6G BRAINS Deliverable D4.2

Modelling of Intelligent IAB and Hybrid User Terminal Optimized by the Multi-agent Deep Reinforcement Learning

Editors:	Dr. Jiancao Hou (Viavi); Israel Koffman (RunEL)
Deliverable nature:	Report
Dissemination level: (Confidentiality)	Public PU
Contractual delivery date:	M24
Actual delivery date:	21.12.2022
Suggested readers:	Public
Version:	1.0
Total number of pages:	63
Keywords:	IAB, Reinforcement Learning, Cell Free, Device to Device

Abstract:

This document updates the progress made in the different research areas related to 6G Access Architecture performed in the 6G BRAINS project. The updated research areas in this document include:

- Artificial Intelligence (AI) based scheduler for a 5G Cell-Free (CF) networks with Integrated Access and Back-haul (IAB)
- Grant-free NOMA for massive Machine-Type Communication
- Hybrid user terminal modelling for the D2D enabled cooperative network
- Intelligent IAB with Beam Steering based on User Location
- Advanced Test and Simulation Tools supporting 6G BRAINS research

Disclaimer

The research leading to these results has received funding from the European Union Horizon 2020 Programme under grant agreement number 101017226 *– 6G BRAINS – H2020-ICT-2020-2. The content of this document reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.*

Impressum

[Full project title]	Bring Reinforcement-learning Into Radio Light Network for Massive Connections
[Short project title]	6G BRAINS
[Number and title of work-package]	WP4: AI Driven High Dynamic Ultra-dense D2D Cell Free Network Architecture
[Number and title of task(s)]	T4.2: Multi-spectrum Links enabled Intelligent IAB and AI driven Beam Scheduler (Leader-REL) T4.4: AI Enabled User Terminal Modelling for the D2D-enhanced Resource Allocation Approach (Leader: ULEC)
[Document title]	D4.2: Modelling of Intelligent IAB and Hybrid User Terminal Optimized by the Multi-agent Deep Reinforcement Learning
[Editor: Name, company]	Israel Koffman, RunEL and Dr. Jiancao Hou, Viavi
[Work-package leader: Name, company]	Israel Koffman, RunEL
[Estimation of PM spent on the Deliverable]	9 PM

Executive summary

6G-BRAINS explores different architectures and essential modules for B5G and 6G ultra dense and dynamic cellular networks. In the document we describe the design of an Artificial Intelligence (AI) based scheduler for a 5G Cell-Free (CF) Networks with Integrated Access and Backhaul (IAB) and intelligent beam steering. The ultra-dense dynamic CF Network includes a cluster of Device to Device (D2D) User terminals (UE) and human-centric control interfaces.

The Cell free Integrated Access and Backhaul scheduler is subdivided into two parts, namely: (1) IAB Bandwidth Allocation, (2) Routing solution.

The IAB Resource Allocation finds the optimal way to divide the spectrum between the Backhaul and Access requirements of the different Donors and the Nodes in the Network using supervised learning AI method with CQI, DL and UL profile and connected base-station for **input parameters and** the rate each link must support and its efficiency for **cost function**. AI based Scheduler considers two approaches, namely: Cloud-based (Centralized) and distributed based. The research seeks to determine and quantify the trade-off between the benefits of the two approaches. The Supervised Learning model architecture uses 2 sub-models, namely: **Big model**: is implemented in the IAB-donor (cloud based) and divides the total bandwidth resources between all IAB (including itself) components (IAB Donor and Nodes); **Small model**: is implemented all the IAB components and allocated the bandwidth resources between the access (UL and DL links of the UEs) and the backhaul (UL and DL links to the served Nodes). This Supervised Learning AI Model then requires training.

The Routing solution is based on Multi-Agent Deep Reinforcement Learning (MA-DRL) for a fully synchronized time-slotted wireless network with the objective to find the optimal route from each BS (Donor or Nodes) to each user in terms of: Packet Error Probability (PER) for the whole packet trajectory; Maintenance of Quality of Service (QoS) requirements; Network Congestion Management including Queue management and Fairness. Two approaches have been examined in our previous report (D4.1), namely: a centralized model using the Dijkstra's algorithm; a decentralized solution using Q-Routing algorithm; to determine their benefits. Here we designed a decentralized routing algorithm called MA-RAC that learns from experience while interacting with a simulated 5G environment that outperforms the Q Routing scheduler previously introduced.

Efficient users clustering and power assignment is presented for a dynamic cell-free network that can cluster NOMA users over the limited spectrum resources and allocate optimal transmit power to each user in the cluster. The clustering finds a group of NOMA-based D2D users that can be scheduled on the same resource blocks forming a cluster. Parent points are distributed following a homogeneous Poisson Point Process, which every parent point (RmUE) is uniformly distributed in the certain area and then offspring points (mUEs) around one parent point to form a cluster.

Deep Reinforcement Learning (DRL) is used to predict the power, mobile users' frequency, users' location or beam trajectories, where the environment provides a reward to the agent for every interaction and the agent aims to select the right action for the next interaction to maximize the discounted reward over a time horizon. The goal is to design a DRL system that jointly optimizes the clustering of UEs and the beamforming vectors to obtain the optimal beamforming vector that maximizes the throughput and minimizes loss.

Artificial Intelligence (AI) is incorporated with beamforming and millimetre Wave (mmWave) enabling intelligent beam steering based on user locations.

Deep Learning Integrated Reinforcement Learning (DLIRL) algorithm is proposed for the beamforming solution to overcome the problems associated with mmWave like blockage impacting the coverage, reliability of highly mobile links, latency overheads associated with the highly mobile users in dense mmWave deployments requiring frequent hand-offs.

The DLIRL beamforming or beam steering track the user locations who is moving at a speed of 25 km/hour and based on that its location is changing in terms of latitude and longitude to perform beam steering towards the moving user. The beam steering experiment is implemented using MATLAB 2021a and data for DNN training is obtained via ray tracing with MATLAB's Site Viewer. The training for DLIRL is based on the impulse response of the received signal at the coordinated Base Stations from isotropic transmissions from user equipment, which is jointly received at the coordinating IABs, and used for performing training of the useful information about the surrounding environment because of transmitted signal's interaction with the surrounding environment. The beamforming, the power control and interference coordination is jointly carried out at the IABs to enhance the performance of the 5G network.

We acknowledge that telecommunications service providers have an urgent need to reduce operational costs while supporting the rapid introduction of new services and products and identifying and leveraging monetization opportunities. AI/ML has emerged as a powerful technology that can support these needs.

6G BRAINS recognizes that in order to transition to an industrialization phase and enable mass adoption of AI/ML, dedicated AI/ML research in specific Cellular Network modules is required. The research results provided in this document show the potential benefits of using this technology in 6G areas like Schedulers, Integrated Access and Backhaul, Device to Device Networks, Beam Forming and Human/Machine Interfaces. Furthermore, it is our opinion that AI/ML should be adopted at additional levels of the network architecture.

List of authors

Company	Author	Contribution
RunEL	Israel Koffman, Baruch Globen	Editor, Reviewer Chapter 1 and Chapter 2
Brunel University	Geoffrey Eappen, John Cosmas, Raj Nilavalan	Chapter 4
Leicester University	Guoqing Xia, Bohan Li and Huiyu Zhou	Chapter 3 and Chapter 6
VIAVI	Jiancao Hou	Chapter 5
Eurescom	Anastasius Gavras	Summary, Review
Thales	Alexandre Kazmierowski	Review

Table of Contents

1 Introduction 11
1.1 Introduction and Objective of this document 11
2 AI Based Cell Free Scheduler with IAB 13
2.1 Definition 13
2.2 General 14
2.2.1 Organisation and Notations 14
2.3 Problem Formulation 15
2.4 Theoretical Background 16
2.4.1 Mathematical Formulation 16
2.4.2 Routing Model 21
2.4.3 The Proposed Multi-Agent Relational Actor-Critic Routing Algorithm 24
2.5 Experimentation Results 26
2.5.1 Connectivity Influence 28
2.5.2 Experiment Results for Changing Load 29
2.5.3 Ablation Study 30
2.6 Conclusions and future work 31
3 Hybrid User Terminal Modelling for the High Dynamic Ultra-Dense D2D CF Network 32
3.1 Grant-free NOMA for massive Machine-Type Communication 32
3.1.1 Algorithm Definition 32
3.1.2 General Model 32
3.1.3 NOMA method for Massive Machine Type Communications (mMTC) 33
3.1.4 Conclusions and recommendation for future research 37
3.2 Hybrid user terminal modelling for the D2D enabled cooperative network 37
3.2.1 Model definitions 37
3.2.2 Technology background 38
3.2.3 Problems and methods 39
3.2.4 Simulation implementation and testing 42
3.2.5 Conclusions and recommendation for future research 44
4 Intelligent Beam Steering algorithm based on User Location 45
4.1 Autonomous Beam Steering 45
4.1.1 DRIDL algorithm definition 45
4.1.2 Simulation Implementation and Testing 46

4.2 Orthogonal Time Frequency Space (OTFS) Sensing of Distance 51
4.2.1 Definition 51
4.2.2 Simulation Implementation and Testing 51
4.3 Conclusions and recommendation for future research 53
5 Advanced Test and Simulation Tools supporting 6G BRAINS research 54
5.1 Definition 54
5.2 Simulation Implementation and Testing 56
6 Summary Conclusions and Recommendations 60
6.1 Deliverable Summary 60
6.2 Future work plans 61
7 References 62

Abbreviations

5G	Fifth Generation (mobile/cellular networks)
5G PPP	5G Infrastructure Public Private Partnership
6G BRAINS	Bring Reinforcement-learning Into Radio Light Network for Massive Connections
AI	Artificial Intelligence
AP	Access Point
AWGN	Additive White Gaussian Noise
B5G	Beyond 5 G
BS	Base Station
CAPEX	Capital Expenditure
CF	Cell-free
CQI	Channel Quality Information
CQT	Constant Q Transform
CSI	Channel state information
CSP	Communication Service Provider
D2D	Device to Device
DDPG	Deep Deterministic Policy Gradient
DLIRL	Deep Learning Integrated Reinforcement Learning
DQL & DDQL	Deep Q-Learning and Double DQL
DSP	Digital Signal Processor
FPGA	Field Programmable Gate Array
GPU	Graphics Processing Unit
IAB	Integrated Access and Backhaul
IoT	Internet of Things
KPI	Key Performance Indicator

LDA	Linear Discriminant Analysis
M2M	Machine to Machine
MA-DRL	Multi Agent Deep Reinforcement Learning
MDP	Markov decision process
MLP	Multilayer Perception
MVC	Model View Controller
NOMA	Non-Orthogonal Multiple Access
OPEX	Operational Expenditure
PCP	Poisson Cluster Process
PDSCH	Physical Downlink Shared Chanel
PEP	Packet Error Probability
QoE	Quality of Experience
QoS	Quality of Service
R&D	Research and Development
RAN	Radio Access Network
RmUE & mUE	Remote mobile UE and mobile UE
SARSA	State- Action-Reward-State-Action
SDN	Software Defined Networks
SINR	Signal to Interference + Noise Ratio
STFT	Short Time Fourier Transform
UE	User Equipment
UL & DL	Uplink and Downlink

1 Introduction

1.1 Introduction and Objective of this document

Communications service providers (CSPs) strive for relentless efficiency, business agility to address new revenue opportunities, and to meet or exceed customer expectations through a superior experience. This continues with the introduction of 5G programmable networks [1] which enable new revenue-creating opportunities through both enhanced user experience as well as the tailoring of telecommunications networks to provide differential services for both existing and new types of enterprise customers (e.g., Industry 4.0, Automotives, Fixed Wireless etc). The introduction of new technologies as well as additional services for customers, the densification of networks to support macro and micro coverage, and the need to ensure services with differing requirements significantly increases the complexity.

Artificial intelligence (AI) technologies have already matured to the point where CSPs have been applying them to their networks, often starting with non-time-critical processes, and are now applying them to the sensitive parts of their networks that directly impact user experience. The increased complexity of networks due to more services, new network technologies, and massive network densification further necessitates the application of AI in telecommunications networks as operations become more complex.

AI technologies can make many CSPs' system functions more capable as well as enable new system functions and approaches. Some example applications include:

- improving network performance though better **radio scheduling** optimized to the operator business model and users' requirements
- improving assurance of offered services and resources, moving from reactive to proactive — even in the face of increasing network complexity and heterogeneity
- improving optimization and use of existing resources, such as spectrum, transport, cloud infrastructure and network functionality
- improving experience management through both increased customers understanding as well as increased tailoring of the offered experience
- improving product and service definition, design, planning and offerings
- improving network and performance planning (such as radio, data centre location and transport)

The maturing capabilities of AI have resulted in increased attention within standardization and open-source communities, both from a purely technology evolution perspective as well as from an architecture definition perspective. While open source and standardization are enablers for increased AI adoption, the fragmentation which occurs in the early phases of industry specification can hinder adoption due to the uncertainty it creates, which occurs between different industry bodies as well as in different groups within industry bodies.

Consequently, CSPs are facing a number of challenges today regarding which standards to follow, which aspects of open source should be utilized directly or via vendors, how to increase industry alignment for scale while simultaneously allowing for differentiation, how to leverage the scale of public cloud providers, how to collect and manage data, and how to support the Life Cycle Management (LCM) of AI models.

The objective of this document is to update the progress performed in the work package 4 Research during the second year of the 6G BRAINS project in the area of innovative Access

Network research and concepts that are investigated by the beneficiaries that are part of the WP-4 team including:

- Artificial Intelligence (AI) based scheduler for a 5G Cell-Free (CF) Networks with Integrated Access and Backhaul (IAB) by RunEL
- Grant-free NOMA for massive Machine-Type Communication by the University of Leicester (ULEC)
- Hybrid user terminal modelling for the D2D enabled cooperative network by ULEC
- Intelligent IAB with Beam Steering based on User Location by Brunel University
- Advanced Test and Simulation Tools supporting 6G BRAINS research by Viavi.

2 AI Based Cell Free Scheduler with IAB

2.1 Definition

The research process of an AI based scheduler for Cell Free Cellular Networks with Integrated Access and Backhaul (IAB) capability was initially reported in Deliverable D4.1 [17] Chapter 2 of the 6G BRAINS project, this Chapter updates the initial report with the advancement performed in the last 12 months of the project.

The Cell Free Scheduler is comprised of two serial AI based components:

- The Spectrum allocation module based on a distributed Supervised Learning method
- The Routing module based on a distributed Multiple Agent Deep Reinforcement Learning method

In this report the second component (Routing module) of the serial scheduler is described in detail and the significant improvement over the results reported in D4.1 are highlighted. The research is performed by RunEL with collaboration of the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Be'er-Sheva, Israel.

The following Figure 1 describes the Scheduler main mission in modern cellular networks, the large number of different parameters that influence its performance that makes it impossible to find a deterministic model that can deliver an optimal scheduling result; and therefore, we need to search for an AI based solution

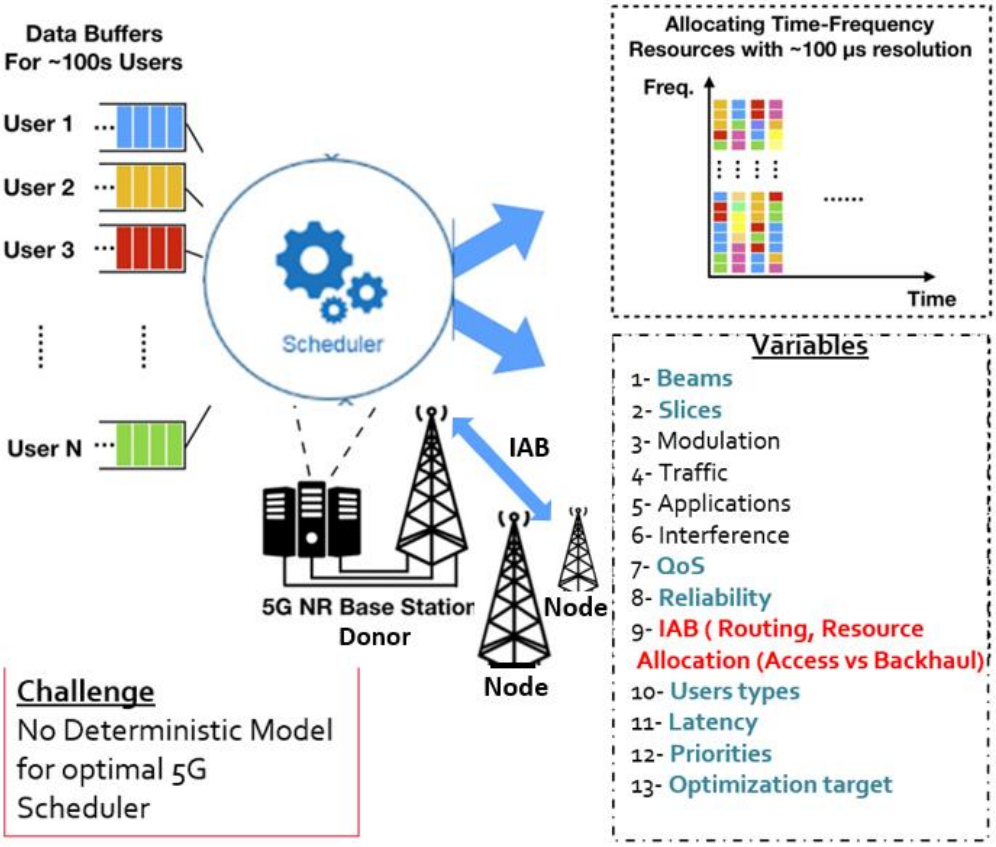


Figure 1: Cell Free Scheduler with IAB

2.2 General

The growing use of wireless communications and the limitations of the electromagnetic spectrum have sparked the development of more efficient methods to improve network management techniques. To support those requirements, The 3rd Generation Partnership Project (3GPP) defines a standard, New Radio (NR), which proposes novel designs and technologies to comply with the requirements for fifth-generation (5G) networks [2]. As part of this new protocol, the 3GPP defines new bands at the millimetre Wave (mmWave) frequencies. By utilizing the severe path losses in mmWave frequencies, operators can exploit the spatial diversity and increase data rates over traditional congested bands. Due to the nature of those frequencies, operating at them introduces new physical challenges, such as severe path and penetration losses. To overcome these physical challenges, we can increase the density of the network and use beam-forming methods [3]. In spite of the potential benefits of increasing network density, the deployment and operation of fibre between the Next-Generation Node Base Station (gNB) and the core network would lead to high costs. Integrated Access and Backhaul (IAB) is a promising solution for successful 5G adoption because, as part of this solution, only a fraction of the gNBs are connected to the traditional fibre-like infrastructures [4]. Thus, we can reduce redundant deployment and operational costs by using the advantage of spatial diversity. The fiber-connected gNBs are called IAB donors, while the rest of the gNBs are called IAB nodes, and they use a multi-hop wireless connection for the backhaul traffic. Although IAB networks reduce deployment and operation costs, ensuring reliable network performance is an open research problem due to the highly nonstationary characteristic of this kind of network. Dynamic topology, shared wireless channels, and limited node capabilities are factors that need to be considered to support those requirements. The main paradigm in this context is network's congestion control via routing, in which each destination might have multiple paths and base stations monitor the network conditions to choose their next hop. There are two main approaches for implementing routing algorithms in a wireless network: A centralized approach and a distributed approach. In the centralized approach, there is a central network processor, which is a single point of contact for path selection, whereas in the distributed approach every node makes next-hop decisions based only on its own observations without knowing other nodes' decisions. In practical implementations, due to bandwidth limitations and multi-hop structure, the information sharing operation is limited only to the base station's neighbourhood, thus, the base-station (BS) can only observe a part of the current network state, which implies that when operating in a distributed manner, next hop transmission decisions are based only on partial observations. In this deliverable, we focus on the design of a distributed routing algorithm for an IAB based networks.

2.2.1 Organisation and Notations

The rest of this Chapter is organized as follows: Section 2.3 details the problem formulation and assumptions; Section III motivates and discusses the rationale for the chosen Multi-Agent Reinforcement Learning (MARL) approach and details the proposed algorithm. Section 2.4 reports simulation results, including a comparison with approaches proposed in previous works and with the optimal scheme (when possible). These results clearly demonstrate the advantages of the proposed approach over other approaches. Lastly, Section 2.5 concludes this work.

Throughout this work, we use N to denote natural numbers, bold letters, e.g., X to denote vectors, and Xi denotes the i'th element in the vector X, $i \geq 0$. Calligraphic letters used to denote sets, e.g., X, and the cardinality of a set denoted by $|\cdot|$, e.g., $|\mathcal{X}|$ is the cardinality of the set X.

2.3 Problem Formulation

We consider a multi-hop IAB wireless network with IAB donors, IAB nodes and multiple User Equipment (UE) [4]. IAB donors are wired to the core network, whereas IAB nodes use wireless communication to backhaul their traffic to the core network via a multi-hop connection. Both IAB donor and node provides an access and backhaul interfaces for UE and IAB nodes, respectively.

We model this network by an undirected weighted graph $G = (N, L, d)$, where N, L denotes the set of nodes and wireless links, respectively, and $d : L \rightarrow N$ assigns delay to each wireless link. There are three sets present in N, a set D of IAB donors, a set B of IAB nodes and a set U of UEs, e.g., $N = D \cup B \cup U$. Each of the nodes $n \in D \cup B$ is equipped with an independent buffer queue Q_n, a transceiver with beam-forming capability and routing ability. Each of the links $(n,m) \in L$ is a bidirectional link between node n and node m, portraying a time-varying wireless channel. We assume that time is slotted by $t \in N$, and for simplification, we assume that packets are constant in length and that transmission rates are limited to transmit integer numbers of packets per slot. As another simplification, we represent the wireless link's spectral efficiency as a delay between the two nodes of the graph using the aforementioned mapping. The reason we assume this is that links with different degrees of spectral efficiency will require different numbers of transmissions to transfer the same amount of data, so using a low spectral efficiency link instead of one with a high spectral efficiency will produce a larger delay.

