ABSTRACT

Title of Dissertation: DESIGN AND OPTIMIZATION OF 5G-&-BEYOND HYBRID COMMUNICATIONS SYSTEMS

Nariman Torkzaban
Doctor of Philosophy, 2023

Dissertation Directed by: Professor John S. Baras
Department of Electrical and Computer Engineering

5G and beyond communication systems are envisaged to fulfill three key promises that enable novel use cases and applications such as telemedicine, augmented reality/virtual reality (AR/VR), smart manufacturing, autonomous vehicles (AVs), etc. These three key promises are i) Enhanced mobile broadband (eMBB), ii) Ultra-reliable low latency Communications (URLLC), and iii) Massive machine-type communications (mMTC). In other words, 5G is required to achieve key performance indicators (KPIs) in terms of low latency, massive device connectivity, consistent quality of service (QoS), and high security. For instance, user bit-rates up to 10 Gbps and round-trip times (RTTs) as small as 1–10 ms are demanded in specific application scenarios in 5G. Toward achieving the 5G key promises, it is essential to utilize the capacity of all sorts of communications networks (terrestrial, space, aerial) and supporting technologies (SDN, NFV, etc.) simultaneously, leading to the so-called hybrid communication networks as opposed to the traditional stand-alone ones. This signifies the importance of a seamless integration and configuration policy tailored to specific use cases and QoS requirements of 5G and beyond

services and will spawn several challenging design and optimization problems from the control and management to the physical layer of next-generation systems. In this thesis, we will address such critical problems in the course of 9 chapters.

In the second chapter, we study the benefits of incorporating trust into decision-making for resource provisioning in next-generation communications networks. In this regard, we study the trust-aware service chain embedding problem for enhancing the reliability of virtual network function (VNF) placement on the trusted infrastructure. The problem of placing the VNFs on the NFV infrastructure (NFVI) and establishing the routing paths between them, according to the service chain template, is termed SFC embedding. The objectives and constraints for the optimization problem formulation of SFC embedding may vary depending on the corresponding network service. We introduce the notion of trustworthiness as a measure of security in SFC embedding and thus network service deployment. We formulate the resulting trust-aware SFC embedding problem as a Mixed Integer Linear Program (MILP). We relax the integer constraints to reduce the time complexity of the MILP formulation and obtain a Linear Program (LP). We investigate the trade-offs among the two formulations, seeking to strike a balance between results accuracy and time complexity.

The space-air-ground integrated network (SAGIN) offers potential benefits that are not possible otherwise, including global coverage, low latency, and high reliability. On the other hand, the heterogeneity of the integrated network with non-unified interfaces, and the diversity of 5G use cases with large-scale applications highlight the need for a unified management structure and a dynamic resource allocation policy that are both scalable and flexible enough to handle the increasing complexity. In the third chapter, on one hand, we optimize the integration of the hybrid network by deployment of satellite gateways on the ground segment of the network to

ensure proper connection between the layers with minimum latency, and on the other hand, we aim at providing a seamless management and control scheme for the hybrid network utilizing the capacities of the supportive technologies, software-defined networking (SDN) and network function virtualization (NFV); In particular, we study the problem of SDN controller placement with the goal maximizing the reliability of the hybrid network.

In the fourth chapter, we propose trust as a metric to measure the trustworthiness of the FL agents and thereby enhance the security of the FL training. We first elaborate on trust as a security metric by presenting a mathematical framework for trust computation and aggregation within a multi-agent system. We then discuss how this framework can be incorporated within an FL setup introducing the trusted FL algorithm for both centralized and decentralized FL. Next, we propose a framework for decentralized FL in UAV-enabled networks which involves the placement of the UAVs while ensuring the connectivity of the network of deployed UAVs.

We dedicate the remaining chapters to studying the novel design problems and the key technologies for the physical layer of next-generation wireless systems with an emphasis on millimeter-wave communications, massive MIMO, and hybrid beamforming. We introduce a novel antenna configuration called twin-ULA (TULA) and its composite configurations to generate sharp beams with maximal and uniform gain. We introduce a novel beam alignment technique to maximize the utility of transmission in the presence of multipath, efficiently utilize reconfigurable intelligent surfaces (RIS) to enhance mmWave coverage in urban environments, and synchronize and calibrate in distributed massive MIMO networks for 6G systems, where the synchronization involves the carrier frequency offset estimation and compensation, and the calibration involves mitigating reciprocity mismatches in digital and analog RF chains of the access points (APs) implementing hybrid beamforming, enabling efficient downlink channel estimation.

DESIGN AND OPTIMIZATION OF 5G AND BEYOND HYBRID COMMUNICATIONS SYSTEMS

by

Nariman Torkzaban

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Professor John S. Baras, Chair/Advisor
Dr. Mohammad A. (Amir) Khojastepour
Professor Richard J. La
Professor Behtash Babadi
Professor Bruce L. Golden

Dedication

To my best and oldest friends, Shahzad and Behzad.

Acknowledgments

First and foremost, I would like to express my gratitude to my advisor, Professor John S. Baras, for granting me the opportunity to delve into some of the most interesting, challenging, and impactful research fields of our time. It has been a pleasure to work with and learn from such an extraordinary individual who possesses a brilliant mind and a big heart.

I would also like to extend my gratitude to Professor Richard J. La, Professor Behtash Babadi, Dr. Mohammad A. (Amir) Khojastepour, and Professor Bruce L. Golden for serving on my dissertation committee. Their feedback during various stages has been instrumental in the improvement and completion of this work. I am greatly thankful to Mrs. Kim Edwards, for her generous and timely assistance with the administrative aspects of my work.

I wish to emphasize that my dissertation work would not have reached its current level of excellence without the exceptional mentorship of Dr. Khojastepour during my time as an intern at NEC Laboratories, America, Inc., and long beyond that. He tirelessly spent hours with me almost every day for more than 1.5 years in various stages of my work, from initial brainstorming sessions to tackling complex theorems. The technical rigor evident in the final five chapters of this thesis owes a significant debt to his guidance.

Furthermore, I extend my gratitude to my mentor, Yubing Jian, and my manager, Arnaud Meylan, during my internship at Qualcomm Inc. They provided invaluable insights and profound knowledge, equipping me with the tools to succeed in a thriving corporate environment.

My colleagues and lab mates at the university have enriched my graduate life in many ways and deserve a special mention. To begin with, I am immensely thankful to Professor Chrysa Papagianni for her invaluable mentorship. Her extensive expertise, profound wisdom, and instructive feedback have played a crucial role in shaping my journey along this path. I would also like to thank my old officemates, Siddharth Bansal, Omkar Ninawe, Sandeep Damera, Dr. Usman Amin Fiaz, Dr. Fatima Alimardani, Dr. Christos Mavridis, Dr. Nilesh Suriyarachchi, Dr. Faizan M. Tariq, and Dr. Leda Apergi who made the office an enjoyable place to work. Among all, Asim Zoulkarni deserves a special mention. He has been not only a great colleague but also a true friend to me, especially towards the end of my time at the University. I would also like to extend my heartfelt thanks to Hamed S. (Abdollah) and Parisa N., as well as Farshid S. and Sorour S. for providing me with the chance to cherish the wonderful gift of their friendship. Of course, my life-long friends Soroush G. and his wife Mojan J., Soroush O., and Erfan B. should always be acknowledged as part of anything I may accomplish.

I owe my deep thanks to my cousin Negar, who was always supportive of me as family and did not let me feel that I was away from home. Words cannot express the gratitude I owe to my lovely family- my beautiful mother, and my kind father- who have always stood by me, and have pulled me through against impossible odds at times even from a far distance.

Last but not least, I would like to express my deepest gratitude to my long-time partner, Anousheh, who was also my senior in our research group. Apart from the insightful discussions with her, the results of which will follow in this dissertation, her unwavering support, love, and forgiveness have been my rock throughout this journey. Her encouragement, understanding, and belief in me were my pillars of strength, and I am incredibly fortunate to have her by my side. She has truly been at the heart of my every moment and every endeavor, and I look forward to a

future filled with more adventures and accomplishments together.

This dissertation is based on research supported by the Office of Naval Research (ONR) under grant N00014-17-1-262, Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR00111990027, and grants by Leidos Corporation, and Lockheed Martin Corporation.

It is impossible to remember all, and I apologize to those I have inadvertently left out.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

5G	Fifth Generation (of wireless communication)
6G	Sixth Generation (of wireless communication)
ACI	Angular Coverage Interval
ACK	Acknowledgment
ADC	Analog-to-Digital Converter
AGC	Automatic Gain Control
AI	Artificial Intelligence
AoA	Angle of Arrival
AoD	Angle of Departure
AP	Access Point
AWGN	Additive White Gaussian Noise
BBU	Baseband Beamforming Unit
BA	Beam Alignment
BER	Bit Error Rate
BF	Beamforming
BS	Base Station
BW	Bandwidth
CFO	Carrier Frequency Offset
CIR	Channel Impulse Response
CDF	Cumulative Distribution Function
CDNs	Content Delivery Networks
CQI	Channel Quality Indicator
CoMP	Coordinated Multi-Point Transmission
CP	Cyclic Prefix
CPS	Cyber-Physical Systems
CRB	Cramer-Rao Bound
CPU	Central Processing Unit
CS	Central Server
CSI	Channel State Information
DAC/ADC	Digital-to-Analog Converter / Analog-to-Digital Converter
DA	Deterministic Annealing
DC	Data Center
DFL	Decentralized Federated Learning
DFT	Discrete Fourier Transform

DL	Deep Learning
EPC	Evolved Packet Core
ETSI	European Telecommunications Standards Institute
EWMA	Exponential Weighted Moving Average
Fed-Avg	Federated Averaging
FFT	Fast Fourier Transform
FL	Federated Learning
FW	Firewall
Gbps	Gigabits per second
GES	Generalized Exhaustive Search
GEO	Geosynchronous Equatorial Orbit
GS	Gap Subcarriers
HBF	Hybrid Beamforming
Hz	Hertz
ICI	Inter-Carrier Interference
IDS	Intrusion Detection Systems
I-BA	Interactive Beam Alignment
IID	Independent and Identically Distributed
IoT	Internet of Things
InP	Infrastructure Provider
ISTN	Integrated Satellite-Terrestrial Network
ITU	International Telecommunication Union
ISLs	Inter-Satellite Links
JGCP	Joint Gateway and Controller Placement
KPB-SCE	K-Shortest Path-Based Service Chain Embedding
KPIs	Key Performance Indicators
LCMV	Linearly Constrained Minimum Variance
LC-QAM	Low Complexity Quadrature Amplitude Modulation
LEO	Low Earth Orbit
LMMSE	Linear Minimum Mean Squared Error
LP	Linear Programming
LTE	Long-Term Evolution
MCF	Multi-Commodity Flow Allocation
MIMO	Multiple-Input, Multiple-Output
MMSE	Minimum Mean Square Error
MU	Mobile User
MU-MIMO	Multi-User Multiple Input Multiple Output
NACK	Negative Acknowledgment
NAT	Network Address Translation
NFs	Network Functions

NFV	Network Function Virtualization
NI-BA	Non-Interactive Beam Alignment
NMSE	Normalized Mean-Square Error
NP	Non-deterministic Polynomial
NSF	National Science Foundation
NS	Network Slice
NS	Null Subcarriers
NLOS	Non-Line-of-Sight
OFDM	Orthogonal Frequency-Division Multiplexing
OMP	Orthogonal Matching Pursuit
PA/LNA	Power Amplifier / Low-Noise Amplifier
PAPR	Peak-to-Average Power Ratio
PB-SCE	Path-Based Service Chain Embedding
PB-TASCE	Path-Based Trust-Aware Service Chain Embedding
PDF	Probability Density Function
PDP	Power Delay Profile
PSN	Phase Shifter Network
QoS	Quality of Service
RF	Radio Frequency
RIS	Reconfigurable Intelligent Surface
SAGINs	Satellite-Aided Ground Integrated Networks
SA	Simulated Annealing
SB	Scanning Beam
SFO	Sampling Frequency Offset
SER	Symbol Error Rate
SINR	Signal-to-Interference-plus-Noise Ratio
SNR	Signal-to-Noise Ratio
SISO	Single-Input Single-Output
SVD	Singular Value Decomposition
TB	Transmit Beam
TDD	Time-Division Duplex
TULA	Twin ULA
UAV	Unmanned Aerial Vehicle
ULA	Uniform Linear Array
UPA	Uniform Planar Array
UP-DeFeL	UAV Deployment for Decentralized Federated Learning
UR	Uncertainty Region
UE	User Equipment
UL-DL	Uplink-Downlink
ULA	Uniform Linear Array

VNE	Virtual Network Embedding
VNF	Virtual Network Function
VNF-FG	VNF Forwarding Graph
ZFBF	Zero-Forcing Beamforming

Chapter 1: Introduction

In a world where information drives our interconnected society, the ongoing development of communication systems plays a crucial role in technological advancement. This exploration delves into the intricacies of 5G, which is already revolutionizing the way we communicate, and extends into the promising domains of 6G and beyond. The interdisciplinary nature of the problems that arise as communication systems evolve, has attracted researchers from various fields including but not limited to Mathematics, Control and Systems Theory, Computer Science, and Machine Learning to address the multifaceted challenges in the design and optimization of next-generation networks. In this dissertation, we develop a set of tools, theories, and algorithms to model, design, and optimize the next-generation communication systems, that lie in the intersection of aforementioned fields. In this chapter, we first introduce some fundamental concepts that are essential to this thesis' foundation and then proceed with presenting the contributions and organization of the chapters.

1.1 Trust-aware Resource Provisioning

Given the escalating intricacy characterizing modern cyber-physical systems (CPS), the imperative to formulate an innovative framework for modeling, analyzing, and predicting their behaviors has become increasingly evident. This endeavor assumes heightened significance in

light of recent advancements in the Internet of Things (IoT), coupled with the promises of 5G to facilitate extensive machine-to-machine (M2M) communications. In this evolving landscape, tightly coupled next-generation CPS devices are poised to engage in collaborative efforts, leveraging sophisticated sensing, computing, and communication capabilities on an amplified scale enabling them to realize a wide range of applications and use cases, involving data collection, processing, and decision-making, from healthcare, vehicular networks, and smart manufacturing, to 5G service provisioning, and content delivery. All these applications heavily rely on the constant exchange of collected raw data and processed information between the collaborating agents, as opposed to the traditional case where data were collected and processed at a centralized entity. Therefore, with the heterogeneity and the large scale of the CPS, as well as the paramount importance of devising a seamless management and control scheme dealing with privacy and security threats becomes a pivotal concern.

The fact that the information is crowd-sourced by the CPS agents, to a large extent, eliminates the risk of the existence of a single point of failure and contributes to the resilience of the network, but at the same time demonstrates the need to establish trust relationships between the agents that are exchanging information. More specifically, apart from ensuring the security of communications between the network agents, it is essential to answer the following questions: (i) whether an agent refuses to share its information with other agents due to privacy concerns or conflict of interest. (ii) whether an agent manipulates the received data before processing. (iii) whether an agent intentionally or unintentionally, shares incorrect information with the rest of the network? etc.; In other words, it is essential to establish to what extent each agent of the network can be trusted. Clearly, such mechanisms of trust contribute essentially to the resilience of networked cyber-physical systems (Net-CPS).

Within the context of Net-CPS, we interpret trust as a relation between network entities that may interact or collaborate in groups toward achieving various goals. These relations are set up and updated based on the evidence generated from the previous collaboration of the agents within a protocol. Suppose the collaboration has been contributive towards the achievement of a specific goal (positive evidence). In that case, the parties accumulate their trust perspective towards one another, and otherwise (negative evidence), trust will decrease between them. Trust estimates have input to decisions such as access control, resource allocation, agent participation, and so on. The method by which trust is computed and aggregated within the network may depend on the specific application, however, we enumerate the central differences in the terminology of how the trust computation and aggregation are employed:

Centralized vs. Decentralized: Under the *centralized* regime, all the network entities rely on a central trusted party that estimates the trustworthiness level of each entity and updates all the network nodes. In this sense, all the nodes are forced to agree on the degree to which each entity is trusted as dictated by the central provider. On the other hand, under the *decentralized* approach, each user is responsible for calculating its opinion on the level of trustworthiness for each entity it might be interested in. This distinction however is irrelevant to the fashion trust is computed and only relates to the semantics of trust. For instance, under a decentralized regime, a user may utilize a distributed approach for computing the trust of its target.

Global vs. Local: *Local* trust is the opinion that a trustor node has towards a trustee and is generated depending on the first-hand evidence gathered based on local interactions, however, *global* trust is formed by combining the first-hand evidence and the opinions of other nodes about the specific trustee and is usually more accurate. In fact, the local exchange of the local observations is used towards obtaining global trust.

Proactive vs. Reactive: Under a *proactive* regime, the entities manage to keep the trust estimates updated, while under a *reactive* regime, the trust estimates are computed only when they are required. The proactive scheme is not communication efficient as a large bandwidth needs to be consumed to keep the trust values updated; therefore a reactive scheme is usually preferred unless the frequency by which trust decisions are made is comparable to the frequency of the local trust updates.

Direct vs. Indirect: Directed trust is obtained via interaction through direct communication with another agent. However, *indirect* trust is a trust relationship between two entities that have not interacted in the past. Establishing an indirect trust relationship heavily relies on the assumption that trust has the *transitivity* property which is not necessarily the case in any application.

1.2 Space-Air-Ground Integrated Networks

Space-Air-Ground Integrated Networks (SAGIN) represent an advanced and integrated communication framework that encompasses various domains: space, aerial (such as drones or aircraft), and terrestrial (ground-based) networks. The core idea behind SAGIN is to seamlessly connect these diverse components to create a robust and versatile communication infrastructure. Space-based assets, typically satellites, form the backbone of the space segment. Satellites in orbit are equipped with communication equipment that can relay data across vast distances. They provide global coverage and are especially useful for connecting remote and inaccessible areas. Space networks are instrumental in supporting applications like global internet access, Earth observation, and satellite-based navigation. The aerial component of SAGIN includes

unmanned aerial vehicles (UAVs), drones, aircraft, or high-altitude balloons. These platforms can be deployed quickly to cover specific areas or events. They serve various purposes, such as disaster response, surveillance, or temporary network coverage augmentation. The terrestrial component includes traditional cellular networks, Wi-Fi, and other ground-based infrastructure. Ground networks serve as the last-mile connectivity for end-users and play a crucial role in providing consistent connectivity.

By integrating multiple layers, SAGIN is inherently resilient. If one component encounters issues, communication can be seamlessly routed through another layer. This redundancy ensures high availability. SAGIN networks can cover a wide range of areas, from urban environments to remote and under-developed regions. This broad coverage is especially valuable for disaster response, rural connectivity, and global communication. The system can be easily scaled up or down based on requirements. For example, in a crowded event or during a disaster, additional aerial platforms or ground-based equipment can be deployed to meet increased demand. SAGIN networks can adapt to changing conditions. For example, aerial platforms can be re-positioned to address shifting communication needs in real-time. As technology advances, SAGIN networks are becoming more feasible and practical, and they represent a vital part of the evolving landscape of communication systems.

1.3 Cell-free Massive MIMO Systems

Cell-free Massive MIMO is an innovative and advanced wireless communication system architecture that departs from the traditional cellular network structure. In a Cell-free Massive MIMO system, the concept of ”cells” is replaced with a more distributed and collaborative app-

roach, offering significant benefits in terms of coverage, capacity, and reliability. In traditional cellular networks, base stations (cell towers) are responsible for providing coverage to specific geographic areas. In Cell-free Massive MIMO, the network deploys a large number of distributed access points (APs) equipped with antennas throughout the coverage area. These distributed APs are typically smaller and closer together than traditional cell towers.

Unlike traditional cellular systems where base stations operate independently, in Cell-free Massive MIMO, all distributed APs collaborate and coordinate their signals. This cooperation is achieved through advanced signal processing techniques. Each distributed AP is equipped with a large number of antennas, often hundreds or more. This massive antenna array allows for spatial multiplexing, enabling multiple users to be served concurrently over the same frequency resources. The distributed APs are typically connected to a centralized signal processing unit, which manages the coordination and signal processing tasks. The centralized processing enables efficient beamforming and interference management.

Cell-free Massive MIMO focuses on serving users directly, regardless of their location within the coverage area. This user-centric approach optimizes signal strength and quality for each user, leading to improved performance. Through cooperation and coordination, Cell-free Massive MIMO minimizes interference between users. This results in a cleaner and more efficient communication environment, which enhances network capacity and reliability.

The distributed nature of the system and the cooperation among APs ensure more uniform and extensive coverage, including in areas with challenging propagation conditions. Massive MIMO enables spatial multiplexing, allowing the system to support a large number of users with high data rates simultaneously. By minimizing interference and optimizing beamforming, Cell-free Massive MIMO offers robust and reliable communication, even in crowded and interference-

prone environments. The architecture is highly flexible and can adapt to changing user densities and traffic patterns. It can be easily scaled by deploying additional APs. By optimizing signal transmission and reducing interference, Cell-free Massive MIMO systems are more energy-efficient compared to traditional cellular networks.

Cell-free Massive MIMO represents a promising evolution in wireless communication systems, particularly suited for dense urban environments, industrial IoT applications, and scenarios with high user mobility. Its ability to provide consistent and high-quality connectivity, along with improved capacity, positions it as a key technology in the transition to 6G and beyond.

1.4 Millimeter-wave Communications

Millimeter wave (mmWave) communication is a wireless communication technology that operates in a high-frequency range, typically in the spectrum between 30 gigahertz (GHz) and 300 GHz. This frequency range is higher than traditional microwave frequencies, which are used in many wireless communication systems. Millimeter waves are characterized by short wavelengths, typically in the millimeter range, which is where the name ”millimeter wave” comes from.

The frequency range used in mmWave communication is significantly higher than that of conventional wireless communication, such as 4G and 5G, which typically operate in the sub-6 GHz and 3.5 GHz ranges. The high frequencies of mmWave enable the transmission of large amounts of data due to their broader bandwidth. Millimeter wave frequencies offer wide bandwidths that can support extremely high data rates. This makes mmWave technology suitable for applications that demand ultra-fast data transfer, such as high-definition video streaming,

augmented reality, virtual reality, and large-scale data downloads. One limitation of mmWave communication is its relatively short propagation range. Millimeter waves are more susceptible to atmospheric absorption and obstacles like buildings and vegetation, which can attenuate the signals. As a result, mmWave communication is often used for shorter-range, high-capacity applications, including point-to-point links and localized wireless networks. Millimeter waves are sensitive to obstacles, so they often require a clear line of sight between the transmitter and receiver. This characteristic makes mmWave technology suitable for applications like fixed wireless access, where there is a direct line of sight between a rooftop antenna and a user's premises. To overcome the challenges of obstacles and limited range, mmWave systems often use multi-beam technology, where multiple narrow beams are formed to track the user's device or move data between different points. Beamforming and beam steering are essential techniques to maintain the connection as the user moves. Millimeter wave technology is often used in small cell deployments, where a dense network of small, low-power base stations is placed throughout an area to provide localized high-capacity coverage. This approach is commonly used in urban environments and venues with high data demand. Millimeter wave communication is a key component of 5G and future wireless communication systems. It plays a significant role in delivering high data rates and low latency for next-generation applications. Beyond consumer applications, mmWave technology is also used for wireless backhaul in telecommunications networks, connecting cellular base stations to the core network infrastructure. This helps support the increasing data traffic demands of mobile networks.

Millimeter-wave communication has the potential to revolutionize wireless communication by enabling ultra-fast data transfer rates and low-latency connections. mmWave is a critical component of 5G and is expected to be even more essential in the development of future wireless

communication standards, including 6G.

1.5 Reconfigurable Intelligent Surfaces

Reconfigurable intelligent surface (RIS), also known as intelligent reflecting surface (IRS) is a novel technology in the field of wireless communications and signal processing. RIS is a two-dimensional surface comprising a large number of small passive reflecting elements, such as antennas or meta-material-based units, that can be electronically controlled to manipulate the electromagnetic waves in a way that optimizes wireless communication links. The key idea behind RIS is to control the phase and amplitude of the reflected signals from these elements to enhance the performance of wireless communication systems. RIS can focus signals in desired directions, which can increase the signal strength at the receiver and improve the signal-to-noise ratio. It can redirect signals around obstacles or towards specific receivers, overcoming obstacles like buildings or other structures that would otherwise block direct line-of-sight communication. RIS can be used to cancel out interference from other wireless sources, leading to cleaner and more reliable communication. By strategically placing RIS units, it's possible to extend the coverage of wireless networks and fill in coverage gaps. RIS can reduce the power consumption of wireless devices by improving the link quality, allowing for lower transmission power levels. RIS technology is being researched and developed for various applications, including 5G and beyond, the Internet of Things (IoT), and wireless communication in challenging environments. It has the potential to revolutionize wireless communication by providing more efficient and flexible ways to manage the propagation of electromagnetic waves in wireless networks.

1.6 Contributions and Organization of the Dissertation

This dissertation is organized into 9 chapters covering different interrelated design and optimization problems for next-generation hybrid networks.

In Chapter 2, we introduce the notion of trustworthiness as a measure of security in SFC embedding and thus network resource provisioning. Next, we incorporate this notion into the service deployment problem by defining the trustworthiness of the service network nodes, links, substrate network hosts, and the interconnecting paths to form the trust-aware service chain embedding problem. We introduce and formulate two variants of the problem, i.e. the *link-based* and the *path-based* trust-aware service chain embedding problems as mixed integer-linear programs (MILP), and then provide approximate models based on linear programming (LP) relaxation and rounding, and column generation with selecting k-shortest candidate substrate paths for hosting each virtual link, to reduce the complexity of the model. We validate the performance of our methods through simulations and conduct a discussion on evaluating the methods and some operation trade-offs.

In Chapter 3, we study two challenging optimization problems that arise while considering the deployment of the space-air-ground integrated networks (SAGINs), among which the optimal satellite gateway deployment problem is of significant importance. We propose a joint satellite gateway placement and routing strategy for the terrestrial network to minimize the overall cost of gateway deployment and traffic routing while adhering to the average delay requirement for traffic demands. Although traffic routing and gateway placement can be solved independently, the dependence between the routing decisions for different demands makes it more realistic to solve an aggregated model instead. Moreover, with the increasing interest in the software-

defined integration of 5G networks and satellites, the existence of an effective scheme for optimal placement of SDN controllers is essential. We discuss the interrelation between the two problems above and propose suitable methods to solve them under various network design criteria. We first provide a MILP model for solving the joint problem, and then motivate the decomposition of the model into two disjoint MILPs. We then show that the resulting problems can be modeled as the optimization of submodular set functions and solved efficiently with provable optimality gaps.

In Chapter 4, we propose trust as a metric to measure the trustworthiness of the FL agents and thereby enhance the security of the FL training. We first elaborate on trust as a security metric by presenting a mathematical framework for trust computation and aggregation within a multi-agent system. We then discuss how this framework can be incorporated within CFL and DFL setups introducing trust-aware FL algorithms. UAVs are envisioned as vital components of next-generation networks, driving diverse applications across various domains. UAV-enabled networks typically encompass diverse edge devices, including UAVs, ground stations, sensors, and other IoT devices, all generating vast amounts of data. Therefore, FL in UAV-enabled networks has attracted much attention over the past several years. However, most of the approaches in the state-of-the-art suffer from a single point of failure as well as massive communication overhead and excessive delay due to the existence of a central aggregator. The rest of the chapter proposes a framework for decentralized FL in UAV-enabled networks. Our approach consists of two parts; i) First, we propose a UAV placement scheme while ensuring the connectivity of the network of deployed UAVs. ii) Next, we use a consensus-based approach for decentralized FL among the UAV agents. We verify the effectiveness of our approach via extensive numerical simulation.

We allocate Chapter 5 to the design and optimization of cell-free (distributed) massive MIMO networks that are envisioned to realize cooperative multi-point transmission in next-

generation wireless systems. In this context, for efficient cooperative hybrid beamforming, the cluster of access points (APs) needs to obtain precise estimates of the uplink channel to perform reliable downlink precoding. However, due to the radio frequency (RF) impairments between the transceivers at the two en-points of the wireless channel, full channel reciprocity does not hold which results in performance degradation in the cooperative hybrid beamforming (CHBF) unless a suitable reciprocity calibration mechanism is in place. We propose a two-step approach to calibrate any two hybrid nodes in the distributed MIMO system. We then present and utilize the novel concept of *reciprocal tandem* to propose a low-complexity approach for jointly calibrating the cluster of APs and estimating the downlink channel. Finally, we validate our calibration technique's effectiveness through numerical simulation.

In Chapter 6, we propose a novel scheme for codebook design for hybrid beamforming in mmWave systems. We highlight the shortcomings of uniform linear arrays (ULAs) in generating perfect beams, i.e., beams with maximal uniform gain and sharp edges, and propose a solution based on a novel antenna configuration, namely, twin-ULA (TULA). Consequently, we propose two antenna configurations based on TULA: Delta and Star. We pose the problem of finding the beamforming coefficients as a continuous optimization problem for which we find the analytical closed-form solution by a quantization/aggregation method. Thanks to the derived closed-form solution the beamforming coefficients can be easily obtained with low complexity. We note that to efficiently search for the beam to serve a user and to jointly serve multiple users it is often required to use a composite beam which consists of multiple disjoint lobes. A composite beam covers multiple desired angular coverage intervals (ACIs) and ideally has maximum and uniform gain (smoothness) within each desired ACI, negligible gain (leakage) outside the desired ACIs, and sharp edges. We propose an algorithm for designing such an ideal composite codebook

by providing an analytical closed-form solution with low computational complexity. There is a fundamental trade-off between the gain, leakage, and smoothness of the beams. Our design allows us to achieve different values in such trade-offs based on changing the design parameters Through numerical analysis, we illustrate the effectiveness of the proposed antenna structure and beamforming algorithm to reach close-to-perfect beams.

In Chapter 7, we introduce an innovative beam alignment (BA) scheme tailored to mmWave communications. The mmWave technology relies on precisely aligning narrow beams with the channel Angle of Departures (AoDs) to maximize beamforming gain. While most existing BA schemes in the literature primarily consider single-path channels, real-world scenarios involve multiple resolvable paths with varying AoDs. As a result, traditional BA methods may not perform optimally in the presence of multipath or may not fully leverage this diversity to enhance robustness. Our proposed BA schemes address this challenge by transmitting probing packets via a set of scanning beams and gathering feedback for all scanning beams at the end of the probing phase from each user. We formulate the BA scheme as an optimization problem aimed at minimizing the expected average transmission beamwidth under different policies, where the policy maps received feedback to the set of transmission beams (TB). To expand the range of feasible feedback sequences, we demonstrate the unique structure of the scanning beams, referred to as the Tulip Design, and accordingly reformulate the minimization problem with a set of linear constraints and a reduced number of variables, efficiently solved using a greedy algorithm. We also explore the trade-off between the gain of SBs and TBs. Higher SB gain enhances SB penetration, while increased TB gain improves communication link performance. However, TB performance is contingent on the set of SBs. We show that by expanding the coverage of each SB, which in turn reduces its penetration, we create opportunities to find sharper TBs and amplify

beamforming gain. We quantify this trade-off through a trade-off curve, demonstrating that the Tulip Design attains the optimal curve for single dominant path channels. Additionally, we provide closed-form solutions for special cases of this trade-off curve and introduce an algorithm with performance evaluations to elucidate the need for further optimization of SB sets in state-of-the-art beam search algorithms.

In Chapter 8, we study the design of RIS for extending the coverage map of next-generation networks. A mmWave coverage map consists of blind spots due to shadowing and fading especially in dense urban environments. Beam-forming employing massive MIMO is primarily used to address high attenuation in the mmWave channel. Due to their ability to manipulate the impinging electromagnetic waves in an energy-efficient fashion, RISs are considered a great match to complement the massive MIMO systems in realizing the beam-forming task and hence effectively filling in the mmWave coverage gap.

We propose a novel RIS architecture, namely RIS-UPA where the RIS cells are arranged in a Uniform Planar Array (UPA). We show how RIS-UPA can be used in an RIS-aided MIMO system to fill the coverage gap in mmWave by forming beams of a custom footprint, with optimized main lobe gain, minimum leakage, and fairly sharp edges. Further, we propose a configuration for RIS-UPA that can simultaneously support multiple two-way communication pairs. We obtain theoretical low-complexity closed-form solutions for our design and validate our findings through extensive numerical experiments.

Low-complexity estimation and correction of carrier frequency offset (CFO) are essential in orthogonal frequency division multiplexing (OFDM). Therefore, in Chapter 9, we propose a low-overhead blind CFO estimation technique based on cyclic prefix (CP), in multi-input multi-output (MIMO)-OFDM systems. We propose to use antenna diversity for CFO estimation. Given

that the RF chains for all antenna elements at a communication node share the same clock, the carrier frequency offset (CFO) between two points may be estimated by using the combination of the received signal at all antennas. We improve our method by combining the antenna diversity with time diversity by considering the CP for multiple OFDM symbols. We provide a closed-form expression for CFO estimation and present algorithms that can considerably improve the CFO estimation performance at the expense of a linear increase in computational complexity. We validate the effectiveness of our estimation scheme via extensive numerical analysis.

Chapter 2: Trust-aware Service Chain Embedding

2.1 Overview

With the emergence of NFV and SDN, communication networks are becoming increasingly agile. Supported features such as programmability, IT virtualization, and open interfaces, enable operators and infrastructure providers to launch a network service rapidly and in a more cost and operation-efficient manner, allowing for shorter time-to-market cycles while driving down both operational and capital costs.

Essentially, NFV implements NFs through software virtualization techniques (e.g., running on containers or virtual machines) and runs them on commodity hardware [1]. Examples of such VNFs include firewalls, load balancers, IDS, caches, etc. and more recently mobile network functions [2]. These virtual appliances can be instantiated and scaled on demand, using cloud scaling technologies, to match the service demand requirements and utilize the underlying distributed computing, storage, and networking resources most efficiently. To support more complex services, a sequence of NFs may be stitched together, creating an SFC, and can be represented by a graph containing VNFs as nodes and the traffic demand between those VNFs, termed VNF-FG [3]. SDN on the other hand decouples the control plane from the underlying data plane, allowing programmability of the network control function using standard, open interfaces. Network intelligence is logically centralized, using software-based SDN Controllers that maintain a global

view of the network. SDN and NFV are complementary technologies, and each one can leverage the other to improve the flexibility and agility of networks and service delivery.

Using NFV and SDN, networks can be sliced into multiple dedicated, mutually isolated, end-to-end logical networks composed of several customizable software-defined functions, tailored to a given application or service [4] [5] [6]. A network slice is self-contained, ensuring resource isolation, including all necessary functions and capabilities, appropriately chained together, following the NFV service function chaining principle [7], to address the requirements of the corresponding services/applications [5] [8] [9]. VNFs and network elements are configured in each network slice to meet such requirements, enabling at the same time new business opportunities, by providing support for multi-service and/or multi-tenancy. That is primarily the reason why network slicing has evolved from a simple fixed network overlay concept to a fundamental feature of the emerging 5G systems [10] [7] [2].

While the vision is very compelling from an infrastructure, operation, and business perspective, the implementation of network slices poses multiple challenges, many of which are inherent to the enabling technologies, specific to the shared physical medium, or associated with the application context. To create a network slice [11] [2] and accommodate the needs of a particular service, we must be able to dynamically allocate, install, program, and configure all the service-specific network functions and chain these functions together, as prescribed by the network service graph.

Towards the deployment of SFCs, the SFC embedding problem becomes of paramount importance. The problem entails the placement of VNFs on the physical hosts and establishing the routing paths between them, according to the SFC template. Therefore considering the network slice as an SFC or a collection of SFCs, efficient service chain resource allocation

(service chain embedding or VNF-FG embedding) becomes of paramount importance. Service chain embedding is an NP-hard problem, as a generalization of the Virtual Network Embedding problem [12]. Appropriate approaches are needed for allocating network and computing resources to the requested SFC. Solving the service chain embedding problem requires the identification and adoption of appropriate resource allocation policies, resulting from the identified strategic objective(s) to optimize. Depending on the network service, such objective(s) may be related to QoS, profit maximization, fault tolerance, load balancing, energy saving, security, etc. Many of these can be expressed as hard constraints in the problem formulation e.g., end-to-end delay in [13]. In this chapter, we also consider constraints and objectives motivated by security recommendations laid out to protect systems supporting multi-tenancy [14] [15].

In particular, similar to [16], we capture the notion of security by integrating "trust weights" into the service chain embedding problem. These weights indicate the trustworthiness of a host system, based on its interactions with the other network entities, location (e.g., physically secured location), the level of hardening for the host, etc. Thus, by using such weights, we enhance the trustworthiness of the service chain deployment and its resilience to attacks. We assume that such trust weights are developed based on monitoring and configuration/deployment data and are disseminated via appropriate methods so that they are timely available to the entity (e.g., orchestrator, domain controller) that performs the embedding process.

Trust weights are also considered for the VNFs in the service chain, expressing their respective requirements in terms of security, e.g., VNFs performing mission-critical operations must be hosted on a trusted infrastructure. Hence, the proposed trust-aware service chain embedding approach ensures matching the requested security level of each VNF in the chain, with the security level offered by the servers that will host them.

Though the incorporation of trust seems intuitively fit for resource allocation problems spanning multiple administrative domains, due to the uncertainty in dealing with multiple InPs, we believe that it befits also the case of a single InP, due to the distributed heterogeneous nature of the NFVI. For example, edge NFVI PoPs are less reliable/secure than the NFVI in the central office or core DCs; in a single NFVI PoP, the security levels of the hosts may not be homogeneous (e.g., adoption of security zones, as collections of resources that share a common security exposure or security risk). The main contributions of the chapter are as follows.

- We incorporate trust into the service chain embedding problem as a metric for capturing security. We take into account the trustworthiness requirements for both the NFs and the edges between them. Similarly, we assign trust values to the substrate network paths as well as the substrate hosts to model the trustworthiness of the NFV infrastructure.

- We propose a link-based model for the problem when trust-based requirements only impose constraints on the NFs and the physical servers hosting them.

- We model the link-based problem as a MILP and present an approximation algorithm based on LP relaxation and rounding to obtain near-optimal solutions for the MILP in polynomial time.

- By augmenting the notion of trust to the virtual links and the substrate paths, we present a path-based formulation for the problem.

- We model the path-based problem as a MILP and propose a column generation-based method to obtain near-optimal solutions for the MILP in polynomial time.

2.2 Related Work

In this section, we briefly discuss related work on service chain embedding in general and then we focus on studies on security and trust in VNF-FG embedding.

The definition and approaches for the VNF-FG embedding problem vary, depending on several design decisions such as: (i) embedding objectives such as QoS [17], profit maximization [18], fault tolerance [19], load balancing [20], energy efficiency [21] etc.; (ii) solution strategy such as exact [22], heuristic [23] or meta-heuristic [24]; (iii) the technology domain where VNF-FG embedding is applied such as radio access networks [25], LTE/EPC [26], mobile core network [13], cloud networks [27]; (iv) type of resources such as CPU and BW [13], processing time and buffer capacity [28], ternary content-addressable memory [29]; (v) single administrative [30] domain or multi-domain approaches [31]. Authors in [12] provide a taxonomy of SFC embedding approaches.

Trust in NFV has been considered relatively recently, with studies such as [32] that discuss the challenges associated with incorporating trust into NFVI, mentioning also the need to apply policies for binding VNFs to certain (trusted) NFV platforms depending on the platforms' configurations. The requirement for placement of NFVs onto trusted platforms/hosts is more important given the emergence of 5G for supporting vertical markets (e.g., healthcare). The problem of security awareness in resource allocation has only been addressed in the context of VNE. Liu et al. [33] abstract the security requirements of virtual networks to quantifiable security levels and formulate the problem with security constraints related to the security demands of the virtual resources and the minimum security level required for a virtual node to be hosted on a substrate node. Similar to the above we consider trust values of the substrate nodes and trust demands for

the VNFs for the placement of VNFs on trusted servers. To the best of our knowledge, this is the first formulation for trust-aware service chain embedding.

2.3 Models and Definitions

2.3.1 Trust Model

Trust is a commonly used concept in a variety of application domains, such as e-commerce, peer-to-peer systems, cloud computing, and social networks. Trust management systems are implemented to define and gather trust-related evidence, and evaluate, establish, and manage trust relationships between entities in distributed systems. However, this chapter does not delve into the specifics of trust modeling or trust assessment. We assume that the infrastructure provider has efficient mechanisms for distributing trust evidence related to the underlying infrastructure. There are various ways to quantify trust; in different trust schemes, continuous or discrete numerical values have been assigned to measure the level of trustworthiness of a network entity. For example, in [16] the trustworthiness of nodes in multi-hop wireless networks is defined as a continuous value in $[0, 1]$. In [34] reputation-based trust for experimental infrastructures in a federated environment is defined in the same fashion. Following the same rationale, we denote that trust takes a continuous numerical value in $[0, 1]$. The trust estimate is updated periodically; we denote the update period as $T_{interval}$. During the update period, denoted by $[\tau - T_{interval}, \tau]$, the trust evaluation mechanism provides the trust estimate of node u denoted as t'_u. We use an exponential weighted moving average (EWMA), to characterize the node's time-varying trustworthiness over time, as it reacts faster to recent updates in trust values. Hence, the new derived

trust value for node u at time τ is given by

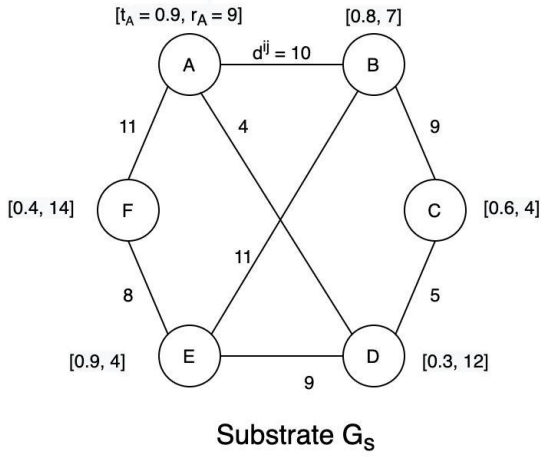
$$t_u(\tau) = (1-\alpha)t_u(\tau - T_{interval}) + \alpha\, t'_u \tag{2.1}$$

where $\alpha \in [0,1]$ is a constant weight indicating the preference between the updated and historic samples of trust values.

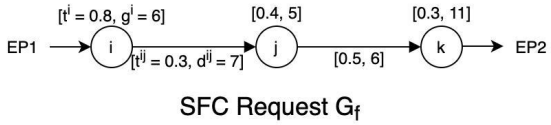
2.3.2 Network Model

The substrate network, is modeled as an undirected graph $G_s = (N_s, E_s)$, while the request network is modeled as a directed graph $G_f = (N_f, E_f)$. Each substrate node $u \in N_s$ has a residual processing capacity r_u, and each substrate link $(u,v) \in E_s$ has a BW capacity of c_{uv}, while the CPU requirement of request node $i \in N_f$ and the BW demand of a request link $(i,j) \in E_f$ are represented by g^i and d^{ij} accordingly. We denote by t_u the trustworthiness of the substrate node $u \in N_s$, and by t^i the trust requirement of the request node $i \in N_f$, while this metric for a request link $(i,j) \in E_f$ is denoted by t^{ij} and for a substrate path p by t_p, where p is a connected set of edges in the substrate graph. As in [14], trust takes fractional numerical value in $[0,1]$. We note that the trustworthiness of a substrate path can be any function (according to a specific use-case or methodology) of the trust values corresponding to the links and nodes belonging to that path. We define the following components of the path-based formulation to facilitate the description of the model:

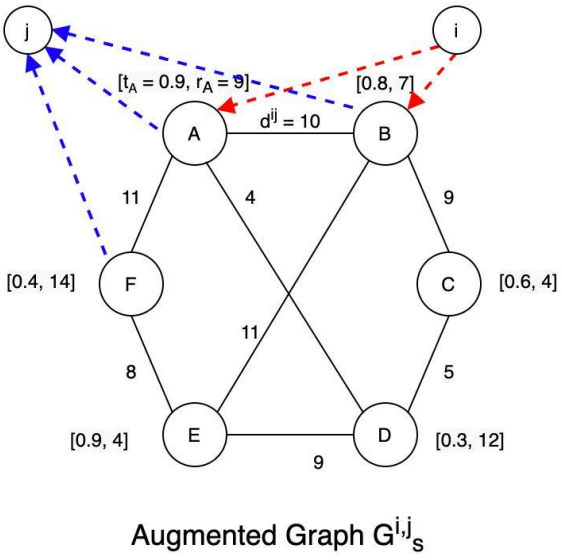
Definition 1 ***Augmented Graph.*** *For a commodity (virtual link) $k = (i,j)$ where $ij \in E_f$ we denote by $G_s^k = (N_s^k, E_s^k)$, the augmented graph corresponding to commodity k, whereby for every node $u \in N_s$ that is eligible for hosting request node i, the directed* ***augmented edge*** (i,u)



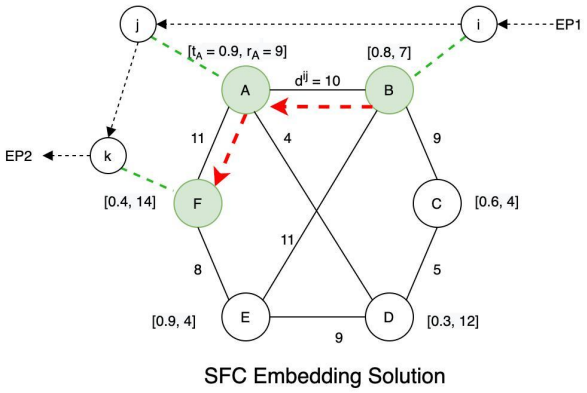
(a) Substrate Graph $G_s = (N_s, E_s)$



(b) Request Graph $G_f = (N_f, E_f)$



(c) $G_s^{ij} = (N_s^{ij}, E_s^{ij})$ Augmented Substrate Graph



(d) Trust-Aware SFC Embedding Solution

Figure 2.1: Trust-aware SFC Embedding Network Models

*is added to E_s. Similarly, for every node u that is eligible for hosting the request node j, the directed **augmented edge** (u, j) will be added to E_s. Hence, for the augmented graph, we will have:*

$$N_s^k = N_s \cup \{i, j\}$$

$$E_s^k = E_s \cup \{iu | u \in N_s \quad and \quad t_u \geq t^i \quad and \quad r_u \geq g^i\}$$

$$\cup \{uj | u \in N_s \quad and \quad t_u \geq t^j \quad and \quad r_u \geq g^j\}$$

Furthermore, we denote by $G_s^a = (N_s^a, E_s^a)$ the augmented graph corresponding to the request graph $G_f = (N_f, E_f)$, which contains all the nodes and links in all of the augmented graphs for all the commodities.

Definition 2 ***Augmented Path.*** *For a commodity $K = (i, j)$ where $(i, j) \in E_f$ we denote by p^k from i to j, a generic augmented path corresponding to commodity k, where the initial and the final links are augmented edges corresponding to commodity k. In other words, once we remove the initial and final edge from p^k the result will be a path of the original graph G_s. We further denote by $\mathcal{P}^k$, the set of all augmented paths corresponding to commodity k, and by $\mathcal{P}$ the set of all augmented paths.*

For instance, fig. 2.1c shows an augmented graph for commodity (virtual link) (i, j) in the Request Graph depicted in fig. 2.1b, that is going to be placed on the substrate network shown in Fig. 2.1a. Furthermore. each of the directed paths in the augmented graph depicted in Fig. 2.1c, that start with a **red** edge and end with a **blue** edge, is an augmented path corresponding to

commodity (i, j).

2.4 Problem Description

Agile service deployment of the requested network service at the operator's infrastructure involves placement of its constituent VNFs and in-sequence routing through them as prescribed by the Service Chain. Therefore, decisions must be made on where functions should be placed among the available NFVI and how traffic must be routed among them. NFVI covers all hardware and software resources that comprise the NFV environment; it includes network connectivity between locations, e.g., between data centers and public clouds or NFVI Points of Presence (NFVI-PoPs). Physical resources include computing, storage, and network hardware providing processing, storage, and connectivity for VNFs through the virtualization layer that sits just above the hardware and abstracts the physical resources. The optimization challenge at hand is to efficiently allocate the host and BW resources to the Service Graphs; in other words, to efficiently map the Service Graphs to the operator's NFV infrastructure adhering to the capacity constraints of the physical resources, as well as performance and resilience objectives and constraints. Performance and resilience objectives and constraints (e.g., delay budget between NFs, etc.) are dictated by the corresponding network service to be implemented, while additional objectives related to the operators' effective use of the infrastructure such as load balancing and energy conservation may be included in the problem formulation.

SFC embedding bears many similarities to the Virtual Network Embedding (VNE) problem since chained NFs can be seen as directional graphs to be embedded into a substrate network. However, the two problems call for different solutions since i) VNE requests are undirected

graphs as opposed to service chains that define an ordered set of VNFs and traffic must flow accordingly through this predefined ordered set. ii) VNFs exhibit ingress/egress bit-rate variations due to specific VNF operations (e.g., compression as coding, transcoding in video optimizers, etc.). iii) With VNE we primarily observe one-level mappings, where virtual network requests are mapped to the substrate network. However, in NFV environments we may have two-level mappings, i.e. service function chaining requests to VNF instances and then VNF instances to physical networks as vNFS may be shared between service chains that constitute the network service.

2.4.1 Trust-aware SFC Embedding

Considering trust-aware embedding of service chains, VNFs apart from their computing requirements can have also explicit requirements in terms of security, expressed via the respective trust values in the *Request*. Both can be expressed as constraints in the corresponding resource allocation problem. Thus, the trust-aware embedding solution in Fig. 2.2 ensures that: (i) the computing and BW resources, allocated to the service chain, do not exceed the residual capacities of servers and links, respectively and, (ii) each VNF in the chain is assigned to a server that matches (at minimum) its requested trust level. For example, considering VNF j, even though servers w and x can both support its computing requirements, only server w can provide (at minimum) the required trust level. In other words, the trustworthiness of the server should be equal to or higher than the one requested. Binding the SFC End Points (EPs) to substrate ingress and egress nodes has been omitted for readability purposes.

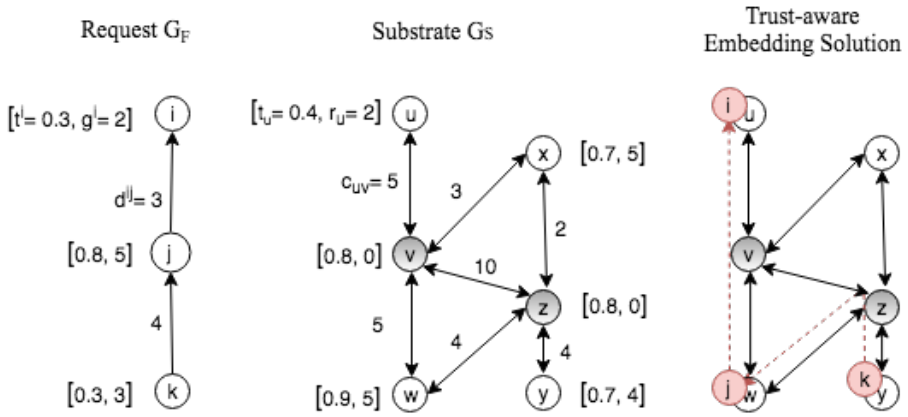


Figure 2.2: Example of trust-based SFC embedding.

2.4.2 Path-based Trust-aware SFC Embedding

We note that the method in [14] is *link-based*; i.e. i) The flow decision variables are link-to-link, and ii) Provided in the output is the assignment of each network request link to a set of substrate links that are guaranteed to generate valid continuous substrate paths by suitable flow formation and conservation constraints. However, in this chapter, we represent the SFC embedding problem by a *path-based* model; i.e. i) The flow decision variables are link-to-path; and ii) In the output, the request links are directly assigned to the pre-selected substrate paths.

One of the main contributions of this chapter is to propose a path-based model for the SFC embedding problem (PB-SCE) which provides multiple advantages over the traditional link-based formulation used in [14]. Firstly, a path-based formulation allows for the integration of various network and routing policies within the service chain embedding framework with low complexity. For instance, PB-SCE can be simply augmented by a path pre-selection phase to admit requirements such as traffic splitting, guaranteeing maximum delay or cost, or even assuring the existence of (disjoint) backup paths.

Moreover, within the path-based framework many of the design metrics that would enforce non-linear constraints to the link-based formulation (e.g. reliability, trust, availability, etc.), can simply be computed along the network paths in an offline fashion and be input to the path-based formulation. For instance, it is not possible in the link-based formulation in [14] to incorporate a linear constraint for capturing the trust requirement of each virtual edge; however, in the path-based model, it is straightforward to compute the trustworthiness of a network path following the corresponding trust aggregation policy and then input it to the model as a linear constraint. This is the main motivation for introducing the path-based trust-aware SFC embedding (PB-TASCE) model.

Finally, we note that in the context of trust-aware service chain embedding, a path-based model allows for abstracting out the method by which the trustworthiness of the infrastructure is computed and aggregated. Precisely, considering a path-based approach that only requires the trust values assigned to the paths, disregarding how this value is computed based on the same for the underlying components, allows for the application of our method in different settings, where the trust needs to be modeled differently [35]. For instance, interpreting trust as a multiplicative metric will lead to a different trustworthiness judgment for a path compared to the case where the trustworthiness of a path is computed as the minimum of the trustworthiness of all its edges.

2.5 Link-based Trust-aware Service Chain Embedding

2.5.1 MILP Formulation

To formulate the trust-aware service chain embedding problem, we consider:

- the set of binary variables $\mathbf{x}$, where x_u^i expresses the assignment of VNF i to the substrate

node u

- the set of binary variables **f**, where f_{uv}^{ij} expresses the amount of BW from substrate link (u, v) allocated to VNF-FG edge (i, j) of the SFC.

The problem is formulated as follows:

Objective:

$$\text{Min.} \sum_{i \in N_F} \sum_{u \in N_S} t_u g^i x_u^i + \phi \sum_{ij \in E_F} \sum_{uv \in E_S} f_{uv}^{ij} \quad s.t. \tag{2.2}$$

Placement Constraints:

$$\sum_{u \in N_S} x_u^i = 1, \ \forall i \in N_F \tag{2.3}$$

Trust Constraints:

$$(t_u - t^i) x_u^i \geq 0, \ \forall i \in N_F, \forall u \in N_S \tag{2.4}$$

Flow Constraints:

$$\sum_{uv \in E_S} (f_{uv}^{ij} - f_{vu}^{ij}) = d^{ij}(x_u^i - x_u^j), \ \forall ij \in E_F, u \in N_S \tag{2.5}$$

Capacity Constraints:

$$\sum_{i \in N_F} g^i x_u^i \leq r_u, \ \forall u \in N_S \tag{2.6}$$

$$\sum_{ij \in E_F} f_{uv}^{ij} \leq c_{uv}, \ \forall uv \in E_S \tag{2.7}$$

Domain Constraints:

$$f_{uv}^{ij} \geq 0, \ \forall ij \in E_F, uv \in E_S \tag{2.8}$$

$$x_u^i \in \{0, 1\}, \ \forall i \in N_F, u \in N_S \tag{2.9}$$

Constraint (2.21) ensures that each NF is exactly placed on one server. Constraint (2.4) ensures that each VNF is placed on a server that can support its requested trust level. Constraint (2.5) enforces flow conservation; i.e. the sum of all inbound and outbound traffic in switches and servers that do not host VNFs should be zero. Moreover, this condition ensures that for a given pair of assigned nodes i, j (VNFs), there is a path in the network graph where the VNF-FG edge (i, j) has been mapped. Constraints (2.24) and (2.25) guarantee that the allocated computing and BW resources do not exceed the residual capacities of servers and links, respectively. Constraints (2.27), (2.26) express the domain constraints for the variables **x** and **f**.

The objective function (2.2) attempts to minimize the corresponding embedding cost. The first term of this function represents the amount of CPU resources multiplied by the trust estimate of each assigned server. This term is minimized, if servers with lower security levels are preferred. Coupled with the set of constraints (2.4), it leads to solutions where the trust level of the servers

hosting the requested VNFs is close (but over) to the requested one. Assuming that increased trust levels for the substrate resources are used to instantiate the service chain will lead to higher provisioning costs; using the trust estimates at the objective function keeps this cost to the minimum feasible value. The second term of the objective function expresses the accumulated BW assigned to the VNF-FG edges. The term is minimized if all VNF-FG edges are mapped onto single-hop links. The parameter $\phi = \frac{1}{\sum_{ij \in E_F} d_{ij}} \left(\sum_{i \in N_F} g^i \right)$ is used for the normalization of CPU and BW units.

2.5.2 LP Relaxation and Rounding Algorithm

We derive the LP model from the original MILP model by relaxing the integrality of the binary **x** variables. Thus, the domain constraints in MILP are replaced by the following:

$$0 \leq x_u^i \leq 1 \; \forall i \in N_F, u \in N_S, \quad f_{uv}^{ij} \geq 0 \; \forall ij \in E_F, uc \in E_S$$

As the non-binary x_u^i values do not represent mappings between the VNFs and the servers, we use a deterministic rounding technique to obtain integer values for the variables **x** and embed the SFC requests. The pseudo-code for the LP rounding is shown in Algorithm 1. The LP solver (**Solve_LP(..)**) is called iteratively. At each iteration (Lines 2-12), one dimension x_u^i of the current LP solution is rounded. The selected dimension is the maximum fractional value among **x** that, if rounded to 1, still satisfies the capacity constraint of the corresponding substrate node. The rest of the corresponding x_v^i, $\forall v \in N_S \setminus \{u\}$ fractional solutions for the particular VNF will be zero, due to constraint (2.21). In the case that the capacity constraint can not be satisfied for any of the fractional solutions, the request will be rejected (*Sol=false*).

Given the mapping of the VNFs to the infrastructure (integral **x** values) we solve the multi-commodity flow allocation (MCF) problem, to determine the routing paths between them, as prescribed by the SFC (Lines 14-17). The algorithm will terminate if there is no solution found for the LP (*Sol=false*), or if there are no remaining dimensions of the solution to be rounded (*Sol=true*). The *Sol* flag determines the rejection/acceptance of the request (Lines 18-23).

Algorithm 1 LP Rounding

1: **repeat**
2: $\{x_u^i, f_{uv}^{ij}\} \leftarrow$ **Solve_LP(..)**
3: If solution exists *Sol:= true*, Otherwise *Sol:= false*
4: $X \leftarrow \{x_u^i | x_u^i \notin \{0,1\}\}$
5: **if** $X \neq 0$ **then**
6: $\{i_0, u_0\} \leftarrow argmax_{\{i \in N_f, u \in N_s | g_i \leq r_u\}} X$
7: **if** $\{i_0, u_0\}$ exists **then**
8: **Add LP Constraint** $x_{u_0}^{i_0} = 1$
9: **else**
10: *Sol=false*
11: **end if**
12: **end if**
13: **until** $(X = 0) \vee (Sol = false)$
14: **if** $Sol = true$ **then**
15: $\{x_u^i, f_{uv}^{ij}\} \leftarrow$ **Solve_MCF(..)**
16: If solution exists *Sol:= true*, Otherwise *Sol:= false*
17: **end if**
18: **if** *Sol:= true* **then**
19: **return** $\{x_u^i, f_{uv}^{ij}\}$ {Accept the request}
20: **else**
21: $\{\forall x_u^i := 0, \forall f_{uv}^{ij} := 0\}$
22: **return** $\{x_u^i, f_{uv}^{ij}\}$ {Deny the request}
23: **end if**

Contrary to the MILP formulation that is computationally very expensive to solve or even intractable for large problem instances, the relaxed linear program can be solved in polynomial time. The LP solver is invoked at most $|N_F| + 1$ times, as every iteration, up to $|N_F|$, leads (ideally) to the mapping of a VNF, while the MCF algorithm at the end provides us with the corresponding flow allocation.

2.6 Path-based Trust-aware Service Chain Embedding

In this section, we propose the path-based problem formulation for trust-aware service chain embedding. To this end, we define two sets of variables as follows,

- **x**, denotes the set of binary variables x_u^i which express the assignment of VNF i to substrate node u.
- **f**, denotes the set of continuous variables f_p which express the amount of flow passing through the augmented path $p \in \mathcal{P}$ in the augmented substrate graph.

2.6.1 Mixed Integer Linear Programming Formulation

We start with a MILP formulation as follows which contains all the service requirements as hard constraints.

PB-TASCE Objective:

$$\textbf{Minimize} \sum_{p \in \mathcal{P}} c_p f_p + \gamma \sum_{i \in N_f} \sum_{u \in N_s} t_u x_u^i \tag{2.10}$$

Placement Constraints:

$$\sum_{u \in N_S} x_u^i = 1, \quad \forall i \in N_F \tag{2.11}$$

$$\sum_{p \in \mathcal{P}^{ij}} f_p = d^{ij}, \quad \forall ij \in E_f \tag{2.12}$$

$$\sum_{p \in \mathcal{P}: iu \in p} f_p \leq x_u^i M, \quad \forall i \in N_f, u \in N_s \tag{2.13}$$

Trust Constraints:

$$(t_u - t^i)x_u^i \geq 0, \quad \forall i \in N_F, \forall u \in N_S \tag{2.14}$$

$$(t_p - t^{ij})f_p \geq 0, \quad \forall k \in E_f, p \in \mathcal{P}^k \tag{2.15}$$

Capacity Constraints:

$$\sum_{i \in N_F} g^i x_u^i \leq r_u, \quad \forall u \in N_S \tag{2.16}$$

$$\sum_{p:uv \in p} f_p \leq c_{uv}, \quad \forall uv \in E_s \tag{2.17}$$

Domain Constraints:

$$x_u^i \in \{0, 1\}, \quad \forall i \in N_F, u \in N_S \tag{2.18}$$

$$f_p \geq 0, \quad \forall p \in \mathcal{P} \tag{2.19}$$

The objective function (9.5) is the weighted sum of the flow embedding (BW) and server assignment (processing) costs with γ being the normalization factor to determine the balance between the two terms of the objective function. The processing cost corresponding to each substrate server is proportional to its trust value, i.e. the more trustworthy servers are more expensive. The constraint set (2.11) ensures that each VNF-FG node is placed on exactly one substrate node. Constraints in set (2.12) make sure that the traffic demand of each VNF-FG link will be allocated to this commodity using as many augmented paths as needed, while constraints (2.13) enforce that no flow passes through the paths inclusion of which in the SFC embedding

solution is disallowed by the node assignment policy, where M is a large enough constant. The constraint sets (2.14), and (2.15) guarantee that the trust requirements of each virtual link and each virtual node are satisfied, while the sets (2.16), and (2.17) guarantee that the allocated CPU and BW resources by each substrate node and link do not exceed their residual capacity, respectively. The constraints in sets (2.18), and (2.19) are the domain constraints corresponding to variable sets **x** and **f** respectively. We note that removing constraints (2.14), and (2.15) from the last model gives the baseline PB-SCE model.

We note that the PB-TASCE model cannot be used efficiently in realistic settings with large-scale networks due to not being scalable. First, as a generalization of the VNE problem, the SFC embedding problem is NP-hard. Second, the space complexity of the model is mostly driven by the size of the path set $\mathcal{P}$. Indeed, for a complete substrate graph, the set $\mathcal{P}$ may contain as many as $(e|E_f|/2)(|N_s|!)$ paths [36]. Even, for a sparse network graph, the size of the set of augmented paths for each virtual edge may grow exponentially in $|N_s|$. Therefore, due to constraint (2.15), the size of the constraints set grows exponentially with the scale of the network. Similarly, the number of path variables f_p grows exponentially with the scale of the network. Therefore, in the rest of this section, we propose a column-generation-based solution to tackle this issue by systematically including only the set of paths that may be chosen in the final solution to carry non-zero traffic, i.e. $f_p > 0$. We modify the PB-SCE and the PB-TASCE models to contain only the k-shortest augmented paths for each commodity. This will result in lower complexity at the expense of suboptimal results. Opting for different values of k one can adjust the performance of the algorithm and seek for a suitable value of k to seek a balance between complexity and result accuracy. We will explore this trade-off in detail in the evaluation section. We refer to these new models as KPB-SCE and KPB-TASCE in order.

2.6.2 Column Generation Method

Column generation (CG) is an effective technique for solving large linear programs (LPs), especially when the number of decision variables is substantially higher than the number of constraints. This is because in the final optimal LP solution only as many variables as the number of constraints may appear with non-zero values. Therefore, it is possible to start by solving the considered LP with only a subset of the decision variables and then gradually *generate* variables (i.e. *columns*) that have the potential to improve the value of the objective function.

The procedure consists of iterative steps including solving a restricted master problem (RMP) followed by a pricing sub-problem (PS). Initially, the RMP is formulated as the original problem with a limited set of decision variables such that an initial feasible solution to the LP can be obtained efficiently by solving the initial RMP. Once the primal RMP is solved the dual variables corresponding to the RMP constraints will be computed and utilized in formulating the PS. The objective function of the PS for each decision variable represents the *reduced cost* of that decision variable. The reduced cost of a decision variable in linear programming is the amount by which the objective function coefficient of that variable would have to improve (increase for a maximization problem, decrease for a minimization problem) before that variable would become non-zero in the optimal solution. In other words, it is the amount by which the objective function coefficient of the variable would have to increase (or decrease) to make it beneficial to include the variable in the optimal solution. Therefore, in the case of a minimization problem, the PS amounts to minimizing the PS objective function to find the variable with the most negative reduced cost and then adding that to the RMP variable set. Next, the RMP and the PS are solved again and this procedure continues until no variable can be found with a negative reduced cost.

2.6.3 Column-Generation-based Solution

Following the discussions in the last subsection, the RMP is formulated as,

$$\textbf{Minimize} \sum_{p \in \bar{\mathcal{P}}_{trusted}} c_p f_p + \gamma \sum_{i \in N_F} \sum_{u \in N_S} t_u x_u^i \tag{2.20}$$

$$\sum_{u \in N_S} x_u^i = 1, \quad \forall i \in N_F \tag{2.21}$$

$$\sum_{p \in \bar{\mathcal{P}}_{trusted}^{ij}} f_p = d^{ij}, \quad \forall ij \in E_F \tag{2.22}$$

$$\sum_{p \in \bar{\mathcal{P}}_{trusted} : iu \in p} f_p \leq x_u^i M, \quad \forall i \in N_F, u \in N_F \tag{2.23}$$

$$\sum_{i \in N_F} g^i x_u^i \leq r_u, \quad \forall u \in N_S \tag{2.24}$$

$$\sum_{p \in \bar{\mathcal{P}}_{trusted} : uv \in p} f_p \leq c_{uv}, \quad \forall uv \in E_s \tag{2.25}$$

$$x_u^i \in [0, 1], \quad \forall i \in N_F, u \in N_S \tag{2.26}$$

$$f_p \geq 0, \quad \forall p \in \bar{\mathcal{P}}_{trusted} \tag{2.27}$$

where $\bar{\mathcal{P}}_{trusted}^{ij} \subseteq \mathcal{P}_{trusted}^{ij}$ i.e. the set of those augmented paths in the augmented graph G^{ij} that satisfy the trust constraints and it holds that $\cup_{ij \in E_F} \bar{\mathcal{P}}_{trusted}^{ij} = \bar{\mathcal{P}}_{trusted}$. Therefore, con-

straints (2.14) and (2.15) are removed from the RMP.

2.6.3.1 Initial Solution

The CG procedure starts from an initial RMP with an initial set of variables -preferably the smallest possible one- that yield a feasible solution to the RMP. However, in general, it is not trivial to find such an instance of the RMP. In the path-based service chain embedding problem, for a general service request $G_F = (N_F, E_F)$ and a substrate $G_S = (N_S, E_S)$ it is not possible to deterministically establish a set $\bar{\mathcal{P}}_{trusted,init}$ such that the corresponding RMP is feasible. To tackle this issue, we extend the $G_S = (N_S, E_S)$ graph by adding a copy of the service request graph to form a new substrate graph $G'_S = (N'_S, E'_S)$ and then constructing the new augmented graph $G'^a_S = (N'^a_S, E'^a_S)$. Furthermore, we assign large weights to the augmented edges in G'^a_S. An example of this graph is shown in Fig. 2.3. We propose a *dummy* initial solution for path-based service chain embedding by placing the request graph on its copy in the $G'_S = (N'_S, E'_S)$ graph. The large weights assigned to the dummy augmented edges ensure that if the relaxed service chain embedding problem is feasible, none of the dummy placements will end up in the final solution once the CG procedure terminates.

2.6.3.2 Addition of Columns

To formulate the pricing sub-problem to generate new columns (paths) that can improve the objective value of the RMP, we need to utilize the dual problem corresponding to the RMP

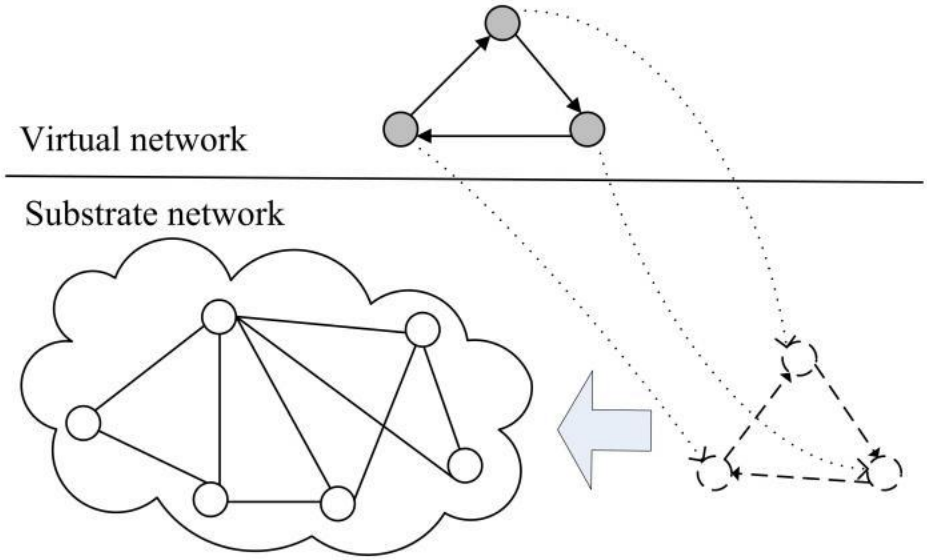


Figure 2.3: Initial Dummy Solution

denoted by DRMP. The DRMP is given as follows,

$$\text{Maximize } \sum_{i} \alpha_i + \sum_{ij} d_{ij}\sigma_{ij} - \sum_{u} r_u\beta_u - \sum_{uv} c_{uv}\gamma_{uv} \tag{2.28}$$

$$\sigma_{ij} - \sum_{u\in N_s:(i,u)\in p} \tau_{iu} - \sum_{or(u,i)\in p} \tau_{iu} - \sum_{uv\in p} \gamma_{uv} \leq c_p \quad \forall p \in \tilde{\mathcal{P}}^{ij}, \forall ij \in E_f \tag{2.29}$$

$$\alpha_i + M\tau_{iu} - g^i\beta_u \leq 0 \quad \forall i \in N_f, u \in N_s \tag{2.30}$$

$$\tau_{iu}, \beta_u, \gamma_{uv} \geq 0, \quad \alpha_i, \sigma_{ij} \text{ unrestricted} \tag{2.31}$$

where the variable sets α, β, γ,σ, and τ correspond to constraint sets in the last MILP in respective order. The following lemma is useful in obtaining the pricing sub-problem.

Lemma 3 *The optimal solution to the relaxed path-based service chain embedding problem is*

obtained at an RMP instance where there is no $p \in \mathcal{P}$, which violates constraint (2.29) from the DRMP problem.

Proof. Let RMP^* denote the RMP instance at which the optimal solution to the relaxed path-based service chain embedding problem is obtained. Let $\mathcal{P}^*$ be the set of generated paths at this instance and let RMP^* denote the corresponding dual problem. Further, let $\overline{RMP}$ and $\overline{DRMP}$ be the primal and dual RMP instances corresponding to the set of generated paths $\overline{\mathcal{P}}$. Let $OPT(.)$ denote the optimal solution to each problem instance. It holds for any $\overline{\mathcal{P}} \subseteq \mathcal{P}^*$ that,

$$OPT(\overline{RMP}) \geq OPT(RMP^*) = OPT(DRMP^*) \leq OPT(\overline{DRMP}) \tag{2.32}$$

where the middle equality follows from the strong duality theorem, the LHS inequality follows from the fact that the optimal solution to the minimization problem $\overline{RMP}$ is a feasible solution of RMP^*, and the RHS inequality holds since the instance $\overline{DRMP}$ contains fewer constraints than the $DRMP^*$ model. Now, consider the specific instance of $\overline{RMP}$, where none of the (2.29) constraints are violated when evaluated for all $p \in \mathcal{P}$. Then the corresponding $(\alpha, \beta, \gamma, \sigma)_{\overline{\mathcal{P}}}$ and its associated $(x, f)_{\overline{\mathcal{P}}}$ vectors will be feasible solutions to the $DRMP^*$ and RMP^* problems respectively, because none of the constraints of the $DRMP^*$ model are violated. Then, it is straightforward to write:

$$OPT(\overline{RMP}) = OPT(\overline{DRMP}) \leq OPT(DRMP^*) = OPT(RMP^*) \tag{2.33}$$

Combining (2.32), and (2.33), the desired result follows. Conversely, it holds that in the optimal solution to a linear program, all the basic and non-basic variables have greater than or

equal to zero reduced costs. It is straightforward to show that equation (2.29) represents the reduced cost of each f_p variable for each $p \in \mathcal{P}$. Therefore, in the optimal solution to the relaxed service chain embedding problem the constraints (2.29) are satisfied for all $p \in \mathcal{P}$.

2.7 Link-based Model Performance Evaluation

In this section, we evaluate the efficiency of the proposed trust-aware SFC embedding method; we benchmark the LP algorithm against the MILP one. Specifically, we provide an overview of the performance evaluation setup (3.2.3.2), and discuss the evaluation results (3.2.3.3).

2.7.1 Performance Evaluation Setup

For the simulations, we use an event-based simulator implemented in Java ([26, 29]), including an SFC and DC topology generator. We use CPLEX for our MILP models based on the branch-and-cut method. We used the dual simplex method for the LP problem. Our tests are carried out on a server with an Intel i5 CPU at 2.3 GHz and 8 GB of main memory.

NFV Infrastructure. We have generated a 3-layer fat-tree network topology for the DC, consisting of 16 pods. In our evaluations, we use only one portion (or zone) of the DC consisting of 4 (out of 16) pods, utilizing 2 switches per layer and the corresponding set of 2 servers per ToR switch. For each server, we consider 8 cores running at 2 GHz. Each server's initial utilization is uniformly distributed $U(0.3, 0.6)$. The ToR-to-Server link capacity is 8 Gbps while the inter-rack link capacity is 16 Gbps. The initial trustworthiness of the substrate nodes is drawn from a uniform distribution $U(0.2, 1)$.

Service Chains. We generate VNF-FGs based on three service chain templates: (i) a chain

handling traffic that needs to pass through a particular sequence of VNFs, *i.e.,* a NAT and an FW followed by an IDS; (ii) the second template reflects the case where traffic in a service chain is split by a particular VNF, according to some predefined policy, *e.g.,* load balancing; (iii) the last template corresponds to a bifurcated path with a single endpoint reflecting cases where one part of the traffic needs to be encrypted while another part needs to pass through a firewall [37]. For each chain, the number of requested VNFs, inbound traffic demands, and requested trust level are uniformly distributed, $U(5, 9)$, $U(50, 100)$ Mbps, and $U(0.2, 0.8)$ respectively. The CPU demand of each VNF is derived from the inbound traffic rate and the respective VNF profile (cycles per packet). Resource profiles are available for a wide range of VNFs [38, 39], while profiling techniques can be employed for processing workloads whose computing requirements are not known in advance [39].

Evaluation Scenarios. We conducted two sets of simulations for two distinct cases. In the first case, the trustworthiness of each substrate node does not change over time (*Experiment A: Static Trustworthiness*), while in the second one, the trust estimate is updated at intervals of 500-time units, according to the description in Section 9.3 (*Experiment B: Dynamic Trustworthiness*). The purpose of the 1^{st} experiment is to benchmark the approximation algorithm against the optimal embedding approach, while the 2^{nd} one aims to showcase the impact of the trustworthiness updates in the embedding process. The new trust estimates (t'_u) are drawn from a uniform distribution $U(0.4, 1)$. The α parameter in formula (1) is tuned to 0.25 [40]. For both cases, requests arrive according to a Poisson process, with an average rate of 4 requests per 100 time units, each having an exponentially distributed lifetime with an average of 1,000 time units. Every simulation is executed for 6,000 requests and repeated for 10 iterations [29].

