Delay-Guaranteed Path Selection and Scheduling in IAB Networks

Pooria Seyed Eftetahi, *Student Member, IEEE,* Lin Cai, *Fellow, IEEE,* Xiangyu Ren, *Student Member, IEEE*

***Abstract*—Integrated access and backhaul (IAB) is a promising solution to improve coverage at low deployment costs. In IAB networks, due to wireless channel variations, guaranteeing delay for delay-sensitive applications is a major challenge. Given the random traffic arrivals and channel variations, using a central controller for packet-level delay management becomes infeasible due to the added delay from the central controller. In this paper, we propose a distributed cross-layer method to provide delay-guaranteed path selection and scheduling for the IAB network where priority queues and weighted round robin are adopted to deliver differentiated services. Our goal is to determine the optimal path and scheduling decisions to maximize the total utility of the IAB network. We deploy an iterative approach in a distributed manner to solve the maximization problem at each IAB-node. Through simulation, we show that our proposed solution guarantees the delay at the packet-level while achieving considerable gains in terms of delay and packet delivery ratio compared to the state-of-the-art.**

***Index Terms*—Integrated Access and Backhaul (IAB), path selection, scheduling, delay-guarantee, renewal optimization**

I. Introduction

THE amount of mobile traffic worldwide is anticipated to increase 670 times from 2010 by 2030 [1]. One strategy to address the increased demand is through network densification, which has the potential to substantially improve both network capacity and coverage [2]. However, this approach encounters major drawbacks such as scalability and cost. One solution to address them is the deployment of wireless backhaul, known as integrated access and backhaul (IAB) [3]. For 5G NR, IAB has been standardized and recognized as a cost-effective alternative to wired backhauling [4].

Future wireless systems, such as 5G Beyond and 6G, are anticipated to support global connectivity [5] as there is a growing demand for providing reliable communication services in remote and rural regions. Opting for wireless backhaul instead of wired optical fiber presents an economically appealing solution to extend connectivity to these underserved areas and facilitate deployment.

To cover large-scale remote areas, deployment of IAB with multi-hop backhauling can provide a flexible range extension and deployment [6]. Also, it enables wireless backhauling around obstacles [7]. The 3rd Generation Partnership Project (3GPP) Rel-16 supports multi-hop topologies based on spanning trees (ST) and directed acyclic graphs (DAG) [8]. While multi-hop IAB networks improve the coverage, it would increase the network latency due to added hops. Increasing the number of hops introduces other challenges including path selection, scheduling, and ensuring latency guarantees.

Latency is a crucial Quality of Service (QoS) requirement in many emerging applications such as augmented reality (AR), virtual reality (VR), and real-time control of IoT. With the prevailing best-effort Internet services, the transmission of packets is susceptible to issues like packet loss and random latency, mainly due to factors like network congestions [9]. Furthermore, the wireless nature of IAB networks can exacerbate the issue. The best-effort design in existing network routing protocol cannot deal with the diverse needs of data-sensitive applications and the wireless channel nature, making it undesirable for delay-sensitive applications.

It is imperative to incorporate a prioritization mechanism within networks, establishing a framework for Differentiated Services (DiffServ) that not only caters to the distinct requirements of various applications but also enhances overall network resource efficiency. In this regard, IAB networks should differentiate the delay-sensitive (DS) packets from non-delay-sensitive (NDS) packets, and guarantee the required latency for DS packets. However, employing priority queues poses new challenges such as managing queues, scheduling paths, and selecting routes. The path selection and queue management algorithm should consider both network capacity and traffic dynamics to ensure delay-guaranteed services.

Researchers have extensively explored ways to ensure reliable, delay-guaranteed services for delay sensitive applications by tackling routing and scheduling challenges in wired networks. Deterministic network (DetNet) standards led by the IETF DetNet working group seek to offer delay-guaranteed services based on predetermined flows by investigating explicit data paths [10]. Each deterministic flow's data traffic is transported with assured delay and minimal variation in delay through the allocation of reserved resources in a centralized manner [11]. However, in IAB networks, the throughput of individual links undergoes constant changes due to wireless channel variations. These fluctuations can impact the latency experienced on each link. In addition, resource reservation methods for traffic with random arrival rates have led to low efficiency. Flow-based traffic engineering solutions lack the capability to ensure packet-level delay guarantees and they cannot leverage the capacity of available paths to balance the loads and manage congestion.

The problem of routing for delay-guaranteed service can be formulated as an optimization problem with the objective of maximizing the number of in-time delivered packets. Optimizing routing for delay-guaranteed services in wireless networks, particularly in IAB networks is an open issue. Many existing optimization solutions encounter scalability issues, especially in dynamic environments like IAB networks, where the wireless channel undergoes constant changes.

Motivated by the above challenge, We focus on the backhaul routing and scheduling in the IAB network, which does not include the radio access part. We propose a scheduling and path selection solution named Delay-Guaranteed Path selection and Scheduling (DGPS) considering the delay budget of packets, queuing delay, and varying transmission rates of wireless links. We refer to this procedure as the forwarding decision. To achieve an effective forwarding decision, DGPS utilizes both static network topology and link information, and dynamic traffic and wireless channel information. To ensure fast response to traffic dynamics and channel changes, DGPS adopts a distributed approach. As a result, each IAB-node will be able to adapt its forwarding decision according to real-time observation of the traffic, queue, and wireless channel condition.

