 4.2, we show how to convert (P1) to another optimization problem with only a linear number of variables. Note that the solution to (P1) only provides the optimal fraction of time each node should spend in sleep, listen, and transmit states, but *does not* indicate how the nodes can make their individual sleep, listen, and transmit decisions locally. Therefore, in Section 4.3, we focus on the design of EconCast that makes these decisions based on (P1).

4.2 Oracle Throughput

In this section, we present an equivalent LP formulation for (P1) in a clique network which only has a linear number of variables. We also derive both an upper and a lower bound for the

oracle groupput in non-clique topologies which will be used later for evaluating the performance of EconCast in non-clique topologies.

Recall that α_i and β_i are the fraction of time node i spends in listen and transmit states, respectively. We can rewrite the constraints in (P1) as follows

$$\alpha_i L_i + \beta_i X_i \leq \rho_i, \quad \forall i \in \mathcal{N}, \tag{4.9}$$

$$\alpha_i + \beta_i \leq 1, \quad \forall i \in \mathcal{N}, \tag{4.10}$$

$$\sum_{i \in \mathcal{N}} \beta_i \leq 1. \tag{4.11}$$

Specifically, (4.9) is the usual power constraint on each node $i \in \mathcal{N}$, and (4.10) is due to the fact that a node can only operate in one state at any time. We remark that energy-constrained nodes can only be awake for very small fractions of time (i.e., $\alpha_i + \beta_i \ll 1$), and therefore (4.10) may be redundant. Finally, collision-free operation in a clique network where *at most* one transmitter can be present at any time imposes (4.11), which bounds the sum of the transmit fractions by 1.

4.2.1   Oracle Groupput in a Clique

To maximize the groupput (4.1), it suffices that any node only listens when there is another transmitter, since listening when no one transmits wastes energy. Namely, the fraction of time node i listens cannot exceed the aggregate fraction of time all other nodes transmit, i.e.,

$$\alpha_i \leq \sum_{j \neq i} \beta_j, \ \forall i \in \mathcal{N}. \tag{4.12}$$

Since a node only listens when there exists *exactly one* transmitter, every listen counts as a reception, and the groupput of a node (i.e., the throughput it receives from all other nodes) is simply the fraction of time it spends in listen state, α_i. Therefore, the groupput in a clique network simplifies to $\sum_{i \in \mathcal{N}} \alpha_i$. The *oracle groupput*, denoted by $\mathcal{T}_g^\star$, can be obtained by solving the following maximization problem

$$\text{(P2)} \ \mathcal{T}_g^\star := \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \sum_{i \in \mathcal{N}} \alpha_i, \ \text{subject to (4.9)} - \text{(4.12)}. \tag{4.13}$$

(P2) is an LP consisting of $2N$ variables and $(3N + 1)$ constraints (i.e., solving for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ given inputs of N, $\boldsymbol{\rho}$, $\mathbf{L}$, and $\mathbf{X}$). On a conventional laptop running Matlab, this computation for thousands of nodes takes seconds. Moreover, we show that the oracle groupput obtained by solving (P2) is indeed *achievable* by an oracle which can schedule nodes' listen and transmit periods. This result is summarized in the following lemma.

Lemma 4.1 *The (rational-valued) solution $(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ to (P2) can be feasibly scheduled by an oracle in a fixed-size slotted environment via a periodic schedule, (perhaps) after a one-time energy accumulation interval.*

Proof 4.1 *The proof can be found in Appendix 4.8.1.*

In *homogeneous* networks (i.e., $\rho_i = \rho$, $L_i = L$, $X_i = X$, $\forall i \in \mathcal{N}$) where nodes are sufficiently energy-constrained (i.e., (4.9) dominates (4.10)), the solution to (P2) is given by[4]

$$\beta^\star = \rho/(X + (N-1)L), \ \alpha^\star = (N-1)\beta^\star, \ \mathcal{T}_g^\star = N\alpha^\star.$$

4.2.2 Oracle Anyput in a Clique

The oracle anyput is obtained based on the observation that a transmission only occurs when there is *at least* one listener. Define variables $\chi_{i,j}$ as the fraction of time node j receives a transmission from node i, for the following constraints

$$\beta_i \leq \sum_{j \neq i} \chi_{i,j}, \quad \forall i \in \mathcal{N}, \tag{4.14}$$

$$\alpha_j = \sum_{i \neq j} \chi_{i,j}, \quad \forall j \in \mathcal{N}. \tag{4.15}$$

The *oracle anyput*, denoted by $\mathcal{T}_a^\star$, can be obtained by solving the following maximization problem

$$(\mathsf{P3}) \quad \mathcal{T}_a^\star := \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \ \sum_{i \in \mathcal{N}} \beta_i \tag{4.16}$$
$$\text{subject to} \quad (4.9) - (4.11), (4.14), \text{and } (4.15).$$

---

[4]We show that in the optimal solution, the equalities hold for equations (4.9) and (4.12). The details are in [145].

First, (4.14) ensures that when node i transmits, there is always *at least* one other node than can receive this transmission. Then, (4.15) makes sure that in the optimal schedule, the fraction of time node j listens is large enough to cover all the transmissions it receives. Therefore, (P3) maximizes the anyput by ensuring that every transmission is received by *at least* one node. In *homogeneous* networks, the closed-form solution to (P3) is given by

$$\beta^\star = \alpha^\star = \rho/(X + L), \ \mathcal{T}_a^\star = N\beta^\star.$$

4.2.3 Oracle Groupput in Non-cliques

The problem formulations (P1)–(P3) so far have assumed a clique network. Obtaining the exact maximum groupput for non-cliques (denoted by $\mathcal{T}_{nc}^\star$) is difficult. This is because a node may receive simultaneous transmissions from two nodes which are not within communication range of each other. As explained before, listen and transmit events are rare within energy-constrained nodes. Therefore, the likelihood of simultaneous transmissions is small and it is expected to have minimal impact on the throughput.

We present both an upper bound $\overline{\mathcal{T}_{nc}^\star}$ and a lower bound $\underline{\mathcal{T}_{nc}^\star}$ on the maximum groupput in non-clique topologies. In the scenarios where $\overline{\mathcal{T}_{nc}^\star}$ and $\underline{\mathcal{T}_{nc}^\star}$ are the same, the exact maximum groupput $\mathcal{T}_{nc}^\star$ can be obtained. The lower bound $\underline{\mathcal{T}_{nc}^\star}$ is obtained by solving (P2) but replace constraint (4.12) by

$$\alpha_i \leq \sum_{i \in \mathcal{N}(i)} \beta_j, \ \forall i \in \mathcal{N},$$

where $\mathcal{N}(i)$ is the set of neighboring nodes of node i. This ensures that the fraction of time node i listens cannot exceed the sum of its neighboring nodes' fractions of transmissions. The upper bound $\overline{\mathcal{T}_{nc}^\star}$ is obtained by solving (P2) in which the constraint (4.11) is removed. This allows overlapping transmissions which can possibly happen in non-cliques. Numerical results show that with certain topologies, $\overline{\mathcal{T}_{nc}^\star} = \underline{\mathcal{T}_{nc}^\star}$ holds, resulting in the exact maximum groupput $\mathcal{T}_{nc}^\star$. In Section 4.5.6, we compute $\mathcal{T}_{nc}^\star$ and evaluate the performance of EconCast in non-clique topologies.

Node	1	2	3	4
Power Budget: ρ_i (mW)	0.005	0.01	0.05	0.1
Awake (%): $\alpha_i^\star + \beta_i^\star$	0.5	1.0	5.0	10.0
Transmit when Awake (%)	20.0	22.0	53.6	65.7

Table 4.2: A simple example in a heterogeneous network.

4.3 Distributed Protocol

In this section, we describe EconCast from the perspective of a single node that transitions between sleep, listen, and transmit states, under a power budget. Since we focus on a single node i, in parts of this section, we drop the subscript i of previously defined variables for notational compactness.

4.3.1 A Simple Heterogeneous Example

To better understand the challenges faced in designing EconCast, consider a simple example of 4 nodes, all having identical listen and transmit power consumption $L_i = X_i = 1\,\mathrm{mW}$ ($i = 1, 2, 3, 4$), but different power budgets ρ_i, as indicated in Table 4.2. Table 4.2 also shows the percentage of time each node spends in listen and transmit states $(\alpha_i^\star, \beta_i^\star)$ ($i = 1, 2, 3, 4$) such that the groupput is maximized by solving (P1). It also shows the percentage of time each node spends in transmit state *when awake* (i.e., $\frac{100 \cdot \beta_i^\star}{\alpha_i^\star + \beta_i^\star}\%$).

If, instead, all nodes have the same power budget of $\rho_i = 0.1\,\mathrm{mW}$, the percentage of time each node spends in transmit state when awake is 25% (with $\alpha_i^\star = 0.075$, $\beta_i^\star = 0.025$, $i = 1, 2, 3, 4$). Note that in the above example, the power budget of node 4 remains unchanged but changes in other nodes' power budgets shift the percentage of time it should transmit when awake from 25% to 65.7%. This clearly shows that the partitioning of a node's power budget among listen and transmit states is highly dependent on other nodes' properties. *However, we will show that if a node does not know the properties of its neighbors, an optimal configuration can be obtained without explicitly solving (P1).*
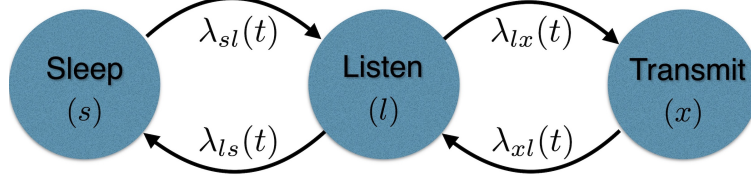
Figure 4.1: The node's states and transition rates.

4.3.2 Protocol Description

To clearly present EconCast, we start from a theoretical framework and slowly build on it to address practicalities. As mentioned in Section 4.1, a node can be in one of three states: sleep (s), listen (l), and transmit (x). As depicted in Fig. 4.1, it must pass through listen state to transition between sleep and transmit states. The time duration a node spends in state u before transitioning to state v is exponentially distributed with rate $\lambda_{uv}(t)$. These *transition rates* can be adjusted over time. We remark that sending packets with exponentially distributed length (i.e., a node transitions from transmit state to listen state with a rate λ_{xl}) is impractical. However, it can be shown that this is equivalent to continuously transmitting back-to-back unit-length packets with probability $(1 - \lambda_{xl})$ if $\lambda_{xl} \in [0, 1]$, which is indeed the case in EconCast.

To maximize the groupput or anyput, EconCast can operate in *groupput mode* or *anyput mode*, respectively. The throughput as a function of $\pi_{\mathbf{w}}$ (see (4.5)) is controlled by appropriately adjusting the transition rates between different states of each node. EconCast determines in a distributed manner how these adjustments are performed over time. Roughly speaking, each node adjusts its transition rates $\lambda_{uv}(t)$ based on limited information that can be obtained in practice, including

- Its power consumption levels in listen and transmit states, L and X, and energy storage level $b(t)$.

- A sensing of transmit activity of other nodes over the channel (CSMA-like carrier sensing).

- A count of other active listeners (in groupput mode), $c(t)$, or an indicator of whether there are any active listeners (in anyput mode), $\gamma(t)$. In practice, $c(t)$ and $\gamma(t)$ may not be accurate. We denote $\hat{c}(t)$ and $\hat{\gamma}(t)$ as their estimated values.

Note that in EconCast, unlike in previous work such as Panda [17], each node *does not* need to know the number of nodes in the network, N, and the power budgets and consumption levels of other nodes. Furthermore, a node *does not* need to know its power budget ρ explicitly (e.g., in the case of energy harvesting [69]), although this knowledge can be incorporated, if available.

Under EconCast, a node sets $\lambda_{sl}(t)$ as an increasing function of the available stored energy, $b(t)$, to more aggressively exit sleep state. Furthermore, it sets $\lambda_{lx}(t)$ as an increasing function of the number of listeners, $\hat{c}(t)$, to enter transmit state more frequently when more nodes are listening. We will describe how these functions are chosen in Section 4.3.5.

4.3.3 Estimating Active Listeners: Pings

We now discuss the estimation of $\hat{c}(t)$ or $\hat{\gamma}(t)$. Recall from Section 4.1 that nodes can send out periodic *pings* that any other listener can receive. The pings need not carry any explicit information and are potentially significantly cheaper and shorter than control packet transmissions (e.g., an ACK). Therefore, they consume less power and take much less time than a minimal data transmission.

Consider the case in which all nodes are required to send pings at a pre-determined rate and the power consumption is accounted for in the listening power consumption L. In such a case, a fellow listener detecting such pings (e.g., using a simple energy detector) can use the count of such pings in a given period of time, or the inter-arrival times of pings, to estimate the number of active listeners $c(t)$. Estimating $\gamma(t)$ is even easier by detecting the existence of any ping. Although poor estimates are expected to affect the throughput achieved by EconCast, this performance degradation is not significant, as we will show in Section 4.5.5.

4.3.4 Two Variants of EconCast

We now address the incorporation of the estimates $\hat{c}(t)$ and $\hat{\gamma}(t)$ into EconCast. We present two versions of EconCast which only differ when a node is in transmit state:

- EconCast-C (the capture version): a node may "capture" the channel and transmit for an exponential amount of time (i.e., several back-to-back packets). When each packet transmission is completed, the transmitter listens for pings for a fixed-length *pinging interval*. Each successful recipient of the transmission initiate one ping at time chosen uniformly at random on this interval. The transmitter then estimates $\hat{c}(t)$ or $\hat{\gamma}(t)$ based on the count of pings received and adjusts $\lambda_{xl}(t)$ (as described in Section 4.3.5). In Section 4.6.3, we discuss the experimental implementation of this process.

- EconCast-NC (the non-capture version): a node always releases the channel after one packet transmission. Each node continuously pings and receives pings from other nodes when listening, estimates $\hat{c}(t)$ or $\hat{\gamma}(t)$, and adjusts $\lambda_{lx}(t)$ (as described in Section 4.3.5).

EconCast-C is significantly easier to implement since the estimates are only needed for the transmitter right after each packet transmission. The probability that the same transmitter will continue transmitting depends on the estimates $\hat{c}(t)$ or $\hat{\gamma}(t)$. Therefore, our implementation and experimental evaluations in Section 4.6 focus on EconCast-C.

4.3.5 Setting Transition Rates

Consider a node running EconCast. Time is broken into intervals of length τ_k $(k = 1, 2, \cdots)$. The k-th interval is from time t_{k-1} to time t_k and we let $t_0 = 0$. EconCast takes input of two internal variables:

- η is a multiplier which is updated at the beginning of each time interval. Let $b[k]$ $(k = 0, 1, \cdots)$ denote the energy storage level at the end of the k-th time interval. Let $(\cdot)^+$ denote $\max(0, \cdot)$ and $\eta[k]$ is updated as follows

$$\eta[k] = \left(\eta[k-1] - \delta_k/\tau_k \cdot (b[k] - b[k-1]) \right)^+, \tag{4.17}$$

in which $\delta_k \in (0, 1)$ is a step size and $b[k] = b(t_k)$. We use square brackets here to imply that the multiplier $\eta[k]$ remains constant for $t \in [t_k, t_{k+1})$.

- $A(t)$ is the carrier sensing indicator of a node, which is 1 when the node does not sense

63

any ongoing transmission, and is 0 otherwise. Carrier sensing forces a node to "stick" to its current state. When receiving an ongoing transmission, a node in listen state will not exit the listen state until it finishes receiving the full transmission, and a node in sleep state will not leave the sleep state (i.e., it enters the listen state but immediately leaves when it senses the ongoing transmission).

The transition rates are described as follows (the superscripts C and N denote EconCast-C and EconCast-NC), in which the two throughput modes only differ in $\lambda_{xl}(t)$ (for the capture version) or in $\lambda_{lx}(t)$ (for the non-capture version). For *groupput* maximization, at any time t in the k-th interval,

$$\lambda_{sl}(t) = A(t) \cdot \exp(-\eta[k]L/\sigma), \tag{4.18a}$$

$$\lambda_{ls}(t) = A(t), \tag{4.18b}$$

$$\lambda_{lx}^{C}(t) = A(t) \cdot \exp(\eta[k](L - X)/\sigma), \tag{4.18c}$$

$$\lambda_{lx}^{N}(t) = A(t) \cdot \exp(\eta[k](L - X)/\sigma + \hat{c}(t)/\sigma), \tag{4.18d}$$

$$\lambda_{xl}^{C}(t) = \exp(-\hat{c}(t)/\sigma), \tag{4.18e}$$

$$\lambda_{xl}^{N}(t) = 1. \tag{4.18f}$$

For *anyput* maximization, $\hat{c}(t)$ is replaced with $\hat{\gamma}(t)$. Theorem 4.1 below states the main result of this chapter and the proof is in Section 4.4.

Theorem 4.1 *Let $\sigma \to 0$ and select parameters δ_k and τ_k properly (e.g., $\delta_k = 1/[(k+1)\log(k+1)]$ and $\tau_k = k$). Under perfect knowledge of $c(t)$ or $\gamma(t)$, the average throughput of EconCast ($\mathcal{T}_g$ or $\mathcal{T}_a$) converges to the oracle throughput ($\mathcal{T}_g^\star$ or $\mathcal{T}_a^\star$) given by (P1).*

4.3.6 Stability and Choice of σ, δ_k, and τ_k

EconCast is adaptive and, as expected, it must deal with the tradeoff of adapting "quickly but poorly" to "optimally but slowly". This adaptation manifests itself into the parameters σ, δ_k, and τ_k. When σ is decreased, the throughput increases, as we will describe in Section 4.4. However,

the *burstiness* also increases with respect to decreased σ. The burstiness is a characteristic of communication involving multiple packets that are successfully received in bursts. In Section 4.5, we describe how the burstiness can be analyzed and measured.

Under a given value of σ, each node continuously adjusts the rates $\lambda_{uv}(t)$ based on its multiplier η according to (4.17), which is a function of the ratio δ_k/τ_k. Small δ_k/τ_k ratios (e.g., small step sizes and large intervals) make smaller changes of η over time, and lead to longer convergence time to the "right" multiplier values. In contrast, larger δ_k/τ_k ratios make η oscillate more wildly near the optimal value, such that the performance of EconCast is further from the optimal. Careful choices of the parameters is required for guaranteed convergence (as stated in Theorem 4.1). In practice, as we set for the simulations and experimentation, $\delta_k = \delta$ and $\tau_k = \tau$ for some small constant δ and large constant τ are sufficient.

4.4   Proof of Theorem 4.1

The proof of Theorem 4.1 is based on a Markov Chain Monte Carlo (MCMC) approach [62, 146] from statistical physics and consists of three parts: (i) We compute the steady state distribution of the network Markov chain under EconCast with fixed Lagrange multiplier vector $\boldsymbol{\eta} = [\eta_i]$; (ii) We present an alternative concave optimization problem whose optimal value approaches that of (P1) as $\sigma \to 0$ and show that the steady state distribution of EconCast is indeed the optimal solution to this alternative optimization problem when the Lagrange multipliers are chosen optimally; (iii) We show that nodes operating EconCast update their Lagrange multipliers locally according to a "noisy" gradient descent. Based on a stochastic approximation argument, the updated Lagrange multipliers will converge to their optimal values with proper choices of step sizes and interval lengths as given in Theorem 4.1.

Part (i): Steady State Distribution. The following lemma describes the network state distribution generated by EconCast when $\boldsymbol{\eta}$ freezes.

Lemma 4.2 *With fixed $\boldsymbol{\eta}$, the network Markov chain, resulted from overall interactions among the*

Algorithm 4.1 Gradient Descent Algorithm

Input parameters: σ, $\boldsymbol{\rho}$, $\mathbf{L}$, and $\mathbf{X}$
Initialization: $\alpha_i(0) = \beta_i(0) = \eta_i(0) = 0$, $\forall i \in \mathcal{N}$
  1: **for** $k = 1, 2, \cdots$ **do**
  2:    $\delta(k) = 1/k$, compute $\boldsymbol{\pi}(k)$ from (4.19) using $\boldsymbol{\eta} = \boldsymbol{\eta}(k)$
  3:    **for** $i = 1, 2, \cdots, N$ **do**
  4:        Update $\eta_i(k)$, $\alpha_i(k)$, and $\beta_i(k)$ according to (4.23) and (4.24)
---

nodes according to the transition rates (4.18), has the steady state distribution

$$\pi_{\mathbf{w}}^{\boldsymbol{\eta}} = \frac{1}{Z^{\boldsymbol{\eta}}} \exp \left[\frac{1}{\sigma} \left(\mathcal{T}_{\mathbf{w}} - \sum_{i:w_i=l} \eta_i L_i - \sum_{i:w_i=x} \eta_i X_i \right) \right], \tag{4.19}$$

where $Z^{\boldsymbol{\eta}}$ is a normalizing constant so that $\sum_{\mathbf{w} \in \mathcal{W}} \pi_{\mathbf{w}}^{\boldsymbol{\eta}} = 1$.

Proof 4.2 *The proof can be found in Appendix 4.8.2.*

Part (ii): An Alternative Optimization. We then present an optimization problem (P4) as follows

$$(\mathsf{P4}) \max_{\boldsymbol{\pi}} \quad \sum_{\mathbf{w} \in \mathcal{W}} \pi_{\mathbf{w}} \mathcal{T}_{\mathbf{w}} - \sigma \sum_{\mathbf{w} \in \mathcal{W}} \pi_{\mathbf{w}} \log \pi_{\mathbf{w}} \tag{4.20}$$
$$\text{subject to} \quad (4.6), \ (4.7), \text{and } (4.8),$$

where σ is the positive constant used in EconCast (the counterpart in statistical physics is the temperature in systems of interacting particles). Note that (P4) is a concave maximization problem and as $\sigma \to 0$, the optimal value of (P4) approaches that of (P1). To solve (P4), consider the Lagrangian $\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\eta})$ formulated by moving the power constraint (4.6) into the objective (4.20) with a Lagrange multiplier $\eta_i \geq 0$ for each node i, i.e.,

$$\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\eta}) = \sum_{\mathbf{w} \in \mathcal{W}} \pi_{\mathbf{w}} \mathcal{T}_{\mathbf{w}} - \sigma \sum_{\mathbf{w} \in \mathcal{W}} \pi_{\mathbf{w}} \log \pi_{\mathbf{w}} - \sum_{i \in \mathcal{N}} \left[\eta_i (\alpha_i L_i + \beta_i X_i - \rho_i) \right]. \tag{4.21}$$

In view of (4.7) and (4.8), it can be shown that with fixed $\boldsymbol{\eta}$, the optimal $\boldsymbol{\pi}^{\boldsymbol{\eta}} = [\pi_{\mathbf{w}}^{\boldsymbol{\eta}}]$ that maximizes $\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\eta})$ is exactly given by (4.19). Therefore, if EconCast knows the optimal Lagrange multiple vector $\boldsymbol{\eta}^{\star}$, it can start with $\boldsymbol{\eta}^{\star}$ and the steady state distribution generated by EconCast will converge to the optimal solution to (P4).

In order to find $\boldsymbol{\eta}^\star$, consider the dual function $\mathcal{D}(\boldsymbol{\eta}) = \mathcal{L}(\boldsymbol{\pi}^{\mathbf{n}}, \boldsymbol{\eta})$ over $\boldsymbol{\eta} \succeq \mathbf{0}$ (here $\mathbf{0}$ is an N-dimensional zero vector and $\succeq$ denotes component-wise inequality). Interestingly, it can be shown that the partial derivative of $\mathcal{D}(\boldsymbol{\eta})$ with respect to η_i is simply given by

$$\partial \mathcal{D}/\partial \eta_i = \rho_i - (\alpha_i L_i + \beta_i X_i), \tag{4.22}$$

which is the difference between the power budget ρ_i and the average power consumption of node i. Therefore, the dual can be minimized by using a gradient descent algorithm with inputs of step size $\delta_k > 0$, $\boldsymbol{\rho}$, $\mathbf{L}$, and $\mathbf{X}$, which generates a state probability $\boldsymbol{\pi}(k)$ ($k = 1, 2, \cdots$). This algorithm is described in Algorithm 4.1 along with the following equations

$$\eta_i(k) = \left[\eta_i(k-1) - \delta_k \cdot (\rho_i - \alpha_i(k)L_i - \beta_i(k)X_i) \right]^+, \tag{4.23}$$

$$\alpha_i(k) = \sum_{\mathbf{w} \in \mathcal{W}_i^l} \pi_{\mathbf{w}}^{\boldsymbol{\eta}(k)}, \ \beta_i(k) = \sum_{\mathbf{w} \in \mathcal{W}_i^x} \pi_{\mathbf{w}}^{\boldsymbol{\eta}(k)}. \tag{4.24}$$

Hence, with the right choice of step size δ_k (e.g., $\delta_k = 1/k$), $\boldsymbol{\pi}(k)$ converges to the optimal solution to (P4).

Part (iii): Convergence Analysis. To arrive at a distributed solution, instead of computing the quantities α_i and β_i directly according to (4.24) (which is centralized with high complexity), nodes operating EconCast approximate the difference between the power budget and the average power consumption (4.22) by observing the dynamics of their energy storage level. Specifically, each node i can update its Lagrange multiplier $\eta_i(k)$ based on the difference between its energy storage levels at the end and the start of an interval of length τ_k, divided by τ_k, as described by (4.17). Therefore, η_i is updated according to a "noisy" gradient descent. However, it follows from stochastic approximation (with Markov modulated noise) that by choosing step sizes and interval lengths as given in Theorem 4.1, these noisy updates will converge to $\boldsymbol{\eta}^\star$ as $k \to \infty$ (see e.g., Theorem 1 of [147]). As mentioned in Section 4.3.6, the choice of parameters σ, δ_k, and τ_k will affect the tradeoff between convergence time and the performance of EconCast. The detailed proof uses similar techniques as in the proof of [147, Theorem 1] with minor modifications and can be found in the technical report [145].

4.5 Numerical Results

In this section, we evaluate the throughput and latency performance of EconCast when operating in groupput and anyput modes. We use the following notation: (i) $\mathcal{T}_g^\star$ ($\mathcal{T}_a^\star$) is the oracle groupput (anyput) obtained by solving (P1) or, equivalently, (P2), (ii) $\mathcal{T}_g^\sigma$ ($\mathcal{T}_a^\sigma$) is the achievable groupput (anyput) of EconCast with a given value of σ obtained by solving (P4), and (iii) $\widetilde{\mathcal{T}_g^\sigma}$ ($\widetilde{\mathcal{T}_a^\sigma}$) is the groupput (anyput) of EconCast obtained via simulations with a given value of σ. For brevity, we ignore the subscripts of $\mathcal{T}^\sigma$ when describing results that are general for both groupput and anyput.

4.5.1 Setup

We consider clique networks[5] with $\sigma \in \{0.1, 0.25, 0.5\}$. The nodes' power budgets and consumption levels correspond to the energy harvesting budgets and ultra-low-power transceivers in [136, 144, 148]. Note that the performance of EconCast only depends on the *ratio* between the listen or transmit power and the power budget. For example, nodes with $\rho = 10\,\text{uW}, L = X = 500\,\text{uW}$ behave exactly the same as nodes with $\rho = 1\,\text{mW}, L = X = 50\,\text{mW}$. Therefore, the oracle throughput applies and EconCast can operate in very general settings.

In the simulations, each node has a constant power input at the rate of its power budget, and adjusts the transition rates based on the dynamics of its energy storage level. Since nodes perform carrier sensing when waking up, there are no simultaneous transmissions and collisions. We also assume that the packet length is $1\,\text{ms}$ and that nodes have accurate estimate of the number of listeners or the existence of any active listeners, i.e., $\hat{c}(t) = c(t)$ or $\hat{\gamma}(t) = \gamma(t)$.

The simulation results show that $\widetilde{\mathcal{T}^\sigma}$ perfectly matches $\mathcal{T}^\sigma$ for $\sigma \in \{0.25, 0.5\}$. For $\sigma = 0.1$, $\widetilde{\mathcal{T}^\sigma}$ does not converge to $\mathcal{T}^\sigma$ within reasonable time due to the bursty nature of EconCast, as will be described in Section 4.5.4. Therefore, we evaluate the throughput performance of EconCast by comparing $\mathcal{T}^\sigma$ to $\mathcal{T}^\star$ with varying σ in both heterogeneous and homogeneous networks. Specifically, *homogeneous networks* consist of nodes with the same power budget and consumption levels,

---

[5]We evaluate the throughput performance of EconCast-C in non-clique topologies in 4.5.6.
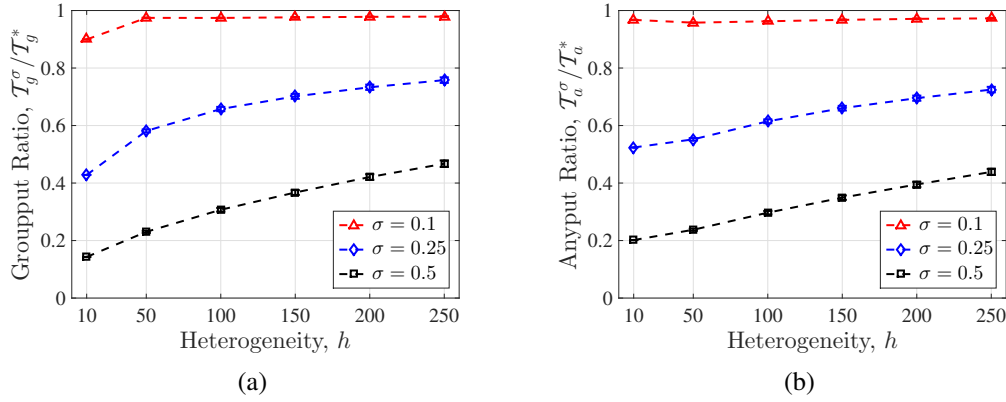
Figure 4.2: Sensitivity of the achievable throughput normalized to the oracle throughput, $\mathcal{T}^\sigma/\mathcal{T}^\star$, for: (a) groupput and (b) anyput, to the heterogeneity of the power budget, ρ, and power consumption levels, L and X.

i.e., $\rho_i = \rho, L_i = L, X_i = X, \forall i \in \mathcal{N}$. The simulation results also confirm that nodes running EconCast consume power on average at the rate of their power budgets.

4.5.2 Heterogeneous Networks – Throughput

One strength of EconCast is its ability to deal with *heterogeneous* networks. Fig. 4.2 shows the groupput and anyput achieved by EconCast normalized to the corresponding oracle groupput and anyput (i.e., $\mathcal{T}^\sigma/\mathcal{T}^\star$), for heterogeneous networks with $N = 5$ and $\sigma \in \{0.1, 0.25, 0.5\}$. Intuitively, higher values of $\mathcal{T}^\sigma/\mathcal{T}^\star$ indicate better performance of EconCast.

Along the x-axis, the *network heterogeneity*, denoted by h, is varied from 10 to 250 at discrete points. The relationship between the network heterogeneity and the values of h is as follows: (i) for each node i, L_i and X_i are independently selected from a uniform distribution on the interval $[510 - h, 490 + h]$ uW, (ii) for each node i, a variable h' is first sampled from the interval $[-\log\frac{h}{100}, \log h]$ uniformly at random, and then ρ_i is set to be $\exp(h')$. Therefore, the energy budget ρ_i varies from $100/h$ to h uW. As a result, for any h, L_i and X_i have mean values of 500 uW, and ρ_i has median of 10 uW but its mean increases as h increases. Note that a homogeneous network is represented by $h = 10$.

The y-axis indicates for each value of h, the mean and the 95% confidence interval of the

ratios $\mathcal{T}^{\sigma}/\mathcal{T}^{\star}$ averaged over 1,000 heterogeneous network samples. Specifically, in each network sample, each node i samples (ρ_i, L_i, X_i) according to the processes described above. Figs. 4.2(a) and 4.2(b) show that the network heterogeneity with respect to both the nodes' power budgets and consumption levels increases as h increases. Fig. 4.2 also shows that the throughput ratio $\mathcal{T}^{\sigma}/\mathcal{T}^{\star}$ increases as σ decreases, and approaches 1 as $\sigma \to 0$. Furthermore, with increased heterogeneity of the network, the throughput ratio has little dependency on the network heterogeneity h but heavy dependency on σ. In general, the groupput and anyput ratios are similar except for homogeneous networks ($h = 10$). In such networks, the anyput ratio is slightly higher than the groupput ratio. This is due to the fact that nodes have the same values of ρ_i, L_i, and X_i). Therefore, determining the existence of any active listeners, $\gamma(t)$, is easier than determining the number of active listeners, $c(t)$.

4.5.3 Homogeneous Networks – Throughput and Comparison to Related Work

We now evaluate the throughput of EconCast in *homogeneous* networks with $N = 5$, $\rho = 10\,\mathrm{uW}$, $L + X = 1\,\mathrm{mW}$, and $\sigma \in \{0.1, 0.25, 0.5\}$. We also compare the groupput achieved by EconCast to three protocols in related work: *Panda* [17], *Birthday* [55], and *Searchlight* [50], which operate under stricter assumptions than EconCast. In particular:

- The probabilistic protocols Panda and Birthday both require a homogeneous set of nodes and a priori knowledge of N. The throughput of Panda and Birthday is computed as described in [17] and [55], respectively.

- The deterministic protocol Searchlight is designed for minimizing the worst case pairwise discovery latency, which does not directly address multi-party communication across a shared medium. However, the discovery latency is closely related to the throughput, since the inverse of the average latency is the throughput. Hence, maximizing throughput is equivalent to minimizing the average discovery latency. We derive an upper bound on the throughput of Searchlight by multiplying the pairwise throughput by $(N - 1)$. This is assuming that all other $(N - 1)$ nodes will be receiving when one node transmits. However, in practice the

70
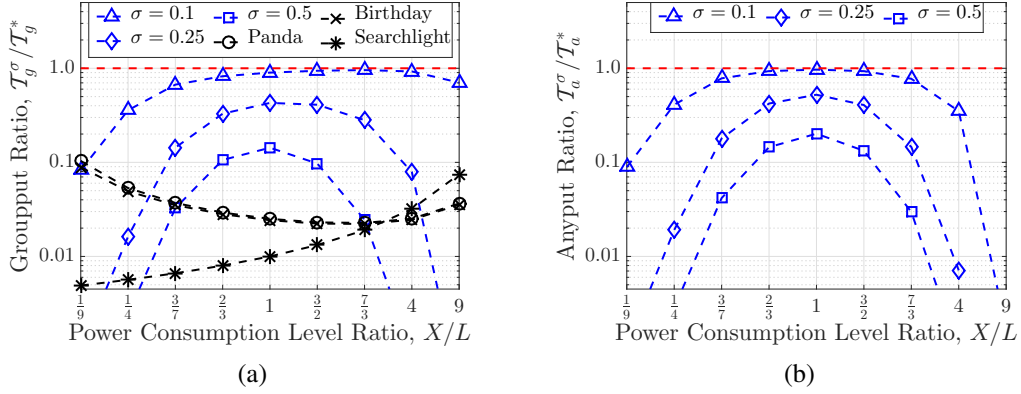
Figure 4.3: Throughput performance of EconCast-C when operating in: (a) groupput mode and (b) anyput mode, with $N = 5$, $\rho = 10\,\text{uW}$, $L + X = 1\,\text{mW}$, and $\sigma \in \{0.1, 0.25, 0.5\}$, as a function of X/L.

throughput is likely to be lower unless all the nodes are synchronized and coordinated.

Figs. 4.3(a) and 4.3(b) present, respectively, the groupput and anyput achieved by EconCast normalized to the oracle groupput and anyput, as a function of the power consumption ratio X/L, with $N = 5$, $\rho = 10\,\text{uW}$, and $L + X = 1\,\text{mW}$. Fig. 4.3(a) also presents the throughput achieved by Panda, Birthday, and Searchlight[6] protocols. The horizontal dashed lines at 1 represent the oracle groupput and anyput. Note that with $L = X = 500\,\text{uW}$, the ratio $\mathcal{T}_g^\sigma / \mathcal{T}_g^\star$ achieved by EconCast outperforms that of Panda by $6\times$ and $17\times$ with $\sigma = 0.5$ and $\sigma = 0.25$, respectively. In particular, the groupput ratio $\mathcal{T}_g^\sigma / \mathcal{T}_g^\star$ significantly outperforms that of prior art for $X \approx L$. The simulation results, which will be discussed later, also verify this throughput improvement.

Fig. 4.3 shows that for a given value of X/L, $\mathcal{T}^\sigma$ approaches $\mathcal{T}^\star$ with decreasing σ, as expected (see Section 4.3). Moreover, for each value of σ, the throughput ratio $\mathcal{T}^\sigma / \mathcal{T}^\star$ increases as the power consumption ratio X/L is closer to 1. This is realistic for current commercial low-power radios that have symmetric power consumption levels in listen and transmit states. This is due to the fact that (i) with small X/L values, nodes enter transmit state infrequently, since listening is expensive and they must pass the listen state to enter the transmit state, and (ii) with large X/L values, nodes spend energy transmitting even when there are no other nodes listening (e.g.,

---

[6]For Searchlight protocol, we compare its throughput upper bound to $\mathcal{T}_g^\star$.
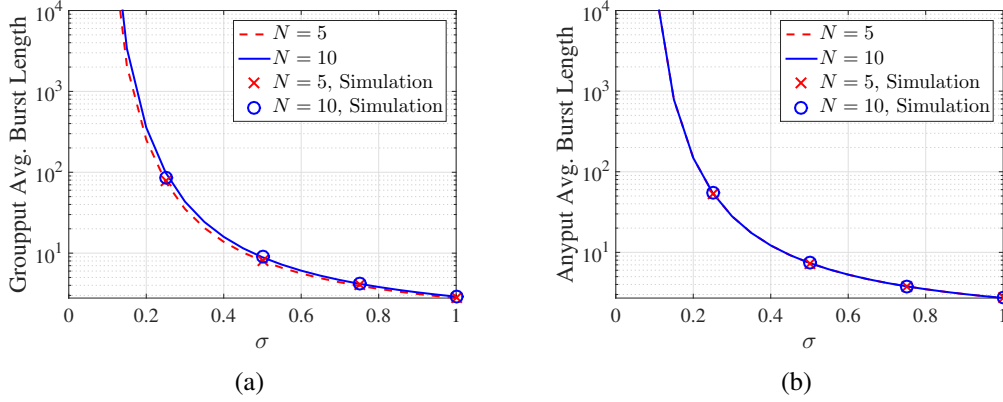
Figure 4.4: Analytical (curves) and simulated (markers) average burst length of EconCast-C when operating in: (a) groupput mode and (b) anyput mode, with $N \in \{5, 10\}$, $\sigma \in \{0.25, 0.5\}$, $\rho = 10\,\text{uW}$, and $L = X = 500\,\text{uW}$.

$c(t) = 0$). In particular, anyput degrades with large X/L values, since anyput depends on the *existence* of any active listeners when some node is transmitting. Therefore, when listening is expensive, the fact that multiple nodes listen simultaneously does not improve anyput. We believe that any distributed protocol will suffer from such performance degradation since, unlike Panda, Birthday, and Searchlight, nodes in a fully distributed setting *do not* have any information about the properties of other nodes in the network.

4.5.4 Burstiness and Latency

The results until now suggest allowing $\sigma \to 0$. While reducing σ improves throughput, it considerably increases the *burstiness* of communication, as mentioned in Section 4.3. We measure the burstiness by the *average burst length*, which is defined as the average number of back-to-back packets successfully received by a node without exiting the listen state. It is not hard to see that the average burst length in the anyput mode is independent of the network size, N, and only depends on σ, while the average burst length in the groupput mode depends on both N and σ (see Appendix 4.8.3 for the detailed analytical derivation). In general, increased burstiness means that the long-term average throughput can be achieved under the given power budgets, but the variance of the throughput is more significant over short term intervals.
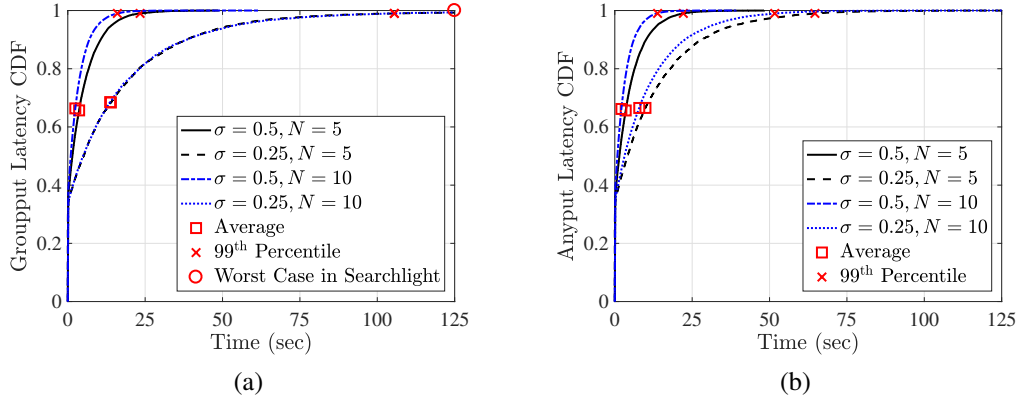
Figure 4.5: The CDF, mean, and 99th percentile latency of EconCast-C when operating in: (a) groupput mode and (b) anyput mode, with $N \in \{5, 10\}$, $\sigma \in \{0.25, 0.5\}$, $\rho = 10\,\text{uW}$, and $L = X = 500\,\text{uW}$.

Figs. 4.4(a) and 4.4(b) show the average burst length of EconCast (in log scale) when operating in groupput and anyput modes, respectively, in homogeneous networks with $N \in \{5, 10\}$, $\rho = 10\,\text{uW}$, $L = X = 500\,\text{uW}$, and varying σ. Values are obtained using the analytical results (4.27)–(4.28) derived in Appendix 4.8.3 (curves) and contrasted with simulations at specific values of σ (markers). Aside from showing that the simulation results and the analytical results are well matched, Fig. 4.4 also demonstrates that reducing σ dramatically increases burstiness. For example, with $\sigma = 0.25$ and $N = 10$, a node running EconCast in groupput mode has an average burst length of 85, and this value is increased to 4×10^5 with $\sigma = 0.1$. This explains why $\widetilde{\mathcal{T}_g^\sigma}$ does not converge within reasonable time with $\sigma = 0.1$ (see Section 4.5.1). Comparing Fig. 4.4(a) with Fig. 4.4(b), it can be seen that the groupput average burst length increases more rapidly than the anyput average burst length as σ decreases. Moreover, Fig. 4.4(b) shows that the anyput average burst length is independent of N, which corresponds to the analysis in Appendix 4.8.3. The reason is that the burst length of EconCast in anyput mode only depends on $\gamma(t)$, which always equals to 1, if the transmission is successful. We remark that reducing the burstiness is a subject of future work.

A second metric of interest is the communication *latency*, which we define as the average time interval between consecutive bursts that are successfully received by a node from some other

node. This metric, together with the burstiness, can fully characterize the throughput achieved by EconCast. Particularly, we can divide time into cycles, where each cycle contains a successfully received burst and the interval followed from this burst until the start of the next successfully received burst. Then, in homogeneous networks, the throughput delivered to a single node is equal to the ratio between the average burst length and the average cycle length.

Figs. 4.5(a) and 4.5(b) present the CDF latency of EconCast when operating in groupput and anyput modes obtained via simulations, and indicate both the average and the 99th percentile latency values. The homogeneous networks considered are with $N \in \{5, 10\}$, $\sigma \in \{0.25, 0.5\}$, $\rho = 10\,\mathrm{uW}$, and $L = X = 500\,\mathrm{uW}$. Fig. 4.5(a) also shows the pairwise worst case latency of Searchlight computed from [50] under the same power budget and consumption levels.[7]

Fig. 4.5 shows that the latency of EconCast in both groupput and anyput modes (i) increases as σ decreases and (ii) decreases as N increases. The first observation is due to the fact that reducing σ increases the burstiness, which causes higher variation in the nodes' energy storage levels, and thus longer sleep durations. The second observation is due to the fact that as the network size increases, each node is more likely to receive a packet from some other node in the network (note that we do not specify the source of the packet).

Comparing Fig. 4.5(a) with Fig. 4.5(b), it is observed that EconCast operating in anyput mode has slightly lower average latency than in groupput mode. However, with a smaller σ value (i.e., $\sigma = 0.25$), the 99th percentile latency of EconCast when operating in anyput mode is significantly lower than that in groupput mode. This results from the fact that the average burst length of EconCast in anyput mode depends on the existence of any listening nodes, whose value is always less than or equal to the number of listening nodes considered in groupput mode. For all parameters considered, the 99th percentile groupput latency is within 110 seconds, outperforming the Searchlight pairwise worst case latency bound of 125 seconds. Note that although EconCast has a non-zero probability of having any latency, its latency is below the worst case latency of Searchlight in most cases (over 99%).

---

[7]This is computed with slot length of 50 ms and a beacon (packet) length of 1 ms as was done in [52].
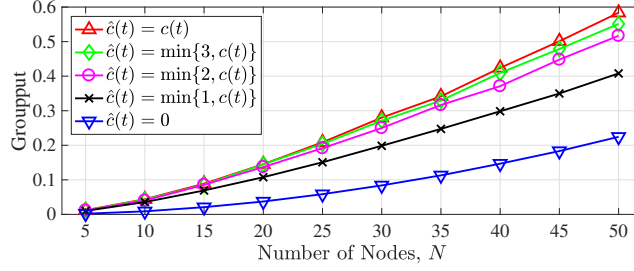
Figure 4.6: Simulated throughput of EconCast-C when operating in groupput mode based on both accurate and inaccurate estimates of the number of listeners, $\hat{c}(t)$, with $\sigma = 0.5$, $\rho = 10\,\text{uW}$, $L = X = 500\,\text{uW}$, and varying N.

4.5.5 Groupput Sensitivity to the Estimated Number of Listeners

As described in Section 4.3.3, the performance of EconCast in groupput mode depends on the estimated number of listeners, $\hat{c}(t)$. Estimating $c(t)$ becomes more challenging in larger networks, due the possibility of more pings being sent during the pinging interval (see Section 4.1.2). However, inaccurate estimates do not significantly degrade the groupput performance of EconCast. To explore the effect of inaccurate estimate $\hat{c}(t)$, we evaluate EconCast in the groupput mode while $\hat{c}(t) \neq c(t)$. Fig. 4.6 presents the groupput achieved by EconCast via simulations based on both accurately and inaccurately estimated number of listeners. The homogeneous networks considered are with varying N, $\sigma = 0.5$, $\rho = 10\,\text{uW}$, and $L = X = 500\,\text{uW}$.

Fig. 4.6 shows that the groupput achieved by EconCast using estimate $\hat{c}(t) = \min\{2, c(t)\}$ is very close to the groupput, $\widetilde{\mathcal{T}_g^\sigma}$, achieved using the ground truth $\hat{c}(t) = c(t)$. Moreover, the groupput is within 30% of $\widetilde{\mathcal{T}_g^\sigma}$ with $\hat{c}(t) = \min\{1, c(t)\}$ for all values of N considered. This implies that EconCast still has good groupput performance with very rough estimates.[8] Note that such a rough estimate can be obtained easily by a simple energy detector even when multiple pings collide since the transmitter does not need to correctly decode each ping individually. Therefore, the groupput performance of EconCast is not very sensitive to the estimated number of listeners, $\hat{c}(t)$.

---

[8]Setting $\hat{c}(t) = \min\{1, c(t)\}$ is similar to the anyput mode in which the transmitter only detects the existence of any listeners.
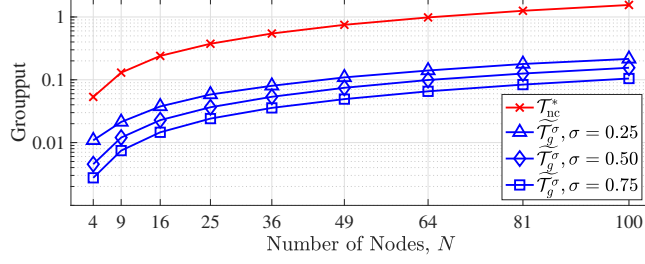
Figure 4.7: The oracle groupput, $\mathcal{T}_{\mathrm{nc}}^{\star}$, and the throughput of EconCast-C when operating in group-put mode obtained via simulations, $\widetilde{\mathcal{T}_g^{\sigma}}$, in grid topologies with varying N, $\sigma \in \{0.25, 0.5, 0.75\}$, $\rho = 10\,\mathrm{uW}$, and $L = X = 500\,\mathrm{uW}$.

4.5.6 Groupput Evaluation in Non-clique Topologies

We now compute the oracle groupput for non-clique topologies (derived in Section 4.2.3) and evaluate the groupput achieved by EconCast in such scenarios via simulations with accurate estimate $\hat{c}(t) = c(t)$. Since simultaneous transmissions can happen in non-clique topologies, none of the transmissions will be counted as throughput in the simulations.

We use grid topologies with varying number of nodes, N, in which each node has at most 4 neighbors. For example, $N = 25$ represents a 5×5 grid. Fig. 4.7 presents the oracle groupput, $\mathcal{T}_{\mathrm{nc}}^{\star}$, for grid topologies, and the throughput achieved by EconCast in groupput mode obtained via simulations with with varying σ and N. Note that for all the grid topologies considered, the upper and lower bounds of $\mathcal{T}_{\mathrm{nc}}^{\star}$ (see Section 4.2.3) are the same, providing the exact oracle groupput.

Fig. 4.7 shows that EconCast achieves 14% – 22% of the oracle groupput, $\mathcal{T}_{\mathrm{nc}}^{\star}$, with $\sigma = 0.25$. Although increasing σ leads to lower groupput, it can be observed that as N increases, the groupput approaches 10% of $\mathcal{T}_{\mathrm{nc}}^{\star}$ with $\sigma = 0.5$. Despite the fact that the groupput cannot be obtained for $\sigma = 0.1$, achieving 10% – 20% of $\mathcal{T}_{\mathrm{nc}}^{\star}$ is remarkable given the fact that EconCast operates in a distributed manner.

4.6 Experimental Evaluation

To experimentally evaluate the performance of EconCast-C,[9] we implement it using the TI eZ430-RF2500-SEH node [149].[10] In this section, we first describe the energy measurements performed on the nodes running EconCast-C. Then, we describe the method by which nodes can estimate the number of listening nodes. Finally, we experimentally evaluate the performance of EconCast-C.

4.6.1 Experimental Setup

The TI eZ430-RF2500-SEH node is equipped with: (i) an ultra-low-power MSP430 micro-controller and a CC2500 wireless transceiver operating at 2.4 GHz at 250 kbps, (ii) a solar energy harvester (SEH-01) that converts ambient light into electrical energy, and (iii) a 1 mF capacitor to power up the transceiver board. Despite its drawbacks which will be discussed below, it can be used for evaluation by extending the length of the shortest allowable data transmission.

We consider power budgets of $\rho \in \{1\,\text{mW}, 5\,\text{mW}\}$. From our measurements, a node spends $L = 67.08\,\text{mW}$ in the listen state and $X = 56.29\,\text{mW}$ in the transmit state.[11] The power consumption levels are very similar from node to node. Recall from Section 4.5 that the performance of EconCast depends on the ratio between the power consumption levels and budget. Therefore, our experimental results will be similar to experiments when both the power consumption levels and budget are scaled down (e.g., a network of nodes with $\rho \in \{10\,\text{uW}, 50\,\text{uW}\}$, $L = 670\,\text{uW}$, and $X = 560\,\text{uW}$).

Each node is programmed with its ρ, L, and X as the input of EconCast-C. The nodes' main drawbacks include (i) inaccurate readings of the energy storage level (i.e., the voltage of the on-board capacitor) which are sensitive to the environment, and (ii) the fact that the 1 mF capacitor cannot support multiple packet transmissions. Due to these drawbacks, we implement (via soft-

---

[9]See Section 4.3.4 the reasons for only implementing EconCast-C.

[10]A demonstration of the testbed was presented in [18].

[11]This corresponds to a $-16\,\text{dBm}$ transmission power, at which nodes within the same room typically have little or no packet loss.

ware) a virtual battery at each node. The virtual battery emulates the node's energy storage level based on its sleep, listen, transmit activities, as well as the energy consumption of sending and receiving the real pings, and is used for updating the Lagrange multiplier according to (4.17). We show in the following section that in practice, a node running EconCast-C using this virtual battery is indeed consuming power at a rate close to its power budget.

4.6.2 Energy Consumption Measurements

To accurately measure the power consumption of the nodes, we disable the on-board solar cell, and attach a large pre-charged capacitor ($C_{\text{cap}} = 5\,\text{F}$) that stores energy in advance (similar power consumption measurements were used in [150]). The energy consumed is computed by

$$E_{\text{consumed}} = 0.5 C_{\text{cap}} \cdot \left(V_{t_0}^2 - V_{t_1}^2 \right), \tag{4.25}$$

where V_{t_0} and V_{t_1} are the measured power voltage values of the capacitor at t_0 and t_1. The empirical average power consumption, $\overline{P}$ (mW), is then computed by

$$\overline{P} = E_{\text{consumed}} / \left(t_1 - t_0 \right). \tag{4.26}$$

Note that even with such a big capacitor, a node with a power budget of $1\,\text{mW}$ ($5\,\text{mW}$) has a lifetime of only 135 (27) minutes with $V_{t_0} = 3.6\,\text{V}$ and $V_{t_1} = 3.0\,\text{V}$, which represent its stable working voltage range.

To measure the power consumption of the nodes, we charge the capacitor to $V_{t_0} = 3.6\,\text{V}$ and log the readings of V_{t_1} after 30 minutes using a multimeter. The empirical average power consumption is computed from (4.25) and (4.26) for $\sigma \in \{0.25, 0.5\}$ and is averaged using 60 runs. Because L and X do not account for some additional energy usage,[12] the *actual power consumption*, $\overline{P}$, is in fact a small fraction higher than the *target power budget*, ρ. Irrespective of σ, the measurement results show that $\overline{P}$ exceeds ρ by 11% for $\rho = 1\,\text{mW}$, and by 4% for $\rho = 5\,\text{mW}$. Note that the capacitor used has an equivalent energy storage as a $0.6\,\text{mAh}$ battery and EconCast works well at

---

[12]The additional energy usage includes the energy consumption of sending and receiving the real pings, and the energy consumed in powering up the regulator circuitry, etc.
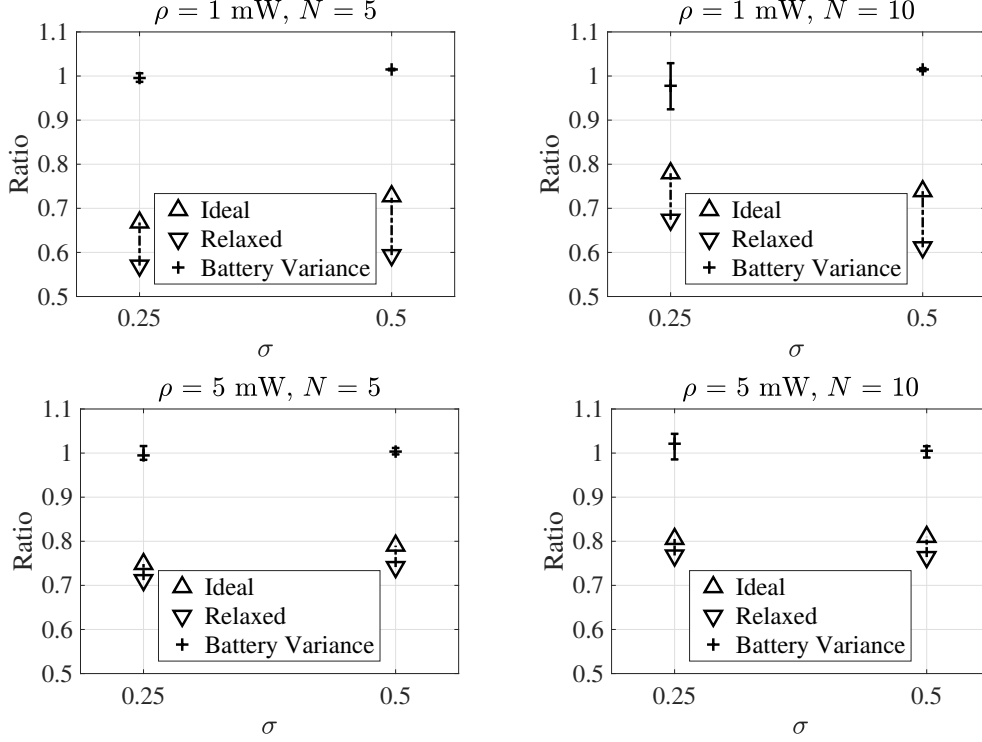
Figure 4.8: Points marked as "Ideal" ("Relaxed") represent ratio of experimental throughput normalized to the achievable throughput obtained by using the target power budget (actual power consumption) and points marked as "Battery Variance" present the average, minimum, and maximum ratios of power consumption normalized to target power budget, with $N \in \{5, 10\}$, $\rho \in \{1\,\mathrm{mW}, 5\,\mathrm{mW}\}$, and $\sigma \in \{0.25, 0.5\}$.

this battery size. Such a battery can be scaled down if the nodes' power budget is scaled down.

Observing the empirical power consumption of the nodes, we compute the achievable throughput by solving (P4) using both the actual power consumption, $\overline{P}$, and the target power budget, ρ, denoted by $\overline{\mathcal{T}_g^\sigma}$ and $\mathcal{T}_g^\sigma$, respectively. In Section 4.6.4, we compare the experimental throughput to both $\overline{\mathcal{T}_g^\sigma}$ and $\mathcal{T}_g^\sigma$. Having verified the power consumption of the nodes, we replace the capacitor with AAA batteries,[13] allowing the experiments to run for longer times.

4.6.3 Practical Pinging

To enable practical pinging in EconCast-C, a short, fixed-length *pinging interval* is introduced after each packet transmission. During this interval, the transmitter listens for pings and recipients

---

[13]The constant voltage of AAA batteries limits the ability to measure the power consumption of the nodes.

of the previous packet send a short ping at a random time uniformly distributed within the interval. The transmitter then estimates the number of listeners, $\hat{c}(t)$, by counting the pings it receives, and adjusts the transition rate, $\lambda_{xl}^C(t)$, according to (4.18e).

Ideally, each ping should be much shorter than both the pinging interval and the packet length in order to reduce the collisions between pings, and for the transmitter to successfully receive it. Therefore, we use pings of length $0.4\,\mathrm{ms}$, which is the shortest packet that can be sent by a node. Based on this, we empirically set the pinging interval to $8\,\mathrm{ms}$ and each data packet to $40\,\mathrm{ms}$. Namely, the energy consumption per ping is 1% of the consumption for a data packet.

4.6.4 Performance Evaluation

We consider homogeneous networks with $N \in \{5, 10\}$, $\rho \in \{1\,\mathrm{mW}, 5\,\mathrm{mW}\}$, $\sigma \in \{0.25, 0.5\}$, and nodes are located in proximity. One additional listening node (a 6th or 11th node) is also present but only as an observer and is connected to a PC via a USB port. Each data packet contains the node ID and information about the number of packets it has received from each other node. The observer node reports all received packets to the PC for storage and post processing. Each experiment is conducted for up to 24 hours. The experimental throughput is computed by dividing the duration of successful transmissions by the experiment duration.

Throughput evaluation: Fig. 4.8 presents the ratio of the experimentally obtained throughput, $\widetilde{\mathcal{T}_g^\sigma}$, normalized to the achievable throughput $\mathcal{T}_g^\sigma$ and $\overline{\mathcal{T}_g^\sigma}$ (see Section 4.6.2). Separate charts represent the results for differing number of nodes, N, and power budget, ρ. Points marked "Ideal" show the experimental throughput normalized to the achievable throughput computed by solving (P4) with the target power budget ρ (i.e., $\widetilde{\mathcal{T}_g^\sigma}/\mathcal{T}_g^\sigma$). Points marked "Relaxed" show the experimental throughput normalized to the achievable throughput computed by solving (P4) with the actual power consumption $\overline{P}$ (i.e., $\widetilde{\mathcal{T}_g^\sigma}/\overline{\mathcal{T}_g^\sigma}$). As expected, $\overline{\mathcal{T}_g^\sigma}$ is higher than $\mathcal{T}_g^\sigma$, resulting in a lower throughput ratio.

Fig. 4.8 shows that despite the practical limitations (e.g., packet collisions, inaccurate clocks, and inaccurate estimates of the number of active listeners, $\hat{c}(t)$) faced when running EconCast-C

$(N, \rho \text{ (mW)})$	(5,1)	(10,1)	(5,5)	(10,5)
$\widetilde{\mathcal{T}_g^\sigma}/\mathcal{T}_g^\sigma$ (%)	66.78	77.96	74.84	80.53
$\mathcal{T}_{\text{Panda}}/\mathcal{T}_g^\sigma$ (%)	6.24	9.64	19.35	35.63
$\widetilde{\mathcal{T}_g^\sigma}/\mathcal{T}_{\text{Panda}}$	10.76	8.09	3.87	2.26

Table 4.3: Experimental throughput of EconCast-C compared to computed throughput of Panda (all normalized to the achievable throughput $\mathcal{T}_g^\sigma$), with $\sigma = 0.25$ and varying (N, ρ).

on real hardware, the ratio $\widetilde{\mathcal{T}_g^\sigma}/\overline{\mathcal{T}_g^\sigma}$ is between $57\% - 77\%$ ($\widetilde{\mathcal{T}_g^\sigma}/\mathcal{T}_g^\sigma$ is between $67\% - 81\%$) for all settings considered. Moreover, Table 4.3 shows the improvement of EconCast-C over the throughput of Panda computed according to [17], denoted by $\mathcal{T}_{\text{Panda}}$, under the same power consumption levels and budget, with $\sigma = 0.25$. It can be seen that with power budget of $\rho = 1\,\text{mW}$, the experimental throughput of EconCast-C outperforms the analytically computed throughput of Panda by $8\times-11\times$.

We remark that getting a higher experimental throughput ratio is limited by the following reasons. First, there is an $8\,\text{ms}$ pinging interval (see Section 4.6.3) after each packet transmission which effectively reduces the number of bits delivered. Second, collisions of pings or failed decodings of pings result in inaccurate estimates of the number of listeners. Third, the low-power clock used by a node during its sleep state drifts and can be affected by its environment.

Power consumption: In Section 4.6.2, we show that the virtual battery, which is updated based on the energy consumption of both data packets and real pings, is valid for evaluating the actual power consumption of the node. Fig. 4.8 also presents the mean, minimum, and maximum power consumption of the virtual battery normalized to the target power budget ρ. Specifically, a value of 1 means that a node consumes power on average at the rate of its power budget throughout the experiment, and a higher value means that a node consumes power at the rate which is higher than its power budget.

The results show that nodes running EconCast-C consume power at rates which are within 7% and 3% of the target power budget with $\sigma = 0.25$ and $\sigma = 0.5$, respectively. This is because smaller value of σ increases the communication burstiness (see Section 4.5.4), resulting in larger variance of the nodes' virtual battery levels.

# of Listeners	0	1	2	3	4
$\rho = 1\,\mathrm{mW}$ (%)	89.03	9.69	1.28	0.00	0.00
$\rho = 5\,\mathrm{mW}$ (%)	59.21	31.22	8.22	1.24	0.11

Table 4.4: Distribution of number of pings (active listeners) received after each packet transmission with $N = 5$, $\sigma = 0.25$, and varying ρ.

Collection of Pings: An important input to EconCast-C is the estimates of number of active listeners, $\hat{c}(t)$, based on which the transmitter decides the probability to continuously transmit. Larger values of $\hat{c}(t)$ lead to longer average burst length and can potentially significantly increase the throughput. For example, receiving 1 ping, the transmitter continuously transmits a packet with probability 0.8647 with $\sigma = 0.5$. This probability increases to 0.9817 with $\sigma = 0.25$, which substantially increases the burstiness. Also, with lower power budget, a successful transmission happens more rarely and it becomes harder to collect pings.

Table 4.4 presents the distribution of number of pings (i.e., number of active listeners) received by the transmitter after each packet transmission, during experiments of $N = 5$, $\sigma = 0.25$, and $\rho \in \{1\,\mathrm{mW}, 5\,\mathrm{mW}\}$. It can be shown that with a higher power budget, the nodes are more active and the transmitter has higher probability to receive more pings. On the other hand, with lower power budget, the transmitter almost never receives more than three pings in a 5-node experiment, resulting in lower throughput as illustrated in Fig. 4.8.

4.7 Conclusion

In this chapter, we generalized the ND rate maximization problem described in Chapter 3 and considered the problem of maximizing the broadcast groupput and anyput among a set of energy-constrained nodes with heterogeneous power budgets and listen and transmit power consumption levels. We provided efficient methods to obtain oracle groupput and oracle anyput for a given set of heterogeneous nodes. We developed the EconCast distributed protocols that control the nodes' transitions among sleep, listen, and transmit states. We analytically showed that heterogeneous nodes using the protocols (without any a priori knowledge regarding the number of nodes in the

network, and power budgets and consumption levels) can achieve the oracle groupput and anyput in a limiting sense (when $\sigma \to 0$). We evaluated the throughput performance of EconCast numerically and through extensive simulations, and compared it to the state of the art. We also considered the design tradeoffs in relation to σ and the impact of σ on the burstiness and throughput. Finally, we experimentally evaluated EconCast using commercial-off-the-shelf energy harvesting nodes, thereby demonstrating its practicality.

4.8 Appendices

4.8.1   Appendix – Proof of Lemma 4.1

We describe here a schedule that works assuming each packet transmission has a fixed transmit length of θ, though the proof can be extended to varying transmission lengths. Therefore, time can be broken into slots of length θ (by the oracle) and nodes sleep, listen, or transmit on a per-slot basis. Assume that the optimal solution $(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ to (P2) yields rational values for all $\alpha_i^\star$ and $\beta_i^\star$. The period size of the oracle schedule P is set to the least common denominator over all solution variables in $(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$. Hence, during each period, node i listens for $\alpha_i^\star P$ slots and transmits for $\beta_i^\star P$ slots. The slots during the period can be assigned arbitrarily by the oracle to transmitters (e.g., weighted round-robin or in-order) and (4.11) ensures that there will be sufficient slots. Once the slots for the transmitters are assigned, each listener can then choose their $\alpha_i^\star P$ slots in which they listen from the set of transmit slots assigned to other transmitters, and (4.12) ensures these are sufficient as well (note that multiple listeners are permitted for a single transmitter slot).

If the periodic scheduler is launched immediately, some nodes may not have the harvested (or budgeted) energy to perform all listen and transmit tasks within the first period (e.g., it may be assigned to transmit and listen early on, and recoup the energy during later slots). If such a case, we simply delay the initial iteration, allowing all nodes to harvest (or budget) $\rho_i P \theta$ amount of energy for that initial period. Then the nodes have enough energy to repeat all the subsequent periods since the energy node i spends during the k-th period is $(\alpha_i L_i P \theta + \beta_i X_i P \theta)$, and the energy it accumulates during this period is $\rho_i P \theta$. Hence, it follows from (4.9) that no more energy is spent

83

than is accumulated (budgeted). Therefore, there is sufficient energy to repeat the period. ∎

4.8.2 Appendix – Proof of Lemma 4.2

For the capture version EconCast-C, we prove that the transition rates (4.18a), (4.18b), (4.18c), and (4.18e) will drive the network Markov chain to a steady state with distribution (4.19), by checking that the *detailed balanced equations* hold. We assume $\sigma = 1$ and drop the constant term $1/\sigma$ for brevity. For network state $\mathbf{w}$, define $\mathcal{N}_s(\mathbf{w})$, $\mathcal{N}_l(\mathbf{w})$, and $\mathcal{N}_x(\mathbf{w})$ as the sets of nodes in sleep, listen, and transmit states, respectively, and their cardinalities as $N_s(\mathbf{w})$, $N_l(\mathbf{w})$, and $N_x(\mathbf{w})$, respectively. Note that $N_s(\mathbf{w}) + N_l(\mathbf{w}) + N_x(\mathbf{w}) = N$ and $N_x(\mathbf{w}) \in \{0, 1\}$. We also use $\mathbf{w} = (\mathcal{N}_s(\mathbf{w}), \mathcal{N}_l(\mathbf{w}), \mathcal{N}_x(\mathbf{w}))$ to denote network state $\mathbf{w}$ and $r(\mathbf{w}, \mathbf{w}')$ to denote the transition rate from state $\mathbf{w}$ to $\mathbf{w}'$. We consider the following cases.

Case 1: If node i is in sleep state ($w_i = s$), the only transition that can happen is to transition into listen state when the channel is clear, i.e.,

$$\mathbf{w} = (\mathcal{N}_s(\mathbf{w}), \mathcal{N}_l(\mathbf{w}), \varnothing) \to \mathbf{w}' = (\mathcal{N}_s(\mathbf{w}) \setminus \{i\}, \mathcal{N}_l(\mathbf{w}) \cup \{i\}, \varnothing).$$

In this case, $r(\mathbf{w}, \mathbf{w}') = \exp(-\eta_i L_i)$ and $r(\mathbf{w}', \mathbf{w}) = 1$.

Case 2: If node i is in listen state ($w_i = l$), and transitions to sleep state, i.e.,

$$\mathbf{w} = (\mathcal{N}_s(\mathbf{w}), \mathcal{N}_l(\mathbf{w}), \varnothing) \to \mathbf{w}' = (\mathcal{N}_s(\mathbf{w}) \cup \{i\}, \mathcal{N}_l(\mathbf{w}) \setminus \{i\}, \varnothing).$$

In this case, $r(\mathbf{w}, \mathbf{w}') = 1$ and $r(\mathbf{w}', \mathbf{w}) = \exp(-\eta_i L_i)$.

Case 3: If node i is in listen state ($w_i = l$) and transitions to transmit state, i.e.,

$$\mathbf{w} = (\mathcal{N}_s(\mathbf{w}), \mathcal{N}_l(\mathbf{w}), \varnothing) \to \mathbf{w}' = (\mathcal{N}_s(\mathbf{w}), \mathcal{N}_l(\mathbf{w}) \setminus \{i\}, \{i\}).$$

In this case, $r(\mathbf{w}, \mathbf{w}') = \exp(\eta_i(L_i - X_i))$ and $r(\mathbf{w}', \mathbf{w}) = \exp(-N_l(\mathbf{w}'))$.