Once an IAB node or UE is activated, it is connected to an already active node, i.e., either an IAB donor or another IAB node which has a path to an IAB donor. Thus, we build the network topology in an iterative greedy fashion similarly to [5] where we set constraints over the maximal number of IAB parents (Pmax parent), IAB children (Cmax children), number of users each base station has (Umax children) and the number of associated base-station each user has (Umax parent). It should be noted that by using the following topology generation scheme, we receives a connected graph, i.e., there is a path from any base-station to any other node in the network. We assume that all our nodes, operate at mmWave bands for both, backhaul and access transmission and reception (in-band backhauling) with beam-forming capabilities. Therefore, similarly to [6], we disregard the interference between non-assigned nodes since narrow beam mmWave frequencies have a power limit rather than interference limit.

In our network, F represents the set of information flows - that is, each $f_{ij} \in F$, $i, j \in N$ denotes an information flow between nodes i and j, dictating the amount of traffic between i^{th} and j^{th} nodes. In order to model the stochastic nature of packet's arrival process, we use a random Poisson process $\{X_i\}_{i=0}^{\infty}$ with parameter λ which we term as our network load. At each time-slot, we sample our arrival process, which resolves with the number of generated packets. Each packet is assigned a time limit when it is generated. We refer to this time as Time-To-Live (TTL). Upon expiration of this TTL, the packet is dropped. The packets are then distributed among the network's base-stations using the following paradigm. Each donor receives a certain number of packets, corresponding to 80 percent of his available wireless bandwidth,

while the remaining packets are distributed in a uniform random manner among all the network's base stations.

Afterwards, each base-station $i \in D \cup B$ uniformly sample a destination ˜ d for each of her newly injected packets from the available destinations, i.e., ˜ $d \in \{j, \forall f_{i,j} \in F | i \text{ is constant}\}$. A final step is to push the new packets through the queue $\{Q_n\}_n \in D \cup B$ of each base station, where each queue contains packets waiting for transmission. The base-stations queue packets in an unlimited-sized prioritized queue based on TTL. Accordingly, the base-station always processes packets on top of the prioritized queue first.

We denote Ni as the set of neighbours of node $i \in N$, i.e., $(i, j) \in L$, $\forall j \in N_i$, and let $N,K \in N$ be the number of channels and activated base stations, respectively. At each time step, each base-station $i \in D \cup B$ extracts a set of packets $Pi \in Q_i$, $|P_i| \leq N$. This is followed by deciding where to send each packet $p \in P_i$, which means that the i^{th} base-station have to choose one destination from Ni for each packet p ∈ Pi. In our model, users may move or change their base-station associations between two consecutive time slots, which would resolve in a change of the network topology. In addition, the edges delay are slowly varying around their initial values to modulate the changes in the wireless link.

2.4 Theoretical Background

This paragraph presents our mathematical formulation of the problem in detail. We begin by exploring Markov Decision Processing and Partially Observed Markov Decision Processing and their application to the multi-agent scenario, Stochastic Game and Partially Observed Stochastic Game, respectively. We then define our problem as a Partially Observed Stochastic Game and introduce the tools we used in our algorithm from the field of Multi-agent Reinforcement Learning. We conclude by explaining our algorithm in detail.

2.4.1 Mathematical Formulation

2.4.1.1 Markov Decision Process (MDP)

Modelling sequential decision-making problems using Markov Decision Processes is a tool for simulating sequential interactions between a single decision maker and the environment.

The following steps (Figure 2) describe the decision-making process between the agent and the environment, according to this model:

1) At time t environment is in state St and agent chooses action At
2) At time t+1 environment makes a transition to state St+1 and responds with a reward Rt+1
3) $t \leftarrow t + 1$
4) Go back to 1.

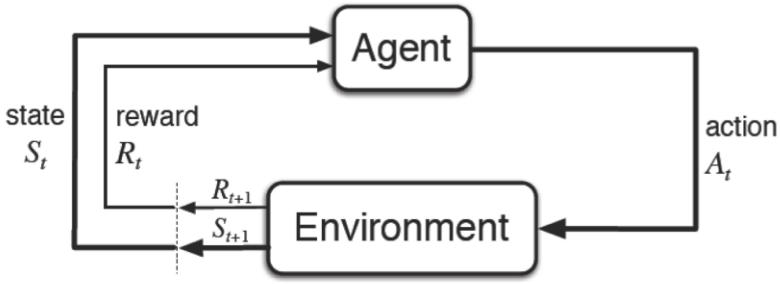


Figure 2: Markov Decision Process Framework

Under this framework, the agent's Goal is to select actions that maximize expected cumulative discounted reward G_t. Where we define G_t as follows,

$$G_t = \sum_{n=1}^{\infty} \gamma^n \mathrm{R}_{t+n+1}, \gamma \in [0, 1)$$

The discount factor γ determines how much immediate rewards are favoured over more distant rewards. Next, in theoretical terms, we can define a Markov Decision Process as a tuple < S,A,R,P, p_0, γ > where each component represents:

1) S is the discrete set of environment states.
2) A finite set of actions.
3) R : S × A × S → R is the reward function of the agent.
4) Initial state probability: $p_0(s) = Pr(S_0 = s)$
5) State transition probability matrix P, $p(s'|a, s) = P(S_{t+1} = s'|S_t = s, A_t = a) = Pa_{ss'}$
6) a discount factor γ ∈ [0, 1]

2.4.1.2 Partially Observe Markov Decision Process

A partially observable Markov decision process (POMDP) is a generalization of a Markov decision process (MDP). A POMDP models an agent decision process in which it is assumed that the system dynamics are determined by an MDP, but the agent cannot directly observe the whole underlying state as we can see in Figure 3. Instead, it must maintain a sensor model (the probability distribution of different observations given the underlying state) and the underlying MDP. Unlike the policy function in MDP which maps the underlying states to the actions, POMDP's policy is a mapping from the observations (or belief states) to the actions. Formally, a POMDP is a tuple < S,A,R,P,Ω,O, p_0, γ > where each component represents:

1) S is the discrete set of environment states.
2) A Finite set of actions.
3) R: S × A × S → R is the reward function of the agent. Set of rewards R ∈ R
4) State transition probability matrix P, $p(s'|a, s) = Pr(St+1 = s'|St = s,At = a) = Pa_{ss'}$
5) Observation conditional probability matrix Ω, $p(o|a, s) = Pr(O_{t+1} = o|S_t = s,At = a)$
6) Initial state probability: $p_0(s) = Pr(S_0 = s)$
7) a discount factor γ ∈ [0, 1]

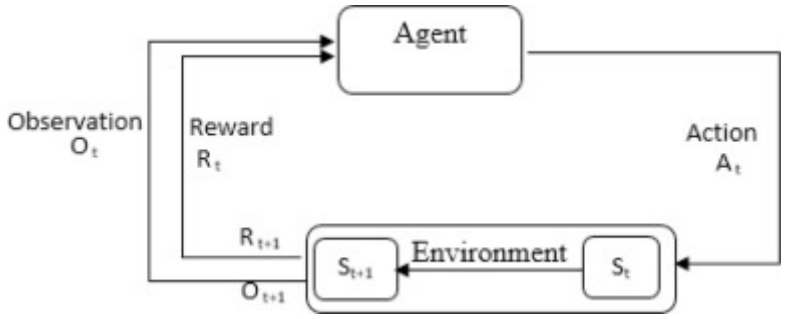


Figure 3: Partially Observe Markov Decision Process Framework

2.4.1.3 Stochastic Game

The generalization of the Markov decision process to the multi-agent case is the stochastic game. Similar to MDP, it is a tool for simulating sequential interactions between multiple decision makers and a single environment as we can see in Figure 4. At the beginning of each step the game is in some state. The agents select actions, and each agent receives a payoff based on the current state, the actions of the other agents, and their own action. Formally, a stochastic game (SG) is a tuple < S, A_1, . . . ,A_n , P, R_1, . . . ,R_n, p_0, γ > where each component represents:

1) n is the number of agents.
2) S is the discrete set of environment states.
3) $\{A_i\}^n_{i=1}$ are the discrete sets of actions available to the agents, yielding the joint action set
   $A = A_1 \times \cdots \times A_n$.
4) State transition probability matrix $P : S \times A \times S \rightarrow [0, 1]$, $p(s'|a, s) = P_r(S_{t+1} = s'|S_t = s, A_t = a) = P^a_{ss'}$
5) $R_i : S \times A \times S \rightarrow R$, i = 1, . . . , n are the reward functions of the agents.
6) Initial state probability: $p_0(s) = Pr(S_0 = s)$
7) a discount factor $\gamma \in [0, 1]$

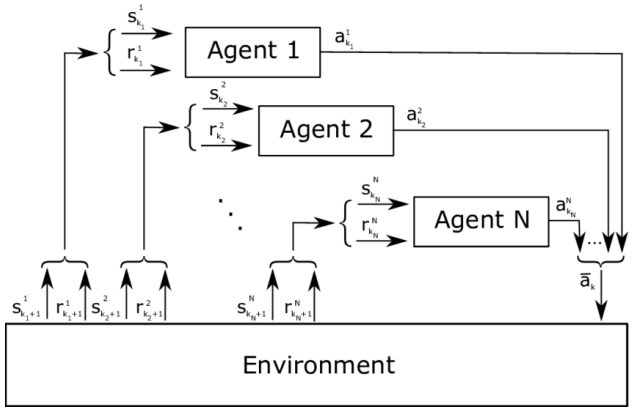


Figure 4: Multi-Agent Markov Decision Process Framework

In the multi-agent case, the state transitions are the result of the joint action of all the agents, $a = [a_1, \ldots, a_n]^T$, $a \in A$, $a_i \in A_i$. Consequently, the rewards ri and the returns G_i also depend on the joint action. The policies $\pi i : S \times A_i \rightarrow [0, 1]$ form together the joint policy Π. As extending stochastic games to a partially observed scenario is exactly the same as extending MDP to POMDP, we omit this explanation from this document. Next, we formulate our problem as a Partially-Observe Stochastic Game. Prior to detailing our algorithm, we briefly review the tools

we used from the field of multi-agent reinforcement learning to motivate our selection of this algorithmic approach.

2.4.1.4 Q-Learning

Q-learning is a model-free RL algorithm. which means, that it does not assumes any prior knowledge over the MDP model. When applied to an MDP < S,A,R,P, p_0, γ >, and under certain assumptions, this algorithm obtains the optimal policy in the sense of maximizing the expected accumulated discounted reward for any given initial state [[7], Ch. 6]. The Q-learning algorithm is a value-based RL algorithm, which means that it computes the optimal action-value function for finding the optimal policy. Let A denote the set of actions, S denote the set of states, and let $q_\pi(s, a)$, $v_\pi(s)$, $s \in S$, $a \in A$, denote the action-value and value function, respectively. Where, action-value function qπ(s, a) represents the expected accumulated discounted reward starting from state s, picking action a, and following policy π afterwards, and value $v_\pi(s)$ function represents expected accumulated discounted reward starting from state s, and following policy π. The term $\gamma \in [0, 1)$, denotes the discount factor. Because we consider the case of infinite time-horizon problem, then [[7], Ch. 3]

$$q_\pi(s,a) \triangleq \mathbb{E}_\pi \left\{ G_t | \mathbf{S}_t = s, A_t = a \right\}, \; v_\pi(s) \triangleq \mathbb{E}_\pi \left\{ q_\pi(\mathbf{S}_t, A_t) | \mathbf{S}_t = s \right\}.$$

The optimal policy π^*, is a policy that satisfies

$$q_*(s,a) \triangleq q_{\pi^*}(s,a) \geq q_\pi(s,a)$$

for any policy π and for every possible state-action pair, $(s, a) \in S \times A$. The optimal policy can be obtained easily from the optimal action-value function, $q_*(s, a)$, as

$$\pi^*(s) = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \{ q_*(s,a) \}$$

The Q Learning algorithm iteratively estimates the optimal action-value function for each valid state-action pair in an online manner as follow: At each time step $t \in N$, the agent observes a state $s \in S$, selects an action $a \in A$, receives a reward r for executing the selected action $a \in A$, and observes the next state $s' \in S$. Then, the estimation of the corresponding $q_*(s, a)$, referred to as the Q-value and denoted as Q(s, a), is updated according to the update rule:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \cdot \left(r + \gamma \cdot \max_{a' \in \mathcal{A}} \{ Q(s',a') \} - Q(s,a) \right),$$

for some $\alpha \in (0, 1)$ referred as the learning rate. To explore various state-action pairs, the action a is selected according to an ϵ-greedy policy, meaning that most of the time the selected action maximizes the estimated optimal action-value function, whereas in the rest of the time the action is selected randomly from the set of all valid actions. Mathematically, the agent at state $s \in S$, selects an action

$$a = \underset{a' \in \mathcal{A}}{\operatorname{argmax}} \{ Q(s,a') \}$$

with probability $1 - \epsilon$, and a uniformly random action from all possible actions in state s, with probability ϵ. According to [[7], Ch. 6], this algorithm is proven to converge to $q_*(s, a)$ with probability 1 if all of the state-action pairs are visited infinitely often, and a variant of the usual stochastic approximation conditions is satisfied. Although our agent is guaranteed to converge to the optimal policy by following the ϵ-greedy exploration technique, this exploration method might increase the minimal number of interactions between the agent and the environment due to inefficient exploration of actions (it samples actions uniformly without considering any

prior knowledge). For this problem of exploration vs exploitation, there are some alternative methods that aim to optimize exploration by considering prior knowledge. For example, in policy gradient methods, the policy and therefore the exploration are optimized directly by using the agent's prior knowledge.

2.4.1.5 Policy Gradient and Actor Critic

Policy Gradient based methods are also a model-free RL algorithms but unlike the aforementioned Q-Learning, these methods instead learn a parameterized policy that can select actions without consulting a value function. [[7], Ch. 13]. Let < S,A,R,P, p0, γ >, denote our MDP. In policy gradient methods, the policy can be parameterized in any way, as long as $\pi(a|s; \theta)$, $\theta \in R^L$ is differentiable with respect to its parameters θ for every state-action pair (s, a) ∈ S × A, that is, as long as $\nabla_\theta \pi(a|s, \theta)$ exists and is always finite. Because we consider the case of infinite time-horizon problem, then our goal is to find a policy $\pi_{\theta\star}$ such that it maximizes our expected discounted rewards [[7], Ch. 13],

$$\theta^\star \triangleq \arg\max_\theta J(\theta) = \arg\max_\theta \mathbb{E}_{\pi_\theta}[G_0] = \arg\max_\theta \sum_{s\in\mathcal{S}} \mu(s) \sum_{a\in A} \pi(a|s;\theta) Q_{\pi_\theta}(s,a),$$

where μ is the steady-state distribution under policy

$$\pi,\ \mu(s) = \lim_{t\to\infty} \mathbb{P}[S_t = s | A_{0:t} \sim \pi],$$

which is assumed to exist and to be independent of the initial state (an ergodicity assumption). The basic idea behind policy gradient-based methods is to improve the policy π performance by using the gradient of the objective function J(θ) with respect to θ. These methods aim to maximize performance by updating their parameters iteratively using a 1st order approximation method known as gradient ascent. i.e.,

$$\theta_{t+1} \leftarrow \theta_t + \eta \cdot \nabla_\theta J(\theta),$$

For some η ∈ (0, 1) which is referred as the learning-rate. Policy gradient theorem [[7], Ch. 13] given below is the fundamental result which underlies those popular algorithms:

$$\nabla_\theta J(\theta) = \sum_{s\in\mathcal{S}} \mu(s) \sum_a Q_{\pi_\theta}(s,a) \nabla_\theta \pi(a|s;\theta),$$

Following both [[7], Ch. 13] and [8] we have,

$$\nabla_\theta J(\theta) = \sum_{s\in\mathcal{S}} \mu(s) \sum_a Q_{\pi_\theta}(s,a) \nabla_\theta \pi(a|s;\theta) = \mathbb{E}_{\pi_\theta}\Big[G_t \nabla_\theta \log\pi(a_t|s_t;\theta)\Big],$$

A policy parameter update algorithm that uses the following unbiased gradient estimator is known as the REINFORCE algorithm [8],

$$\nabla_\theta J(\theta) \approx G_t \nabla_\theta \log\pi(a_t|s_t;\theta)$$

REINFORCE algorithm is not applicable in infinite horizon scenario since we have to wait until we get to terminate state to estimate the full step return Gt. Because we are dealing with a scenario of infinite horizon, we will also need to estimate Gt. Fortunately, we can do so by integrating additional unbiased estimator, i.e.,

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}\Big[G_t \nabla_\theta \log\pi(a_t|s_t;\theta)\Big] = \mathbb{E}_{\pi_\theta}\Big[(R_t + \gamma \cdot v_{\pi_\theta}(s_{t+1})) \nabla_\theta \log\pi(a_t|s_t;\theta)\Big]$$

Additionally, due to the estimation process, this method tends to suffer from high variance; however, we can reduce a baseline function that will not affect the gradient estimation in any way, assuming the baseline is independent of the current action [[7], Ch. 13]. Thus, to reduce the variance of the estimated stochastic gradient, we introduce a baseline b(s) = $v_{\pi\theta}$ (s). This type of method is called 'Actor-Critic'. In a simple term, Actor-Critic is a Temporal Difference (TD) version of policy gradient. In the general case any function parameterization has two networks: Actor and Critic. The actor decided which action should be taken and critic inform the actor how good was the action and how it should adjust. The learning of the actor is based on policy gradient approach. In comparison, critics evaluate the action produced by the actor by computing the value function.

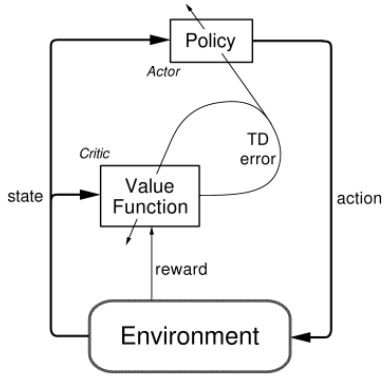


Figure 5: Actor Critic Framework

The Actor-Critic-learning algorithm (Figure 5) iteratively optimize the policy parameters while simultaneously estimate the corresponding value function for each valid state in an online manner as follow: At each time step t ∈ N, the agent observes a state s ∈ S, selects an action a ∈ A, receives a reward r for executing the selected action a ∈ A, and observes the next state s' ∈ S. First, we use the critic's evaluation to update the policy parameters, and we update them according to the following rule:

$$\theta \leftarrow \theta + \eta \cdot \nabla_\theta \log(\pi(a|s;\theta)) \cdot \big(r + \gamma \cdot V(s') - V(s)\big),$$

Then, the estimation of the corresponding $v_\pi(s)$, referred to as the State-value and denoted as V (s), is updated according to the update rule:

$$V(s) \leftarrow V(s) + \alpha \cdot \big(r + \gamma \cdot V(s') - V(s)\big),$$

for some α, η ∈ (0, 1) referred as the learning rates. Action-value methods have no natural way of finding stochastic optimal policies, whereas policy approximating methods can, which we will empirically show later that is our desirable policy behaviour. Following this, we define our mathematical formulation of the problem.

2.4.2 Routing Model

2.4.2.1 State

Let S be the global state space of our network and O_i be the i^{th} base-station observation space, where

$$\mathcal{S} \triangleq \times_{i=0}^{K-1} \mathcal{O}_i.$$

Next, we define our agent's observation, which is handcrafted from three kinds of features:

1) f_{packet} - features that are related to the packet's information.
2) $f_{base-station}$ - features that are related to the current base-station.
3) $f_{neighborhood}$ - features that are related to the current base-station neighbourhood.

By using those features, we aim to increase the agent's knowledge regarding the global network state. Throughout the following definitions we disregard the time index in favour of simplicity, moreover, let p, b represent the packet and the base-station, respectively.

1) Packet's Features: These are features that are derived from the current packet, to reduce the observation space complexity we introduce the mapping $l : N \rightarrow \{0, 1\}^K$,

$$l(n)[\![i]\!] \triangleq \begin{cases} 1 & \text{if } n \in \mathcal{N}_i \\ 0 & \text{if } n \notin \mathcal{N}_i \end{cases}$$

Where $i \in D \cup B$, by using this mapping, we can approximate the representation of each destination based on his base-station association. We also use the packet's time sensitivity, which is expressed as its time to live value. In conclusion:

$$f_{packet}(p) \triangleq \{l(p.destination), p.TimeToLive\}$$

2) Local Base-Station Features: These are features that are derived from the current base station, over here we use the base-station queue length. i.e.,

$$f_{base-station}(b) \triangleq \{Q_b\}$$

3) Neighbourhood Features: This feature set is derived from the current base-station's neighbourhood. We first collect the local features from all base-stations that are in the local neighbourhood, then aggregate those features using a size-invariant mapping such as Min(·), Max(·) and Mean(·). In this manner, it is like the phase of features aggregation in a graph neural network layer.

$$f_{neighborhood}(b) \triangleq \{Mean(\{|Q_i|\}_{i\in\mathcal{N}_b}), Min(\{|Q_i|\}_{i\in\mathcal{N}_b}), Max(\{|Q_i|\}_{i\in\mathcal{N}_b})\}$$

Let

$$\mathbf{O}_i \in \mathcal{O}_i$$

be an observation of the i^{th} base-station, where the j'^{th} element in Oi, Oi,j , represents our hand-crafted observation which is related to jth channel. Accordingly, let

$$F : (\cup_{i\in\mathcal{D}\cup\mathcal{B}}\mathcal{P}_i) \times (B \cup D) \rightarrow \mathbb{R}^{\dim(\mathbf{O}_{i,j})}$$

be our handcrafted feature mapping, thus, we can define $O_{i,j} \triangleq F(p_j , i) \triangleq f_{packet}(p_j) \cup f_{base-station}(i) \cup f_{neighborhood}(i)$ where $p_j \in Pi$ is the packet that the i^{th} base-station attempts to send over his j^{th} channel.

