This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2022.3200673

Routing and Resource Allocation for IAB Multi-Hop Network in 5G Advanced

Hao Yin, *Graduate Student Member, IEEE*, Sumit Roy, *Fellow, IEEE*, Liu Cao, *Graduate Student Member, IEEE*

***Abstract*—Integrated access and backhaul (IAB) is a novel feature for extending the network coverage in 5G cellular networks, based on sharing/efficient allocation of owner's spectrum traditionally reserved for access. However, since ultra-reliability and low latency (URLLC) requirements are a key component of 5G advanced services, provisioning such services present stringent challenges for IAB multi-hop network design. To fulfill the URLLC requirements in the IAB network, we propose a cross-layer design on routing and resource allocation under the current 3rd Generation Partnership Project (3GPP) 5G standards. We first formulate a routing problem for the IAB multi-hop network, which minimizes the latency while satisfying the reliability requirement. Subsequently, we present a reinforcement learning (RL) framework to solve the resource allocation and routing problem based on the local information of each agent (IAB node) in the environment. Afterward, we propose a novel entropy-based RL algorithm with federated learning (FL) mechanism to improve the overall performance as well as accelerate the convergence speed. Via the simulation, the proposed algorithm outperforms baseline algorithms from the latency and reliability perspective, respectively. Meanwhile, the convergence speed with the proposed algorithm also improves by using FL.**

***Index Terms*—5G, IAB, multi-hop, routing, resource allocation**

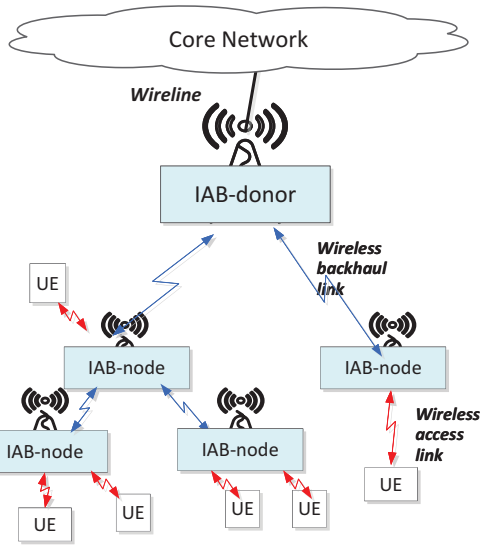


Fig. 1: Illustration of IAB in 3GPP Release 16.

I. Introduction

With growing demand for wireless access in support of new applications, next evolution of cellular networks will not only provision for network capacity, but also in conjunction with a significantly lower application delays, while maintaining coverage [2]. 5th generation (5G) and beyond deployments promise to increase average link speeds relative to 4G Long Term Evolution (LTE) by 10x (and peak rates by 100x). Coupled with edge computing (locating compute, storage and associated network functions close to the end-user) that promises to reduce network latency by 2 orders of magnitude, new mobile broadband use cases, e.g., Augmented (AR) and Virtual Reality (VR) enabled user devices and automation in industrial manufacturing and transportation/logistics based on Vehicle-to-Everything (V2X) networking, are expected. In pursuit of higher network capacity, frequencies above 24 GHz have been identified for Radio Access Networks (RANs) - so-called Frequency Range 2 (FR2) or millimeter-wave (mmWave) band - for 5G networks, for meeting the demands from traffic growth that is challenging the capacity of access networks below 6 GHz. Large channel bandwidths (several hundred MHz) available at those mainly underutilized spectrum portions thereby have unleashed plenty of opportunities to deliver the RAN Gbps-throughput promise [3]. However, mmWaves exhibit unfavorable propagation characteristics such as high isotropic losses and marked susceptibility to blockages and signal attenuation [4]. Indeed, mmWave deployments are typically coverage-limited, leading to denser deployments for hot-spot (high-demand) style scenarios [5]. Clearly, achieving such dense 5G deployments (even if localized) incurs significant capital expenditures (CAPEX) such as the fiber construction and site acquisition costs.

In order to provide a technically effective and economically viable solution to the required network densification, wireless backhaul solutions for 5G networks have recently emerged as a viable strategy. Notably, 3GPP Release 16 specifications introduced a new multi-hop wireless access architecture, Integrated Access and Backhaul (IAB), a wireless backhaul solution in which the access and backhaul links share the same hardware, protocol stack, and also spectrum. As illustrated in Fig. 1, IAB uses relaying among infrastructure nodes (IAB-nodes) to extend the coverage for the mobile edge users to the base station (IAB-donor) that is connected by high bandwidth wireline to the 5G core. In addition to the cost reduction, other factors are motivating the implementation of IAB networks. 1) Joint utilization of FR2 by access and backhaul: With 5G and

This work was supported in part by NSF CCRI Award 2016379. This work was presented in part at the IEEE Vehicular Technology Conference (VTC) 2021-Fall, [DOI: 10.1109/VTC2021-Fall52928.2021.9625389] [1]. (*Corresponding author: Liu Cao*.)

The authors are with the Department of Electrical and Computer Engineering, University of Washington, Seattle, WA 98195 USA (e-mail: haoyin@uw.edu; sroy@uw.edu; liucao@uw.edu).

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2022.3200673

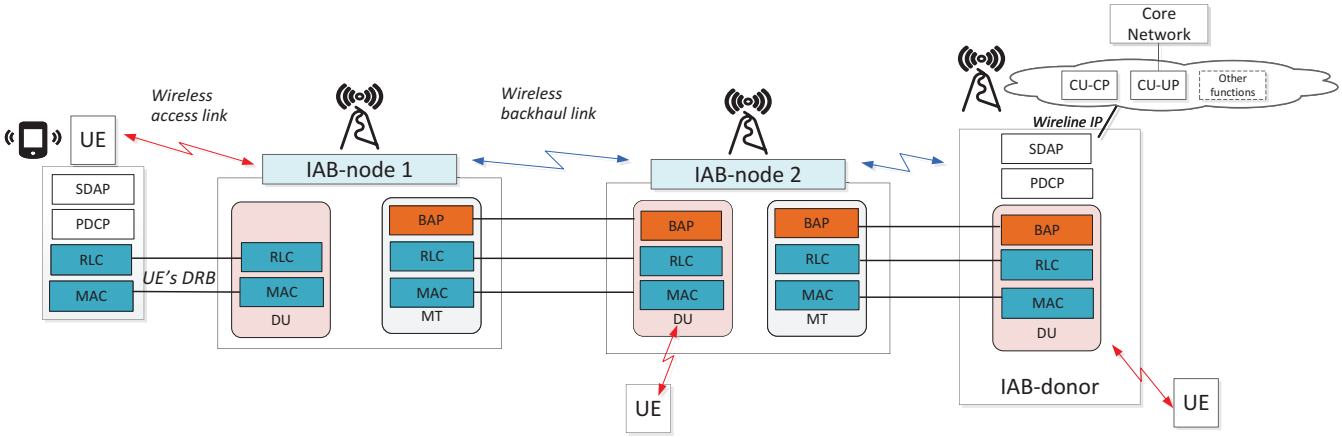


Fig. 2: Protocol stack for UE-access using IAB-relaying with the BAP layer in 3GPP IAB architecture 1a.

beyond, the access links will also operate in the mmWave spectrum. The spectrum range, which was previously used for backhauling, will also be used by the access links. As a result, there may be a conflict of interest between the access and backhaul links, which requires standardization. 2) Supporting NLOS backhauling: With low-height access points installed on, e.g., lamp posts, there exists a probability for blockage. Thus non-line-of-sight (NLoS) communication is also taken into account in the backhaul links [6].

We focus on the wireless multi-hop access and backhaul network where the different links reuse the same frequency band, leading to potential mutual interference-limited capacity considerations. Therefore, proper management of the radio resource allocation is fundamental to operate this network to fulfill the URLLC requirements of different 5G new advanced services. In particular, since the proposed media access control (MAC) solution is based on the Time Division Multiple Access (TDMA) method, it involves the optimization of routing paths and scheduling of directional transmissions along with established links [3]. As a result, this paper investigates the routing and resource allocation for 5G IAB multi-hop network where the URLLC requirements are highly strict.

A. Related Work

Recent implementations of multi-hop wireless networks include the well-known IEEE 802.11s enhancement to the base CSMA/CA medium access control (MAC) protocol. Such 802.11s AP meshes were used for range extension to end users, typically limited to a few hops to manage the resulting increased interference due to frequency reuse among co-channel links.

As far as cellular networks, 3GPP defined a version of LTE relays [7] limited to two-hop communications, that did not achieve significant commercial success. Conversely, 5G NR is a beam-based air interface relying on dedicated reference symbols and channels, alleviating many of the constraints LTE suffered. Because mmWave transmissions are highly directional, interference is naturally mitigated with appropriately elevated BS locations. Thus, mmWave multi-hop based in IAB can be designed more like a wired multi-hop network with switches and hubs than a conventional wireless multi-hop system. The backhaul traffic is routed to the donor node, with scheduling at each hop, effectively managing the network interference.

In this work, we focus on *joint* routing and resource allocation for IAB multi-hop networking in 5G NR. The performance analysis of such mmWave multi-hop network is relatively recent, beginning with initial work in [8], [9]. [10] presented an analytical framework for IAB-enabled cellular networks on the coverage and performance. In addition, the study [11] proposed a global traffic allocation scheme to achieve the low latency requirements in multi-hop transmission. In [12], a novel joint incentive and resource allocation design were proposed for the IAB problem. However, the primary shortcoming of all above works is that they do not address the URLLC requirements in 5G networks. Due to a lack of global network status information and imperfect sensing, achieving URLLC latency bounds in multi-hop networks is a largely unsolved problem.