Evaluation Metrics. We use the following metrics for the evaluation of the two service chain

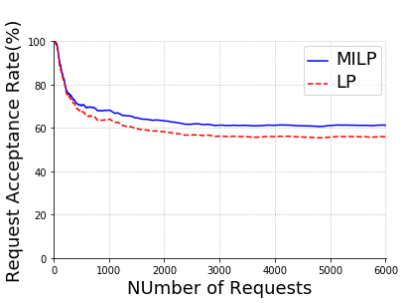
mapping methods:

- **Request Acceptance Rate** is the rate of successfully embedded requests divided by the total number of requests.
- **Resource Utilization** is the amount of CPU or BW units allocated for all the embedded requests.
- **Resource Revenue per Request** is the average amount of CPU or BW units specified in the requests that have been embedded successfully.
- **Cumulative Resource Revenue** is the cumulative amount of CPU or BW units in requests that have been embedded successfully, divided by the total number of embedding requests (successful or not successful).
- **Cumulative Resource Cost** is the cumulative amount of CPU or BW units allocated to requests that have been embedded successfully, divided by the total number of embedding requests (successful or not successful).

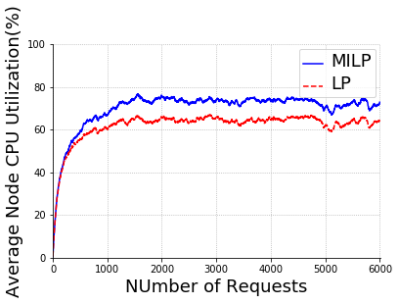
2.7.2 Evaluation Results

2.7.2.1 Experiment A: Static Trustworthiness

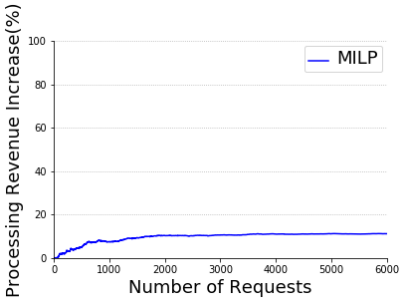
Fig. 3.2a illustrates the rate of accepted requests for the LP and the MILP approaches. The acceptance rate of the LP, as expected, lags in a steady state by approximately 5%. The trend in the acceptance rate is in line with the average CPU utilization statistics across DCs, depicted in Fig. 3.2b. CPU utilization is below 80% in steady state for both approaches. The higher



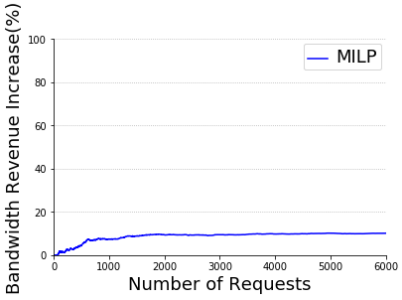
(a) Acceptance Rate



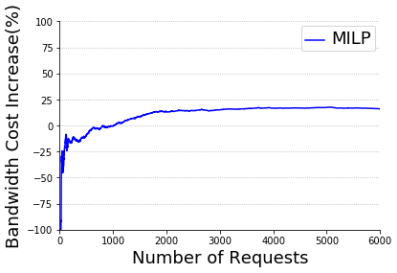
(b) CPU Utilization



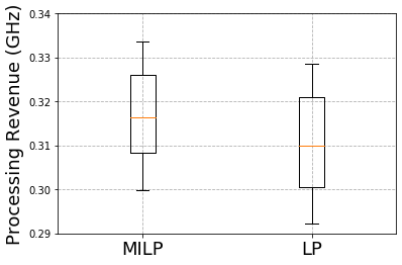
(c) Inc. Cum. CPU Revenue to LP



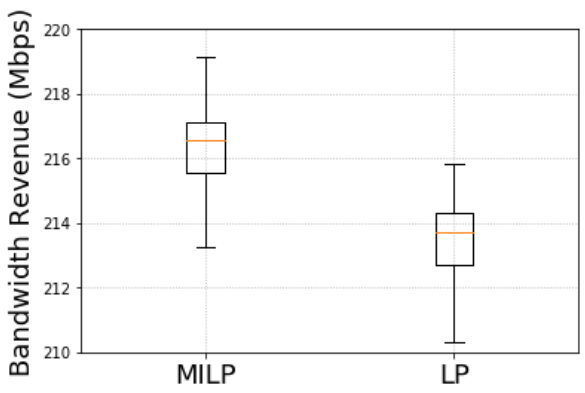
(d) Inc. Cum. BW Revenue to LP



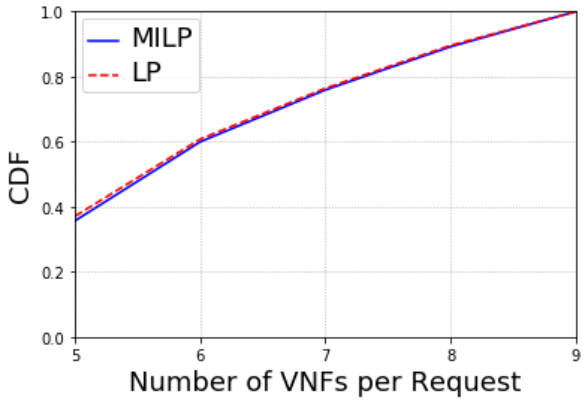
(e) Inc. Cum. BW Cost



(f) CPU Revenue per Request

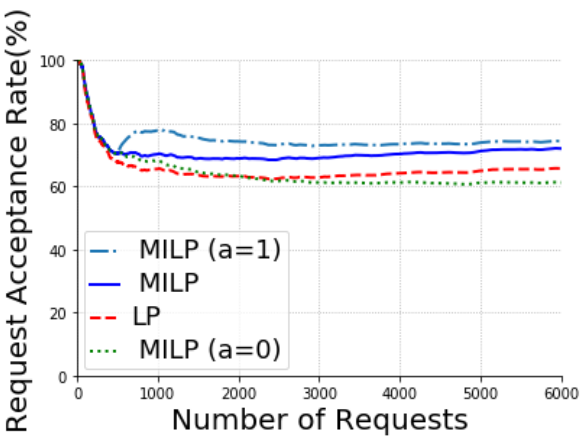


(g) BW Revenue per Request

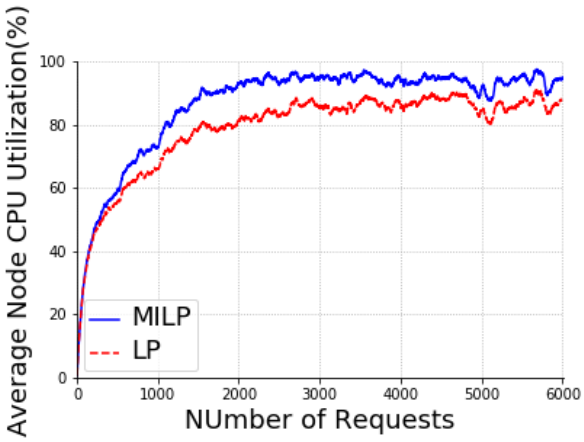


(h) CDF of Accepted Requests

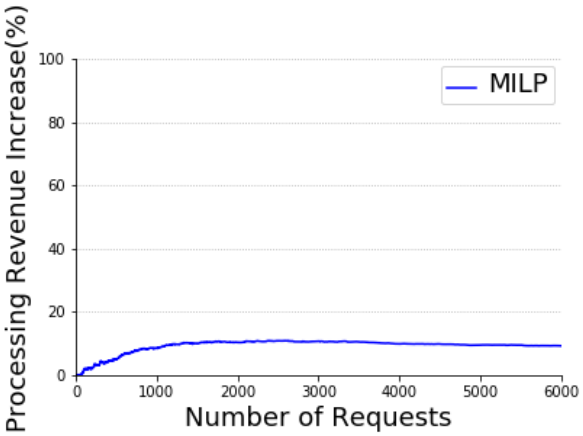
Figure 2.4: Link-based Model Experiment A Numerical Results



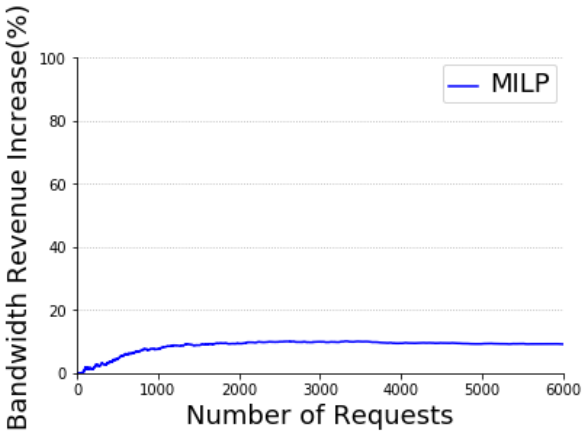
(a) Acceptance Rate



(b) CPU Utilization



(c) Incremental Cumulative CPU Revenue to LP



(d) Incremental Cumulative BW Revenue to LP

Figure 2.5: Link-based Model Experiment B Numerical Results

utilization levels (9%) achieved by the MILP stem from the higher request acceptance rate. BW utilization is on average less than 50%, hence not included here due to lack of space. Service request denials are mainly caused due to the inability of the substrate to match the required trust values of the VNFs. In particular, over time, trustworthy servers become saturated while less utilized servers cannot support the required trust levels of the SFCs, leading to service request denials.

Fig. 3.2c and Fig. 2.4d present the percentage increase of the cumulative CPU and BW revenue for the MILP, as opposed to the approximation algorithm. In particular, the MILP generates in a steady state approximately 10% more CPU and BW revenue, compared to the LP. The result matches the above-mentioned gap in the request acceptance rate between the two approaches.

Fig. 2.4e presents the percentage difference for the cumulative BW cost between the MILP and the LP. The BW cost for the MILP in a steady state is approximately 16% higher than the LP. The percentage is not analogous to the percentage difference in accepted requests, as resources get fragmented over time, hence longer paths are used for SFC embedding. Since the MILP accepts more requests, the fragmentation is more severe. Consequently, higher per-request cost is associated with the MILP. Moreover, in a transient state, the noted difference is due to the sub-optimal solutions of the LP. Collocation of the requested VNFs renders the BW cost to zero; the approximation algorithm starts to employ multiple hosts for the embedding (hence not all VNFs of the SFC are collocated) at request #10 as opposed to the MILP that does so after request #25.

Fig. 2.4f and Fig. 2.4g present the minimum, first quartile, median, third quartile, and maximum of the CPU and BW revenue per request, for the two approaches. Fig. 2.4h depicts the CDF of the service chains' size (number of VNFs) successfully embedded by the two programs.

We notice that the LP follows the same trend as the MILP, in terms of the the sizes of the requests it accepts. Moreover, the difference in revenue per request is less than 2%, for CPU and BW. Both results prove the efficiency of the approximation algorithm.

2.7.2.2 Experiment B: Dynamic Trust Estimates

Fig. 2.5a illustrates the rate of accepted requests for the LP and MILP, for dynamic trust estimates. Two additional α values are used for the MILP to showcase two extreme cases; in the first case, the trustworthiness of a node relies on the historical data (α=0), while in the second case, it is equal to the trust estimate for the particular time interval (α =1). Fig. 2.5b denotes the average CPU utilization of the substrate servers for the LP and MILP. Regarding the different α values for the MILP, we noticed no difference before the first trust update. Thereupon, since on average the trustworthiness of the servers increases, the graphs corresponding to (α = 0.25) and (α = 1) diverge from the static-trust one (α = 0) leading to higher acceptance rates. The increase, as expected, is gradual for α = 0.25, while over time converges to the graph corresponding to (α = 1).

We observe a similar trend to *Experiment A*, concerning the acceptance ratio between the two programs. The difference in steady state is around 6%. Moreover, we notice an increase in the acceptance rate for both programs over time, due to the gradual increase in the trustworthiness of the servers, according to the experiment setup. This leads to an increase in the acceptance rate by 10% compared to the static scenario. This increase is also witnessed by the corresponding increase in utilization; increased trustworthiness of the infrastructure leads to a higher acceptance rate and saturation of computing resources, as trust constraints are more easily satisfied.

Figs. 2.5c and 2.5d, present the percentage increase of the cumulative CPU and BW revenue for the MILP, as opposed to the approximation algorithm. Results exhibit a similar trend as *Experiment A* and are per the above-mentioned gap between the acceptance rates of the two approaches.

2.8 Path-based Model Performance Evaluation

In this section we compare the performance of the proposed path-based models in general with the link-based model in [14], present the outcome of our service chain embedding scheme under both node and link trust constraints, and provide the performance evaluation results for the aforementioned approximation methods. We first describe the simulation environment setup and scenarios and then proceed with presenting the evaluation results.

2.8.1 Evaluation Scenarios

We carry out two distinct sets of experiments to evaluate the performance of the proposed schemes. In the first set of experiments we compare the performance of the path-based service chain embedding method to that of the link-based scheme in [14] from different perspectives and report the results. We run the KPB-SCE model for different values of k and benchmark them against the link-based method. None of the trust constraints are in place for this experiment.

The second set of experiments deals with service chain embedding under both node and link trust constraints. More precisely, this set of experiments compares the performance of the PB-TASCE model to that of the baseline PB-SCE, and PB-SCE with node constraints to highlight how the integration of trust constraints impacts the performance of the SFC embedding methods.

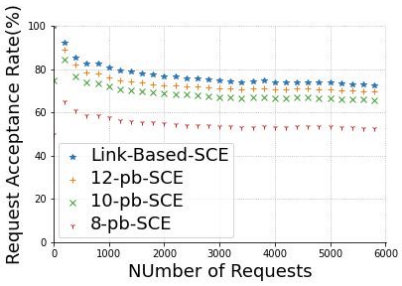
2.8.2 Evaluation Results

1) *Experiment A:* Fig. 3.2a shows the comparison between the performance of link-based SCE MILP model of [14], and the proposed k-pb-SCE algorithms, for different values of $k = 8, 10, 12$. As fig. 3.2a depicts, as k increases and more paths are included in the solution space, the performance of the k-pb-SCE algorithm increases. The change from $k = 8$ to $k = 10$ is more obvious than the change from $k = 10$ to $k = 12$. The 8-pb-SCE method on average admits around 55% of the requests while 10-pb-SCE, 12-pb-SCE, and the link-based SCE, accept 67%, 70% and 74% of the requests in order.

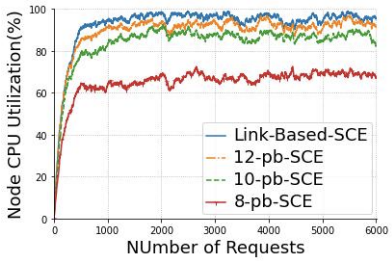
Fig. 3.2b compares the CPU utilization of the substrate servers. As expected, the higher the request acceptance ratio is the higher the CPU utilization will be, as more processing resources are consumed. In a steady state, in the case of the link-based SCE approach, on average more than 95% of the processing resources are consumed. The 12-pb-SCE can almost keep up to this level, while the CPU utilization for 8-pb-SCE remains as low as around 70%.

Fig. 3.2c and 2.6d, depict the percentage difference between the per-request processing and BW revenue generated by the path-based approximation methods and the link-based method. By fig. 3.2c, the processing revenue generated by the k-pb-SCE methods remains within 10% of the optimal link-based methods. Moreover, as fig. 2.6d suggests, in a steady state, the 8-pb-SCE method provides around 14% less BW revenue compared to that of the optimal link-based method. This value can be mitigated to 9% and 4% by taking 10, and 12 shortest paths for each commodity in the solution space.

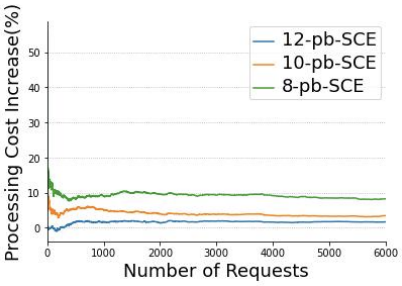
Fig. 2.6e and 2.6f show the per-request profile of the BW cost and the BW revenue. Firstly, we observe a significant difference between the BW cost and revenue for admitted requests which



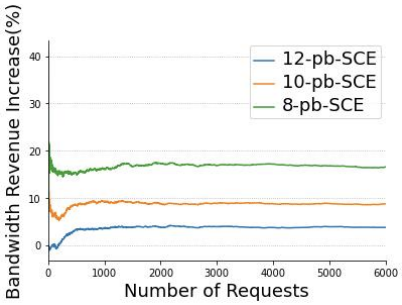
(a) Acceptance Ratio



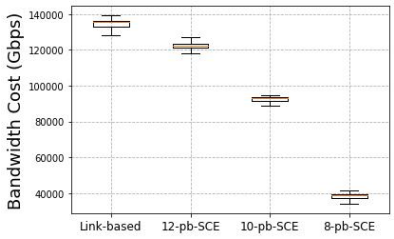
(b) CPU Utilization



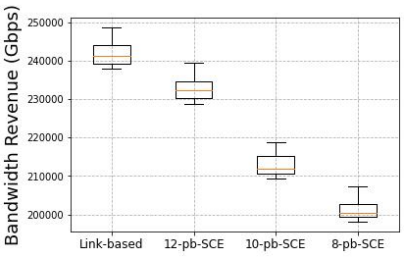
(c) CPU Revenue to Link-Based



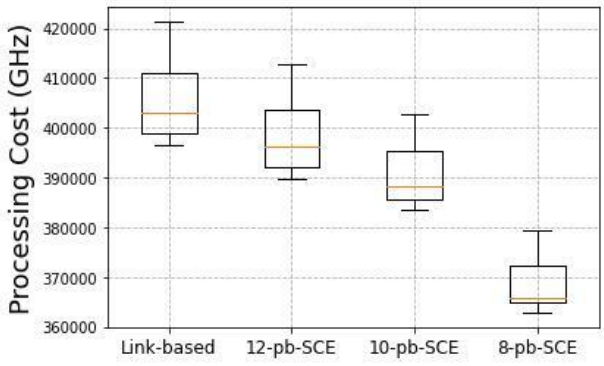
(d) BW Revenue to Link-Based



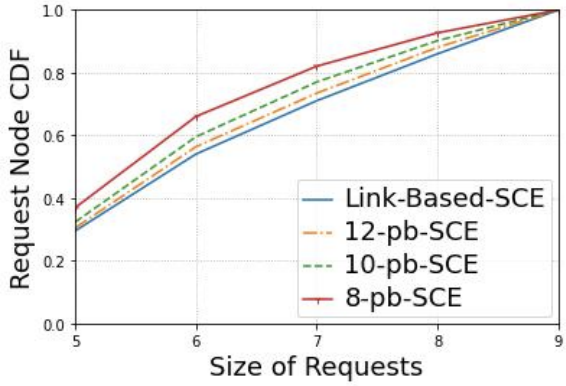
(e) BW Cost per Request



(f) BW Revenue per Request



(g) Processing Cost per Request



(h) CDF of Accepted Requests

Figure 2.6: Path-based Model Experiment A Numerical Results

stems from the fact that different functions can be collocated on the substrate servers which will induce zero BW consumption and therefore zero BW cost. Moreover, we observe that the more optimal the algorithm is, the more it is successful in admitting more costly network requests. This is because when more substrate paths are injected into the solution space as the value of parameter k increases, more efficient options are there for placing each request link, in the request embedding decision-making process.

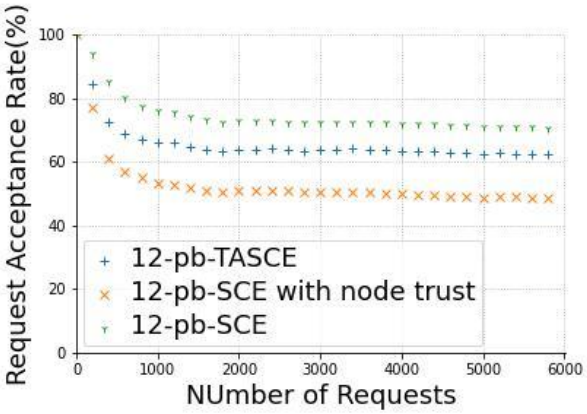
The box plot for the per-request processing cost is depicted in Fig. 2.6g. It can be seen that the per-request processing cost for the approximation methods remains within 10% of that of the link-based method confirming the observation of Fig. 3.2c.

Fig. 2.6h shows the CDF of the number of VNFs in each SFC that is admitted by the SFC embedding mechanisms; i.e. the population of SFCs from certain sizes that are successfully placed on the substrate network. As expected, switching from $k = 8$ to $k = 10$ and then $k = 12$, the profile of the admitted service chain sizes converges to that of the link-based approach, which further confirms the effectiveness of our approximate embedding methods.

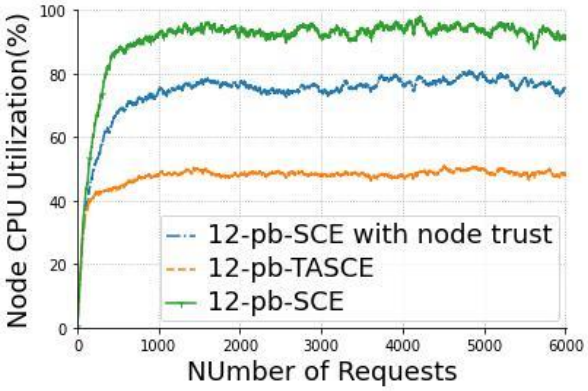
2) *Experiment B:*

Fig. 2.7a elaborates on the impact of incorporating trust into the path-based SCE model. As this figure suggests, for $k = 12$, the addition of trust requirements for the request nodes (i.e. constraint (2.4)) may reduce the performance of the SFC embedding method by 10% on average in the steady state. Furthermore, when the link trust requirements are integrated within the SFC embedding framework(i.e. the 12-pb-TASCE model) the acceptance ratio diminishes by another 10%.

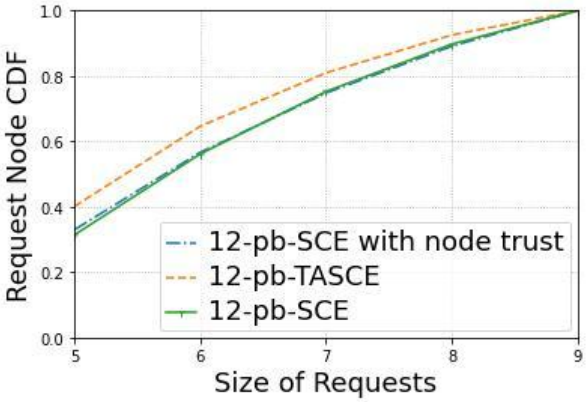
The impact of the natural decline in the acceptance ratio, when taking into account the trust constraints can be observed in the server CPU utilization profile in Fig. 2.7b as well. One can



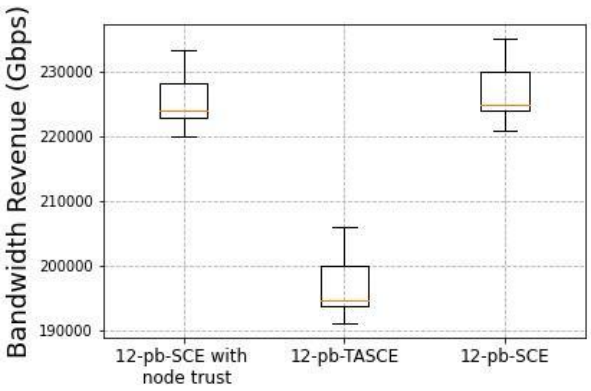
(a) Acceptance Ratio



(b) CPU Utilization



(c) CDF of Accepted Requests



(d) BW Revenue per Request

Figure 2.7: Path-based Model Experiment B Numerical Results

observe that there is an 18% and a further 25% decline in the substrate CPU utilization, associated with the addition of trust requirements for request nodes and links in the respective order. We note that the performance drop caused by the node trust constraints is quite natural in that the substrate nodes with lower trustworthiness host request nodes less frequently, but the severe drop in CPU utilization due to the inexistence of trustworthy substrate paths is quite more interesting; the reason being that the probability of rejecting a larger request (with more nodes and links) is higher since due to the link trust constraints, it gets more unlikely to find feasible substrate paths for each request link when the request size increases.

To further investigate the impact of the size of requests in the embedding decision, we tested the performance of the 12-pb-TASCE algorithm when only requests with 5 VNFs arrive. We then repeated the same experiment for the requests of only 9 VNFs. In the former case, we observed an increase of around 10% in the CPU utilization, while in the latter this parameter dropped by around 12%.

This observation is well-aligned with Fig. 2.7c as well which suggests that the 12-pb-TASCE method tends to admit the smaller requests compared to the case when there are no restrictions on the trustworthiness of the substrate paths.

Finally, fig. 2.7d shows the impact of the restrictions on the trustworthiness of substrate paths, compared to the two other cases. We observe that the per-request BW revenue remains almost the same when there are only restrictions on node trustworthiness since the set of feasible paths in the solution space does not change while when the path trust constraints are introduced the BW revenue diminishes by around 15 to 20 percents.

2.9 Conclusions

In this chapter, we introduced the trust-aware service chain embedding problem, for delivering secure network service deployment on a trustworthy infrastructure. We introduced frameworks for both the link-based and path-based trust-aware service chain embedding instances. For the link-based model, we introduce a MILP formulation, enforcing trustworthiness through appropriate trust-related constraints and trust weights in the objective function of the optimization problem. We then relaxed the integer constraints and used a deterministic rounding approach to obtain a polynomial-time solution algorithm. For the path-based model, we started with a baseline formulation. Then we provided a formulation for the approximate problem by taking into account only k-shortest paths candidates for each virtual link in a column generation framework. We finally incorporated the trust constraints for both virtual nodes and links and evaluated the efficiency of our algorithm through simulations and numerical results.

Chapter 3: Ground Segment Optimization in Space-Ground Hybrid Networks

3.1 Overview

5G mobile communication systems are required to achieve KPIs in terms of low latency, massive connectivity, consistent QoS, and high security. For instance, user bit rates up to 10 Gbps and RTTs as small as 1-10 ms are demanded in specific application scenarios in 5G. Moreover, due to the significantly increasing traffic volume, the number of registered users, and novel provisioned use-cases such as cloud computing, IoT, massive-data applications, etc., it has become evident over the past few years that to achieve the 5G key promises, it is essential to take advantage of the full capacity of all communications types & segments (e.g. terrestrial, aerial, and space) as well as supporting technologies (e.g. SDN, NFV, etc.) simultaneously, otherwise the traditional stand-alone terrestrial networks will fail to achieve the key projected promises. Based on the above discussion, towards the inevitable convergence of the above paradigms and technologies, an optimized integration and configuration policy, that is tailored to specific use cases and corresponding QoS requirements, becomes of paramount importance. Moreover, given that only less than 60% of the world has access to stable communications, the importance of adopting and integrating satellite communications has been recognized and supported by standardization bodies such as 3GPP, ITU [41], and ETSI [42].

The SAGIN [43], depicted in 3.1 offers potential benefits which are not possible other-

wise, including global coverage, low latency, and high reliability. In particular, satellites can replace, extend, or complement the terrestrial networks, in rural and hard-to-reach areas, or where the existence of communications infrastructure is costly or even infeasible in use cases such as mountains and marine communications. Furthermore, satellites can offload the terrestrial networks by accommodating the delay-insensitive applications, allowing the terrestrial segment to survive when there is a surge in the traffic load. Finally, due to their global coverage, satellites can provide a reliable and seamless back-haul for the aerial segment [44], and also the monitoring and control applications for IoT, vehicular networks, etc. On the other hand, despite providing the advantages of the three different segments, new challenges are also introduced in the integrated network due to the limitations of each of the layers, including but not limited to, complicated end-to-end resource provisioning due to the additional resource constraints, high control complexity due to the different dynamics of each segment, non-unified interfaces between the layers, etc. which significantly impact the decisions regarding traffic routing, spectrum allocation, mobility management, QoS and traffic management, etc. Additionally, in the integrated network, as there are multiple available paths along which data traffic can be routed, optimized path selection is important to satisfy the Quality of Service (QoS) requirements of the traffic flows and improve the utilization of network resources [45]. These challenges, together with the diversity of 5G use cases with large-scale applications, highlight the importance of a unified management and control structure, and a dynamic resource allocation policy which are both scalable and flexible enough to handle the increasing complexity. Existing satellite networks employ a decentralized management architecture, scheduled link allocation, and static routing strategies, which make it difficult to support flexible traffic scheduling adapting to the changes in traffic demands [46] [47]. Complex handover mechanisms should be in place at the gateway nodes, while their high-power

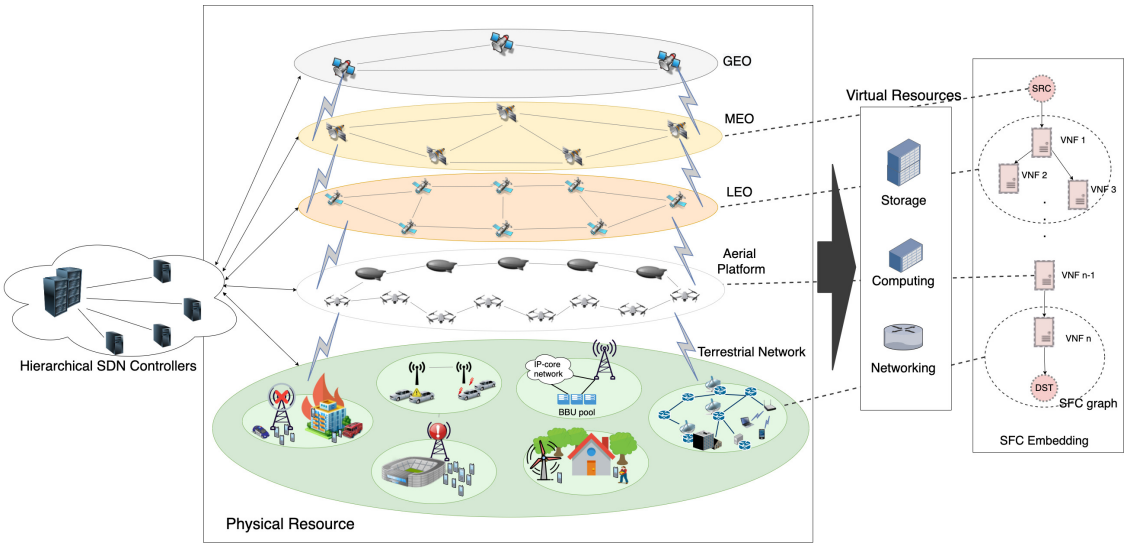


Figure 3.1: Space-Air-Ground Integrated Network (SAGIN)

consumption requirements need to be also taken into consideration. The key to address these issues is in the co-existence of two different but complementary supporting technologies namely, Software-Defined Networking (SDN), and Network Function Virtualization (NFV).

SDN allows for separating the control logic of the network from its forwarding logic and realizes a centralized management policy. This not only allows for a simple realization of the forwarding layer but also paves the way for dynamic configuration of control and management policies. With NFV, the network functions are decoupled from the proprietary hardware and realized through software. This both reduces the hardware cost and also by implementing the network functions on virtual machines or containers, provides an easier and more flexible provisioning of scalable solutions driving higher profitability for the network providers.

Towards realizing the SDN/NFV-enabled SAGIN two important optimization problems arise immediately; i) In the architectures concerned with GSO satellites, due to high delivered throughput per satellite, a large number of gateways are required; sometimes exceeding a couple of dozens. ii) Moreover, once the gateway deployment policy is decided, it is of paramount

importance to develop a smart and adaptive mechanism to handle the user hand-overs between the gateways or LEO satellites, traffic routing, load balancing, etc. Due to their abstract view of the network, SDN controllers are the best fit for this purpose. Thus, it becomes essential to formulate an optimization problem for deciding the minimum number of gateways and SDN controllers and their optimal location within the SAGIN.

While approaching the above-mentioned problems, depending on the design requirements, multiple objectives such as cost minimization, load balancing, mission offloading, reliability maximization, etc. can be sought under various network design constraints, such as guarantee requirements on the end-to-end latency, network security or availability, fault tolerance, etc [48]. In this chapter, we investigate two important problems to optimize the ground segment of SAGIN.

- Joint Satellite Gateway Placement and Routing for ISTNs. [49]
- Joint Satellite Gateway & SDN Controller Placement in SDN-enabled SAGINs. [50] [51]

3.2 Joint Satellite Gateway Placement and Routing for Integrated Satellite-Terrestrial Networks

Based on the extra degrees of freedom in ISTN deployment, brought forth by the softwarization and programmability of the network, the satellite gateway placement problem is of paramount importance for the space-ground integrated network. The problem entails the selection of an optimal subset of the terrestrial nodes for hosting the satellite gateways while satisfying a group of design requirements. Various strategic objectives can be pursued when deciding the optimal placement of the satellite gateways including but not limited to cost reduction, latency minimization, reliability assurance, etc. [52], [53], [54]. In what follows, we propose a method

for the cost-optimal deployment of satellite gateways on the terrestrial nodes. Particularly, we aim to minimize the overall cost of gateway deployment and traffic routing, while satisfying latency requirements. To this end, we formulate the problem as a mixed-integer linear program (MILP), with latency bounds as hard constraints. We derive an approximation method from our MILP model to significantly reduce the time complexity of the solution, at the expense of sub-optimal gateway placements, and investigate the corresponding trade-off. Furthermore, to reduce latency and processing power at the gateways, we impose a (varying) upper bound on the load that can be supported by each gateway. It is important to note that, traffic routing and facility placement are usually solved as different problems, but assigning a demand point to a facility without considering the other demand points might not be realistic. Given the significant interrelation of the two problems we develop and solve a single aggregated model instead of solving the two problems successively.

3.2.1 Network Model and Problem Description

We model the terrestrial network as an undirected $G = (V, E)$ graph where $(u, v) \in E$ if there is a link between the nodes $u, v \in V$. Let $J \subseteq V$ be the set of all potential nodes for gateway placement and $I \subseteq V$ be the set of all demand points. We note that the sets I and J are not necessarily disjoint. A typical substrate node $v \in V$ may satisfy one or more of the following statements: (i) node v is a gateway to the satellite, (ii) node v is an initial demand point of the terrestrial network, (iii) node v relays the traffic of other nodes, to one or more of the gateways. The ideal solution will introduce a set of terrestrial nodes for gateway placement which results in the cheapest deployment & routing, together with the corresponding routes from all the demand

points to the satellite, while satisfying the design constraints.

Regarding the delay of the network, we consider only propagation delay. The propagation delay of a path in the network is the sum of the propagation delays over its constituting links. A GEO satellite is considered in the particular system model [52]. Let d_{uv} represent the contribution of the terrestrial link (u, v) to the propagation delay of a path that contains that link. The propagation delay from a gateway to the satellite is constant [53]. Moreover, we consider multi-path routing for the traffic demands. Therefore, we define a flow $i \rightarrow j$ as the fraction of the traffic originated at node $i \in I$, routed to the satellite through gateway $j \in J$.

Once all the flows corresponding to node i are determined, all the routing paths from node i to the satellites are defined. Let c_j denote the cost associated with deploying a satellite gateway at node j and c_{uv} the bandwidth unit cost for each link (u, v). We also define a_i as the traffic demand rate of node i which is to be served by the satellite. The capacity of each link (u, v) is denoted by q_{uv} while the capacity of the gateway-satellite link is q_j for gateway j. Table 8.1 summarizes all the notations used for the parameters and variables throughout the section.

Variables	Description
y_j	The binary decision variable of gateway placement at node j
x_{ij}	The fraction of traffic demand originating at i passing through gateway j
f_{uv}^{ij}	The amount of traffic originating at i assigned to gateway j passing through the link (u, v)
Parameters	**Description**
$G = (V, E)$	Terrestrial network graph
J	The set of potential nodes for gateway placement
I	The set of demand points
c_j	The cost associated with deploying a satellite gateway at node j
c_{uv}	Bandwidth unit cost for link (u, v)
a_i	Traffic demand rate of node i
q_{uv}	Capacity of link (u, v)
q_j	Capacity of gateway-satellite link for gateway j
d_{uv}	Propagation delay of the terrestrial link (u, v)
d_{max}	Maximum allowed average delay for each terrestrial node

Table 3.1: System model parameters and variables

Since the propagation latency is usually the dominant factor in determining the network delay [54], we first develop a joint satellite gateway placement and routing (JSGPR) MILP formulation to minimize the cumulative gateway placement cost with hard constraints on the average propagation delay for the traffic of each demand point. Following that, we will derive a variant of JSGPR with load balancing (JSGPR-LB) which aims at mutually optimizing the overall cost and the load assigned to all the deployed gateways. Finally, we use an LP-based approximation approach to reduce the time complexity of the proposed scheme at the cost of a sub-optimal gateway placement.

3.2.2 Problem Formulation

3.2.2.1 JSGPR Initial MILP Formulation

Inspired by [55], we formulate a baseline for the JSGPR problem as the capacitated facility location-routing problem considering:

- The set of binary variables **y**, where y_j expresses the placement of a gateway at node j.
- The set of continuous variables **x**, where x_{ij} expresses the fraction of traffic demand originating at i passing through gateway j.
- The set of continuous variables **f**, where f_{uv}^{ij} expresses the amount of traffic demand originating at i, assigned to gateway j passing through the link (u, v).

We note that for a gateway u, the variable f_{uu}^{iu} represents the amount of traffic which is originated at node i and is forwarded to the satellite through the gateway placed at node u. The

resulting MILP formulation is as follows:

$$\textbf{Minimize} \quad \sum_{j \in J} c_j y_j + \sum_{i \in I} \sum_{j \in J} \sum_{(u,v) \in E} c_{uv} f_{uv}^{ij} \tag{3.1}$$

Demand Constraints:

$$\sum_{j \in J} x_{ij} = 1, \quad \forall i \in I \tag{3.2}$$

Feasibility Constraints:

$$x_{ij} \leq y_j, \quad \forall i \in I, j \in J \tag{3.3}$$

Capacity Constraints:

$$\sum_{i \in I} a_i x_{ij} \leq q_j y_j, \quad \forall j \in J \tag{3.4}$$

$$\sum_{i \in I} \sum_{j \in J} f_{uv}^{ij} \leq q_{uv} \quad \forall (u,v) \in E \tag{3.5}$$

Flow Constraints:

$$a_i x_{ii} + \sum_{v \in V:(i,v) \in E} \sum_{j \in J} f_{iv}^{ij} = a_i, \quad \forall i \in I \tag{3.6}$$

$$\sum_{v \in V:(v,u) \in E} f_{vu}^{ij} - \sum_{v \in V:(u,v) \in E} f_{uv}^{ij} = a_i x_{iu} \quad \forall i \in I, j \in J, u \in V, u \neq i \tag{3.7}$$

Domain Constraints:

$$y_j \in \{0,1\}, \quad \forall j \in J \tag{3.8}$$

$$x_{ij} \in [0,1], \quad \forall i \in I, j \in J \tag{3.9}$$

$$f_{uv}^{ij} \geq 0, \quad \forall (u,v) \in E, i \in I, j \in J \tag{3.10}$$

The objective function minimizes the total cost comprised of two terms; the first one represents the cost of installing and operating a gateway at location j, aggregated over the total number of gateways used. The second term corresponds to the transport/connection cost from the demand originating at i to its assigned gateway j, aggregated for all demands.

Constraints set (3.2) assures that traffic demands are supported by the selected gateways. Feasibility constraints set (3.3) make sure that demands are only served by open gateways. Constraints (3.6) and (3.7) enforce the flow conservation. The domains of y_i, x_{ij} and f_{uv}^{ij} variables are defined in constraints (3.8), (3.9), (3.10), respectively.

In the aforementioned formulation, we can express the set of x_{ij} variables, utilizing the corresponding last-hop flow variables:

$$x_{ij} = \frac{\sum_{u \in V} f_{uj}^{ij}}{a_i}, \quad \forall i \in I, j \in J \tag{3.11}$$

We can replace the set of x_{ij} variables as in (3.11). The new MILP model for JSGPR is presented in the next subsection.

3.2.2.2 JSGPR Final MILP Formulation

The new MILP formulation is as follows:

$$\textbf{Minimize} \quad \sum_{j \in J} c_j y_j + \phi \sum_{i \in I} \sum_{j \in J} \sum_{(u,v) \in E} c_{uv} f_{uv}^{ij} \tag{3.12}$$

Demand Constraints:

$$\sum_{j \in J} \sum_{u \in V} f_{uj}^{ij} = a_i, \quad \forall i \in I \tag{3.13}$$

Feasibility Constraints:

$$\sum_{u \in V} f_{uj}^{ij} \leq y_j a_i, \quad \forall i \in I, \forall j \in J \tag{3.14}$$

Capacity Constraints:

$$\sum_{i \in I} \sum_{u \in V} f_{uj}^{ij} \leq q_j y_j, \quad \forall j \in J \tag{3.15}$$

$$\sum_{i \in I} \sum_{j \in J} f_{uv}^{ij} \leq q_{uv}, \quad \forall (u,v) \in E \tag{3.16}$$

Flow Constraints:

$$\sum_{u \in V:(u,i) \in E} f_{ui}^{ii} + \sum_{v \in V:(i,v) \in E} \sum_{j \neq i \in J} f_{iv}^{ij} = a_i, \quad \forall i \in I \tag{3.17}$$

$$\sum_{v \in V:(u,v) \in E} f_{vu}^{ij} - \sum_{v \in V:(u,v) \in E} f_{uv}^{ij} = \sum_{u \in V} f_{uu}^{iu} \quad \forall i \in I, j \in J, u \in V, u \neq i \tag{3.18}$$

Domain Constraints:

$$y_j \in \{0,1\}, \quad \forall j \in J \tag{3.19}$$

$$f_{uv}^{ij} \geq 0, \quad \forall (u,v) \in E, i \in I, j \in J \tag{3.20}$$

where $\phi = \frac{1}{(\sum_{i \in I} a_i)}$ is the normalization factor between the two terms of the objective function. Additionally, to meet the average delay requirement for each demand point i we will impose a new constraint exploiting the corresponding flow variables:

$$\sum_{j \in J} \sum_{(u,v) \in E} \frac{f_{uv}^{ij}}{a_i} d_{uv} \leq d_{max} \quad \forall i \in I \tag{3.21}$$

where d_{max} is the maximum allowed average delay for each terrestrial node. We note that we have ignored the delay of the terrestrial-satellite link in the above calculation since it is a constant term as explained in section 9.2.

3.2.2.3 JSGPR-LB MILP Formulation

In the aforementioned specification, the assigned load to a gateway is solely bounded by the capacity constraints (3.15) and (3.16). We define a new single decision variable l_{max} to represent the maximum traffic assigned to a gateway, common for all the selected gateways. We add l_{max} as an additional term to the objective function. The new objective function is described in (3.22):

$$\textbf{Minimize} \quad (\sum_{j \in J} c_j y_j + \phi \sum_{i \in I} \sum_{j \in J} \sum_{(u,v) \in E} c_{uv} f_{uv}^{ij}) + \alpha l_{max} \tag{3.22}$$

where α is a constant factor determining the balance between the two terms of the objective function. Also, we change constraints (3.15) as following:

$$\sum_{i \in I} \sum_{u \in V} f_{uj}^{ij} \leq l_{max} \quad \forall j \in J \tag{3.23}$$

where $l_{max} \leq q_j \quad \forall j \in J$. We will call this last model, $JSGPR - LB$. The resulting optimization problem aims to balance the load of the gateways in conjunction with the cost of the gateway deployment [56], [57].

3.2.2.4 LP Relaxation and Approximation Algorithm

Since the MILP model is known to be NP-hard, the problem is intractable for larger-scale networks [55]. For the aforementioned JSGPR and JSGPR-LB MILP formulations, the optimal fractional solution is computed for the problem via linear programming relaxation. The relaxed problem can be solved by any suitable linear programming method, in polynomial time. A rounding technique is applied, similar to [58] to obtain the integer solution of the MIP problem. The resulting multi-commodity flow allocation problem is solved to identify the routing paths.

3.2.3 Performance Evaluation

In this section, we evaluate the performance of our satellite gateway placement method. We first describe the simulation environment setup and scenarios, then we review the performance evaluation results.

3.2.3.1 Performance Evaluation Setup

We evaluate the performance of our JSGPR and JSGPR-LB approaches on multiple real network topologies publicly available at the Topology Zoo [59]. The five different topologies we consider are listed in table 3.4. The link lengths and capacities are extracted from the topology zoo. The propagation delays are calculated based on the lengths of the links with the propagation velocity of $C = 2 \times 10^8 m/s$ [60]. The value of d_{max} is set to $10ms$, and the deployment cost for each node is taken from a uniform random generator $U(500, 1000)$. The unit bandwidth cost for all the links is set to be equal to 1. Also, q_j is set to $240Mbps$ for all the gateways. To develop and solve our MILP and LP models we use the CPLEX commercial solver. Our tests are carried out on a server with an Intel i5 CPU at 2.3 GHz and 8 GB of main memory.

3.2.3.2 Evaluation Scenarios

We have conducted two sets of experiments for two different cases. In the first case, we benchmark our LP-based approximation algorithm against the MILP-based optimal one for JSGPR, whereas the second case aims to observe the impact of our load-balancing approach on the gateway placement. For both cases and all topologies, if c^i_{max} is the maximum capacity among all the outgoing links of node i, the traffic rate originating at node i is taken from a

Topology	Nodes	Links
Sinet	13	18
Ans	18	25
Agis	25	32
Digex	31	35
Bell Canada	48	64

Table 3.2: Summary of the studied topologies

uniform distribution $U(\frac{2c^i_{max}}{3}, c^i_{max})$. Specifically, we will use the following metrics to evaluate the performance of our algorithm:

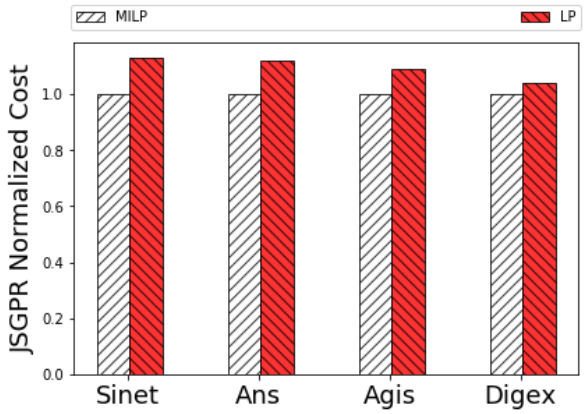
- **Solver Runtime** is the amount of time the CPLEX solver takes to solve the generated MILP or LP instances.
- **Average Delay** as explained in section 9.2.
- **Total Cost** is the total cost of deploying the satellite gateways and routing the traffic.
- **Gateway Load** is the amount of traffic flow assigned to the gateways after the gateway placement problem is solved.

3.2.3.3 Evaluation Results

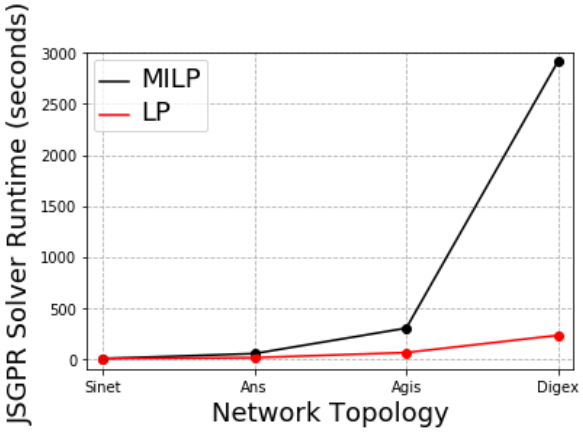
Experiment A - Approximation vs. Exact Method

Fig. 3.2a reflects the average normalized total cost for both the optimal MILP-based approach and the LP-based approximation of JSGPR. Due to the suboptimal placement, the LP-based approach results in additional deployment costs within the range of at most 13% of the optimal placement, but as the scale of the network gets larger this gap decreases. For Digex the approximate approach leads to about 4% increase in the deployment cost.

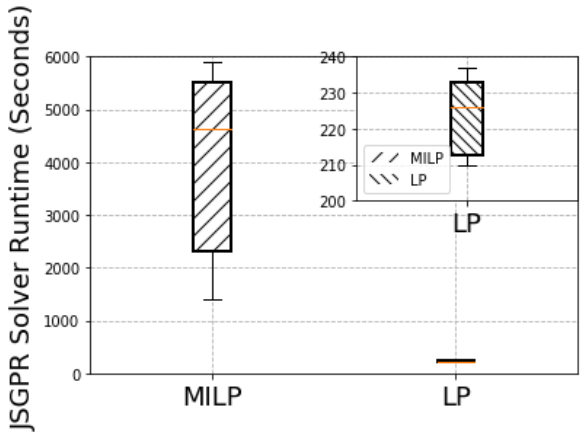
Fig. 3.2b depicts the average runtime for both the MILP and LP formulations for the 4 topologies. We note that for larger networks, CPLEX failed to provide the solution for the MILP problem, while our LP model continued to solve the problem within the expected time limit. Particularly, for the Bell-Canada topology, in 40% of the runs, CPLEX was not able to find any feasible solutions within the first 10 hours while the approach via approximation could provide



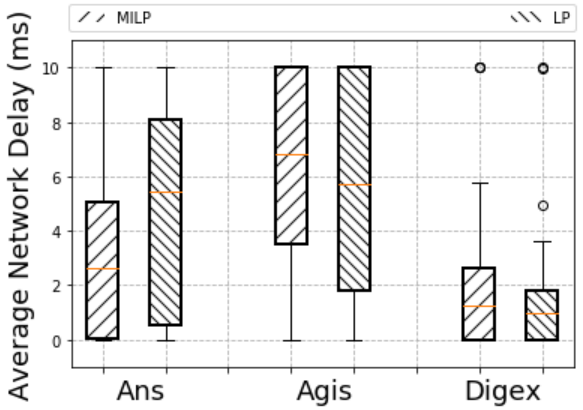
(a) Exp-A: Normalized Total Cost



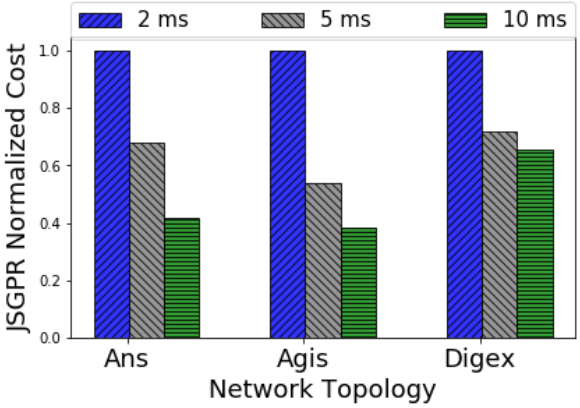
(b) Exp-A: Average Solver Runtime



(c) Exp-A: Solver Runtime for Digex

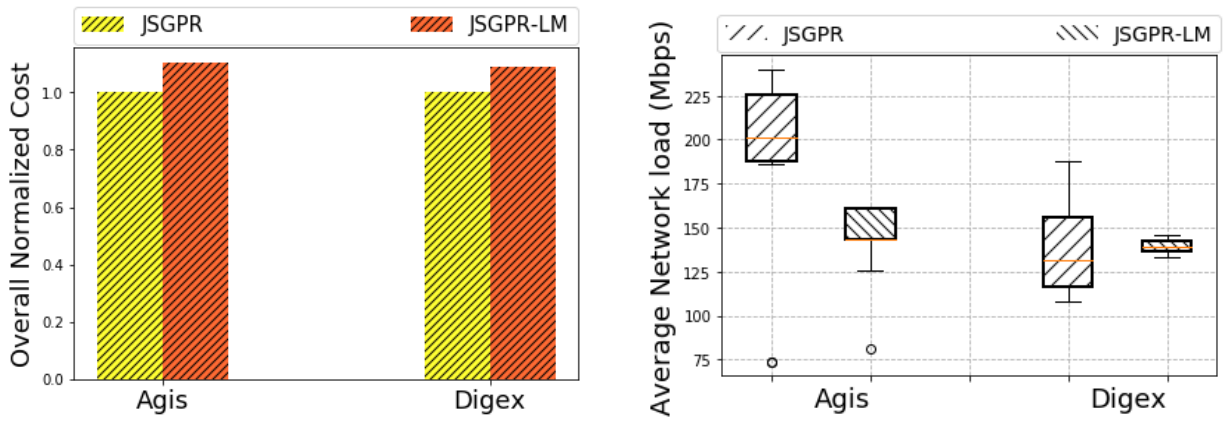


(d) Exp-A: Average Delay ($d_{max} = 10ms$)



(e) Exp-A: Total Cost with Varying Delay Bound

Figure 3.2: Experiment A - Approximation vs. Exact Method



(a) Exp-B: Normalized Total Cost

(b) Exp-B: Gateway Load

Figure 3.3: Experiment B - The Impact of Load Minimization

the suboptimal placement in less than 15 minutes. The solver runtime for Digex is shown in 3.2c. The average runtime for the LP is 230 seconds while for the MILP, it is around 3000 seconds.

Fig. 3.2d, represents the average delay for Ans, Agis, and Digex. The average of the expected experienced delay under the MILP model is $2.95, 6.18,$ and 2.15 seconds, while this value under the LP model is $4.68, 5.6,$ and 1.65 seconds. We note that the suboptimal procedure leads to additional deployed gateways in Agis and Digex, which in return will have the nodes end up closer to the gateways and consequently, experience lower average delay. Further, an insightful observation is that, in Agis, d_{max} is relatively a tight upper bound on the average delay experienced by each node. Therefore, the delay profile is pushed towards its upper bound, whereas, in Digex, due to the low density of links, more gateways are required to be deployed on the terrestrial nodes, which will make the gateways available more quickly; Therefore, the delay profile is inclined towards its lower bound.

Fig. 3.2e depicts the normalized cost of the JSGPR problem for different values of d_{max}. For instance, as indicated by this figure, if in Digex, a delay of $10ms$ is tolerable instead of $2ms$, a 35% reduction in cost results; Similarly, upgrading the service from a delay of $5ms$ to $2ms$ in

Agis, will almost double the cost.

Overall, the aforementioned figures illustrate that the performance of our approximation algorithm is very close to the exact approach, but with the important advantage of reduced time complexity which shows the efficiency of our proposed approximation method.

Experiment B - The Impact of Load Minimization

Fig. 3.3a shows the profile of the average load on the gateways for JSGPR, and JSGPR-LB, considering the Agis and Digex topologies. As expected, JSGPR-LB is more costly, since to evenly share the load between the gateways, a larger number of gateways will be required.

Fig. 3.3b depicts the profile of the load assigned to the gateways. In both depicted topologies, the load profile under JSGPR-LB is very thin and concentrated over its average, proving the efficiency of the formulation. Lower load on the placed gateways is achieved by the sub-optimal placement of the gateways (due to the larger number of gateways) which results in a more expensive gateway placement. As depicted in Fig. 3.3a the total cost of gateway placement in the studied topologies for JSGPR-LB is above and within the range of 16% of the optimal placement cost achieved by JSGPR.

3.3 Joint Satellite Gateway & SDN Controller Placement in SDN-enabled 5G-Satellite Hybrid Networks

Within the framework of SAGIN, the control plane is mainly in charge of making the routing decisions, conducting the traffic hand-offs between the satellite gateways and satellite switches, ensuring the service QoS, and delivering the necessary instructions to the SDN-enabled switches, leaving only the simple forwarding task to the data plane switches. Given the central

role of the SDN controllers in SDN-enabled SAGIN, it becomes important to maintain reliable communication paths between the SDN controllers and the SDN-enabled switches.

The major contributions of this section are as follows:

- We model the joint satellite gateway and SDN controller placement (JGCP) problem as a MILP. We consider two variants of this problem with two different sets of objectives. (i) Jointly maximizing the reliability of network-to-gateway and network-to-controller assignments. (ii) Minimizing the synchronization cost of the SDN controllers, jointly with minimizing the latency of the network-to-gateway and network-to-controller assignments.

- Inspired by [61], we provide a realistic scheme for modeling the synchronization overhead between the SDN controllers given the location and the network segment (terrestrial or space) in which each of the SDN controllers resides.

- We decompose the MILP model into two disjoint MILPs and then show that the resulting models lie in the framework of submodular optimization. Then we apply two approximation methods for solving these two models that run efficiently in time and provide provable theoretical optimality gaps.

- We conduct extensive numerical experiments to evaluate the performance of the provided methods and algorithms. We use publicly available real-world scenarios and various simulation settings for the performance evaluation tasks.

3.3.1 System Model

We consider an SDN-enabled hybrid 5G-satellite network as in Fig. 3.1 that consists of two logical segments; the *data plane*, and the *control plane*. The backhaul SDN-enabled switches reside within the data plane, where the users deliver their generated traffic to 5GC through the Radio Access Network (RAN); i.e. gNBs, small cells, etc. The high-throughput GEO satellite is also an SDN-enabled switch. The control plane realizes a logically centralized but physically distributed control scheme for network management where the SDN controllers reside on top of physical hardware. The control plane maintains a global view of the network and performs the network management over all portions of the network, (i.e. core, backhaul, access, etc.). The space segment mainly consists of SDN-enabled GEO satellites which communicate with one another through laser links. The communication between the SDN-enabled components in the backhaul, RAN, and 5GC network relies on fiber, while the communication between the space and ground segments is done through the satellite gateways or RNs. We focus only on the optimization of the ground segment, therefore we do not consider the mobility of the satellite layer and the inter-satellite links (ISLs) in our analysis.

We model the terrestrial network as an undirected $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ graph where $(u, v) \in \mathcal{E}$ if there is a link between the nodes $u, v \in \mathcal{V}$. Let $\mathcal{G} \subseteq \mathcal{V}$ be the set of all potential nodes for gateway placement and $\mathcal{K} \subseteq \mathcal{V}$ be the set of all potential switches for controller placement. We note that the sets $\mathcal{G}$ and $\mathcal{K}$ are not necessarily disjoint. A typical substrate node $v \in \mathcal{V}$ may satisfy one or more of the following statements: (i) A satellite gateway is placed at node v, (ii) node v is an initial demand point of the terrestrial network, (iii) node v hosts an SDN controller. Various objectives may be sought when solving the joint satellite gateway and SDN

controller placement problem from reliability maximization, latency minimization, gateway and controller load balancing, etc., depending on the design requirements. The optimal solution will introduce two sets of terrestrial locations $g_1, g_2, ..., g_m$, and $k_1, k_2, ..., k_n$ for gateway placement and controller placement, in respective order, together with the node-to-gateway, and node to controller assignment.

Regarding the delay of the network, we consider only propagation latency. The propagation latency of a path in the network is the sum of the propagation delays over its constituting links. A GEO satellite is considered in the particular system model similar to [62], [63], [64], and [65]. Let d_{uv} represent the contribution of the terrestrial link (u, v) to the propagation delay of a path that contains that link. The propagation delay from a gateway to the satellite is constant.

Furthermore, let us define the reliability r_{ij}^s of each terrestrial-satellite path P_{ij}^s from node $i \in \mathcal{V}$ to satellites s which passes through gateway $j \in \mathcal{G}$ as in [66], where reliability of a path is modeled as the product of the reliability along its constituting components; i.e.

$$r_{ju}^s = (1 - P_e^{sg}) \prod_{e \in P_{ju}^s} (1 - P_e) \prod_{u \in P_{ju}^s} (1 - P_v) \quad \forall j \in \mathcal{G}, u \in \mathcal{V} \tag{3.24}$$

Similarly, the reliability of a controller-terrestrial node path is defined as the product of reliability along its constituting components as follows.

$$r_{ku} = \prod_{e \in P_{ku}} (1 - P_e) \prod_{v \in P_{ku}} (1 - P_v) \quad \forall k \in \mathcal{K}, u \in \mathcal{V} \tag{3.25}$$

Table 8.1 summarizes the notations reserved for variables and parameters frequently used.

Variables	Description
x_j	The binary decision variable of gateway placement at node j
y_k	The binary decision variable of controller placement at node k
w_{jv}	The binary decision variable for the assignment of node v to gateway j
z_{jv}	The binary decision variable for the assignment of node v to controller j
Parameters	**Description**
$\mathcal{T} = (\mathcal{V}, \mathcal{E})$	Terrestrial network graph
$\mathcal{G}$	The set of potential nodes for gateway placement
$\mathcal{K}$	The set of potential nodes for controller placement
m_{max}	The maximum number of permitted gateways for deployment
k_{max}	The maximum number of permitted controllers for deployment
d_{uv}	The propagation delay of the terrestrial link (u, v)
r^s_{jv}	The reliability of the shortest path from node v to satellite s passing through gateway j
r_{kv}	The reliability of the shortest path from node v to SDN controller k
P_e	Failure probability of edge e from the terrestrial network
P_v	Failure probability of node v from the terrestrial network
P^{sg}_e	Failure probability of gateway-satellite edge e
$U(w, x, y, z)$	Utility of the joint gateway and controller placement
$V(w, x, y, z)$	Cost of the joint gateway and controller placement

Table 3.3: System model parameters and variables

3.3.2 Optimization Model & Solution Approach

We propose a two-step solution for the deployment of satellite gateways and the placement of the SDN controllers due to various sets of design requirements as hard constraints. In the first step, we model the gateway deployment problem as a MILP and determine the optimal placement policy for gateways, and then assuming the placement of gateways is computed and given as input to the controller placement phase, we formulate the SDN controller placement problem again as a MILP and determine the optimal controller placement policy. We use the sub (super)-modularity property and sub-modular optimization techniques that are described in detail in subsection 3.3.2.2.

3.3.2.1 Problem MILP Formulation

As explained in section 3.4, within the framework of SAGIN, various objectives may be pursued by solving the problem at hand. One may want to minimize the cost of operation, the overall number of deployed equipment, and the latency between the equipment and the demand points, or maximize the reliability of the chosen paths. In its most general form, we wish to maximize a composite utility function which is a function of the placement of the gateways, the assignment of demand points to gateways, the placement of controllers, and the assignment of demand points to controllers. In other words, a total utility function is going to be maximized that is determined by the choice of the deployment, and the assignment policies. i.e.

$$Maximize \quad U(w, x, y, z) \tag{3.26}$$

with the following sets of decision variables:

- The set of binary assignment decision variables **w** where $w_{ij} = 1$, if the traffic of node j is assigned to the gateway placed at node i.
- The set of binary placement decision variables **x** where $x_i = 1$, if a gateway is placed at node i.
- The set of binary assignment decision variables **z** where $z_{ij} = 1$, if node j is assigned to the controller placed at node i.
- The set of binary placement decision variables **y** where $y_i = 1$, if a controller is placed at node i.

and subject to various design requirements enumerated below. Each terrestrial node, whether it is a controller or not, has to be assigned to a gateway. i.e.,

$$\sum_{j \in \mathcal{G}} w_{jv} = 1 \quad \forall v \in \mathcal{V} \tag{3.27}$$

and the assignment of the terrestrial node v to node j is only valid if, in the resulting placement policy, a gateway is located at node j.

$$w_{jv} \leqslant x_j \quad \forall j \in \mathcal{G}, v \in \mathcal{V} \tag{3.28}$$

Each node has to be assigned to a controller. i.e.,

$$\sum_{k \in \mathcal{K}} Z_{kv} = 1 \quad \forall v \in \mathcal{V} \tag{3.29}$$

And again this assignment has to be valid:

$$Z_{kv} \leqslant y_k \quad \forall v \in \mathcal{V}, k \in \mathcal{K} \tag{3.30}$$

In the case that the number of resources is limited, there is a maximum number of gateways and SDN controllers available. Therefore the number of facilities in the deployment policy should not exceed the maximum available:

$$\sum_{j \in \mathcal{G}} x_j \leqslant g_{max} \tag{3.31}$$

$$\sum_{j \in \mathcal{G}} y_k \leqslant k_{max} \tag{3.32}$$

where g_{max} and k_{max} are the maximum number of available gateways and SDN controllers to be deployed.

All constraints mentioned above, possibly together with other scenario-specific constraints may be embedded in the optimization problem depending on the design requirements and the a-priori knowledge of the network. The utility function U can be a composite of multiple utility functions U_i. For instance, we can consider the utility obtained by the minimization of the number of deployed gateways and the utility obtained by maximizing the reliability of the controller gateway paths. Similarly, objectives inspired by balancing the load on the gateways and/or the controllers can be taken into account, etc. Corresponding to each utility function we can think of a cost function V by simply negating the sign of the utility function. Therefore, the utility maximization problem can be also pursued as a cost-minimization one. Let us present the baseline MILP formulation for the joint satellite gateway deployment and SDN controller placement as follows:

In the case that the average reliability of the network is considered, we may have the corresponding utility function as:

$$U^{rel}(w,x,y,z) = \frac{1}{|\mathcal{V}|}(\sum_{j\in\mathcal{G}}\sum_{v\in\mathcal{V}} r^s_{jv} w_{jv} + \sum_{k\in\mathcal{K}}\sum_{v\in\mathcal{V}} r_{kv} z_{kv}) \tag{3.33}$$

where, the first term corresponds to the average reliability of node-to-gateway assignments, and the second term captures the same for the assignment to SDN controllers. We note that within the context of SDN, the SDN controllers need to communicate information with one another regarding the state of the network to maintain a global view of the network. This is often denoted as the *synchronization* cost in the literature. A poor configuration of the SDN-

enabled network may result in huge excessive overhead contributing negatively to the cost of operation. It is therefore desirable to minimize the overhead corresponding to the synchronization between the SDN controllers. In [61], the authors provide a realistic model for capturing the synchronization cost between the SDN controllers at the edge which we adapt in this work with some modifications:

$$V^C(w,x,y,z) = \sum_{m,n\in\mathcal{K}} l^{(1)}_{mn} y_m y_n + \sum_{m,n\in\mathcal{K}} y_m y_n l^{(2)}_{mn} (\sum_{v\in\mathcal{V}} z_{mv}) + \sum_{m\in\mathcal{K}} y_m l^{(3)}_m \tag{3.34}$$

where the first summation captures the constant communication cost between two controllers, and the second term captures the cost that is dependent on the load of each controller, i.e. the number of users that are assigned to that controller. Finally, the third part of the synchronization cost captures the synchronization between the controllers on the ground segment to those in the space segment. In the following, we assume that the constant cost between two ground controllers is mainly a function of the latency between them, i.e. $l^{(1)}_{mn} = d_{mn}$, the variable cost depends on the load of the controllers through a constant $l^{(2)}_{mn} = l^{con}$, and the inter-segment cost depends on the distance of each controller to the gateway it is assigned to. In other words, the synchronization cost from equation (3.35), can be written as:

$$V^C_{sync}(w,x,y,z) = \sum_{m,n\in\mathcal{K}} d_{mn} y_m y_n + \sum_{m,n\in\mathcal{K}} y_m y_n l^{con} (\sum_{v\in\mathcal{V}} z_{mv}) + \sum_{m\in\mathcal{K}} y_m (\sum_{j\in\mathcal{G}} d_{jm} w_{jm}) \tag{3.35}$$

Therefore, the objective to minimize the average network latency, and the SDN controller

synchronization overhead can be modeled as:

$$V(w,x,y,z) = \sum_{j\in\mathcal{G}} x_j + \alpha \sum_{j\in\mathcal{G}} \sum_{v\in\mathcal{V}} d_{jv} w_{jv} + \psi[\sum_{k\in\mathcal{K}} \sum_{v\in\mathcal{V}} d_{kv} z_{kv} + \beta V^{C}_{sync}(w,x,y,z)] \quad (3.36)$$

where α, β, and ψ, are the corresponding factors to balance the emphasis of the objective function on different terms. Therefore, we formulate the joint satellite gateway placement and SDN controller placement in an SDN-enabled SAGIN for reliability maximization as a MILP as follows:

$$\text{Maximize} \quad U^{rel}(w,x,y,z) \quad (3.37)$$

$$\text{subject to:} \quad (3.27), (3.28), (3.29), (3.30), (3.31), (3.32) \quad (3.38)$$

$$x_j \in \{0,1\}, \quad \forall j \in \mathcal{G} \quad (3.39)$$

$$w_{jv} \in \{0,1\}, \quad \forall v \in \mathcal{V}, \quad \forall j \in \mathcal{G} \quad (3.40)$$

$$z_{kv} \in \{0,1\}, \quad \forall k \in \mathcal{K}, \quad \forall v \in \mathcal{V} \quad (3.41)$$

$$y_k \in \{0,1\}, \quad \forall k \in \mathcal{K} \quad (3.42)$$

To formulate the problem for minimizing the average network latency and overhead, we have to apply some standard linearization methods as in [61], to linearize the synchronization cost $V^{C}_{sync}(w,x,y,z)$. Namely, let us define a new set of binary variables ($\theta_{mn} \in \{0,1\}$) to replace $\{y_m y_n\}$ term.

Then we add the following linear constraints:

$$\theta_{mn} \leq y_m, \quad \forall m, n \in \mathcal{K} \tag{3.43}$$

$$\theta_{mn} \leq y_n, \quad \forall m, n \in \mathcal{K} \tag{3.44}$$

$$\theta_{mn} \geq y_m + y_n - 1, \quad \forall m, n \in \mathcal{K} \tag{3.45}$$

Similarly, we introduce $(\phi_{mnv} \in \{0, 1\})$, and $(\mu_{mj} \in \{0, 1\})$ to replace $\{y_m y_n z_{mv}\}$, and, $\{y_m w_{jm}\}$ in respective order, and add the following constraints:

$$\phi_{mnv} \leq \theta_{mn}, \quad \forall m, n \in \mathcal{K}, v \in \mathcal{V} \tag{3.46}$$

$$\phi_{mnv} \leq y_m, \quad \forall m, n \in \mathcal{K}, v \in \mathcal{V} \tag{3.47}$$

$$\phi_{mnv} \geq y_m + \theta_{mn} - 1, \quad \forall m, n \in \mathcal{K}, v \in \mathcal{V} \tag{3.48}$$

$$\mu_{mj} \leq y_m, \quad \forall m \in \mathcal{K}, j \in \mathcal{G} \tag{3.49}$$

$$\mu_{mj} \leq w_{jm}, \quad \forall m \in \mathcal{K}, j \in \mathcal{G} \tag{3.50}$$

$$\mu_{mj} \geq y_m + w_{jm} - 1, \quad \forall m \in \mathcal{K}, j \in \mathcal{G} \tag{3.51}$$

towards achieving the following MILP formulation.

$$\text{Mainimize} \quad V^{lat+ovrh.}(w, x, y, z) \tag{3.52}$$

$$\text{subject to:} \quad (3.27) - (3.30), \text{and } (3.43) - (3.51) \tag{3.53}$$

$$x_j \in \{0, 1\}, \quad \forall j \in \mathcal{G} \tag{3.54}$$

$$w_{jv} \in \{0, 1\}, \quad \forall v \in \mathcal{V}, \quad \forall j \in \mathcal{G} \tag{3.55}$$

$$z_{kv} \in \{0, 1\}, \quad \forall k \in \mathcal{K}, \quad \forall v \in \mathcal{V} \tag{3.56}$$

$$y_k \in \{0, 1\}, \quad \forall k \in \mathcal{K} \tag{3.57}$$

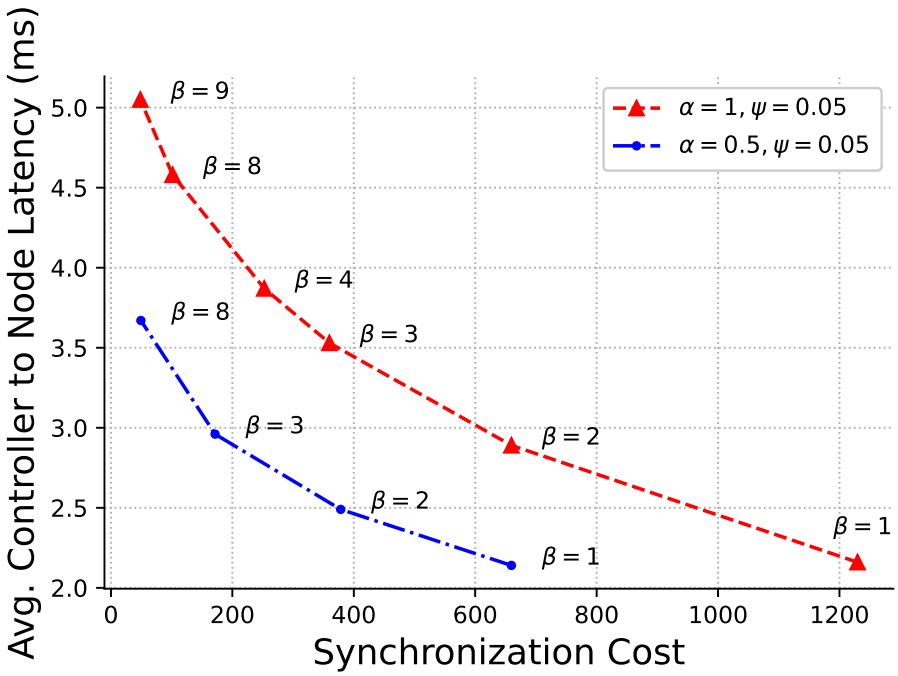


Figure 3.4: Trade-off between the synchronization cost and average controller-to-node latency

Fig. 3.4, illustrates the impact of parameters α, and β, on the solution of the problem, while keeping ψ constant, where the trade-off between the synchronization cost and the average

controller-to-node latency is depicted for two different cases of α. When $\alpha = 0.5$ is set, by increasing the value of β, more emphasis is put on minimizing the synchronization cost compared to the average controller-to-node latency. A similar trend is observed when setting $\alpha = 1$, but with the curve shifted upwards; since increasing the value of α, will put more emphasis on optimizing the gateway placement policy.

The joint satellite gateway and SDN controller placement problem is an instance of the facility location problem that is known to be $\mathcal{NP}$-hard. Commercial MILP solvers such as CPLEX can be used directly to solve such models for small instances using well-established methods such as branch& bound, branch& cut, etc., however due to the large number of variables, the problem will soon become intractable for large instances, and such methods will tend to be ineffective. Therefore, to overcome this issue, it is crucial to come up with approximation methods that scale well with the size of the network and generate acceptable solutions both in terms of time results accuracy and time complexity. In the next subsection, we give a short review of submodular optimization and then will justify that while specifying the utility/cost functions, our model can suitably fit into the submodular optimization framework subject to constraints of known types. We then apply two interesting algorithms from the submodular optimization literature to efficiently solve our problems.

3.3.2.2 Sub (Super)-modular Optimization Methods

Let us start with the definition of submodular functions.

Definition 4 ***Submodular Functions.*** *Let finite set G of elements be the ground set. Then a function $f : 2^G \rightarrow \mathbb{R}$ over the ground set is said to be submodular if for all subsets $A, B \subseteq G$, it*

holds that $f(A) + f(B) \geq f(A \cup B) + f(A \cap B)$.

Equivalently, f *is said to be submodular if for all subsets* $A, B \subseteq G$*, with* $A \subseteq B$ *and every element* $i \in G \backslash B$ *it holds that* $f(A \cup \{i\}) - f(A) \geq f(B \cup \{i\}) - f(B)$.

This intuitively means that for a submodular set function, adding an element to a subset will result in diminishing returns with increasing the subset size. We also note that if for all subsets $A, B \subseteq G$, with $A \subseteq B$ it holds that $f(A) \leq f(B)$, then f is called a *monotone submodular function*. Moreover, it is worth noting that, if f, g are submodular functions then $[f + g], [kf, k > 0], [-f]$, are submodular, submodular, and supermodular in order.

Given the diminishing return property of the submodular functions, many utility functions can suitably fit in this class. Therefore, motivated by the natural application of submodularity property in real-world scenarios such as welfare maximization, social networks, information gathering, feature selection, etc., optimization problems involving submodular/supermodular functions have developed a lot of interest among the research community. Especially, over the past decade, a lot of interesting methods have been proposed for approximately solving the submodular/supermodular optimization problems subject to a variety of constraints while providing acceptable optimality bounds. Within the SDN research community, several papers have used submodular optimization to model and solve multiple resource allocation problems [67] [68] [61]. Moreover, in [69], a very interesting taxonomy of such problems in the mobile edge computing (MEC) framework, along with an insightful discussion on their submodularity property is provided. In this section, we show that the utility functions corresponding to the objectives of our interest are submodular or the corresponding cost functions can be modeled as supermodular functions. Then we appeal to two interesting algorithms from the submodular

optimization literature to efficiently come up with approximate solutions. Let us first state two theorems that we will use in the subsequent sections.