2.4.2.2 Action

Naturally, as our problem is to find the optimal path for each packet, we define

$$\mathcal{A}_i \triangleq \times_{n=0}^{N-1}\mathcal{N}_i$$

as the action set for the i^{th} base-station. Accordingly, the action set of the entire network

$$\mathcal{A} \triangleq \times_{i=0}^{K-1}\mathcal{A}_i.$$

When the action

$$A_j^{(i)}(t) = l \in \mathcal{A}_i$$

represents that the i[th] base-station is trying to transmit at the j[th] channel to l[th] node at time step t ∈ N.

2.4.2.3 Reward

Let Ri : S × A_i → **R**− be the reward function of the i[th] base-station. Also, let

$$R_j^{(i)}(t) \in \mathbb{R}^-$$

be the reward of the i[th] base-station for the transmission of packet p_j ∈ Pi to node *l* over the j[th] channel at time step t ∈ N. Similarly, to [9], we define our reward as the following equation:

$$R_j^{(i)}(t) = -(q_i(p_j) + d((i,l)))$$

Specifically, $q_i(p_j)$ represents the duration of packet p_j waiting at node i queue before transmission and $d((i, l))$ represents the transmission delay between node *i* and node *l*.

2.4.2.4 Modelling as a stochastic game

A reactive routing scheme is used in a multi-hop network to dynamically minimize packet delay while ensuring that packets reach their destination on time. There might be multiple hops, links with inefficient spectral efficiency, and nodes with overloaded queues in a packet's path, all of which may cause delay. Following our natural intuition, we begin by formulating the problem as a Partially Observed Stochastic Game, which we describe using the following tuple:

$$< \mathcal{S}, \mathcal{O}_1, \ldots, \mathcal{O}_K, \mathcal{A}_1, \ldots, \mathcal{A}_K, \mathcal{P}, \Omega, R_1, \ldots, R_K, p_0, \gamma >$$

Thus, this problem can be seen as a multi-agent problem, with K different agents/algorithms, each representing another base-station in our network. There are various metrics to measure or estimate the congestion within the network. In our scenario we define our congestion estimation using the following metrics:

- Packet latency - The time it takes for a packet to travel from its source to its destination.
- Arrival Ratio - The percentage of packets that made it to their destination successfully.

This multi-objective problem aims to minimize the packet latency while simultaneously maximizing the arrival ratio. Therefore, it may suffer from a Pareto-front, which means that optimizing with respect to one objective, leads to sub-optimal solution in another objective [10]. Despite the fact that the network performance measurements are well defined, an individual agent does not necessarily have access to their signals. For example, arrival ratio represents a metric which is dependent at the entire network, due to the multi-hop structure of the network, at each time slot an individual can't even obtain a good estimation of this value. We define K policies, each represents another agent and denoted by:

$$\pi_i : \mathcal{O}_i \times \mathcal{A}_i \rightarrow [0,1], i \in \{0, ..., K-1\}$$

The policies

$$\{\pi_i\}_{i=0}^{K-1}$$

form together the joint policy Π : S × A → [0, 1].

Finally, our objective is to derive an RL-based algorithm that identifies the joint policy Π that maximize the expected accumulated discounted reward over an infinite time horizon, i.e.

$$\Pi^* = \underset{\Pi}{\operatorname{argmax}} \Big\{ \mathrm{E}_{\Pi} \{ \sum_{l=0}^{\infty} \gamma^l \sum_{k=0}^{K-1} \sum_{n=0}^{N-1} R_n^{(k)}(t+l+1) | \mathbf{S}(t) \} \Big\},$$

for a discount factor γ ∈ [0, 1).

2.4.3 The Proposed Multi-Agent Relational Actor-Critic Routing Algorithm

Thriving for simplicity, we initially attempted to solve this task using standard routing-based reinforcement learning techniques such as Q-routing and Full-Echo Q-Routing [11]. Due to the challenges this task exhibits, namely partial observable, multi-agent optimization and highly dynamic topology, these methods did not manage to generalize well. The reason is that each agent suffers from performance degradation due to very partial observations which do not provide sufficient information about the entire network state and other agent policies. We proposed an algorithm to deal with those problems by using partial information exchange through the neighbourhood of each base-station ("Relational"), thus leading to cooperation between different agents where agents only share information with their neighbours. Our model leverages the agent's prior knowledge into the policy optimization process, reducing the need to explore multiple options, and allowing the agent to be more aware of his neighbourhood status through a sophisticated state representation. Through enhancing the agent's knowledge of his neighbourhood, we are aiming to improve the agent's estimation of the global network state.

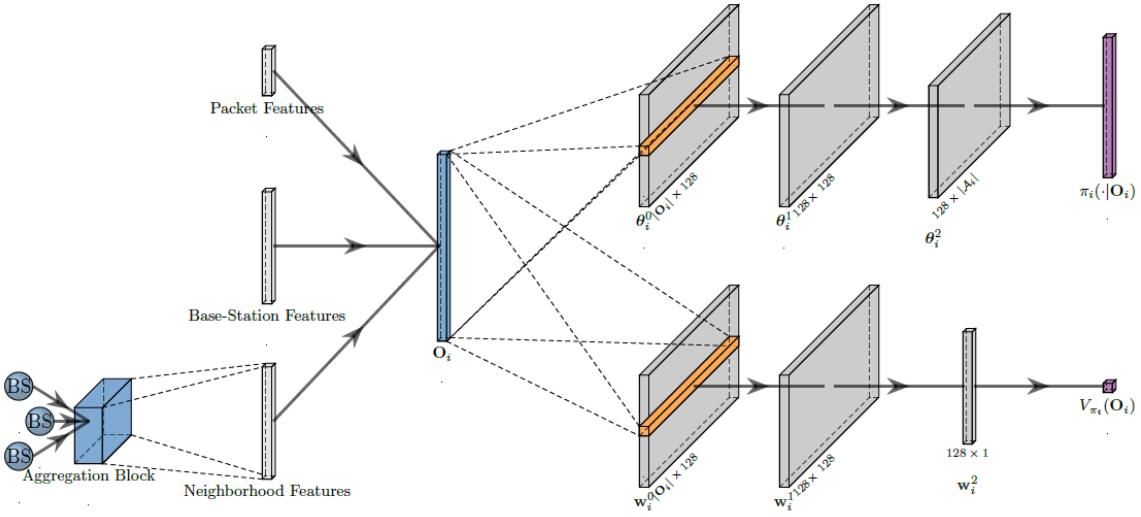


Figure 6: Relational Actor and Critic Neural Architecture of the ith Agent.

In this algorithm, each base-station is using an iterative online on policy method, Actor- Critic. In this scheme, the actor decided which action should be taken and critic inform the actor how good was the action and how it should adjust. Critic is used to learn an estimated representation of the network state based on the current observation, and then the actor uses this information to update its policy. Every node i ∈ DUB in the network represents its own strategy through its actor. For every agent we approximate both actor and critic using a neural network which we denote as

$$\pi_i(\cdot;\boldsymbol{\theta}_i), \hat{V}_{\pi_i}(\cdot;\mathbf{w}_i), \boldsymbol{\theta}_i \in \mathbb{R}^{L_1}, \mathbf{w}_i \in \mathbb{R}^{L_2}, L_1, L_2 \ll |\mathcal{A}| \cdot |\mathcal{S}|,$$

respectively, as depicted in Figure 6 above. Considering this representation, the node strategy determines the action *a* at node *i* by sampling the actor distribution $\pi_i(\cdot; \theta_i)$. As soon as node *i* sends a packet p, destined for node u, to one of its neighbouring nodes y ∈ N_i, he receives a feedback through the ACK signal that contains the neighbour y critic's estimation and the aforementioned reward $R^{(i)}$, ˆ $V_{\pi y}$ (F(*p, y*);$\mathbf{w}_y$) and $-(q_i(p)+d((i, y))$, respectively. The Critic's value essentially estimates the remaining time in the journey of the packet p starting from node y and destined to node u while following policy π_y.

Critic is updated through minimizing the following objective *L* w.r.t. w_i,

$$\mathcal{L}(\pi_i) = \mathbb{E}_{\Pi}\left[\sum_{n=0}^{N-1}\left(\hat{V}_{\pi_i}(\mathbf{O}_{n,i};\mathbf{w}_i) - (R_n^{(i)} + \gamma \cdot \hat{V}_{A_n^{(i)}}(F(p_n, A_n^{(i)});\mathbf{w}_{A_n^{(i)}}))\right)^2\right],$$

Where each critic updates its parameters using gradient descent method,

$$w_i \leftarrow w_i - \alpha \cdot \nabla_{\mathbf{w}_i}\mathcal{L}(\pi_i)$$

Next, the following approximation is proposed to approximate the objective gradient with respect to θ_i,

$$\nabla_{\boldsymbol{\theta}_i} J(\Pi) \approx \mathbb{E}_{\Pi}\left[\sum_{n=0}^{N-1}\nabla_{\boldsymbol{\theta}_i}\log\left(\pi(A_n^{(i)}|\mathbf{O}_{n,i};\boldsymbol{\theta}_i)\right)\cdot\left(R_n^{(i)}+\gamma\cdot\hat{V}_{A_n^{(i)}}\left(F(p_n, A_n^{(i)});\mathbf{w}_{A_n^{(i)}}\right)-\hat{V}_{\pi_i}\left(\mathbf{O}_{n,i};\mathbf{w}_i\right)\right)\right]$$

Where we neglect the time indexing for notional simplicity. Afterwards, each actor updates his policy using gradient ascent method,

$$\boldsymbol{\theta}_i \leftarrow \boldsymbol{\theta}_i + \eta\nabla_{\boldsymbol{\theta}_i} J(\Pi)$$

It should be mentioned that in a general partially observe stochastic game setting solved using multi-agent reinforcement learning techniques, as considered here, the transition probabilities *P* are unknown and only partial observations are available for each agent. A moving target poses one of the challenges in such a scenario. In contrast to a single-agent system, in which the state transition of the underlying environment is influenced only by the actions of a single agent, a multi-agent system is affected by the coordinated actions of all agents [12]. Consequently, under the setting of multi-agent, the assumptions of the single-agent algorithms are violated, whereby the MDP property becomes inoperative since the underlying environment is no longer stationary for the individual agents. As a result, the distribution of outcomes *s'* differs for unique policies *Π* and *Π'* over *N* updates for a given set of state-action pairs, i.e.,

$$\mathbb{P}(s'|s, a, \pi_0, \pi_1, \ldots, \pi_{K-1}) \neq \mathbb{P}(s'|s, a, \pi_0', \pi_1', \ldots, \pi_{K-1}')$$

Additionally, in multi-agent-based systems, a common problem is multiple Nash equilibrium points [13] within the joint policy space, which may resolve with convergence to a less desirable, local optima strategy solution [13]. As a result, convergence to the optimal policy is not guaranteed theoretically. In practice, however, it achieves very good performance even in various POMDP models with infinitely large state space. For example, the work of [14], developed an Actor-Critic algorithm for teaching multiple agents how to play Starcraft games directly from screen images, and achieved very good performance in various stages. In conclusion, at each time step *t ∈ N,* each agent simultaneously selects *N* actions (each for a different channel), where each action is chosen according to the following rule,

$$\mathbf{A}_j^{(i)}(t) = a, \text{w.p. } \pi_i(a|F(p_j, i);\boldsymbol{\theta}_i), \forall a \in \mathcal{A}_i,$$

The steps of the proposed Multi-Agent-Actor-Critic-Router algorithm are summarized below.

Algorithm 1 The Multi Agent Relational Actor-Critic Router Algorithm for Simultaneously Optimize Routing Strategy over Multiple Channels

for agent $k = 0, 1, \ldots, K-1$ **do**
  Initialize Actor and Critic weights θ_k, w_k.
  Observe $\mathrm{O}_k(1)$.
end for
for time step $t = 1, 2, \ldots$ **do**
  for Channel $n = 0, 1, \ldots, N-1$ **do**
    for BS $k = 0, 1, \ldots, K-1$ **do**
      $\mathbf{A}_n^{(k)}(t) = a$, w.p. $\pi_k(a|\mathrm{O}_{k,n}(t); \boldsymbol{\theta}_k), \forall a \in \mathcal{A}_k$,
    end for
  end for
  Execute actions.
  for BS $k = 0, 1, \ldots, K-1$ **do**
    Obtain the rewards $\mathrm{R}(t+1)$ associated with the k^{th} base-station.
    for channel $n = 0, 1, \ldots, N-1$ **do**
      Set target $y_n^k = R_n^{(k)}(t+1) + \gamma \cdot \hat{V}_{A_n^{(k)}}\left(F(p_n, A_n^{(k)}); \mathrm{w}_{A_n^{(k)}}\right)$.
    end for
  end for
  Update critic's and actor's parameters based on Eq. 7 and Eq. 8.
  for BS $k = 0, 1, \ldots, K-1$ **do**
    $\boldsymbol{\theta}_k \leftarrow \boldsymbol{\theta}_k + \eta \cdot \left(\frac{1}{N}\sum_{n=0}^{N-1} \nabla_{\boldsymbol{\theta}_k} \log\left(\pi(A_n^{(k)}|\mathrm{O}_{n,k}; \boldsymbol{\theta}_k)\right) \cdot \left(y_n^k - \hat{V}_{\pi_k}(\mathrm{O}_{n,k}; \mathrm{w}_k)\right)\right)$
    $\boldsymbol{w}_k \leftarrow \boldsymbol{w}_k - \alpha \cdot \nabla_{\boldsymbol{w}_k}\left(\frac{1}{N}\sum_{n=0}^{N-1}\left(\hat{V}_{\pi_k}(\mathrm{O}_{n,k}; \boldsymbol{w}_k) - y_n^k\right)^2\right)$
  end for
end for

2.5 Experimentation Results

During this section, we report the results of experiments carried out in order to demonstrate the importance of network routing in an IAB network as well as to test and evaluate the performance of the proposed MA-RAC algorithm. MA-RAC was implemented as described in above in MA-RAC Section 2.4. In the following, all the metrics we mentioned in Section 2.2 are used as the figure-of-merit for evaluating the performance of the different algorithms. We study here the performance of 2 different versions of MA-RAC, the first version ('Relational A2C') uses mutual weights between the different agents for both Actor and Critic, i.e., $w_i = w \cap \theta_i = \theta . \forall_i \in \{0, \ldots K-1\}$. As for the second version ('Dec-Rec-Relational A2C'), it functions in a decentralized manner, i.e., each base-station uses its own set of weights for both Actor and Critic. In order to suppress the non-stationary issue, a time dependence component was incorporated into this model.

In this study, MA-RAC's performance is evaluated against the performance of 5 other algorithms:

1) Centralized-Routing: In this policy, each agent may observe the full system state, and at each time step selects the shortest path (also considering the delay induced by queue of other agents) to the packet's destination.

2) Minimum-Hop-Routing: According to this policy, each agent observes the topology state, but not the internal queue states of the base-station, and at each time step selects the shortest path to the packet's destination.

3) Back-Pressure: In this policy, each node observes its own queues and the queues in its current neighbours (Node stores queues for each possible destination). Then, for a given destination the node calculates the differentiation between his queue and his neighbours' queues. Using this information, the node chooses which node to send the next packet. For the sake of fair comparison, we have developed a version of this algorithm that takes the message's direction into account as well.

4) Q-Routing: In this policy, each node is using an off policy iterative method, Q-learning [11]. In this case, Q-learning is used to first learn a representation of the network state in terms of Q-values and then use these values to make routing decisions. Each node n in the network represents its own view of the network state through its Q-table. Given this representation of the state, the action a at node n is to find the best neighbour node to deliver a packet which results to lower latency for the packet to reach its destination. As soon as node n sends a packet, destined for node d, to one of its neighbouring nodes *y*, node *y* sends its best estimate

$$\max_{a' \in \mathcal{A}_y} Q^{(y)}(z, d)$$

for the destination *d* back to node *n* over the ACK signal. This value essentially estimates the remaining time in the journey of the packet. Upon receiving $Q^{(y)}(z, d)$ node n computes the new estimate for $Q^{(n)}(y, d)$ as follows:

$$\hat{Q}^{(n)}(y, d) = \max_{a' \in \mathcal{A}_y} Q^{(y)}(a', d)$$

$Q^{(n)}(y, d)$ is node n's best estimated delay that a packet would take to reach its destination node d from node n when sent via its neighbouring node y excluding any time that this packet would spend in node n's queue, and including the total waiting time and transmission delay over the entire path that it would take starting from node y. Q-value is modified by the following formula:

$$Q^{(n)}_{new}(y, d) = Q^{(n)}_{old}(y, d) + \alpha \cdot \left(r + \gamma \cdot \hat{Q}^{(n)}(y, d) - Q^{(n)}(y, d)\right)$$

5) Hybrid-Routing: A Q-Routing agent is trained simultaneously with an on policy iterative method in this routing algorithm [15]. In this case, Q-learning is used to learn a representation of the network state in terms of Q-values and then Hybrid routing uses these values to update the agent's policy parameters by using actor-critic method. As soon as node *i* extract a packet destined to node *d* from its queue, the node selects its next action based on sampling his policy distribution, $\pi_i(d, \cdot; \theta_i)$. Then, the node sends the packet to one of its neighbouring nodes *y*. The corresponding Q-value is updated based on Q-Routing update rule, and then its policy parameters θ_i are updated according to the following formula:

$$\theta_i \leftarrow \theta_i + \alpha \cdot \nabla_{\boldsymbol{\theta}_i} \log \pi_i(y, d; \boldsymbol{\theta}_i) \cdot \left(r + \gamma \max_{a' \in \mathcal{A}_y} Q^{(y)}(a', d) - \max_{a \in \mathcal{A}_i} Q^{(i)}(a, d)\right)$$

To evaluate our algorithm performance, we have developed a gym-based simulated IAB environment [16]. The simulation takes place over a 2-dimensional grid. Table 1 and Table 2 describe network and algorithms hyper parameters, respectively.

Table 1: Simulation Network Hyper-parameters

$\|\mathcal{D}\|$	$\|\mathcal{B}\|$	$\|\mathcal{U}\|$	N	P^{max}_{parent}	$C^{max}_{children}$	$U^{max}_{children}$	U^{max}_{parent}	TTL
1	9	100	1	3	3	35	3	50

Table 2: Algorithm's Hyper-parameters

γ	α	η	ϵ_D	ϵ_{min}
0.995	0.0001	0.0001	0.9999	0.01

The performance of the different algorithms was obtained by averaging the outcomes of 5 independent experiments for each algorithm at each scenario.

Figure 7 illustrates a network composed of 1 channel with 1 IAB Donor, 9 IAB Nodes, 100 users, and a TTL duration of 50 time-slots.

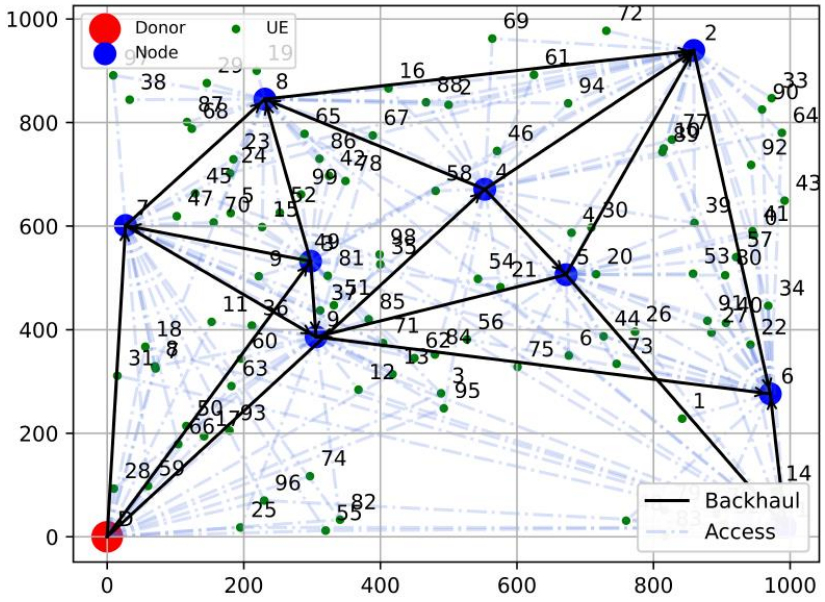
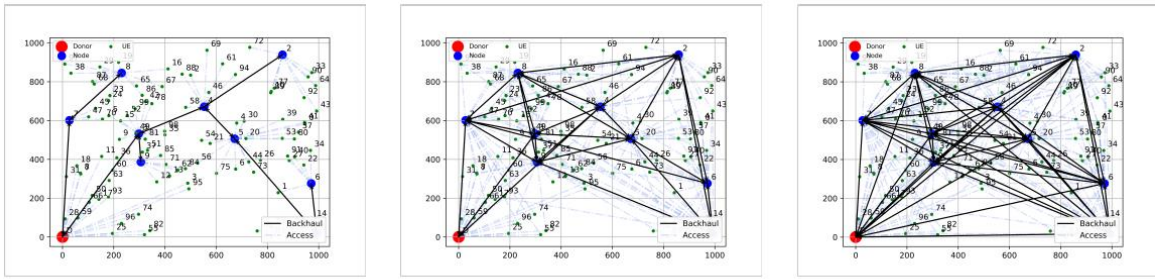


Figure 7: Network Topology Illustration

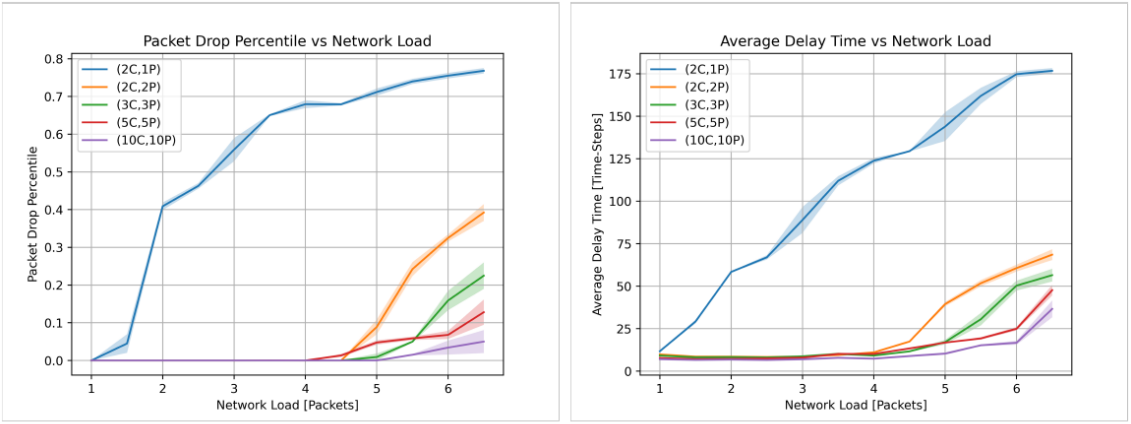
2.5.1 Connectivity Influence

In this experiment, we have measured the influence of network's connectivity over the performance in terms of the metrics we've mentioned in Section II. In order to change the network's connectivity, we have modified the constraints which dictates the number of parents each IAB node / User may have and the number of devices (IAB children / Users) that each IAB can support. Based on the results illustrated at Figure 8, we can infer that higher network connectivity is necessary to support the expected 5G's high load and low latency, but we must still consider that there is a trade-off between increased connectivity and interference between nearby base stations. Due to this conclusion, this study will examine how load-balancing techniques can improve the aforementioned metrics while taking partial connectivity into account.