Towards intelligent operations and scheduling, there has been a growing interest in the application of artificial intelligence (AI) strategies in 5G NR. In tandem with the new broad and complex features offered by the new IAB protocols, data-driven approaches would enable their optimal usage for real-world applications to achieve the quality, reliability, latency, and efficiency requirements like URLLC. Among all the AI strategies, reinforcement learning (RL) is designed to learn from the environment by exploring the underlining connection of different parameters, which has shown a good fit and performance gain in the scheduling and resource allocation problem [13]–[15]. An optimization-aided DRL-based framework was developed in [16] to aim at maximizing the eMBB data rate subject to a URLLC reliability constraint in resource slicing problem. [17] proposed an Advantage Actor-Critic (A2C) based RL approach on the IAB resource allocation algorithm, which was able to cope with the dynamics of the link status in mmWave 5G IAB networks. The aforementioned works mainly apply the DRL algorithms to obtain a deterministic policy in the wireless communication optimization problem. However, convergence speed and model generalization are

significant concerns regarding the development in the real world. This paper introduces a model-free off-policy DRL algorithm based on maximum entropy reinforcement learning, Soft Actor-Critic (SAC) [18], to accelerate the convergence speed and also align with the current 3GPP standard. In this paper, we mainly focus on the cross-layer modeling on the resource allocation and routing in the IAB multi-hop network. The main contributions of this paper are as follows:

- **We formulate a routing optimization problem for the IAB multi-hop network that minimizes the transmission latency while satisfying the reliability requirement, respectively.** We analyze the multi-hop network under the current 3GPP 5G standard and propose an optimal routing algorithm with global information.
- **We present a deep reinforcement learning (DRL) framework to solve the routing and resource allocation problem based on the local information in the multi-hop network.** The proposed framework only requires the local information to optimize the routing paths and resource scheduling of directional transmissions along with established links.
- **We propose a novel entropy-based reinforcement learning (RL) algorithm with federated learning (FL) mechanism to accelerate the convergence speed.** The computation complexity of the proposed algorithm is investigated in terms of convergence speed and runtime. This architecture can be further extended to other similar DRL-based decision-making scenarios.

This paper is organized as follows: A brief recap of IAB architecture and routing and resource allocation mechanism is summarized in Section II. Section III analyzes the latency and reliability for multi-hop transmission and formulate an optimal routing problem. In Section IV, we present a DRL framework for the resource allocation and routing problem, which our proposed SAC algorithm can further solve with the FL mechanism. Section V compares performances between our proposed algorithm and baseline algorithms from different perspectives based on ns-3 simulation results. Finally, Section VI draws the conclusions.

II. System Architecture

A. Integrated Access and Backhaul Architectures in 5G NR

To cope with the need for appropriate backhaul rates for small cell networks, 3GPP first proposed a study item on IAB in [19]. The physical-layer specification of IAB were completed in 2019, and higher-layer protocols and architecture were completed in 3GPP Rel-16 [20]. Further enhancements (e.g., mobile IAB) have been carried out in 3GPP Rel-17. However, despite the consensus about IAB's ability to reduce costs, designing a high-performance IAB network is still an open research challenge [21].

As Fig. 2 shows, two types of wireless links constitute the IAB network: access and backhaul links. An access link connects UE and an IAB node or IAB donor, while a backhaul link exists between IAB parent and IAB child node. The IAB functionality requires two network entities: IAB-donor and IAB-node(s). An IAB-donor is a gNB that provides network access to UEs via a network of backhaul and access links. The IAB donor is split into a centralized unit (CU), which terminates the Packet Data Convergence Protocol (PDCP) and the Radio Resource Control (RRC) protocol as well as a distributed unit (DU) that terminates the lower protocols, i.e., Radio Link Control (RLC), Medium Access Control (MAC) and the physical (PHY) [20]. The motivation of the CU/DU functional split in the IAB donor is that all time-sensitive functionalities, e.g., scheduling, fast re-transmission, segmentation, etc., can be realized in the DU, i.e., close to the radio and the antenna. At the same time, it is possible to centralize the more minor time-sensitive radio functionalities in the CU [22]. The IAB-donor connects to the IAB-node(s) using the 5G access interface and is connected to the Core Network (CN). A Backhaul Adaptation Protocol (BAP) layer is added above the RLC layer to include routing information and allow for hop-by-hop forwarding. The IAB node comprises mobile termination (MT) and DU functionalities. The IAB node connects to an upstream IAB node or an IAB donor's DU via MT function, while it also provides wireless backhaul for the downstream IAB nodes and UEs via the DU function. Note that IAB nodes can be cascaded without a technical limit to the number of IAB nodes. Therefore, it is important to consider the latency and reliability of multi-hop transmission.

The 3GPP standard proposed five different configurations for IAB architecture, with various levels of decentralization of the network and backhauling functionalities [21]. In this paper, we consider Architecture 1a, which has more chance to be selected for future standardization according to the 3GPP nomenclature [21]. It should be noted that this architecture not only leverages CU/DU-split architecture but also adds an adaptation layer that replaces the IP functionality to hold wireless routing information enabling hop-by-hop forwarding [19]. Fig. 2 shows Architecture 1a, where multiple IAB-nodes use wireless backhaul, while IAB-donors have fiber connectivity toward the core network. In this architecture, the IAB donor is the node that serves the IAB nodes and other UEs that are directly connected to it. Each IAB node has a mobile termination (MT) function which connects to a parent DU (IAB donor DU or another IAB node DU) and a DU function that serves UEs or the MT functions of child IAB nodes. Such a configuration yields the most limited impact on the core network and signaling overhead and the lowest relay complexity and processing requirements [22]. Compared with other architectures, Architecture 1a implements a functional split of the radio protocol stack (the split happens at the RLC layer), with the control and upper layers in the IAB-donor CU and the lower layers in the DUs of the IAB-nodes. Therefore, the RRC, SDAP, and PDCP layers reside in the CU, while RLC, MAC, and PHY are in the DUs. An additional adaptation layer manages the routing on top of RLC, enabling the end-to-end connection between DUs and CU.

B. Packet transmission over wireless backhaul

1) Multi-hop forwarding and routing with BAP layer: The multi-hop forwarding is newly enabled via the IAB-specific BAP, inserted as a specific header in the RAN layer 2 stack.

Consider a general IAB network as illustrated in Fig. 2. Several IAB nodes are transmitting backhaul traffic to the IAB-donor over the wireless link and then forward to the core network. Since the IAB-donor assigns a unique L2 address (BAP address) to each IAB node that it controls. After the initialization, the IAB donor will know the existing IAB nodes inside its network. Then each IAB node is able to know the total number of nodes as well as its neighbor nodes in the current network under the global configuration of the IAB donor. For the transmission between the IAB nodes and IAB donor, the BAP header will include the source and destination ID as well as an optional path ID. Each IAB node has its routing table (configured by the IAB donor) containing the next hop identifier for each BAP ID. The routing tables for the downlink (DL) and uplink (UL) directions can be different, used by DU or MT parts separately.

2) Features for IAB networks on PHY, MAC and RLC layer: The physical layer of IAB is intended to support in-band backhauling with the same carrier frequencies for both the NR backhaul links and the access links. The in-band operation comes with a half-duplex constraint, implying that the IAB-MT part of an IAB node cannot receive while its collocated DU is transmitting and vice versa to avoid intra-site interference. Therefore, a strict time-domain separation is required between transmission and reception phases within each IAB node. At the MAC layer, the IAB-nodes support flexible resource allocation for both DL and UL, which is thus similar to the normal UE allocation. An IAB network attempts to schedule the wireless resources to meet each UE bearer's requirement regardless of the number of hops a given UE is away from the Donor DU. The scheduler on the wireless backhaul link can distinguish the Quality of Service (QoS) profiles associated with different RLC channels. It may also apply information regarding the number of hops a packet needs to traverse, in addition to the QoS profile of the bearers, in order to provide hop-agnostic performance. Backhaul (BH) channel is a logistic mapping for transporting packets between IAB nodes/donor. Different packets from UEs will map to a single BH RLC channel that is established only between two IAB entities.

With the features provided by the IAB network, we can design the routing and resource allocation algorithms. Wireless backhaul links are vulnerable to blockage, e.g., due to moving objects such as vehicles, seasonal changes (foliage), or infrastructure changes (new buildings). Such vulnerability also applies to physically stationary IAB-nodes. In addition, traffic variations can create uneven load distribution on wireless backhaul links, leading to local link or node congestion [22]. Therefore, topology adaptation for physically fixed relays shall be supported to enable the robust operation and dynamic routing, which is still a challenge for the IAB networks.

In the 5G network, the channel conditions are obtained as follows: the receiver (RX) reports channel information, i.e., channel status information (CSI), to help the transmitter (TX) to determine the MCS and the resources for packet transmission. Following the 3GPP standard, this method is also applied to multi-hop networks. The feedback from the RX node contains the channel quality indicator (CQI) that estimates at the RX node and instructs the TX node to select a corresponding MCS for a certain block error rate (BLER). The CQI feedback is based on the CSI reference signal of each subband (which contains several contiguous RB). The TX node then uses the CQI value to determine the MCS for each transmission.

C. Node management and Routing

In the IAB routing mechanism, each IAB node is assigned a unique address (BAP address), and the IAB donor CU configures a routing table at each IAB node to direct the flow of traffic based on these node addresses. A mechanism is established within the IAB network to help forward it via multiple intermediate IAB nodes between the IAB donor and a specific UE from a packet perspective. It includes the route selection and the next-hop destination at each IAB node once a route is selected. However, due to the multi-hop nature of IAB networks, the backhaul link failure may occur on intermediate IAB nodes along a transmission path which is caused by the differences in their effective link capacities (i.e., different SINRs). In addition, high latency will also be incurred because of the different congestion conditions on intermediate IAB nodes. Therefore, an optimal route needs to be selected between the IAB donor and the specific UE in the IAB networks based on the reliability and latency requirements.