Case 4: If node i is in transmit state ($w_i = x$), the only transition that can happen is to transition

to listen state when its transmission is finished, i.e.,

$$\mathbf{w} = (\mathcal{N}_s(\mathbf{w}), \mathcal{N}_l(\mathbf{w}), \{i\}) \to \mathbf{w}' = (\mathcal{N}_s(\mathbf{w})\{i\}, \mathcal{N}_l(\mathbf{w}) \cup \{i\}, \varnothing).$$

In this case, $r(\mathbf{w}, \mathbf{w}') = \exp(-N_l(\mathbf{w}))$ and $r(\mathbf{w}', \mathbf{w}) = \exp(\eta_i(L_i - X_i))$.

For each case, $\pi_{\mathbf{w}} \cdot r(\mathbf{w}, \mathbf{w}') = \pi_{\mathbf{w}'} \cdot r(\mathbf{w}', \mathbf{w})$ holds and similar detailed balance equations hold for the non-capture version EconCast-NC. Therefore, we complete the proof. ∎

4.8.3   Appendix – Burstiness Analysis of EconCast

To derive the average burst length (denoted by B) of EconCast-C, we use $\pi_{\mathbf{w}}^{\star}$ to denote the optimal solution to (P4) and define $\mathcal{W}' = \{\mathbf{w} \in \mathcal{W} : \nu_{\mathbf{w}} = 1, c_{\mathbf{w}} \geq 1\}$, i.e., the set of states with successfully received bursts. Recall that for a given value of σ, the optimal value of (P4) is exactly $\mathcal{T}^{\sigma}$. According to (4.18e), for a given state $\mathbf{w} \in \mathcal{W}'$, the average burst length of EconCast-C in groupput mode is $\exp(c_{\mathbf{w}}/\sigma)$. Therefore, during a (long enough) time duration of T, the average number of bursts received by all the nodes can be computed by $\sum_{\mathbf{w} \in \mathcal{W}'} \frac{T \cdot \pi_{\mathbf{w}}^{\star}}{\exp(c_{\mathbf{w}}/\sigma)}$, and the average burst length of EconCast in groupput mode, B_g, is given by

$$B_g = (\text{Avg. Total Burst Length})/(\text{Avg. Number of Bursts})$$
$$= \frac{T \cdot \sum_{\mathbf{w} \in \mathcal{W}'} \pi_{\mathbf{w}}^{\star}}{\sum_{\mathbf{w} \in \mathcal{W}'} \frac{T \cdot \pi_{\mathbf{w}}^{\star}}{\exp(c_{\mathbf{w}}/\sigma)}} = \frac{\sum_{\mathbf{w} \in \mathcal{W}'} \pi_{\mathbf{w}}^{\star}}{\sum_{\mathbf{w} \in \mathcal{W}'} \frac{\pi_{\mathbf{w}}^{\star}}{\exp(c_{\mathbf{w}}/\sigma)}}. \tag{4.27}$$

Similarly, the average burst length of EconCast-C in anyput mode is computed by replacing $c_{\mathbf{w}}$ with $\gamma_{\mathbf{w}}$ in (4.27). Since $\gamma_{\mathbf{w}} = 1$ always holds for $\mathbf{w} \in \mathcal{W}'$, B_a simplifies to

$$B_a = \frac{\sum_{\mathbf{w} \in \mathcal{W}'} \pi_{\mathbf{w}}^{\star}}{\sum_{\mathbf{w} \in \mathcal{W}'} [\pi_{\mathbf{w}}^{\star} \cdot \exp(-1/\sigma)]} = \exp(1/\sigma). \tag{4.28}$$

This shows that the anyput average burst length is independent of the number of nodes, N, and only depends on σ. ∎

Part II

Full-Duplex Wireless

Chapter 5: Wideband Full-duplex Wireless via Frequency-Domain Equalization: Design and Experimentation

Full-duplex (FD) wireless – simultaneous transmission and reception on the same frequency channel – is an emerging communication paradigm that can significantly improve spectrum efficiency at the physical (PHY) layer and provide many other benefits at the higher layers [8, 9]. As mentioned in Sections 1.1 and 2.2, despite extensive work on the development of FD radios and systems using using laboratory bench-top equipment, a *compact IC-based* FD radio design is necessary for supporting FD in hand-held devices such as cell phones and tablets.

In this chapter, we design and implement an RF canceller leveraging the technique of frequency-domain equalization (FDE) (which can be realized in IC implementations) using discrete components on a printed circuit board (PCB), as shown in Fig. 5.1(a). We refer to it as the FDE-based *PCB canceller*, which emulates its RFIC counterpart and facilitates the evaluation of the canceller configuration scheme and the experimentation using SDRs in a network with multiple FD nodes. In Chapter 9, we will present the integration of FDE-based FD radios in the open-access COSMOS testbed [13, 14], which allows the community to experiment with wideband compact FD wireless.

We first present the problem formulation and background on RF cancellers (Section 5.1). We present a realistic PCB canceller model and an optimized canceller configuration scheme, which allows efficient adaption of the canceller to environmental changes (Section 5.2). The PCB canceller model is experimentally validated and is shown to have high accuracy (Section 5.3). We implement an FDE-based FD radio by integrating the PCB canceller with an NI USRP SDR, as depicted in Fig. 5.1(b), which achieves 95 dB overall SIC across 20 MHz real-time bandwidth and an average FD link throughput gain of $1.85\times$–$1.91\times$. Using our testbed (see Fig. 5.1(c)), we extensively evaluate the network-level FD gain and confirm analytical results in two types of networks:
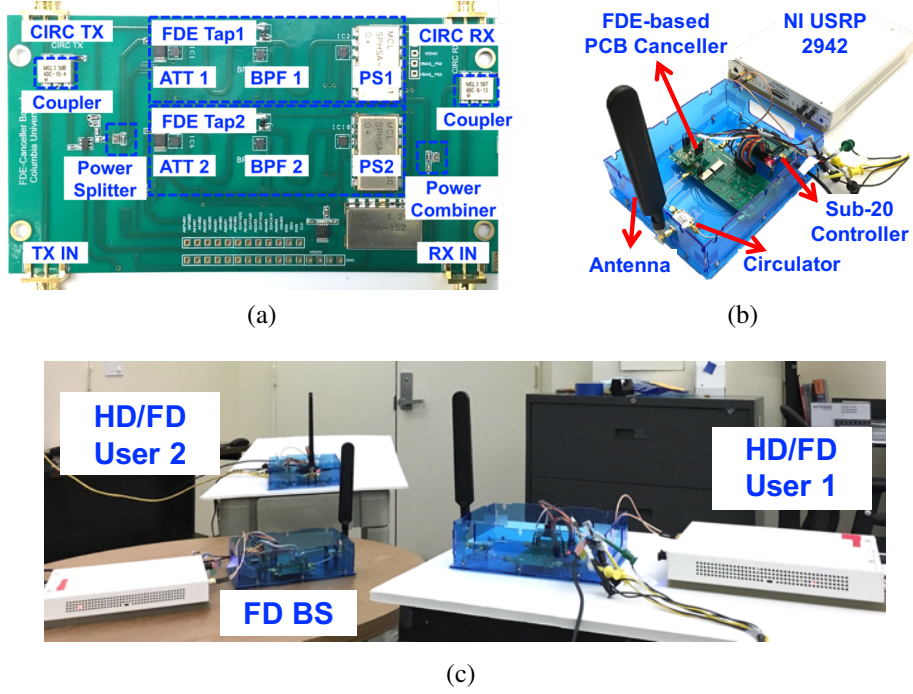
Figure 5.1: (a) The frequency-domain equalization- (FDE-) based wideband RF canceller implemented using discrete components on a printed circuit board (PCB), (b) the implemented FDE-based full-duplex (FD) radio, and (c) the experimental testbed consisting of an FD base station (BS) and 2 users that can operate in either half-duplex (HD) or FD mode.

(i) *UL-DL networks* consisting of one FD base station (BS) and two half-duplex (HD) users with inter-user interference (IUI), and (ii) *heterogeneous HD-FD networks* consisting of one FD BS and co-existing HD and FD users (Section 5.4). Finally, we numerically evaluate the FDE-based cancellers based on measurements and validated canceller models, and discuss various canceller design tradeoffs (Section 5.5).

This chapter's contributions were previously presented in [22, 24]. Some of the results in this chapter are based on joint work with Ph.D. students Mahmood Barrani Dastjerdi and Jin Zhou, and undergraduate student Jackson Welles, who contributed to the design and implementation of the FDE-based RF canceller. In a recent joint work with Aravind Nagulu, the RF canceller presented in this chapter was integrated with a CMOS circulator to implement FD radios with high power (e.g., multi-watt) handling [25].

$\|z\|, \angle z$	Amplitude and phase of a complex value $z = x + jy$ $(x, y \in \mathbb{R})$, where $\|z\| = \sqrt{x^2 + y^2}$ and $\angle z = \tan^{-1}\left(\frac{y}{x}\right)$
B	Total wireless bandwidth/desired RF SIC bandwidth
K, k	Total number of frequency channels and channel index
f_k	Center frequency of the kth frequency channel
M	Number of FDE taps in an FDE-based RF canceller
$H_{\text{SI}}(f_k)$	Frequency response of the antenna interface
$H^{\text{P}}(f_k)$	Frequency response of the FDE-based PCB canceller
$H_i^{\text{P}}(f_k)$	Frequency response of the ith FDE tap in the PCB canceller
$A_i^{\text{P}}, \phi_i^{\text{P}}$	Amplitude and phase controls of the ith FDE tap in the PCB canceller
$C_{\text{F},i}, C_{\text{Q},i}$	Digitally tunable capacitors that control the center frequency and quality factor of the ith FDE tap in the PCB canceller

Table 5.1: Nomenclature

5.1 Background and Formulation

In this section, we review concepts related to FD wireless and RF canceller configuration optimization. We also present different RF canceller designs and specificaully the design of the FDE-based RF canceller. Summary of the main notation is provided in Table 5.1.

5.1.1 FD Background and Notation

Fig. 5.2 shows the block diagram of a single-antenna FD radio using a circulator at the antenna interface. Due to the extremely strong SI power level and the limited dynamic range of the analog-to-digital converter (ADC) at the RX, a total amount of 90–110 dB SIC must be achieved across the antenna, RF, and digital domains. Specifically, (i) SI suppression is first performed at the antenna interface, (ii) an RF SI canceller then taps a reference signal at the output of the TX power amplifier (PA) and performs SIC at the input of the low-noise amplifier (LNA) at the RX, and (iii) a digital SI canceller further suppresses the residual SI.

Consider a wireless bandwidth of B that is divided into K orthogonal frequency channels. The channels are indexed by $k \in \{1, \ldots, K\}$ and denote the center frequency of the kth channel by f_k.[1] We denote the antenna interface response by $H_{\text{SI}}(f_k)$ with amplitude $\|H_{\text{SI}}(f_k)\|$ and phase

---

[1] We use discrete frequency values $\{f_k\}$ since in practical systems, the antenna interface response is measured

89
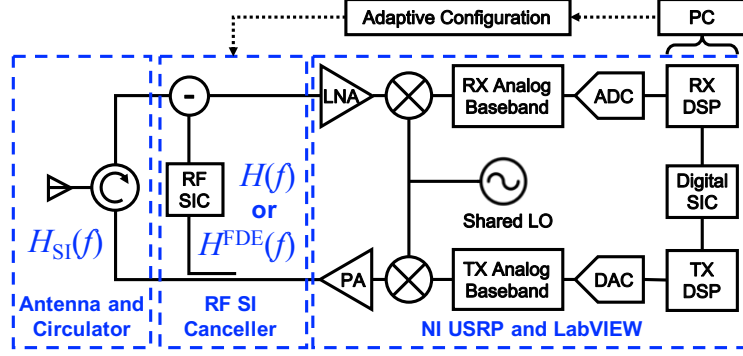
Figure 5.2: Block diagram of a full-duplex (FD) radio.

$\angle H_{\mathrm{SI}}(f_k)$. Note that the actual SI channel includes the TX-RX leakage from the antenna interface as well as the TX and RX transfer functions at the baseband from the perspective of the digital canceller. Since the chapter focuses on achieving wideband RF SIC, we use $H_{\mathrm{SI}}(f_k)$ to denote the antenna interface response and also refer to it as the *SI channel*. We refer to *TX/RX isolation* as the ratio (in dB, usually a negative value) between the residual SI power at the RX input and the TX output power, which includes the amount of TX/RX isolation achieved by both the antenna interface and the RF canceller/circuitry. We then refer to *RF SIC* as the absolute value of the TX/RX isolation. We also refer to *overall SIC* as the total amount of SIC achieved in both the RF and digital domains. The antenna interface used in our experiments typically provides a TX/RX isolation of around $-20\,\mathrm{dB}$.

5.1.2   Problem Formulation

Ideally, an RF canceller is designed to best emulate the antenna interface, $H_{\mathrm{SI}}(f_k)$, across a desired bandwidth, $B = [f_1, f_K]$. We denote by $H(f_k)$ the frequency response of an RF canceller and by $H_{\mathrm{res}}(f_k) := H_{\mathrm{SI}}(f_k) - H(f_k)$ the *residual SI channel response*. The optimized RF canceller

at discrete points (e.g., per OFDM subcarrier). However, the presented model can also be applied to cases with continuous frequency values.

configuration is obtained by solving (P1):

$$\text{(P1) min: } \sum_{k=1}^{K} |H_{\text{res}}(f_k)|^2 = \sum_{k=1}^{K} |H_{\text{SI}}(f_k) - H(f_k)|^2$$

subject to: constraints on configuration parameters of $H(f_k)$, $\forall k$.

The RF canceller configuration obtained by solving (P1) minimizes the residual SI power referred to the TX output. One main challenge associated with the design of the RF canceller with response $H(f_k)$ to achieve wideband SIC is due to the highly frequency-selective antenna interface, $H_{\text{SI}}(f_k)$. Moreover, an efficient RF canceller configuration scheme needs to be designed so that the canceller can adapt to time-varying $H_{\text{SI}}(f_k)$.

5.1.3   RF Canceller Designs

Delay line-based RF cancellers. An RF canceller design introduced in [81] involves using M delay line taps. Specifically, the ith tap is associated with a time delay of τ_i, which is *pre-selected* and *fixed* depending on the selected circulator and antenna, and an amplitude control of A_i. Since the Fourier transform of a delay of τ_i is $e^{-2\pi f \tau_i}$, an M-tap delay line-based RF canceller has a frequency response of $H^{\text{delay-line}}(f_k) = \sum_{i=1}^{M} A_i e^{-j2\pi f_k \tau_i}$. The configurations of the amplitude controls, $\{A_i\}$, are obtained by solving (P1) with $H(f_k) = H^{\text{delay-line}}(f_k)$. In [81], an RF canceller of $M = 16$ delay line taps is implemented. In [83], a similar approach is considered with $M = 3$ and an additional phase control, ϕ_i, on each tap, resulting in an RF canceller model of $H^{\text{delay-line}}(f_k) = \sum_{i=1}^{3} A_i e^{-j(2\pi f_k \tau_i + \phi_i)}$. As mentioned in Section 2.2, although such cancellers can achieve wideband SIC, this approach is more suitable for large-form-factor nodes than for compact/small-form-factor implementations.

Amplitude- and phase-based RF cancellers. A compact design that is based on an amplitude- and phase-based RF canceller realized in an RFIC implementation is presented in [151]. This canceller has a single-tap with one amplitude and frequency control, (A_0, ϕ_0), which can emulate the antenna interface, $H_{\text{SI}}(f_k)$, at *only one* given cancellation frequency f_1 by setting $A_0 = |H_{\text{SI}}(f_1)|$ and $\phi_0 = \angle H_{\text{SI}}(f_1)$. The same design is also realized using discrete components on a PCB (with-
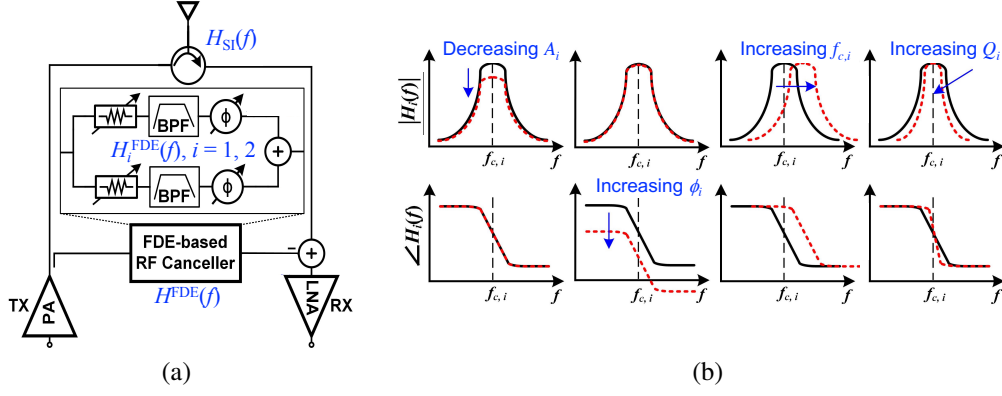
Figure 5.3: (a) Block diagram of an FDE-based RF canceller with $M = 2$ FDE taps, and (b) illustration of amplitude and phase responses of an ideal 2nd-order bandpass filter (BPF) with amplitude, phase, center frequency, and quality factor (i.e., group delay) controls.

out using any length delay lines), and is integrated in the ORBIT testbed for open-access FD research [38]. However, this type of RF cancellers has limited RF SIC perfromacne and bandwidth, since it can only emulate the antenna interface at a single frequency.

An FDE-based RF canceller. A compact design that can achieve significantly enhanced performance and bandwidth of RF SIC is based on the technique of frequency-domain equalization (FDE) and was implemented in an RFIC [152]. Fig. 5.3(a) shows the diagram of an FDE-based canceller, where parallel reconfigurable bandpass filters (BPFs) are used to emulate the antenna interface response across wide bandwidth. We denote the frequency response of a general FDE-based RF canceller consisting of M FDE taps by

$$H^{\text{FDE}}(f_k) = \sum_{i=1}^{M} H_i^{\text{FDE}}(f_k), \tag{5.1}$$

where $H_i^{\text{FDE}}(f_k)$ is the frequency response of the ith FDE tap containing a reconfigurable BPF with amplitude and phase controls. Theoretically, any m^{th}-order RF BPF ($m = 1, 2, \dots$) can be used. Fig. 5.3(b) illustrates the amplitude and phase of a 2nd-order BPF with different control parameters. For example, increased BPF quality factors result in "sharper" BPF amplitudes and increased group delay. Since it is shown [152] that a 2nd-order BPF can accurately model the FDE N-path filter, the frequency response of an FDE-based RFIC canceller with M FDE taps is given

by

$$H^{\mathrm{I}}(f_k) = \sum_{i=1}^{M} H_i^{\mathrm{I}}(f_k) = \sum_{i=1}^{M} \frac{A_i^{\mathrm{I}} \cdot e^{-j\phi_i^{\mathrm{I}}}}{1 - jQ_i \cdot (f_{\mathrm{c},i}/f_k - f_k/f_{\mathrm{c},i})}. \tag{5.2}$$

Within the ith FDE tap, $H_i^{\mathrm{I}}(f_k)$, A_i^{I} and ϕ_i^{I} are the amplitude and phase controls, and $f_{\mathrm{c},i}$ and Q_i are the center frequency and quality factor of the 2nd-order BPF (see Fig. 5.3(b)). In the RFIC canceller, $f_{\mathrm{c},i}$ and Q_i are adjusted through a reconfigurable baseband capacitor and transconductors, respectively.

As Fig. 5.3(b) and (5.2) suggest, one FDE tap features four degrees of freedom (DoF) so that the antenna interface, $H_{\mathrm{SI}}(f_k)$, can be emulated *not only in amplitude and phase, but also in the slope of amplitude and the slope of phase (i.e., group delay)*. Therefore, the RF SIC bandwidth can be significantly enhanced through FDE when compared with the amplitude- and phased-based RF cancellers.

5.2 Design and Optimization

In this section, we present our design and implementation of an FDE-based canceller using discrete components on a PCB (referred to as the *PCB canceller*). Recall that the motivation is to facilitate integration with an SDR platform, the experimentation of FD at the link/network level, and integration with open-access wireless testbeds. We then present a realistic PCB canceller model, which is validated (Section 5.3) and used in the experimental and numerical evaluations (Sections 5.4 and 5.5).

5.2.1 FDE PCB Canceller Implementation

Fig. 5.1(a) and Fig. 5.3(a) show the implementation and block diagram of the PCB canceller with 2 FDE taps. In particular, a reference signal is tapped from the TX input using a coupler and is split into two FDE taps through a power divider. Then, the signals after each FDE tap are combined and RF SIC is performed at the RX input. Each FDE tap consists of a reconfigurable 2nd-order BPF, as well as an attenuator and phase shifter for amplitude and phase controls. We
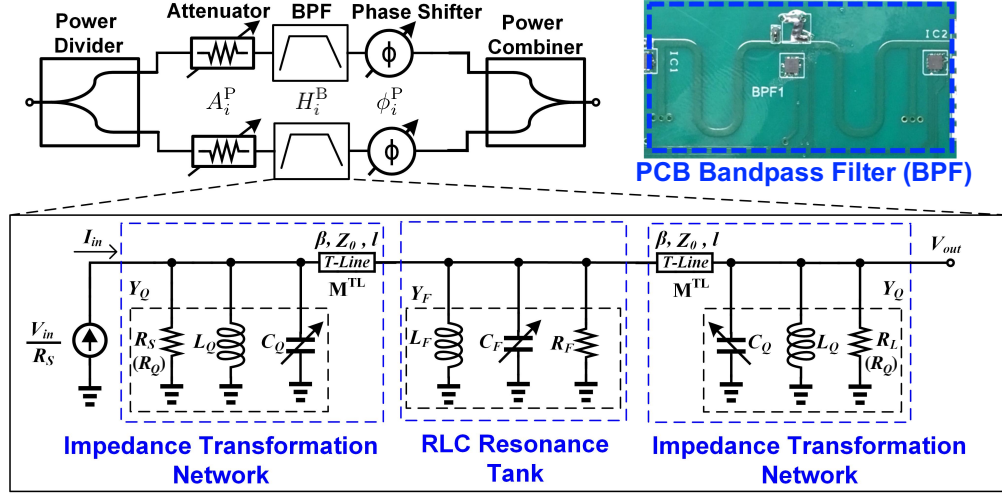
Figure 5.4: Block diagram of the implemented 2 FDE taps in the PCB canceller (see Fig. 5.3(a)), each of which consists of an RLC bandpass filter (BPF), an attenuator for amplitude control, and a phase shifter for phase control.

refer to the BPF here as the *PCB BPF* to distinguish from the one in the RFIC canceller (5.2). The PCB BPF (with size of $1.5\,\mathrm{cm} \times 4\,\mathrm{cm}$, see Fig. 5.4) is implemented as an RLC filter with impedance transformation networks and is optimized around $900\,\mathrm{MHz}$ operating frequency.[2] When compared to the N-path filter used in the RFIC canceller [152] that consumes certain amount of DC power, this discrete component-based passive RLC BPF has zero DC power consumption and can support higher TX power levels. Moreover, it has a lower noise level than the RFIC implementation.

The PCB BPF center frequency in the ith FDE tap can be adjusted through the capacitor, $C_{\mathrm{F},i}$, in the RLC resonance tank. In order to achieve a high and adjustable BPF quality factor, impedance transformation networks including transmission-lines (T-Lines) and digitally tunable capacitors, $C_{\mathrm{Q},i}$, are introduced. In our implementation, $C_{\mathrm{F},i}$ consists of two parallel capacitors: a fixed $8.2\,\mathrm{pF}$ capacitor and a Peregrine Semiconductor PE64909 digitally tunable capacitor (4-bit) with a $0.12\,\mathrm{pF}$ resolution. For $C_{\mathrm{Q},i}$, we use the Peregrine Semiconductor PE64102 digitally tunable capacitor (5-bit) with a $0.39\,\mathrm{pF}$ resolution. In addition, the programmable attenuator has a tuning range of 0–$15.5\,\mathrm{dB}$ with a $0.5\,\mathrm{dB}$ resolution, and the passive phase shifter is controlled by an 8-bit digital-

---

[2]We select $900\,\mathrm{MHz}$ around the Region 2 902–$928\,\mathrm{MHz}$ ISM band as the operating frequency but the approach can be easily extended to other bands (e.g., $2.4\,\mathrm{GHz}$) with slight modification of the hardware design and proper choice of the frequency-dependent components.

to-analog converter (DAC) and covers full $360°$ range.

5.2.2 FDE PCB Canceller Model

Ideally, the PCB BPF has a 2nd-order frequency response from the RLC resonance tank. However, in practical implementation, its response deviates from that used in the FDE-based RFIC canceller (5.2). Therefore, there is a need for a valid model tailored for evaluating the performance and optimized configuration of the PCB canceller. Based on the circuit diagram in Fig. 5.4, we derive a realistic model for the frequency response of the PCB BPF, $H_i^{\mathrm{B}}(f_k)$, given by (the details can be found in Appendix 5.7)

$$
\begin{aligned}
H_i^{\mathrm{B}}(f_k) =R_s^{-1}\Big[& j\sin(2\beta l)Z_0 Y_{\mathrm{F},i}(f_k)Y_{\mathrm{Q},i}(f_k) + \cos^2(\beta l)Y_{\mathrm{F},i}(f_k) + 2\cos(2\beta l)Y_{\mathrm{Q},i}(f_k) \\
& + j\sin(2\beta l)/Z_0 + 0.5j\sin(2\beta l)Z_0(Y_{\mathrm{Q},i}(f_k))^2 - \sin^2(\beta l)Z_0^2 Y_{\mathrm{F},i}(f_k)(Y_{\mathrm{Q},i}(f_k))^2 \Big]^{-1},
\end{aligned}
\tag{5.3}
$$

where $Y_{\mathrm{F},i}(f_k)$ and $Y_{\mathrm{Q},i}(f_k)$ are the admittance of the RLC resonance tank and impedance transformation networks, i.e.,

$$
\begin{cases}
Y_{\mathrm{F},i}(f_k) = 1/R_{\mathrm{F}} + j2\pi C_{\mathrm{F},i}f_k + 1/(j2\pi L_{\mathrm{F}}f_k), \\
Y_{\mathrm{Q},i}(f_k) = 1/R_{\mathrm{Q}} + j2\pi C_{\mathrm{Q},i}f_k + 1/(j2\pi L_{\mathrm{Q}}f_k).
\end{cases}
\tag{5.4}
$$

In particular, to have perfect matching with the source and load impedance of the RLC resonance tank, R_{S} and R_{L} are set to be the same value of $R_{\mathrm{Q}} = 50\,\Omega$ (see Fig. 5.4). β and Z_0 are the wavenumber and characteristic impedance of the T-Line with length l (see Fig. 5.4). In our implementation, $L_{\mathrm{F}} = 1.65\,\mathrm{nH}$, $L_{\mathrm{Q}} = 2.85\,\mathrm{nH}$, $\beta l \approx 1.37\,\mathrm{rad}$, and $Z_0 = 50\,\Omega$.

In addition, other components in the PCB canceller (e.g., couplers and power divider/combiner) can introduce extra attenuation and group delay, due to implementation losses. Based on the S-Parameters of the components and measurements, we observed that the attenuation and group delay introduced, denoted by A_0^{P} and τ_0^{P}, are constant across frequency in the desired bandwidth. Hence, we empirically set $A_0 = -4.1\,\mathrm{dB}$ and $\tau_0 = 4.2\,\mathrm{ns}$. Recall that each FDE tap is also associated with

amplitude and phase controls, A_i^{P} and ϕ_i^{P}, the PCB canceller with two FDE taps is modeled by

$$H^{\mathrm{P}}(f_k) = A_0^{\mathrm{P}} e^{-j2\pi f_k \tau_0^{\mathrm{P}}} \cdot \left[\sum_{i=1}^{2} A_i^{\mathrm{P}} e^{-j\phi_i^{\mathrm{P}}} H_i^{\mathrm{B}}(f_k) \right], \tag{5.5}$$

where $H_i^{\mathrm{B}}(f_k)$ is the PCB BPF model given by (5.3). As a result, the ith FDE tap in the PCB canceller (5.5) has configuration parameters $\{A_i^{\mathrm{P}}, \phi_i^{\mathrm{P}}, C_{\mathrm{F},i}, C_{\mathrm{Q},i}\}$, featuring 4 DoF.

5.2.3 Optimization of Canceller Configuration

Based on (P1), we now present a general FDE-based canceller configuration scheme that jointly optimizes all the FDE taps in the canceller.[3] Although our implemented PCB canceller has only 2 FDE taps, both its model and the configuration scheme can be easily extended to the case with a larger number of FDE taps, as described in Section 5.5.

The inputs to the FDE-based canceller configuration scheme are: (i) the PCB canceller model (5.5) with given number of FDE taps, M, (ii) the antenna interface response, $H_{\mathrm{SI}}(f_k)$, and (iii) the desired RF SIC bandwidth, $f_k \in [f_1, f_K]$. The optimized canceller configuration is obtained by solving (P2).

(P2) min: $\displaystyle\sum_{k=1}^{K} \left| H_{\mathrm{res}}^{\mathrm{P}}(f_k) \right| = \sum_{k=1}^{K} \left| H_{\mathrm{SI}}(f_k) - H^{\mathrm{P}}(f_k) \right|^2$

subject to: $A_i^{\mathrm{P}} \in [A_{\min}^{\mathrm{P}}, A_{\max}^{\mathrm{P}}]$, $\phi_i^{\mathrm{P}} \in [-\pi, \pi]$, $C_{\mathrm{F},i} \in [C_{\mathrm{F,min}}, C_{\mathrm{F,max}}]$, $C_{\mathrm{Q},i} \in [C_{\mathrm{Q,min}}, C_{\mathrm{Q,max}}]$, $\forall i$.

Note that (P2) is challenging to solve due to its non-convexity and non-linearity, as opposed to the linear program used in the delay line-based RF canceller [81]. This is due to the specific forms of the configuration parameters in (5.5) such as (i) the higher-order terms introduced by f_k, and (ii) the trigonometric term introduced by the phase control, ϕ_i^{P}. In addition, the antenna interface response, $H_{\mathrm{SI}}(f_k)$, is also frequency-selective and time-varying.

In general, it is difficult to maintain analytical tractability of (P2) (i.e., to obtain its optimal solution in closed-form). However, in practice, it is unnecessary to obtain the global optimum

---

[3]The RFIC canceller presented in [152] is configured based on heuristics. In Section 5.5, we show that the optimized configuration scheme can significantly improve the RFIC canceller performance.

	Highest Q-Factor	Lowest Q-Factor
Highest Center Freq.	Set 1: $(C_{F,min}, C_{Q,min})$	Set 3: $(C_{F,min}, C_{Q,max})$
Lowest Center Freq.	Set 2: $(C_{F,max}, C_{Q,min})$	Set 4: $(C_{F,max}, C_{Q,max})$

Table 5.2: Four (C_F, C_Q) configurations used in the validations.

to (P2) as long as the performance achieved by the obtained local optimum is sufficient (e.g., at least 45] dB RF SIC is achieved). In this work, the local optimal solution to (P2) is obtained using a MATLAB solver. The detailed implementation and performance of the optimized canceller configuration are described in Section 5.4.2.

5.3 Model Validation

Validation of the PCB BPF. We first experimentally validate the PCB BPF model, $H_i^B(f_k)$, given by (5.3). The ground truth is obtained by measuring the frequency response (using S-Parameters measurements) of the PCB BPF using a test structure, which contains only the BPF, to avoid the effects of other components on the PCB. The measurements are conducted with varying (C_F, C_Q) configurations and the result of each configuration is averaged over 20 measurement instances.[4] The BPF center frequency is measured as the frequency with the highest BPF amplitude, and the BPF quality factor is computed as the ratio between the center frequency and the 3 dB bandwidth around the center frequency.

The PCB BPF has a *fixed* quality factor of 2.7, achieved by using only the RLC resonance tank. By setting $C_Q = C_{Q,max}$ and $C_Q = C_{Q,min}$ (see Section 5.2.1), the measured lowest and highest achievable BPF quality factors are 9.2 and 17.8, respectively. This shows an improvement in the PCB BPF quality factor tuning range of 3.4×–6.6×, achieved by introducing the impedance transformation networks. Similarly, by setting $C_F = C_{F,max}$ and $C_F = C_{F,min}$, the PCB BPF has a center frequency tuning range of 18 MHz.

Fig. 5.5 presents the modeled and measured amplitude and phase responses of the PCB BPF with 4 different (C_F, C_Q) configurations (see Table 5.2) which cover the entire tuning range of the

---

[4]We drop the subscript i, since both PCB BPFs behave identically.
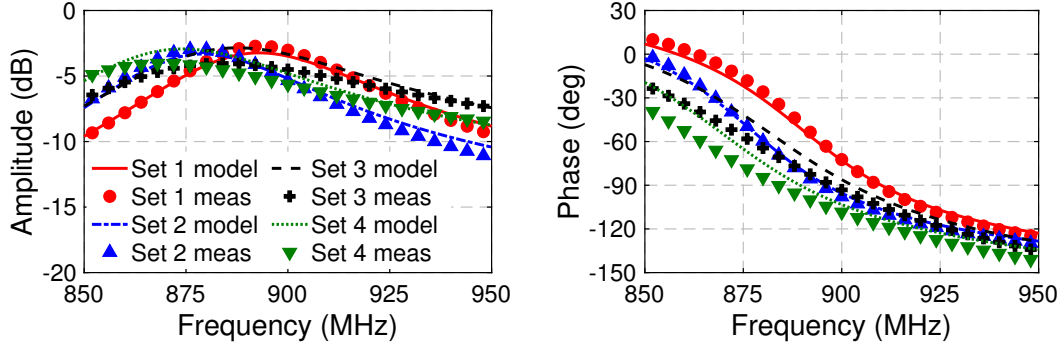
Figure 5.5: Modeled and measured amplitude and phase responses of the implemented PCB BPF with varying (C_F, C_Q) indicated in Table 5.2.

	Set 1	Set 2	Set 3	Set 4
Max Difference in Amplitude	0.6 dB	0.8 dB	1.2 dB	1.2 dB
Max Difference in Phase	7.4°	5.0°	13.8°	15.1°

Table 5.3: Accuracy of the PCB BPF model with varying (C_F, C_Q) indicated in Table 5.2.

BPF center frequency and quality factor. Table 5.3 summarizes the maximum differences between the measured and modeled amplitude and phase of the PCB BPF in the range of 850–950 MHz. In particular, the PCB BPF model (5.3) matches very closely with the measurements at the highest BPF quality factor value, where the maximum differences between the measured and modeled amplitude and phase in Set 1/2 are 0.6/0.8 dB and 7.4°/5.0°, respectively. At the lowest BPF quality factor value, the differences in Set 3/4 are 1.2/1.2 dB and 13.8°/15.1°, respectively, still showing the validity of the PCB BPF model. The same level of accuracy of the PCB BPF model (5.3) is also observed for other (C_F, C_Q) configurations within their tuning ranges.

Validation of the PCB canceller. We use the same experiments as in the PCB BPF validation to validate the PCB canceller model with 2 FDE taps, $H^\text{P}(f_k)$, given by (5.5). We consider two cases for controlled measurements: (i) only one FDE tap is active, and (ii) both FDE taps are active. Note that the programmable attenuators have a maximal attenuation of only 15.5 dB (see Section 5.2.1) and at this maximal attenuation, signals can still leak through the FDE tap, resulting in inseparable behaviors between the two FDE taps.

To minimize the effect of the second FDE tap, we set the 1st FDE tap at its highest amplitude
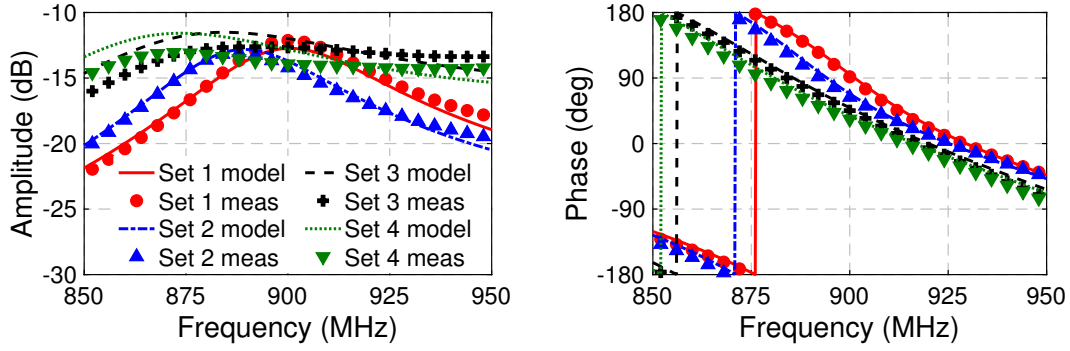
Figure 5.6: Modeled and measured amplitude and phase responses of the PCB canceller, where only the 1st FDE tap is active, with varying (C_F, C_Q) indicated in Table 5.2.

	Set 1	Set 2	Set 3	Set 4
Max Difference in Amplitude (1st tap)	1.0 dB	0.9 dB	1.7 dB	1.5 dB
Max Difference in Phase (1st tap)	2.2°	7.6°	7.2°	11.4°
Max Difference in Amplitude (2nd tap)	1.6 dB	1.6 dB	2.5 dB	2.2 dB
Max Difference in Phase (2nd tap)	6.4°	5.1°	11.8°	9.7°

Table 5.4: Accuracy of the PCB canceller model, where only the 1st or 2nd FDE tap is active, with varying (C_F, C_Q) indicated in Table 5.2.

(i.e., lowest attenuation value of A_1^P) with varying values of $(C_{F,1}, C_{Q,1})$ while setting the 2nd FDE tap with the lowest amplitude (i.e., highest attenuation value of A_2^P). Fig. 5.6 shows the modeled and measured amplitude and phase responses of the PCB canceller in this case, i.e., only the 1st FDE tap is active. The corresponding maximum differences are summarized in Table 5.4. The results show that at the highest BPF quality factor value, the maximum differences in amplitude and phase for Set 1/2 are 1.0/0.9 dB and 2.2°/7.6°, respectively. At the lowest BPF quality factor value, the maximum differences for Set 3/4 are 1.7/1.5 dB and 7.2°/11.4°, respectively, still validating the PCB canceller model. Table 5.4 also shows similar results in the case where only the 2nd FDE tap is active by setting highest attenuation value of A_1^P and lowest attenuation value of A_2^P. The measurements are repeated with different $\{A_i^P, \phi_i^P, C_{F,i}, C_{Q,i}\}$ settings for $i = 1, 2$, and all the results demonstrate the same level of accuracy of the PCB canceller model (5.5) (see Fig. 5.7 for an example).

5.4 Experimental Evaluation

In this section, we discuss the integration of the PCB canceller described in Section 5.2 with an SDR testbed. Then, we present extensive experimental evaluation of the FDE-based FD radio at the node, link, and network levels.

5.4.1 Implementation and Testbed

FDE-based FD radio and the SDR testbed. Figs. 5.1(b) and 5.1(c) depict our FDE-based FD radio design (whose diagram is shown in Fig. 5.2) and the SDR testbed. A 698–960 MHz swivel blade antenna and a coaxial circulator with operating frequency range 860–960 MHz are used as the antenna interface. We use the NI USRP-2942 SDR with the SBX-120 daughterboard operating at 900 MHz carrier frequency, which is the same as the operating frequency of the PCB canceller. As mentioned in Section 5.2.1, our PCB canceller design can be easily extended to other operating frequencies. The USRP has a measured noise floor of -85 dBm at a fixed receiver gain setting.[5]

We implemented a full OFDM-based PHY layer using NI LabVIEW on a host PC with a real-time RF bandwidth of $B = 20$ MHz.[6] The baseband complex (IQ) samples are streamed between the USRP and the host PC through a high-speed PCI-Express interface. The OFDM symbol size is 64 samples (subcarriers) with a cyclic-prefix ratio of 0.25 (16 samples). Throughout the evaluation, $\{f_k\}_{k=1}^{K=52}$ is used to represent the center frequency of the 52 non-zero subcarriers. The OFDM PHY layer supports various modulation and coding schemes (MCSs) with constellations from BPSK to 64QAM and coding rates of 1/2, 2/3, and 3/4, resulting in a highest (HD) data rate of 54 Mbps. The digital SIC algorithm with a highest non-linearity order of 7 is implemented in LabVIEW to further suppress the residual SI signal after RF SIC.[7]

In total, our testbed consists of 3 FDE-based FD radios, whose performance is experimentally

---

[5]This USRP receiver noise floor is limited by the environmental interference at around 900 MHz. The USRP has a true noise floor of around -95 dBm at the same receiver gain setting, when not connected to an antenna.

[6]We consider a general OFDM-based PHY but do not consider the specific frame/packet structure defined by the standards (e.g., LTE or WiFi PHY).

[7]The digital SIC algorithm is based on Volterra series and a least-square problem, which is similar to that presented in [81]. We omit the details here due to limited space.

evaluated at the node, link, and network levels. Regular USRPs (without the PCB canceller) are also included in scenarios where additional HD users are needed.

Optimized PCB canceller configuration. The optimized PCB canceller configuration scheme is implemented on the host PC and the canceller is configured by a SUB-20 controller through the USB interface. For computational efficiency, the PCB canceller response (5.5) (which is validated in Section 5.3 and is independent of the environment) is pre-computed and stored. The canceller configuration has the following steps.

1. Measure the real-time antenna interface response, $H_{\text{SI}}(f_k)$, using a preamble of 2 OFDM symbols by dividing the received preamble by the known transmitted preamble in the frequency domain;

2. Solve for an initial PCB canceller configuration using optimization (P2) based on the measured $H_{\text{SI}}(f_k)$ and the canceller model (5.5) (see Section 5.2.3). The returned configuration parameters are rounded to their closest possible values based on hardware resolutions (see Section 5.2.1);

3. Perform a finer-grained local search and record the optimal canceller configuration (usually around 10 iterations).

In our design, the optimized PCB canceller configuration can be obtained in less than 10 ms on a regular PC with quad-core Intel i7 CPU via a non-optimized MATLAB solver. Specifically, 100 instance runs of the optimized canceller configuration scheme require an average runtime of 6.7 ms with a standard deviation of 0.4 ms. Assuming that the canceller needs to be configured once per second, this is less than 1% overhead. We note that a C-based optimization solver and/or an implementation based on FPGA/look-up table can significantly improve the performance of the canceller configuration scheme and is left for future work.

5.4.2 Node-Level: Microbenchmarks

Optimized PCB canceller response and RF SIC. We set up an FDE-based FD radio running the optimized PCB canceller configuration scheme and record the canceller configuration, measured
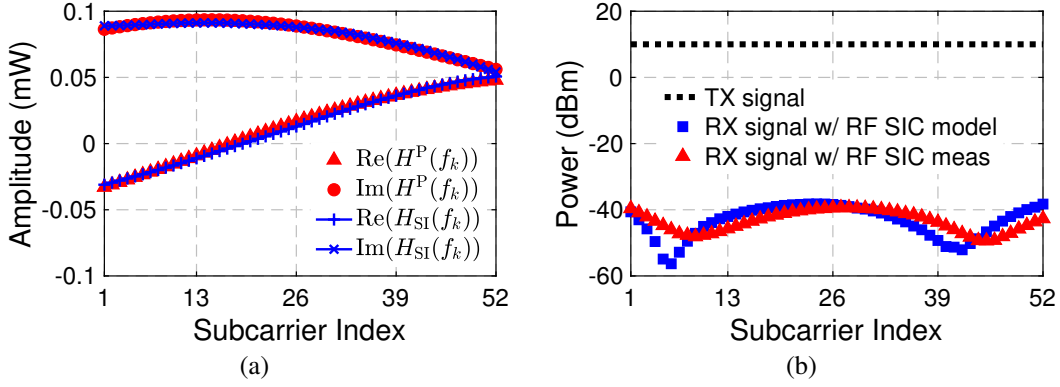
Figure 5.7: (a) Real and imaginary parts of the optimized PCB canceller response, $H^{\mathrm{P}}(f_k)$, vs. real-time SI channel measurements, $H_{\mathrm{SI}}(f_k)$, and (b) modeled and measured RX signal power after RF SIC at 10 dBm TX power. An average 52 dB RF SIC across 20 MHz is achieved.
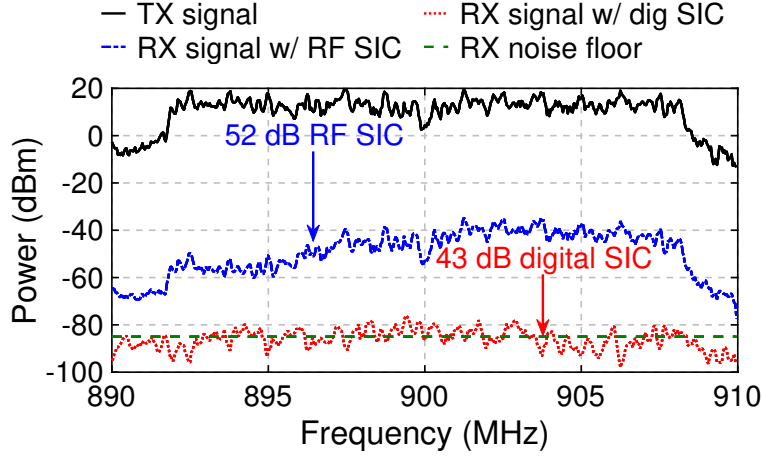
Figure 5.8: Power spectrum of the received signal after SIC in the RF and digital domains with 10 dBm average TX power, 20 MHz bandwidth, and -85 dBm receiver noise floor.

$H_{\mathrm{SI}}(f_k)$, and measured residual SI power after RF SIC. The recorded canceller configuration is then used to compute the PCB canceller response using (5.5).

Fig. 5.7(a) shows an example of the optimized PCB canceller response, $H^{\mathrm{P}}(f_k)$, and the measured antenna interface response, $H_{\mathrm{SI}}(f_k)$, in real and imaginary parts (or I and Q). It can be seen that $H^{\mathrm{P}}(f_k)$ closely matches with $H_{\mathrm{SI}}(f_k)$ with maximum differences in amplitude and phase are only 0.5 dB and 2.5°, respectively. This confirms the accuracy of the PCB canceller model and the performance of the optimized canceller configuration. Fig. 5.7(b) shows the modeled (computed by subtracting the modeled canceller response from the measured $H_{\mathrm{SI}}(f_k)$) and measured

RX signal power after RF SIC at $+10\,$dBm TX power. The results show that the FDE-based FD radio achieves an average $52\,$dB RF SIC across $20\,$MHz bandwidth, from which $20\,$dB is obtained from the antenna interface isolation. Similar performance is observed throughout the experimental evaluation. Note that the profile of RF SIC is similar to that presented in the sensitivity analysis of the FDE-based cancellers (see Section 5.5).

Overall SIC. We measure the overall SIC achieved by the FDE-based FD radio including the digital SIC in the same setting as described above, and the results are presented in Fig. 5.8. It can be seen that the FDE-based FD radio achieves an average $95\,$dB overall SIC across $20\,$MHz, from which $52\,$dB and $43\,$dB are obtained in the RF and digital domains, respectively. Recall from Section 5.4.1 that the USRP has noise floor of $-85\,$dBm, the FDE-based RF radio supports a maximal average TX power of $10\,$dBm (where the peak TX power can go as high as $20\,$dBm). We use TX power levels lower than or equal to $10\,$dBm throughout the experiments, where all the SI can be canceled to below the RX noise floor.

5.4.3   Link-Level: SNR-PRR Relationship

We now evaluate the relationship between link SNR and link packet reception ratio (PRR). We setup up a link with two FDE-based FD radios at a fixed distance of 5 meters with equal TX power. In order to evaluate the performance of our FD radios with the existence of the PCB canceller, we set an FD radio to operate in HD mode by turning on only its transmitter or receiver. We conduct the following experiment for each of the 12 MCSs in both FD and HD modes, with varying TX power levels. In particular, the packets are sent over the link simultaneously in FD mode or in alternating directions in HD mode (i.e., the two radios take turns and transmit to each other). In each experiment, both radios send a sequence of 50 OFDM streams, each OFDM stream contains 20 OFDM packets, and each OFDM packet is 800-Byte long.

We consider two metrics. The *HD (resp. FD) link SNR* is measured as the ratio between the average RX signal power in both directions and the RX noise floor when both radios operate in HD (resp. FD) mode. The *HD (resp. FD) link PRR* is computed as the fraction of packets successfully

| (a) Code rate 1/2 | (b) Code rate 3/4 |
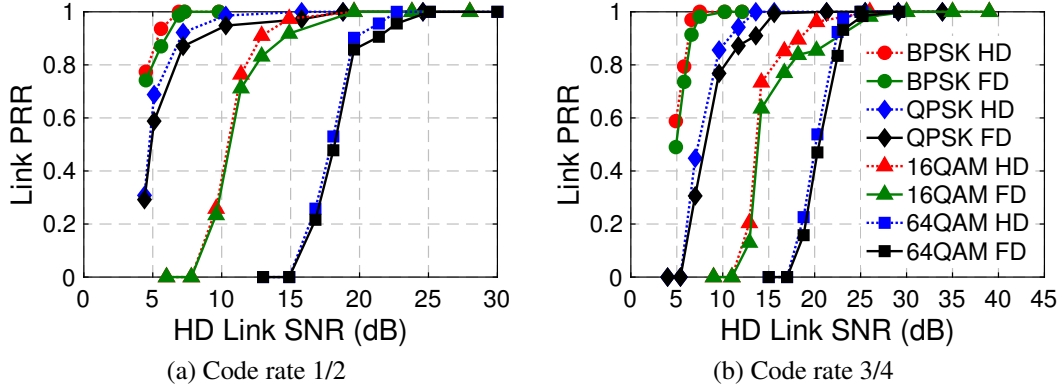
Figure 5.9: HD and FD link packet reception ratio (PRR) with varying HD link SNR and modulation and coding schemes (MCSs).

sent over the HD (resp. FD) link in each experiment. We observe from the experiments that the HD and FD link SNR and PRR values in both link directions are similar. Similar experiments and results were presented in [153] for HD links.

Fig. 5.9 shows the relationship between link PRR values and HD link SNR values with varying MCSs. The results show that with sufficient link SNR values (e.g., 8 dB for BPSK-1/2 and 28 dB for 64QAM-3/4), the FDE-based FD radio achieves a link PRR of 100%. With insufficient link SNR values, the average FD link PRR is 6.5% lower than the HD link PRR across varying MCSs. This degradation is caused by the link SNR difference when the radios operate in HD or FD mode, which is described later in Section 5.4.4. Since packets are sent simultaneously in both directions on an FD link, this average PRR degradation is equivalent to an average FD link throughput gain of $1.87\times$ under the same MCS.

5.4.4 Link-Level: SNR Difference and FD Gain

Experimental setup. To thoroughly evaluate the link level FD throughput gain achieved by our FD radio design, we conduct experiments with two FD radios with 10 dBm TX power, one emulating a base station (BS) and one emulating a user. We consider both line-of-sight (LOS) and non-line-of-sight (NLOS) experiments as shown in Fig. 5.10. In the LOS setting, the BS is placed at the end of a hallway and the user is moved away from the BS at stepsizes of 5 meters up to a distance of
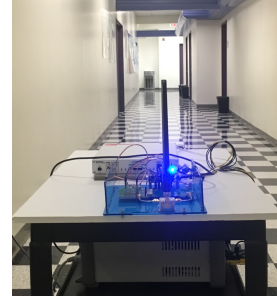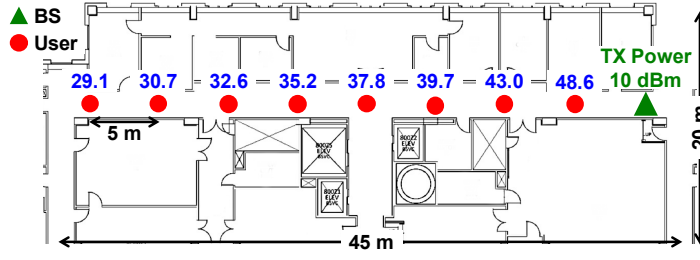
(a) LOS deployment and an FD radio in a hallway
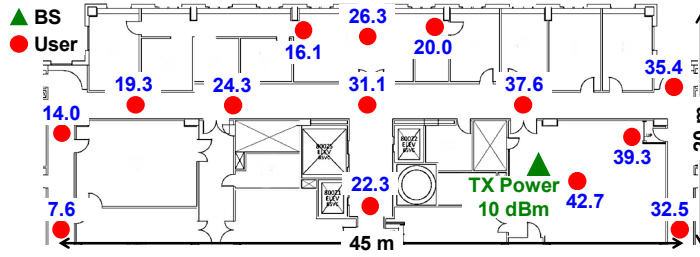
(b) NLOS deployment and an FD radio in a lab environment

Figure 5.10: (a) Line-of-sight (LOS), and (b) non-line-of-sight (NLOS) deployments, and the measured HD link SNR values (dB).

40 meters. In the NLOS setting, the BS is placed in a lab environment with regular furniture and the user is placed at various locations (offices, labs, and corridors). We place the BS and the users at about the same height across all the experiments.[8] The measured HD link SNR values are also included in Fig. 5.10. Following the methodology of [81], for each user location, we measure the *link SNR difference*, which is defined as the absolute difference between the average HD and FD link SNR values. Throughout the experiments, link SNR values between 0–50 dB are observed.

Difference in HD and FD link SNR values. Fig. 5.11 shows the measured link SNR difference as a function of the HD link SNR (i.e., for different user locations) in the LOS and NLOS experiments, respectively, with 64QAM-3/4 MCS. For the LOS experiments, the average link SNR difference is 0.6 dB with a standard deviation of 0.2 dB. For the NLOS experiments, the average link SNR difference is 0.6 dB with a standard deviation of 0.3 dB. The SNR difference has a higher variance in the NLOS experiments, due to the complicated environments (e.g., wooden desks and

---

[8]In this work, we emulate the BS and users without focusing on specific deployment scenarios. The impacts of different antenna heights and user densities, as mentioned in [154], will be considered in future work.

(a) LOS Experiment (b) NLOS Experiment
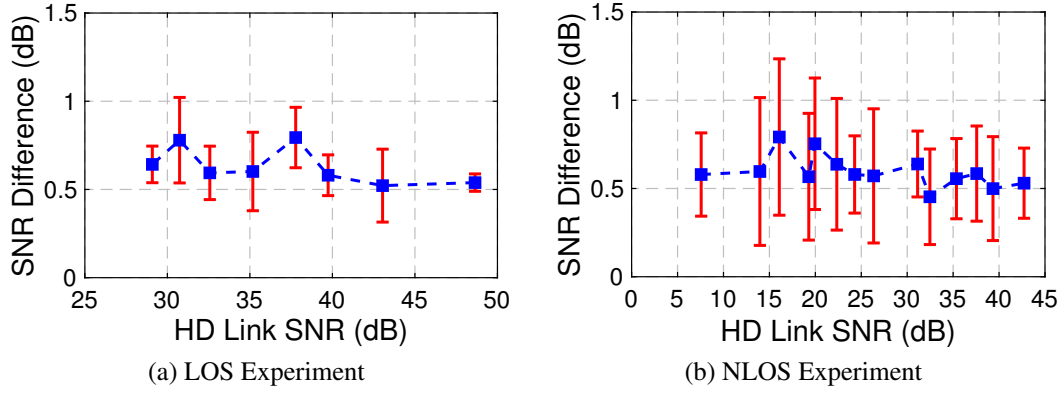
Figure 5.11: Difference between HD and FD link SNR values in the (a) LOS, and (b) NLOS experiments, with 10 dBm TX power and 64QAM-3/4 MCS.

(a) (b)
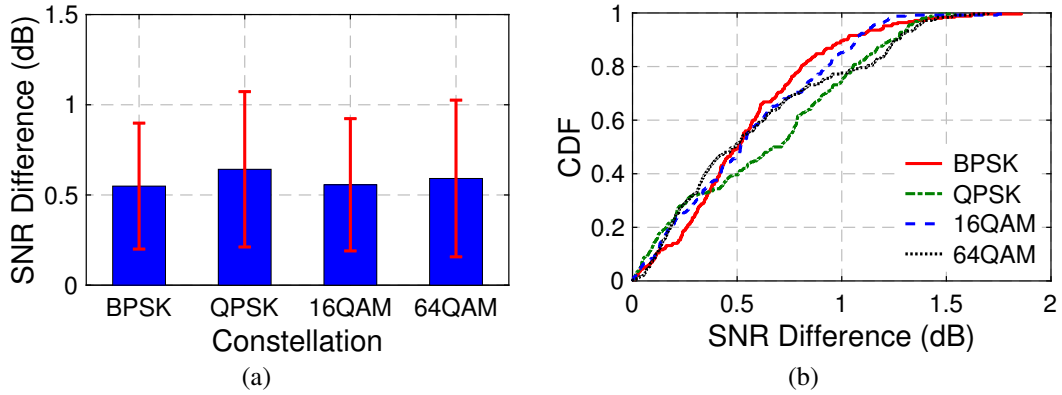
Figure 5.12: Difference between HD and FD link SNR values with 10 dBm TX power under varying constellations: (a) mean and standard deviation, (b) CDF.

chairs, metal doors and bookshelves, etc.). In both cases, the link SNR difference is minimal and uncorrelated with user locations, showing the robustness of the FDE-based FD radio.

Impact of constellations. Fig. 5.12 shows the measured link SNR difference and its CDF with varying constellations and 3/4 coding rate. It can be seen that the link SNR difference has a mean of 0.6 dB and a standard deviation of 0.4 dB, both of which are uncorrelated with the constellations.

FD link throughput and gain. For each user location in the LOS and NLOS experiments, the HD (resp. FD) link throughput is measured as the highest average data rate across all MCSs achieved by the link when both nodes operate in HD (resp. FD) mode . The FD gain is computed as the ratio between FD and HD throughput values. Recall that the maximal HD data rate is 54 Mbps, an FD link data rate of 108 Mbps can be achieved with an FD link PRR of 1.
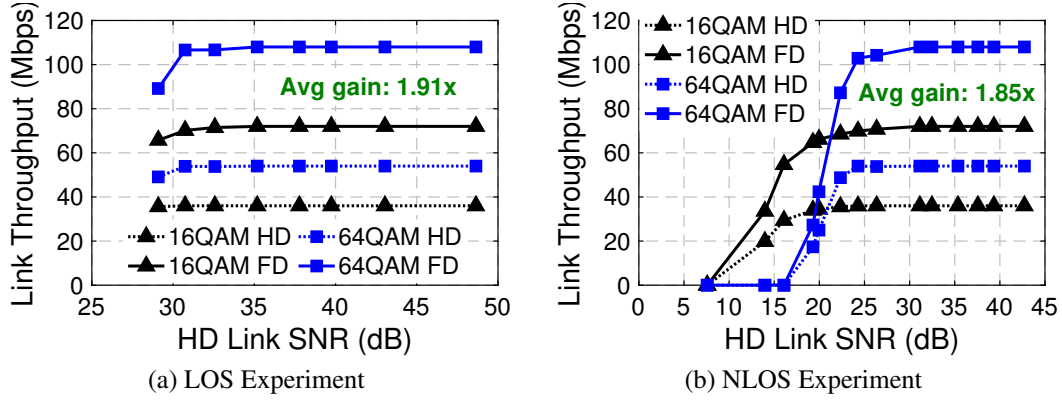
Figure 5.13: HD and FD link throughput in the (a) LOS, and (b) NLOS experiments, with $10\,\mathrm{dBm}$ TX power and 16QAM-3/4 and 64QAM-3/4 MCSs.

Fig. 5.13 shows the average HD and FD link throughput with varying 16QAM-3/4 and 16QAM-3/4 MCSs, where each point represents the average throughput across 1,000 packets. The results show that with sufficient link SNR values (e.g., $30\,\mathrm{dB}$ for 64QAM-3/4 MCS), the FDE-based FD radios achieve an *exact* link throughput gain of $2\times$. In these scenarios, the HD/FD link always achieves a link PRR of 1 which results in the maximum achievable HD/FD link data rate. With medium link SNR values, where the link PRR less than 1, the average FD link throughput gains across different MCSs are $1.91\times$ and $1.85\times$ for the LOS and NLOS experiments, respectively. We note that if higher modulation schemes (e.g., 256QAM) are considered and the corresponding link SNR values are high enough for these schemes, the HD/FD throughput can increase (compared to the values in Fig. 5.13). However, considering such schemes is not required in order to evaluate the FDE-based cancellers and the FD gain.

5.4.5 Network-Level FD Gain

We now experimentally evaluate the network-level throughput gain introduced by FD-capable BS and users. The users can significantly benefit from the FDE-based FD radio suitable for hand-held devices. We compare experimental results to the analysis (e.g., [95]) and demonstrate practical FD gain in different network settings. Specifically, we consider two types of networks as depicted in Fig. 5.14: (i) *UL-DL networks* with one FD BS and two HD users with inter-user interference
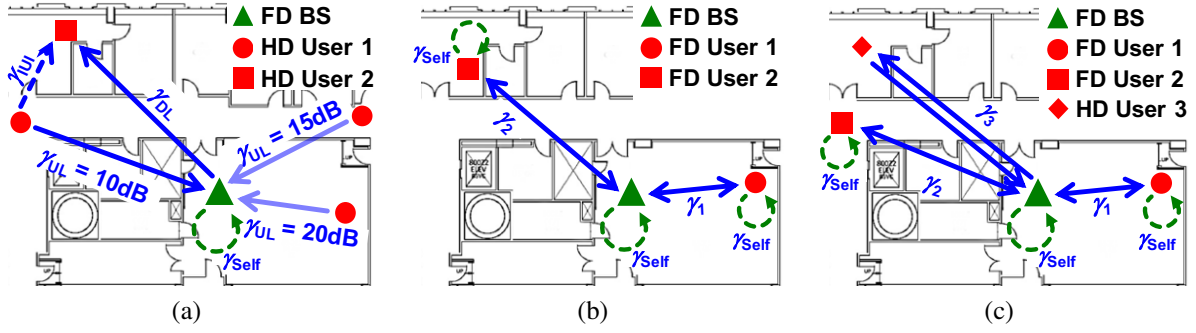
Figure 5.14: An example experimental setup for: (a) the UL-DL networks with varying γ_{UL} and γ_{DL}, (b) heterogeneous 3-node network with one FD BS and 2 FD users, and (c) heterogeneous 4-node networks with one FD BS, 2 FD users, and one HD user.

(IUI), and (ii) *heterogeneous HD-FD networks* with HD and FD users. Due to software challenge with implementing a real-time MAC layer using a USRP, we apply a TDMA setting where each (HD or FD) user takes turn to be activated for the same period of time.

UL-DL Networks with IUI. We first consider UL-DL networks consisting of one FD BS and two HD users (indexed 1 and 2). Without loss of generality, in this setting, user 1 transmits on the UL to the BS, and the BS transmits to user 2 on the DL (see Fig. 5.14(a)).

Analytical FD gain. We use Shannon's capacity formula $r(\gamma) = B \cdot \log_2(1 + \gamma)$ to compute the *analytical throughput* of a link under bandwidth B and (HD) link SNR γ. If the BS is only HD-capable, the network throughput in a UL-DL network when the UL and DL share the channel in a TDMA manner with equal fraction of time is given by

$$r_{\text{UL-DL}}^{\text{HD}} = \frac{B}{2} \cdot \log_2\left(1 + \gamma_{\text{UL}}\right) + \frac{B}{2} \cdot \log_2\left(1 + \gamma_{\text{DL}}\right), \tag{5.6}$$

where γ_{UL} and γ_{DL} are the UL and DL SNRs, respectively. If the BS is FD-capable, the UL and DL can be simultaneously activated with an analytical network throughput of

$$r_{\text{UL-DL}}^{\text{FD}} = B \cdot \log_2(1 + \frac{\gamma_{\text{UL}}}{1 + \gamma_{\text{Self}}}) + B \cdot \log_2(1 + \frac{\gamma_{\text{DL}}}{1 + \gamma_{\text{IUI}}}), \tag{5.7}$$

where: (i) $\left(\frac{\gamma_{\text{DL}}}{1+\gamma_{\text{IUI}}}\right)$ is the signal-to-interference-plus-noise ratio (SINR) at the DL HD user, and (ii) γ_{Self} is the residual self-interference-to-noise ratio (XINR) at the FD BS. We set $\gamma_{\text{Self}} = 1$ when
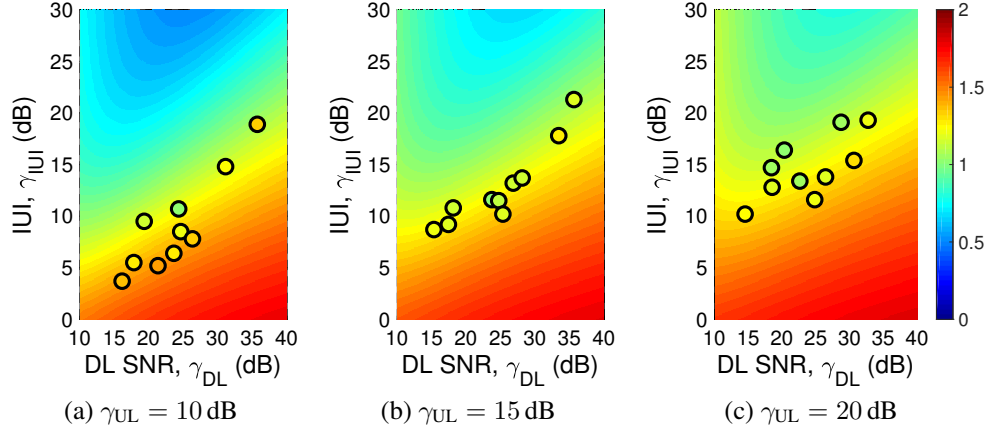
Figure 5.15: Analytical (colored surface) and experimental (filled circles) network throughput gain for UL-DL networks consisting of one FD BS and two HD users with varying UL and DL SNR values, and inter-user interference (IUI) levels: (a) $\gamma_{\text{UL}} = 10\,\text{dB}$, (b) $\gamma_{\text{UL}} = 15\,\text{dB}$, and (c) $\gamma_{\text{UL}} = 20\,\text{dB}$. The baseline is the network throughput when the BS is HD.

UL SNR, γ_{UL}	Analytical FD Gain	Experimental FD Gain
10 dB	1.30×	1.25×
15 dB	1.23×	1.16×
20 dB	1.22×	1.14×

Table 5.5: Average FD Gain in UL-DL Networks with IUI.

computing the analytical throughput. Namely, the residual SI power is no higher than the RX noise floor (which can be achieved by the FDE-based FD radio, see Section 5.4.2). The *analytical FD gain* is then defined as the ratio $(r^{\text{FD}}_{\text{UL-DL}}/r^{\text{HD}}_{\text{UL-DL}})$. Note that the FD gain depends on the coupling between γ_{UL}, γ_{DL}, and γ_{IUI}, which depend on the BS/user locations, their TX power levels, etc.

Experimental FD gain. The experimental setup is depicted in Fig. 5.14(a), where the TX power levels of the BS and user 1 are set to be $10\,\text{dBm}$ and $-10\,\text{dBm}$, respectively. We fix the location of the BS and consider different UL SNR values of $\gamma_{\text{UL}} = 10/15/20\,\text{dB}$ by placing user 1 at three different locations. For each value of γ_{UL}, user 2 is placed at 10 different locations, resulting in varying γ_{DL} and γ_{IUI} values.

Fig. 5.15 shows the analytical (colored surface) and experimental (filled circles) FD gain, where the analytical gain is extracted using (5.6) and (5.7), and the experimental gain is computed using the measured UL and DL throughput. It can be seen that smaller values of γ_{UL} and lower ratios

between γ_{DL} and γ_{IUI} lead to higher throughput gains in both analysis and experiments. The average analytical and experimental FD gains are summarized in Table 5.5. Due to practical reasons such as the link SNR difference and its impact on link PRR (see Section 5.4.3), the experimental FD gain is within 93% of the analytical FD gain. The results confirm the analysis in [95] and demonstrate the practical FD gain achieved in wideband UL-DL networks without any changes in the current network stack (i.e., only bringing FD capability to the BS). Moreover, performance improvements are expected through advanced power control and scheduling schemes.

Heterogeneous 3-Node Networks. We consider heterogeneous HD-FD networks with 3 nodes: one FD BS and two users that can operate in either HD or FD mode (see an example experimental setup in Figs. 5.1(c) and 5.14(b)). All 3 nodes have the same $0\,\text{dBm}$ TX power so that each user has symmetric UL and DL SNR values of γ_i ($i = 1, 2$). We place user 1 at 5 different locations and place user 2 at 10 different locations for each location of user 1, resulting in a total number of 50 pairs of (γ_1, γ_2).

Analytical FD gain. We set the users to share the channel in a TDMA manner. The analytical network throughput in a 3-node network with zero, one, and two FD users is respectively given by

$$r^{\text{HD}} = \frac{B}{2} \cdot \log_2\left(1 + \gamma_1\right) + \frac{B}{2} \cdot \log_2\left(1 + \gamma_2\right), \tag{5.8}$$

$$r^{\text{HD-FD}}_{\text{User } i \text{ FD}} = B \cdot \log_2\left(1 + \frac{\gamma_i}{1 + \gamma_{\text{Self}}}\right) + \frac{B}{2} \cdot \log_2\left(1 + \gamma_{\bar{i}}\right), \tag{5.9}$$

$$r^{\text{FD}} = B \log_2 \cdot \left(1 + \frac{\gamma_1}{1 + \gamma_{\text{Self}}}\right) + B \log_2 \cdot \left(1 + \frac{\gamma_2}{1 + \gamma_{\text{Self}}}\right), \tag{5.10}$$

where $\gamma_{\text{Self}} = 1$ is set. We consider both FD gains of $(r^{\text{HD-FD}}_{\text{User } i \text{ FD}}/r^{\text{HD}})$ (i.e., user i is FD and user $\bar{i} \neq i$ is HD), and $(r^{\text{FD}}/r^{\text{HD}})$ (i.e., both users are FD).

Experimental FD gain. For each pair of (γ_1, γ_2), experimental FD gain is measured in three cases: (i) only user 1 is FD, (ii) only user 2 is FD, and (iii) both users are FD. Fig. 5.16 shows the analytical (colored surface) and experimental (filled circles) FD gain for each case. We exclude the results with $\gamma_i < 3\,\text{dB}$ since the packets cannot be decoded, resulting in a throughput of zero (see Fig. 5.9).