The main contributions of this paper are three-fold:

- First, to ensure the E2E packet-level delay, we adopt a cross-layer approach that involves the collaboration of link layer queue management and routing sublayer. We formulate our forwarding decision problem as a utility maximization problem considering the latency constraint of each packet.
- Second, we propose a distributed approach to solve the mentioned problem. Each IAB-node runs the proposed algorithm to select the optimal forwarding decision for each packet by considering the delay requirement, current queue length, and the throughput of each wireless link. The proposed algorithm maximizes the reward using an iterative method
- Third, to verify the performance of the proposed solution, extensive simulations have been conducted. Simulation results show that DGPS can support delay-guaranteed services, and improve the packet delivery ratio compared to the existing state-of-the-art.

The rest of the paper is organized as follows. In Section II, we introduce the related work and IAB architecture. In Section III, the system model and problem formulation are described. Section IV presents the proposed DGPS solution. Section V presents the simulation results, followed by the concluding remarks and further research issued in Section VI.

II. Related Work and IAB architecture

A. Related work

Recently, IAB networks have attracted much attention, particularly given the deployment of millimetre waves (mmWave). There are several works [12]–[14] which have deployed stochastic geometry to model and analyze wireless backhaul systems. For instance, Saha et al. [12] established an analytical framework to characterize the rate coverage probability by leveraging stochastic geometry tools. In addition, there are several studies focusing on resource allocation problem [15]–[17]. Reference [17] suggested a distributed stochastic way to jointly solve the problem of resource allocation and path selection in DAG multi-hop multi-path IAB networks. They deployed stochastic optimization tools to solve the problem. They found the problem's deterministic equivalent problem by replacing the random variables with their expected values.

A limited number of studies have investigated path selection, scheduling, or their combination with other aspects of the network. In [18], authors developed multi-hop multipath scheduling to perform path selection and rate allocation using stochastic optimization. They used reinforcement learning to find the best path and converted the non-convex rate allocation to a convex problem using a successive convex approximation approach.

An optimization framework which can consider traffic demand variations by employment of constraints and multiple relaxations was developed in [19]. Nevertheless, the mentioned solution necessitates manual intervention from experienced experts and involves time-consuming iterations. This approach becomes impractical in situations where IAB-nodes demand real-time results and We experience fluctuations in the wireless channel. Authors in [20] investigated the routing problem of IAB networks with the goal of minimizing the latency while meeting the reliability requirement using entropy-based reinforcement learning with federated learning.

In [21] an outdoor in-band mesh architecture for the IAB network was explored. The research focused on incorporating a joint routing and power control mechanism, both centralized and semi-centralized, under the constraints of Quality of Service (QoS). This approach is implemented through a logical controller that oversees network operations at the MAC layer.

In [22], authors developed a scheduling scheme in a multi-hop network to minimize the end-to-end (E2E) delay. They considered two scenarios: 1) with the global knowledge of queues and 2) limited feedback of queues at each frame; They solved the problem of sectorizing and interference by using conflict graphs. They used a deep deterministic policy gradient algorithm from reinforcement learning to learn the delay-minimizing scheduling policy.

[23] proposed a risk-averse reinforcement learning approach for IAB mmWave networks to enhance reliability. [24] introduced a game-theoretic approach where access points (APs) selfishly optimize their own latency. Using the Shapley value method, achieving improved system performance in terms of delay and fairness. However, their focus is on reducing the latency and not providing any form of guarantee.

Majority of previous studies have attempted to execute path selection and scheduling in IAB networks with an emphasis on increasing throughput or reducing latency, while guaranteeing the delay remains an open issue. Also, due to the vast number of parameters used in optimization, the majority of the suggested optimization solutions are rather complex, and not desirable for DS packet delivery. Furthermore, delay-guaranteed routing and scheduling should employ a distributed method instead of depending on a centralized controller to ensure fast response to network dynamics. In this way, every IAB-node would be able to make adaptive decisions based on real-time observations of queue status and channel situation. This approach can reduce complexity and overhead, leading to a quicker and more efficient decision-making process.

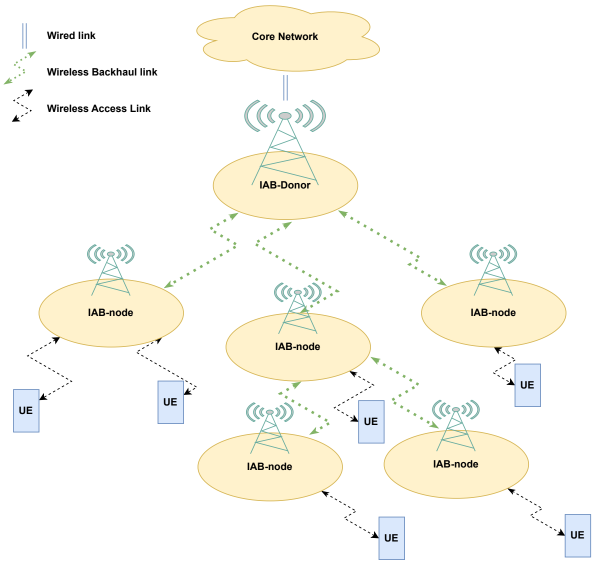


Fig. 1: 3GPP's illustration of IAB architecture diagram in Stand-Alone mode.

B. IAB architecture and protocol stack

3GPP initially introduced a study item related to wireless backhaul to handle the increasing demand in data rate, referred to as LTE relaying [8]. However, due to the lack of spectrum, it did not attract enough attention.