Theorem 5 ([70]) *There exists a* $(1 - 1/e - \varepsilon)$*-approximation algorithm for maximizing a monotone non-negative submodular function with cardinality constraints. The algorithm has a* $\mathcal{O}(\frac{n}{\varepsilon} \log \frac{n}{\varepsilon})$ *time complexity.*

Theorem 6 ([71]) *There exists a* $1/2$*-approximation randomized greedy algorithm for maximizing a non-negative submodular function, which runs in linear time.*

3.3.2.3 Problem Decomposition

The joint placement problem specified in equations (3.93)-(3.96) contains variables corresponding to the placement of gateways and the placement of controllers. This will render a rather complicated problem with non-linear constraints due to the unknown location of both the gateways and the controllers. Therefore, to simplify the problem we adopt a sequential two-step approach. Since the location of the gateways has a great impact on both the latency and the reliability requirements, we first determine the optimal placement of the gateways and then choose the best controller placement policy accordingly. Hence, we can decompose the utility function in (3.93) and write:

$$U(w, x, y, z) = U_g(w, x) + U_c^{wx}(y, z) \tag{3.58}$$

where the first term is concerned only with the placement of the gateways and the second term corresponds to the utility obtained by placing the SDN controllers given that the location and

the assignment of gateways are determined. Similarly, if a cost function is to be minimized we obtain:

$$V(w, x, y, z) = V_g(w, x) + V_c^{wx}(y, z) \tag{3.59}$$

Under this decomposition, we are dealing with two sub-problems: (i) *Satellite Gateway Deployment Problem*, and (ii) *SDN Controller Placement Problem*. In the next subsections, we address these two problems sequentially.

3.3.2.4 Optimal Gateway Placement

Following the discussion in section 3.3.2.1, multiple objectives may be sought by deciding the optimal gateway placement policy. For instance, the network provider may intend to minimize the average network latency while at the same time minimizing the number of gateways deployed. Thus, an optimal gateway placement policy minimizes the aggregated cost function. Given the above discussion, we formulate the gateway placement problem for minimizing the aggregated cost (delay-oriented) with the objective function:

$$V_g(w, x) = V_g^1(w, x) + \alpha V_g^2(w, x) \tag{3.60}$$

where,

$$V_g^1(w, x) = \sum_{j \in \mathcal{G}} x_j \tag{3.61}$$

$$V_g^2(w, x) = \sum_{v \in \mathcal{V}} \sum_{j \in \mathcal{G}} d_{jv} w_{jv} \tag{3.62}$$

The corresponding MILP model is as follows:

$$\text{Minimize} \quad V_g(w, x) \tag{3.63}$$

$$\text{subject to:} \quad (3.27), (3.28) \tag{3.64}$$

$$x_j \in \{0, 1\}, \quad \forall j \in \mathcal{G} \tag{3.65}$$

$$w_{jv} \in \{0, 1\}, \quad \forall v \in \mathcal{V}, \quad \forall j \in \mathcal{G} \tag{3.66}$$

Next, we argue by the following theorem that if the placement of gateways is known, then the optimal assignment policy can be uniquely determined and then use this observation to show that the cost function $V_g(w, x)$ is supermodular.

Lemma 7 *Given the placement of the gateways x, the optimal assignment policy w^* can be uniquely determined; i.e. there exists a deterministic function $g : G \to W$ for which $w^* = g(x)$.*

Proof. Once the placement of gateways is determined, $V_g^1(w, x)$ is already decided. Let $\mathcal{X} \subseteq \mathcal{G}$ be the set of nodes where a gateway is placed. In other words,

$$\mathcal{X} = \{j \in \mathcal{G} : x_j = 1\} \tag{3.67}$$

To minimize the objective function (3.60) the optimal assignment policy is as follows:

$$w^*_{jv} = \mathbb{1}_{\{j = \arg\min_{j \in \mathcal{X}} d_{jv}\}}, \quad \forall j \in G, v \in \mathcal{V}. \tag{3.68}$$

which implies $w^* = g(x)$. ■

Theorem 8 *The cost function $V_g(w, x)$ is supermodular.*

Proof. Let $\mathcal{A}, \mathcal{B} \in \mathcal{G}$ be two arbitrary sets of locations corresponding to two separate gateway placement policies such that $\mathcal{A} \subseteq \mathcal{B}$. Let us update the policies $\mathcal{A}$, and $\mathcal{B}$ by adding a new gateway at location $g \in \mathcal{G} \setminus \mathcal{B}$. For any set function f, let $R_g^{\mathcal{A}}(f)$ be the amount of change in the cost function by adding the new gateway g to policy $\mathcal{A}$. $R_g^{\mathcal{B}}(f)$ can be defined in a similar fashion. i.e. $R_g^{\mathcal{A}}(f) = f(\mathcal{A} \cup \{g\}) - f(\mathcal{A})$. By the definition of the supermodular functions, a function f is supermodular if $R_g^{\mathcal{A}} \leq R_g^{\mathcal{B}}$. $V_g^1(w, x)$ only depends on the placement policy, and the marginal return of adding any new gateway is constant. Thus, $R_g^{\mathcal{A}} = R_g^{\mathcal{B}}$ and $V_g^1(w, x)$ is modular and therefore supermodular. ■

To prove the supermodularity of $V_g^2(w, x)$, let $j_v^{\mathcal{A}} =: \arg\min_{j \in \mathcal{A}} d_{jv}$ be the location of the gateway to which node $v \in \mathcal{V}$ is assigned under gateway placement policy $\mathcal{A}$. The definition for $j_v^{\mathcal{B}}$ follows similarly. Moreover, let us consider the set of all nodes that can contribute to reducing the cost of gateway placement by switching the recently introduced gateway:

$$\sigma(\mathcal{V}) = \{v \in \mathcal{V} : d_{gv} \leq \min_{j \in \mathcal{B}} d_{jv}\} \tag{3.69}$$

Then we can write:

$$R_g^{\mathcal{B}} - R_g^{\mathcal{A}} = \sum_{v \in \mathcal{V}} \min(0, (d_{gv} - d_{j_v^{\mathcal{B}} v})) - \sum_{v \in \mathcal{V}} \min(0, (d_{gv} - d_{j_v^{\mathcal{A}} v})) \tag{3.70}$$

$$\geq \sum_{v \in \sigma(\mathcal{V})} (d_{gv} - d_{j_v^{\mathcal{B}} v}) - \sum_{v \in \sigma(\mathcal{V})} (d_{gv} - d_{j_v^{\mathcal{A}} v}) \tag{3.71}$$

$$= \sum_{v \in \sigma(\mathcal{V})} (d_{j_v^{\mathcal{A}} v} - d_{j_v^{\mathcal{B}} v}) \geq 0. \tag{3.72}$$

where each node v contributes to the cost reduction if and only if it can get closer to its nearest

gateway by switching to the newly added gateway g. The lower bound in equation 3.71, comes from the fact that it only includes a subset of such nodes for policy $\mathcal{A}$; therefore potential cost savings may be left out of equation 3.71.

As explained in section 3.3.2.2, literature on the approximation algorithms for optimizing submodular functions is very rich where most of the interesting results are derived for the case of non-negative functions. Also, it is worth noting that minimizing a supermodular function may be approached by first casting the optimization problem as a non-negative submodular maximization and then utilizing the corresponding approximation algorithms. With such an approach, the authors in [61] have previously used the algorithm mentioned in theorem 6, for SDN controller placement at the edge. Similarly, let $\bar{V}_g$ be the maximum of equation (3.60). Thus, minimizing (3.60) is tantamount to maximizing

$$\hat{V}_g(\mathcal{X}) = \bar{V}_g - (V_g^1(\mathcal{X}) + V_g^2(\mathcal{X})) \tag{3.73}$$

that is a non-negative submodular function, in which the dependence on w is dropped according to theorem 7. Algorithm 2 summarizes how the $(1/2)$-approximation procedure occurs.

Algorithm 2 (1/2)-approximation greedy algorithm

Input: $\hat{v}_g : 2^{\mathcal{G}} \rightarrow \mathbb{R}_+$
Output: $(\bar{X}, \hat{v}_g(\bar{X}))$

Initialize $\underline{X} = \emptyset, \quad \bar{X} = \{1\}^{|\mathcal{G}|}$
for $j \in \mathcal{G}$:
  $\bar{\Delta} = max(\hat{V}_g(\bar{X}) - \hat{V}_g(\bar{X} \setminus \{i\}), 0)$
  $\underline{\Delta} = max(\hat{V}_g(\underline{X}) - \hat{V}_g(\underline{X} \cup \{i\}), 0)$
  Set $\underline{X} = \underline{X} \cup \{i\}$ with probability $\frac{\underline{\Delta}}{(\underline{\Delta}+\bar{\Delta})}$
  otherwise
  Set $\bar{X} = \bar{X} \setminus \{i\}$
end
return $(\bar{X}, \hat{v}_g(\bar{X}))$

The algorithm starts by taking two extreme cases and then decides on the placement of a gateway at each location in an iterative fashion. Before ith iteration begins, a gateway is present in ith location by the policy $\bar{X}$ and absent by the policy $\underline{X}$. The algorithm computes the contribution of the inclusion/exclusion of a gateway in ith location and makes a randomized choice accordingly. After iteration i ends both the policies agree on the inclusion/exclusion of a gateway at location i. Hence, when the execution of the algorithm finishes the two policies will be the same.

It is important to note that in realistic cases the number of facilities and the amount of resources are limited. Therefore, it makes sense to formulate the gateway placement optimization problem under an additional *cardinality* constraint. Let us consider the reliability-oriented utility function. i.e.

$$U_g(w,x) = \sum_{v\in\mathcal{V}} \sum_{j\in\mathcal{G}} r^s_{jv} w_{jv} \tag{3.74}$$

Therefore, the MILP model will be as follows:

$$\text{Maximize} \quad U_g(w,x) \tag{3.75}$$

$$\text{subject to:} \quad (3.27), (3.28), (3.31) \tag{3.76}$$

$$x_j \in \{0,1\}, \quad \forall j \in \mathcal{G} \tag{3.77}$$

$$w_{jv} \in \{0,1\}, \quad \forall v \in \mathcal{V}, \quad \forall j \in \mathcal{G} \tag{3.78}$$

In the next theorem, we claim that $U_g(w,x)$ is a non-negative *submodular* function that is monotone as well.

Theorem 9 *$U_g(w,x)$ is a monotone submodular function.*

Proof. Similar to the proof of theorem 8, consider any $\mathcal{A}, \mathcal{B} \in \mathcal{G}$ as two arbitrary sets of locations corresponding to two separate gateway placement policies such that $\mathcal{A} \subseteq \mathcal{B}$ and suppose a new gateway is going to be deployed at location $g \in \mathcal{G}$. To see that $U_g(w, x)$ is an increasingly monotone function, observe that:

$$\forall v \in \mathcal{V} : r^s_{j^{\mathcal{B}}_v v} \geq r^s_{j^{\mathcal{A}}_v v} \tag{3.79}$$

where $j^{\mathcal{A}}_v$, and $j^{\mathcal{B}}_v$ maximize the reliability of assignment for node v. The assertion follows by summing the LHS and the RHS of the last inequality, over $v \in \mathcal{V}$. To show the submodularity of $U_g(w, x)$, we need to compare $R^{\mathcal{B}}_g$ and $R^{\mathcal{B}}_g$. With a similar logic define:

$$\sigma(\mathcal{V}) = \{v \in \mathcal{V} : r^s_{gv} \geq \max_{j \in \mathcal{B}} r^s_{jv}\} \tag{3.80}$$

It holds that:

$$R^{\mathcal{B}}_g - R^{\mathcal{A}}_g \leq \sum_{v \in \sigma(\mathcal{V})} (r^s_{gv} - r^s_{j^{\mathcal{B}}_v v}) - \sum_{v \in \sigma(\mathcal{V})} (r^s_{gv} - r^s_{j^{\mathcal{A}}_v v}) = \sum_{v \in \sigma(\mathcal{V})} (r^s_{j^{\mathcal{A}}_v v} - r^s_{j^{\mathcal{B}}_v v}) \leq 0 \tag{3.81}$$

Therefore, by the definition of submodularity, $U_g(w, x)$ is submodular. ■

To approximately solve this problem, we use theorem 5. Algorithm 3 outlines the corresponding procedure based on theorem 5.

The algorithm starts by finding the most beneficial single gateway placement, i.e. a single gateway that provides the highest average network reliability among all others. Then according to this maximum contribution, it iteratively picks a value for minimum contribution and adds all the gateways that can be more beneficial than that, all in a single iteration. In other words, at

Algorithm 3 $(1 - 1/e - \varepsilon)$-approximation algorithm

Input: $U_g : 2^{\mathcal{G}} \rightarrow \mathbb{R}_+, s \in \{1, ..., |\mathcal{G}|\}, 0 < \varepsilon < 1$
Output: $(X, U_g(X))$

```
Initialize: X ← ∅, and d ← max_{j∈G} U_g({j})
for (w = d; w ≥ (ε/n)d; w ← w(1 − ε))
    for j ∈ G
        if |X ∪ {j}| ≤ s and U_g(X ∪ {j}) − U_g(X) ≥ w
            X ← X ∪ {j}
        endif
    end
end
return (X, U_g(X))
```

each iteration, all those locations that can pass the minimum contribution level, are eligible for placement of a gateway. The minimum contribution decreases in each iteration according to a pre-defined step-size and therefore the criterion for eligibility becomes less strict over time. If at any step before the end of the loop, the number of the deployed gateways reaches that maximum allowed s, the algorithm terminates immediately.

3.3.2.5 Optimal Controller Placement

Once the placement of the gateways is determined, we need to choose the optimal controller placement policy accordingly. Since controllers are expensive resources, we wish to deploy as few controllers as possible. Furthermore, the paths from gateways to controllers must be reliable enough, to prevent disconnections between the control and the data plane. Finally, since the SDN controllers need to maintain a global view of the network, they will have to synchronize by communicating information regarding the global state of the network. As explained in section 3.3.2.1, regardless of how this is achieved, the rate of communication needed between any pair of controllers is proportional to the load of each controller, in addition to the constant rate.

Therefore, the cost of this communication needs to be taken into account which can be a function of the location of the controllers, the latency between them, etc.; The objective function of the SDN controller placement problem (latency/overhead oriented), consists of four parts:

$$V_c^{wx}(y,z) = V_{c_1}^{wx}(y,z) + \beta[V_{c_2}^{wx}(y,z) + V_{c_3}^{wx}(y,z) + V_{c_4}^{wx}(y,z)] \tag{3.82}$$

where,

$$V_{c_1}^{wx}(y,z) = \sum_{k\in\mathcal{K}}\sum_{v\in\mathcal{V}} d_{kv}z_{kv} \tag{3.83}$$

$$V_{c_2}^{wx}(y,z) = \sum_{m,n\in\mathcal{K}} d_{mn}y_m y_n \tag{3.84}$$

$$V_{c_3}^{wx}(y,z) = \sum_{m,n\in\mathcal{K}} y_m y_n l^{con}(\sum_{v\in\mathcal{V}} z_{mv}) \tag{3.85}$$

$$V_{c_4}^{wx}(y,z) = \sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{X}} d_{jm}y_m \tag{3.86}$$

Therefore, considering the analysis in section 3.3.2.1 regarding the linearization of the objective function, we obtain the following MILP model for the SDN controller placement problem:

$$\text{Minimize} \sum_{k\in\mathcal{K}}\sum_{v\in\mathcal{V}} d_{kv}z_{kv} + \beta[\sum_{m,n\in\mathcal{K}} d_{mn}\theta_{mn} + \sum_{m,n\in\mathcal{K}}\sum_{v\in\mathcal{V}} l^{con}\phi_{mnv} + \sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{X}} d_{jm}y_m] \tag{3.87}$$

$$\text{subject to:} \quad (3.29), (3.30), (3.43) - (3.48) \tag{3.88}$$

$$z_k \in \{0,1\}, \quad \forall k \in \mathcal{K} \tag{3.89}$$

$$y_{jk} \in \{0,1\}, \quad \forall k \in \mathcal{K}, \quad \forall j \in \mathcal{X} \tag{3.90}$$

In the next theorem, we show that the function $V_c^{wx}(y,z)$ is sub-modular, and accordingly the $(1/2)$-approximation method can be used to approximately minimize it.

Theorem 10 *The cost function $V_c^{wx}(y,z)$ is supermodular.*

Proof. We will first show the supermodularity of each of the 4 terms. As usual, consider any $\mathcal{A},\mathcal{B} \in \mathcal{K}$ such that $\mathcal{A} \subseteq \mathcal{B}$ to be two sets of nodes selected to host SDN controllers by two different controller placement policies, where $k_v^{\mathcal{A}} =: \arg\min_{k\in\mathcal{A}} d_{kv}$ for all $v \in \mathcal{V}$, is the controller that node v is assigned to, by the assignment policy implied by the placement $\mathcal{A}$. $k_v^{\mathcal{B}}$ can be similarly defined for set $\mathcal{B}$. By introducing a new SDN controller to both of the policies at any arbitrary location $\kappa \in \mathcal{K} \setminus \mathcal{B}$ and substituting the parameters z_{kv}, d_{kv}, $k_v^{\mathcal{A}}$, and $k_v^{\mathcal{B}}$ with their counterparts in the statement of the proof of theorem 8, it will follow that $R_\kappa^{\mathcal{B}} \geq R_\kappa^{\mathcal{A}}$; hence $V_{c_1}^{wx}(y,z)$ is a supermodular function. The supermodularity of the function $V_{c_2}^{wx}(y,z)$ is immediate, since once adding a new controller to sets $\mathcal{A}$, and, $\mathcal{B}$, since the set $\mathcal{B}$ is bigger, the new controller has to communicate with more controllers compared to when it is added to set $\mathcal{A}$; therefore $R_\kappa^{\mathcal{B}} \geq R_\kappa^{\mathcal{A}}$. The function $V_{c_3}^{wx}(y,z)$, is modular. To see this we can change the order of summations in $V_{c_3}^{wx}(y,z)$ and write:

$$V_{c_3}^{wx}(y,z) = \sum_{v\in\mathcal{V}} \Big(\sum_{k\in\mathcal{K}: z_{kv}=0} l^{con} y_k \Big) \tag{3.91}$$

Now, observe that adding a new element to sets $\mathcal{A}$, and, $\mathcal{B}$, will only contribute l^{con} for each $v \in \mathcal{V}$, regardless of the size of the set. Therefore $R_\kappa^{\mathcal{B}} - R_\kappa^{\mathcal{A}} = 0$, and the function is modular. Similarly, $V_{c_3}^{wx}(y,z)$ is modular, since the marginal return by adding a new controller is proportional to the distance of that controller to its closest gateway, regardless of the cardinality of the set of the controllers. Therefore, given that modularity is a special case of supermodularity,

$V_c^{wx}(y,z)$ as the sum of supermodular and modular functions is supermodular and the assertion follows. ■

Given that in the SDN-enabled hybrid network, the SDN controllers perform the task of network management, any failures in the nodes or links of the network may disconnect the data and the control plane and block the way of the instructions from controllers to the SDN switches which in turn will result in the performance degradation of the hybrid network. Therefore, it makes sense to formulate a reliable controller placement problem to maximize the average reliability of the control paths. i.e.

$$U_c^{wx}(y,z) = \sum_{k\in\mathcal{K}}\sum_{v\in\mathcal{V}} r_{kv} z_{kv} \tag{3.92}$$

The MILP model is as follows:

$$\text{Maximize} \quad U_c^{wx}(y,z) \tag{3.93}$$

$$\text{subject to:} \quad (3.29), (3.30), (3.32) \tag{3.94}$$

$$z_{kv} \in \{0,1\}, \quad \forall k \in \mathcal{K}, \quad \forall v \in \mathcal{V} \tag{3.95}$$

$$y_k \in \{0,1\}, \quad \forall k \in \mathcal{K} \tag{3.96}$$

Next, we show that the function $U_c^{wx}(y,z)$ is submodular, and therefore, the algorithm mentioned in theorem 5 can be applied for approximately solving the corresponding MILP model.

Theorem 11 *The function $U_c^{wx}(y,z)$ is monotone and submodular.*

Proof. The proof is exactly similar to that of theorem 9 when z_{kv}, $R_\kappa^{\mathcal{A}}$, $R_\kappa^{\mathcal{B}}$, r_{kv}, $k_v^{\mathcal{A}}$, and $k_v^{\mathcal{B}}$ replace their counterparts in the statement of the proof. ■

Table 3.4: Network Topology Settings

Topology	Nodes	Links
Nsfnet	13	15
Sinet	13	18
Ans	18	25
Aarnet	19	24
Agis	25	32
Digex	31	35
Chinanet	42	86
Bell Canada	48	64
Tinet	53	89

Table 3.5: Failure Probability Settings

	P_v terrestrial nodes	P_e terrestrial links	$P_{e_{sg}}$ satellite links
Case 1	[0,0.05]	[0,0.02]	[0,0.02]
Case 2	[0,0.06]	[0,0.04]	[0,0.03]
Case 3	[0,0.07]	[0,0.06]	[0,0.04]
Case 4	[0,0.08]	[0,0.08]	[0,0.05]

3.3.3 Performance Evaluation

In this section, we evaluate the performance of the proposed gateway deployment and the controller placement methods. We first explain the evaluation setup and the simulation environment and then provide the corresponding results for each method.

3.3.3.1 Evaluation Setup & Metrics

Similar to [65], for our simulations, we utilize multiple real network topologies publicly available at the Internet Topology Zoo [59]. In addition to the previous topologies used in [65] we use 4 new networks to be able to compare the performance of our algorithms to that of the literature. The complete list of topologies that we have considered is listed in table 3.4. The lengths of network links are extracted from the Topology Zoo, based on which the corresponding

latency value for each link is computed similarly to [65]. To compute the shortest paths between each pair of network nodes, we adopted an implementation of the Yen's algorithm [72]. To facilitate the comparison of the algorithms with literature we apply the same approach as in [64] to compute the failure probability for each network component. We randomly generate the failure probabilities for terrestrial nodes, terrestrial links, and the satellite link, in 4 different settings. Table 3.5 lists the range of failure probabilities in different cases for each network component. For solving the MILP models we use CPLEX commercial solver, and conduct all the experiments on an Intel Xeon processor at 3.5 GHz and 16 GB of main memory. Unless otherwise stated, each experiment is repeated 100 times and the results are averaged over all runs.

To evaluate the performance of the proposed methods we use multiple metrics as follows:

- **Utility Function Value** is the overall value of the utility function provided by the solution of the JGCP problem.
- **Cost Function Value** is the overall value of the cost function provided by the solution of the JGCP problem.
- **Average node-to-facility Latency** is the average node-to-gateway or node-to-controller latency experienced by the terrestrial network.
- **Average Network node-to-facility Reliability** is the average reliability of the shortest path between each terrestrial node and a gateway or a controller.
- **Number of Deployed Facilities** is the total number of satellite gateways or controllers that are deployed on the terrestrial network as a result of the JGCP.

- **Solver Run-time** is the amount of time it takes for the solver to generate the solution to the JGCP problem.

3.3.3.2 Numerical Results

We conduct 4 experiments to evaluate the performance of our algorithms for problems discussed in section 9.3. The first two experiments correspond to the gateway placement problem.

In *Exp. A*, the latency-oriented gateway placement problem is considered with the objective of jointly minimizing the number of deployed gateways and the average network-to-gateway latency. We discuss the trade-off between the number of deployed gateways and the resulting average network-to-gateway latency.

Exp. B aims at maximizing the average reliability of node-to-controller paths, while the number of gateways to be deployed is limited not to exceed a given maximum. We compare our results with the algorithms mentioned in [64], and also show how our method works under different settings for network failure probability.

The next two experiments address the SDN controller placement problem. *Exp. C* evaluates the performance of our method in jointly minimizing the synchronization cost between the SDN controllers, and the average node-to-controller latency, while *Exp. D* considers the reliability-oriented SDN controller placement problem.

Fig. 3.5 compares the performance of the $(1/2)$-approximation greedy algorithm with that of the optimal MILP approach. In particular, apart from the overall objective value, the latency of the network nodes to the closest gateway is depicted in the graph, together with the number of gateways used in the resulting deployment policy. It is interesting to observe that the overall

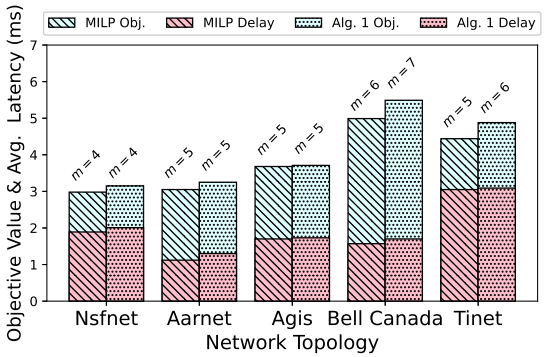


Figure 3.5: Exp A. Objective function & average node-to-gateway latency

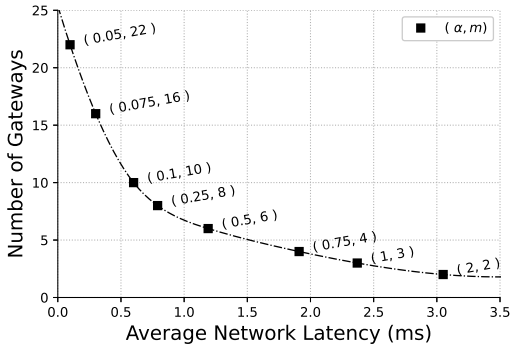
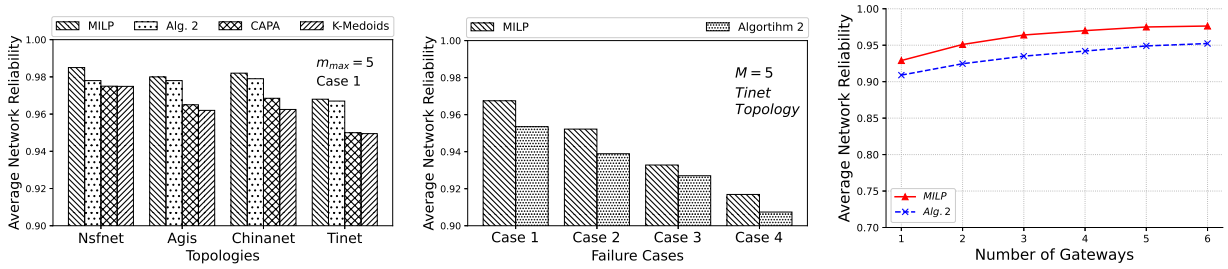


Figure 3.6: Exp A. Impact of α on the gateway deployment policy

value for the objective function for the approximation method remains within 10% of the optimal MILP-based method. Moreover, the terrestrial nodes only experience at most 5% increase in the tolerated latency to reach their closest gateway.

Fig. 3.6 shows how the trade-off between the two terms of the objective function can be controlled by changing the value of the parameter α i.e. how tuning α can balance the emphasis of the gateway deployment policy on minimizing the number of gateways used or minimizing the average node-to-gateway latency. For instance, lowering the value of α from 2 to 1 will almost half the amount of average node-to-gateway delay at the expense of doubling the number of deployed gateways.

Fig. 3.7a shows the result of running *Exp. B* when a maximum of 5 gateways is allowed to be deployed in the network. We benchmark the performance of Algorithm 3, in terms of average node-to-gateway reliability, against that of the approaches mentioned in [64] and report the results. Similar to [64], we average our results when the failure probabilities of the network components follow *Case 1* of table 3.5. Our experiments show that all the methods perform reasonably well while our approach results in a performance that is slightly closer to the optimal



(a) Different topologies (b) Different failure conditions (c) Different number of gateways

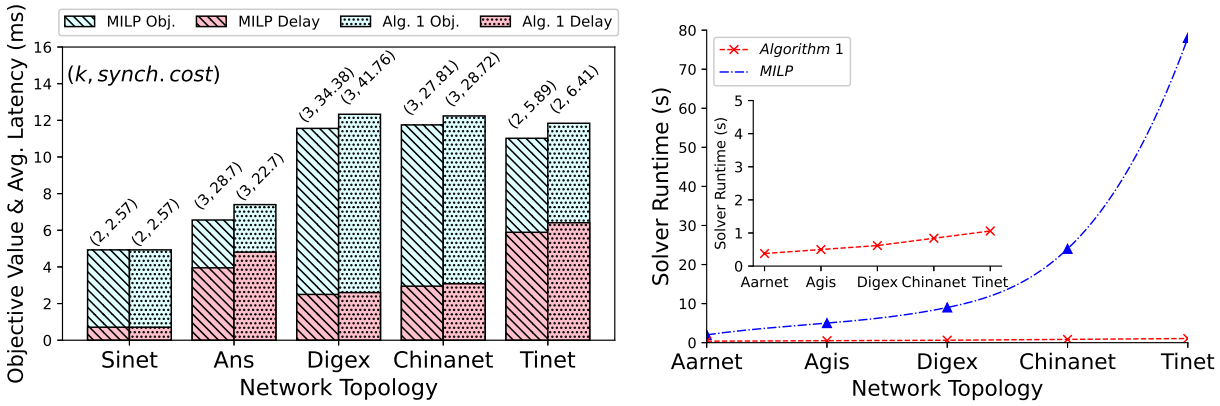
Figure 3.7: Exp B: Average node-to-gateway reliability

MILP.

For different failure cases, we run *Exp. B*, under 4 different scenarios for network failure probability. In Fig. 3.7a, we show the result of changing the failure settings on *Tinet* topology, while the maximum number of deployed gateways cannot exceed $m_{max} = 5$. Naturally, when the network components are more prone to failure, the average reliability decreases. It is observed that under all scenarios, the solutions generated by our approach remain close to the optimal solution.

Fig. 3.7c, shows the effect of changing the maximum number of gateways allowed in the gateway deployment policy. It is observed that our method follows the performance of the optimal model by at most a 3% gap in terms of maximizing the average network reliability.

Fig. 3.8a, depicts the performance of running Exp. C, where the approximate method using Algorithm 2 for SDN controller placement is bench-marked against the MILP-based optimal one. Our results indicate that the performance of the approximate approach remains within 10% of optimality both in terms of the value of the objective function and the average node-to-controller latency. The number of controllers placed and the corresponding synchronization cost are annotated on top of each topology studied. We further note that the approximate method



(a) Obj. value and average node-to-controller latency

(b) Comparison of time complexity

Figure 3.8: Exp C. Jointly minimizing the synch. cost and the average node-to-controller latency

maintains the same number of controllers as in the optimal solution.

Fig. 3.8b illustrates the comparison between the time complexity of the MILP model (using CPLEX with branch-and-cut) and the submodular optimization algorithms. It is observed as the scale of the network increases the solver runtime for solving the MILP increases exponentially, while the submodular optimization evolves linearly in terms of time which further confirms the theoretical result for its time complexity. For *Tinet* topology, the MILP solver takes about $70-80$ seconds to find the optimal solution while Algorithm 2 can run 100 times in less than 2 seconds.

Finally, Fig. 3.9 depicts the result of running Algorithm 3, for maximizing the reliability of nod-to-controller paths. The maximum number of controllers placed is enforced to not exceed $k_{max} = 5$. As indicated by our results the approximate method performs reasonably well by generating near-optimal results.

Overall, the performance evaluation experiments and analysis reveal that for all discussed variants of the satellite gateway deployment and the SDN controller placement, the approximate methods discussed in this section can provide solutions that are very close to optimal ones in

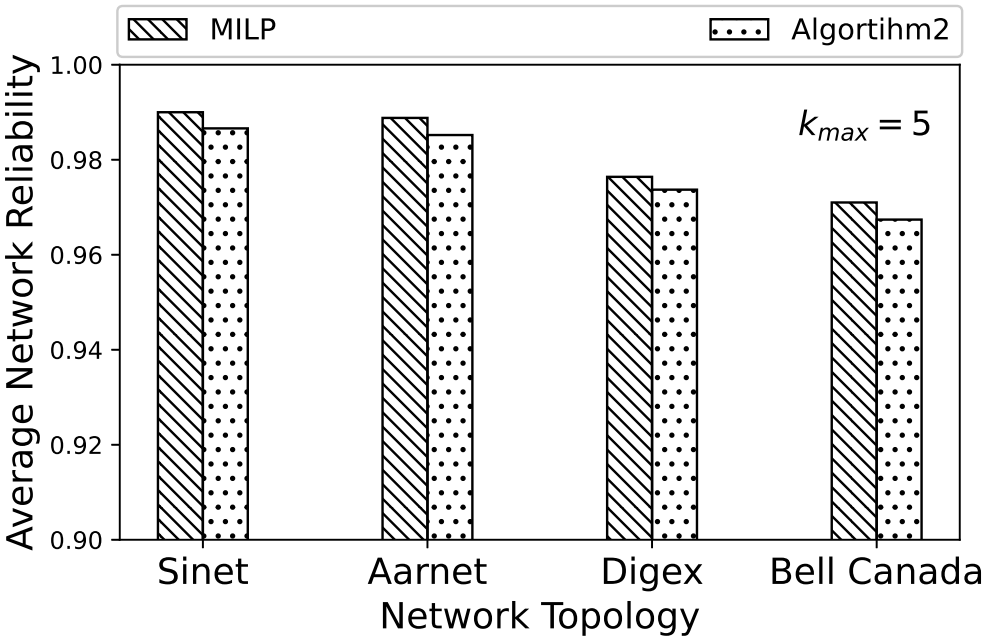


Figure 3.9: Exp D. Average node-to-controller reliability

terms of accuracy, while extensively, lowering the corresponding time complexity.

3.4 Related Work

Although the gateway placement problem over general networks is studied extensively, the satellite gateway deployment and the SDN controller placement problems in SDN-enabled SAGINs are relatively new problems. In [43], the authors have very well explained why the traditional gateway placement methods are not suitable to satisfactorily address these problems. Over the past few years, there have been interesting works concerning these two problems in different settings and from several perspectives; The gateway placement with GEO satellites has received a lot of attention while the attention towards LEO constellations has increased a lot only in the past few years. Some works have primarily considered the optimization of the ground segment by optimizing the deployment of ground stations, while some other works were only

concerned with the deployment of gateways and controllers on the aerial or space layer. Several papers have studied gateway diversity for optimizing HTS architectures. These works typically aim at optimizing the smart gateway configuration [73], gateway assignment [74], paving the way towards optical feeder links, and dealing with atmospheric turbulence [75].

A line of research target gateway deployment and SDN controller placement in LEO constellations. In [76], the authors have proposed a genetic algorithm for the deployment of multiple gateways in a large constellation, taking into account the gateway-satellite connectivity, hop count, and the potential location of gateways, to achieve satisfactory performance in terms of load balance, latency, and traffic peak, towards an optimized layout. A joint gateway assignment and routing method is proposed in [77] with the goal of congestion avoidance in mega-constellations. For SDN controller placement, the research works differ mostly in the location of placing the controllers each of which provides some benefits and some shortcomings. Some papers decide to place the controllers on the ground segment, some place them on the LEO SDN-enabled satellite switches [78] [79], and some on the GEO layer [80], while some other works propose hierarchical controller architectures comprising ground stations, LEO, MEO, and GEO-layer controllers [81], [82], [83]. In [82], [78], and, [79], the authors consider the controller placement problem in both static and dynamic modes, where in the former the controller placement and satellite-to-controller assignments remain unchanged, while in the latter the number of the controllers, their locations, and therefore the assignments vary concerning change in demand and traffic pattern over time. Further in [78], and [79], the flow setup time is adopted as a metric which makes the problem statement realistic as optimizing the flow setup time is a major concern in SDN-enabled networks.

Within the context of SAGIN, various objectives can be sought when solving the satellite

gateway deployment problem that may be inspired by different requirements such as cost or latency minimization, reliability awareness, route optimization, load balancing, etc. The problem is usually formulated in the literature as a combinatorial optimization problem and is solved exactly or approximately. The authors in [62], aim at minimizing the average user-to-gateway latency over the network, given reliability constraints, however, they only take into account the shortest paths from user to gateway, and therefore no routing scheme is provided. They solve the problem using a Particle Swarm Optimization (PSO) algorithm and compare their approach to the brute-force greed search method in terms of time complexity. In [64], only maximization of reliability is considered, where the authors use a simple clustering algorithm and compare their method with the result of optimal enumeration to justify the lower time complexity at the expense of sub-optimal results.

The joint gateway deployment and controller placement in SAGIN have also received increasing attention over the past few years. The authors in [66], formulate a joint deployment of satellite gateways and SDN controllers to maximize the average reliability with hard constraints on user-to-satellite delay. They propose an iterative approach that relies on simulated annealing and clustering, where in each iteration first the current gateway placement policy and then the controller placement are updated towards the convergence. In [84], the same problem under similar settings and with similar objectives has been taken with the only difference that the simulated annealing approach from [66] is augmented with a portioning phase (separately w.r.t gateways & controllers) to render several sub-problems of smaller size. Finally, in [85], several meta-heuristic approaches namely, simulated annealing, double simulated annealing, and genetic algorithm for the same problem have been considered and their performance is compared.

Despite incorporating many interesting ideas, most of the mentioned works assume that the

optimal number of equipment to deploy is known as a-priori, do not consider the routing overhead and gateway load, and do not employ any realistic scheme for the control plane. Although these simplifications may render more tractable problems allowing for more interesting and smart solutions, but are prone to become unusable in some real-world scenarios.

In [63], this issue is mitigated to some extent as the authors take into account the capacity constraints of satellite links and formulate a problem towards maximizing the average reliability which they solve using a greedy algorithm. In [65], a joint gateway placement and traffic routing is formulated as an instance of the capacitated facility location-routing problem which does not make any assumption on the optimal number of gateways and also provides a routing solution that does not contradict the capacity constraints and is not limited to the shortest paths using multi-commodity flow allocation.

3.5 Conclusions

we introduced the joint satellite gateway placement and routing problem over an ISTN, for facilitating the terrestrial-satellite communications while adhering to propagation latency requirements, in a cost-optimal manner. We also balanced the corresponding load between selected gateways. To yield a polynomial solution time, we relaxed the integer variables and derived an LP-based rounding approximation for our model. Further, we considered several variants of the joint satellite gateway and SDN controller placement (JGCP) in SDN-enabled 5G-Satellite hybrid networks. We separately considered the average reliability of assignments and the latency & overhead of the communications between SDN controllers as the objectives of the problem. We proposed a MILP model and a submodularity-based optimization framework for

solving the problem and then evaluated the performance of the proposed methods using simulation. Our results confirm the effectiveness of our approach both in terms of results accuracy and time complexity.

Within the framework of 5G-Satellite integration, the SDN controller placement problem for LEO constellations becomes more important due to the more frequent need for hand-offs between the satellite switches; and at the same time is more challenging due to the dynamically changing network topology and potentially large number of SDN controllers required in both the terrestrial and the space layer. Moreover, instead of placing the satellite gateways on the terrestrial nodes, aerial platforms can be a choice leading to a cross-layer network design problem. This approach provides more flexibility to adapt to the evolving network topology, but becomes more challenging in terms of the more complicated flow routing decisions and also the partly-intermittent nature of the aerial layer communications. These two problems remain within the topics of our future research.

Chapter 4: Trust-aware Federated Learning: A Multi-agent UAV-enabled Networks Scenario

4.1 Overview

AI systems, especially those employing machine learning (ML) and deep learning (DL) models, typically undergo training on data that belongs to and is managed by different stakeholders distributed across separate data repositories. To tackle the challenge of safeguarding data privacy in AI applications, Federated Learning (FL) [86] was introduced as an innovative distributed ML approach that facilitates collaborative model development among agents such as organizations, mobile phones, sensors or drones, while safeguarding sensitive data privacy by ensuring that each participant retains control over their local data. In addition, FL enhances communication efficiency by exchanging the updated model parameters of participating agents instead of their dataset.

Two main FL setups have been proposed in the literature, namely centralized FL (CFL) and decentralized FL (DFL). In the CFL, individual devices carry out local training and transmit their updated local models to a central server (e.g. a server at the network edge or co-located with gNB in 5G networks), all while keeping their private raw data secure. The server aggregates parameters from local models to refine the global FL model before redistributing the enhanced

model back to the devices. The process is repeated for a predefined number of rounds or until a desired accuracy level is achieved. As a leading algorithm in this setting, *Federated Averaging (FedAvg)* [87] runs Stochastic Gradient Descent (SGD) in parallel on a small subset of the total agents and averages the local updates only once in a while.

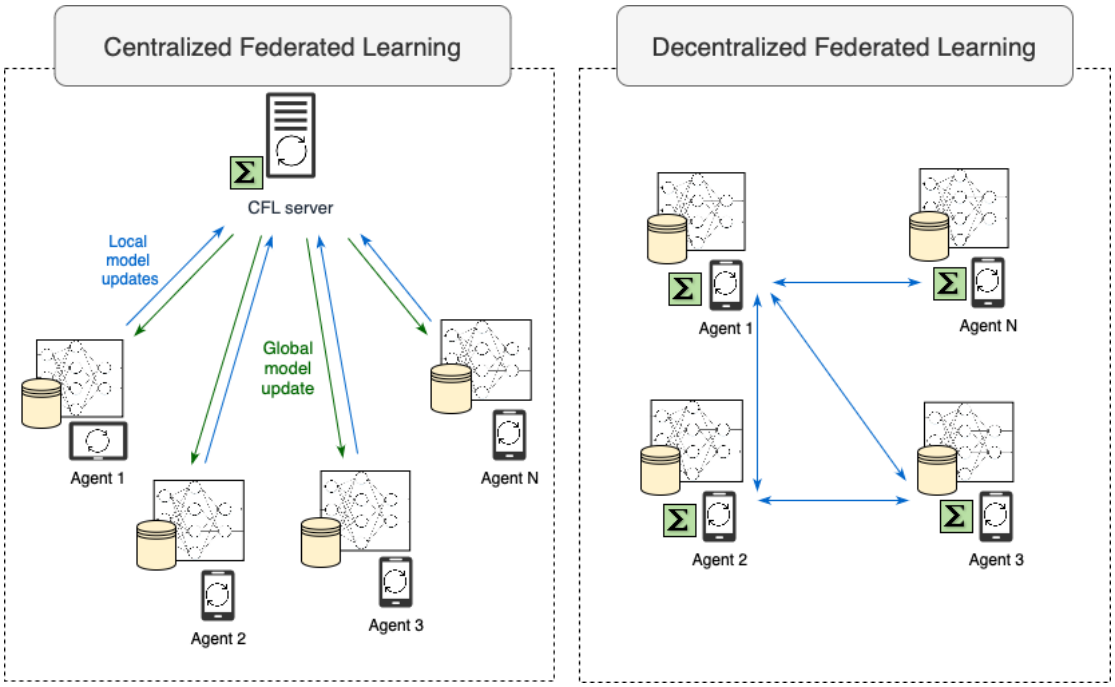


Figure 4.1: Centralized FL vs. Decentralized FL

Despite enhancing the privacy of the network agents and the efficiency of the communications, CFL is susceptible to security vulnerabilities and may encounter challenges meeting the security prerequisites in real-world applications and use cases. The central server, which distributes, aggregates, and updates ML models, are attractive target for malicious actors seeking to manipulate models to generate biased outcomes or achieve specific results [88]. Even with a trusted central server, security issues such as data poisoning, model poisoning, and adversarial attacks, can introduce bias in the model predictions. Lack of scalability, connectivity guarantees, and the existence of single-point-of-failure in CFL settings necessitates an FL paradigm that does

not rely on a central server for model aggregation and distribution. Moreover, next-generation networks are expected to be enhanced by new forms of decentralized and infrastructure-less communication schemes and device-to-device (D2D) multi-hop connections such as in UAV-aided networks [89]. Within this scope, novel approaches are required to address decentralized (server-less) FL (DFL). While several research activities have focused on distributed learning algorithms [90], due to the special features of an FL setup in which the data is generated locally and remains decentralized and because of the communication efficiency considerations, many existing approaches developed for distributed learning do not apply to DFL. Authors in [91] extend the CFL approach for massively dense IoT networks where the agents perform training steps on their local data using SGD and consensus-based methods. At each consensus step, agents transmit their local model update to their one-hop neighbors. Each agent fuses the received messages from its neighbors and then feeds the result to SGD. Fig. 4.1 shows the architectures of CFL and DFL settings. Similar to CFL, serverless FL faces major security challenges such as model and data poisoning attacks from compromised agents.

The fact that the information is crowd-sourced by the FL agents in a distributed FL, highlights the need to establish trust relationships between the FL agents. More specifically, apart from ensuring the security of communications between the FL agents, answering the following questions is of paramount importance: (i) whether an agent refuses to share its information with the FL process due to privacy concerns or conflict of interest. (ii) whether an agent manipulates the received data before processing. (iii) whether an agent intentionally or unintentionally, shares incorrect information with the rest of the network? etc. [92] [93]. In other words, it is essential to establish to what extent each agent of the network and its model updates can be trusted.

To this end, we propose *trust-aware FL* mechanisms for both CFL and DFL systems. In the

proposed schemes, trust is interpreted as a relation between different network entities that may interact or collaborate in groups toward achieving various specific goals. These relations are set up and updated based on the evidence generated when the agents collaborate within a previous protocol. If the collaboration has been contributive towards the achievement of the specific goal (positive evidence), the parties accumulate their trust perspective towards one another, otherwise (negative evidence), trust will decrease between them. Trust estimates have input to decisions such as access control, resource allocation, agent participation, and so on. The method by which trust is computed and aggregated within the network may depend on the specific application. In [93], the authors enumerate the central differences in the terminology of trust computation and aggregation. In this chapter, we propose an attack-tolerant consensus-based FL algorithm by incorporating the trust concept into the consensus step [94].

The integration of UAVs into modern communication networks has ushered in a new era of connectivity and data-driven capabilities. UAVs have transcended their traditional roles in surveillance and reconnaissance to play pivotal roles in enhancing the coverage, capacity, and intelligence of communication networks and to become integral components of various applications, including disaster management, precision agriculture, environmental monitoring, and even last-mile delivery services. As these UAV-enabled networks continue to expand, the need for efficient, secure, and intelligent data processing and decision-making becomes paramount. UAV-enabled networks typically encompass a diverse array of edge devices, including UAVs, ground stations, sensors, and other IoT devices, all generating vast amounts of data [95]. Managing and harnessing this data is a non-trivial task, often constrained by limited bandwidth, energy, and computational resources, particularly in remote or dynamic environments. Federated Learning (FL) offers a transformative paradigm for distributed machine learning. Instead of

centralizing data in a single location, it allows training machine learning models collaboratively across decentralized devices while keeping data localized. This decentralized approach aligns perfectly with the inherent characteristics of UAV-enabled networks, where data privacy, latency, and adaptability are paramount concerns. The potential benefits of applying FL to UAV-enabled networks are far-reaching. FL enables UAVs to continuously improve their performance and intelligence by collectively learning from data generated across the network. Moreover, FL enhances privacy, as sensitive data remains on-device, reducing the risk of data breaches and ensuring compliance with data protection regulations. Additionally, it can significantly reduce the communication overhead associated with transmitting large volumes of data to centralized servers, thereby conserving valuable bandwidth resources and mitigating latency issues.

4.2 Trust Aggregation Model

4.2.1 Trust Aggregation Framework

In this section, we present two schemes for propagating and aggregating the trust estimates within a network of CPS devices (4.2). The first scheme corresponds to the case where there is no central entity involved in estimating the trustworthiness of the network agents, and the nodes participate in direct computation of trust to obtain local trust estimates on the other peers, using the locally-available first-hand evidence they have gathered, the recorded history they have stored from the past observations, and the knowledge they obtain by sensing the environment. Once all agents form their local views, they will participate in the local exchange of their local trust estimates to form more accurate global values for the trustworthiness of the networked agents. Then, the obtained global trust model can be used in the corresponding trust-aware applications.

Within the second scheme, however, the global trust values are obtained in an indirect fashion. There exists a central trusted party that is constantly monitoring the network and communicating with the CPS agents to gather evidence on their state. The CPS agents may share their local view on their neighbors with the central entity which may be used in computing the trust estimates by the central entity. Once the central party calculates the trust values of the CPS agents using the information it has gathered, it will push the relevant information to each agent. The calculated trust values can be used by the central entity to perform centralized trust-aware decision-making or can be used by each agent to participate in local or distributed trust-aware protocols. In what follows, we will formalize the above discussion to mathematically model the processing, propagating, and aggregation of the trust values. The components of our model mostly follow and rely on the discussion in [96].

We model the network of agents at time instance k, as an undirected graph given by $\mathcal{G}^{(k)} = (\mathcal{N}^{(k)}, \mathcal{L}^{(k)})$ where $\mathcal{N}$ is the set of nodes and for $n, m \in \mathcal{N}^{(k)}$, $\mathcal{L}^{(k)}$ contains all links $(m, n)^{(k)}$ where agents m, and n can communicate with one another at time instance k. We denote this graph as the *communication graph* at time instance k. Let $\mathcal{N}_i^{(k)}$ be the set of neighbors of node i at time step k. Apart from the communication relationship, we define *local trust* relationships between nodes $i, j \in \mathcal{N}^{(k)}$. Let $\tau_{ij}^{(k)}$, and $t_{ij}^{(k)}$ be the local and global view of node i on trustworthiness of the node j at time instance k in respective order. We may ignore the index k whenever doing so does not lead to confusion.

4.2.2 Local Trust Model

To formalize the definition of local trust, let us define $X_{ij}^{(k)}$ to be a random variable denoting the reputation that node j has in the perspective of node i in time instance k. $X_{ij}^{(k)}$ follows a Beta distribution with parameters $\alpha_{ij}^{(k)}$, and $\beta_{ij}^{(k)}$. The beta distribution is often used for modeling reputation or trustworthiness because it can effectively capture uncertainty and variability in reputation scores or ratings. In the context of reputation systems and online reviews, Beta distribution is a popular choice due to its flexibility and suitability for representing the distribution of user ratings or opinions.

Define $r_{ij}^{(k)} = \alpha_{ij}^{(k)} - 1$, and $s_{ij}^{(k)} = \beta_{ij}^{(k)} - 1$ that determine the number of times up to round k, that node j's behavior is benign and malicious in the perspective of node i, in respective order. The details of how $r_{ij}^{(k)}$, and $s_{ij}^{(k)}$, are obtained depending on the specific scenario and are to be explicitly mentioned in the next sections. We let $\tau_{ij}^{(k)}$ to be precisely the expected value of the reputation random variable in the Beta system $X_{ij}^{(k)}$. Formally, we have:

$$f_{X_{ij}^{(k)}}(x; \alpha_{ij}^{(k)}, \beta_{ij}^{(k)}) = \left(\frac{\Gamma\left(\alpha_{ij}^{(k)} + \beta_{ij}^{(k)}\right)}{\Gamma\left(\alpha_{ij}^{(k)}\right)\Gamma\left(\beta_{ij}^{(k)}\right)}\right) \cdot \left(x^{\alpha_{ij}^{(k)}-1}\left(1-x\right)^{\beta_{ij}^{(k)}-1}\right) \tag{4.1}$$

$$\tau_{ij}^{(k)} = \mathbb{E}\left[X_{ij}^{(k)}\right] = \frac{r_{ij}^{(k)} + 1}{r_{ij}^{(k)} + s_{ij}^{(k)} + 2} \tag{4.2}$$

Intuitively, the evolution of r, and s parameters needs to be in a way that the more recent information receives more relative importance compared to the older ones. Therefore, we define $0 < \rho_1 < \rho_2$ as forgetting factors to control the balance between old and new terms.

$$r_{ij}^{(k+1)} = \rho_1 r_{ij}^{(k)} + I_{ij}^{(k+1)} \tag{4.3}$$

$$s_{ij}^{(k+1)} = \rho_2 s_{ij}^{(k)} + 1 - I_{ij}^{(k+1)}, \tag{4.4}$$

The function $I_{ij}^{(k+1)} \in [0, 1]$ models the instantaneous perspective of node i on the behavior of node j in $(k+1)^{th}$ round.

4.2.3 Global Trust Model

At each instance, k, within the local trust model, each node i computes its local trust for all nodes $j \in \mathcal{N}_i$ in the communication graph. To make more accurate estimates, node i will need to take into account the opinions of other network nodes who have first-hand evidence of one node j's behavior. Following the approach in [96], node i computes iteratively its global trust estimate for node j, i.e. $t_{ij}^{(k)}$ using the opinions of its neighbors as:

$$t_{ij}^{m} = \begin{cases} 1 & \text{if } i = j \\ \sum_{l \in \mathcal{N}_i, l \neq j} w_{il} t_{lj}^{m-1} & \text{if } i \neq j \end{cases} \tag{4.5}$$

where $w_{il} = \frac{\tau_{il}}{\sum_{l \in \mathcal{N}_i, l \neq j} \tau_{il}}$. In other words, node i pays more attention to the opinions of those of its neighbors whom it trusts more. We note again that the global trust computation is an iterative process that is going to be embedded in each iteration of the trust-aware protocol. Therefore, to avoid any confusion we have used the iteration counter m for this process. Here, we have dropped the superscript k as we assume the value of local trust remains constant within the loop of computing the global trust.

In the next two sections, we will show how the above trust framework can be used in practice to enhance the security of the CPS-based protocols. We will select as case studies two recent challenging problems from the state of the art, and consider two corresponding well-known algorithms for them, where security is a big challenge in these protocols. We will then augment those protocols with our trust inference and aggregation framework and show how trust can be used to improve the security of such protocols.

4.3 Trust-aware Federated Learning

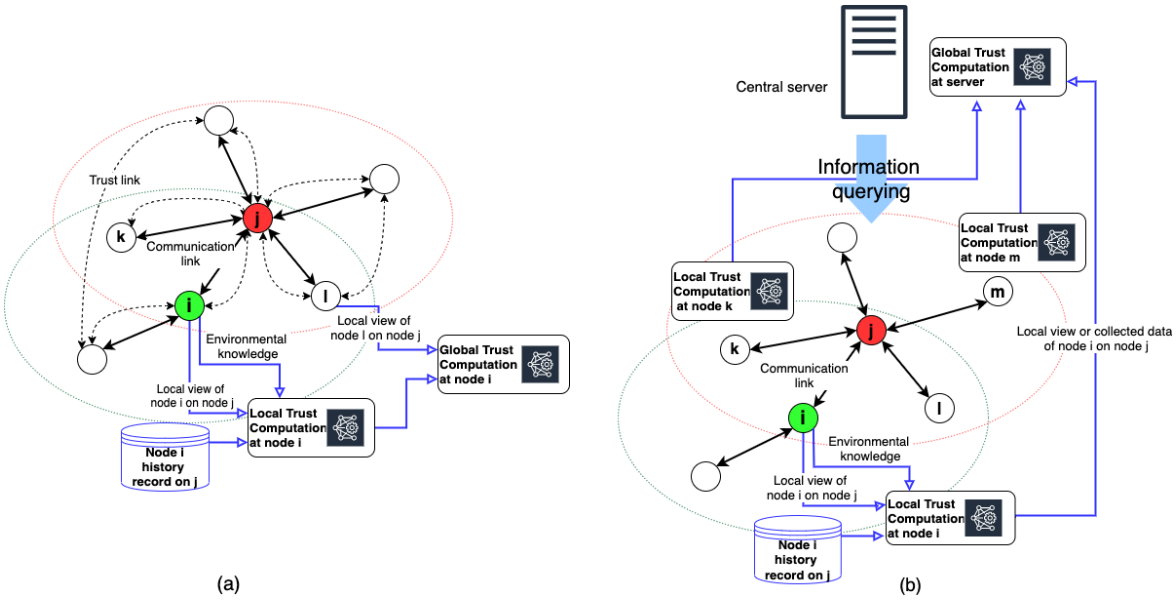


Figure 4.2: Trust aggregation framework in (a) decentralized and (b) centralized regimes

In this section, we present our proposed frameworks, trust-aware CFL (TCFL) and trust-aware DFL (TDFL) in detail. As discussed in previous sections, in the next-generation networked systems with high scale and heterogeneity, it is very impractical for the edge devices (sensors, drones, mobile phones, etc.) to transmit their collected data to a remote data center for centralized machine learning tasks, due to the limited communication resources and privacy constraint.

Therefore, the recent trend of high-stake applications such as target recognition in drones, augmented/virtual reality, autonomous vehicles, etc. requires a novel paradigm change calling for distributed, low-latency, and reliable ML at the network edge (referred to as edge ML [97]). FL is a promising approach to tackle the above challenges. However, there exist several security issues in an FL mechanism arising mainly from the participation of a fleet of possibly unreliable or compromised devices during the training time. According to [98], attacks in FL can be classified into two categories: (i) *data inference attacks*, in which the adversary tries to infer information about the users' private data, and (ii) *model performance attacks*, which includes data poisoning, update poisoning, and model evasion attacks. In data poisoning, the adversary subverts the user's dataset in the local training. The goal of the adversary in update poisoning is to alter the model updates transmitted to the central server. The model evasion attack refers to the data alternation at the inference time. In this chapter, we focus on the second type of attack. In the following, we first elaborate CFL and DFL architectures and then discuss the question of evaluating the trust decision $I_{ij}^{(k)}$ introduced in section 4.2 for TCFL and TDFL.

4.3.1 Trust-aware Centralized Federated Learning

In CFL, each agent has a local dataset that is utilized to compute the updates to the current global model. The local updates (gradients) are then aggregated and fed back to the participating agents by a central server. This process is repeated until a desired accuracy level or a specified number of rounds is achieved. Each iteration (communication round) includes a local training phase resulting in the generation of local model parameters and a communication phase to transmit local models to the central server as the following steps explain:

- **User selection**: The central server selects a set of users at each communication round, considering different factors such as available bandwidth, energy budget, computing power, etc.

- **Model broadcast**: The selected users download the most recent global model.

- **Local training**: Each selected user performs local computation and generates an update to the model parameter using the local data.

- **Model Aggregation**: The central server aggregates the collected local model updates and then generates a global model.

In the following, we first describe the federated optimization problem implicit in CFL and then explain our proposed trust-aware mechanism. Consider a network consisting of a set $\mathcal{N}$ of N agents and one central server that collaboratively trains an ML model. Each agent $i \in \mathcal{N}$ has D_i data samples and the total number of samples is D. The dth sample is denoted by $(\mathbf{x}_d, \mathbf{y}_d)$ where $\mathbf{x}_d \in \mathbb{R}^{D_{in} \times 1}, \mathbf{y}_d \in \mathbb{R}^{D_{out} \times 1}$, $d = 1, ..., D$. We assume that $\mathcal{P}_i$ is the set of indexes of data points on agent i and the data collected by different agents have the same distribution (**IID** assumption). The local dataset $\mathcal{D}_i = \{(\mathbf{x}_d, \mathbf{y}_d), d \in \mathcal{P}_i\}$ is used to train a local model $\mathcal{M}_i$ parameterized by $\mathbf{w}_i$. Let $\mathcal{M}$ and $\mathbf{w}$ be the global model and the global parameter vector. The objective of the FL training is to minimize:

$$F(\mathbf{w}) = \sum_{i \in \mathcal{N}} q_i F_i(\mathbf{w}) = \sum_{i \in \mathcal{N}} q_i \frac{1}{D_i} \sum_{d \in \mathcal{P}_i} f(\mathbf{w}, \mathbf{x}_d, \mathbf{y}_d) \tag{4.6}$$

where $f(\mathbf{w}, \mathbf{x}_d, \mathbf{y}_d)$ is the loss function capturing the accuracy of the FL model, and q_i is the weight of the ith device such that $q_i \geq 0$ and $\sum_i q_i = 1$. The objective of FL is to minimize

(4.6). As a leading algorithm in this setting, *Federated Averaging (FedAvg)* [87] runs Stochastic Gradient Descent (SGD) in parallel on a small subset of the total agents and averages the local updates only once in a while. Although *FedAvg* is shown to stabilize the convergence and ensure privacy to a great extent, since the central server has limited control over the behavior of the participating agents, it is vulnerable to adversary behaviors such as poisoning attacks as shown in [99]. To address this issue, we explore the incorporation of trust into *FedAvg*. The proposed algorithm is referred to as *Trust-aware CFL (TCFL)* and is described in the pseudo-code of Algorithm 4.

Let $t_{Si}^{(k)}$ be the opinion of the central server on the trustworthiness of agent i at round k. The server obtains the value of $t_{Si}^{(k)}$ based on the model described in section 4.2. Moreover, the trust evaluation method proposed for FL setups is discussed in section 4.3.3. We denote by c, μ_k and B the fraction of chosen agents at each round of algorithm, the learning rate, and the batch size respectively. At each round of *TCFL*, the server first selects $n = cN$ number of agents randomly. This is enforced by the fact that due to the large number of agents and the intermittent and bandwidth-constrained communications between the agents and the central server particularly in the case of wireless communications, it is not practical for all agents to participate in the model update at all rounds. The selected agents then can download the current global model and start the local training process denoted by the *ModelUpdate* procedure. Each agent i transmits a message to the server, denoted by $\mathbf{m}_i^{(k)}$. For a benign agent, $\mathbf{m}_i^{(k)} = \mathbf{w}_i^{(k)}$, while an adversary sends a message different from the update computed by running SGD on its local data. Finally, the server aggregates the local model updates received from the selected agents according to their trustworthiness and their dataset size reflected as the coefficients in step 7.

Algorithm 4 *Trust-aware CFL*

Input: $n = c.N$, $\{q_i, i \in \mathcal{N}\}$, μ_k, B

Initialize $\mathbf{w}_i^{(0)}$
for each round $k = 1, 2, ...$ **do**
  Server selects a subset S_k of n agents at random (agent i is chosen with probability q_i)
  Server transmits $\mathbf{w}^{(k)}$ to all (chosen) agents
  Each agent $i \in S_k$ computes $\mathbf{w}_i^{(k+1)} = ModelUpdate(\mathbf{w}^{(k)}, \mu_k)$ and sends a message $\mathbf{m}_i^{(k+1)}$ to the server
  Server updates the trust values for each agent i: $t_{Si}^{(k+1)} \leftarrow ComputeTrust(i, \{\mathbf{m}_i^{(k+1)}, i \in \mathcal{N}\})$
  Server aggregates the local updates: $\mathbf{w}^{(k+1)} \leftarrow \sum_{i \in S_k} \frac{D_i t_{Si}^{(k+1)}}{\sum_{i \in S_t} D_i t_{Si}^{(k+1)}} \mathbf{m}_i^{(k+1)}$
end for
$ComputeTrust(i, \{\mathbf{m}_i^{(k)}, i \in \mathcal{N}\})$:
  Server computes $I_i^{(k)}$ from (4.8) or (4.11)
  Server computes $t_{Si}^{(k)}$ according to (4.2)
$ModelUpdate(\mathbf{w}^{(k)}, \mu_k)$:
  Initialize $\mathbf{\Psi}_{i,k} \leftarrow \mathbf{w}^{(k)}$
  $\mathcal{B}$ = mini-batch of size B
  for $b \in \mathcal{B}$ **do**
    $\mathbf{\Psi}_{i,k} \leftarrow \mathbf{\Psi}_{i,k} - \mu_k \nabla f(\mathbf{\Psi}_{i,k})$
  end for
  Return $\mathbf{w}_i^{(k)} \leftarrow \mathbf{\Psi}_{i,k}$

4.3.2 Trust-aware Decentralized Federated Learning

The main drawbacks of CFL are the scalability, connectivity, and single-point-of-failure issues. Moreover, next-generation networks are expected to be enhanced by new forms of decentralized and infrastructure-less communication schemes and device-to-device (D2D) multi-hop connections such as in UAV-aided networks [89]. Within this scope, novel approaches are required to address decentralized (serverless) FL.

In [91], the CFL approach is extended for massively dense IoT networks that do not rely on a central server. Agents perform training steps on their local data using SGD and consensus-based methods. At each consensus step, agents transmit their local model update to their one-hop neighbors. Each agent fuses the received messages from its neighbors and then feeds the result to SGD. We propose an attack-tolerant consensus-based FL algorithm by incorporating the trust concept into the consensus step. The proposed approach is given in Algorithm 5. In particular, let $t_{ij}^{(k)}$ denote the trustworthiness of agent $j \in \mathcal{N}_i$ evaluated at i. Similar to the centralized trusted FL approach, in step 7, agent i aggregates the received updates from its neighbors with the weights of $\frac{D_j t_{ij}^{(k)}}{\sum_{j \in \mathcal{N}_i} D_j t_{ij}^{(k)}}$, i.e. the neighbors of i with higher trust values contribute more to the aggregated model at i. Then, each agent updates its model independently using SGD on its local data.

4.3.3 Trust Evaluation Method

In this section, we elaborate on the methods that are used in evaluating the trustworthiness of the network agents. These methods are embedded into the trust evaluation scheme used in the trust model in section 4.2. Throughout the FL protocols, in each iteration, the model parameters or their updates, corresponding to the local models of the agents are communicated within the

Algorithm 5 *Trust-aware DFL*

Input: $\mathcal{N}_i$, μ_k, ϵ_k

Initialize $\mathbf{w}_i^{(0)}$
for each round $k = 1, 2, ...$ **do**
  Agent i receives the messages $\{\mathbf{m}_j^{(k)}\}_{j \in \mathcal{N}_i}$
  Agent i updates the trust values for all its neighbors:
  $t_{i,j}^{(k)} \leftarrow ComputeTrust(i, j, \{\mathbf{m}_j^{(k)}, j \in \mathcal{N}_i\})$
  $\mathbf{\Psi}_{i,k} \leftarrow \mathbf{m}_i^{(k)}$
  for each agents $j \in \mathcal{N}_i$ **do**
    $\mathbf{\Psi}_{i,k} \leftarrow \mathbf{\Psi}_{i,k} + \epsilon_k \frac{D_j t_{i,j}^{(k)}}{\sum_{j \in \mathcal{N}_i} D_j t_{i,j}^{(k)}} (\mathbf{m}_j^{(k)} - \mathbf{w}_i^{(k)})$
  end for
  Each agent i computes:
  $\mathbf{w}_i^{(k+1)} \leftarrow ModelUpdate(\mathbf{\Psi}_{i,k}, \mu_k)$ and sends $\mathbf{m}_i^{(k+1)}$ to all its neighbors
end for

$ComputeTrust(i, j, \{\mathbf{m}_i^{(k)}, i \in \mathcal{N}\})$:
  Agent i computes $I_{ij}^{(k)}$ from (4.8) or (4.11)
  Agent i computes its local trust for j ($\tau_{ij}^{(k)}$) from (4.2)
  Agent i computes its global trust for j ($t_{ij}^{(k)}$) from (4.5) $ModelUpdate(\mathbf{w}_k, \mu_k)$:
  Initialize $\mathbf{\Psi}_{i,k} \leftarrow \mathbf{w}^{(k)}$
  $\mathcal{B}$ = mini-batch of size B
  for $b \in \mathcal{B}$ **do**
    $\mathbf{\Psi}_{i,k} \leftarrow \mathbf{\Psi}_{i,k} - \mu_k \nabla f(\mathbf{\Psi}_{i,k})$
  end for
  Return $\mathbf{w}_{i,k} \leftarrow \mathbf{\Psi}_{i,k}$

network. These weights play an important role in determining the trust level of the agents. We enumerate multiple methods that assign trust values to the agents based on the communicated weight updates:

- *Clustering-based Method*: In this method, the trustor entity i compares the messages it has received from trustee j, to all the other messages it has received from the other parties. Formally, for each neighbor j, party i computes:

$$\mathrm{dev}_{ij}^{(k)} = \sum_{l \in \mathcal{N}_i^+} \frac{\left|\left| w_l^{(k)} - w_j^{(k)} \right|\right|_2^2}{\left|\mathcal{N}_i^+\right|} \tag{4.7}$$

  where $\left|\mathcal{N}_i^+\right|$ is the set of the neighbors of agent i containing itself. Then, for each trustee j, it will benchmark the value of $dev_{ij}^{(k)}$ against a multiple of the median of all the deviations:

$$I_{ij}^{(k)} = \begin{cases} 1 & dev_{ij}^{(k)} \leq th_i * \text{ median } \left(\left\{\mathrm{dev}_{ij}^{(k)}\right\}\right) \\ 0 & o.w. \end{cases} \tag{4.8}$$

  This way, by adjusting the value of th_i at iteration k, the trustor party can decide not to trust those parties from which it has received too far-away messages.

- *Distance-based Method*: In this method, the trustor party i computes the distance between its local model and the model at trustee j, and uses this distance as a measure to assign trust values to its neighbors. Formally, party i computes the distance between the message it has received from party j in the previous and the current iterations; i.e.

$$d_{ij}\left(w_i^{(k-1)}, m_j^{(k-1)}\right) = \left|\left| w_i^{(k-1)} - m_j^{(k-1)} \right|\right|_2 \tag{4.9}$$

and,

$$d_{ij}\left(w_i^{(k-1)}, m_j^{(k)}\right) = \left\| w_i^{(k-1)} - m_j^{(k)} \right\|_2 \tag{4.10}$$

Then party i computes the difference between the two computed distances and decides on the value of $I_{ij}^{(k)}$ as follows:

$$I_{ij}^{(k)} = \mathbb{I}_{\left\{d_{ij}\left(w_i^{(k-1)}, m_j^{(k-1)}\right) - d_{ij}\left(w_i^{(k-1)}, m_j^{(k)}\right) \geq 0\right\}} \tag{4.11}$$

In other words, if node j has a benign behavior, then in one iteration of the protocol, its local model must have shifted towards the local model of party i. If this is not the case then party j has to be malicious or must be communicating an incorrect message to i.

4.4 UAV Deployment for Decentralized Federated Learning

The baseline for the UAV deployment problem is formulated as follows.

UP-DeFeL: Minimize N subject to

$\exists \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$:

$$\max_{v \in \mathcal{V}} \min_{n \in [N]} \|\mathbf{v} - \mathbf{x}_n\| \leq R_g \tag{4.12}$$

$$\max_{j \in \{2,\ldots,N\}} \min_{i<j} \|\mathbf{x}_j - \mathbf{x}_i\| \leq R_a \tag{4.13}$$

$$\mathbf{x}_n \in \mathbb{R}^2, \quad \forall n \in [N] \tag{4.14}$$

where $||.||$ is the l^2 norm operator, $\mathbf{v} \in \mathbb{R}^2$ is the Cartesian coordinates of the ground subject $v \in \mathcal{V}$ which is known a-priori and $\mathbf{x}_n \in \mathbb{R}^2$ is the n-th decision variable pair corresponding to the n-th

deployed UAV. Constraint set (4.12) ensures that each ground subject is within the air-ground communication range of the closest deployed UAV, where R_g is the ground radius of each UAV beam, while constraint set (4.13) enforces each deployed UAV is within the aerial communication range of at least one of the previously deployed UAVs where R_a is the aerial communication range of each UAV. The optimization model aims to cover all the ground subjects with a minimum number of deployed UAVs N while ensuring the deployed UAVs form a connected graph.

The exact solution to **UP-DeFeL** can only be obtained by performing an exhaustive search on different ways each ground node can be assigned to each deployed UAV for different values of N. The above model is an instance of the Euclidean disk cover problem which belongs to the class of $\mathcal{NP}$-hard problems [100]. Therefore, in the next subsection, we present a meta-heuristic approach based on deterministic annealing (DA) for solving an approximation of the above model that can generate near-optimal solutions with low time complexity.

4.4.1 Approximate UP-DeFeL Model

We leverage the following lemma in our approximation procedure.

Lemma 12 *For any real numbers* $\{s_n > 0\}_{n=1}^{N}$ *and large enough real number* $\alpha > 0$*, it holds that*

$$\max(s_1, \ldots, s_n) \cong (s_1^{\alpha} + \ldots + s_n^{\alpha})^{\frac{1}{\alpha}} \tag{4.15}$$

Using the above lemma, we can approximate constraints (4.12) and (4.13) to obtain the

approximate **UP-DeFel** model i.e. **A-UP-DeFel** as follows.

A-UP-DeFeL: Minimize N subject to

$\exists \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$:

$$\sum_{v \in \mathcal{V}} d_\alpha \left(\mathbf{v}, \mathbf{x}_{a_1(v)}\right) \leqslant R_g^\alpha \tag{4.16}$$

$$\sum_{n=2}^{N} d_\beta \left(\mathbf{x}_n, \mathbf{x}_{a_2(n)}\right) \leqslant R_a^\beta \tag{4.17}$$

$$\mathbf{x}_n \in \mathbb{R}^2, \quad \forall n \in [N] \tag{4.18}$$

where $a_1(v)$ is the function that assigns the ground subject $v \in \mathcal{V}$ to the closest deployed UAV, $a_1(v)$ is the function that assigns a UAV to the closest lower number deployed UAV, and the distortion functions $d_\alpha(.,.)$ and $d_\beta(.,.)$ are given by

$$d_\alpha \left(\mathbf{v}, \mathbf{x}_n\right) = \left\|\mathbf{v} - \mathbf{x}_n\right\|^\alpha \tag{4.19}$$

$$d_\beta \left(\mathbf{x}_j, \mathbf{x}_i\right) = \min_{i<j} \left\|\mathbf{x}_j - \mathbf{x}_i\right\|^\beta \tag{4.20}$$

A-UP-DeFeL is a non-convex constrained clustering problem. We first explain and then use the DA [101] procedure to solve this problem for near-optimal solutions.

4.4.2 Deterministic Annealing

DA is a clustering technique used to allocate a vast collection of data points, denoted as $\mathcal{X}$, to a limited set of centers $\mathcal{Y}$ by minimizing an average distortion function given by $D = \sum_{x \in \mathcal{X}} p(x) d(x, y(x))$, where $p(x)$ represents the likelihood of the data point x. Instead of a hard

clustering problem that can get stuck in local minima, DA introduces a soft clustering approach. Here, every data point $x \in \mathcal{X}$ might correspond to several centers $y \in \mathcal{Y}$ via *association fractions* described as $p(y|x)$. Based on the association fra distortion function can be reformulated as $D = \sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x)d(x,y)$. The ultimate goal of the DA method is to minimize this distortion function across varying levels of randomness, which are steered by the entropy function $H(X,Y)$. In more exact terms, DA seeks to minimize the target function $F = D - TH(Y|X)$ across a range of *temperature* levels T, starting at high temperatures and methodically reducing the T value.