(a) Single Parent Topology (No Routing). (b) 3 Parents Topology (c) Full-Connectivity Topology (Mesh).

Figure 8: Visualization of Different Connectivity Scenarios.

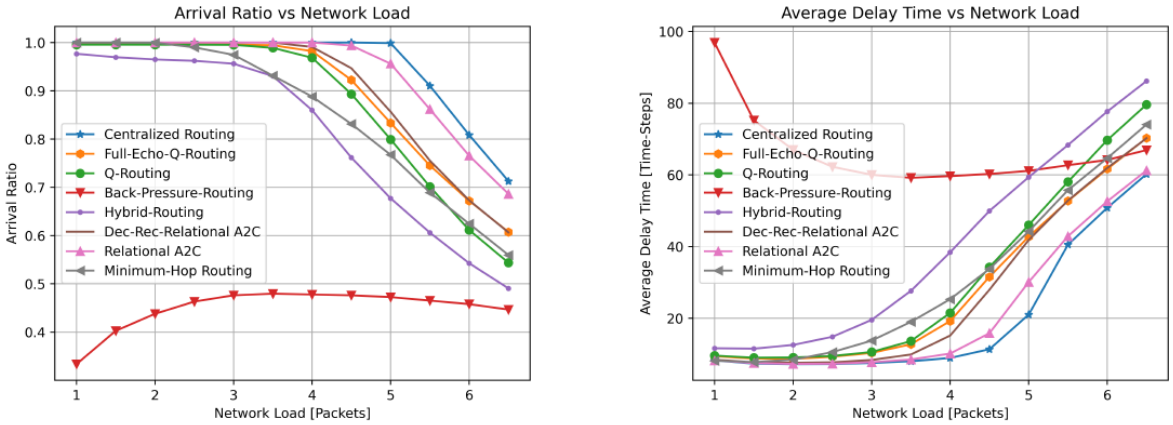


(a) Packet Drop Percentile Performance. (b) Average Delay Time Performance.

Figure 9: Performance illustration of different network's connectivity cases.

2.5.2 Experiment Results for Changing Load

In this experiment, we have measured the influence of network's load over the performance in terms of the metrics we've mentioned in Section 2.2. The parameter λ of the Poisson distribution has been modified in order to change the network's load. The parameter indicates the average number of packets generated in each time-slot by the IABs. To modify λ, we scanned various loads successively from bottom-to-top, and then from top-to-bottom. The results presented are an average of 10 different measurements for each load across five different network topologies. Based on the results illustrated at Figure 10, we can determine that although acting in a decentralized manner, both versions of MA-RAC's algorithm has managed to achieve superior performance than the other benchmarks. Furthermore, the mutual weight version achieved similar performance to the Centralized-Routing algorithm. Moreover, due to the significant gap between both versions of MA-RAC and the traditional hybrid algorithm we also conclude that an increased neighbourhood information is essential for performance in such a dynamic scenario.

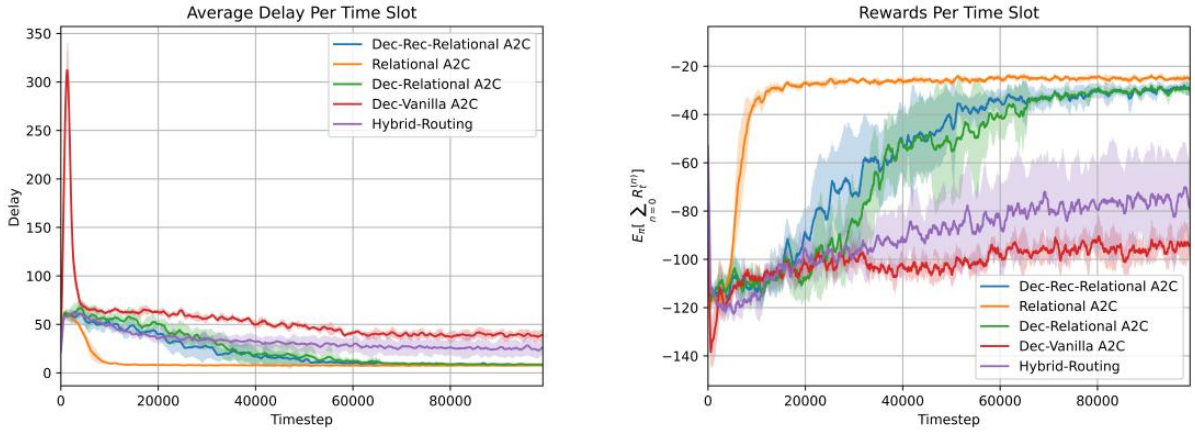


(a) Arrival Rate Performance. (b) Average Delay Time Performance.

Figure 10: Performance illustration of different routing algorithms under different loads

2.5.3 Ablation Study

We investigated how different components affected the performance of the algorithm in this experiment. We benchmark our baseline to be the Hybrid-Routing algorithm [15]. As a first extension we offer to use function approximation to approximate methods instead of the previous tabular representation of both Actor and Critic function. We term this method as 'Dec-Vanilla A2C'. As we can see from both Figure 11 and Figure 12, using this extension does not yield any performance improvements and even caused performance degradation. This result is reasonable and occurred due to the tabular nature of this problem state-space. Thus, we offer to extend our state-space to achieve better representation of the current network scenario as we proposed in Section 2.4. We term this method as 'Dec-Relational Actor-Critic'. As we can see from both train and test phases, Figure 11 and Figure 12, this extension improved the agent's performance. As another extension, we offer to combat the non-stationary issue with integration of time-dependency component into our model Figure 6, which we have located right after the first embedding layer for both Actor and Critic. We term this extension as 'Dec-Rec-Relational Actor-Critic'. Figure 11 and Figure 12 show that although this extension increased the model's complexity, it improved the convergence time and slightly improved performance during the test phase. Furthermore, we propose to improve the algorithm performance further by allowing agents to share information by means of network weights, which we identify as 'Relational A2C'. We are able to show in Figure 11 and Figure 12 that this method works better than all the previous extensions while remaining decentralized.



(a) Average Delay Time Performance Through Training. (b) Reward Performance Through Training.

Figure 11: Performance of the different algorithmic components during the training phase.

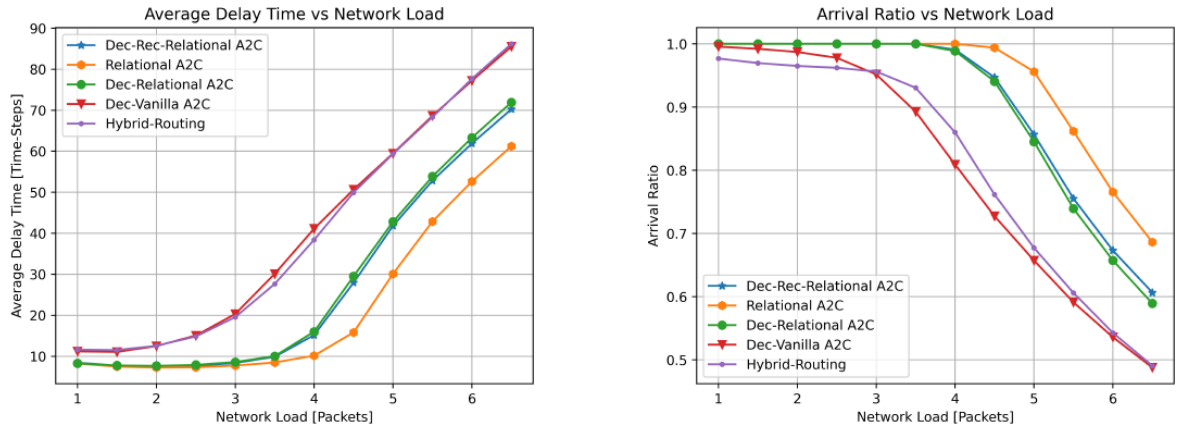


Figure 12: Performance of the different algorithmic components during the test phase.

2.6 Conclusions and future work

An advanced decentralized routing algorithm for 5G-NR IAB is proposed in this document to reduce congestion in the network. Our study showed that network routing is critical to reduce congestion in such an IAB network. Further, we designed a decentralized routing algorithm called MA-RAC that learns from experience while interacting with a simulated 5G environment. We benchmark our algorithm against the traditional reinforcement learning approach (Q-Routing) and a lower bound that we derive from a centralized solution. We conclude that our routing algorithm outperforms the traditional Q-Routing algorithm in different network scenarios, and it achieves results that are similar to those of a centralized solution.

The future work planed is to combine the simulations of the “Resource Allocation” scheduler reported in D4.1 and the Routing scheduler reported herein. In addition, the WP-4 team will try to combine the outcome of the UE dynamic simulation tool described in chapter 5 below with the “Resource Allocation scheduler” to obtain more realistic training data for the scheduler simulations.

3 Hybrid User Terminal Modelling for the High Dynamic Ultra-Dense D2D CF Network

3.1 Grant-free NOMA for massive Machine-Type Communication

3.1.1 Algorithm Definition

An adaptive matching pursuit (AMP) algorithm is proposed for the joint user activity detection and signal recovery for the grant-free (GF) non-orthogonal multiple access (NOMA) in terms of the massive machine-type communication (mMTC) scenario. The AMP algorithm takes the frame-wise block-sparsity into consideration and applies the matching pursuit method into the block-sparsity-based framework. It also employs a dynamic user sparsity decision method without needing the prior information of the noise level. The detailed algorithm implementation steps are summarized in the following.

3.1.2 General Model

For the mMTC scenarios, typical communication characteristics include small-data size transmission per device, high energy efficiency requirement, and sporadic transmission.

Considering these characteristics, many researchers have recently tried the grant-free schemes, by utilizing its virtue of avoiding complicated grant scheduling and signalling between the users and BS, thus enabling reduced signalling overhead and latency in comparison with the grant-based schemes.

The receiver design of the power domain-NOMA (PD-NOMA) depends on the received power difference among users for the effective data decoding based on the successive interference cancellation (SIC) technique. Two key factors would prevent the effective application of the PD-NOMA in grant-free scenarios. Firstly, without a loop-locked power control, guaranteeing sufficient power difference among received signals of multiple users is a challenge. To this end, an open-loop power control is developed, by using a multiple-agent double deep Q-network aided GF NOMA algorithm to determine the optimal transmit power levels. This multi-agent DRL algorithm relies on the predetermined distances between the users and the BS, which is somewhat unpractical due to the random location distributions of massive users. Secondly, a deterministic near-far situation is usually needed for the receiver to complete successful decoding. But in grant-free access, the random near-far phenomenon caused by the random user activity at one time slot would prevent the implementation of SIC. Thus, the code domain-based GF NOMA has been widely exploited by researchers, such as the spreading-based GF NOMA.

The general system model of spreading-based GF NOMA can be described as follows. Each user activates with a certain probability and selects the available channel resources. For each active user, the source bits are firstly modulated according to a certain modulation type. Then a symbol sequence is obtained by mapping the modulated symbol according to user-specific spreading signature. Finally, the joint user activity detection and signal reconstruction are realized by using advanced detection methods at the receiver.

The spreading-based NOMA utilizes specific spreading codes, which can be divided into sparse codes and non-sparse ones. Generally, to well decode the transmitted signals of the active users, the rank of the spreading matrix needs to be greater than $2s$ with s denoting the sparsity level of the active users. A more accurate condition for compressed sensing-based detection

can be referred to the restricted isometric property. An example of sparse spreading-based NOMA scheme, (sparse code multiple access) SCMA encoder is defined as a map from log2(*M*) bits to a K-dimensional complex codebook of size *M*. The K-dimensional complex codewords of a codebook are sparse vectors with N non-zero entries. K represents the spreading factor of the system and N denotes the dimension of the multi-dimensional constellation used in SCMA. The source bits are mapped into a K-dimensional codeword selected from the codebook and sparsely transmitted on *K* radio resources (e.g., OFDMA subcarriers). Note that the main difference between the SCMA and (low-density signature) LDS-based NOMA is that SCAM considers multi-dimensional constellation in spreading operation while the classic LDS scheme uses single dimensional constellation modulation only.

The compressive sensing (CS)-based multiple user detection (MUD) can be used for the grant-free NOMA. Far fewer samples than that required by the Nyquist criteria can be used to estimate the sparse signal by using the compressive sensing technique, thus reducing the computational complexity and calculation resources. The CS-based MUD algorithms at the receiver identify active users by detecting the positions of non-zero elements of the estimated transmitted symbol vector in noise-less scenario or the given number of largest entries in noisy scenario. A compressive sampling matching pursuit (CoSaMP) algorithm and a subspace pursuit (SP) algorithm are proposed with low complexity and excellent robustness to noise. Note that the main difference between CoSaMP and SP is the number of user indexes added into the candidate user set at each iteration. Low complexity MPA-based receiver can be employed for further data recovery by making full use of the sparsity of low-density MA signature structure (e.g., SCMA, LDS).

In addition, some block-sparsity-based methods are studied for GF NOMA. Firstly, considering that the users usually transmit signal in consecutive slots, the temporal correlation should be considered. Based on this, a frame-wise (block) sparsity model is applied, where the user activity remains constant over an entire data frame. A threshold-aided block sparsity adaptive subspace pursuit (TA-BSASP) and a cross-validation-based block sparsity adaptive subspace pursuit (CVA-BSASP) are respectively proposed by applying the block sparsity structure into the CoSaMP algorithm. The threshold-aided method can approach the oracle performance with the prior knowledge of the noise floor. The cross-validation based method is more practical with no prior knowledge of both the sparsity level and the noise floor. A complexity-reduced enhanced block-coordinate-descent based method is developed for MUD by using the block sparsity, which could reduce inter-user interference by pruning a majority of inactive users in advance and reduce complexity by removing the duplicate matrix multiplications.

3.1.3 NOMA method for Massive Machine Type Communications (mMTC)

3.1.3.1 NOMA method description

The code domain NOMA is considered in a single-antenna BS-based system to support massive connectivity in mMTC with limited channel resources. All users are also with a single antenna. As shown in Figure 13, massive users are distributed over the BS, while only a small part of user are active and transmit data at a slot. To support the massive connectivity for mMTC, we consider the overloading system with the number of subcarriers *K* less than the number of users *Q*. For an uplink transmission, one user first modulates its source bits to data symbols using a certain modulation scheme, e.g., M-QAM. Then, the modulated symbols spread onto

different channels based on the given spreading signatures. Finally, the spreading sequences of all users are superimposed at the BS.

Assuming the activities of the users are constant in one frame, the block sparsity can be utilized for modelling the received signal. The received signal over one frame can be denoted as:

$$\boldsymbol{Y} = \boldsymbol{G}\boldsymbol{X} + \boldsymbol{V}$$

where $\boldsymbol{G}$ is the equivalent channel matrix with each column $\boldsymbol{g}_q$ being the equivalent channel vector of the corresponding user q, $\boldsymbol{X}$ is the transmitted signal matrix with its column $\boldsymbol{x}_t$ as the transmitted signal of all users at the slot t, and $\boldsymbol{V}$ is the additive noise with the matching dimension. The target is to recover the transmitted signal $\boldsymbol{X}$ from the noisy measurements $\boldsymbol{Y}$. To apply the frame-wise block sparsity into the compressed sensing-based signal recovery, we vectorize the received signal as,

$$\mathrm{y} = \widetilde{\boldsymbol{G}}\boldsymbol{x} + \boldsymbol{v}$$

where $\widetilde{\boldsymbol{G}} = \boldsymbol{G} \otimes \boldsymbol{I}_T$ with $\otimes$ denoting the kronecker product operation, $\boldsymbol{x}$ is the vectorization of the transpose of $\boldsymbol{X}$, with its q^{th} block vector $\boldsymbol{x}_q$ denoting the transmitted signal of user *q, and* T is the number of slots in one frame. The optimization problem is formulated as

$$\operatorname{argmin} \varepsilon(\boldsymbol{x}) = \left\|\boldsymbol{y} - \widetilde{\boldsymbol{G}}\boldsymbol{x}\right\|_2^2$$

$$\|\boldsymbol{x}_t\|_0 \leq \bar{s}$$

where $\bar{s}$ is the maximum sparsity level. Therefore, the problem is to recover the transmitted signals $\boldsymbol{x}$ from the received signal $\boldsymbol{y}$ with the constraint of sparsity level.

The existing methods usually assume the known user sparsity level or determine the user sparsity level based on the prior noise floor, which are unpractical in most circumstances. Thus, it is necessary to design a flexible user sparsity decision method for the grant-free access in the mMTC scenario. The cross-validation aided method has recently been used to determine the user sparsity. However, the performance by using this method is unstable and can be affected by the number of the samples for the cross validation.

Figure 13: The massive machine type communication

3.1.3.2 Algorithm development

The compressed sensing-based method can be used to detect the user activity and recover the signal from the noisy measurements **y.** The iterations of the proposed adaptive matching pursuit (AMP) algorithm for the joint user activity detection and signal recovery are summarized in Algorithm 1 below. We define the vector $\boldsymbol{x}[q, T]$ as the q^{th} $T \times 1$ vector block of $\boldsymbol{x}$ and define the matrix $\widetilde{\boldsymbol{G}}[q, T]$ as the matrix block of $\widetilde{\boldsymbol{G}}$ constituted by consecutive columns from index (q-1)T+1 to index qT. Further, $\boldsymbol{x}[\Lambda, T]$ and $\widetilde{\boldsymbol{G}}[\Lambda, T]$ denote the sub-vector and sub-

matrix by selecting their respective blocks according to the indexes from the set Λ, respectively. The finding function $F(V,a)$ selects the indexes of the first a largest elements of an ordered set/vector V. The function $vec^{-1}(\boldsymbol{x},T)$ is used to unvectorize a vector into the corresponding transmitted signal matrix $\boldsymbol{X}$. The notation $\|\boldsymbol{r}\|_2$ denotes the l_2 norm of a matrix or a vector $\boldsymbol{r}$.

With known modulation constellation of the transmitted signals, the range of the l_2 norms of the transmitted signals of the active users in a given sampling duration (usually a frame) is generally smaller than a determined threshold, while the range would be relatively large if the inactive users are falsely detected (or deemed as active mistakenly). In fact, with the sampling duration T long enough, the range would converge to 1 if all symbols are transmitted with equal probability. We can speculate the falsely detected users exist when the range $\hat{\gamma}_s$ of the l_2 norms of the estimated signals is higher the threshold $\acute{\gamma}$, denoting the upper bound for the range under a limited sampling duration. Herein, the step 14 is mainly used for determining the sparsity levels higher than the real one under which the falsely detected inactive users must exist. The sparsity set in Algorithm 1 is $S = \{1,2,\dots,\bar{s}\}$ and the user set is $Q = \{1,2,\dots,Q\}$. The step 15 is due to the fact that the energy of the residual would gradually decrease with the sparsity level s increasing up to the real one.