One fundamental aspect of 5G involves leveraging high-frequency transmission carriers, such as millimeter-wave (mmWave), to enable the utilization of a potentially larger spectrum. Although mmWave can achieve a higher transmission data rate, it limits the coverage area of each Base station (BS) due to faster decay of signals over distance, and more severe shadowing effect. Therefore, there is a requirement for a denser deployment of BSs. Nevertheless, while the deployment of IABs is beneficial, creating a high-performing IAB network remains an ongoing challenge [25].

As illustrated in Fig. 1, an IAB network consists of two components: IAB-node(s) and IAB-donor. According to 3GPP technical report [7], IAB-node refers to a BS which enables wireless access for User Equipments (UEs) while also wirelessly backhauling the associated access traffic. IAB-donor is defined as the BS that provides the connection between UEs and the core network while also providing wireless backhauling capabilities to IAB nodes.

In this structure, each IAB node is equipped with a distributed unit (DU) and a mobile-termination (MT). The MT enables the IAB node to establish connections with an upstream IAB node or the IAB donor. Through the DU, the IAB node establishes radio link control (RLC) channels to UEs and the MTs of downstream IAB nodes. The IAB-donor is also equipped with a DU to provide support for UEs and MTs of downstream IAB nodes.

Additionally, the IAB donor is equipped with a centralized unit (CU) that oversees the DUs of all IAB nodes, including its own DU. The CU can be divided to two parts: the control plane (CU-CP), which is responsible for hosting the radio resource control and the control plane part of the packet data convergence protocol (PDCP) [4]; and the user plane (CU-UP), which hosts the user plane part of both the PDCP and the service data adaptation protocol (SDAP) [4]. The assumption is made that the DUs on an IAB node are exclusively served by a single IAB donor [7]. The DU part of IAB-donor encompasses the RLC, Medium Access Control (MAC), and Physical layer (PHY) protocols.

The 3GPP introduced five different architecture diagrams which are different in interfaces and additional functionality to perform multi-hop forwarding. A promising architecture is 1a [25], which is depicted in Fig. 2. This architectural design utilizes the CU/DU split architecture and incorporates an adaptation layer for hop-by-hop forwarding and backhauling [7]. The IAB donor's CU/DU functional split is motivated by the concentration of time-sensitive tasks, such as scheduling, and fast re-transmission in the DU, strategically positioned in proximity to the radio part. Simultaneously, it facilitates the centralization of less critical, time-sensitive radio functionalities within the CU [26].

As illustrated in Fig. 2, the network layer in IAB-nodes is replaced with a sublayer named backhaul adaptation protocol (BAP). It replaces the IP functionality of the standard F1-stack [7]. Within the IAB-node, the BAP sublayer comprises a BAP entity at the MT function and one separate but collocated BAP entity at the DU function. on the IAB-donor-DU, the BAP sublayer encompasses only a single BAP entity [27]. Each BAP entity consists of both a transmitting and a receiving part. BAP is responsible for routing of the next hop, Determination of BAP destination and path for packets from the upper layer, data transfer, etc [27].

The IAB-donor allocates different L2 addresses referred to as BAP addresses to each controlled IAB node. Once it is initialized, the IAB-donor becomes aware of the IAB-nodes within its network. Consequently, under the IAB donor's global knowledge, each IAB node can ascertain both the total number of IAB-nodes and its neighboring IAB-nodes within the network. The BAP header used in transmission between IAB nodes and the IAB donor carries source ID, destination IDs, and the optional path ID. Each IAB node maintains an individual routing table specifying the next hop identifier for every BAP ID. Once an IAB-node receives a packet, the BAP sub-layer checks the destination ID. If the IAB-node is the destination, it will be forwarded to higher layers. otherwise, it will be forwarded to DU part of the IAB to transmit the packet based on its forwarding table.

III. System Model

A. Network Deployment

As illustrated in Fig. 3, we considered three tiers of BSs in the IAB network, including one macro base station as the IAB-donor which has wired backhaul to the core, and a set of small base stations as IAB-nodes $\mathcal{B} = \{1, ..., B\}$ which connect to neighbor nodes using wireless backhaul. The first tier of the IAB network consists of the IAB donor; the second tier consists of IAB-nodes which have a direct link to the IAB-donor and the last tier is the IAB-nodes which do not have

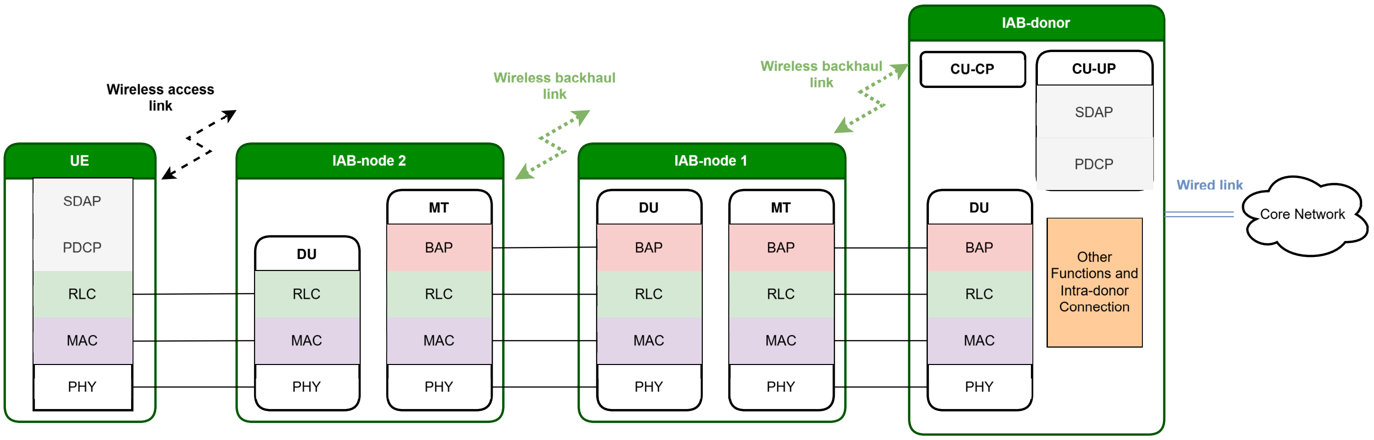


Fig. 2: Example of the protocol stack for UE access with the BAP layer in 3GPP's architecture 1a, responsible for managing routing among IAB-nodes.