4.4.3 DA solution for A-UP-DeFeL

The overall distortion function D and the conditional entropy $H(\mathcal{U}|\mathcal{V})$ are respectively given by

$$D = \sum_{v\in\mathcal{V}} p(v) \sum_{u\in\mathcal{U}} p(u|v) d_\alpha(u,v) + \lambda \sum_{n=2}^{N} d_\beta\left(u_n, u_m\right) \tag{4.21}$$

$$H(\mathcal{U}|\mathcal{V}) = -\sum_{v\in\mathcal{V}} p(v) \sum_{u\in\mathcal{U}} p(u|v) \log p(u|v) \tag{4.22}$$

where $\lambda > 0$ is the coefficient that balances the emphasis of the objective function on the first and second terms, $0 \leq p(u|v) \leq 1$ for $v \in \mathcal{V}$ and $u \in \mathcal{U}$ are the association fractions that need to satisfy constraints $\sum_{u\in\mathcal{U}} p(u|v) = 1$ for all $v \in \mathcal{V}$ and $p(u) = \sum_{v\in\mathcal{V}} p(v)p(u|v)$ for all $u \in \mathcal{U}$. We also note that if all the ground subjects are of the same priority then $p(v)$ follows the uniform distribution, i.e. $p(v) = \frac{1}{|\mathcal{V}|}$ for all $v \in \mathcal{V}$. Accordingly, the target function F is minimized when

the association fractions follow the Gibbs distribution:

$$p(u|v) = \frac{\exp\left(-\frac{d_\alpha(v,u)}{T}\right)}{\sum_{u\in\mathcal{U}} \exp\left(-\frac{d_\alpha(v,u)}{T}\right)} \qquad \forall v \in \mathcal{V}, u \in \mathcal{U} \tag{4.23}$$

in which case we have,

$$F^* = -T \sum_{v\in\mathcal{V}} p(v) \log\left(\sum_{u\in\mathcal{U}} \exp\left(-\frac{d_\alpha(v,u)}{T}\right)\right) + \lambda \sum_{n=2}^{N} d_\beta(u_n, u_m) \tag{4.24}$$

where F^* is the minimum of F. Then, the optimal locations for the UAVs are obtained by setting the partial derivatives of F^* for $\mathbf{x}_u$ vectors to zero for all $u \in \mathcal{U}$ where $\mathbf{x}_u = (x_u^{(0)}, x_u^{(1)})$ in the Cartesian plane. After straightforward calculus, it follows that for $i \in \{0, 1\}$,

$$\begin{aligned} x_u^{(i)} = & \frac{\alpha \sum_{u\in\mathcal{U}} v^{(i)} g_\alpha(v,u)}{\alpha \sum_{u\in\mathcal{U}} g_\alpha(v,u) + \lambda\beta(g_\beta(u) + \sum_{w\in\mathcal{W}_u} g_\beta(w))} \\ & + \frac{\lambda\beta(x_{a_2(u)}^{(i)} g_\beta(u) + \sum_{w\in\mathcal{W}_u} x_w^{(i)} g_\beta(w))}{\alpha \sum_{u\in\mathcal{U}} g_\alpha(v,u) + \lambda\beta(g_\beta(u) + \sum_{w\in\mathcal{W}_u} g_\beta(w))} \end{aligned} \tag{4.25}$$

where $\mathcal{W}_u = \{w \in \mathcal{U} : a_2(w) = u\}$ and for all $v \in \mathcal{V}, u \in \mathcal{U}$ we define the functions

$$g_\alpha(v,u) \doteq p(v)p(u|v)d_\alpha(v,u)^{1-2/\alpha} \tag{4.26}$$

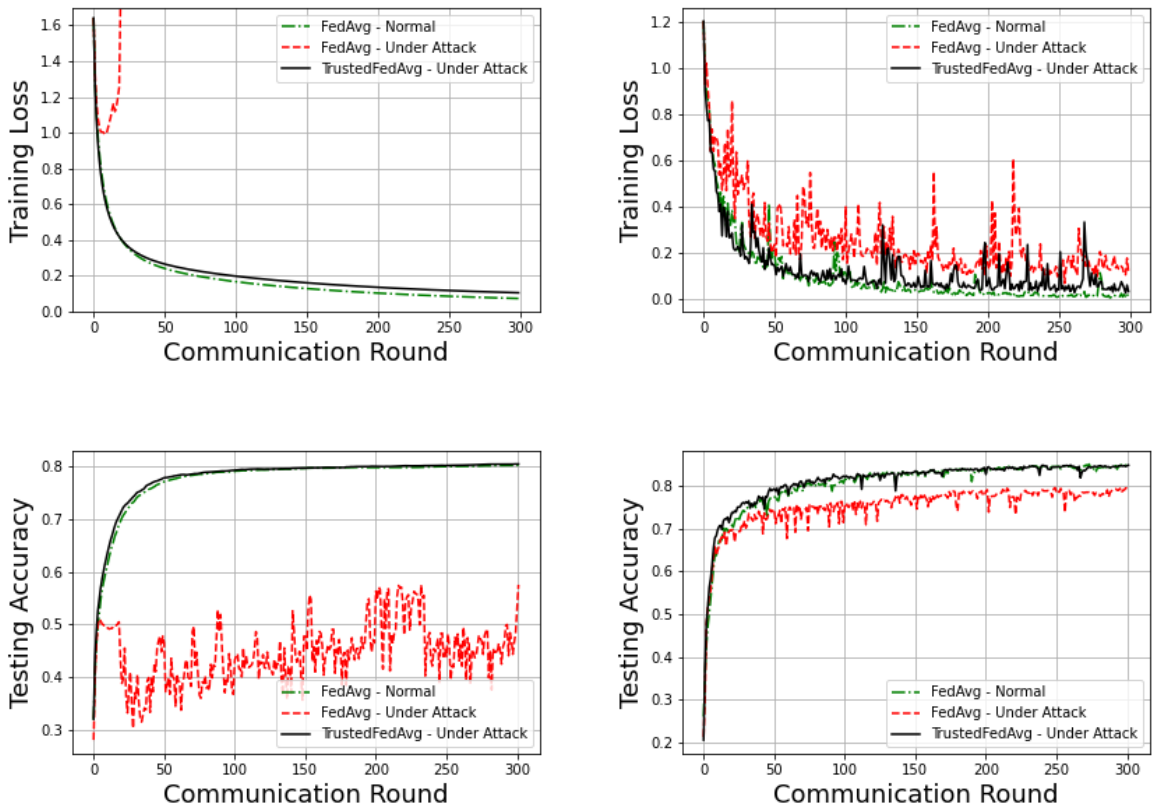
$$g_\beta(u) \doteq d_\beta\left(u, a_2(u)\right)^{1-2/\beta}. \tag{4.27}$$

4.5 Performance Evaluation

For the simulation setup, we adopt the settings of [91]. Our implementation of the FL process and the validation dataset are both based on [102]. We implement an attack in the presence of 80 nodes participating in the FL task with 10% of the nodes being corrupt. The attacker parties will generate the poisoned model by choosing the weights randomly in the range $(a * w_{min}, a * w_{max})$ where w_{min} and w_{max} are the minimum and maximum of the weights they receive from their neighbors. We implement this attack under two *mild* and *hard* settings corresponding to the case where $a = 1$, and $a = 2$ respectively. We have depicted 120 epochs of the process when the attack is mild. Fig. 4.4a shows how incorporating trust into the decentralized federated learning framework can protect the protocol from being invaded by malicious parties. In the absence of the trust mechanism, under corrupt network agents, the validation loss will not converge to the correct value corresponding to when the nodes are operating normally. Therefore, the performance of the trained model on test data degrades significantly, even when implementing the mild version of the attack. However, when the trust model is in place the trained model will converge to that of the normal implementation. We note that the validation loss after the 120 epoch converges to 0.14, and 0.18 for the normal (without any attacks) and trust-aware models, and diverges from 1.2 for the attacked unprotected model. The 0.14 validation loss corresponds to a 90% of accuracy on the test data.

4.5.1 Experimental Setup

Dataset: The data sample d characterized by $(\mathbf{x}_d, \mathbf{y}_d)$ has the input of $\mathbf{x}_d$, a 512-point FFT of the beat signals obtained from the radar echoes and averaged over ten consecutive frames. Each input



(a) MNIST-MLP (b) MNIST-CNN

Figure 4.3: Evaluation results on MNIST. Top: mean training loss; bottom: testing accuracy

FFT measurement is labeled by one of the $C = 8$ classes representing the distance between the cooperating robot and the human worker. Data distribution is non-IID as the local data samples collected independently by each device may contain only a subset of labels. Each device obtains $D_i = 25$ samples. The size of the validation dataset is $D = 16000$.

ML model: The considered learning model is a 2NN with a first fully connected (FC) layer of 32 hidden nodes of dimension 512×32 followed by a ReLu layer and a second FC layer of dimension $32 \times C$ followed by a softmax layer. We note that although the considered model is simple, it is a realistic use case for IoT devices with limited computation and power capabilities.

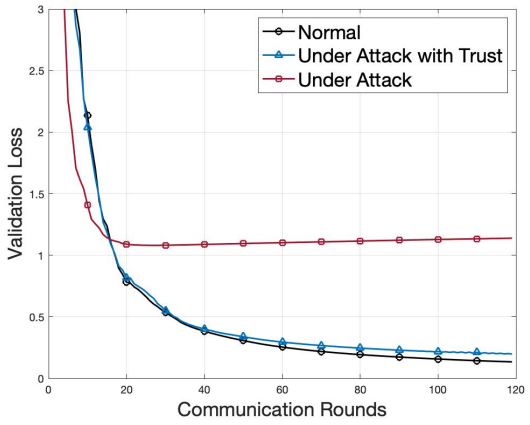
Attack model: We implement a model poisoning attack in the presence of 80 devices participa-

ting in the FL task where $p = 5\%, 10\%$ and 20% of the nodes are corrupt. The attacker parties will generate the poisoned model by choosing and transmitting the weights randomly in the range $(q * w_{min}, q * w_{max})$, where w_{min} and w_{max} are the minimum and maximum of the weights they receive from their neighbors. We assume that $q = 2$.

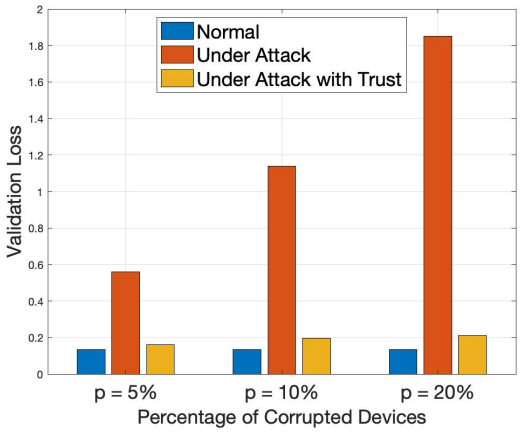
4.5.2 Trusted Decentralized FL: Numerical Results

Effect of the trust on resilience against attacks: In Fig. 4.4a, we compare the performance of the proposed trusted FL algorithm in the presence of the attacked agents, with both the normal (no attack) system and under attack system without trust incorporation. The validation loss is shown for 120 communication rounds and we assume that 10% of the agents are under attack. Fig. 4.4a illustrates how incorporating trust into the decentralized federated learning framework can protect the protocol from being invaded by malicious parties. In the absence of the trust mechanism and under corrupt network agents, the validation loss will not converge to the correct value corresponding to when all the agents are operating normally. Therefore, the performance of the trained model on test data degrades significantly, even with a moderate model poisoning attack. However, when the trust mechanism is in place, the trained model will converge to that of the normal implementation. We note that the validation loss after 120 epoch converges to 0.14, and 0.18 for the normal (without any attacks) and the attacked trust-aware processes, and diverges from 1.89 for the attacked unprotected process. We note that the value of 0.14 validation loss corresponds to a 90% of accuracy on the test data.

Impact of the percentage of the compromised agents: In Fig. 4.4b, the validation loss of the normal system, the system under attack with trust, and the system under attack without



(a) Effect of trust on resilience against attacks



(b) Impact of the percentage of compromised agents

trust consideration is depicted for three different values of p, corresponding to mild ($p = 5\%$), moderate ($p = 10\%$) and severe ($p = 20\%$) model poisoning attacks. It is observed that while the validation loss of the attacked system is significantly higher than the normal scenario, the proposed trusted decentralized FL algorithm can improve the performance of the trained model noticeably for all three values of p. In particular, we note that while the validation loss increases from 0.56 for a mild attack scenario to 1.89 and 1.85 for the moderate and severe attack cases respectively, incorporating trust in the decentralized FL algorithm results in a maximum loss value of 0.21 for the severe attack scenario.

Fig. 4.5 evaluates the performance of the decentralized FL algorithm by benchmarking the validation loss profile for 140 epochs against other learning algorithms including FedAvg, and on-device ML. The decentralized FL algorithms are compared for parameter k, i.e. the number of neighbors that each UAV can use in the consensus-based learning process. FedAvg corresponds to the centralized FL where we assume there exists a centralized aggregator that is reachable by all the active UAVs in each round. On-device ML corresponds to the case that there is no federation, i.e. each UAV only uses the data and the model of its own. The results

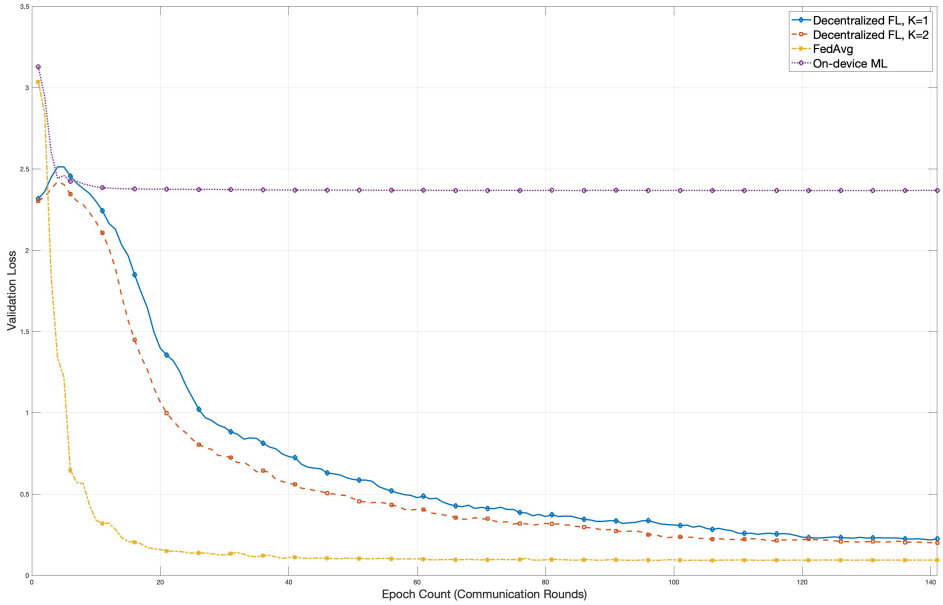


Figure 4.5: Comparing different ML techniques validation loss for the MNIST task

are depicted when averaged over 100 runs and over all the UAVs. It is observed that given the insufficient amount of data and the non-i.i.d data profile at each UAV, the on-device ML method converges to a very high validation loss, compared to any of the FL methods. Also, it is observed that validation loss for both the decentralized FL methods converges to a value that is slightly higher than that of the FedAvg. This confirms the successful performance of the decentralized FL technique for the MNIST task.

The comparison of the decentralized FL methods for different values of k in Fig. 4.6, reveals the impact of the number of neighbors used per device in each round on the performance of the decentralized FL method. The box plots in Fig. 4.6 compare the validation loss profile for the UAVs at certain epoch numbers for the case of $k = 1$ and $k = 2$. It is observed that as the UAVs are allowed to communicate with $k = 2$ neighbors in each round, the validation

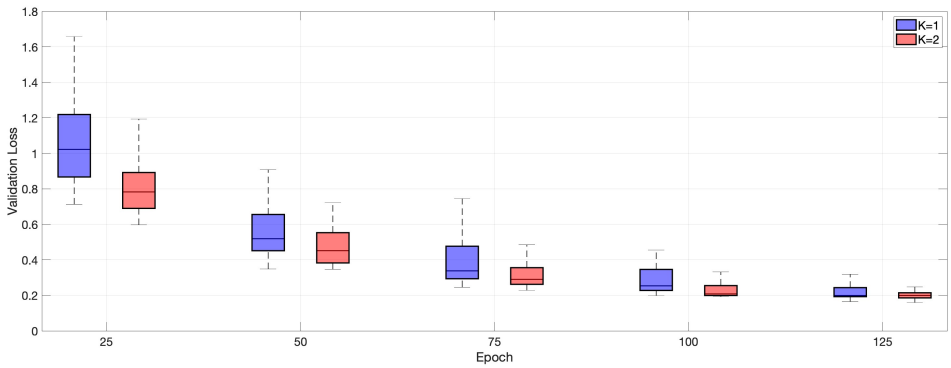


Figure 4.6: Impact of the number of allowed per-device neighbors in each communication round

loss converges faster to its final value. In this experiment, for example, $k = 2$ will result in convergence in 120 epochs, while $k = 1$ needs at least 30 more epochs to converge. However, no notable difference in the final value of validation loss is observed for the two cases.

Chapter 5: Channel Reciprocity Calibration for Hybrid Beamforming in Cell-free Massive MIMO Systems

5.1 Overview

Massive MIMO is considered an enabling technology towards realizing 5G and beyond communications, since a large number of antennas per base station allows for extended coverage, effective beamforming [103] [104], higher data rates per users [105], and better spectral efficiency [106]. For this to be practical on a large scale, effective techniques are required to reduce the hardware cost and power consumption of such systems. A hybrid beamforming transceiver is an effective solution for low-cost massive MIMO by reducing the number of radio-frequency (RF) chains and hence expensive elements such as digital-to-analog/analog-to-digital converters (DAC/ADC), mixers, etc., and introducing simple phase shifters [107]. Further, massive MIMO is to be widely used in the physical layer of the next-generation wireless systems, due to its potential for increasing the network capacity and user bit rates [106]. To mitigate the inter-cell interference resulting from the densification of the networks deploying massive MIMO, the distributed massive MIMO (a.k.a. cell-free massive MIMO) technology is proposed. However, such network densification technique contributes to the inter-cell-interference (ICI) effect which negatively impacts the performance of the wireless communications system. To overcome this

issue, the concept of distributed massive MIMO systems is envisioned as a potential technology for realizing 6G multi-antenna systems [108].

A distributed massive MIMO system is a scalable [105] implementation of coordinated multi-point transmission (CoMP), where a multitude of access points (APs) that are geographically separated, co-operate towards jointly serving the mobile users (MUs) to increase the received signal quality and system throughput [109]. Each AP has its local oscillator and may employ a large array of antennas. The APs are interconnected and connected to a central server (CS) through wired backhauling (e.g. Ethernet). To enable beamforming to a user in the downlink the nodes which are involved in the downlink transmission need to know the downlink channel to a particular user. To avoid feedback overhead, the AP may be able to estimate the downlink channel if the downlink and uplink are at the same frequency. Therefore, the recommended scheme for distributed MIMO systems is a time-division duplex (TDD).

Under the TDD scheme, each AP is enabled to estimate the uplink channel using the uplink pilot symbols received from the user equipment (UEs). Exploiting the reciprocity of the UL-DL channel, the uplink estimates can be utilized for estimating the downlink channel. This is typically termed a reciprocity-based channel estimation in the literature. Under ideal channel reciprocity, once the uplink channel is estimated the same estimate can be used as the equivalent for the downlink channel. However, in practice, full channel reciprocity does not hold in general due to the non-reciprocity in the RF chains of the transceivers at the channel endpoints. To overcome this issue, calibration techniques must be used to tune the transceiver's hardware for reciprocity-based channel estimation to become feasible. Several types of signaling exist in the literature for reciprocity calibration. A line of research proposes to perform calibration utilizing the bidirectional channel between the APs and the UEs [110]. Some other approaches perform

internal calibration at the APs [111] [112], while some works consider over-the-air signaling between the APs. Given that bidirectional signaling between the APs and the UEs highly depends on the quality of the UEs links and that deploying cables for the sake of calibration between the distributed APs is not practical, it is widely supported that the third calibration approach is the right solution for distributed MIMO systems [108] [113] [114].

Fig. 5.1 depicts the structure of a hybrid beamformer. Regardless of the imperfection in the phase shifter network (PSN), the reciprocity mismatch at hybrid beamforming systems stems from the imperfection in the hardware at the *digital* and the *analog* RF chains. More precisely, at the digital RF chain, the digital-to-analog converter, and analog-to-digital converter (DAC/ADC), and the analog chain, the power amplifier, and the low-noise amplifier (PA/LNA) require calibration for channel reciprocity. Therefore, although the literature on reciprocity calibration for fully-digital beamforming is rich, reciprocity mismatch in hybrid beamforming systems needs to be addressed differently. Reciprocity calibration for hybrid beamforming in TDD-based massive MIMO systems was first addressed in [115], where the authors used an internal calibration method where they considered a PSN with sub-array-based architecture. As opposed to this pioneering work, and similar to [116] [108] [106], in this chapter, we considered a fully-connected PSN. Within the context of distributed massive MIMO systems, the authors in [108] proposed a maximum likelihood (ML)-based calibration relying on joint beam sweep by all APs in the network. The authors consider a single mismatch parameter per AP and only consider fully digital beamforming. As opposed to this work, in this chapter, we consider a single mismatch coefficient per digital RF chain which allows for a much more precise reciprocity calibration [117] [118]. However, among the prior art, at the macroscopic level our approach remains closest to the one presented in [107] where the authors decompose the channel recip-

rocity calibration in TDD hybrid beamforming systems into two disjoint problems. They first come up with a closed-form solution for the calibration of the digital chain and then formulate and solve an optimization problem to calibrate the analog chain. In [116] calibration of the PSN is considered. Having taken this imperfection into account in a later work, the authors calibrate the analog and digital chain in the hybrid massive MIMO system [106]. In this chapter, we decompose the channel reciprocity calibration task for hybrid beamforming into two sub-problems. The first sub-problem entails calibrating the digital chain up to an unknown scaling factor. Having obtained the calibration parameters of the digital chain, the second sub-problem aims at calibrating the analog RF chain. We note that compared to the prior art, our approach has minimal overhead in terms of the number of communication rounds required for estimating the calibration parameters. The main contributions of the chapter are as follows:

- We present a novel two-step low-complexity scheme for calibrating any two nodes in a distributed MIMO system.
- We introduce the novel concept of *reciprocal tandem* that is a cornerstone in the joint reciprocity calibration of the APs in a distributed MIMO system.
- Utilizing the concept of reciprocal tandem, we augment the two-step approach for reciprocity calibration of two nodes with a third step, and propose a low-complexity three-step scheme for the joint calibration of the cluster of APs in the distributed MIMO network.
- We demonstrate how our reciprocity calibration technique enables the cluster of APs to perform the cooperative hybrid beamforming task utilizing numerical simulation.

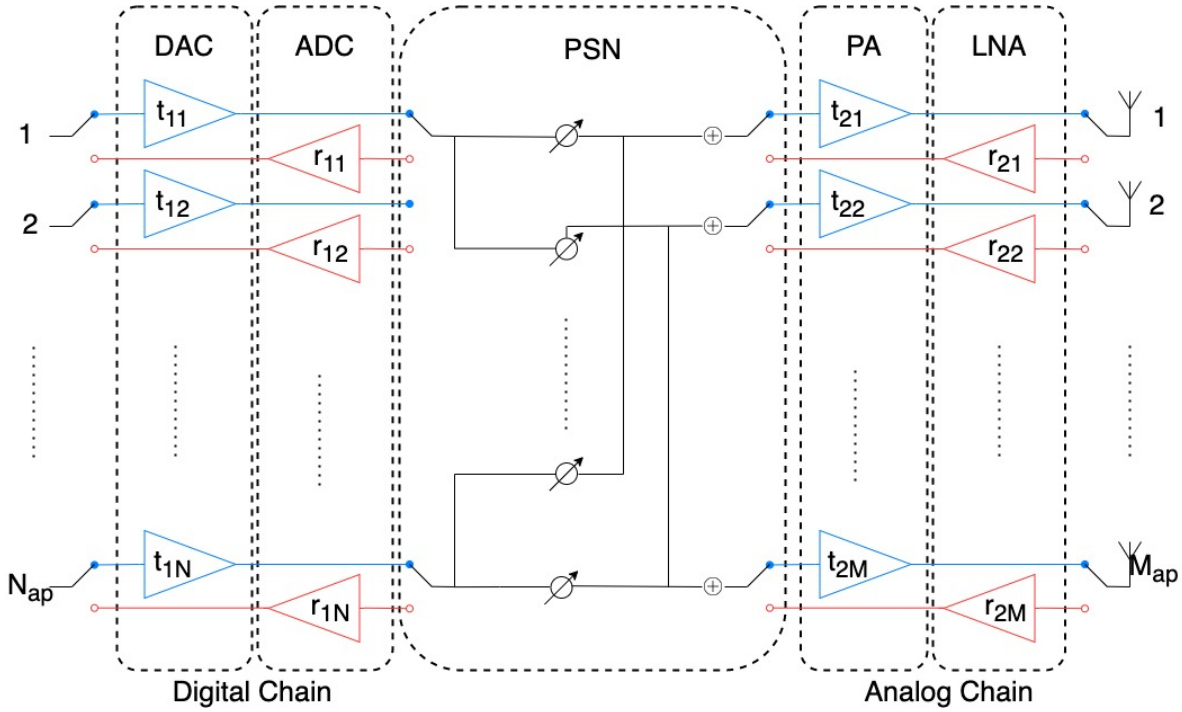


Figure 5.1: Hybrid Beamforming System Model

5.2 System model

We consider a distributed mmWave MIMO system where a cluster of cooperative multi-antenna APs employing hybrid beamforming, jointly serve mobile users (MUs).

5.2.1 Beamforming Model

The hybrid beamformer consists of a first-level digital beamformer and a second-level analog beamformer. Each AP is equipped with N_{ap} digital RF chains which are fed by the output of the digital beamformer for each transmitted stream. In turn, the output of the digital RF chain goes through an analog beamformer which is connected to M_{ap} analog RF chains each of which is connected to an antenna element where the antenna elements are arranged in a uniform linear array (ULA) structure. The analog beamformers are usually comprised of phase-only vectors

which are implemented by a fully connected phase shift network (PSN) [106] between the output of the digital RF chains and the input of the analog RF chains. In a subarray-based structure [115] the output of each digital RF chain feeds a disjoint set of analog RF chains (or antennas), hence, the analog beamformer has a block diagonal structure. We consider a general analog beamformer in this paper (except for a practical constraint discussed later in Section 9.3) that can treat either of the fully connected or subarray-based structures as a special case. At the MUs, there are M_{mu} and N_{mu} analog and digital RF chains, respectively.

5.2.2 Channel Model

The mmWave channel between each AP and each MU is assumed to have only a few spatial clusters $L \ll N_{\text{ap}}, N_{\text{mu}}$, where L is the number of scatterers, determining the number of paths between each AP and each MU. We consider a geometric channel model that is given by,

$$\mathbf{H} = \sqrt{\frac{N_{\text{ap}} N_{\text{mu}}}{L}} \sum_{\ell=1}^{L} \alpha_\ell \mathbf{a}_{\text{mu}}\left(\phi_\ell\right) \mathbf{a}_{\text{ap}}^{T}\left(\theta_\ell\right) \tag{5.1}$$

where $\alpha_l \sim \mathcal{CN}\left(0, \sigma_\alpha^2\right)$ is the gain of the $l-$th path, $\mathbf{a}_{\text{mu}}\left(\phi_\ell\right) \in \mathbb{C}^{N_{\text{mu}}}$, and $\boldsymbol{a}_{\text{ap}}\left(\theta_\ell\right) \in \mathbb{C}^{N_{\text{ap}}}$ denote the directivity vectors of the MUs and the APs at the angles of arrival (AoA) $\phi_\ell \in [-\pi/2, \pi/2)$ to the MUs, and the angles of departure (AoD) $\theta_\ell \in [-\pi/2, \pi/2)$ from the APs. The directivity vectors $\boldsymbol{a}_{\text{ap}}$, and $\boldsymbol{a}_{\text{mu}}$ are given by,

$$\begin{aligned}
\mathbf{a}_{\text{ap}}\left(\theta_\ell\right) &= \left[1, e^{-j\frac{2\pi d}{\lambda}\sin\theta_\ell}, \cdots, e^{-j\frac{2\pi d}{\lambda}(N_{\text{ap}}-1)\sin\theta_\ell}\right]^T, \\
\mathbf{a}_{\text{mu}}\left(\phi_\ell\right) &= \left[1, e^{-j\frac{2\pi d}{\lambda}\sin\phi_\ell}, \cdots, e^{-j\frac{2\pi d}{\lambda}(N_{\text{mu}}-1)\sin\phi_\ell}\right]^T,
\end{aligned} \tag{5.2}$$

where λ is the carrier wavelength and d denotes the distance between every two adjacent antenna elements and is set as $d = \lambda/2$. The effective channel between any AP and MU is a function of the mmWave channel and the transfer functions of the RF chains, the PSNs, and the analog beamformers on both sides. The uplink (UL) and the downlink (DL) channels are given by,

$$\mathbf{H}_{DL} = \mathbf{R}_{\text{mu},1}\mathbf{B}^T\mathbf{R}_{\text{mu},2}\mathbf{H}\mathbf{T}_{\text{ap},2}\mathbf{F}\mathbf{T}_{\text{ap},1}, \tag{5.3}$$

$$\mathbf{H}_{UL} = \mathbf{R}_{\text{ap},1}\mathbf{F}^T\mathbf{R}_{\text{ap},2}\mathbf{H}^T\mathbf{T}_{\text{mu},2}\mathbf{B}\mathbf{T}_{\text{mu},1} \tag{5.4}$$

where $\mathbf{T}_{\text{ap},1}, \mathbf{R}_{\text{ap},1} \in \mathbb{C}^{N_{\text{ap}} \times N_{\text{ap}}}$ are the *digital reciprocity calibration matrices* denoting the frequency responses of the transmit and receive digital RF chains (DAC/ADC) at the AP, and similarly $\mathbf{T}_{\text{ap},2}, \mathbf{R}_{\text{ap},2} \in \mathbb{C}^{M_{\text{ap}} \times M_{\text{ap}}}$ are the *analog reciprocity calibration matrices* which denote the transmit and receive frequency responses of the analog chains (PA/LNA) at the AP, respectively.

Similarly, at the MU, $\mathbf{T}_{\text{mu},1}, \mathbf{R}_{\text{mu},1} \in \mathbb{C}^{N_{\text{mu}} \times N_{\text{mu}}}$ denote the frequency responses of the transmit and receive digital RF chains, and $\mathbf{T}_{\text{mu},2}, \mathbf{R}_{\text{mu},2} \in \mathbb{C}^{M_{\text{mu}} \times M_{\text{mu}}}$ are the transmit and receive frequency responses of the analog chain. We note that all these matrices are diagonal with the diagonal elements modeling the gain and phase characteristics of each of the chain elements. The off-diagonal entries model the cross-talk between the RF hardware that is assumed to be almost zero under proper RF design [119]. We assume that all the transceivers store codebooks for the hybrid beamforming task where each codebook consists of codewords that each represent a beamforming vector. The $\mathbf{F} \in \mathbb{C}^{M_{\text{ap}} \times N_{\text{ap}}}$ and $\mathbf{B} \in \mathbb{C}^{M_{\text{mu}} \times N_{\text{mu}}}$ are known the beamforming matrices at the AP and the MU, respectively where each beamforming matrix consists of N_t beamforming vectors. The matrices $\mathbf{B}$ and $\mathbf{F}$ model the beamformers that precode the input analog streams that are then amplified and sent to the analog chains. We note that each AP or MU uses the same

matrices for beamforming both when transmitting and receiving. The transmission in the DL direction can be modeled by,

$$\mathbf{y}_{\text{DL}} = \mathbf{H}_{\text{DL}}\mathbf{x}_{\text{DL}} + \mathbf{z}_{\text{DL}} \tag{5.5}$$

$$\mathbf{y}_{\text{UL}} = \mathbf{H}_{\text{UL}}\mathbf{x}_{\text{UL}} + \mathbf{z}_{\text{UL}} \tag{5.6}$$

where $\mathbf{x}_{\text{DL}} \in \mathbb{C}^{N_{\text{ap}}}$ and $\mathbf{x}_{\text{UL}} \in \mathbb{C}^{N_{\text{mu}}}$ are the input streams to the digital RF chains (i.e., the output of the baseband processing unit (BBU) which could in turn be digitally precoded symbols) the AP and the MU, respectively. In the DL direction, the AP may use a digital precoder $\mathbf{W} \in \mathbb{C}^{N_{\text{ap}} \times N_s}$, such that $\mathbf{x}_{\text{DL}} = \mathbf{W}\mathbf{s}$ where $\mathbf{s} \in \mathbb{C}^{N_s}$ is the vector of digital symbols transmitted from the AP with $N_s \leq N_{\text{ap}}$ being the number of data streams, $\mathbf{y}_{\text{DL}} \in \mathbb{C}^{M_{\text{mu}}}$ and $\mathbf{y}_{\text{UL}} \in \mathbb{C}^{M_{\text{ap}}}$ are the vectors of the received signal streams at the BBU in the DL and UL directions, respectively, and $\mathbf{z}_{\text{DL}} \in \mathbb{C}^{M_{\text{mu}}}$ and $\mathbf{z}_{\text{UL}} \in \mathbb{C}^{M_{\text{ap}}}$ are the vectors of the additive white Gaussian noise (AWGN) with the distributions $\mathbf{z}_{\text{DL}} \sim \mathcal{CN}\left(\mathbf{0}, \sigma_{\text{z}}^2\mathbf{I}_{M_{\text{mu}}}\right)$ and $\mathbf{z}_{\text{UL}} \sim \mathcal{CN}\left(\mathbf{0}, \sigma_{\text{z}}^2\mathbf{I}_{M_{\text{ap}}}\right)$.

5.3 Reciprocity Calibration between two nodes

In this section, we address calibration between two nodes S and S'. Without loss of generality, we assume each node employs M analog RF chains and N digital RF chains. The goal is to estimate a combination of the digital and analog calibration matrices in the transmit and receive paths for both nodes in such a way that the channel estimation in one direction can be used in the reverse direction. To this end, we propose a two-step approach for the calibration process where in the first step we estimate the digital reciprocity calibration matrices and in the second step we estimate a combination of the analog reciprocity matrices. As becomes evident

in the following, there is no sub-optimality in breaking the calibration process into two steps, and our approach takes advantage of the estimated parameters in the first step in the second step to minimize the number of pilot transmissions required for calibration. This is very important to minimize the time spent on calibration not only to reduce the calibration overhead but also to make sure that the channel variation is negligible during the calibration process.

In the first step, the digital reciprocity calibration matrices in the transmit and receive paths for either of the nodes are estimated up to a scaling factor. Without loss of generality, as we will discuss later, we use the calibration parameter corresponding to the first antenna element as the scaling factor. In the second step, we use the estimation of the digital calibration matrices in formulating the problem of finding a combination of analog calibration matrices that are required to benefit from channel reciprocity. Although no claim of optimality is made, the proposed approach is built on a particular pilot transmission and analog beamformer selection to minimize the number of transmissions required to perform the reciprocity calibration. This is particularly important because, during the transmission of a group of such pilots, the channel variation is assumed to be negligible. Moreover, reciprocity calibration has to be performed periodically to capture the effect of variation in the properties of the elements in the analog and digital RF chains. Hence, an efficient calibration process with minimum possible time should be devised to reduce the overhead of the calibration.

We note that even during pilot transmissions, hybrid beamforming is employed by using proper analog beamformers (i.e., transmit and receive beamformers). A proper analog beamformer usually uses the entire transmit antenna array, i.e., the weight corresponding to each antenna element is nonzero. This property trivially holds in the special case that the analog beamformers are implemented by a PSN. This property is important in practice to ensure good beamforming

gain, as in theory one may suggest using a trivial beamformer that activates only one antenna to expedite the calibration process to a linear time based on the number of antennas. Although in abstract theoretical analysis, such beamformers may seem to be the most efficient in terms of reducing the calibration time, in practice, an analog beamformer that uses a single antenna on average transmits (or receives) a fraction of $1/M$ power in comparison to a beamformer which uses the entire antenna array that consists of M antenna elements. As a result of this important practical limitation on the analog beamformers, we will see that even though the first step, i.e., the digital chain reciprocity calibration, may be performed in a time that is in the linear order of the number of digital RF chains, the second step, i.e., the analog chain reciprocity calibration between two nodes, requires a time that is quadratic in the order of the number of analog RF chains, i.e., number of antenna elements.

5.3.1 Digital Chain Reciprocity Calibration

To estimate the digital transmit matrix $\mathbf{T}_1$ at node S and digital receive matrix $\mathbf{R}'_1$ at node S', we transmit N consecutive pilot signals from N different digital RF chains of node S with the digital reciprocity calibration parameters $t_{1i}, i = 1, \ldots, N$. In all these transmissions, we use an arbitrary but unique beamforming vector, say $\mathbf{f}_1$. In other words, we use the pilot signal $\mathbf{s} = [s]$ and $\mathbf{F} = \mathbf{f}_1$ in (5.5). On the receiver side, the node S' will receive on all its digital RF chains using the $M \times N$ receive beamforming matrix $\mathbf{B}$ that is defined as $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_1, \ldots, \mathbf{b}_1]$, i.e., consisting of N column vectors b_1. The received signal for the i-th transmission is given by

$$\mathbf{y}'_i = \mathbf{R}'_1 \mathbf{B}^T \mathbf{R}'_2 \mathbf{H} \mathbf{T}_2 \mathbf{f}_1 t_{1i} s + z_{DL}, \quad \forall i = 1 \ldots N \tag{5.7}$$

where $\mathbf{R}'_1$ is a diagonal matrix with the diagonal elements with the diagonal elements $r'_{1i}, i = 1, \ldots, N$ represented by the vector $\mathbf{r}'_1 = [r'_{11}, \ldots, r'_{1N}]$. Similarly, for the diagonal matrix $\mathbf{T}_1$, $\mathbf{T}_2$, and $\mathbf{R}'_2$ we define $\mathbf{t}_1 = [t_{11}, \ldots, t_{1N}]$, $\mathbf{t}_2 = [t_{21}, \ldots, t_{2N}]$, and $\mathbf{r}'_2 = [r'_{21}, \ldots, r'_{2N}]$. Further, define the scaling factor $h \doteq \mathbf{b}_1^T \mathbf{R}'_2 \mathbf{H} \mathbf{T}_2 \mathbf{f}_1$. Therefore, $\mathbf{R}'_1 \mathbf{B}^T \mathbf{R}'_2 \mathbf{H} \mathbf{T}_2 \mathbf{f}_1 t_{1i}$ can be estimated from (5.7) as,

$$\tilde{\mathbf{y}}'_i = \mathbf{R}'_1 h t_{1i}, \quad \forall i = 1 \ldots N \tag{5.8}$$

Note that for all $k, l \in 1..N$, the relation between any r'_{1k} and r'_{1l} can be easily found as $r'_{1k}/r'_{1l} = \tilde{y}'_{ik}/\tilde{y}'_{il}$ through any of the N equations in (5.8) where $\tilde{y}'_{ik}$ is the k-th element of the vector $\tilde{\mathbf{y}}'$. Therefore, the $\mathbf{R}'_1$ matrix can be estimated up to a scaling factor. However, there are N such noisy observations that may each result in different noisy estimates for $\mathbf{R}'_1$. A similar argument goes for $\mathbf{T}_1$. To get a more reliable estimate of both $\mathbf{R}'_1$ and $\mathbf{T}_1$, we formulate the following optimization problem to find first-level reciprocity calibration matrices.

Problem 13 *(First-level Reciprocity Calibration)*

The transmit and receive matrices $\mathbf{T}_1$, and $\mathbf{R}'_1$ can be found up to a scaling factor as the solution to the following least-square minimization problem.

$$\mathbf{T}_1, \mathbf{R}'_1 = \arg\min \sum_{i=1}^{N} \sum_{j=1}^{N} \left\| y'_{ij} - r'_{1i} h t_{1j} \right\|^2 \tag{5.9}$$

Solving Problem 13, enables us to estimate the digital transmit matrix $\mathbf{T}_1$ at node S and digital receive matrix $\mathbf{R}'_1$ at node S' with N^2 observations that are achieved through only N

Algorithm 6 *Digital Chain Calibration*

Input: $s, \mathbf{F} = \mathbf{f}_1, \mathbf{B} = [\mathbf{b}_1, \mathbf{b}_1, \ldots, \mathbf{b}_1]$.

1: **For** $i = 1$ to N:

2: Tx: Send s with digital chain t_{1n} and beamformer $\mathbf{F}$.

3: Rx: Receive vector $\mathbf{y}'_i$ on all RF chains with beamformer $\mathbf{B}$ according to:

$$\mathbf{y}'_i = \mathbf{R}'_1 \mathbf{B}^T \mathbf{R}'_2 \mathbf{H} \mathbf{T}_2 \mathbf{f}_1 t_{1i} s + z_{DL}$$

4: **End For**

5: Define $h \doteq \mathbf{b}_1^T \mathbf{R}'_2 \mathbf{H} \mathbf{T}_2 \mathbf{f}_1$. **Solve**

$$\{t_{1i}, r'_{1i}\}_{i=1}^N = \arg \min_{\{t_{1i}, r'_{1i}\}_{i=1}^N} \sum_{i=1}^N \sum_{j=1}^N \left\| y'_{ij} - r'_{1i} h t_{1j} \right\|^2$$

6: **Return** $\mathbf{T}_1 = \{t_{1i}\}_{i=1}^N$, and $\mathbf{R}'_1 = \{r'_{1i}\}_{i=1}^N$.

transmissions. Similarly, the digital transmit matrix $\mathbf{T}'_1$ at node S' and digital receive matrix $\mathbf{R}_1$ at node S can also be determined with additional N pilot transmission from node S' that results in additional N^2 observations.

5.3.2 Analog Chain Reciprocity Calibration

Having determined the $\mathbf{R}_1$, $\mathbf{T}_1$, $\mathbf{R}'_1$, and $\mathbf{T}'_1$ matrices up to an unknown scaling factor, in this section, we focus on estimating the receive and transmit matrices of the analog chain. To do so, we perform a structured pilot transmission over different digital transmit chains, and for each such transmission, we receive on all N digital receive chains as described in the following. We select M transmit beamformer $\mathbf{f}_i$, $i = 1, \ldots, M$ and $\mathbf{b}_i, i = 1, \ldots, M$ from the transmit and receive codebook, respectively, such that the beamforming matrices $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_M]$ and $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_M]$ are full rank. For each $\mathbf{f}_i$, $i = 1, \ldots, M$, we perform $k = 0, 1, \ldots, \lceil M/N \rceil - 1$ transmissions with transmit beamformer $\mathbf{f}_i$ from chain 1, and receive beamformers given by

$\mathbf{b}_{1+kN}, \mathbf{b}_{2+kN}, \ldots, \mathbf{b}_{N+kN}$ for the k^{th} transmission with the beamformer $\mathbf{f}_i$, where $\mathbf{b}_m = \mathbf{b}_1$ for $m > M$. Hence, after $M\lceil M/N \rceil$ transmission we gather M^2 observations that we arrange in a $M \times M$ matrix $\mathbf{Y}$. Using the model in the downlink direction (5.5), we can write

$$\mathbf{y}'_{ki} = \mathbf{R}'_1 \mathbf{B}_k^T \mathbf{R}'_2 \mathbf{H} \mathbf{T}_2 \mathbf{f}_i t_{11} s + \mathbf{z}'_{ki}, \tag{5.10}$$

for all $k = 1, \ldots, \lceil M/N \rceil$ and $i = 1, \ldots, M$ where the estimated values for $\mathbf{R}'_1$ is already computed in the last step up to a scaling factor r'_{11}. The j^{th}, $j = 1, \ldots, N$ row of the matrix product $\mathbf{R}'_1 \mathbf{B}_k$ for $k = 1, \ldots, \lceil M/N \rceil$ is given by a row vector

$$r'_{1j} \mathbf{b}^T_{N(k-1)+j} = r'_{11} (r'_{1j}/r'_{11}) \mathbf{b}^T_{N(k-1)+j} \doteq r'_{11} \tilde{\mathbf{b}}^T_m \tag{5.11}$$

where $m = N(k-1)+j$. Stacking the these row vectors $\tilde{\mathbf{b}}^T_m$ for $m = 1, \ldots, M$ results in an $M \times M$ matrix $\tilde{\mathbf{B}}$. Considering $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_M]$ which consists of all M transmit beamforming vectors and the modified received beamforming vectors in $\tilde{\mathbf{B}}$, we would get an aggregate $M \times M$ observation matrix $\mathbf{Y}'$ as

$$\mathbf{Y}' = r'_{11} \tilde{\mathbf{B}}^T \mathbf{R}'_2 \mathbf{H} \mathbf{T}_2 \mathbf{F} t_{11} s' + \mathbf{Z}' \tag{5.12}$$

Similarly, in the uplink direction, we can write,

$$\mathbf{Y} = r_{11} \tilde{\mathbf{F}}^T \mathbf{R}_2 \mathbf{H}^T \mathbf{T}'_2 \mathbf{B} t'_{11} s + \mathbf{Z}. \tag{5.13}$$

Let $\tilde{\mathbf{Y}}'$ and $\tilde{\mathbf{Y}}$ be the estimation of $r'_{11} \tilde{\mathbf{B}}^T \mathbf{R}'_2 \mathbf{H} \mathbf{T}_2 \mathbf{F} t_{11}$ and $r_{11} \tilde{\mathbf{F}}^T \mathbf{R}_2 \mathbf{H}^T \mathbf{T}'_2 \mathbf{B} t'_{11}$ based on (5.12)

and (5.24), respectively. We have the following two observations for the channel,

$$\tilde{\mathbf{H}}_1 = (r'_{11}t_{11})^{-1}\mathbf{R}_2'^{-1}\tilde{\mathbf{B}}^{-T}\tilde{\mathbf{Y}}'\mathbf{F}^{-1}\mathbf{T}_2^{-1} \tag{5.14}$$

$$\tilde{\mathbf{H}}_2 = (r_{11}t'_{11})^{-1}\mathbf{T}_2'^{-1}\mathbf{B}^{-T}\tilde{\mathbf{Y}}^{T}\tilde{\mathbf{F}}^{-1}\mathbf{R}_2^{-1} \tag{5.15}$$

Now, define $\mathbf{X} \doteq \tilde{\mathbf{B}}^{-T}\tilde{\mathbf{Y}}'\mathbf{F}^{-1}$ and $\mathbf{Z} \doteq \mathbf{B}^{-T}\tilde{\mathbf{Y}}^{T}\tilde{\mathbf{F}}^{-1}$, where $\mathbf{X}, \mathbf{Z} \in \mathbb{C}^{M\times M}$ are completely known. Let x_{ij} and z_{ij} be the elements of the $\mathbf{X}$ and $\mathbf{Z}$ matrices for all $i, j = 1 \dots M$. We have,

$$\frac{\tilde{h}_{1,ij}}{\tilde{h}_{2,ij}} = \frac{x_{ij}r_{2i}'^{-1}t_{2j}^{-1}}{z_{ij}t_{2i}'^{-1}r_{2j}^{-1}}.\frac{r_{11}'^{-1}t_{11}^{-1}}{t_{11}'^{-1}r_{11}^{-1}} \approx 1. \quad \forall i, j = 1 \dots M \tag{5.16}$$

Let us define $\beta = r_{11}^{-1}t_{11}'^{-1}/(r_{11}'^{-1}t_{11}'^{-1})$, and $\alpha_i = r_{2i}t_{2i}^{-1}$, $\alpha'_i = r'_{2i}t'^{-1}_{2i}$. We have

$$\frac{\alpha_j}{\alpha'_i} \approx \beta\frac{z_{ij}}{x_{ij}}. \quad \forall i, j = 1 \dots M \tag{5.17}$$

This means that given α_1, we can find all α_i's and α'_i's based on a single scaling factor that is α_1. Please note that the product of $\mathbf{R}_2\mathbf{T}_2^{-1}$ is a diagonal matrix where the diagonal elements are given by $[\alpha_1, \alpha_2, \dots, \alpha_M]$. Similarly, $\mathbf{R}'_2\mathbf{T}_2'^{-1}$ the follows a diagonal structure with the diagonal elements given as $[\alpha'_1, \alpha'_2, \dots, \alpha'_M]$. To minimize the estimation error, the estimation of each fraction is not performed individually and we formulate the problem as a joint least square optimization which involves all fractions to get the best estimates of the second-level calibration matrices as follows.

Problem 14 *(Second-level Reciprocity Calibration)*

The second-level analog chain matrices can be found as the solution to the following least-square

Algorithm 7 *Analog Chain Calibration*

Input: $s, \mathbf{R}_1, \mathbf{T}_1, \mathbf{R}'_1, \mathbf{T}'_1$, Full-Rank $\mathbf{F}$, Full-Rank $\mathbf{B}$ such that $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_M], \mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_M]$

1: **For** $\mathbf{f}_i, i = 1$ to M:
2: **For** $k = 0 \ldots \lceil M/N \rceil - 1$:
3: Tx: Send s on digital chain t_{11} with beamformer $\mathbf{f}_i$
4: Rx: Receive $\mathbf{y}'_{ki}$ on all digital chains with beamformer $\mathbf{B}_k = [\mathbf{b}_{1+kN}, \mathbf{b}_{2+kN}, \ldots, \mathbf{b}_{N+kN}]$ according to:

$$\mathbf{y}'_{ki} = \mathbf{R}'_1 \mathbf{B}_k^T \mathbf{R}'_2 \mathbf{H} \mathbf{T}_2 \mathbf{f}_i t_{11} s + \mathbf{z}'_{ki}$$

5: **End For**
6: **End For**
7: Stack the rows of $\mathbf{R}'_1 \mathbf{B}_k$, $k = 1 \ldots \lceil M/N \rceil - 1$ to get $r'_{11} \tilde{\mathbf{B}}$
8: Obtain $\mathbf{Y}' = r'_{11} \tilde{\mathbf{B}}^T \mathbf{R}'_2 \mathbf{H} \mathbf{T}_2 \mathbf{F} t_{11} s' + \mathbf{Z}'$.
9: Repeat steps 1-6 for the uplink direction and obtain

$$\mathbf{Y} = r_{11} \tilde{\mathbf{F}}^T \mathbf{R}_2 \mathbf{H}^T \mathbf{T}'_2 \mathbf{B} t'_{11} s + \mathbf{Z}.$$

10: Let $\beta = r_{11}^{-1} t_{11}'^{-1} / (r_{11}'^{-1} t_{11}^{-1})$, $\alpha_i = r_{2i} t_{2i}^{-1}$, $\alpha'_i = r'_{2i} t'^{-1}_{2i}$.
11: Obtain matrices $\mathbf{X} \doteq \tilde{\mathbf{B}}^{-T} \tilde{\mathbf{Y}}' \mathbf{F}^{-1}$ and $\mathbf{Z} \doteq \mathbf{B}^{-T} \tilde{\mathbf{Y}}^T \tilde{\mathbf{F}}^{-1}$.
12: **Solve**

$$\{\alpha_i, \alpha'_i\}_{i=1}^M = \arg \min_{\alpha_i, \alpha'_j, i,j \in [M]} \sum_{i=1}^M \sum_{j=1}^M \| x_{ij} \alpha_j - \beta z_{ij} \alpha'_i \|^2$$

13: **Return** $\mathbf{R}_2 \mathbf{T}_2^{-1} = diag\{\alpha_i\}_{i=1}^M$, $\mathbf{R}'_2 \mathbf{T}'^{-1}_2 = diag\{\alpha'_i\}_{i=1}^M$

optimization problem.

$$\{\alpha_i, \alpha'_i\}_{i=1}^M = \arg \min_{\alpha_i, \alpha'_j, i,j \in [M]} \sum_{i=1}^M \sum_{j=1}^M \| x_{ij} \alpha_j - \beta z_{ij} \alpha'_i \|^2 \tag{5.18}$$

Having found the best estimates of α_i and α'_i parameters, the mismatch calibration matrices $\mathbf{R}_2 \mathbf{T}_2^{-1}$ and $\mathbf{R}'_2 \mathbf{T}'^{-1}_2$ can be easily formed and found up to a scaling factor as the second-level calibration matrices.

5.4 Reciprocity-based DL Channel Estimation

We note that the ultimate goal of reciprocity calibration is indeed finding the downlink channel estimate based on the uplink channel estimation using the pilots transmitted by the users. In this section, we first study the DL channel estimation with a single AP and then proceed to the case of multiple APs.

5.4.1 Downlink Channel Estimation with a Single AP

Using $\mathbf{F}_{\text{mu}}$ as the transmit beamformer at the AP and $\mathbf{B}_{\text{mu}}$ as the receive beamformer at the MU in the DL direction, and similarly using $\mathbf{B}_{\text{bs}}$ as the transmit beamformer at the MU and $\mathbf{F}_{\text{bs}}$ as the receive beamformer at the AP in the UL direction the signal model at the MU and the AP is given by,

$$\mathbf{Y}_{\text{mu}} = \mathbf{R}_1' \mathbf{B}_{\text{mu}}^T \mathbf{R}_2' \mathbf{H} \mathbf{T}_2 \mathbf{F}_{\text{mu}} \mathbf{T}_1 s + \mathbf{Z}_{mu} \tag{5.19}$$

$$\mathbf{Y}_{\text{ap}} = \mathbf{R}_1 \mathbf{F}_{\text{ap}}^T \mathbf{R}_2 \mathbf{H}^T \mathbf{T}_2' \mathbf{B}_{\text{ap}} \mathbf{T}_1' s + \mathbf{Z}_{ap} \tag{5.20}$$

Since the digital calibration matrices are known up to a scaling factor, the problem reduces to finding the effective DL channel $\mathbf{H}_{\text{DL}}^{\text{eff}} \doteq \mathbf{R}_2' \mathbf{H} \mathbf{T}_2$ based on observations of $\mathbf{Y}_{ap}$ which involves the effective uplink channel $\mathbf{H}_{\text{UL}}^{\text{eff}} \doteq \mathbf{R}_2 \mathbf{H}^T \mathbf{T}_2'$. By simple algebra, one can find the effective DL channel matrix in terms of the effective UL channel matrix as

$$\mathbf{H}_{\text{DL}}^{\text{eff}} = \left(\mathbf{R}_2' \mathbf{T}_2'^{-1}\right) \left(\mathbf{H}_{\text{UL}}^{\text{eff}}\right)^T \left(\mathbf{R}_2 \mathbf{T}_2^{-1}\right)^{-1} \tag{5.21}$$

Now, it remains to estimate $\mathbf{H}_{\text{UL}}^{\text{eff}}$. We select M_{mu} transmit beamformer $\mathbf{b}_i$, $i = 1, \ldots, M_{\text{mu}}$ and M_{ap} receive beamformers $\mathbf{f}_i, i = 1, \ldots, M_{\text{ap}}$ from the transmit and receive codebook, such that the matrices $\mathbf{F} = \left[\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_{M_{\text{ap}}}\right]$ and $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{M_{\text{mu}}}]$ are full rank. For each $\mathbf{b}_i$, $i = 1, \ldots, M_{\text{mu}}$, we perform $k = 0, 1, \ldots, \lceil M_{\text{ap}}/N_{\text{ap}} \rceil - 1$ transmissions with transmit beamformers $\mathbf{b}_i$ from chain 1, and receive beamformers $\mathbf{f}_{1+k\lceil M_{\text{ap}}/N_{\text{ap}} \rceil}, \mathbf{f}_{2+k\lceil M_{\text{ap}}/N_{\text{ap}} \rceil}, \ldots, \mathbf{f}_{N_{\text{ap}}+k\lceil M_{\text{ap}}/N_{\text{ap}} \rceil}$ for the k^{th} transmissions with the beamformer $\mathbf{b}_i$, where $\mathbf{b}_m = \mathbf{b}_1$ for $m > M_{\text{mu}}$. Hence, after $M_{\text{mu}}\lceil M_{\text{ap}}/N_{\text{ap}} \rceil$ transmission we gather $M_{\text{mu}}M_{\text{ap}}$ observations that we arrange in a $M_{\text{mu}} \times M_{\text{ap}}$ matrix $\mathbf{Y}$. Using the uplink model (5.6), we can write

$$\mathbf{y}_{ki} = \mathbf{R}_1 \mathbf{F}_k^T \mathbf{R}_2 \mathbf{H} \mathbf{T}_2' \mathbf{b}_i t_{11}' s + \mathbf{z}_{ki}', \tag{5.22}$$

for all $k = 1, \ldots, \lceil M_{\text{ap}}/N_{\text{ap}} \rceil$ and $i = 1, \ldots, M_{\text{mu}}$ where the estimated values for $\mathbf{R}_1$ is already computed up to a scaling factor r_{11}. The j^{th}, $j = 1, \ldots, N_{\text{ap}}$ row of the matrix product $\mathbf{R}_1\mathbf{F}_k$ for $k = 1, \ldots, \lceil M_{\text{ap}}/N_{\text{ap}} \rceil$ is given by a row vector

$$r_{1j}\mathbf{f}_{N_{\text{ap}}(k-1)+j}^T = r_{11}(r_{1j}/r_{11})\mathbf{f}_{N_{\text{ap}}(k-1)+j}^T \doteq r_{11}\tilde{\mathbf{f}}_m^T \tag{5.23}$$

where $m = N_{\text{ap}}(k-1) + j$. Stacking these row vectors $\tilde{\mathbf{f}}_m^T$ for $m = 1, \ldots, M_{\text{ap}}$ results in an $M_{\text{ap}} \times M_{\text{ap}}$ matrix $\tilde{\mathbf{F}}$. Considering $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_{M_{\text{ap}}}]$ which consists of all M_{ap} transmit beamforming vectors and the modified received beamforming vectors in $\tilde{\mathbf{F}}$, we would get an aggregate $M_{\text{ap}} \times M_{\text{ap}}$ observation matrix $\mathbf{Y}$ as

$$\mathbf{Y} = r_{11}\tilde{\mathbf{F}}^T \mathbf{R}_2 \mathbf{H}^T \mathbf{T}_2' \mathbf{B} t_{11}' s + \mathbf{Z}. \tag{5.24}$$

where $\mathbf{B}$ is comprised of putting M_{mu} transmit beamforming vectors used by MU each in separate transmissions in one row to get an $M_{\text{mu}} \times M_{\text{mu}}$ matrix and $\tilde{\mathbf{F}}^T$ is comprised of the transpose of the modified received beamforming vectors in one column to get an $M_{\text{ap}} \times M_{\text{ap}}$ matrix. We note that the M_{mu} transmit beamforming vectors, and the M_{ap} receive beamforming vectors are chosen such that $\mathbf{B}$ and $\mathbf{F}$ (and hence, $\tilde{\mathbf{F}}$) are invertible. Let $\tilde{\mathbf{Y}}$ be the estimation of $r_{11}\tilde{\mathbf{F}}^T\mathbf{R}_2\mathbf{H}^T\mathbf{T}'_2\mathbf{B}t'_{11}$. Hence, the effective uplink channel matrix can be written as

$$\mathbf{H}_{\text{UL}}^{\text{eff}} = \left(\left(\mathbf{r}_{11}\tilde{\mathbf{F}}^T\right)^{-1}\tilde{\mathbf{Y}}\left(\mathbf{B}\mathbf{t}'_{11}\right)^{-1}\right)^T \tag{5.25}$$

5.4.2 Downlink Channel Estimation with Multiple APs

Contrary to the case of beamforming from a single AP, where achieving the calibration between the MU and the AP was enough to obtain the downlink channel estimate by combining the equations (5.21) and (5.25), we show that when multiple APs co-operate towards jointly serving an MU a third calibration step must be performed between the co-operative APs in addition to the two calibration steps outlined in the previous section. Let us consider the scenario involving two collaborating APs, namely, AP_1, and AP_2. The extension of the results to the case of multiple APs is straightforward. Suppose each of the APs estimates its DL channel to the MU separately, by performing calibration between the AP and the MU as discussed in section 5.4.1. Since the calibration matrices are only known up to a scaling factor, the DL received signal model between

each AP and the MU is given by,

$$\mathbf{y}_{\text{DL, 1}} = c_1\mathbf{H}_{\text{DL,1}}\mathbf{s} + \mathbf{z}_{DL,1} \tag{5.26}$$

$$\mathbf{y}_{\text{DL, 2}} = c_2\mathbf{H}_{\text{DL,2}}\mathbf{s} + \mathbf{z}_{DL,2} \tag{5.27}$$

where c_1 and c_2 are the unknown coefficients of the estimated downlink channel between AP_1, AP_2 and the user, respectively. The overall downlink channel from both APs is obtained by collecting all the column vectors of $c_1\mathbf{H}_{\text{DL,1}}$ and $c_2\mathbf{H}_{\text{DL,2}}$ into a single channel matrix $\mathbf{H}_{\text{DL}} = [c_1\mathbf{H}_{\text{DL,1}}, c_2\mathbf{H}_{\text{DL,2}}]$. It is important to note that in general, the coefficients c_1 and c_2 are unknown and uncorrelated parameters. Therefore, to perform cooperative beamforming we need to at least estimate the ratio c_2/c_1. Following equation (5.19), it can be easily inferred that

$$c_1 = r_{11}^{mu} t_{11}^1 \sigma_2^{mu} (\sigma_2^1 t_{11}^{mu} r_{11}^1)^{-1} \tag{5.28}$$

$$c_2 = r_{11}^{mu} t_{11}^2 \sigma_2^{mu} (\sigma_2^2 t_{11}^{mu} r_{11}^2)^{-1} \tag{5.29}$$

where r_{11}^{mu}, t_{11}^{mu}, and σ_2^{mu} are the unknown scaling factors (embedded in the first elements) of the calibration matrices $\mathbf{R}_1^{mu}$, $\mathbf{T}_1^{mu}$, and $\mathbf{R}_2^{mu}(\mathbf{T}_2^{mu})^{-1}$ for the MU. Similarly, r_{11}^k, t_{11}^k, and σ_2^k are the unknown scaling factors (embedded in the first elements) of the calibration matrices $\mathbf{R}_1^k$, $\mathbf{T}_1^k$, and $\mathbf{R}_2^k(\mathbf{T}_2^k)^{-1}$ for the k-th AP, $k = 1, 2$. Define $c \doteq c_1/c_2$, we have

$$c = \frac{r_{11}^{mu} t_{11}^1 \sigma_2^{mu} (\sigma_2^1 t_{11}^{mu} r_{11}^1)^{-1}}{r_{11}^{mu} t_{11}^2 \sigma_2^{mu} (\sigma_2^2 t_{11}^{mu} r_{11}^2)^{-1}} = \frac{t_{11}^1 (\sigma_2^1 r_{11}^1)^{-1}}{t_{11}^2 (\sigma_2^2 r_{11}^2)^{-1}} \tag{5.30}$$

It is observed that the coefficient c is not a function of the calibration parameters of the

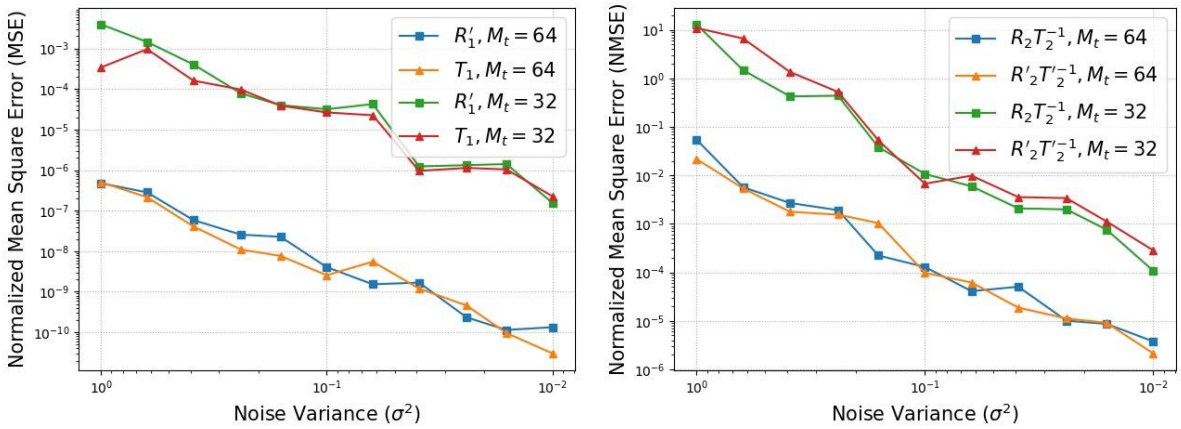
MU and only depends on the calibration parameters of the APs. Suppose the two-step calibration process between the APs is performed and the two APs have estimated the calibration matrices up to a scaling factor. We show that the coefficient c in (5.30) can be obtained by taking a third step in the calibration process. In the following, we show employing the notion of reciprocal tandem may facilitate the estimation of this ratio efficiently by only two pilot transmissions, one from each AP. AP_1 transmits a pilot signal from its first digital RF chain using the beamforming vector $\mathbf{f}_1^1$ that is received by AP_2 on its first digital RF chain using beamforming vector $\mathbf{b}_1^2$. In the reverse direction, AP_2 transmits a pilot signal from its first digital RF chain using $\mathbf{b}_2^2$ which is the scaled version of the reciprocal tandem of $\mathbf{b}_1^2$ and AP_1 receives the signal in its first digital RF chain using the beamforming vector $\mathbf{f}_2^1$ which is the scaled version of the reciprocal tandem of $\mathbf{f}_1^1$. Since $\mathbf{R}_2^2(\mathbf{T}_2^2)^{-1}$ is known up to a scaling factor, the scaling factor is taken as its first diagonal element which is denoted by σ_2^2 and hence the estimated matrix can be written as $\mathbf{R}_2^2(\mathbf{T}_2^2)^{-1}/\sigma_2^2$. As a result, we can compute $\mathbf{b}_2^2 = (\sigma_2^2)^{-1}\check{\mathbf{b}}_1^2 = (\sigma_2^2)^{-1}\mathbf{R}_2^2(\mathbf{T}_2^2)^{-1}\mathbf{b}_1^2$. Similarly, we can compute $\mathbf{f}_2^1 = (\sigma_2^1)\check{\mathbf{f}}_1^1 = ((\sigma_2^1)^{-1}\mathbf{R}_2^1(\mathbf{T}_2^1)^{-1})^{-1}\,\mathbf{f}_1^1$. Let the observations in both directions be $\tilde{y}_{12}$ and $\tilde{y}_{12}$. We get,

$$\tilde{y}_{12} = r_{11}^2 \left(\mathbf{b}_1^2\right)^T \mathbf{R}_2^2\mathbf{H}\mathbf{T}_2^1\mathbf{f}_1^1 t_{11}^1 \tag{5.31}$$

$$\tilde{y}_{21} = r_{11}^1 \left(\mathbf{f}_2^1\right)^T \mathbf{R}_2^1\mathbf{H}^T\mathbf{T}_2^2\mathbf{b}_2^2 t_{11}^2 \tag{5.32}$$

Transposing the RHS of (5.32) and replacing $\check{\mathbf{b}}_2^2$ and $\check{\mathbf{f}}_2^1$ with their equivalent values, we can rewrite it as

$$\tilde{y}_{21} = t_{11}^2(\sigma_2^2)^{-1} \left(\mathbf{b}_1^2\right)^T \mathbf{R}_2^2\mathbf{H}\mathbf{T}_2^1\sigma_2^1\mathbf{f}_1^1 r_{11}^1 \tag{5.33}$$



(a) First-level Calibration NMSE

(b) Second-level Calibration NMSE

Figure 5.2: Reciprocity Calibration Normalized MSE (NMSE)

The ratio of the scalar observations $\tilde{y}_{12}$ and $\tilde{y}_{21}$, would give

$$\tilde{y}_{12}/\tilde{y}_{21} = \frac{t^1_{11}(\sigma^1_2)^{-1}(r^1_{11})^{-1}}{t^2_{11}(\sigma^2_2)^{-1}(r^2_{11})^{-1}} \tag{5.34}$$

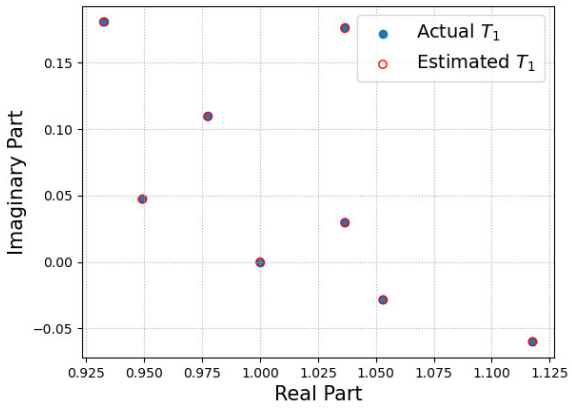
that can directly be used to estimate the parameter c. It is easy to extend this analysis to the case of multiple APs by picking one AP as the reference and calibrating the rest of the APs with respect to the reference one.

5.5 Performance Evaluation

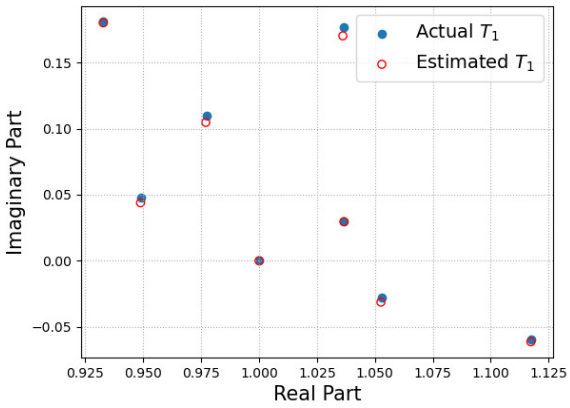
In this section, we first describe the simulation setup and then proceed to the analysis of the numerical results.

5.5.1 Simulation Setup and Parameters

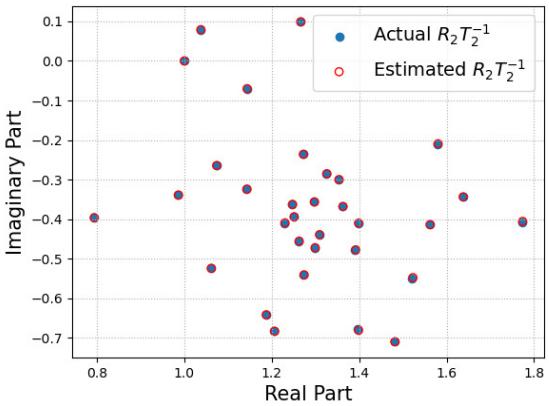
We consider two nodes with $M_t = 32, 64$ and $M_r = 32, 64$ antenna elements and $N_t = M_t/4$, and $N_r = M_r/4$ digital RF chains, respectively. We consider a multi-path channel model



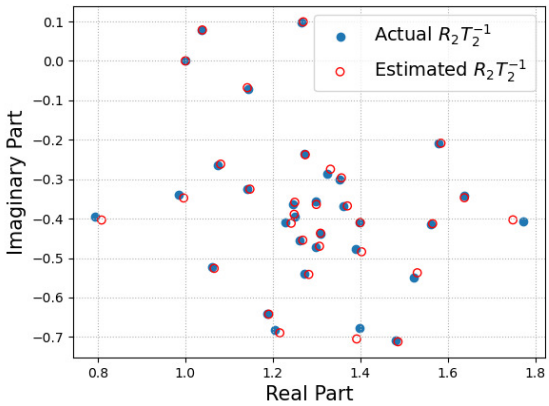
(a) First-level Calibration Performance, $\sigma = 0.01$.



(b) First-level Calibration Performance, $\sigma = 0.1$



(c) Second-level Calibration Performance, $\sigma = 0.01$



(d) Second-level Calibration Performance, $\sigma = 0.1$

Figure 5.3: Channel Reciprocity Calibration Scheme Performance

with $L = 4$ paths. We assume the gain of each path $\alpha_\ell, \ell = 1, \ldots, 4$ follows a Gaussian distribution with mean zero and variance $\sigma_\alpha^2 = 1$. The AoAs/AoDs are sampled from a uniform distribution $\{\theta_k, \phi_k\} \sim \mathcal{U}(-\pi/2, \pi/2)$. We assume the reciprocity mismatch gains follow a log-normal distribution, i.e. $\left\{\ln|t_{i,n}|, \ln|r_{i,n}|, \ln\left|t'_{i,n}\right|, \ln\left|r'_{i,n}\right|\right\} \sim \mathcal{N}(0, \sigma), i \in \{1, 2\}, n = 1 \ldots N$, for standard deviation σ. Similarly, the phase of the mismatch parameters follows a uniform distribution $\left\{\angle t_{i,n}, \angle r_{i,n}, \angle t'_{i,n}, \angle r'_{i,n}\right\} \sim \mathcal{U}(-\pi/16, \pi/16)$. Our tests are carried out on a server with an Intel i9 CPU at 2.3 GHz and 16 GB of main memory.

5.5.2 Reciprocity Calibration between Two nodes

In this experiment, we implement the calibration process described in sections 5.3.1 and 5.3.2 under different circumstances. Fig. 5.2(a, b) depict the normalized mean-square error (NMSE) resulting from the calibration technique in estimating each of the calibration matrices for the digital and analog RF chains, versus the transmission noise variance, respectively. The actual and estimated calibration matrices are normalized with respect to the first element of the matrix to ignore the constant scaling factor in calculating the NMSE. We observe that in both steps, as the number of antennas per AP increases the NMSE curves are shifted downwards (NMSE improved by a factor of 1000), as there will be more observations which allow for making a more precise estimate of the calibration matrices.

Fig. 5.3 provides a visualization of the effectiveness of our proposed calibration approach under various noise variances. In Fig. 5.3, we consider $M = 32$ antenna elements and $N = 8$ digital RF chains at each AP. The scatter plots in Fig. 5.3(a,b) show the performance of the first step of the calibration process when estimating the calibration matrix T_1 for $\sigma = 0.01$ and $\sigma = 0.1$, respectively. It is observed that, in either case, the calibration parameters are obtained with high accuracy and as the noise variance gets smaller the estimation error reduces. In the second step of the calibration, one is interested in $R_2 T_2^{-1}$ instead of the calibration matrices R_2 and T_2 itself. Fig. 5.3(c,d) presents similar results for the second step of the calibration.

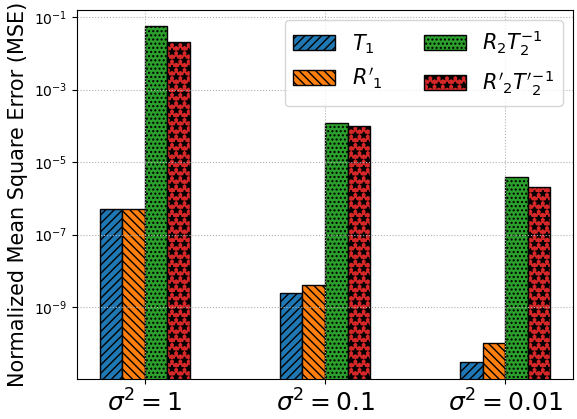


Figure 5.4: Cal. MSE vs. Noise Variance

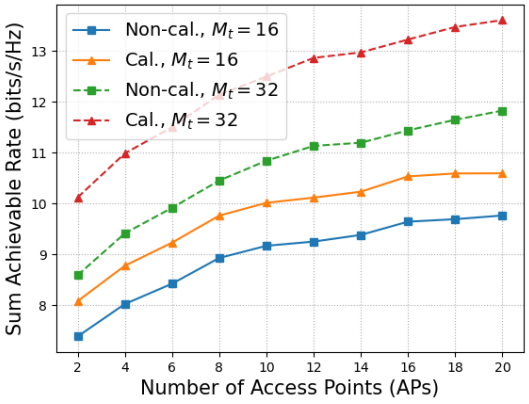


Figure 5.5: Sum Rate under varying K

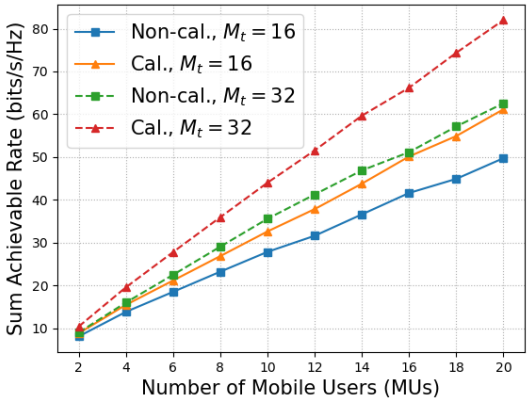
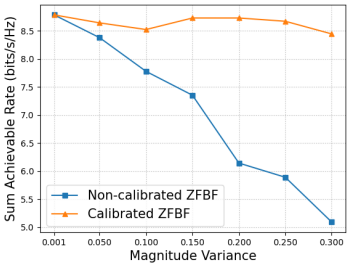
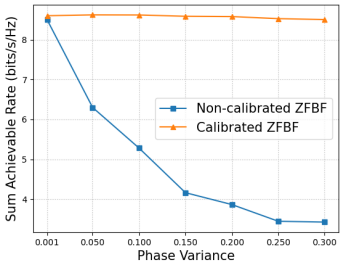


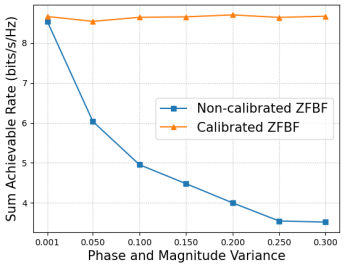
Figure 5.6: Sum Rate under varying U



(a) Varying magnitude



(b) Varying phase



(c) Varying phase and magnitude

Figure 5.7: Sum Rate Performance under varying Mismatch with $K = U = 2$

5.5.3 Co-operative Zero-forcing Beamforming (ZFBF)

Consider a multi-AP system with K APs each employing N antennas that are serving a group of U MUs under ZFBF. The received signal at user u is given by,

$$\mathbf{y} = \mathbf{H}_{\text{DL}}^{T}\mathbf{W}\mathbf{D}\mathbf{s} + \mathbf{z}_{\text{DL}} \tag{5.35}$$

where $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_U] \in \mathbb{C}^{KN \times U}$, and $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_U] \in \mathbb{C}^{KN \times U}$ denote the downlink channel matrix, and precoding matrix, $\mathbf{D} = \text{diag}\left(\sqrt{P_1}, \ldots, \sqrt{P_U}\right) \in \mathbb{R}^{U \times U}$ is the diagonal power matrix, and $\mathbf{s} = [s_1, \ldots, s_U]^T \in \mathbb{C}^U$ is the transmit signal. Under zero-force precoding, for every user u, the corresponding precoder $\mathbf{w}_u$ is orthogonal to all the channel vectors $\mathbf{h}_v$ associated with other users $v \neq u$. This way the interference resulting from the desired transmission for a user is suppressed for other users when the aggregate signal passes through the channel. In the matrix form, the orthogonality condition can be stated as $\mathbf{H}^T\mathbf{W} = \mathbf{Q}$ where $\mathbf{Q}$ is usually picked as the Identity matrix. Therefore, the optimal $\mathbf{W}$ can be found as the pseudo-inverse of matrix $\mathbf{H}^T$ as $\mathbf{W} = \mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}$. In this experiment, we simulate a cooperative ZFBF scenario in a multi-user setup with and without calibration. We consider the achievable sum rate of the users to compare the performance of ZFBF with and without calibration. The sum rate is computed as $r = \sum_{u=1}^{U} \log(1 + SNR_u)$, where SNR_u is the output signal-to-noise ratio corresponding to user u. When no calibration technique is in place, the beamforming precoders are decided based on the uplink channel estimate which is not necessarily a reasonable approximation of the downlink channel depending on the level of imperfections. This will result in lower output SNR and therefore lower user sum rate compared to the case where calibration is in place and the

downlink channel is accurately estimated.