The results show that with small link SNR values, the experimental FD gain is lower than the
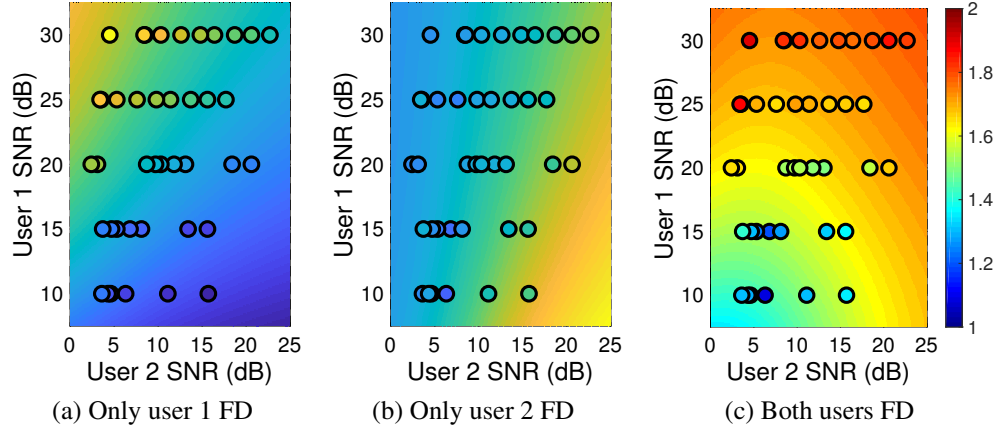
Figure 5.16: Analytical (colored surface) and experimental (filled circles) network throughput gain for 3-node networks consisting of one FD BS and two users with varying link SNR values: (a) only user 1 is FD, (b) only user 2 is FD, and (c) both users are FD. The baseline is the network throughput when both users are HD.
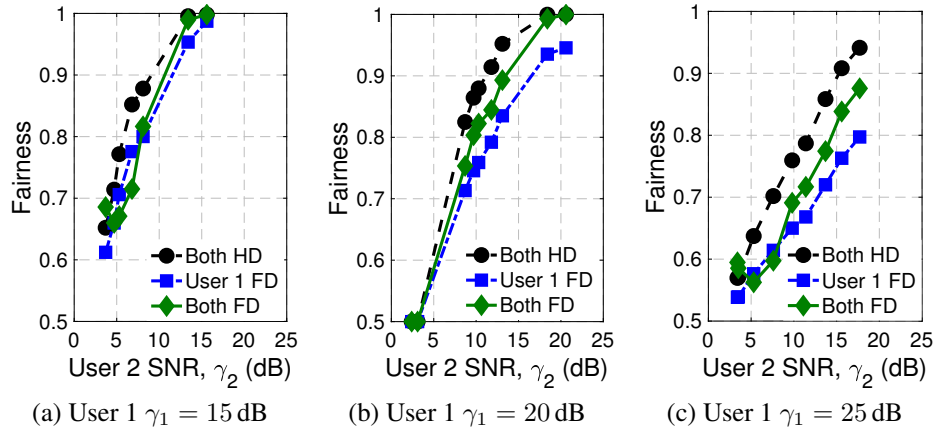
Figure 5.17: Measured Jain's fairness index (JFI) in 3-node networks where both users are HD, User 1 is FD, and both users are FD, with varying (γ_1, γ_2).

analytical value due to the inability to decode the packets. On the other hand, with sufficient link SNR values, the experimental FD gain exceeds the analytical FD gain. This is because setting $\gamma_{\mathrm{Self}} = 1$ in (5.9) and (5.10) results in a 3 dB SNR loss in the analytical FD link SNR, and thereby in a lower throughput. However, in practice, the packets can be decoded with a link PRR of 1 with sufficient link SNRs, resulting in exact twice number of packets being successfully sent over an FD link. Moreover, the FD gain is more significant when enabling FD capability for users with higher link SNR values.
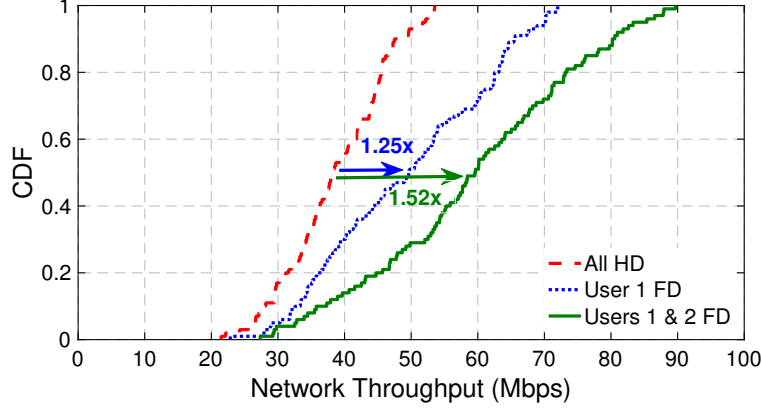
Figure 5.18: CDF of the experimental network throughput in 4-node networks when zero, one, or two users are FD-capable.

Another important metric we consider is the fairness between users, which is measured by the Jain's fairness index (JFI). In the considered 3-node networks, the JFI ranges between 0.5 (worst case) and 1 (best case). Fig. 5.17 shows the measured JFI when both users operate in HD mode, user 1 operates in FD mode, and both users operate in FD mode. with varying user SNR values (γ_1, γ_2). The results show that introducing FD capability to both users results in an average degradation in the network JFI of only 5.6/4.4/7.4% for $\gamma_1 = 15/20/25$ dB (averaged across varying γ_2), while the average network FD gains are 1.32/1.58/1.73$\times$ (see Fig. 5.16), respectively. In addition, the JFI increases with more balanced user SNR values, which is as expected. For example, under the same value of γ_1, increased value of γ_2 (with $\gamma_2 < \gamma_1$) leads to improved JFI, whose value approaches to 1 as γ_2 approaches γ_1.

Heterogeneous 4-Node Networks. We experimentally study 4-node networks consisting of an FD BS and three users with 10 dBmd TX power (see an example experimental setup in Fig. 5.14(c)). The experimental setup is similar to that described above. 100 experiments are conducts where the 3 users are placed at different locations with different user SNR values. For each experiment, the network throughput is measured in three cases where: (i) zero, (ii) one, and (iii) two users are FD-capable.

Fig. 5.18 shows the CDF of the network throughput of the three cases, where the measured link SNR varies between 5–45 dB.Overall, the median network throughput is increased by 1.25/1.52$\times$
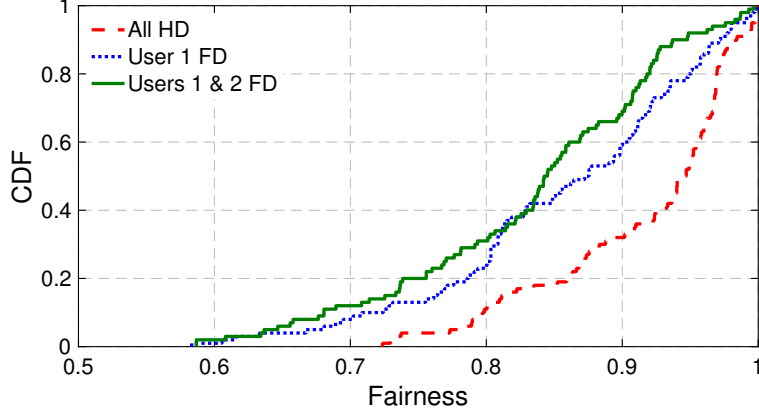
Figure 5.19: CDF of the measured Jain's fairness index (JFI) in 4-node networks when zero, one, or two users are FD-capable.
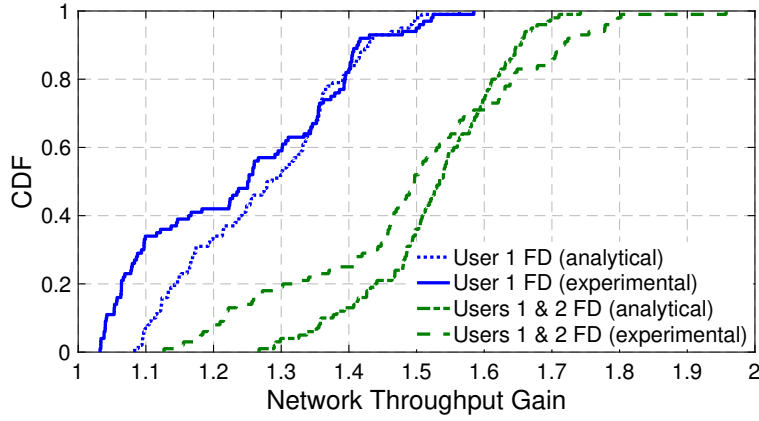
Figure 5.20: CDF of the analytical and experimental network throughput gains in 4-node networks when one or two users are FD-capable.

when one/two FD users become FD-capable. Fig. 5.19 plots the CDF of the corresponding JFI, where although introducing FD-capable users results in lower values of the experimental JFI, the median JFI values degrade by only 0.06 and 0.10 with one and two FD users, respectively. Moreover, Fig. 5.20 shows the CDF of both the analytical and experimental network throughput gains in 4-node networks when one or two users are FD-capable. In particular, the median analytical and experimental network throughput gains have a difference of only 4% and 3% when one and two users are FD-capable. These trends and results show that in a real-world environment, the total network throughput increases as more users become FD-capable, and the improvement is more significant with higher user SNR values. Note that we only apply a TDMA scheme and a more advanced MAC layer (e.g., [29, 102]) has the potential to improve the FD gain and fairness

performance.

5.5 Numerical Evaluation

In this section, we numerically evaluate and compare the performance of the FDE-based RFIC [152] and PCB cancellers *based on measurements and validated models*. We confirm that the PCB canceller emulates its RFIC counterpart and show that the optimized canceller configuration scheme can significantly improve the performance of the RFIC canceller. We also evaluate the performance of FDE-based cancellers with respect to the number of FDE taps, M, and desired RF SIC bandwidth, B, and discuss various design tradeoffs.

5.5.1 Setup

We use a real, practical antenna interface response, $H_{\mathrm{SI}}(f_k)$, measured in the same setting as described in Section 5.4.1, and consider $M \in \{1, 2, 3, 4\}$ and $B \in \{20, 40, 80\}$ MHz. We only report the RF SIC performance with up to 4 FDE taps since, as we will show, this case can achieve sufficient RF SIC up to 80 MHz bandwidth.[9]

We use (5.2) to both model and evaluate the RFIC canceller with configuration parameters $\{A_i^{\mathrm{I}}, \phi_i^{\mathrm{I}}, f_{\mathrm{c},i}, Q_i\}$, since it is shown that a 2nd-order BPF can accurately model the FDE N-path filter [152]. Similar to (P2) (see Section 5.2.3), the optimized RFIC canceller configuration can be obtained by solving (P3) with $H^{\mathrm{I}}(f_k)$ given by (5.2).

$$(\mathsf{P3}) \ \min: \ \sum_{k=1}^{K} \left| H_{\mathrm{res}}^{\mathrm{I}}(f_k) \right| = \sum_{k=1}^{K} \left| H_{\mathrm{SI}}(f_k) - H^{\mathrm{I}}(f_k) \right|^2$$

subject to: $A_i^{\mathrm{I}} \in [A_{\min}^{\mathrm{I}}, A_{\max}^{\mathrm{I}}], \ \phi_i^{\mathrm{I}} \in [-\pi, \pi], \ f_{\mathrm{c},i} \in [f_{\mathrm{c,min}}, f_{\mathrm{c,max}}], \ Q_i \in [Q_{\min}, Q_{\max}], \ \forall i.$

Note that in [152], there is no optimization of the RFIC canceller configuration, and the canceller is configured based on a heuristic approach. As we will show, the optimized canceller scheme outperforms the heuristic approach by an order of magnitude in terms of the amount of RF SIC

---

[9]We select typical values of $20/40/80$ MHz as the desired RF SIC bandwidth, since the circulator has a frequency range of 100 MHz.
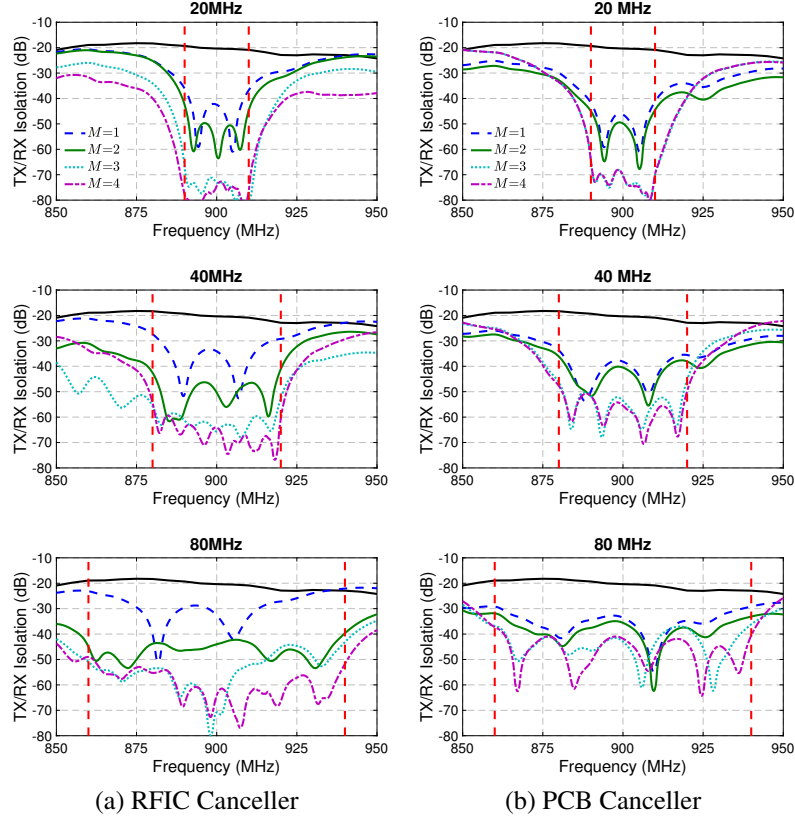
Figure 5.21: TX/RX isolation of the antenna interface (black curve) and with the RFIC and PCB cancellers with varying number of FDE taps, M, and desired RF SIC bandwidth, B, *in the ideal case*.

achieved.

The implemented PCB canceller includes only $M = 2$ FDE taps due to its design (see Section 5.2). However, it is practically feasible to include more parallel FDE taps. For numerical evaluation purposes, we model the PCB canceller with $M > 2$ FDE taps by extending the validated model (5.5) with symmetric FDE taps (i.e., all BPFs in the FDE taps behave identically). Although the canceller configuration scheme has a computational complexity of 4^M (i.e., four DoF per FDE tap), we will show that $M = 4$ taps can achieve sufficient amount of RF SIC in realistic scenarios.

In practice, the canceller configuration parameters cannot be arbitrarily selected from a continuous range as described in (P2) and (P3). Instead, they are often restricted to discrete values given the resolution of the corresponding hardware components. To address this problem, we evaluate

the canceller models in both the *ideal* case and the case *with practical quantization constraints*. The canceller configuration with quantization constraints are obtained by rounding the configuration parameters returned by solving (P2) or (P3) to their closest quantized values.

In particular, the RFIC canceller has the following constraints: $\forall i$, $A_i^{\mathrm{I}} \in [-40, -10]\,\mathrm{dB}$, $\phi_i^{\mathrm{I}} \in [-\pi, \pi]$, $f_{\mathrm{c},i} \in [875, 925]\,\mathrm{MHz}$, and $Q_i \in [1, 50]$. When adding practical quantization constraints, we assume that the amplitude A_i^{I} has a 0.25 dB resolution. For ϕ_i^{I}, $f_{\mathrm{c},i}$, and Q_i, an 8-bit resolution constraint is introduced, which is equivalent to 256 discrete values spaced equally in the given range. These constraints are practically selected and can be easily realized in an IC implementation. The PCB canceller model has following constraints: $\forall i$, $A_i^{\mathrm{P}} \in [-15.5, 0]\,\mathrm{dB}$, $\phi_i^{\mathrm{P}} \in [-\pi, \pi]$, $C_{\mathrm{F},i} \in [0.6, 2.4]\,\mathrm{pF}$, and $C_{\mathrm{Q},i} \in [2, 14]\,\mathrm{pF}$. When adding the quantization constraints, we consider 0.5 dB, 0.12 pF, and 0.39 pF resolution to A_i^{P}, $C_{\mathrm{F},i}$, and $C_{\mathrm{Q},i}$, respectively, and ϕ_i^{P} has an 8-bit resolution. These numbers are consistent with our implementation and experiments (see Sections 5.2.1 and 5.4).

5.5.2 Performance Evaluation and Comparison between the RFIC and PCB Cancellers

Fig. 5.21 shows the TX/RX isolation achieved by the RFIC and PCB cancellers with optimized canceller configuration, with varying M and B in the ideal case (i.e., without quantization constraints). It can be seen that: (i) under a given value of B, a larger number of FDE taps results in higher average RF SIC, and (ii) for a larger value of B, more FDE taps are required to achieve sufficient RF SIC. For example, the ideal RFIC and PCB cancellers with 2 FDE taps can achieve an average 50/46/42 dB and 50/42/35 dB RF SIC across 20/40/80 MHz bandwidth, respectively.

Fig. 5.22 shows the TX/RX isolation achieved by the RFIC and PCB cancellers with optimized canceller configuration under practical quantization constraints. Comparing to Fig. 5.21, the results show a performance degradation due to limited hardware resolutions, which is more significant as M increases. This is because a larger value of M introduces a higher number of DoF with more canceller parameters that can be flexibly controlled. As a result, the RF SIC performance is more sensitive to the coupling between individual FDE tap responses after quantization. The

116
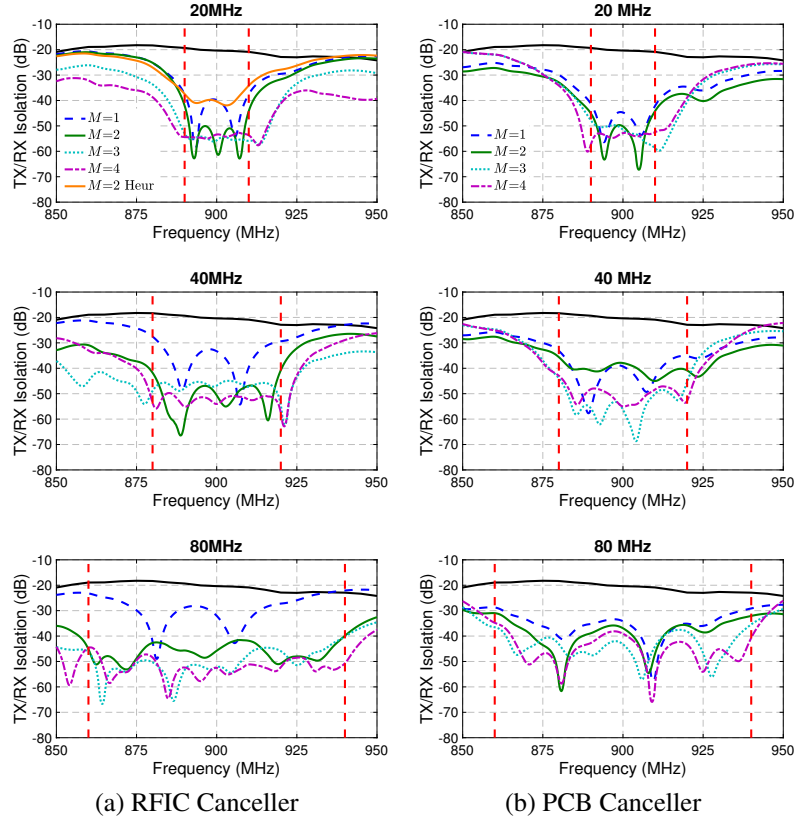
Figure 5.22: TX/RX isolation of the antenna interface (black curve) and with the RFIC and PCB cancellers with varying number of FDE taps, M, and desired RF SIC bandwidth, B, *under practical quantization constraints*.

results show that under practical constraints, the RFIC and PCB cancellers with 4 FDE taps can still achieve an average 54/50/45 dB and 52/45/39 dB RF SIC across 20/40/80 MHz bandwidth, respectively. Fig. 5.22 also shows that the RFIC canceller under the optimized configuration scheme achieves a 10 dB higher RF SIC compared with that achieved by the heuristic approach described in [152] (labeled "Heur").

It is interesting to observe that the RF SIC profile of the PCB canceller with 2 FDE taps is very similar to our experimental results (see Fig. 5.7 in Section 5.4.2). It is also worth noting that, in practice, adding more FDE taps cannot improve the amount of RF SIC in some scenarios (e.g., with 20 MHz bandwidth), which is limited by the quantization constraints. However, performance improvement is expected by relaxing these constraints (e.g., through using components with higher resolutions and/or wider tuning ranges).

117

	PCB (this work)	RFIC [152]
Center Frequency	900 MHz	1.37 GHz
# of FDE Taps	2	2
Antenna Interface	A single antenna and a circulator	A TX/RX antenna pair
Antenna Isolation	20 dB	35 dB
Canceller SIC (20 MHz)	32 dB	20 dB
Canceller Configuration	Optimization (P2)	Heuristic
Digital SIC	43 dB	N/A
Evaluation	Node/Link/Network	Node

Table 5.6: Comparison between the PCB and RFIC cancellers.

Table 5.6 shows the comparison between our implemented PCB canceller and the RFIC canceller presented in [152]. To summarize, we numerically show that the performance of the RFIC and PCB cancellers is similar. The results based on measurements and validated canceller models confirm that the PCB canceller accurately emulates its RFIC counterpart, and that the FDE-based approach is valid and suitable for achieving wideband RF SIC in small-form-factor devices.

5.6 Conclusion

We designed and implemented a PCB canceller using the FDE technique, which was shown to achieve wideband RF SIC in compact nodes. We presented a PCB canceller model and a scheme for optimizing the canceller configuration. We experimentally evaluated the performance of the FDE-based FD radio at the node, link, and network levels using an SDR testbed. We also compared the RFIC and PCB implementations and discussed various design tradeoffs of the FDE-based canceller. In Chapter 6, we extend from single-antenna radios to multi-antenna systems, and focus on enabling FD operation with phased arrays. Recall that our network-level experiments described in Section 5.4.5 only applied a TDMA scheme, and a more advanced MAC layer is needed to achieve improved performance. Therefore, we further explore this problem in Chapter 7, where we develop an efficient and distributed scheduling algorithm for heterogeneous networks with co-existing HD and FD users. In Chapter 9, we present the integration of the FDE-based FD radios in the COSMOS testbed [13, 14].

5.7 Appendix: The PCB BPF Model

We use transmission (ABCD) matrix to derive $H_i^{\mathrm{B}}(f)$, given by (5.3). From Fig. 5.4 and $Y_{\mathrm{F}}(f_k)$ and $Y_{\mathrm{Q}}(f_k)$ in (5.4),

$$
\begin{bmatrix} V_{\mathrm{in}} \\ I_{\mathrm{in}} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ Y_{\mathrm{Q}}(f_k) & 1 \end{bmatrix} \cdot \mathbf{M}^{\mathrm{TL}} \cdot \begin{bmatrix} 1 & 0 \\ Y_{\mathrm{F}}(f_k) & 1 \end{bmatrix} \cdot \mathbf{M}^{\mathrm{TL}} \cdot \begin{bmatrix} 1 & 0 \\ Y_{\mathrm{Q}}(f_k) & 1 \end{bmatrix} \cdot \begin{bmatrix} V_{\mathrm{out}} \\ I_{\mathrm{out}} \end{bmatrix}
$$

$$
:= \begin{bmatrix} M_A^{\mathrm{BPF}}(f_k) & M_B^{\mathrm{BPF}}(f_k) \\ M_C^{\mathrm{BPF}}(f_k) & M_D^{\mathrm{BPF}}(f_k) \end{bmatrix} \cdot \begin{bmatrix} V_{\mathrm{out}} \\ I_{\mathrm{out}} \end{bmatrix}, \tag{5.11}
$$

where $\mathbf{M}^{\mathrm{TL}}$ is the ABCD matrix of a T-Line with wavenumber β, characteristic impedance Z_0, and length l, i.e.,

$$
\mathbf{M}^{\mathrm{TL}} = \begin{bmatrix} \cos\left(\beta l\right) & j Z_0 \sin\left(\beta l\right) \\ j \sin\left(\beta l\right)/Z_0 & \cos\left(\beta l\right) \end{bmatrix}. \tag{5.12}
$$

With the parameters described in Section 5.2.1, the frequency response of the implemented PCB BPF, $H_i^{\mathrm{B}}(f_k)$, is given by

$$
H_i^{\mathrm{B}}(f_k) = \frac{V_{\mathrm{out}}(f_k)}{V_{\mathrm{in}}(f_k)} = \frac{1}{R_{\mathrm{S}}} \cdot \frac{V_{\mathrm{out}}(f_k)}{I_{\mathrm{in}}(f_k)} = \frac{1}{R_{\mathrm{S}}} \cdot \frac{1}{M_C^{\mathrm{BPF}}(f_k)}.
$$

Plugging (5.4) and (5.12) into (5.11) yields the model $H_i^{\mathrm{B}}(f)$.

Chapter 6: Wideband Full-Duplex Phased Array with Joint Transmit and Receive Beamforming: Optimization and Rate Gains

In the previous chapter, we presented the design and experimentation with *single-antenna compact full-duplex (FD) radios* leveraging the technique of frequency-domain equalization (FDE). As mentioned in Section 1.1, combining FD with *multi-antenna* systems has the potential to further improve the data rate and increase the communication range. In this chapter, we show that *an FD phased array can simultaneously achieve wideband RF self-interference cancellation (SIC) with minimal Tx beamforming (TxBF) and Rx beamforming (RxBF) gain losses, and improved FD rate gains*. This is achieved by carefully selecting the Tx and Rx analog beamforming weights (a.k.a., beamformers) or, in other words, TxBF and RxBF can be *repurposed* to support wideband RF SIC and enable FD operation in phased arrays.

We first provide background on phased arrays and beamforming (Section 6.1). We consider network scenarios as depicted in Fig. 6.1, where a base station (BS) equipped with an N-element FD phased array communicates with: (i) a single-antenna user which is HD- or FD-capable (the *BS-User* case), or (ii) with another FD BS (the *BS-BS* case). We derive the data rate gain resulting from using FD, and the FD link objective is to *maximize the FD data rate gains* in the two considered cases (Section 6.2). Using motivating examples, we demonstrate the FD rate gains introduced by FD phased arrays employing TxBF and RxBF in various network settings (Section 6.3). We formulate an optimization problem to jointly determine the optimal FD Tx and Rx beamformers. Due to its non-convexity and analytical intractability, we then present alternative convex optimization problems by leveraging the structural properties of the SI channel matrix and its coupling with the Tx and Rx beamformers. We develop an iterative algorithm that efficiently solves the alternative optimization problems with provable performance guarantees (Section 6.4).
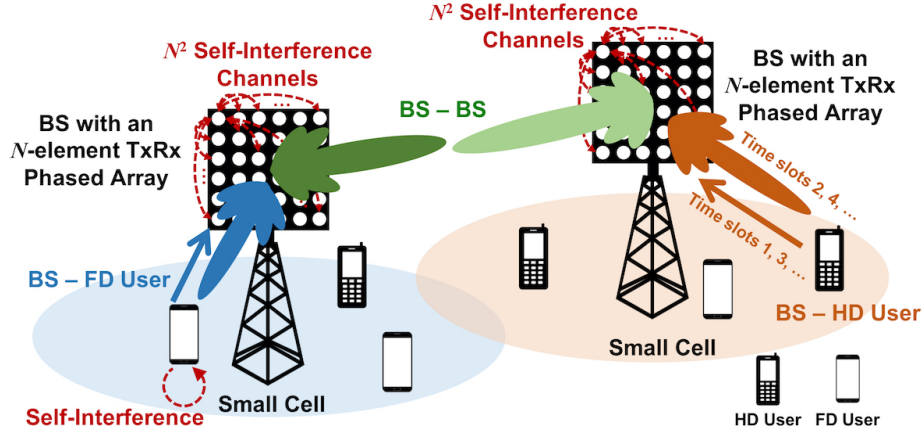
Figure 6.1: Example network scenarios where a base station (BS) is equipped with an N-element phased array employing Tx and Rx beamforming. Considered use cases: (i) *BS-User*: uplink-downlink (UL-DL) transmission between the BS and a user in HD (orange) or FD (blue) mode, and (ii) *BS-BS*: bidirectional transmission between two BSs in HD (beamforming in dark/light green in alternate time slots) or FD (simultaneous beamforming in dark/light green) mode.

We numerically evaluate the network performance and the corresponding FD rate gains when using the iterative algorithm, based on SI channel measurements and datasets (Section 6.5). Extensive evaluations are conducted under different network scenarios and a wide range of number of antennas, antenna array geometries, and Tx power levels. Specifically, the results show that an FD phased array with 9/36/72 elements can cancel the total SI power to below the noise floor with sum TxBF and RxBF gain losses of 10.6/7.2/6.9 dB, even at Tx power level of 30 dBm (note that the conventional HD TxBF and RxBF can provide sum TxBF and RxBF gains of 19.1/30.1/37.1 dB). Moreover, the *FD rate gains* in the BS-User case are at least 1.33/1.66/1.68× with $N = 9/36/72$. The FD rate gains in the BS-BS case are at least 1.53× with $N \geq 36$ in all considered SNR regimes. We also demonstrate the efficiency of the developed iterative algorithm.

This chapter's contributions were previously presented in [27] and are based on joint work with Ph.D. student Mahmood Baraani Dastjerdi. The developed optimization framework was also integrated with an 8-element FD phased array RFIC that we presented in [28], which is in collaboration with Ph.D. students Mahmood Baraani Dastjerdi and Negar Reiskarimian.

N	Number of antenna/Tx/Rx elements
B	Total wireless bandwidth/desired RF SIC bandwidth
K, k	Number of frequency channels and channel index
f_k	Center frequency of the kth frequency channel
$\mathsf{Re}[z], \mathsf{Im}[z]$	Real and imaginary parts of $z \in \mathbb{C}$
$\lvert z \rvert, \angle z$	Amplitude and phase of $z = \lvert z \rvert \cdot e^{\angle z} \in \mathbb{C}$
P_{t}	Total power of an N-element Tx array
$P_{\text{nf}}, N P_{\text{nf}}$	Noise floor of a single Rx element and an N-element Rx array
$\mathbf{w} = [w_n]$	Tx beamforming weights or Tx beamformer, $\mathbf{w} \in \mathbb{C}^N$
$\mathbf{v} = [v_n]$	Rx beamforming weights or Rx beamformer, $\mathbf{v} \in \mathbb{C}^N$
$H_{mn}(f_k)$	Self-interference channel between the nth Tx element and the mth Rx element, $H_{mn}(f_k) \in \mathbb{C}$
$\mathbf{H}(f_k)$	Self-interference channel matrix at f_k, $\mathbf{H}(f_k) \in \mathbb{C}^{N \times N}$
$\mathbf{s}_{\text{t}} = \mathbf{s}(\phi_{\text{t}}, \theta_{\text{t}})$	Steering vector in the desired Tx beam-pointing direction $(\phi_{\text{t}}, \theta_{\text{t}})$
$\mathbf{s}_{\text{r}} = \mathbf{s}(\phi_{\text{r}}, \theta_{\text{r}})$	Steering vector in the desired Rx beam-pointing direction $(\phi_{\text{r}}, \theta_{\text{r}})$
$A_{\text{t}}(\phi, \theta), a_{\text{t}}$	Tx array factor in (ϕ, θ), and in the Tx beam-pointing direction
$A_{\text{r}}(\phi, \theta), a_{\text{r}}$	Rx array factor in (ϕ, θ), and in the Rx beam-pointing direction
$g_{\text{t}}, g_{\text{r}}$	TxBF and RxBF gains in the Tx and Rx beam-pointing directions

Table 6.1: Nomenclature

6.1 Background on Phased Arrays and (Half-Duplex) Beamforming

In this section, we follow [155] and provide background on phased arrays and beamforming in the half-duplex setting. Beamforming is a technique that uses an antenna phased array to achieve directional signal transmission or reception. For completeness, we describe Tx and Rx phased arrays, and Tx and Rx beamforming (TxBF and RxBF). We provide an overview of the steering vector and Tx/Rx arrays, followed by the important characteristics of beamforming including array factor, beam pattern, and beamforming gain. The notation is summarized in Table 6.1.

Steering Vector. Denote by N the number of antennas in the array (see Fig. 6.2(a)). A spherical coordinate system is considered where the azimuth and elevation angles are denoted by ϕ and θ, respectively. Let $s_n(\phi, \theta)$ be the relevant phase delay experienced by a plane wave as it departs/reaches the nth Tx/Rx element in the spatial direction of (ϕ, θ). Consider an example N-element rectangular antenna array with N_x rows and N_y columns ($N_x \cdot N_y = N$), where the antennas are indexed as shown in Fig. 6.2(a). Assuming half-wavelength spacing between adjacent

(a)   **(b)**

LNA: Low Noise Amplifier, PA: Power Amplifier, ATT: Variable Attenuator, PS: Phase Shifter
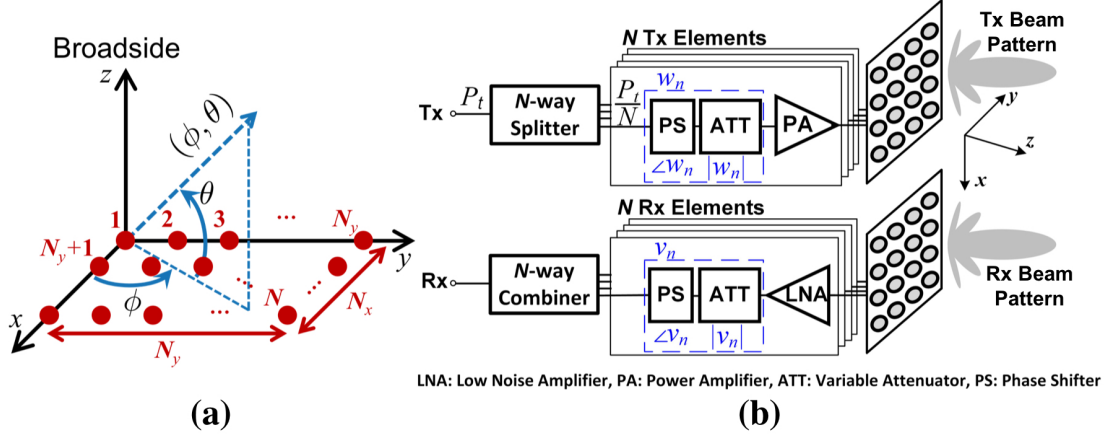
Figure 6.2: (a) An example N-element rectangular antenna array in a spherical coordinate system, (b) block diagrams of N-element Tx (top) and Rx (bottom) phased arrays in the HD setting.

antennas, $s_n(\phi, \theta)$ for the nth element at location (n_x, n_y), where $n = (n_x - 1)N_y + n_y$, is

$$s_n(\phi, \theta) = e^{j\pi[(n_x-1)\cos\theta\cos\phi + (n_y-1)\cos\theta\sin\phi]}, \ \forall n_x, n_y.$$

The *steering vector* in the direction of (ϕ, θ) is then given by $\mathbf{s}(\phi, \theta) = [s_n(\phi, \theta)] \in \mathbb{C}^N$, which depends on the antenna array geometry.

Transmit (Tx) and Receive (Rx) Arrays, and Analog Beamformer. As illustrated in Fig. 6.2(b), in an N-element Tx array with total Tx power of P_{t}, each Tx element consists of a power amplifier (PA), a variable attenuator (ATT), and a phase shifter (PS). Symmetrically, in an N-element Rx array, each Rx element consists of a low noise amplifier (LNA), a variable attenuator, and a phase shifter. Denote by P_{nf} the noise floor of a single Rx element. Then, the Rx array has an array noise floor of $N \cdot P_{\text{nf}}$ due to the aggregated noise from all Rx elements. We assume the followings: (i) with an ideal power splitter, each Tx element has a maximum Tx power of P_{t}/N, and (ii) the LNA of each Rx element has a unit gain of 1.

An *analog beamformer* is the set of complex-valued weights applied to each element *relative* to that of the first element. Specifically, denote the weight applied to the nth Tx (resp. Rx) element by $w_n = |w_n| \cdot e^{j\angle w_n}$ (resp. $v_n = |v_n| \cdot e^{j\angle v_n}$) with amplitude $|w_n|$ (resp. $|v_n|$) and phase $\angle w_n$ (resp. $\angle v_n$). In particular, $|w_n|^2, |v_n|^2 \leq 1$ represent the variable gain on the nth Tx/Rx element

controlled by the attenuator. $\angle w_n, \angle v_n \in [-\pi, \pi]$ represent the phase on the nth Tx/Rx element controlled by the phase shifter. The vectors $\mathbf{w} = [w_n] \in \mathbb{C}^N$ and $\mathbf{v} = [v_n] \in \mathbb{C}^N$ are called the *Tx and Rx (analog) beamformers*, respectively.

Array Factor, Beam Pattern, and Beamforming Gain. An analog beamformer features $(N-1)$ degrees of freedom (DoF), which are typically configured to alter the *beam pattern* (see Fig. 6.2(b)) to enhance the signal directivity and/or to suppress interference by constructing nulls. The far-field Tx and Rx *array factors* in the direction of (ϕ, θ), which quantify the effect of combining weighted transmitting Tx and receiving Rx elements, are given by [155]

$$
\begin{cases}
A_{\text{t}}(\phi, \theta) = \sum\limits_{n=1}^{N} (s_n(\phi, \theta) \cdot w_n) = \mathbf{s}^{\top}(\phi, \theta) \cdot \mathbf{w} \in \mathbb{C}, \\
A_{\text{r}}(\phi, \theta) = \sum\limits_{n=1}^{N} (s_n(\phi, \theta) \cdot v_n) = \mathbf{s}^{\top}(\phi, \theta) \cdot \mathbf{v} \in \mathbb{C}.
\end{cases}
$$

The corresponding far-field Tx and Rx *beam patterns* are defined as

$$
E_{\text{t}}(\phi, \theta) = |A_{\text{t}}(\phi, \theta)|^2 = |\mathbf{s}^{\top}(\phi, \theta) \cdot \mathbf{w}|^2, \ E_{\text{r}}(\phi, \theta) = |A_{\text{r}}(\phi, \theta)|^2 = |\mathbf{s}^{\top}(\phi, \theta) \cdot \mathbf{v}|^2. \tag{6.1}
$$

The *Tx beamforming (TxBF) gain* (resp. *Rx beamforming (RxBF) gain*) is defined as the power gain of the Tx (resp. Rx) signal in the far-field normalized to the maximum total Tx (resp. Rx) power. Denote by $G_{\text{t}}(\phi, \theta)$ and $G_{\text{r}}(\phi, \theta)$) the TxBF gain and RxBF gain in the spatial direction of (ϕ, θ), respectively. They are given by

$$
G_{\text{t}}(\phi, \theta) = E_{\text{t}}(\phi, \theta)/N = |\mathbf{s}^{\top}(\phi, \theta) \cdot \mathbf{w}|^2/N, \ G_{\text{r}}(\phi, \theta) = E_{\text{r}}(\phi, \theta)/N = |\mathbf{s}^{\top}(\phi, \theta) \cdot \mathbf{v}|^2/N. \tag{6.2}
$$

Denote the *desired* Tx and Rx beam-pointing directions by $(\phi_{\text{t}}, \theta_{\text{t}})$ and $(\phi_{\text{r}}, \theta_{\text{r}})$. Let $\mathbf{s}_{\text{t}}$ and $\mathbf{s}_{\text{r}}$ be the steering vectors in the desired Tx and Rx beam-pointing directions, given by

$$
\mathbf{s}_{\text{t}} = \mathbf{s}(\phi_{\text{t}}, \theta_{\text{t}}), \ \mathbf{s}_{\text{r}} = \mathbf{s}(\phi_{\text{r}}, \theta_{\text{r}}).
$$

Then, the (complex-valued) Tx and Rx array factors in the desired beam-pointing directions, de-

noted by a_t and a_r, are then given by

$$a_\mathrm{t} = A_\mathrm{t}(\phi_\mathrm{t}, \theta_\mathrm{t}) = \mathbf{s}_\mathrm{t}^\top \mathbf{w}, \ a_\mathrm{r} = A_\mathrm{r}(\phi_\mathrm{r}, \theta_\mathrm{r}) = \mathbf{s}_\mathrm{r}^\top \mathbf{v}. \tag{6.3}$$

Therefore, the TxBF and RxBF gains in $(\phi_\mathrm{t}, \theta_\mathrm{t})$ and $(\phi_\mathrm{r}, \theta_\mathrm{r})$, denoted by g_t and g_r, are given by

$$g_\mathrm{t} = |a_\mathrm{t}|^2/N, \ g_\mathrm{r} = |a_\mathrm{r}|^2/N. \tag{6.4}$$

Note that since g_t and g_r are the normalized power gains introduced by TxBF and RxBF, respectively, they *do not* depend on the absolute power levels of the Tx and Rx signals.

Conventional (Half-Duplex) TxBF and RxBF. It is known that by setting $\mathbf{w} = \mathbf{s}_\mathrm{t}^*$ (resp. $\mathbf{v} = \mathbf{s}_\mathrm{r}^*$), a maximum Tx (resp. Rx) array factor of N and a maximum TxBF (resp. RxBF) gain of N in the desired beam-pointing direction can be achieved [155], i.e.,

$$a_\mathrm{t}^\mathrm{max} = \mathbf{s}_\mathrm{t}^\top \mathbf{s}_\mathrm{t}^* = N, \ a_\mathrm{r}^\mathrm{max} = \mathbf{s}_\mathrm{r}^\top \mathbf{s}_\mathrm{r}^* = N, \ g_\mathrm{t}^\mathrm{max} = g_\mathrm{r}^\mathrm{max} = N. \tag{6.5}$$

We refer to $\mathbf{w}_\mathrm{conv} = \mathbf{s}_\mathrm{t}^*$ (resp. $\mathbf{v}_\mathrm{conv} = \mathbf{s}_\mathrm{r}^*$) as the *conventional HD Tx (resp. Rx) beamformers*.

6.2 Model, Rates, and Objective

In this section, we present the model of an FD phased array node combining TxBF and RxBF, as well as the corresponding FD TxBF and RxBF gains. Then, we discuss the data rate gains obtained by FD and the FD link objective.

6.2.1 FD Phased Array Model

We consider an FD phased array node as depicted in Fig. 6.3(a). A BS is equipped with an N-element *FD phased array*, where each antenna is *shared* between a pair of Tx and Rx elements via a circulator, and where TxBF and RxBF are applied. As shown in Fig. 6.1, a user has *a single antenna* and does not apply beamforming. We use b and u in the subscript to denote the BS and user. Consider a wireless bandwidth of B that is divided into K orthogonal frequency channels indexed by $k \in \{1, \ldots, K\}$, where f_k is the center frequency of the kth channel. Without loss of
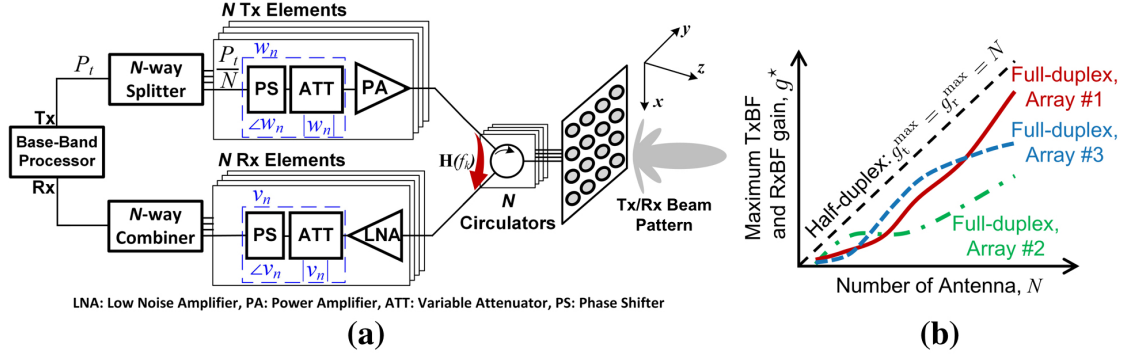
Figure 6.3: (a) Block diagram of an N-element FD phased array with SI channel matrix, $\mathbf{H}(f_k)$, where a Tx and an Rx phased arrays are connected to the antennas through circulators, (b) illustration of the relationship between the maximum TxBF and RxBF gains, $g_t^\star = g_r^\star = g^\star$, that can be achieved by an FD phased array after repurposing TxBF and RxBF for wideband RF SIC, and the number of antennas, N, for various array geometries.

generality, we assume an OFDM-based system with bandwidth B and K subcarriers.

SI Channel Matrix. Similar to [90], we let $H_{nn}(f_k), \forall n$, denote the frequency response of the SI channel from the nth Tx element to the nth Rx element at frequency $f_k, \forall k$, and let $H_{mn}(f_k), \forall m \neq n$, denote the frequency response of the cross-talk SI (CTSI) channel from the nth Tx element to the mth Rx element at frequency $f_k, \forall k$. We then denote $\mathbf{H}(f_k) = [H_{mn}(f_k)] \in \mathbb{C}^{N \times N}$ as the *SI channel matrix* of the TxRx array at frequency $f_k, \forall k$, consisting of all the N^2 SI and CTSI channels, as shown in Fig 6.1. We assume $|H_{mn}(f_k)| \leq 1, \forall i, j, k$, due to the propagation loss of the Tx signal. Ideally, due to channel reciprocity, $H_{mn}(f_k) = H_{nm}(f_k), \forall k$, for any Tx and Rx element pair (n, m), resulting in $\mathbf{H}(f_k) = \mathbf{H}^\top(f_k)$. However, as we will show in Section 6.5.1, this does not hold in a realistic environment, due to effects such as imperfections of antenna matching and environmental interference and noise.[1]

Use Cases. We consider four use cases as depicted in Fig. 6.1:

(i) *BS-User (HD or FD)*: uplink-downlink (UL-DL) transmission between a BS and a user in HD (orange) or FD (blue) mode,

(ii) *BS-BS (HD or FD)*: bidirectional transmission between two BSs in HD (beamforming in

---

[1]Note that our model and results also apply to an FD phased array where the Tx and Rx arrays are equipped with *seperate* antennas, which requires twice many antenna elements (as illustrated in Fig. 6.3(a)). Such separated Tx and Rx antennas usually provide better isolation (i.e., smaller values of $|H_{mn}(f_k)|$) than a shared antenna interface.

dark/light green in alternate time slots) or FD (simultaneous beamforming in dark/light green) mode.

The cases where the BSs or the user operate in HD mode and the channel is shared in a TDD manner are considered as the *benchmark HD cases*. When operating in FD mode, a BS applies simultaneous TxBF and RxBF when communicating with an FD user or another FD BS. We also assume the BS has the information about the direction of an intended user or another BS.

Self-Interference-to-Noise Ratio (XINR) under TxBF and RxBF. For an FD node, XINR is defined as the ratio between the residual SI power after analog and digital SIC and the noise floor. Denote by $\gamma_{bb}(f_k)$ and $\gamma_{uu}(f_k)$ the XINR of the BS and the user at frequency f_k, respectively, when operating in FD mode. Since we focus is on the SIC at the BS with TxBF and RxBF, we assume that a user can always cancel its SI to below the noise floor, i.e., $\gamma_{uu}(f_k) \leq 1, \forall k$ [22, 31]. Accordingly, a user transmits at the maximum possible power level when operating in either HD or FD mode.

For the BS, denote by $P_{\text{SI}}^{\text{bf}}(f_k)$ the SI power under TxBF and RxBF, and by SIC_{dig} the amount of achievable digital SIC.[2] Recall that the BS has a noise floor of (NP_{nf}), the XINR is given by

$$\gamma_{bb}(f_k) = \frac{P_{\text{SI}}^{\text{bf}}(f_k)}{SIC_{\text{dig}} \cdot (NP_{\text{nf}})}, \ \forall k. \tag{6.6}$$

We now derive $\gamma_{bb}(f_k)$ under joint TxBF and RxBF. Without loss of generality, we assume a wideband (e.g., OFDM) transmit symbol $x(f_k) \in \mathbb{C}$ with unit power of $|x(f_k)|^2 = 1, \forall k$. The received SI symbol, denoted by $y(f_k) \in \mathbb{C}$, is given by

$$y(f_k) = \sum_{m=1}^{N}\sum_{n=1}^{N}\left[w_n H_{mn}(f_k)v_m \right] \cdot \sqrt{\frac{P_{\text{t}}}{N}} \cdot x(f_k) + z = \sqrt{\frac{P_{\text{t}}}{N}} \cdot \mathbf{v}^{\top}\mathbf{H}(f_k)\mathbf{w} \cdot x(f_k) + z, \ \forall k,$$

where z is the additive white Gaussian noise. Then, the SI power is given by

$$P_{\text{SI}}^{\text{bf}}(f_k) = |y(f_k)|^2 = |\mathbf{v}^{\top}\mathbf{H}(f_k)\mathbf{w}|^2 \cdot \frac{P_{\text{t}}}{N}, \ \forall k. \tag{6.7}$$

---

[2]We assume that SIC_{dig} is constant across frequency, since delay taps can be easily generated in the digital domain [81].

Plugging (6.7) into (6.6) yields the XINR of the BS:

$$\gamma_{bb}(f_k) = \frac{|y(f_k)|^2}{SIC_{\mathrm{dig}} \cdot (NP_{\mathrm{nf}})} = \frac{|\mathbf{v}^\top \mathbf{H}(f_k)\mathbf{w}|^2}{SIC_{\mathrm{dig}}NP_{\mathrm{nf}}} \cdot \frac{P_{\mathrm{t}}}{N}, \ \forall k. \tag{6.8}$$

6.2.2   FD Beamformers and Beamforming Gains

A maximum TxBF and RxBF gain of N can be achieved under the conventional HD Tx and Rx beamformers, $\mathbf{w}_{\mathrm{conv}}$ and $\mathbf{v}_{\mathrm{conv}}$ (see (6.5)). However, as we will show in Section 6.5.3, the XINR under $\mathbf{w}_{\mathrm{conv}}$ and $\mathbf{v}_{\mathrm{conv}}$ is significant, i.e., $\gamma_{bb}(f_k) \gg 1, \forall k$. Therefore, using these beamforming weights is impractical for a BS operating in FD mode.

To support FD operation, we aim to achieve wideband RF SIC in an FD phased array *only* through manipulating the TxBF and RxBF weights, $\mathbf{w}$ and $\mathbf{v}$, thereby not requiring specific RF canceller hardware or circuitry. Specifically, we aim to achieve both $\gamma_{bb}(f_k) \leq 1, \forall k$, and the maximum FD data rate gain by properly selecting $\mathbf{w}$ and $\mathbf{v}$. However, these weights, termed as the *FD Tx and Rx beamformers*, may not achieve the maximum TxBF and RxBF gain of N as in the conventional HD setting. Accordingly, we define the optimal FD TxBF and RxBF gain as follows.

Definition 6.1 (Optimal FD TxBF and RxBF Gain) *For a given FD phased array with $\mathbf{H}(f_k)$ and P_t, the optimal FD TxBF and RxBF gain is the equal maximum TxBF and RxBF gains that can be achieved while satisfying $\gamma_{bb}(f_k) \leq 1, \forall k$.*

Denote by $g_{\mathrm{t}}^\star = g_{\mathrm{r}}^\star = g^\star$ the optimal FD TxBF and RxBF gain, and by $a_{\mathrm{t}}^\star$ and $a_{\mathrm{r}}^\star$ the corresponding optimal *FD Tx and Rx array factors* (the reason for setting $g_{\mathrm{t}}^\star = g_{\mathrm{r}}^\star$ is described in Section 6.2.4). From (6.4), $g_{\mathrm{t}}^\star = |a_{\mathrm{t}}^\star|^2/N$ and $g_{\mathrm{r}}^\star = |a_{\mathrm{r}}^\star|^2/N$. To quantify the performance of FD TxBF and RxBF, we present the following definition.

Definition 6.2 (TxBF and RxBF Gain Losses) *The TxBF gain loss is the ratio between the maximum HD TxBF gain and the optimal FD TxBF gain, i.e., $(N/g_t^\star)$. Symmetrically, the RxBF gain loss is $(N/g_r^\star)$.*

The TxBF and RxBF gain losses are typically represented in dB. For example, a $3\,\mathrm{dB}$ TxBF gain loss means that the far-field Tx power is reduced by half, since $g_\mathrm{t}^\star = 0.5N \Rightarrow 10\log_{10}(N/g_\mathrm{t}^\star) = 3\,\mathrm{dB}$. Similarly, a $6\,\mathrm{dB}$ TxBF gain loss corresponds to $g_\mathrm{t}^\star = 0.25N$.

Fig. 6.3(b) pictorially illustrates the relationship between $g^\star$ and the number of antennas, N, for various array geometries, where different antenna arrays may have different $g^\star$. In particular, for a given FD phased array with $\mathbf{H}(f_k)$ and Tx power level P_t, there exists a pair of optimal FD Tx and Rx beamformers that achieves $g^\star$ while satisfying $\gamma_{bb}(f_k) \leq 1, \forall k$. The FD phased array will experience higher TxBF and RxBF gains with increased value of N, since a larger number of Tx and Rx weights can be manipulated.

6.2.3   Sum Link Rate and FD Rate Gain

We now derive the HD and FD data rates and the effect of the optimal FD TxBF and RxBF gain on the FD rates. Denote by γ the link SNR without beamforming. We use Shannon's capacity formula to compute the rate on a link with bandwidth, B, and link SNR, γ. Since the TxBF and RxBF gains, g_t and g_r, are independent of the absolute power of the Tx and Rx signals (see Section 6.1), the link SNR improvement introduced by beamforming equals to g_t and g_r in the desired Tx and Rx beam-pointing directions, respectively.

For the BS-User case, we denote by γ_{bu} ($u \rightarrow b$) and γ_{ub} ($b \rightarrow u$) the UL and DL SNRs, respectively. For the BS-BS case, we index the BSs by b_1 and b_2 and denote the link SNRs by $\gamma_{b_1 b_2}$ ($b_2 \rightarrow b_1$) and $\gamma_{b_2 b_1}$ ($b_1 \rightarrow b_2$). The sum of the HD link rates in both cases, when the BSs and user operate in HD mode and share the channel in a TDD manner equally (i.e., each link is activated for 50% of the time), are given by

$$r_{\text{BS-User}}^{\text{HD}} = \frac{B}{2} \log_2 \left(1 + N\gamma_{bu}\right) + \frac{B}{2} \log_2 \left(1 + N\gamma_{ub}\right), \tag{6.9}$$

$$r_{\text{BS-BS}}^{\text{HD}} = \frac{B}{2} \log_2 \left(1 + N^2\gamma_{b_1 b_2}\right) + \frac{B}{2} \log_2 \left(1 + N^2\gamma_{b_2 b_1}\right). \tag{6.10}$$

In particular, the UL and DL SNR improvements in the BS-User case (6.9) result from the maximum RxBF and TxBF gains of N in the desired beam-pointing directions. Similarly, the SNR

improvements in the BS-BS case (6.10), which are factors of $g_\text{t}^\text{max} g_\text{r}^\text{max} = N^2$ for both $\gamma_{b_1 b_2}$ and $\gamma_{b_2 b_1}$, stem from the combined TxBF and RxBF gains of both BSs.

When the BSs and user operate in FD mode, the link SNRs are affected by both the degraded TxBF and RxBF gains, $g_\text{t}^\star$ and $g_\text{r}^\star$, and the frequency-dependent XINR of the BS and user, $\gamma_{bb}(f_k)$ and $\gamma_{uu}(f_k)$, respectively. As a result, the sum of the FD link rates in both use cases are

$$r_\text{BS-User}^\text{FD} = \frac{B}{K} \cdot \sum_{k=1}^{K} \log_2 \left(1 + \frac{g_\text{r}^\star \gamma_{bu}}{1 + \gamma_{bb}(f_k)} \right) + \log_2 \left(1 + \frac{g_\text{t}^\star \gamma_{ub}}{1 + \gamma_{uu}(f_k)} \right), \tag{6.11}$$

$$r_\text{BS-BS}^\text{FD} = \frac{B}{K} \cdot \sum_{k=1}^{K} \log_2 \left(1 + \frac{g_\text{t}^\star g_\text{r}^\star \gamma_{b_1 b_2}}{1 + \gamma_{bb}(f_k)} \right) + \log_2 \left(1 + \frac{g_\text{t}^\star g_\text{r}^\star \gamma_{b_2 b_1}}{1 + \gamma_{bb}(f_k)} \right). \tag{6.12}$$

Due to the coupling between $g_\text{t}^\star$, $g_\text{r}^\star$, and $\gamma_{bb}(f_k)$ through the Tx and Rx beamformers, $\mathbf{w}$ and $\mathbf{v}$ (see (6.8)), and the frequency-dependent $\mathbf{H}(f_k)$, maximizing (6.11) and (6.12) presents numerous challenges. To allow analytical tractability, we approximate the FD sum rates in (6.11) and (6.12) by setting $\gamma_{uu}(f_k) = \gamma_{bb}(f_k) = 1, \forall k$. We refer to the approximated FD sum rates as $\widetilde{r}_\text{BS-User}^\text{FD}$ and $\widetilde{r}_\text{BS-BS}^\text{FD}$, respectively, and they are given by

$$\widetilde{r}_\text{BS-User}^\text{FD} = B \cdot \left[\log_2 \left(1 + \frac{g_\text{r}^\star \gamma_{bu}}{2} \right) + \log_2 \left(1 + \frac{g_\text{t}^\star \gamma_{ub}}{2} \right) \right], \tag{6.13}$$

$$\widetilde{r}_\text{BS-BS}^\text{FD} = B \cdot \left[\log_2 \left(1 + \frac{g_\text{t}^\star g_\text{r}^\star \gamma_{b_1 b_2}}{2} \right) + \log_2 \left(1 + \frac{g_\text{t}^\star g_\text{r}^\star \gamma_{b_2 b_1}}{2} \right) \right]. \tag{6.14}$$

Since we aim to achieve $\gamma_{bb}(f_k) \leq 1$ (0 dB) for an FD phased array and recall that $\gamma_{uu}(f_k) \leq 1$, $\widetilde{r}_\text{BS-User}^\text{FD}$ and $\widetilde{r}_\text{BS-BS}^\text{FD}$ are *lower bounds* of the FD sum rates, $r_\text{BS-User}^\text{FD}$ and $r_\text{BS-BS}^\text{FD}$, in (6.11)–(6.12).

In the rest of the chapter, we focus on maximizing $\widetilde{r}_\text{BS-User}^\text{FD}$ and $\widetilde{r}_\text{BS-BS}^\text{FD}$. We define the *FD rate gain* in the BS-User (resp. BS-BS) case as the ratio between the FD sum rate lower bound and the HD sum rate, i.e., $\widetilde{r}_\text{BS-User}^\text{FD}/r_\text{BS-User}^\text{HD}$ (resp. $\widetilde{r}_\text{BS-BS}^\text{FD}/r_\text{BS-BS}^\text{HD}$).

6.2.4 FD Link Objective

Our objective is to maximize the FD rate gains in the two FD use cases. Based on the observation above, our goal is to obtain the optimal FD Tx and Rx beamformers, $\mathbf{w}$ and $\mathbf{v}$, that maximize

TxBF and RxBF gains, g_t and g_r, while achieving sufficient amount of RF SIC, i.e.,

$$\gamma_{bb}(f_k) \leq 1, \ \forall k \Leftrightarrow |\mathbf{v}^\top \mathbf{H}(f_k)\mathbf{w}|^2 \cdot \frac{P_\mathrm{t}}{N} \leq SIC_\mathrm{dig} \cdot (NP_\mathrm{nf}) := N\beta, \ \forall k, \tag{6.15}$$

where $\beta := SIC_\mathrm{dig}P_\mathrm{nf}$ is a constant independent of N and P_t. Moreover, in order to (i) preserve the phase of the signals (i.e., the Tx and Rx array factors, a_t and a_r, should only contain real parts), and (ii) to balance the TxBF and RxBF gains in the far-field (i.e., $g_\mathrm{t} = g_\mathrm{r}$ so that the Tx and Rx signals do not experience different power level changes), the following constraints are set based on (6.3)–(6.4):

$$\begin{cases} \text{(i) } a_\mathrm{t}, a_\mathrm{r} \in \mathbb{R} \Leftrightarrow \mathsf{Im}[a_\mathrm{t}] = \mathsf{Im}[a_\mathrm{r}] = 0 \text{ (no phase offset)}, \\[2mm] \text{(ii) } g_\mathrm{t} = g_\mathrm{r} \Leftrightarrow a_\mathrm{t} = a_\mathrm{r} \text{ (balanced TxBF and RxBF gains)}. \end{cases} \tag{6.16}$$

Next, we first present the benefits introduced by FD phased arrays in terms of FD rate gains (Section 6.3). Then, we present the corresponding problem formulation (Section 6.4).

6.3 Motivating Examples

In this section, we illustrate the FD rate gains obtained by FD phased arrays with joint TxBF and RxBF, where the FD rate gains are computed as described in Section 6.2.3. We provide several examples illustrating that higher values of the optimal FD TxBF and RxBF gain lead to increased FD rate gains. According to Sections 6.2.2 and 6.2.4, we let $g_\mathrm{t}^\star = g_\mathrm{r}^\star = g^\star$ with the corrsponding $a_\mathrm{t}^\star = a_\mathrm{r}^\star = a^\star \in \mathbb{R}$ satisfying $g^\star = (a^\star)^2/N$ (see (6.16)). We will describe how to obtain $a^\star$ (and thus $g^\star$) in Section 6.4.

BS-User Case: We first consider equal UL and DL SNR values of $\gamma_{bu} = \gamma_{ub} = \gamma$ in low, medium, and high SNR regimes with $\gamma \in \{0, 15, 30\}\,\mathrm{dB}$. These values correspond to typical Rx signal levels in an LTE network, where the user is at the edge, middle, and center of the small cell. Fig. 6.4 presents the FD rate gain with varying number of antennas, $N \in \{4, 8, \ldots, 128\}$, and the *normalized* optimal FD Tx and Rx array factor $a^\star/N \in [0, 1]$.

Fig. 6.4 shows, for example, that a 16-element FD phased array can achieve FD rate gains of
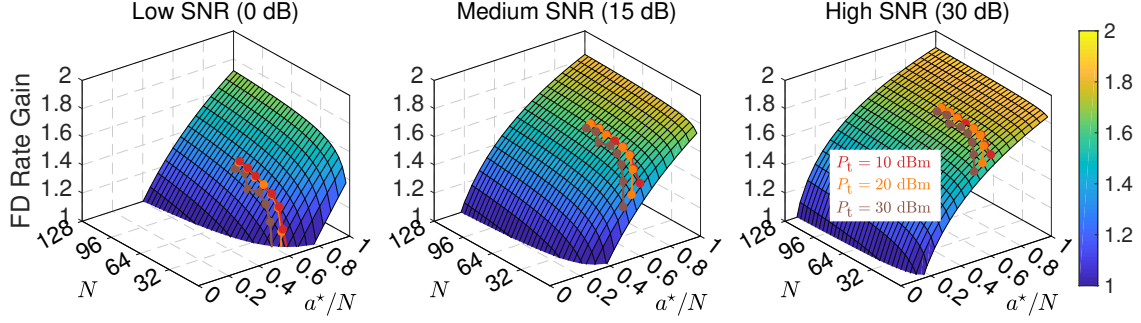
Figure 6.4: FD rate gain in the BS-User case with varying number of antennas, N, and the normalized optimal FD Tx and Rx array factors, $a^\star/N$, for different link SNR values $\gamma \in \{0, 15, 30\}$ dB. The relationships between $a^\star/N$ and N, based on the Argos traces [156], are also overlayed on the surface with $P_t \in \{10, 20, 30\}$ dBm.
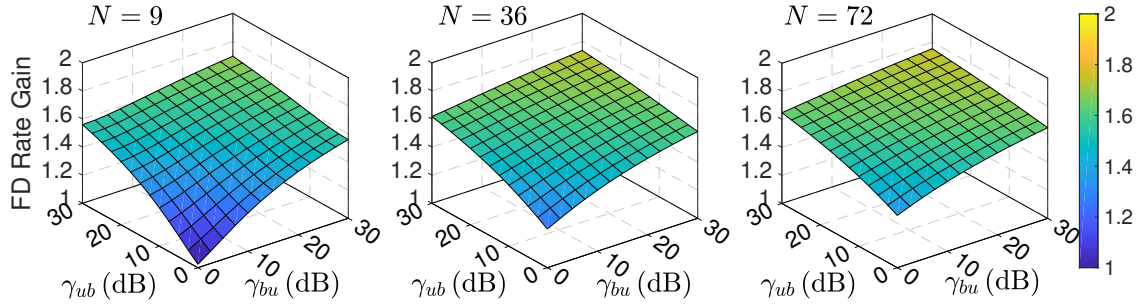
Figure 6.5: FD rate gain in the BS-User case with varying UL and DL SNR values, γ_{bu} and γ_{ub}, respectively, with $N \in \{9, 36, 72\}$ and 3 dB TxBF and RxBF gain loss (i.e., $g^\star = N/2$).

$1.13/1.55/1.71\times$ in low/medium/high SNR regimes, with 3 dB TxBF and RxBF gain loss (i.e., $g^\star = 0.5N$ and $a^\star = \frac{\sqrt{2}}{2}N$). These rate gains increase to $1.35/1.63/1.75\times$ when $N = 64$. Moreover, with 6 dB TxBF and RxBF gain losses, a 64-element FD phased array can achieve FD rate gains of $1.05/1.46/1.62\times$ in low/medium/high SNR regimes. As we will show in Section 6.5, a TxBF and RxBF gain loss of 6 dB is sufficient to achieve $\gamma_{bb}(f_k) \leq 1, \forall k$, in most considered scenarios. It is also interesting to note that under sufficient link SNR values, an FD phased array with more antennas provides marginal improvements on the FD rate gain. For example, in the medium SNR regime, $N = 64/128$ elements can achieve FD rate gains of $1.63/1.66\times$, respectively (namely, doubling the number of antennas provides an improvement of only 3%).

To provide a practical example of the FD rate gain, we overlay in Fig. 6.4 curves that represent relationships between $a^\star/N$ and N using $\mathbf{H}(f_k)$ from the Argos dataset [156] with different Tx power levels (the details will be described in Section 6.5.4 and shown in Fig. 6.12). For example,
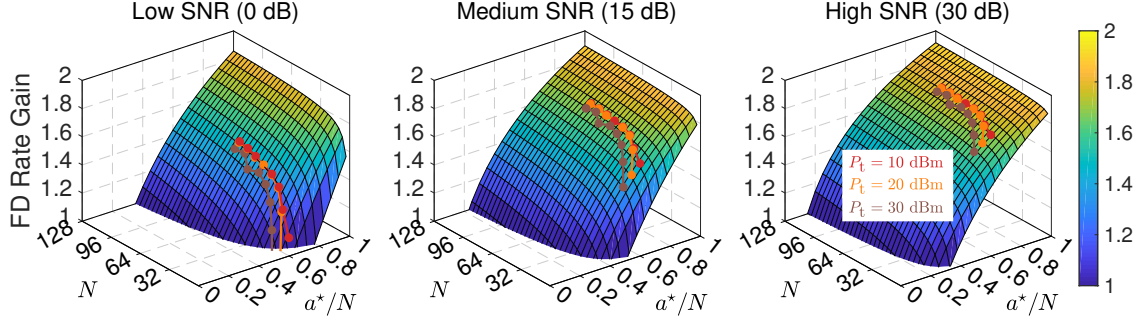
Figure 6.6: FD rate gain in the BS-BS case with varying number of antennas, N, and the normalized optimal FD Tx and Rx array factors, $a^\star/N$ for different link SNR values $\gamma \in \{0, 15, 30\}$ dB. The relationships between $a^\star/N$ and N, based on the Argos traces [156], are also overlayed on the surface with $P_t \in \{10, 20, 30\}$ dBm.

the Argos array with $P_t = 20$ dBm can achieve $a^\star/N = 0.66/0.69/0.70$ for $N = 18/36/72$. In addition, the Argos array with $P_t = 20$ dBm and $N = 36$ experiences only 3.1 dB TxBF and RxBF gain losses each, and can achieve maximum FD rate gains of $1.27/1.60/1.72\times$ in low/medium/high SNR regimes. Importantly, the curves show that in a *given FD phased array* with given values of N and P_t, there exists an upper limit of $a^\star$ on a_t and a_r that *any* FD Tx and Rx beamformers cannot exceed. Since the FD rate gain increases as a function of $g^\star = (a^\star)^2/N$ (see (6.13)–(6.14)), designing FD Tx and Rx beamformers that reach the upper limit of $a^\star$ will result in the maximum achievable FD rate gain.

We also evaluate the FD rate gain with asymmetric UL and DL SNR values, a scenario which is more common in realistic network settings. Fig. 6.5 presents the FD rate gain with varying UL and DL SNR values, γ_{bu} and γ_{ub}, with $N \in \{9, 36, 72\}$ and with 3 dB TxBF and RxBF gain loss. The results show that FD rate gains of 1.28–$1.74\times/1.37$–$1.75\times$ can be achieved with $N = 36/72$ under all considered UL and DL SNR values. Note that the FD rate gain also increases as a function of the number of antennas, N.