TABLE I: Key notations

Symbol	Description
$\mathcal{B}$, $\mathcal{L}$	Set of IAB-nodes and Wireless links
$\mathcal{B}_i$	Set of the neighbours of IAB-node i
Ω_i	Set of feasible paths from IAB-node i
$\mathcal{H}_i$	Set of forward decisions
q	Queue type
$b^q_{i,j}$	Buffer size of IAB-node i to j of queue q
T_{start}, t_{DB}	Generation time and Delay Budget of the packet
$h^q_{i,j}$	Forward decision at IAB-node i to j using queue q
TB, $pktsz$	Transport Block and Packet size
μ	Numerology
t_I	Transmission time interval
$T^{tr}_{i,j}$, $T^q_{i,j}$	Transmission and Queuing Delay from i toward j
w_q	Round Robin weight for queue q
$Q^q_{i,j}$, $T^{u,q}_{i,j}$	Queue length and upper bound of q from i toward j
$T_{hop}(r_j)$	Per hop delay budget
$G^q_{i,j}[t]$	Achieved reward

a direct link to the IAB-donor. The network is modelled as a graph $\mathcal{G} =< \mathcal{B}, \mathcal{L} >$. Here, $\mathcal{L}$ is the set of wireless links among BSs.

Let $\mathcal{B}_i$ be the set of the neighbours of IAB-node i with size B_i. Assume that there is a directive wireless link between each IAB-node and its neighbour using directional antennas. Each IAB-node can simultaneously communicate with multiple IAB-nodes using hybrid beamforming and multi-user MIMO techniques. Spatial multiplexing allows multiple signals to be separated in the spatial domain [28]. Also, IAB-nodes are in in-band operational mode which results in a half-duplex mode. The in-band operation is subject to a half-duplex limitation, meaning that the IAB MT component within an IAB-node cannot receive while its corresponding DU is transmitting. Similarly, vice versa holds true to prevent intra-site interference [20]. In addition, we assume perfect beam alignment and noise-limited system due to the deployment of mmWave band which reduces the interference in neighbour IAB-nodes [1] [29].

B. Queue Model and Delay Budget

We consider the downlink transmission of DS and NDS packets with different delay requirements. In each IAB-node, two queues are deployed at each port to offer DiffServ, a delay-guaranteed queue (DGQ) denoted by $q = 1$, and a best-effort queue (BEQ) by $q = 2$.

Let the buffer size at the output port of IAB-node i toward IAB-node j of each queue be denoted by $b^q_{i,j}$. Given $b^q_{i,j}$ and assuming that the link throughput remains stable over a small period of time, an IAB-node can determine the upper bound of queuing delay by dividing the buffer length over the minimum throughput associated with the queue. For instance, if the buffer size is 96 Kb and the throughput is 1.5 Mbps, then the queuing delay upper bound would be 64 milliseconds.

Since wireless backhaul links of IAB-nodes are susceptible to blockage, e.g. moving objects and seasonal change [20], the throughput of each link changes over time which alters the delay upper bound. We assume each IAB-node periodically broadcasts the delay upper bound of its queues to its upstream IAB-nodes, which are along the paths toward the IAB-donor. The frequency of these broadcasts depends on network dynamics, such as physical channel conditions. Typically, the broadcast interval aligns with the channel coherence time, which ranges from tens to hundreds of milliseconds [30], ensuring timely updates while minimizing network overhead.

To guarantee the per-hop delay upper bound of each DGQ and prevent BEQ starvation, the weighted round robin (WRR) scheduling method is adopted [31]. WRR is a scheduling algorithm that allocates different portions of service time to queues based on predefined weights. Queues with higher weights receive more time or resources, allowing them to process more data, while still ensuring all queues are served in a round-robin manner. Thus, each queue can be ensured to have a minimum portion of link resource based on its weight when it is non-empty, and the link resource can be used by others if the queue is empty.

In the WRR scheduling approach, each queue is allocated a portion of the service time which is its weight. The buffer size and weight assigned to each queue can determine the maximum queueing delay of the corresponding queue. For

[1] Moreover, to avoid loop problem we only consider inter tier transmission i.e. the IAB-nodes in a same tier cannot both transmit and receive simultaneously during the same time slot.