Fig. 5.4 shows the high accuracy of our calibration technique under different noise levels by depicting the mean squared error (MSE) of the estimated calibration matrices where all the estimated matrices are normalized with respect to their first elements. Fig. 5.5 shows how the user sum rate evolves by increasing the number of APs when there are 2 MUs targeted. seen that the channel reciprocity calibration enhances the user sum rate by 20% and 30% for $M_t = 16$ and $M_t = 32$, respectively, when averaged over different cases on the number of APs in the distributed massive MIMO network. Fig. 5.7b investigates the same effect as the number of MUs increases where there are 2 APs in place. We observe that as the number of targeted MUs increases the impact of calibration becomes more visible. We conclude that channel reciprocity calibration in multi-user scenarios is drastically important, and its importance gains more attention in densely populated environments.

Fig. 5.7 measures the impact of our calibration scheme on the performance of ZFBF at different levels of reciprocity mismatch where there are 2 MUs and 2 APs. In figures 5.7a and 5.7b, the magnitude and the phase of reciprocity mismatch are changing such that the intensity of the hardware imperfection increases. We observe that as these imperfections worsen (both in magnitude and phase), the performance of the non-calibrated ZFBF degrades and the users achieve a lower rate. however, when our calibration technique is in place the achieved rate remains constant. Fig. 5.7c shows this effect when the magnitude and the phase of the mismatch coefficients tend to deviate from their ideal values simultaneously. The trend in this figure is closer to the trend in Fig. 5.7b which indicates that the CHBF is more vulnerable to phase mismatch.

Chapter 6: Codebook Design for mm-wave Beamforming in 5G and Beyond

6.1 Overview

With the exponential growth in the number and the diversity of broadband applications in next-generation communication systems, the ever-increasing need to explore higher bandwidths reveals the pivotal role of mmWave communications in future wireless networks. However, given the high path loss and poor scattering associated with mmWave communications, effective beamforming techniques integrating a large number of antennas (massive MIMO) are required to ensure satisfactory QoS. The 5th generation of wireless cellular communication systems relies on massive MIMO as a key component in the realization of the physical layer. Estimating the channel state information (CSI) at the receiver is a bottleneck in massive MIMO systems since the overhead of such estimation wipes out the potential gain that can be achieved through such knowledge. Therefore, two-layer beamforming is used to split the beamforming task into two steps; (i) first-layer or baseband beamforming over which the CSI is obtained by inserting pilots into the transmitted streams, and (ii) second-layer beamforming which generates a fixed set of beams chosen from a codebook.

Therefore, feedback-based beamforming techniques are employed for efficient mmWave communications. On the other hand, the cost of having one RF chain per transmit antenna is prohibitive. Hence, not only does the multitude of RF chains for massive MIMO systems

incur a considerable cost and power consumption, but also it is not going to be used to achieve multiplexing gain as the full CSI is not practically available. Therefore more practical system designs consider reducing the number of RF chains to achieve lower power consumption and lower cost. Such designs that only employ a single RF chain are denoted as analog beamforming structures in the literature, while hybrid beamforming is reserved for designs consisting of a few RF chains. In hybrid beamforming, the beamformer consists of two layers. At the first layer, the baseband beamforming unit (BBU) performs digital beamforming, i.e. controls both the gain and the phase of the input symbol, while at the second layer, the radio remote head (RRH) only performs phase shifts, i.e. realizes the analog beamforming.

In the literature, some works mainly focus on the channel estimation problem either by using feedback from the receiver [120] [121] [122], or by using prior assumptions on the channel statistics such as the channel covariance matrix, its sparsity, low rank, etc. [123] [124] [125]. To relax the channel estimation challenges, a line of research suggests employing a hierarchical multi-level beamforming and codebook design, where the AoD/AoA interval is split into disjoint regions and is searched sequentially at increasing resolution using multi-level codebooks and sounding the corresponding beamformers [122] [126] [127]. The beam search problem is modeled and solved in [122] as a Neyman-Pearson (NP) detection problem with the objectives of maximizing the data rate and spectral efficiency, while the authors in [122] propose a DFT-based hybrid codebook design by the objective of minimizing the gain variance at main lobe, and the authors in [123] [126] aim at minimizing the distance of the designed gain from the ideal one. More recently, another line of work has started efforts in employing machine learning-based methods for channel estimation and hybrid beamforming in MIMO systems [128] [129] [130] [131].

Due to the nature of mmWave channels as being largely line of sight and having only a

few dominant paths, physical beamforming (beam steering) is a practical and effective way. The communication beams are designed to have maximum gain toward the direction of the angle of departure of the user channel. Often, a beam search procedure, e.g., sequential beam search [132], hierarchical beam search [122], interactive [133], and non-interactive beam search [134], is used to find the best communication beam. The proposed algorithms in the literature often ignore the effect of multipath. Recent works [135] have shown that it is in fact possible to exploit multipath to our benefit to find more robust beams that are less susceptible to blockage and shadowing. The key idea in [135] is to design a *composite beam* that has multiple lobes that cover the dominant paths of the user channel.

In this chapter, we study the codebook design problem considering both the digital and hybrid beamforming approaches [136] [137]. Under the fully digital scheme, We first consider a uniform linear array (ULA) of antennas and design a codebook with the objective of minimizing the mean-square error (MSE) between the desired (ideal) beam gain computed a priori, and the resulting beam gain from the devised beamforming vector. Next, having the digital beamforming codebook, we design the hybrid beamforming codebook using the simple yet effective orthogonal matching pursuit (OMP) algorithm. In contrast to the methods described in [126], and [123] that utilize the quantization of the angular space, our method results in a super low complexity closed-form solution that is found analytically. Moreover, we argue that the beam generated by a ULA is intrinsically two-sided and will extend symmetrically with respect to the axis of the ULA, i.e., the line passing from the base of antenna elements. This will result in a significant design suboptimality as (i) the beams cannot cover only the desired angular interval, and (ii) due to the redundant propagation in the symmetric lobe, lowers the gain on the main arc. To deal with this issue, we introduce a novel antenna structure, namely, *twin*-ULA (TULA) which consists of

two ULAs that are placed in the 2D plane parallel to each other along one axis and with a fixed distance along the other axis. We show that our new design under this setting not only suppresses the undesired beams but also improves the gain of the desired main lobe. TULA is different from the uniform planar antenna arrays (UPA) in the following senses: (1) TULA is not uniform (2) the plane of UPA is orthogonal to the azimuth plane while the plane of TULA is the azimuth plane. Indeed, to shape the elevation angle in addition to the azimuth angle, we introduce the novel structure of twin-UPA which consists of two UPA's with parallel planes, however, the analysis of this structure is out of the scope of this dissertation. The driven low-complexity analytical solution for the beamforming in ULA, TULA is also extended to two more elaborate antenna structures namely delta and star configurations. The analytical solution eliminates the need for an exhaustive search (e.g., see [123] [126] for ULA) for optimizing the beams.

Furthermore, we propose an algorithm to design *composite beams*, i.e., beams that are comprised of multiple non-neighboring angular coverage intervals (ACIs), say in the azimuth direction, of possibly different widths [138] [139]. The composite beams are not only important as data communication beams, but they can also facilitate the beam search process. A codebook of composite beams, i.e. a *composite codebook*, is designed for a set of composite beams that are defined over a set of desired ACIs *ACI set*. Each entry of the codebook is a beamforming vector that generates a composite beam defined as a beam that covers a union of disjoint ACIs out of a set of all desired ACIs.

Such a composite codebook can be used in a variety of applications in next-generation mmWave communications such as user tracking [140], target monitoring [141], 5G positioning, two-way communications [142], design of reconfigurable intelligent surfaces [143], UAV-enabled networks [89], etc. The composite codebook design problem may be also viewed as a generalized

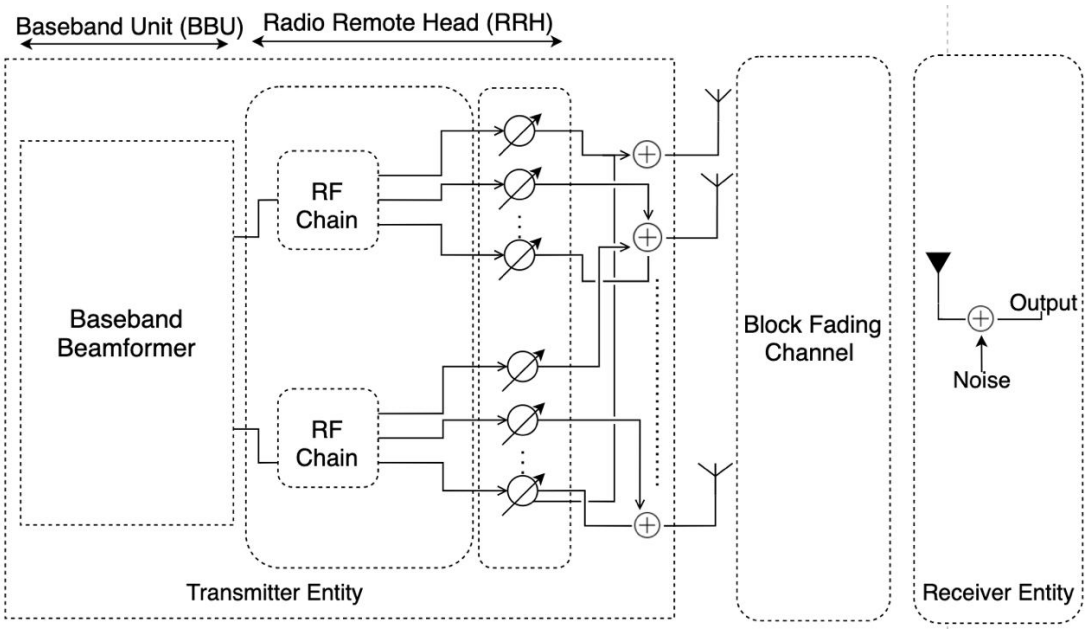


Figure 6.1: System Model

version of the codebook design problem [126] [122] where the angular range under study is divided into equal-length ACIs and each codeword is supposed to cover only a single ACI.

6.2 System Model

We consider a multiple-input single-output (MISO) system consisting of a base station (BS) as the transmitter, and a mobile user (MU) as the receiver as in Fig. 7.1. Given the practical limits of the MU equipment, stemming from its small antenna size, it makes sense to focus our attention on only optimizing the transmitter side. Nonetheless, most of the ideas presented in this chapter can also be applied to the receiver side. We consider M_t antennas at the transmitter and a single antenna at the receiver.

6.2.1 Channel Model

The channel model is given as,

$$y = \sqrt{\rho}\mathbf{h}^H\mathbf{c}s + n \tag{6.1}$$

with $s \in \mathbb{C}$ being the input symbol satisfying the power constraint (i.e. $\mathrm{E}\left[|s|^2\right] \leq 1$), $\mathbf{c} \in \mathbb{C}^{M_t}$ the unit-norm beamforming vector, $\mathbf{h} \in \mathbb{C}^{M_t}$ the block fading channel vector, and ρ the system signal-to-noise ratio (SNR), and the complex additive white Gaussian noise $n \sim \mathcal{CN}(0, 1)$.

6.2.2 Beamforming Model

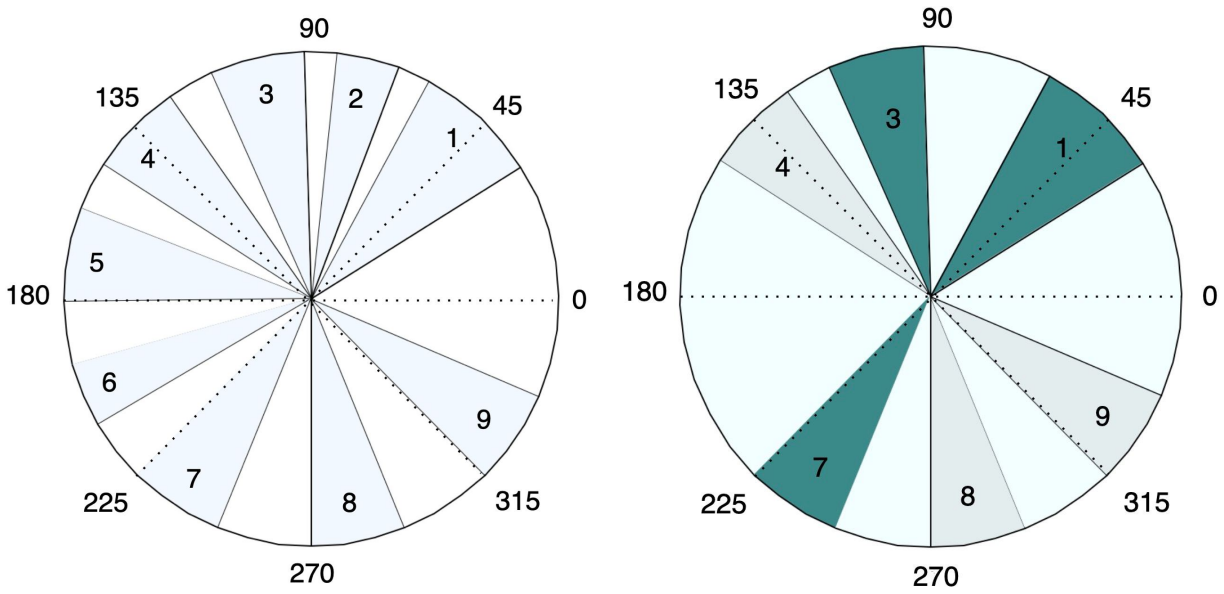
We assume under fully-digital beamforming, both the gain and the phase of the antenna elements, i.e. beamforming coefficients, can be controlled using a single streamed transmission, i.e., $N_{RF} = 1$. On the other hand, the RRH may only consist of phased-array antennas realizing analog beamforming, where only the phase of the antenna elements can be controlled. Under hybrid beamforming, we assume the RRH exploits phased array antennas, where multiple streams i.e. $N_{RF} > 1$, may be transmitted through multiple RF chains. In this case, the beamformer $\mathbf{c}$ is given by $\mathbf{c} = \mathbf{F}\mathbf{v}$ where $\mathbf{F} = [\mathbf{f}_1, \cdots, \mathbf{f}_{N_{RF}}] \in \mathbb{C}^{M_t \times N_{RF}}$ is the analog beamsteering matrix and $\mathbf{v} \in \mathbb{C}^{N_{RF}}$ denotes the symbol transmitted by N_{RF} streams. The first-layer digital beamformer at the BBU controls both the gain and the phase and generates N_{RF} streams and the second-layer beamformer is a phased array antenna. Nonetheless, it is important to note that in either scenario the pilot transmission and channel estimation can only be done on the baseband. By the unit-norm constraint on $\mathbf{c}$, it must hold that $\|\mathbf{F}\mathbf{v}\| = 1$. Moreover, by using phased-array antennas, the

beam-steering matrix $\mathbf{F}$ realizes only a set of phase shifts, i.e., all the vectors $\mathbf{f}_n, n = 1 \dots N_{RF}$ are subject to an *equal gain* constraint defined as $|\mathbf{f}_n^{(m)}| = 1, m = 1 \dots M_t$. We assume each codebook $\mathcal{C}$, consists of $Card(\mathcal{C}) = Q$ codewords. Once a codeword $\mathbf{c}_q$ is designed under the fully digital scheme, $\mathbf{F}_q$, and $\mathbf{v}_q$ must be found subject to the above constraints to realize the hybrid beamforming scheme.

We consider the design of physical beams that are steered in the azimuth plane where the beams are supposed to cover one (or multiple disjoint) ACIs. An ACI covering the angular range from θ_b^s to θ_b^f is denoted by $\omega_b = [\theta_b^s, \theta_b^f)$ where $\theta_b^c = (\theta_b^s + \theta_b^f)/2$, and $\lambda_b = |\theta_b^s - \theta_b^f|$ are the center and beamwidth associated with the beam lobe covered by this ACI. We assume $\theta \in [-\pi, 0]$. Further, let us introduce the change of variable $\psi = \pi cos\theta$. We have $\psi \in [-\pi, \pi]$, and each beam is represented by $\omega_b^\psi = [\psi_b^s, \psi_b^f)$ in the $\psi -$ domain, where $\psi_b^a = \pi cos\theta_b^a$, $a \in \{s, f\}$. Further define $\delta_b = |\psi_b^s - \psi_b^f|$. For the rest of the paper, we prefer to work with the beams over the $\psi -$ domain, unless otherwise stated.

Let $\mathcal{A}$ denote the set of ACIs defined for a given codebook design problem. Let each element of $\mathcal{A}$ be denoted by an index. A beam is denoted by $\mathcal{B}(B) = \{w_b\}_{b \in B}$, where B is the set of indices of non-neighboring ACIs. A set of ACIs is non-neighboring if any pair of ACIs in that set are not neighbors, i.e., the starting angle of one beam is not equal to the ending angle of the other beam. A beam is called *single beam* (*composite beam*) if $|B| = 1$ ($|B| > 1$).

Example 1. One example of this setting is shown in fig. 6.2. Particularly, fig. 6.2a depicts a



(a) Angular Coverage Intervals (ACIs) (b) Two Composite Beams

Figure 6.2: Example of the Codebook Design Problem Settings

hypothetical ACI set given as the input to the codebook design problem where we have,

$$\mathcal{A} = \Big\{[\frac{\pi}{6}, \frac{\pi}{3}), [\frac{19\pi}{48}, \frac{23\pi}{48}), [\frac{\pi}{2}, \frac{5\pi}{8}), [\frac{11\pi}{16}, \frac{13\pi}{16}), [\frac{7\pi}{8}, \pi), [\frac{-7\pi}{8}, \frac{-19\pi}{24}), [\frac{-3\pi}{4}, \frac{-5\pi}{8}), [\frac{-\pi}{2}, \frac{-3\pi}{8}), [\frac{-\pi}{4}, \frac{-\pi}{8})\Big\}$$

and Fig. 6.2b depicts two potentially desired composite beams $\mathcal{B}(B_1), \mathcal{B}(B_2) \subseteq \mathcal{A}$. We have,

$$\mathcal{B}(B_1) = \Big\{[\frac{\pi}{6}, \frac{\pi}{3}), [\frac{\pi}{2}, \frac{5\pi}{8}), [\frac{-3\pi}{4}, \frac{-5\pi}{8})\Big\}, \ \mathcal{B}(B_2) = \Big\{[\frac{11\pi}{16}, \frac{13\pi}{16}), [\frac{-\pi}{2}, \frac{-3\pi}{8}), [\frac{-\pi}{4}, \frac{-\pi}{8})\Big\}$$

where, $B_1 = \{1, 3, 7\}$, and $B_2 = \{4, 8, 9\}$.

Example 2. One may be interested in designing a codebook where the beamwidth of each beam, whether single or composite, has a resolution of say $\pi/8$ and is not larger than $\pi/2$. In this case, the ACI set $\mathcal{A}$ consists of all ACIs like $\omega_b = [\theta_b^s, \theta_b^f)$ with beamwidth $\lambda_b \in \{\pi/8, \pi/4, 3\pi/8, \pi/2\}$

and $\theta_b^s = k\pi/8$, $k \in \mathbb{Z}_{\geq 0}$.

The first example shows the intuition behind our beamforming setting given an arbitrary ACI list. The second example highlights the case where the ACIs are overlapping. Further, we say a vector $\mathbf{g}$ of length L is a *geometric vector* with parameter ρ if and only if it can be written as $\mathbf{g} = [1, e^{j\rho}, \cdots, e^{j(L-1)\rho}]$.

6.2.3 Antenna Array Model

We assume the angle of departure (AoD) θ, is uniformly distributed over the range $[-\pi, 0]$. Suppose the antennas are placed at $\mathbf{r}_m$, $m = 0, \ldots, M_t - 1$, for some $2D$ vector $\mathbf{r}_m$. For every beamforming vector $\mathbf{c}$, the gain of the antenna array at every AoD θ is proportional to

$$G(\mathbf{c}, \theta) = \left| \sum_{m=0}^{M_t-1} c_m e^{-j\frac{2\pi}{\lambda}[\cos\theta, \sin\theta]\mathbf{r}_m} \right|^2 \tag{6.2}$$

In this chapter, we consider various configurations for the antenna array, namely, *ULA*, *TULA*, *Delta(Δ)*, and *Star* (see Fig. 6.3) that are to be specified shortly.

A ULA is defined as $\mathbf{r}_m = md\mathbf{e}_x$, for some $d \in \mathbb{R}^+$, thus we have

$$G^{ula}(\mathbf{c}, \theta) = \left| \sum_{m=0}^{M_t-1} c_m e^{j\frac{2\pi md}{\lambda}\cos\theta} \right|^2 \tag{6.3}$$

In this chapter, we consider a half-wavelength ULA, i.e. $d = \frac{\lambda}{2}$. Furthermore, let us introduce a change of variable, $\psi = \pi\cos\theta$ where $\psi \in [-\pi, \pi]$. It is then straightforward to write

$$G^{ula}(\psi, \mathbf{c}) = \left| \mathbf{d}_{ula,M_t}^H(\psi)\mathbf{c} \right|^2 \tag{6.4}$$

where

$$\mathbf{d}_{ula,M_t}(\psi) = [1, e^{j\psi}, \ldots, e^{j(M_t-1)\psi}]^T \tag{6.5}$$

A TULA of M_t antennas is defined as,

$$\begin{aligned} \mathbf{r}_m &= md\mathbf{e}_x, \quad m = 0, \ldots, \frac{M_t}{2} - 1 \\ \mathbf{r}_m &= md\mathbf{e}_x + d_y\mathbf{e}_y, \quad m = \frac{M_t}{2}, \ldots, M_t - 1 \end{aligned} \tag{6.6}$$

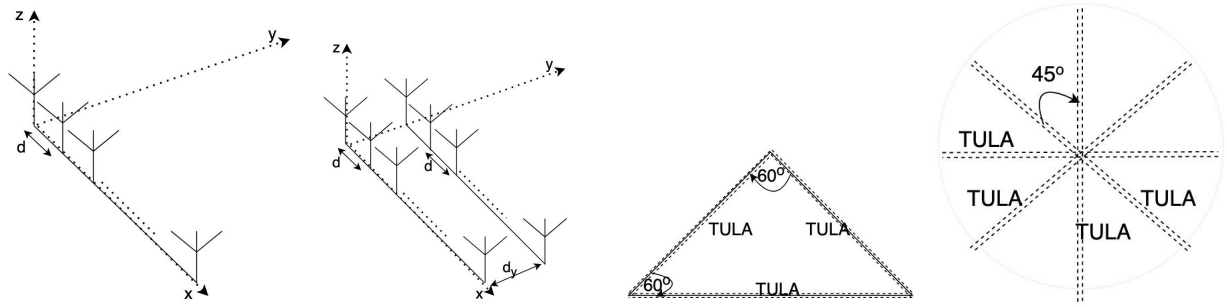
In this chapter, we set $d_y = \frac{\lambda}{3}$, and introduce another auxiliary variable $\phi = \frac{2\pi}{3}\sin\theta$. In the following equation (6.2) for the gain of the TULA, we can derive

$$G^{tula}(\mathbf{c}, \theta) = \left|\mathbf{d}^H_{tula,M_t}\mathbf{c}\right|^2 \tag{6.7}$$

where

$$\mathbf{d}_{tula,M_t}(\psi, \phi) = \left[\mathbf{d}^T_{ula,\frac{M_t}{2}}(\psi), \quad e^{j\phi}\mathbf{d}^T_{ula,\frac{M_t}{2}}(\psi)\right]^T \tag{6.8}$$

As fig. 6.3 depicts, a Delta (Δ) configuration of antennas is defined as an equilateral triangle consisting of 3 TULA of antennas, one on each side. A Star configuration consists of 4 co-centric TULA legs each of which being a $\frac{\pi}{4}$-rotation of its predecessor. The spacing between the antennas remains as explained before. In the next section, we formulate the codebook design problem for a ULA and explain the motivation behind using the other configurations.



(a) ULA (b) TULA (c) Delta(Δ) (d) Star

Figure 6.3: Antenna Array Model

6.3 Single-beam Codebook Design Problem Formulation

For convenience, in this section, we will drop the index *ula* from the expression of the array factor $\mathbf{d}_{ula,M_t}(\psi)$ and gain $G^{ula}(\psi, \mathbf{c})$. Before presenting the formulation of the codebook design problem, we introduce some new notations. We divide the angular range of $\theta \in [-\pi, 0]$ into equal-length beams ω_q for $q \in \{1, \cdots, Q\}$. i.e.

$$\omega_q = [\theta_{q-1}, \theta_q), \quad \theta_q = -\pi + \frac{\pi}{Q} q$$

Corresponding to ω_q intervals, there exists ν_q ranges with respect to ψ such that,

$$\nu_q = [\psi_{q-1}, \psi_q), \quad \psi_q = -\pi \cos(\frac{\pi}{Q} q)$$

Under the reference gain as in (6.4) and using Parseval's theorem [144], we will have:

$$\int_{-\pi}^{\pi} G(\psi, \mathbf{c}) d\psi = 2\pi \left\|\mathbf{c}\right\|^2 = 2\pi \tag{6.9}$$

Let $G_{\text{ideal },q}(\psi)$ denote the desired ideal gain which is supposed to be constant on ν_q and zero otherwise. It must hold that,

$$\int_{-\pi}^{\pi} G_{\text{ideal },q}(\psi)d\psi = \int_{\nu_q} td\psi + \int_{[-\pi,\pi]\setminus\nu_q} 0d\psi = (\psi_q - \psi_{q-1})t = 2\pi \tag{6.10}$$

which in turn will give:

$$G_{\text{ideal },q}(\psi) = \frac{2\pi}{(\psi_q - \psi_{q-1})}\mathbb{1}_{\nu_q}(\psi), \quad \psi \in [-\pi, \pi] \tag{6.11}$$

For each specific beam q, we aim to design the codeword so as to mimic the ideal gain computed in equation (6.11). Therefore, the plain codebook design problem is formulated as the minimization of an MSE as follows:

$$\mathbf{c}_q^{opt} = \underset{\mathbf{c},\|\mathbf{c}\|=1}{\arg\min} \int_{-\pi}^{\pi} \left|G_{\text{ideal },q}(\psi) - G(\psi, \mathbf{c})\right| d\psi \tag{6.12}$$

By uniformly sampling on the range of ψ we can rewrite the optimization problem in (6.43) as follows,

$$\mathbf{c}_q^{opt} = \underset{\mathbf{c},\|\mathbf{c}\|=1}{\arg\min} \left[\lim_{L\to\infty} \sum_{p=1}^{Q}\sum_{\ell=1}^{L} \frac{\delta_p \left|G_{\text{ideal },q}\left(\psi_{p,\ell}\right) - G\left(\psi_{p,\ell}, \mathbf{c}\right)\right|}{L}\right] \tag{6.13}$$

where for $q = 1 \ldots Q$,

$$\delta_q = \psi_q - \psi_{q-1}, \quad \psi_{q,\ell} = \psi_{q-1} + \frac{\delta_q(\ell - 0.5)}{L}$$

We can write equation (6.13), as

$$\mathbf{c}_q^{opt} = \underset{\mathbf{c},\|\mathbf{c}\|=1}{\arg\min} \lim_{L\to\infty} \frac{1}{L} \left|\mathbf{G}_{\text{ideal },q} - \mathbf{G}(\mathbf{c})\right| \tag{6.14}$$

where,

$$\mathbf{G}(\mathbf{c}) = \left[\delta_1 G\left(\psi_{1,1}, \mathbf{c}\right) \ldots \delta_Q G\left(\psi_{Q,L}, \mathbf{c}\right)\right]^T \in \mathbb{Z}^{LQ} \tag{6.15}$$

$$\mathbf{G}_{\text{ideal },q} = \left[\delta_1 G_{\text{ideal },q}\left(\psi_{1,1}\right) \ldots \delta_Q G_{\text{ideal },q}\left(\psi_{Q,L}\right)\right]^T \in \mathbb{Z}^{LQ} \tag{6.16}$$

Note that it holds that

$$\mathbf{G}_{\text{ideal },q} = 2\pi \left(\mathbf{e}_q \otimes \mathbf{1}_{L,1}\right) \tag{6.17}$$

with $\mathbf{e}_q \in \mathbb{Z}^Q$ being the standard basis vector for the q-th axis among Q ones. Now, note that $\mathbf{1}_{L,1} = \mathbf{g} \odot \mathbf{g}^*$ for any equal gain $\mathbf{g} \in \mathbb{C}^L$. Therefore, we can write:

$$\mathbf{G}_{\text{ideal },q} = 2\pi \left(\mathbf{e}_q \otimes \left(\mathbf{g} \odot \mathbf{g}^*\right)\right) = \left(\sqrt{2\pi}\left(\mathbf{e}_q \otimes \mathbf{g}\right)\right) \odot \left(\sqrt{2\pi}\left(\mathbf{e}_q \otimes \mathbf{g}\right)\right)^* \tag{6.18}$$

Similarly, it is straightforward to observe,

$$\mathbf{G}(\mathbf{c}) = \left(\mathbf{D}^H \mathbf{c}\right) \odot \left(\mathbf{D}^H \mathbf{c}\right)^* \tag{6.19}$$

where

$$\mathbf{D} = \left[\sqrt{\delta_1}\mathbf{D}_1 \cdots \sqrt{\delta_Q}\mathbf{D}_Q\right] \in \mathbb{C}^{M_t \times LQ} \tag{6.20}$$

$$\mathbf{D}_q = \left[\mathbf{d}_{M_t}\left(\psi_{q,1}\right) \cdots \mathbf{d}_{M_t}\left(\psi_{q,L}\right)\right] \in \mathbb{C}^{M_t \times L}. \tag{6.21}$$

Comparing the expressions (8.28), (6.18), and (8.36), one can show that the optimal choice of $\mathbf{c}_q$ in (8.28) is the solution to the following optimization problem for a proper choice of $\mathbf{g}_q$.

Problem 15 *Given an equal-gain vector $\mathbf{g}_q \in \mathbb{C}^L$, $q \in \{1, \cdots, Q\}$, find vector $\mathbf{c}_q \in \mathbb{C}^{M_t}$ such that*

$$\mathbf{c}_q = \underset{\mathbf{c}, \|c\|=1}{\arg\min} \lim_{L \to \infty} \left\| \sqrt{2\pi}\left(\mathbf{e}_q \otimes \mathbf{g}_q\right) - \mathbf{D}^H \mathbf{c} \right\|^2 \tag{6.22}$$

However, we now need to find the optimal choice of $\mathbf{g}_q$ that minimizes the objective in (8.28). Using (6.18), and (8.36), we have the following optimization problem.

Problem 16 *Find equal-gain $\mathbf{g}_q \in \mathbb{C}^L$, $q = 1, \ldots, Q$, such that*

$$\mathbf{g}_q = \underset{\mathbf{g}}{\arg\min} \left\| abs(\mathbf{D}^H \mathbf{c}_q) - \sqrt{2\pi} abs(\mathbf{e}_q \otimes \mathbf{g}) \right\|^2 \tag{6.23}$$

where $abs(.)$ denotes the element-wise absolute value of a vector.

Hence, the codebook design for a system with fully digital beamforming capability is found by solving Problem 17 for proper choice of $\mathbf{g}_q$ obtained as a solution to Problem 24. The

codebook for Hybrid beamforming is then found as

$$\mathbf{F}_q, \mathbf{v}_q = \arg\min_{\mathbf{F},\mathbf{v}} \|\mathbf{F}\mathbf{v} - \mathbf{c}_q\|^2 \tag{6.24}$$

where the columns of $\mathbf{F}_q \in \mathbb{C}^{M_t \times N_{RF}}$ are equal-gain vectors and $\mathbf{v}_q \in \mathbb{C}^{N_{RF}}$. The solution may be obtained using a simple, but yet effective suboptimal algorithm such as orthogonal matching pursuit (OMP) [126] [122]. In the next section, we continue with the solutions of problems 17 and 24.

6.4 Proposed Single-beam Codebook Design Method

6.4.1 Single-beam Codebook Design under ULA Setting

Note that the solution to problem 17 is the limit of the sequence of solutions to a least-square optimization problem as L goes to infinity. Specifically, for each L we find that,

$$\mathbf{c}_q^{(L)} = \sqrt{2\pi}(\mathbf{D}\mathbf{D}^H)^{-1}\mathbf{D}\left(\mathbf{e}_q \otimes \mathbf{g}_q\right) \tag{6.25}$$

$$\mathbf{c}_q^{(L)} = \sigma \mathbf{D}_q \mathbf{g}_q \tag{6.26}$$

where $\sigma = \frac{\sqrt{2\pi\delta_q}}{L\sum_{p=1}^{2^B}\delta_p} = \frac{1}{L}\sqrt{\frac{\delta_q}{2\pi}}$, noting that it holds that,

$$\mathbf{D}\mathbf{D}^H = \left(L\sum_{p=1}^{2^B}\delta_p\right)\mathbf{I}_{M_t} \tag{6.27}$$

Even though Problem 17 admits a nice analytical closed-form solution, doing so for the

Problem 24 is not a trivial task, especially because the objective function is not convex. However, the convexification of the objective problem (8.41) in the form of

$$\mathbf{g}_q = \arg\min_{\mathbf{g}} \left\| = \frac{1}{2\pi L}\mathbf{D}^H\mathbf{D}(\mathbf{e}_q \otimes \mathbf{g}) - \mathbf{e}_q \otimes \mathbf{g} \right\|^2 \tag{6.28}$$

and using $\mathbf{c}_q^{(L)}$ from (8.43) leads to an effective solution for the original problem. Indeed, it can be verified by solving the optimization problem (8.45) numerically that a close-to-optimal solution admits the following form.

$$\mathbf{g}_q = \begin{bmatrix} 1 & \alpha^{\eta} & \cdots & \alpha^{\eta(L-1)} \end{bmatrix}^T \tag{6.29}$$

for some real η, and $\alpha = e^{j(\frac{\delta_q}{L})}$. In the following, we use the analytical form (8.46) for $\mathbf{g}_q$ for the rest of our derivations. This solution would not be the optimal solution for the original problem (8.41). However, it provides a near-optimal solution with the added benefits of allowing to (i) find the limit of the solution as L goes to infinity, and (ii) express the beamforming vectors in closed form, as it will be revealed in the following discussion. An analytical closed-form solution for $\mathbf{c}_q$ can be found as follows.

$$\mathbf{c}_q{}^{(L)} = \sigma \sum_{l=1}^{L} g_{q,l}\mathbf{d}_{M_t}(\psi_{q,l}) = \begin{bmatrix} \sigma \sum_{l=1}^{L} g_{q,l} & \cdots & \sigma \sum_{l=1}^{L} g_{q,l} e^{j(M_t-1)\psi_{q,l}} \end{bmatrix}^T \tag{6.30}$$

Choosing $\mathbf{g}_q$ as in (8.46), the m-th element of the vector $\mathbf{c}_q = \lim_{L\to\infty} \mathbf{c}_q{}^{(L)}$, i.e. $c_{q,m}$, is

given by

$$c_{q,m} = \frac{1}{\sqrt{2\pi}} \lim_{L\to\infty} \frac{1}{L} \sum_{l=0}^{L-1} g_{q,l} e^{j(m\psi_{q,l+1})} = \frac{1}{\sqrt{2\pi}} \lim_{L\to\infty} \frac{1}{L} \sum_{l=0}^{L-1} g_{q,l} e^{jm(\psi_{q-1}+\frac{\delta_q(\ell+0.5)}{L})} \tag{6.31}$$

After some basic manipulations, we get,

$$c_{q,m} = \frac{1}{\sqrt{2\pi}} e^{jm\psi_{q-1}} \int_0^1 e^{j\xi x} dx = \frac{e^{j(m\psi_{q-1}+\frac{\xi}{2})}}{\sqrt{2\pi}} sinc(\frac{\xi}{2\pi}) \tag{6.32}$$

where, $\xi = \delta_q(\eta + m)$ and $sinc(t) = \frac{sin(\pi t)}{\pi t}$. However, note that the design under ULA antenna configuration suffers from multiple shortcomings. (a) First, due to the behavior of ULA, any beam is symmetric with respect to the ULA axis (i.e., the line passing through the base of the antennas). This is observed easily by noticing that the reference gain equation (6.4) is an even function of θ. This means that any beam designed using ULA is two-sided, e.g., if there is a codeword designed to excite the beam covering $[\frac{\pi}{8}, \frac{\pi}{4}]$, this codeword will excite the symmetric beam $[\frac{-\pi}{4}, -\frac{\pi}{8}]$ as well. (b) Second, the gains corresponding to the codewords covering different beams vary considerably. Even the ideal gain of the beams with equal beamwidth varies based on the angular position of the beam (e.g., see equation (6.11)). (c) Third, as we will show in section 9.4, the beams that are close to the direction of the antennas are neither so sharp nor so stable in terms of gain. To deal with shortcoming (a), we introduce a novel configuration for antenna elements, namely, TULA.

6.4.2 Single-beam Codebook Design under TULA Setting

For each codeword $\mathbf{c}_{q,twin}$, let us explicitly write the expression for the reference gain as before,

$$G(\theta, \mathbf{c}_{q,twin}) = \left|\mathbf{d}^H_{twin,M_t}(\theta)\mathbf{c}_{q,twin}\right|^2 \tag{6.33}$$

where

$$\mathbf{d}_{twin,M_t}(\theta) = \left[\mathbf{d}^T_{\frac{M_t}{2}}(\theta), e^{j(\frac{2\pi}{3}\sin(\theta))}\mathbf{d}^T_{\frac{M_t}{2}}(\theta)\right]^T \tag{6.34}$$

We also set,

$$\mathbf{c}_{q,twin} = \left[\mathbf{c}^T_{q,twin,\frac{M_t}{2}}, e^{j\beta}\mathbf{c}^T_{q,twin,\frac{M_t}{2}}\right]^T \tag{6.35}$$

for some $\mathbf{c}_{q,twin,\frac{M_t}{2}} \in \mathbb{C}^{\frac{M_t}{2}}$ and $\beta \in \mathbb{R}$. Therefore,

$$G(\theta, \mathbf{c}_{q,twin}) = \left|\mathbf{d}^H_{\frac{M_t}{2}}(\theta)\mathbf{c}_{q,twin,\frac{M_t}{2}}\right|^2 \left|(1 + e^{j(\beta - \frac{2\pi}{3}\sin(\theta))})\right|^2$$

Further, we define

$$L(\theta) = \left|1 + e^{j(\beta - \frac{2\pi}{3}\sin(\theta))}\right| = \left|\cos(\frac{\beta}{2} - \frac{\pi}{3}\sin(\theta))\right| \tag{6.36}$$

Observing equation (6.63), we have decomposed the TULA gain into two parts. We use the first part to generate a gain similar to the desired one designed with ULA and the second part

to meet the desired isolation level. More precisely, we set

$$c_{q,twin,m} = e^{j(m\psi_{q-1}+\frac{\xi}{2})} sinc(\frac{\xi}{2\pi}), m = 0 \dots \frac{M_t}{2} - 1 \tag{6.37}$$

To capture the isolation requirement, we define the *isolation factor* $0 \leq \mu < 1$ as follows,

$$\mu = \int_{\omega_q} \frac{L(-\theta)}{L(\theta)} d\theta = \int_{\omega_q} \frac{\cos(\frac{\beta}{2} + \frac{\pi}{3}\sin(\theta))}{\cos(\frac{\beta}{2} - \frac{\pi}{3}\sin(\theta))} d\theta \tag{6.38}$$

to denote the level of isolation between each ω_q and its counterpart. Therefore, given any desired isolation level μ, it only remains to solve (6.65) for β and plug into equation (6.62), to complete the TULA codebook design. By the isolation factor incorporated into the codebook design framework, we can effectively solve shortcoming (a) and generate one-sided beams. This design will provide another advantage by significantly improving the gain of the desired beams as we will illustrate in section 9.4 utilizing numerical results. Nonetheless, the shortcomings (b) and (c) are still outstanding in the TULA framework. To address the shortcomings (b) and (c), we propose two more antenna configurations: Delta (Δ) and Star. Both configurations utilize multiple TULA models at the same time where each TULA is responsible for forming a number of the beams. In other words, each of the TULAs will be excited only with those codewords that generate the sharpest and most stable beams that at the same time generate high gains at that interval. Next, we show how this approach resolves the gain variance issue and also improves the sharpness of the peripheral beams.

6.5 Composite Codebook Design Problem Formulation

Let $\mathbf{c}_B$ be the codeword corresponding to the beam $\mathcal{B}(B)$. Using the Parseval's theorem [145], it is straightforward to write,

$$\int_{-\pi}^{\pi} G(\psi, \mathbf{c}) d\psi = 2\pi \|\mathbf{c}\|^2 = 2\pi \tag{6.39}$$

Applying the last equation to the ideal gain corresponding to the codeword $\mathbf{c}_B$ we get,

$$\begin{aligned} \int_{-\pi}^{\pi} G_{\text{ideal},B}(\psi) d\psi &= \int_{\mathcal{B}(B)} t d\psi + \int_{[-\pi,\pi]\setminus\mathcal{B}(B)} 0 d\psi \\ &= \sum_{b\in B} \int_{\omega_b^{\psi}} t d\psi = \sum_{b\in B} \delta_b t = 2\pi \end{aligned} \tag{6.40}$$

Therefore, $t = \frac{2\pi}{\Delta_B}$, where $\Delta_B = \sum_{b\in B} \delta_b$. It follows that,

$$G_{\text{ideal},B}(\psi) = \frac{2\pi}{\Delta_B} \mathbb{1}_{\mathcal{B}^{\psi}(B)}(\psi), \quad \psi \in [-\pi, \pi] \tag{6.41}$$

We wish to find the optimal configuration $\mathbf{c}_B$ for the antenna array such that $G(\psi, \mathbf{c})$ becomes the most accurate estimate of $G_{\text{ideal},B}(\psi)$. To this end, we formulate the codebook design problem as an MSE as follows,

$$\mathbf{c}_B^{opt} = \underset{\mathbf{c}, \|\mathbf{c}\|=1}{\arg\min} \int_{-\pi}^{\pi} \left| G_{\text{ideal},B}(\psi) - G(\psi, \mathbf{c}) \right| d\psi \tag{6.42}$$

To solve the above optimization problem, we rewrite the integral in (6.42) as the equivalent

infinite series as follows,

$$\mathbf{c}_B^{opt} = \underset{\mathbf{c},\|\mathbf{c}\|=1}{\arg\min} \left[\lim_{L\to\infty} \frac{2\pi}{L} \sum_{\ell=1}^{L} \left| G_{\text{ideal},B}\left(\psi_\ell\right) - G\left(\psi_\ell, \mathbf{c}\right) \right| \right] \tag{6.43}$$

where, $\boldsymbol{\psi} = [\psi_1, \cdots, \psi_L]$ is the vector corresponding to the sampled $\psi-$ domain, i.e., $\psi_\ell = -\pi + \ell(\frac{2\pi}{L})$, $\ell = 1, \cdots, L$. Also, let $\boldsymbol{\psi}_b$ be the elements of $\boldsymbol{\psi}$ that lie in ω_b. It holds that, $|\boldsymbol{\psi}| = L$, and ψ_b is comprised of $|\boldsymbol{\psi}_b| = L_b$, $b \in B$ consecutive elements of ψ. We can rewrite the optimization problem in equation (6.43) as follows,

$$\mathbf{c}_B^{opt} = \underset{\mathbf{c},\|\mathbf{c}\|=1}{\arg\min} \lim_{L\to\infty} \frac{1}{L} |\mathbf{G}_{\text{ideal},B} - \mathbf{G}(\mathbf{c})| \tag{6.44}$$

where,

$$\mathbf{G}(\mathbf{c}) = \left[G\left(\psi_1, \mathbf{c}\right) \ldots G\left(\psi_L, \mathbf{c}\right)\right]^T \in \mathbb{Z}^L \tag{6.45}$$

$$\mathbf{G}_{\text{ideal},B} = \left[G_{\text{ideal},B}\left(\psi_1\right) \ldots G_{\text{ideal},B}\left(\psi_L\right)\right]^T \in \mathbb{Z}^L \tag{6.46}$$

Observe that each beam $\mathcal{B}(B)$, divides the angular range into $2|B|$ regions, $|B|$ of which cover the desired ACIs. Define $\mathbf{e}_B(b) \in \mathbb{Z}^{2|B|}$ to be the standard basis vector corresponding to the representation of beam b in the set $\{1, \cdots, 2|B|\}$. For instance, in the example described by fig. 6.2b, we have $e_B([\frac{\pi}{2}, \frac{5\pi}{8})) = [0,0,1,0,0,0]^T$, or $e_B([\frac{-\pi}{4}, \frac{\pi}{8})) = [0,0,0,0,1,0]^T$. Utilizing this notation we write,

$$\mathbf{G}_{\text{ideal},B} = \sum_{b\in B} \frac{2\pi}{\Delta_B} \left(\mathbf{e}_B(b) \otimes \mathbf{1}_{L_b,1}\right) \tag{6.47}$$

Now, observe that for any equal gain $\mathbf{g} \in \mathbb{C}^{L_b}$ it holds that $\mathbf{1}_{L_b,1} = \mathbf{g} \odot \mathbf{g}^*$. Therefore, for such choice of $\mathbf{g}$ we can write:

$$\begin{aligned}
\mathbf{G}_{\text{ideal},B} &= \sum_{b \in B} \frac{2\pi}{\Delta_B} \left(\mathbf{e}_B(b) \otimes (\mathbf{g}_b \odot \mathbf{g}_b^*) \right) \\
&= \sum_{b \in B} \left(\gamma_b \left(\mathbf{e}_B(b) \otimes \mathbf{g}_b \right) \right) \odot \left(\gamma_b \left(\mathbf{e}_B(b) \otimes \mathbf{g}_b \right) \right)^* \\
&= \left(\sum_{b \in B} \gamma_b \left(\mathbf{e}_B(b) \otimes \mathbf{g}_b \right) \right) \odot \left(\sum_{b \in B} \gamma_b \left(\mathbf{e}_B(b) \otimes \mathbf{g}_b \right) \right)^*
\end{aligned} \tag{6.48}$$

Similarly, one can easily verify that,

$$\mathbf{G}(\mathbf{c}) = \left(\mathbf{D}^H \mathbf{c} \right) \odot \left(\mathbf{D}^H \mathbf{c} \right)^* \tag{6.49}$$

where $\mathbf{D} = \left[\mathbf{d}_{M_t}(\psi_1) \cdots \mathbf{d}_{M_t}(\psi_L) \right] \in \mathbb{C}^{M_t \times L}$. Given the special form of the equations (8.28), and (8.35), and their usage in the optimization problem entailed in (8.36), it is straightforward to conclude that $\mathbf{c}_B^{opt}$ is the solution to the following optimization problem for appropriate choices of $\mathbf{g}_b$.

Problem 17 *Given any set of equal-gain vectors $\mathbf{g}_b \in \mathbb{C}^{L_b}$, $b \in B$ find vector $\mathbf{c}_B \in \mathbb{C}^{M_t}$ such that*

$$\mathbf{c}_B = \underset{\mathbf{c}, \|c\|=1}{\arg\min} \lim_{L \to \infty} \left\| \sum_{b \in B} \gamma_b \left(\mathbf{e}_B(b) \otimes \mathbf{g}_b \right) - \mathbf{D}^H \mathbf{c} \right\|^2 \tag{6.50}$$

However, in order to find the optimal solution to the optimization problem in (8.28), we need to

find the optimal choices of $\mathbf{g}_b$, $b \in B$. Utilizing (8.35), and (8.36), the following optimization problem arises.

Problem 18 *Find a set $\mathcal{G}_B$ of equal-gain $\mathbf{g}_b \in \mathbb{C}^{L_b}$, such that*

$$\mathcal{G}_B = \arg\min_{\mathcal{G}} \left\| abs(\mathbf{D}^H \mathbf{c}_B) - abs(\sum_{b \in B} \gamma_b (\mathbf{e}_B(b) \otimes \mathbf{g}_b)) \right\|^2 \tag{6.51}$$

where $\mathcal{G}_B = \{\mathbf{g}_b | b \in B\}$, and $abs(.)$ denotes the element-wise absolute value of a vector.

Therefore, under the fully-digital scheme, by solving problems 24, and 17 the optimal configuration $\mathbf{c}_B^{opt}$ for composite beam $\mathcal{B}(B)$ can be obtained. However, under the hybrid beamforming regime, the optimal configuration is found as

$$\mathbf{F}_B, \mathbf{v}_B = \arg\min_{\mathbf{F},\mathbf{v}} \|\mathbf{F}\mathbf{v} - \mathbf{c}_B^{opt}\|^2 \tag{6.52}$$

under the equal gain condition on the columns of $\mathbf{F}_B$. Several simple heuristic approaches exist in the literature to obtain near-optimal solutions to the above problem including the effective orthogonal matching pursuit (OMP) [126] [122] algorithm. In the next section, we continue with the solution to problems 17, and 24.

6.6 Proposed Composite Codebook Design Method

In this section, we propose our approach for composite codebook design under the ULA and TULA settings respectively.

6.6.1 Composite Codebook Design under ULA Setting

Observe that each fixed value of L problem 17 falls in the class of least-square optimization problems. Therefore, as L tends to infinity, $\mathbf{c}_B$ is obtained as the limit of the solutions to this problem. For each L the solution is given by,

$$\mathbf{c}_B^{(L)} = \sum_{b\in B} \gamma_b (\mathbf{D}\mathbf{D}^H)^{-1}\mathbf{D}\left(\mathbf{e}_B(b)\otimes \mathbf{g}_b\right) \tag{6.53}$$

$$\mathbf{c}_B^{(L)} = \frac{1}{L}\sum_{b\in B} \gamma_b \mathbf{D}(\mathbf{e}_B(b)\otimes \mathbf{g}_b) \tag{6.54}$$

where it holds that,

$$\mathbf{D}\mathbf{D}^H = \sum_{l=1}^{L} \mathbf{d}_{M_t}\left(\psi_l\right)\mathbf{d}_{M_t}^H\left(\psi_l\right) = \sum_{l=1}^{L}\mathbf{I}_{M_t} = L\mathbf{I}_{M_t}$$

Define $\mathbf{\Gamma}_{\mathrm{B}} = \sum_{b\in B}\gamma_b(\mathbf{e}_B(b)\otimes \mathbf{g}_b))$. Replacing $\mathbf{\Gamma}_{\mathrm{B}}$ and (8.43) in equation (8.41) we get,

$$\mathcal{G}_B = \arg\min_{\mathcal{G}} \left\| abs(\frac{\mathbf{D}^H\mathbf{D}}{L}\mathbf{\Gamma}_{\mathrm{B}}) - abs(\mathbf{\Gamma}_{\mathrm{B}})\right\|^2. \tag{6.55}$$

Even though Problem 17 admits a nice analytical closed-form solution, doing so for the Problem 24 is not a trivial task, especially due to the fact that the objective function is not convex. However, the convexification of (8.41) in the form of

$$\mathcal{G}_{\mathrm{B}} = \arg\min\nolimits_{\mathcal{G}} \left\| \left(\frac{1}{L}\mathbf{D}^H\mathbf{D} - \mathbf{I}_{LQ}\right)\mathbf{\Gamma}_{\mathrm{B}}\right\|^2 \tag{6.56}$$

leads to an effective solution for the original problem. Indeed, it can be verified by solving

the optimization problem (8.45) numerically that a close-to-optimal solution admits the form of geometric vectors $\mathbf{g}_b$, of lengths L_b, and parameters $\rho_b = \frac{\eta\delta_b}{L}$, $b \in B$, for some real value η.

We use this analytical form for $\mathbf{g}_b$ for the rest of our derivations. This solution would not be the optimal solution for the original problem (8.41). However, it provides a near-optimal solution with added benefits of allowing to (i) find the limit of the solution as L goes to infinity, (ii) express the beamforming vectors in closed form, as it will be revealed in the following discussion, and (iii) change η in order to design beams with different qualities such as beamforming gain, smoothness, and leakage as it is explained in more details in section 9.4. Let us expand the expression for $\mathbf{c}_B^{(L)}$ from equation (8.43) to get,

$$\mathbf{c}_B{}^{(L)} = \sum_{b\in B} \frac{\gamma_b}{L} \left(\sum_{\ell=1}^{L_b} g_{b,\ell} \mathbf{d}_{M_t}(\psi_{b,\ell}) \right) \tag{6.57}$$

Further, define

$$\mathbf{c}_b^{(L)} = \frac{\gamma_b}{L} \sum_{\ell=1}^{L_b} g_{b,\ell} \mathbf{d}_{M_t}(\psi_{b,\ell}), \quad b \in \mathrm{B} \tag{6.58}$$

The m-th element of the vector $\mathbf{c}_b = \lim_{L\to\infty} \mathbf{c}_b{}^{(L)}$, i.e. $c_{b,m}$, is given by

$$\begin{aligned} c_{b,m} &= \sqrt{\frac{2\pi}{\Delta_B}} \lim_{L\to\infty} \frac{1}{L} \sum_{\ell=0}^{L_b-1} g_{b,\ell} e^{jm\psi_{b,\ell}} \\ &= \frac{\delta_b}{\sqrt{2\pi\Delta_B}} \lim_{L_b\to\infty} \frac{1}{L_b} \sum_{\ell=0}^{L_b-1} g_{b,\ell} e^{jm\psi_{b,\ell}} \end{aligned} \tag{6.59}$$

For large enough L, we can write $\psi_{\ell,b} = \psi_b^s + \ell(\frac{\delta_b}{L_b})$ to get

$$\begin{aligned} c_{b,m} &= \frac{\delta_b}{\sqrt{2\pi\Delta_B}} \lim_{L_b\to\infty} \frac{1}{L_b} \sum_{\ell=0}^{L_b-1} e^{j\frac{\ell\eta\delta_b}{L_b}} e^{jm(\psi_b^s+\ell\frac{\delta_b}{L_b})} \\ &= \frac{\delta_b}{\sqrt{2\pi\Delta_B}} e^{jm(\psi_b^s)} \lim_{L\to\infty} \frac{1}{L_b} \sum_{\ell=0}^{L_b-1} \alpha^{(\eta+m)\ell} \\ &= \frac{\delta_b}{\sqrt{2\pi\Delta_B}} e^{jm(\psi_b^s)} \int_0^1 \alpha^{(\eta+m)L_b x} dx \end{aligned} \tag{6.60}$$

where $\alpha = e^{j\frac{\delta_b}{L_b}}$. After a few straightforward steps, we get,

$$c_{b,m} = \frac{\delta_b}{\sqrt{2\pi\Delta_B}} e^{j(m\psi_b^s+\frac{\xi}{2})} sinc(\frac{\xi}{2\pi}) \tag{6.61}$$

where $\xi = \delta_b(\eta + m)$. We note that the ULA antenna structure inherently generates two-sided lobes and hence inefficient beams due to (i) having beam lobes in undesired scopes, and (ii) having lower effective beam gain in desired ACIs. In the following, we discuss a beam design using a TULA antenna structure. The TULA structure not only generates single-sided beams but also improves the beam gain (by almost 3 dB) [145].

6.6.2 Composite Codebook Design under TULA Setting

For some $\beta \in \mathbb{R}$ define the codeword under the TULA configuration as,

$$\mathbf{c}_{B,twin} = \left[\mathbf{c}^T_{B,twin,\frac{M_t}{2}}, e^{j\beta}\mathbf{c}^T_{B,twin,\frac{M_t}{2}}\right]^T \tag{6.62}$$

Under such codeword, we can rewrite the corresponding reference gain as,

$$G(\theta, \mathbf{c}_{B,twin}) = L(\theta)^2 \left| \mathbf{d}^H_{\frac{M_t}{2}}(\theta) \mathbf{c}_{B,twin,\frac{M_t}{2}} \right|^2 \tag{6.63}$$

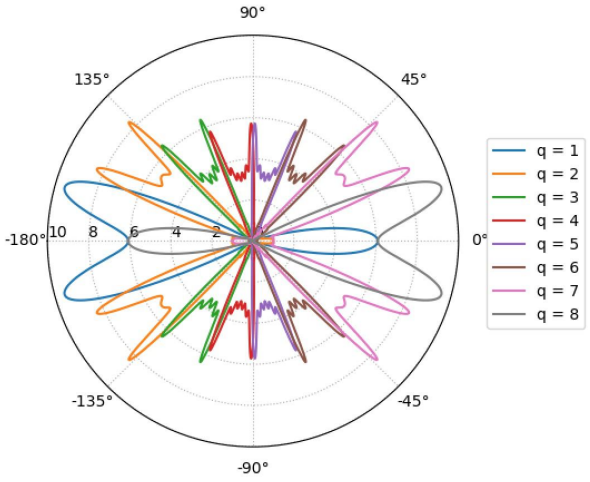
where we define,

$$L(\theta) = \left| 1 + e^{j(\beta - \frac{2\pi}{3}\sin(\theta))} \right| = \left| \cos(\frac{\beta}{2} - \frac{\pi}{3}\sin(\theta)) \right| \tag{6.64}$$

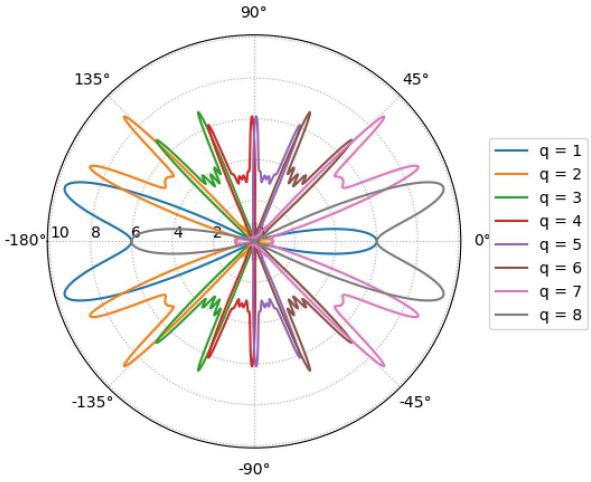
We note that the gain in (6.63) consists of two terms, one being the reference gain of a ULA of $M_t/2$ antennas (by setting the same codeword as in (6.61)) and one being a function of the parameter β. As before, we use the first term to obtain the beamforming gain but use the second term to provide the required level of isolation between each beam and its mirrored counterpart to resolve the inefficiency of the ULA structure. To capture the isolation requirement, we define the *isolation factor* $0 \leq \mu < 1$ as follows

$$\mu = \int_{\omega_b} \frac{L(-\theta)}{L(\theta)} d\theta = \int_{\omega_b} \frac{\cos(\frac{\beta}{2} + \frac{\pi}{3}\sin(\theta))}{\cos(\frac{\beta}{2} - \frac{\pi}{3}\sin(\theta))} d\theta \tag{6.65}$$

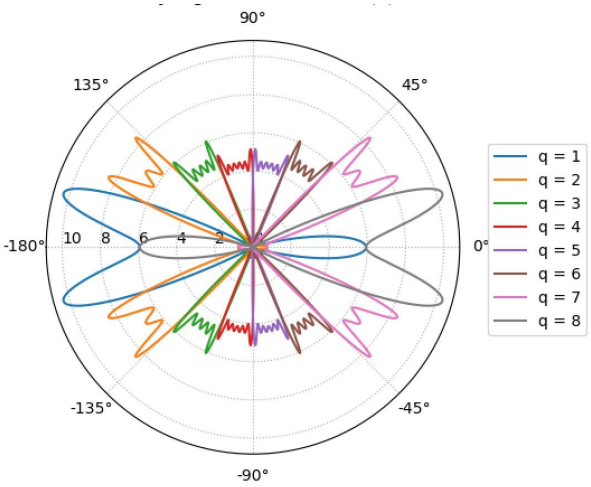
to denote the level of isolation between each ω_b and its mirrored counterpart. Opting for a small enough value for μ, optimal β has to be obtained by numerically solving equation (6.65). Plugging in the obtained β into equation (6.62) completes the codebook design under the TULA structure.



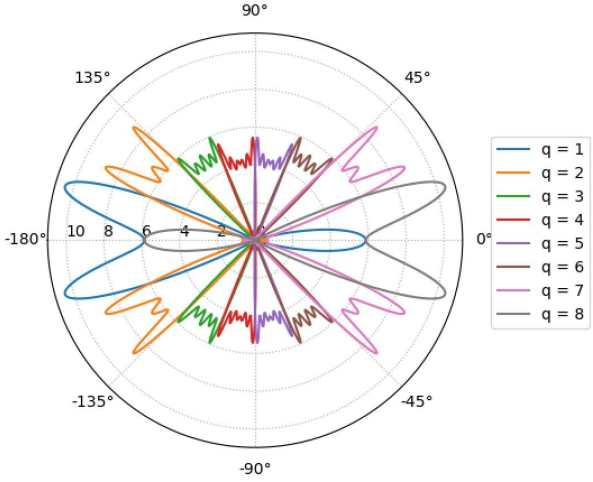
(a) Codebook in [120]



(b) Codebook in [120]

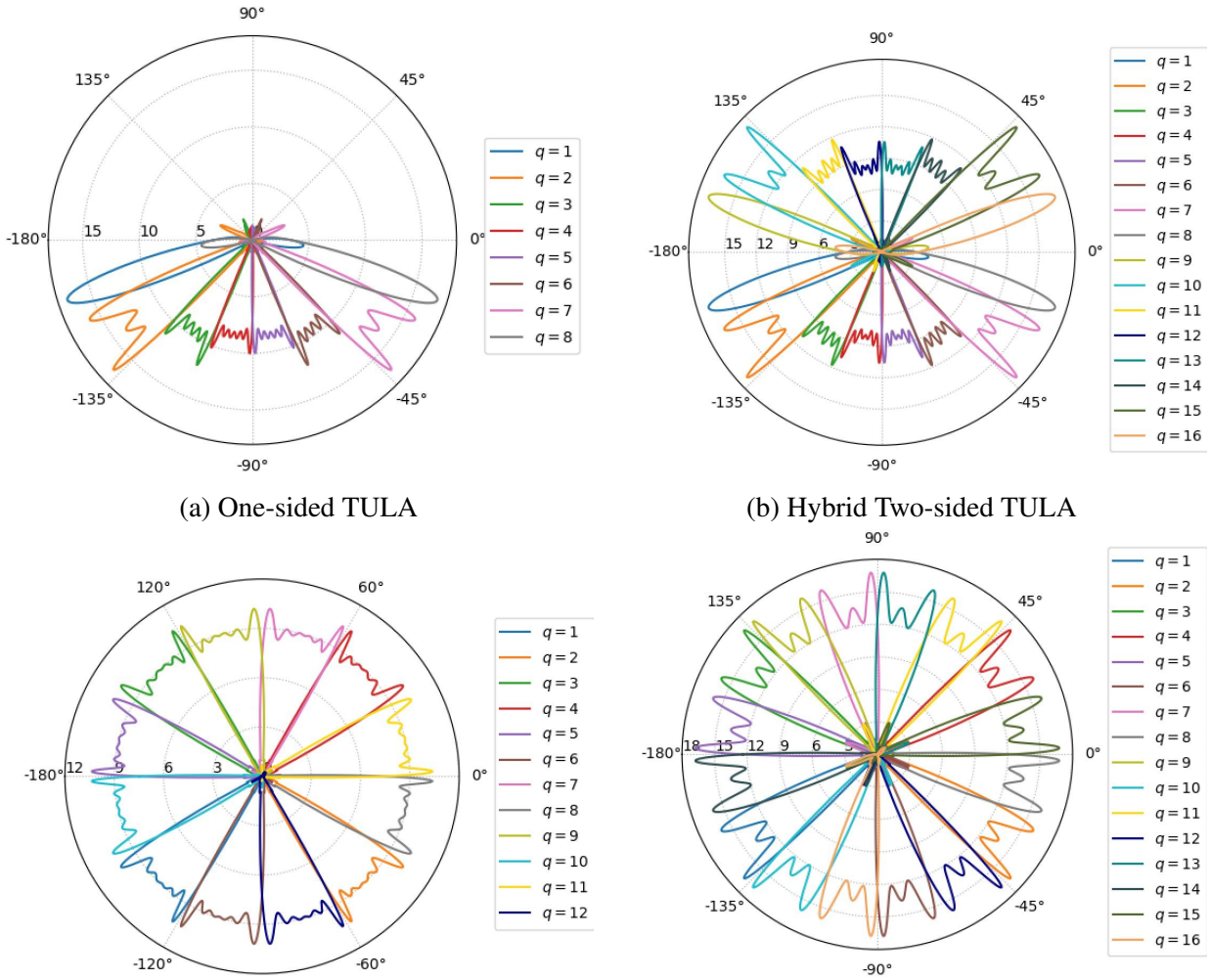


(c) Proposed Codebook



(d) Proposed Codebook

Figure 6.4: ULA patterns, $M_t = 32$, $Card(\mathcal{C}) = 8$, (a)(c) Fully-digital, (b)(d) Hybrid.



(a) One-sided TULA

(b) Hybrid Two-sided TULA

(c) Δ Configuration

(d) Star Configuration

Figure 6.5: TULA-based patterns, (a)(d) Fully-digital, $Card(\mathcal{C}) = 16$, (b) Hybrid, $Card(\mathcal{C}) = 16$, (c) Fully-digital, $Card(\mathcal{C}) = 12$.

6.7 Performance Evaluation

In this section, we evaluate the effectiveness of our proposed codebooks by means of simulation under various structures ULA, TULA, Star, and Delta.

6.7.1 Single-beam Codebook Design Problem Numerical Results

A scenario is considered where the whole azimuth angular range $[-\pi, \pi]$ is uniformly split into equal beam widths each of which is excited by one of the codewords. Fig. (6.4a), and (6.4b) correspond to the codebook design in [120] which achieves equal beamwidths. These figures were regenerated based on the design in [120] using $L = 300$ and $\mathbf{g} = 1_{L,1}$. Using the same antenna configuration (ULA), Fig. (6.4c), and (6.4d) depict our design which improves on the design in [120] in terms of the reference gain by choosing the vector $\mathbf{g}$ as in (8.46). Both the design in [120] and the design in this paper generate beams with fairly sharp edges and relatively stable gain in the coverage area.

In order to overcome shortcoming (a), we presented the idea of TULA. Fig. (8.4a) depicts the first 8 beams of a codebook with 16 beamforming vectors under the TULA structure with $\mu = 0.01$. It can be seen that TULA is very effective in generating one-sided beams. Fig. (8.4b) depicts all 16 beams of the same codebook under the hybrid beamforming regime. It is observed that not only does TULA solve shortcoming (a) but also it provides considerable gain improvement on both sides (almost double). This can be interpreted as removing the undesired signal in the undesired angular interval on the other side of the antenna axis and contributing to the gain generated at the desired angular interval. Figures (6.4b), and (6.4d) show the hybrid beamforming design for $N_{RF} = 4$ of the corresponding (6.4a), (6.4c), respectively. It can be seen that with

the codebook size of up to 16, the loss of using hybrid beamforming instead of fully digital beamforming is not noticeable. Fig. (8.4c) and (6.5d) illustrate the beam patterns of a codebook of size 12 and a codebook of size 16 using the Delta and the Star configurations respectively. It is observed that the beam patterns have relatively sharp edges, almost similar gain across the beams, relatively constant gain over the beamwidth, and precisely splitting the azimuth angle into equal-sized beams. It is understood that as the number of beams in the codebook increases, the sharpness of the beams decreases and in order to design the beams with the same sharpness, the number of antennas has to increase.

6.7.2 Composite Codebook Design Problem Numerical Results

We consider the ACI set and corresponding indices as in Example 1. Fig. 6.6 depicts the composite beam corresponding to $\mathcal{B}(\{1,3\})$ designed for ULA, where the beam gain is depicted in dB. It is observed that the beams have smooth gains within the desired ACIs with very sharp edges and negligible out-of-band leakage. As illustrated in Fig. 6.7, by using TULA antenna structure, our beam design technique is capable of covering beams $\mathcal{B}(B_1)$, and $\mathcal{B}(B_2)$ as in Example 1 with high stable gain, while resolving the two-sided beam issue arises in ULA.

To quantify three main qualities of a beam, i.e., the gain, leakage, and smoothness we define three performance metrics, namely the *in-band average gain* of the beam, the *out-band average gain*, and its *in-band variance*, respectively. Fig. 6.8 presents the evaluation of these metrics for a single beam centered around $\theta = \frac{\pi}{2}$ versus its beam width. Note that, the amount of in-band variance and out-band average gain are negligible, which confirms the high quality of the beams generated by our design. Moreover, as intuition also suggests, it is observed that the beam

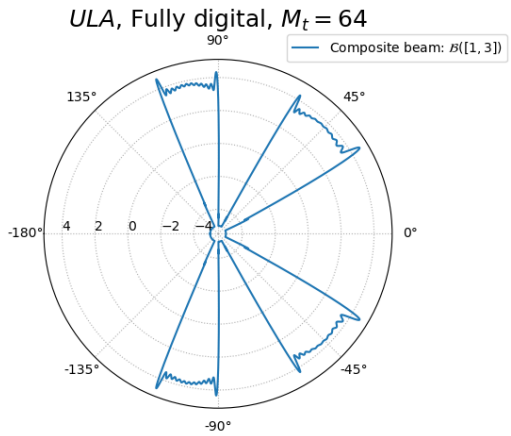


Figure 6.6: Fully-digital, ULA

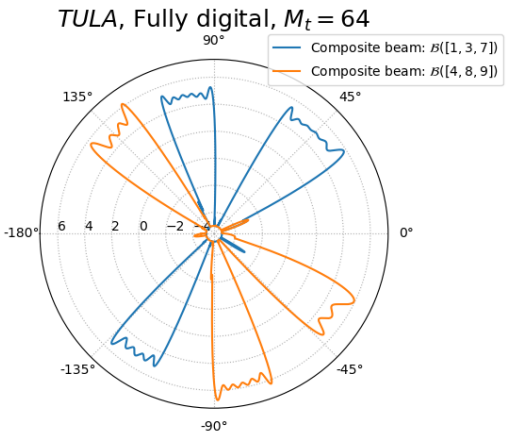


Figure 6.7: Fully-digital, TULA

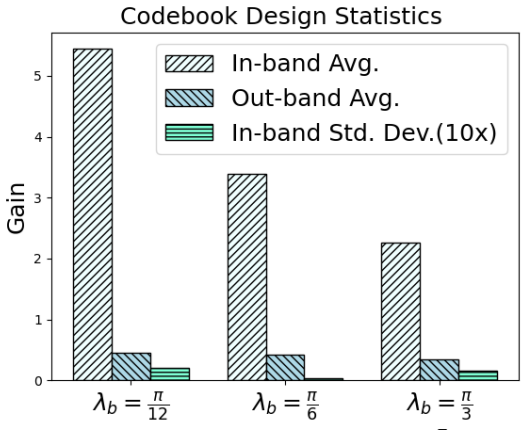


Figure 6.8: Beam quality vs. λ_b

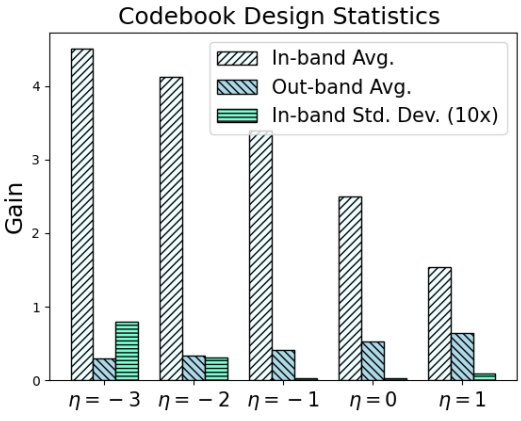
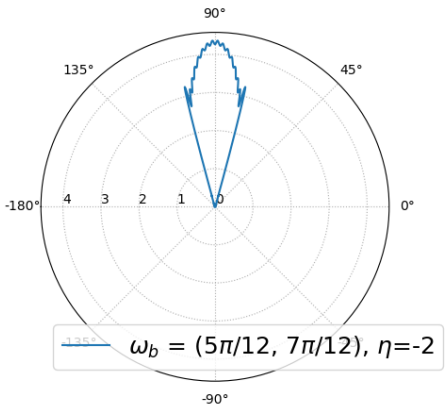
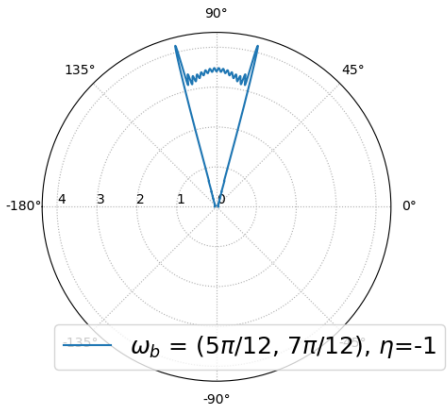


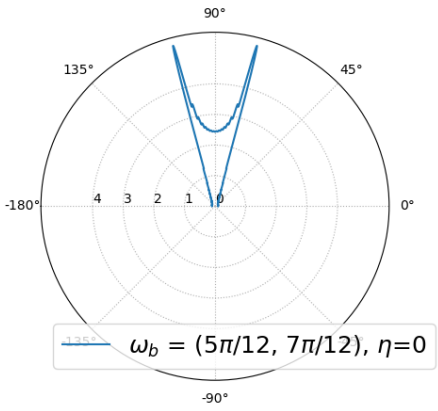
Figure 6.9: Beam quality vs. η



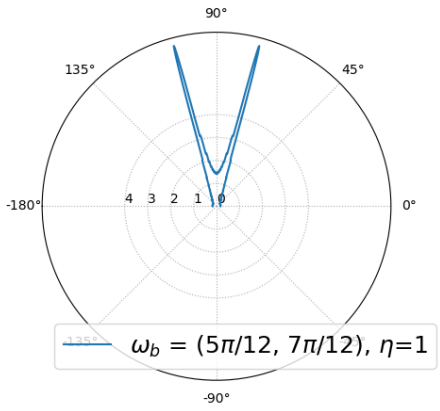
(a) Single-beam shape, $\eta = -2$



(b) Single-beam shape, $\eta = -1$



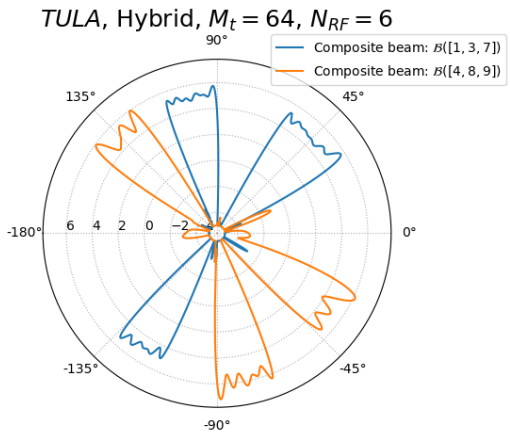
(c) Single-beam shape, $\eta = 0$



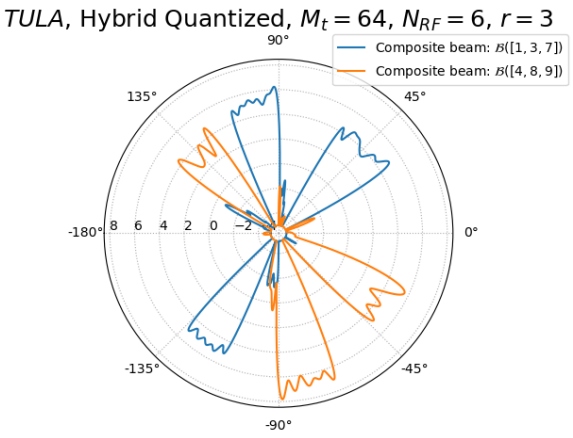
(d) Single-beam shape, $\eta = 1$

Figure 6.10: Single-beam shape for varying η

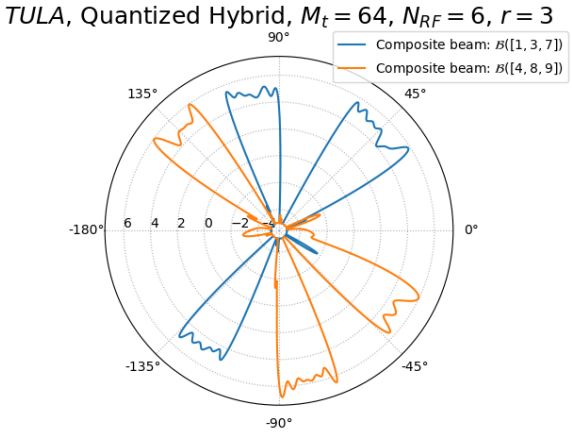
gain is almost inversely proportional to its beam width. Another important design parameter is η introduced in the definition of the geometric vector form for $\mathbf{g}$. Fig. 6.9 shows how varying the value of η impacts the beam quality measures, for a single-beam centered around $\theta = \frac{\pi}{2}$, with $\lambda_b = \frac{\pi}{6}$. Of course, based on the design parameter η, there is a three-way trade-off between the smoothness, gain, and leakage. For the rest of this section, we have used $\eta = -1$ which results in the smoothest beam gain with an acceptable in-band gain. To provide a better envisioning of the beam shape for the reader, Fig. 6.10 shows how changing the value of η impacts the beam pattern for a single beam.