BS-BS Case: Fig. 6.6 plots the FD rate gain in the BS-BS case with the same setting as used in the BS-User case. The results show that a 16-element FD phased array can achieve FD rate gains of $1.25/1.53/1.66\times$ in low/medium/high SNR regimes, with only 3 dB TxBF and RxBF gain loss. These gains increase to $1.49/1.64/1.72\times$ with $N = 64$. The curves representing the relationships between $a^\star/N$ and N using the Argos traces are overlayed in Fig. 6.6, and similar observations as
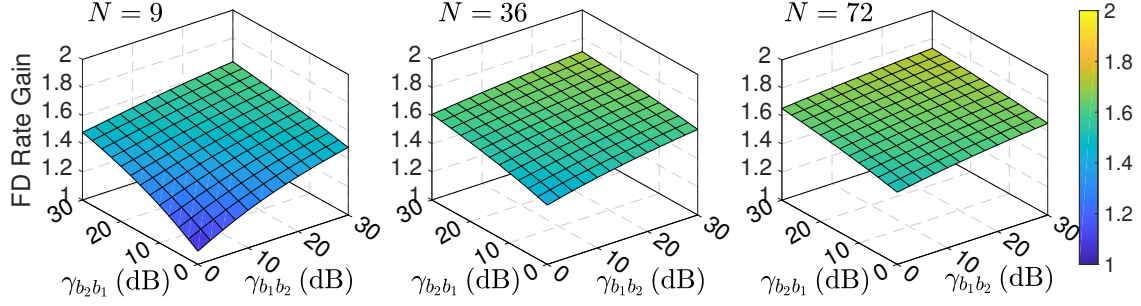
Figure 6.7: FD rate gain in the BS-BS case with varying link SNR values, $\gamma_{b_2 b_1}$ and $\gamma_{b_1 b_2}$, respectively, with $N \in \{9, 36, 72\}$ and 3 dB TxBF and RxBF gain loss (i.e., $g^\star = N/2$).

in the BS-User case are also relevant in the BS-BS case. In both use cases, the FD rate gain does not approach $2\times$, due to the XINR at the BS and user, $\gamma_{bb} = \gamma_{uu} = 1$.

Fig. 6.7 shows the FD rate gain with varying link SNR values between the two BSs, $\gamma_{b_1 b_2}$ and $\gamma_{b_2 b_1}$, with $N \in \{9, 36, 72\}$ and with 3 dB TxBF and RxBF gain loss at both BSs. In particular, FD rate gains of 1.42–$1.71\times$/1.51–$1.73\times$ can be achieved with $N = 36/72$ under all considered link SNR values. Moreover, the FD rate gain is more significant than that in the BS-User case with lower link SNR values, where the link SNR can be largely increased by simultaneous TxBF and RxBF applied at both BSs.

Findings. As indicated above, although an FD phased array experiences TxBF and RxBF gain losses in the desired beam-pointing directions to achieve $\gamma_{bb}(f_k) \leq 1, \forall k$, the network can still achieve significant FD rate gains in various settings. Therefore, it is important to obtain the optimal Tx and Rx array factors, $a_{\mathrm{t}}^\star$ and $a_{\mathrm{r}}^\star$, in an FD phased array in order to achieve high FD rate gain in the considered use cases. The results also reveal an interesting phenomenon: with increased number of antennas, N, minimal TxBF and RxBF gain losses (e.g., 3 dB) do not affect the achieved FD rate gains. In addition, higher TxBF and RxBF gain losses are needed to achieve sufficient SIC under different Tx power level, P_{t}, and bandwidth, B, requirements, as we will show in Section 6.5. This tradeoff opens up a possibility of designing FD phased arrays with different geometries and/or values of N under varying system requirements (e.g., Tx power levels, link SNRs, bandwidth).

6.4 Formulation and Optimization

In this section, we formulate an optimization problem based on the FD link objective described in Section 6.2.4. Due to the non-convexity and computational complexity of the problem, we then present an alternative formulation whose solution can be efficiently obtained using an iterative algorithm with provable performance guarantees.

6.4.1 Problem Formulation

Based on the FD link objective defined in Section 6.2.4, the following problem jointly determines the FD Tx and Rx beamformers.

$$(\textsf{Opt-TxRx}) \ a^\star := \max_{\mathbf{w}, \mathbf{v}} \ a \tag{6.17}$$

$$\text{s.t.} : \textsf{Re}[\mathbf{s}_\mathrm{t}^\top \mathbf{w}] = a, \ \textsf{Im}[\mathbf{s}_\mathrm{t}^\top \mathbf{w}] = 0, \tag{6.18}$$

$$\textsf{Re}[\mathbf{s}_\mathrm{r}^\top \mathbf{v}] = a, \ \textsf{Im}[\mathbf{s}_\mathrm{r}^\top \mathbf{v}] = 0, \tag{6.19}$$

$$P_{\mathrm{SI}}^{\mathrm{bf}}(f_k) = |\mathbf{v}^\top \mathbf{H}(f_k) \mathbf{w}|^2 \cdot \tfrac{P_\mathrm{t}}{N} \le N\beta, \ \forall k, \tag{6.20}$$

$$|w_n|^2 \le 1, \ |v_n|^2 \le 1, \ \forall n. \tag{6.21}$$

Specifically, the objective (6.17) is to maximize the Tx and Rx array factors, subject to the following constraints (see Section 6.2.4): (i) the Tx and Rx array factors in the desired beam-pointing directions, $(\phi_\mathrm{t}, \theta_\mathrm{t})$ and $(\phi_\mathrm{r}, \theta_\mathrm{r})$, have equal *real* responses, (6.18)–(6.19), (ii) the residual SI power at any frequency is suppressed to below the array noise floor, i.e., $\gamma_{bb}(f_k) \le 1, \forall k$, (6.20), and (iii) the amplitude of the beamforming weight on each Tx/Rx element is at most one (see Section 6.1), (6.21). *Essentially, the Tx and Rx beamformers that are obtained as a solution are different from the conventional HD beamfomers such that the total SI power is canceled to below the array noise floor with minimal TxBF and RxBF gain loss.* In order words, TxBF and RxBF are *repurposed* for achieving wideband RF SIC. We note that $(\textsf{Opt-TxRx})$ can also be extended to include general complex-valued desired array patterns. Throughout the chapter, we practically set $P_\mathrm{nf} = -90\,\mathrm{dBm}$

and $SIC_{\text{dig}} = 40\,\text{dB}$.[3] According to (6.15), $\beta = P_{\text{nf}} \cdot SIC_{\text{dig}} = -90\,\text{dBm} + 40\,\text{dB} = -50\,\text{dBm} = 10^{-5}\,\text{mW}$.

Note that (6.18)–(6.19) are linear constraints, and (6.21) is a convex (non-linear) constraint. Moreover, (Opt-TxRx) always has a feasible solution where $\mathbf{w}$ and $\mathbf{v}$ have very small amplitudes. However, in general, (Opt-TxRx) is a *non-convex* optimization problem whose solution poses several challenges. The non-convexity stems from the coupling between $\mathbf{w}$ and $\mathbf{v}$ through $\mathbf{H}(f_k)$ in (6.20) where $\mathbf{H}(f_k)$ is not a Hermitian matrix, and not even a symmetric matrix in realistic environments (see Section 6.2.1). Moreover, the problem becomes computationally expensive to solve using existing solvers (which may only return a local optimum) with increased number of antennas, N, especially for massive-antenna systems and large-scale phased arrays. For benchmarking purposes, in Section 6.5, we use the nonlinear programming solver from MATLAB to solve (Opt-TxRx) and denote the returned (possibly only locally) optimal FD Tx and Rx array factors by $a_{\text{t}}^\star$ and $a_{\text{r}}^\star$, respectively.[4]

6.4.2 Observation and Intuition

Due to the intractability of (Opt-TxRx), we now describe alternative optimization problems which are based on the observation below. Note that the SI power in (6.20) can be written as,

$$
\begin{aligned}
P_{\text{SI}}^{\text{bf}}(f_k) &= |\mathbf{v}^\top \mathbf{H}(f_k)\mathbf{w}|^2 \cdot \frac{P_{\text{t}}}{N} = \left(\mathbf{v}^\top \mathbf{H}(f_k)\mathbf{w}\right)^\dagger \cdot \left(\mathbf{v}^\top \mathbf{H}(f)\mathbf{w}\right) \cdot \frac{P_{\text{t}}}{N} \\
&= \mathbf{w}^\dagger \Big(\underbrace{\mathbf{H}^\dagger(f_k)\mathbf{v}^*\mathbf{v}^\top \mathbf{H}(f_k)}_{:=\mathbf{H}_{\mathbf{v}}(f_k)} \Big)\mathbf{w} \cdot \frac{P_{\text{t}}}{N} = \mathbf{w}^\top \mathbf{H}_{\mathbf{v}}(f_k)\mathbf{w} \cdot \frac{P_{\text{t}}}{N}, \ \forall k.
\end{aligned}
$$

It can be seen that with a *fixed* Rx beamformer, $\mathbf{v}$,

- $\mathbf{H}_{\mathbf{v}}(f_k)$ is a *Hermitian* matrix, i.e., $\mathbf{H}_{\mathbf{v}}(f_k) = \mathbf{H}_{\mathbf{v}}^\dagger(f_k), \forall k$;

- $\mathbf{H}_{\mathbf{v}}(f_k)$ is *positive semidefinite* since, for any non-zero Tx beamformer, $\mathbf{w}$, the SI power

---

[3]Recent work has achieved $SIC_{\text{dig}} = 43/50\,\text{dB}$ [22, 81] which leads to more relaxed requirements on the amount of RF SIC.

[4]Due to the non-convexity of (Opt-TxRx), we use the same $a_{\text{t}}^\star$ and $a_{\text{r}}^\star$ as in Section 6.2.2 to denote the numerically obtained solution to (Opt-TxRx) using existing solvers.

cannot be negative, i.e.,

$$\mathbf{w}^\dagger \mathbf{H_v}(f_k)\mathbf{w} \cdot \tfrac{P_\mathrm{t}}{N} \geq 0, \ \forall k, \ \forall \mathbf{w} \in \mathbb{C}^N \text{ and } \mathbf{w} \neq \mathbf{0}.$$

Therefore, the optimal Tx beamformer that maximizes the Tx array factor, a_t, given a fixed $\mathbf{v}$, can be obtained by solving:

$$(\mathsf{P1}) \ \max_{\mathbf{w}} \ a_\mathrm{t}, \quad \text{subject to: } \mathsf{Re}[\mathbf{s}_\mathrm{t}^\top \mathbf{w}] = a_\mathrm{t}, \ \mathsf{Im}[\mathbf{s}_\mathrm{t}^\top \mathbf{w}] = 0,$$

$$\mathbf{w}^\dagger \mathbf{H_v}(f_k)\mathbf{w} \cdot \tfrac{P_\mathrm{t}}{N} \leq N\beta, \ \forall k,$$

$$|w_n|^2 \leq 1, \ \forall n.$$

Unlike (Opt-TxRx), this is a *quadratically constrained convex program*, since $\mathbf{H_v}(f_k)$ is a Hermitian matrix. Symmetrically, the optimal Rx beamformer that maximizes the Rx array factor given a fixed Tx beamformer $\mathbf{w}$, can be obtained by solving:

$$(\mathsf{P2}) \ \max_{\mathbf{v}} \ a_\mathrm{r}, \quad \text{subject to: } \mathsf{Re}[\mathbf{s}_\mathrm{r}^\top \mathbf{v}] = a_\mathrm{r}, \ \mathsf{Im}[\mathbf{s}_\mathrm{r}^\top \mathbf{v}] = 0,$$

$$\mathbf{v}^\dagger \mathbf{H_w}(f_k)\mathbf{v} \cdot \tfrac{P_\mathrm{t}}{N} \leq N\beta, \ \forall k,$$

$$|v_n|^2 \leq 1, \ \forall n.$$

(P1) and (P2) are convex programs that can be solved efficiently via existing solvers (e.g., CVX). Intuitively, an algorithm that updates $\mathbf{w}$ and $\mathbf{v}$ by iteratively solving (P1) and (P2) can be applied, i.e., solving for $\mathbf{v}$ given fixed $\mathbf{w}$, and then solving for an updated $\mathbf{w}$ with the newly obtained $\mathbf{v}$. However, the obtained TxBF and RxBF gains can be largely imbalanced since $\mathbf{w}$ and $\mathbf{v}$ are updated *independently*.

6.4.3 The Iterative Algorithm

We now present an iterative algorithm (described in Algorithm 6.1) that *simultaneously* maximizes and balances the Tx and Rx array factors.[5] Let $\kappa \in \mathbb{Z}$ be the index of iteration. Let $\mathbf{w}^{(0)}$ and

---

[5]The idea of the iterative algorithm is similar to that presented in [157].

$\mathbf{v}^{(0)}$ be the initial Tx and Rx beamformers with corresponding Tx and Rx array factors of $a_{\mathrm{t}}^{(0)}$ and $a_{\mathrm{r}}^{(0)}$, respectively. Let $\mathbf{w}^{(\kappa)}$ and $a_{\mathrm{t}}^{(\kappa)}$ (resp. $\mathbf{v}^{(\kappa)}$ and $a_{\mathrm{r}}^{(\kappa)}$) be the optimal Tx (resp. Rx) beamformer and Tx (resp. Rx) array factor obtained by the iterative algorithm in the κth iteration. For $\kappa \in \mathbb{Z}$, we define the following two objective functions.

$$F_{\mathrm{t}}^{(\kappa+1)}(a_{\mathrm{t}}) = a_{\mathrm{t}} - \alpha_{\kappa+1} \cdot (a_{\mathrm{t}} - a_{\mathrm{r}}^{(\kappa)})^2, \ F_{\mathrm{r}}^{(\kappa+1)}(a_{\mathrm{r}}) = a_{\mathrm{r}} - \alpha_{\kappa+1} \cdot (a_{\mathrm{r}} - a_{\mathrm{t}}^{(\kappa+1)})^2, \qquad (6.22)$$

where $\alpha_{\kappa+1}$ is the *step size*. Essentially, in the $(\kappa + 1)^{\mathrm{th}}$ iteration, a penalty term is introduced, which is the square of the difference between the Tx and Rx array factors with a weighting factor of $\alpha_{\kappa+1}$. Therefore, Tx and Rx array factors with a larger difference will prevent their individual value from increasing rapidly.

To allow analytical tractability and easy implementation of the developed iterative algorithm, it is important to properly select: (i) the initial Tx and Rx beamformers, $\mathbf{w}^{(0)}$ and $\mathbf{v}^{(0)}$, and (ii) the step size sequence, $\{\alpha_{\kappa}\}_{\kappa \in \mathbb{Z}}$. In particular, we set:

$$\mathbf{w}^{(0)} = \frac{\beta^{1/4} \cdot \mathbf{w}_{\mathrm{conv}}}{2P_{\mathrm{t}}^{1/4} N^{1/2}}, \ \mathbf{v}^{(0)} = \frac{\beta^{1/4} \cdot \mathbf{v}_{\mathrm{conv}}}{2P_{\mathrm{t}}^{1/4} N^{1/2}}, \ \alpha_{\kappa} = \frac{1}{\kappa^2}, \ \forall \kappa, \qquad (6.23)$$

since this choice of $\mathbf{w}^{(0)}$ and $\mathbf{v}^{(0)}$ satisfies constraint (6.20) in (Opt-TxRx), i.e.,

$$|(\mathbf{v}^{(0)})^{\top} \mathbf{H}(f_k) \mathbf{w}^{(0)}|^2 \cdot \frac{P_{\mathrm{t}}}{N} \leq \left[\frac{\beta^{1/4}}{2P_{\mathrm{t}}^{1/4} N^{1/2}} \right]^4 \cdot |\mathbf{v}_{\mathrm{conv}}|^2 \cdot |\mathbf{w}_{\mathrm{conv}}|^2 \cdot \frac{P_{\mathrm{t}}}{N} = \frac{1}{2} \cdot N\beta < N\beta.$$

We also note that the above choices of $\mathbf{w}^{(0)}$, $\mathbf{v}^{(0)}$, and $\{\alpha_{\kappa}\}$ are *not unique*. For example, any step size sequence $\{\alpha_{\kappa}\}$ satisfying $1 = \alpha_1 \geq \alpha_2 \geq \cdots > 0$ also suffices. In Section 6.5, we will evaluate the effect of $\{\alpha_{\kappa}\}$ on the solution obtained by the iterative algorithm.

6.4.4 Performance Analysis

In this section, we analyze the performance of the iterative algorithm. We first present Lemma 6.1 about the structural properties of the objective functions of (Opt-Tx) and (Opt-Rx) in (6.22). Then, we state the main results in Proposition 6.1. The proofs of Lemma 6.1 and Proposition 6.1 can be found in Appendices 6.7.1 and 6.7.2, respectively.

Algorithm 6.1 The Iterative Algorithm

Input and Initialization: N, $P_{\rm t}$, $\mathbf{H}(f_k), \forall k$, $\mathbf{s}_{\rm t} = \mathbf{s}(\phi_{\rm t}, \theta_{\rm t})$, and $\mathbf{s}_{\rm r} = \mathbf{s}(\phi_{\rm r}, \theta_{\rm r})$. Initial values of Tx and Rx beamformers $\mathbf{w}^{(0)}$ and $\mathbf{v}^{(0)}$, respectively. The step size sequence, $\{\alpha_\kappa\}_{\kappa \in \mathbb{Z}}$.

For $\kappa = 0, 1, \cdots$ **do**

1: Obtain $\mathbf{w}^{(\kappa+1)}$ with given $\mathbf{v}^{(\kappa)}$ and $a_{\rm r}^{(\kappa)}$ by solving the following optimization problem, (Opt-Tx):

$$(\text{Opt-Tx}) \; a_{\rm t}^{(\kappa+1)} := \arg\max_{\mathbf{w}} \; F_{\rm t}^{(\kappa+1)}(a_{\rm t}) = a_{\rm t} - \alpha_{\kappa+1} \cdot (a_{\rm t} - a_{\rm r}^{(\kappa)})^2$$

$$\text{s.t.} : \mathrm{Re}[\mathbf{s}_{\rm t}^\top \mathbf{w}] = a_{\rm t}, \; \mathrm{Im}[\mathbf{s}_{\rm t}^\top \mathbf{w}] = 0, \; \mathbf{w}^\dagger \mathbf{H}_{\mathbf{v}^{(\kappa)}}(f_k)\mathbf{w} \cdot \tfrac{P_{\rm t}}{N} \leq N\beta, \; \forall k, \; |w_n|^2 \leq 1, \; \forall n.$$

2: Obtain $\mathbf{v}^{(\kappa+1)}$ with given $\mathbf{w}^{(\kappa+1)}$ and $a_{\rm t}^{(\kappa+1)}$ by solving the following optimization problem, (Opt-Rx):

$$(\text{Opt-Rx}) \; a_{\rm r}^{(\kappa+1)} := \arg\max_{\mathbf{v}} \; F_{\rm r}^{(\kappa+1)}(a_{\rm r}) = a_{\rm r} - \alpha_{\kappa+1} \cdot (a_{\rm r} - a_{\rm t}^{(\kappa+1)})^2$$

$$\text{s.t.} : \mathrm{Re}[\mathbf{s}_{\rm r}^\top \mathbf{v}] = a_{\rm r}, \; \mathrm{Im}[\mathbf{s}_{\rm r}^\top \mathbf{v}] = 0, \; \mathbf{v}^\dagger \mathbf{H}_{\mathbf{w}^{(\kappa+1)}}(f_k)\mathbf{v} \cdot \tfrac{P_{\rm t}}{N} \leq N\beta, \; \forall k, \; |v_n| \leq 1, \; \forall n.$$

3: Iterate until the Tx and Rx array factor improvements are within δN, i.e.,

$$\max\left\{ a_{\rm t}^{(\kappa+1)} - a_{\rm t}^{(\kappa)}, a_{\rm r}^{(\kappa+1)} - a_{\rm r}^{(\kappa)} \right\} \leq \delta \cdot N. \tag{6.24}$$

Lemma 6.1 *Under the iterative algorithm,* $\forall \kappa \in \mathbb{Z}$,

$$F_t^{(\kappa+1)}(a_t^{(\kappa+1)}) \geq F_t^{(\kappa+1)}(a_t^{(\kappa)}), \; F_r^{(\kappa+1)}(a_r^{(\kappa+1)}) \geq F_r^{(\kappa+1)}(a_r^{(\kappa)}).$$

Recall that $g_{\rm t}^{(\kappa)} = (a_t^{(\kappa+1)})^2/N$ and $g_{\rm r}^{(\kappa)} = (a_r^{(\kappa+1)})^2/N$. We have the following proposition.

Proposition 6.1 *With initial Tx and Rx beamformers,* $\mathbf{w}^{(0)}$ *and* $\mathbf{v}^{(0)}$, *and step size,* α_κ, *given in (6.23), under the iterative algorithm,*

$$a_t^{(\kappa+1)} \geq a_t^{(\kappa)} \text{ and } a_r^{(\kappa+1)} \geq a_r^{(\kappa)}, \; \forall \kappa \in \mathbb{Z}. \tag{6.25}$$

Furthermore, the corresponding TxBF and RxBF gains satisfy, $\forall \kappa \in \mathbb{Z}$,

$$g_t^{(\kappa+1)} \geq g_t^{(\kappa)}, \; g_t^{(\kappa+1)} \geq g_r^{(\kappa)}, \; g_t^{(\kappa+1)} + g_r^{(\kappa+1)} \geq g_t^{(\kappa)} + g_r^{(\kappa)}. \tag{6.26}$$

Corollary 6.1 *The iterative algorithm is guaranteed to terminate in at most* $\lceil 1/\delta \rceil$ *iterations.*

Proof 6.1 *Proposition 6.1 states that the Tx and Rx array factors obtained by Algorithm 6.1 after each iteration,* $a_t^{(\kappa)}$ *and* $a_r^{(\kappa)}$, *are always* monotonically non-decreasing. *From the termination condition (6.24) and the fact that* $a_t^{max} = a_r^{max} = N$ *(see (6.5)), the iterative algorithm is guaranteed to terminate within at most* $\lceil N/(\delta \cdot N) \rceil = \lceil 1/\delta \rceil$ *iterations.*
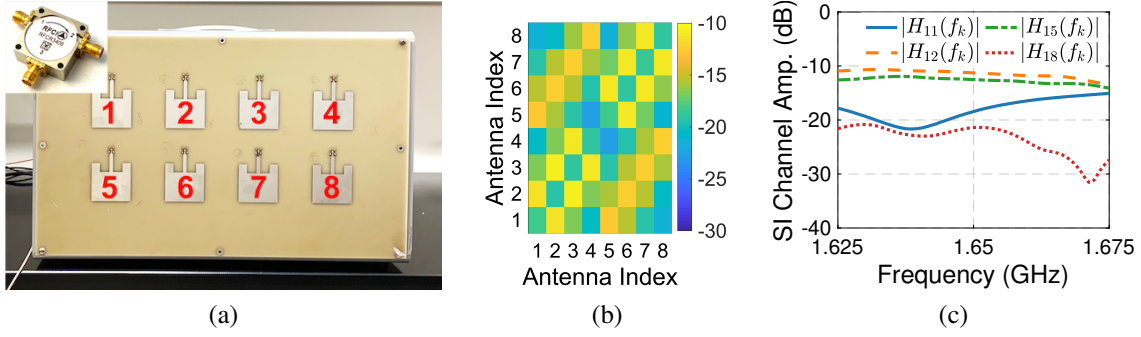
Figure 6.8: (a) A customized 1.65 GHz 8-element rectangular array and the RFCR3206 circulator, (b) the measured $|H_{mn}(f_k)|$ at 1.65 GHz, and (c) the amplitudes of some example SI channels, $|H_{mn}(f_k)|$.

6.5 Measurement-based Evaluation

In this section, we first describe the measurements, datasets, and setup. Then, we numerically evaluate the performance of an FD phased array with TxBF and RxBF, and the corresponding FD rate gains. We also discuss various design tradeoffs.

6.5.1 Measurements and Datasets

Since currently large-scale Tx and Rx phased array nodes are not widely available, our evaluations are based on $\mathbf{H}(f_k)$ from measurements and traces. In particular, we consider two antenna arrays with different array geometries, number of antennas, N, and bandwidth, B.

A Customized Rectangular Array with Circulators. We custom designed a 1.65 GHz 8-element rectangular antenna array using a slot loop antenna structure as shown in Fig. 6.8(a). The spacing between adjacent antennas is half-wavelength. An RF-CI RFCR3406 circulator [158] is also included (see Fig. 6.3). The frequency responses of the antenna array and the circulator are measured using a vector network analyzer at frequencies between 1.625–1.675 GHz ($B = 50$ MHz), from which the SI channel matrix, $\mathbf{H}(f_k) \in \mathbb{C}^{8 \times 8}$, is constructed. Figs. 6.8(b) and 6.8(c) respectively plot the measured $|H_{mn}(f_k)|$ at $f_k = 1.65$ GHz, and example SI channels, $|H_{mn}(f_k)|$, with high frequency-selectivity.

The Argos Hexagonal Array [156, 159]. We also leverage the publicly available Argos dataset
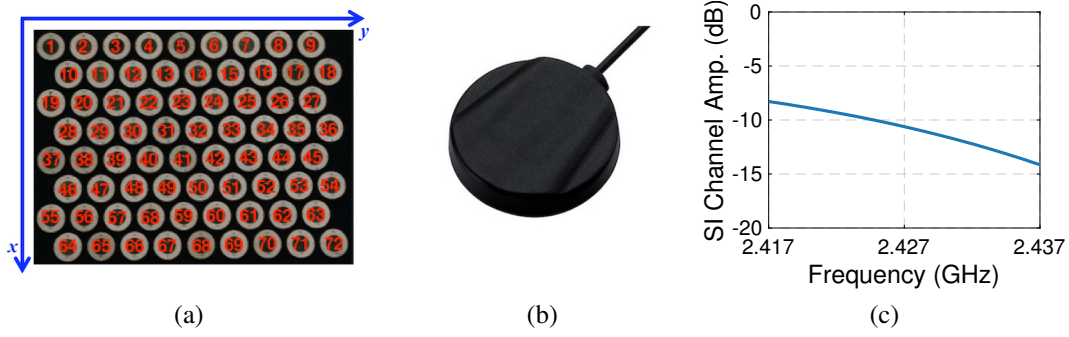
Figure 6.9: (a) The 2.4 GHz 72-element Argos hexagonal array with SI channel measurements reported in [90, 156], (b) the Taoglas 2.4 GHz circular antenna, and (c) the measured antenna amplitude response used for $|H_{nn}(f_k)|, \forall n$.

from [156, 159]. The Argos platform consists of 72 circular patch antennas at 2.4 GHz placed in a hexagonal grid consisting of 8 rows and 9 columns, with 0.6-wavelength spacing between adjacent elements (see Fig. 6.9(a)). The SI channel matrix, $\mathbf{H}(f_k) \in \mathbb{C}^{72 \times 72}$, is measured using a WARPv3 platform with $B = 20$ MHz bandwidth and $K = 64$ subcarriers (52 non-zero subcarriers). With such a large number of antennas, uniform linear arrays (ULAs) and hexagonal planar arrays with different values of N can be constructed by taking a subset of the measurements.

However, the Argos platform employs *seperate* Tx and Rx antennas and does not contain circulators. Therefore, $\mathbf{H}(f_k)$ is missing the diagonal elements, $H_{nn}(f_k), \forall n$. To complete $\mathbf{H}(f_k)$, we measure the frequency response of a Taoglas 2.4 GHz circular antenna [160] (see Fig. 6.9(b)). Using the completed $\mathbf{H}(f_k)$, we generate hexagonal arrays with $N \in \{9, \ldots, 72\}$ by considering the top $\{1, \ldots, 8\}$ rows of the Argos array. Note that case of $N = 9$ corresponds to a ULA.

Steering Vectors of the Rectangular and Argos Arrays. The steering vector of the rectangular array is computed as described in Section 6.1. The steering vector of the Argos hexagonal array is given by ($n = (n_x - 1)N_y + n_y$)

$$
s_n(\phi, \theta) = \begin{cases} e^{j \frac{6\pi}{5}[(n_x - 1)\cos\theta\cos\phi + \frac{\sqrt{3}}{2}(n_y - 1)\cos\theta\sin\phi]}, & n_y \text{ odd}, \\ e^{j \frac{6\pi}{5}[(n_x - \frac{1}{2})\cos\theta\cos\phi + \frac{\sqrt{3}}{2}(n_y - 1)\cos\theta\sin\phi]}, & n_y \text{ even}. \end{cases}
$$

Our evaluations using the Argo dataset is with $B = 20$ MHz as provided [156, 159]. For higher bandwidth up to $B = 50$ MHz, we use the rectangular array measurements. For compactness of

(a) $\alpha_\kappa = 1/\kappa^2$

(b) $\alpha_\kappa = 1/\kappa$

(c) $\alpha_\kappa = 1$
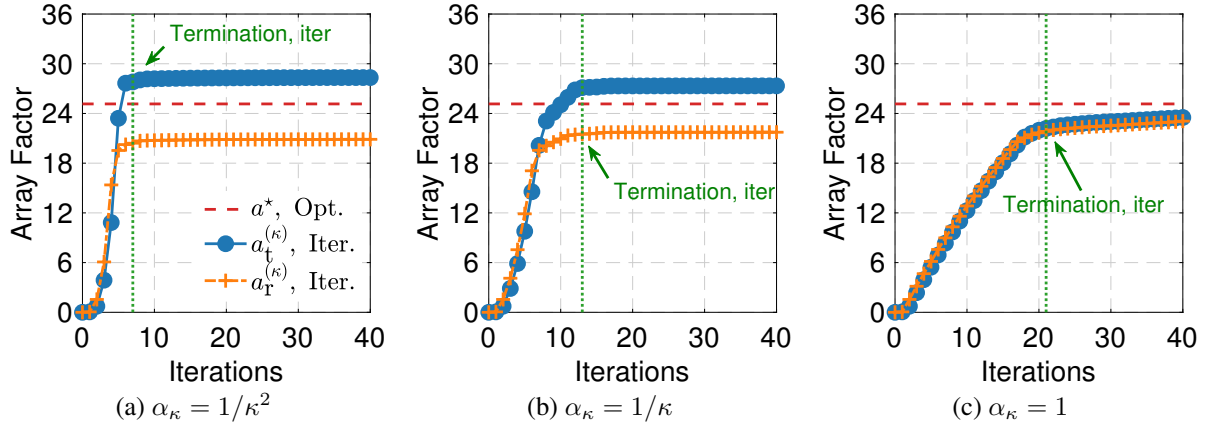
Figure 6.10: The optimal Tx and Rx array factors, $a^\star$, and the Tx and Rx array factors obtained the by iterative algorithm, $a_t^{(\kappa)}$ and $a_r^{(\kappa)}$, with $N = 36$, $P_t = 20\,\text{dBm}$, and different step sizes, $\{\alpha_\kappa\}$.

presentation, unless mentioned otherwise, we use $N \in \{8, 9, 18, \ldots, 72\}$ to correspond to different array geometries. Note that in both antenna arrays, the measured $\mathbf{H}(f_k)$ is neither Hermitian nor symmetric although Fig. 6.8(b) presents some level of symmetry.

6.5.2 Setup

TxBF and RxBF in HD and FD Modes. We consider TxBF and RxBF in the *front side* of the antenna array with $\phi \in [-180°, 180°]$ and $\theta \in [0°, 90°]$. Specifically, the array *broadside* corresponds to the direction of $\theta = 90°, \forall \phi$ (see Fig. 6.2(a)). We consider $P_t \in \{10, 20, 30\}$ dBm[6], and $P_{\text{nf}} = -90\,\text{dBm}$ and $SIC_{\text{dig}} = 40\,\text{dB}$ (see Section 6.2.4). The FD rate gains are computed as described in Section 6.2.3.

The following TxBF and RxBF schemes are considered:

(1) *Conventional HD TxBF and RxBF* (**Conv.**), which is based on (6.5) and achieves $a_t^{\text{max}} = a_r^{\text{max}} = N$ and $g_t^{\text{max}} = g_r^{\text{max}} = N$;

(2) *Only TxBF* (**TxBF**), which is similar to the approach presented in [91] but adapted to our FD phased array model (see Section 6.2). Since [91] considers a narrowband system where only TxBF is repurposed, we set the Rx beamformer in this scheme to be $\mathbf{v}_{\text{conv}}$, which maximizes the RxBF gain in the main beam-pointing direction, and optimize for the Tx beamformer

---

[6]These values correspond to the typical Tx power levels of a BS in a small/micro cell.
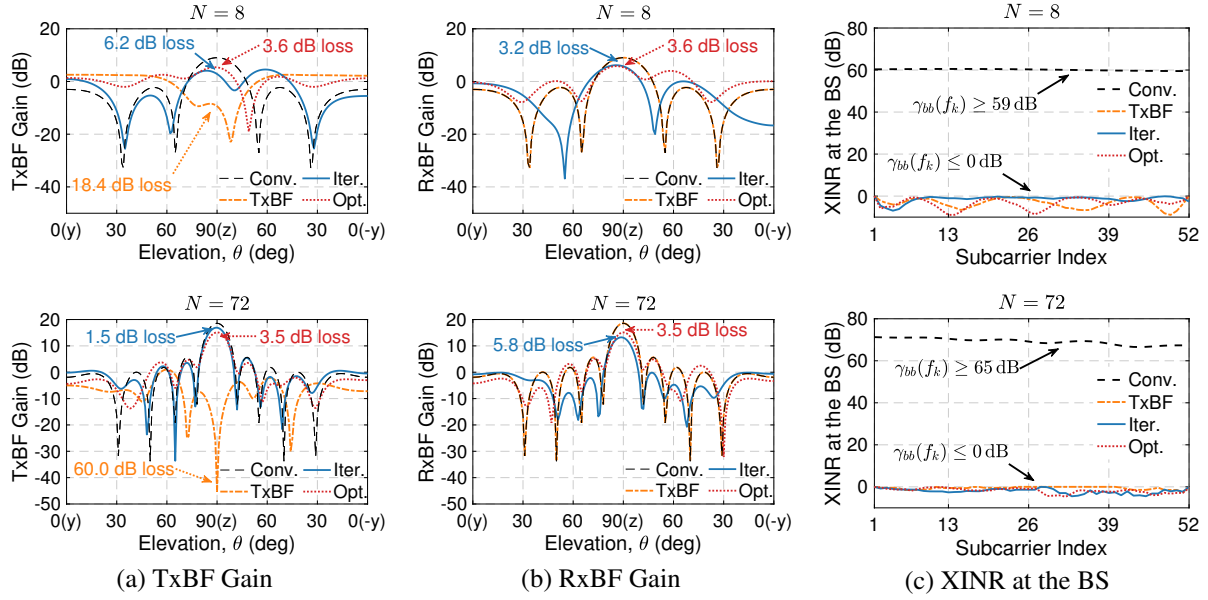
Figure 6.11: TxBF and RxBF gains (on the y-z plane), and the resulting XINR under different TxBF and RxBF schemes with the rectangular array ($N = 8$) and the Argos array ($N = 72$), $P_t = 30\,\mathrm{dBm}$, $B = 20\,\mathrm{MHz}$, and the Tx and Rx beam-pointing directions in the array broadside (z-axis).

across the desired wide bandwidth;

(3) *Optimal FD TxBF and RxBF* (**Opt.**), which is based on solving (Opt-TxRx). Recall that the obtained (equal) optimal Tx and Rx array factor is denoted by $a^\star$, and corresponds to the (equal) optimal TxBF and RxBF gain of $g^\star$;

(4) *Iterative FD TxBF and RxBF* (**Iter.**), which is based on the iterative algorithm with $\delta = 0.01$ set in the termination step. Denote by $\widetilde{a}_t$ (resp. $\widetilde{a}_r$) the returned Tx (resp. Rx) array factor and by $\widetilde{g}_t$ (resp. $\widetilde{g}_r$) the corresponding TxBF (resp. RxBF) gain, respectively.

The evaluations are performed using a laptop with a quad-core Intel i7 CPU and 16GB RAM. For **Opt.**, we apply the nonlinear solver in MATLAB.[7] For **Iter.**, we apply the MATLAB CVX solver for solving the convex (Opt-Tx) and (Opt-Rx) in each iteration of Algorithm 6.1. For **TxBF**, we also apply the CVX solver given the convexity of the optimization problem.

Selecting the Step Size, $\{\alpha_\kappa\}$. To study the impact of $\{\alpha_\kappa\}$, we remove the termination condition

---

[7]The returned solution to (Opt-TxRx) may be locally optimal due to its non-convexity.

in the iterative algorithm and record the obtained $a_{\mathrm{t}}^{(\kappa)}$ and $a_{\mathrm{r}}^{(\kappa)}$ (see Algorithm 6.1). We consider $N = 36$, $P_{\mathrm{t}} = 20\,\mathrm{dBm}$, and three different step sizes satisfying the conditions specified in Section 6.4.3: (i) $\alpha_\kappa = 1/\kappa^2$, (2) $\alpha_\kappa = 1/\kappa$, and (iii) $\alpha_\kappa = 1$ (constant). Fig. 6.10 plots both the obtained $a_{\mathrm{t}}^{(\kappa)}$ and $a_{\mathrm{r}}^{(\kappa)}$ over iterations, κ, and the optimal Tx and Rx array factor, $a^\star$ (red dashed line). The results show that under all considered three choices of $\{\alpha_\kappa\}$, $a_{\mathrm{t}}^{(\kappa)}$ and $a_{\mathrm{r}}^{(\kappa)}$ converge within 25 iterations. However, $a_{\mathrm{t}}^{(\kappa)}$ and $a_{\mathrm{r}}^{(\kappa)}$ become more imbalanced with more aggressive step sizes (e.g., $\alpha_\kappa = 1/\kappa^2$). The results for other values of N and P_{t} also reveal similar trends. Therefore, we empirically set $\alpha_\kappa = 1/\kappa^2$, which achieves fast termination (e.g., less than 10 iterations for all values of N, P_{t}, and B considered) and for which the obtained $\widetilde{g}_{\mathrm{t}}$ and $\widetilde{g}_{\mathrm{r}}$ are relatively well balanced.

6.5.3   XINR and Gain Loss under TxBF and RxBF

We now evaluate the XINR at the BS, $\gamma_{bb}(f_k)$, under different TxBF and RxBF schemes. We consider both the customized rectangular array ($N = 8$) and the Argos array ($N = 72$), with $P_{\mathrm{t}} = 30\,\mathrm{dBm}$, $B = 20\,\mathrm{MHz}$, and the desired Tx and Rx beam-pointing directions in the array broadside (z-axis). Fig. 6.11 plots the TxBF and RxBF gains (see (6.2)) and the resulting XINR, $\gamma_{bb}(f_k)$, under the considered TxBF and RxBF schemes. It can be seen that **Conv.** results in extremely high XINR of $\gamma_{bb}(f_k) \geq 65\,\mathrm{dB}$, $\forall k$, thereby FD operation at the BS cannot be supported. Both **Opt.** and **Iter.** are able to cancel the SI power to below the array noise floor, i.e., $\gamma_{bb}(f_k) \leq 1$ (0 dB). The corresponding TxBF and RxBF gain losses under **Opt./Iter.** are only 3.6/6.2 dB and 3.6/3.2 dB for $N = 8$. The TxBF and RxBF gain losses are reduced to 3.5/1.5 dB and 3.5/5.8 dB for $N = 72$.

Fig. 6.11 also shows that the **TxBF** scheme leads to TxBF gain losses of 18.4/60.0 dB for 8/72 elements, respectively, which are significantly larger than those in the other schemes. This is because with a *fixed* Rx beamformer that aims to achieve the maximum RxBF gain (as applied in **TxBF**), more attenuated Tx beamformer weights are required to cancel the strong SI. Therefore, we do not include **TxBF** in the rest of the evaluations since these largely degraded TxBF gains will

144
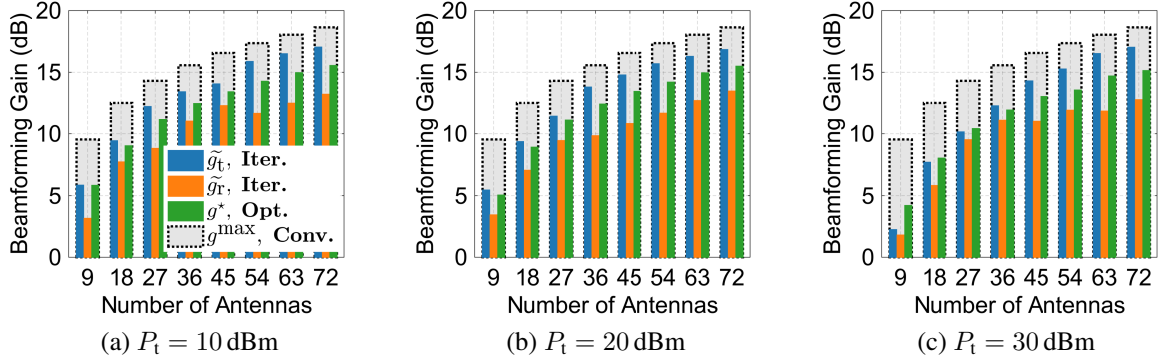
Figure 6.12: TxBF and RxBF gains under different TxBF and RxBF schemes with varying number of antennas, N, and $P_t \in \{10, 20, 30\}$ dBm.

lead to poor FD rate gains.

6.5.4 FD TxBF and RxBF Gains and Rate Gain

FD TxBF and RxBF Gains. We first evaluate the FD TxBF and RxBF gains. We consider Tx and Rx beam-point directions in the array broadside with $N \in \{9, 18, \ldots, 72\}$ and $P_t \in \{10, 20, 30\}$ dBm. Fig. 6.12 plots the optimal FD TxBF and RxBF gain, $g^\star$, and the iterative FD TxBF and RxBF gains, $\widetilde{g}_t$ and $\widetilde{g}_r$, respectively. The conventional HD TxBF and RxBF gains of N are also plotted. The results show that for a given number of antennas, N, the TxBF and RxBF gain losses are more significant with increased Tx power level, P_t. For a given value of P_t, the TxBF and RxBF gain losses decrease with a larger number of antennas, N. For example, under the **Iter.** scheme, an FD phased array with $P_t = 20$ dBm and $N = 72$ experiences 1.7/5.1 dB TxBF/RxBF gain losses, respectively. These values are only marginally changed to 1.8/5.7 dB with $N = 36$. It can also be seen that **Iter.** achieves relative balanced FD TxBF and RxBF gains across varying N. Specifically, $\widetilde{g}_t$ and $\widetilde{g}_r$ are always within ± 2.8 dB of the optimal FD TxBF and RxBF gain, $g^\star$. Moreover, $\gamma_{bb}(f_k) \leq 1, \forall k$, can be achieved with at most 8.0/8.2/11.6 dB *sum* TxBF and RxBF gain loss for $P_t = 10/20/30$ dBm when $N \geq 18$.

BS-BS Case. We consider the FD rate gain when both BSs face each other in the array braodside. Fig. 6.13 plots the FD rate gains under the **Iter.** scheme with varying P_t and link SNR values.

(a) Low SNR (0 dB)   (b) Medium SNR (15 dB)   (c) High SNR (30 dB)
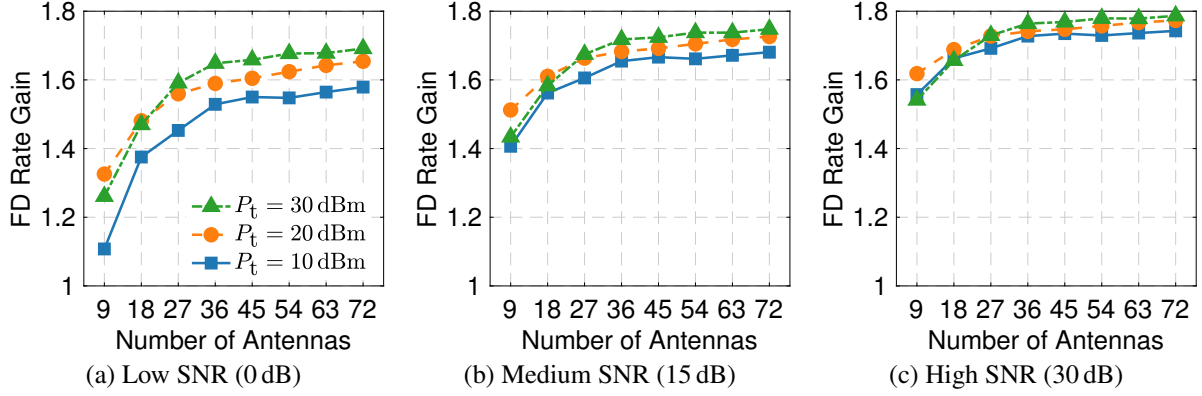
Figure 6.13: FD rate gain in the BS-BS case when both BSs face each other in the broadside, with varying number of antennas, N, and $P_{\mathrm{t}} \in \{10, 20, 30\}$ dBm, in the low/medium/high SNR regimes.

Avg. ratio b/w FD rate gains under $(\widetilde{g}_{\mathrm{t}}, \widetilde{g}_{\mathrm{r}})$ and $g^\star$	$N = 9$	$N = 18$	$N = 27$	$N = 36$	$N \geq 45$
Low SNR (0 dB)	0.89	0.95	0.97	0.98	≥ 0.98
Medium SNR (15 dB)	0.93	0.97	0.98	0.99	≥ 0.99
High SNR (30 dB)	0.95	0.98	0.98	0.99	≥ 0.99

Table 6.2: Average ratio between the FD rate gains achieved under **Iter.** (with $\widetilde{g}_{\mathrm{t}}$ and $\widetilde{g}_{\mathrm{r}}$) and **Opt.** (with $g^\star$).

The results show that although the FD phased array experiences TxBF and RxBF gain losses to achieve $\gamma_{bb}(f_k) \leq 1, \forall k$, an FD rate gain of at least $1.53\times$ can be achieved with $N \geq 36$ in all SNR regimes. Also, the FD rate gain improves with increased values of both N and the link SNR. Moreover, when the number of antennas is large, further increasing N introduces only marginal FD rate gain since the SI power is already canceled to below the noise floor with a smaller value of N (see Section 6.3).

To compare the performance of the **Iter.** and **Opt.** schemes, Table 6.2 summarizes the average ratio between the FD rate gains achieved by **Iter.** (with $\widetilde{g}_{\mathrm{t}}$ and $\widetilde{g}_{\mathrm{r}}$) and **Opt.** (with $g^\star$). The results show that the FD rate gains achieved under $\widetilde{g}_{\mathrm{t}}$ and $\widetilde{g}_{\mathrm{r}}$ are very close to that achieved under $g^\star$, in all considered scenarios, where the average ratio is at least 89%.

BS-User Case with Spatially Distributed Users. To evaluate the FD rate gain, we consider spatially distributed users in the directions of $\phi \in [-180°, 180°]$ and $\theta \in [0°, 90°]$ with respect to
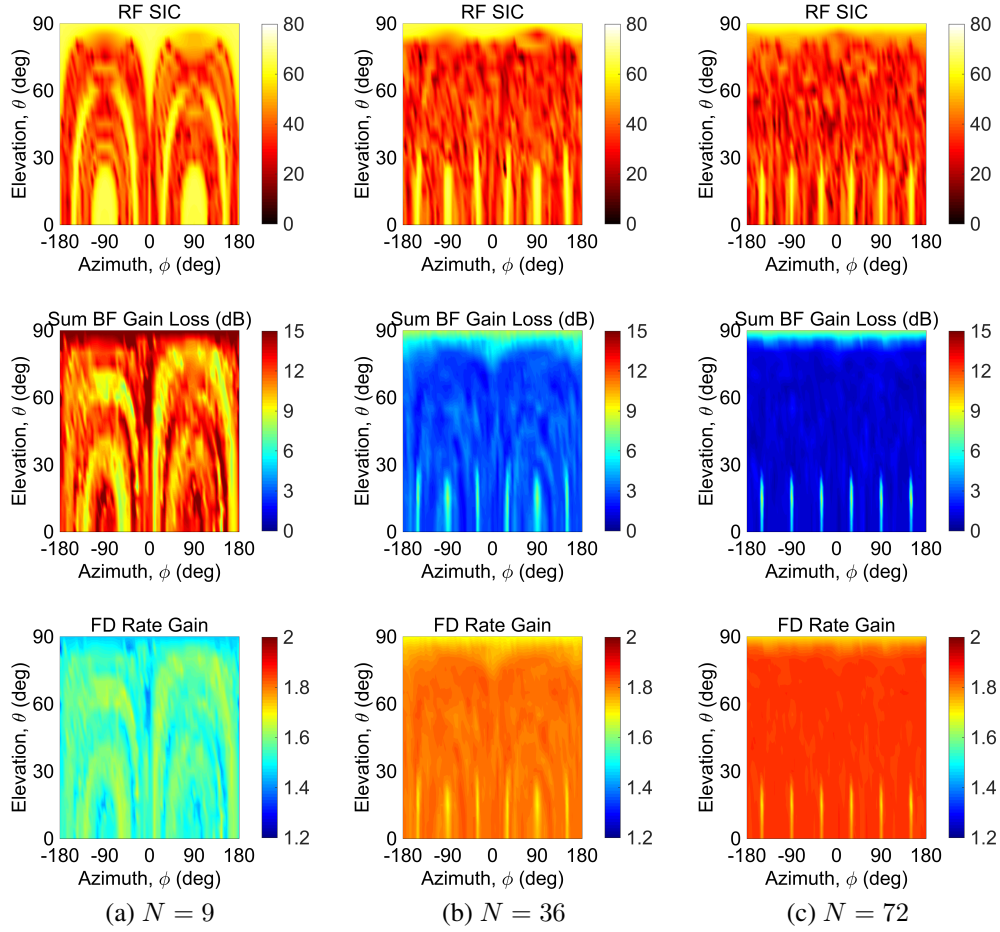
Figure 6.14: Spatial distributions of the sum TxBF and RxBF gain loss (top row), and the FD rate gain in the BS-User case when the user is in different spatial directions from the FD BS (bottom row), with $N \in \{9, 36, 72\}$, $P_{\mathrm{t}} = 30\,\mathrm{dBm}$, and $\gamma_{bu} = \gamma_{ub} = 0\,\mathrm{dB}$.

the BS. The BS applies the **Iter.** scheme with the desired Tx and Rx beam-pointing directions equal to the user direction. We consider the low SNR regime with $N \in \{9, 36, 72\}$ and $P_{\mathrm{t}} = 30\,\mathrm{dBm}$ (see Section 6.3). Fig. 6.14 plots the spatial distributions of (i) the amount of RF SIC achieved at the BS by the **Iter.** scheme, (ii) the sum TxBF and RxBF gain loss for achieving $\gamma_{bb}(f_k) \le 1, \forall k$, and (iii) the resulting FD rate gain with low UL and DL SNR values of $0\,\mathrm{dB}$.

The results show that the sum TxBF and RxBF gain loss varies across all spatial directions, since the total SI power depends on both the array geometry and the beam-pointing directions. The SI power is the strongest (i.e., the required amount of RF SIC is the highest) in the array broadside (z-axis) and in the direction of adjacent antennas close to the array x-y plane (e.g.,

147

Number of Antennas, N	9	18	27	36	45	54	63	72
Runtime Improvements	0.99×	1.72×	2.41×	2.12×	2.70×	3.18×	5.51×	6.00×

Table 6.3: Runtime improvements of the iterative algorithm (Algorithm 6.1) over directly solving the non-convex problem (Opt-TxRx).

$\phi = \pm 90°$ for $N = 9$ and $\phi = \pm 30/90/150°$ for $N \in \{36, 72\}$, with very small values of θ). Yet, the **Iter.** scheme is still able to achieve $\gamma_{bb}(f_k) \leq 1$ (more than 65 dB RF SIC) under $P_t = 30$ dBm with maximum sum TxBF and RxBF gain losses of 9.7/8.6 dB for $N = 36/72$. Overall, the FD rate gains are at least 1.33/1.66/1.68× for $N = 9/36/72$, and when the user is not in the direction of the strongest SI power, the FD rate gains can be increased to 1.68/1.83/1.87× for $N = 9/36/72$. Note that the considered UL and DL SNR values are only 0 dB, and higher link SNR values would also increase FD rate gain at the same sum TxBF and RxBF gain loss.

Efficiency of the Iterative Algorithm. We also compare the performance of the **Iter.** and **Opt.** schemes in terms of the time consumed to obtain the Tx and Rx beamformers. We perform 100 runs of solving (Opt-TxRx) and of the iterative algorithm in all considered values of N and P_t, and measure the average running times. Table 6.3 summarizes the improvements in the average running time of **Iter.** over **Opt.**. The results show that with $N = 9$, both schemes have similar running times. However, as N increases, the **Iter.** scheme achieves 2–6× runtime improvements compared to the **Opt.** scheme, since the latter is solving the non-convex problem (Opt-TxRx).

Effect of the Bandwidth, B. Lastly, we evaluate the effects of the desired RF SIC bandwidth, B, on the FD TxBF and RxBF using the measurements of the 8-element rectangular array with circulators (recall that the Argos dataset is only with $B = 20$ MHz). We consider the **Iter.** scheme with $B \in \{10, \ldots, 50\}$ MHz and $P_t \in \{10, 20, 30\}$ dBm. Fig. 6.15 plots the sum TxBF and RxBF gain loss and the corresponding FD rate gain in both the BS-User and BS-BS cases. The results show that, even with only 8 elements, an FD phased array can achieve $\gamma_{bb}(f_k) \leq 1, \forall k$, for up to $B = 50$ MHz at $P_t = 10$ dBm, where the sum TxBF and RxBF gain loss is at most 8.5 dB (TxBF/RxBF gain loss of 4.4/4.1 dB). The sum TxBF and RxBF gain loss increases to 12.3 dB with $P_t = 20$ dBm. However, although higher TxBF and RxBF gain losses are required in
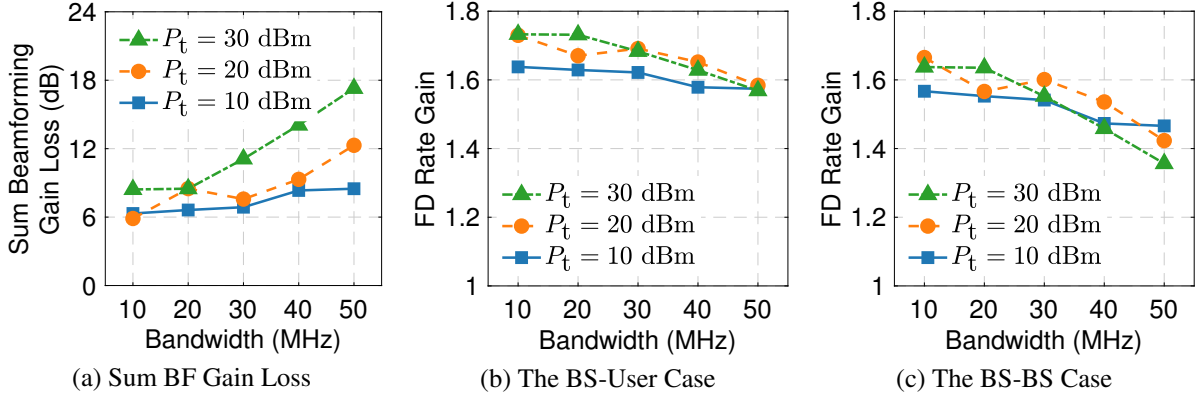
Figure 6.15: Sum TxBF and RxBF gain loss, and FD rate gains in the BS-User and BS-BS cases with $N = 8$ and varying desired RF SIC bandwidth, $B \in \{10, \ldots, 50\}$ MHz, and $P_t \in \{10, 20, 30\}$ dBm.

scenarios with increased P_t and B, an 8-element FD phased array is able to achieve FD rate gains of at least $1.47/1.42/1.36\times$ under $P_t = 10/20/30$ dBm, with bandwidth of up to $B = 50$ MHz.

6.6 Conclusion

In this chapter, we considered FD phased arrays repurposing TxBF and RxBF for achieving wideband RF SIC. We formulated optimization problems to obtain the maximum FD TxBF and RxBF gains and developed an iterative algorithm to efficiently solve the optimization problems. Using measurements and datasets, we extensively evaluated the performance of the FD phased array and the resulting FD rate gains in various network settings. In [28], we implemented an 8-element FD phased array RFIC integrating the developed optimization framework, where 100 dB overall SIC was achieved across the RF and digital domains at 100 mW average Tx power (WiFi-like) over 20 MHz bandwidth.

6.7 Appendices

6.7.1 Appendix – Proof of Lemma 6.1

Since $\mathbf{w}^{(\kappa)}$ and $\mathbf{v}^{(\kappa)}$ are optimal solutions to optimization problems (Opt-Tx) and (Opt-Rx) in the κth iteration, it holds that

$$
\begin{cases}
\text{Re}[\mathbf{s}_t^\top \mathbf{w}^{(\kappa')}] = a_t^{(\kappa')}, \ \text{Im}[\mathbf{s}_t^\top \mathbf{w}^{(\kappa')}] = 0, \ \text{Re}[\mathbf{s}_r^\top \mathbf{v}^{(\kappa')}] = a_r^{(\kappa')}, \ \text{Im}[\mathbf{s}_r^\top \mathbf{v}^{(\kappa')}] = 0, \ (\kappa' = \kappa, \kappa+1), \\
(\mathbf{v}^{(\kappa)})^\dagger \mathbf{H}_{\mathbf{w}^{(\kappa)}}(f_k)\mathbf{v}^{(\kappa)} \cdot \frac{P_t}{N} \leq N\beta, \ (\mathbf{w}^{(\kappa+1)})^\dagger \mathbf{H}_{\mathbf{v}^{(\kappa)}}(f_k)\mathbf{w}^{(\kappa+1)} \cdot \frac{P_t}{N} \leq N\beta, \ \forall k.
\end{cases}
$$

Assume by contradiction that $F_t^{(\kappa+1)}(a_t^{(\kappa+1)}) < F_t^{(\kappa+1)}(a_t^{(\kappa)})$, we can select a new Tx beamformer $\widetilde{\mathbf{w}}^{(\kappa+1)} = \mathbf{w}^{(\kappa)}$ with an increased value of $F_t^{(\kappa+1)}$ while satisfying all the constraints in (Opt-Tx). Similarly, $F_r^{(\kappa+1)}(a_r^{(\kappa+1)}) \geq F_r^{(\kappa+1)}(a_r^{(\kappa)})$ also holds. $\qquad\square$

6.7.2 Appendix – Proof of Proposition 6.1

The proof is based on induction on κ and the structural properties of the objective functions (6.22). By taking the second derivatives of (6.22), it can be seen that

- $F_t^{(\kappa+1)}(a_t)$ is symmetric w.r.t. $\psi_t^{(\kappa+1)} = a_r^{(\kappa)} + \frac{1}{2\alpha_{\kappa+1}}$ and is monotonically increasing for $a_t \in (0, \psi_t^{(\kappa+1)}]$ and monotonically decreasing for $a_t \in (\psi_t^{(\kappa+1)}, \infty)$,

- $F_r^{(\kappa+1)}(a_r)$ is symmetric w.r.t. $\psi_r^{(\kappa+1)} = a_t^{(\kappa+1)} + \frac{1}{2\alpha_{\kappa+1}}$ and is monotonically increasing for $a_r \in (0, \psi_r^{(\kappa+1)}]$ and is monotonically decreasing for $a_r \in (\psi_r^{(\kappa+1)}, \infty)$.

The proof is based on the induction of κ.

<u>Base Case</u> ($\kappa = 0$): Let $\mathbf{w}^{(0)}$ and $\mathbf{v}^{(0)}$ be the initial Tx and Rx beamformers given by (6.23) with initial array factors:

$$
a_t^{(0)} = \mathbf{s}_t^\top \mathbf{w}^{(0)} = \tfrac{1}{2} \cdot \beta^{1/4} N^{1/2} P_t^{-1/4}, \ a_r^{(0)} = \mathbf{s}_r^\top \mathbf{v}^{(0)} = \tfrac{1}{2} \cdot \beta^{1/4} N^{1/2} P_t^{-1/4}.
$$

For $\kappa = 1$, $F_t^{(1)}(a_t)$ is monotonically increasing for $a_t \in (0, \psi_t^{(1)}]$. Since $\frac{P_t}{N} \geq N\beta$, it holds that

$$
\psi_t^{(1)} = a_r^{(0)} + \tfrac{1}{2\alpha_1} = \tfrac{1}{2}(\beta^{1/4} N^{1/2} P_t^{-1/4} + 1) \geq \beta^{1/4} N^{1/2} P_t^{-1/4}.
$$

One can then select $\mathbf{w}^{(1)} = 2\mathbf{w}^{(0)}$, which satisfies

$$a_{\mathrm{t}}^{(1)} = \mathbf{s}_{\mathrm{t}}^{\top}\mathbf{w}^{(1)} = \beta^{1/4}N^{1/2}P_{\mathrm{t}}^{-1/4}.$$

Since $|H_{mn}(f_k)| \leq 1$ (see Section 6.2), it holds that

$$|(\mathbf{v}^{(0)})^{\top}\mathbf{H}(f_k)\mathbf{w}^{(1)}|^2 \cdot \tfrac{P_{\mathrm{t}}}{N} = \left|\sum_{m=1}^{N}\sum_{n=1}^{N}(w_n H_{mn}(f_k)v_m)\right|^2 \cdot \tfrac{P_{\mathrm{t}}}{N}$$

$$\leq \left|N^2 \cdot \tfrac{\beta^{1/4}}{N^{1/2}P_{\mathrm{t}}^{1/4}} \cdot \tfrac{\beta^{1/4}}{2N^{1/2}P_{\mathrm{t}}^{1/4}}\right|^2 \cdot \tfrac{P_{\mathrm{t}}}{N} = \tfrac{N\beta}{4} < N\beta.$$

Hence, there $\exists\, \mathbf{w}^{(1)}$ such that $a_{\mathrm{t}}^{(1)} > a_{\mathrm{t}}^{(0)}$ and all constraints in (Opt-Tx) are satisfied. Similarly, $a_{\mathrm{r}}^{(1)} > a_{\mathrm{r}}^{(0)}$ holds.

Inductive Step ($\kappa \geq 1$): Assuming that $a_{\mathrm{t}}^{(0)} \leq \cdots \leq a_{\mathrm{t}}^{(\kappa)}$ and $a_{\mathrm{r}}^{(0)} \leq \cdots \leq a_{\mathrm{r}}^{(\kappa)}$, we need to prove that $a_{\mathrm{t}}^{(\kappa)} \leq a_{\mathrm{t}}^{(\kappa+1)}$ and $a_{\mathrm{r}}^{(\kappa)} \leq a_{\mathrm{r}}^{(\kappa+1)}$. We consider two cases.

Case 1: $a_{\mathrm{t}}^{(\kappa)} \leq a_{\mathrm{r}}^{(\kappa)}$. First, we show that $a_{\mathrm{t}}^{(\kappa+1)} \geq a_{\mathrm{t}}^{(\kappa)}$. Assume by contradiction that $a_{\mathrm{t}}^{(\kappa+1)} < a_{\mathrm{t}}^{(\kappa)} \leq a_{\mathrm{r}}^{(\kappa)}$, and $a_{\mathrm{t}}^{(\kappa+1)}$ is achieved by $\mathbf{w}^{(\kappa+1)}$. Since $F_{\mathrm{t}}^{(\kappa+1)}(a_{\mathrm{t}})$ is monotonically increasing for $a_{\mathrm{t}} \in (0, \psi_{\mathrm{t}}^{(\kappa+1)}]$, it must hold that $F_{\mathrm{t}}^{(\kappa+1)}(a_{\mathrm{t}}^{(\kappa+1)}) < F_{\mathrm{t}}^{(\kappa+1)}(a_{\mathrm{t}}^{(\kappa)})$, contradicting Lemma 6.1.

Second, we show that $a_{\mathrm{r}}^{(\kappa+1)} \geq a_{\mathrm{r}}^{(\kappa)}$. Assume by contradiction that $a_{\mathrm{r}}^{(\kappa+1)} < a_{\mathrm{r}}^{(\kappa)}$. Recall that $F_{\mathrm{r}}^{(\kappa+1)}(a_{\mathrm{r}})$ has an axis of symmetry at $\psi_{\mathrm{r}}^{(\kappa+1)} = a_{\mathrm{t}}^{(\kappa+1)} + \frac{1}{2\alpha_{\kappa+1}}$, and is monotonically increasing for $a_{\mathrm{r}} \in (0, \psi_{\mathrm{r}}^{(\kappa+1)}]$. We consider the following cases based on the relationships between $a_{\mathrm{r}}^{(\kappa)}$, $a_{\mathrm{t}}^{(\kappa+1)}$, and $a_{\mathrm{r}}^{(\kappa+1)}$, as depicted in Fig. 6.16.

Case (i): If $a_{\mathrm{r}}^{(\kappa)} \leq \psi_{\mathrm{r}}^{(\kappa+1)}$ (see Fig. 6.16(a)).

(i) Assume by contradiction that $a_{\mathrm{r}}^{(\kappa+1)} < a_{\mathrm{r}}^{(\kappa)} \leq \psi_{\mathrm{r}}^{(\kappa+1)}$, it holds that $F_{\mathrm{r}}^{(\kappa+1)}(a_{\mathrm{r}}^{(\kappa+1)}) < F_{\mathrm{r}}^{(\kappa+1)}(a_{\mathrm{r}}^{(\kappa)})$, which contradicts Lemma 6.1.

Case (ii)–(iv): If $g_{\mathrm{r}}^{(\kappa)} > \psi_{\mathrm{r}}^{(\kappa+1)}$ (see Fig. 6.16(b)).

(ii) If $a_{\mathrm{r}}^{(\kappa+1)} < 2\psi_{\mathrm{r}}^{(\kappa+1)} - a_{\mathrm{r}}^{(\kappa)} < \psi_{\mathrm{r}}^{(\kappa+1)} < a_{\mathrm{r}}^{(\kappa)}$, it is easy to see that one can select $\widetilde{\mathbf{v}}^{(\kappa+1)} = \mathbf{v}^{(\kappa)}$ that satisfies all the constraints and yields a higher value of the objective function $F_{\mathrm{r}}^{(\kappa+1)}(a_{\mathrm{r}})$.

(iii) If $2\psi_{\mathrm{r}}^{(\kappa+1)} - a_{\mathrm{r}}^{(\kappa)} \leq a_{\mathrm{r}}^{(\kappa+1)} < \psi_{\mathrm{r}}^{(\kappa+1)} < a_{\mathrm{r}}^{(\kappa)}$, there exist a real number $\xi \in (0,1)$ such that
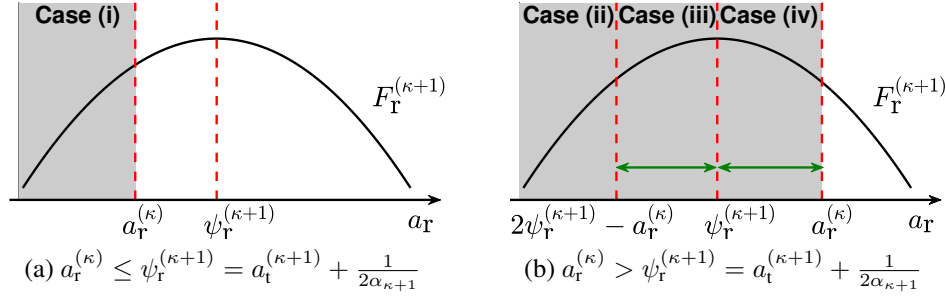
Figure 6.16: Illustration of Case 1 in the proof, where $a_{\mathrm{r}}^{(\kappa+1)} < a_{\mathrm{r}}^{(\kappa)}$ (shaded area) is assumed by contradiction.

$\xi a_{\mathrm{r}}^{(\kappa+1)} + (1-\xi)a_{\mathrm{r}}^{(\kappa)} = \psi_{\mathrm{r}}^{(\kappa+1)}$. Let $\widetilde{\mathbf{v}}^{(\kappa+1)} = \xi\mathbf{v}^{(\kappa+1)} + (1-\xi)\mathbf{v}^{(\kappa)}$. It holds that

$$\widetilde{a}_{\mathrm{r}}^{(\kappa+1)} = \xi a_{\mathrm{r}}^{(\kappa+1)} + (1-\xi)a_{\mathrm{r}}^{(\kappa)} = \psi_{\mathrm{r}}^{(\kappa+1)}, \ F_{\mathrm{r}}^{(\kappa+1)}(\widetilde{a}_{\mathrm{r}}^{(\kappa+1)}) > F_{\mathrm{r}}^{(\kappa+1)}(a_{\mathrm{r}}^{(\kappa+1)}),$$

$$|\widetilde{v}_{n}^{(\kappa+1)}|^2 = |\xi v_{n}^{(\kappa+1)} + (1-\xi)v_{n}^{(\kappa)}|^2 \le \xi^2 + (1-\xi)^2 + 2\xi(1-\xi) = 1,$$

$$(\widetilde{\mathbf{v}}^{(\kappa+1)})^{\dagger}\mathbf{H}_{\mathbf{w}^{(\kappa+1)}}(f_k)\widetilde{\mathbf{v}}^{(\kappa+1)} \cdot \frac{P_{\mathrm{t}}}{N} < \left[\xi \cdot (\mathbf{v}^{(\kappa+1)})^{\dagger}\mathbf{H}_{\mathbf{w}^{(\kappa+1)}}(f_k)\mathbf{v}^{(\kappa+1)}\right.$$

$$\left. + (1-\xi)\cdot(\mathbf{v}^{(\kappa)})^{\dagger}\mathbf{H}_{\mathbf{w}^{(\kappa+1)}}(f_k)\mathbf{v}^{(\kappa)}\right] \cdot \frac{P_{\mathrm{t}}}{N} < N\beta, \ \forall k,$$

where the last inequality comes from the fact that $\mathbf{H}_{\mathbf{w}^{(\kappa+1)}}(f_k)$ is Hermitian and positive semidefinite, and thus $(\mathbf{x}^{\dagger}\mathbf{H}_{\mathbf{w}^{(\kappa+1)}}(f_k)\mathbf{x})$ is convex with respect to $\mathbf{x}$.