It should be noted that each IAB is only able to acquire the local information regarding all its nearby IAB nodes; by contrast, IAB donor obtains the information from all IAB nodes in an IAB network. When an IAB node obtains new local information, the relevant routing decisions (routing tables) will be globally renewed/reconfigured by the IAB donor. Meanwhile, in order to avoid congestion-related packet drops among the IAB-nodes, the routing within IAB networks is also supported for both UL and DL directions which can happen on different nodes and links between the IAB donor and a specific UE. Although the link failure or the congestion problem can be handled by higher-layer protocols, e.g., Transmission Control Protocol (TCP), the scope of the impacted nodes will extend well beyond the RAN/IAB network. Furthermore, if packets are dropped due to congestion in the IAB network, the TCP congestion avoidance and slow start mechanisms may be triggered, and the end-to-end performances can be significantly impaired.

The problem of routing in IAB networks has been investigated under different cases [23]–[25]. On the other hand, deep reinforcement learning (DRL) algorithm-based solutions have been previously proposed for different (non-IAB) network use-cases/topology [26]–[28]. While some DRL was also applied in the IAB-based networks from different perspectives [29], [30]. The DRL can be also an appropriate approach to cope with previous issues in the multi-hop transmission problem. The main reason is that each agent (node) in the IAB network can be constructed to find a route that maximizes the expected reward through interaction with the real-time environment. Under the DRL algorithm, the selected route(s) in the UL/DL direction can consider both the latency and reliability requirements. Meanwhile, since such optimal route(s) can be achieved

with a fast convergence speed, the DRL approach is still applicable when some changes happen abruptly in the IAB network, e.g., the leaving or coming UEs or the change of IAB network topology.

III. System Model

This section provides an analysis of the latency and reliability requirements in multi-hop networks for the edge users according to the current 3GPP NR standards. We follow the analysis in Section II about the resource allocation and routing problem. Then based on the analysis, we formulate the problem as an optimal routing problem with the central knowledge that could be solved with Dijkstra's algorithm.

A. Latency for Multi-hop transmission

A key degree of freedom in 5G for resource allocation is the flexible numerology (u) [31] that allows sub-carrier spacing to scale as $2^u \times 15$ kHz to provide a balance between different service requirements. LTE system latency in the user plane is typically measured as a multiple of Transmission Time Interval (TTI). The analysis of NR can reuse the same approach but with different system parameters due to enhanced hardware capability and numerology, summarised by [32], [33]. The NR TTI length is equal to the slot length[1].

For a direct transmission from the base station to user equipment (UE), the total delay depends upon 4 components - i) queuing time before transmission T_{que}, ii) stack processing time at source (gNB) $T^{\text{s}}_{\text{proc}}$, iii) transmission time T_{trans}, and iv) processing time at destination (UE) $T^{\text{dest}}_{\text{proc}}$. All these delay components are a multiple of ***TTI*** (same length as slot) as suggested above. Then the total delay of the direct transmission is calculated in Eq. (1).

$$T^{\text{dir}}_{\text{delay}} = T_{\text{que}} + T^{\text{s}}_{\text{proc}} + T^{\text{dest}}_{\text{proc}} + T_{\text{trans}}, \tag{1}$$

where the processing delay is normally $T_{\text{proc}} = T^{\text{s}}_{\text{proc}} + T^{\text{dest}}_{\text{proc}} = 4 \cdot \textit{TTI}$ [34]. The transmission delay is related to both the packet size pkt required to be transmitted in current slot and the modulation and coding scheme (MCS). For a given MCS, the transmission block size TB can be determined [2]. Then the transmission time is calculated as $T_{\text{trans}} = \lceil \frac{pkt}{TB} \rceil \cdot \textit{TTI}$.

With regard to the multi-hop transmission, the packet in the relay nodes, which is forwarded to the next node, does not need to go through all the stack procedures as discussed above. By introducing the CU/DU split architecture, the processing time is reduced since there is no need to go pass all the whole L2 and L3 stacks compared with the none split option. We assume that each relay node would immediately forward the packet to the next node, the latency caused by each relay is thereby half of the processing time plus the transmission time, i.e., $T_{\text{relay}} = \frac{1}{2} T_{\text{proc}} + T^{\text{relay}}_{\text{trans}}$. Thus, the total delay from the source to destination through n relay nodes is calculated in Eq. (2):

$$\begin{aligned} T_{\text{delay}} &= T^{\text{dir}}_{\text{delay}} + n \cdot T_{\text{relay}} \\ &= T_{\text{que}} + \frac{n+2}{2} T_{\text{proc}} + \sum_{i=1}^{n+1} T_{\text{trans}}(i), \end{aligned} \tag{2}$$

where $T_{\text{trans}}(i)$ is the transmission time at the link i.

B. Resource allocation and Reliability

In 5G network, the receiver (RX) reports channel status information (CSI) to help the transmitter (TX) to determine the MCS and which RB to transmit. This method is also applied for the multi-hop networks. The feedback from the RX sides contains the channel quality indicator (CQI) that estimates at the RX and instructs the TX to select a corresponding MCS for a certain block error rate (BLER). The CQI feedback is based on the CSI reference signal of each subband (contains several contiguous RB). The TX then uses the CQI value to determine the MCS for each transmission. In a multi-hop network, the nodes are not all scheduled by a center controller, thus the collision and interference may happen when two close nodes transmit at the same time. On the basis of the procedures analysis above, the reliability is composed of two parts, the collision probability p_c [3] and the BLER p_b on each link. Regarding a multi-hop transmission with n relays (total $n+2$ nodes and $n+1$ transmissions), we assume the transmission between two nodes is independent, the probability of a successful multi-hop transmission $\mathbb{P}$ is thus given by

$$\mathbb{P} = \prod_{i=1}^{n+1} (1 - p_b(i))(1 - p_c(i)) = \prod_{i=1}^{n+1} p_s(i)(1 - p_c(i)), \tag{3}$$

where $p_b(i)$ is pre-configured for the i^{th} link and $p_s(i) = 1 - p_b(i)$ denotes the probability of a successful transmission expected by the current configuration. In this paper, we consider the 3GPP channel model [36]. It supports the modeling of wireless channels between 0.5 and 100 GHz by means of a stochastic Spatial Channel Model (SCM), in which a single instance of the channel matrix $\mathbf{H}$ is computed according to random distributions for large scale fading parameters (i.e., the delay profile, the angles of arrival and departure, and the shadowing) and for the small scale fading (i.e., for small variations in the channel, for example, as given by the Doppler spread).

C. Routing and Graph Model

Without loss of generality, the connection and links between different devices could be modeled as a graph. In this graph $G = (\mathcal{V}, E)$, the vertex $\mathcal{V}$ represents the network devices (nodes) and the edges E represent the communication links between the pairs of network nodes. Each node is aware of the connection thereof to neighbour nodes and associated channel conditions, and then according to the configuration, it can obtain the p_b in the link. We first consider the routing problem,

[1]For example, the TTI is $\frac{1}{2^3}$ ms when using numerology 3.

[2]The detailed calculation is shown in 3GPP standard [35].

[3]The collision happens when two close nodes select the same subband at the same time for packet transmission.

that the destination node (serving as a controller) have the perfect knowledge of the network topology and the channel conditions. A further assumption is that there is no hidden terminal problem and the edge nodes are far enough not to interfere with each other thus there is no collision due to the RB selection, i.e., $p_c = 0$. Then this could be formulated as a problem of finding the path q of minimum latency for a given reliability constraint σ. Notably, the retransmission procedure can improve the reliability, thus, the formulation is conducted to minimum latency as the objective and find a path q from the set $\mathcal{Q}_{S,D} = \{q_{S,D} | \forall q_{S,D} \in G\}$ of all paths connecting Source (S) to Destination (D) for a given MCS.

Problem 1 (Optimal Routing Problem)**.**

$$\min_{q \in \mathcal{Q}_{S,D}} \quad T_{\text{delay}}(q), \quad \textit{s.t.}\ \mathbb{P}(q) \geq \sigma. \tag{4}$$

Notice that the given constraint is related to the reliability, $\mathbb{P}$, expressed in Eq. (4), which has considered the resource constraint (i.e., collision probability and BLER). The constrained optimization problem in (4) can be transformed into an unconstrained problem by applying the Lagrange multiplier method and expressions as follows with immediately scheduling (no queuing delay):

$$\begin{aligned} \min_{q \in \mathcal{Q}_{S,D}} \quad & T_{\text{delay}}(q) - \mu \log(\mathbb{P}(q)) \\ & = T_{\text{proc}} + \sum_{i \in q} (\frac{1}{2} T_{\text{proc}} + T_{\text{trans}}(i) + \mu \cdot \log(\frac{1}{p_s(i)}) \\ & = C + \sum_{i \in q} c(i, \mu), \end{aligned} \tag{5}$$

where $\mu \geq 0$ is the Lagrange multiplier, C is a constant and $c(i,\mu) \triangleq \frac{1}{2} T_{\text{proc}} + T_{\text{trans}}(i) + \mu \cdot \log(\frac{1}{p_s(i)})$. In terms of a particular Lagrange multiplier μ, the optimal path q^* can be obtained by the Dijkstra's algorithm [37] by setting $c(i,\mu)$ as the weight of each link. Then the problem is reduced to finding a solution of μ^* by the following lemma and algorithm.