Algorithm 1 The adaptive matching pursuit algorithm

Input: The received signals $\boldsymbol{y}$, equivalent channel matrices $\tilde{\boldsymbol{G}}$, the number of the consecutive time slots $\mathcal{T}$, the maximum user sparsity level $\bar{s}$, and the maximum iteration $\mathcal{L}$ for user detection.
Output: Reconstructed sparse signal $\boldsymbol{X}$

1: (Support initialization) The initial support set $\Gamma_{init} = \varnothing$.
2: **for** sparsity $s = 1$ to $\bar{s}$ **do**
3: (Residual and support initialization) Initial iteration index $\iota = 1$, $\boldsymbol{r}(\iota) = \boldsymbol{y}$, $\Gamma(1) = \Gamma_{init}$.
4: **repeat**
5: (Support estimation) $\Lambda = \Gamma(\iota) \cup \mathcal{F}(\{\|\tilde{\boldsymbol{G}}^{\mathrm{H}}[q,\mathcal{T}]\boldsymbol{r}(\iota)\|_2^2\}_{\mathcal{Q}}, s)$.
6: (LS estimation) $\boldsymbol{w}[\Lambda,\mathcal{T}] = (\tilde{\boldsymbol{G}}[\Lambda,\mathcal{T}])^{\dagger}\boldsymbol{y}$, $\boldsymbol{w}[\mathcal{Q}\setminus\Lambda,\mathcal{T}] = 0$.
7: (Support pruning) $\Gamma(\iota+1) = \mathcal{F}(\{\|\boldsymbol{w}[q,\mathcal{T}]\|_2^2\}_{\mathcal{Q}}, s)$.
8: (Signal estimation) $\boldsymbol{x}_\iota[\Gamma(\iota+1),\mathcal{T}] = (\tilde{\boldsymbol{G}}[\Gamma(\iota+1),\mathcal{T}])^{\dagger}\boldsymbol{y}$, $\boldsymbol{x}^\iota[\mathcal{Q}\setminus\Gamma(\iota+1),\mathcal{T}] = 0$.
9: (Residual update) $\boldsymbol{r}(\iota+1) = \boldsymbol{y} - \tilde{\boldsymbol{G}}\boldsymbol{x}^\iota$, $\iota = \iota+1$.
10: **until** $\|\boldsymbol{r}(\iota)\|_2^2 \geq \|\boldsymbol{r}(\iota-1)\|_2^2$ or $\iota - 1 = \mathcal{L}$.
11: (Sparsity update) $\boldsymbol{x}_s = \boldsymbol{x}_{\iota-2}$, $\varepsilon_s = \|\boldsymbol{r}(\iota-1)\|_2^2$, $\Gamma_s = \Gamma(\iota-1)$ and $\Gamma_{init} = \Gamma_s$.
12: (Range update) $\hat{\boldsymbol{X}} = [\mathrm{vec}^{-1}(\boldsymbol{x}_s,\mathcal{T})]^{\mathrm{T}}$, $\hat{\gamma}_s = \max_{q\in\Gamma_s}\|\hat{\boldsymbol{x}}_q\|_2 / \min_{q\in\Gamma_s}\|\hat{\boldsymbol{x}}_q\|_2$.
13: **end for**
14: (Candidate sparsity set) $\mathcal{S}_c = \mathcal{S}\setminus\{s\in\mathcal{S}:\hat{\gamma}_s > \acute{\gamma}\}$.
15: (Sparsity decision) $s_o = \arg\min_{s\in\mathcal{S}_c}\varepsilon_s$,
16: (Active set) $\Gamma = \Gamma_{s_o}$.
17: (Signal recovery) $\boldsymbol{X} = [\mathrm{vec}^{-1}(\boldsymbol{x}_{s_o},\mathcal{T})]^{\mathrm{T}}$.

3.1.3.3 Simulation Implementation and Testing

In this section, simulations are carried out to evaluate the normalized mean squared error (NMSE), the user detection error rate (DER), and the total symbol error rate (SER) of grant-

free AMP-based mMTC. The NMSE is defined as $NMSE = \frac{\|\hat{\boldsymbol{X}}-\boldsymbol{X}\|}{\|\boldsymbol{X}\|}$ where $\boldsymbol{X}$ is the transmitted signal matrix. The DER is defined as $DER = \frac{N_m+N_f}{\boldsymbol{N}}$ where N_m is the number of active users missed to be detected and N_f is the number of inactive users falsely detected to be active. The SER is defined as $SER = \frac{(N_m+N_f)T+N_d}{(N_a+N_f)T}$ where N_a is the number of the active users and N_d is the number of the symbol errors of the detected users.

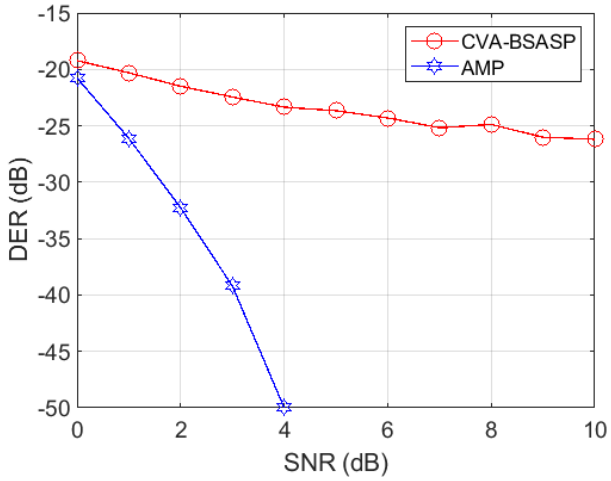


Figure 14: The detection error rate with respect to the input SNR

The benchmark considered is the cross-validation aided block-sparsity adaptive subspace pursuit algorithm (CVA-BSASP) which does not require any prior information either. Without loss of generality, we consider 40 users in total, with 5 active users in one frame. The number of the subcarriers used for the spreading-based NOMA is 20.

The detection error rate is shown in Figure 14. We could find that the proposed AMP algorithm presents significantly improved detection accuracy as compared to the CVA-BSASP. Moreover, the DER sharply decreases with the increase of the input SNR when using the proposed AMP and the detection error can even be eliminated when the SNR is higher than 4dB. However, the DER for the CVA-BSASP decreases slowly and even converge with the increase of the input SNR.

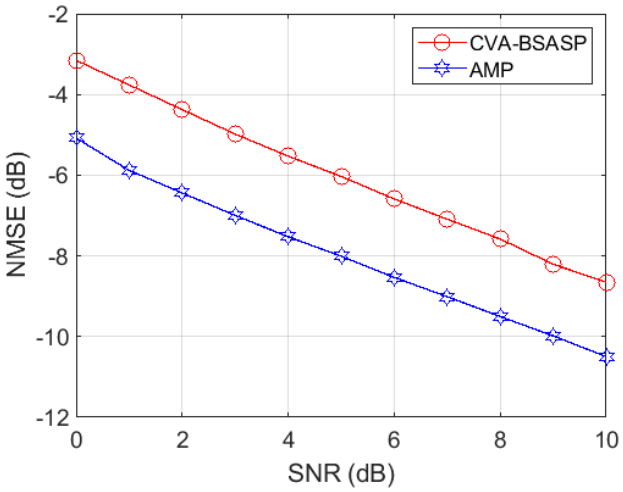


Figure 15: The normalized mean squared error with respect to the input SNR

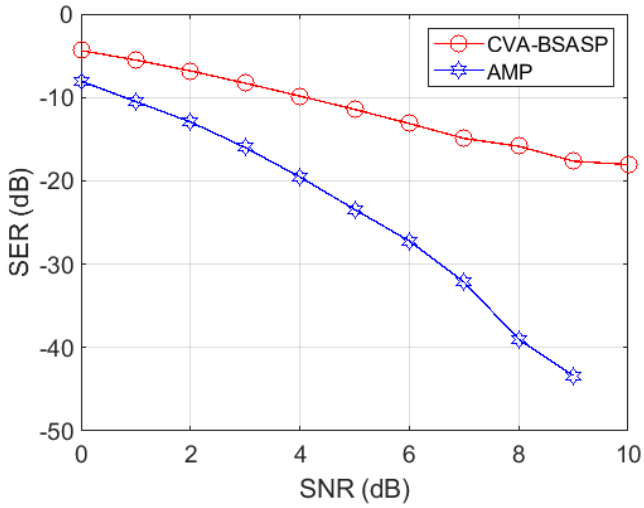


Figure 16: The symbol error rate with respect to the input SNR

Figure 15 and Figure 16 illustrate the NMSE and the SER with respect to the input SNRs, respectively. The results show that there is an approximate 2dB difference of NMSEs between the proposed AMP and the CVA-BSASP algorithm. However, the AMP outperforms the CVA-BSASP in terms of the SER, especially when the SNR is high. When the SNR is larger than 9dB, the symbol error can be eliminated too.

3.1.4 Conclusions and recommendation for future research

A block-sparsity-based adaptive matching pursuit algorithm is proposed for the joint user activity detection and signal recovery for the grant-free access in mMTC scenarios. The proposed method utilizes a novel user sparsity decision method with only the modulation constellation of the transmitted signal as the prior information, enabling the applicability of the proposed adaptive matching pursuit method.

The proposed method is designed based on the single-antenna base station and single-antenna users. In the future work, the multiple-antenna base station can be considered and the spatial division multiple access can be integrated with the spreading-based non-orthogonal multiple access for the system overloading in the massive users' communication scenarios.

3.2 Hybrid user terminal modelling for the D2D enabled cooperative network

3.2.1 Model definitions

A device-to-device (D2D) enabled cooperative network is considered for cell-free uplink communication. As shown in Figure 17, N D2D clusters are predetermined by existing clustering methods, such as K-means methods. The external users' equipment (EUEs) in one D2D cluster transmit their signals by NOMA principle to the cellular user equipment (CUE), which then combines the received signal and its own signal and transmit it to the BS. Beamforming and SIC decoding will be applied at the BS by exploiting the differences of the effective channel gains between clustered users (including EUEs and CUE) and the BS. Assume the base station is equipped with M antenna elements while both CUE and EUE are with single antenna, i.e., MISO system.

A two-phase transmission is applied in this cooperative cell-free uplink system. First, the EUEs transmit their massage signals to the CUE by NOMA in each cluster. The CUEs combine the received signal and its own signal by reallocating transmission power for them. Secondly, the CUEs transmit the superimposed signal to the BS which implements beamforming and SIC to decode the signals of respective users.

The closed-form signal-to-interference-plus-noise ratio (SINR) can be derived for both CUE and EUEs in each cluster with given beamforming weights and power allocation ratios. For the beamforming, the conjugate beamforming and zero-forcing beamforming methods can be utilized. The power allocation optimisation can be regarded as a Markov decision process, and a novel multi-agent deep reinforcement learning (DRL) scheme is designed to solve it. To meet the user fairness, the reward of the DRL environment is set to be the minimum SINR over all UEs.

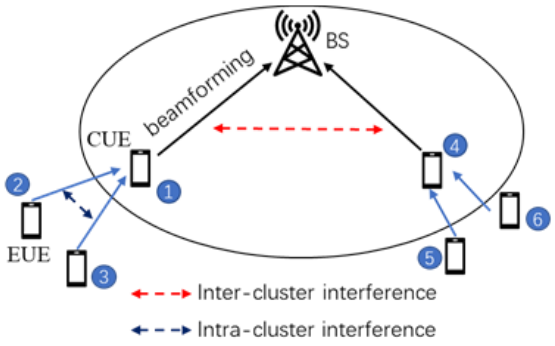


Figure 17: D2D enabled cell-free network

3.2.2 Technology background

mMTC widely exists in fifth-generation (5G) networks. The emerging NOMA scheme is attracting considerable attention due to its capacity to support massive connectivity in numerous applications including multimedia applications and the IoT [18]. However, with the number of the UE and tall infrastructures increasing, UEs, due to long distance or blockage by some obstacles, may not access the base station (BS) directly in cellular communication.

Recently, distributed antenna-based cell-free (CF) wireless network is being considered as a solution, where a large number of UEs in a geographical area will be served simultaneously in NOMA scenarios by a large number of spatially distributed access points (antennas), which coordinate with a centralized processing unit. The distributed antenna-based schemes face many challenges now. The first is to guarantee the synchronisation at distributed antennas for transmitting and receiving signals. The second is to dynamically determine the set of distributed antennas that serve the UEs near them and manage the interference between adjacent antenna sets in downlink mode or between adjacent users in uplink mode.

On the other hand, emerging cooperative NOMA device-to-device (D2D) communication is gradually applied for the downlink performance enhancement for far users within the cell coverage where the near cellular user functions as a relay. Two kinds of scenarios are classified according to if there is a direct communication link between the BS and far users. For the direct link scenario, the near user or central user plays a role of assistant, where in the first phase, the BS broadcasts signals using the NOMA protocol to a central user and a cell-edge user, and in the second phase, the central user helps the BS cooperatively relay signals intended for the cell-edge user. For the scenario without a direct link, the central user

functions as an enabler, where the BS broadcasts the superimposed signals to the central user in the first phase and the central user decodes and forwards the message signal for the far user in the second phase. Besides, this cooperative mode is also used in the cognitive network, where the secondary user shares the same frequency spectrum with the primary user by assisting the primary user communication as a combine-and-forward relay.

3.2.3 Problems and methods

3.2.3.1 Problem description

Inspired by the cooperative NOMA D2D communication, an uplink cell-free multiple-input-single-output (MIMO) network is developed by using the cellular UE (CUE) as a relay between the external UEs (EUEs) and the cell BS, shown in Figure 18. For this cell-free uplink communication system, three parts need to be considered, i.e., the clustering configuration of UEs (including one CUE and a couple of EUEs), transmit power allocation for UEs in each cluster based on NOMA and the beamforming at the BS. Many existing clustering methods for an NOMA network can be used for the considered uplink system, including match theory and k-means. After clustering, the closed-form expression of the signal-to-interference-plus-noise ratio (SINR) of each NOMA UE can be derived, based on the given beamforming weights, power allocation (PA) factors and the successive interference cancellation (SIC) decoding order. Let $\gamma_{n,k}$ denote the SINR of user k of cluster n.

Considering the SINR proportional to the user rate, rate maximisation is equivalent with optimising the beamforming and the PA. Many good beamforming methods are used in cellular or cell-free networks, including zero-forcing beamforming/precoding, conjugate beamforming/precoding and deep reinforcement learning scheme (DRL) based beamforming. Besides, a deep learning-based uplink power controlling method is proposed for rate maximization based on different criteria, i.e., max-sum, max-min and max-product. The max-min optimization aims to provide uniform service to all UEs for user fairness.

The conjugate and zero-forcing beamforming methods can be directly utilized. Therefore, the main problem in this scenario is to design an intelligent DRL system for the transmitting power allocation for the users in each cluster. Good power allocation schemes serve for the successive interference cancellation by sufficiently utilizing the channel differences, e.g., the random channel fading, the path loss and the shadowing loss caused by the spatial distributions.

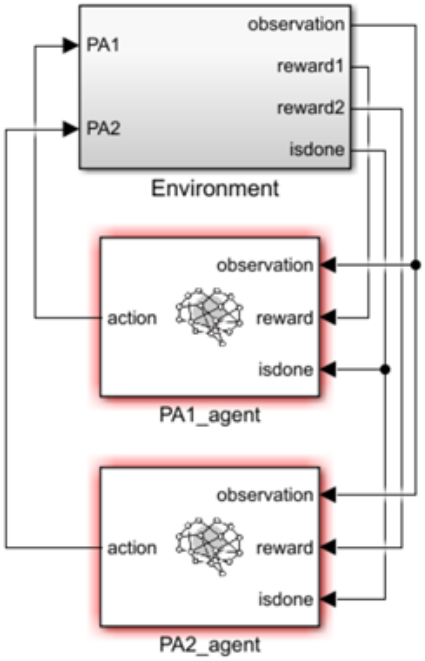


Figure 18: Multi-agent DRL interaction model

3.2.3.2 Algorithm development

Two DRL schemes are proposed, i.e., the single agent DRL and the multi-agent DRL scheme. When using the single agent, the power allocation coefficients for the users from all clusters are the output of the single agent network, while the power allocation for each cluster is learned by one agent by the multi-agent DRL scheme. Taking two agents as an example, the multi-agent DRL power allocation scheme is illustrated in Figure 19, where all agents collaboratively update their parameters by interacting with the environment.

Deep deterministic policy gradient (DDPG), as a DRL method, provides a solution to manage the problem with continuous state space and continuous action space. The DDPG agent consists of critic network $Q(s,a|\theta_Q)$, actor network $\mu(s|\theta_\mu)$ and their respective target networks $Q'(s,a|\theta_{Q'})$ and $\mu'(s|\theta_{\mu'})$. The state for the agent n for cluster n, $\boldsymbol{s}_n = \left[s_{n,1}, s_{n,2}, \dots, s_{n,K}\right]^T$ where state $s_{n,k}$ denotes the SINR of user *k* of cluster *n* and similarly the action set $\boldsymbol{a}_n = \left[p_{n,1}, p_{n,2}, \dots, p_{n,K}, \eta_n\right]^T$ where action $p_{n,k}$ denotes the power allocation ratio of user *k* for data transmission of cluster *n* and action η_n denotes the power allocation ratio for data forwarding of the cluster head of cluster *n*. K is the number of the users in each cluster n.

Two tricks are employed to stabilize the training of the DDPG actor-critic architecture.

1) the experience replay buffer to train the critic.
2) target networks for both the actor and the critic which are updated using the periodic Polyak averaging, i.e.,

$$\theta_{Q'}(t+t_0) = (1-\delta)\theta_{Q'}(t) + \delta\theta_Q(t)$$
$$\theta_{\mu'}(t+t_0) = (1-\delta)\theta_{\mu'}(t) + \delta\theta_\mu(t)$$

with t denoting time step, t_0 denoting update period and δ in [0, 1] denoting the averaging factor.

We also consider the exploration-exploitation policy by adding a stochastic noise onto the action output of DDPG agent at each time step, i.e., $\boldsymbol{a}_n(t) = \mu\left(\boldsymbol{s}_n(t)|\theta_\mu(t)\right) + \boldsymbol{v}(t)$. Note that each element of $\boldsymbol{a}_n(t)$ still needs to be limited within [0,1]. At each sample time step t, the noise value $\boldsymbol{v}(t)$ is updated using the following formula, where the initial value $\boldsymbol{v}(0)$ is defined as a zero vector $\boldsymbol{0}$,

$$\boldsymbol{v}(t+1) = \boldsymbol{v}(t) + \chi\left(\overline{\boldsymbol{v}} - v(t)\right) + \varepsilon(t)\boldsymbol{\omega}$$

where $\overline{\boldsymbol{v}}$ denotes the mean of $\boldsymbol{v}(t)$, the constant χ specifies how quickly the noise model output is attracted to the mean, $\varepsilon(t)$ is the standard deviation of $v(t)$ and $\boldsymbol{\omega}$ is a random vector satisfying the standard Gaussian distribution.

At each sample time step, the standard deviation decays as shown in the following code.

$$\varepsilon(t+1) = \varepsilon(t)(1-\epsilon)$$

with $0 < \epsilon \leq 1$ denoting the standard deviation decaying rate.

The critic network has two inputs, i.e., state input (SINRs) and action input (power allocation ratios) which have different orders of magnitudes. The power allocation ratios themselves are within [0,1]. Thus, we add a softmax layer after the state input to normalise them into the range of [0,1]. Similarly, we also add a softmax layer after the state input of the actor network. The output layer of the actor network is a sigmoid layer to ensure the power allocation vector $\boldsymbol{a}$ in the range [0,1].

3.2.3.3 Environment design

For the reward calculation of DRL environment, firstly determine the SIC decoding order in the order of decreasing arrived power of NOMA users. Secondly, calculate the SINRs (states) of NOMA users for each cluster by using SIC. Finally, we select the minimum SINR over users in each cluster *n* as the reward of the current iteration, i.e.,

$$r_n = \min\left(\boldsymbol{s}_n\right)$$

The DDPG based PA training process is given in Algorithm 2 with the discount factor β in [0,1]. Note that we have deleted the subscript *n* for corresponding notations where no ambiguities, such as writing $\boldsymbol{s}_n$ as $\boldsymbol{s}$.