(iv) If $\psi_{\mathrm{r}}^{(\kappa+1)} < a_{\mathrm{r}}^{(\kappa+1)} < a_{\mathrm{r}}^{(\kappa)}$, let $\widetilde{\mathbf{v}}^{(\kappa+1)} = \xi\mathbf{v}^{(\kappa+1)}$, where $\xi = \frac{\psi_{\mathrm{r}}^{(\kappa+1)}}{a_{\mathrm{r}}^{(\kappa+1)}} < 1$. Since $\widetilde{\mathbf{v}}^{(\kappa+1)}$ is linearly scaled down from $\mathbf{v}^{(\kappa+1)}$ by a factor of ξ, we have

$$\widetilde{a}_{\mathrm{r}}^{(\kappa+1)} = \xi a_{\mathrm{r}}^{(\kappa+1)} < a_{\mathrm{r}}^{(\kappa+1)}, \ F_{\mathrm{r}}^{(\kappa+1)}(\widetilde{a}_{\mathrm{r}}^{(\kappa+1)}) > F_{\mathrm{r}}^{(\kappa+1)}(a_{\mathrm{r}}^{(\kappa+1)}), \ |\widetilde{v}_{n}^{(\kappa+1)}|^2 = \xi^2 \cdot |v_{n}^{(\kappa+1)}|^2 < 1, \ \forall n,$$

$$(\widetilde{\mathbf{v}}^{(\kappa+1)})^{\dagger}\mathbf{H}_{\mathbf{w}^{(\kappa+1)}}(f_k)\widetilde{\mathbf{v}}^{(\kappa+1)} \cdot \frac{P_{\mathrm{t}}}{N} = \xi^2 \cdot (\mathbf{v}^{(\kappa+1)})^{\dagger}\mathbf{H}_{\mathbf{w}^{(\kappa+1)}}(f_k)\mathbf{v}^{(\kappa+1)} \cdot \frac{P_{\mathrm{t}}}{N} < N\beta, \ \forall k.$$

This contradicts the fact that $\mathbf{v}^{(\kappa+1)}$ is the optimal solution since $\widetilde{\mathbf{v}}^{(\kappa+1)}$ yields $F_{\mathrm{r}}^{(\kappa+1)}(\widetilde{a}_{\mathrm{r}}^{(\kappa+1)}) > F_{\mathrm{r}}^{(\kappa+1)}(a_{\mathrm{r}}^{(\kappa+1)})$ while satisfying constraints in (Opt-Rx).

Now the only marginal case left is $a_{\mathrm{r}}^{(\kappa+1)} = \psi_{\mathrm{r}}^{(\kappa+1)}$. Note that

$$a_{\mathrm{t}}^{(\kappa)} \le a_{\mathrm{t}}^{(\kappa+1)}, \text{ and } a_{\mathrm{r}}^{(\kappa+1)} = \psi_{\mathrm{r}}^{(\kappa+1)} = a_{\mathrm{t}}^{(\kappa+1)} + \frac{1}{2\alpha_{\kappa+1}} < a_{\mathrm{r}}^{(\kappa)}.$$

Since $a_{\mathrm{t}}^{(\kappa+1)} \ge a_{\mathrm{t}}^{(\kappa)}$ and $\alpha_{\kappa+1} \le \alpha_{\kappa}$, we have $\psi_{\mathrm{r}}^{(\kappa+1)} \ge \psi_{\mathrm{r}}^{(\kappa)}$. Therefore, one can select $\widetilde{\mathbf{v}}^{(\kappa)} =$

152

$\frac{a_r^{(\kappa+1)}}{a_r^{(\kappa)}} \cdot \mathbf{v}^{(\kappa)}$, which yields a higher value of the objective function $F_r^{(\kappa)}(a_r)$ since it is decreasing for $a_r \in [a_r^{(\kappa+1)}, a_r^{(\kappa)})$. This contradicts the fact that $\mathbf{v}^{(\kappa)}$ is optimal in the κth iteration and the proof of **Case 1** completes.

Case 2: $a_t^{(\kappa)} > a_r^{(\kappa)}$. The proof is similar to **Case 1** and thus is omitted here. $\qquad\square$

Chapter 7: Hybrid Scheduling in Heterogeneous Half- and Full-Duplex Wireless Networks

In Chapters 5 and 6, we presented our results on enabling full-duplex (FD) in single- and multi-antenna systems. In this chapter, we consider infrastructure-based random-access networks (e.g., IEEE 802.11) consisting of an FD access point (AP) and both half-duplex (HD) and FD users in a single collision domain, and focus on the design and performance evaluation of scheduling algorithms for such heterogeneous HD-FD networks.

There are three main approaches to wireless scheduling algorithms that can guarantee maximum throughput: (i) Maximum Weight Scheduling (MWS) [161], (ii) Greedy Maximal Scheduling (GMS) [162], and (iii) Queue-Length-based Random-Access Algorithms (e.g., Q-CSMA) [163, 164]. In this chapter, we show that a combination of the two latter approaches guarantees maximum throughput and provides good delay performance in heterogeneous HD-FD networks. We first introduce the network model and preliminaries (Section 7.1). Then, we show by using the notion of Local Pooling [162, 165] that GMS is throughput-optimal in the considered HD-FD networks. However, since GMS is fully centralized, we leverage ideas from distributed Q-CSMA to develop the *Hybrid*-GMS (H-GMS) algorithm that combines centralized GMS with distributed Q-CSMA (Section 7.2). The main feature of the proposed H-GMS algorithm is that instead of approximating MWS (as done in "traditional" Q-CSMA), it approximates GMS.

The design of H-GMS leverages the fact that in infrastructure-based networks, the AP has access to all the downlink (DL) queues and can resolve the contention among the DL queues (e.g., using longest-queue-first). In contrast, the users do not have access to any DL queues or to any other uplink (UL) queues, and therefore, must share the medium in a distributed manner, while ensuring FD operation when possible. We prove the throughput optimality of H-GMS (namely, it

can support any rate vector in the capacity region of heterogeneous HD-FD networks) by using the fluid limit technique (Section 7.3). The results are general, and, unlike most existing work, apply to a wide range of weight functions. We also present variants of H-GMS with different degrees of centralization and derive two lower bounds on the average queue length (Section 7.4). These lower bounds serve as benchmarks when evaluating the delay performance of H-GMS.

We demonstrate the benefits of introducing FD-capable users into an all-HD network in terms of both network and individual throughput gains (Section 7.5). For example, when all links have equal arrival rate, the throughput gain of the HD-FD network over the all-HD network increases with the number of FD users, and it reaches a gain of 2 when all users are FD-capable. We also demonstrate that it is generally possible for all users to experience improved individual throughput at the cost of lowering the priority of FD users, revealing an interesting *fairness-efficiency tradeoff*. Finally, we present extensive simulation results to evaluate the different variants of the H-GMS algorithm and compare them to the classical Q-CSMA algorithm (Section 7.6). The results show H-GMS achieves 16–$30\times$ better delay performance and improves fairness between HD and FD users by up to 50% compared to the fully distributed Q-CSMA algorithm.

This chapter's contributions were previously presented in [29, 30, 166] and the work is in collaboration with Ph.D. student Jelena Diakonikolas.

7.1 Network Model and Preliminaries

7.1.1 Network Model

We consider a single-channel, *heterogeneous* wireless network consisting of one AP and N users, with a UL and a DL between each user and the AP. The set of users is denoted by $\mathcal{N}$. The AP is FD, while N_F of the users are FD and $N_H = N - N_F$ are HD. Without loss of generality, we index the users by $[N] = \{1, 2, \cdots, N\}$ where the first N_F indices correspond to FD users and the remaining N_H indices correspond to HD users. The sets of FD and HD users are denoted by $\mathcal{N}_F$ and $\mathcal{N}_H$, respectively. We consider a collocated network where the users are within the communication range of each other and the AP. The network can be represented by a directed star

graph $G = (\mathcal{V}, \mathcal{E})$ with the AP at the center and two links between AP and each user in both directions. Thus, we have $\mathcal{V} = \{AP\} \cup \mathcal{N}$ (with $|\mathcal{V}| = 1 + N$) and $|\mathcal{E}| = 2N$.

7.1.2 Traffic Model, Schedule, and Queues

We assume that time is slotted and packets arrive at all UL and DL queues according to some independent stochastic process. For brevity, we will use superscript $j \in \{u, d\}$ to denote the UL and DL of a user. Let l_i^j denote link j (UL or DL) of user i, each of which is associated with a queue Q_i^j. We use $A_i^j(t) \leq A_{\max} < \infty$ to denote the number of packets arriving at link j (UL or DL) of user i in slot t. The arrival process is assumed to have a well-defined long-term rate of $\lambda_i^j = \lim_{T \to +\infty} \frac{1}{T} \sum_{t=1}^{T} A_i^j(t)$. Let $\boldsymbol{\lambda} = [\lambda_i^u, \lambda_i^d]_{i=1}^N$ be the arrival rate vector on the ULs and DLs.

All the links are assumed to have capacity of one packet per time slot and the SIC at all the FD-capable nodes is *perfect*.[1] A *schedule* at any time slot t is represented by a vector

$$\mathbf{X}(t) = [X_1^u(t), X_1^d(t), \cdots, X_N^u(t), X_N^d(t)] \in \{0, 1\}^{2N},$$

where $X_i^u(t)$ (resp. $X_i^d(t)$) is equal to 1 if the UL (resp. DL) of user i is scheduled to transmit a packet in time slot t and $X_i^u = 0$ (resp. $X_i^d = 0$), otherwise. We denote the set of all feasible schedules by $\mathcal{S}$. Let $\mathbf{e}_i \in \{0, 1\}^{2N}$ be the ith basis vector (i.e., an all-zero vector except the ith element being one). Since a pair of UL and DL of the same FD user can be activated at the same time, we have:

$$\mathcal{S} = \{\mathbf{0}\} \cup \{\mathbf{e}_{2i-1}, \mathbf{e}_{2i}, \forall i \in \mathcal{N}\} \cup \{\mathbf{e}_{2i-1} + \mathbf{e}_{2i}, \forall i \in \mathcal{N}_F\}.$$

Choosing $\mathbf{X}(t) \in \mathcal{S}$, the queue dynamics are described by:

$$Q_i^j(t) = [Q_i^j(t-1) + A_i^j(t) - X_i^j(t)]^+, \ \forall t \geq 1,$$

where $[\cdot]^+ = \max(0, \cdot)$. $\mathbf{Q}(t) = [Q_i^u(t), Q_i^d(t)]_{i=1}^N$ denotes the queue vector, and $\mathbf{1}(\cdot)$ denotes the indicator function.

---

[1] We remark that imperfect SIC can also be incorporated into the model by letting the corresponding link capacity be $c_i^j \in (0, 1)$. For simplicity and analytical tractability, we assume $c_i^j = 1$, $\forall i \in \mathcal{N}$, throughout this chapter.

7.1.3 Capacity Region and Throughput Optimality

The capacity region of the network is defined as the set of all arrival rate vectors for which there exists a scheduling algorithm that can stabilize the queues. It is known that, in general, the capacity region is the convex hull of all feasible schedules [161]. Therefore, the capacity region of the heterogeneous HD-FD network is given by $\Lambda_{\text{HD-FD}} = \text{Co}(\mathcal{S})$, where $\text{Co}(\cdot)$ is the convex hull operator. It is easy to see that this capacity region can be equivalently characterized by the following set of linear constraints[2]:

$$\Lambda_{\text{HD-FD}} = \left\{ \boldsymbol{\lambda} \in [0,1]^{|\mathcal{E}|} : \sum_{i \in \mathcal{N}_F} \max\{\lambda_i^{\text{u}}, \lambda_i^{\text{d}}\} + \sum_{i \in \mathcal{N}_H} (\lambda_i^{\text{u}} + \lambda_i^{\text{d}}) \leq 1 \right\}. \qquad (7.1)$$

Let a network in which all the users and the AP are only HD-capable be the *benchmark all-HD network*, whose capacity region is given by $\Lambda_{\text{HD}} = \text{Co}(\mathbf{e}_1, \cdots, \mathbf{e}_{2N})$, or equivalently

$$\Lambda_{\text{HD}} = \left\{ \boldsymbol{\lambda} \in [0,1]^{|\mathcal{E}|} : \sum_{i \in \mathcal{N}} (\lambda_i^{\text{u}} + \lambda_i^{\text{d}}) \leq 1 \right\}. \qquad (7.2)$$

A scheduling algorithm is called *throughput-optimal* if it can keep the network queues stable for all arrival rate vectors $\boldsymbol{\lambda} \in \text{int}(\Lambda)$, where $\text{int}(\Lambda)$ denotes the interior of Λ.

To compare $\Lambda_{\text{HD-FD}}$ with Λ_{HD} and quantify the network throughput gain when a certain number of HD users become FD-capable, similar to [167], we define the *capacity region expansion function* $\gamma(\cdot)$ as follows. Given $\boldsymbol{\lambda}_0$ on the Pareto boundary of Λ_{HD}, the capacity region expansion function at point $\boldsymbol{\lambda}_0$, denoted by $\gamma(\boldsymbol{\lambda}_0)$, is defined as

$$\gamma(\boldsymbol{\lambda}_0) = \sup\{\zeta > 0 : \zeta \cdot \boldsymbol{\lambda}_0 \in \Lambda_{\text{HD-FD}}\}. \qquad (7.3)$$

$\gamma(\cdot)$ can be interpreted as a function that scales an arrival rate vector on the Pareto boundary of Λ_{HD} to a vector on the Pareto boundary of $\Lambda_{\text{HD-FD}}$, as N_F users become FD-capable. It is not hard to see that $\gamma : \Lambda_{\text{HD}} \to [1, 2]$.

---

[2]It is straightforward to only use linear inequalities, by replacing $\max\{\lambda_i^{\text{u}}, \lambda_i^{\text{d}}\}$ with λ_i and adding linear inequalities $\lambda_i^{\text{u}} \leq \lambda_i, \lambda_i^{\text{d}} \leq \lambda_i$.

7.2 Scheduling Algorithms and Main Result

There are three main approaches to wireless scheduling algorithms that can guarantee maximum throughput:

Maximum Weight Scheduling (MWS) [161], which relies on the queue length information and schedules non-conflicting links with the maximum total queue length. In contrast to the all-HD networks where only a single link can be scheduled at a time, in the considered setting the UL and DL of any FD user can be scheduled simultaneously. Thus, to implement MWS, queue length information needs to be shared between each FD user and the AP, which requires significant overhead.

Greedy Maximal Scheduling (GMS) [162], which is a centralized policy that greedily selects the link with the longest queue, disregards all conflicting links, and repeats the process. Typically, GMS has better delay performance than MWS and Q-CSMA. Although GMS is equivalent to MWS in an all-HD network, in general, it is not equivalent to MWS and thus, is not throughput-optimal in general topologies.

Queue-Length-based Random-Access Algorithms (e.g., Q-CSMA) [163, 164], which are fully distributed and do not require sharing of the queue length information between the users and the AP. These algorithms have been shown to be throughput-optimal. However, they generally experience poor delay performance and suffer from excessive queue lengths.

In this section, we first briefly introduce MWS in the considered HD-FD networks. We then use Local Pooling [162, 165] to prove that GMS is throughput-optimal in the considered networks, and therefore, MWS [161] is unneeded. Based on that, we present the hybrid scheduling algorithm, H-GMS – a decentralized version of GMS that leverages ideas from distributed Q-CSMA [163, 164]. H-GMS is tailored for heterogeneous HD-FD networks and uses information about the DL queues that is available at the AP, but does not require global information about the UL queues. We state the main result (Theorem 7.1) about the throughput optimality of H-GMS and describe its various implementations with different levels of centralization. We later show (in Section 7.6)

Algorithm 7.1 MWS for HD-FD Networks (in slot t)

---

1. Initialize $\mathbf{X}(t) = \mathbf{0}$.

2. Let $\hat{Q}_i(t) = Q_i^{\mathrm{u}}(t) + Q_i^{\mathrm{d}}(t)$, $\forall i \in \mathcal{N}_F$. Select user $i^\star$ by (break ties uniformly at random)

$$i^\star = \arg\max \left\{ \max_{i \in \mathcal{N}_F} \{\hat{Q}_i(t)\}, \max_{i \in \mathcal{N}_H} \{Q_i^{\mathrm{u}}(t), Q_i^{\mathrm{d}}(t)\} \right\}.$$

3. • If $i^\star \in \mathcal{N}_F$, set $X_i^{\mathrm{u}}(t) = X_i^{\mathrm{d}}(t) = 1$;
   • If $i^\star \in \mathcal{N}_H$, set $X_i^j(t) = 1$ if $Q_i^j(t) \geq Q_i^{\{\mathrm{u,d}\}\backslash\{j\}}(t)$. Break ties uniformly at random.

4. Use $\mathbf{X}(t)$ as the transmission schedule in slot t.

Algorithm 7.2 GMS for HD-FD Networks (in slot t)

---

1. Initialize $\mathbf{X}(t) = \mathbf{0}$.

2. Select link $l^\star \in \mathcal{E}$ with the largest queue length (i.e., $l^\star = \arg\max_{i \in \mathcal{N}, \, j \in \{\mathrm{u,d}\}} \{Q_i^j(t)\}$). If the longest queue is not unique, break ties uniformly at random.

3. • If $l^\star = l_i^{\mathrm{u}}$ or l_i^{d} for some $i \in \mathcal{N}_F$, set $X_i^{\mathrm{u}}(t) = X_i^{\mathrm{d}}(t) = 1$;
   • If $l^\star = l_i^j$ for some $i \in \mathcal{N}_H$ and $j \in \{\mathrm{u,d}\}$, set $X_i^j(t) = 1$.

4. Use $\mathbf{X}(t)$ as the transmission schedule in slot t.

---

that these variants of H-GMS have different delay performance.

7.2.1   Centralized Max-Weight Scheduling (MWS)

We first describe the throughput-optimal MWS (see Algorithm 7.1), where in each time slot t, the schedule $\mathbf{X}(t) \in \mathcal{S}$ with the maximum sum queue length is selected.

7.2.2   Centralized Greedy Maximal Scheduling (GMS)

We now show that a GMS (see Algorithm 7.2) is throughput-optimal in *any* collocated heterogeneous HD-FD network, independent of the values of N_F and N_H. In both MWS and GMS, a pair of FD UL and DL is always scheduled at the same time, as such a schedule yields a higher throughput than scheduling only the UL or only the DL.

Proposition 7.1 *The Greedy Maximal Scheduling (GMS) algorithm is throughput-optimal in any collocated heterogeneous HD-FD network.*

The proof (see Appendix 7.8.1) is based on [162, Theorem 1], [165], and the fact that the interference graph of *any* collocated heterogeneous HD-FD network satisfies the Overall Local Pooling (OLoP) conditions, which guarantee that GMS is throughput-optimal.

7.2.3 Hybrid-GMS (H-GMS) Algorithm

We now present a hybrid scheduling algorithm, H-GMS, which combines the concepts of GMS and Q-CSMA [163, 164]. Instead of approximating MWS [161] in a decentralized manner (as in traditional Q-CSMA), H-GMS approximates GMS, which is easier to decentralize in the considered HD-FD networks. H-GMS leverages the existence of an AP to resolve the contention among the DL queues, since the AP has explicit information about these queues and can select one of them (e.g., the longest queue). Thus, effectively at most one DL queue needs to perform Q-CSMA in each time slot. On the other hand, since users are unaware of the UL and DL queue states of other users and at the AP, every user needs to perform Q-CSMA in order to share the channel distributedly. Therefore, the number of possible participants under H-GMS in each slot is at most $(N + 1)$. Moreover, we show that this hybrid approach yields much better delay performance than Q-CSMA while still achieving throughput optimality.

Algorithm 7.3 presents the pseudocode for H-GMS, which operates as follows. Each slot t is divided into a short control slot and a data slot. The control slot contains only two control mini-slots (independent of the number of users, N). We refer to the first mini-slot as the *initiation mini-slot* and to the second one as the *coordination mini-slot*. H-GMS has three steps: (1) Initiation, (2) Coordination, and (3) Data transmission, as explained below.

(1) Initiation. By the end of slot $(t - 1)$, the AP knows $\mathbf{X}(t - 1)$ since every packet transmission has to be sent from or received by the AP. If $\mathbf{X}(t - 1) = \mathbf{0}$ (i.e., idle channel), then the AP starts an initiation in slot t using the initiation mini-slot as follows. First, the AP centrally finds the index of the user with the longest DL queue, i.e., $i^\star(t) = \arg\max_{i \in \mathcal{N}} Q_i^{\mathrm{d}}(t)$. If multiple DLs

Algorithm 7.3 H-GMS Algorithm (in slot t)

$-$ If $\mathbf{X}(t-1) = \mathbf{0}$:

1. In the initiation mini-slot, the AP computes $i^\star = \arg\max_{i \in \mathcal{N}} Q_i^{\mathrm{d}}(t)$. If multiple DL queues have the same length, break ties according to some deterministic rule. The AP chooses an initiator link $\mathrm{IL}(t)$ from $\mathcal{L}(t) = \{l_1^{\mathrm{u}}, \cdots, l_N^{\mathrm{u}}, l_{i^\star}^{\mathrm{d}}\}$ according to an access probability distribution $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_N, \alpha_{\mathrm{AP}}]$.

2. If $\mathrm{IL}(t) = l_{i^\star}^{\mathrm{d}}$, the AP sets:

   - $X_{i^\star}^{\mathrm{d}}(t) = 1$ with probability $p_{i^\star}^{\mathrm{d}}(t)$, or $X_{i^\star}^{\mathrm{d}}(t) = 0$ with probability $\bar{p}_{i^\star}^{\mathrm{d}}(t) = 1 - p_{i^\star}^{\mathrm{d}}(t)$;

   - In the coordination mini-slot, AP broadcasts a control packet containing the information of $\mathrm{IL}(t)$ and user $i^\star$ sets $X_{i^\star}^{\mathrm{u}}(t) = X_{i^\star}^{\mathrm{d}}(t) \cdot \mathbf{1}(i^\star \in \mathcal{N}_F)$;

3. If $\mathrm{IL}(t) = l_i^{\mathrm{u}}$ for some $i \in \mathcal{N}$, in the coordination mini-slot, the AP broadcasts the information of $\mathrm{IL}(t)$ and user i sets:

   - $X_i^{\mathrm{u}}(t) = 1$ with probability $p_i^{\mathrm{u}}(t)$, or $X_i^{\mathrm{u}}(t) = 0$ with probability $\bar{p}_i^{\mathrm{u}}(t) = 1 - p_i^{\mathrm{u}}(t)$;

   - In the same coordination mini-slot, user i sends a control packet containing this information to the AP if $i \in \mathcal{N}_F$, and AP sets $X_i^{\mathrm{d}}(t) = X_i^{\mathrm{u}}(t)$;

4. At the beginning of the data slot,

   - AP activates DL i if $X_i^{\mathrm{d}}(t) = 1$;

   - User i activates it UL if $X_i^{\mathrm{u}}(t) = 1$;

$-$ If $\mathbf{X}(t-1) \neq \mathbf{0}$, set $\mathrm{IL}(t) = \mathrm{IL}(t-1)$. Repeat Steps 2–4.

have equal (largest) queue length, it breaks ties according to some deterministic rule. Then, the AP randomly selects an initiator link $\mathrm{IL}(t)$ from the set $\mathcal{L}(t) = \{l_1^{\mathrm{u}}, \cdots, l_N^{\mathrm{u}}, l_{i^\star}^{\mathrm{d}}\}$ according to an *access probability* distribution $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_N, \alpha_{\mathrm{AP}}]$ satisfying: (i) $\alpha_i > 0, \forall i \in \mathcal{N}$, and $\alpha_{\mathrm{AP}} > 0$, and (ii) $\alpha_{\mathrm{AP}} = 1 - \sum_{i=1}^N \alpha_i$. We refer to α_i and α_{AP} as the access probability for user i and the AP, respectively. Therefore,

$$\mathrm{IL}(t) = \begin{cases} l_i^{\mathrm{u}}, & \text{with probability } \alpha_i, \ \forall i \in \mathcal{N}, \\ l_{i^\star}^{\mathrm{d}}, & \text{with probability } \alpha_{\mathrm{AP}}, \end{cases} \tag{7.4}$$

i.e., $\mathrm{IL}(t)$ is either a UL or the DL with the longest queue. If $\mathbf{X}(t-1) \neq \mathbf{0}$, set $\mathrm{IL}(t) = \mathrm{IL}(t-1)$.

(2) Coordination. In the coordination mini-slot, if the DL of user $i^\star$ is selected as the initiator link ($\mathrm{IL}(t) = l_{i^\star}^{\mathrm{d}}$), the AP sets $X_{i^\star}^{\mathrm{d}}(t) = 1$ with probability $p_{i^\star}^{\mathrm{d}}(t)$. Otherwise, it remains silent. If the AP decides to transmit on DL $l_{i^\star}^{\mathrm{d}}$ (i.e., $X_{i^\star}^{\mathrm{d}}(t) = 1$), it broadcasts a control packet containing the

information of $\mathrm{IL}(t)$ and user $i^\star$ sets $X_{i^\star}^{\mathrm{u}}(t) = 1$ if and only if $i^\star \in \mathcal{N}_F$.

If the UL of user i is selected as the initiator link ($\mathrm{IL}(t) = l_i^{\mathrm{u}}$ for some $i \in \mathcal{N}$), the AP broadcasts the information of $\mathrm{IL}(t)$ and user i sets $X_i^{\mathrm{u}}(t) = 1$ with probability $p_i^{\mathrm{u}}(t)$. Otherwise, user i remains silent. If user i is FD-capable and decides to transmit (i.e., $X_i^{\mathrm{u}}(t) = 1$), it sends a control packet containing this information to the AP and the AP sets $X_i^{\mathrm{d}}(t) = 1$.[3] Note that the real-time information of *all* the UL queue lengths is *not* shared or available at the AP.

The transmission probability of the link is selected depending on its queue size $Q_i^j(t)$ at the beginning of slot t. Specifically, similar to [163, 164], link l_i^j chooses logistic form

$$p_i^j(t) = \frac{\exp\left(f(Q_i^j(t))\right)}{1 + \exp\left(f(Q_i^j(t))\right)}, \ \forall i \in \mathcal{N}, \ \forall j \in \{\mathrm{u}, \mathrm{d}\}, \tag{7.5}$$

where $f(\cdot)$ is a positive increasing function (to be specified later), called the *weight function*. Further, if an FD initiator UL (or DL) decides to stop transmitting (after packet transmission in the last slot), it again sends a short coordination message which stops further packet transmissions at the DL (or UL) or the same FD user.

(3) Data transmission. After steps (1)–(2), if either a pair of FD UL and DL or an HD link (UL or DL) is activated, a packet is sent on the links in the data slot. The initiator link then starts a new coordination in the subsequent control slot which either leads to more packet transmissions or stops further packet transmissions at the links involved in the schedule.

Remark: The initiation step in H-GMS is described as a polling mechanism where the AP draws a link $\mathrm{IL}(t)$ from $\mathcal{L}(t)$ according to the access probability distribution $\boldsymbol{\alpha}$. Alternatively, the initiation step can be described in a distributed fashion using an extra mini-slot as follows: user i sends a short initiation message with probability α_i. If AP receives the message, it sends back a clear-to-initiate message and sets $\mathrm{IL}(t) = l_i^{\mathrm{u}}$, otherwise (i.e., in case of collision or idleness) $l_{i^\star}^{\mathrm{d}}$ is selected as the initiator link by the AP. This effectively emulates polling user i with probability $\widetilde{\alpha}_i = \alpha_i \prod_{i' \neq i}(1 - \alpha_{i'})$ and AP with probability $\widetilde{\alpha}_{\mathrm{AP}} = 1 - \sum_{i=1}^{N} \widetilde{\alpha}_i$.

---

[3]This operation can be done in the same coordination mini-slot since FD user i can simultaneously receive the control packet ($\mathrm{IL}(t) = l_i^{\mathrm{u}}$) from the AP and send its control packet ($X_i^{\mathrm{u}}(t) = 1$) back to the AP.

7.2.4 Main Result: Throughput Optimality of H-GMS

The system state under H-GMS evolves as a Markov chain $(\mathbf{X}(t), \mathbf{Q}(t))$. The following theorem states our main result regarding the positive recurrence of this Markov chain (throughput optimality of H-GMS).

Theorem 7.1 *For any arrival rate vector* $\boldsymbol{\lambda} \in \mathrm{int}(\Lambda_{HD\text{-}FD})$*, the system Markov chain* $(\mathbf{X}(t), \mathbf{Q}(t))$ *is positive recurrent under H-GMS (Algorithm 7.3). The weight function* $f(\cdot)$ *in* (7.5) *can be any nonnegative increasing function such that* $\lim_{x \to \infty} f(x)/\log x < 1$*, or* $\lim_{x \to \infty} f(x)/\log x > 1$ *(including* $f(x) = x^\beta$*,* $\beta > 0$*).*

Establishing Theorem 7.1 is not trivial due to the coupling between $\mathbf{X}(t)$ and $\mathbf{Q}(t)$: The dynamics of the schedule process $\mathbf{X}(t)$ is governed by the queue process $\mathbf{Q}(t)$, while at the same time, the dynamics of $\mathbf{Q}(t)$ depends on $\mathbf{X}(t)$. Depending on the functional shape of the weight function $f(\cdot)$, this coupling gives rise to vastly different behaviors for the Markov chain $(\mathbf{X}(t), \mathbf{Q}(t))$. For functions $f(\cdot)$ that grow slower than $\log(\cdot)$, the convergence of the schedule process $\mathbf{X}(t)$ occurs on a much faster time-scale ("fast mixing") compared to the time-scale of changes in the queue process $\mathbf{Q}(t)$. For more aggressive functions $f(\cdot)$, the convergence of $\mathbf{X}(t)$ occurs on a much slower time-scale ("slow mixing") compared to the time-scale of changes in $\mathbf{Q}(t)$. Nevertheless, Theorem 7.1 states that the system Markov chain is stable (positive recurrent) for a wide range of weight functions. We provide a proof of Theorem 7.1 in Section 7.3 based on the analysis of the fluid limits of the system under the H-GMS algorithm.

7.2.5 Variants of the H-GMS Algorithm

In this subsection, we introduce three variants of the H-GMS algorithm, which differ only in Step 1 of Algorithm 7.3.

- H-GMS (Algorithm 7.3): The AP selects the *longest DL*.

- H-GMS-R: The AP selects a DL *uniformly at random*, i.e., $i^\star \sim \mathrm{Unif}(1, \cdots, N)$ (in step 1 of Algorithm 7.3).

- H-GMS-E: Exactly the same as H-GMS except for the access probability being set according to the following:

$$\widetilde{\alpha}_i \propto \max\left\{\widetilde{Q}_i^{\mathrm{u}}/(\sum_{i'=1}^{N}\widetilde{Q}_{i'}^{\mathrm{u}} + Q_{i^\star}^{\mathrm{d}}), \; \alpha_{\mathrm{th}}\right\}, \; \forall i \in \mathcal{N},$$

$$\widetilde{\alpha}_{\mathrm{AP}} \propto \max\left\{Q_{i^\star}^{\mathrm{d}}/(\sum_{i'=1}^{N}\widetilde{Q}_{i'}^{\mathrm{u}} + Q_{i^\star}^{\mathrm{d}}), \; \alpha_{\mathrm{th}}\right\},$$

where $\widetilde{Q}_i^{\mathrm{u}}$ an estimate of UL queue length of user i. Specifically, when a user transmits on the UL, it includes its queue length in the packets and the AP updates $\widetilde{Q}_i^{\mathrm{u}}$ using this information contained in the last (i.e., most recently) received packet from user i on the UL. Then, $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_N, \alpha_{\mathrm{AP}}]$ is obtained after normalization, i.e.,

$$\alpha_i = \frac{\widetilde{\alpha}_i}{\sum_{i'=1}^{N}\widetilde{\alpha}_{i'} + \widetilde{\alpha}_{\mathrm{AP}}}, \; \forall i \in \mathcal{N}, \; \alpha_{\mathrm{AP}} = \frac{\widetilde{\alpha}_{\mathrm{AP}}}{\sum_{i'=1}^{N}\widetilde{\alpha}_{i'} + \widetilde{\alpha}_{\mathrm{AP}}}.$$

A minimum access probability $\alpha_{\mathrm{th}} > 0$ has been introduced to ensure that each link is selected with a non-zero probability. Otherwise, an HD UL l_i^{u} ($\forall i \in \mathcal{N}_H$) with a zero queue-length estimate would never be selected by the AP (i.e., $\widetilde{Q}_i^{\mathrm{u}} = 0$ and thus $\widetilde{\alpha}_i = 0$), and the AP would never receive any updated information of $\widetilde{Q}_i^{\mathrm{u}}$ since $\widetilde{\alpha}_i$ would remain zero.

The access probability distribution $\boldsymbol{\alpha}$ is *non-adaptive* in H-GMS and H-GMS-R, and is *adaptive* in H-GMS-E. As we will see in Section 7.6, the adaptive choice of $\boldsymbol{\alpha}$ helps balance the queue lengths between FD and HD users.

7.3 Proof of Theorem 7.1 via Fluid Limits

We prove Theorem 7.1 based on the analysis of the fluid limits of the system under H-GMS (Algorithm 7.3). The proof has three parts: (i) existence of the fluid limits (Lemma 7.1), (ii) deriving the fluid limit equations for the various choices of $f(\cdot)$ (Lemma 7.3), and (iii) proving the stability of the queues in the fluid limits using a Lyapunov method, which implies the stability of the original stochastic process.

Part (i): Definition and Existence of Fluid Limits. Consider a scaled process $\mathbf{Q}^{(r)}(t)$ where

$\mathbf{Q}^{(r)}(t) = \mathbf{Q}(rt)/r$. Note that the queue process $\mathbf{Q}$ is scaled in both time and space by a factor $r > 0$. To avoid technical difficulties, we can simply work with a continuous process by linear interpolation among the values at integer time points. Suppose the scaled process, with $r > 0$, starts from an initial state $\mathbf{Q}^{(r)}(0)$. Any (possibly random) limit $\mathbf{q}(t)$ of the scaled process $\mathbf{Q}^{(r)}(t)$ as $r \to \infty$ is called a fluid limit. The process $\mathbf{Q}^{(r)}(t)$ can be constructed as follows. At any time $t \geq 0$,

$$\mathbf{Q}^{(r)}(t) = \mathbf{Q}^{(r)}(0) + \overline{\mathbf{A}}^{(r)}(t) - \overline{\mathbf{S}}^{(r)}(t), \tag{7.6}$$

where for any user $i \in \mathcal{N}$ with UL or DL $j \in \{\mathrm{u},\mathrm{d}\}$,

$$\overline{A}_i^{j\,(r)}(t) = \frac{1}{r}\sum_{\tau=1}^{rt} A_i^j(\tau), \ \ \overline{S}_i^{j\,(r)}(t) = \frac{1}{r}\sum_{\tau=1}^{rt} X_i^j(\tau) \cdot \mathbf{1}(Q_i^j(\tau) > 0).$$

Similarly, we denote by $\mathbf{a}(t)$ and $\mathbf{s}(t)$ the limits of the scaled processes $\overline{\mathbf{A}}^{(r)}(t)$ and $\overline{\mathbf{S}}^{(r)}(t)$ as $r \to \infty$, respectively. The following lemma shows that the scaled process converges to the fluid limit in a weak convergence sense, in the metric of uniform norm on compact time intervals. It is possible to show a stronger convergence (i.e., almost sure convergence uniformly over compact time intervals) in the case of $\lim_{x \to \infty} f(x)/\log x < 1$; nevertheless, the weak convergence is sufficient for our proofs.

Lemma 7.1 (Existence of Fluid Limits) *Suppose $\mathbf{Q}^{(r)}(0) \to \mathbf{q}(0)$. Then any sequence r has a subsequence such that $(\mathbf{Q}^{(r)}(t), \overline{\mathbf{A}}^{(r)}(t), \overline{\mathbf{S}}^{(r)}(t)) \Rightarrow (\mathbf{q}(t), \mathbf{a}(t), \mathbf{s}(t))$ along the subsequence. The sample paths $(\mathbf{q}(t), \mathbf{a}(t), \mathbf{s}(t))$ are Lipschitz continuous and thus differentiable almost everywhere with probability one.*

Proof 7.1 *The proof is standard and follows from Lipschitz continuity of the scaled process, see, e.g., [168].*

Part (ii): Fluid Limit Equations under H-GMS. Recall that the schedule $\mathbf{X}(t)$ at time t is determined after the Initiation and Coordination steps of Algorithm 7.3. Let $Y(t)$ indicate the initiator link which is activated in slot t. Let $i^\star = \arg\max_{i \in \mathcal{N}} Q_i^{\mathrm{d}}(t)$, then the state space of $Y(t)$ can be

labeled as $S_Y = \{0, 1, \cdots, N, i^\star\}$, where $Y(t) = 0$ means no link is active, $Y(t) = i^\star$ means DL $l_{i^\star}^d$ is active, and $Y(t) = i$, for $i \in \{1, \cdots, N\}$, means UL l_i^u is active. We further use $\{Y^{\mathbf{Q}}(t)\}_{t \geq t_0}$ to denote the dynamics of $Y(t)$, assuming a fixed queue length vector $\mathbf{Q}(t) = \mathbf{Q}(t_0) = \mathbf{Q}$ for all times $t \geq t_0$. Under the H-GMS algorithm, $\{Y^{\mathbf{Q}}(t)\}_{t \geq t_0}$ evolves as an irreducible and aperiodic Markov chain over the state space S_Y. If $Y^{\mathbf{Q}}(t) = i$ for an i which is an initiator UL or DL of an FD user $i \in \mathcal{N}_F$, then the other link of the same FD user will follow the initiator link and become active as well under H-GMS. Due to the activation/deactivation coordination among the initiator link and the follower link, adding the possible follower link does not change the subsequent dynamics of the Markov chain $Y^{\mathbf{Q}}(t)$ under fixed $\mathbf{Q}$.

Let $\mathbf{P}^{\mathbf{Q}} = [P(s, s')]$ be the transition probability matrix of $Y^{\mathbf{Q}}(t)$, where $P(s, s')$ is the transition probability from state $s \in S_Y$ to $s' \in S_Y$. Then, under Algorithm 7.3, we have

$$
\begin{aligned}
P(0, i) &= \alpha_i p_i^u, \ P(i, i) = p_i^u, \ P(i, 0) = \bar{p}_i^u, \ \forall i \in \mathcal{N} \\
P(0, i^\star) &= \alpha_{\text{AP}} p_{i^\star}^d, \ P(i^\star, i^\star) = p_{i^\star}^d, \ P(i^\star, 0) = \bar{p}_{i^\star}^d, \\
P(0, 0) &= 1 - \sum_{i=1}^{N} P(0, i) - P(0, i^\star).
\end{aligned} \tag{7.7}
$$

Lemma 7.2 *The steady-state distribution of Markov chain $Y^{\mathbf{Q}}(t)$ is given by*

$$
\begin{aligned}
\pi^{\mathbf{Q}}(i) &= \alpha_i \exp(f(Q_i^u))/Z, \ i \in S_Y \setminus \{0, i^\star\}; \\
\pi^{\mathbf{Q}}(i^\star) &= \alpha_{AP} \exp(f(Q_{i^\star}^d))/Z, \ \pi^{\mathbf{Q}}(0) = 1/Z,
\end{aligned} \tag{7.8}
$$

where Z is the normalizing constant and $f(\cdot)$ is the weight function from (7.5).

The proof of Lemma 7.2 is in Appendix 7.8.2. The following corollary is immediate as the result of Lemma 7.2 and the fact that $Y(t)$ uniquely determines $\mathbf{X}(t)$ by (possible) activation of both the UL and DL of an FD user in the coordination step.

Corollary 7.1 *Let $\mathbf{f}_i = \mathbf{e}_{2i-1} + \mathbf{e}_{2i}$, $i \in \mathcal{N}_F$, be an FD bi-directional transmission schedule, and $\mathbf{h}_i^u = \mathbf{e}_{2i-1}$ ($\mathbf{h}_i^d = \mathbf{e}_{2i}$), $i \in \mathcal{N}_H$, be an HD UL (DL) transmission schedule. Given a fixed queue*

vector $\mathbf{Q}(t) = \mathbf{Q}$, *in steady state, if* $i^{\star} \in \mathcal{N}_F$,

$$\mathbb{P}\left\{\mathbf{X} = \mathbf{f}_{i^{\star}}\right\} = [\alpha_{AP}\exp(f(Q_{i^{\star}}^d)) + \alpha_{i^{\star}}\exp(f(Q_{i^{\star}}^u))]/Z,$$

$$\mathbb{P}\left\{\mathbf{X} = \mathbf{f}_i\right\} = \alpha_i\exp(f(Q_i^u))/Z, \ \forall i \in \mathcal{N}_F, \ i \neq i^{\star},$$

$$\mathbb{P}\left\{\mathbf{X} = \mathbf{h}_i^u\right\} = \alpha_i\exp(f(Q_i^u))/Z, \ \forall i \in \mathcal{N}_H.$$

Otherwise, if $i^{\star} \in \mathcal{N}_H$,

$$\mathbb{P}\left\{\mathbf{X} = \mathbf{f}_i\right\} = \alpha_i\exp(f(Q_i^u))/Z, \ \forall i \in \mathcal{N}_F,$$

$$\mathbb{P}\left\{\mathbf{X} = \mathbf{h}_i^u\right\} = \alpha_i\exp(f(Q_i^u))/Z, \ \forall i \in \mathcal{N}_H,$$

$$\mathbb{P}\left\{\mathbf{X} = \mathbf{h}_{i^{\star}}^d\right\} = \alpha_{AP}\exp(f(Q_{i^{\star}}^d))/Z,$$

where Z *and* $f(\cdot)$ *are as in Lemma 7.2.*

Consider a fluid sample path under our H-GMS algorithm. Suppose $\mathbf{q}(t) = \mathbf{q} \neq \mathbf{0}$ at a time t. This implies that for r large enough, all the queues with non-zero fluid limit $q_i^j > 0$ are of size $Q_i^j = \mathcal{O}(q_i^j r)$ in the original process, while all the queues with zero fluid limit are of size $Q_i^j = o(r)$ in the original process. Therefore, taking the limit $r \to \infty$ in (7.8), and noting that the weight function $f(\cdot)$ is a positive increasing function of the queue size, it follows that

$$\pi^{\mathbf{Q}}(i) \to 0 \text{ if } q_i^u = 0, \ i \in S_Y \setminus \{0, i^{\star}\},$$

$$\pi^{\mathbf{Q}}(i^{\star}) \to 0 \text{ if } q_{i^{\star}}^d = 0, \ \pi^{\mathbf{Q}}(0) = 0.$$

This shows that a queue with a zero fluid limit *cannot* initiate transmission in steady state. Consequently,

$$\mathbb{P}\left\{\mathbf{X} = \mathbf{f}_i\right\} \to 0, \text{ if } \max\{q_i^u, q_i^d\} = 0, \ i \in \mathcal{N}_F,$$

$$\mathbb{P}\left\{\mathbf{X} = \mathbf{h}_i^u\right\} \to 0, \text{ if } q_i^u = 0, \ i \in \mathcal{N}_H,$$

$$\mathbb{P}\left\{\mathbf{X} = \mathbf{h}_i^d\right\} \to 0, \text{ if } q_i^d = 0, \ i \in \mathcal{N}_H.$$

Hence, in steady state, with high probability, the Markov chain $\mathbf{X}(t)$ never activates an HD link

with empty fluid limit queue or an FD link whose both UL and DL queues are empty, i.e., it chooses a *Maximal Schedule* over the non-zero fluid queues (note that the returned schedule might not be a MWS schedule). However, as mentioned in Section 7.2.4, the Markov chain $\mathbf{X}(t)$ might not always be at its steady state due to coupling between $\mathbf{X}(t)$ and $\mathbf{Q}(t)$. This coupling gives rise to qualitatively different fluid limits, depending on the time-scale of convergence of the schedule process compared to the time-scale of the changes in the queue process. For weight functions $f(\cdot)$, such that $\lim_{r\to\infty} f(r)/\log r < 1$, the schedule process $\mathbf{X}(t)$ is always close to its steady state at the fluid scale; while for functions $f(\cdot)$ with $\lim_{r\to\infty} f(r)/\log r > 1$, this does not happen. Nevertheless, in both cases, the following Lemma establishes a set of equations that the fluid limit sample paths under H-GMS algorithm must satisfy. The equations do not uniquely describe the fluid limit process but are sufficient to establish stability in our setting.

Lemma 7.3 (Fluid Limit Equations) *Consider any nonnegative increasing weight function $f(\cdot)$ in (7.5), such that $\lim_{x\to\infty} f(x)/\log x < 1$, or $\lim_{x\to\infty} f(x)/\log x > 1$ (including $f(x) = x^{\beta}$, $\beta > 0$). Let $\widehat{q}_i(t) = \max\{q_i^u(t), q_i^d(t)\}$, for $i \in \mathcal{N}_F$. At any regular point t (i.e., any point where the derivatives of all the functions exist), for any $j \in \{u, d\}$,*

$$q_i^j(t) = q_i^j(0) + a_i^j(t) - s_i^j(t), \; i \in \mathcal{N} \tag{7.9}$$

$$a_i^j(t) = \lambda_i^j t, \; s_i^j(t) = \int_0^t \mu_i^j(\tau) \, d\tau, \; \mu_i^j(t) \in [0, 1], \tag{7.10}$$

$$\mu_i^j(t) \cdot \mathbf{1}(q_i^j(t) = 0, \mathbf{q}(t) \neq \mathbf{0}) = 0, \; i \in \mathcal{N}_H, \tag{7.11}$$

$$\mu_i^j(t) \cdot \mathbf{1}(\widehat{q}_i(t) = 0, \mathbf{q}(t) \neq \mathbf{0}) = 0, \; i \in \mathcal{N}_F, \tag{7.12}$$

$$if \; q_i^j(t) = \widehat{q}_i(t), \; \mu_i^j(t) = \max\{\mu_i^u(t), \mu_i^d(t)\}, \; i \in \mathcal{N}_F, \tag{7.13}$$

$$if \; \mathbf{q}(t) \neq \mathbf{0}, \; then \; \sum_{i \in \mathcal{N}_F} \max\left\{\mu_i^u(t), \mu_i^d(t)\right\} + \sum_{i \in \mathcal{N}_H} \left(\mu_i^u(t) + \mu_i^d(t)\right) = 1. \tag{7.14}$$

The proof of Lemma 7.3 is provided Appendix 7.8.3. Essentially, (7.9)–(7.10) hold for any scheduling algorithm and their proof is standard. $\mu_i^j(t)$ is the rate that queue $q_i^j(t)$ is served at time t in the fluid limit. (7.11)–(7.14) imply that at any time, H-GMS chooses a maximal sched-

ule from the queues that are *non-zero* in the fluid limit (i.e., $\mathbf{q}(t) \neq \mathbf{0}$). However, the choice of maximal schedule could be random over the space of such maximal schedules.

Part (iii): Stability of the Queues in the Fluid Limit. The following proposition proves the stability of the queues in the fluid limit, which completes the proof of Theorem 7.1.

Proposition 7.2 *Starting from an initial queue size $\mathbf{q}(0)$, there is a deterministic finite time T by which all the queues at the fluid limit will reach zero.*

Proof 7.2 *Let $\widehat{q}_i(t) = \max\{q_i^u(t), q_i^d(t)\}$, $i \in \mathcal{N}_F$. Consider the Lyapunov function*

$$V(\mathbf{q}(t)) = \sum_{i \in \mathcal{N}_F} \widehat{q}_i(t) + \sum_{i \in \mathcal{N}_H} \left(q_i^u(t) + q_i^d(t)\right).$$

Let $\mathcal{U}_H^j(t) := \{i \in \mathcal{N}_H : q_i^j(t) > 0\}$, $j \in \{u, d\}$, and $\mathcal{U}_F(t) := \{i \in \mathcal{N}_F : \widehat{q}_i(t) > 0\}$. Suppose $V(\mathbf{q}(t)) > 0$ (i.e., $\mathbf{q}(t) \neq \mathbf{0}$). Then based on the fluid limit equations (7.11)–(7.14):

(i) The network is draining some subsets $\mathcal{P}_H^u(t) \subseteq \mathcal{U}_H^u(t)$, $\mathcal{P}_H^d(t) \subseteq \mathcal{U}_H^d(t)$, and $\mathcal{P}_F(t) \subseteq \mathcal{U}_F(t)$ of non-zero queues,

(ii) $\widehat{q}_i(t)$ for user $i \in \mathcal{P}_F(t)$ is always drained at rate $\max\{\mu_i^u(t), \mu_i^d(t)\}$,

(iii) $\sum_{i \in \mathcal{P}_F(t)} \max\{\mu_i^u(t), \mu_i^d(t)\} + \sum_{i \in \mathcal{P}_H^u(t)} \mu_i^u(t) + \sum_{i \in \mathcal{P}_H^d(t)} \mu_i^d(t) = 1$.

Hence, using (7.9)–(7.10) and properties (i)–(iii) above,

$$\frac{dV(\mathbf{q}(t))}{dt} \leq \sum_{i \in \mathcal{N}_F} \max\{\lambda_i^u, \lambda_i^d\} + \sum_{i \in \mathcal{N}_H} (\lambda_i^u + \lambda_i^d) - \sum_{i \in \mathcal{P}_F(t)} \max\{\mu_i^u(t), \mu_i^d(t)\}$$

$$- \sum_{i \in \mathcal{P}_H^u(t)} \mu_i^u(t) - \sum_{i \in \mathcal{P}_H^d(t)} \mu_i^d(t)$$

$$= \sum_{i \in \mathcal{N}_F} \max\{\lambda_i^u, \lambda_i^d\} + \sum_{i \in \mathcal{N}_H} (\lambda_i^u + \lambda_i^d) - 1 \leq -\delta,$$

where the last inequality is due to the fact that $\boldsymbol{\lambda} \in \mathrm{int}(\Lambda_{HD\text{-}FD})$, by the assumption of Theorem 7.1. Thus, there must exist a small $\delta > 0$ such that $\boldsymbol{\lambda}/(1 - \delta) \in \Lambda_{HD\text{-}FD}$. Therefore, $V(\mathbf{q}(t))$ will hit zero in finite time $T = V(\mathbf{q}(0))/\delta$, and in fact remains zero afterwards.

Proposition 7.2 implies the stability (positive recurrence) of the original Markov chain $(\mathbf{X}(t), \mathbf{Q}(t))$

169

in a similar fashion as [169] (note that the component $\mathbf{X}(t)$ lives in a finite state space). This completes the proof of Theorem 7.1.

Remark: We emphasize that, the proposed H-GMS algorithm approximates GMS in a distributed manner, where the fluid limits are largely different from that of the classical Q-CSMA (which approximates MWS in a distributed manner). Further, we establish throughput optimality of H-GMS for a broad family of (almost) any increasing weight function $f(\cdot)$.

7.4 Lower Bounds on the Average Queue Length

In this section, we analyze the delay performance of H-GMS in terms of the average queue length in order to provide a benchmark for the performance evaluation in Section 7.6. In particular, we derive two lower bounds: (i) a fundamental lower bound that is independent of the scheduling algorithms, and (ii) an improved lower bound tailored for the developed H-GMS and H-GMS-R.[4] In Section 7.6.2, we numerically evaluate these lower bounds and compare them to the average queue length achieved by various scheduling algorithms.

We adopt the following notation. Given a set of links $\mathcal{L}$, we use $\lambda_\mathcal{L} = \sum_{l \in \mathcal{L}} \lambda_l$ to denote the sum of arrival rates, and use $\mathcal{Q}_\mathcal{L} = \sum_{l \in \mathcal{L}} \mathbb{E}[Q_l]$ to denote the expected sum of queue lengths of $\mathcal{L}$ in steady state. The average queue length in a given heterogeneous HD-FD network, $(\mathcal{N}, \mathcal{E})$, is defined by

$$\overline{Q} = \frac{1}{|\mathcal{E}|} \cdot \sum_{l \in \mathcal{E}} \mathbb{E}[Q_l] = \frac{\mathcal{Q}_\mathcal{E}}{2N}. \tag{7.15}$$

Therefore, finding a lower bound on $\overline{Q}$ is equivalent to finding a lower bound on $\mathcal{Q}_\mathcal{E}$.

7.4.1   A Fundamental Lower Bound

We first derive a fundamental lower bound on $\overline{Q}$ that is independent of the chosen (possibly centralized) scheduling algorithm, based on the following result.

---

[4]The analysis can possibly be extended to H-GMS-E by incorporating its *time-varying and queue-dependent* access probability. We leave this analysis for future work.

Proposition 7.3 ([170, Proposition 4.1]) *With independent packet arrivals, the expected sum of queue lengths in a clique $\mathcal{C}$ under any scheduling policy satisfies*

$$\mathcal{Q}_{\mathcal{C}} = \sum_{l \in \mathcal{C}} \mathbb{E}[Q_l] \geq \sum_{l \in \mathcal{C}} \frac{\lambda_l + \mathrm{Var}\,[A_l] - \lambda_l \lambda_{\mathcal{C}}}{2\,(1 - \lambda_{\mathcal{C}})} := \mathcal{Q}_{\mathcal{C}}^{LB}.$$

Note that $\mathcal{Q}_{\mathcal{C}}^{\mathrm{LB}}$ is equivalent to the sum of queue lengths in a standard single-server GI/D/1 queue in clique $\mathcal{C}$. In order to obtain a tight fundamental lower bound in the heterogeneous HD-FD networks, one needs to find the largest clique of links, $\mathcal{E}_{\mathrm{max}}$, with the maximal sum of arrival rates. In particular, we divide $\mathcal{E}$ into two disjoint sets $\mathcal{E} = \mathcal{E}_{\mathrm{max}} \cup \mathcal{E}_{\mathrm{min}}$:

$$\begin{cases} \mathcal{E}_{\mathrm{max}} = \{l_i^j : \forall i \in \mathcal{N}_F \text{ if } \lambda_i^j \geq \lambda_i^{\bar{j}}\} \cup \{l_i^{\mathrm{u}}, l_i^{\mathrm{d}} : \forall i \in \mathcal{N}_H\}, \\ \mathcal{E}_{\mathrm{min}} = \{l_i^j : \forall i \in \mathcal{N}_F \text{ if } \lambda_i^j < \lambda_i^{\bar{j}}\}, \end{cases}$$

where $\{\bar{j}\} = \{\mathrm{u}, \mathrm{d}\} \setminus \{j\}$ and we break ties uniformly at random if $\lambda_i^{\mathrm{u}} = \lambda_i^{\mathrm{d}}$ for $\forall i \in \mathcal{N}_F$. Essentially, $\mathcal{E}_{\mathrm{max}}$ includes the UL *and* DL of each HD user, and the higher arrival rate link (UL *or* DL) of each FD user. As a result, $\lambda_{\mathcal{E}_{\mathrm{max}}}$ approaches 1 as $\boldsymbol{\lambda}$ approaches the boundary of $\Lambda_{\mathrm{HD\text{-}FD}}$ (see (7.1)). The following proposition gives the fundamental lower bound on the average queue length in the heterogeneous HD-FD networks.

Proposition 7.4 *A fundamental lower bound on the average queue length in the considered heterogeneous HD-FD networks, denoted by $\overline{Q}_{Fund}^{LB}$, is given by*

$$\overline{Q} \geq \overline{Q}_{Fund}^{LB} := \frac{\mathcal{Q}_{\mathcal{E}_{max}}^{LB}}{2N}, \tag{7.16}$$

where $\overline{Q}$ is the average queue length defined in (7.15), and $\mathcal{Q}_{\mathcal{E}_{max}}^{LB}$ is given by Proposition 7.3 for clique $\mathcal{E}_{max}$.

Proof 7.3 *Since a pair of FD UL and DL will always be activated at the same time, it holds that for $\forall i \in \mathcal{N}_F$, $\mathbb{E}[Q_i^j] \geq \mathbb{E}[Q_i^{\bar{j}}]$ if $\lambda_i^j \geq \lambda_i^{\bar{j}}$. By assigning the FD UL/DL with a higher arrival rate to $\mathcal{E}_{max}$, we construct a maximal clique, $\mathcal{E}_{max}$, with the maximal possible sum or arrival rates. Although it is possible that two queues from both $\mathcal{E}_{min}$ and $\mathcal{E}_{max}$ are served simultaneously, it is still*

guaranteed that

$$\mathcal{Q}_{\mathcal{E}} \geq \mathcal{Q}_{\mathcal{E}_{max}} \geq \mathcal{Q}_{\mathcal{E}_{max}}^{LB}, \tag{7.17}$$

and Proposition 7.4 follows directly.

7.4.2 An Improved Lower Bound under H-GMS and H-GMS-R

We now derive an improved lower bound on $\overline{Q}$ for the considered heterogeneous HD-FD networks taking into account the characteristics of the developed H-GMS and H-GMS-R (e.g., the access probability α and the transmission probability $p(\cdot)$). The result is stated in the following proposition.

Proposition 7.5 *Let $p^{-1}(\cdot)$ be the inverse of the transmission probability $p(\cdot)$ given by (7.5). Let $\lambda_{min} = \min_{i \in \mathcal{N}}\{\lambda_i^u, \lambda_i^d\}$ be the minimum link arrival rate and $\alpha_{max} = \max\{\alpha_1, \cdots, \alpha_N, \alpha_{AP}\}$ be the maximum access probability. The average queue length under H-GMS and H-GMS-R is lower bounded by $\overline{Q}_{H\text{-}GMS}^{LB}$ given by*

$$\overline{Q} \geq \overline{Q}_{H\text{-}GMS}^{LB} := \max\left\{\overline{Q}_{Fund}^{LB}, \ \left(1 - \frac{N_F}{2N}\right) \cdot p^{-1}\left(\frac{\lambda_{min}/\alpha_{max}}{1 - \lambda_{\mathcal{E}_{max}} + \lambda_{min}/\alpha_{max}}\right)\right\}, \tag{7.18}$$

where $\overline{Q}_{Fund}^{LB}$ is given in Proposition 7.4.

Proof 7.4 *The proof is based on the workload decomposition rules [171] and can be found in Appendix 7.8.4.*

Remark: Note that (7.18) applies to any variant of H-GMS with fixed access probability α. The lower bound $\overline{Q}_{H\text{-}GMS}^{LB}$ depends on: (i) the ratio between the link arrival rate and access probability $\frac{\lambda_{min}}{\alpha_{max}}$, and (ii) the weight function $f(\cdot)$ (through $p(\cdot)$). A more aggressive $f(\cdot)$ results in a lower value of $\overline{Q}_{H\text{-}GMS}^{LB}$. The lower bound can be also applied to H-GMS-E by setting $\alpha_{max} = 1$; however, this will result in a loose lower bound as it ignores the adaptive behavior of α.
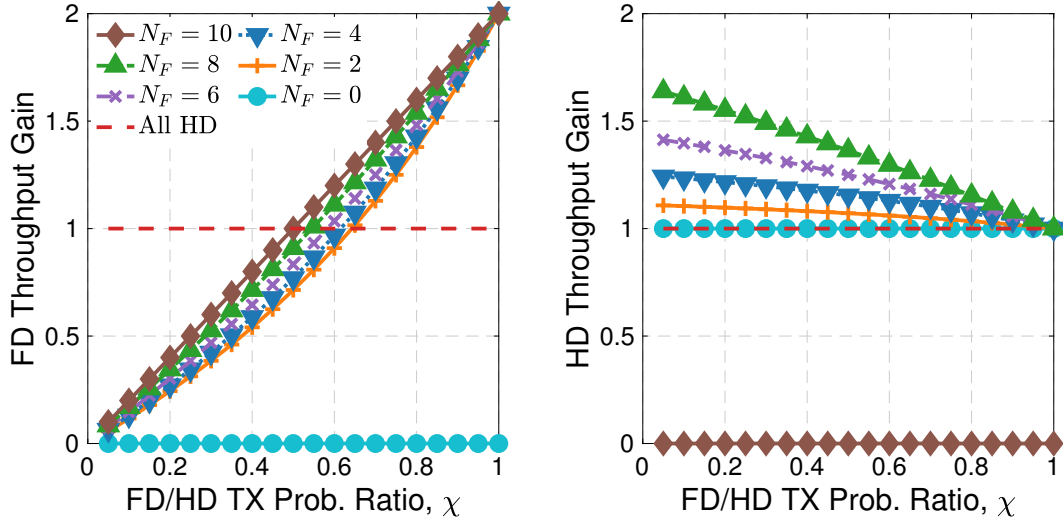
Figure 7.1: Throughput gain of FD and HD users when the throughput is compared to the individual throughput of an HD user in the all-HD network under the static H-GMS-R algorithm, with $N = 10$, $N_F \in \{0, 2, \cdots, 10\}$, and $p_h = 0.5$.

7.5 Benefits of Introducing FD-Capable Nodes

In this section, we illustrate the benefits of introducing FD-capable nodes into all-HD networks, in terms of obtained throughput gains. The throughput gains can be expressed for individual users or the network (i.e., the sum rates). We define the network (individual) throughput gain as the ratio between the achievable network (individual) throughput in a heterogeneous HD-FD network and that in an all-HD network with the same total number of users.

For simplicity and illustrative purposes, consider a static version of H-GMS-R, with access probabilities $\boldsymbol{\alpha} = \frac{1}{1+N} \cdot \mathbf{1}$ (see Algorithm 7.3 and Section 7.2.5), and fixed transmission probabilities $p_f^{\mathrm{u}} = p_f^{\mathrm{d}} = p_f$, $p_h^{\mathrm{u}} = p_h^{\mathrm{d}} = p_h \in (0, 1)$ for FD and HD users in (7.5), respectively. By analyzing the Markov chain (similar to Lemma 7.2) under fixed $\boldsymbol{\alpha}$, p_f, and p_h, the network throughput (i.e., sum rates) of the heterogeneous HD-FD network, $S_{\mathrm{HD\text{-}FD}}$, is given by

$$S_{\mathrm{HD\text{-}FD}} = \frac{\frac{2N_F}{N} \frac{p_f}{1-p_f} + \frac{N_H}{N} \frac{p_h}{1-p_h}}{1 + \frac{N_F}{N} \frac{p_f}{1-p_f} + \frac{N_H}{N} \frac{p_h}{1-p_h}}. \tag{7.19}$$

Note that the throughput of the benchmark all-HD network is simply $S_{\mathrm{HD}} = p_h$. If $p_f = p_h = p$ (i.e., FD and HD users transmit with the same probability when they capture the wireless channel),

(7.19) becomes $S_{\text{HD-FD}} = (1 + \frac{N_F}{N}) \cdot p$. This implies that under the static H-GMS-R, the network throughput gain achieved by the HD-FD network is $(1 + \frac{N_F}{N}) \in [1, 2]$, which increases with respect to the number of FD users, N_F.

Assigning equal transmission probabilities results in FD users having $2\times$ throughput compared to the HD users. We can balance the throughput obtained by FD and HD users by assigning different transmission probabilities. Let $p_h = p$ and $p_f = \chi \cdot p$ for some *transmission probability ratio* χ. In order to balance the individual throughput of FD and HD users, we lower the priority of FD transmissions by choosing $\chi \in (0, 1]$.

We numerically evaluate the individual user throughput gain. We consider both the benchmark all-HD network (with transmission probability $p_h = p$) and HD-FD networks with $N = 10$ and vary $N_F \in \{0, 2, \cdots, 10\}$ in the latter. We select constant $p_h = 0.5$ and $p_f = \chi \cdot p_h$ with varying $\chi \in (0, 1]$. Fig. 7.1 plots individual throughput gains of an FD or HD user. As Fig. 7.1 suggests, if FD and HD users are assigned equal transmission probabilities ($\chi = 1$), an FD user gets $2\times$ throughput compared to an HD user. If the transmission probability of the FD users is lowered (by decreasing χ), the throughput of FD and HD users is more balanced. For example, with $\chi = 0.75$, the individual throughput gains of FD and HD users are 43% and 20%, respectively.

The results reveal an interesting phenomenon: when N_F is sufficiently large, at the cost of slightly lowering the priority of FD users, even HD users can experience throughput improvements. This opens up the possibility of designing wireless protocols with different fairness-efficiency tradeoffs by setting different priorities among FD and HD users.

7.6 Simulations

In this section, we evaluate the performance of different scheduling algorithms in heterogeneous HD-FD networks via simulations. We focus on (i) network-level *delay* performance (represented by the long-term average queue length per link), and (ii) *fairness* between FD and HD users (represented by the relative delay performance between FD and HD users).

7.6.1 Setup

Throughout this section, we consider heterogeneous HD-FD networks with one FD AP and 10 users ($N = 10$), with a varying number of FD users, N_F.[5] We choose a rate vector $\mathbf{v} = [v_i^{\mathrm{u}}, v_i^{\mathrm{d}}]_{i=1}^N$ on the boundary of the capacity region $\Lambda_{\text{HD-FD}}$ (see Section 7.1.3) and consider arrival rates of the form $\boldsymbol{\lambda} = \rho \mathbf{v}$, in which $\rho \in (0, 1)$ is the *traffic intensity*. Note that as $\rho \to 1$, $\boldsymbol{\lambda}$ approaches the boundary of $\Lambda_{\text{HD-FD}}$. Since we focus on the fairness between FD and HD users, we assume equal UL and DL arrival rates over all the users. Therefore, for $j \in \{\mathrm{u}, \mathrm{d}\}$, we use $v_f = v_i^j$, $\forall i \in \mathcal{N}_F$, and $v_h = v_i^j$, $\forall i \in \mathcal{N}_H$, to denote the equal UL and DL arrival rates assigned to FD and HD users, respectively. For an *equal arrival rate* model, we have $v_f = v_h = 1/(N_F + 2N_H)$ computed using (7.1).

The packet arrivals at each link l_i^j follow an independent Bernoulli process with rate λ_i^j. For each algorithm under a given traffic intensity, ρ, we take the average over 10 independent simulations, each of which lasts for 10^6 slots. For simplicity, we refer to the "queue length of an FD (resp. HD) user" as the sum of its UL and DL queue lengths, and only compare the average queue length between FD and HD users without distinguishing between individual UL and DL. The considered algorithms include:

- MWS, GMS: The centralized MWS and GMS algorithms;

- H-GMS, H-GMS-R, and H-GMS-E: Three variants of the H-GMS algorithm as described in Section 7.2.5;

- Q-CSMA: The standard distributed Q-CSMA algorithm from [164], in which each link (UL or DL) performs channel contention independently and the AP does not leverage the central DL queue information.

In the last four distributed algorithms, the transmission probability of link l in slot t is selected as $p_l(t) = \frac{\exp\left(f(Q_l(t))\right)}{1+\exp\left(f(Q_l(t))\right)}$ where the weight function $f(x) = \log\left(1 + x\right)$ (i.e., $p_l(t) = \frac{1+Q_l(t)}{2+Q_l(t)}$). We set $\boldsymbol{\alpha} = \frac{1}{1+N} \cdot \mathbf{1}$ for H-GMS and H-GMS-R, and $\alpha_{\text{th}} = 0.01$ for H-GMS-E (see Section 7.2.5). We

---

[5]The results for heterogeneous HD-FD networks with a different number of users, N, are similar, and thus, omitted.
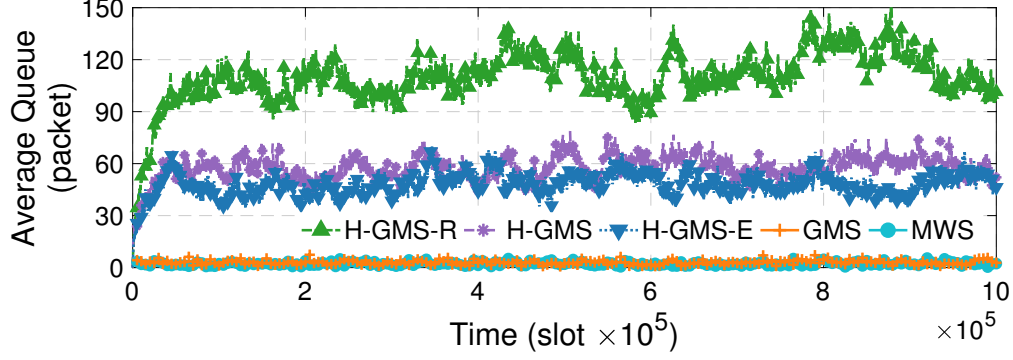
Figure 7.2: Sample path of average queue length per link under different scheduling algorithms for a heterogeneous HD-FD network with $N_F = N_H = 5$, and *very high* traffic intensity $\rho = 0.95$.

will show that different degrees of centralization at the AP result in performance improvements of H-GMS over the classical Q-CSMA in terms of both delay and fairness. We also consider effects of different weight functions in Section 7.6.4.

7.6.2 Delay Performance

We first consider the queue length dynamics under various scheduling algorithms in an HD-FD network with $N_F = N_H = 5$ and traffic intensity $\rho = 0.95$. This implies that $v_f = v_h = 1/15$, corresponding to a capacity region expansion value of $\gamma = 4/3$ (see Section 7.1.3 with $v_h = 1/20$ in the all-HD network). Fig. 7.2 plots the sample paths of the average queue length of the network (i.e., averaged over all the ULs and DLs) under different algorithms. The result for Q-CSMA algorithm is omitted since, as we will see shortly, its average queue length is at least one order of magnitude larger than those achieved by other algorithms.

Fig. 7.3 plots the average queue length with varying traffic intensities in HD-FD networks with $N = 10$ and $N_F \in \{0, 5, 10\}$. Recall that in the equal arrival rate model, the relationship between the link packet arrival rate and traffic intensity is $\lambda_i^j = \rho/(N_F + 2N_H)$, $\forall i \in \mathcal{N}$, $\forall j \in \{u, d\}$. Fig. 7.3 shows that the capacity region of the HD-FD networks expands with increased value of N_F. Compared with Figs. 7.3(a), Figs. 7.3(b) and 7.3(c) show a capacity region expansion value of $\gamma = 4/3$ for $N_F = 5$, and $\gamma = 2$ for $N_F = 10$, respectively.