Lemma 1. *The optimal $q^*(\mu)$ solves the problem 1, when there exists a μ that achieves $q^*(\mu) = \sigma$.*

Proof: Please see the proof in App. A. ∎

Lemma 1 introduces a way to obtain the optimal μ^* by finding the one that satisfies $\mathbb{P}(q^*(\mu)) = \sigma$. A bisection approach is employed with exponential convergence rate to find the optimal μ^* as shown in Algorithm 1 in App. B.

For each link i, the $T_{\text{trans}}(i)$ is also related to the MCS choice from the possible set M_{cs}. The weight now used in the unconstrained optimization (5) become

$$\tilde{c}(i,\mu) \triangleq \min_{M_{cs} \in \{1,2,\dots\}} \frac{1}{2} T_{\text{proc}} + T_{\text{trans}}(i) + \mu \cdot \log(\frac{1}{p_s(i)}). \tag{6}$$

The optimal MCS $M^*_{cs}(i,\mu^*)$ then can be obtained according to Lemma 1 by

$$M^*_{cs}(i,\mu^*) = \arg \min_{M_{cs} \in \{1,2,\dots\}} \frac{1}{2} T_{\text{proc}} + T_{\text{trans}}(i) + \mu \cdot \log(\frac{1}{p_s(i)}). \tag{7}$$

Recalling the BAP layer that contains the information of the path from the source node to the destination, the nodes can build up the network topology from the history information. Notice that increasing the relay nodes in the path will increase delay and decrease the reliability, so we can use Dijkstra's algorithm with the same weight $c(i)$ to avoid the backwards path and the neighbours with more hops. In this way, we narrow down the original routing and resource allocation problem so that IAB node selects the next hop with corresponding MCS and slots.

IV. Deep Reinforcement Learning

In section III, we show an optimal routing solution with the assumption of the perfect network knowledge and no interference among nodes for the resource allocation in the uplink (UL) and downlink (DL), respectively. Though the algorithm requires global information, we can apply it to reduce the complex routing problem based on the graph to the neighbor selection and resource allocation problem. **Problem 1** considering Dijkstra in the previous section is used to determine an optimal routing scheme. Since the location of IAB nodes keeps fixed in the IAB networks, we thereby adopt the solution to **Problem 1** as the pre-configured/default routing setup for the IAB networks. Afterward, we propose SAC to solve the joint optimization problem between resource allocation and possibly to change the routing by choosing a different neighbor. Each IAB node initially utilizes the pre-configured routing setup for the routing problem and then re-selects its neighbors to modify the optimal routing path when it learns from the environment. As a result, the proposed DRL algorithm helps each IAB node update its resource allocation scheme and periodically overwrites the previous routing setup. Therefore, resource allocation and routing portions run simultaneously under the proposed DRL algorithm.

The challenges to addressing the joint optimization problem between resource allocation and routing for the IAB network are summarized as follows: 1) **Dynamic channel**: with 5G and beyond, the mmWave channel is changing rapidly, which is vulnerable to penetration, attenuation, and blockage in the IAB networks. 2) **Dynamic UE traffic**: Since UEs in the IAB network are primarily mobile users, UE traffic is also quite dynamic in terms of UE packet throughput and leaving/coming of UE nodes. 3) **Imperfect feedback**: The CSI feedback that an IAB node receives could be inaccurate due to the delay or the imperfect channel estimation. In summary, the IAB nodes may be relatively fixed and static, while the wireless environments change from time to time, so we also need a dynamic resource allocation algorithm. The DRL-based algorithm is able to cope with these challenges. This is because each IAB node aided by DRL can learn the potential patterns (aimed for dynamic changes) as well as the bias (aimed for imperfect feedback) from the environment. By contrast, traditional routing algorithms, such as meta-heuristic algorithms, only aim for the routing problem rather than the joint optimization of the routing and the resource allocation. Thus they cannot cope with the previously illustrated dynamic changes and correct the bias incurred by the imperfect feedback. Therefore, with traditional routing algorithms, the overall performance of IAB networks in such a dynamic environment will significantly degrade.

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2022.3200673

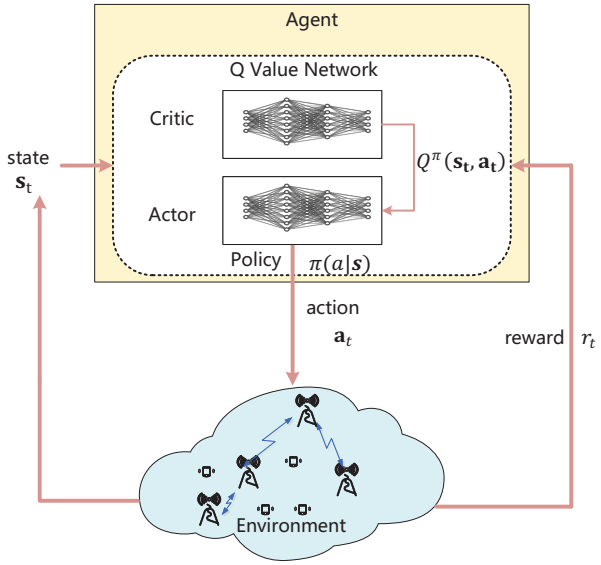


Fig. 3: Illustration of DRL Architecture.

A. Deep Reinforcement Learning

As illustrated in Fig. 3, the framework of reinforcement learning consists of agents and environments where each agent interacts with each other. In the multi-hop network, each IAB node is considered as an agent, and the wireless channel and transmission results are regarded as the environment that presents a collection of the channel states and the feedback for each transmission. Meanwhile, each IAB node is able to know the total number of nodes in the current network based on the global configuration of the IAB donor. Each node is assumed to allocate the resources based on the Time Division Multiple Access (TDMA) method.

At each time slot t, each IAB node can acquire the related information of its neighboring IAB nodes in the UL/DL direction respectively. The obtained information from each neighboring IAB node includes the channel information (i.e., CSI), total latency T_{delay}, and reliability (the probability of success $\mathbb{P}$). If we denote the total number of nodes in the multi-hop network is M, the state $\mathbf{s}_t$ of the IAB node $m(1 \leq m \leq M)$ at slot t is given by

$$\mathbf{s}_t^m = \begin{bmatrix} CSI_1^{UL}, & T_{1,delay}^{UL}, & \mathbb{P}_1^{UL} \\ CSI_1^{DL}, & T_{1,delay}^{DL}, & \mathbb{P}_1^{DL} \\ \vdots & \vdots & \vdots \\ CSI_M^{UL}, & T_{M,delay}^{UL}, & \mathbb{P}_M^{UL} \\ CSI_M^{DL}, & T_{M,delay}^{DL}, & \mathbb{P}_M^{DL} \end{bmatrix}, \tag{8}$$

where $T_{n,delay}$ and $\mathbb{P}_n$ of node $n(1 \leq n \leq M)$ in the UL/DL direction are expressed in Eq. (2) and Eq. (3). Note that if a node does not belong to the neighboring node set of node m (including node m), all parameters related to this node in $\mathbf{s}_t^m$ will be set to a special value to indicate such a relation. As a result, each IAB node has the state with the same dimension, however, only the information from the neighboring nodes are truly effective in its state matrix.

The node m then takes action, i.e., choosing the best neighboring nodes to transmit its packets. Since the channel condition, latency and reliability vary at each neighboring node. The best chosen neighboring nodes may be different each time. To solve the problem above, we thereby implement the DRL algorithm where the state $\mathbf{s}_t^m$ of node m is regarded as the input while the output is an action score list. The action score list of node m at time slot t includes the scores of all nodes in the IAB network, which is expressed in Eq. (9).

$$\mathbf{a}_t^m = \begin{bmatrix} a_1^{UL}, & \cdots, & a_n^{UL}, & \cdots, & a_M^{UL} \\ a_1^{DL}, & \cdots, & a_n^{DL}, & \cdots, & a_M^{DL} \end{bmatrix}, \tag{9}$$

where a_n denotes the score of the node n in the UL/DL direction. This also benefits the model transfer and relay selection in the following discussion. Once the action score list is updated, node m chooses the node(s) with the highest UL/DL transmissions scores. For the UL resource allocation, the node m chooses some nodes with the highest scores and schedules the corresponding time slots for the chosen nodes to transmit their packets. Meanwhile, node m chooses the nodes with the highest scores for the DL resource allocation and then forwards their own packets to the chosen nodes immediately. Note that each node buffers a different number of packets to be transmitted in the UL direction, the chosen neighboring nodes need to transmit the packets in their buffers in order of node priority [4]. Therefore, when the node m chooses l nodes with the highest scores in the UL direction, l will depend on the number of packets as well as the packet size in the buffer of the higher priority node.

The decision function taken by each node is determined by the policy π_θ, where θ is the parameter of the policy π. There are many different RL algorithms to find and improve the policy π, while the objective of the standard RL is to maximize the expected sum of rewards from time t:

$$R_t(\pi) = \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \pi_\theta} \left[\sum_{k=t}^{\infty} \gamma^{(k-t)} r_k \right], \tag{10}$$

where $\gamma \in [0, 1]$ is the discount factor used to avoid the accumulated reward to be infinity, and $r(s_t, a_t)$ is the reward by taking action a_t at state s_t. In reinforcement learning, the transition of the state $\mathbf{s}_t$ and reward r_t are stochastic and modelled as a Markov decision process (MDP), where the transition probability of state $\mathbf{s}_{t+1}$ depends only on the last state $\mathbf{s}_t$ and the action a_t taken by the agent. Therefore, each transition from $\mathbf{s}_t$ to $\mathbf{s}_{t+1}$ can be characterized by a conditional probability $p(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)$. The reward r_t is used to guide the training and improve the policy. In the IAB multi-hop network, each node can calculate the reward based on the feedback (ACK/NACK) from the environment each time.