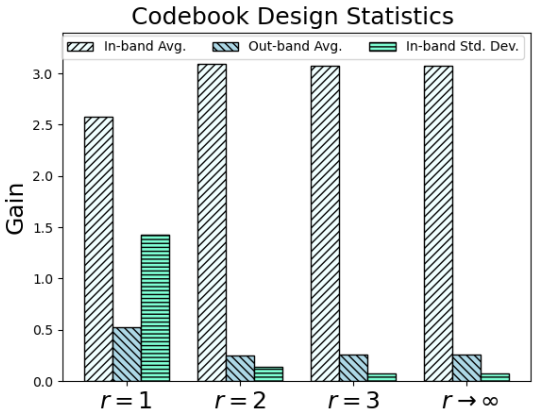
(a) Hybrid



(b) Hybrid-Quantized



(c) Quantized-Hybrid



(d) Beam quality vs. r

Figure 6.11: Effect of quantization on hybrid beamforming using TULA

Figures 6.6-6.10 illustrate the simulation results for a system that is capable of fully-digital beamforming, i.e., where each antenna element is wired to a single RF chain. However, to reduce the number of RF chains hybrid beamforming is employed, where only a few (say 6) RF chains control a second layer phased array feed line to the antennas. Hence a beamformer $\mathbf{c}$ is approximated by $\mathbf{Fv}$ where columns of matrix $\mathbf{F}$ correspond to the phase-only elements. Although suboptimal, an efficient algorithm to find $\mathbf{F}$ from $\mathbf{c}$ is OMP. Fig. 6.11a shows the result of running the OMP algorithm where $N_{RF} = 6$ RF chains are used. However, we note that the codebook has to be stored with finite resolution, say r bits, for each phase shift corresponding to the entries of $\mathbf{F}$. A naive approach is to quantize the entries of $\mathbf{F}$ obtained using OMP using r bits and the result of such approach for $r = 3$ is given in Fig. 6.11b. Comparing Fig. 6.11a and 6.11b, it is observed that such inevitable quantization considerably affects the beam shape and beam gain, hence it suggests that we require finer quantization. However, a more elaborate technique can be used where the quantization is performed in each step of the OMP algorithm. The difference is in the former approach we find the hybrid beamforming matrix $\mathbf{F}$ and then quantize it, i.e., hybrid-quantized, while in the latter approach, we perform quantization at each step, i.e., quantized-hybrid. Using a quantized-hybrid approach, Fig. 6.11c shows that the beam pattern and its gain for $r = 3$ is very similar to that of Fig. 6.11a where there is no limit on the resolution, i.e., $r = \infty$. Finally, Fig. 6.11d shows the effect of increasing the quantization resolution in the quality of the beams generated by the quantized-hybrid approach. It is observed that at $r = 3$, the quality of the generated beams levels that of the hybrid scheme, denoted by $r \to \infty$.

6.8 Conclusions

Despite ULA's advantages such as generating sharp beams, it has a few shortcomings particularly due to its inherent two-sided beams. We proposed a novel TULA structure that can effectively generate one-sided beams. TULA is not only capable of doubling the number of beams in the codebook, it provides a much higher gain for each respective beam in comparison to the ULA structure. Using TULAs in triangle (Δ) and star configurations, we have shown that the entire azimuth angle can be covered by a codebook of beams with equal beam widths, equal gain, and fairly sharp edges. We studied the composite codebook design problem and illustrated how multiple disjoint ACIs with different beam widths can be covered with a single codeword. We highlighted the inefficiencies of ULA in forming arbitrary composite beams and showed that how we can overcome these inefficiencies by employing a novel antenna structure, namely TULA. We derived a low-complexity analytical closed-form solution for the composite codebook design problem for both the ULA and the TULA case and confirmed the validity of our theoretical findings by means of numerical experiments.

Chapter 7: Multi-user Beam Alignment for 5G and Beyond

7.1 Overview

In pursuance of larger bandwidth that is required for realizing one of the main promises of 5G, i.e. enhanced mobile broadband (eMBB), millimeter wave (mmWave) communications is a key technology due to the abundance of unused spectrum available at mmWave frequency ranges [146]. However, high path loss and poor scattering associated with mmWave communications lead to intense shadowing and severe blockage, especially in dense urban environments. These are among the major obstacles to increasing data rates in such high-frequency bands. To tackle these issues effective beamforming (BF) techniques are required to avoid the power leakage to undesired directions using directional transmission patterns, i.e., narrow beams [147]. Furthermore, several experimental results demonstrate that the mmWave channel usually consists of a few components (a.k.a spatial clusters) [148]. Therefore, it is essential to align the devised narrow transmission beams with the direction of the channel components. The problem of aligning the directions of the beams with the angle of departure (AoD) associated with clusters of the channel, is termed as the *beam alignment (BA)* problem. In the literature, the beam alignment problem is also indexed as *beam training* or *beam search*. Devising effective beam alignment schemes is essential since a slight deviation of the transmitted beam AoDs from the mmWave channel clusters may result in a severe drop in the beamforming gain [149] [140].

Beam alignment schemes may be categorized as *exhaustive search (ES)* and *hierarchical search (HS)*. Under the ES scheme, a.k.a beam sweeping, the angular search space is divided into multiple angular coverage intervals (ACIs) each covered by a beam. Then the beam with the highest received signal strength at the receiver is chosen [150] [151]. To yield narrower beams in ES, the number of beams increases which results in larger beam sweeping overhead. The HS scheme lowers the overhead of the beam search by first scanning the angular search space with coarser beams and then gradually finer beams [152] [153] [122]. The BA procedure may happen in one of the two modes, i.e. *interactive* BA (I-BA), and *non-interactive* BA (NI-BA). In the NI-BA mode, the transmitter sends the scanning packets in the scanning phase and receives the feedback from the users after the scanning phase is over, while in the I-BA mode, the transmitter receives the feedback for the previously transmitted scanning pilots during the scanning phase and can utilize this information in the rest of the scanning phase. Most of the prior art on I-BA are limited to single-user scenarios while NI-BA schemes can handle multi-user scenarios as the set of scanning beam does not change or depend on the received feedback from the users.

The process of beam search relies on sending beams out of a set of scanning beams (SB) and tailoring a data transmission beam (TB) based on the received feedback. In this chapter, we address the problem of beam alignment in a multipath environment [154] [155]. We formulate the BA scheme as minimizing the expected average transmission beamwidth under different policies. The policy is defined as a function from the set of received feedback to the set of transmission beams (TB). To maximize the number of possible feedback sequences, we prove that the set of scanning beams (SB) has a special form, namely, *Tulip Design*. Consequently, we rewrite the minimization problem with a set of linear constraints and a reduced number of variables which is solved by using an efficient greedy algorithm.

Further, we discuss a fundamental trade-off between the gain of the SBs and TBs. The higher the gain of an SB, the better the penetration of the SB, and the higher the gain of TB the better the communication link performance. However, TB depends on the set of SBs, and by increasing the coverage of each SB, there is more opportunity to find a sharper TB to increase its beamforming gain. This means that the beamforming gain and hence the penetration of the SB is reduced. We define a quantitative measure for such a trade-off in terms of a trade-off curve. We prove a fundamental result by finding the class of SB set designs namely *Tulip design* which achieves this fundamental trade-off curve for channels with a single dominant path. We also find closed-form solutions for the trade-off curve for special cases and provide an algorithm with its performance evaluation results to find the trade-off curve in general. Our results reveal that the state-of-the-art beam search algorithm should further optimize their SB sets based on this fundamental trade-off between the SB penetration and TB gain.

7.2 System model

We consider a mmWave communications scenario with a single base station (BS) and an arbitrary number of mobile users (MUs), say N, where prior knowledge on the value of N may or may not be available at the BS. The BA procedure aims at obtaining the accurate AoDs corresponding to the downlink mmWave channel from the BS to the users. Under the BA procedure, the BS transmits probing packets in different directions via various *scanning beams (SBs)* and receives feedback from all the users, based on which the BS computes a *transmission beam (TB)* for each user.

7.2.1 Channel Model

Unlike prior art, we consider multipath in the transmission from the BS to the MUs. More precisely, we assume the mmWave channel from the BS to each MU contains a maximum of p resolvable paths where each *resolvable path* corresponds to a possible AoD of the channel. Let $\Psi_j = \{\Psi_{ij}\}_{i=1}^{p}$ denote the random AoD vector corresponding to the channel between the BS and the j^{th} MU, where Ψ_{ij} represents the AoD of the i^{th} path. Denote by $f_{\Psi_j}(\psi_{1j}, \ldots, \psi_{pj})$, defined over $\mathcal{D} \subset (0, 2\pi]^p$, the probability density function (PDF) of Ψ_j. The PDF $f_{\Psi_j}(.)$ encapsulates the knowledge about the AoD of the j^{th} user before the BA procedure or may act as a priority function over the angular search domain. Such information may be inferred from previous beam tracking, training, or alignment trials. A uniform distribution is tantamount to the lack of any prior knowledge or priority over the search domain.

7.2.2 Beamforming Model

We consider a multi-antenna base station with an antenna array of large-size realizing beams of high resolution. For power efficiency, we assume hybrid beamforming techniques are in effect in the BS deploying only a few RF chains. Further, we adopt a *sectored antenna model* where each beam is modeled by the constant gain of its main lobe, and the angular coverage interval (ACI) it covers. Such models are widely adopted in the literature for modeling the beamforming gain and the directivity of mmWave transmitters.

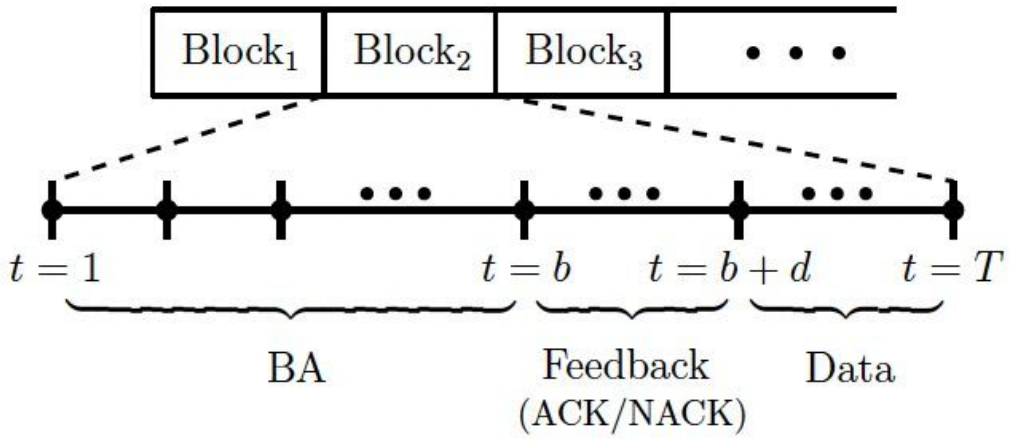


Figure 7.1: Time-slotted System Model

7.2.3 Time-slotted System Model

We consider a system operating under the time division duplex (TDD) and the NI-BA schemes, with frames of length T. Each frame consists of T equal slots. In each frame, the first b slots are dedicated to the transmission of the probing packets, denoted by *scanning time-slots (STS)* and the next d slots denoted by *feedback time-slots (FTS)* are allocated to receiving the users' feedback that may arrive through a side channel or according to any random access mode. Finally, the last $T + b + d$ slots are reserved for data transmission, namely *data transmission time-slots (DTS)*.

7.2.4 Beam Alignment Model

The objective of the BA scheme is to generate the narrowest possible TBs for the data transmission phase for each user to produce beams of higher gain and quality. In other words, utilizing the feedback provided by the users in the FTS to the SBs transmitted by the BS in the STS, the BS aims at localizing the AoD of each user to minimize the *uncertainty region (UR)* for

each AoD. Let $\mathcal{B} = \{\Phi_i\}_{i=1}^{b}$ be the set of STS scanning beams where Φ_i denotes the ACI of the SB sent over time-slot $i \in [b]$. The feedback provided by each user to each SB is binary. If the AoD corresponding to at least one of the resolvable paths in the channel from the BS to the MU is within the ACI of the SB, then the MU will receive the probing packet sent via that SB and feedback an acknowledgment (ACK). Otherwise, the feedback of the MU will be considered as a negative acknowledgment (NACK) indicating none of the user AoDs lie in the ACI of the SB. Once the FTS ends, the BS will determine the TBs using the SBs and the feedback sequences provided by the users according to the BA *policy*. The BA policy is formally defined as a function from the set of feedback sequences to the set of TBs.

In the next section, we first elaborate on the BA policy and then provide the BA problem formulation.

7.3 Problem Formulation

7.3.1 Preliminaries

The BA policy determines how the direction of the TBs is calculated. This decision naturally considers the UR of the AoDs of each user channel. In this paper, we consider four different policies that differ based on how they define the URs of the AoDs and whether they require the exact number of spatial clusters or not.

7.3.1.1 general policies

We define two general policies that do not require any information regarding the number of spatial clusters, namely i) spatial diversity (SD) policy, and ii) beamforming (BF) policy. The

SD policy aims at generating TBs with minimal angular span that cover all the angular intervals that may contain a resolvable path, while the BF policy promises to generate TBs that cover at least one resolvable path but further reduce the angular span of the TBs. The advantage of the SD policy to the BF policy is its resilience against the potential failure or blockage of one or some of the spatial clusters as long as at least one resolvable path remains, while the BF policy has the advantage of producing much higher beamforming gains compared to that of the SD policy but it is vulnerable to path blockage. This will introduce an interesting trade-off between connectivity maintenance and high beamforming gain.

7.3.1.2 path-based policies

If the exact number of spatial clusters, p, is known, each of the above policies may be improved by further lowering the span of the resultant TBs. We denote the corresponding two new policies by p-SD policy and p-BF policy respectively. In the following, we state the expressions for the URs corresponding to each of the mentioned SD and BF policies. Let $B^j_{\mathcal{P}}(\mathcal{B}, \mathbf{s})$ denote the UR of the j^{th} user providing the feedback sequence $\mathbf{s}$ under the policy $\mathcal{P} \in \{\text{SD}, \text{BF}, \text{p}-\text{SD}, \text{p}-\text{BF}\}$ and the SB set $\mathcal{B}$. For instance, $B^j_{SD}(\mathcal{B}, \mathbf{s})$ is the minimal angular span that covers all the resolvable paths for the j^{th} user. Similarly, $B^j_{BF}(\mathcal{B}, \mathbf{s})$ is the minimal angular span that covers at least one resolvable path for the j^{th} user. Further, let the *positivity set* $A^j(\mathbf{s}) \subseteq [b]$ be the set of all indices corresponding to the SBs that are acknowledged by the j^{th} user. Define the *negativity set* $N^j(\mathbf{s}) \subseteq [b]$ in a similar fashion for the not acknowledged SBs. To facilitate the statement of the URs and the subsequent following discussions we define the notion

of the *component beam (CB)*. The CB ω_A is defined as,

$$\omega_A = \cap_{i\in A}\Phi_i \setminus \cup_{A\subset T\subset[b],A\neq T} \cap_{i\in T}\Phi_i \tag{7.1}$$

It is straightforward to show that $\omega_A \cap \omega_T = \varnothing$ for any $A \neq T$, and $\Phi_i = \cup_{A,i\in A}\omega_A$ for all $i \in [b]$. We define the CB set as $\mathcal{C} = \{\omega_A, \omega_A \neq \varnothing, A \subset [b]\}$. Obviously, $\mathcal{B}$ can be generated from $\mathcal{C}$ and vice versa.

Note that if the j^{th} user sends an ACK in response to the SB Φ_i, this would mean that $\Theta_i(\mathbf{s}) \doteq \Phi_i$ has at least one resolvable path. On the other hand, a NACK would mean that no resolvable paths reside in Φ_i and therefore, any resolvable path should exist in $\Theta_i(\mathbf{s}) \doteq \mathcal{D} - \Phi_i$, and $B^j_{\mathcal{P}}(\mathcal{B}, \mathbf{s}) \in \mathcal{D} - \Phi_i$ for the above-mentioned policies. Having this in mind, we can explicitly express the uncertainty region for the general policies as follows.

$$B^j_{\text{SD}}(\mathbf{s}) = \big(\cup_{i\in A(\mathbf{s})}\Theta_i(\mathbf{s})\big) \cap \big(\cap_{i\in N(\mathbf{s})}\Theta_i(\mathbf{s})\big) \tag{7.2}$$

$$B^j_{\text{BF}}(\mathbf{s}) = \Theta_k(\mathbf{s}) \cap \big(\cap_{i\in N(\mathbf{s})}\Theta_i(\mathbf{s})\big) \tag{7.3}$$

where $k = \arg\min_{l\in A(\mathbf{s})} \big|\Theta_l(\mathbf{s}) \cap \big(\cap_{i\in N(\mathbf{s})}\Theta_i(\mathbf{s})\big)\big|$. For the path-based policies, having the luxury of the knowledge of the exact value of p, we can improve the SD and the BF policies to p-SD and p-BF, respectively. For the simplicity of presentation, we only express the improved policies for $p = 2$.

We define $\mathcal{W}_A = \{\{C, C'\} \text{ s.t. } \omega_C, \omega_{C'} \in \mathcal{C},\ C \cup C' = A\}$, $\mathcal{V}_A = \cup_{V\in\mathcal{W}_A} V$ and $n = |\mathcal{W}_A|$. Let $\bigotimes \mathcal{W}_A$ denote the Cartesian product of all elements of $\mathcal{W}_A$ where each element of $\bigotimes \mathcal{W}_A$ is

a n-tuple. For a n-tuple $T = (t_1, t_2, \ldots, t_n)$, we define $\text{UNION}(T) = \cup_{i=1}^{n} t_i$. We have

$$B^j_{2-\text{SD}}(\mathbf{s}) = \bigcup_{V \in \mathcal{V}_A} \omega_V \tag{7.4}$$

$$B^j_{2-\text{BF}}(\mathbf{s}) = \text{UNION}(T^*), T^* = \arg \min_{T \in \bigotimes \mathcal{W}_A} |\text{UNION}(T)| \tag{7.5}$$

Note that for the special case of $p = 1$ all the mentioned policies collapse into one. Next, we will present the BA problem formulation.

7.3.2 Problem Formulation

We assume there are N users that are prioritized according to the weight vector given by $\{c_j \geq 0\}_{j=1}^{N}, \sum_{j=1}^{N} c_j = 1$. Let $\mathcal{U} = \{u_k\}_{k=1}^{M}$ denote the range of the policy function $B^j_{\mathcal{P}}(\mathcal{B}, \mathbf{s})$. In other words, the TBs resulting from the BA scheme may take any value in the set $\mathcal{U}$. The expected value of the average beamwidth resulting from the BA scheme for policy $\mathcal{P}$ is

$$\bar{U}_{\mathcal{P}}(\mathcal{B}) = \sum_{j=1}^{N} c_j \mathbb{E}\left[|B_{\mathcal{P}}(\mathbf{s})|\right], \quad \text{where,} \tag{7.6}$$

$$\mathbb{E}\left[|B_{\mathcal{P}}(\mathbf{s})|\right] = \sum_{k=1}^{M} |u_k| \, \mathbb{P}\left\{B_{\mathcal{P}}(\mathbf{s}) = u_k\right\} \tag{7.7}$$

and $|u_k|$ denotes the Lebesgue measure of the u_k. Note that u_k may be a finite union of multiple intervals in which case $|u_k|$ will be the sum of their widths. Given the value of b the objective of the BA scheme is to design $\{\Phi_i\}_{i=1}^{b}$ such that the expected average TB beamwidths as in (7.6)

gets minimized. i.e.,

$$\{\Phi_i^*\}_{i=1}^b = \arg\min_{\{\Phi_i\}_{i=1}^b} \bar{U}_{\mathcal{P}}\left(\{\Phi_i\}_{i=1}^b\right) \tag{7.8}$$

As shown in [133], it is straightforward to establish that a multi-user NI-BA problem can be posed as a single-user NI-BA by casting the weighted average of the users' PDFs as a prior on the AoD of a single user.

$$f_\Psi(\psi) = \sum_{j=1}^{N} c_j f_{\Psi_j}(\psi), \quad \psi \in \mathcal{D} \tag{7.9}$$

Therefore, we solve the problem for the single-user case with the PDF as in (7.9) and remove the index j from the notations.

Let P_A be the probability of receiving a binary feedback sequence with the positivity set A. Using the inclusion-exclusion principle we can express P_A as follows,

$$\begin{aligned} P_A = &\left(\sum_{C\subset A} g\left(\omega_C\right)\right)^p - \sum_{B\subset A^{(L-1)}} \left(\sum_{C\subset B} g\left(\omega_C\right)\right)^p \\ &+ \sum_{B\subset A^{(L-2)}} \left(\sum_{C\subset B} g\left(\omega_C\right)\right)^p - \ldots + (-1)^{(L+1)} \sum_{B\subset A^{(1)}} \left(\sum_{C\subset B} g\left(\omega_C\right)\right)^p \end{aligned} \tag{7.10}$$

where $g\left(\omega_C\right) = \int_{\psi\in\omega_C} f_\Psi(\psi)d\psi$, and $A^{(\ell)}, \ell \in [L]$ is the set of all subsets of A with size ℓ. Further, let $\lambda_{\mathcal{P}}(A)$ be the width of the TB resulting from the feedback sequence $\mathbf{s}$ with the positivity set A. The objective function (7.6) can be rewritten as,

$$\bar{\lambda} \doteq \sum_{A\subset[b]} \lambda_{\mathcal{P}}(A) P_A. \tag{7.11}$$

where $\lambda_{\mathcal{P}}(A)$ for mentioned policies is expressed as,

$$\lambda_{SD}(A) = \sum_{C \subset A} \lambda(\omega_C) \tag{7.12}$$

$$\lambda_{BF}(A) = \min_{i \in A} \sum_{C, i \in C, C \subset A} \lambda(\omega_C) \tag{7.13}$$

$$\lambda_{2-SD}(A) = \sum_{V \in \mathcal{V}_A} \lambda(\omega_V) \tag{7.14}$$

$$\lambda_{2-BF}(A) = \sum_{i=1}^{n} \lambda(\omega_{t_i}), \text{ where } T^* = (t_1, \ldots, t_n) \tag{7.15}$$

The optimized scanning beam set $\mathcal{B}^*$ is obtained from $\mathcal{C}^*$ where

$$\mathcal{C}^* = \arg\min_{\mathcal{C}} \bar{\lambda} \tag{7.16}$$

7.4 Proposed Beam Alignment Scheme

In this section, we propose our solution to the mentioned optimization problem. A set of SB is called *generalized exhaustive search (GES)* if and only if for any i and j, $\Phi_i \cap \Phi_j = \varnothing$. A set of SB is called *exhaustive search (ES)* if and only if it is GES and $\lambda(\omega_i) = \lambda(\omega_j)$. A contiguous beam is denoted by its angular coverage interval (ACI), e.g., the beam Φ_i is denoted as $[s_i, e_i)$. A *composite beam* is defined as a beam with multiple disjoint ACIs. As the number of ACIs increases, the sharpness of the beams deteriorates. For the scanning beams, it is desirable to use the sharpest beams, hence, we use contiguous beams (beams with single ACIs) as scanning beams. It is not hard to show that b scanning beams generate at most $2b$ CBs due to the possible intersection of multiple scanning beams. Out of the possible set of scanning beams, some are

more appropriate. Since the policy is a function of the set of feedback sequences, it is desirable to maximize the size of the set of feedback sequences. Hence, we first pose the following question: "What is *the most distinguishable* set of scanning beams, i.e., the set of beams which can generate the maximum number of possible feedback sequences?"

To answer this question, we define a special form for the set of scanning beams, namely, *Tulip design* for which we have proved it generates the maximum number of feedback sequences for $p = 1$ and $p = 2$. While, we strongly believe that the same is true for $p \geq 3$, we do not have formal proof. Hence, any results that are presented in the evaluation section for $p \geq 3$ are merely the results obtained under the assumption of using the Tulip design.

Definition 19 *Tulip design is given by a set of contiguous SBs* $\mathcal{B} = \{\Phi_i\}, i \in [b]$ *where each beam may only have an intersection with its adjacent beams except for* Φ_1 *and* Φ_b *for which the intersection might be nonempty. This means* $\Phi_i \cap \Phi_j = \varnothing$, $1 < |i - j| < b - 1$.

Theorem 20 *Among the set of contiguous scanning beams, a set of scanning beams with Tulip design generates the maximal number of possible feedback sequences for the channel with* $p = 1$ *and* 2*, for an arbitrary distribution of channel AoD that is nonzero on any points in the range* $[0, 2\pi)$.

Proof. Please see appendix A.

Under the Tulip design, the CB set takes a special form given by $\mathcal{C}_{\text{eff}} = \mathcal{C}^1_{\text{eff}} \cup \mathcal{C}^2_{\text{eff}}$ where $\mathcal{C}^1_{\text{eff}} = \{\omega_i\}_{i=1}^b$ and $\mathcal{C}^2_{\text{eff}} = \{\omega_{i,i\oplus 1}\}_{i=1}^b$. Clearly, $|\mathcal{C}_{\text{eff}}| = 2b$. By using Tulip design, we can reformulate the optimization problem (7.16) in terms of the starting and ending point of the ACI of the SB Φ_i, i.e., $\Phi_i = [x_i, y_i), i \in [b]$. Hence, we have $\omega_i = [y_{i\ominus 1}, x_{i\oplus 1})$ and $\omega_{i,i\oplus 1} = [x_i, y_{i\ominus 1})$.

We have

$$\mathcal{C}^*_{\text{eff}} = \arg\min_{\mathcal{C}_{\text{eff}}} \bar{\lambda} \tag{7.17}$$

$$x_{i+1} \geq x_i, \quad \forall i \in [b-1] \tag{7.18}$$

$$y_{i+1} \geq y_i, \quad \forall i \in [b-2] \tag{7.19}$$

$$x_{i+2} \geq y_i \geq x_{i+1}, \quad \forall i \in [b-2] \tag{7.20}$$

$$x_1 \leq y_{b-1} \leq 2\pi + x_1 \tag{7.21}$$

$$y_b \leq 2\pi + x_2 \tag{7.22}$$

$$2\pi + x_1 \leq y_b \tag{7.23}$$

where constraints (7.18)-(7.23) ensure the validity of the Tulip design. The optimization problem (7.17) is generally nonlinear. For instance, for uniform distribution on the AoD of the user, the objective function (7.17) is a polynomial function of the order $(p+1)$ of the beamwidth of the CBs. We propose a greedy algorithm to solve the BA optimization problem that is pseudo-coded as follows in *Algorithm 8*. The *Greedy-SA* algorithm starts by discretizing the angular domain $\mathcal{D} = [0, 2\pi]$ to get the ground set G_N, the quantized version of the angular range consisting of N points. It then randomly picks $2b$ points from the ground set and forms the initial CB set $\mathcal{C}_{\text{eff}}$. Hence, the initial value of $\bar{\lambda}$ can be easily computed using (7.11). The *Greedy-SA* algorithm makes repeated calls to the *Modify-Sol* routine pseudo-coded in *Algorithm 9* to improve the quality of the CB set $\mathcal{C}_{\text{eff}}$, i.e. to reduce the value of $\bar{\lambda}$. Each time the *Modify-Sol* routine is called it performs the following sequence of operations. The routine generates the set $perm$ of all random tuples (p, q, r) where z_p and z_q are two of the points in the set $\{z_i\}_{i=1}^{2b}$ and r

Algorithm 8 *Greedy-SA*

Input: $N, b, \mathcal{P}, f_\Psi(\psi), done = \varnothing$

$G_N \doteq$ a set of N points in $[0, 2\pi]$
$\{z_i\}_{i=1}^{2b} \doteq$ a random ordered set of points from G_N
$\{z_i\}_{i=1}^{2b} \Leftrightarrow \mathcal{C}_{\text{eff}} \doteq \{(z_i, z_{i+1 \mod b}), i \in [2b]\}$
Compute $\bar{\lambda}$ from (7.11) using $f_\Psi(\psi)$ and $\mathcal{C}_{\text{eff}}$
while not $done$ **do**
    $(done, \bar{\lambda}, \mathcal{C}_{\text{eff}})$ = *modify-sol*$(G_N, \bar{\lambda}, \mathcal{C}_{\text{eff}})$
end while
Return $\bar{\lambda}, \mathcal{C}_{\text{eff}}$

Algorithm 9 *Modify-Sol*

Input: $G_N, \bar{\lambda}, \mathcal{C}_{\text{eff}}, s = True, count = 0$

$dir = \{forward, backward\}$, $\bar{\lambda}_{old} = \bar{\lambda}$
$perm =$ Shuffle $\{(p, q, r) | p, q \in [2b], r \in dir, p \leq q\}$
repeat
    Orderly select next tuple (p, q, r) from $perm$
    Slide $\{z_i\}_{i=p}^{q}$ in $r \in dir$ direction on points in G_N
    Compute $\bar{\lambda}_{new}$ from (7.11) using $f_\Psi(\psi)$ and $\mathcal{C}_{\text{eff}}$
    if $\bar{\lambda}_{new} \geq \bar{\lambda}_{old}$ **then**
        $count++$
    end if
until $(count = 2b^2 + b) \vee (\bar{\lambda}_{new} < \bar{\lambda}_{old})$
if $count = 2b^2 + b$ **then**
    Return $(True, \bar{\lambda}_{new}, \mathcal{C}_{\text{eff}})$
else
    Return *modify-sol* $(G_N, \bar{\lambda}_{new}, \mathcal{C}_{\text{eff}})$
end if

denotes a direction in the set $\{forward, backward\}$. It then repeatedly picks one such tuple and then slides the window $\{z_i\}_{i=p}^{q}$ over the ground set G_N in direction r and computes the new value for $\bar{\lambda}$, namely $\bar{\lambda}_{new}$. The first time $\bar{\lambda}_{new}$ goes lower than its old value, the routine records $\bar{\lambda}_{new}$ and the corresponding $\mathcal{C}_{\text{eff}}$ and calls itself again with these new values. The *Greedy-SA* algorithm terminates when all the points $\{z_i\}_{i=1}^{2b}$ are stable. In other words, when there are no tuples $(p, q, r) \in perm$ which improves the value $\bar{\lambda}$ from equation (7.11) by moving the window $\{z_i\}_{i=p}^{q}$.

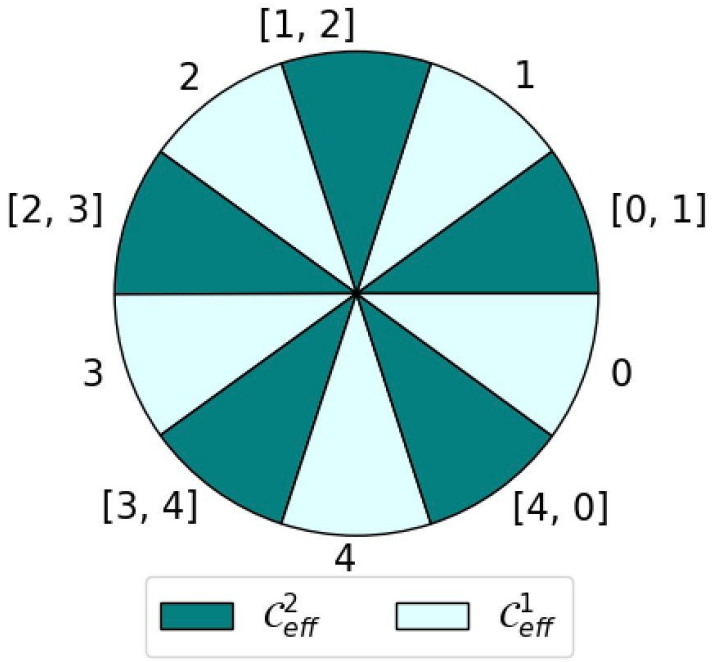


Figure 7.2: Example of a Tulip design for $b = 5$

7.5 Trade-off Curve

An example of a Tulip design is given in Fig. 7.2. Any scanning beam in the Tulip design consists of three parts, namely two side lobes (common between two neighboring scanning beams) and one middle lobe. We denote by each component a *component beam* (CB) and define the set $\mathcal{C}$ as the CB set. We denote by the middle lobes the CBs of the first type and by the side lobes the CBs of the second type. In the above example we have,

$$\mathcal{C}^1 = \{0, 1, 2, 3, 4\}, \mathcal{C}^2 = \{[0, 1], [1, 2], [2, 3], [3, 4], [4, 0]\} \tag{7.24}$$

Next, we define the notion of a trade-off curve.

7.5.1 Trade-off curve

Let $\lambda_{\mathcal{P}}(s)$ denote the beamwidth of TB for the received feedback sequence s and $\eta_i = |\Phi_i|$ denote the beamwidth of the SB i, i.e., Φ_i . Consider a measure $\mu(.)$ which is a function from a set of real numbers to a single real number. In this paper, we particularly are interested in two measures: max and average. We define the max function as $\mu_1(\{\eta_i\}_{i=1}^b) = \max(\{\eta_i\}_{i=1}^b)$ and the average function as $\mu_2(\{\eta_i\}_{i=1}^b) = (1/b)\sum_{i=1}^b \eta_i$. The tradeoff curve for the policy $\mathcal{P}$, using b scanning beams for the channel with p path is defined as the minimum value of $\mu(\{\lambda_{\mathcal{P}}(s)\}_{s\in S})$ for any given value of $\mu(\{\eta_i\}_{i=1}^b)$ where S is the set of all possible feedback sequences.

To find the trade-off curve, one needs to consider all possible SB designs in general. In this section, we provide an interesting result on the class of SB designs that outperform all other designs in terms of the trade-off between the SB and TB sizes. In particular, we prove that for the channel with a single dominant path $p = 1$, the trade-off curve is always achieved by an SB design that follows a tulip design for both max and average measures. The results for multi-path channels are out of the scope of this paper. We note that for $p = 1$ both the diversity and beamforming policies collapse into one. As mentioned before, different measures for the SBs and TBs can be taken into account for demonstrating the trade-off. For the max measure, we consider the pair $(\eta_{max}, \lambda_{max})$ where λ_{max} represents the maximum resulting beamwidth of the beams in the TB set, and η_{max} denotes the maximum beamwidth for the set of SBs. The trade-off curve may be interpreted as follows. For a given value of η_{max} there is a minimum achievable λ_{max}^{min}. The *trade-off curve* is defined as the set of all achievable $(\eta_{max}, \lambda_{max}^{min})$. Also, for a given value of λ_{max} there is a minimum achievable η_{max}^{min}. Alternatively, the *trade-off curve* is composed of all achievable $(\eta_{max}^{min}, \lambda_{max})$. For the average measure, we consider the pair $(\bar{\eta}, \bar{\lambda})$ where $\bar{\lambda}$ represents

the average resulting beamwidth of the beams in the TB set, and $\bar{\eta}$ denotes the average beamwidth for the set of SBs. We have the following theorem.

Theorem 21 *The optimal (i.e., generating the trade-off curve) BA scheme is achieved by an SB set that follows the Tulip design.*

proof. Each SB can be represented by its coverage interval which is an arc on the Trigonometric circle. Hence, if we move in a counterclockwise direction each SB has a starting point and an ending point. A *marker* is defined as any starting point or ending point. Please note that in general multiple starting or ending points may be concurrent (i.e., the same) in which case such marker is called *non-singular* marker. If a marker only represents a single starting point or a single ending point it is called a *singular* marker. The proof consists of multiple steps.

Step 1: We exchange each of the non-singular markers with multiple singular markers which are only a differential amount apart in such a way that all starting points are moved in clockwise directions and all ending points are moved in counterclockwise direction. Please note that in this process the ordering between the starting points or between the ending points is not important.

Step 2: There are exactly $2b$ markers and hence $2b$ arcs on the circle. Each SB is comprised of one or more arcs that are adjacent. An arc may be shared between multiple SBs. The *order* of an arc is the number of scanning beams that intersect with that arc. Since each SB is contiguous and each marker is singular, the order of the arcs is alternatively odd and even on the circle. It can be easily verified that the order of the arc in the Tulip design is alternatively 1 and 2. For the original design, we associate a Tulip design where the corresponding arcs in both designs are either odd or even (See Fig. 7.2).

On one hand, we have

$$\mu_2(\{\eta_i\}_{i=1}^{b}) = (1/b)\sum_{i=1}^{b}\eta_i = (1/b)\sum_{i=1}^{2b}\text{arc}_i n_i$$

where arc_i is the length of the arc i and n_i is its order. Similarly, for $\mu_2(\{\eta_i\}_{i=1}^{b})$ we note that each SB in the Tulip design is entirely covered by at least one SB in the original design. To see this, we note that each SB in Tulip design consists of three arcs, say, $i-1, i, i+1$ where their orders n_{i-1}, n_i, n_{i+1}, are even, odd, and even, respectively. We note that $|n_{i-1} - n_i| = 1$ and $|n_{i+1} - n_i| = 1$ since each marker is singular and hence the sets of the beams that cover any two adjacent arcs exactly differ by one beam. Either, we have $n_{i-1} > n_i > n_{i+1}$ in which case all the beams that cover the arc $i+1$ would cover the arcs $i-1$ and i which means that in the original design, there is a beam that covers all three arcs $i-1, i, , i+1$. Hence, in the Tulip design the beam that covers exactly the arcs $i-1, i, i+1$ is covered by a beam in the original design. If $n_{i-1} > n_i$, $n_i < n_{i+1}$, all the beams that cover the arc i have to cover the arc $i-1$ and $i+1$. If $n_{i-1} < n_i < n_{i+1}$, all the beams that cover the arc $i-1$ have to cover the arc i and $i+1$. Finally, if $n_{i-1} < n_i$, $n_i > n_{i+1}$, all the beams that cover the arc $i-1$ are the same as the beams that cover the arc $i+1$ and they cover the arc i as well, but the arc i covers exactly one different beam. Hence, it is immediate that the average (or max) of SB for a tulip design that matches the same marker positions is not more than the average (or max) of SB for the original design. On the other hand, we note that for both the original design and the Tulip design the position of the markers is the same. Moreover, the set of TBs for Tulip design is all possible $2b$ arcs. Since any other beam design including the original design can at most distinguish the same $2b$ arcs, hence, any measure function (e.g., max or mean) on the set of TBs for the Tulip design is always less

than or equal to that of the original design. This means that the trade-off curve is achieved by the Tulip design.

Step3: We note that after replacing the original SBs with the Tulip design as described in Step 2, it is possible to undo the operation of Step 1 and revert back by collapsing the corresponding group of singular markers into a single marker, for each non-singular marker. This would not affect the correctness of the arguments in step 2. This can also be interpreted as taking the limit over the differential value δ as it goes to zero which obviously collapses the corresponding set of singular points to its original non-singular point. We also note that the introduction of Step 1 is necessary for the way that the Tulip design is compared with the original design. Step 3 will also reveal that the design that achieves the trade-off curve might be a special form of the Tulip design where, e.g., some arcs are diminishing to zero.

We note that the above theorem holds for any channel distribution, i.e., the distribution of the paths. For the case of the multi-path channel, i.e., $p \geq 2$, Theorem 21 does not necessarily hold true for an arbitrary policy. The reason lies in the fact that the TB is in general composite, i.e., it consists of multiple arcs, and the composition of TBs is a function of the design. Therefore, in the evaluation section, we present the achievable trade-off curve for the Tulip design in comparison to that of the other designs without any optimality claim.

7.5.2 Analytical Results for Uniform Distribution

The trade-off curve for uniform distribution somewhat exhibits the worst-case performance. In this section, we drive the trade-off curve for both max and average measures under uniform distribution. The results can also facilitate the visualization of the trade-off curve for more

general cases even though the shape and the endpoints of the trade-off curve vary for different distributions. First, by using Theorem 21 we derive the closed-form solution for the trade-off curve.

Theorem 22 *The trade-off curve for the average measure* $(\bar{\eta}, \bar{\lambda}) = \big(\mu_2(\{\eta_i\}_{i=1}^b), \mu_2(\{\lambda_{\mathcal{P}}(s)\}_{s\in S})\big)$ *and under uniform distribution follows a polynomial from order* 2.

Proof. Consider a Tulip design with $2b$ CBs where $\{\alpha_i\}_{i=1}^b$ and $\{\beta_i\}_{i=1}^b$ denote the CBs of degree 1 and 2 respectively. For a given $\mu_2(\{\eta_i\}_{i=1}^b)$ it holds that,

$$\mu_2(\{\eta_i\}_{i=1}^b) = 1/b \sum_{i=1}^{b} \eta_i = (1/b) \sum_{i=1}^{b} (\alpha_i + 2\beta_i) =$$
$$(1/b) \sum_{i=1}^{b} (\alpha_i + \beta_i) + (1/b) \sum_{i=1}^{b} \beta_i = \frac{2\pi}{b} + (1/b) \sum_{i=1}^{b} \beta_i \tag{7.25}$$

Hence

$$\sum_{i=1}^{b} \alpha_i = 4\pi - b\bar{\eta} \quad \text{and} \quad \sum_{i=1}^{b} \beta_i = b\bar{\eta} - 2\pi \tag{7.26}$$

For uniform distribution on the users' AoD, we can write

$$\bar{\lambda} = 1/2\pi \sum_{i=1}^{b} (\alpha_i^2 + \beta_i^2) \tag{7.27}$$

For a given $\bar{\eta}$, the summation $\sum_{i=1}^b \alpha_i$ and $\sum_{i=1}^b \beta_i$ is fixed, hence, it is easy to verify that the minimum value of $\bar{\mu}$ in (7.27) is obtained when $\alpha_i = \alpha$ and $\beta_i = \beta$ for all $i \in [b]$ for proper

values of α and β. We have,

$$\bar{\lambda} = b/2\pi \left((\alpha+\beta)^2 - 2\alpha\beta\right) \tag{7.28}$$

Following equation (7.26) we get $\alpha = 2\gamma - \bar{\eta}$ and $\beta = \bar{\eta} - \gamma$. Replacing these values in (7.28) it immediately follows that

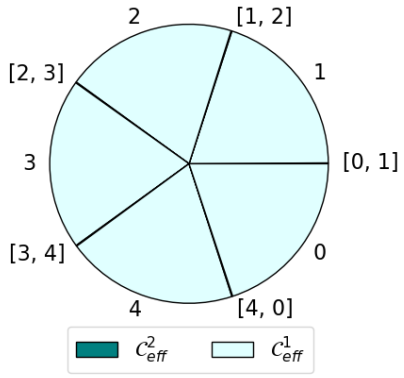
$$\bar{\lambda} = (2/\gamma)\bar{\eta}^2 - 6\bar{\eta} + 5\gamma \tag{7.29}$$

Hence, $\bar{\lambda}$ is a polynomial of degree 2 in terms of η. ■

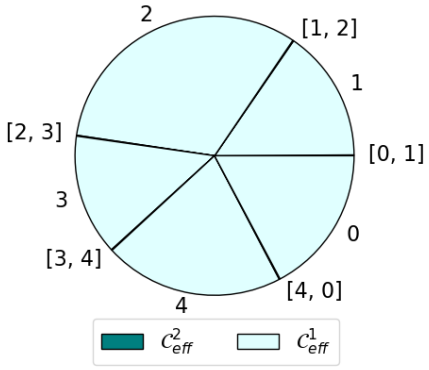
We note that for any SB design, we have $\gamma \leq \bar{\eta} \leq 2\gamma$. On the other hand, the analytical curve in (7.29) has its minimum in $\bar{\eta} = 1.5\gamma$ and it is symmetric with respect to this point. Since a good operating point for $(\bar{\eta}, \bar{\lambda})$ is where both $\bar{\eta}$ and $\bar{\lambda}$ are minimized, it is trivial that the trade-off curve only for $\gamma \leq \bar{\eta} \leq 1.5\gamma$ is of interest. For example for $\gamma \leq \bar{\eta} \leq 1.5\gamma$, both $(\bar{\eta}, \bar{\lambda})$ and $(3\gamma - \bar{\eta}, \bar{\lambda})$ are on the trade-off curve and obviously $(\bar{\eta}, \bar{\lambda})$ is much more efficient than $(3\gamma - \bar{\eta}, \bar{\lambda})$. Hence, we focus our attention to the curve between $\gamma \leq \bar{\eta} \leq 1.5\gamma$.

7.6 Performance Evaluation

In this section, we evaluate the performance of our beam alignment scheme in the multi-path environment for mentioned policies, i.e., $\mathcal{P} \in \{SD, BF, p-SD, p-BF\}$, and also examine the notion of trade-off curve by means of numerical simulations. We characterize the solutions for different policies. We strive to provide additional insights based on our extensive simulations. In practice the user channels have only a few resolvable paths, hence, we mainly focus on $p = 2, 3$.

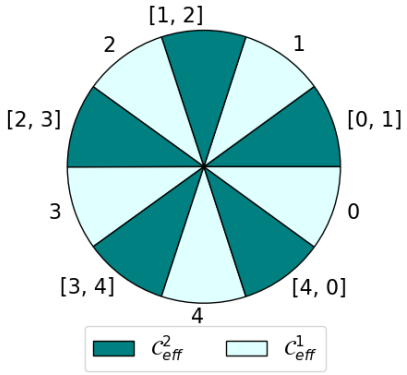


(a) SD Policy, $\bar{\lambda} = 2.26$

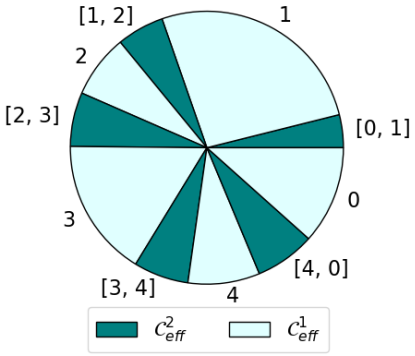


(b) BF Policy, $\bar{\lambda} = 1.145$

Figure 7.3: $p = 2$, $b = 5$, $N = 1000$, Uniform PDF



(a) $2 - SD$ Policy, $\bar{\lambda} = 1.822$



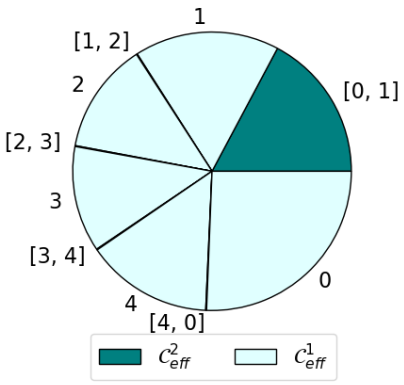
(b) $2 - BF$ Policy, $\bar{\lambda} = 0.836$

Figure 7.4: $p = 2$, $b = 5$, $N = 1000$, Uniform PDF

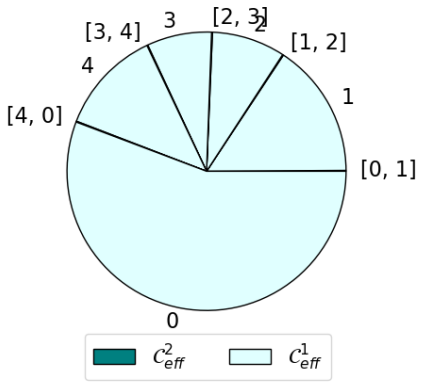
We consider uniform distribution, the cut-normal distribution, i.e., $\mathcal{N}(\mu = \pi, \sigma = 1)$ that is truncated beyond the range $(0, 2\pi)$.

Fig. 7.3-7.5 demonstrate the result of BA under the Tulip design for $b = 5$ and $p = 2$. Each figure shows the CB set that is the output of running algorithm 8 under the given BA policy and the given average PDF of all users. The labels corresponding to each CB are tagged on the corresponding arc.

Fig. 7.3 depicts the result of the BA scheme under the SD and the BF policies where



(a) SD Policy, $\bar{\lambda} = 1.76$



(b) BF Policy, $\bar{\lambda} = 0.71$

Figure 7.5: $p = 2$, $b = 5$, $N = 1000$, Cut-Normal PDF

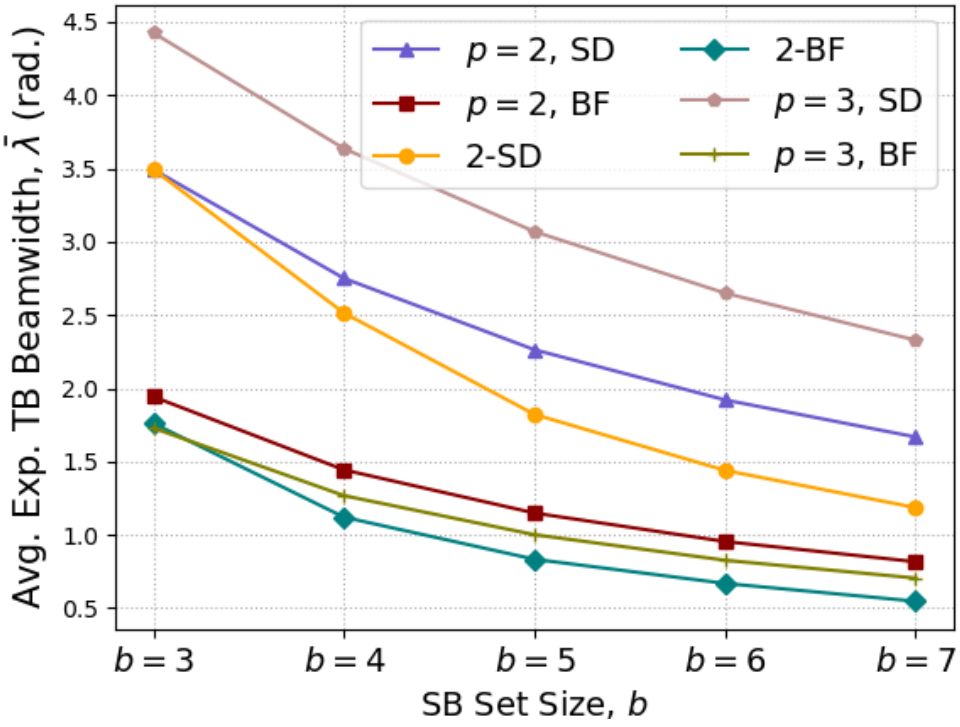


Figure 7.6: Impact of varying the size of SB set on the BA solution

uniform angular distribution is considered for the AoD of the users. This PDF can be interpreted as having no prior information on the AoD of the users. As intuitively expected, it is observed that the solution to the BA under the SD policy is an ES. In other words, the component beams in the set $\mathcal{C}_{\text{eff}}^2$ take zero lengths, and the beams in $\mathcal{C}_{\text{eff}}^1$ take equal lengths. We observe that the solution to the BA under the BF policy is not ES anymore but it takes the form of a GES. Through extensive simulations, we have observed that, for uniform distribution and $b > 2$, the solution to BF and SD policies are always a GES, and ES, respectively. However, it is not trivial to analytically prove this result. Please note that the solution for BF policy and $b = 2$ is not GES. One can easily verify that the optimal solution for BF policy is not unique as any permutations of the b arc would have the same $\bar{\lambda}$, nonetheless, the minimum value of $\bar{\lambda}$ is unique. Indeed, the presented greedy algorithm always converges to the same minimum value for any initialization. We note that the average expected TB beamwidth under the BF policy is less than that of the SD policy which means that the BF policy has a higher beamforming gain. Given that the solution to BF is GES and the permutations of the arc are not important, one can write the objective function explicitly in terms of the length of the arcs in the form of a polynomial of degree b. The solution which minimizes this polynomial can be numerically found. For example, for BF policy, the solution in terms of the length of the arcs for $b = 3, 4, 5$ is $(1.36, 1.93, 2.99)$, $(1.13, 1.27, 1.52, 2.36)$, $(0.89, 0.97, 1.09, 1.31, 2.02)$ with minimum values of 1.94, 1.44, 1.14, respectively. While in SD policy the length of the arcs are equal, in BF policy the length of the arcs vary due to the fact that in BF policy we take the arc with minimum ACI when we receive positive feedback on more than one SB.

Fig. 7.4 depicts the BA result under the p-SD and the p-BF policies for $p = 2$. It is observed that under the 2-SD policy all the CBs take equal length. It should be pointed out that

for the 2-BF policy the greedy algorithm 8 converges to multiple (but usually a small number of) solutions by using different initialization. Hence, we take the minimum of such a solution. We observe that the value of $\bar{\lambda}$ decreases under the 2-SD and the 2-BF policies compared to that of the SD and the BF policies and this is due to the luxury of having the information on the value of p.

Fig. 7.5 shows the result of BA when a cut-normal PDF $\mathcal{N}(\pi, 1)$ truncated between $[0, 2\pi)$ is considered. It is observed that the SD policy is not GES anymore while the BF policy still generates a GES solution. A cut normal distribution on average PDF of the users may be interpreted as having prior knowledge about the users' AoDs. For example, the mean and variance of the users' AOD from prior beam alignment frames may be used to fit a cut normal distribution on the PDF of the users' AoDs in the current frame. Comparing $\bar{\lambda}$ for uniform and cut-normal distributions both for SD and BF policies illustrates that the prior knowledge about the users' AoDs results in a smaller $\bar{\lambda}$.

Fig. 7.6 demonstrates the expected average beamwidth of the users, $\bar{\lambda}$, for different policies as a function of the size of the SB set b. For the channels with uniform distribution and $p = 2$ paths, it is observed that BF policy results in a smaller $\bar{\lambda}$ than the SD policy, hence, achieves higher beamforming gain. The same is true for the channel with $p = 3$. Notably, if the channel has a larger number of paths, say $p = 3$ vs. $p = 2$, the $\bar{\lambda}$ for BF policy decreases as having additional paths allows for choosing the minimum among additional possibilities. However, the SD policy would need to cover more paths and hence has a larger $\bar{\lambda}$. Fig. 7.6 also shows that using the information about the number of paths in $2-SD$ and $2-BF$ policies versus SD and BF policies would result in a better performance, i.e., smaller $\bar{\lambda}$, respectively. Finally, we note that $\bar{\lambda}$ reduces for all policies as the number of scanning beams increases.

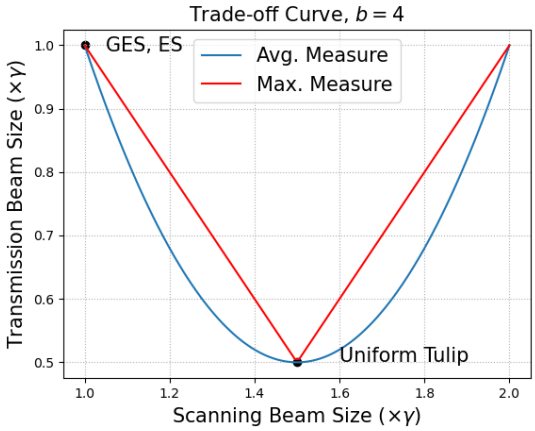


Figure 7.7: Theoretical curve for Uniform dist.

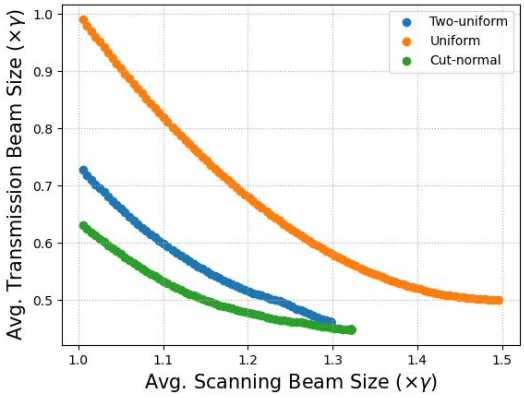


Figure 7.8: Empirical curves for general dist.

7.6.1 Trade-off Curve Numerical Analysis

The trade-off curve for arbitrary distribution in general may not admit an analytical form. Theorem 21 proves useful in this case by restricting the set of scanning beams to Tulip design. We use Algorithm 8 to compute the points on the trade-off curve for a given value of $\bar{\eta} = \bar{\eta}_0$. We start by setting $\bar{\eta}_0 = \gamma$ and increase its value by a small value δ at each step. At each step for a given value of $\bar{\eta}_0$, we find the minimum $\bar{\lambda}$ by simply iterating Algorithm 8 for a given number of trials where at least one such starting set of scanning beams defined by $\{z_i\}_{i=1}^{2b}$ at each step is the one that correspond to the minimum value of $\bar{\lambda}$ in previous step. As we discussed earlier, the trade-off curve for the uniform distribution achieves its minimum at $\bar{\eta} = 1.5\gamma$, and any other distribution beside uniform achieves its minimum at $\bar{\eta} < 1.5\gamma$. Hence, it is not necessary to search for the value of $\bar{\eta} > 1.5\gamma$. After the trade-off curve reaches its minima say at $\bar{\eta}^*$ increasing the value of $\bar{\eta} > \bar{\eta}^*$ would not improve the curve. This is verified through our numerical evaluation as well. For empirical measurements, We consider uniform distribution, the cut-normal distribution, i.e., $\mathcal{N}(\mu = \pi, \sigma = 1)$ that is truncated beyond the range $(0, 2\pi)$, and the piece-wise uniform

distribution with two segments $A = [0, 2\pi/3]$ with $p(A) = 3/4$, and $B = [2\pi/3, 2\pi]$ with $p(B) = 1/4$. Fig. 7.8 illustrates the trade-off curve for $b = 5$ beams using average measures evaluated based on the proposed algorithm for these three distributions. As derived in Section 7.5 the trade-off curve for uniform distribution is a convex curve that starts from $\bar{\eta} = \gamma$ and admits its minimum at $\bar{\eta} = 1.5\gamma$. It is easy to verify that the trade-off curve for the average measure is convex since time sharing between every two SB sets can be used to generate points on the line connecting the performance points of these two SB sets. The trade-off curves for the same number of beams b for non-uniform distributions are always strictly below that of the uniform distribution. This can be interpreted as uniform distribution as the worst case in terms of trade-off. We note that the other two trade-off curves for both piece-wise uniform distribution, and cut-normal distribution in Fig. 7.8 start from $\bar{\eta} = \gamma$ and take minima at a point for which $\bar{\eta} < 1.5\gamma$. Fig. 7.7 depicts the theoretical trade-off for uniform distribution on the user AoD for both measures discussed earlier. The curve corresponds to the Avg. and Max. measures follow parabolic (equation (7.29)) and linear curves, respectively. We note that the point (γ, γ) on the Avg. trade-off curve corresponds to the GES scheme and on the Max. trade-off curve to the ES scheme.

7.7 Conclusion

We studied the non-interactive multi-user beam alignment problem in mmWave systems while considering the effect of multi-path. We introduced the *Tulip design* for the scanning beams in the probing phase and proved its optimality in terms of maximizing the achievable number of feedback sequences when the scanning beams are contiguous. We modeled beam alignment as an optimization problem under different policies and proposed a greedy algorithm to find the optimal

BA scheme. We characterized the solutions of our BA scheme through numerical experiments and presented our observations. We reveal a fundamental trade-off between the SBs and TBs gains and define the notion of the trade-off curve. We show a fundamental class of SB set design, namely *Tulip Design* achieves the trade-off curve. We present a closed-form solution for the trade-off curve for special channels with uniform distribution on AoA. For general distributions, we provided an algorithm and evaluation results to find the trade-off curve. Our results emphasize that the state-of-the-art beam search algorithms should further optimize their SB sets based on this fundamental trade-off between the SB penetration and TB gain.

Chapter 8: RIS-aided mm-Wave Beam-forming for Two-way Communications of Multiple Pairs

8.1 Overview

The next generation of wireless communication systems aims to address the ever-increasing demand for high throughput, low latency, better quality of service, and ubiquitous coverage. The abundance of bandwidth available at the mmWave frequency range, i.e., $[20, 100]$ GHz, is considered a key enabler towards the realization of the promises of next-generation wireless communication systems. However, communication in mmWave suffers from high path loss and poor scattering and diffraction. Therefore, mmWave signals are vulnerable to blockages, especially in urban areas. Since the channel in mmWave is mostly line-of-sight (LoS), i.e., a strong LoS path and very few and much weaker secondary components, the mmWave coverage map includes *blind spots* as a result of shadowing and blockage. Beam-forming employing massive MIMO is primarily used to address the high attenuation in the mmWave channel. In addition to beam-forming, relaying can be designed to generate constructive superposition and enhance the received signals at the receiving nodes. Equipping MIMO communication systems with Reconfigurable Intelligent Surfaces (RISs) may extend the capacity of mmWave by covering the blind spots and providing diverse reception at the receiving nodes.

RIS [156] [157] [158] is denoted as a potential enabling technology for realizing 6G-and-beyond [159] [160], due to its great potential for manipulating the impinging electromagnetic waves and artificially shaping the wireless propagation environment in a cost-effective and energy-efficient manner. The wireless propagation environment in current communication systems is considered to be uncontrollable and stochastic, and therefore, the existing wireless system design methods and principles (e.g. mmWave communication systems, massive MIMO, etc.) are traditionally developed on a reactive basis to adapt to this stochastic behavior. An RIS typically consists of a large number of low-cost reflecting elements arranged in planar artificial metasurfaces. Each RIS cell is capable of manipulating the phase and the amplitude of incident electromagnetic waves in response to real-time external signals provided by a smart controller. The programmability of the RIS enables them to flexibly modulate RF signals without the need to use any mixers, analog phase-shifters, analog-to-digital/digital-to-analog converter, etc. [161] [162]. This not only drives the RIS hardware cost and energy consumption down but also allows for the design of RF chain-free wireless transceivers [163] [164]. Therefore, either as active transceivers or passive reflectors, RIS may be a promising solution to revolutionize the design of the physical layer in next-generation wireless communication systems.

8.1.1 Related work

RISs have attracted a lot of attention both from industry and academia, over the past few years. The effort toward realizing the prospects of achieving RIS-aided wireless communications initially started with theoretical developments based on mathematical models. More recently, real-world trials with prototyping and field experiments have been pursued more seriously than

before in the literature. Several funding agencies have heavily invested in theoretical development, designing, prototyping, and testing the intelligent metamaterial surfaces. Since 2012, the National Science Foundation (NSF) in the US, and the European Commission, Horizon 2020, have spent millions of dollars on projects tackling different aspects of the metasurfaces and integrating the RIS with next-generation networks [165].

As far as experimental trials are concerned, extensive efforts have been made. The Japanese network operator NTT DoCoMo in collaboration with its partners has repeated a few successful trials over the past couple of years, demonstrating how their designed transparent dynamic metasurface can improve the power of the radio signal at 28 GHz [166]. In [167], the authors have proposed a prototype for RIS-enabled wireless communications with an RIS consisting of 1100 elements operating at 5.8 GHz. They show their passive reflect-array design provides significant gain improvement in point-to-point wireless communications through indoor and outdoor field trials. Another prototype is proposed in [168] for modulating the incident waves based on single-carrier Quadrature Phase Shift Keying (QPSK) to design RIS-based transceivers achieving a 2.048 Mbps data rate for video streaming. The prototype in [163] achieves an RIS-based RF-chain free transceiver.

RISs are promising to be deployed in a wide range of communications scenarios, such as high throughput MIMO communications [164] [169], ad-hoc networks, e.g., UAV communications [170] [171], physical layer security [172], etc. In radar, deployment of RIS with the judicious design of phase shifts has shown improvement in the estimation of the radar cross-section [173] and moving target [174]. Apart from the work focusing on theoretical performance analysis of RIS-enabled systems [175] [176], a considerable amount of work has been dedicated to optimizing such an integration, mostly focusing on the phase optimization of RIS elements

[177] [178] [179] to achieve various goals such as maximum received signal strength, maximum spectral efficiency, etc. For more information on the challenges and opportunities associated with RIS, we refer interested readers to [180] [181] and the references therein. In this chapter, we present an RIS-aided architecture to facilitate two-way communications [143] [182] [183] [184]. Most of the prior art in RIS-aided two-way communications, focuses on single-beam communications [142] [185]. However, in this chapter, we propose the idea of RIS-aided multi-beam-forming. The idea of RIS-aided multi-beam-forming, i.e. beam-forming with multiple disjoint lobes was first introduced in [143], where the authors aimed at covering the mmWave blind spots by designing sharp beams covering various ranges of solid angle. The codebook design problem for such beamforming was addressed in [186]. In [187], the authors employ dual beamforming for short-range target monitoring.

8.1.2 Main contributions

In this chapter, we consider a communication scenario between a transmitter, e.g., the Base Station (BS), and terrestrial end users through a passive RIS that reflects the received signal from the transmitter toward the users. Hence, the users that are otherwise in blind spots of network coverage become capable of communicating with the base station through the RIS that is serving as a passive reflector (passive relay) maintaining communication links between the BS and the users. Given the geospatial variance among the locations of the end users served by the same wireless system, the RIS may have to simultaneously accommodate users that lie in different angular intervals that are widely separated with a satisfactory Quality of Service (QoS). In what we refer to as *multi-beamforming*, we particularly address the design of beams consisting of

multiple disjoint lobes to cover different blind spots using sharp, high gain, and effective beam patterns. In the following, we summarize the main contributions of this paper:

- **[RIS-UPA to UPA Transformation]** We present simple yet important properties of RIS with UPA structure (RIS-UPA) when used as a beam-former in Section 8.3. Accordingly, we present a transformation between the beam-former design problem in UPA and RIS-UPA which allows us to directly borrow the design for UPA beam-forming and through a transformation use it for RIS-UPA beam-forming.

- **[Multi-beamforming]** We present a new beam-forming design technique termed as *multi-beamforming* aiming to design beams with multiple disjoint lobes. The multi-beamforming design inherently depends on the solid angle (say Ω_1 in Fig. 8.1) at which the incident wave activates the RIS elements. The proposed beam-forming design easily adapts to changes in Ω_1 and we provide a visualization as to how the beam would change in response to change in Ω_1.

- **[Custom footprint]** The proposed method has the flexibility to design a beam with a custom footprint. The beam footprint may be defined as the cross-section of the beam lobes with the sphere (say at beam gains within the half-power (3dB) point and maximum gain).

- **[Compound beams]** We design the parameters of the RIS to achieve multiple disjoint beams covering various ranges of a solid angle. The designed beams are fairly sharp, have almost uniform gain in the desired Angular Coverage Interval (ACI), and have negligible power transmitted outside the ACI.

- **[Complexity]** Thanks to the analytical closed-form solutions for the multi-beamforming design, the proposed solution bears very low computational complexity even for an RIS with massive array sizes.

- **[Multilink]** We provide RIS beam-forming design in multilink scenarios where different pairs of transmitters and receivers are communicating simultaneously with the help of the same RIS. Moreover, we establish a connection between multilink beamforming and multi-beamforming.

- **[Evaluation]** Through numerical evaluation we show that by using a passive RIS, multi-beamforming can simultaneously cover multiple ACIs. Moreover, multi-beamforming provides tens of dB power boost w.r.t. the single-beam RIS design.

8.1.3 Notations

Throughout this paper, $\mathbb{C}$, $\mathbb{R}$, and $\mathbb{Z}$ denote the set of complex, real, and integer numbers, respectively, $\mathcal{CN}(m, \sigma^2)$ denotes the circular symmetric complex normal distribution with mean m and variance σ^2, $[a, b]$ is the closed interval between a and b, $[m]$ is the set of m positive integers less than or equal to m, $[(m, n)]$ is the set of all $m \times n$ integer pairs with the first element less than m and the second element less than n, $\mathbf{1}_{a,b}$ is the $a \times b$ all ones matrix, $\mathbf{I}_N$ is the $N \times N$ identity matrix, $\mathbb{1}_{[a,b)}$ is the indicator function, $\|\cdot\|$ is the 2 -norm, $\|\cdot\|_\infty$ is the infinity-norm, $|\cdot|$ may denote cardinality if applied to a set and 1-norm if applied to a vector, $\odot$ is the Hadamard product, $\otimes$ is the Kronecker product, the operator diag$\{.\}$ when applied to a square matrix takes the vector of its diagonal elements, and when applied to a vector of elements, forms a diagonal matrix of that vector, $\mathbf{A}^H$, and $\mathbf{A}_{a,b}$ denote conjugate transpose, and $(a, b)^{th}$ entry of $\mathbf{A}$ respectively. We

have summarized the list of variables and parameters frequently used in the paper in table 8.1.

Table 8.1: Frequently-used parameters and variables

Variables	Description
$\boldsymbol{\Theta}$	The decision variables representing the coefficients of RIS elements manipulating the impinging electromagnetic waves
$\boldsymbol{\lambda}$	The decision variables representing the overall impact of the RIS when excited from an incident solid AoA of Ω_1
$\mathbf{c}$	The normalized version of λ, or equivalently, the normalized beam-forming vector corresponding to a Uniform Planar Array (UPA) of antennas
$\mathbf{g}$	Equal-gain vector that has to be optimally chosen when deciding the optimal configuration $\mathbf{c}$.
$\Gamma(\Omega_1, \Theta, \Omega)$	The gain of an RIS configured by $\boldsymbol{\Theta}$ at AoD Ω when excited from an incident angle Ω_1.
$G(\xi, \zeta, \boldsymbol{\lambda})$	The beam-forming gain of a UPA of antennas configured by $\boldsymbol{\lambda}$ at AoD $\psi = [\xi, \zeta]$
η	The design parameter embedded in $\mathbf{g}$ determining the quality of the beams formed by the multi-beamforming approach.
Parameters	**Description**
$\mathbf{H}_t$	The effective channel matrix from the transmitter to the RIS
$\mathbf{H}_r$	The effective channel matrix from the RIS to the receiver
$\mathbf{a}_M(\Omega)$	The array response vector of an array of M antennas or an RIS consisting of M elements at the solid angle Ω
ρ_t	The gain of the Line-of-Sight (LoS) path from the transmitter to the RIS
ρ_r	The gain of the Line-of-Sight (LoS) path from the RIS to the receiver
Ω_t	The Angle of Departure (AoD) from the transmitter towards the RIS
Ω_r	The Angle of Arrival (AoA) from the RIS at the receiver
Ω_1	The Angle of Arrival (AoA) from the transmitter at the RIS
Ω_2	The Angle of Departure (AoD) from the RIS towards the receiver
$\tau(\Omega)$	The operator τ maps the solid angle $\Omega = [\phi, \theta]$ to $\psi = [\xi, \zeta] = \tau(\Omega)$
$\mathbf{d}_M(\psi)$	The directivity vector of an array of antennas at direction ψ
$\mathcal{B}$	The total angular range under study in the Ω domain
$\mathcal{B}^\psi$	The total angular range under study in the ψ domain
$\mathcal{D}_n$	The n-th receive zone in the n-th communication pair
$\mathcal{A}_n$	The minimal index set of the particles covering the n-th receive zone $\mathcal{D}_n$
$\mathcal{B}^\psi_{p,q}$	The area covered by the (p, q)-th particle in the ψ-domain
$\mathbf{e}_{p,q}$	The vector with a 1 in the (p, q)-th entry out of $[(Q_v, Q_h)]$ ones and zero everywhere else

8.2 System model

8.2.1 Channel model

Consider a communications system with a multi-antenna BS with M_t antenna elements as a transmitter and a multi-antenna receiver with M_r antenna elements. The MIMO system is aided by a multi-element RIS consisting of M elements arranged in an $M_h \times M_v$ grid in the form of a UPA as shown in Fig. 8.1, where M_h and M_v are the number of elements in the horizontal and vertical directions, respectively. The received signal $\mathbf{y} \in \mathbb{C}^{M_r}$ as a function of the transmitted

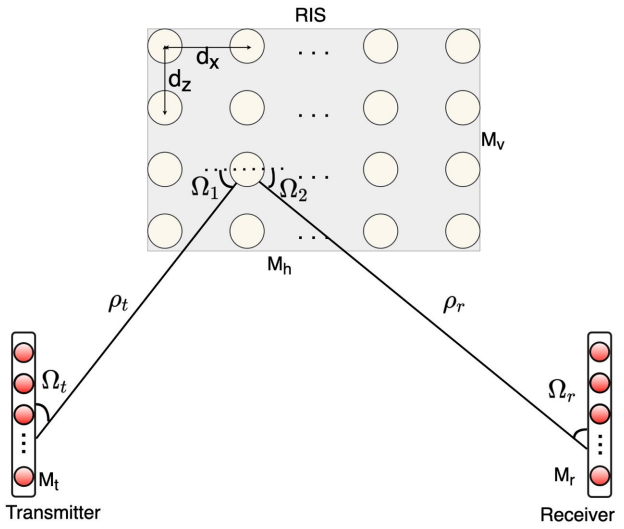


Figure 8.1: System model

signal $\mathbf{x} \in \mathbb{C}^{M_t}$ can be written as,

$$\mathbf{y} = (\mathbf{H}_r \mathbf{\Theta} \mathbf{H}_t)\mathbf{x} + \mathbf{z} \tag{8.1}$$

where $\mathbf{z}$ is the noise vector, with each element of $\mathbf{z}$ is drawn from a complex Gaussian distribution $\mathcal{CN}(0, \sigma_n^2)$, $\mathbf{H}_t \in \mathbb{C}^{M \times M_t}$ and $\mathbf{H}_r \in \mathbb{C}^{M_r \times M}$ are the channel matrices between each party and the RIS. We assume that the RIS consists of elements for which both the phase θ_m and the gain β_m (in form of the attenuation of the reflected signal) of each element, say m, may be controlled and $\mathbf{\Theta} \in \mathbb{C}^{M \times M}$ is a diagonal matrix where the element (m, m) denotes the coefficient $\beta_m e^{j\theta_m}$ of the m^{th} element of the RIS. Assuming a LoS channel model both between the transmitter and the RIS and between the RIS and the receiver and using the directivity vectors at the transmitter,

the RIS, and the receiver, the effective channel matrices can be written as,

$$\mathbf{H}_r = \mathbf{a}_{M_r}(\Omega_r)\rho_r \mathbf{a}_M^H(\Omega_2) \tag{8.2}$$

$$\mathbf{H}_t = \mathbf{a}_M(\Omega_1)\rho_t \mathbf{a}_{M_t}^H(\Omega_t) \tag{8.3}$$

where $\mathbf{a}_M(\Omega)$ is the array response vector of an RIS with elements in a UPA structure (RIS-UPA), Ω_t and Ω_2 are the solid Angles of Departure (AoD) of the transmitted beams from the transmitter and the RIS and Ω_1 and Ω_r are the solid Angles of Arrival (AoA) of the received beams at the RIS and the receiver, respectively.