Figs. 7.2 and 7.3 show that, as expected from Theorem 7.1, all the considered algorithms are
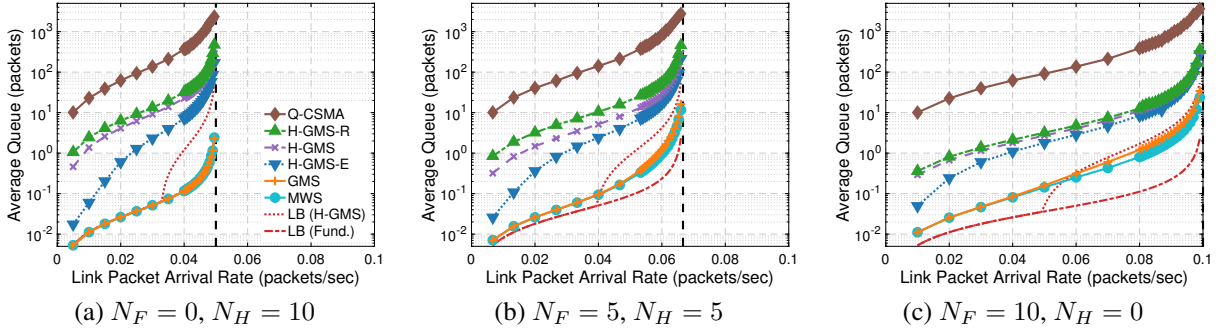
Figure 7.3: Long-term average queue length per link in a heterogeneous HD-FD network with $N = 10$ and equal arrival rates, under different scheduling algorithms and varying number of FD users, N_F: (a) $N_F = 0$, (b) $N_F = 5$, and (c) $N_F = 10$. Both the fundamental and improved lower bounds on the delay are also plotted according to (7.16) and (7.18). The capacity region boundary in each HD-FD network is illustrated by the vertical dashed line.

throughput-optimal – they stabilize all network queues. The fully-centralized MWS and GMS have the best delay performance but require high-complexity implementations. Among distributed algorithms, Q-CSMA [164] has the worst delay performance due to the high contention intensity introduced by a total of $2N$ contending links. By "consolidating" the N DLs into one DL that participates in channel contention, H-GMS-R, H-GMS, and H-GMS-E achieve at least 9–$16\times$, 16–$30\times$, and 25–$50\times$ better delay performance than Q-CSMA, respectively, under different traffic intensities ρ. In particular, H-GMS and H-GMS-E have similar delay performance which is better than for H-GMS-R, since the AP leverages its central information to always select the longest queue DL for channel contention. However, H-GMS and H-GMS-E provide different fairness among FD and HD users due to the choice of access probability distribution α (that is constant for the former and depends on the queue-length estimates for the latter), as we show below.

Fig. 7.3 also presents both the fundamental and improved lower bounds on the delay, $\overline{Q}_{\text{Fund}}^{\text{LB}}$ and $\overline{Q}_{\text{H-GMS}}^{\text{LB}}$, given by (7.16) and (7.18), respectively. The turning point of $\overline{Q}_{\text{H-GMS}}^{\text{LB}}$ where it starts to deviate from $\overline{Q}_{\text{Fund}}^{\text{LB}}$ is because of the $\max(\cdot)$ operator in (7.18). As Fig. 7.3 suggests, the fundamental lower bound, $\overline{Q}_{\text{Fund}}^{\text{LB}}$ is very close to the average queue length obtained by MWS and GMS (they indeed match perfectly in the all-HD network with $N_H = N = 10$). However, in heterogeneous HD-FD networks, $\overline{Q}_{\text{H-GMS}}^{\text{LB}}$ provides a much tighter lower bound on the average queue length

achieved by H-GMS, especially with high traffic intensities.

7.6.3   Fairness

Our next focus is on the fairness performance of H-GMS. Here, we define fairness between FD and HD users as the *ratio between the average queue length of FD and HD users*. We use this notion since, intuitively, if an FD user experiences lower average delay (i.e., queue length) than an HD user, then introducing FD capability to the network will imbalance the service rate both users get. Ideally, we would like the proposed algorithms to achieve good fairness performance in the considered HD-FD networks. Similarly, we define fairness between ULs and DLs as the *ratio between the average UL and DL queue lengths* to evaluate the effects of different levels of centralization at the AP when operating H-GMS.

Equal Arrival Rates. We first evaluate the fairness under different distributed algorithms under the *equal arrival rate model*. We focus on traffic intensity regime of $\rho \in [0.5, 1)$ since, as shown in Fig. 7.3, all links have very small queue lengths with low traffic intensities (e.g., the average queue length is less the 10 packets with $\rho = 0.5$).

Fig. 7.4(a) plots the fairness between FD and HD users in an HD-FD network with $N_F = N_H = 5$ and varying traffic intensity, ρ. It can be observed that H-GMS-R has the worst fairness performance since the DL participating in the channel contention is selected uniformly at random by the AP. When the traffic intensity is low or moderate, Q-CSMA and H-GMS achieve similar fairness of about 0.5. This is because under equal arrival rates, FD queues are about half the length of the HD queues due to the fact that they are being served about twice as often (i.e., an FD bi-directional transmission can be either activated by the FD UL or DL due to the FD PHY capability). When the traffic intensity is high, both H-GMS and H-GMS-E have increased fairness performance since the longest DL queue will be served more often due to the central DL queue information at the AP. Furthermore, H-GMS-E outperforms H-GMS since, under H-GMS-E, the AP not only has explicit information of all the DL queues, but also has estimated UL queue lengths that can be used to better assign the access probability distribution α.

(a) b/w FD and HD users   (b) b/w ULs and DLs
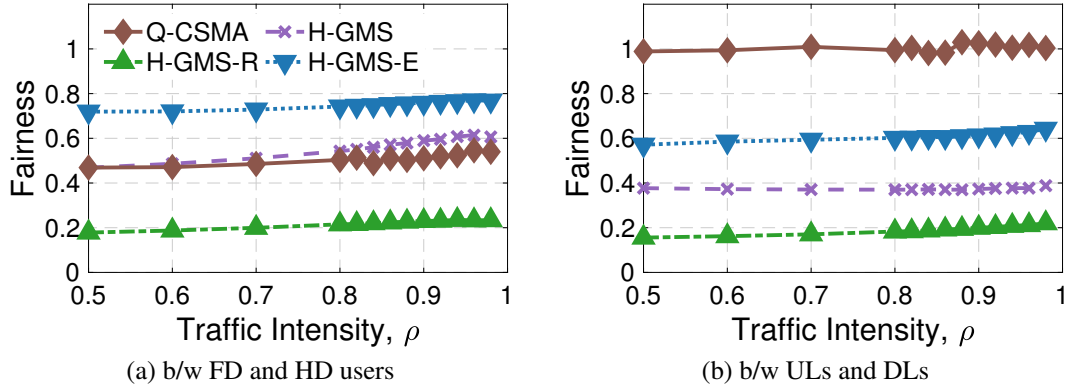
Figure 7.4: Long-term average queue length ratio between (a) FD and HD users, and (b) ULs and DLs with varying traffic intensity, in an HD-FD network with $N_F = N_H = 5$ and equal arrival rates.

(a) Traffic intensity $\rho = 0.8$ (b) Traffic intensity $\rho = 0.95$
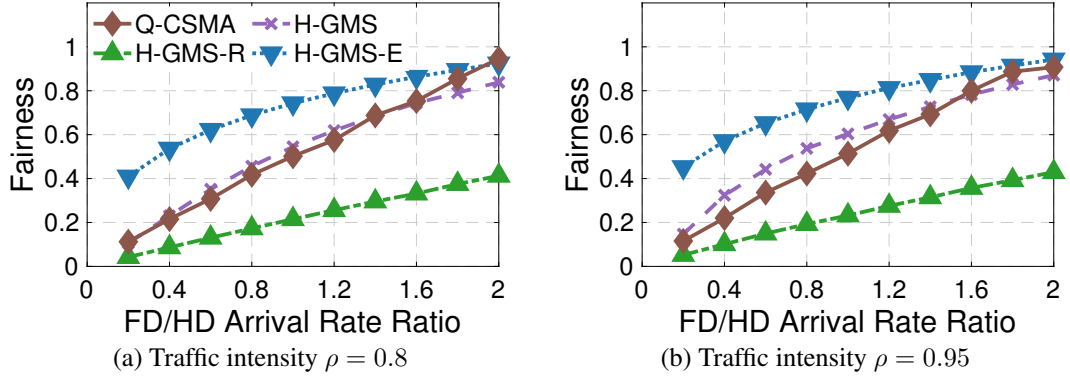
Figure 7.5: Long-term average queue length ratio between FD and HD users in a heterogeneous HD-FD network with $N_F = N_H = 5$ and varying ratio between FD and HD arrival rates, with (a) moderate ($\rho = 0.8$), and (b) high ($\rho = 0.95$) traffic intensities.

Fig. 7.4(b) presents the fairness between ULs and DLs with the same network setting. It can be seen that Q-CSMA has the best fairness performance of around 1 since all the $2N$ link have equal access probability. The fairness by H-GMS-R between FD and HD users, and between ULs and DLs, are almost identical, and are always the worst among all variants of H-GMS. On the other hand, H-GMS-E still has the best fairness performance among all variants of H-GMS by leveraging the information on estimated UL queue lengths.

Different Arrival Rates. We also evaluate the fairness under different distributed algorithms with *different arrival rates* between FD and HD users. Let σ be the ratio between the arrival rates on FD and HD links. It is easy to see that if we assign $v_f = \sigma/(\sigma N_F + 2N_H)$ and $v_h = 1/(\sigma N_F + 2N_H)$,
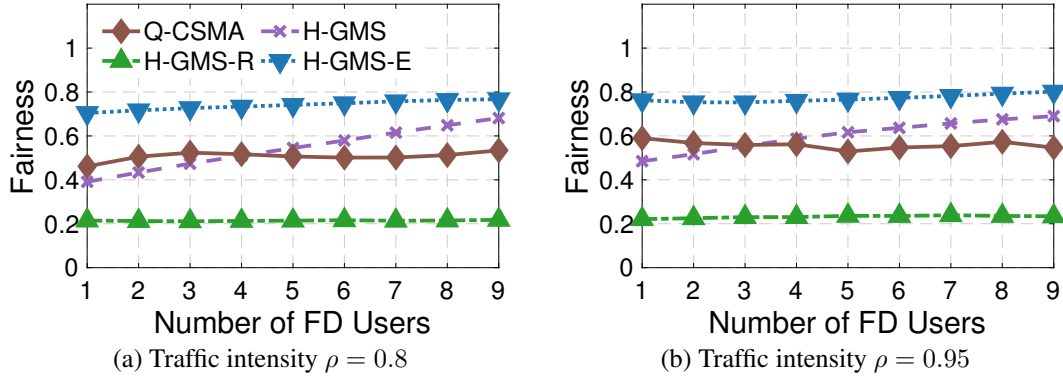
Figure 7.6: Long-term average queue length ratio between FD and HD users in a heterogeneous HD-FD network with $N_F = N_H = 5$ and varying $N_F \in \{1, 2, \cdots, N-1\}$, with (left) moderate ($\rho = 0.8$), and (right) high ($\rho = 0.95$) traffic intensities.

then $\mathbf{v}$ is on the boundary of $\Lambda_{\text{HD}-\text{FD}}$. In this case, we have a capacity region expansion value of $\gamma = 1 + \sigma N_F / (\sigma N_F + 2N_H)$, which depends on both N_F and σ (see Section 7.1.3).

Fig. 7.5 plots the fairness between FD and HD users with varying σ under moderate ($\rho = 0.8$) and high ($\rho = 0.95$) traffic intensities on the x-axis. It can be observed that as the packet arrival rate at FD users increases, the FD and HD queue lengths are better balanced. When $\sigma = 2$, FD and HD users have almost the same average queue length since the FD queues are served twice as often as the HD queues under Q-CSMA, H-GMS, and H-GMS-E. It is interesting to note that the fairness under Q-CSMA and H-GMS is almost a linear function with respect to the arrival rate ratio, σ. This is intuitive since, as the FD queues are served about twice as often as the HD queues, increased arrival rates will result in longer queue lengths at the FD users. Moreover, since the FD and HD queues have about the same queue length when σ approaches 2, H-GMS-E does not further improve the fairness since it generates an access probability distribution that is approximately a uniform distribution.

Impact of the Number of FD Users, N_F. We now evaluate the fairness between FD and HD users with varied number of FD (or equivalently, HD) users under the equal arrival rate model. We vary $N_F \in \{1, 2, \cdots, 9\}$. Fig. 7.6 plots the fairness between FD and HD users under moderate ($\rho = 0.8$) and high ($\rho = 0.95$) traffic intensities.

As Fig. 7.6 suggests, the fairness depends on the number of FD users, N_F, only under H-GMS.

This is because under equal arrival rate, FD users have about half the queue lengths compared with HD users. As N_F increases, the number of HD DLs at the AP (those with relatively larger queue length) decreases and as a result, the AP is very likely to select an HD DL or UL under the H-GMS algorithm, resulting in larger average queue length at the FD users. In addition, H-GMS-E resolves this issue by taking into account the UL queue length estimates. Therefore, the FD users that have smaller queues will be selected with a lower probability so that the longer HD queues will be served at a higher rate. In addition, as N_F increases, H-GMS achieves better fairness than that of the classical Q-CSMA by approximating the GMS (instead of MWS as Q-CSMA does) in a distributed manner. Moreover, H-GMS-E has the best fairness performance which is independent of the value of N_F.

7.6.4 Impact of the Weight Function, $f(x)$

We now evaluate the delay performance of H-GMS under different weight functions and compare it to Q-CSMA. Recall from Theorem 7.1 that H-GMS is throughput-optimal for a broad family of weight functions, $f(x)$, and the relationship between $f(\cdot)$ and the transmission probability $p(\cdot)$ is given by (7.5). In particular, we consider the following weight functions:

- $f(x) = \frac{1}{2}\log(1+x)$: $\lim_{x\to\infty}\frac{f(x)}{\log x} = \frac{1}{2} < 1$;

- $f(x) = \log(1+x)$: $\lim_{x\to\infty}\frac{f(x)}{\log x} = 1$;

- $f(x) = \sqrt{x}$: $\lim_{x\to\infty}\frac{f(x)}{\log x} = \infty$ ($\beta = \frac{1}{2}$);

- $f(x) = x$: $\lim_{x\to\infty}\frac{f(x)}{\log x} = \infty$ ($\beta = 1$).

Fig. 7.7 plots the average queue length with varying traffic intensity in an HD-FD network with $N_F = N_H = 5$ and equal arrival rates, and with different weight functions, $f(x)$, as listed above. For each considered $f(\cdot)$, we consider all four distributed algorithms listed in Section 7.6.1. The results in the case with $f(x) = \log(1+x)$ are shown in Fig. 7.3(b). Table 7.1 summarizes the improvements in the delay performance achieved by variants of H-GMS compared to Q-CSMA, with the considered weight functions and moderate ($\rho = 0.8$) and extremely high ($\rho = 0.98$) traffic
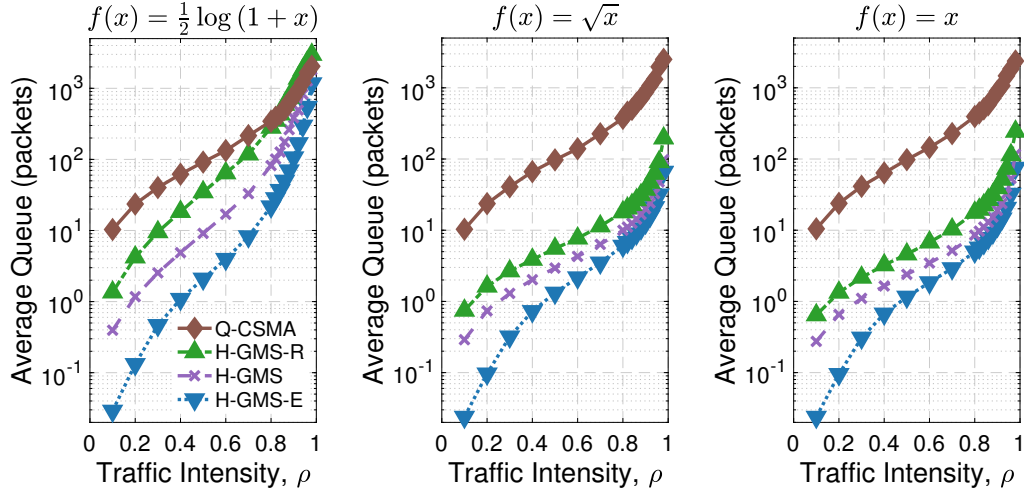
Figure 7.7: Long-term average queue length in a heterogeneous HD-FD network with $N_F = N_H = 5$ and equal arrival rates, with different weight functions and $p_l(t) = \frac{\exp(f(Q_l(t)))}{1+\exp(f(Q_l(t)))}$. The results in the case with $f(x) = \log(1+x)$ are shown in Fig. 7.3(b).

Weight Function, $f(x)$	$\frac{1}{2}\log(1+x)$		$\log(1+x)$		x	
Traffic Intensity, ρ	0.8	0.98	0.8	0.98	0.8	0.98
$\frac{\overline{Q}_{\text{Q-CSMA}}}{\overline{Q}_{\text{H-GMS-R}}}$	1.2	0.7	14.4	8.5	22.3	9.8
$\frac{\overline{Q}_{\text{Q-CSMA}}}{\overline{Q}_{\text{H-GMS}}}$	4.2	1.1	28.4	16.2	46.2	20.4
$\frac{\overline{Q}_{\text{Q-CSMA}}}{\overline{Q}_{\text{H-GMS-E}}}$	15.8	1.7	52.8	25.4	79.2	31.8

Table 7.1: Improvements in the delay performance achieved by H-GMS compared with Q-CSMA under three different weight functions with different aggressiveness and, with moderate ($\rho = 0.8$) and extremely high ($\rho = 0.98$) traffic intensities.

intensities.[6]

The results show that all the scheduling algorithms are throughput-optimal under different choices of $f(x)$ that satisfy the conditions in Theorem 7.1. Overall, the delay performance of Q-CSMA in HD-FD networks has less dependency on $f(x)$ than H-GMS, and variants of H-GMS (especially H-GMS and H-GMS-E) achieve significantly improved delay performance. Moreover, the delay improvement achieved by H-GMS over the classical Q-CSMA becomes more significant with a "more aggressive" weight function. For example, H-GMS with a sublinear/linear weight

---

[6]The results in the cases with $f(x) = \sqrt{x}$ and $f(x) = x$ are almost identical (see Fig. 7.7) and thus omitted in Table 7.1.

function ($f(x) = x^\beta$ with $\beta \in \{\frac{1}{2}, 1\}$) achieves 10–20× better delay than with a logarithmic weight function $f(x) = \frac{1}{2}\log(1 + x)$. This highlights the importance of the selection of $f(x)$ in the design of H-GMS.

7.7 Conclusion

We presented a hybrid scheduling algorithm, H-GMS, for heterogeneous HD-FD infrastructure-based networks. H-GMS is distributed at the users and leverages different degrees of centralization at the AP to achieve good delay performance, and we proved the throughput optimality of H-GMS based on the fluid limit technique. We also derived lower bounds on the average queue length to evaluate the delay performance of H-GMS. We further illustrated various aspects of the performance of H-GMS and compared it to the classical Q-CSMA through extensive simulations. We also illustrated benefits and fairness-efficiency tradeoffs arising from incorporating FD users into existing HD networks.

7.8 Appendices

7.8.1 Appendix – Proof of Proposition 7.1

The proof is based on the structural properties of the interference graph of the heterogeneous HD-FD network. The interference (or conflict) graph is defined as $G_\mathrm{I} = (\mathcal{V}_\mathrm{I}, \mathcal{E}_\mathrm{I})$, where $\mathcal{V}_I$ is the set of network links, and there is an edge between link l_i and link l_j if they interfere with each other. Clearly, the interference graph of a collocated all-HD network is a clique. For the collocated HD-FD network, $\mathcal{V}_\mathrm{I} = \{v_i^j, i \in \mathcal{V}, j \in \{\mathrm{u}, \mathrm{d}\}\}$, where v_i^j corresponds to link j (UL or DL) of user i. Since a pair of FD UL and DL can be simultaneously activated, G_I is a complete graph with N_F edges missing, each of which has endpoints $(v_i^\mathrm{u}, v_i^\mathrm{d})$ for $\forall i \in \mathcal{N}_F$.

It has been shown in [162] that the Greedy Maximal Scheduling (GMS) is throughput-optimal if the interference graph G_I satisfies the so called *Overall Local Pooling (OLoP)* condition. We use the following definition and result from [165].

Definition 7.1 (Co-strongly perfect graph) *A graph G is co-strongly perfect, if and only if G contains a clique that intersects every maximal independent set in G.*

Proposition 7.6 ([165, Definitions 2.2, 2.3, and 5.1]) *Every graph that is co-strongly perfect satisfies OLoP.*

We now prove Proposition 7.1. To show that G_I is co-strongly perfect, if suffices to find a clique contained in G_I that intersects every maximal independent set in G_I. Recall that G_I is a complete graph with N_F edges missing. Let $K = \{v_n^u, n \in \mathcal{N}\} \cup \{v_m^d, m \in \mathcal{N}_H\} \subseteq \mathcal{V}_I$ with $|K| = N_F + 2N_H$. It is easy to see that the induced graph $G(K)$ on K is a clique. In addition, note that the maximal independent set in G_I can be (i) $\{v_m^u\}$ or $\{v_m^d\}$ for some $m \in \mathcal{N}_H$, or (ii) $\{v_n^u, v_n^d\}$ for some $n \in \mathcal{N}_F$ (there are a total number of $(N_F + 2N_H)$ such maximal independent sets). Thus, $G(K)$ intersects with every maximal independent set in G_I, which implies that G_I is co-strongly perfect and satisfies OLoP. Hence, the centralized GMS algorithm (described in Algorithm 7.2) is throughput-optimal in *any* collocated heterogeneous HD-FD network.

7.8.2 Appendix – Proof of Lemma 7.2

Under fixed $\mathbf{Q}(t) = \mathbf{Q}$, (7.7) is the transition probability matrix of the discrete time Markov chain $Y^{\mathbf{Q}}$. Recall that the state space of $Y(t)$ is $S_Y = \{0, 1, \cdots, N, i^\star\}$, where $i^\star = \arg\max_{i \in \mathcal{N}} Q_i^d$. The detailed balance equations for $Y^{\mathbf{Q}}$ are:

$$\pi^{\mathbf{Q}}(0) \cdot P(0, i) = \pi^{\mathbf{Q}}(i) \cdot P(i, 0), \ \forall i \in S_Y \setminus \{0\}. \tag{7.20}$$

Recall that the transmission probability is given by (7.5); from (7.7) and (7.20), we have

$$\pi^{\mathbf{Q}}(i) = \alpha_i \frac{p_i^u}{\overline{p}_i^u} \cdot \pi^{\mathbf{Q}}(0) = \alpha_i \exp\left(f(Q_i^u)\right) \cdot \pi^{\mathbf{Q}}(0), \ \forall i \in S_Y \setminus \{0, i^\star\},$$

$$\pi^{\mathbf{Q}}(i^\star) = \alpha_{\text{AP}} \frac{p_{i^\star}^d}{\overline{p}_{i^\star}^d} \cdot \pi^{\mathbf{Q}}(0) = \alpha_{\text{AP}} \exp\left(f(Q_{i^\star}^d)\right) \cdot \pi^{\mathbf{Q}}(0).$$

Normalizing $\sum_{s \in S_Y} \pi^{\mathbf{Q}}(s) = 1$ yields the steady-state distribution (7.8) in Lemma 7.2, in which

$$Z = 1 + \sum_{i \in S_Y \setminus \{0, i^\star\}} \alpha_i \exp\left(f(Q_i^{\mathrm{u}})\right) + \alpha_{\mathrm{AP}} \exp\left(f(Q_{i^\star}^{\mathrm{d}})\right).$$

7.8.3 Appendix – Proof of Lemma 7.3: Fluid limit Equations

Equations (7.9)–(7.10) hold for any scheduling algorithm and their proof is standard. Equation (7.9) is obtained by taking the limit in (7.6). Equation (7.10) is by applying the Strong Law of Large Numbers to the arrival process. Further, by the Lipschitz continuity of s, the derivative of s_i^j (denoted by $\mu_i^j(t)$) exists at any regular point t (almost everywhere) and is bounded by its Lipschitz constant (less than one). Equations (7.11)–(7.14) are specific to H-GMS, and we prove them below. We consider two cases depending on the choice of the weight function $f(\cdot)$.

Case 1: $\lim_{x \to \infty} f(x)/\log x = b \in (0, 1)$. Recall from Section 7.3 that Markov chain $\{Y^{\mathbf{Q}(t)}(s)\}_{s \geq t}$ denotes the dynamics of $Y(s)$, assuming a fixed $\mathbf{Q}(s) = \mathbf{Q}(t)$ for all $s \geq t$. Consider a fluid sample path under the H-GMS algorithm. Suppose $\mathbf{q}(t) \neq \mathbf{0}$ at a regular point t. By Lipschitz continuity, we can find a short interval $(t, t + \epsilon)$, such that $\mathbf{q}(\tau)$ is approximately constant ($\approx \mathbf{q}$) for $\forall \tau \in (t, t + \epsilon)$, its actual change being of order ϵ for non-zero queues. This implies that for r large enough, all the queues with non-zero fluid limit $q_i^j > 0$ are of size $\mathcal{O}(q_i^j r)$ in the original process, while all the queues with zero fluid limit are of size $o(r)$ in the original process. Therefore, taking the limit $r \to \infty$ in (7.8), it follows that for any $\mathbf{Q}(\tau)$, $\tau \in (rt, rt + r\epsilon)$, $\pi^{\mathbf{Q}(\tau)} \to \widetilde{\pi}^{\mathbf{q}}$, where

$$\widetilde{\pi}^{\mathbf{q}}(i) = \alpha_i (q_i^{\mathrm{u}})^b / \widetilde{Z}^{\mathbf{q}}, \ i \in S_Y \setminus \{0, i^\star\}, \ \text{and} \ \widetilde{\pi}^{\mathbf{q}}(i^\star) = \alpha_{\mathrm{AP}} (q_{i^\star}^{\mathrm{d}})^b / \widetilde{Z}^{\mathbf{q}}, \ \widetilde{\pi}^{\mathbf{q}}(0) = 1/\widetilde{Z}^{\mathbf{q}},$$

$\widetilde{Z}^{\mathbf{q}} = 1 + \sum_{i'=1}^{N} \alpha_{i'} (q_{i'}^{\mathrm{u}})^b + \alpha_{\mathrm{AP}} (q_{i^\star}^{\mathrm{d}})^b$, and the probabilities are zero for queues which are 0 at the fluid limit. This shows that, with high probability, a queue with a zero fluid limit *cannot* initiate transmission in steady-state. Hence, in equilibrium, the Markov chain never activates an HD link with empty fluid limit queue or an FD link with empty (both) UL and DL queues. Next, we argue that at any $\tau \in (rt, rt + r\epsilon)$, the Markov chain $Y^{\mathbf{Q}(\tau)}$ is at its equilibrium distribution $\pi^{\mathbf{Q}(\tau)} = \widetilde{\pi}^{\mathbf{q}}$ as $r \to \infty$.

Proposition 7.7 (Mixing time of Markov chain $Y^{\mathbf{Q}}$) *Let ν_τ and π denote the instantaneous and the equilibrium distribution of Markov chain $\{Y^{\mathbf{Q}}(\tau)\}_{\tau \geq 1}$, respectively. Given $0 < \zeta < 1$, the mixing time is defined as*

$$T_{mix}(\zeta) := \inf \left\{ \tau \geq 1 : \sup_{s \in S_Y} |\nu_\tau(s) - \pi(s)| \leq \zeta \right\}.$$

Let $\alpha_{min} = \min_i\{\alpha_i\}$ and $Q_{max} = \max_{i,j}\{Q_i^j\}$. Then

$$T_{mix}(\zeta) \leq \frac{2 \exp\left(f(Q_{max})\right)}{\alpha_{min}} \cdot \left[\log\left(\frac{2}{\zeta \alpha_{min}}\right) + f(Q_{max}) \right].$$

Proof 7.5 *The proof follows the application of Raleigh Theorem to characterize the second largest eigenvalue modulus (SLEM) of the transition probability matrix of the Markov chain $Y^{\mathbf{Q}}$. The analysis is similar to [172, Lemma 5] with minor modifications and is omitted.*

Hence for the Markov chain $Y^{\mathbf{Q}(rt)}$, the mixing time is $T_{\text{mix}}(1/r) = \mathcal{O}(r^b \log r)$. This shows that for $b < 1$, the mixing time is sub-linear in r which completely vanishes when taking the average at the fluid scale, i.e.,

$$\frac{1}{\epsilon}\left(s_i^{\mathrm{u}}(t+\epsilon) - s_i^{\mathrm{u}}(t)\right) \approx \frac{1}{r\epsilon} \sum_{\tau=rt}^{rt+r\epsilon} \mathbf{1}(Y^{\mathbf{Q}(tr)}(\tau) = i) \rightarrow \widetilde{\pi}^{\mathbf{q}}(i), \text{ as } r \rightarrow \infty$$

where the second convergence is almost surely by the Ergodic Theorem. This indicates that $\mu_i^{\mathrm{u}}(t) = \widetilde{\pi}^{\mathbf{q}}(i), \forall i \in \mathcal{N}$, and similarly, $\mu_{i^\star}^{\mathrm{d}}(t) = \widetilde{\pi}^{\mathbf{q}}(i^\star)$. This implies that $\mu_i^j(t) = 0$ for $i \in \mathcal{N}_H$ if $q_i^j(t) = 0$, $j \in \{\mathrm{u}, \mathrm{d}\}$, which establishes (7.11). Similarly, considering the coordination among the activation of a pair of FD UL and DL, $\mu_i^j(t) = 0$, $j \in \{\mathrm{u}, \mathrm{d}\}$, $i \in \mathcal{N}_F$, if $\max(q_i^{\mathrm{u}}(t), q_i^{\mathrm{d}}(t)) = 0$, giving (7.12). Also, (7.13) stems from the fact that once an FD UL (or DL) initiates the transmission at rate μ_i^{u} (or μ_i^{d}), the corresponding DL (or UL), if *non-zero*, can follow and transmit at the same rate. Finally, (7.14) is due to the fact that if $\mathbf{q}(t) \neq \mathbf{0}$, no queue that is empty in the fluid limit can initiate transmission at a positive rate and thus the non-empty queues transmit at the maximum sum rate of 1.

Case 2: $\lim_{x \to \infty} f(x)/\log x = b > 1$. The analysis in this case is similar to the analysis of aggressive CSMA algorithms in [173, 174]. Suppose that $\mu_i^j(t) > 0$ and $q_i^j(t) > 0$ for some

186

initiator queue. This implies that for some ϵ, $X_i^j(\tau) = 1$ for $\forall \tau \in (rt, r(t + \epsilon))$, and $Q_i^j(\tau) \geq (q_i^j(t) - \epsilon)r$. The probability that this queue releases the channel after one packet transmission is less than $(Q_i^j(\tau))^{-b}$ which is $\mathcal{O}(r^{-b})$ for $b > 1$. The probability that the link releases the channel during any time $\tau \in (rt, r(t + \epsilon))$ is thus less than $\sum_{\tau=rt}^{r(t+\epsilon)} r^{-b}$ which is $\mathcal{O}(r^{1-b})$ which goes to 0 as $r \to \infty$. This shows that at the fluid limit, if $\mu_i^j(t) > 0$ and $q_i^j(t) > 0$, then $\mu_i^j(t) = 1$. Hence, any positive period of transmission, no matter how short, must be followed by full transmission at rate 1 until the queue has drained on the fluid scale. Furthermore, when the queue hits zero, another non-zero queue will capture the channel without any capture delay (the proof is similar to that of [174, Lemmas 8 and 9]).

This implies that in the heterogeneous HD-FD network, whenever the initiator queue q_i^j belongs to an HD user i, the queue q_i^j drains at full rate 1 until it becomes empty at fluid scale. Whenever the initiator queue q_i^j ($j \in \{u, d\}$) belongs to an FD user i, both its UL and DL queues q_i^u and q_i^d can drain at the maximum rate of 1, until the initiator queue hits zero, at which point both queues release the channel (due to the coordination among a pair of FD UL and DL in Algorithm 7.3). Whenever an HD or FD user releases the channel, another HD or FD user will capture the channel immediately and start transmission at full rate. The choice of which user and which queue captures the channel is randomized over non-zero queues according to access probabilities $\boldsymbol{\alpha}$ and whether $i^\star(t) \in \mathcal{N}_F$ or $i^\star(t) \in \mathcal{N}_H$. Nevertheless, as long as $\mathbf{q}(t) \neq \mathbf{0}$, an HD link with non-zero queue or an FD link with at least one non-zero queue (either UL or DL) will be activated at full rate. This shows that the fluid limits still satisfy (7.11)–(7.14).

Hence, the fluid limit equations hold for both cases.

7.8.4 Appendix – Proof of Proposition 7.5

Recall from Section 7.4, $\overline{Q} = \frac{\mathcal{Q}_{\mathcal{E}_{\max}} + \mathcal{Q}_{\mathcal{E}_{\min}}}{2N} \geq \frac{\mathcal{Q}_{\mathcal{E}_{\max}}}{2N}$. Since the queueing dynamics in $\mathcal{E}_{\max}$ and $\widetilde{\mathcal{E}_{\max}}$ are *not* independent due to the existence of FD users, we wish to find a lower bound on $\mathcal{Q}_{\mathcal{E}_{\max}}$. Denote $Q_{l,\mathcal{E}_{\max}}$ as the queue length of link l at an arbitrary epoch during a *non-serving* interval for the clique $\mathcal{E}_{\max}$. Denote $\mathcal{Q}_{\widetilde{\mathcal{E}_{\max}}} = \sum_{l \in \mathcal{E}_{\max}} \mathbb{E}[Q_{l,\mathcal{E}_{\max}}]$. From the workload decomposition rule [171]

187

applied to a discrete time GI/G/1 system in clique $\mathcal{E}_{\max}$, we have [175]

$$\mathcal{Q}_{\mathcal{E}_{\max}} = \sum_{l \in \mathcal{E}_{\max}} \mathbb{E}[Q_l] = \mathcal{Q}_{\mathcal{E}_{\max}}^{\text{LB}} + \sum_{l \in \mathcal{E}_{\max}} \mathbb{E}[Q_{l,\mathcal{E}_{\max}}] = \mathcal{Q}_{\mathcal{E}_{\max}}^{\text{LB}} + \mathcal{Q}_{\widetilde{\mathcal{E}_{\max}}}. \tag{7.21}$$

Since $\mathcal{Q}_{\mathcal{E}_{\max}}^{\text{LB}}$ and $\mathcal{Q}_{\widetilde{\mathcal{E}_{\max}}}$ are both non-negative, and an immediate lower bound on $\overline{Q}$ is given by

$$\overline{Q} = \frac{\sum_{l \in \mathcal{E}} \mathbb{E}[Q_l]}{2N} = \frac{\mathcal{Q}_{\mathcal{E}}}{2N} \geq \frac{\mathcal{Q}_{\mathcal{E}_{\max}}}{2N} \geq \frac{\mathcal{Q}_{\mathcal{E}_{\max}}^{\text{LB}}}{2N}. \tag{7.22}$$

A key observation to derive the improved lower bound is that assuming the system is stable, in each time slot, the probability that link l transitions from idle state to active state (i.e., *link l is activated*) equals the probability it transitions from active state back to idle state (i.e., *link l is deactivated*). Therefore, for any link $l \in \mathcal{E}_{\max}$,

$$\mathbb{P}\{l \text{ is activated}\} = \mathbb{P}\{l \text{ is deactivated}\}. \tag{7.23}$$

Let ∂_l denote the set of conflicting links of l including link l itself and recall that the access probability $\boldsymbol{\alpha}$ is *fixed* under H-GMS and H-GMS-R. For $\forall l \in \mathcal{E}_{\max}$,

$$\mathbb{P}\{l \text{ is activated}\} = \mathbb{E}[\alpha_l \cdot p(Q_l) \cdot \mathbf{1}(X_{l'}(t) = 0, \forall l' \in \partial_l)]$$

$$\leq \mathbb{E}[\alpha_l \cdot p(Q_l) \cdot \mathbf{1}(X_{l'}(t) = 0, \forall l' \in \mathcal{E}_{\max})] = \alpha_l \mathbb{E}[p(Q_{l,\mathcal{E}_{\max}})] \cdot \mathbb{P}\{X_{l'}(t) = 0, \forall l' \in \mathcal{E}_{\max}\}$$

$$= \alpha_l \mathbb{E}[p(Q_{l,\mathcal{E}_{\max}})] \cdot (1 - \sum_{l' \in \mathcal{E}_{\max}} \mathbb{P}\{X_{l'}(t) = 1\}) = \alpha_l \mathbb{E}[p(Q_{l,\mathcal{E}_{\max}})] \cdot (1 - \sum_{l' \in \mathcal{E}_{\max}} \pi_{l'}), \tag{7.24}$$

where $\pi_{l'}$ is the steady state probability of link l' being active. Similarly,

$$\mathbb{P}\{l \text{ is deactivated}\} = \mathbb{E}[(1 - p(Q_l)) \cdot \mathbf{1}(X_l(t) = 1] = (1 - \mathbb{E}[p(Q_l)]) \cdot \pi_l. \tag{7.25}$$

Plugging (7.24) and (7.25) into (7.23) yields

$$\alpha_l \mathbb{E}[p(Q_{l,\mathcal{E}_{\max}})] \cdot (1 - \sum_{l' \in \mathcal{E}_{\max}} \pi_{l'}) \geq (1 - \mathbb{E}[p(Q_l)]) \cdot \pi_l$$

$$\Leftrightarrow \frac{\mathbb{E}[p(Q_{l,\mathcal{E}_{\max}})]}{1 - \mathbb{E}[p(Q_l)]} \geq \frac{\pi_l/\alpha_l}{1 - \sum_{l' \in \mathcal{E}_{\max}} \pi_{l'}} \geq \frac{\lambda_l/\alpha_l}{1 - \lambda_{\mathcal{E}_{\max}}}, \tag{7.26}$$

where the last inequality comes from the fact that in steady state, $\lambda_l \leq \pi_l$ for $\forall l \in \mathcal{E}_{\max}$. Recall

the definitions of $\lambda_{\min}$ and $\alpha_{\max}$ from Proposition 7.5, it is easy to see that $\min_{l \in \mathcal{E}_{\max}} \lambda_l \geq \lambda_{\min}$ and $\max_{l \in \mathcal{E}_{\max}} \alpha_l \leq \alpha_{\max}$ (under both H-GMS and H-GMS-R. Applying (7.26) to all $l \in \mathcal{E}_{\max}$, we obtain

$$
\begin{aligned}
\sum_{l \in \mathcal{E}_{\max}} \mathbb{E}[p(Q_{l,\mathcal{E}_{\max}})] &\geq \frac{1}{1 - \lambda_{\mathcal{E}_{\max}}} \cdot \sum_{l \in \mathcal{E}_{\max}} \frac{\lambda_l}{\alpha_l} \left(1 - \mathbb{E}[p(Q_l)]\right) \\
&\geq \frac{1}{1 - \lambda_{\mathcal{E}_{\max}}} \cdot \frac{\lambda_{\min}}{\alpha_{\max}} \cdot \sum_{l \in \mathcal{E}_{\max}} \left(1 - \mathbb{E}[p(Q_l)]\right) \\
&\geq \frac{1}{1 - \lambda_{\mathcal{E}_{\max}}} \cdot \frac{\lambda_{\min}}{\alpha_{\max}} \cdot |\mathcal{E}_{\max}| \cdot \left(1 - \frac{\sum_{l \in \mathcal{E}_{\max}} \mathbb{E}[p(Q_l)]}{|\mathcal{E}_{\max}|}\right) \\
&\geq \frac{1}{1 - \lambda_{\mathcal{E}_{\max}}} \cdot \frac{\lambda_{\min}}{\alpha_{\max}} \cdot |\mathcal{E}_{\max}| \cdot \left(1 - p\left(\frac{\mathcal{Q}_{\mathcal{E}_{\max}}}{|\mathcal{E}_{\max}|}\right)\right),
\end{aligned} \tag{7.27}
$$

where the last inequality comes from applying Jensen's inequality to the concave increasing function $p(\cdot)$, i.e.,

$$
\frac{\sum_{l \in \mathcal{E}_{\max}} \mathbb{E}[p(Q_l)]}{|\mathcal{E}_{\max}|} \leq p\left(\frac{\mathcal{Q}_{\mathcal{E}_{\max}}}{|\mathcal{E}_{\max}|}\right).
$$

In addition, the left-hand-side of (7.27) can be upper bounded using Jensen's inequality,

$$
\sum_{l \in \mathcal{E}_{\max}} \mathbb{E}[p(Q_{l,\mathcal{E}_{\max}})] \leq |\mathcal{E}_{\max}| \cdot p\left(\frac{\mathcal{Q}_{\widetilde{\mathcal{E}_{\max}}}}{|\mathcal{E}_{\max}|}\right) \leq |\mathcal{E}_{\max}| \cdot p\left(\frac{\mathcal{Q}_{\mathcal{E}_{\max}}}{|\mathcal{E}_{\max}|}\right), \tag{7.28}
$$

where the last inequality is due to $\mathcal{Q}_{\widetilde{\mathcal{E}_{\max}}} \leq \mathcal{Q}_{\mathcal{E}_{\max}}$ (see (7.21)). Putting together (7.27) and (7.28) yields

$$
\mathcal{Q}_{\mathcal{E}_{\max}} \geq |\mathcal{E}_{\max}| \cdot p^{-1}\left(\frac{\lambda_{\min}/\alpha_{\max}}{1 - \lambda_{\mathcal{E}_{\max}} + \lambda_{\min}/\alpha_{\max}}\right),
$$

and as a result,

$$
\overline{Q} \geq \frac{\mathcal{Q}_{\mathcal{E}_{\max}}}{2N} \geq \left(1 - \frac{N_F}{2N}\right) \cdot p^{-1}\left(\frac{\lambda_{\min}/\alpha_{\max}}{1 - \lambda_{\mathcal{E}_{\max}} + \lambda_{\min}/\alpha_{\max}}\right). \tag{7.29}
$$

Combining (7.22) and (7.29) leads to (7.18), completing the proof.

Part III

Experimentation and Measurements in the COSMOS Testbed

Chapter 8: Introduction to the COSMOS Testbed

Emerging technologies such as millimeter-wave (mmWave), optical x-haul networking, and edge cloud will enable radio access links with ultra-high bandwidth (1 Gbps or higher) and ultra-low latency (at the order of 1 ms), and therefore, usher in a new era of applications. As mentioned in Chapters 1 and 2, city-scale programmable multi-technology testbeds are important for allowing the academic research community as well as the wireless industry to explore these technologies in the most demanding real-world environments.

This chapter presents an overview of the NSF PAWR COSMOS testbed [14] which is designed and being deployed in West Harlem (New York City). Based on the context set in this chapter, in Chapters 9–11, we present specific example pilot COSMOS experiments on open-access full-duplex wireless and converged optical-wireless x-haul networking, as well as extensive 28 GHz channel measurements in the COSMOS testbed area.

We first outline COSMOS' multi-layered computing and network architecture (Section 8.1), where the testbed relies on *open-source software* and *allows for different levels of programmability and control* from the user devices to the cloud (see Fig. 1.4(a)). We then describe the design of COSMOS' key building blocks (Section 8.2), which include: (i) software-defined radios (SDRs) and the IBM 28 GHz programmable phased array antenna modules, (ii) optical transport network, and (iii) core and edge cloud. To provide experimenters with sufficient scale and realism, *COSMOS is being deployed in West Harlem, a representative dense urban environment, with diverse radio propagation and pedestrian/vehicle/traffic characteristics*. We describe COSMOS' 3-phase envisioned deployment plan, which covers a 1 sq. mile dense urban area and includes infrastructure and mobile nodes with three different form factors for different deployment scenario (Section 8.3).

This chapter's contributions were previously presented in [13], and are based on results of a large scale collaboration between Rutgers University, Columbia University, and NYU in partner-

ship with New York City (NYC), City College of New York (CCNY), IBM, University of Arizona, and Silicon Harlem.

8.1 COSMOS Design and Architecture

8.1.1 Multi-Layered Computing Architecture

To provide the required capabilities to experimenters, COSMOS must address new requirements including support for mmWave communications, wideband radio signal processing (500 MHz or more), effective virtualization of radio resources, low latency front- and back-haul, tightly coupled edge cloud, and real-world city-scale deployment. To meet the system requirements and to address these challenges, COSMOS' architecture is based on a *multi-layered computing system* (see Fig. 1.4(a)). It is built in a bottom-up manner with commodity components, programmable hardware, and open-source software. Three types of hardware components are included: (i) SDR nodes (*user devices* and *radio hardware*), (ii) edge cloud servers (*radio cloud*), and (iii) *general purpose cloud*. In particular, COSMOS includes three types of radio nodes (see Section 8.2): *large* (rooftop installed wide-area base stations [BSs]), *medium* (building side or lightpole-mounted micro-cellular BSs), and *small* (fixed or mobile).

COSMOS also integrates edge cloud technology, including commodity CPUs/GPUs/FPGAs, for achieving computing speeds needed to support cloud radio access networks (C-RANs), network function virtualization (NFV), and low-latency cloud applications. Signal processing and NFV can be flexibly partitioned between a local SDR with CPU/FPGA assists (i.e., radio hardware), and a remote C-RAN with CPU/GPU/FPGA assists (i.e., radio cloud). These two computing layers are backed up by the general purpose cloud layer that serves network- and application-layer functions associated with an experiment.

The remote accessibility and open programmability of COSMOS (see Section 8.2) allows users to orchestrate their experiments involving various wired/wireless resources by flexibly configuring the network topologies and computing chains. For instance, Fig. 1.4(a) shows the data paths used in example experiments supported by COSMOS' multi-layered computing architecture (see Sec-
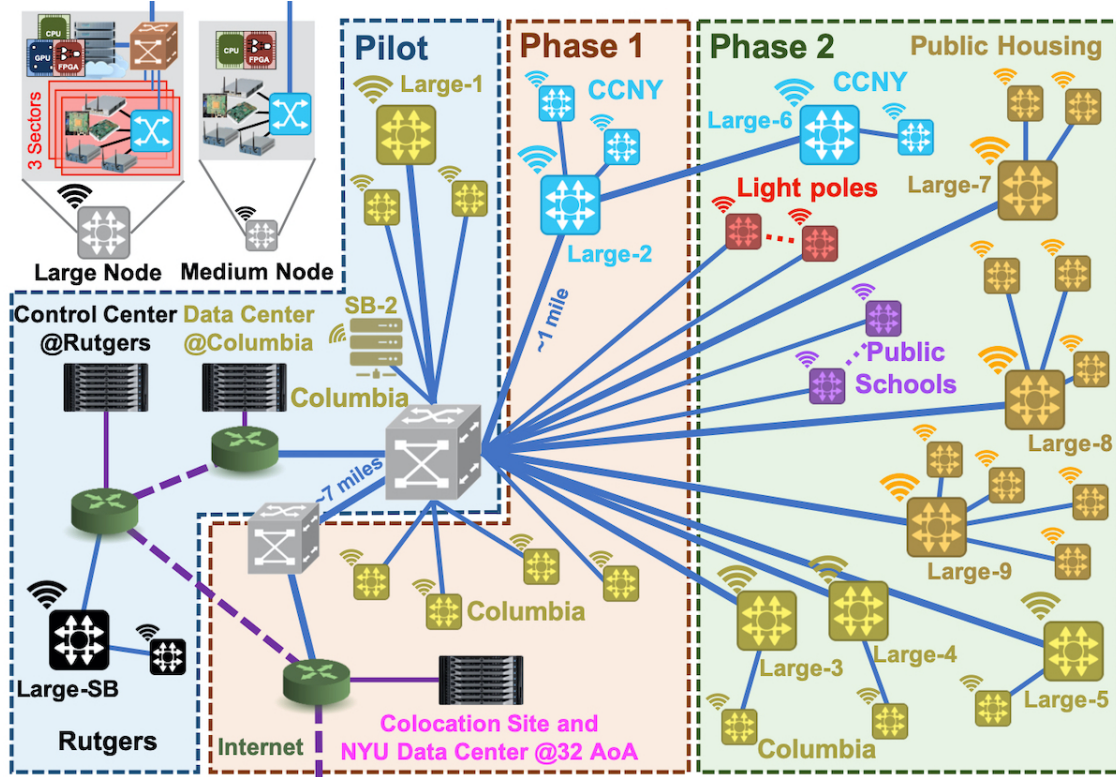
Figure 8.1: COSMOS' network infrastructure, connectivity, and envisioned deployment phases. Components include: large and medium nodes whose colors correspond to the deployment sites, sandboxes (SBs) at Rutgers and Columbia, core optical switching at Columbia (large gray cube), optical switching at the remote data center of NYU in the 32 AoA colocation site (small gray cube), compute at Columbia, NYU data center and Rutgers, optical switching within a large node (brown cube), and the underlying optical transport network (see Section 8.2.3 and Fig. 11.1 in Chapter 11 for details).

tions 9 and 11 for the example experiments): (1) local computing by a medium node, (2) remote computing for a medium node at a nearby large node's servers or at the central computing servers, (3) local computing by large node's computing servers, and (4) remote computing for a large node at the central computing servers. Note that COSMOS can also support a broader set of experiments not limited to the four paths mentioned above (e.g., local computing by a large node using its CPU/FPGA).

8.1.2 Network Infrastructure & Connectivity

To keep pace with the significantly increased wireless link bandwidth and to effectively integrate the emerging C-RANs, COSMOS is designed to incorporate a fast programmable core network for providing connections across different computing layers. COSMOS' core network consists of mostly 100 Gbps+ fiber interconnected with an SDN switching fabric for both minimum latency and flexibility in setting up experimental network topologies. It also integrates advanced optical switching technology based on wavelength-division multiplexing (WDM) switch fabrics and radio over fiber (RoF) interfaces to achieve ultra-low latency connections to edge and central clouds.

As illustrated in Fig. 8.1, the COSMOS testbed includes sites at 4 universities: Rutgers, Columbia, NYU, and CCNY. In particular, Columbia (both the Morningside and Manhattanville campuses) and CCNY are located at the boundary of the targeted West Harlem deployment area of ~1 sq. mile (see also Fig. 1.4(b)). NYU is located in Manhattan and Brooklyn with its data center housed at a colocation site at 32 Avenue of the Americas (32 AoA), which is in Manhattan and is ~7 miles away from the Columbia Morningside campus.

COSMOS is designed to provide an ultra-low latency interconnect between the large/medium nodes and the edge/central cloud resources. Fig. 8.1 shows COSMOS' customized optical fronthaul network, where large and medium nodes (to be deployed at Columbia, CCNY, and NYC assets in West Harlem) will be connected to the core optical switching in the data center at Columbia. As we will describe in Section 8.2.3, based on different experimental requirements, a variety of x-haul optical and networking topologies can be overlaid on the fiber plant using SDN.

The optical network connects the computing rack located at Columbia data center to the collocation site at 32 AoA. The control center (located at Rutgers) is connected to the testbed over the Internet, and includes servers for hosting the user portal, scheduling/virtualization services, experiment management services, data repository, etc. Fig. 8.1 also outlines the phases of the envisioned COSMOS' deployment plan, whose associated challenges (e.g., nodes will be densely deployed in Columbia residential buildings, NYC public housing/schools, and other NYC assets, covering an

194

Advanced PHY techniques: large/distributed MIMO and adaptive beamforming, full-duplex, multi-connectivity, coordinated scheduling, etc.
Conclusively verifying mmWave feasibility for mobility services
Heterogeneity, multi-homing and densification in cellular networks
Dynamic spectrum access (e.g., distributed protocols, directional sharing of mmWave frequencies)
Latency reduction in mobile network PHY/MAC and network layers
Integration of optical x-haul technology with wireless technologies
Clean slate architectures for mobile networks
Adaptive multicast for crowded venues
Edge cloud integration with wireless networks

Table 8.1: Potential research areas that could be studied using the COSMOS testbed.

area of ∼1 sq. mile) are described in Section 8.3.

8.1.3 Supported Experiments

As mentioned in Chapter 1 and above, a major challenge in designing COSMOS is to enable several new classes of wireless experiments that are not supported by existing testbeds. Table 8.1 summarizes the potential research areas that could be studied using the COSMOS testbed at different layers of the protocol stack (PHY to application) and in different network domains (e.g., wireless, wired, and optical). Prior to the community release of COSMOS, we are internally developing a representative set of example experiments that are used to drive the design and validate the capabilities and usability of the testbed. These experiments can also help other researchers envision and plan their own experiments. In Chapters 9 and 11, we present two of our COSMOS pilot experiments on open-access full-duplex wireless (based on the results in Chapter 5) and converged optical-wireless x-haul networking, respectively.

8.2 COSMOS Technology Building Blocks

We now describe COSMOS' key building blocks that have been designed to cope with the challenges outlined above.
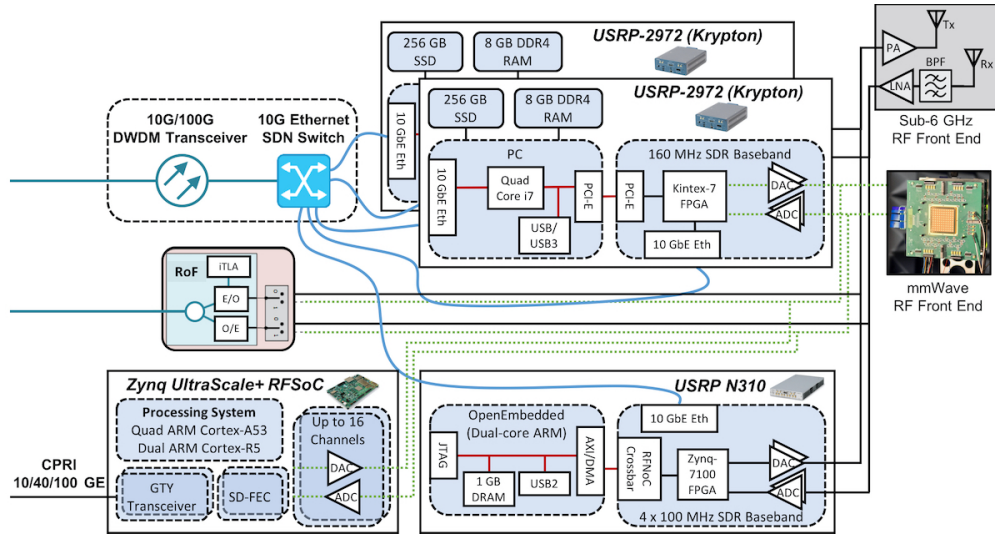
Figure 8.2: Example block diagram of a medium node or a sector of a large node, where different node configurations include a *subset* of the major components (e.g., sub-6/28 GHz RF front ends, FPGA-assisted SDRs, and radio over fiber [RoF] interface).

8.2.1 Software-Defined Radio (SDR) Nodes

The COSMOS SDR nodes in small/medium/large form factors include the same general components (e.g., CPU, FPGA-assisted SDRs, and antennas). They differ by the number of antennas supported, RF bands/bandwidth covered, physical size, and the level of SDR processing provided. Fig. 8.2 shows an example block diagram of an SDR node, where different node configurations include a *subset* of the components. To support various experimental capabilities required by the research community, all the radio nodes offer full flexibility of spectrum use in 400 MHz–6 GHz with a subset of the nodes equipped with 28 GHz mmWave capability (see Section 8.2.2). The nodes also incorporate RF sensing and measurement capabilities to support spectrum use and propagation studies, and operate in conjunction with the edge/central cloud to support different levels of remote processing.

Small Node. The *small* nodes are portable and tailored for installation in a vehicle or carried in a backpack or a cart, and can be used as fixed or mobile clients and for monitoring functions such as spectrum measurement. The hand-held USRP E312 is the most compact. Alternatively, a near-portable platform based on the Intel NUC and USRP B210/B205mini-i or a larger variant based
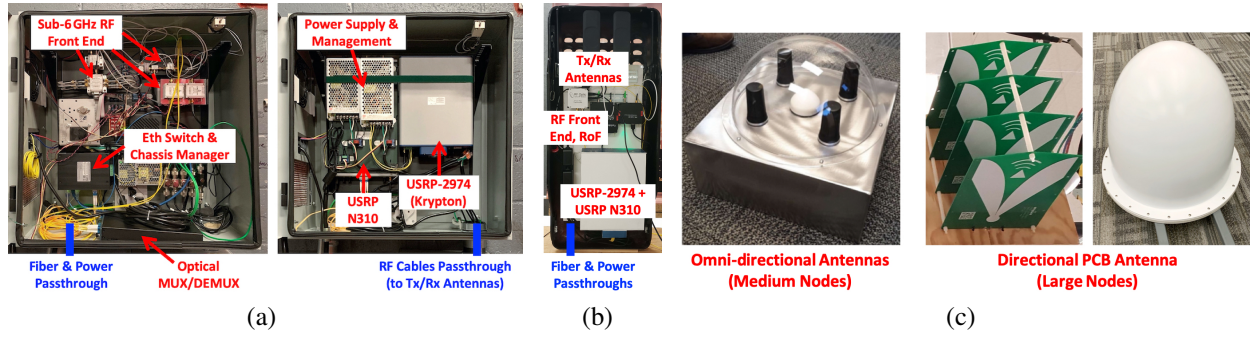
Figure 8.3: (a) Developed and deployed *medium* node (or a sector of a *large* node) with the components mounted on the top shelf (left) and bottom (right) shelf, (b) prototype of the light version of the *medium* node for lightpole mounting (*medium-light*) with the components layout, and (c) sub-6 GHz wideband antennas used for the medium (left) and large (right) node.

on a standard ITX motherboard can provide improved performance.

Medium Node. The *medium* nodes are suitable for street-level building-side and lightpole mounting. They can serve as infrastructure micro-cellular BSs or access points (APs). Fig. 8.3(a) shows a deployed medium node with major modules including USRP-2974/N310, RF front end, Ethernet switches, and fiber connectors (see also Fig. 8.2). Each medium node has a separate antenna package with 4 wideband omni-directional antennas (see Fig. 8.3(c)). A light version of the medium node (*medium-light*) with RoF interface (see Fig. 8.3(b)) will be used for lightpole mounting.

Large Node. The *large* nodes are designed to realize higher powered BSs for wide-area coverage, and are typically deployed on rooftops at 100 feet or higher. A large node consists of *three sectors*, each including identical components as a medium node (with the same three design configurations), and 4 wideband directional antennas, as shown in Fig. 8.3(c). The three sectors are connected to the large node's servers through optical and Ethernet switches.

Sub-6 GHz Transceiver. The core of sub-6 GHz transceiver is based on two types of COTS SDRs. The smaller units (e.g., USRP E312) are associated with reduced baseband bandwidth of up to 50 MHz using USB3 interfaces. The more powerful units, e.g., USRP-2974 as shown in Fig. 8.2, have dual RF daughterboards with 160 MHz baseband and large Xilinx FPGAs.

Since these SDRs span the entire sub-6 GHz band with RF front end complex consisting of power amplifiers (PAs), low-noise amplifiers (LNAs), optional duplexers, etc., the antennas have
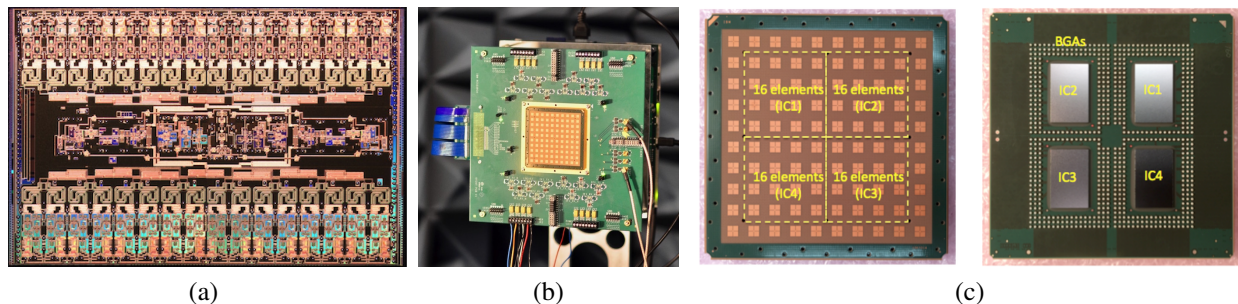
Figure 8.4: (a) The die photograph of the IBM 28 GHz phased array IC implemented in 130 nm SiGe BiCMOS technology, (b) the silicon-based 28 GHz phased array antenna module (PAAM) mounted on a test board, and (c) top and bottom views of a fully assembled 28 GHz PAAM [128].

to be modular to provide efficient coverage. The large/medium nodes have a number of these front ends to support experimental requirements (e.g., full-duplex wireless [8, 31], 900 MHz and 2.4–2.5/2.5–2.7/3.7–4.2/5.1–5.9 GHz). Antennas are selected using an RF switching complex, supporting up to 8 RF front end modules per USRP daughterboard. The RF switching and front ends are controlled by the SDR through the GPIO interface or through an USB/Ethernet-based remote controller.

8.2.2 28 GHz Phased Array Antenna Module

Another key building block of COSMOS is the IBM 28 GHz phased array antenna module (PAAM) [128], which can enable unique mmWave experimentation at the PHY, link, and network layers with multi-beam support and agile beam steering capability, which is currently not supported by other testbeds. Each PAAM employs 4 phased array ICs (see Fig. 8.4(a)) fabricated in 130 nm SiGe BiCMOS technology and 64 dual-polarized antennas that provide eight 16-element or two 64-element concurrent beams with complex modulation formats (e.g., 256QAM). The ICs are assembled with an antenna-in-package array to form the PAAM (see Figs. 8.4(b) and 8.4(c)), which was extensively tested and characterized in lab environments.

We adopt two design configurations using USRP-2974 and the Zynq UltraScale+ RFSoC platform for integration with a subset of COSMOS' radio nodes. Moreover, the baseband will be developed in phases with the initial phase targeting a minimal configuration for 28 GHz bands,

which are likely to be allocated in 100 MHz component carriers with a maximum allocation of 4 carriers. Once the PAAMs are integrated in COSMOS, their multi-beam support and agile beam steering capability can enable experimentation with unique mmWave wireless links and networks based on hybrid beamforming and MIMO. However, since mmWave radio signals experience high path loss, the operation of mmWave networks will require accurate channel models designed for specific deployment sites. In order to inform the deployment of mmWave radios, in Chapter 10, we present our extensive 28 GHz channel measurements in the COSMOS testbed deployment area and results on the coverage characterization and achievable data rates.

8.2.3   Optical Transport Network

A critical building block of COSMOS is the optical transport network and its unique ability to emulate optical front/mid/back-haul (x-haul) networks which, as mentioned in Section 2.3, is currently not supported by other testbeds. It is used to enable high bandwidth and low latency networking and to support experimentation with emerging and future wireless technologies and applications. Using WDM with wavelength switching in colorless, reconfigurable optical add-drop multiplexer (ROADM) unit[1], COSMOS enables a wide range of x-haul networks with different topologies in a real metropolitan setting. In Chapter 11, we describe the detailed design and deployment of COSMOS's optical network and a pilot experiment on wireless handover via optical switching for C-RAN applications.

8.2.4   Core and Edge Cloud

Fig. 8.5 shows COSMOS' cloud architecture including both the core and edge cloud. The edge cloud computing sites contain three types of computing resources: CPUs, GPUs, and FPGAs in order to provide flexible and powerful signal processing capabilities, while also supporting general purpose computing tasks required for different applications. Software tools are provided to support parallel execution on combinations of CPU/GPU/FPGA. The software also allows conve-

---

[1]A ROADM unit enables add, block, pass, or redirect of wavelengths to remotely provision and reconfigure traffic for a WDM network.
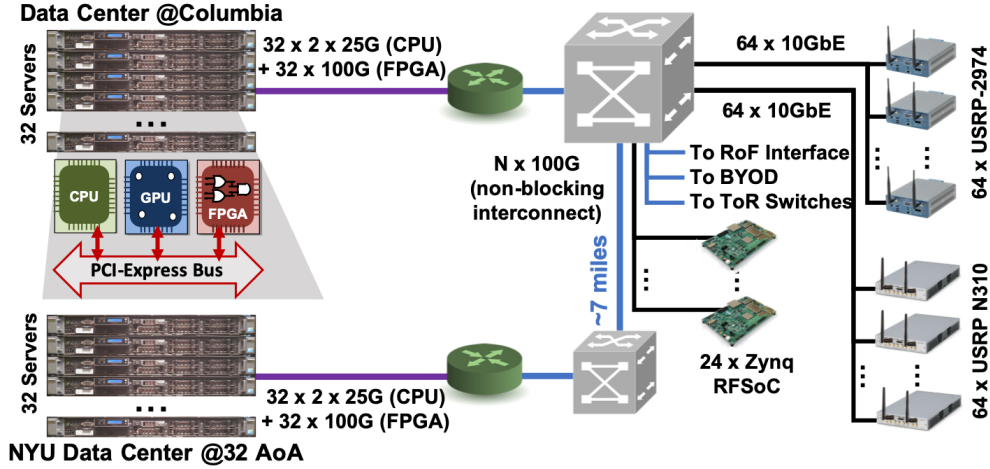
Figure 8.5: Logical topology of COSMOS' cloud architecture overlayed on its optical topology, with aggregated link capacities indicated.

nient software/services development and deployment on a virtualized platform, and access to the "bare-metal" for developing applications/services that need to take full advantage of the ultra-high bandwidth and ultra-low latency of the testbed.

COSMOS' SDN framework integrates native and agent-based control of both wired and wireless resources. This allows SDN experiments to implement application-driven control of optical and data networking functionalities, and radio resources. Further, the SDN framework supports virtualization and allows for logical separation of the same radio or network resource into multiple distinct networks with their own topology and routing protocol [176].

8.3 COSMOS Deployment Plan

To provide realistic environment to experimenters, COSMOS is designed and being deployed to facilitate research and experimentation with advanced wireless technologies in real-world scenarios. Fig. 8.6 shows the vision for the full planned deployment (which is a more detailed version of Fig. 1.4(b)), where a metro dark fiber network and radio nodes (at both sub-6 GHz and millimeter-wave frequencies) will be deployed in a dense urban area of ∼1 sq. mile (covering about 15 city blocks and about 5 city avenues) in the West Harlem neighborhood of Manhattan, in partnership with NYC and the local community. The testbed area includes Columbia University (the Morning-
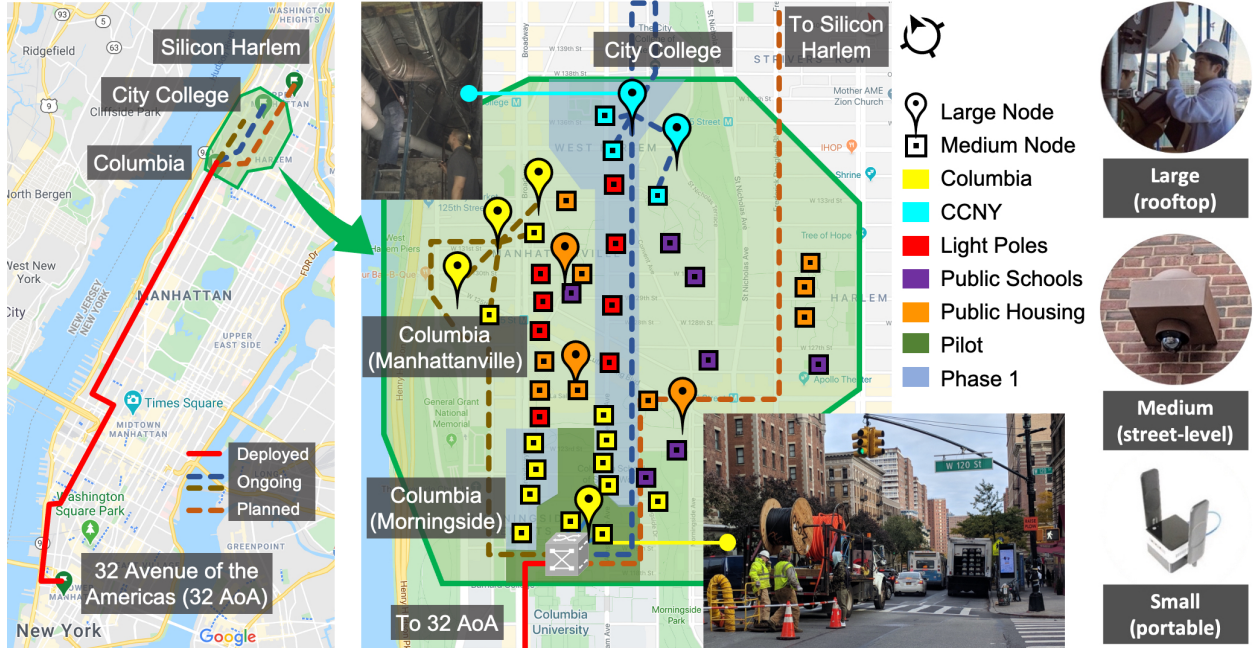
Figure 8.6: (Left) COSMOS' envisioned deployment area of ~1 sq. mile in West Harlem, NYC, covering about 15 city blocks and about 5 city avenues. The deployed and planned dark fiber network is also labeled on the map, which includes a dark fiber ring with ~14 miles distance between Columbia University and 32 Avenue of the Americas (32 AoA) in downtown NYC. (Right) The tentative deployment sites of infrastructure radio nodes in large/medium/small form factors, and the dark fiber interconnections across different sites. The grey cube represents COSMOS' optical switching core located in the Computing Research Facilities (CRF) at Columbia University's Computer Science building (details can be found in Chapter 11 and Fig. 11.1).

side and Manhattanville campuses), City College of New York (CCNY), as well as public housing buildings, public schools, and city lightpoles. In particular, the envisioned deployment includes 9 large nodes (macro-cellular BSs), installed on rooftops, ~40 medium nodes (access points [APs] or micro-cellular BSs), installed at street level (building side, security guard booths, and lightpoles), and ~200 small (near-portable) nodes, used as fixed or mobile devices throughout the testbed.

The location of the testbed was selected due to its proximity to campuses (where buildings can be used for deployment, students can be opt-in users, and university-controlled vehicles circulate), concentration of NYC assets, and the existence of an engaged community of educators, organizers, and entrepreneurs. Fig. 1.4(b) illustrates the vision for the full planned deployment with ~49 infrastructure radio nodes (9 large and ~40 medium) at the conclusion of Phase 2 (see Section 8.1). Large radio nodes (macro-cellular BSs) will be installed on rooftops. Medium nodes

Figure 8.7: (a) The construction of the Pilot Phase large node (Large–1) on an 18th floor rooftop at Columbia, (b)–(c) Pilot Phase deployment of a large/medium node on the 18th/2nd floor (see also Fig. 8.3), and (d) the central optical core switching (including the Calient S320 space switch and several ROAMD units) in the data center at Columbia.

(APs or micro-cellular BSs) will be installed at street level (building side, security guard booths, and lightpoles). Finally, ~200 small (near-portable) nodes will be used as fixed or mobile devices at Columbia, CCNY, and the local community, including in vehicles (e.g., public safety cars and shuttles) and used for educational purposes. These nodes will serve as experimental network users and could be used by university students and staff, high school students, and public housing residents. Dark fiber to the various locations as indicated in Fig. 8.1 will be provided/deployed by Columbia, CCNY, NYC, and potentially by the PAWR industry consortium.