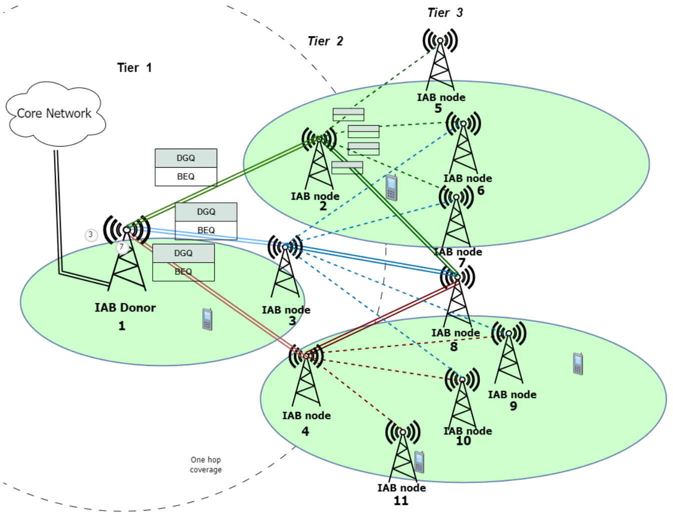


Fig. 3: Illustration of different paths in an IAB network

example, if the link rate is R Mbps at TTI t_s and the buffer size is b^q KB, the upper bound delay for queue q can be calculated as $T^{u,q} = \frac{b^q}{0.125(R \times w_q)}$, where w_q is the weight assigned to the queue. Increasing the weight w_q will reduce the upper bound delay for queue q, and vice versa.

Also, it is assumed that every packet carries the time stamp of its generation time T_{start} and its delay budget, t_{DB}, on its header. Each IAB-node has to choose the best path and queue, also known as the forwarding decision, for each packet. To simplify, we define the set of scheduling and path choices available for each packet at IAB-node i as the "forward decision set", denoted by $\mathcal{H}_i$ and each individual scheduling choice as a "forwarding decision" denoted by $h^q_{i,j}$. In this regard, we can show the set of forward decisions for each packet at IAB-node i as follows:

$$\mathcal{H}_i = \{h^1_{i,1}, h^2_{i,1}, h^1_{i,2}, h^2_{i,2}, \ldots, h^q_{i,j}, \ldots, h^2_{i,B_i}\}, \tag{1}$$

where $h^q_{i,j}$ is the decision to forward a packet at IAB-node i toward IAB-node j using queue q, and B_i represents the number of neighbouring IAB-nodes that IAB-node i can transmit packets to.

C. Channel Model

Given the static position of IAB-donor and IAB-nodes, we assume the link throughput over a small period of time remains stable, while it can change from time to time due to shadowing or fading. We can apply the Markov model to describe the channel. Each state represents a range of SNR, and the steady-state probability and state transition probabilities of the Markov channel are determined based on the first-order and second-order statistics of the channel [32].

We use the Rayleigh fading channel to calculate the steady state distribution and state transition probabilities. The instantaneous SNR, γ is divided into a finite number of levels, denoted by Γ_i where i shows the state. Let S_i denoted the ith state and the set of all states are represented by $\mathcal{S} = \{S_1, S_2, \ldots, S_M\}$. The steady-state probability of $\gamma \in S_i$ is given by:

$$\pi_i = \int_{\Gamma_i}^{\Gamma_{i+1}} p(\gamma)d\gamma = \exp(-\frac{\Gamma_i}{\gamma_0}) - \exp(-\frac{\Gamma_{i+1}}{\gamma_0}), \tag{2}$$

where γ_0 is the mean of the received SNR, and $p(\gamma)$ is the probability density function (pdf) of γ which has an exponential distribution. Assuming that the channel stays in the same state for the duration of coherence time t_{coh}, the transition probability could be approximated as:

$$P_{i,i+1} = \frac{N(\Gamma_{i+1}) \times t_{coh}}{\pi_i}, \quad i = 1, 2, \ldots, M-1, \tag{3}$$

$$P_{i,i-1} = \frac{N(\Gamma_i) \times t_{coh}}{\pi_i}, \quad i = 2, 3, \ldots, M, \tag{4}$$

$$P_{i,i} = 1 - P_{i,i+1} - P_{i,i-1}. \tag{5}$$

Here $N(\Gamma_i)$ is the level-crossing rate of the Rayleigh fading envelope at Γ_i [33]:

$$N(\Gamma_i) = \sqrt{2\pi} f_D \sqrt{\frac{\Gamma_i}{\gamma_0}} \exp{(-\frac{\Gamma_i}{\gamma_0})}, \tag{6}$$

where f_D is the Doppler shift.

D. MCS Selection

One of the key features of 5G NR is adaptive modulation and coding scheme (MCS). Every user reports channel state information (CSI) to its gNB as feedback [34]. CSI reports can be triggered by channel variations or broadcast periodically in accordance with the channel coherence time. These CSI reports ensure that each IAB-node has up-to-date channel information. The reporting mechanisms follow existing standards, such as those defined in 5G NR [35], which include both periodic and aperiodic updates. The BS schedules downlink data transmissions such as MCS, number of transmission layers, and Multiple-Input Multiple-Output (MIMO) precoding based on the user's feedback on channel state information accordingly.

Having different MCS due to channel variations can alternate the network performance in terms of throughput and latency. To illustrate the impact of this change, different states in the Markov channel model are mapped to appropriate MCS.

Based on the chosen MCS and subcarrier spacing, we can calculate the transport block (TB) size and throughput of each state. According to 3GPP's technical specification, there are two approaches to calculate TB [34]. One method determines the precise amount of TB that is employed in this work, while the other estimates the maximum size of TB. In our study, we utilized the precise method to calculate the TB size, which incorporates factors such as code rate, modulation order, number of layers, and physical resource blocks. A comprehensive explanation of this methodology can be found in [34].