Algorithm 2 DDPG based PA method

Randomly initialize critic and actor with $\theta_Q(0)$ and $\theta_\mu(0)$, respectively
Initialize target network: $\theta_{Q'}(0) = \theta_Q(0)$ and $\theta_{\mu'}(0) = \theta_\mu(0)$
Initialize replay buffer R and $t = 0$
for Episode $e = 1$ to E **do**
  for Step $b = 1$ to B **do**
    For observation $\boldsymbol{s}(t)$, select action $\boldsymbol{a}(t) = \mu(\boldsymbol{s}(t)|\theta_\mu(t)) + \boldsymbol{v}(t)$
    Execute action $\boldsymbol{a}(t)$. Observe the reward $r(t)$ and next observation $\boldsymbol{s}(t+1)$
    Store the experience $(\boldsymbol{s}(t), \boldsymbol{a}(t), r(t), \boldsymbol{s}(t+1))$ in the experience buffer R
    Sample a random minibatch of I transitions $(\boldsymbol{s}(u), \boldsymbol{a}(u), r(u), \boldsymbol{s}(u+1))$ from R
    Set $y(u) = r(u) + \beta(Q'(\boldsymbol{s}(u+1), \mu'(\boldsymbol{s}(u+1)|\theta_{\mu'}(u))|\theta_{Q'}(u)))$
    Update the critic with the loss: $L = \sum_u (y(u) - Q(\boldsymbol{s}(u), \boldsymbol{a}(u)|\theta_Q(u)))^2$
    Update the actor using the sampled policy gradient: $\Delta_{\theta_\mu} J = \frac{1}{I} \sum_u \Delta_{\boldsymbol{a}} Q(\boldsymbol{s}, \boldsymbol{a}|\theta_Q)|_{\boldsymbol{s}=\boldsymbol{s}(u), \boldsymbol{a}=\mu(\boldsymbol{s})} \Delta_{\theta_\mu} \mu(\boldsymbol{s}(u)|\theta_\mu)$
    Update $t = t + 1$
    Update $v(t)$ according to (20)
    if $\mathrm{mod}(t, t_0) = 0$ **then**
      Update the target networks with (18) and (19)
    end if
  end for
end for

3.2.4 Simulation implementation and testing

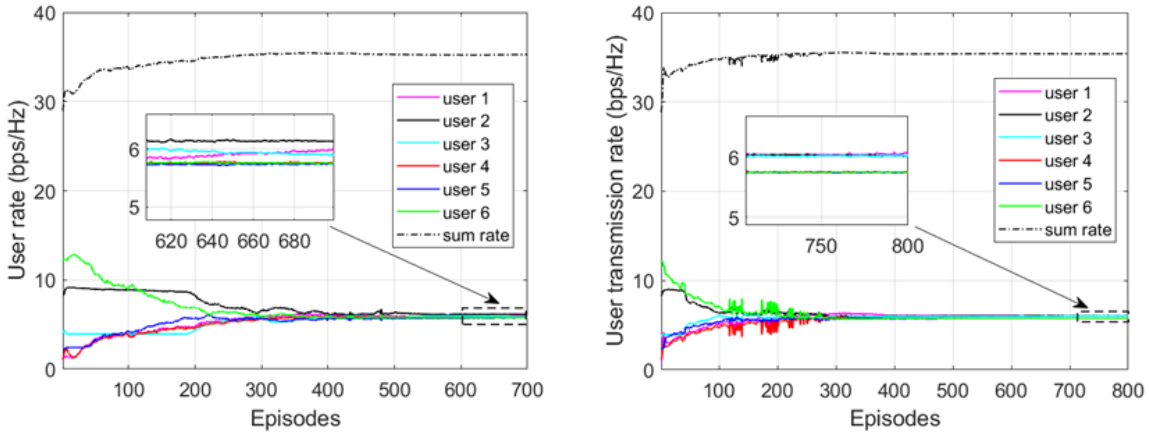
Without losing generality, we consider two D2D clusters in a cellular network that occupy the same frequency spectrum resource, as shown in Figure 17. There are three users in each D2D cluster, including one CUE and two EUEs. The DDPG agent consists of one critic and one actor network. The critic network has two inputs, i.e., state input and action input. As stated earlier, the SoftMax layer is used in both critic and actor network following their respective state input layers for normalisation. Similarly, the sigmoid layer is used as the output layer of the actor network to normalise the estimated power allocation parameters. Besides, the critic network has three fully-connected hidden layers i.e., $258 \times 128 \times 64$, with each followed by a leaky Elu activation layer. The actor network also has three hidden layers ($128 \times 64 \times 32$) followed by leaky Elu activation layers.

We consider the sub-6GHz communication herein. The available transmission powers of all users are assumed to be same, say 20dBm for example. The path loss (in dB) is characterized by the alpha-beta-gamma (ABG) model. Let d_1 and d_4 respectively denote the distance between the corresponding CUE and the BS, and d_2, d_3, d_5 and d_6 respectively denote the distance between the EUE with corresponding CUE. Assume the small-scale random channel fading follows independent but not identically distributed (i.n.d) Nakagami distribution with spreading and shape parameters *M=1* and $\Omega = 1$, respectively. With the receiver noise PSD and the noise figure n_f, the noise power is $p_n = \mathrm{PSD} \times B + n_f(dBm)$. Without loss of generality, the other simulation parameters are given in Table 3.

Table 3: The simulation parameters

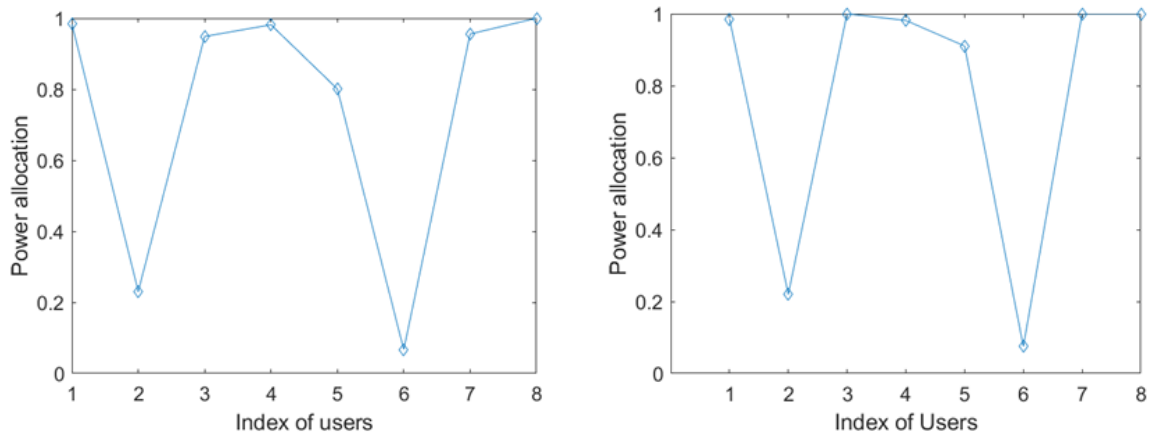
THE SIMULATION PARAMETERS

name	d_1(m)	d_2(m)	d_3(m)	d_4(m)	d_5(m)	d_6(m)
value	20.00	19.58	19.65	25.00	18.00	11.42
name	f(Hz)	B(Hz)	σ^2(dBm/Hz)	n_f(dBm)	$\mathcal{M}$	Ω
value	$5.8e9$	$20e6$	-174	10	1	1



(a) (b)

Figure 19: Learning curves of user rates using the DRL power allocation method: (a) single-agent; (b) multi-agent



(a) (b)

Figure 20: Power allocation using the DRL method: (a) single-agent DRL; (b) multi-agent DRL

As shown in Figure 19, both single-agent DRL and multi-agent DRL method converge within limited episodes. We also observe that the user rates in each cluster converge to a very similar value. This is because for the SIC decoding method, the performance improvement of one user usually implies the performance degradation of the other users until achieving the goal maximizing the worst-case user rate for each cluster. But, for the single agent DRL method, the worst-case user rate is computed over all user clusters and thus the minimizing the worst-case user rate works well for the cluster where the worst-case belongs. So, the user rates of that cluster converge to a similar value while the user rates in other clusters need only to be higher than the worst-case user rate of that cluster.

The higher η_n can lead to higher SINRs for all users in cluster *n*. So, the optimal values $\eta_{n,o}$, *n*=1,2 should be 1. Figure 20 shows the power allocation ratios $p_{n,k}$ for data transmission of each user and the total power allocation η_n for data forwarding of each cluster. The values of each subfigure represent $p_{1,1}$,$p_{1,2}$,$p_{1,3}$,$p_{2,1}$,$p_{2,2}$,$p_{2,3}$,η_1,η_2, respectively, from index 1 to 8. As shown in Figure 20 (a), we have $\eta_1 = 0.956$ which is slightly less than the optimal value 1 because the worst-case user is in cluster 2. But, for the multi-agent DRL in Figure 20 (b), every cluster minimizes its own worst-case user rate, so the values for η_n, *n*=1,2 reach the optimal value 1. We also observe from Figure 20 that the PA ratios of all CUEs ($p_{1,1}$ and $p_{2,1}$) are larger than 0.5, so the received power of the CUE at BS is larger than the received total power of the EUEs in any cluster. Besides, the channel gain between EUE 3 and CUE 1 is larger than that between EUE 2 and CUE 1 and the channel gain between EUE 5 and CUE 4 is larger than that between EUE 6 and CUE 4. We find from Figure 20 that when using zero-forcing beamforming, the PA ratios of EUEs follow $p_{1,3}$> $p_{1,2}$ and $p_{2,2}$2> $p_{2,3}$, i.e., more power allocated to users with higher channel gains for better SIC.

3.2.5 Conclusions and recommendation for future research

In this section, we considered the D2D relay enabled uplink cell-free communication system where the external user equipment accessed the cell BS through the cellular user relay. For effective decoding at the base station, we considered beamforming and a DDPG based power allocation method for worst-case user rate maximisation. Finally, a SIC decoding method was used at the base station based on the different arrived power strengths with given beamforming and power allocation parameters. The simulation results verified the effectiveness of the proposed DRL method for guaranteeing the user fairness through the worst-case rate maximisation.

For the future work, the energy harvesting device can be set up at the relay user to increase the energy efficiency and prolong its service life. In addition, the sum rate maximisation under given individual QoS constraints is effective for improving spectral efficiency. Finally, the aim of maximizing the ergodic rate is also considerable for reducing computational consumption and latency, where the power allocation and beamforming need to be calculated only once within a large-scale coherence time, especially in scenarios with high mobility or high frequency communication usually with a tiny small-scale coherence time.

4 Intelligent Beam Steering algorithm based on User Location

4.1 Autonomous Beam Steering

4.1.1 DRIDL algorithm definition

The Deep Learning Integrated Reinforcement Learning (DLIRL) algorithm is proposed for comprehending intelligent beam steering in Beyond Fifth Generation (B5G) networks as shown in the Figure 21 below [19]. The smart base stations in B5G networks aim to steer the beam towards appropriate user equipment based on the acquaintance of isotropic transmissions. The foremost methodology is to optimize beam direction through reinforcement learning that delivers significant improvement in Signal to Noise Ratio (SNR). This includes alternate path finding during path obstruction and steering the beam appropriately between smart base station and user equipment. The DLIRL is realized through supervised learning with deep neural networks and deep Q learning schemes. The proposed algorithm comprises of an online learning phase for training the weights and a working phase for carrying out the prediction. Results confirm that the performance of the B5G system is improved considerably as compared to its counterparts with a spectral efficiency of 11 bps/Hz at SNR=10 dB for a bit error rate performance of 10^{-5}. As compared to reinforced learning and deep neural network with a deviation of ±3° and ±5°, respectively, the DLIRL beamforming display a deviation of ±2°. Moreover, the DLIRL can track the user equipment and steer the beam in its direction with an accuracy of 92%.

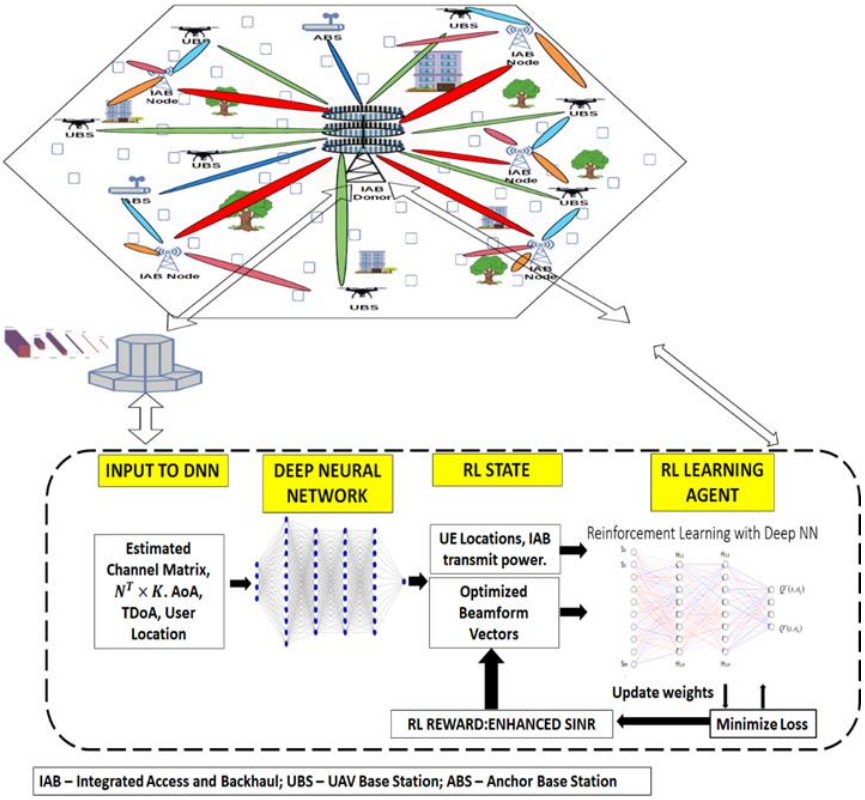


Figure 21: DLIRL for Intelligent Beam Steering

4.1.2 Simulation Implementation and Testing

The beamforming is implemented for the downlink scenario of the data transmission from IAB to UE. The simulation environment is first configured and then the proposed DLIRL beamforming is executed. The simulation parameters employed in this work are as shown in Table 4. The data set of the channel model and the channel parameters for the simulation is generated via MATAB 2021 site viewer-based ray tracing. The Shooting and Bouncing Ray (SBR) based ray-tracing model is employed for the LoS and NLoS communication. The Orthogonal Frequency Division Multiplexing (OFDM) system is employed for symbols transmission. The considered OFDM size is 1024. The DNN architecture has a total of 6 interconnected layers including 4 hidden layers and 1 input and an output layer. The DNN has a total I*k number of inputs and Train number of outputs. The considered data has a set size of two hundred thousand samples and a batch size of two hundred. To have comparative analysis of the proposed algorithm with the existing conventional beamforming techniques, we have used SE and the BER as the metrics. Figure 22 and Figure 23 show the SE for different SNR values received at UE. The simulations were carried out for 30 runs comprising 1000 iterations each. The depicted graph values are averaged values obtained in the simulation environment. For the simulation environment, the IABs are installed on the buildings played in the x-y plane of the 3D environment. The IAB's antennae are facing the street on the y-z plane. The antenna transmit power is considered at 30 dBm. The UEs are mobile and are installed with a single antenna. For each beam coherence time, the UE locations are updated in the x-y plane. During the training period, the UE uplink transmit power is set at 30 dBm.

Table 4: Simulation Parameters

S.No	Parameter	Specification
1	IABs/IAB count (N)	4
2	IAB Antenna array	Uniform Planar Array
3	IAB Antenna Specification	32x8
4	User Equipment (UE) setup	Deployed in a rectangular grid of dimension 40mx60m, resolution 0.1m.
5	DNN Activation Unit	ReLU (Rectified Linear Unit)
6	DNN dropout rate	0.5%
7	DNN batch size	100
8	Python Libraries	Keras with Tensorflow backend
9	System Bandwidth	0.5 GHz
10	OFDM Subcarriers	1024
11	Sampling Factor	1
12	Multipaths	7

Figure 22 visualizes that the DLIRL beamforming has achieved better spectral efficiency as compared to the existing conventional beamformers in [19]. As seen from the curves, the spectral efficiency with analogue beamforming is found to be around 2 bps/Hz, and close

convergence is observed between ZF hybrid precoding, MMSE hybrid precoding, and Kalman hybrid precoding techniques. However, the MSE based fully digital precoding displays improved spectral efficiency as compared to the above-mentioned precoding techniques. For an SNR of 5 dB, the DLIRL based beamforming technique displays an improvement of 77.5%, 60%, 50%, 50%, and 33.3% as compared to the analogue beamforming, ZF hybrid precoding, MMSE hybrid precoding, Kalman hybrid precoding, and MSE based fully digital precoding techniques, respectively. The spectral efficiency is achieved for the multipath scenario considering both LoS and NLoS, total multipath considered for evaluation of Figure 22 is 10 and total IAB antenna elements are 256.

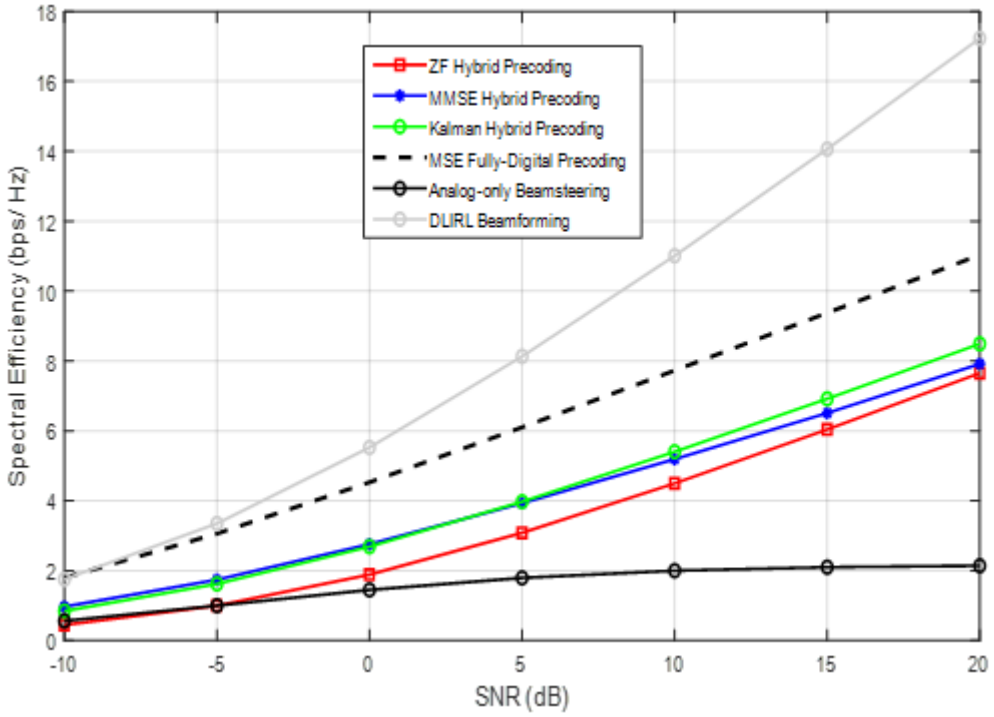


Figure 22: Comparative analysis of SE with reference to IAB SNR

Moreover, as the antenna size increases the performance of DLIRL gets better as compared to the DNN and RL beamformer. Figure 23 shows the comparison between DNN, RL, and DLIRL-based beamformers for different transmitting antenna elements. For instance, for the number of IAB antenna elements equal to 104, the increase in spectral efficiency employing DLIRL-based beamforming is found to be 53.33% and 51.66% more efficient as compared to DNN and RL based beamforming techniques, respectively. The effect of BER for IAB with 4 transmit antenna elements has been displayed in Figure 24. Here, the performance of the system for different MIMO schemes is compared with DLIRL based beamforming scheme. For a BER of 10^{-4}, the proposed DLIRL based beamforming techniques require an Eb/N0 of 7 dB. Alternatively, the system without diversity, Alamouti, and OSTBC schemes require an Eb/N_0 of 10 dB, 13.3 dB, and beyond 20 dB, respectively.

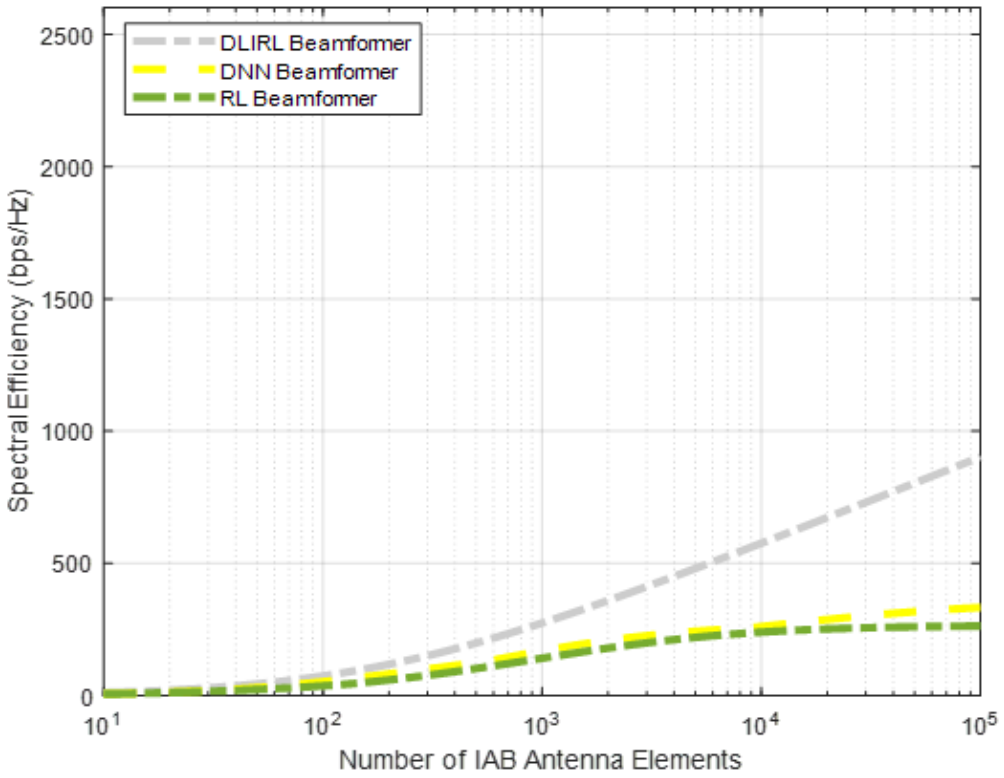


Figure 23: Comparative analysis of SE with reference to increase in number of antennas

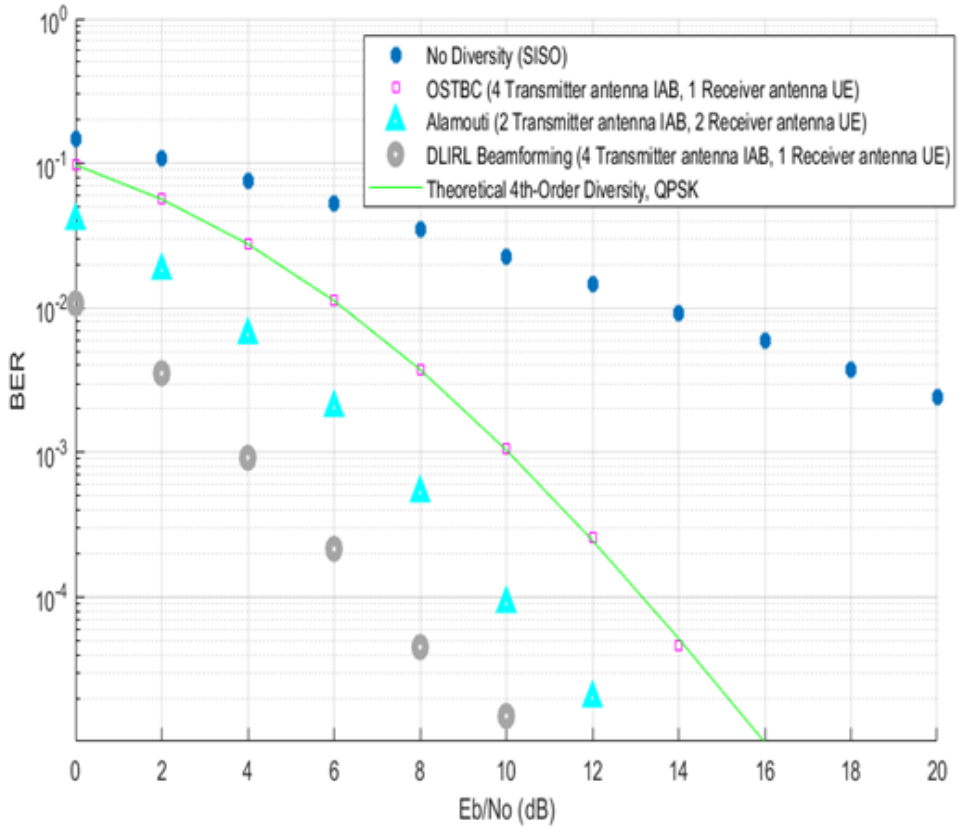


Figure 24: Comparative Analysis using BER

Table 5: SE Performance Evaluation

Beamforming Technique	SE (bps/Hz) at SNR=10 dB
DLIRL Beamforming	11
Analog Beamforming	5
MSE Digital Beamforming	8
Kalman-hybrid precoding	5.7
Minimum Mean Squared Error (MMSE) Hybrid Beamforming	5.1
Zero Forcing Hybrid Beamforming	4.2

The quantitative analysis of the proposed scheme concerning SE and BER is presented in Table 5 and Table 6 respectively. There is a drastic improvement in the SE and BER using DLIRL Beamforming studied at SNR=10 dB shown in both tables.