The node selection's objective is to meet the latency and reliability constraints in the UL/DL direction. After each action in a direction, the environment returns a reward to the agent to evaluate such an action. Therefore, the reward function that guides learning should be consistent with the objective. In our framework, the reward function consists of two components: the latency component and the reliability

[4] After the node with the highest score transmits all the packets in its buffer, the node with the second highest score is the allowed to transit packets. The same procedure is applied for the remaining chosen nodes in the UL direction.

component. Let o_t denotes the feedback at time slot t from the environment: $o_t = 0$ when ACK is received, otherwise $o_t = 1$ when NACK is received. For the latency component, if the ACK is received within the latency constraint τ (i.e., $T_{\text{delay}} < \tau$), a successful transmission happens with a positive reward which is expressed as $\tau - T_{\text{delay}}$. This indicates the smaller T_{delay} is, the higher the reward can be returned from the environment. However, if the NACK (or timeout) happens (i.e., $T_{\text{delay}} \geq \tau$ [5]), $\tau - T_{\text{delay}}$ is also used to quantify the impact on the reward. If the remaining time within the latency constraint is long enough for accommodating a re-transmission, i.e., $\frac{\tau - T_{\text{delay}}}{T_{\text{delay}}} > 1$, due to the higher latency, a positive but lower reward will be returned for a successful re-transmission. Otherwise, a negative reward will be returned. For the reliability component, the more re-transmissions a packet needs, the lower the returned reward of this packet. Considering both components, the reward r_t in the UL/DL direction is thereby expressed as:

$$r_t = \psi_d \left(\frac{(\tau - T_{\text{delay}})}{(T_{\text{delay}})^{o_t}} + (-1)^{o_t} \right) - \psi_r (K_{\text{trans}} - 1), \quad (11)$$

where τ is the latency constraint. K_{trans} is the total number of transmissions for the same packet. ψ_d, and ψ_r are the coefficients which determine the weight of the latency and the reliability component, respectively. In order to obtain a long term performance which successfully achieves the URLLC requirements, both the immediate rewards and future rewards should be taken into consideration as the RL objective in Eq. (10). Note that Eq. (11) is applied for both UL and DL direction. However, the reward r_t in the UL/DL direction are different even though the link used for UL and DL between the agent and its neighbouring node is the same, this is because the channel conditions and latency/reliability requirements are different in two directions.

B. Soft Actor-Critic

The brittle convergence properties and the requirements for meticulous hyperparameter tuning at different RL algorithms environments limit such methods' applicability to a complex, real-world domain like the routing and resource allocation problem for IAB multi-hop network. Most RL algorithms applied in current wireless network problems, like Deep Q Network (DQN) and Deep Deterministic Policy Gradient (DDPG), always obtain a deterministic policy, i.e., the policy only considers one optimal action for a given state. However, it is hard to generalize the property to other similar environments. Besides, the policy for routing and resource allocation in the IAB network is not always unique. Thus it is natural to consider a more robust algorithm with a stochastic policy for the model generalization in our resource allocation and routing problem.

In this section, we introduce Soft Actor-Critic (SAC), a model-free off-policy deep reinforcement learning algorithm based on maximum entropy reinforcement learning [18]. Instead of maximizing the expected sum of rewards in Eq. (10), the SAC algorithm introduces the entropy component into the objective at time t with the discount factor:

$$J_t(\pi) = \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \rho_\pi} \left[\sum_{k=t}^{\infty} \gamma^{(k-t)} \mathbb{E} \left[r_k + \alpha \mathcal{H} \left(\pi \left(\cdot \mid \mathbf{s}_k \right) \right) \mid \mathbf{s}_k, \mathbf{a}_k \right] \right], \quad (12)$$

where the temperature parameter α controls the degree of randomness of the optimal strategy and the importance of entropy relative to the reward, and $\mathcal{H}\left(\pi\left(\cdot \mid \mathbf{s}_t\right)\right)$ is the entropy of each action obtained by the policy. We use $\rho_\pi\left(\mathbf{s}_t, \mathbf{a}_t\right)$ and $\rho_\pi\left(\mathbf{s}_t\right)$ to denote the state and state-action marginals of the trajectory distribution induced by a policy $\pi(\mathbf{a}_t \mid \mathbf{s}_t)$.

The SAC algorithm consists of an actor-critic architecture with separate policy and value function networks as illustrated in Fig. 3. The actor updates the policies based on the policy gradient method, and the objective of the critic part is to evaluate the policy that the learning algorithm searches. More specifically, the SAC algorithms aim to use deep neural networks to learn the basic two functions - the policy function π_θ with parameter θ and the soft Q-function Q_ω with parameter ω.

The Q-function in typical RL algorithms is defined as a cumulative discounted reward by taking action $\mathbf{a}_t$ at state $\mathbf{s}_t$, and can be calculated using the Bellman equation []. In the maximum entropy reinforcement learning framework, we then regard the entropy as part of the reward to calculate the soft Q-function.

$$Q^\pi\left(\mathbf{s}_t, \mathbf{a}_t\right) = r\left(\mathbf{s}_t, \mathbf{a}_t\right) + \gamma \mathbb{E}_{(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}) \sim \rho_\pi}[V\left(\mathbf{s}_{t+1}\right)], \quad (13)$$

where $V(\mathbf{s}_t)$ is the value function defined as

$$V^\pi(\mathbf{s}_t) = \mathbb{E}_{(\mathbf{a}_t) \sim \pi}[Q^\pi\left(\mathbf{s}_t, \mathbf{a}_t\right) - \alpha \log \pi\left(\mathbf{s}_t, \mathbf{a}_t\right)]. \quad (14)$$

The soft Q-function parameters can be trained to minimize the soft Bellman residual

$$J_Q(\omega) = \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \mathcal{D}} \left[\frac{1}{2} \left(Q_\omega - \left(r_t + \gamma \mathbb{E}_{\mathbf{s}_{t+1} \sim \pi} \left[V_{\bar{\omega}}\left(\mathbf{s}_{t+1}\right) \right] \right) \right)^2 \right], \quad (15)$$

where the value function is implicitly parameterized through the soft Q-function parameters via Eq. (14), and it can be optimized with stochastic gradient

$$\hat{\nabla}_\omega J_Q(\omega) = \nabla_\omega Q_\omega \left(Q_\omega - \left(r_t + \gamma \left(Q_{\bar{\omega}} - \alpha \log \left(\pi_\theta \left(\mathbf{a}_{t+1} \mid \mathbf{s}_{t+1} \right) \right) \right) \right) \right) \quad (16)$$

The update makes use of a target soft Q-function with parameters $\bar{\omega}$ obtained as an exponentially moving average of the soft Q-function weights, which is shown to stabilize training [38].

C. Federated Learning

Each IAB node in the coverage of the IAB donor first uses the described SAC algorithm for the routing and resource allocation with random initialization of the neural network (NN) weight θ. However, some nodes are likely initialized with worse NN weights. There may arise an issue where the weights of NN in these nodes may never converge due to the faster change in the environment, such as channel condition. As a result, the applied SAC algorithm does not function in some nodes. To cope with such an issue, we consider adding

[5]The reason $\tau \geq T_{\text{delay}}$ is that the timer will timeout and set NACK before T_{delay} reaches τ.

one more mechanism to this network, Federated Learning (FL) [39], [40], which refers to learning a high-quality global model based on decentralized data storage for many nodes. Note we initialize the states of the IAB-nodes inside the same networks with the same size, thus we can use the FL algroithms to average the NN weights. FL has been shown to be a fast convergent method in distributed networks. We thereby propose our FL mechanism at the IAB donor side based on FedAvg [41].

V. Performance Evaluation

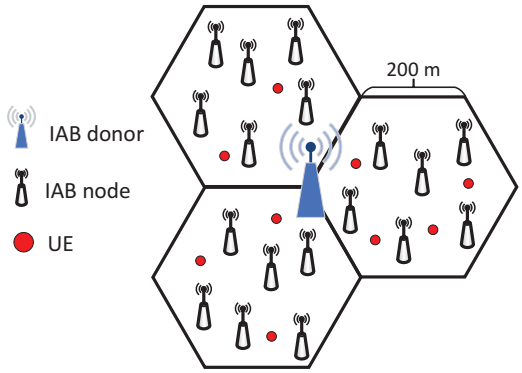


Fig. 4: Simulation Scenario.

In this section, we analyze the performance of our proposed methods. We conduct the simulation based on the homogeneous scenario (urban micro) based on 3GPP standard [19]. As shown in Fig. 4, we consider three hexagonal grids with the IAB-donor located in the center, and six IAB nodes are located inside each grid. We also use the 3GPP channel model in mmWave for the links among IAB nodes and UEs. The major parameters for the channel model are summarized in Table I. The UEs are dropped independently with uniform distribution and connected to the closest IAB nodes. The UEs randomly walk within its IAB nodes' coverage and move with a speed of 80% indoor (3km/h), 20% outdoor (30km/h) as suggested in the standard. We change the number of UEs to generate different traffic loads in the simulation. We adapt the FTP model 3 as the traffic model where the packet size is 0.1 Mbytes while the packet's arrival follows a Poisson distribution with a mean of 100/3 per second. In addition, the ratio of access DL/UL traffic is 4:1. We set the traffic type as the VR/AR traffic with the expected latency less than 5 ms and reliability of 0.999 successful rates as defined in the 3GPP standard [42].