The deployment consists of three phases (see Fig. 8.1):

Pilot Phase (completed in May 2019, see Fig. 8.7): a proof of concept phase that targets functional verification of the designs for all three node types as well as initial performance evaluation (e.g., radio coverage planning). It included the deployment of 2 large, 3 medium, and 30 small nodes. It also included the control center at Rutgers, the central optical core switching (a Calient S320 space switch and 6 ROADM units) and computing (ToR switches and servers) at Columbia, the dark fiber to the colocation site at 32 AoA, as well as two sandboxes that provide isolated testing environments. In particular, the sandbox at Rutgers includes the Large–SB node, a medium node, and compute. It also includes a pair of Sivers IMA 60 GHz radios and a pair of InterDigital Edge-Haul 60 GHz radios. The sandbox at Columbia includes several SDRs, ToR switches, servers, and

the alpha version of the IBM 28 GHz PAAMs (see Section 8.2.2).

Phase 1 (expected during 2021): deployment of 3 large, 8 medium, 60 small nodes (all sub-6 GHz and with mmWave at about 20 nodes) in both Columbia and CCNY campus buildings as well as Columbia residential buildings. In addition, optical switching and computing will be deployed at the NYU data center.

Phase 2 (timeline for completion based on availability of resources): complete deployment of all nodes (all sub-6 GHz and with mmWave at ~25% of the nodes).

Moreover, FCC has recently designated part of the COSMOS testbed area as one of the country's first two Innovation Zones [177] with allowable frequency bands of 2500-2690 MHz, 3700–4200 MHz, 5850–5925 MHz, 5925–7125 MHz, 27.5–28.35 GHz, and 38.6–40.0 GHz. Under this initiative, parties will have the flexibility to conduct multiple non-related experiments under a single authorization within a defined geographic area while protecting incumbent services against harmful interference. This initiative also allows experimental program license holders which are licensed to operate elsewhere to use the NYC Innovation Zone.

8.4   Conclusion

In this chapter, we presented the preliminary experience gained in designing a 1 sq. mile city-scale *programmable* advanced wireless testbed, COSMOS, which incorporates SDRs, mmWave radios, an optical transport network, and edge cloud. We outlined our approaches to addressing many of the challenges and provides insight into the research areas that can be studied in the testbed once operational. We also discussed the unique challenges of deploying the testbed in a dense urban environment. In the following chapters, we present 2 illustrative example experiments, which consider different network layers, different data paths, and different network domains (see Fig. 1.4(a)). Tutorials outlining these experiments are available at [178]. In particular, in Chapter 9, we present an experiment on *open-access full-duplex wireless (data paths 1 and 2)*. In Chapter 10, we present extensive 28 GHz channel measurements in the COSMOS testbed area. In Chapter 11, we present an experiment on *converged optical-wireless x-haul networking (data paths 3 and 4)*.

Chapter 9: Open-Access Full-Duplex Wireless in the COSMOS Testbed

As mentioned in Chapter 1 and Part II, full-duplex (FD) wireless has drawn significant attention [8, 31, 76] as an enabler of next-generation wireless networks. Within the Columbia Flex-ICoN project [11], we focus on the design of and experimentation with FD radios and systems grounded in integrated circuit (IC) implementations, which are suitable for hand-held land form-factor-constrained devices [31]. In [36], we presented the 1st-generation (Gen-1) *narrowband* FD radio and an FD link, featuring 40 dB RF SIC across 5 MHz. This Gen-1 RF SI canceller emulates its RFIC counterpart [151], whose modeling and analysis were presented in [95]. In Chapter 5, we presented the 2nd-generation (Gen-2) *wideband* FD radio which leverages the technique of frequency-domain equalization (FDE) to achieve over 50 dB RF SIC across 20 MHz, again emulating its RFIC counterpart [151].

In order to allow the broader community to experiment with FD wireless, we integrated the two generations of FD radios in the open-access ORBIT [12] and COSMOS wireless testbeds, as shown in Fig. 9.1. As mentioned in Chapter 5, since interfacing an RFIC canceller with a software-defined radio (SDR) presents numerous technical challenges, we implemented the RF cancellers on printed circuit boards (PCBs) to facilitate cross-layered experiments with an SDR platform. We first integrated an improved version of the Gen-1 RF canceller with a USRP N210 SDR in the ORBIT testbed [12]. We then integrated two Gen-2 RF cancellers (an improved version of that presented in Chapter 5) with USRP2 SDRs in the COSMOS testbed.

In this chapter, we first describe the design of open-access FD radios (Section 9.1). We then present the cross-layered (hardware and software) design and integration of the remotely accessible Gen-1 *narrowband* FD radio in the ORBIT testbed (Section 9.2) and Gen-2 *wideband* FD radios in the COSMOS testbed (Section 9.3). Finally, we present two example experiments implemented in GNU Radio [179] that are available to experimenters (Section 9.4): (i) real-time digital SIC, and

(a) Gen-1 narrowband FD radio in the ORBIT testbed.   (b) Gen-2 wideband FD radio in the COSMOS testbed.
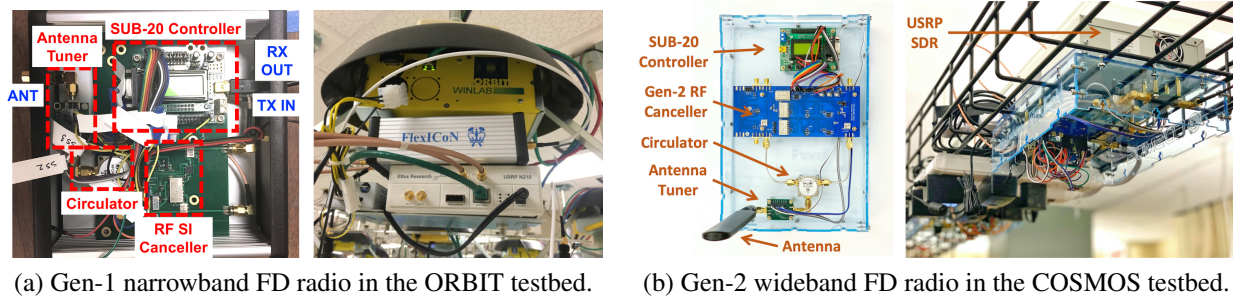
Figure 9.1: The two generations of integrated FD radios. Each FD radio consists of an antenna, a customized Gen-1/2 RF canceller box, a USRP SDR, and a remotely accessible compute node. The RF canceller boxes each contain an antenna tuner, a circulator, a canceller PCB, and a SUB-20 controller.

(ii) measurements of the link-level packet reception ratio (PRR). These experiments show that our Gen-1/Gen-2 FD radios integrated in ORBIT/COSMOS can achieve an overall amount of 80 dB SIC across 5/12.5 MHz bandwidth, respectively. These integrated remotely accessible FD radios and example experiments also can be further extended to more complicated networking scenarios.

The results in this chapter were presented in [37, 38] and are based on joint work with Ph.D. students Mahmood Baraani Dastjerdi, Manav Kolhi, Jin Zhou, M.S. student Guy Farkash, and undergraduate student Jackson Welles. Jakub Kolodziejski, Michael Sherman, and Ivan Seskar provided support for the integration of the open-access FD radios in the ORBIT and COSMOS testbeds.

9.1 Design of the FlexICoN Open-Access FD Radios

Fig. 9.2 shows the block diagram shared by both generations of FD radios. It consists of three main components: a Gen-1/2 RF canceller box, a USRP SDR, and a remotely accessible compute node. The SDRs used in ORBIT and COSMOS are the USRP N210 and USRP2, respectively, both of which have the same RF frontend. These SDRs can operate up to 25 MHz bandwidth, and are controlled via GNU Radio [179] and UHD [180].

Both Gen-1 and Gen-2 canceller boxes share the same overall design as shown in Figs. 9.1 and 9.2. The boxes have four components: an antenna tuner, a circulator, a SUB-20 controller, and
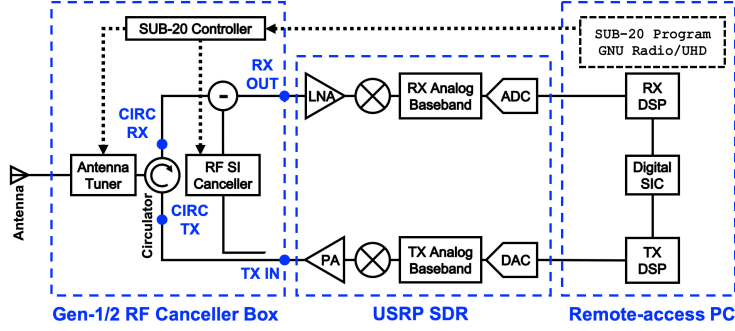
Figure 9.2: Block diagram of the integrated FD radios with three main components: (i) a FlexICoN RF canceller box, (ii) a USRP software-defined radio (SDR), and (iii) a remotely accessible host PC on which experiments are run.
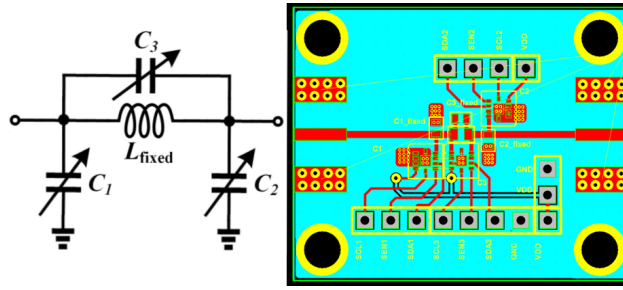
Figure 9.3: Circuit diagram and PCB layout of the programmable antenna tuner as part of the Gen-1/Gen-2 RF canceller box.

an RF canceller PCB which differs between Gen-1 and Gen-2.

Programmable Antenna Tuner. In order for the circulator to better match with the varying impedance of the antenna due to environmental changes around 900 MHz operating frequency, we also designed and implemented a programmable antenna tuner. Fig. 9.3 shows the circuit diagram and the PCB implementation of the antenna tuner. In particular, a π-network with lossless inductor (L) and digitally tunable capacitors (C_i) is used for impedance transformation. We use a fixed chip inductor with inductance $L_{\text{fixed}} = 5.1$ nH and the Peregrine Semiconductor 5-bit PE64909 digitally tunable capacitors for C_i ($i = 1, 2, 3$). By programming the capacitors with code values CAPi ($i = 1, 2, 3$), different antenna interface impedance matching can be achieved. The configuration ranges of the tunable capacitors are CAPi $\in \{0, 1, \cdots, 16\}$, $\forall i = 1, 2, 3$.

Coaxial Circulator. An RF-CI RFCR3204 coaxial circulator [181] is used, with operating frequency between 860–960 MHz.

206
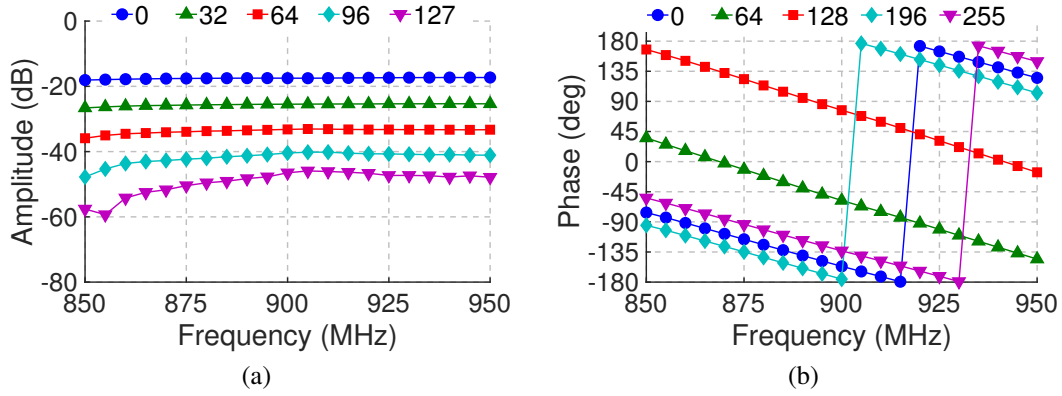
Figure 9.4: Measured amplitude and phase response of the Gen-1 RF canceller PCB with (a) varying attenuation values, and (b) phase shift values.

SUB-20 Controller. As Fig. 9.2 shows, a DIMAX SUB-20 multi-interface USB adapter [182] connected to the remote-access PC is used to program the tunable components on the RF canceller and antenna tuner through SPI. The SUB-20 SPI is configured to operate at the maximal SPI clock of 8 MHz. The SUB-20 control for both the Gen-1 and Gen-2 canceller PCBs is handled through a customized GNU Radio out-of-tree (OOT) module (see Section 9.4.1 for details).

RF Canceller PCB. In order to meet the USRP Rx front-end linearity and the analog-to-digital converter (ADC) dynamic range requirements, sufficient SIC in the RF domain is needed before digital SIC is performed. Therefore, the RF canceller PCB must provide up to 30 dB RF SIC in addition to the 20–25 dB provided by the circulator. The RF canceller taps a reference signal from the output of the power amplifier (PA) at the Tx side, and SIC is performed at the input of the low-noise amplifier (LNA) at the Rx side. The difference between the two generations of canceller boxes is the circuit design of the RF canceller PCB, as discussed in Sections 9.2.1 and 9.3.1.

9.2 Gen-1 Narrowband FD Radio in the ORBIT Testbed

The Gen-1 FD radio is depicted in Fig. 9.1(a), including the components described in Section 9.1. In particular, the RF canceller PCB is a *narrowband* frequency-flat amplitude- and phase-based canceller, which is an improved version of that presented in [36]. In this section, we present the design and implementation of the Gen-1 FD radio, and its integration in the ORBIT testbed.

Figure 9.5: Measured Tx/Rx isolation of the Gen-1 RF canceller box with and without turning on the RF canceller. The RF canceller box with the circulator and the RF canceller provides 40 dB RF SIC across 5 MHz bandwidth.

9.2.1 Gen-1 RF Canceller PCB

The amplitude- and phase-based RF canceller is implemented using discrete components on a PCB and is optimized around 900 MHz operating frequency.[1] The reference signal is tapped through a 6 dB Mini-Circuits ADC-6-13+ directional coupler whose amplitude and phase are subsequently adjusted before SIC is performed at the Rx.

For amplitude adjustment, a 7-bit SKY12343-364LF digital attenuator is used, and a Mini-Circuits passive SPHSA-152+ phase-shifter is used for phase adjustment. The SPHSA-152+ is controlled by an 8-bit TI-DAC081S101 digital-to-analog converter (DAC). The attenuator and DAC have 3 V supply voltage and the phase shifter has a reference voltage of 12 V. Both the attenuator phase shifter are programmed through the SUB-20 controller over SPI. The available parameter configuration ranges for the attenuator `ATT` (<u>ATT</u>uation) and phase shifter `PS` (<u>P</u>hase <u>S</u>hift) are:

$$\mathtt{ATT} \in \{0, 1, \cdots, 127\}, \mathtt{PS} \in \{0, 1, \cdots, 255\}.$$

Fig. 9.4 shows the amplitude and phase measurements of the RF canceller with varying `ATT` values (under fixed `PS` = 0) and with varying `PS` values (under fixed `ATT` = 0). The RF canceller has an amplitude tuning range between -48 dB and -17 dB, and a phase tuning range spanning

---

[1]In this implementation, we select 900 MHz operating frequency but this approach can be easily extended to other frequencies (e.g., 2.4/5 GHz).
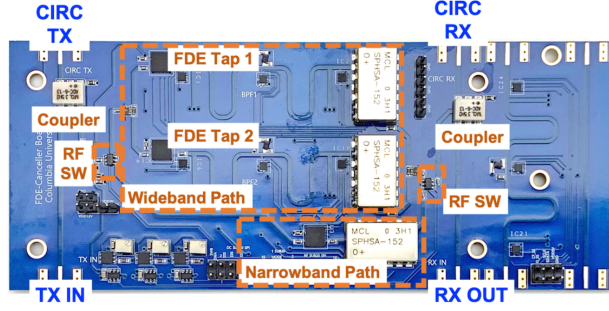
Figure 9.6: The Gen-2 RF canceller PCB.

(a) ATT   (b) PS   (c) CF   (d) QF
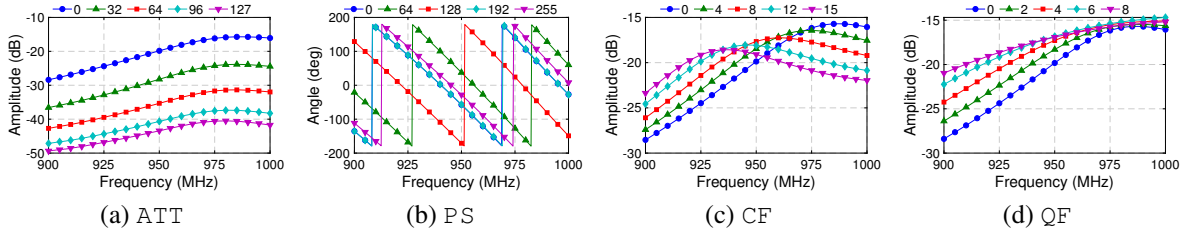
Figure 9.7: Measured response of one FDE tap with varying ATT, PS, CF and QF values. The second FDE tap shares the same behavior.

the entire 360 deg range. This matches well to the 31.75 dB stated in the SKY12343 specification and full 360 deg range expected for the SPHSA-152+. Fig. 9.5 shows an example of the measured Tx/Rx isolation (measured between Tx IN and Rx OUT ports shown in Fig. 9.2), where 40 dB RF SIC is achieved across 5 MHz bandwidth when the canceller PCB is turned on.

9.2.2 Integration in the ORBIT Testbed

An ORBIT node equipped with the Gen-1 RF canceller box is depicted in Fig. 9.1(a). We use node11-10 (the yellow box in Fig. 9.1(a)) in the ORBIT main grid with a USRP N210 SDR. In particular, the RF canceller box Tx IN/Rx OUT ports are connected to the USRP Tx/Rx ports respectively, and the RF canceller box ANT port is connected to an Apex II multi-band antenna. The USRP has a receiver noise floor of -86 dBm at 20 MHz bandwidth.

We developed a node image (flexicon-orbit-v3.ndz) which contains the two example GNU Radio FD experiments. These experiments run on the node11-10 compute node. The example experiments are discussed in Section 9.4.

(a) With $50\,\Omega$ termination. (b) With real antenna.
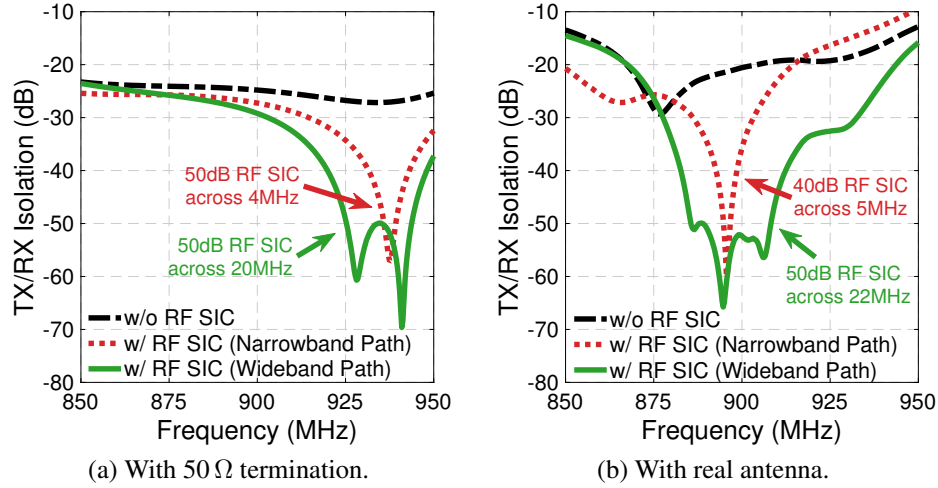
Figure 9.8: Measured RF SIC achieved by the Gen-2 RF canceller PCB without turning on the canceller, and with the narrowband (Gen-1)/wideband (FDE) path on, when the circulator antenna port is: (a) terminated by $50\,\Omega$, and (b) connected to an antenna.

9.3 Gen-2 Wideband FD Radios in the COSMOS Testbed

The Gen-2 FD radio is depicted in Fig. 9.1(b). The Gen-2 RF canceller box includes an FDE-based RF canceller PCB that is an improved version of the canceller described in [22].[2] This canceller PCB can achieve enhanced cancellation performance over a significantly wider bandwidth compared to the Gen-1 canceller PCB, and thus allows for experimentation with wider signal bandwidth. In this section, we present the design and implementation of the Gen-2 FD radio, and its integration in the COSMOS testbed.
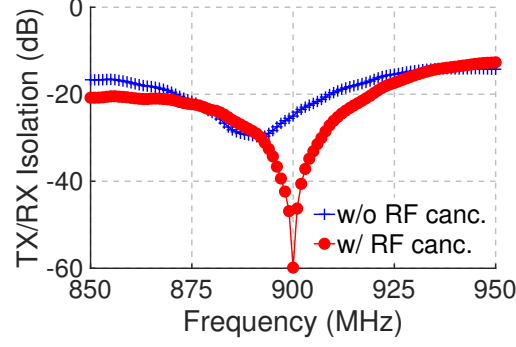
9.3.1 Gen-2 RF Canceller PCB

The Gen-2 RF canceller illustrated in Fig. 9.6 is implemented using discrete components on a PCB and is optimized around a $900\,\mathrm{MHz}$ operating frequency.[3] The reference signal coupled from the Tx is first fed through an Analog Devices (ADI) HMC374 LNA before passing through an ADI HMC221B RF switch . This RF switch switches the reference signal between two paths: a Gen-1 narrowband path (identical to the Gen-1 RF canceller described in Section 9.2.1) and the wideband

---

[2]The detailed design and evaluation of this canceller can be found in [22]

[3]As before, the design can be modified to any target frequency desired.

FDE path. In particular, the FDE path contains two parallel FDE taps, split and combined using Anaren PD0810J5050S2HF power dividers. Each FDE tap consists of a tunable bandpass filter (BPF) with amplitude and phase control, and is used to perform FDE of the SI channel.

Table 9.1 summarizes the tunable components on the Gen-2 canceller PCB. Each of the two paths contains an amplitude and phase control, identical to those on the Gen-1 canceller. Furthermore, BPF is controlled by two tunable capacitors, which control the center frequency and quality factor of the filter. The center frequency capacitor is the PE64102, tunable between 1.88 pF and 14.0 pF. The quality factor capacitor is the PE64909, tunable between 0.6 pF and 2.35 pF. The Gen-1 path contains an amplitude and phase control operating independently of the Gen-2 components. Together, these components provide a total of 2^{46} possible configurations for the Gen-2 path and 2^{15} configurations for the Gen-1 path. The SUB-20 controller is used to configure the Gen-2 canceller PCB over SPI.

Fig. 9.7 shows the effect of varying each parameter on the frequency response of one individual BPF. In each case, only one parameter is varied, while the others are kept at the lowest value. Fig. 9.8 shows the achievable RF SIC of the wideband FDE path. >50 dB RF SIC is achieved across 20 MHz bandwidth. This is 10 dB higher RF SIC achieved over $4\times$ the bandwidth when compared to the Gen-1 path.

9.3.2 Integration in the COSMOS Testbed

The Gen-2 RF canceller box is integrated in the indoor COSMOS Sandbox 2 located at Columbia University. The antenna port is connected to a Larsen SPDA24700 LTE antenna [183], and the canceller box is connected to a USRP2 SDR. We integrate two Gen-2 radios, which form an FD link with 5 ft distance between antennas. Both Gen-2 FD radios are connected to the remotely accessible compute node: the SDR is connected over Ethernet and the SUB-20 controller over USB. The integration is shown in Fig. 9.1(b) and both identical Gen-2 FD radios are shown within the context of COSMOS Sandbox 2 in Fig. 9.9. The two USRP2s are synchronized over a MIMO cable.

The remotely accessible compute node is equipped with an i7 3770 8-core processor and 16 GB

Scope	Control	Name	Tuning range
PCB	RF Switch	SW	$\{0, 1\}$
Gen-2 Path	Amplitude	ATTi	$\{0, \cdots, 127\}$
	Phase	PSi	$\{0, \cdots, 255\}$
	Center Frequency	CFi	$\{0, \cdots, 15\}$
	Quality Factor	QFi	$\{0, \cdots, 31\}$
Gen-1 Path	Amplitude	ATT2	$\{0, \cdots, 127\}$
	Phase	PS2	$\{0, \cdots, 255\}$

Table 9.1: Summary of tunable components for the Gen-2 RF Canceller. $i \in \{0, 1\}$ represents the two FDE taps on the Gen-2 path.

RAM, and runs Ubuntu 16.04, GNU Radio 3.7, and UHD 3.14 [179, 180]. This specification gives the compute node the capability of running real-time experiments with a graphical user interface (GUI) above 10 MHz bandwidth. If the GUI is omitted, the bandwidth can be increased up to 25 MHz, which is the limit of the USRP2 SDRs. Experiments conducted without the GNU Radio GUI still support output via the terminal.

9.4 Remote Experimentation With GNU Radio

Recall from Chapter 1 that the core feature of these integrated FD radios is that they may be remotely accessed by experimenters. The process for remotely accessing the FD radios is the same for both ORBIT and COSMOS: the experimenter logs into the testbed from their local machine using SSH. The tutorial with detailed instructions can be found at [184]. X11 GUI forwarding is supported, allowing the experimenter to use GNU Radio as if it were running on the local machine.

Example experiments implemented in GNU Radio are provided for both the ORBIT and COSMOS FD radios. These example experiments run in *real-time*, meaning that the experimenter can observe results visualized without offline processing. The example experiments employ an OFDM PHY layer [185] with variable bandwidth. The example experiments described below can be used to benchmark several performance metrics of the ORBIT Gen-1 and COSMOS Gen-2 FD radios, including node-level SIC and link-level PRR.
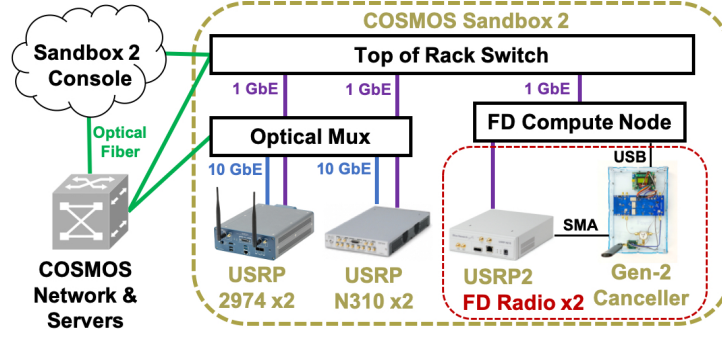
Figure 9.9: The architecture of the COSMOS Sandbox 2 (see also Fig. 8.1), which includes (i) the remotely accessible console and two FD radios, each consisting of a USRP2 SDR and the Gen-2 canceller box, connected to the FD compute node, and (ii) other remotely accessible SDRs used for experimentation with various technologies.

9.4.1 Experiment 1: Node-Level Digital SIC

In this experiment, the experimenter can view a visualization of the real-time performance of one FD radio when transmitting OFDM packets [185]. The following data is visualized: (i) the time domain Rx signal after RF SIC and after digital SIC, (ii) the power spectrum of the Rx signal after RF SIC and after digital SIC, (iii) the digital SIC filter taps, and (iv) the constellations after RF SIC and after digital SIC.

As mentioned above, the provided GNU Radio experiments operate in real-time. This is accomplished by the use of customized C++ out-of-tree (OOT) blocks, which are optimized for real-time performance. The implemented OOT blocks are available at [186] and described below:

SUB-20 Control. A core part of this experiment is the ability of the experimenter to manually configure the canceller PCB to achieve different RF SIC profiles. The experimenter has access to a GUI with controls that allow the full range of input described in Table 9.1. The numerical values set by these GUI controls are input to the SUB-20 control OOT block, which then uses the SUB-20 API [182] to generate the appropriate SPI frames to be transmitted to the RF canceller PCB. In total for the Gen-1 RF canceller, 5 two-byte frames are transmitted to fully program the board, and for the Gen-2 RF canceller, 14 two-byte frames are transmitted. With an 8 MHz SPI clock, one configuration of the Gen-1 and Gen-2 canceller PCBs takes 10 us and 30 us, respectively. Therefore, different RF SIC profiles may be visualized in real time.
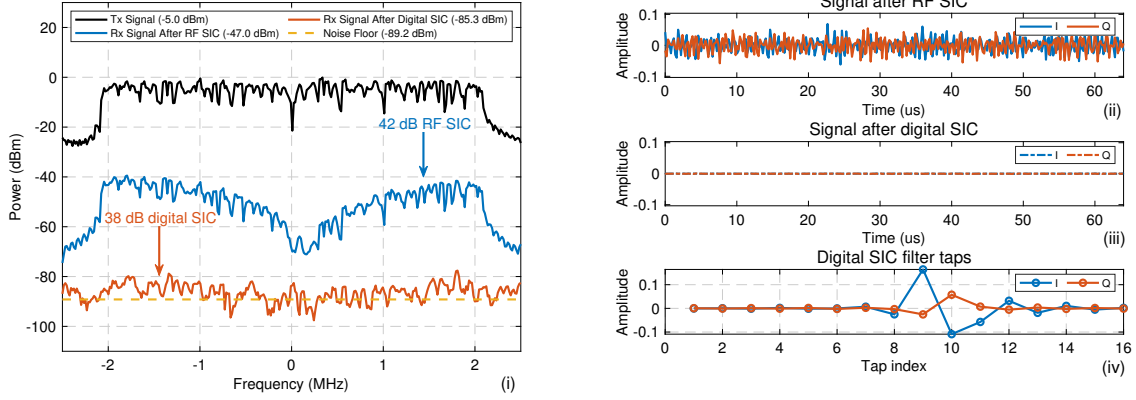
(a) Gen-1 FD radio. Carrier frequency is the 909 MHz ISM band, and the bandwidth is 5 MHz.
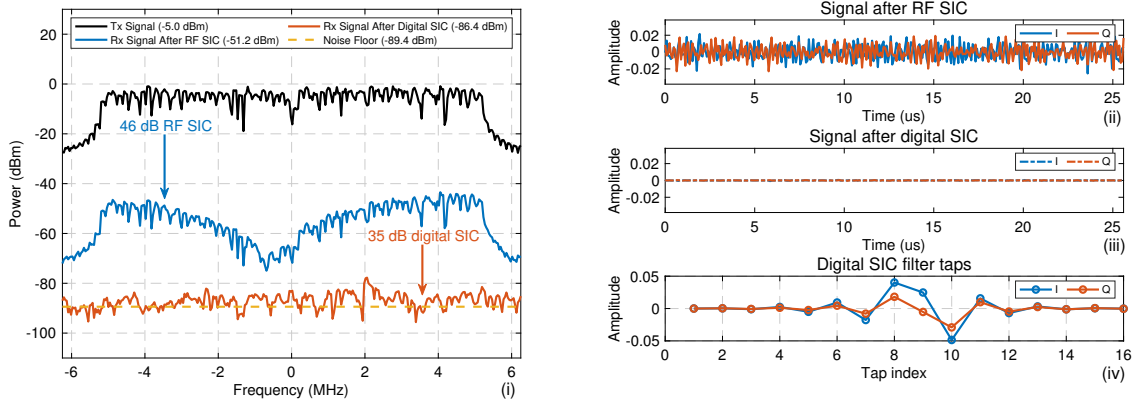
(b) Gen-2 FD radio. Carrier frequency is the 915 MHz ISM band, and the bandwidth is 12.5 MHz.

Figure 9.10: Node-level SIC performance for (a) ORBIT Gen-1 FD radio, and (b) COSMOS Gen-2 FD radio. Data shown are (i) power spectrum of the received signal after SIC in the RF and digital domains; (ii), (iii) time domain signals after RF and after digital SIC; (iv) digital SIC filter taps. The average Tx power is -5 dBm, and the receiver noise floor is -89 dBm.

Digital SIC. Digital SIC is performed on a per-packet level using a least-squares algorithm to estimate the SI channel in the time domain. The channel is estimated using additional pilot OFDM symbols. The OOT block performs both the least-squares channel estimation and digital SIC with the following computation: $\hat{y_{Rx}} = y_{Rx} - \mathbf{A}\hat{h}$. Here, $\hat{y_{Rx}}$ is the residual SI after RF and digital SIC, y_{Rx} is the signal after RF SIC, $\mathbf{A}$ is the the Toeplitz matrix constructed from the transmitted signal, and $\hat{h}$ is the estimated SI channel.

Packet Encapsulation. The implemented digital SIC algorithm requires an additional set of pilot symbols prepended to every packet. This OOT block prepends these symbols, as well as adds zero padding between packets, so that the packet transmission rate is reduced and can be handled by
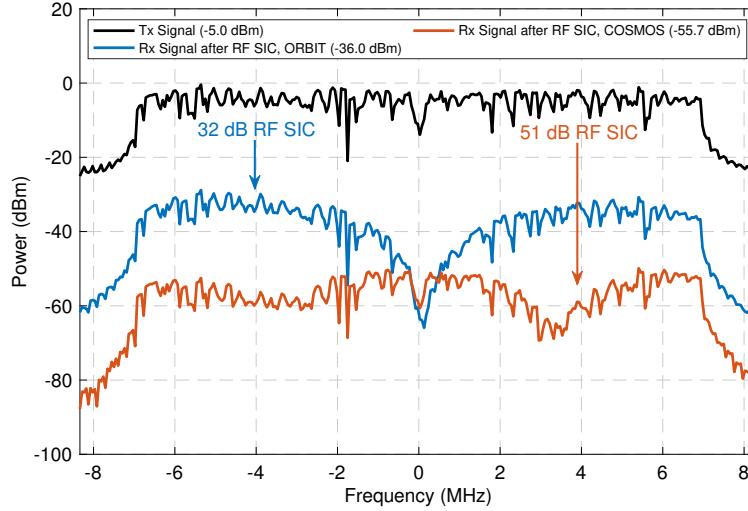
Figure 9.11: RF SIC achieved over 16.6 MHz for the Gen-1 and Gen-2 FD radios. The Gen-2 wideband FD radio can achieve a greater RF SIC over a much wider bandwidth.

remote-access compute node controlling the FD radio.

Constellation Diagram. A key visualization of the Rx signals before and after digital SIC is the constellation diagrams. A clear constellation should still be visible after RF SIC, as sufficient cancellation of the Tx signal has not yet been achieved. At this stage, the SI can often still be decoded. After digital SIC, there should be discernible constellation, only noise, as the SI has been cancelled close to the noise floor.

SIC Performance. Fig. 9.10 shows the node-level SIC performance for the Gen-1 FD radio in ORBIT and one of the Gen-2 FD radios integrated in COSMOS. By configuring the RF cancellers appropriately, the Gen-1 FD radio can achieve 35–40 dB RF SIC across 3–5 MHz, and the Gen-2 FD radio can achieve 45–50 dB RF SIC across 15–20 MHz. In both cases, the RF SIC is followed by 35–40 dB of digital SIC for around 80 dB overall SIC. We will replace the USRP2s with higher performance USRP 2974s, which should improve the amount of digital SIC achieved.

The RF SIC performance alone of the integrated Gen-1 and Gen-2 FD radios is illustrated over a wider bandwidth in Fig. 9.11, which clearly shows the Gen1/Gen2 narrowband/wideband performances. Across the 16.6 MHz bandwidth shown in the figure, the Gen-1 FD radio can achieve 30 dB RF SIC, whereas the Gen-2 FD radio can achieve 50 dB. For the Gen-1 FD radio, most of the RF SIC outside of 5MHz bandwidth is achieved by the circulator alone.

215

9.4.2 Experiment 2: Link-Level FD Packet Reception Ratio (PRR)

The experimenter can also use the two Gen-2 FD radios to perform link-level experimentation. In this experiment, link PRR as a function of the signal-to-noise ratio (SNR) is measured. For simplicity, one Gen-2 FD radio is run in FD mode, and the other is run in half-duplex (HD) mode. Using only one radio in FD mode alleviates the computational load on the remote-access compute node, and the link-level PRR can still be measured. The canceller PCB of the radio in FD mode must first be configured to achieve the required level of SIC. The experimenter can accomplish this by running Experiment 1 and using the SUB-20 OOT block and its GUI to generate the desired RF SIC profile. Digital SIC is achieved using the custom OOT block described in Experiment 1.

The experiment runs in two modes: HD or FD. In both modes, The Tx power of the HD radio is swept during the experiment. In HD experiment mode, the FD radio does not transmit and only receives, and in FD experiment mode, the FD radio both receives and transmits with a constant Tx power. At each HD radio Tx power used in the experiment, a measurement with 1,024 OFDM packets (with QPSK modulation and no channel coding) is taken.

We consider two performance metrics. The first is the *HD/FD link SNR*, which is measured as the ratio between the average Rx signal power and the Rx noise floor when the experiment is run in HD or FD mode. The HD/FD link SNR is measured for every packet, and then the average is taken as the HD/FD link SNR value for the measurement. The second metric is the *HD/FD link PRR*, which is calculated as the percentage of the 1,024 transmitted packets that are successfully received and decoded by the FD radio. One link PRR is calculated for every link SNR.

Fig. 9.12 plots the measured HD and FD link PRR as a function of the HD link SNR when both radios operate in HD and FD modes. The results show that with sufficient link SNR values, the Gen-2 FD radios can achieve a link-level PRR of 1. This corresponds to a FD rate gain of exactly 2×. With insufficient link SNR values, there is a reduction in PRR of 25% on average, which corresponds to a FD rate gain of 1.5×. The lower PRR values are a result of degradation in link SNR when operating in FD mode. This lower *FD link SNR* is caused by the imperfect SIC achieved on the Gen-2 FD radios.
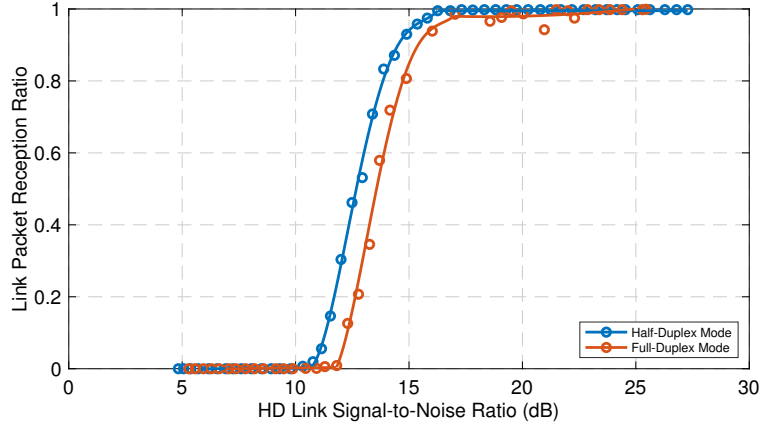
Figure 9.12: HD and FD link packet reception ratio (PRR) for QPSK-modulated, uncoded OFDM packets with varying HD link signal-to-noise ratio (SNR).

9.5 Conclusion

In this chapter, we presented our cross-layered (hardware and software) design and implementation of the first open-access, remotely accessible FD radios which are integrated with the ORBIT and COSMOS wireless testbeds. The presented example experiments serve not only as a tutorial for experimenters, but also a foundation upon which to build further experimentation for any desired scenario. We believe that the integrated FD radios and example experiments can facilitate further hands-on research in FD wireless. This experiment also showcases the capability of integrating customized experimental hardware in an open-access wireless testbed, where the testbed's infrastructure (e.g., SDRs and edge cloud servers) can be leveraged for different types of experiments.

Chapter 10: 28 GHz Channel Measurements in the COSMOS Testbed Area

Millimeter-wave (mmWave) communication is a core technology for 5G and future wireless networks, where the widely available spectrum at mmWave frequencies can be leveraged to achieve very high data rates (above Gbps) in both cellular and Wi-Fi systems. To allow the research community to experiment with state-of-the-art mmWave radios, a key building block of the COSMOS testbed is the IBM 28 GHz PAAM, which can enable unique mmWave experimentation currently not supported by other testbeds (see Section 8.2.2). However, as mentioned in Chapter 1 and Section 8.2.2, the operation of mmWave networks requires accurate channel models designed for specific types of deployment site.

In this chapter, we report extensive 28 GHz channel measurements in the COSMOS testbed area (see Figs. 1.4(b) and 8.6), which is a representative urban street canyon environment with both line-of-sight (LOS) and non-light-of-sight (NLOS) links. Our measurement campaign includes over 1,500 link measurements (with over 24 million power measurements) collected along 13 sidewalks adjacent to the 3 considered measurement sites. We first describe the measurement equipment, environment, and collected dataset (Section 10.1). We then present and discuss the measurement results on the path gain values and their fitted lines, and the effective azimuth beamforming gain (Section 10.2). The results show that, even in the NLOS scenario, the measured path gain values are on average 5–10 dB higher than that provided by the 3GPP 38.901 urban canyon NLOS model [15].

Finally, based on the measurement results, we consider the link SNR values and theoretical data rates that can be achieved on individual sidewalks (Section 10.3). In particular, link SNR values of at least 15 dB can be achieved on all sidewalks up to 194 m link distance. We believe that our measurement results can provide insight into the deployment of the 28 GHz phased array antenna modules (PAAMs) (developed by IBM and Ericsson [128]), in the COSMOS testbed. The

results can also serve as the benchmark for the experimentation with these modules once deployed.

Some of the results in this chapter were presented in [41, 42] and are based on joint work with Dr. Dmitry Chizhik, Dr. Jinfeng Du, Dr. Rodolfo Feick, Dr. Reinaldo Valenzuela, and Ph.D. student Manav Kohli. Undergraduate students Tianyi Dai, Angel Daniel Estigarribia, Kimberly Santiago, and Jackson Welles contributed to the measurement campaign.

10.1 28 GHz Channel Measurements: Equipment, Environments, and Dataset

10.1.1 Measurement Equipment

To maximize data collection speed and link budget, we use a 28 GHz portable narrowband channel sounder with a transmitter (Tx) and a rotating receiver (Rx) (more details can be found in [42]). In particular, the Tx is equipped with an omni-directional antenna (with 0 dBi gain) and emits a 28 GHz continuous-wave tone at +22 dBm power. The Rx is equipped with a horn antenna with 24 dBi gain and 10° half-power beamwidth, which is connected to a low-noise amplifier, a mixer, and a USB power meter with a bandwidth of 20 kHz and an effective noise figure of 5 dB. The Rx is mounted on a rotating platform, which allows for a full 360° angular scan in the azimuth plane every 500 ms (at rotational speed of 120 rpm). Using an onboard Raspberry Pi, the Rx records power samples at a rate of 740 samples per second. Both the Tx and Rx have free running local oscillators (LOs) with a frequency accuracy of at least 10^{-7}, and are powered by batteries to support portable measurements.

The channel sounder system is calibrated in the lab and in an anechoic chamber to ensure an absolute power accuracy of 0.15 dB. The full dynamic range of the Rx is 50 dB (from the Rx noise floor to its 1 dB compression point). The Rx also includes two switchable power amplifiers which can further extend the dynamic range to 75 dB (when both power amplifiers are switched on). Together with the use of removable attenuators at the Tx, the channel sounder supports a range of measurable path gain with at least 10 dB SNR from -62 dB to -137 dB. This corresponds to 1 m range in free space to 200 m range with 30 dB excess loss. The measurable path gain extends to -161 dB with the 24 dBi Rx horn antenna.
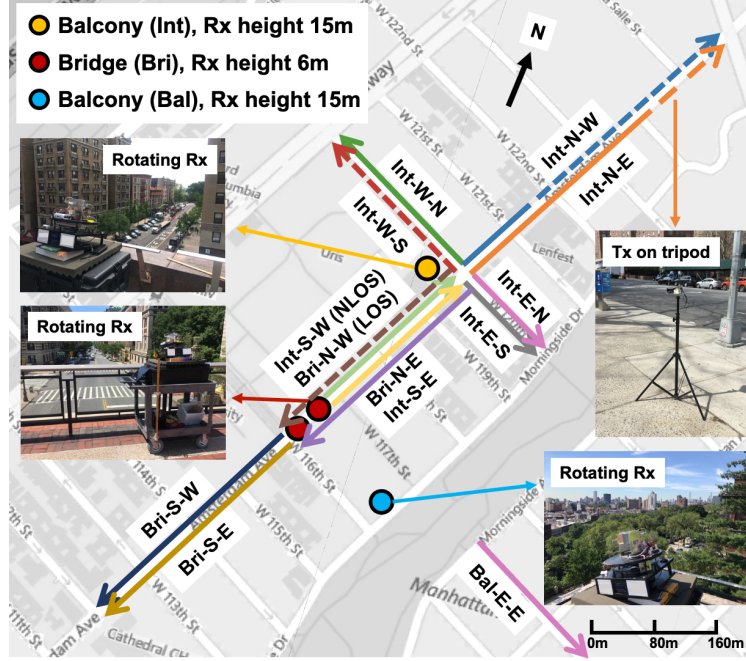
Figure 10.1: Locations and example street views of the three measurement sites in the COSMOS testbed deployment area. The receiver (Rx) is placed at a fixed location in each site (marked by a circle), and the transmitter (Tx) is moved along the measurement sidewalks (marked by lines). Solid and dashed lines correspond to sidewalks (or part of the sidewalks) where the Tx and Rx are in line-of-sight (LOS) and non-line-of-sight (NLOS), respectively.

10.1.2 Measurement Environment

We perform extensive 28 GHz channel measurements in the deployment area of the PAWR COSMOS testbed [13, 14], which is currently being deployed in West Harlem, New York City. Fig. 10.1 shows the map of the measurement area and example street views in the measurement locations. The considered area is a representative *urban street canyon* environment containing mostly concrete and brick buildings with heights of 10–60 m (3–16 stories). The vegetation in the measurement area consists of sparse thin trees separated equally on both sides on the city streets, representative of Northeastern U.S. cities. As shown in the map, the area is bordered by two parks (Riverside Park and Morningside Park).

We consider 3 measurement sites representing different deployment scenarios of a mmWave small cell BS, as shown in Fig. 10.1:

- Intersection Balcony (**Int**): the 4$^{\text{th}}$ floor balcony (with a height of 15 m) in the SW corner

Sidewalk Name	Sidewalk Color	Sidewalk Group	Sidewalk Length (m)	Meas. Step Size (m)	# of Meas. Links	Slope (n)	Intercept (b [dB])	RMS (σ [dB])	Median Azi. Gain (dBi)	10th-perc. Azi. Gain (dBi)
Int-N-E	Orange	Int-LOS	507	3/6 (near/far)	101	−3.5	−36.8	4.3	14.1	12.3
Int-W-N	Dark Green	Int-LOS	256	3	79	−2.6	−52.2	4.4	14.2	13.2
Int-S-E	Light Purple	Int-LOS	317	3	93	−3.4	−35.5	4.6	14.2	13.4
Int-E-N	Pink	Int-LOS	146	1.5	85	−2.3	−60.3	4.5	14.2	13.1
Int-E-S	Gray	Int-LOS	146	1.5	88	−2.8	−49.5	3.2	14.1	12.4
Int-N-W	Light Blue	Int-NLOS	509	3	139	−3.6	−36.0	3.6	13.1	11.8
Int-W-S	Red	Int-NLOS	256	3	69	−3.1	−47.5	3.1	11.6	10.4
Int-S-W	Brown	Int-NLOS	317	3	100	−3.6	−39.2	3.4	12.9	11.2
Bri-N-E	Yellow	Bri	219	3	65	−2.3	−60.0	3.9	12.6	11.3
Bri-N-W	Light Green	Bri	219	3	70	−2.6	−52.5	4.3	13.4	11.7
Bri-S-E	Dark Yellow	Bri	280	3/6 (near/far)	84	−2.5	−55.7	5.5	12.8	11.6
Bri-S-W	Dark Blue	Bri	280	3/6 (near/far)	87	−2.2	−59.8	4.0	13.2	11.4
Bal-E-E	Purple	Bal	488	3	156	−3.4	−47.2	5.8	13.9	12.8

Table 10.1: Summary of (i) the measurement campaign setup on 13 sidewalks in 3 sites (Intersection Balcony [Int], Bridge [Bri], and Open Balcony [Bal]), as shown in Fig. 10.1, and 4 sidewalk groups (Int-LOS, Int-NLOS, Bri, and Bal), (ii) the slope (n), intercept (b), and root mean square (RMS, σ) values for the path gain fitted lines, and (iii) the median and 10th-percentile azimuth gain values.

of the intersection of Amsterdam Avenue and West 120th street.[1] This location emulates the scenario where the BS is deployed on building roof/side in an urban street canyon;

- Bridge (**Bri**): the bridge (with a height of 6 m) crossing above Amsterdam Avenue between West 116th and 117th streets. This emulates the scenario where the BS is deployed on a lightpole in the middle of a two-way avenue;

- Open Balcony (**Bal**): the balcony of the Faculty House (with a height of 15 m) facing Morningside Park, where the neighboring buildings are 4–16 stories high. This emulates the scenario where the BS is deployed on the building roof and/or building side facing an open-space area.[2]

The channel measurements are collected on 8/4/1 sidewalks[3] in sites Int/Bri/Bal, respectively. In Fig. 10.1, each sidewalk is represented by a line with an arrow pointing from the corresponding Rx in the site. On most sidewalks, direct LOS (*solid* lines in Fig. 10.1) exists between the Tx and Rx despite possible blockages such as foliage and trees, street signs, lightpoles, etc. Although LOS with these types of blockage is typically considered as NLOS, we categorize these sidewalks as LOS to distinguish from the "true" NLOS case caused by the street canyon environment (e.g.,

---

[1]The COSMOS pilot nodes are located at this intersection.

[2]Note that this deployment scenario is common throughout Manhattan such as in streets adjacent to Riverside Park, Central Park, and St. Nicholas Park (where the COSMOS nodes in City College of New York are expected to be deployed).

[3]We differentiate two sidewalks of the same street with different labels, e.g., Int-N-E and Int-N-W correspond to the East and West sides of Int-N (i.e., Amsterdam Avenue between West 120th and 125th streets), respectively.

strong blockage by concrete buildings) as illustrated by the *dashed* lines in Fig. 10.1.

We categorize three sidewalks in group Int-NLOS where: (i) Tx on Int-W-S (red line) and Int-S-W (brown line) are right underneath the Rx with no direct LOS, and (ii) Tx on Int-N-W (light blue line) is blocked from the Rx by the buildings on the same side of the street. By exploring the angle-of-arrival (AoA) of the angular spectrum at the Rx, it can be observed that most of the received signal power comes from diffraction and/or reflections off building edges/walls on sidewalks in the Int-NLOS group.[4]

For each site, the rotating Rx with a 10° horn antenna is used to emulate the BS with better angular resolution in order to capture signals arriving from all azimuth directions. The Tx with an omni-directional antenna, emulating a user equipment (UE), is placed on a tripod and is moved along the sidewalks to obtain link measurements as a function of the link distance. On most sidewalks, the Tx is moved away from the Rx at a step size of 3 m in short distances, and at a step size of 6 m in long distances (i.e., $d > 250\,\mathrm{m}$).[5] Two exceptional sidewalks are Int-E-N/S, where the Tx is moved away from the Rx at a smaller step size of 1.5 m to ensure enough number of measurement links on shorter sidewalks.

For a given pair of Tx and Rx locations, we record one link measurement which lasts for 20 seconds and consists of 40 full 360° azimuth scans. The link distance of each measurement is calculated using the 3D geometric coordinates obtained from Google Earth with terrain characteristics (e.g., the elevation/slope of each sidewalk) taken into account. For example, sidewalks Int-N-W and Int-N-E are going downhill from south to the north, and the resulting relative height between the Rx and Tx is 15/41 m at the nearest/furthest (southernmost/northernmost) location, respectively.

Table 10.1 summarizes our measurement campaign on the 13 sidewalks during March–June, 2019, where the sidewalk color corresponds to the colored lines in Fig. 10.1. Over 1,500 link measurements were collected. For each link measurement, over 15,000 power samples were collected

---

[4]Further explorations and studies on the effects of diffraction and reflections in urban street canyon environments is a subject of our future research.

[5]In the measurement area, a street block is approximately 80 m long, and a street tile size is 1.5 m × 1.5 m.
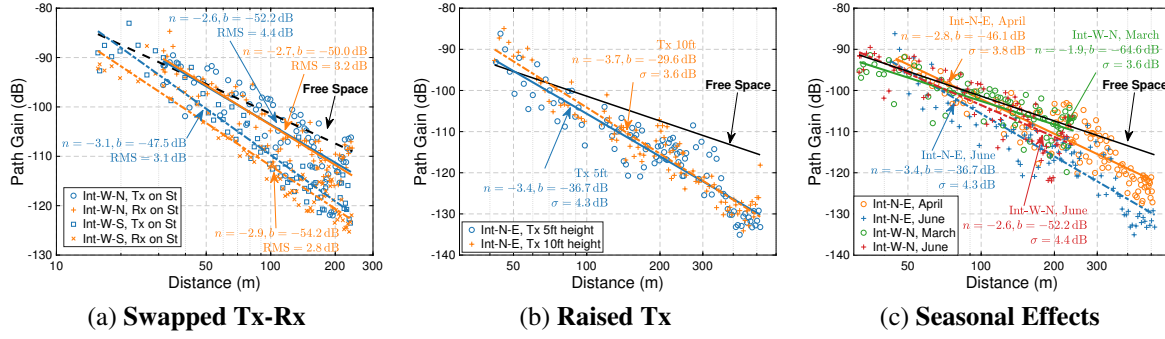
Figure 10.2: Measured path gain values and their fitted lines for understanding the effects of the measurement setup and environments. Considered effects include: (a) **Swapped Tx-Rx**: Relative locations between the Tx and Rx, (b) **Raised Tx**: Height of the Tx, and (c) **Seasonal Effects**: Vegetation depth and foliage on the sidewalks.

during 20 seconds as a function of both time and azimuth angle, from which the path gain and effective azimuth gain can be computed (see Section 10.1.3). Overall, a total number of over 24 million power measurements were collected.

10.1.3   Path Gain and Azimuth Gain

As described above, the rotating Rx with a horn antenna can record the received signal power as a function of both time and the azimuth angle. Following the methodology described in [42], we now briefly describe the calculation for two considered performance metrics: (i) path gain, PG, and (ii) effective azimuth gain, G_{az}, which quantifies the effects of scattering environments on the signal angular spread, and thus the effective beamforming gain.

Denote by $P(d, \phi)$ the instantaneous received power measurement at link distance of $d\,(\mathrm{m})$ in the azimuth direction of ϕ. It is shown in [42] that the average power for each link measurement with distance d, denoted by $\overline{P}(d)$ (obtained by averaging $P(d, \phi)$ over the azimuth angle, ϕ, using a directional antenna) equals to that measured using an omni-directional antenna, i.e.,

$$\overline{P}_{\mathrm{horn}}(d) = \frac{1}{2\pi} \int_{\phi=0}^{2\pi} P(d, \phi)\, \mathrm{d}\phi = \overline{P}_{\mathrm{omni}}(d). \tag{10.1}$$

The path gain as a function of the link distance, $\mathsf{PG}(d)$, is then computed as the received signal power level when removing the Tx power, P_{Tx}, and the antenna elevation gain, G_{el} (recall that the

Rx is rotating on the azimuth plane), i.e.,

$$\mathsf{PG}(d) = \frac{\overline{P}_{\text{horn}}(d)}{P_{\text{Tx}} \cdot G_{\text{el}}}, \tag{10.2}$$

where G_{el} is the Rx antenna elevation gain measured in an anechoic chamber. Accordingly, the effective azimuth beamforming gain is defined as the peak-to-average ratio of the receive power measurements, which is given by

$$G_{\text{az}}(d) = \frac{\max_\phi\{P(d, \phi)\}}{\overline{P}_{\text{horn}}(d)}. \tag{10.3}$$

The measured path gain values are linearly curve-fitted by

$$\mathsf{PG}(d) = 10n \cdot \log_{10} d \ + b + \mathcal{N}(0, \sigma) \text{ (dB)}, \tag{10.4}$$

where n, b, and σ are the slope, intercept, and root mean square (RMS) of the fitted line. We refer to these parameters of a fitted line by their variables throughout the chapter. The obtained values of n, b, and σ for the sidewalks are summarized in Table 10.1 and are discussed in Section 10.2. Note that the *Friis free space* path gain line is with $n = -2$, $b = -61.4$ dB, and $\sigma = 0$ dB.

10.2 Measurement Results

We now present and discuss the measurement results of the path gain values and the effective azimuth beamforming gain in the 3 measurement sites (see Fig. 10.1 and Table 10.1).

10.2.1 Effects of Setup and Environments

Conducting ideal and controllable channel measurements in an urban street canyon environment is challenging, due to various factors such as the measurement setup and environments. In order to better understand the possible effects of these factors on the results, we first conduct three basic sets of measurements to study the following effects: (i) **Swapped Tx-Rx**: relative locations between the Tx and Rx, (ii) **Raised Tx**: height of the Tx, and (iii) **Seasonal Effects**: vegetation depth and foliage on the sidewalks.

Swapped Tx-Rx. Recall that we use the Rx to emulate the BS due to its capability to capture received power across the full $360°$ azimuth plane. Ideally, swapping the placement of the omni-directional Tx antenna and a rotating directional Rx antenna is expected to result in the same average power with proper compensation of the vertical beam pattern effects. We first conduct measurements to verify this and to understand the sensitivity to swapped Tx and Rx locations.[6] In particular, we consider sidewalks Int-W-N and Int-W-S with the Tx (resp. the rotating Rx) placed on the intersection balcony and the street (resp. the street and intersection balcony).[7]

Fig. 10.2(a) shows the measured path gain values and their fitted lines for the four measurements on two sidewalks. It can be seen that the link measurements on the same street but with swapped Tx and Rx locations are similar, where the differences in n (slope) and b (intercept) of the fitted lines are 0.1/0.2 and 2.2/6.7 dB for Int-W-N/Int-W-S, respectively. Due to the more complex environmental effects in the NLOS channels, the measured path gains on Int-W-S are on average lower than that on Int-W-N.These results show that swapping the Tx and Rx locations has only minimal effects on the path gain. Therefore, throughout the rest of the measurements, we use the rotating Rx with better angular resolution to emulate the BS with heights of 6–15 m (see Section 10.1.2).

Raised Tx. We evaluate the effects of the Tx height by placing the Tx on a tripod at 1.5/3.0 m and measuring the path gains on sidewalk Int-N-E.[8] Fig. 10.2(b) shows the measured path gains, whose fitted lines for both Tx heights differ by only 2.5 dB and 0.8 dB at the near ($d = 50$ m) and far ($d = 500$ m) ends, respectively. Although at different heights, the Tx experiences slightly different channel characteristics (e.g., blockage by trees), the measured path gains are similar, with less than 3 dB difference between the fitted lines over link distances of 40–500 m.

Seasonal Effects. Finally, we evaluate the seasonal effects during the period of March–June, 2019. In particular, we conducted measurements on Int-N-E and Int-W-N in two different months (all in sunny days) where the foliage and the depth of the tree leaves vary between early spring (no leafs)

---

[6]Channel reciprocity is valid when swapping Tx and Rx while leaving antennas in place. Here, the Tx and Rx antennas are also swapped together with the platform.

[7]In Table 10.1, the rows of Int-W-N and Int-W-S are with Rx on the intersection balcony.

[8]In Table 10.1, the row of Int-N-E is with Tx at 1.5 m height.
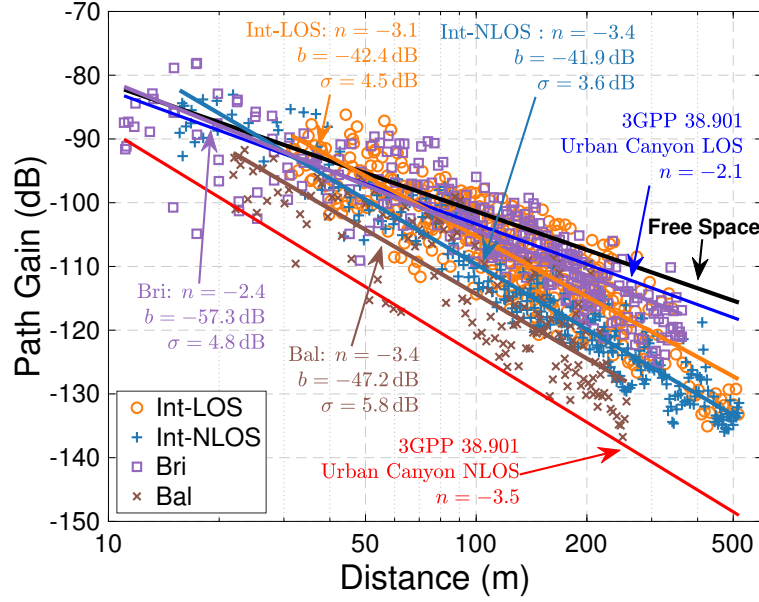
Figure 10.3: Measured path gain values and the corresponding fitted lines for different sidewalk groups (Int-LOS, Int-NLOS, Bri, and Bal).

and summer (full leafs).[9] Fig. 10.2(c) shows the measured path gains and their fitted lines. The results show that the differences between months are minimal for d up to 250 m. Note that for $d > 250$ m on Int-N-E, the path gain values measured in the summer (June) drop by around 10 dB compared to that measured in early spring (April). This could possibly stem from the fact that the aggregated effects of the foliage along the street is more significant in the summer. We plan to conduct more comprehensive measurements in the fall/winter to further investigate the effects of foliage on the path gains.

Overall, we believe that the measurements are good representative of the three effects on the corresponding streets, and hence can provide insight into these effects on other sidewalks in the same area. Yet, we plan to conduct more measurements in the future for further empirical validation. We also plan to study and evaluate other effects including traffic loads and weather, etc.
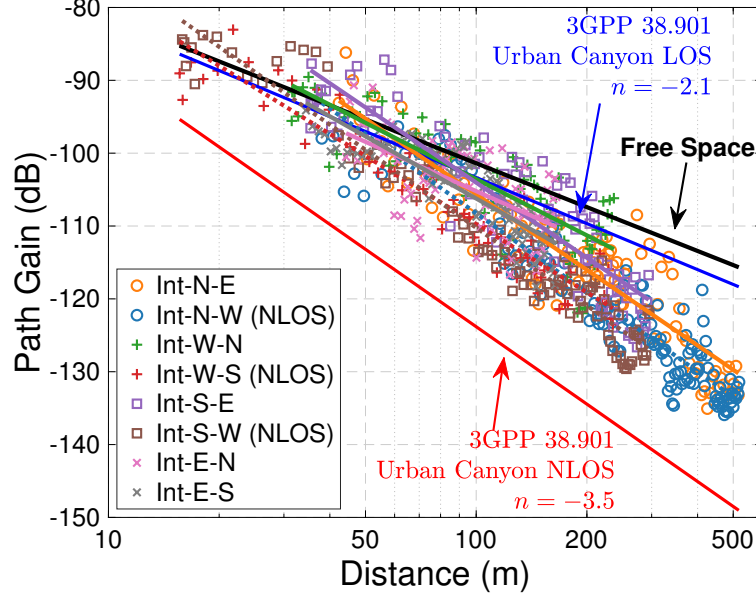
Figure 10.4: Measured path gain values and the corresponding fitted lines for individual sidewalks in the Intersection Balcony (Int).

10.2.2 Path Gain

We now study different deployment scenarios where the rotating Rx is placed in sites Int/Bri/Bal. These cases correspond to different mmWave BS deployment scenarios as described in Section 10.1.2.

As shown in Table 10.1, the sidewalks are categorized into four groups (see also Fig. 10.1). Fig. 10.3 shows the measured path gains for each group (in markers) and their fitted lines using the 1,216 link measurements collected. The results show that Int-LOS and Bri have comparable path gains since similar effects of the urban street canyon environment apply in both cases. However, these effects are less significant for Int-NLOS and Bal, where lower measured path gains are observed. We conjecture that this is because of the reduced Rx power that could have been reflected from building sides/walls. Moreover, most path gains fall between the 3GPP 38.901 urban street canyon LOS and NLOS models [15]. Specifically, the values of n, b, and σ for groups Int-LOS/Int-NLOS/Bri/Bal are $-3.1/-3.4/-2.4/-3.4$, $-42.4/-41.9/-57.3/-47.2$ dB, and $4.5/3.6/4.8/5.8$ dB, respectively.

We also consider individual sidewalks in Int and Fig. 10.4 plots the measured path gain values

---

[9]In Table 10.1, the rows of Int-N-E and Int-W-N correspond to their measurements conducted in June 2019.
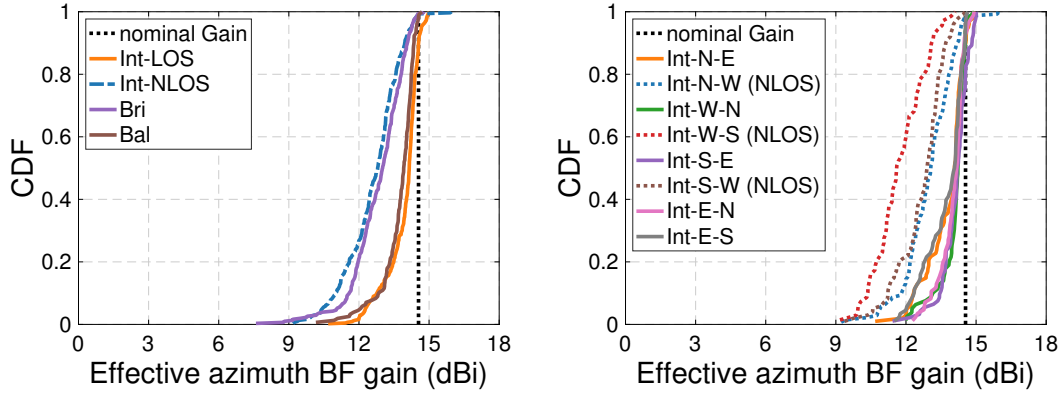
Figure 10.5: CDFs of the effective azimuth beamforming gain for the sidewalk groups (left) and for individual sidewalks in Intersection Balcony (Int) (right).

and their fitted lines. The corresponding parameters for the fitted lines are summarized in Table 10.1. It can be seen that the measured path gain values are mostly 10 dB higher than the 3GPP 38.901 urban canyon NLOS model, and over 5 dB lower than the 3GPP 38.901 urban canyon LOS model for distances beyond 100 m.

10.2.3 Effective Azimuth Beamforming Gain

As mentioned in Section 10.1.3, the effective azimuth gain characterizes the corresponding beamforming gain under the effects of scattering environments on the angular spread of the signal. Fig. 10.5 plots the CDFs of the effective azimuth beamforming gain of groups Int-LOS/Int-NLOS/Bri/Bal (left) and of individual sidewalks at Int (right). In particular, he median azimuth gains for Int-LOS/Int-NLOS/Bri/Bal are 14.2/12.8/13.0/13.9 dBi, respectively, which are very similar across different sites. Both the median and 10th-percentile azimuth gains for individual streets in Int are summarized in Table 10.1. It can be seen that: (i) the 10th-percentile azimuth gain is at most 1.8 dB lower than the median azimuth gain across all sidewalks, and (ii) the beamforming gain degradation is more significant on NLOS sidewalks (e.g., the median azimuth beamforming gain is always within 1.7/2.9 dB of the nominal value for sidewalks with LOS/NLOS, respectively).
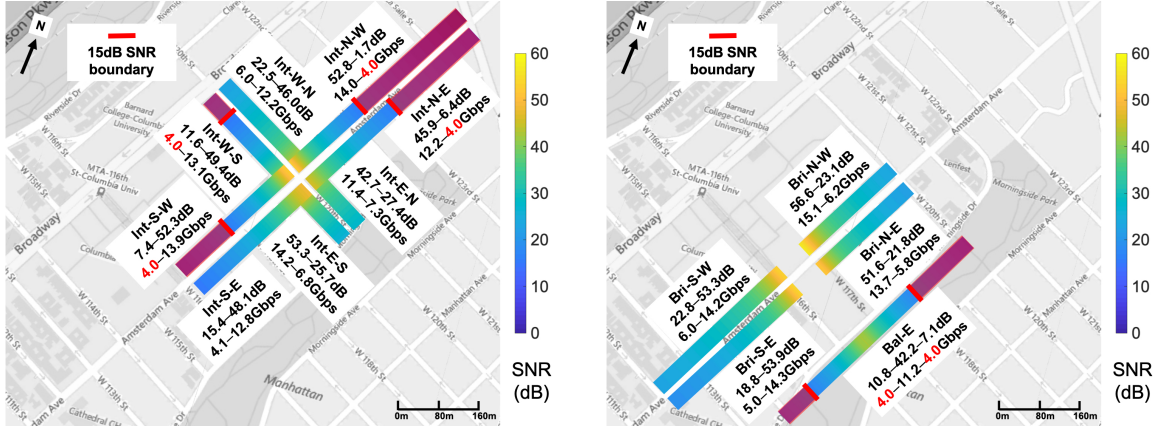
Figure 10.6: Heatmap of median link SNR values, which are generated using the path gain model (10.4) with slope (n), intercept (b), and RMS (σ) values given in Table 10.1, on individual sidewalks in the COSMOS testbed deployment area. The corresponding Shannon rates are also included for individual sidewalks, where UE locations with link SNR values lower than 15 dB are ignored in the rate computation (shaded red area).

10.3 Signal-to-Noise Ratio (SNR) Coverage and Achievable Data Rates

Empirical path gain models based on extensive measurements (such as those discussed in Sections 10.1 and 10.2) will allow performance evaluation of mmWave networks in various settings. In this section, we consider link SNR values that can be achieved in the COSMOS testbed deployment area. These values can provide insight into characterizing the coverage at 28 GHz and the achievable data rates.

We assume the following parameters. The BS has a constant Tx power of $P_{\text{Tx}} = +28$ dBm and a maximum Tx gain of 23 dBi.[10] For each sidewalk, the effective Tx gain, denoted by G_{Tx}, is obtained by subtracting the median azimuth gain degradation indicated in Table 10.1 from the maximum Tx gain. The UE is assumed to have a nominal Rx gain of $G_{\text{Rx}} = 11.0$ dBi and a noise figure of 10 dB. We consider signal bandwidth of 800 MHz. The resulting Rx noise floor can be calculated as $P_{\text{nf}} = -174 + 10\log_{10}(800 \times 10^6) + 10 = -75$ dBm.