Table 6: BER Performance Evaluation

Diversity Scheme	BER at SNR=10 dB
No Diversity, Single Input Single Output (SISO)	$10^{-1.8}$
Orthogonal Space Time Block Coding (OSTBC), 1x4 MIMO Transmit Diversity	10^{-3}
Alamouti, 2x2 MIMO Diversity	10^{-4}
DLIRL Beamforming, 1x4 Transmit Diversity	10^{-5}

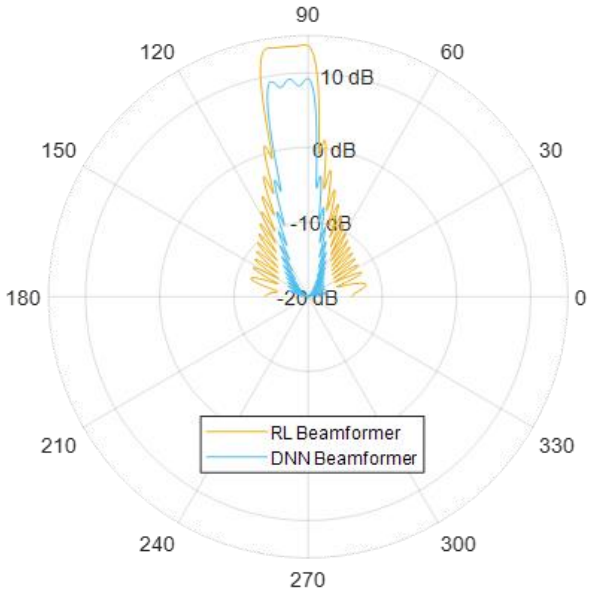


Figure 25: Beamform towards UE using DNN and RL separately

To answer the question of whether the proposed scheme can learn the beamforming, we have simulated Figure 25 and Figure 26. The simulation graphs in Figure 25 and Figure 26 are also the report of the average value for the 30 runs with 1000 iterations each. The proposed scheme can project the beam towards the UE positioned at 100 degrees in the northwest direction of the antenna placement. From Figure 25 and Figure 26, it is estimated that the DLIRL based beamformer is better than its counterparts DNN and RL in steering the beam towards the UE placed at 101.5° normal to the antenna placement of IAB. The proposed DLIRL beamforming has Angle of Departure (AoD) towards UE location with a deviation of ±2°, whereas RL has a deviation of ±3° and DNN's deviation is ±5°.

The DLIRL is capable of performing efficient beamforming due to the effective training. It is vital to have comparative analysis of the DLIRL with existing DNN and RL algorithm in terms of training validation accuracy, training loss, number of iterations and epochs. Figure 25 sheds light on the validation accuracy of the proposed (DLIRL) and existing (DNN and RL) training algorithms. For the training we employed 20 epochs, 160 iterations, and 100 runs. Each run comprised 20 epochs and each epoch had 8 iterations. From the validation accuracy as shown in Figure 26 it can be inferred that the proposed DLIRL due to its optimized amalgamation of DNN and RL has better training accuracy as compared to the DNN and RL. These training accuracy results are clearly reflected in the beamforming effectiveness as shown in Figure 25 and Figure 26.

The proposed DLIRL is effective in getting trained in a smaller number of samples as shown in Figure 26. The SE of DLIRL is comparable to DNN and RL for very few samples (100). Above the 100 training samples the performance of the DLIRL is better than its counterparts.

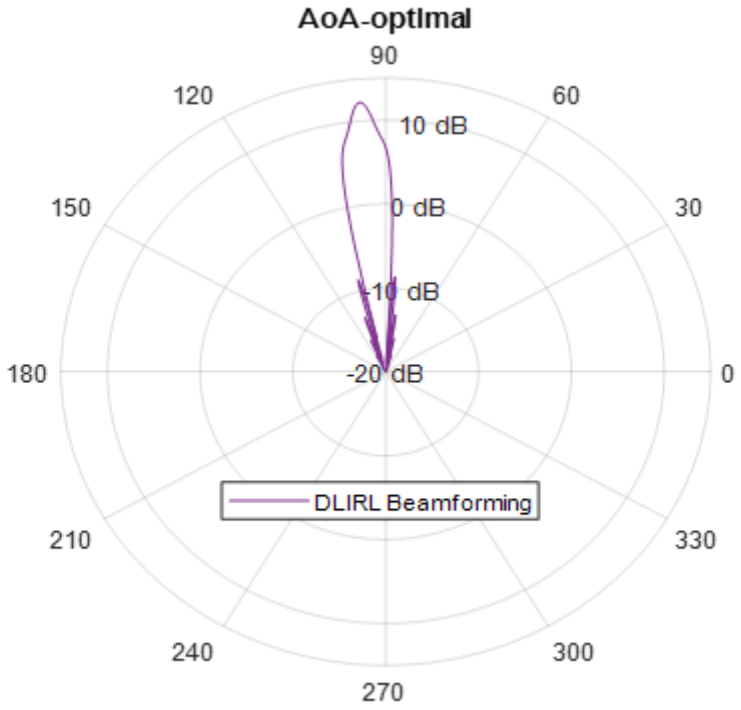


Figure 26: Beamform towards UE using DLIRL

4.2 Orthogonal Time Frequency Space (OTFS) Sensing of Distance

4.2.1 Definition

The idea of OTFS is reduce the disadvantage associated with the OFDM and enhance its pros. OTFS (Orthogonal Time Frequency Space) modulates each information (e.g., QAM) symbol onto one of a set of two dimensional (2D) orthogonal basis functions that span the bandwidth and time duration of the transmission burst or packet [20]. The modulation basis function set is specifically derived to directly represent the dynamics of the time- varying multipath channel. OTFS can be implemented as a pre- and post-processing block to filtered OFDM systems, thus enabling architectural compatibility with LTE. OTFS transforms the time-varying multipath channel into a time-independent two-dimensional channel in the Delay-Doppler domain that directly represents the geometry of the various reflectors composing the wireless link. In this way, OTFS eliminates the difficulties in tracking time-varying fading, particularly in high-speed vehicle communications. Due to its ability to extract the full diversity of channel across time and frequency, OTFS enables linear scaling of throughput with the number of antennas in moving vehicle applications. In addition, since the Delay-Doppler channel representation is very compact, OTFS enables dense and flexible packing of reference signals, a key requirement to support the large antenna arrays used in massive MIMO applications.

4.2.2 Simulation Implementation and Testing

OTFS works in delay Doppler domain rather than time-frequency domain. The delay Doppler domain representation of channel converts time-variant channel to time-invariant channel as shown in Figure 27 below.

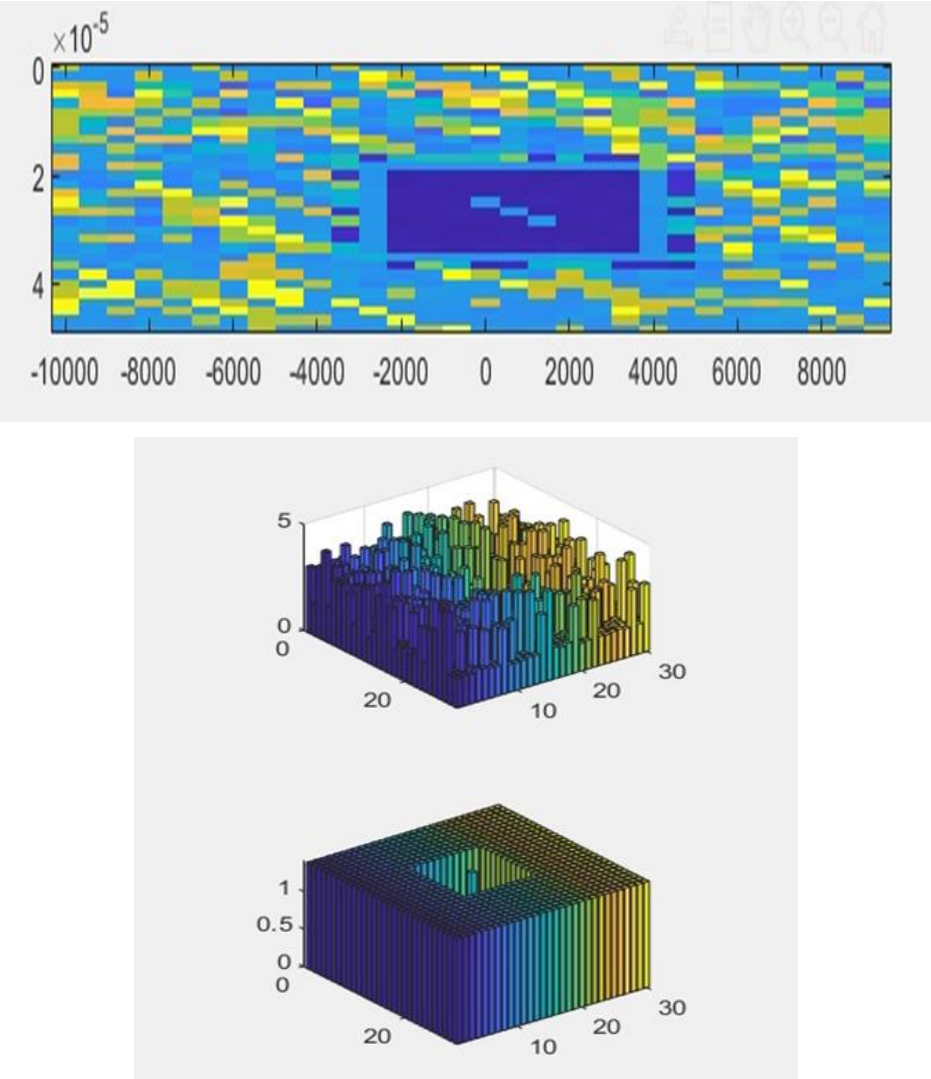


Figure 27: a. OTFS channel images plot, b. OTFS channel impulse response

BER analysis for the OTFS and OFDM modulated signals: We can visualize from the comparative analysis of the OTFS and OFDM modulated signal for fast moving user equipment. The BER for OTFS modulated signals is high because its performance is not degraded even with the fast motion of the user equipment and the bit error and bits loss is less as compared to the conventional OFDM (see Figure 28).

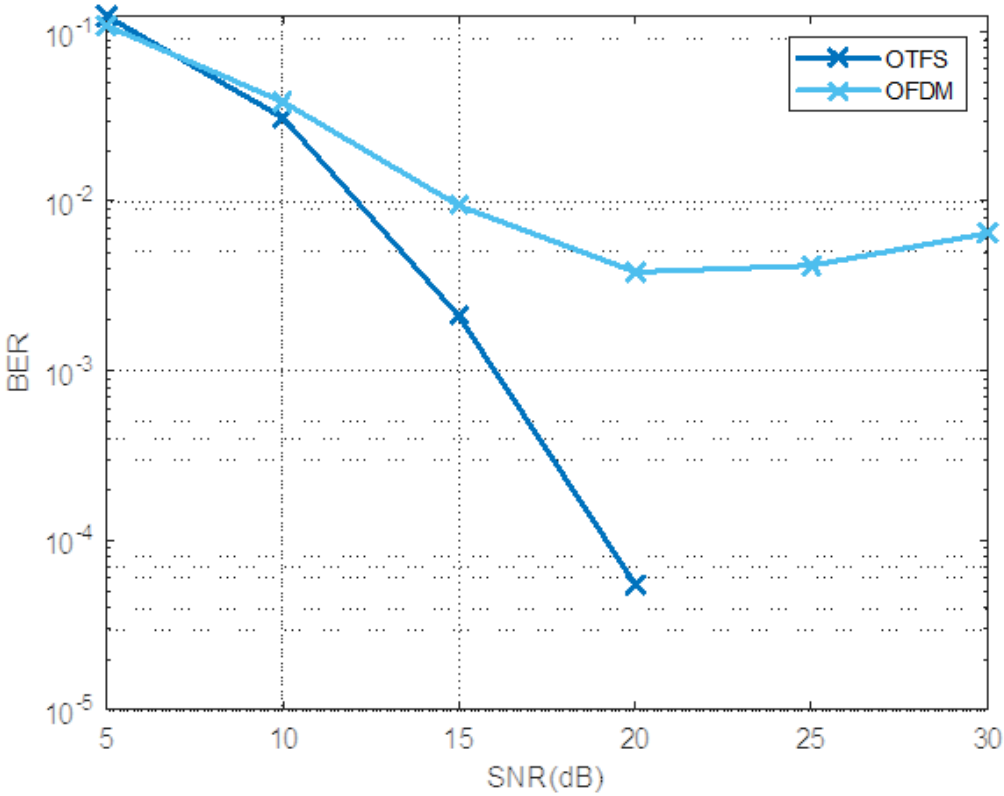


Figure 28: BER analysis of OTFS and OFDM modulated signal at high Speed (Doppler Frequencies)

4.3 Conclusions and recommendation for future research

OTFS is suitable for serving radio channels from fast moving user equipment units with high Doppler Frequencies and may also be suitable for measuring location from Angle of Arrival using its impulse response and distance by using its delay.

Future research involves exploring how OTFS can be used for measuring location from Angle of Arrival using its impulse response and distance using its delay.

5 Advanced Test and Simulation Tools supporting 6G BRAINS research

Following the development of cell free IAB scheduler and radio resource allocation algorithms by partners, our main focus in this section is to develop advanced test and simulation tools for multi-UE mobility behaviour modelling and the corresponding dynamic channel state information (CSI) evaluation and analysis. The tools allow to configure up to thousand UEs' mobile behaviour and capture their CSI logs. The captured CSI logs, including UE location, signal to noise radio (SNR), channel quality indicator (CQI), precoding matrix indicator (PMI), rank indicator (RI), etc, have potential to be used as the input of the above developed IAB scheduler and radio resource allocation algorithms for AI model training.

5.1 Definition

The developed test and simulation tools are mainly for cellular downlink communications. The development for uplink communications support is planned for the next step research. Figure 29 provides the brief description of the considered system model.

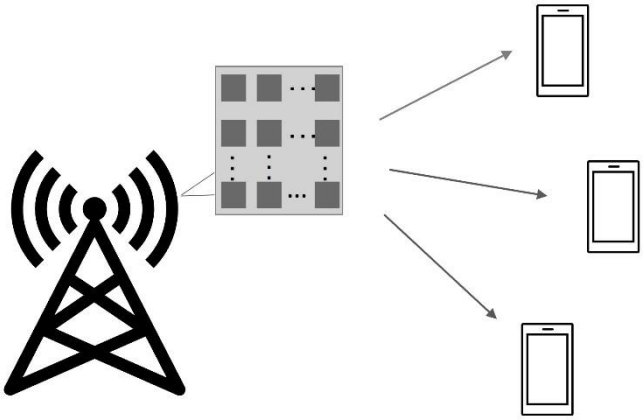


Figure 29: System model

As shown in Figure 29, the base station (BS) is equipped with N transmit antenna elements, which are arranged in a $N_1 \times N_2$ size panel of cross-polarised antenna pairs. Hence, $N = 2N_1N_2$, where N_1 is the number of columns, N_2 is the number of rows and 2 is the number of polarisations. The set of precoding matrices, also known as codebook, are set up and each precoding matrix is formed from basic beamforming vectors. Each beamforming vector is a discrete Fourier transform (DFT) vector of length N_1N_2 with oversampling factors O_1 and O_2, where $O_1 = 4$ and

$$O_2 = \begin{cases} 1 \ \ if \ N_2 = 1; \\ 4 \ \ otherwise. \end{cases}$$

Thus, for the above system, there are $K = O_1O_2N_1N_2$ beamforming vectors in total. $O_1O_2N_1N_2$ beams point to $O_1O_2N_1N_2$ distinct directions. Depend on antenna structure at BS antenna panel, the angle width in azimuth and zenith dimensions, such as in the bounded area of directions, are not lower than the gain of the boresight direction subtract 3dB, which is also called 3dB azimuth and zenith beamwidths. The 3dB azimuth bandwidth, denoted by $\emptyset_{3\text{dB}}$, and 3dB zenith beamwidth denoted by $\theta_{3\text{dB}}$ are the model parameters specified by end users. Furthermore, the gain to a direction is a composition of two factors, the gain due to directional antenna elements and the gain from a beamforming technique. This composite gain is represented by a single input by end users, G, which takes the value of $10\log_{10}(N_1N_2)$dB by

default. Meanwhile, the model also takes the spherical coordinate of the UE, $(r, \emptyset, \theta)$, as the model input, where r is the distance between the BS and the UE, $\emptyset$ is the UE azimuth angle, and θ is the UE zenith angle. Provided that the Cartesian coordinate of the UE relative to the BS is (x, y, z), we have

$$r = \sqrt{x^2 + y^2 + z^2};$$

$$\emptyset = \arg(x + jy);$$

$$\theta = \arccos\left(\frac{z}{\sqrt{x^2+y^2+z^2}}\right),$$

where $\arg(\cdot)$ denotes the argument (phase) of a complex number and $\arccos(\cdot)$ is the inverse cosine function.

The developed model models the large scale (i.e., slow fading pathloss) and small scale (i.e., multipath fast fading) radio propagation effects. Slow fading pathloss model of each link follows the simple free-space pathloss (FSPL) model, where the pathloss exponent is set to 2 and no shadowing effect is considered. The reason we only support pathloss exponent of 2 is because there is little benefit/effect for the modelling feature. Meanwhile, we do not introduce shadowing, aiming to avoid difficulty in a case that one wants to test a behaviour requiring a deterministic outcome. Fast fading, which is based on a stochastic model, should bring in sufficient degree of uncertainty for a scattering environment, and it will be used to reflect through the CSI and hybrid automatic repeat request (HARQ) acknowledgement (ACK)/negative-acknowledgement (NACK) feedback reporting sent to the BS. In combination with the channel SNR, channel gain coefficients are used to generate instantaneous SNR, which is then used directly to map to RI, CQI and used to populate HARQ ACK/NACK.

We assume that mobility model always has sufficient line-of-sight (LoS) and believe that non-LoS (NLoS) is more benefited to the signal processing designs at the receiver. On the other hand, in carrying out a system test in order to see the performance of the whole network assisted by a beamforming technique, NLoS might cause cumbersome. Without LoS path, the desirable beams could be any beams making our mobility model unnecessarily complicated in implementation. In addition, we limit our model to support only the cases for which the LoS power is not lower than NLoS power. In representing the LoS-to-NLoS power ratio by a parameter K, provided by user, the model requires that *K* must not be less than 0dB. With this assumption, instead of scanning over all possible beams, a search in the neighbours of LoS beam is sufficient to find the best beams, and hence the most suitable precoding matrices.

In our model, fast fading component adopts the tapped delay line (TDL) models following the 3GPP specification. Each TDL model is associated with a channel delay profile, namely TDL-A, TDL-B, TDL-C, TDL-D or TDL-E. When a channel delay profile is specified to a link between the BS antenna and a UE's antennas, the channel is time dispersive and its impulse response has several random components, each is a Jakes process, at different delays. The variances of those random components are allocated according to the delay profile and the sum of variances is *K*dB lower than the deterministic component cause by LoS path. Since each channel delay profile specified in the 3GPP standard only defines the powers at normalised delays, a parameter called delay scale needs to be set to derive the actual delays for all the taps.