The gains of the LoS paths from the transmitter to the RIS and from the RIS to the receiver are denoted by ρ_t and ρ_r, respectively. Note that an arbitrary solid angle Ω specifies a pair of elevation and azimuth angles (ϕ, θ). Therefore, in the above definitions we set $\Omega_a = [\phi_a, \theta_a]$, $a \in \{1, 2, t, r\}$. Further, assuming no pairing between the RIS elements, $\mathbf{\Theta}$ will be a diagonal matrix specified as

$$\mathbf{\Theta} = \text{diag}\{[\beta_1 e^{j\theta_1}, \ldots, \beta_M e^{j\theta_M}]\} \tag{8.4}$$

where $\beta_i \in [0, 1]$ and $\theta_i \in [0, 2\pi]$. Using equations (8.1)-(8.3), the contribution of the RIS to the channel matrix when excited from incident angle Ω_1 for a receiver at a given solid angle solid angle Ω is given by

$$\Gamma(\Omega_1, \mathbf{\Theta}, \Omega) = \mathbf{a}_M^H(\Omega)\mathbf{\Theta}\mathbf{a}_M(\Omega_1) = \mathbf{a}_M^H(\Omega)\boldsymbol{\lambda} \tag{8.5}$$

where we define $\boldsymbol{\lambda} = \mathbf{\Theta}\mathbf{a}_M(\Omega_1)$, $\boldsymbol{\lambda} \in \mathbb{C}^M$.

8.2.2 RIS model

Consider a RIS consisting of $M_v \times M_h$ antenna elements forming a UPA structure that is placed at the x-z plane, where $M \doteq M_v M_h$ and the z-axis corresponds to the horizon. Let d_z, and d_x denote the distance between the antenna elements in the z and x axis, respectively. Throughout this paper, we assume that the transmitter and the receiver are within the far field of the RIS. As opposed to the near-field regime, the diversity gain distribution in the far field does not depend on the distance between the transmitter and the RIS [188]. RIS-enabled communications in the near field require a thoroughly different design. For an example of such a specification, please refer to [189]. The array response vector of an RIS-UPA can be found in a similar way to that of a UPA. At a solid angle $\Omega = [\phi, \theta]$, we have

$$\mathbf{a}_M(\Omega) = \left[1, e^{j\frac{2\pi}{\lambda}\mathbf{r}_\Omega \mathbf{r}_1}, \ldots, e^{j\frac{2\pi}{\lambda}\mathbf{r}_\Omega \mathbf{r}_{M-1}}\right]^T \in \mathbb{C}^M \tag{8.6}$$

where we define $\mathbf{r}_\Omega = [\cos\phi\cos\theta, \cos\phi\sin\theta, \sin\phi]$, and $\mathbf{r}_m = (m_h d_x, 0, m_v d_z)$ to respectively denote the direction corresponding to the solid angle Ω, and the location of the m-th RIS element corresponding to the antenna placed at the position (m_v, m_h) with $m = m_v M_h + m_h$.

Further, we define a transformation of variables as follows. For a solid angle $\Omega = [\phi, \theta]$, define $\psi = [\xi, \zeta]$

$$\xi = \frac{2\pi d_z}{\lambda}\sin\phi, \quad \zeta = \frac{2\pi d_x}{\lambda}\sin\theta\cos\phi \tag{8.7}$$

Introducing the new variables into equation (8.6), it is straightforward to write,

$$\mathbf{a}_M(\Omega) = \mathbf{d}_M(\psi) = \mathbf{d}_{M_v}(\xi) \otimes \mathbf{d}_{M_h}(\zeta) \in \mathbb{C}^M \tag{8.8}$$

where $\mathbf{d}_M$ denotes the directivity vector of the RIS, and the directivity vectors $\mathbf{d}_{M_a}$, $a \in \{v, h\}$ are defined as follows.

$$\begin{aligned} \mathbf{d}_{M_v}(\xi) &= \left[1, e^{j\xi} \cdots e^{j(M_v-1)\xi}\right]^T \in \mathbb{C}^{M_v} \\ \mathbf{d}_{M_h}(\zeta) &= \left[1, e^{j\zeta} \cdots e^{j(M_h-1)\zeta}\right]^T \in \mathbb{C}^{M_h} \end{aligned} \tag{8.9}$$

where $\psi_v = \xi$, and $\psi_h = \zeta$. Finally, let $\mathcal{B}$ be the angular range for Ω under our interest defined as follows,

$$\mathcal{B} = \left[-\phi^{\mathrm{B}}, \phi^{\mathrm{B}}\right) \times \left[-\theta^{\mathrm{B}}, \theta^{\mathrm{B}}\right) \tag{8.10}$$

Accordingly, let $\mathcal{B}^{\psi}$ be the angular range under interest in the (ξ, ζ) domain given by

$$\mathcal{B}^{\psi} = \left[-\xi^{\mathrm{B}}, \xi^{\mathrm{B}}\right) \times \left[-\zeta^{\mathrm{B}}, \zeta^{\mathrm{B}}\right) \tag{8.11}$$

In this chapter, we set $d_x = d_z = \frac{\lambda}{2}$, $\phi^{\mathrm{B}} = \frac{\pi}{4}$, and $\theta^{\mathrm{B}} = \frac{\pi}{2}$, hence $\xi \in [-\pi\frac{\sqrt{2}}{2}, \pi\frac{\sqrt{2}}{2})$, and $\zeta \in [-\pi, \pi)$. To formalize the variable transformation introduced in (8.7), we define the transformation operator $\tau : \mathcal{B} \longrightarrow \mathcal{B}^{\psi}$ as $\tau([\phi, \theta]) = [\xi, \zeta]$.

8.3 General Properties of RIS as beamformer

For a RIS-UPA excited by emission from solid angle Ω_1, we can write for the reference gain at any direction Ω,

$$\begin{aligned} |\Gamma(\Omega_1, \Theta, \Omega)| &= |\mathbf{a}_M^H(\Omega)\Theta\mathbf{a}_M(\Omega_1)| \\ &= \left|\sum_{m=0}^{M-1} \theta_{m,m} e^{-j(m_v\xi + m_h\zeta)} e^{j(m_v\xi_1 + m_h\zeta_1)}\right| \\ &= \left|\sum_{m=0}^{M-1} \theta_{m,m} e^{j(\tau(\Omega_1)-\tau(\Omega))\mathbf{m}}\right| \end{aligned} \tag{8.12}$$

where we define $\mathbf{m} = [m_v, m_h]^T$. In the following, we present three fundamental facts together with their interpretation, that are useful for our subsequent developments in the next sections. The proofs for all these facts are straightforward and can be achieved by basic calculus. First of all, from the identity,

$$|\Gamma(\Omega_1, \text{diag}\{\mathbf{a}_M^H(\Omega_1)\Theta\}, \Omega)| = |\Gamma(0, \Theta, \Omega)| \tag{8.13}$$

we note that the gain of a UPA with beam-forming matrix Θ at any solid angle Ω is equal to that of a RIS-UPA excited from an AOA Ω_1 with parameter matrix $\text{diag}(\mathbf{d}_M^H(\psi_1)\Theta)$. Second, for any solid angle Ω_2, with $\tau(\Omega_2) = \psi_2$, it holds that,

$$|\Gamma(\Omega_1, \Theta, \Omega)| = |\Gamma(\Omega_1, \text{diag}\{\mathbf{a}_M^H(\Omega_2)\Theta\}, \Omega')| \tag{8.14}$$

where $\tau(\Omega') = \tau(\Omega) - \tau(\Omega_2)$. The last identity implies that for two RIS-UPAs with parameter

matrices Θ and $\text{diag}\{\mathbf{d}_M^H(\psi_2)\Theta\}$ that are excited from the same AoA Ω_1, the gain patterns are just a rotation of each other such that the gain at direction Ω for the first one is equal to the gain at the direction of Ω' for the second one, where Ω' is as denoted above. The same holds for a UPA by setting $\Omega_1 = 0.$, i.e., the gain pattern of two UPAs with parameters Θ and $\text{diag}\{\mathbf{d}_M^H(\psi_2)\Theta\}$ are related by a rotation of each other such that the gain at direction Ω for the first one is equal to the gain at the direction of Ω' for the second one. Finally, it is straightforward to show,

$$|\Gamma(\Omega_1, \Theta, \Omega)| = |\Gamma(\Omega_2, \Theta, \Omega^{''})| \tag{8.15}$$

where $\tau(\Omega'') = \tau(\Omega) + \tau(\Omega_2) - \tau(\Omega_1)$. By virtue of the last identity, we note that for two RIS-UPAs with the same parameters Θ that are excited from two different AoA Ω_1 and Ω_2 the gain patterns are just a rotation of each other such that the gain at direction Ω for the first one is equal to the gain at the direction of Ω'' for the second one, where Ω'' is obtained as above. As we will see in Section 8.4.2, we use this property to transform a Multi-Transmitter Multi-Receiver (MTMR) communication system into a Single-Transmitter Multi-Receiver (STMR) system. Consequently, it will enable us to design a single RIS to accommodate communications between multiple disjoint pairs, possibly with large angular separations.

Please note that if a user is located at the solid angle Ω with respect to the RIS, then the AoA of the incident wave at the RIS when the user is transmitting is defined as $\Omega = [\phi, \theta]$. However, due to the change in the direction of the beam, the reflected beam from an RIS towards the user is directed at angle $\Omega^r = [-\phi, \theta + \pi]$. Hence, from (8.7) and by the definition of the operator τ, we have $\tau(\Omega^r) = -\tau(\Omega)$. From (8.12), for an RIS which is excited from an incident angle with

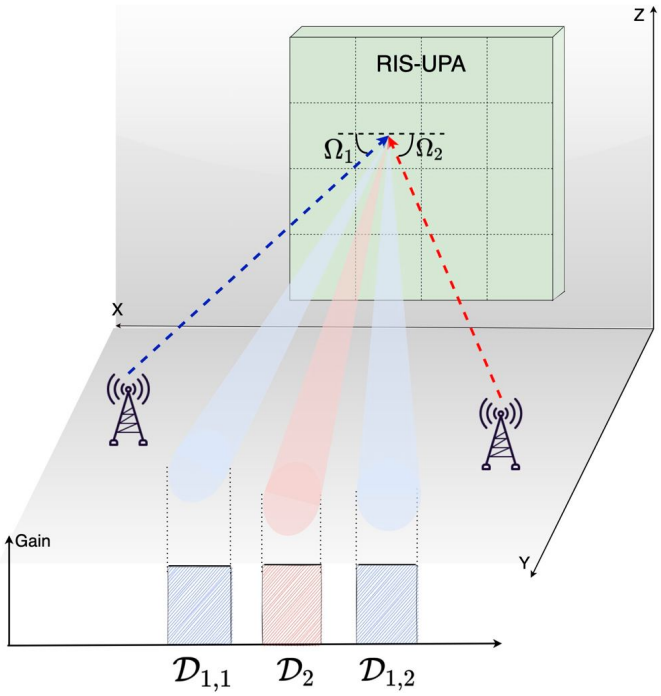


Figure 8.2: RIS-enabled two-way communications

AoA of $\Omega_1 = [\phi_1, \theta_1]$ reflects the signal with AoD of $\Omega_2 = [\phi_2, \theta_2]$, we have

$$|\Gamma(\Omega_1, \Theta, \Omega_2^r)| = |\Gamma(\Omega_2, \Theta, \Omega_1^r)| \tag{8.16}$$

where $\Omega_1^r = [-\phi_1, \theta_1 + \pi]$ and $\Omega_2 = [-\phi_2, \theta_2 + \pi]$. which means that the RIS has the same gain at the AoD of Ω_1^r when it is excited from the incident angle with AoA of Ω_2^r. This is trivially equivalent to the channel reciprocity property which indicates the possibility of two-way communications between each pair.

8.4 Problem Formulation

In this section, we show how RIS can be used to facilitate two-way communications between multiple pairs, simultaneously. We will first present the description of the problem

and explain our approach. We then proceed with the formulation of the problem and propose our solution.

8.4.1 Problem description

We study the RIS design problem for RIS-aided two-way communications where multiple communications pairs are involved. Each pair realizes a MIMO system consisting of a multi-antenna transmitter (e.g. multi-antenna BS), and a possibly multi-antenna receiver (e.g. multi-antenna mobile user). We assume there is no LoS channel between the pairs and the mmWave channel is assisted by an RIS with elements that are arranged according to a UPA structure.

Formally, we consider N communications pairs specified by $(\Omega_n, \mathcal{D}_n)$, $n = 1, \ldots, N$. The n-th transmitter is emitting at the RIS with an AoA of Ω_n, and the n-th receiver may reside within the receive zone $\mathcal{D}_n$, i.e., $\mathcal{D}_n$ represents the range of AoDs from the RIS that may cover the n-th receiver. For the n-th receive zone, we denote each continuous range of such AoDs by an Angular Coverage Interval (ACI). We note that each receive zone may comprise one or more ACIs. Fig. 8.2 depicts an example of such a setup for $N = 2$ pairs, where receive zone $\mathcal{D}_1 = \mathcal{D}_{1,1} \cup \mathcal{D}_{1,2}$ consists of two ACIs. As far as the n-th communications pair is concerned, the *ideal* RIS must be configured in such a way that when excited from solid angle Ω_n it covers the receive zone $\mathcal{D}_n$ uniformly, with high and sharp gains, while leaving minimal gain leakage to other intervals. In RIS-aided MTMR two-way communications, we aim to design the RIS such that all pairs are satisfied simultaneously with a high QoS. The simplest instance of the above problem is when $N = 1$, i.e. when there is only one pair. We denote this instance by STMR.

Furthermore, we note that (i) the boundaries of the ACIs are determined based on the

potential locations of the receivers, and, (ii) we wish to consider the ACIs of minimal size to avoid sacrificing the gain. Therefore, the footprint of the receive zones on spherical coordinates may become of arbitrary shape. This may introduce additional complexity to the RIS design problem. To resolve this issue, let us uniformly divide $\mathcal{B}^{\psi}$ into $Q = Q_v Q_h$ subregions, where Q_v and Q_h determine the division resolution in the vertical and horizontal directions, respectively. We denote each such subregion by a *particle* $\mathcal{B}^{\psi}_{p,q}$ that is specified as,

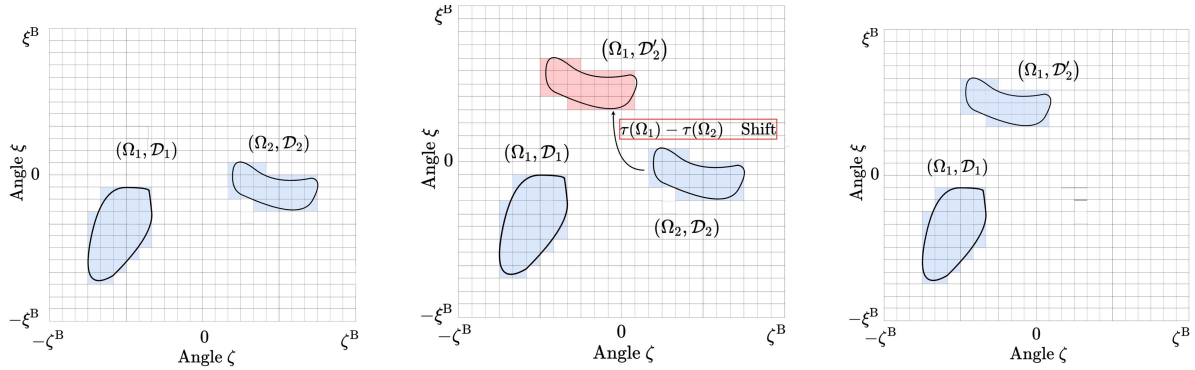
$$\mathcal{B}^{\psi}_{p,q} = \nu^{p}_{v} \times \nu^{q}_{h}, \quad p \in [Q_v], q \in [Q_h] \tag{8.17}$$

where $\nu^{p}_{v} = [\xi^{p-1}, \xi^{p}]$, and $\nu^{q}_{h} = [\zeta^{q-1}, \zeta^{q}]$ defining,

$$\xi^{p} = -\xi^{\mathrm{B}} + p\delta_v, \quad \zeta^{q} = -\zeta^{\mathrm{B}} + q\delta_h \tag{8.18}$$

with $\delta_v = \frac{2\xi^{\mathrm{B}}}{Q_v}$, and $\delta_h = \frac{2\zeta^{\mathrm{B}}}{Q_h}$. Further, define for all (p, q) pairs the notation $\delta_{p,q} = \delta_v \delta_h$. We wish to cover each receive zone $\mathcal{D}_n$ in the RIS-enabled MTMR problem with the smallest set of particles, i.e., $\mathcal{D}_n \sim \bigcup_{(p,q)\in\mathcal{A}_n} \mathcal{B}^{\psi}_{p,q}$, with $\mathcal{A}_n$ being the smallest set of index pairs (p, q) that beams $\mathcal{B}^{\psi}_{p,q}$ collectively cover $\mathcal{D}_n$. The union of $\mathcal{B}^{\psi}_{p,q}$ is in fact approximating the shape of the desired receive zone $\mathcal{D}_n$, where the resolution of the approximation is set by the pair (Q_v, Q_h). By using larger values of Q_v, and Q_h, one can opt for finer particles and boost the quality of the approximation at the expense of solving a larger optimization problem. Explicitly, we have

$$\mathcal{A}_n = \arg\min_{\{\hat{\mathcal{A}} | \mathcal{D}_n \subseteq \bigcup\limits_{(p,q)\in\mathcal{A}} \mathcal{B}_{p,q}\}} |\hat{\mathcal{A}}|, \quad n \in [N]. \tag{8.19}$$



(a) MTMR (b) Transforming MTMR to STMR (c) STMR

Figure 8.3: Transforming two-way MTMR to STMR communications

Further define $\mathcal{A} = \bigcup_{n=1}^{N} \mathcal{A}_n$. Next, we will show how every instance of MTMR can be cast as an STMR problem. Then we pose the two-way communications problem in the STMR case, as a composite beam-forming problem under the UPA antenna structure. We then propose a low-complexity closed-form for the optimal solution to the last problem.

8.4.2 Transformation between MTMR and STMR

Consider N communications pairs $(\Omega_n, \mathcal{D}_n)$, $n = 1, \ldots, N$, where Ω_n is the AoA of the n-th transmitter to the RIS-UPA plane, and $\mathcal{D}_n$ denotes the n-th receive zone. For an arbitrary RIS-UPA configuration $\hat{\Theta}$, and for any communications pair, by equation (8.15), we get

$$|\Gamma(\Omega_n, \hat{\Theta}, \Omega)| = |\Gamma(\Omega_1, \hat{\Theta}, \tilde{\Omega})|, \quad \forall n \in [N]. \tag{8.20}$$

where $\tau(\tilde{\Omega}) = \tau(\Omega) + \tau(\Omega_1) - \tau(\Omega_n)$. This would mean under the ψ-domain that if $\hat{\Theta}$ is optimized to cover the angular interval $\mathcal{D}_n$, when excited from an AoA of Ω_n, same configuration when excited from an AoA of Ω_1 will cover the angular interval $\tilde{\mathcal{D}}_n$ that is a shifted version of $\mathcal{D}_n$, by $\tau(\Omega_1) - \tau(\Omega_n)$. Therefore, instead of solving the RIS-enabled MTMR two-way communications

problem under transmission pairs $(\Omega_n, \mathcal{D}_n)$, $n \in [N]$, we propose to solve an RIS-enabled STMR two-way communications problem with the transmission pair $(\Omega_1, \tilde{\mathcal{D}})$, where $\tilde{\mathcal{D}} = \bigcup_{n=1}^{N} \mathcal{D}_n$. Let $\tilde{\boldsymbol{\Theta}}^*$ be the optimal configuration of the RIS-UPA derived for solving the STMR problem. Then a straightforward use of equation (8.15) in the reverse order for each n, will verify that $\boldsymbol{\Theta} = \tilde{\boldsymbol{\Theta}}^*$ is the optimal RIS-UPA configuration for the MTMR case. An example of the above transformation is illustrated for $N = 2$ in Fig. 8.3. Fig 8.3(a) depicts the receive zones $\mathcal{D}_n$, $n = 1, 2$, over the ψ-domain. Each receive zone only consists of one ACI and the particle resolution is set to $(Q_v, Q_h) = (24, 24)$. Using equation (8.15), the receive zone $\mathcal{D}_2$ will get shifted by $\tau(\Omega_1) - \tau(\Omega_2)$ to form $\mathcal{D}_2'$. Fig. 8.3(b) depicts this transition. As the result of this transformation Fig. 8.3(c) plots the corresponding STMR receive zone $\tilde{\mathcal{D}} = \mathcal{D}_1 \cup \mathcal{D}_2'$ that is comprised of two ACIs.

8.4.3 Relationship between RIS-UPA and UPA-antenna beam-forming

RIS-UPA refers to an RIS with UPA structure while UPA-antenna refers to a regular multi-element antenna array where the elements are arranged in the same UPA structure. In this section, we clarify the relation between the beam-forming gain of an RIS-UPA and its UPA-antenna counterpart.

We start by explicitly writing $\boldsymbol{\lambda}$ in equation (8.5) in terms of its elements as follows

$$\boldsymbol{\lambda} = [\lambda_{0,0}, \ldots, \lambda_{0,M_h-1}, \lambda_{1,0}, \ldots, \lambda_{M_v-1,M_h-1}] \tag{8.21}$$

where λ_{m_v,m_h} corresponds to the element located at position (m_v, m_h) in the UPA grid. Using

the expressions (8.6)-(8.9), for an RIS-UPA that is excited from an incident angle Ω_1, we have

$$\lambda_{m_v,m_h} = \beta_{m_v,m_h} e^{-j(\theta_{m_v,m_h} - m_v\xi_1 - m_h\zeta_1)} = \beta_{m_v,m_h} e^{-j(\theta_{m_v,m_h} - \tau(\Omega_1)\mathbf{m})}. \tag{8.22}$$

We note that $\boldsymbol{\lambda}$ depends on the AoA of the incident beams at the RIS, i.e., Ω_1, as well as the RIS parameters $\boldsymbol{\Theta}$. Using equations (8.5) and (8.8), we can restate the reference gain of the RIS in direction $\psi = [\xi, \zeta]$ by explicitly defining $G(\xi, \zeta, \boldsymbol{\lambda})$ as follows

$$G(\xi, \zeta, \boldsymbol{\lambda}) = \left| \left(\mathbf{d}_{M_v}(\xi) \otimes \mathbf{d}_{M_h}(\zeta) \right)^H \boldsymbol{\lambda} \right|^2. \tag{8.23}$$

On the other hand, the gain of UPA-antenna and the feed coefficients $\mathbf{c}$ is given by

$$G(\xi, \zeta, \mathbf{c}) = \left| \left(\mathbf{d}_{M_v}(\xi) \otimes \mathbf{d}_{M_h}(\zeta) \right)^H \mathbf{c} \right|^2 \tag{8.24}$$

that has a clear similarity.

This means that to design the RIS-UPA for the STMR problem with receive zone $\mathcal{D}$ we can use the multi-beamforming design framework to cover the ACIs included in $\mathcal{D}$ for the UPA-antenna [143]. In particular, a RIS-UPA with parameters $\boldsymbol{\lambda}$ and a UPA-antenna with beam-forming parameters $\mathbf{c}$ have the same beam-forming gain pattern if UPA structures are the same, and $\boldsymbol{\lambda} = \mathbf{c}$. Hence, a RIS-UPA which is excited from the solid angle Ω_1 has the same beam-forming gain as its UPA-antenna counterpart if $\boldsymbol{\Theta} = \text{diag}\{\mathbf{c}^T \odot \mathbf{a}_M^H(\Omega_1)\}$. In the following, we address the design of beam-forming coefficients of $\mathbf{c}$ for an antenna with a UPA structure which can then be used to design a RIS-UPA.

For any normalized beam-forming vector $\mathbf{c}$, it is straightforward to show that

$$\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} G\left(\xi,\zeta,\mathbf{c}\right) d\xi d\zeta = (2\pi)^2 \tag{8.25}$$

that can be interpreted as the power conservation law [144]. Note that the dependence between variables ξ and ζ can be resolved using the approximation in [190]. The power conservation law (8.25) allows us to define an optimization problem for the UPA multi-beamforming design in terms of a normalized gain pattern where the total gain is divided by $(2\pi)^2$. In the next section, we define the multi-beamforming design problem as the core of designing our proposed RIS structure.

8.4.4 Multi-beamforming design problem formulation

We wish to design beam-formers that provide high, sharp, and constant gain within the desired ACIs and zero gain everywhere else. We have then for the ideal gain corresponding to such beam-former $\mathbf{c}$ that,

$$\iint_{\mathcal{B}^{\psi}} G_{\mathcal{D}}^{\text{ideal}}\left(\xi,\zeta\right) d\xi d\zeta = \sum_{i=1}^{k}\iint_{\mathcal{D}_i} t d\xi d\zeta = \sum_{(p,q)\in\mathcal{A}}\iint_{\mathcal{B}_{p,q}^{\psi}} t d\xi d\zeta = \sum_{(p,q)\in\mathcal{A}} \delta_{p,q} t = (2\pi)^2 \tag{8.26}$$

where $\delta_{p,q} = \delta_v \delta_h$ denotes the area of the (p,q)-th beam in the (ξ,ζ) domain. Therefore, we can derive $t = \frac{(2\pi)^2}{|\mathcal{A}|\delta_{p,q}}$. It holds that,

$$G_{\mathcal{D}}^{\text{ideal}}\left(\xi,\zeta\right) = \frac{(2\pi)^2}{|\mathcal{A}|\delta_{p,q}} \mathbb{1}_{\mathcal{D}}\left(\xi,\zeta\right) \tag{8.27}$$

Using the beam-former $\mathbf{c}$ we wish to mimic the deal gain in equation (8.27). Therefore, we formulate the following optimization problem,

$$\mathbf{c}_{\mathcal{D}}^{opt} = \arg\min_{\mathbf{c},\|\mathbf{c}\|=1} \iint_{\mathcal{B}^{\psi}} \left| G_{\mathcal{D}}^{\text{ideal}}\ (\xi,\zeta) - G\left(\xi,\zeta,\mathbf{c}\right) \right| d\xi d\zeta \tag{8.28}$$

By partitioning the range of (ξ,ζ) into predefined intervals and then uniformly sampling with the rate (L_v, L_h) per interval along both axes, we can rewrite the optimization problem as follows,

$$\begin{aligned}\mathbf{c}_{\mathcal{D}}^{opt} &= \arg\min\nolimits_{\mathbf{c},\|\mathbf{c}\|=1} \sum_{r=1}^{Q_v}\sum_{s=1}^{Q_h} \iint_{\mathcal{B}_{r,s}^{\psi}} \left| G_{\mathcal{D}}^{\text{ideal}}\ (\xi,\zeta) - G\left(\xi,\zeta,\mathbf{c}\right) \right| d\xi d\zeta \\ &= \lim_{L_h,L_v\to\infty} \sum_{r=1}^{Q_v}\sum_{s=1}^{Q_h}\sum_{l_v=1}^{L_v}\sum_{l_h=1}^{L_h} \frac{\delta_v\delta_h}{L_hL_v} \left| G_{\mathcal{D}}^{\text{ideal}}\ (\xi_{r,l_v},\zeta_{s,l_h}) - G\left(\xi_{r,l_v},\zeta_{s,l_h},\mathbf{c}\right) \right| \end{aligned} \tag{8.29}$$

where,

$$\xi_{r,l_v} = \xi^{r-1} + l_v\frac{\delta_v}{L_v}, \quad \zeta_{s,l_h} = \zeta^{s-1} + l_h\frac{\delta_h}{L_h} \tag{8.30}$$

We can rewrite equation (8.29) as,

$$\mathbf{c}_{\mathcal{D}}^{opt} = \arg\min_{\mathbf{c},\|\mathbf{c}\|=1} \lim_{L_h,L_v\to\infty} \frac{1}{L_hL_v} \left| \mathbf{G}_{\mathcal{D}}^{\text{ideal}} - \mathbf{G}(\mathbf{c}) \right| \tag{8.31}$$

where,

$$\mathbf{G}(\mathbf{c}) = \delta_{p,q}\left[G\left(\xi_{1,1},\zeta_{1,1},\mathbf{c}\right)\cdots G\left(\xi_{Q_v,L_v},\zeta_{Q_h,L_h},\mathbf{c}\right)\right]^T \tag{8.32}$$

and,

$$\mathbf{G}_{\mathcal{D}}^{\text{ideal}} = \delta_{p,q} \left[G_{\mathcal{D}}^{\text{ideal}} \left(\xi_{1,1}, \zeta_{1,1} \right) \cdots G_{\mathcal{D}}^{\text{ideal}} \left(\xi_{Q_v,L_v}, \zeta_{Q_h,L_h} \right) \right]^T \tag{8.33}$$

Unfortunately, the optimization problem in (8.31) does not admit an optimal closed-form solution as is, due to the absolute values of the complex numbers existing in the formulation. However, note that,

$$\mathbf{G}_{\mathcal{D}}^{\text{ideal}} = \sum_{(p,q)\in\mathcal{A}} \delta_{p,q} \frac{(2\pi)^2}{|\mathcal{A}|\delta_{p,q}} \left(\mathbf{e}_{p,q} \otimes \mathbf{1}_{L,1} \right) = \frac{(2\pi)^2}{|\mathcal{A}|} \sum_{(p,q)\in\mathcal{A}} \mathbf{e}_{p,q} \otimes \mathbf{1}_{L,1} \tag{8.34}$$

with $\mathbf{e}_{p,q} \in \mathbb{Z}^Q$ being the standard basis vector for the (p,q)-th axis among (Q_v, Q_h) pairs. Now, note that $\mathbf{1}_{L,1} = \mathbf{g} \odot \mathbf{g}^*$ for any equal gain $\mathbf{g} \in \mathbb{C}^L$ where $L = L_h L_v$. An equal-gain vector $\mathbf{g} \in \mathbb{C}^L$ is a vector where all elements have equal absolute values (in this case, equal to 1). Therefore, we can write:

$$\begin{aligned} \mathbf{G}_{\mathcal{D}}^{\text{ideal}} &= \sum_{(p,q)\in\mathcal{A}} \frac{(2\pi)^2}{|\mathcal{A}|} \left(\mathbf{e}_{p,q} \otimes (\mathbf{g} \odot \mathbf{g}^*) \right) = \frac{(2\pi)^2}{|\mathcal{A}|} \sum_{(p,q)\in\mathcal{A}} \left(\mathbf{e}_{p,q} \otimes \mathbf{g} \right) \odot \left(\mathbf{e}_{p,q} \otimes \mathbf{g} \right)^* \\ &= \left(\sum_{(p,q)\in\mathcal{A}} \frac{2\pi}{\sqrt{|\mathcal{A}|}} \left(\mathbf{e}_{p,q} \otimes \mathbf{g} \right) \right) \odot \left(\sum_{(p,q)\in\mathcal{A}} \frac{2\pi}{\sqrt{|\mathcal{A}|}} \left(\mathbf{e}_{p,q} \otimes \mathbf{g} \right) \right)^* \end{aligned} \tag{8.35}$$

Also, it is straightforward to write,

$$\mathbf{G}(\mathbf{c}) = \left(\mathbf{D}^H \mathbf{c} \right) \odot \left(\mathbf{D}^H \mathbf{c} \right)^* \tag{8.36}$$

where, $\mathbf{D}^H = \sqrt{\delta_v \delta_h}(\mathbf{D}_v^H \otimes \mathbf{D}_h^H)$, and for $a \in \{v, h\}$, and $b \in [Q_a]$ we have,

$$\mathbf{D}_a = [\mathbf{D}_{a,1}, \cdots, \mathbf{D}_{a,Q_a}] \in \mathbb{C}^{M_a \times L_a Q_a} \tag{8.37}$$

where,

$$\mathbf{D}_{v,b} = [\mathbf{d}_{M_v}(\xi_{b,1}), \cdots, \mathbf{d}_{M_v}(\xi_{b,L_v})] \in \mathbb{C}^{M_v \times L_v} \tag{8.38}$$

$$\mathbf{D}_{h,b} = [\mathbf{d}_{M_h}(\zeta_{b,1}), \cdots, \mathbf{d}_{M_h}(\zeta_{b,L_h})] \in \mathbb{C}^{M_h \times L_h} \tag{8.39}$$

Comparing the expressions (8.31), (8.35), and (8.36), one can show that the optimal choice of $\mathbf{c}_{\mathcal{D}}$ in (8.28) is the solution to the following optimization problem for proper choices of $\mathbf{g}_{p,q}$.

Problem 23 *Given equal-gain vectors $\mathbf{g}_{p,q} \in \mathbb{C}^L$, for $(p,q) \in \mathcal{A}$ find vector $\mathbf{c}_{\mathcal{D}} \in \mathbb{C}^M$ such that*

$$\mathbf{c}_{\mathcal{D}} = \underset{\mathbf{c}, \|c\|=1}{\arg\min} \lim_{L \to \infty} \left\| \sum_{(p,q) \in \mathcal{A}} \frac{2\pi}{\sqrt{|\mathcal{A}|}} (\mathbf{e}_{p,q} \otimes \mathbf{g}_{p,q}) - \mathbf{D}^H \mathbf{c} \right\|^2 \tag{8.40}$$

However, we now need to find the optimal choices of $\mathbf{g}_{p,q}$ that minimize the objective in (8.31). Using (8.35) and (8.36), we have the following optimization problem.

Problem 24 *Find equal-gain vectors $\mathbf{g}_{p,q}^* \in \mathbb{C}^L$, $(p,q) \in \mathcal{A}$ such that*

$$< \mathbf{g}_{p,q}^* >_{(p,q)\in\mathcal{A}} = \underset{<\mathbf{g}_{p,q}>_{(p,q)\in\mathcal{A}}}{\arg\min} \left\| abs(\mathbf{D}^H \mathbf{c}_{\mathcal{D}}) - \frac{2\pi}{\sqrt{|\mathcal{A}|}} abs\Big(\sum_{(p,q)\in\mathcal{A}} \mathbf{e}_{p,q} \otimes \mathbf{g}_{p,q} \Big) \right\|^2 \tag{8.41}$$

where $abs(.)$ denotes the element-wise absolute value of a vector.

In the next section, we continue with the solution of problems 23, and 24.

8.5 Proposed Multi-beamforming Design Solution

Note that the solution to Problem 23 is the limit of the sequence of solutions to a least-square optimization problem as L goes to infinity. For each L we find that,

$$\mathbf{c}_{\mathcal{D}}^{(L)} = \sum_{(p,q)\in\mathcal{A}} \frac{2\pi}{\sqrt{|\mathcal{A}|}} (\mathbf{D}\mathbf{D}^H)^{-1}\mathbf{D}\left(\mathbf{e}_{p,q}\otimes\mathbf{g}_{p,q}\right) \tag{8.42}$$

$$\mathbf{c}_{\mathcal{D}}^{(L)} = \sum_{(p,q)\in\mathcal{A}} \sigma(\mathbf{D}_{v,p}\otimes\mathbf{D}_{h,q})\mathbf{g}_{p,q} \tag{8.43}$$

where $\sigma = \frac{2\pi\sqrt{\delta_v\delta_h}}{LQ\delta_v\delta_h\sqrt{|\mathcal{A}|}} = \frac{2\pi}{LQ\sqrt{\delta_v\delta_h|\mathcal{A}|}}$, noting that it holds that,

$$\mathbf{D}\mathbf{D}^H = \delta_v\delta_h(\mathbf{D}_v\otimes\mathbf{D}_h)(\mathbf{D}_v^H\otimes\mathbf{D}_h^H) = \delta_v\delta_h LQ \doteq \kappa \tag{8.44}$$

Even though Problem 23 admits a nice analytical closed-form solution, doing so for the Problem 24 is not a trivial task, especially because the objective function is not convex. However, the convexification of the objective problem in the form of

$$\begin{aligned} <\mathbf{g}_{p,q}^*>_{(p,q)\in\mathcal{A}} &= \underset{<\mathbf{g}_{p,q}>_{(p,q)\in\mathcal{A}}}{\arg\min} \left\| \mathbf{D}^H\mathbf{c}_{\mathcal{D}} - \frac{2\pi}{\sqrt{|\mathcal{A}|}} \sum_{(p,q)\in\mathcal{A}} \mathbf{e}_{p,q}\otimes\mathbf{g}_{p,q} \right\|^2 \\ &= \underset{<\mathbf{g}_{p,q}>_{(p,q)\in\mathcal{A}}}{\arg\min} \left\| \left(\kappa\mathbf{D}^H\mathbf{D} - \mathbf{I}_{LQ}\right) \sum_{(p,q)\in\mathcal{A}} \mathbf{e}_{p,q}\otimes\mathbf{g}_{p,q} \right\|^2 \end{aligned} \tag{8.45}$$

leads to an effective solution for the original problem. Indeed, it can be verified by solving the optimization problem (8.45) numerically that the solution admits the form (8.46) in the following

conjecture.

Conjecture 25 *The minimizer of* (8.45) *is in the form of*

$$\mathbf{g}_{p,q}^{*}=\begin{bmatrix} 1 & \alpha_v\alpha_h & \cdots & \alpha_v^{(L_v-1)}\alpha_h^{(L_h-1)} \end{bmatrix}^T, (p,q)\in\mathcal{A} \tag{8.46}$$

for some η_v, η_h *where* $\alpha_a = e^{j(\frac{\eta_a}{L_a})}$, $a\in\{v,h\}$.

Except for some special cases, we have not been able to analytically prove this conjecture in its entirety. In the following, we use the analytical form (8.46) for $\mathbf{g}_{p,q}^{*}$ for the rest of our derivations. This solution would not be the optimal solution for the original problem (8.41). However, it provides a near-optimal solution with the added benefits of allowing us to (i) find the limit of the solution as L goes to infinity, and (ii) express the beam-forming vectors in closed form, as it will be revealed in the following discussion. An analytical closed-form solution for $\mathbf{c}_{\mathcal{D}}$ can be found as follows. It holds that,

$$\begin{aligned}\mathbf{c}_{\mathcal{D}}{}^{(L)} &= \sum_{(p,q)\in\mathcal{A}}\left(\sum_{(l_v,l_h)=(1,1)}^{(L_v,L_h)}\sigma g_{p,q,l_v,l_h}\mathbf{d}_{M_t}\left(\xi_{p,l_v},\zeta_{q,l_h}\right)\right)\\ &= \sum_{(p,q)\in\mathcal{A}}\left(\sum_{(l_v,l_h)=(1,1)}^{(L_v,L_h)}\sigma g_{q,p,l_v,l_h}\left[1,\cdots,e^{j\mu_{p,q,l_v,l_h}^{M_v-1,M_h-1}}\right]^T\right)\end{aligned} \tag{8.47}$$

where $\mu_{p,q,l_v,l_h}^{m_v,m_h}=(m_v\xi_{p,l_v}+m_h\zeta_{q,l_h})$. We can then write for the $(m_v,m_h)^{th}$ component of the beam-former $\mathbf{c}_{\mathcal{D}}$,

$$c_{p,q,m_v,m_h}=\lim_{L_h,L_v\to\infty}\frac{1}{L_hL_v}\sum_{(p,q)\in\mathcal{A}}\sum_{(l_h,l_v)=(1,1)}^{(L_h,L_v)}g_{p,q,l_v,l_h}e^{j\mu_{p,q,l_v,l_h}^{M_v-1,M_h-1}} \tag{8.48}$$

Using equation (8.30), we can rewrite (8.48) as,

$$c_{p,q,m_v,m_h} = \frac{2\pi}{Q} e^{j\chi_{p-1,q-1}^{m_v,m_h}} \left(\frac{1}{L_v} \lim_{L_v \to \infty} \sum_{l_v=1}^{L_v} e^{j\frac{\eta_v + m_v \delta_v}{L_v} l_v} \right) \left(\frac{1}{L_h} \lim_{L_h \to \infty} \sum_{l_h=1}^{L_h} e^{j\frac{\eta_h + m_h \delta_h}{L_h} l_h} \right) \tag{8.49}$$

to get,

$$\begin{aligned} c_{\mathcal{D},m_v,m_h} &= \sum_{(p,q)\in\mathcal{A}} \frac{2\pi}{Q} e^{j\chi_{p-1,q-1}^{m_v,m_h}} \int_0^1 e^{j\xi_v x} dx \int_0^1 e^{j\xi_h x} dx \\ &= \sum_{(p,q)\in\mathcal{A}} \frac{2\pi}{Q} e^{j(\zeta_{p-1,q-1}^{m_v,m_h} + \frac{\xi_v+\xi_h}{2})} sinc(\frac{\xi_v}{2\pi}) sinc(\frac{\xi_h}{2\pi}) \end{aligned} \tag{8.50}$$

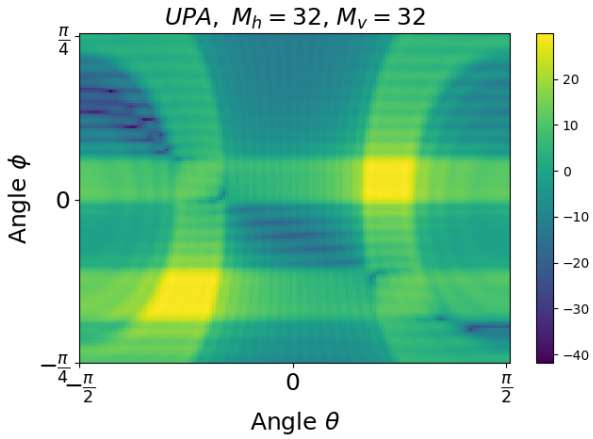
with $\chi_{p,q}^{m_v,m_h} = (m_v \xi^p + m_h \zeta^q)$, and $\xi_a = \delta_a m_a + \eta_a$, for $a \in \{v, h\}$. Now that the closed-form expression for $\mathbf{c}_{\mathcal{D}}$, and therefore, $\boldsymbol{\lambda}$ is known, for a RIS that is excited from the solid angle Ω_1, the RIS parameters at the antenna placed at location (m_v, m_h) can be easily computed. More precisely, we get,

$$\beta_{m_v,m_h} = |\mathbf{c}_{\mathcal{D},m_v,m_h}| \tag{8.51}$$

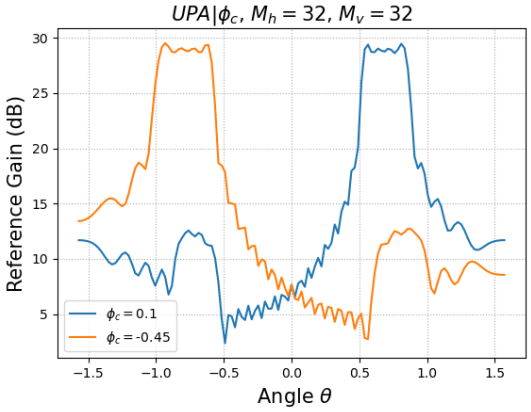
$$\theta_{m_v,m_h} = \underline{/\mathbf{c}_{\mathcal{D},m_v,m_h}} + m_v \xi_1 + m_h \zeta_1 \tag{8.52}$$

The solution given by (8.51) and (8.52) is optimized to find a beam-forming pattern that is the closest to the desired normalized beam pattern. However, in practice, the norm of $\mathbf{c}$ depends on the power P that can be inserted by amplifiers (active elements), say $\|\mathbf{c}\| \leq P$. Hence for an RIS-UPA with active elements the gains β_{m_v,m_h} are scaled by $\sqrt{P}$. Also, the solution given by (8.51) and (8.52) may be further tailored for the case that the RIS elements are passive.

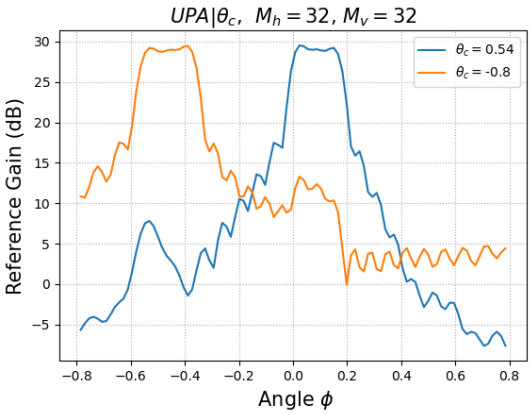
If the gain control for the passive RIS elements is still possible, the absolute value of the



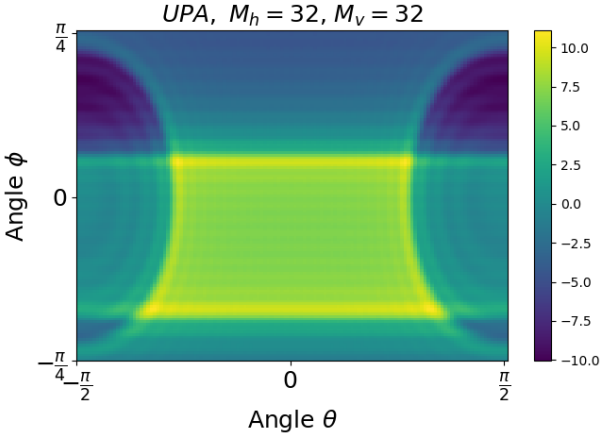
(a) Dual-beam 3D UPA pattern



(b) UPA pattern cut at ϕ_c



(c) UPA pattern cut at θ_c



(d) Singular-beam 3D UPA pattern

Figure 8.4: RIS-UPA beam patterns for multi-beamforming settings

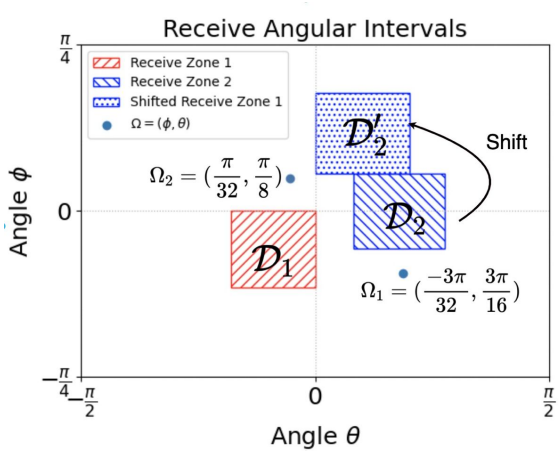
gain for each element may not exceed 1 (a value less than 1 corresponds to an attenuation). To maximize the power reflected by the RIS, we scale the gains β_{m_v,m_h} so that their maximum is equal to one, i.e., $\beta_{m_v,m_h} = |\mathbf{c}_{\mathcal{D},m_v,m_h}|/\|\mathbf{c}\|_\infty$. Finally, in the case that gaining control (attenuation) at the RIS with passive elements is not feasible, we have $\beta_{m_v,m_h} = 1$. In the next section, we evaluate the effectiveness of our RIS beam-forming design approach through numerical experiments.

8.6 Performance Evaluation

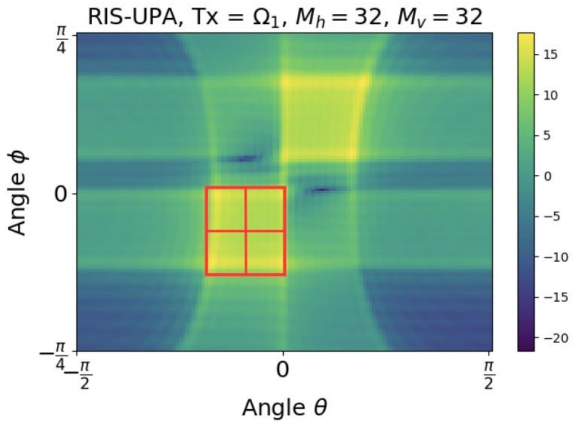
In this section, we evaluate the performance of our multibeam design framework.

8.6.1 Multibeam design

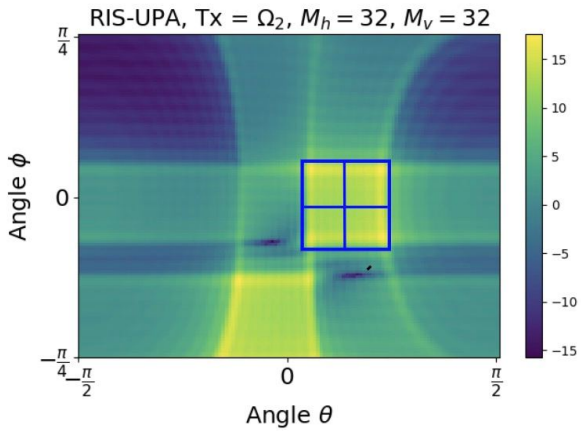
First, we consider a dual-beam design problem which comprises two lobes with centers at directions $(-8\pi/32, -5\pi/32)$ and $(7\pi/32, \pi/32)$ for the pairs of the solid angle (ϕ, θ) with the beamwidth equal to $\pi/16$. We divide both the ψ_h, and the ψ_v range uniformly into $Q_h = 16$, and $Q_v = 16$ regions resulting in $Q = 256$ equally-shaped units in (ψ_v, ψ_h) domain. We cover each desired beam with the smallest number of the designed units to provide uniform gain at the desired angular regions. Figures 8.4(a)-(c) depict the beam pattern of the dual beam obtained through our design where all angles are measured in radians. Fig. 8.4(a), shows the heat map corresponding to the gain of the reflected beam from the RIS for the designed dual-beam. The gains are computed in dB. It can be seen that the designed beam-former generates two disjoint beams with an almost uniform gain over the desired ACIs. It is also observed that the beams sharply drop outside the desired ACIs and effectively suppress the gain everywhere



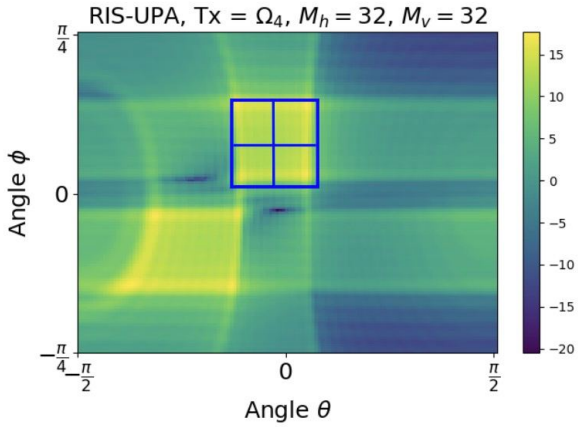
(a) Receive angular intervals



(b) RIS pattern at incident angle Ω_1

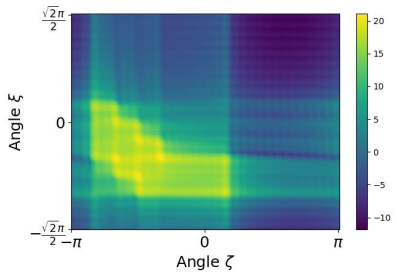


(c) RIS pattern at incident angle Ω_2

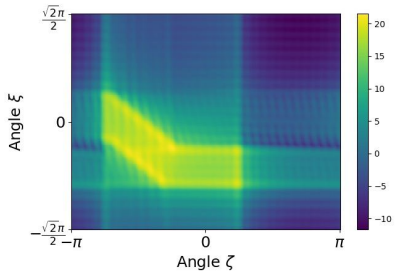


(d) RIS pattern at incident angle Ω_4

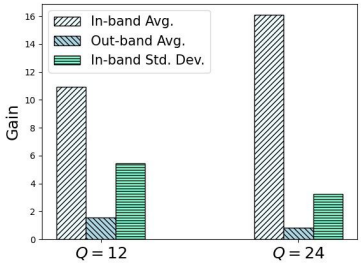
Figure 8.5: RIS-UPA beam patterns for MTMR settings



(a) $(Q_h, Q_v) = (12, 12)$



(b) $(Q_h, Q_v) = (24, 24)$



(c) Beam pattern statistics

Figure 8.6: Effect of resolution on the beam quality

outside the ACI. In order to quantify the suppression and the leakage out of the desired ACI we depict the cross-section of the gain pattern at a fixed elevation angle ϕ_c for two values of $\phi_c \in \{-8\pi/32, 7\pi/32\}$ located inside the two lobes of the designed dual beam in Fig. 8.4(b). Similarly, Fig. 8.4(c) shows the cross-section of the beam pattern at a fixed azimuth angle θ_c for two values of $\theta_c \in \{-5\pi/32, \pi/32)\}$. Both Fig. 8.4(b) and Fig. 8.4(c) confirm the sharpness of both lobes of the designed dual-beam and can be used to find the beamwidth of each lobes at an arbitrary fraction from its maximum values, e.g,, the 3dB beamwidth or 10dB beamwidth. Indeed, there is a negligible difference between 3dB and 10dB beamwidth which clarifies the sharpness of the beams. From Fig. 8.4(b) and Fig. 8.4(c), it is also observed that the gain within the ACI is almost uniform. Nonetheless, we should emphasize the fact that the shape of the lobes of the beam that are centered at different solid angles may suffer from slight deformation as seen by Fig. 8.4(a). This phenomenon worsens as the corresponding lobes of the beams get too close to the plane of the RIS.

8.6.2 Comparison of multibeam and single beam

In order to compare the performance of our multi-beam design to a single-beam design, we consider a beam with a single lobe that is capable of covering the same two regions as in the dual-beam design. Fig. 8.4(d), shows the heat map corresponding to the gain of the reflected beam from an RIS for the corresponding single beam that is optimized based on our design. As was the case for multi-beam, this figure also shows that for a single beam, our design generates an almost uniform and fairly sharp beam. However, comparing Fig. 8.4(a) and Fig. 8.4(d), we observe that in the desired ACI the multi-beamforming procedure enhances the gain by about 20

dB over the beams with an optimized single lobe. The reason for this difference is that according to the power conservation axiom, with a constant input power level, the narrower each beam is the higher the reference gain over the area it covers. In fact, if we target two disjoint ACIs with a single beam, the beam must be wide enough to cover both areas. This will result in dissipating the power in angles that are not intended. However, the muti-beam design allows for generating two disjoint beams, each being narrow enough to only cover the intended ACI. Therefore, the power will not be wasted in undesired directions.

8.6.3 Two-way multi-link communications

We consider two links. The first link is between the transmitter located as $\Omega_1 = (\frac{-3\pi}{32}, \frac{3\pi}{16})$ and a receiver that is in the ACI $\mathcal{D}_1 = [-0.36, 0] \times [-0.56, 0]$ and the second link is between the transmitter located as $\Omega_2 = (\frac{\pi}{32}, \frac{\pi}{8})$ and a receiver that is in the ACI $\mathcal{D}_1 = [-0.18, 0.18] \times [0.26, 0.86]$. Each of the receive zones $\mathcal{D}_1$ and $\mathcal{D}_2$ is comprised of 4 particles in an $Q_v \times Q_h = 8 \times 8$ grid. Fig. 8.5(a) depicts the angular location of the two transmitters and the ACI for the receivers. In order to design a beam-former that covers $\mathcal{D}_1$ when transmitting from angular position Ω_1 and covers $\mathcal{D}_2$ when transmitting from angular position Ω_2, we first find a composite beam with a unified incident angle, say Ω_1. This means that we find the ACI $\mathcal{D}_2'$ when the RIS array is excited by a beam at incident angle Ω_1 which is equivalent to the ACI $\mathcal{D}_2$ where the same array is excited from the incident angle Ω_2. The transformed ACI $\mathcal{D}_2'$ is depicted in Fig 8.5(a). The heat map of the designed beams is depicted in figures 8.5(b)-(d) where the RIS is excited from angles Ω_1, Ω_2, and Ω_4, respectively. Figures 8.5(b)-(c) illustrate that when the RIS is excited from an incident angle Ω_1 and Ω_2, the corresponding ACI $\mathcal{D}_1$ and $\mathcal{D}_2$ are respectively covered by

the designed beam as illustrated in the respective figures, however, in either case, another angular region is also covered by the beam which is not necessary and could be considered as possible wastage of the power. We note that this happens due to the fact that the unwanted ACI when the RIS is excited from the incident angle Ω_1 is indeed generating the desired ACI when the RIS is excited from the incident angle Ω_2. Finally, we consider Ω_4 to be an angular point in the ACI $\mathcal{D}_2$, e.g., we take Ω_4 to be the center of the ACI $\mathcal{D}_2$. Fig. 8.5(d) shows that if the RIS is excited from an incident angle Ω_4, e.g., when a user in ACI $\mathcal{D}_2$ is replying, then it will be received by the corresponding transmitter which is located at angular position Ω_2. In figures 8.5(b)-(d) the windows show the positioning of the grids that lie on the particles covering the corresponding desired ACIs.

8.6.4 Beams with arbitrary shape (footprint)

Here, we illustrate the possibility of designing a beam with an arbitrary pattern, or more precisely, an arbitrary footprint. We aim to design a beam that covers the receiving zone $\mathcal{D}$ that is specified as follows.

$$\mathcal{D} = \begin{cases} -\frac{\sqrt{2}}{2}(\zeta + \frac{11}{6}) < \xi < -\frac{\sqrt{2}}{2}(\zeta + \frac{3}{2}) & \text{if } \frac{-5\pi}{6} \leq \zeta \leq \frac{-\pi}{3} \\ \frac{-7\sqrt{2}\pi}{24} < \xi < \frac{-3\sqrt{2}\pi}{24} & \text{if } \frac{-\pi}{3} < \zeta \leq \frac{\pi}{3} \end{cases} \tag{8.53}$$

We consider two possible quantization levels of the beam footprint with the resolution $Q_h = Q_v = 12$ and $Q_h = Q_v = 24$. Fig. 8.6(a) and Fig. 8.6(b) illustrate the heat map of the designed beam. The finer the resolution, the better the beam's shape approximation. Fig. 8.6(c) provides the quantitative comparison between these two resolutions; it shows the higher resolution increases the overall beam-forming gain, lowers the leakage, and generates

smoother gain.

8.7 Conclusions

An RIS can be incorporated into mmWave communications to fill the coverage gaps in the blind spots of the mmWave system. We proposed a novel approach for designing RISs, namely RIS-UPA, where the RIS elements are arranged according to a UPA structure. We proposed a configuration for the elements of a RIS-UPA that enables the coverage of multiple disjoint angular intervals simultaneously. On this ground, we showed that an RIS-UPA-assisted MIMO system can support multiple two-way communication pairs simultaneously. We established that the RIS-aided multiple-pair scenario can be transformed into a single-pair scenario and then by appealing to the similarities of RIS-UPA and UPA beam-forming we argued that we can borrow the principles of UPA multi-beamforming design to obtain closed-form low-complexity solutions for the RIS design problem. Both our theoretical results and numerical experiments demonstrate that our RIS configuration can form beams of custom footprints and will result in sharp, high, and stable gains within the desired ACIs regardless of their spatial locations, while effectively suppressing all the undesired out-of-band components.

Chapter 9: Blind Cyclic Prefix-based Carrier Frequency Offset Estimation in MIMO-OFDM Systems

9.1 Overview

Orthogonal frequency-division Multiplexing (OFDM) has been widely adopted in wireless communications standards as a strong multi-carrier modulation technique due to its spectral efficiency and resistance against frequency selective fading. OFDM is considered an enabling technology for the fifth generation (5G) and beyond communications systems, complementing multi-input multi-output (MIMO). However, OFDM is vulnerable against carrier frequency offset (CFO) which is among the major impairments [191] between radio-frequency (RF) transceivers that are caused by the frequency mismatch between the local oscillators at the RF transceivers or by the Doppler shift. Such an impairment destroys the orthogonality among the subcarriers and results in severe performance degradation in multi-carrier systems. Therefore, it is essential to design CFO estimation and compensation techniques to avoid the decline in bit error rate (BER) at the receiver. To be more precise, CFO usually consists of two components when normalized to subcarrier spacing; (i) the *integral* part that results in a circular shift in the indices of the subcarriers, resulting in frequency ambiguity, and (ii) the *fractional* part that impacts the orthogonality of the subcarriers provoking inter-carrier interference (ICI). The integral CFO

estimation has been studied in the literature [192] [193] [194] [195]. However, most of the efforts in the literature including the present work, have been focused on fractional CFO estimation and compensation. See [196] for a comprehensive survey on CFO estimation and compensation techniques and an illustrative example of how CFO is specified in wireless communications standards.

CFO estimation approaches in the prior art can be mostly classified into two categories; (i) data-oriented, and (ii) non-data-oriented. The approaches of the first type, either employ time domain training symbol sequences [197] [198] [199] or rely on extensive use of frequency domain pilots [200] [201] [202]. In order to obtain high accuracy, such approaches will have to use a large number of training sequences in the time domain or occupy a large bandwidth in the frequency domain, introducing an extra overhead and inevitably causing the performance degradation of MIMO OFDM systems. For this reason, the non-data-oriented (a. k. a. blind) CFO estimation techniques, have gained increasing attention over the past decade.

A group of blind CFO estimators, utilize the null subcarriers (NS) in the OFDM blocks [203] [204]. The NS-based blind estimators exploit the fact that the ICI resulting from the CFO will show up at the null subcarriers and can be used to estimate and correct the CFO parameters. NS-based methods usually formulate the CFO estimation as a polynomial optimization problem over multiple blocks of OFDM symbols and then employ ESPIRIT-like or MUSIC-like search algorithms, or polynomial rooting methods to solve the optimization problem. Given the complexity of the search methods and the need for multiple OFDM blocks for accurate estimation, the NS-based methods are of higher complexity. In [204], the authors introduce the novel concept of gap subcarriers (GS) and exploit this concept to approximate the CFO estimation objective with a cosine function. Then the parameters of the cosine function are determined uniquely by observing the value of the cost function at three different trial CFO values. Then the CFO parameter can

be easily estimated. This method skips the large overhead of the search methods and complex traditional polynomial rooting techniques.

Another category of blind CFO estimators exploits the structure of the cyclic prefix (CP) to design low-complexity CFO estimators [205] [206]. the performance of the CP-based techniques may decline when the channel becomes more frequency-selective. In [205], a CFO estimation technique relying on the remodulation of the received signals at the receiver end is proposed in multipath environments. Specifically, the theoretical mean-squared error (MSE) for CFO estimation is presented as a closed-form solution. The authors carry out the Cramer-Rao Band (CRB) on the MSE of CFO estimation for multipath channels and based on these theoretical analyses propose a fine CFO estimation technique that is of low complexity.

In this chapter, we propose a low-complexity blind CFO estimation approach for a MIMO system in the uplink direction, where each mobile user (MU) may be served by one or multiple access points (APs) [207]. The main contributions of the paper are as follows.

- We propose to use antenna diversity for CFO estimation. Given that the RF chains for all antenna elements at a communication node share the same clock, the CFO between two points may be estimated by using the combination of the received signal at all antennas.
- Our proposed scheme also combines antenna diversity with time diversity by considering the CP for multiple OFDM symbols.
- We provide a low-complexity closed-form expression and derive the Cramer-Rao lower bound for CFO estimation.
- We provide an algorithm that can considerably improve the CFO estimation performance at the expense of a linear increase in computational complexity.

- We derive the Cramer-Rao lower bound for the CFO estimation which is based on the observations of the received time domain signals of the CP at the beginning and at the end of each OFDM symbol at all available antennas.

9.2 System Model

We consider an OFDM system, where the AP and MU are employing MIMO for which the antenna elements at each node operate according to a common local oscillator. Let N denote the size of the Discrete Fourier Transform (DFT) L denote the size of the CP and $\tilde{N} = N + L$ is the total symbol length including the CP. The k-th time domain vector of the transmitted signal is given by

$$\mathbf{x}_k = [x[\tilde{N}k], x[\tilde{N}k+1], \ldots, x[\tilde{N}(k+1)-1)]]^T \tag{9.1}$$

where $x[i] = x[i+N], \forall i, 1 \leq i \mod \tilde{N} \leq L$. The time series signal can be interpreted as blocks of length $\tilde{N}$ including the CP. We assume that the number of the channel taps does not exceed L and hence the channel taps for the m-th antenna in the time domain may be represented by

$$h_0^{(m)}, h_1^{(m)}, \ldots, h_{L-1}^{(m)}.$$

Without considering the effect of CFO and noise, the received signal at antenna m at time i can be written as

$$r^{(m)}[i] = \sum_{l=0}^{L-1} x[i-l]h_l^{(m)} \tag{9.2}$$

which depends on the current and the past $L-1$ time domain transmitted signals due to the effect of the multipath channel. Let $\theta_k = N\Delta f/f_s$ denote the normalized CFO for the k-th OFDM symbol with respect to the first received time domain signal in the time frame k where the CFO is Δf in Hz and f_s is the sampling frequency. We only consider fractional CFO, i.e., $-0.5 \leq \theta_k \leq 0.5$. We note that the normalized CFO in a multi-user scenario where each OFDM symbol may be transmitted from a different user may be different for different received time domain vectors. Nonetheless, if two adjacent symbols k and $k+1$ belong to the same user $\theta_k = \theta_{k+1}$, but the first received time domain signal $x[(k+1)\tilde{N}]$ for the received vector $k+1$ is rotated by $\mathbf{e}^{j2\pi\frac{k\tilde{N}}{N}\theta_k}$ with respect to the first received time domain signal $x[k\tilde{N}]$ of the received vector k. Let $\psi[i]$ denote the CFO for the i-th received signal which can be written as $\psi[k\tilde{N}+i] = e^{j2\pi\theta_k i/N}$ with respect to the first symbol of the $\mathbf{x}_k$. By considering the effect of CFO and noise, the received signal at antenna m at time i may be written as

$$r^{(m)}[i] = \sum_{l=0}^{L-1} \psi[i-l]x[i-l]h_l^{(m)} + z^{(m)}[i] \tag{9.3}$$

where $z[i]$ is the AWGN noise for the received signal at time i with the variance of σ_z^2.

9.3 Proposed CFO Estimation Technique

Consider the k-th OFDM symbol received at the m-th antenna element. Let ξ be a variable that will later be useful for CFO estimation and define the vector $\mathbf{y}_k^{(m)}(\xi)$ of length $2L$ with ℓ-th

element given by,

$$y_k^{(m)}[\ell](\xi) = \left(r^{(m)}[k\tilde{N} + l] - e^{j2\pi\xi} r^{(m)}[k\tilde{N} + N + l]\right) \tag{9.4}$$

The auto-correlation function of the vector $\mathbf{y}_k^{(m)}(\xi)$ in the expanded form is given by,

$$J_{k,m}(\xi) = \mathbb{E}\left\{\sum_{l=0}^{2L-1} y_k^{(m)}[\ell](\xi) y_k^{(m)}[\ell](\xi)^*\right\} \tag{9.5}$$

where $\rho = e^{j2\pi\xi}$, and $(.)^*$ is the Hermitian operator. We note that due to the definition of CP and the fact that the transmitted signal is random with zero mean, for $0 \leq i, j \leq N + L - 1$, we have

$$\mathbb{E}\{x[k\tilde{N} + i]x[k\tilde{N} + j]\} = \sigma_x^2 \delta(|i - j| \mod N) \tag{9.6}$$

where $\delta(.)$ is the Kronecker delta function and σ_x^2 is the power of the time domain signal per sample. Using equations (9.3) and (9.6), after straightforward algebraic operations followed by the expansion of (9.5), $J_{k,m}(\xi)$ is simplified as

$$J_{k,m}(\xi) = 2L\eta^{(m)}(1 - cos(2\pi(\xi - \theta_k)) + 2L\sigma_z^2 \tag{9.7}$$

where $\eta^{(m)} = \sigma_x^2 \sum_{l=0}^{L-1} |h_l^{(m)}|^2$ is independent of ξ. The simplified form of the auto-correlation function in (9.7) reveals that the function achieves its minimum with respect to ξ at $\xi = \theta_k$. Therefore, it makes sense to minimize equation (9.5) as a cost function to obtain ξ as the estimate for the CFO parameter θ_k. If the receiver is equipped with M antennas, the cost function can be

redefined as the summation over all antennas

$$J_k^M(\xi) = \sum_{m=1}^{M} J_{k,m}(\xi) = 2L \left(\sum_{m=1}^{M} \eta^{(m)} \right) (1 - cos(2\pi(\xi - \theta_k)) + 2LM\sigma_z^2. \tag{9.8}$$

Clearly, the cost function $J_k^M(\xi)$ for multiple antennae also has its minimum at $\xi = \theta_k$.

9.3.1 Coarse Estimation

For large values of M, the cost (9.8) is empirically given by,

$$J_k^M(\xi) \approx \sum_{m=1}^{M} \sum_{l=0}^{2L-1} \left(y_k^{(m)}[\ell](\xi) \right) \left(y_k^{(m)}[\ell](\xi) \right)^*. \tag{9.9}$$

The minimizer of (9.9) is found by setting its derivative with respect to ξ to zero which would uniquely provide the CFO estimate as,

$$\xi = \frac{1}{2\pi} \angle \sum_{m=1}^{M} \sum_{l=0}^{2L-1} r^{(m)}[k\tilde{N} + l] r^{(m)}[k\tilde{N} + N + l]^* \tag{9.10}$$

If multiple OFDM symbols are received from a single user (without loss of generality, denoted by symbols k, $k = 1, \ldots, K$), the cost function can be better approximated as

$$J^{K,M}(\xi) \approx \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{l=0}^{2L-1} \left(y_k^{(m)}[\ell](\xi) \right) \left(y_k^{(m)}[\ell](\xi) \right)^* \tag{9.11}$$

and the CFO estimate would be found as,

$$\xi = \frac{1}{2\pi} \angle \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{l=0}^{2L-1} r^{(m)}[k\tilde{N} + l] r^{(m)}[k\tilde{N} + N + l]^* \tag{9.12}$$

The above solution for $M = 1$ antenna is similar to the solution found in [205] for coarse CFO estimation using K OFDM symbols. The solution given by (9.12) relies on the independence between the time domain transmitted symbols as well as the independence between the channel coefficients. The larger the ensemble over which the summation is calculated, the better the approximation. Hence, increasing the number of OFDM symbols K or the number of antennas M would increase the estimation performance.

9.3.2 Fine Estimation

We note that for each k, $1 \leq k \leq K$ and m, $1 \leq m \leq M$ the objective function is comprised of $2L$ terms which contribute to the objective function of the form

$$J_{k,m}(\xi) = A - Bcos(2\pi(\xi - \theta_k) \tag{9.13}$$

We note that for such a convex objective function, the larger the second derivative the lower the estimation error of ξ. The objective function in (9.11) can be written as

$$J^{K,M}(\xi) = \sum_{l=0}^{2L-1} \mathbb{E}\left\{\left(y_k^{(m)}[\ell](\xi)\right)\ \left(y_k^{(m)}[\ell](\xi)\right)^*\right\} = \sum_{l=0}^{2L-1} R(l) \tag{9.14}$$

and therefore, the second derivative of (9.11) is given by the summation of the second derivatives of $R(l)$, i.e.,

$$\frac{\partial^2}{\partial \xi^2} J^{K,M}(\xi) = \sum_{l=0}^{2L-1} \frac{\partial^2}{\partial \xi^2} R(l) \tag{9.15}$$

Note that, $R(l)$ can be stated in the summation form as

$$R(l) = \frac{1}{M} \sum_{m=1}^{M} R^{(m)}(l) \tag{9.16}$$

where,

$$R^{(m)}(l) = 2(\eta^{(m)} - \eta_l^{(m)}\, cos(2\pi(\xi - \theta_k))) - 2\sigma_z^2 \tag{9.17}$$

with,

$$\begin{cases} \eta_l^{(m)} = \sigma_x^2 \sum_{i=0}^{l} |h_i^{(m)}|^2 & 0 \leq i \leq L-1 \\ \eta_l^{(m)} = \sigma_x^2 \sum_{i=l-l}^{L-1} |h_i^{(m)}|^2 & L \leq i \leq 2L-1 \end{cases}$$

Hence, $R(l)$ is obtained as

$$R(l) = 2\sum_{m=1}^{M} \eta^{(m)} - 2\sum_{m=1}^{M} \eta_l^{(m)} cos(2\pi(\xi - \theta_k)) + 2\sigma_z^2 \tag{9.18}$$

and its second derivative with respect to ξ is given by

$$\frac{\partial^2}{\partial \xi^2} R^{(m)}(l) = 8\pi^2 cos(2\pi(\xi - \theta_k))) \sum_{m=1}^{M} \eta_l^{(m)} \tag{9.19}$$

Comparing (9.17) and (9.19), it is noted that the smaller the value of $R(l)$ the larger its derivative. Hence, one can improve the estimation performance by using a summation over a proper subset of indices, i.e., $\{0, 1, \ldots, 2L-1\}$, that contribute to the objective function (9.14). Let us denote this subset by $S(\Lambda)$. The parameter Λ is the size of the set $S(\Lambda)$ which is important to be chosen such that the estimation error is

minimized. For a given index subset $S(\Lambda)$, the objective function is modified as

$$J^{K,M}(\xi) = \sum_{l \in S(\Lambda)} R(l) = \sum_{l \in S(\Lambda)} \mathbb{E}\left\{ \left(y_k^{(m)}[\ell](\xi) \right) \left(y_k^{(m)}[\ell](\xi) \right)^* \right\} \tag{9.20}$$

where $R(l)$ can be empirically found as

$$R(l) \approx \sum_{k=1}^{K} \sum_{m=1}^{M} \left(y_k^{(m)}[\ell](\xi) \right) \left(y_k^{(m)}[\ell](\xi) \right)^* \tag{9.21}$$

The fine estimate is then given by

$$\xi = \frac{1}{2\pi} \angle \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{l \in S(\Lambda)} r^{(m)}[k\tilde{N} + l] r^{(m)}[k\tilde{N} + N + l]^* \tag{9.22}$$

Algorithm 10 provides the steps required to perform a fine CFO estimation for a pre-selected value of Λ. This algorithm may be run for a fixed value of Λ. For example, $\Lambda = L$ provides a good estimation as will be shown later in the evaluation section. However, the algorithm can be further improved by finding the value of Λ adaptively. The detailed of adaptive-fine CFO estimation is given in Algorithm 11. This algorithm provides significant improvement with only two iterations. Please refer to the results in the evaluation section.

Algorithm 10 *Fixed-Fine-Estimate*

Input: Λ, $1 \leq \Lambda \leq 2L - 1$

1: Compute the coarse estimate of θ_k using (9.12).
2: Calculate $R(l)$, $0 \leq l \leq 2L - 1$ using the approximation in (9.21).
3: find the set $S(\Lambda)$ which contains the indices of Λ smallest values of $R(l)$.
4: $\hat{\theta} \leftarrow$ Calculate the fine estimate via equation (9.22).
5: **Return** $\hat{\theta}$.