We consider UE locations at a step size of 1 m on each sidewalk. The median path gain for each BS and UE pair with a link distance of d is obtained based on the path gain equation (10.4)

---

[10]The resulting maximum total EIRP of +51 dBm is comparable to that of the 28 GHz phased array antenna modules [128] to be deployed in COSMOS.

with slope (n), intercept (b), and RMS (σ) values given by Table 10.1. The link SNR value at each UE location with link distance of d is computed as

$$\mathsf{SNR}(d) = P_{\text{Tx}} + G_{\text{Tx}} + \mathsf{PG}_{\text{med}}(d) + G_{\text{Rx}} - P_{\text{nf}}. \tag{10.5}$$

Fig. 10.6 shows the resulting link SNR values for individual sidewalks using color bars, where the link SNR value decreases as link distance increases. The range of the achievable median SNR values is also labeled for each sidewalk, with a cutoff point where the link SNR drops to below 15 dB. In particular, 5 sidewalks have link SNR values lower than 15 dB at the far end, and the smallest cufoff distance is 194 m on Int-S-W. Fig. 10.6 also shows the corresponding Shannon rates that can be achieved on individual sidewalks, where the UE locations with link SNR values lower than 15 dB are ignored (shaded red area). The results illustrate the theoretically achievable data rates in the considered area.

10.4   Conclusion

In this chapter, we presented the results of an extensive 28 GHz channel measurement campaign in the COSMOS testbed deployment area. We thoroughly studied 3 sites and obtained measurements of over 1,500 links on 13 sidewalks, some of which were measured multiple times in different settings. We studied the effects of swapped Tx and Rx locations, Tx heights, and seasonal effects on the path gain and concluded that these effects are insignificant in most scenarios. Moreover, we studied the path gain values, the effective azimuth beamforming gain, and median SNR values that can be achieved on the 13 sidewalks. Overall, the results can inform the deployment of the 28 GHz IBM PAAMs, described in Section 8.2.2, in the COSMOS testbed.

Chapter 11: Converged Optical-Wireless Experiments

To experimentally evaluate converged optical-wireless technologies and SDN control in a real-world setting, a key infrastructure element in the COSMOS testbed is the *dark fiber-based programmable optical x-haul network*, which allows Physical layer research extending throughout the city deployment (see Sections 8.2.3 and 8.3). In this chapter, we present the architecture design and deployment of COSMOS' programmable optical network (Section 11.1). We then demonstrate a covnerged optical-wireless experimentation, where we consider the scenario of wireless handover (HO) in C-RANs and propose a scheme of *dynamic capacity allocation* of optical/wireless links to support varying network bandwidth requirements (Section 11.2). We implement and experimentally evaluate the proposed HO scheme using COSMOS' optical and radio resources.

The chapter's contributions were previously presented in [13, 43–45] and the work is in colalboration with Craig Gutterman, Artur Minakhmetov, Jiakai Yu, and Shengxiang Zhu.

11.1 Programmable Optical Network in the COSMOS Testbed

The architecture design of the COSMOS' optical network makes use of WDM and optical space switching to provide two important capabilities: (i) *flexible experimentation and network topology reconfiguration* of a large numbers of radio and computing resource connections, and (ii) *multi-layered optical x-haul networking* for experimentation on novel optical devices, systems, SDN/NFV optical control planes, and optical architectures.

Fig. 11.1 shows the diagram of COSMOS' *envisioned* core optical switching architecture and the corresponding optical switching for large radio nodes in the bottom inset. In particular, a Calient S320 320×320 multicast space switch (SS) [187] forms the core of COSMOS' optical network in the main computing research facilities (CRF) located in the Mudd Building at Columbia
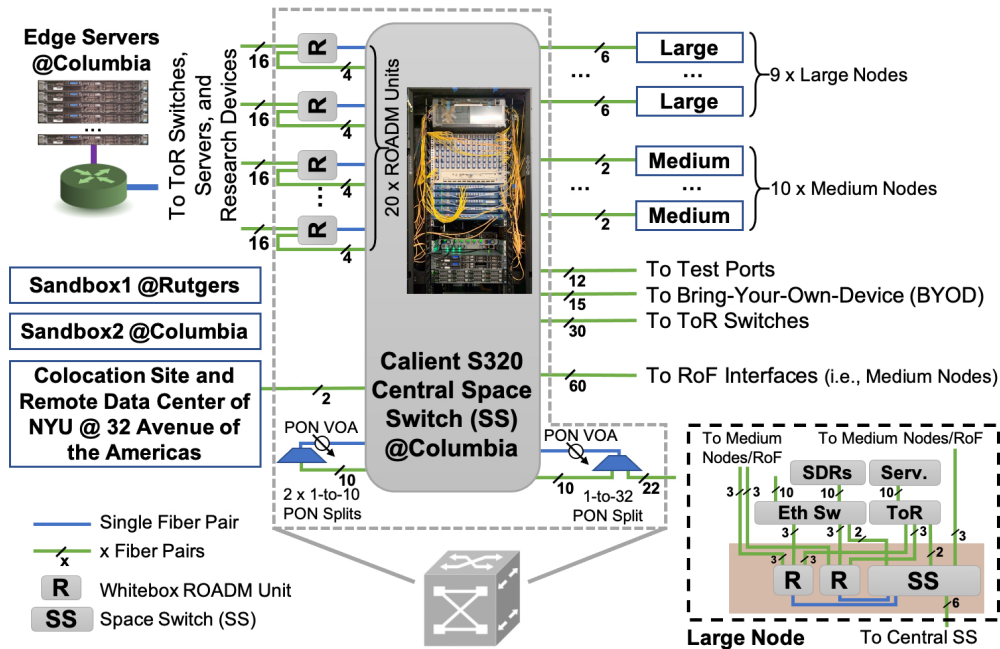
Figure 11.1: The *envisioned* core optical switching architecture of the COSMOS testbed and the switching architecture of a large node.

University. Dark fiber pairs connect the central S320 SS with similar smaller 16×16 SSs at each of the large nodes. These SSs allow for remote and automated re-fibering of connections and devices throughout the testbed. WDM is provided by the whitebox reconfigurable optical add drop multiplexing (ROADM) units [188] connected to the SSs. Other devices attached to the central S320 SS include splitters for passive optical networks (PONs), top-of-rack (ToR) switches, test equipment, and different experimental hardware. Using these capabilities, the fiber pairs between any two devices (e.g., a radio node and an edge server) can be configured for combinations of point to point, PON, and ROADM/WDM networks. WDM can also provide basic capacity expansion to the links, supporting up to 96 channels, which are fully transparent and flex-grid configurable to support 10/100 Gbps wavelengths planned for the testbed as well as custom user-supplied signals. As shown in Fig. 11.1, many radio nodes will also be equipped with analog radio over fiber (RoF) capabilities in addition to the digital RoF used throughout.

The central Calient S320 SS is planned to have 48 fiber pairs connecting 9 large radio nodes in different locations (tentative locations are shown in Fig. 8.6) including several Columbia University
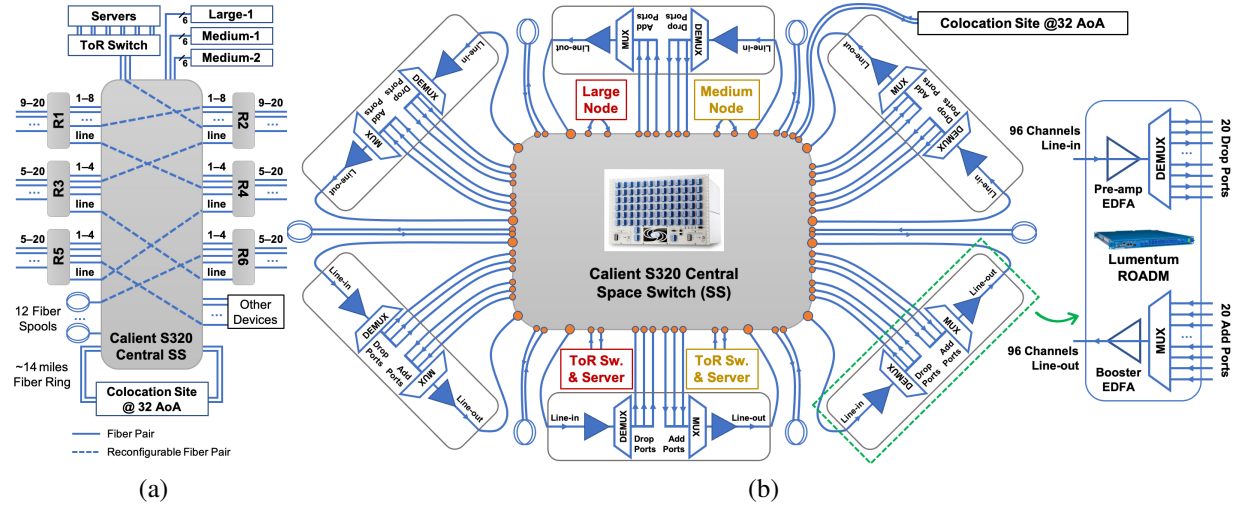
Figure 11.2: The *current* deployment and configuration of COSMOS' optical network: (a) optical connectivity topology consisting of a subset of devices shown in Fig. 11.1, and (b) main components in the optical layer including one Calient S320 central space switch (SS) and six Lumentum ROADM units (only 3 pairs of connections are shown between the SS and add/drop ports).

buildings, CCNY buildings, and city locations. Another two fiber pairs are used to connect to the data center at 32 AoA in downtown Manhattan, which is ∼7 miles south of the testbed area. The fully equipped testbed will include 20 programmable ROADM units connected to the central S320 SS via 20×4 wavelength filtered add/drop fiber pairs. Another 20×16 add/drop pairs will be used to connect to computing resources or research devices. The 20 line side fiber pairs connect to the SS for WDM transport over the dark fiber pairs to other nodes. These ROADM units can be reconfigured to support various requirements from connected servers and research devices through top-layer user applications orchestrated by a dedicated COSMOS optical SDN controller. There are also additional fiber pairs for connections to the medium nodes, commercial hardware, test ports, and other devices.

Large radio nodes are constructed with a 16×16-port space switch with 2×1 fiber pairs for 2 ROADM units, 6 pairs for direct Columbia CRF connections, 2 pairs for Ethernet switches, 2 pairs for ToR switches, and 3 pairs for medium nodes/remote radio heads (RRHs). For medium nodes, currently there are two different configurations. One will use 6 fiber pairs to carry 1550 nm sub-6 GHz radio and spectral monitoring signals, and mmWave/RoF signals. The other configura-

tion is implemented with only 2 fiber pairs. One pair will use a fixed 8-channel MUX/DEMUX to support carrying all of these signals on one fiber and the other fibers are left for WDM expansion and mmWave/RoF signals. All node control information is carried on a dedicated 1300 nm signal introduced through a red/blue filter or using a 1300 nm port on the device, added and dropped prior to the ROADM units. A combination of short reach and long reach pluggable transceivers at 10/100 Gbps will be used throughout the testbed. Long reach pluggables include full band tunability for dense WDM operation. For multi-hop amplifier transmission, Menara IPG pluggables with built in optical transport network (OTN) wavewrapper including forward error correction, are used.

Fig. 11.2 shows the current optical network connections and configuraion of the COSMOS testbed, and the main components in the optical layer. In particular, the central S320 SS has connections to six Lumentum ROADM units (each with two uni-directional EDFAs and WSSs), one ToR switch (which further connects to edge servers and radios nodes), the colocation site at 32 AoA, as well to 12 fiber spools and other devices (e.g., comb source and optical telemetry). A fully transparent x-haul optical transport network can be configured using the Calient S320 SS and the Lumentum ROADM units. As illustrated in Fig. 11.2(b), wavelengths carrying signals can be fed (i) into the line-in port, pumped by the Pre-amp EDFA, and independently switched to any drop port of the DEMUX WSS, or (ii) into any add port of the MUX WSS, boosted by the Booster EDFA, and aggregated into the line-out port. In addition, each optical wavelength can not only carry its independent signal stream for services, but can also be used to support protocols like SONET/SDH, OTN, CPRI, etc., for wireless signaling between RRHs and BBUs.

11.2   Optical Switching based Wireless Handover (HO) in C-RAN

The use of WDM in front-haul networks suggests adopting multiple wavelength-specific PONs or point-to-point links that connect BBUs and RRHs. These links are provisioned for each RRH requiring high capacity [189], and can be dynamically reconfigured by an SDN controller as part of an x-haul network. As mentioned in Section 2.3, using optical networks to facilitate wireless
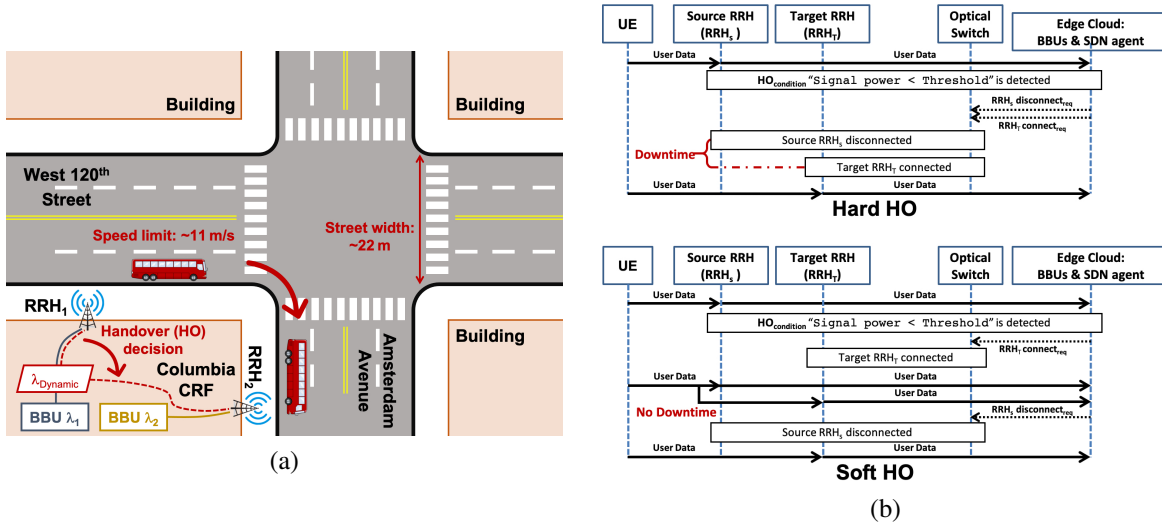
Figure 11.3: (a) The considered handover (HO) scheme, illustrated at the COSMOS Pilot deployment site, where a vehicle taking a turn at an intersection is served by two remote radio heads (RRHs) through dynamic optical switching and wavelength assignment. (b) The signaling flow for the considered HO modes: *hard HO* ("break before make"), where the new connection to the target RRH is made *after* the connection to the source RRH is broken, and *soft HO* ("make before break"), where the new connection to the target RRH is made *before* the connection to the source RRH is broken.

HO has been investigated before [133–135], and each of these approaches is a candidate for experimental evaluation in COSMOS. A key element for each is the ability to coordinate optical switching operations in response to wireless HOs.

We consider the application of wireless HO between RRHs connected to a mobile UE that requires exceptionally large capacity with deterministic low latency. An example application scenario is shown in Figs. 11.3(a), where vehicles move in a dense urban area and passengers use high data rate applications (e.g., real time multi-view vehicle video analytics). When the UE is moving from the coverage area of one RRH to another, the capacity of the underlying optical link needs to be *dynamically and seamlessly* re-allocated. To avoid provisioning high capacity throughout the network, a channel on a dedicated wavelength can be *optically switched* from one link to the other link using either the optical SSs or the ROADM units. The COSMOS SDN control environment is well suited for the study of such multi-technology SDN control scenarios. In this pilot experiment, SDN control was implemented for the edge cloud, including the Ethernet switches and BBUs along

with the optical switching.

11.2.1 Hard HO ("Break before Make") and Soft HO ("Make before Break")

We consider two HO modes, whose signaling flows are shown in Fig. 11.3(b). Moreover, the HO decision on optical switching from the source RRH to the target RRH is made based on the wireless signal power threshold. In particular:

- *Hard HO* ("break before make"): the SDN can assign the UE only a single wavelength, that would "follow" the UE from the source RRH to the target RRH. The switching consists of changing the SS configuration, always keeping one port connected to the link leading to the edge cloud and changing ports to the link leading to the RRHs. A downtime period exists during the hard HO process, during which the UE is not connected to any RRH.

- *Soft HO* ("make before break"): uses at least two wavelengths or independent fiber channels at the time of HO. Specifically, new connections from the edge cloud to a RRH are made before the old connections are broken. Therefore, compared with hard HO, soft HO has *zero* downtime but uses additional network resources.

11.2.2 Implementation in the COSMOS Testbed

We implemented and evaluated the wireless HO via optical switching in the COSMOS testbed. Fig. 11.4 shows the experimental setup, which involves two RRHs (RRH_1 and RRH_2), located on two sides of the Columbia CEPSR Building, which emulate the two COSMOS medium nodes (see Fig. 8.7). In particular, each RRH_i ($i = 1, 2$) is emulated using two USRP N210 SDRs, with one assigned a *fixed* capacity ($USRP_{i,\text{fix}}$) and one assigned a *dynamic* capacity ($USRP_{i,\text{dyn}}$). The link between $USRP_{i,\text{fix}}$ and the edge cloud is always live, whereas the link between $USRP_{i,\text{dyn}}$ and the edge cloud can be dynamically allocated via optical switching. We use a USB-powered USRP B205-mini SDR as the UE. For simplicity we consider only upstream traffic in this experiment, which does not limit the validity of our observations.

All USRPs are connected to a Pica8 ToR switch via 1 Gbps Ethernet. The Pica8 ToR switch has
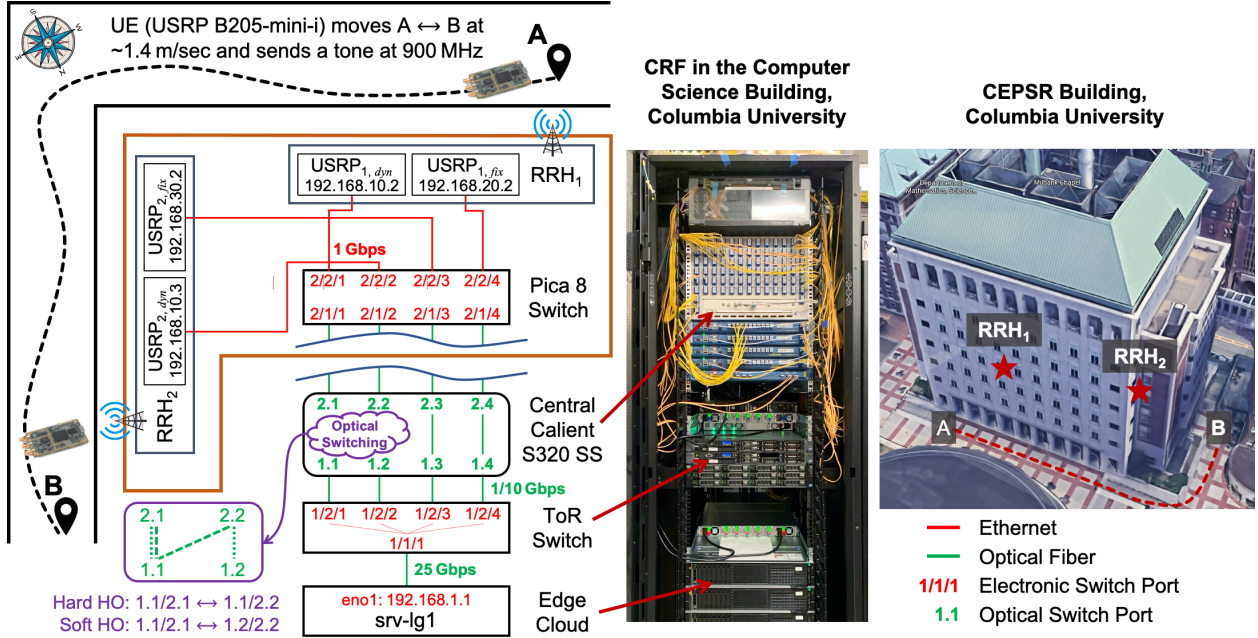
Figure 11.4: Setup of the C-RAN wireless handover (HO) experiment in the COSMOS testbed: Two RRHs located on each side of the Columbia CEPSR building, each consisting of 2 USRP SDRs. One USRP has a dedicated channel with *fixed* capacity through an optical link, while the other USRP has a *dynamic* optical channel allocated by an SDN controller via optical switching performed at the Calient S320 SS. The user equipment (UE) moves around the building corner on the campus level and passes between two RRHs, which invokes HO based on the HO signaling flows shown in Fig. 11.3.

four bi-directional optical transceivers (tunable C-Band SFP+s with 1/10 Gbps in 1530–1565 nm) connected to the Calient S320 SS located in the Columbia CRF. Specifically, three ports are used for hard HO and an extra 4th port is required for soft HO. The S320 SS routes lightpath from the Pica8 ToR towards another ToR switch that has a 25 Gbps connection to edge cloud servers. To avoid switching transients, we use direct point-to-point WDM connections over the SS as opposed to the ROADM units. Moreover, the SS is used for switching due to its faster SDN controlled switching time ($\sim$100 ms compared to $\sim$1 s for the ROADM units).

We used the open-source GNU Radio [179] software to implement the edge cloud functionalities, which include the SDN controller/agent and the remote processing of signal I/Q samples received by the RRHs. Our software runs on a COSMOS edge cloud server ("srv-lg1" in Fig. 11.4), and controls all USRPs through the depicted links and the SS through a TELNET tunnel. In particular, it sends commands over an out-of-band control channel to the SS to implement the hard/soft

HO based on the signaling flow shown in Fig. 11.4.

The UE transmits a continuous wave tone at 0.5 MHz frequency at 900 MHz carrier frequency. The RRHs receive radio signals over 5 MHz bandwidth at around 900 MHz and stream the received I/Q samples to the edge cloud server "srv-lg1", and the received signal strength at $\text{USRP}_{i,\text{fix}}$ is measured at 0.1 ms intervals. A switching/HO event is triggered if the received signal strength becomes lower than a threshold (-57 dBm), and the dynamic capacity is switched from the source RRH to the target RRH (i.e., from $\text{USRP}_{i,\text{dyn}}$ to $\text{USRP}_{j,\text{dyn}}$).

11.2.3 Experimental Evaluation and Measurement Results

We consider two performance metrics: (i) *HO time*, which is the time duration between the triggering and completion of the HO event, and (ii) *downtime*, which is the duration in which no $\text{USRP}_{i,dyn}$ is connected. We conducted 15 trials for hard HO and 14 trials for soft HO using 10 Gbps optical links between the edge cloud and USRPs. We also conducted indoor experiments with 1 Gbps optical links, where HO conditions are emulated in software.

Our experimental results show that the average switching time for hard HO is 0.957/1.830 sec with 1/10 Gbps links, respectively, where the corresponding downtime is 0.914/1.647 sec. Similarly, the average switching time for soft HO is 0.724/1.614 sec with 1/10 Gbps links, respectively, with *zero* downtime. The HO switching time is limited by the time it takes for the optical transceivers to establish the link connections, which is at the order of seconds [190]. In order to confirm this, we performed measurements of the optical link recovery time using WireShark. The results show that it takes <50 ms (set by the SS switching time) to change the S320 SS port configurations from state ON to OFF and then immediately from OFF to ON. The time difference between 1 Gbps and 10 Gbps is due to the fact that the 1 Gbps links are fixed rate, and thus no rate-negotiation phase is needed during the link establishment. We note that the long HO times can be reduced by using burst-mode transceivers, and we will also consider using optical switching using the ROADM units in future work.
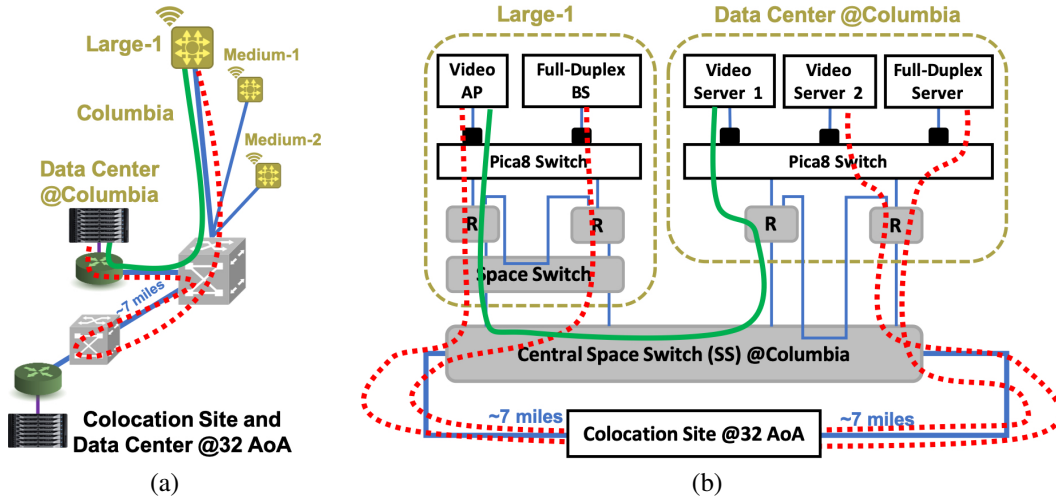
Figure 11.5: (a) Example of optical x-haul network topologies that can be created by COSMOS' optical transport network with SDN control, and (b) setup of the optical-wireless x-haul experiment and the corresponding optical paths.

11.2.4 Remote Signal Processing and Adaptive Video Streaming supported by Optical Switching

Using COSMOS' optical network and its dark fiber infrastructure deployed in Manhattan, we also prototyped C-RAN functionalities including remote processing of digitized RoF signals from a custom full-duplex (FD) radio, as described in Chapter 5, and on-demand optical switching of video streams from different servers.

Recall from Section 8.3, dark fiber already connects Columbia University to a large co-location data center at 32 AoA (7 miles south of the deployment area). Fig. 11.5 shows the overall system design and diagram of the COSMOS x-haul network for prototyping the desired C-RAN functionalities. In particular, a wideband FD radio implemented using a USRP SDR, described in Chapter 5, serves as a small cell (SC) node at Columbia University. The SC node communicating with user-end wireless devices sends complex I/Q data through the dark fiber network for remote processing at the edge-cloud server using NI LabVIEW. On the x-haul side, real-time USRP-server communication is realized over a metro optical network using dark fiber provided by New York City and ZenFi Networks (see Fig. 8.6). In addition, the Adaptive Multicast Services (AMuSe) application [191] simultaneously operates on the x-haul network where the wireless AP receives

video streams from different servers through on-demand optical switching managed by a Ryu based SDN controller [44] (black path to server 1 or green path across 32 AoA to server 2).

In this method, rather than placing all the computational functionalities at the radio node, having a reconfigurable and dynamic optical layer C-RAN capability allows for offloading the node's functionality based on application requirements (e.g., baseband I/Q data streams or multicast video) to a remote compute cluster. Due to its high capacity (10s of Gbps) and low latency (less than 1 ms), such an x-haul optical system can support high-performance computing tasks, and flexible network topologies for next-generation wireless devices. With different application requests, the SDN controlled x-haul optical system automatically switches the optical path either to the 32 AoA dark fiber network or to the Pica8 switch, enabling an optically switched bi-directional communication link between the AMuSe AP and the servers. Error free operation of the amplified dark fiber network was demonstrated along with switching of the AMuSe video server between near and far locations.

11.3 Conclusion

We presented the design and implementation of a field-deployed programmable optical network as a key building block of the city-scale COSMOS advanced wireless testbed. A multi-dimensional optical infrastructure built using a city-deployed dark fiber network and advanced optical switching technologies (i.e., high-port count space switching and tunable wavelength-routed switching) is fully controlled by an SDN controller and converged with SDRs and edge cloud resources. We demonstrated the integration of optical and wireless network switching using a C-RAN wireless handover experiment.

Chapter 12: Conclusions

In this thesis, we considered algorithms and experimentation for future wireless networks based on a cross-layered approach. We made contributions to several critical components, including Internet-of-Things (IoT) systems and full-duplex (FD) wireless, as well as to the experimentation with advanced wireless technologies in the city-scale PAWR COSMOS testbed.

In Part I, we first focused on the Physical layer and MAC protocols for ultra-low-power energy harvesting networks, and studied the supported tracking and monitoring applications for IoT systems. We first designed and optimized the Panda protocol, which maximizes the neighbor discovery rate between homogeneous energy harvesting nodes using a centralized probabilistic approach. We then designed and optimized the EconCast protocol, which maximizes the throughput among heterogeneous energy-constrained ultra-low-power nodes. Using both simulations and real-world experiments with commercial off-the-shelf energy harvesting nodes, we showed that, in terms of throughput, Panda outperforms existing protocols by up to $3\times$ and EconCast further outperforms Panda by up to an order of magnitude.

Then, in Part II, we studied FD wireless and focused on enabling FD on small-form-factor devices, in both single- and multi-antenna systems. In particular, we leveraged the technique of frequency-domain equalization to design, model, and optimize single-antenna compact wideband FD radios that can achieve an overall amount of 95 dB self-interference cancellation (SIC) across 20 MHz bandwidth. We experimentally showed that our prototyped FD radios can achieve a link- and network-level throughput gain of up to $1.91\times$. We also focused on combing FD operation with multi-antenna phased arrays, where simultaneous transmit and receive beamforming is employed to achieve wideband RF SIC leveraging the spatial degrees of freedom. In the MAC layer, we developed the novel Hybrid-Greedy Maximum Scheduling (H-GMS) algorithm, which was proven to achieve the throughput optimality in a distributed manner.

Finally, in Part III, we described the experimentation with advanced wireless technologies and millimeter-wave channel measurements in the city-scale PAWR COSMOS testbed, which is being deployed in a dense urban area in West Harlem, New York City. We provided a brief overview of testbed design and architecture and presented several experiments and a measurement campaign, including: the integration of the first open-access wideband FD radios, a pilot study on converged optical wireless x-haul networking for cloud radio access network (C-RAN) applications, and extensive 28 GHz channel measurements in the testbed area. Our results and experience helped drive and validate the testbed design, and can further inform the testbed's deployment and facilitate the research in different areas.

In summary, this thesis focused on various emerging technologies that can increase network capacity and data rates as well as reduce communication latency, which can support a broad new class of real-time applications and network services. We believe that the contributions made in this thesis are essential for realizing the vision of future wireless networks across all layers of the network stack.

References

[1] S. Mumtaz, A. Alsohaily, Z. Pang, A. Rayes, K. F. Tsang, and J. Rodriguez, "Massive internet of things for industrial applications: Addressing wireless iiot connectivity challenges and ecosystem fragmentation," *IEEE Industrial Electronics Magazine*, vol. 11, no. 1, pp. 28–33, 2017.

[2] Ericsson, *Cellular networks for massive IoT*, `https : / / www . ericsson . com / 48ff1f / assets / local / reports – papers / white – papers / massive_ iot_whitepaper.pdf`, 2020.

[3] E. Dahlman, G. Mildh, S. Parkvall, J. Peisa, J. Sachs, Y. Selén, and J. Sköld, "5G wireless access: Requirements and realization," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 42–47, 2014.

[4] Qualcomm, *5G: The fabric for society*, `https : / / www . qualcomm . com / media / documents/files/5g-vision-use-cases.pdf`, 2018.

[5] ——, *5G evolution and 6G*, `https : / / www . nttdocomo . co . jp / english / binary / pdf / corporate / technology / whitepaper_ 6g / DOCOMO_ 6G_ White_PaperEN_20200124.pdf`, 2020.

[6] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Computer Networks*, vol. 54, no. 15, pp. 2787 –2805, 2010.

[7] M. Gorlatova, P. Kinget, I. Kymissis, D. Rubenstein, X. Wang, and G. Zussman, "Energy harvesting active networked tags (EnHANTs) for ubiquitous object networking," *IEEE Wireless Communications*, vol. 17, no. 6, pp. 18–25, 2010.

[8] A. Sabharwal, P. Schniter, D. Guo, D. W. Bliss, S. Rangarajan, and R. Wichman, "In-band full-duplex wireless: Challenges and opportunities," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 9, pp. 1637–1652, 2014.

[9] H. Krishnaswamy and G. Zussman, "1 chip 2x the bandwidth," *IEEE Spectrum*, vol. 53, no. 7, pp. 38–54, 2016.

[10] *The Columbia EnHANTs project*, `https://enhants.ee.columbia.edu/`.

[11] *The Columbia FlexICoN project*, `https://flexicon.ee.columbia.edu/`.

[12] *Open-access research testbed for next-generation wireless networks (ORBIT)*, `http:// www.orbit-lab.org`.

[13] D. Raychaudhuri, I. Seskar, G. Zussman, T. Korakis, D. Kilper, T. Chen, J. Kolodziejski, M. Sherman, Z. Kostic, X. Gu, H. Krishnaswamy, S. Maheshwari, P. Skrimponis, and C. Gutterman, "Challenge: COSMOS: A city-scale programmable testbed for experimentation with advanced wireless," in *Proc. ACM MobiCom'20*, 2020.

[14] *Cloud enhanced open software defined mobile wireless testbed for city-scale deployment (COSMOS)*, `https://cosmos-lab.org/`, 2020.

[15] *5G: Study on channel model for frequencies from 0.5 to 100 GHz (3GPP TR 38.901 version 14.0.0 Release 14)*, `https://www.etsi.org/deliver/etsi_tr/138900_138999/138901/14.00.00_60/tr_138901v140000p.pdf`, 2017.

[16] R. Margolies, G. Grebla, T. Chen, D. Rubenstein, and G. Zussman, "Panda: Neighbor discovery on a power harvesting budget," in *Proc. IEEE INFOCOM*, 2016.

[17] ——, "Panda: Neighbor discovery on a power harvesting budget," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3606–3619, 2016.

[18] T. Chen, G. Chen, S. Jain, R. Margolies, G. Grebla, D. Rubenstein, and G. Zussman, "Demo abstract: Power-aware neighbor discovery for energy harvesting things," in *Proc. ACM SenSys'16*, 2016.

[19] T. Chen, J. Ghaderi, D. Rubenstein, and G. Zussman, "Maximizing broadcast throughput under ultra-low-power constraints," in *Proc. ACM CoNEXT'16*, 2016.

[20] ——, "Maximizing broadcast throughput under ultra-low-power constraints," *IEEE/ACM Trans. Netw.*, vol. 26, no. 2, pp. 779–792, 2018.

[21] D. Rubenstein, G. Zussman, J. Ghaderi, R. Margolies, T. Chen, and G. Grebla, *Systems and methods for throughput enhancement among ultra-low power wireless network devices*, US Patent 10,200,956, 2019.

[22] T. Chen, M. B. Dastjerdi, J. Zhou, H. Krishnaswamy, and G. Zussman, "Wideband full-duplex wireless via frequency-domain equalization: Design and experimentation," in *Proc. ACM MobiCom'19*, 2019.

[23] T. Chen, J. Zhou, G. Zussman, and H. Krishnaswamy, "Integrated full-duplex radios: System concepts, implementations, and experimentation," in *Full-Duplex Communications for Future Wireless Networks*, Springer, 2020, pp. 299–320.

[24] T. Chen, M. B. Dastjerdi, J. Welles, J. Zhou, H. Krishnaswamy, and G. Zussman, "Poster: Enabling wideband full-duplex wireless via frequency-domain equalization," in *Proc. ACM MobiCom'19*, 2019.

[25] A. Nagulu, T. Chen, G. Zussman, and H. Krishnaswamy, "Non-magnetic 0.18 SOI circulator with multi-watt power handling based on switched-capacitor clock boosting," in *Proc. IEEE ISSCC'20*, 2020.

[26] ——, "Multi-watt, 1-GHz CMOS circulator based on switched-capacitor clock boosting," *IEEE J. Solid-State Circuits (to appear)*, 2020.

[27] T. Chen, M. B. Dastjerdi, H. Krishnaswamy, and G. Zussman, "Wideband full-duplex phased array with joint transmit and receive beamforming: Optimization and rate gains," in *Proc. ACM MobiHoc'19*, 2019.

[28] M. B. Dastjerdi, N. Reiskarimian, T. Chen, G. Zussman, and H. Krishnaswamy, "Full duplex circulator-receiver phased array employing self-interference cancellation via beamforming," in *Proc. IEEE RFIC'18*, 2018.

[29] T. Chen, J. Diakonikolas, J. Ghaderi, and G. Zussman, "Hybrid scheduling in heterogeneous half- and full-duplex wireless networks," in *Proc. IEEE INFOCOM'18*, 2018.

[30] ——, "Hybrid scheduling in heterogeneous half-and full-duplex wireless networks," *IEEE/ACM Trans. Netw.*, vol. 28, no. 2, pp. 764–777, 2020.

[31] J. Zhou, N. Reiskarimian, J. Diakonikolas, T. Dinc, T. Chen, G. Zussman, and H. Krishnaswamy, "Integrated full duplex radios," *IEEE Commun. Mag.*, vol. 55, no. 4, pp. 142–151, 2017.

[32] N. Reiskarimian, T. Dinc, J. Zhou, T. Chen, M. B. Dastjerdi, J. Diakonikolas, G. Zussman, and H. Krishnaswamy, "One-way ramp to a two-way highway: Integrated magnetic-free nonreciprocal antenna interfaces for full-duplex wireless," *IEEE Microw. Mag.*, vol. 20, no. 2, pp. 56–75, 2019.

[33] H. Krishnaswamy, G. Zussman, J. Zhou, J. Marašević, T. Dinc, N. Reiskarimian, and T. Chen, "Full-duplex in a hand-held device—from fundamental physics to complex integrated circuits, systems and networks: An overview of the columbia flexicon project," in *Proc. Asilomar'16*, 2016.

[34] J. Marašević, T. Chen, J. Zhou, N. Reiskarimian, H. Krishnaswamy, and G. Zussman, "Full-duplex wireless: Algorithms and rate improvement bounds for integrated circuit implementations," in *Proc. ACM HotWireless'16*, 2016.

[35] T. Chen, J. Zhou, M. B. Dastjerdi, J. Diakonikolas, H. Krishnaswamy, and G. Zussman, "Demo abstract: Full-duplex with a compact frequency domain equalization-based RF canceller," in *Proc. IEEE INFOCOM'17*, 2017.

[36] T. Chen, J. Zhou, N. Grimwood, R. Fogel, J. Marašević, H. Krishnaswamy, and G. Zussman, "Demo: Full-duplex wireless based on a small-form-factor analog self-interference canceller," in *Proc. ACM MobiHoc'16*, 2016.

[37] M. Kohli, T. Chen, M. B. Dastjerdi, J. Welles, I. Seskar, H. Krishnaswamy, and G. Zussman, "Open-access full-duplex wireless in the ORBIT and COSMOS testbeds," in *Proc. ACM MobiCom'20 Workshop on Wireless Network Testbeds, Experimental evaluation & CHaracterization (WiNTECH)*, 2020.

[38] T. Chen, M. B. Dastjerdi, G. Farkash, J. Zhou, H. Krishnaswamy, and G. Zussman, "Open-access full-duplex wireless in the ORBIT testbed," *arXiv preprint arXiv:1801.03069v2*, 2018.

[39] T. Chen, J. Welles, M. Kohli, M. B. Dastjerdi, J. Kolodziejski, M. Sherman, I. Seskar, H. Krishnaswamy, and G. Zussman, "Programmable optical x-haul network in the COSMOS testbed," in *Proc. IEEE ICNP'19 Workshop on Midscale Education and Research Infrastructure and Tools (MERIT)*, 2019.

[40] T. Chen, M. B. Dastjerdi, G. Farkash, J. Zhou, H. Krishnaswamy, and G. Zussman, "Demo abstract: Open-access full-duplex wireless in the ORBIT testbed," in *Proc. IEEE INFOCOM'18*, 2018.

[41] T. Chen, M. Kohli, T. Dai, A. D. Estigarribia, D. Chizhik, J. Du, R. Feick, R. A. Valenzuela, and G. Zussman, "28 GHz channel measurements in the COSMOS testbed deployment area," in *Proc. ACM MobiCom'19 Workshop on Millimeter-Wave Networks and Sensing System (mmNets)*, 2019.

[42] J. Du, D. Chizhik, R. Feick, M. Rodriguez, G. Castro, R. Valenzuela, *et al.*, "Suburban fixed wireless access channel measurements and models at 28 GHz for 90% outdoor coverage," *arXiv preprint arXiv:1807.03763*, 2018.

[43] A. Minakhmetov, C. Gutterman, T. Chen, C. Ware, L. Iannone, D. Kilper, and G. Zussman, "Experiments on cloud-RAN wireless handover using optical switching in a dense urban testbed," in *Proc. OSA OFC'20, Th2A.25*, 2020.

[44] J. Yu, C. Gutterman, A. Minakhmetov, M. Sherman, T. Chen, S. Zhu, G. Zussman, I. Seskar, and D. Kilper, "Dual use SDN controller for management and experimentation in a field deployed testbed," in *Proc. OSA OFC'20, T3J.3*, 2020.

[45] J. Yu, T. Chen, C. Gutterman, S. Zhu, G. Zussman, I. Seskar, and D. Kilper, "COSMOS: Optical architecture and prototyping," in *Proc. OSA OFC'19, M3G.3 (invited)*, 2019.

[46] C. Gutterman, A. Minakhmetov, J. Yu, M. Sherman, T. Chen, S. Zhu, I. Seskar, D. Raychaudhuri, D. Kilper, and G. Zussman, "Experimentation with full-duplex wireless in the

COSMOS testbed," in *Proc. IEEE ICNP'19 Workshop on Midscale Education and Research Infrastructure and Tools (MERIT)*, 2019.

[47] M. Radi, B. Dezfouli, K. A. Bakar, S. A. Razak, and M. Lee, "Network initialization in low-power wireless networks: A comprehensive study," *The Computer Journal*, vol. 57, no. 8, pp. 1238–1261, 2014.

[48] W. Sun, Z. Yang, X. Zhang, and Y. Liu, "Energy-efficient neighbor discovery in mobile ad hoc and wireless sensor networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 3, pp. 1448–1459, 2014.

[49] P. Dutta and D. Culler, "Practical asynchronous neighbor discovery and rendezvous for mobile sensing applications," in *Proc. ACM SenSys'08*, 2008.

[50] M. Bakht, M. Trower, and R. H. Kravets, "Searchlight: Won't you be my neighbor?" In *Proc. ACM MobiCom'12*, 2012.

[51] A. Kandhalu, K. Lakshmanan, and R. R. Rajkumar, "U-connect: A low-latency energy-efficient asynchronous neighbor discovery protocol," in *Proc. ACM/IEEE IPSN'10*, 2010.

[52] W. Sun, Z. Yang, K. Wang, and Y. Liu, "Hello: A generic flexible protocol for neighbor discovery," in *Proc. IEEE INFOCOM'14*, 2014.

[53] A. Purohit, B. Priyantha, and J. Liu, "WiFlock: Collaborative group discovery and maintenance in mobile sensor networks," in *Proc. ACM/IEEE IPSN'11*, 2011.

[54] L. You, Z. Yuan, P. Yang, and G. Chen, "ALOHA-like neighbor discovery in low-duty-cycle wireless sensor networks," in *Proc. IEEE WCNC'11*, 2011.

[55] M. J. McGlynn and S. A. Borbash, "Birthday protocols for low energy deployment and flexible neighbor discovery in ad hoc wireless networks," in *Proc. ACM MobiHoc'01*, 2001.

[56] W. Ye, J. Heidemann, and D. Estrin, "Medium access control with coordinated adaptive sleeping for wireless sensor networks," *IEEE/ACM Trans. Netw.*, vol. 12, no. 3, pp. 493–506, 2004.

[57] L. Chen, R. Fan, K. Bian, M. Gerla, T. Wang, and X. Li, "On heterogeneous neighbor discovery in wireless sensor networks," *arXiv preprint arXiv:1411.5415*, 2014.

[58] A. El-Hoiydi and J.-D. Decotignie, "Wisemac: An ultra low power MAC protocol for the downlink of infrastructure wireless sensor networks," in *Proc. IEEE ISCC'04*, 2004.

[59] S. Vasudevan, M. Adler, D. Goeckel, and D. Towsley, "Efficient algorithms for neighbor discovery in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 21, no. 1, pp. 69–83, 2013.

[60] M. Chiang, S. H. Low, J. C. Doyle, *et al.*, "Layering as optimization decomposition: A mathematical theory of network architectures," *Proc. IEEE*, vol. 95, no. 1, pp. 255–312, 2007.

[61] J. Ghaderi and R Srikant, "On the design of efficient CSMA algorithms for wireless networks," in *Proc. IEEE CDC'10*, 2010.

[62] J. Ghaderi, S. Borst, and P. Whiting, "Queue-based random-access algorithms: Fluid limits and stability issues," *Stochastic Systems*, vol. 4, no. 1, pp. 81–156, 2014.

[63] P. Marbach and A. Eryilmaz, "A backlog-based CSMA mechanism to achieve fairness and throughput-optimality in multihop wireless networks," in *Proc. Allerton Conf.*, 2008.

[64] L. Jiang and J. Walrand, "A distributed CSMA algorithm for throughput and utility maximization in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 18, pp. 960–972, 2010.

[65] K. Xu, O. Dousse, and P. Thiran, "Self-synchronizing properties of CSMA wireless multi-hop networks," in *Proc. ACM SIGMETRICS'10*, 2010.

[66] S. Chen, P. Sinha, N. B. Shroff, and C. Joo, "A simple asymptotically optimal joint energy allocation and routing scheme in rechargeable sensor networks," *IEEE/ACM Trans. Netw.*, vol. 22, no. 4, pp. 1325–1336, 2014.

[67] J. Taneja, J. Jeong, and D. Culler, "Design, modeling, and capacity planning for micro-solar power sensor networks," in *Proc. IEEE IPSN'08*, 2008.

[68] *EnOcean Energy Harvesting Wireless Sensor Solutions and Networks*, `http://www.enocean.com/`.

[69] R. Margolies, M. Gorlatova, J. Sarik, G. Stanje, J. Zhu, P. Miller, M. Szczodrak, B. Vigraham, L. Carloni, P. Kinget, *et al.*, "Energy-harvesting active networked tags (EnHANTs): Prototyping and experimentation," *ACM Trans. Sens. Netw.*, vol. 11, no. 4, p. 62, 2015.

[70] P. Zhang and D. Ganesan, "Enabling bit-by-bit backscatter communication in severe energy harvesting environments," in *Proc. USENIX NSDI'14*, 2014.

[71] Y.-s. Kuo, P. Pannuto, G. Kim, Z. Foo, I. Lee, B. Kempke, P. Dutta, D. Blaauw, and Y. Lee, "Mbus: A 17.5 pJ/bit/chip portable interconnect bus for millimeter-scale sensor systems with 8 nW standby power," in *Proc. IEEE CICC*, 2014.

[72] A. El-Hoiydi and J.-D. Decotignie, "WiseMAC: An ultra low power MAC protocol for the downlink of infrastructure wireless sensor networks," in *Proc. IEEE ISCC'04*, 2004.

[73] A. Dunkels, "The ContikiMAC radio duty cycling protocol," 2011, Swedish Institute of Computer Science.

[74] *Texas Instruments CC2640 SimpleLink bluetooth smart wireless MCU*, `ti.com/lit/ds/symlink/cc2640.pdf`, 2015.

[75] *Texas Instruments MSP430 solar energy harvesting development tool*, `http://focus.ti.com/docs/toolsw/folders/print/ez430-rf2500-seh.html`, 2012.

[76] K. E. Kolodziej, B. T. Perry, and J. S. Herd, "In-band full-duplex technology: Techniques and systems survey," *IEEE Trans. Microw. Theory Techn.*, vol. 67, no. 7, pp. 3025–3041, 2019.

[77] B. Radunovic, D. Gunawardena, P. Key, A. Proutiere, N. Singh, V. Balan, and G. Dejean, "Rethinking indoor wireless mesh design: Low power, low frequency, full-duplex," in *Proc. IEEE WIMESH'10*, 2010.

[78] J. I. Choi, M. Jain, K. Srinivasan, P. Levis, and S. Katti, "Achieving single channel, full duplex wireless communication," in *Proc. ACM MobiCom'10*, 2010.

[79] M. Jain, J. I. Choi, T. Kim, D. Bharadia, S. Seth, K. Srinivasan, P. Levis, S. Katti, and P. Sinha, "Practical, real-time, full duplex wireless," in *Proc. ACM MobiCom'11*, 2011.

[80] E. Aryafar, M. A. Khojastepour, K. Sundaresan, S. Rangarajan, and M. Chiang, "MIDU: Enabling MIMO full duplex," in *Proc. ACM MobiCom'12*, 2012.

[81] D. Bharadia, E. McMilin, and S. Katti, "Full duplex radios," in *Proc. ACM SIGCOMM'13*, 2013.

[82] M. K. Chung, M. S. Sim, J. Kim, D. K. Kim, and C.-B. Chae, "Prototyping real-time full duplex radios," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 56–63, 2015.

[83] D. Korpi, J. Tamminen, M. Turunen, T. Huusari, Y.-S. Choi, L. Anttila, S. Talwar, and M. Valkama, "Full-duplex mobile device: Pushing the limits," *IEEE Commun, Mag.*, vol. 54, no. 9, pp. 80–87, 2016.

[84] M. Duarte, A. Sabharwal, V. Aggarwal, R. Jana, K. Ramakrishnan, C. W. Rice, and N. Shankaranarayanan, "Design and characterization of a full-duplex multiantenna system for WiFi networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 3, pp. 1160–1177, 2013.

[85] D. Bharadia and S. Katti, "Full duplex MIMO radios," in *Proc. USENIX NSDI'14*, 2014.

[86] B. Chen, V. Yenamandra, and K. Srinivasan, "FlexRadio: Fully flexible radios and networks," in *Proc. USENIX NSDI'15*, 2015.

[87] M. Chung, M. S. Sim, D. Kim, and C.-B. Chae, "Compact full duplex MIMO radios in D2D underlaid cellular networks: From system design to prototype results," *IEEE Access*, vol. 5, pp. 16 601–16 617, 2017.

[88] D. Bharadia and S. Katti, "Fastforward: Fast and constructive full duplex relays," *ACM SIGCOMM CCR*, vol. 44, no. 4, pp. 199–210, 2015.

[89] L. Chen, F. Wu, J. Xu, K. Srinivasan, and N. Shroff, "BiPass: Enabling end-to-end full duplex," in *Proc. ACM MobiCom'17*, 2017.

[90] E. Everett, C. Shepard, L. Zhong, and A. Sabharwal, "Softnull: Many-antenna full-duplex wireless via digital beamforming," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8077–8092, 2016.

[91] E. Aryafar and A. Keshavarz-Haddad, "PAFD: Phased array full-duplex," in *Proc. IEEE INFOCOM'18*, 2018.

[92] E. Ahmed, A. M. Eltawil, and A. Sabharwal, "Rate gain region and design tradeoffs for full-duplex wireless communications," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 3556–3565, 2013.

[93] J. Bai and A. Sabharwal, "Distributed full-duplex via wireless side-channels: Bounds and protocols," *IEEE Trans. Wireless Commun.*, vol. 12, no. 8, pp. 4162–4173, 2013.

[94] S. Goyal, P. Liu, S. S. Panwar, R. A. Difazio, R. Yang, and E. Bala, "Full duplex cellular systems: Will doubling interference prevent doubling capacity?" *IEEE Commun. Mag.*, vol. 53, no. 5, pp. 121–127, 2015.

[95] J. Marasevic, J. Zhou, H. Krishnaswamy, Y. Zhong, and G. Zussman, "Resource allocation and rate gains in practical full-duplex systems," *IEEE/ACM Trans. Netw.*, vol. 25, no. 1, pp. 292–305, 2017.

[96] J. Diakonikolas and G. Zussman, "On the rate regions of single-channel and multi-channel full-duplex links," *IEEE/ACM Trans. Netw.*, vol. 26, no. 1, pp. 47–60, 2018.

[97] G. Zheng, "Joint beamforming optimization and power control for full-duplex MIMO two-way relay channel," *IEEE Trans. Signal Process.*, vol. 63, no. 3, 2015.

[98] Z. Qian, F. Wu, Z. Zheng, K. Srinivasan, and N. B. Shroff, "Concurrent channel probing and data transmission in full-duplex MIMO systems," in *Proc. ACM MobiHoc'17*, 2017.

[99] X. Xie and X. Zhang, "Does full-duplex double the capacity of wireless networks?" In *Proc. IEEE INFOCOM'14*, 2014.

[100] X. Qin, H. Zeng, X. Yuan, B. Jalaian, Y. T. Hou, W. Lou, and S. F. Midkiff, "Impact of full duplex scheduling on end-to-end throughput in multi-hop wireless networks," *IEEE Trans. Mobile Comput.*, vol. 16, no. 1, pp. 158–171, 2016.

[101] S. Wang, V. Venkateswaran, and X. Zhang, "Fundamental analysis of full-duplex gains in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 25, no. 3, pp. 1401–1416, 2017.

[102] J. Y. Kim, O. Mashayekhi, H. Qu, M. Kazandjieva, and P. Levis, "Janus: A novel MAC protocol for full duplex radio," *CSTR*, vol. 2, no. 7, 2013.

[103] M. Duarte, A. Sabharwal, V. Aggarwal, R. Jana, K. Ramakrishnan, C. W. Rice, and N. Shankaranarayanan, "Design and characterization of a full-duplex multiantenna system for WiFi networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 3, pp. 1160–1177, 2014.

[104] S. Goyal, P. Liu, O. Gurbuz, E. Erkip, and S. Panwar, "A distributed MAC protocol for full duplex radio," in *Proc. Asilomar'13*, 2013.

[105] S.-Y. Chen, T.-F. Huang, K. C.-J. Lin, Y.-W. P. Hong, and A. Sabharwal, "Probabilistic medium access control for full-duplex networks with half-duplex clients," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2627–2640, 2017.

[106] A. Sahai, G. Patel, and A. Sabharwal, "Pushing the limits of full-duplex: Design and real-time implementation," *arXiv preprint: 1107.0607*, 2011.

[107] A. Tang and X. Wang, "A-duplex: Medium access control for efficient coexistence between full-duplex and half-duplex communications," *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, pp. 5871–5885, 2015.

[108] Y. Yang and N. B. Shroff, "Scheduling in wireless networks with full-duplex cut-through transmission," in *Proc. IEEE INFOCOM'15*, 2015.

[109] M. A. Alim, M. Kobayashi, S. Saruwatari, and T. Watanabe, "In-band full-duplex medium access control design for heterogeneous wireless LAN," *EURASIP J. Wireless Commun. and Netw.*, no. 1, 2017.

[110] *IEEE 802.11 full duplex topic interest group*, `https://mentor.ieee.org/802.11/dcn/18/11-18-0191-01-0wng-full-duplex-for-802-11.pptx`.

[111] G. Zheng, I. Krikidis, J. Li, A. P. Petropulu, and B. Ottersten, "Improving physical layer secrecy using full-duplex jamming receivers," *IEEE Trans. Signal Process.*, vol. 61, no. 20, pp. 4962–4974, 2013.

[112] D. W. K. Ng, Y. Sun, and R. Schober, "Power efficient and secure full-duplex wireless communication systems," in *Proc. IEEE CNS'15*, 2015.

[113] G. Chen, Y. Gong, P. Xiao, and J. A. Chambers, "Physical layer network security in the full-duplex relay system," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 3, pp. 574–583, 2015.

[114] Y. Liu, Y. Shen, D. Guo, and M. Z. Win, "Network localization and synchronization using full-duplex radios," *IEEE Trans. Signal Process.*, vol. 66, no. 3, pp. 714–728, 2018.

[115] G. Sun, F. Wu, X. Gao, and G. Chen, "PHED: Pre-handshaking neighbor discovery protocols in full duplex wireless ad hoc networks," in *Proc. IEEE GLOBECOM'12*, 2012.

[116] *Bristol Is Open*, `http://www.bristolisopen.com/`.

[117] *ADRENALINE testbed*, `http://networks.cttc.es/ons/adrenaline/sdnnfv-cloud-computing-platform-and-core-network-for-5g-services/`.

[118] *POWDER-RENEW*, `https://powderwireless.net/`.

[119] R. Ricci, E. Eide, and C. Team, "Introducing CloudLab: Scientific infrastructure for advancing cloud architectures and applications," *The USENIX Magazine*, vol. 39, no. 6, pp. 36–38, 2014.

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

[121] G. R. MacCartney, M. K. Samimi, and T. S. Rappaport, "Omnidirectional path loss models in New York City at 28 GHz and 73 GHz," in *Proc. IEEE PIMRC'14*, 2014.

[122] T. S. Rappaport, G. R. MacCartney, M. K. Samimi, and S. Sun, "Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design," *IEEE Trans. Commun.*, vol. 63, no. 9, pp. 3029–3056, 2015.

[123] A. F. Molisch, A. Karttunen, R. Wang, C. U. Bas, S. Hur, J. Park, and J. Zhang, "Millimeter-wave channels in urban environments," in *Proc. EuCAP'16*, 2016.

[124] T. S. Rappaport, F. Gutierrez, E. Ben-Dor, J. N. Murdock, Y. Qiao, and J. I. Tamir, "Broadband millimeter-wave propagation measurements and models using adaptive-beam antennas for outdoor urban cellular communications," *IEEE Trans. Antennas Propga.*, vol. 61, no. 4, pp. 1850–1859, 2012.

[125] C. U. Bas, R. Wang, S. Sangodoyin, S. Hur, K. Whang, J. Park, J. Zhang, and A. F. Molisch, "28 GHz microcell measurement campaign for residential environment," in *Proc. IEEE GLOBECOM'17*, 2017.

[126] Y. Zhang, S. Jyoti, C. R. Anderson, D. J. Love, N. Michelusi, A. Sprintson, and J. V. Krogmeier, "28 GHz channel measurements and modeling for suburban environments," in *Proc. IEEE ICC'18*, 2018.

252

[127] J. Zhang, X. Zhang, P. Kulkarni, and P. Ramanathan, "OpenMili: A 60 GHz software radio with a programmable phased-array antenna," in *Proc. ACM MobiCom'16*, 2016.

[128] B. Sadhu, Y. Tousi, J. Hallin, S. Sahl, S. K. Reynolds, Ö. Renström, K. Sjögren, O. Haapalahti, N. Mazor, B. Bokinge, *et al.*, "A 28-GHz 32-element TRX phased-array IC with concurrent dual-polarized operation and orthogonal phase and gain control for 5G communications," *IEEE J. Solid-State Circuits*, vol. 52, no. 12, pp. 3373–3391, 2017.

[129] S. K. Saha, Y. Ghasempour, M. K. Haider, T. Siddiqui, P. De Melo, N. Somanchi, L. Zakrajsek, A. Singh, R. Shyamsunder, O. Torres, *et al.*, "X60: A programmable testbed for wideband 60 GHz WLANs with phased arrays," *Computer Communications*, vol. 133, 2019.

[130] C. U. Bas, R. Wang, S. Sangodoyin, D. Psychoudakis, T. Henige, R. Monroe, J. Park, J. Zhang, and A. F. Molisch, "Real-time millimeter-wave MIMO channel sounder for dynamic directional measurements," *arXiv preprint arXiv:1807.11921*, 2018.

[131] S. Sur, V. Venkateswaran, X. Zhang, and P. Ramanathan, "60 GHz indoor networking through flexible beams: A link-level profiling," in *Proc. ACM SIGMETRICS'15*, 2015.

[132] D. Steinmetzer, D. Wegemer, M. Schulz, J. Widmer, and M. Hollick, "Compressive millimeter-wave sector selection in off-the-shelf IEEE 802.11ad devices," in *Proc. ACM CoNEXT'17*, 2017.

[133] A. Mukhopadhyay and G. Das, "A ring-based wireless optical network to reduce the handover latency," *J. Lightw. Technol.*, vol. 33, no. 17, pp. 3687–3697, 2015.

[134] A. Koutsaftis, R. Kumar, P. Liu, and S. S. Panwar, "Fast inter-base station ring (FIBR): A new millimeter wave cellular network architecture," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 12, pp. 2699–2714, 2019.

[135] X. Wang, S. Thota, M. Tornatore, S.-S. Lee, H.-H. Lee, S. Park, and B. Mukherjee, "Handover reduction via mobility-prediction-based VPON formation in optical-access-enabled cloud-RAN," in *Proc. IEEE/OSA OFC'15*, 2015.

[136] M. Gorlatova, A. Wallwater, and G. Zussman, "Networking low-power energy harvesting devices: Measurements and algorithms," *IEEE Trans. on Mobile Comput.*, vol. 12, no. 9, pp. 1853–1865, 2012.

[137] *SANYO amorphous solar cell AM-1815*, `farnell.com/datasheets/87124.pdf`, 2008.

[138] K. Tutuncuoglu and A. Yener, "Optimum transmission policies for battery limited energy harvesting nodes," *IEEE Trans. Wireless Commun.*, vol. 11, no. 3, pp. 1180–1189, 2012.

[139] *Texas Instruments MSP430 mixed signal microcontroller*, `ti.com/lit/ds/symlink/msp430f2274.pdf`, 2006.

[140] *Texas Instruments low-cost low-power 2.4 GHz RF transceiver (SWRS 040C)*, `www.ti.com/lit/ds/swrs040c/swrs040c.pdf`, 2015.

[141] C. M. Vigorito, D. Ganesan, and A. G. Barto, "Adaptive control of duty cycling in energy-harvesting wireless sensor networks," in *Proc. IEEE SECON'07*, 2007.

[142] B. Dezfouli, M. Radi, S. Abd Razak, K. Whitehouse, K. A. Bakar, and T. Hwee-Pink, "Improving broadcast reliability for neighbor discovery, link estimation and collection tree construction in wireless sensor networks," *Computer Networks*, vol. 62, pp. 101–121, 2014.

[143] E. Cinlar, "Markov renewal theory," *Advances in Applied Probability*, vol. 1, no. 2, pp. 123–187, 1969.

[144] M. Gorlatova, J. Sarik, G. Grebla, M. Cong, I. Kymissis, and G. Zussman, "Movers and shakers: Kinetic energy harvesting for the internet of things," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 8, pp. 1624–1639, 2015.

[145] T. Chen, J. Ghaderi, D. Rubenstein, and G. Zussman, "Maximizing broadcast throughput under ultra-low-power constraints," *arXiv preprint arXiv:1610.04203v2*, 2017.

[146] A. Proutiere, Y. Yi, T. Lan, and M. Chiang, "Resource allocation over network dynamics without timescale separation," in *Proc. IEEE INFOCOM'10*, 2010.

[147] L. Jiang and J. Walrand, "Convergence and stability of a distributed CSMA algorithm for maximal network throughput," in *Proc. CDC'09*, 2009.

[148] K. Philips, "Ultra low power short range radios: Covering the last mile of the IoT," in *Proc. IEEE ESSCIRC'14*, 2014.

[149] *Texas Instruments eZ430-RF2500-SEH solar energy harvesting development tool user's guide*, 2013.

[150] J. Wang, H. Hassanieh, D. Katabi, and P. Indyk, "Efficient and reliable low-power backscatter networks," in *Proc. ACM SIGCOMM'12*, 2012.

[151] J. Zhou, A. Chakrabarti, P. Kinget, and H. Krishnaswamy, "Low-noise active cancellation of transmitter leakage and transmitter noise in broadband wireless receivers for FDD/co-existence," *IEEE J. Solid-State Circuits*, vol. 49, no. 12, pp. 1–17, 2014.

[152] J. Zhou, T.-H. Chuang, T. Dinc, and H. Krishnaswamy, "Integrated wideband self-interference cancellation in the RF domain for FDD and full-duplex wireless," *IEEE J. Solid-State Circuits*, vol. 50, no. 12, pp. 3015–3031, 2015.

[153] W. Zhou, T. Das, L. Chen, K. Srinivasan, and P. Sinha, "BASIC: Backbone-assisted successive interference cancellation," in *Proc. ACM MobiCom'16*, 2016.

[154] D. López-Pérez, M. Ding, H. Claussen, and A. H. Jafari, "Towards 1 Gbps/UE in cellular systems: Understanding ultra-dense small cell deployments," *IEEE Commun. Surv. Tut.*, vol. 17, no. 4, pp. 2078–2101, 2015.

[155] A. B. Constantine, *Antenna theory: Analysis and design, third edition*. John Wiley & Sons, 2005.

[156] *Argos full-duplex channel measurement dataset*, `http://data.argos.rice.edu/`, 2016.

[157] D. P. Bertsekas and J. N. Tsitsiklis, "Convergence rate and termination of asynchronous iterative algorithms," in *Proc. ACM ICS'89*, 1989.

[158] *RF-CI coaxial SMA connector circulators*, `http://www.rf-ci.com/circulators/coaxial-circulators/sma-connector-thin-circulator-communication-band/`.

[159] C. Shepard, H. Yu, N. Anand, E. Li, T. Marzetta, R. Yang, and L. Zhong, "Argos: Practical many-antenna base stations," in *Proc. ACM MobiCom'12*, 2012.

[160] *Taoglas stingray WA.500w glassmount 2.4/5.8GHz antenna*, `https://www.taoglas.com/product/stingray-wa-500w-glassmount-2-45-8ghz/`.

[161] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Trans. Autom. Control*, vol. 37, no. 12, pp. 1936–1948, 1992.

[162] A. Dimakis and J. Walrand, "Sufficient conditions for stability of longest-queue-first scheduling: Second-order properties using fluid limits," *Adv. Appl. Prob.*, vol. 38, no. 2, pp. 505–521, 2006.

[163] J. Ghaderi and R. Srikant, "On the design of efficient CSMA algorithms for wireless networks," in *Proc. IEEE CDC'10*, 2010.

[164] J. Ni, B. Tan, and R. Srikant, "Q-CSMA: Queue-length-based CSMA/CA algorithms for achieving maximum throughput and low delay in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 20, no. 3, 2012.

[165] B. Birand, M. Chudnovsky, B. Ries, P. Seymour, G. Zussman, and Y. Zwols, "Analyzing the performance of greedy maximal scheduling via local pooling and graph theory," *IEEE/ACM Trans. Netw.*, vol. 20, no. 1, pp. 163–176, 2012.

[166] T. Chen, J. Diakonikolas, J. Ghaderi, and G. Zussman, "Fairness and delay in heterogeneous half-and full-duplex wireless networks," in *Proc. Asilomar Conference on Signals, Systems, and Computers (Asilomar)*, 2018.

[167] J. Marašević and G. Zussman, "On the capacity regions of single-channel and multi-channel full-duplex links," in *Proc. ACM MobiHoc'16*, 2016.

[168] W. Whitt, "Weak convergence of probability measures on the function space c[0,infty)," *Ann. of Math. Stat.*, vol. 41, no. 3, pp. 939–944, 1970.

[169] J. G. Dai, "On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models," *Ann. Appl. Prob.*, pp. 49–77, 1995.

[170] G. R. Gupta and N. B. Shroff, "Delay analysis for wireless networks with single hop traffic and general interference constraints," *IEEE/ACM Trans. Netw.*, vol. 18, no. 2, pp. 393–405, 2010.

[171] O. J. Boxma, "Workloads and waiting times in single-server systems with multiple customer classes," *Queueing Systems*, vol. 5, no. 1-3, 1989.

[172] J. Ghaderi and R Srikant, "The impact of access probabilities on the delay performance of Q-CSMA algorithms in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 21, no. 4, pp. 1063–1075, 2013.

[173] M. Feuillet, A. Proutiere, and P. Robert, "Random capture algorithms fluid limits and stability," in *Proc. IEEE ITA'10*, 2010.

[174] J. Ghaderi, S. Borst, and P. Whiting, "Queue-based random-access algorithms: Fluid limits and stability issues," *Stochastic Systems*, vol. 4, no. 1, pp. 81–156, 2014.

[175] N. Bouman, S. C. Borst, and J. S. van Leeuwaarden, "Delay performance in random-access networks," *Queueing Systems*, vol. 77, no. 2, 2014.

[176] A. Bavier, N. Feamster, M. Huang, L. Peterson, and J. Rexford, "In VINI veritas: Realistic and controlled network experimentation," in *Proc. ACM SIGCOMM'06*, 2006.

[177] *Innovation zones for program experimental licenses in designated portions of new york city and salt lake city*, https://docs.fcc.gov/public/attachments/DA-19-923A1.pdf/, 2019.

[178] *COSMOS tutorials*, https://wiki.cosmos-lab.org/wiki/tutorials/, 2019.

[179] *GNU Radio*, http://gnuradio.org/.

[180] *USRP hardware driver (UHD) software*, `https://github.com/EttusResearch/uhd`.

[181] *RF-CI SMA-F Circulator Model RFCR3204, mechanical drawing and data sheet*, `http://www.rf-ci.com/wp-content/themes/rfci/pdf/sma-connector-thin-circulator-communication/CR3204-OSrA.pdf`.

[182] *DIMAX SUB-20 user manual*, `https://www.xdimax.com/sub20/doc/sub20-man.pdf`.

[183] *Pulse larsen spda24700/2700 swivel blade antenna*, `https://www.wpsantennas.com/SPDA247002700-Swivel-Blade-omnidirection.aspx`.

[184] *Tutorial: Full-duplex wireless in the ORBIT and COSMOS testbeds*, `https://wiki.cosmos-lab.org/wiki/Tutorials/Wireless/FullDuplex`, 2020.

[185] B. Bloessl, M. Segata, C. Sommer, and F. Dressler, "An IEEE 802.11a/g/p OFDM receiver for GNU Radio," in *Proc. SRIF'13*, 2013.

[186] *The Columbia FlexICoN project: Instructions and code*, `https://github.com/Wimnet/flexicon_orbit`, 2020.

[187] *Calient S320 Optical Circuit Switch*, `https://www.calient.net/products/s-series-photonic-switch/`, 2020.

[188] *Lumentum ROADM Graybox*, `https://www.lumentum.com/en/products/dci-roadm-graybox/`, 2020.

[189] M. Ruffini, "Multidimensional convergence in future 5G networks," *J. Lightw. Technol.*, vol. 35, no. 3, pp. 535–549, 2016.

[190] P. Cibula, *Mechanisms influencing BASE-T 1Gb/10Gb time-to-link*, IEEE P802.3bz 2.5/5GBASE-T, 2015.

[191] V. Gupta, C. Gutterman, Y. Bejerano, and G. Zussman, "Experimental evaluation of large scale WiFi multicast rate control," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2319–2332, 2018.