For comparison, we implement a greedy algorithm characterized by Eq. (4)-Eq. (7) as the baseline. Note that the greedy algorithm is an extension of Semi-Persistent Scheduling algorithm defined in 3GPP standard [43], because greedy algorithm always tends to stick with the current transmission policy and adopt a new transmission policy only if some conditions fulfill. In the greedy algorithm, each IAB node selects the next-hop with the best channel condition, i.e., choosing the maximum MCS in the current transmission, and if one transmission fulfills the URLLC requirement, it keeps the transmission on the same nodes until the URLLC requirement will not be satisfied. Besides, we choose the Advantage Actor-Critic (A2C) method proposed in [17] to compare the enhancement of our proposed SAC algorithm. A2C approach can leverage merits of both value based approach and policy gradient and it empirically performs better than other similar RL approaches on coping with dynamic link blockages in a complicated IAB scenario, where each node selects a pattern (a set of links activated in parallel) according to the current policy, and then the links in this pattern are activated and enabled to transmit data.

TABLE I: Simulation Parameters in ns-3.

Parameter	Value
Power	23 dBm
Bandwidth	100 MHz
Channel model	3GPP mmWave channel model
Environment	3GPP Urban Micro (UMi)
UE receiver noise figure	10 dB
Numerology	3
Center frequency	28 GHz
Pathloss model	3GPP MmWave propagation loss model
BS receiver noise figure	7 dB
UE traffic model	FTP model 3

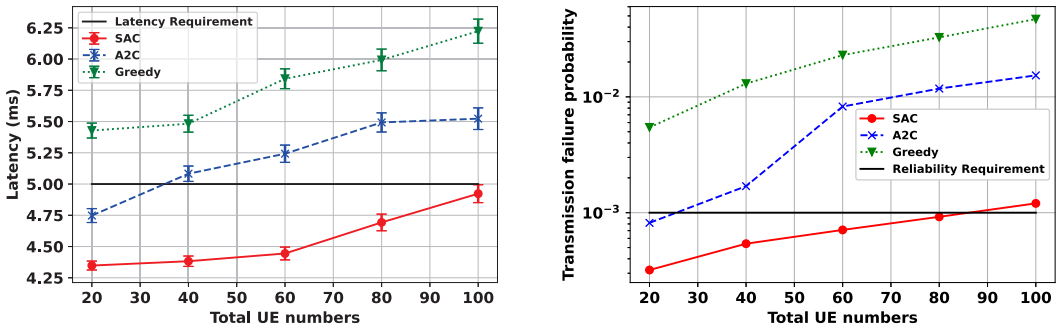
TABLE II: Main SAC hyperparameters.

Parameter	Value
Batch size	1024
Learning rate	1e-3
γ	0.99
Critic NN	(256, 1024, 1024, 256)
Actor NN	(128, 512, 512, 128)

Table. II summarized the main parameters for the SAC networks that we explored. For the other hyper-parameters, we use similar setups from the work [18]. The SAC algorithm is a feasible deep RL toward the real-world setup and is less sensitive to some hyperparameters. During our experiments, the convergence speed and performance of the SAC algorithm mainly depend on the neural network design and related training parameters. For example, we need a more extensive network for the Critic Network because it needs to learn to predict the values of different actions, which requires a more considerable learning ability. Thus we choose a sizeable NN setup to have a better generality in different scenarios. More complex models and other parameters can be explored in future work.

We first provide the algorithms' computational complexity and the average running time for each allocation, which are reflected in Table. III. In the row of computational complexity, N indicates the number of IAB nodes. In contrast, n indicates the maximum number between the total number of sub-carriers that all IAB nodes utilize (as the nodes in the input layer) and the number of nodes in the hidden layer in the neural network

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2022.3200673



(a) Average delay.

(b) Average transmission failure probability.

Fig. 5: Average simulation results of different UE numbers.

[6]. Note all the feedback like CSI and ACK is followed the 3GPP standard from the control link, so there is no additional overhead needed for the DRL algorithm. For the FL, the updates are around 1 minute with less than 2 Mbytes of data; thus, the overhead can be ignored.

TABLE III: Computational complexity.

Algorithm	Greedy	Expert	SAC	A2C
Computational complexity	O(N)	O($N \log N$)	O(n^2)	O(n^2)
10 IAB case (ms)	0.0323	0.0976	0.635	0.548
20 IAB case (ms)	0.0703	0.2413	1.141	1.211

Fig. 5 shows the simulation results of different numbers of UEs. The proposed SAC algorithm outperforms the other two algorithms from latency and reliability perspectives. As Fig. 5(a) shows, the average latency in the three algorithms increases with the increasing number of UEs. This is because the more the number of UEs is, the higher the queuing delay will be induced in T_{delay}^{dir} which is expressed in Eq. (1), leading to higher total latency. Note that both greedy and A2C methods do not meet the target latency requirement under the cases of a large number of UEs. However, our proposed SAC algorithm can always fulfill such a defined requirement, whatever the number of UEs is. Meanwhile, as is shown in Fig. 5(b), the average transmission failure probability with SAC is significantly lower than the other two algorithms, especially when the number of UEs is large. Notice that the proposed SAC algorithm is able to fulfill the target reliability requirement in most cases while the other two algorithms always fail. Besides, the failure probability in the three algorithms increases with the increasing number of UEs. With more UEs, IAB nodes are more likely to be scheduled to transmit at the same slot, causing a higher collision probability p_c. Thus the probability of a successful multi-hop transmission, $\mathbb{P}$ expressed in Eq. (3), will decrease accordingly. As a result, as the number of UEs increases, the average delay in Fig. 5(a) will be impacted by the average transmission failure probability since a successful multi-hop transmission is likely to need more retransmissions.

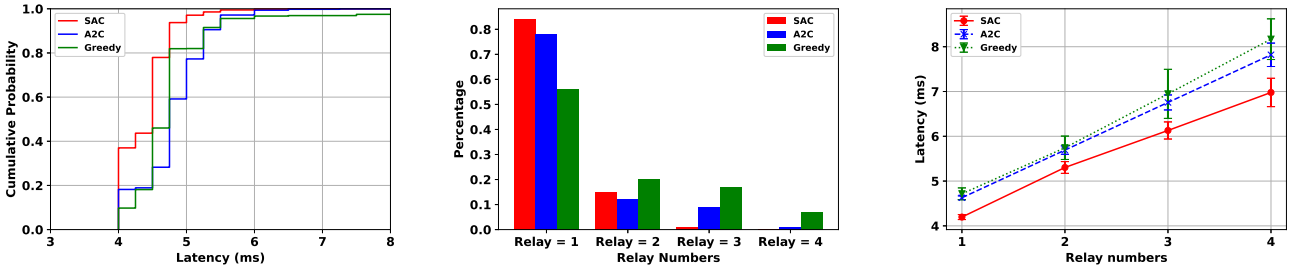
[6]For instance, if there are 10 IAB nodes in total, and the bandwidth size for packet transmission is 12 sub-carriers, i.e., the total number of sub-carriers that all IAB nodes utilize is 120. If the maximum number of neuron in the hidden layer is 512, then $n = \max\{120, 512\} = 512$.

We further explore the relevant performances under a fixed topology with a fixed number of UEs. Fig. 6(a) shows the cumulative distribution function (CDF) of latency for 40 UEs among the discussed algorithms. Since the given latency requirement is 5 ms, it is straightforward to see that almost 99% UEs can satisfy such a requirement under the proposed SAC. By contrast, the corresponding percentages are much lower under the other algorithms. Particularly, the CDF with the Greedy algorithm has a long tail, which indicates that some UEs may suffer from extraordinarily high latency larger than 8 ms. However, this issue is well coped with under the other two algorithms, where the corresponding highest latency is only around 6 ms and 6.5 ms under SAC and A2C, respectively.

Fig. 6(b) shows the distribution of the number of relay nodes used for multi-hop transmissions among three algorithms. Compared with the other two algorithms, the proposed SAC always utilizes fewer relay nodes to complete a multi-hop transmission. Particularly, the percentage of using one relay node with SAC is around 10% higher than that with A2C while around 30% higher than that with greedy. Meanwhile, the number of relay nodes with SAC is up to 3 while it reaches 4 in both greedy algorithm and A2C. From the latency perspective, the fewer the number of relay nodes is utilized, the lower aggregate transmission time and processing time that will be induced in the total delay expressed in Eq. (2), i.e., the smaller the expectation of n will be in Eq. (2). Therefore, with the same number of UEs, the average delay with SAC is always lower than that with the other two algorithms, which has been validated in Fig. 5(a). From the reliability perspective, fewer relay nodes also reduce the number of links in a multi-hop transmission, as a result, one packet transmission is less likely to be impacted by the collision probability or the BLER, which thereby increases $\mathbb{P}$ expressed in Eq. (3). The simulation results can also validate this regarding the average transmission failure probability in Fig. 5(b).

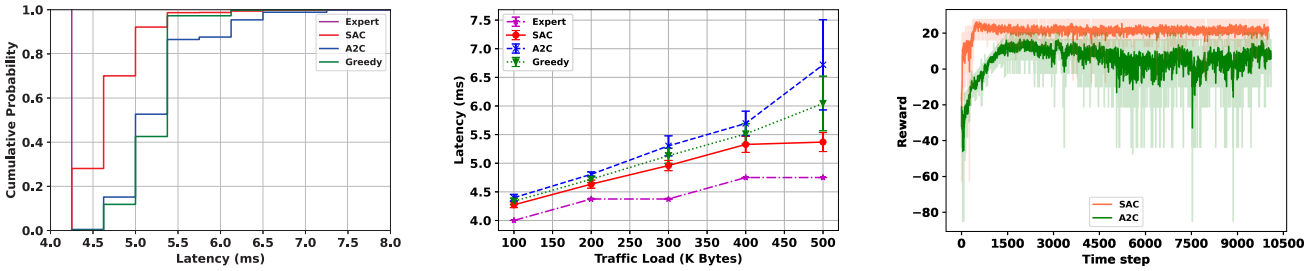
Fig. 6(c) illustrates how the number of relay nodes impacts the average delay under the same topology with the same number of UEs. As we can see, the average delay shows a linear increase with the increasing number of relay nodes among the three algorithms. This can be explained by the fact that the processing time and transmission time included

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2022.3200673



(a) Delay CDF. (b) Percentage of different relay nodes. (c) Delay for different relay numbers.