E. Latency Calculation

One of the key enabler technologies of 5G is the use of flexible numerology (μ) that allows the network to deploy different sub-carrier spacing of $2^{\mu} \times 15$KHz to support different services such as URLLC, mMTC, etc. [36]. 5G-NR latency is mostly measured in terms of multiple transmission time interval (TTI), t_I which is depended on selected μ. The length of NR TTI is equal to slot length which is $\frac{1}{2^{\mu}}$ ms [20]. The E2E delay for direct transmission from the IAB-donor to a user is mostly affected by queuing delay T^{q}, and transmission delay T^{tr}, since the Processing delay is almost constant, and propagation delay is negligible. Thus, the total delay for a multi-hop transmission from IAB-node m to n using path r hops can be calculated by:

$$T^{Total}_{m,n}(r) = \sum_{(i,j)\in r} \left[T^{tr}_{i,j} + T^{q}_{i,j}\right]. \tag{7}$$

The transmission time from IAB-node i toward IAB-node j using queue q can be calculated by $T^{tr}_{i,j} = \lceil \frac{pktsz}{\text{TB}_{i,j} \times w_q} \rceil \times t_I$, where $\text{TB}_{i,j}$ is TB size for the path from IAB-node i to j and w_q is the weight of the round robin for the link departure. Queuing delay can be calculated by using:

$$T^{q}_{i,j} = \sum_{l=1}^{Q^{q}_{i,j}} \frac{(pktsz)_l}{\text{TB}_{i,j} \times w_q} \times t_I, \tag{8}$$

here $pktsz$ shows the packet size and $Q^{q}_{i,j}$ is the queue length which can be calculated as follows:

$$Q^{q}_{i,j}[t+1] = \left[Q^{q}_{i,j}[t] - r^{q}_{i,j}, 0\right]^{+} + a^{q}_{i,j}, \tag{9}$$

where $r^{q}_{i,j} = \text{TB}_{i,j} \times w_q$ is the number of departure and $a^{q}_{i,j}$ is arrival packets, respectively. The delay upper bound for each queue ($T^{u,q}_{i,j}$) can be calculated by:

$$T^{u,q}_{i,j} = \frac{b^{q}_{i,j}}{\text{TB}_{i,j} \times w_q} \times t_I. \tag{10}$$

F. Problem Formulation

Consider the topology shown in Fig. 3. For instance, there are three paths $r_1 := \{\text{IAB } 1 \rightarrow \text{IAB } 2 \rightarrow \text{IAB } 8\}$, $r_2 := \{\text{IAB } 1 \rightarrow \text{IAB } 3 \rightarrow \text{IAB } 8\}$, and $r_3 := \{\text{IAB } 1 \rightarrow \text{IAB } 4 \rightarrow \text{IAB } 8\}$ from the IAB-donor to IAB-node 8. Assume that every one of these paths can satisfy the required latency as demanded by the application. This means that the total delay cost, represented as $c(r)$, for each path, is lower than t_{DB}. Here, $c(r) = \sum_{h\in r} \mathcal{D}^{1}_{h}$ stands for the sum of delay upper bounds of the DGQs, $\mathcal{D}^{1}_{h}$, for every individual hop h at path r.

Each IAB-node (including the IAB-donor) has two queues for each path, and in total, there are six choices to deliver the packet at the IAB donor. However, not all choices can meet the required t_{DB} because of the network dynamics and different TB sizes at each path. Moreover, due to the deployment of WRR, it is crucial to investigate how to use different queues since they have a mutual impact. With the deployment of WRR, a high-priority queue prolongs the queuing delay in the low-priority queue, and vice versa.

Considering the above condition, the optimal decision aims to maximize the network's overall utility considering the delay requirement. In other words, the optimization problem's focus is on identifying a series of appropriate forwarding decisions for each packet, ensuring that the packet's end-to-end delay criteria are met. The optimization problem can be formulated as follows:

$$P0: \max_{h_{\text{DS}}, h_{\text{NDS}} \in \mathcal{H}} \sum_{t=1}^{T} U_{\text{DS}}(P_{\text{DS}}[t] | h_{\text{DS}}) + U_{\text{NDS}}(P_{\text{NDS}}[t] | h_{\text{NDS}}) \tag{11a}$$

$$s.t. \qquad \mathcal{D}(P_{DS}[t]) \leq t_{DB}, \tag{11b}$$

where P_{DS} and P_{NDS} denote arrived packets at their destination at TTI t, respectively. U_{DS} and U_{NDS} represent the network utility function for DS and NDS packets. The network utility function is a mathematical tool used to measure the satisfaction or benefit that users, operators, or applications obtain. h_{DS} and h_{NDS} show the forwarding decision made at each IAB-node, respectively. $\mathcal{D}(.)$ calculates the E2E delay. Eq.(11a) aims to maximize the total utility and (11b) is applied to ensure every received packet has met its delay requirement.

IV. Delay-Guaranteed Path selection and Scheduling (DGPS)

Performing path selection among the IAB-nodes to reach the destination via wireless links is challenging due to dynamic channels. We propose a DGPS algorithm to perform wireless routing and queue selection, based on local queue information and TB size. Our proposed algorithm can be divided into two parts. DGPS initially filters out undesired paths and queues in three steps to reduce time complexity since the implementation of priority queues can increase the time complexity. Then, we aim to identify paths and queues that maximize overall utility

The distributed nature of the proposed method ensures scalability, as each IAB-node independently makes forwarding decisions without relying on a centralized controller. This approach distributes the computational and decision-making load across the network, enabling efficient scaling as the number of IAB-nodes increases compared to existing Internet routing such as OSPF. The decision space grows linearly with the number of neighbouring IAB-nodes rather than the total network size, making it more efficient. Our time complexity analysis confirms that the algorithm remains efficient even in larger networks.