Algorithm 11 *Adaptive-Fine-Estimate*

Input: $ITER$

$iter = 0$
$\hat{\theta} \leftarrow$ the coarse estimate of θ_k using (9.12).
repeat
  Calculate $R(l)$, $0 \leq l \leq 2L-1$ using the approximation in (9.21)
  For $\Lambda = 1:1:2L-1$:
    Compute $S(\Lambda)$ which contains the indices of Λ smallest values of $R(l)$.
    $\phi(\Lambda) \leftarrow$ Calculate the fine estimate via (9.22).
  End For
  $\Lambda^{\text{OPT}} \leftarrow \arg\min_{1 \leq \Lambda \leq 2L-1} |\hat{\theta} - \phi(\Lambda)|$
  $\hat{\theta} \leftarrow \phi(\Lambda^{\text{OPT}})$
  $iter++$
until $iter = ITER$
Return $\hat{\theta}$

9.3.3 Derivation of the Cramer-Rao Bound

In this section, we derive the Cramer-Rao lower bound on the MSE estimate of any non-biased CFO estimator which is based on the observation of $r^{(m)}(i)$ for all m, $1 \leq m \leq M$ and available time index i. First, we note that the channel taps are samples of zero mean circularly symmetric complex Gaussian distributions. Hence, over the ensemble of the channel realizations, $r^{(m)}[i]$ will also follow zero-mean circularly symmetric complex Gaussian distributions. In order to find the joint distribution of $r^{(m)}[i]$, for all available i and m, it suffices to find the covariance matrix. Let X be defined as a vector containing all $r^{(m)}[i]$, for K OFDM symbols, M antennas and $2L$ values of i, i.e., $i = 1, 2, \ldots, L, N+1, N+2, \ldots, N+L$. We have

$$\mathbf{x} = [r^{(1)}[1], \ldots, r^{(1)}[K\tilde{N}], r^{(2)}[1], \ldots, r^{(M)}[K\tilde{N}]]^T \tag{9.23}$$

The entries of the covariance matrix $\mathbb{E}\{XX^*\}$ is given as follows. If $m_1 \neq m_2$ and $i_1 \neq i_2$, $j \neq 0$ and $j \neq N$, we have

$$\mathbb{E}\{r^{(m)}[k\tilde{N}+i](r^{(m)}[k\tilde{N}+i])^*\} = \eta^{(m)} + \sigma_z^2, \tag{9.24}$$

$$\mathbb{E}\{r^{(m)}[k\tilde{N}+i](r^{(m)}[k\tilde{N}+N+i])^*\} = \eta_i^{(m)} e^{j2\pi\theta} \tag{9.25}$$

$$\mathbb{E}\{r^{(m)}[k\tilde{N}+i](r^{(m)}[k\tilde{N}+i+j])^*\} = 0, \tag{9.26}$$

$$\mathbb{E}\{r^{(m_1)}[k\tilde{N}+i](r^{(m_2)}[k\tilde{N}+i])^*\} = 0 \tag{9.27}$$

$$\mathbb{E}\{r^{(m_1)}[k\tilde{N}+i_1](r^{(m_2)}[k\tilde{N}+i_2])^*\} = 0 \tag{9.28}$$

Let us define the vector $\mathbf{v}_{k,i}^{(m)} = [r^{(m)}[k\tilde{N}+i], r^{(m)}[k\tilde{N}+N+i]]^T$, and let $\mathbf{R}_i^{(m)}$ be the covariance matrix of $\mathbf{v}_{k,i}^{(m)}$. The distribution of $\mathbf{x}$ is given by

$$f_X(\mathbf{x},\theta) = \prod_{m,1\leq m\leq M} \left(\prod_{i,0\leq i\leq 2L} \frac{1}{\pi^2 \left|\mathbf{R}_i^{(m)}\right|} \exp\left(-\mathbf{v}_{k,i}^{(m)*} \mathbf{R}_i^{(m)} \mathbf{v}_{k,i}^{(m)}\right) \right)^K \tag{9.29}$$

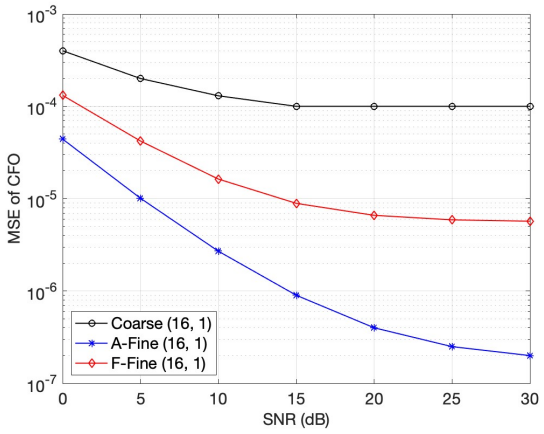
The Cramer-Rao bound can be found by calculating the Fisher information content of the log-likelihood function defined as

$$I(\theta) = -\mathbb{E}\left\{ \frac{\partial^2 \log(f_X(\mathbf{x},\theta))}{\partial\theta^2} \right\} \tag{9.30}$$

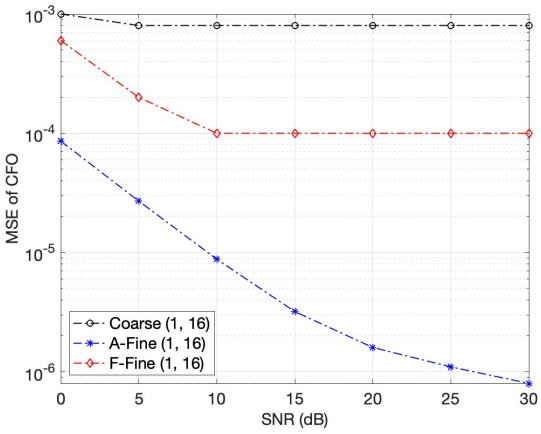
$$= -\mathbb{E}\left\{ \sum_{m,1\leq m\leq M} \sum_{i,0\leq i\leq 2L} \frac{4K\pi^2}{\left|\mathbf{R}_i^{(m)}\right|} \mathbf{v}_{k,i}^{(m)*} \mathbf{Q}_i^{(m)} \mathbf{v}_{k,i}^{(m)} \right\} \tag{9.31}$$

where

$$\mathbf{Q} = \begin{bmatrix} 0 & -\eta_i^{(m)} e^{j2\pi\theta} \\ -\eta_i^{(m)} e^{j2\pi\theta} & 0 \end{bmatrix} \tag{9.32}$$



(a) CFO estimation with $(K, M) = (16, 1)$



(b) CFO Estimation with $(K, M) = (1, 16)$

Figure 9.1: Coarse vs. fine CFO estimation with M antenna elements and K OFDM symbols

Hence, we have

$$I(\theta) = 8K\pi^2 \sum_{m,1\leq m\leq M} \sum_{i,0\leq i\leq 2L} \frac{\eta_i^{(m)^2}}{(\eta^{(m)} + \sigma_z^2)^2 - \eta_i^{(m)^2}} \tag{9.33}$$

The Cramer-Rao bound is then given by the inverse of the fisher information $I(\theta)$.

9.4 Performance Evaluation

In this section, we evaluate the performance of our CFO estimation technique. We first describe the simulation parameters, setup, and evaluation metrics and then proceed with analyzing the numerical results.

9.4.1 Simulation Setup and Parameters

Without loss of generality, we assume the MUs have only a single antenna embedded, while each AP may employ multiple antenna elements. We consider the transmission of symbols from a 16-QAM constellation. Throughout the experiments, we set the DFT size $N = 64$, CP length $L = 16$, and the

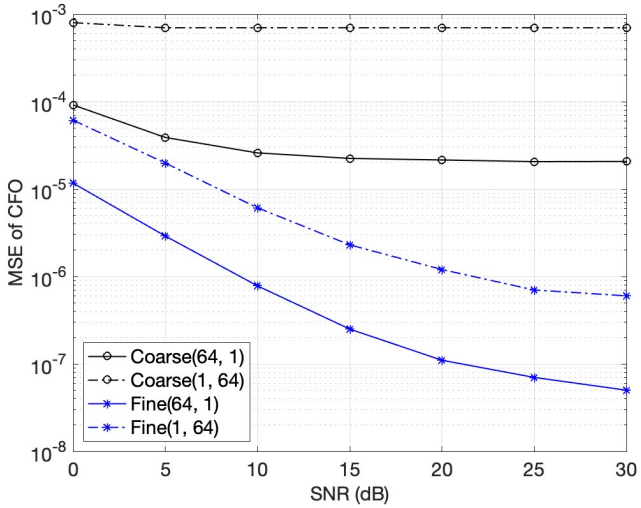


Figure 9.2: Time vs. Antenna Diversity $(K, M) = (64, 1), (1, 64)$

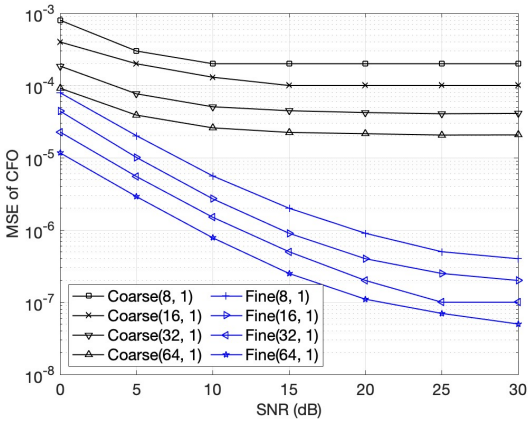


Figure 9.3: Impact of K, for $M = 1$

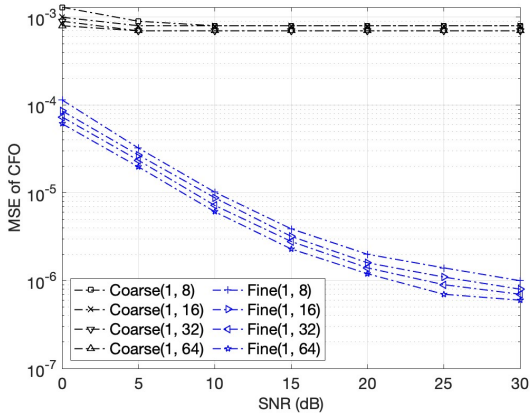
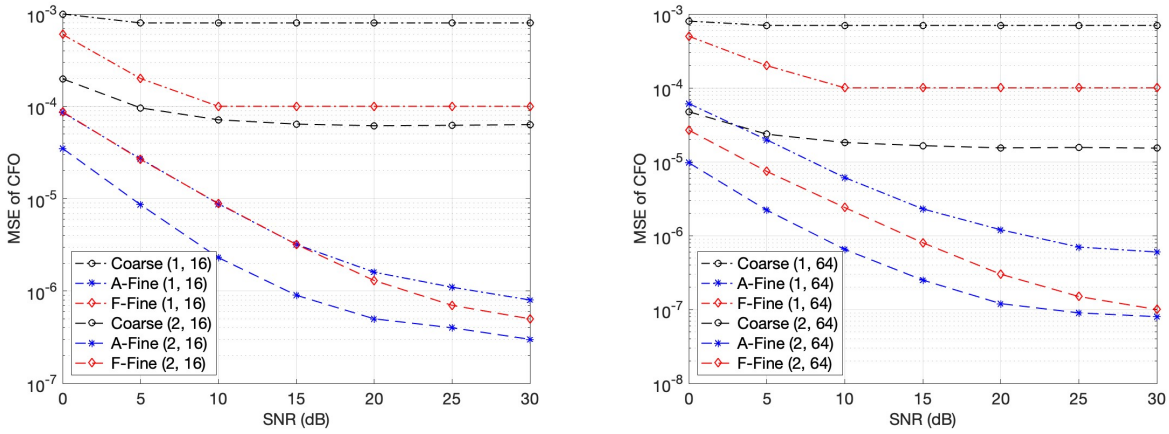


Figure 9.4: Impact of M, for $K = 1$

number of antenna elements per AP $M = 1, 8, 16, 32, 64$. We adopt a multipath Rayleigh fading channel with $T = 5$ taps with adaptive white Gaussian noise (AWGN), and the CFO value is set to 0.295. Our tests are carried out by MATLAB on a server with an Intel i9 CPU at 2.3 GHz and 16 GB of main memory and each simulation is run for $I = 10000$ trials and the results are averaged. We use the following metrics to evaluate the performance of our CFO estimation method:

- **Estimation Mode** may be *coarse* or *fine*. Fine estimation can be implemented in two *adaptive* and *fixed* modes based on the static or dynamic choice of the parameter Λ.



(a) Comparison of coarse and fine CFO estimation (b) Comparison of coarse and fine CFO estimation.

Figure 9.5: Evaluation of combining time and antenna diversity for various K, and M

- **Time Diversity** corresponds to when CFO estimation is carried out across time; i. e. over multiple OFDM symbols.

- **Antenna Diversity** corresponds to the case where CFO estimation is carried out across multiple antenna elements employed on a single AP.

9.4.2 Numerical Results

Fig. 9.1 depicts the MSE of CFO estimation under different modes, i.e. e. coarse, adaptive fine (a-fine), and fixed fine (f-fine). Fig. 9.1a corresponds to the case of time diversity with $K = 16$, i.e. where a single-antenna MU communicates 16 OFDM symbols to a single-antenna AP. While, Fig. 9.1b corresponds to the case of antenna diversity with $M = 16$, i.e. where a single-antenna MU communicates a single OFDM symbol to 16 antenna elements of the AP. The value of $\Lambda = 16$ is set for the fixed fine CFO estimation. It is observed that the performance under adaptive fine estimation is superior to the coarse estimation by at least two orders of magnitude. Also, both the coarse estimation and fixed-fine estimation techniques saturate at some MSE value while the adaptive fine estimation method continues its descending trend even after $30dB$ SNR.

Fig. 9.2 compares the performance of CFO estimation under time diversity and antenna diversity, for $K = M = 64$. It is observed that the MSE plot for CFO estimation under antenna diversity will lie and saturate above the MSE plot under time diversity, for coarse and fine estimation, respectively. The reason for this observation is that, although the channel tap coefficients for the antenna elements at a single AP are uncorrelated, they still follow the same power profile. Therefore, approximating the expectation with a summation needs larger values of M in antenna diversity than K in time diversity.

Fig. 9.3 and 9.4 illustrate the effect of increasing the value of parameters K and M on the performance of CFO estimation, under time diversity and antenna diversity, respectively. It is observed that increasing the value of parameters K and M shifts the CFO MSE plots downwards and improves the performance of the CFO estimation. Further, it is observed that increasing K from 8 to 64 under time diversity improves the performance of CFO estimation by an order of magnitude while increasing M has a less improving effect under the antenna diversity scheme.

Fig. 9.5 shows how combining antenna diversity and time diversity improves the performance of CFO estimation. Fig. 9.5a, and 9.5b show this effect for $M = 16$ and $M = 64$, respectively. Interestingly, we observe that increasing the OFDM symbols results in more improvement in the case of $M = 64$ compared to the case of $M = 16$.

9.5 Conclusions

This paper presented a low-complexity CP-based blind CFO estimation for MIMO-OFDM systems. We used antenna diversity for CFO estimation. Given that the RF chains for all antenna elements at a communication node share the same clock, the carrier frequency offset (CFO) between two points may be estimated by using the combination of the received signal at all antennas. Further, we incorporated the notion of time diversity into the CFO estimation by considering the CP for multiple consecutive OFDM blocks. We defined a cost function employing the correlation of the received OFDM signals at multiple

antennae and multiple OFDM blocks. We proposed algorithms for estimating the CFO with low complexity. Numerical results verify the validity of our proposed CFO estimation technique.

Chapter A: Proof of Theorem (21)

Appendix I

We will start by a few supplementary definitions that are required for the proof of Theorem 20.

Definition 26 *A direction is defined on the azimuth angles, say, counter-clockwise. The ACI as well as the starting and the ending point of a beam is measured with respect to this direction. For example, one may have a beam starting at $3\pi/2$ and ending at $\pi/2$ which is represented as $[3\pi/2, \pi/2)$.*

Definition 27 *Consider a set of points $a_1, a_2, \ldots, a_n$, $n \geq 3$ on a circle. For any three indices $1 \leq i, j, k \leq n$, we write $a_i \prec a_j \prec a_k$ (and read as a_i precedes a_j precedes a_k) if by starting from point a_i and moving in the defined direction (say counter clockwise) we first encounter a_j and then encounter a_k before reaching a_i again. Equivalently, we write $a_k \succ a_j \succ a_i$ and read a_k succeeds a_j succeeds a_i. Please note that the precedence relation requires at least three points to be specified.*

Definition 28 *Consider b beams and denote them by beam i, $i \in B$ where $B = 1, 2, \ldots, b$ denotes the set of all possible beam. For a given subset $C \subset B$, we define a component beam ω_C as a beam (possibly comprised of an interval or a collection of intervals) where each differential arc of the beam contains all the beams in C and does not contain any beam which is not in C. The set C is called the* spawning set *of ω_C and its cardinality is called the* degree *of ω_C. A component beam with spawning set C may or may not exist, and if it exist we may refer to the component beam by its spawning set C. We say a* component beam has a beam i in its intersection *if the beam i belongs to its spawning set.*

Definition 29 *We define the following operators and notations on spawning set. Let a spawning set (or equivalently a component beam) be denoted by capital letters C, D, E, F, X, Y, etc., and a set of spawning sets be represented by $\mathcal{C}, \mathcal{E}$, etc., e.g., $\mathcal{C} = \{C, D\}$, and $\mathcal{E} = \{E, F\}$. The union of two spawning set $C \cup D$ is denoted by CD. The cross product (or product or union product) of two sets of spawning sets $\mathcal{C}$ and $\mathcal{E}$ is denoted by $\mathcal{C} \times \mathcal{E}$ and is defined as a set that contains the union of any pair of sets such that one set belongs to the first and the other one belongs to the second set of spawning sets, i.e., $\mathcal{C} \times \mathcal{E} = \{X, X = X_1X_2, X_1 \in \mathcal{C}, X_2 \in \mathcal{D}\}$. For example $\{C, D\} \times \{E, F\} = \{CE, CF, DE, DF\}$. The multiplication of the set of spawning sets $\mathcal{C}$ by itself may be represented as $\mathcal{C}^2 = \mathcal{C} \times \mathcal{C}$. The higher power $k > 2$ may be defined recursively as $\mathcal{C}^k = \mathcal{C}^{k-1} \times \mathcal{C}$.*

If a set of spawning sets $\mathcal{D}$ contains a single spawning set C it may be represented by C instead of $\mathcal{D}$. Similarly, if a spawning set contains a single beam c, i.e., $C = \{c\}$ it may be directly represented by c instead of C. The definition of the operators $\setminus$ and $\times$ is inherited for the spawning sets and beams as well. For example, we may write $C \times \mathcal{D}$ where in this expression C is interpreted as a set of spawning set which contains a single spawning set C. We denote the set minus operation by $-$, i.e., $C - D = \{x, x \in C, x \notin D\}$. We assume that the operator $\setminus$ has precedence on the operators $\times$, $\cup$, $\cap$, and $-$, e.g., $C \times \mathcal{D}E = C \times (\mathcal{D} \setminus E)$

Lemma 30 *Among the set of contiguous scanning beams, a set of scanning beams with Tulip design generates the maximal number of possible feedback sequences for the channel with $p = 1$ path.*

Proof. The proof relies on the fact that b contiguous beams have total of $2b$ starting and ending points and hence can generate at most $2b$ component beams. On the other hand for $p = 1$ a feedback sequence is valid only and only if the set of the indices of the beam with positive feedback corresponds to an spawning set of a component beam that is realized by the set of scanning beams. Since there are at most $2b$ component beam, there is only $2b$ possible feedback sequences and Tulip design already provides

$2b$ component beams. We note that the Tulip structure is not the only optimal design for $p = 1$.

Lemma 31 *For a channel with $p = 2$ paths, a set of scanning beams with Tulip design generates $1, 3, 7$, and 15 feedback sequences for $b = 1, 2, 3$, and 4 beams, respectively, and generates $2b(b-2)$ feedback sequences for $b > 4$ beams.*

Proof. The proof simply follows from a direct counting argument. There are b component beams of degree 1 with spawning sets $\{1\}, \{2\}, \ldots, \{b\}$ that are directly part of the Tulip design. For $b > 1$, there are $\binom{b}{2}$ feedback sequences of degree 2 corresponding to the beam index sets $\{12\}, \{13\}, \ldots, \{1b\}, \{23\}, \{24\}, \ldots, \{2b\}, \ldots, \{(b-1)b\}$. A degree 2 feedback sequence corresponding to an index set $\{ij\}$ can be generated by selecting path 1 and path 2 inside the component beams $\{i\}$ and $\{j\}$, respectively. For $b > 2$, a feedback sequence of degree 3 may be generated by selecting a path in a component beam of degree 2 and a path in a component beam of degree 1. In the following consider the indices of the beam in mod b. In other words, the beam $b + 1$ is the same as beam 1. There are a total of $b(b-3)$ feedback sequences of type $i(i+1)k$. Consider the set $C_i = \{i(i+1)k, k \neq i, k \neq (i+1)\}$. There are b sets $C_i, i = 1, 2, \ldots, b$, and each set C_i has $b-2$ distinct feedback sequences and there are exactly b feedback sequences of the form $i(i+1)(i+2)$ which appear in all sets $C_i, i = 1, 2, \ldots, b$ exactly twice. Hence the union $\cup_{i \in B} C_i$ contains exactly $b(b-3)$ distinct feedback sequences. Similarly, for $b > 4$, it can be argued that there are exactly $b(b-3)/2$ degree 4 feedback sequences of type $i(i+1)j(j+1)$, $i+1 \succ j$, $j+1 \succ i$ that is generated by having one path in the component beam $i(i+1)$ and another in the component beam $j(j+1)$. For $b = 4$, there is only one feedback sequence 1234 that is generated by having one path in component beam 12 and one path in component beam 34. Summing all the feedback sequences of degree 1 to degree 4 proves the statement of the lemma.

Proof of Theorem 20:

Lemma 30 proves the optimality of Tulip design for a channel with a single path, i.e., $p = 1$. Lemma 20, extends this optimality to the case of $p = 2$, however, the proof is considerably more involved.

Here, we first provide the sketch of the proof and then formally get into the details. Lemma 31, counts the number of feedback sequences that is generated by Tulip design. It is relatively easy to check for small numbers of $b \leq 5$ that no other set of scanning beams can generate more feedback sequences as Tulip design. For the case of $b > 5$, we prove the theorem by using the following steps.

Consider a set of b contiguous scanning beams $\mathcal{B}$ positioned on a circle. Let $\mathcal{C}$ denote the set of component beams generated by the set of scanning beams $\mathcal{B}$.

First, we note that without loss of generality we can assume that set of starting and ending point of each scanning beam interval, i.e., *end point markers*, are disjoint, which means that every starting point and ending point is unique. Suppose not; consider the ordered set of end point markers $x_1 \succ x_2 \succ \ldots \succ x_{2b}$ and assume that two consecutive points x_i and x_j (say x_1 and x_2 or x_{2b} and x_1) are equal $x_i = x_j$ and we have $x_{i-1} \succ x_i \succ x_{j+1}$. Consider two cases: (i) At least one of the points x_i and x_j are an ending point. Let x_j be an ending point. By moving it to point x'_j such that $x_i \succ x'_j \succ x_{j+1}$, one can generate an extra component beams. (ii) Both x_i and x_j are an starting point. By moving x_i to point x'_i such that $x_{i-1} \succ x'_i \succ x_j$, one can generate an extra component beams. In either case (i) or case (ii), by adding a new component beam to the set $\mathcal{C}$, the cardinality of $\mathcal{C}^2$ cannot decrease.

Hence, we can assume that in the maximal case of $\mathcal{C}$, its cardinality is exactly $2b$. In other words, there exist a $\mathcal{C}$ which is maximal, i.e., $|\mathcal{C}^2|$ is maximum when $|\mathcal{C}| = 2b$. It can be easily verified that when considering $2b$ the component beams on the loop, the spawning sets of any two consecutive component beam C_i and C_{i+1} differ only in one element which means that (i) either $C_i \subset C_{i+1}$ or $C_i \supset C_{i+1}$ and (ii) $||C_i| - |C_{i+1}|| = 1$.

Claim 1: Consider a subset $\mathcal{A} = \{A_1, \ldots, A_n\}$ of n elements of the $\mathcal{C}$ that are adjacent in the loop. If we add a new spawning set A to one end of the loop (without loss of generality assume to the right-hand side of the loop), and $|A| < |A_n|$, then $|(\mathcal{A} \cup A)^2| \leq |\mathcal{A}^2| + n - 2$. One of the following four cases might have occurred, and in each case, all three unions $A \cup A_n$, $A \cup A_{n-1}$, and $A \cup A_{n-2}$ are repeated in some other combinations and hence they do not generate a new spawning set in $(\mathcal{A} \cup A)^2$. We have (i) $A_{n-2} \supset A_{n-1} \supset A_n \supset A$ in which case $A \cup A_n = A_n$, $A \cup A_{n-1} = A_{n-1}$, and $A \cup A_{n-2} = A_{n-2}$; (ii) $A_{n-2} \subset A_{n-1} \supset A_n \supset A$ in which case $A \cup A_n = A_n$, $A \cup A_{n-1} = A_{n-1}$, and $A \cup A_{n-2} = A_{n-2}$ or A_{n-1}; (iii) $A_{n-2} \subset A_{n-1} \subset A_n \supset A$ in which case $A \cup A_n = A_n$, $A \cup A_{n-1} = A_n$, and $A \cup A_{n-2} = A$ or A_{n-2}; (iv) $A_{n-2} \supset A_{n-1} \subset A_n \supset A$ in which case $A \cup A_n = A_n$, $A \cup A_{n-1} = A_n$, and $A \cup A_{n-2} = A_n \cup A_{n-2}$. Hence, at most $n - 3$ new subsets of the form $A_i \cup A, 1 \leq i \leq n - 3$ and one new subset corresponding to the new subset A are generated.

Claim 2: Consider a subset $\mathcal{A} = \{A_1, \ldots, A_n\}$ of n elements of the $\mathcal{C}$ that are adjacent in the loop. If we add a new spawning set A to one end of the loop (without loss of generality assume to the right-hand side of the loop), and $|A| > |A_n|$, then $|(\mathcal{A} \cup A)^2| \leq |\mathcal{A}^2| + n - 1$ or $|(\mathcal{A} \cup A)^2| \leq 2n$. One of the following three cases might have occurred. We have (i) $A_{n-1} \subset A_n \subset A$ in which case $A \cup A_n = A$, $A \cup A_{n-1} = A$; (ii) $A_{n-2} \subset A_{n-1} \supset A_n \subset A$ in which case $A \cup A_n = A$ and $A \cup A_{n-2} = A \cup A_{n-1}$; (iii) $\exists i, i \geq 1$ such that $A_i \subset A_{i+1} \supset \ldots \supset A_{n-2} \supset A_{n-1} \supset A_n \subset A$ in which case $A \cup A_i = A \cup A_{i+1}$ and $A \cup A_n = A$; (iv) $A_1 \supset A_2 \supset \ldots \supset A_{n-2} \supset A_{n-1} \supset A_n \subset A$. Hence, for the first three cases at most $n - 2$ new subsets of the form $A_i \cup A$ out of all subsets of $A_i \cup A, 1 \leq i \leq n$ plus one new subset A are generated. Otherwise, in case (iv) the possible different subsets are only in the form of $A_1, A_2, \ldots, A_n, A \cup A_1, A \cup A_2, \ldots, A \cup A_n$ which are 2n in total.

Claim 3: Let F_n is the maximum number of elements in the $|(\mathcal{A})^2|$ where $|A| = n$. It can be directly checked for the small number of $n \leq 5$ that $F_1 = F_2 = F_3 = 1$, $F_4 = 5$, and $F_5 = 8$. Based on the results

of claims 1 and 2, we can prove that $F_{n+2} \leq F_n + 2n - 2$, hence we have

$$\left\{ \begin{array}{ll} F_n \leq \frac{n(n-4)}{2} + 5, & n = 2b, n \geq 4 \\ F_n \leq \frac{(n+1)(n-5)}{2} + 8, & n = 2b-1, n \geq 5 \end{array} \right\} \tag{A.1}$$

Claim 4: Let $\mathcal{A}_i = \{A \in \mathcal{C}, i \in A\}$ and $n_i = |\mathcal{A}_i|$. Obviously the set of component beams in $\mathcal{A}_i = \{A_1, A_2, \ldots, A_{n_i}\}$ are sequentially in order on the loop. Let n be the maximum value of n_i and correspondingly $\mathcal{A} = \mathcal{A}_i$ and $\mathcal{D} = \mathcal{C} - \mathcal{A}_i$. Assume the elements of $\mathcal{A} = \{A_1, \ldots, A_n\}$ and $\mathcal{D} = \{D_1, \ldots, D_m\}$ are in order on the loop such that D_1 and D_m are the neighbors of A_n, A_1, respectively. We have $n + m = 2b$. Since $D_m \subset A_1$ and for the i which maximizes n_i we have $i \in A_1$ but $i \notin D_m$, hence $D_m \subset A_1$. Similarly we have and $D_1 \subset A_{n_i}$, which means that for all $k = 1, 2, \ldots, n_i$, we have $D_m A_i = A_1 A_i$ and $D_1 A_i = A_{n_i} A_i$. This means that there are $2n_i$ elements in $\mathcal{A} \times \mathcal{D}$ which is in common with elements in $\mathcal{A}^2$.

If there is $n_i \geq 6$ then we have $|\mathcal{A} \times \mathcal{D}| \leq n_i(2b - n_i) - 2n_i$ and for even n_i we have

$$|\mathcal{C}^2| \leq |\mathcal{A}^2| + |\mathcal{D}^2| + |\mathcal{A} \times \mathcal{D}| \tag{A.2}$$

$$\leq (\frac{(n_i)(n_i - 4)}{2} + 5) + (\frac{(2b - n_i)(2b - n_i - 4)}{2} + 5)$$

$$+ (n_i(2b - n_i) - 2n_i) < 2b(b - 2), \tag{A.3}$$

and for odd n_i we have

$$|\mathcal{C}^2| \leq (\frac{(n_i + 1)(n_i - 5)}{2} + 8) +$$

$$(\frac{(2b - n_i + 1)(2b - n_i - 5)}{2} + 8)$$

$$+ (n_i(2b - n_i) - 2n_i) < 2b(b - 2). \tag{A.4}$$

Hence, the maximal design cannot have $n_i \geq 6$ since for Tulip structure $|\mathcal{C}^2| = 2b(b-2)$.

Claim 5: Next, assume that $n_i = 5$. We note that D_1 cannot be a subset D_m, otherwise there is a beam which is in at least 6 consecutive component beams $A_n, D_1, D_2, \ldots, D_m$. Hence, we can find a minimum index $1 < k \leq m$ such that $D_1 \not\subset D_k$ for which we have $A_n D_k = A_n D_{k-1}$. Similarly, we have a minimum index l such that $D_m \not\subset D_l$ for which we have $A_1 D_l = A_1 D_{l-1}$. On the other hand, there is a maximum index $1 \leq j < n$ such that $D_1 \not\subset A_j$ otherwise there is a common beam in at least 6 consecutive component beams $A_1, A_2, \ldots, A_n, D_1$. Hence we have $D_1 A_j = D_1 A_{j-1}$. Similarly there a minimal index t such that $D_m \not\subset A_t$ otherwise there is a common beam in at least 6 consecutive component beams $D_m, A_1, A_2, \ldots, A_n$. Hence we have $D_m A_t = D_m A_{t-1}$. This means we have $|\mathcal{A} \times \mathcal{D}| \leq 5(2b-5) - 2 \times 5 - 4$ and

$$|\mathcal{C}^2| \leq (8) + (\frac{(2b-4)(2b-10)}{2} + 8) + (5(2b-5) - 14) < 2b(b-2). \tag{A.5}$$

Hence, the maximal design cannot have $n_i \geq 5$ since for Tulip structure $|\mathcal{C}^2| = 2b(b-2)$.

Claim 6: Next, assume that $n_i = 4$. This case is very similar to the Claim 5 for $n_i = 5$ which the exception that we use the formula of F_n for the even n. We note that D_1 cannot be a subset D_m, otherwise there is a beam which is in at least 5 consecutive component beams $A_n, D_1, D_2, \ldots, D_m$. Hence, we can find a minimum index $1 < k \leq m$ such that $D_1 \not\subset D_k$ for which we have $A_n D_k = A_n D_{k-1}$. Similarly, we have a minimum index l such that $D_m \not\subset D_l$ for which we have $A_1 D_l = A_1 D_{l-1}$. On the other hand, there is a maximum index $1 \leq j < n$ such that $D_1 \not\subset A_j$ otherwise there is a common beam in the at least 5 consecutive component beams $A_1, A_2, \ldots, A_n, D_1$. Hence we have $D_1 A_j = D_1 A_{j-1}$. Similarly there a minimal index t such that $D_m \not\subset A_t$ otherwise there is a common beam in at least 5 consecutive component beams $D_m, A_1, A_2, \ldots, A_n$. Hence we have $D_m A_t = D_m A_{t-1}$. This means we

have $|\mathcal{A} \times \mathcal{D}| \leq 4(2b-4) - 2 \times 4 - 4$ and

$$|\mathcal{C}^2| \leq (5) + (\frac{(2b-4)(2b-8)}{2} + 5)$$
$$+ (4(2b-4) - 12) < 2b(b-2). \tag{A.6}$$

Hence, the maximal design cannot have $n_i \geq 4$ since for Tulip structure $|\mathcal{C}^2| = 2b(b-2)$.

On the one hand, the max n_i is less than or equal to 3. On the other hand for all C_i in $\mathcal{C}$, $|C_i| \geq 1$ and for every two consecutive C_i and C_{i+1} we have $|C_i| = |C_{i+1}| \geq 3$. Hence,

$$\sum_{i=1}^{b} n_i = \sum_{i=1}^{2b} |C_i| >= 3b$$

which means that all n_i have to be equal to 3. This means that the Tulip structure is maximal and indeed the unique optimal solution.

Bibliography

[1] B. Han and et al., "Network function virtualization: Challenges and opportunities for innovations," *IEEE Communications Magazine*, vol. 53, no. 2, pp. 90–97, 2015.

[2] K. Katsalis and et al., "Network slices toward 5g communications: Slicing the lte network," *IEEE Communications Magazine*, vol. 55, no. 8, pp. 146–154, August 2017.

[3] Etsi nfv architectural overview. [Online]. Available: http://www.etsi.org/

[4] A. Gholami and J. S. Baras, "Collaborative cloud-edge-local computation offloading for multi-component applications," in *2021 IEEE/ACM Symposium on Edge Computing (SEC)*, 2021, pp. 361–365.

[5] A. Gholami, K. Rao, W.-P. Hsiung, O. Po, M. Sankaradas, and S. Chakradhar, "Roma: Resource orchestration for microservices-based 5g applications," in *NOMS 2022-2022 IEEE/IFIP Network Operations and Management Symposium*, 2022, pp. 1–9.

[6] A. Gholami, N. Torkzaban, and J. S. Baras, "Mobile network slicing under demand uncertainty: A stochastic programming approach," 2023.

[7] X. Li, "Network slicing with elastic sfc," Huawei Technologies Canada Inc., Tech. Rep., 2016.

[8] K. Rao, W.-P. Hsiung, O. Po, M. Sankaradas, S. Chakradhar, and A. Gholami, "Resource orchestration for microservices-based 5g applications," Feb. 2 2023, uS Patent App. 17/863,685.

[9] A. Gholami, K. Rao, W.-P. Hsiung, O. Po, M. Sankaradas, J. S. Baras, and S. Chakradhar, "Application-specific, dynamic reservation of 5g compute and network resources by using reinforcement learning," in *Proceedings of the ACM SIGCOMM Workshop on Network-Application Integration*, ser. NAI '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 19–25. [Online]. Available: https://doi.org/10.1145/3538401.3546598

[10] J. Ordonez-Lucena, "Network slicing for 5g with sdn/nfv: Concepts, architectures, and challenges," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 80–87, 2017.

[11] N. Nikaein, E. Schiller, R. Favraud, K. Katsalis, D. Stavropoulos, I. Alyafawi, Z. Zhao, T. Braun, and T. Korakis, "Network store: Exploring slicing in future 5g networks," in *Proceedings of the 10th International Workshop on Mobility in the Evolving Internet Architecture (MobiArch '15)*. New York, NY, USA: ACM, 2015, pp. 8–13.

[12] J. G. Herrera and J. F. Botero, "Resource allocation in nfv: A comprehensive survey," *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 518–532, September 2016.

[13] D. Dietrich and et al., "Network function placement on virtualized cellular cores," in *IEEE COMSNETS*, Bangalore, 2017, pp. 259–266.

[14] N. Torkzaban, C. Papagianni, and J. S. Baras, "Trust-aware service chain embedding," in *2019 Sixth International Conference on Software Defined Systems (SDS)*, Rome, Italy, 2019, pp. 242–247.

[15] N. Torkzaban and J. S. Baras, "Trust-aware service function chain embedding: A path-based approach," ser. 2020 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN), 2020, pp. 31–36.

[16] E. Paraskevas, T. Jiang, and J. Baras, "Trust-aware network utility optimization in multihop wireless networks with delay constraints," in *IEEE Control and Automation Conf.*, 2016, pp. 593–598.

[17] M. Huang and et al., "Throughput maximization of delay-sensitive request admissions via virtualized network function placements and migrations," in *IEEE ICC*, Kansas, MO, 2018, pp. 1–7.

[18] Y. Ma and et al., "Profit maximization for admitting requests with network function services in distributed clouds," *IEEE Transactions on Parallel and Distributed Systems*, vol. 30, no. 5, pp. 1143–1157, 1 May 2019.

[19] B. Yang and et al., "Algorithms for fault-tolerant placement of stateful virtualized network functions," in *2018 IEEE International Conference on Communications (ICC)*, Kansas City, MO, 2018, pp. 1–7.

[20] F. Carpio, S. Dhahri, and A. Jukan, "Vnf placement with replication for load balancing in nfv networks," in *IEEE ICC*, Paris, 2017, pp. 1–6.

[21] V. Eramo, M. Ammar, and F. G. Lavacca, "Migration energy aware reconfigurations of virtual network function instances in nfv architectures," *IEEE Access*, vol. 5, pp. 4927–4938, 2017.

[22] I. Jang and et al., "Optimal network resource utilization in service function chaining," in *Proc. IEEE NetSoft*, Seoul, South Korea, Jun. 2016, p. 11–14.

[23] V. Eramo and et al., "An approach for service function chain routing and virtual function network instance migration in network function virtualization architectures," *IEEE/ACM Trans. on Networking*.

[24] C. W. et al., "Towards optimal resource allocation of virtualized network functions for hierarchical datacenters," *IEEE Trans. on Network and Service Management*, vol. 15, no. 4, pp. 1532–1544, Dec. 2018.

[25] C. Chang, N. Nikaein, and T. Spyropoulos, "Radio access network resource slicing for flexible service execution," in *IEEE INFOCOM 2018 - IEEE INFOCOM WKSHPS*, Honolulu, HI, 2018, pp. 668–673.

[26] C. Papagianni, P. Papadimitriou, and J. Baras, "Rethinking service chain embedding for cellular network slicing," in *IFIP Networking*, 2018, pp. 253–261.

[27] J. Soares and S. Sargento, "Optimizing the embedding of virtualized cloud network infrastructures across multiple domains," in *Proc. IEEE ICC*, London, U.K., Jun. 2015, pp. 442–447.

[28] R. M. et al., "Design and evaluation of algorithms for mapping and scheduling of virtual network functions," in *Proc. of IEEE NetSoft*, London, 2015, pp. 1–9.

[29] C. Papagianni, P. Papadimitriou, and J. Baras, "Towards reduced-state service chaining with source routing," in *IEEE CNSM*, 2018, pp. 438–443.

[30] R. C. et al., "Near optimal placement of virtual network functions," in *IEEE INFOCOM*, Hong Kong, China, April 2015.

[31] D. D. et al., "Multi-provider service chain embedding with nestor," in *IEEE Trans. on Network and Service Management*, vol. 14, no. 1, 2017, pp. 91–105.

[32] S. R. et al., "Incorporating trust in nfv: Addressing the challenges," in *2017 20th Conference on Innovations in Clouds, Internet and Networks (ICIN)*, Paris, 2017, pp. 87–91.

[33] S. L. et al., "Security-aware virtual network embedding," in *2014 IEEE ICC*, Sydney, NSW, 2014, pp. 834–840.

[34] A. Kapoukakis and et al., "Reputation-based trust in federated testbeds utilizing user experience," in *IEEE CAMAD*, Athens, 2014, pp. 56–60.

[35] G. Theodorakopoulos and J. S. Baras, "On trust models and trust evaluation metrics for ad hoc networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 2, pp. 318–328, Feb. 2006.

[36] M. Hamann and M. Fischer, "Path-based optimization of nfv-resource allocation in sdn networks," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, Shanghai, China, 2019, pp. 1–6.

[37] M. C. Luizelli and et al., "Piecing together the nfv provisioning puzzle: Efficient placement and chaining of virtual network functions," in *IEEE IFIP*, Ottawa, ON, 2015, pp. 98–106.

[38] M. Dobrescu, K. Argyarki, and S. Ratnasamy, "Toward predictable performance in software packet-processing platforms," in *USENIX NSDI*, San Jose, CA, USA, March 2016.

[39] A. Abujoda and P. Papadimitriou, "Profiling packet processing workloads on commodity servers," in *IFIP WWIC*, Russia, June 2013.

[40] J. M. Lucas and M. S. Saccucci, "Exponentially weighted moving average control schemes: Properties and enhancements," *Technometrics*, vol. 32, pp. 1–29, 1990.

[41] L. Kuang, Z. Feng, Y. Qian, and G. Giambene, "Integrated terrestrial-satellite networks: Part two," *China Communications*, vol. 15, no. 8, pp. iv–vi, 2018.

[42] A. Guidotti, B. Evans, and M. Di Renzo, "Integrated satellite-terrestrial networks in future wireless systems," *International Journal of Satellite Communications and Networking*, vol. 37, no. 2, pp. 73–75, 2019, https://doi.org/10.1002/sat.1292.

[43] J. Liu, Y. Shi, Z. M. Fadlullah, and N. Kato, "Space-air-ground integrated network: A survey," *IEEE Communications Surveys Tutorials*, vol. 20, no. 4, pp. 2714–2741, 2018.

[44] A. Gholami, N. Torkzaban, J. S. Baras, and C. Papagianni, "Joint Mobility-Aware UAV Placement and Routing in Multi-Hop UAV Relaying Systems," ser. Ad Hoc Networks. Springer International Publishing, 2021, pp. 55–69. [Online]. Available: https://doi.org/10.1007/978-3-030-67369-7_5

[45] M. Jia, X. Gu, Q. Guo, W. Xiang, and N. Zhang, "Broadband hybrid satellite-terrestrial communication systems based on cognitive radio toward 5g," *IEEE Wireless Communications*, vol. 23, no. 6, pp. 96–106, 2016.

[46] T. Li, H. Zhou, H. Luo, and S. Yu, "Service: A software defined framework for integrated space-terrestrial satellite communication," *IEEE Transactions on Mobile Computing*, vol. 17, no. 3, pp. 703–716, March 2018.

[47] N. Torkzaban, A. Zoulkarni, A. Gholami, and J. S. Baras, "Capacitated beam placement for multi-beam non-geostationary satellite systems," in *2023 IEEE Wireless Communications and Networking Conference (WCNC)*, 2023, pp. 1–6.

[48] G. Wang, S. Zhou, S. Zhang, Z. Niu, and X. Shen, "Sfc-based service provisioning for reconfigurable space-air-ground integrated networks," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 7, pp. 1478–1489, 2020.

[49] N. Torkzaban, A. Gholami, J. S. Baras, and C. Papagianni, "Joint satellite gateway placement and routing for integrated satellite-terrestrial networks," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, 2020, pp. 1–6.

[50] N. Torkzaban and J. S. Baras, "Controller placement in sdn-enabled 5g satellite-terrestrial networks," in *2021 IEEE Global Communications Conference (GLOBECOM)*, 2021, pp. 1–6.

[51] ——, "Joint satellite gateway deployment & controller placement in software-defined 5g-satellite integrated networks," 2021.

[52] Y. Cao, Y. Shi, J. Liu, and N. Kato, "Optimal satellite gateway placement in space-ground integrated network for latency minimization with reliability guarantee," *IEEE Wireless Communications Letters*, vol. 7, no. 2, pp. 174–177, 2017.

[53] J. Liu, Y. Shi, L. Zhao, Y. Cao, W. Sun, and N. Kato, "Joint placement of controllers and gateways in sdn-enabled 5g-satellite integrated network," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 2, pp. 221–232, 2018.

[54] K. Yang, B. Zhang, and D. Guo, "Partition-based joint placement of gateway and controller in sdn-enabled integrated satellite-terrestrial networks," *Sensors*, vol. 19, no. 12, p. 2774, 2019.

[55] D. Papadimitriou, D. Colle, and P. Demeester, "Mixed integer optimization for the combined capacitated facility location-routing problem," *Annals of Telecommunications*, vol. 73, no. 1-2, pp. 37–62, 2018.

[56] R. Dhaou, L. Franck, A. Halchin, E. Dubois, and P. Gelard, "Gateway selection optimization in hybrid manet-satellite network," in *Intl. Conf. on Wireless and Satellite Systems*. Springer, 2015, pp. 331–344.

[57] R. Kushwah, S. Tapaswi, and A. Kumar, "Enhanced gateway deployment scheme for load balancing in heterogeneous networks," in *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2018, pp. 1–7.

[58] N. Torkzaban, C. Papagianni, and J. S. Baras, "Trust-aware service chain embedding," in *2019 Sixth International Conference on Software Defined Systems (SDS)*. IEEE, 2019, pp. 242–247.

[59] S. Knight, H. X. Nguyen, N. Falkner, R. Bowden, and M. Roughan, "The internet topology zoo," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 9, pp. 1765–1775, 2011.

[60] G. Wang, Y. Zhao, J. Huang, and W. Wang, "The controller placement problem in software defined networking: A survey," *IEEE Network*, vol. 31, no. 5, pp. 21–27, 2017.

[61] Q. Qin, K. Poularakis, G. Iosifidis, and L. Tassiulas, "Sdn controller placement at the edge: Optimizing delay and overheads," ser. IEEE INFOCOM 2018 - IEEE Conference on Computer Communications, 2018, pp. 684–692.

[62] Y. Cao, Y. Shi, J. Liu, and N. Kato, "Optimal satellite gateway placement in space-ground integrated network for latency minimization with reliability guarantee," *IEEE Wireless Communications Letters*, vol. 7, no. 2, pp. 174–177, 2018.

[63] Y. Cao, H. Guo, J. Liu, and N. Kato, "Optimal satellite gateway placement in space-ground integrated networks," *IEEE Network*, vol. 32, no. 5, pp. 32–37, 2018.

[64] Y. Cao, L. Zhao, Y. Shi, and J. Liu, "Gateway placement for reliability optimization in 5g-satellite hybrid networks," ser. 2018 International Conference on Computing, Networking and Communications (ICNC), 2018, pp. 372–376.

[65] N. Torkzaban, A. Gholami, J. S. Baras, and C. Papagianni, "Joint satellite gateway placement and routing for integrated satellite-terrestrial networks," ser. ICC 2020 - 2020 IEEE International Conference on Communications (ICC), 2020, pp. 1–6.

[66] J. Liu, Y. Shi, L. Zhao, Y. Cao, W. Sun, and N. Kato, "Joint placement of controllers and gateways in sdn-enabled 5g-satellite integrated network," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 2, pp. 221–232, 2018.

[67] W. Xia, T. Q. S. Quek, J. Zhang, S. Jin, and H. Zhu, "Resource allocation by submodular optimization in programmable hierarchical c-ran," ser. 2018 IEEE/CIC International Conference on Communications in China (ICCC), 2018, pp. 558–562.

[68] A. K. Tran, M. J. Piran, and C. Pham, "Sdn controller placement in iot networks: An optimized submodularity-based approach," *Sensors*, vol. 19, no. 24, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/24/5474

[69] K. Poularakis, J. Llorca, A. M. Tulino, I. Taylor, and L. Tassiulas, "Joint service placement and request routing in multi-cell mobile edge computing networks," ser. IEEE INFOCOM 2019 - IEEE Conference on Computer Communications, 2019, pp. 10–18.

[70] A. Badanidiyuru and J. Vondrák, *Fast algorithms for maximizing submodular functions*, ser. Proceedings of the 2014 Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1497–1514. [Online]. Available: https://epubs.siam.org/doi/abs/10.1137/1.9781611973402.110

[71] N. Buchbinder, M. Feldman, J. Naor, and R. Schwartz, "A tight linear time (1/2)-approximation for unconstrained submodular maximization," ser. 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science, 2012, pp. 649–658.

[72] E. Martins, Q. V. Pascoal, and M. B. Marta, "A new implementation of yen's ranking loopless paths algorithm." *Quarterly Journal of the Belgian, French and Italian Operations Research Societies*, vol. 1, no. 2, pp. 121–133, 2003.

[73] T. Rossi, M. De Sanctis, F. Maggio, M. Ruggieri, C. Hibberd, and C. Togni, "Smart gateway diversity optimization for ehf satellite networks," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 56, no. 1, pp. 130–141, 2020.

[74] A. J. Roumeliotis, C. I. Kourogiorgas, and A. D. Panagopoulos, "An optimized simple strategy for capacity allocation in satellite systems with smart gateway diversity," *IEEE Systems Journal*, pp. 1–7, 2020.

[75] N. Girault and P.-D. Arapoglou, "Optical feeder link architectures for very HTS: ground segment," ser. International Conference on Space Optics — ICSO 2018, Z. Sodnik, N. Karafolas, and B. Cugny, Eds., vol. 11180, International Society for Optics and Photonics. SPIE, 2019, pp. 2039 – 2049. [Online]. Available: https://doi.org/10.1117/12.2536121

[76] "Multiple gateway placement in large-scale constellation networks with inter-satellite links," *Int J Satell Commun Network*, vol. 39, pp. 47 – 64, 2021.

[77] Q. Chen, X. Chen, L. Yang, S. Wu, and X. Tao, "A distributed congestion avoidance routing algorithm in mega-constellation network with multi-gateway," *Acta Astronautica*, vol. 162, pp. 376 – 387, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0094576518317387

[78] A. Papa, T. De Cola, P. Vizarreta, M. He, C. Mas Machuca, and W. Kellerer, "Dynamic sdn controller placement in a leo constellation satellite network," ser. 2018 IEEE Global Communications Conference (GLOBECOM), 2018, pp. 206–212.

[79] A. Papa, T. de Cola, P. Vizarreta, M. He, C. Mas-Machuca, and W. Kellerer, "Design and evaluation of reconfigurable sdn leo constellations," *IEEE Transactions on Network and Service Management*, vol. 17, no. 3, pp. 1432–1445, 2020.

[80] D. Wei, N. Wei, L. Yang, and Z. Kong, "Sdn-based multi-controller optimization deployment strategy for satellite network," ser. 2020 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), 2020, pp. 467–473.

[81] S. Xu, X. Wang, B. Gao, M. Zhang, and M. Huang, "Controller placement in software-defined satellite networks," ser. 2018 14th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN), 2018, pp. 146–151.

[82] S. Wu, X. Chen, L. Yang, C. Fan, and Y. Zhao, "Dynamic and static controller placement in software-defined satellite networking," *Acta Astronautica*, vol. 152, pp. 49 – 58, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0094576518308269

[83] S. Wu, X. Liu, Q. Chen, J. Guo, L. Yang, Y. Zhao, and C. Fan, "Update method for controller placement problem in software-defined satellite networking," ser. 2019 28th International Conference on Computer Communication and Networks (ICCCN), 2019, pp. 1–7.

[84] D. G. K. Yang, B. Zhang, "Partition-based joint placement of gateway and controller in sdn-enabled integrated satellite-terrestrial networks." *Sensors (Basel)*, vol. 19, no. 12, pp. 2774–2794, 2019.

[85] D. K. Luong, Y. Hu, J. Li, and M. Ali, "Metaheuristic approaches to the joint controller and gateway placement in 5g-satellite sdn networks," ser. ICC 2020 - 2020 IEEE International Conference on Communications (ICC), 2020, pp. 1–6.

[86] https://ai.googleblog.com/2017/04/federated-learning-collaborative.html.

[87] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Comm.-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[88] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, "A survey on security and privacy of federated learning," *Future Generation Computer Systems*, vol. 115, pp. 619–640, 2021.

[89] A. Gholami, N. Torkzaban, J. Baras, and C. Papagianni, "Joint mobility-aware uav placement and routing in multi-hop uav relaying systems," in *Ad Hoc Networks*. Cham: Springer Intl. Pub., 2021, pp. 55–69.

[90] O. Shamir, N. Srebro, and T. Zhang, "Comm.-efficient distributed optimization using an approximate newton-type method," in *International conference on machine learning*. PMLR, 2014, pp. 1000–1008.

[91] S. Savazzi, M. Nicoli, and V. Rampa, "Federated learning with cooperating devices: A consensus approach for massive IoT networks," *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 4641–4654, 2020.

[92] Y. Wang, "Trust quantification for networked cyber-physical systems," *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 2055–2070, 2018.

[93] G. Theodorakopoulos and J. S. Baras, "On trust models and trust evaluation metrics for ad hoc networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 2, pp. 318–328, 2006.

[94] A. Gholami, N. Torkzaban, and J. S. Baras, "On the importance of trust in next-generation networked cps systems: An ai perspective," 2021.

[95] A. Gholami, U. A. Fiaz, and J. S. Baras, "Drone-assisted communications for remote areas and disaster relief," 2019.

[96] X. Liu and J. S. Baras, "Using trust in distributed consensus with adversaries in sensor and other networks," in *17th International Conference on Information Fusion (FUSION)*, 2014, pp. 1–7.

[97] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless network intelligence at the edge," *Proceedings of the IEEE*, vol. 107, no. 11, pp. 2204–2239, 2019.

[98] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.

[99] C. Xie, O. Koyejo, and I. Gupta, "Slsgd: Secure and efficient distributed on-device ml," in *Joint European Conf. on Machine Learning and Knowledge Discovery in Databases*. Springer, 2019, pp. 213–228.

[100] R. J. Fowler, M. S. Paterson, and S. L. Tanimoto, "Optimal packing and covering in the plane are np-complete," *Information Processing Letters*, vol. 12, no. 3, pp. 133–137, 1981. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0020019081901113

[101] K. Rose, E. Gurewitz, and G. Fox, "A deterministic annealing approach to clustering," *Pattern Recognition Letters*, vol. 11, no. 9, pp. 589–594, 1990. [Online]. Available: https://www.sciencedirect.com/science/article/pii/016786559090010Y

[102] "Data repository. (2019). federated learning: Example dataset (fmcw 122ghz radars)," https://github.com/labRadioVision/federated, accessed: 2021-03-12.

[103] N. Torkzaban, M. A. Amir Khojastepour, and J. S. Baras, "Codebook design for composite beamforming in next-generation mmwave systems," in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, 2022, pp. 1545–1550.

[104] N. Torkzaban and M. A. A. Khojastepour, "Codebook design for hybrid beamforming in 5g systems," in *ICC 2022 - IEEE International Conference on Communications*, 2022, pp. 1391–1396.

[105] G. Interdonato, E. Björnson, H. Quoc Ngo, P. Frenger, and E. G. Larsson, "Ubiquitous cell-free massive mimo communications," *EURASIP Journal on Wireless Comm. and Networking*, vol. 2019, no. 1, pp. 1–13, 2019.

[106] X. Wei, Y. Jiang, X. Wang, and C. Shen, "Tx-rx reciprocity calibration for hybrid massive mimo systems," *IEEE Wireless Communications Letters*, vol. 11, no. 2, pp. 431–435, 2022.

[107] R. Nie, L. Chen, Y. Chen, and W. Wang, "Hierarchical-absolute reciprocity calibration for millimeter-wave hybrid beamforming systems," 2022. [Online]. Available: https://arxiv.org/abs/2204.06705

[108] J. Vieira and E. G. Larsson, "Reciprocity calibration of distributed massive mimo access points for coherent operation," in *2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2021, pp. 783–787.

[109] Y.-S. Yang, W.-C. Huang, C.-P. Li, H.-J. Li, and G. L. Stüber, "A low-complexity transceiver structure with multiple cfos compensation for ofdm-based coordinated multipoint systems," *IEEE Transactions on Communications*, vol. 63, no. 7, pp. 2658–2670, 2015.

[110] F. Kaltenberger, H. Jiang, M. Guillaud, and R. Knopp, "Relative channel reciprocity calibration in mimo/tdd systems," in *2010 Future Network & Mobile Summit*, 2010, pp. 1–10.

[111] C. Shepard, H. Yu, N. Anand, E. Li, T. Marzetta, R. Yang, and L. Zhong, "Argos: Practical many-antenna base stations," in *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking*, ser. Mobicom '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 53–64. [Online]. Available: https://doi.org/10.1145/2348543.2348553

[112] J. Vieira, F. Rusek, and F. Tufvesson, "Reciprocity calibration methods for massive mimo based on antenna coupling," in *2014 IEEE Global Communications Conference*, 2014, pp. 3708–3712.

[113] R. Rogalin, O. Y. Bursalioglu, H. Papadopoulos, G. Caire, A. F. Molisch, A. Michaloliakos, V. Balan, and K. Psounis, "Scalable synchronization and reciprocity calibration for distributed multiuser mimo," *IEEE Transactions on Wireless Communications*, vol. 13, no. 4, pp. 1815–1831, 2014.

[114] J. Vieira, F. Rusek, O. Edfors, S. Malkowsky, L. Liu, and F. Tufvesson, "Reciprocity calibration for massive mimo: Proposal, modeling, and validation," *IEEE Tran. on Wireless Comm.*, vol. 16, no. 5, pp. 3042–3056, 2017.

[115] X. Jiang and F. Kaltenberger, "Channel reciprocity calibration in tdd hybrid beamforming massive mimo systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 3, pp. 422–431, 2018.

[116] X. Wei, Y. Jiang, Q. Liu, and X. Wang, "Calibration of phase shifter network for hybrid beamforming in mmwave massive mimo systems," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2302–2315, 2020.

[117] N. Torkzaban, M. A. A. Khojastepour, and J. S. Baras, "Channel reciprocity calibration for hybrid beamforming in distributed mimo systems," in *2023 IEEE Wireless Communications and Networking Conference (WCNC)*, 2023, pp. 1–6.

[118] N. Torkzaban, A. Khojastepour, and J. S. Baras, "Enabling cooperative hybrid beamforming in tdd-based distributed mimo systems," 2023.

[119] X. Jiang, M. **v**Cirkić, F. Kaltenberger, E. G. Larsson, L. Deneire, and R. Knopp, "Mimo-tdd reciprocity under hardware imbalances: Experimental results," in *2015 IEEE International Conference on Communications (ICC)*, 2015, pp. 4949–4953.

[120] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Channel estimation and hybrid precoding for mmwave cellular systems," *IEEE Journ. of Selected Topics in Sig. Proc.*, vol. 8, no. 5, pp. 831–846, 2014.

[121] S. Noh, M. D. Zoltowski, and D. J. Love, "Training sequence design for feedback assisted hybrid beamforming in massive mimo systems," *IEEE Transactions on Communications*, vol. 64, no. 1, pp. 187–200, 2016.

[122] S. Noh, M. D. Zoltowski, and D. Love, "Multi-resolution codebook and adaptive beamforming sequence design for mm-wave beam alignment," *IEEE Tran. on Wireless Comm.*, vol. 16, no. 9, pp. 5689–5701, 2017.

[123] J. Song, J. Choi, S. G. Larew, D. J. Love, T. A. Thomas, and A. A. Ghosh, "Adaptive millimeter wave beam alignment for dual-polarized mimo systems," *IEEE Transactions on Wireless Communications*, vol. 14, no. 11, pp. 6283–6296, 2015.

[124] E. Vlachos, G. C. Alexandropoulos, and J. Thompson, "Massive mimo channel estimation for millimeter wave systems via matrix completion," *IEEE Signal Processing Letters*, vol. 25, no. 11, pp. 1675–1679, 2018.

[125] E. Vlachos, G. C. Alexandropoulos, and J. Thomson, "Wideband mimo channel estimation for hybrid beamforming millimeter wave systems via random spatial sampling," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 5, pp. 1136–1150, 2019.

[126] J. Song, J. Choi, and D. J. Love, "Codebook design for hybrid beamforming in millimeter wave systems," in *2015 IEEE International Conference on Communications (ICC)*, 2015, pp. 1298–1303.

[127] D. Qiao, H. Qian, and G. Y. Li, "Multi-resolution codebook design for two-stage precoding in fdd massive mimo networks," in *2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2017, pp. 1–5.

[128] J. Tao, J. Xing, J. Chen, C. Zhang, and S. Fu, "Deep neural hybrid beamforming for multi-user mmwave massive mimo system," in *2019 IEEE Global Conf. on Sig. and Info. Proc. (GlobalSIP)*, 2019, pp. 1–5.

[129] A. Elbir and K. Mishra, "Joint antenna selection and hybrid beamformer design using unquantized and quantized deep learning networks," *IEEE Tran. on Wireless Comm.*, vol. 19, no. 3, pp. 1677–1688, 2020.

[130] A. M. Elbir and S. Coleri, "Federated learning for hybrid beamforming in mm-wave massive mimo," *IEEE Communications Letters*, vol. 24, no. 12, pp. 2795–2799, 2020.

[131] J. Chen, W. Feng, J. Xing, P. Yang, G. E. Sobelman, D. Lin, and S. Li, "Hybrid beamforming/combining for millimeter wave mimo: A machine learning approach," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 10, pp. 11 353–11 368, 2020.

[132] T. Nitsche, A. Flores, E. Knightly, and J. Widmer, "Steering with eyes closed: mm-wave beam steering without in-band measurement," 2015.

[133] M. Hussain and N. Michelusi, "Energy-efficient interactive beam alignment for millimeter-wave networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 2, pp. 838–851, 2019.

[134] C. N. Barati, S. A. Hosseini, M. Mezzavilla, T. Korakis, S. S. Panwar, S. Rangan, and M. Zorzi, "Initial access in millimeter wave cellular systems," *IEEE Transactions on Wireless Communications*, vol. 15, no. 12, pp. 7926–7940, 2016.

[135] M. A. Amir Khojastepour, S. Shahsavari, A. Khalili, and E. Erkip, "Multi-user beam alignment for millimeter wave systems in multi-path environments," in *2020 54th Asilomar Conference on Signals, Systems, and Computers*, 2020, pp. 549–553.

[136] N. Torkzaban and M. A. A. Khojastepour, "Codebook design for hybrid beamforming in 5g systems," in *ICC 2022 - IEEE International Conference on Communications*, 2022, pp. 1391–1396.

[137] M. Khojastepour and N. Torkzaban, "Codebook design for beamforming in 5g and beyond mmwave systems," Jan. 26 2023, uS Patent App. 17/856,305.

[138] N. Torkzaban, M. A. Amir Khojastepour, and J. S. Baras, "Codebook design for composite beamforming in next-generation mmwave systems," in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, 2022, pp. 1545–1550.

[139] M. Khojastepour and N. Torkzaban, "Codebook design for composite beamforming in next-generation mmwave systems," Mar. 23 2023, uS Patent App. 17/948,624.

[140] N. Michelusi and M. Hussain, "Optimal beam-sweeping and communication in mobile millimeter-wave networks," in *2018 IEEE International Conference on Communications (ICC)*, 2018, pp. 1–6.

[141] M. Nosrati, S. Shahsavari, S. Lee, H. Wang, and N. Tavassolian, "A concurrent dual-beam phased-array doppler radar using mimo beamforming techniques for short-range vital-signs monitoring," *IEEE Transactions on Antennas and Propagation*, vol. 67, no. 4, pp. 2390–2404, 2019.

[142] S. Atapattu, R. Fan, P. Dharmawansa, G. Wang, J. Evans, and T. A. Tsiftsis, "Reconfigurable intelligent surface assisted two–way communications: Performance analysis and optimization," *IEEE Transactions on Communications*, vol. 68, no. 10, pp. 6552–6567, 2020.

[143] N. Torkzaban and M. A. Amir Khojastepour, "Shaping mmwave wireless channel via multi-beam design using reconfigurable intelligent surfaces," in *2021 IEEE Globecom Workshops (GC Wkshps)*, 2021, pp. 1–6.

[144] S. Hur, T. Kim, D. J. Love, J. V. Krogmeier, T. A. Thomas, and A. Ghosh, "Millimeter wave beamforming for wireless backhaul and access in small cell networks," *IEEE Transactions on Communications*, vol. 61, no. 10, pp. 4391–4403, 2013.

[145] N. Torkzaban and M. Khojastepour, "Codebook design for beamforming in 5g and beyond mmwave systems," in *2022 IEEE ICC (Accepted)*, 2021.

[146] S. A. Busari, S. Mumtaz, S. Al-Rubaye, and J. Rodriguez, "5g millimeter-wave mobile broadband: Performance and challenges," *IEEE Communications Magazine*, vol. 56, no. 6, pp. 137–143, 2018.

[147] S. Kutty and D. Sen, "Beamforming for millimeter wave communications: An inclusive survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 2, pp. 949–973, 2016.

[148] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1164–1179, 2014.

[149] T. Nitsche, A. B. Flores, E. W. Knightly, and J. Widmer, "Steering with eyes closed: Mm-wave beam steering without in-band measurement," in *2015 IEEE Conference on Computer Communications (INFOCOM)*, 2015, pp. 2416–2424.

[150] C. N. Barati, S. A. Hosseini, M. Mezzavilla, T. Korakis, S. S. Panwar, S. Rangan, and M. Zorzi, "Initial access in millimeter wave cellular systems," *IEEE Transactions on Wireless Communications*, vol. 15, no. 12, pp. 7926–7940, 2016.

[151] M. Giordani, M. Mezzavilla, C. N. Barati, S. Rangan, and M. Zorzi, "Comparative analysis of initial access techniques in 5g mmwave cellular networks," in *2016 Annual Conference on Information Science and Systems (CISS)*, 2016, pp. 268–273.

[152] V. Desai, L. Krzymien, P. Sartori, W. Xiao, A. Soong, and A. Alkhateeb, "Initial beamforming for mmwave communications," in *2014 48th Asilomar Conf. on Signals, Systems and Computers*, 2014, pp. 1926–1930.

[153] M. Hussain and N. Michelusi, "Throughput optimal beam alignment in millimeter wave networks," in *2017 Information Theory and Applications Workshop (ITA)*, 2017, pp. 1–6.

[154] N. Torkzaban, M. A. Khojastepour, and J. S. Baras, "The trade-off between scanning beam penetration and transmission beam gain in mmwave beam alignment," in *2022 56th Asilomar Conference on Signals, Systems, and Computers*, 2022, pp. 16–21.

[155] ——, "Multi-user beam alignment in presence of multi-path," in *2022 56th Annual Conference on Information Sciences and Systems (CISS)*, 2022, pp. 224–229.

[156] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Transactions on Wireless Communications*, vol. 18, no. 8, pp. 4157–4170, 2019.

[157] C. Liaskos, S. Nie, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, and I. Akyildiz, "A new wireless communication paradigm through software-controlled metasurfaces," *IEEE Communications Magazine*, vol. 56, no. 9, pp. 162–169, 2018.

[158] E. Basar, M. Di Renzo, J. De Rosny, M. Debbah, M.-S. Alouini, and R. Zhang, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116 753–116 773, 2019.

[159] C. Hu, L. Dai, S. Han, and X. Wang, "Two-timescale channel estimation for reconfigurable intelligent surface aided wireless communications," *IEEE Transactions on Communications*, pp. 1–1, 2021.

[160] E. Basar, "Reconfigurable intelligent surface-based index modulation: A new beyond mimo paradigm for 6g," *IEEE Transactions on Communications*, vol. 68, no. 5, pp. 3187–3196, 2020.

[161] Z. Yang, W. Xu, C. Huang, J. Shi, and M. Shikh-Bahaei, "Beamforming design for multiuser transmission through reconfigurable intelligent surface," *IEEE Transactions on Communications*, vol. 69, no. 1, pp. 589–601, 2021.

[162] X. Yuan, Y.-J. A. Zhang, Y. Shi, W. Yan, and H. Liu, "Reconfigurable-intelligent-surface empowered wireless communications: Challenges and opportunities," *IEEE Wireless Communications*, vol. 28, no. 2, pp. 136–143, 2021.

[163] W. Tang, J. Y. Dai, M. Z. Chen, K.-K. Wong, X. Li, X. Zhao, S. Jin, Q. Cheng, and T. J. Cui, "Mimo transmission through reconfigurable intelligent surface: System design, analysis, and implementation," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 11, pp. 2683–2699, 2020.

[164] C. Huang, R. Mo, and C. Yuen, "Reconfigurable intelligent surface assisted multiuser miso systems exploiting deep reinforcement learning," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1839–1850, 2020.

[165] R. Liu, Q. Wu, M. D. Renzo, and Y. Yuan, "A path to smart radio environments: An industrial viewpoint on reconfigurable intelligent surfaces," 2021.

[166] D. Kitayama, Y. Hama, K. Goto, K. Miyachi, T. Motegi, and O. Kagaya, "Transparent dynamic metasurface for a visually unaffected reconfigurable intelligent surface: controlling transmission/reflection and making a window into an rf lens," *Opt. Express*, vol. 29, no. 18, pp. 29 292–29 307, Aug 2021. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-29-18-29292

[167] X. Pei, H. Yin, L. Tan, L. Cao, Z. Li, K. Wang, K. Zhang, and E. Björnson, "Ris-aided wireless communications: Prototyping, adaptive beamforming, and indoor/outdoor field trials," 2021.

[168] W. Tang, X. Li, J. Y. Dai, S. Jin, Y. Zeng, Q. Cheng, and T. J. Cui, "Wireless communications with programmable metasurface: Transceiver design and experimental results," *China Communications*, vol. 16, no. 5, pp. 46–61, 2019.

[169] Q. Nadeem, A. Kammoun, A. Chaaban, M. Debbah, and M.-S. Alouini, "Asymptotic max-min sinr analysis of reconfigurable intelligent surface assisted miso systems," *IEEE Transactions on Wireless Communications*, vol. 19, no. 12, pp. 7748–7764, 2020.

[170] S. Li, B. Duo, X. Yuan, Y.-C. Liang, and M. Di Renzo, "Reconfigurable intelligent surface assisted uav communication: Joint trajectory design and passive beamforming," *IEEE Wireless Communications Letters*, vol. 9, no. 5, pp. 716–720, 2020.

[171] A. Gholami, N. Torkzaban, J. S. Baras, and C. Papagianni, "Joint mobility-aware uav placement and routing in multi-hop uav relaying systems," 2020. [Online]. Available: https://arxiv.org/abs/2009.14446

[172] A. U. Makarfi, K. M. Rabie, O. Kaiwartya, X. Li, and R. Kharel, "Physical layer security in vehicular networks with reconfigurable intelligent surfaces," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, 2020, pp. 1–6.

[173] Z. Esmaeilbeig, K. V. Mishra, and M. Soltanalian, "Irs-aided radar: Enhanced target parameter estimation via intelligent reflecting surfaces," in *2022 IEEE 12th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 2022, pp. 286–290.

[174] Z. Esmaeilbeig, K. V. Mishra, A. Eamaz, and M. Soltanalian, "Cramer-rao lower bound optimization for hidden moving target sensing via multi-irs-aided radar," 2022. [Online]. Available: https://arxiv.org/abs/2210.05812

[175] Y. Han, W. Tang, S. Jin, C.-K. Wen, and X. Ma, "Large intelligent surface-assisted wireless communication exploiting statistical csi," *IEEE Trans. on Vehicular Tech.*, vol. 68, no. 8, pp. 8238–8242, 2019.

[176] M. Jung, W. Saad, Y. Jang, G. Kong, and S. Choi, "Performance analysis of large intelligent surfaces (liss): Asymptotic data rate and channel hardening effects," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 2052–2065, 2020.

[177] S. Abeywickrama, R. Zhang, Q. Wu, and C. Yuen, "Intelligent reflecting surface: Practical phase shift model and beamforming optimization," *IEEE Trans. on Comm.*, vol. 68, no. 9, pp. 5849–5863, 2020.

[178] H. Guo, Y.-C. Liang, J. Chen, and E. G. Larsson, "Weighted sum-rate maximization for reconfigurable intelligent surface aided wireless networks," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3064–3076, 2020.

[179] B. Di, H. Zhang, L. Li, L. Song, Y. Li, and Z. Han, "Practical hybrid beamforming with finite-resolution phase shifters for reconfigurable intelligent surface based multi-user communications," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 4, pp. 4565–4570, 2020.

[180] Y. Liu, X. Liu, X. Mu, T. Hou, J. Xu, M. Di Renzo, and N. Al-Dhahir, "Reconfigurable intelligent surfaces: Principles and opportunities," *IEEE Communications Surveys Tutorials*, pp. 1–1, 2021.

[181] M. A. ElMossallamy, H. Zhang, L. Song, K. G. Seddik, Z. Han, and G. Y. Li, "Reconfigurable intelligent surfaces for wireless communications: Principles, challenges, and opportunities," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 3, pp. 990–1002, 2020.

[182] N. Torkzaban, M. Farajzadeh-Tehrani, J. S. Baras *et al.*, "Ris-aided mmwave beamforming for two-way communications of multiple pairs," *ITU Journal on Future and Evolving Technologies*, vol. 4, no. 1, 2023.

[183] M. Khojastepour and N. Torkzaban, "Reconfigurable intelligent surface beamforming," Mar. 23 2023, uS Patent App. 17/948,751.

[184] ——, "Shaping mmwave wireless channel via multi-beam design using reconfigurable intelligent surfaces," Feb. 9 2023, uS Patent App. 17/863,720.

[185] B. Guo, C. Sun, and M. Tao, "Two-way passive beamforming design for ris-aided fdd communication systems," in *2021 IEEE Wireless Communications and Networking Conference (WCNC)*, 2021, pp. 1–6.

[186] N. Torkzaban, M. A. Amir Khojastepour, and J. S. Baras, "Codebook design for composite beamforming in next-generation mmwave systems," in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, 2022, pp. 1545–1550.

[187] M. Nosrati, S. Shahsavari, S. Lee, H. Wang, and N. Tavassolian, "A concurrent dual-beam phased-array doppler radar using mimo beamforming techniques for short-range vital-signs monitoring," *IEEE Transactions on Antennas and Propagation*, vol. 67, no. 4, pp. 2390–2404, 2019.

[188] J. Xu, Y. Liu, X. Mu, and O. A. Dobre, "Star-riss: Simultaneous transmitting and reflecting reconfigurable intelligent surfaces," *IEEE Communications Letters*, vol. 25, no. 9, pp. 3134–3138, 2021.

[189] Y. Jiang, F. Gao, M. Jian, S. Zhang, and W. Zhang, "Reconfigurable intelligent surface for near field communications: Beamforming and sensing," 2022. [Online]. Available: https://arxiv.org/abs/2204.10114

[190] J. Song, J. Choi, and D. J. Love, "Common codebook millimeter wave beam design: Designing beams for both sounding and communication with uniform planar arrays," *IEEE Transactions on Communications*, vol. 65, no. 4, pp. 1859–1872, 2017.

[191] Z. Mokhtari, M. Sabbaghian, and R. Dinis, "A survey on massive mimo systems in presence of channel and hardware impairments," *Sensors*, vol. 19, no. 1, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/1/164

[192] T. H. Pham, S. A. Fahmy, and I. V. McLoughlin, "Efficient integer frequency offset estimation architecture for enhanced ofdm synchronization," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, no. 4, pp. 1412–1420, 2016.

[193] M. Marey, "Integer cfo estimation algorithm for sfbc-ofdm systems," *IEEE Communications Letters*, vol. 22, no. 8, pp. 1632–1635, 2018.

[194] M. Marey and H. Mostafa, "Maximum-likelihood integer frequency offset estimator for alamouti sfbc-ofdm systems," *IEEE Communications Letters*, vol. 24, no. 4, pp. 777–781, 2020.

[195] J. D. Roth, D. A. Garren, and R. C. Robertson, "Integer carrier frequency offset estimation in ofdm with zadoff-chu sequences," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 4850–4854.

[196] A. Mohammadian and C. Tellambura, "Rf impairments in wireless transceivers: Phase noise, cfo, and iq imbalance – a survey," *IEEE Access*, vol. 9, pp. 111 718–111 791, 2021.

[197] P. Moose, "A technique for orthogonal frequency division multiplexing frequency offset correction," *IEEE Transactions on Communications*, vol. 42, no. 10, pp. 2908–2914, 1994.

[198] T. Schmidl and D. Cox, "Robust frequency and timing synchronization for ofdm," *IEEE Transactions on Communications*, vol. 45, no. 12, pp. 1613–1621, 1997.

[199] M. Ghogho, P. Ciblat, A. Swami, and P. Bianchi, "Training design for repetitive-slot-based cfo estimation in ofdm," *IEEE Transactions on Signal Processing*, vol. 57, no. 12, pp. 4958–4964, 2009.

[200] J. Yu and Y. Su, "Pilot-assisted maximum-likelihood frequency-offset estimation for ofdm systems," *IEEE Transactions on Communications*, vol. 52, no. 11, pp. 1997–2008, 2004.

[201] J. Lei and T.-S. Ng, "A consistent ofdm carrier frequency offset estimator based on distinctively spaced pilot tones," *IEEE Transactions on Wireless Communications*, vol. 3, no. 2, pp. 588–599, 2004.

[202] C. Chen, Y. Chen, Y. Han, H.-Q. Lai, and K. J. R. Liu, "High resolution carrier frequency offset estimation in time-reversal wideband communications," *IEEE Transactions on Communications*, vol. 66, no. 5, pp. 2191–2205, 2018.

[203] F. Gao and A. Nallanathan, "Reply to "a comment on 'blind maximum likelihood cfo estimation for ofdm systems via polynomial rooting'"," *IEEE Signal Processing Letters*, vol. 14, no. 4, pp. 292–292, 2007.

[204] Y. Meng, W. Zhang, G. L. Stüber, and W. Wang, "Blind fast cfo estimation and performance analysis for ofdm," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 10, pp. 11 501–11 514, 2020.

[205] T.-C. Lin and S.-M. Phoong, "A new cyclic-prefix based algorithm for blind cfo estimation in ofdm systems," *IEEE Transactions on Wireless Communications*, vol. 15, no. 6, pp. 3995–4008, 2016.

[206] L. Yang, H. Zhang, Y. Cai, and H. Yang, "Blind carrier frequency offset estimation for mimo-ofdm systems based on the banded structure of covariance matrices for constant modulus signals," *IEEE Access*, vol. 6, pp. 51 804–51 813, 2018.

[207] N. Torkzaban, A. Khojastepour, and J. S. Baras, "Blind cyclic prefix-based cfo estimation in mimo-ofdm systems," 2023.