Fig. 6: Detailed simulation results of 40 UE numbers.



(a) Delay CDF for 300 Kbyte load. (b) Various traffic load. (c) Reward for 300 Kbyte load.

Fig. 7: Simulation results under a fixed topology with 40 UEs and different traffic loads.

in the total delay, as Eq. (2) suggests, is proportional to the number of relay nodes. Notice that although the increase of relay nodes also increases the average delay due to a higher failure probability, the increment of failure probability is not so sensitive, indicating that the increment of average delay is dominated by the processing time and transmission time when a node is added. Besides, SAC always outperforms the other two algorithms on the average delay, regardless of the number of relay nodes. Therefore, the SAC algorithm is more capable of learning from the environment and making stochastic decisions by regarding the entropy as part of the reward in the SAC algorithm. This adapts to different environments and handles the decision-making procedure.

More details under a fixed 40 UEs and different traffic loads are shown in Fig. 7. With the fixed topology, we can then use the graph model proposed in Sec III-C to give the expert solution. We allocate the traffic among the users proportional to the traffic load requiring among the IAB nodes and UEs for both DL and UL. Fig. 7(a) shows the cumulative distribution function (CDF) of delay for traffic load 300 Kbytes among four algorithms. The CDF with SAC grows faster than the other two algorithms, indicating that SAC outperforms the other two algorithms due to the higher ratio of low-latency packets. Particularly, the probability of the delay $\leq$ 5 ms reaches around 95% with SAC, while the corresponding probability with the other two algorithms only reaches around 50%. Note that the CDF with A2C has the longest tail; its variance is thereby the largest among them. Subsequently, we investigate the latency regarding the traffic load, which is shown in Fig. 7(b). Due to the linear increment of traffic load, the latency with each algorithm shows a linear growth in the region of low traffic loads, where the latency with SAC is always the lowest. However, when traffic load is 500 Kbytes, the latency with both A2C and greedy algorithms grows faster because a large traffic load also incurs the increase of collision probability that impacts the total delay.

Fig. 7(c) shows the comparison between SAC and A2C on the reward convergence for traffic load 10 Kbytes. We conducted such a simulation with 10000-time steps to ensure that the returned reward could converge in both algorithms. As we can see, SAC outperforms A2C on the reward convergence in three aspects: 1) Convergence speed: SAC converges quite faster than A2C. More precisely, SAC takes around 700-time steps while A2C takes around 2100 time steps to reach the convergence; 2) Steady reward value: the steady reward value with SAC is at least 30% larger than that with A2C; 3) Stability: the returned reward trend with SAC is more stable than that with A2C after 2100 time steps. Therefore, the stochastic policy generated by the SAC algorithm enhances the ability to transfer knowledge compared with the A2C algorithm. Besides, the SAC algorithm's objective also encourages exploring more possible actions that contribute to faster convergence.

Afterward, we study the impact of FL on the convergence speed of SAC and A2C. As Fig. 8 shows, the reward in both algorithms with FL converges faster than that without FL. In particular, while the convergence speed of two algorithms without FL are quite close, SAC with FL takes around 2500 time steps fewer than that with SAC without FL, and A2C with FL take around 2000 time steps fewer than that with A2C

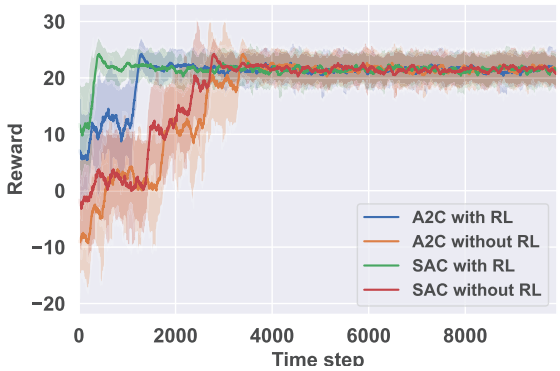


Fig. 8: FL and the convergence speed [7].

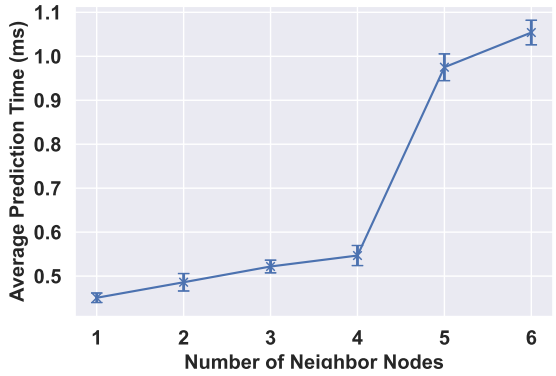


Fig. 9: The average running time.

without FL, indicating that SAC achieves more improvement on A2C after adding FL. Hence the reward with SAC with FL still outperforms A2C with FL. This indicates that The structure of the entropy-based reinforcement learning with federated learning has the potential to be implemented in the radio intelligence controller in 5G and beyond networks.

Furthermore, we implement the SAC algorithm through PyTorch to measure the runtime of selecting actions for a given state. We run such a simulation with the mean packet size is 30K Bytes. The simulation results are shown in Fig. 9. As the number of neighboring nodes increases, the average runtime increases slowly at first. Since the policies are running parallel, the increased time is not so significant based on the comparison of the action score. After the number exceeds 4, however, the running time doubles due to the limitation of the computation resources; thus the policies are no longer running parallel. Note that the algorithm runtime is approximately around 0.5 ms for the case of 1 neighboring node. This performance may be impacted by the efficiency of PyTorch and the computing ability of the computer. The efficiency and computing ability can be further improved in a real system where the proposed algorithm can be implemented on the hardware and software.

[7]The standard deviation over all time steps are 92.96, 74.17, 16.87, and 32.03, respectively (aligning with the same order of algorithms in Fig. 8); while the standard deviation after the convergence point (around 4500) for all algorithms: 12.73, 11.61, 10.95, and 11.34, respectively.

VI. Conclusion

In this paper, we focused on the cross-layer modeling on the routing and resource allocation in the multi-hop IAB network under the latest 5G NR standard. An optimal routing problem that minimized the transmission latency and also satisfied the transmission reliability constraint was first formulated and analyzed. Subsequently, we presented a DRL framework to solve the proposed routing and resource allocation problem in the IAB network based on the local information. Afterwards, we proposed a novel entropy based reinforcement learning algorithm with federated learning mechanism to accelerate the convergence speed as well as decrease the algorithm complexity. The numerical results showed that our proposed algorithm outperformed the existing algorithms on the aspects of latency and reliability from different perspectives.

Since this work provided a general solution to the investigated joint optimization problem in the IAB networks regardless of network topology, this work also hints at the effectiveness of mesh-based communication in IAB networks while it is not yet considered by 3GPP IAB standardization. In the future, it would be interesting to explore how effective the proposed algorithm can be on the meshed IAB networks after such an IAB architecture is standardized. Besides, we would like to consider the industrial aspects of IAB with more realistic setups, i.e, limiting the maximum hops and testing under the open source test-beds.

Appendix

A. Proof of Lemma 1

Proof: When $q^*(\mu) = \sigma$, we have $\log \mathbb{P}(q^*(\mu)) = \log \sigma$. For any other $q^{'} \in \mathcal{Q}$ that satisfies $\mathbb{P}\left(q^{'}(\mu)\right) \geq \sigma$, we have by applying Eq. (5),

$$\begin{aligned} T_{\text{delay}}(q^*(\mu)) - \mu \log \mathbb{P}(q^*(\mu)) &\leq T_{\text{delay}}\left(q^{'}\right) - \mu \log \mathbb{P}\left(q^{'}\right) \\ &\leq T_{\text{delay}}\left(q^{'}\right) - \mu \log \mathbb{P}(\sigma). \end{aligned} \tag{17}$$

From Eq. (17), we then obtain that $T_{\text{delay}}(q^*(\mu)) \leq T_{\text{delay}}\left(q^{'}\right)$. Therefore $q^*(\mu)$ is the optimal solution for the original Problem 1. ■

B. Bisection Approach

The Bisection approach to find the optimal μ^* is shown in Algorithm 1.

Algorithm 1 Bisection approach to find μ^*.

1: $\mu_{\text{low}} = \mu = 0$, $\mu_{\text{up}} = \text{MAX\_MU}$
2: **while** $(\mathbb{P}(q^*(\mu)) < \sigma)$ or $(\mathbb{P}(q^*(\mu)) > \sigma + \eta_0)$ **do**
3: **if** $(\mathbb{P}(q^*(\mu)) < \sigma)$ **then** $\mu_{\text{low}} \leftarrow \mu$
4: **else** $\mu_{\text{up}} \leftarrow \mu$
5: $\mu \leftarrow \frac{\mu_{\text{low}} + \mu_{\text{up}}}{2}$
6: return $\mu^* \leftarrow \mu$

REFERENCES

[1] H. Yin, L. Cao, and X. Deng, "Scheduling and resource allocation for multi-hop URLLC network in 5G sidelink," in *Proc. IEEE 94th Veh. Technol. Conf. (VTC2021-Fall)*. IEEE, 2021, pp. 1–7.

[2] M. Series, "IMT vision–framework and overall objectives of the future development of IMT for 2020 and beyond," *Recommendation ITU*, vol. 2083, p. 0, 2015.