5.2 Simulation Implementation and Testing

The set of input parameters of the developed mobility model is given in Table 7. In this table, the system parameters are considered as parameters applied for all the UEs and expected not to change during a test. The link parameters, which can be changed during a test, are UE specific. For example, when UE follows a moving path defined before a test run, the mobility model will produce the outcomes for discrete values of distance r and direction $\emptyset, \theta$ along the path. Considering the configuration of a UE's speed and direction, and corresponding Doppler frequency can be derived from the geometric model. Specifically, suppose that a UE is in a path from the position (x, y, z) to the position $(\tilde{x}, \tilde{y}, \tilde{z})$ with a constant speed v, the Doppler frequency is given as

$$f_D = \frac{x(\tilde{x}-x) + y(\tilde{y}-y) + z(\tilde{z}-z)}{\sqrt{(x^2+y^2+z^2)[(\tilde{x}-x)^2 + (\tilde{y}-y)^2 + (\tilde{z}-z)^2]}} \frac{vf_c}{c},$$

Where $c = 3 \times 10^8$m/s is the speed of radio propagation. Note that $f_D \leq \frac{vf_c}{c}$ because

$$\frac{x(\tilde{x}-x) + y(\tilde{y}-y) + z(\tilde{z}-z)}{\sqrt{(x^2+y^2+z^2)[(\tilde{x}-x)^2 + (\tilde{y}-y)^2 + (\tilde{z}-z)^2]}} \leq 1,$$

in which the equality holds only if $\frac{\tilde{x}-x}{x} = \frac{\tilde{y}-y}{y} = \frac{\tilde{z}-z}{z}$, i.e., the moving path is on the LoS.

Table 7: Input parameters for the developed mobility model

Parameter Name	Notation	Required Constraint	Default Value
System Parameters			
transmit power (in dB/RE)	P_{TX}		
noise power at UEs (in dB/RE)	I_{oc}		
MCS Table		{64- QAM, 256-QAM}	
carrier frequency	f_c		
numerology	μ	{0, 1, 2, 3}	
system bandwidth [2]	N_{RB}		
antenna panel width	N_1		
antenna panel height	N_2		
antenna plus beamforming gain	G		$10\log_{10}(N_1 N_2)$ dB
CSI azimuth beamwidth	$\phi_{3\text{dB}}$	$0 < \phi_{3\text{dB}} < 180^o$	$\frac{180^o}{N_1}$
CSI zenith beamwidth	$\theta_{3\text{dB}}$	$0 < \theta_{3\text{dB}} < 180^o$	$\frac{180^o}{N_2}$
Link Parameters			
channel delay profile		{TDL-A, TDL-B, TDL-C, TDL-D, TDL-E}	
delay scale	DS	{10ns ,30ns, 100ns, 300ns, 1000ns}	
gNB-UE distance	r		
UE azimuth angle	ϕ		
UE zenith angle	θ		
Doppler frequency	f_D		
LoS-to-NLoS power ratio	K	$K[\text{dB}] \geq 0$ $K = +\infty$ if $f_D = 0$	

From the input parameters as given in Table 7, the mobility model will run online and produce a series of instantaneous CSI reports based on the user's configuration, which include UE's location, UE specified RI, wideband PMI, wideband and sub-band CQI, and HARQ ACK/NACK status when requested.

As mentioned above, CSI reports are triggered by the calculation of SNR of each UE at specific time instance. The way to calculate SNR is the combination of the SNRs for both slow fading model and the fast fading model. For the slow fading model, given the direction $(\emptyset, \theta)$ of a UE, the LoS beam will be taken into account for the SNR calculation and is defined as the beam that maximizes

$$A_{i_{1,1}i_{1,2}}(\emptyset, \theta) = -\min\left\{-\left(A_{H,\emptyset_0}(\emptyset) + A_{V,\theta_0}(\theta)\right), 30\right\},$$

were

$$A_{H,\emptyset_0}(\emptyset) = -\min\left\{12\left(\frac{\emptyset - \emptyset_0}{\emptyset_{3\text{dB}}}\right)^2, 30\right\};$$

$$A_{H,\theta_0}(\theta) = -\min\left\{12\left(\frac{\theta - \theta_0}{\theta_{3\text{dB}}}\right)^2, 30\right\};$$

and $i_{1,1}$ with the size of $[0, O_1N_1 - 1]$ and $i_{1,2}$ with the size of $[0, O_2N_2 - 1]$ are the pair of beam indices. It is worth noting that the directions of all the beams with the same $i_{1,2}$ will have an identical zenith angle. Following the standard FSPL model, the pathloss of slow fading is given by

$$L(r)[\text{dB}] = 35.25 + 20\log_{10}(f_c) + 20\log_{10}(r).$$

In addition, the isotropic SNR is defined as

$$\overline{\text{SNR}}[dBc] = P_{TX} - L(r) - I_{oc},$$

where P_{TX} is the transmit power and I_{oc} is the noise power at UEs. Thus, the slow fading SNR conditioned on a specific beam with the indices $(i_{1,1}, i_{1,2})$ is

$$\overline{\text{SNR}}_{i_{1,1},i_{1,2}}[dBc] = \overline{\text{SNR}} + G + A_{i_{1,1}i_{1,2}}(\emptyset, \theta).$$

Note that the SNR conditioned on LoS beam must be the highest one between all the conditional SNR as we mentioned before.

For the fast fading, the channel condition varies across both the time and frequency, and Doppler frequency will affect the frequency of channel varying with time. The actual values of CSI reports and the generated HARQ ACK/NACK outcomes are the results of mappings from the instantaneous SNR of the transmission layers as we mentioned above. Assume that the signals of two polarisations go through two independent paths. The following form of row-vector channel can be used to model each path in an resource block (RB) is given by

$$\mathbf{h} = \sqrt{\frac{K}{K+1}}\mathbf{h}_{\text{LoS}} + \sqrt{\frac{1}{K+1}}\mathbf{h}_{\text{NLoS}},$$

in which, the first component at the right-hand-side is deterministic LoS and the second component at the right-hand-side is random NLoS. The model in the above equation allows to model the channel quality for the $(i, j)^{th}$ beam at time t for the k^{th} RB as

$$\mathrm{SNR}_{i,j,k}(t) = \overline{\mathrm{SNR}}_{i_{1,1},i_{1,2}} \left| \sqrt{\frac{K}{K+1}} + \sqrt{\frac{1}{K+1}} x_{i,j,k}(t) \right|^2,$$

where $\overline{\mathrm{SNR}}_{i_{1,1},i_{1,2}}$ is the slow fading SNR defined above; $x_{i,j,k}(t)$ is a standard Gaussian random variable for $i \in \{1,2\}$ polarisation index and $j \in \{1,2,\cdots,7\}$ beam index. Note that, as strong LoS is assumed, for generating quantities for CSI reporting, the developed mobility model only considers instantaneous the LoS beam and its 6 neighbouring beams.

The model generates the above instantaneous SNR for 2 polarisations and 7 beams in every time slot and reports CSI, consisting of RI, PMI, CQI etc. To decide RI and wideband PMI, given the last slot with time index t, we use the average SNR across the whole RBs, which is given by

$$\widehat{\mathrm{SNR}}_{i,j}(t) = \sum_{k=1}^{N_{RB}} \mathrm{SNR}_{i,j,k}(t),$$

where N_{RB} denotes the number of RBs for transmission. Depending on the configured modulation and coding scheme (MCS) table, CQI can be linked to the calculated SNR value based on some pre-defined lookup table. Denote that $\gamma(\cdot)$ as the function mapping from a SNR to CQI and $\varphi(\cdot)$ as the function mapping from a CQI to a spectrum efficiency. For rank-1 hypothesis, the spectrum efficiency is given by

$$\varphi_{rank-1} = \max_j \varphi\left[\gamma\left(\widehat{\mathrm{SNR}}_{1,j} + \widehat{\mathrm{SNR}}_{2,j}\right)\right],$$

and for rank-2 hypothesis, the spectrum efficiency is given by

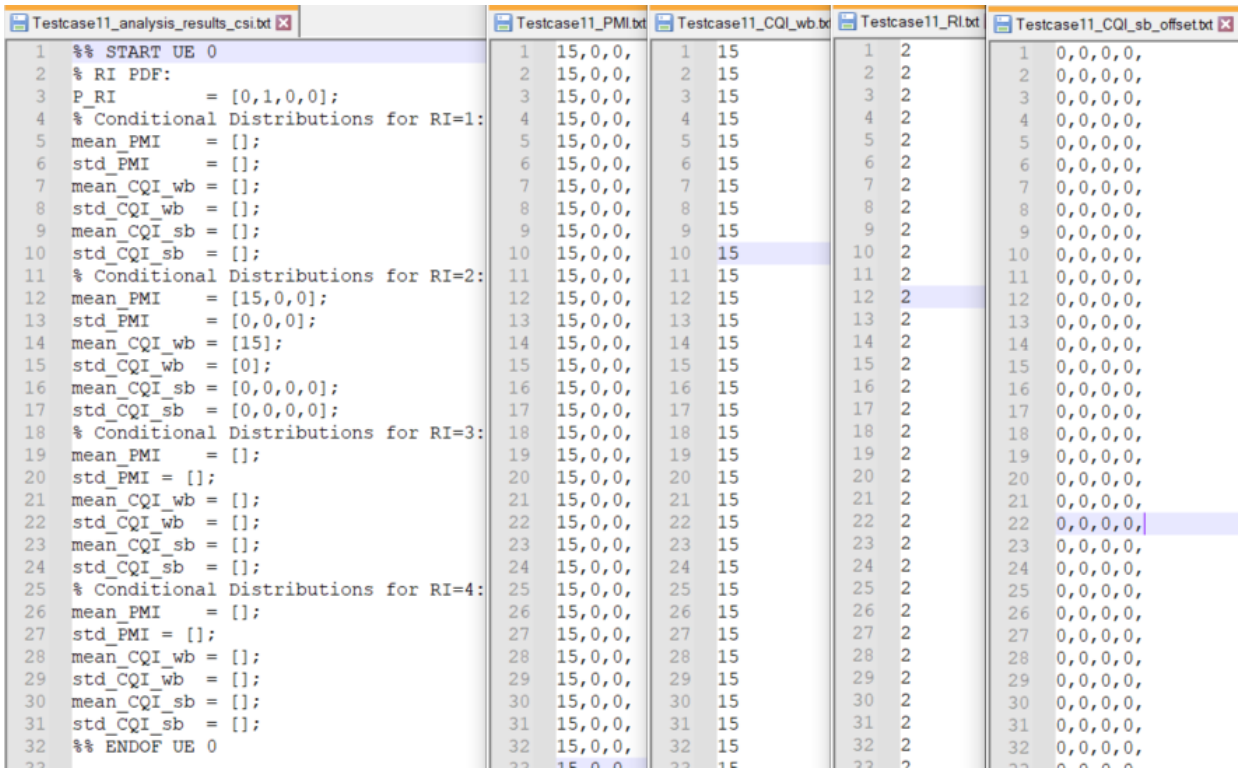
$$\varphi_{rank-2} = \max\left\{2\varphi\left[\gamma \min\left(\widehat{\mathrm{SNR}}_{1,1}, \alpha_2\right)\right], 2\varphi\left[\gamma \min\left(\widehat{\mathrm{SNR}}_{2,1}, \alpha_1\right)\right]\right\},$$

where

$$\alpha_1 = \max_j \widehat{\mathrm{SNR}}_{1,j},$$

$$\alpha_2 = \max_j \widehat{\mathrm{SNR}}_{2,j}.$$

Same strategy is applied to calculate higher rank efficiency if UE is equipped with higher number of antennas. The highest efficiency between different rank hypotheses will decide RI. Following that, the derivation of PMI, i.e., $i_{1,1}$, $i_{1,2}$ and CQI value can also be determined. Figure 30 shows an example of the simulated CSI report with a randomly configured system. For the next step research, we will collaborate with project partners to design and train AI/ML model for radio resource allocation, leveraging the developed CSI reporting tools.



Testcase11_analysis_results_csi.txt

```
%% START UE 0
% RI PDF:
P_RI        = [0,1,0,0];
% Conditional Distributions for RI=1:
mean_PMI    = [];
std_PMI     = [];
mean_CQI_wb = [];
std_CQI_wb  = [];
mean_CQI_sb = [];
std_CQI_sb  = [];
% Conditional Distributions for RI=2:
mean_PMI    = [15,0,0];
std_PMI     = [0,0,0];
mean_CQI_wb = [15];
std_CQI_wb  = [0];
mean_CQI_sb = [0,0,0,0];
std_CQI_sb  = [0,0,0,0];
% Conditional Distributions for RI=3:
mean_PMI    = [];
std_PMI = [];
mean_CQI_wb = [];
std_CQI_wb  = [];
mean_CQI_sb = [];
std_CQI_sb  = [];
% Conditional Distributions for RI=4:
mean_PMI    = [];
std_PMI = [];
mean_CQI_wb = [];
std_CQI_wb  = [];
mean_CQI_sb = [];
std_CQI_sb  = [];
%% ENDOF UE 0
```

Testcase11_PMI.txt

```
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
15,0,0,
```

Testcase11_CQI_wb.txt

```
15
15
15
15
15
15
15
15
15
15
15
15
15
15
15
15
15
15
15
15
15
15
15
15
15
15
15
15
15
15
15
15
```

Testcase11_RI.txt

```
2
2
2
2
2
2
2
2
2
2
2
2
2
2
2
2
2
2
2
2
2
2
2
2
2
2
2
2
2
2
2
2
```

Testcase11_CQI_sb_offset.txt

```
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
0,0,0,0,
```

Figure 30: An example of CSI reporting results

6 Summary Conclusions and Recommendations

6.1 Deliverable Summary

These deliverable details the modelling and analysis of the intelligent IAB and intelligent UE. It is now composed of four parts, that is:

1. the intelligent cell-free IAB scheduler for spectrum allocation and traffic routing,
2. the intelligent UE modelling and decoding methods for D2D enabled cooperative network and the grant-free access network,
3. the location sensing-based intelligent beam scheduler and
4. the advanced test and simulation tools for multi-UE mobility behaviours modelling and channel evaluation.

The main content of the deliverable is described below.

1. The cell free integrated access and backhaul scheduler is subdivided into two parts, namely: (1) IAB bandwidth allocation, (2) Routing solution. The IAB resource allocation finds the optimal way to divide the spectrum between the backhaul and access requirements of the different donors and the nodes in the network using supervised learning AI method with CQI, DL and UL profile and connected base-station for input parameters and the rate each link must support and its efficiency for cost function. The routing solution, that is described in this deliverable in details. is based on multi-agent deep reinforcement learning (MA-DRL) for a fully synchronized time-slotted wireless network with the objective to find the optimal route from each BS (donor or nodes) to each user in terms of: packet error probability (PER) for the whole packet trajectory; maintenance of quality of service (QoS) requirements; network congestion management including queue management and fairness. The Relational Actor and Critic neural architecture is proposed as the best routing solution.

2. The D2D enabled cooperative network consists of the D2D clusters and the IAB node. Each D2D clusters may contain a couple of far users and a near user where the far users directly transmit data to the near user. The near user acts a D2D relay and forwards the received signal and transmit its own signal to the IAB node, respectively. For effective decoding at the base station, beamforming and a DDPG based power allocation method for worst-case user rate maximization are employed. Finally, a SIC decoding method is used at the base station based on the different arrived power strengths resulted from the optimized beamforming and power allocation parameters.

   For the grant-free access network for mMTC, a block-sparsity-based adaptive matching pursuit algorithm is proposed for the joint user activity detection and signal recovery. The proposed method utilizes a novel user sparsity decision method with only the modulation constellation of the transmitted signal as the prior information, enabling its practicability.

3. OTFS transforms the time-varying multipath channel into a time-independent two-dimensional channel in the delay-Doppler domain that directly represents the geometry of the various reflectors composing the wireless link.

   For the location sensing-based intelligent beam scheduler, OTFS is suitable for radio channels from fast moving user equipment with high Doppler frequencies and may

also be suitable for measuring location from angle of arrival using its impulse response and distance using its delay.

4. The advanced test and simulation tools are developed for multi-UE mobility behaviour modelling and the corresponding dynamic channel state information (CSI) evaluation and analysis. The tools allow to configure up to thousand UEs' mobile behaviour and capture their CSI logs. The captured CSI logs, including UE location, SNR, CQI, PMI, RI, etc, have potential to be used as the input of the above developed IAB scheduler and radio resource allocation algorithms for AI model training.

6.2 Future work plans

Following the implementation of the IAB scheduler, the next task will be the replacement of a traditional deterministic scheduler of a 5G network with this AI based scheduler and test how the main network KPIs such as end-to-end latency, reliability, throughput, spectral efficiency, power consumption, QoS, QoE, etc. are affected by the AI based scheduler.

Furthermore, conduct investigation on how the AI based scheduler and the D2D cluster scheduler can be integrated with each other to obtain better overall Network performance.

The cell-free MIMO with D2D UE clusters is planned as the future work, including the user clustering, the access point clustering, the spectrum allocation, the intelligent beamforming and power allocation. The deep reinforcement learning method will be used for solving the highly nonlinear optimization problem and well adapting to the varying channels due to the dynamic clustering and the mobility of the users.

Towards human-centric control interfaces for cellular networks, future work includes the implementation of all the knowledge applied to the presented state machine, validating all defined phases with the presented use case, and completing the intents implementation system and move on to the integration of this system with the voice recognition system.

In the context of intelligent IABs with beam steering based on user location, OTFS transforms the time-varying multipath channel into a time-independent two-dimensional channel in the Delay-Doppler domain that directly represents the geometry of the various reflectors composing the wireless link. The future scope of this work aims at measuring location from Angle of Arrival using its impulse response and distance using its delay by using OTFS.

The advanced test and simulation tools provide the UE behaviors modelling and channel evaluation. The results of the user and channel analysis are promising to be integrated with the cell-free IAB scheduler, the beamforming scheduler and the resource allocation, since these techniques are usually dependent on the user mobility, location, and channel information. In addition, the channel measurements and modelling in multiple frequency bands from partners in work package 3 can also be used for the cell-free scheduler, resource allocation for D2D communication and intelligent beam steering.

Autonomous beam steering in a factory digital twin using Winprop will be performed (as opposed to urban environment using Matlab's Siteview) to obtain beam direction accuracy results. Autonomous beam steering in a factory digital twin using Winprop will be performed (as opposed to urban environment using Matlab's Siteview) to obtain beam direction accuracy results.

7 References

[1] 5G programmable networks: https://www.ericsson.com/en/core-network/5g-core/forms/5g-core-programmabilityunderestimated-opportunity

[2] 3GPP TS 38.300; NR and NG-RAN Overall Description.

[3] B. D. Van Veen and K. M. Buckley, “Beamforming: A versatile approach to spatial filtering,” IEEE assp magazine, vol. 5, no. 2, pp. 4–24, 1988.

[4] 3GPP TR 38.874; Study on integrated access and backhaul.

[5] M. Polese, M. Giordani, A. Roy, D. Castor, and M. Zorzi, “Distributed path selection strategies for integrated access and backhaul at mmwaves,” in 2018 IEEE Global Communications Conference (GLOBECOM), 2018, pp. 1–7.

[6] M. Simsek, O. Orhan, M. Nassar, O. Elibol, and H. Nikopour, “Iab topology design: a graph embedding and deep reinforcement learning approach,” IEEE Communications Letters, vol. 25, no. 2, pp. 489–493, 2020.

[7] R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction. MIT press, 2018.

[8] R. J. Williams, “Simple statistical gradient-following algorithms for connectionist reinforcement learning,” Machine learning, vol. 8, no. 3, pp. 229–256, 1992.

[9] C. J. Watkins and P. Dayan, “Q-learning,” Machine learning, vol. 8, no. 3-4, pp. 279–292, 1992.

[10] P. Ngatchou, A. Zarei, and A. El-Sharkawi, “Pareto multi objective optimization,” in Proceedings of the 13th International Conference on, Intelligent Systems Application to Power Systems, 2005, pp. 84–91.

[11] J. A. Boyan and M. L. Littman, “Packet routing in dynamically changing networks: A reinforcement learning approach,” in Advances in neural information processing systems, 1994, pp. 671–678.

[12] T. Mai, H. Yao, Z. Xiong, S. Guo, and D. T. Niyato, “Multi-agent actor-critic reinforcement learning based in-network load balance,” in GLOBECOM 2020 - 2020 IEEE Global Communications Conference, 2020, pp. 1–6.

[13] J. Hu and M. P. Wellman, “Nash q-learning for general-sum stochastic games,” Journal of machine learning research, vol. 4, no. Nov, pp. 1039–1069, 2003.

[14] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, “Counterfactual multi-agent policy gradients,” in Proceedings of the AAAI conference on artificial intelligence, vol. 32, no. 1, 2018.

[15] S. Zeng, X. Xu, and Y. Chen, “Multi-agent reinforcement learning for adaptive routing: A hybrid method using eligibility traces,” in 2020 IEEE 16th International Conference on Control Automation (ICCA), 2020, pp. 1332–1339.

[16] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, “Openai gym,” arXiv preprint arXiv:1606.01540, 2016.

[17] 6G BRAINS Deliverable D4.1: Koffman, Israel, Globen, Baruch, Eappen, Geoffrey, Cosmas, John, Zhang, Yue, Lu, Ge, Li, Wei, & Gavras, Anastasius. (2021). D4.1 Design and Description of the Intelligant IAB and RmUE/mUE and human-centric control interfaces

over Dynamic Ultra-dense D2D Cell Free Network (1.0). Zenodo. https://doi.org/10.5281/zenodo.6794507

[18] G. Xia, Y. Zhang, L. Ge and H. Zhou, "Deep reinforcement learning based dynamic power allocation for uplink device-to-device enabled cell-free network," 2022 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), 2022, pp. 01-06, doi: 10.1109/BMSB55706.2022.9828568.

[19] Geoffrey Eappen, John Cosmas, T. Shankar ,A. Rajesh , Rajagopal Nilavalan, Joji Thomas "Deep learning integrated reinforcement learning for adaptive beamforming in B5G networks" IET Communications, 28 September 2022 https://doi.org/10.1049/cmu2.12501

[20] Yi Hong, Tharaj Thaj and Emanuele Viterbo "Delay-Doppler Communications – Principles and Applications" Science Direct, 2022