For DGPS to work effectively, each IAB-node must broadcast its updated upper-bound delay, which is calculated based on the TB size, to its parent IAB-nodes along the path toward the IAB-donor.

In our proposed algorithm, exploring multiple paths expands the routing choices, at the cost of higher time and space complexity. For a network with n IAB-nodes and an average of k neighbour IAB-nodes, the time complexity of routing table generation rises to $O(kn \log n)$, compared to $O(n \log n)$ for a single-path routing protocol. The space complexity for maintaining the forwarding table at each router increases from n (in single-path routing) to kn.

A. Queue and Path filtering

Since we have a rich decision space for finding the optimal decision, one of our objectives is to reduce time complexity. In this regard, we reduce decision space through a three-step process. The following processes are listed in Lines 1-6 in Algorithm 1.

Let Ω_i denote the feasible paths from IAB-node i to the destination. In DGPS, each IAB-node finds the shortest path from each of its neighbours to the destination. Thus, if an IAB-node has k neighbors, it will calculate k paths. As the number of neighbours increases, the decision space expands. As all nodes enlarge the routing options effectively, the number of paths explored grows exponentially w.r.t. the network size. This can enhance the overall performance by exploring more routing choices, particularly in denser networks.

In order to filter out the unacceptable routes and queues, upon the arrival of a packet, the information such as t_{DB} and T_{start} will be extracted and the packet will be enqueued to the ingress buffer. Then the IAB-node calculates the remained $t^*_{DB} = T_{start} + t_{DB} - T_{now}$ for the packet and computes $T_{th} = \min\{t^*_{DB}, c(r)^*_{prev}\}$ as the threshold to remove undesirable paths in the following, where $c(r)^*_{prev}$ is the previous IAB-node's computed cost (Lines 1-3).

we first focus on performing path filtering to ensure that the candidate routes can meet the required latency. For each route r, $c(r)$ is computed and compared with T_{th}. If the selected route cannot meet the threshold ($c(r) > T_{th}$), path r is removed from Ω_i. After this filtering, if set Ω_i becomes empty, the IAB-node will drop the packet since it cannot guarantee the delay (Line 4).

Then, we focus on the available queue space for each queue. The available queuing space of each queue is given by $b^q_{i,j} - Q^q_{i,j}$. If the available space is less than $pktsz$, the IAB-node will eliminate that queue from the list of possible choices because there is insufficient space to accommodate the packet, resulting in the packet being dropped (Line 5).

Next, the per-hop delay budget is computed:

$$T_{hop}(r_j) = t^*_{DB} - c(r_j). \tag{12}$$

The IAB-node compares the queuing delay $T^q_{i,j}[t]$ of each $h^q_{i,j} \in \mathcal{H}_i$ with per-hop delay budget $T_{hop}(r_j)$, and removes the forwarding decisions which cannot meet the requirement (Line: 6). At this point, if there is no more available path, the packet would be dropped. Otherwise, it will continue to select the path according to the following utility maximization process.

B. Utility maximization

The IAB-node has a finite set of feasible paths $\tilde{\mathcal{H}}_i$ which can meet the delay requirement. Our goal is to find the best decision to maximize the overall utility. Since our problem is spatiotemporally correlated, it is hard to find the optimal answer. We propose a distributed approach to find the forwarding decision based on the packet's delay budget which can be categorized as a renewal optimization problem [37], [38].

At every TTI, we can use a cost-reward tuple $(T^q_{i,j}[t], G^q_{i,j}[t])$ to represent each forwarding decision $h^q_{i,j} \in \mathcal{H}_i$, where $G^q_{i,j}[t]$ is the achieved reward for executing $h^q_{i,j}[t] \in \mathcal{H}_i$:

$$G^q_{i,j} = \frac{(T_{hop}(r_j) - T^q_{i,j})}{w_q}. \tag{13}$$

Such a reward aims to use the lower priority queues first when they are available and feasible to meet the deadline while reserving the higher priority queues for the impending urgent packets. Then, the infinite horizon reward per time cost ratio for each IAB-node i can be calculated as follows:

$$\theta_i = \lim_{K \to \infty} \frac{\sum_{k=1}^{K} G^q_{i,j}[k]}{\sum_{k=1}^{K} T^q_{i,j}[k]}, \tag{14}$$

where $G^q_{i,j}[k]$ and $T^q_{i,j}[k]$ are achieved reward and queuing latency for the received packet k at IAB-node i, respectively. In this regard, instead of solving P0, we focus on improving network utility through local decision-making at each IAB-node. While it does not guarantee global optimality, it provides scalable and efficient solutions for dynamic environments. we can decouple it into a set of sub-problems, i.e., we can maximize θ_i for each IAB-node $i \in \mathcal{B}$:

$$P1: \quad \max_{h^q_{i,j} \in \tilde{\mathcal{H}}_i} \theta_i, \tag{15a}$$

$$s.t. \quad (T^q_{i,j}[k], G^q_{i,j}[k]) \in \mathcal{U}(\tilde{\mathcal{H}}_i), \quad k \in \{0, 1, \ldots, K\}. \tag{15b}$$

Here, $\mathcal{U}(\tilde{\mathcal{H}_i})$ represents the set of all possible cost-reward tuples in $\tilde{\mathcal{H}}_i$. However, finding the optimal answer to P1 is challenging due to the randomness of network dynamics and the deployment of adaptive MCS at each IAB. To overcome these issues, we deployed a heuristic approach leveraging the iterative algorithm for solving renewal optimization problems in [37], [39] to solve P1 based on previous decisions where θ^* is approximated by repeatedly selecting the cost-reward pair that maximizes the reward gradient i.e., we choose the tuple which maximizes the reward gradient.