[3] B. Zhang, F. Devoti, I. Filippini, and D. De Donno, "Resource allocation in mmwave 5G IAB networks: A reinforcement learning approach based on column generation," *Comput. Netw.*, p. 108248, 2021.

[4] M. Pagin, T. Zugno, M. Polese, and M. Zorzi, "Resource management for 5G NR integrated access and backhaul: a semi-centralized approach," *arXiv:2102.09938*, 2021.

[5] S. M. A. Zaidi, M. Manalastas, H. Farooq, and A. Imran, "Mobility management in emerging ultra-dense cellular networks: A survey, outlook, and future research directions," *IEEE Access*, vol. 8, pp. 183 505–183 533, 2020.

[6] O. P. Adare *et al.*, "Uplink power control in integrated access and backhaul networks," in *Proc. IEEE Int. Symp. Dyn. Spectr. Access Netw. (DySPAN)*. IEEE, 2021, pp. 163–168.

[7] C. Hoymann, W. Chen, J. Montojo, A. Golitschek, C. Koutsimanis, and X. Shen, "Relaying operation in 3GPP LTE: challenges and solutions," *IEEE Commun. Mag.*, vol. 50, no. 2, pp. 156–162, 2012.

[8] J. M. B. da Silva, G. Fodor, and T. F. Maciel, "Performance analysis of network-assisted two-hop D2D communications," in *Proc. IEEE Glob. Commun. Workshops (GC Wkshps)*. IEEE, 12/7/2014 - 12/11/2014, pp. 1050–1056.

[9] J. Huang, Y. Liao, C.-C. Xing, and Z. Chang, "Multi-hop D2D communications with network coding: From a performance perspective," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2270–2282, 2019.

[10] C. Saha, M. Afshang, and H. S. Dhillon, "Bandwidth partitioning and downlink analysis in millimeter wave and backhaul for 5G," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8195–8210, 2018.

[11] G. Yang, M. Haenggi, and M. Xiao, "Traffic allocation for low-latency multi-hop networks with buffers," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 3999–4013, 2018.

[12] Y. Liu, A. Tang, and X. Wang, "Joint incentive and resource allocation design for user provided network under 5G and backhaul networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 2, pp. 673–685, 2020.

[13] C. Zhong *et al.*, "A Deep Actor-Critic Reinforcement Learning Framework for Dynamic Multichannel Access," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 4, pp. 1125–1139, 2019.

[14] H. Yang and X. Xie, "An actor-critic deep reinforcement learning approach for transmission scheduling in cognitive internet of things systems," *IEEE Syst. J.*, vol. 14, no. 1, pp. 51–60, 2019.

[15] H. Yang *et al.*, "Intelligent resource management based on reinforcement learning for ultra-reliable and low-latency IoV communication networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4157–4169, 2019.

[16] M. Alsenwi *et al.*, "Intelligent resource slicing for eMBB and URLLC coexistence in 5G and beyond: A deep reinforcement learning based approach," *arXiv:2003.07651*, 2020.

[17] B. Zhang, F. Devoti, and I. Filippini, "RL-based resource allocation in mmwave 5G IAB networks," in *Proc. Mediterr. Commun. Comput. Netw. Conf. (MedComNet)*, 2020, pp. 1–8.

[18] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018.

[19] 3GPP, "Study on Integrated Access and Backhaul," The 3rd Generation Partnership Project, Tech. Rep. TR38.874, Dec 2018.

[20] ——, "NG-RAN; architecture description," The 3rd Generation Partnership Project, Tech. Rep. TR38.401, Oct 2021.

[21] M. Polese, M. Giordani, T. Zugno, A. Roy, S. Goyal, D. Castor, and M. Zorzi, "Integrated access and backhaul in 5G mmwave networks: Potential and challenges," *IEEE Commun. Mag.*, vol. 58, no. 3, pp. 62–68, 2020.

[22] C. Madapatha *et al.*, "On integrated access and backhaul networks: Current status and potentials," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 1374–1389, 2020.

[23] C. Madapatha, B. Makki, A. Muhammad, E. Dahlman, M.-S. Alouini, and T. Svensson, "On topology optimization and routing in integrated access and backhaul networks: A genetic algorithm-based approach," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 2273–2291, 2021.

[24] A. HasanzadeZonuzy, I.-H. Hou, and S. Shakkottai, "Broadcasting real-time flows in integrated backhaul and access 5G networks," in *Proc. Int. Symp. Model. Optim. Mobile, Ad Hoc, Wireless Netw. (WiOPT)*. IEEE, 2019, pp. 1–8.

[25] B. Zhai, M. Yu, A. Tang, and X. Wang, "Mesh architecture for efficient integrated access and backhaul networking," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*. IEEE, 2020, pp. 1–6.

[26] Y. Xu, W. Xu, Z. Wang, J. Lin, and S. Cui, "Load balancing for ultradense networks: A deep reinforcement learning-based approach," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9399–9412, 2019.

[27] X. Meng, H. Inaltekin, and B. Krongold, "Deep reinforcement learning-based topology optimization for self-organized wireless sensor networks," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[28] X. Chen, R. Proietti, H. Lu, A. Castro, and S. B. Yoo, "Knowledge-based autonomous service provisioning in multi-domain elastic optical networks," *IEEE Commun. Mag.*, vol. 56, no. 8, pp. 152–158, 2018.

[29] W. Lei, Y. Ye, and M. Xiao, "Deep reinforcement learning-based spectrum allocation in integrated access and backhaul networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 3, pp. 970–979, 2020.

[30] Q. Cheng *et al.*, "Deep reinforcement learning-based spectrum allocation and power management for IAB networks," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Wkshps)*. IEEE, 2021, pp. 1–6.

[31] 3GPP, "Physical Channels and Modulation," The 3rd Generation Partnership Project, Tech. Rep. TR38.211, Jan 2020.

[32] Samsung, "4G-5G Interworking White Paper," Tech. Rep., Jul 2017.

[33] H. Yin, L. Zhang, and S. Roy, "Multiplexing URLLC traffic within eMBB services in 5G NR: Fair scheduling," *IEEE Trans. Commun.*, pp. 1–1, 2020.

[34] 3GPP, "Feasibility study for further advancements for E-UTRA (LTE-Advanced)," The 3rd Generation Partnership Project, Tech. Rep. TR36.912, Jul 2020.

[35] ——, "Physical layer procedures for data," The 3rd Generation Partnership Project, Tech. Rep. TR38.214, Jun 2021.

[36] ——, "Study on channel model for frequencies from 0.5 to 100 GHz," The 3rd Generation Partnership Project, Tech. Rep. TR38.901, Jan 2020.

[37] S. Rayadurgam and Y. Lei, *Communication Networks: an Optimization, Control, and Stochastic Networks Perspective*. Cambridge Univ. Press, 2013.

[38] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[39] M. M. Wadu, S. Samarakoon, and M. Bennis, "Joint client scheduling and resource allocation under channel uncertainty in federated learning," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 5962–5974, 2021.

[40] J. Konečnỳ *et al.*, "Federated learning: Strategies for improving communication efficiency," *arXiv:1610.05492*, 2016.

[41] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Statist.*, 2017, pp. 1273–1282.

[42] 3GPP, "Study on scenarios and requirements for next generation access technologies," The 3rd Generation Partnership Project, Tech. Rep. TR38.913, Jul 2020.

[43] ——, "Physical layer procedures," The 3rd Generation Partnership Project, Tech. Rep. TR36.213, Dec 2020.

Hao Yin (Student Member, IEEE) received his B.E. from Huazhong University of Science and Technology in 2019. He is currently pursuing a PhD in Electrical & Computer Engineering from University of Washington. His research focuses on the scheduling and resource allocation algorithms in wireless communication, especially for the next generation 5G and Wi-Fi systems. He also works on applying machine learning algorithms to complex wireless systems to build more intelligent wireless system.

Sumit Roy (Fellow, IEEE) received the B. Tech. degree from the Indian Institute of Technology (Kanpur) in 1983, and the M. S. and Ph.D. degrees from the University of California (Santa Barbara), all in Electrical & Comp. Engineering in 1985 and 1988 respectively, as well as an M. A. in Statistics and Applied Probability in 1988. He was appointed to Integrated Systems Professor (2014-19) of Electrical & Computer Engineering, Univ. of Washington-Seattle where his research and technology transition interests have included design and evaluation of wireless communication and sensor network systems with an emphasis on 5G & beyond technologies, multi-standard inter-networking and coexistence using software-defined networking approaches. He spent 2001-03 at Intel Wireless Technology Lab as a Senior Researcher engaged in systems architecture and standards development for ultra-wideband systems (Wireless PANs) and next generation high-speed wireless LANs. He has been active in IEEE Communications Society in various roles (journal editor and Distinguished Lecturer) and was elevated to IEEE Fellow (2007) for "contributions to multi-user communications theory and cross-layer design of wireless networking standards". He served 2 terms as (elected) member of Executive Committee, National Spectrum Consortium dedicated to efficient spectrum sharing between Federal and commercial networks and is the co-author of IEEE TAES 2016 Best paper award for work on Radar-Comm coexistence. He is presently Program Lead for Innovate Beyond 5G for OUSD R&E 5G-to-xG initiative from 07/2020 https://www.cto.mil/5g/.

Liu Cao (Student Member, IEEE) received the B.E. degree in Electrical Engineering from Jinan University, Guangzhou, China, in 2017, and the M.S. degree in Electrical Engineering from Northwestern University, Evanston, IL, USA, in 2019. He is currently pursuing the Ph.D. degree in Electrical & Computer Engineering at the University of Washington, Seattle, WA, USA. His research interests include 5G NR, V2X and machine learning for wireless communications.