We can convert P1 to the following problem:

$$P2: \quad \max_{h^q_{i,j} \in \tilde{\mathcal{H}}_i} \quad \psi[k] = G^q_{i,j}[k] - \delta_i[k] T^q_{i,j}[k], \tag{16a}$$

$$s.t. \quad (T^q_{i,j}[k], G^q_{i,j}[k]) \in \mathcal{U}(\tilde{\mathcal{H}}_i), \quad k \in \{0, 1, \ldots, K\}, \tag{16b}$$

where δ_i will be updated after each decision as follows:

$$\delta_i[k+1] = \left[\delta_i[k] + \mu(G^q_{i,j}[k] - \delta_i[k] T^q_{i,j}[k])\right]^{\delta_i^{max}}_{\delta_i^{min}}. \tag{17}$$

Here, $[.]^{\delta_{max}}_{\delta_{min}}$ is the normalization function between the range $[\delta_{min}, \delta_{max}]$, and μ denotes the step size (Line: 7-14).

Algorithm 1 DGPS Algorithm

```
Input: $\mathcal{H}_i$ and $\Omega_i$ at node $i$
Output: Forwarding decision $h^{q*}_{i,j}(Path, Queue)$
Upon a packet arrival, extract $t_{DB}$ and $T_{start}$ and enqueue it to ingress buffer
Calculate the remained delay budget:
    $t^*_{DB} = T_{start} + t_{DB} - T_{now}$
Define $T_{th} = \min\{t^*_{DB}, c(r)^*_{prev}\}$
Path filtering using $T_{th}$ and $C(r)$
Eliminating queues without enough space
Removing $h^q_{i.j}$ based on $T_{hop}(r_j)$ and $T^q_{i,j}[t]$
for each $h^k_{i,j} \in \tilde{\mathcal{H}}_i$ do
    Calculate $G^q_{i,j}$ using Eq. 13
    Define $\psi[k] = G^q_{i,j}[k] - \delta_i[k]T^q_{i,j}[k]$
    if $\psi^*[k] \geq \psi[k]$ then
        $\psi^*[k] = \psi[k], \quad path = j, \quad queue = q$
    end if
end for
update $\delta_i[k+1]$ using (17)
```

Architecture 1a for IAB networks, which is deployed in this work, assumes that the DUs on an IAB-node are served by only one IAB-donor [7]. However, this method can be extended to additional tiers and IAB-donors due to its distributed nature. In DGPS, each IAB-node independently makes forwarding decisions, making it well-suited for implementation in larger networks.

V. Performance Evaluation

In this section, we evaluate the performance of our proposed routing and scheduling protocol. In DGPS algorithm, two queues are deployed at each IAB-node for each link, which can be easily extended to more priority queues. Our work is compared with two routing protocols, the open shortest path first (OSPF) protocol [40] which calculates the shortest path among IAB-nodes to efficiently distribute the data using Dijkstra's algorithm, and the equal-cost multipath (ECMP) protocol [41] which can use multiple parallel paths with equal cost (in terms of a static cost metric) to improve load balancing. The link cost metrics in these algorithms are the delay upper bound based on the selected MCS and they do not use prioritized queues for each link. Also, we assume that these algorithms are able to drop packets if they exceed the specified delay budget. This prevents wasting queue space on delivering outdated packets.

A. Network Configuration

We consider a network of eleven IAB-nodes in a DAG topology with one IAB-donor as illustrated in Fig. 3. Because of the deployment of the DAG topology, each IAB-node is guaranteed to have at least one path available. The simulation parameters are summarized in TABLE II and TABLE III.

We divided the received SNR into four different levels and mapped each level to the appropriate MCS [42]. Our deployed MCSs are $[\text{QPKS}, 602/1024]$, $[\text{16QAM}, 490/1024]$, $[\text{64QAM}, 466/1024]$, $[\text{64QAM}, 873/1024]$. We considered our network is using numerology ($\mu = 2$) and we are using 66 resource blocks (RB) and there are 12 subcarriers per RB. Furthermore, we used normal cyclic prefix and two physical downlink shared channel (PDSCH) transmission layers. The TB for each path is calculated based on these MCS and configurations. For instance, the TB size for $[\text{16QAM}, 490/1024]$ based on this configuration is equal to 19992 bits. Furthermore, due to the higher transmission power of the IAB-donor, being a macro base station, compared to other IAB-nodes, which are small base stations [43], results in a higher steady-state probability for employing higher MCS. The transition probabilities between the states are presented in TABLE IV.

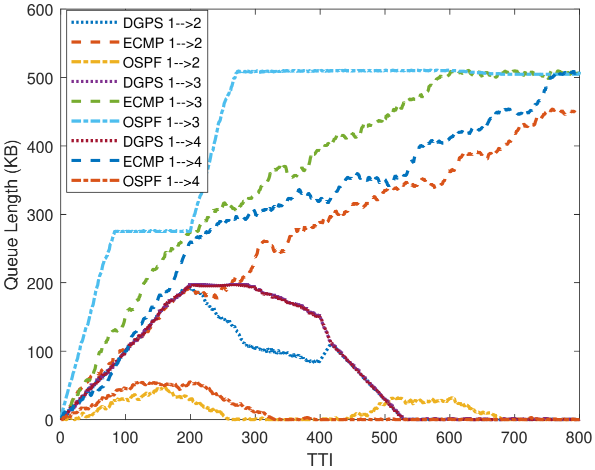


Fig. 4: Queue length of IAB-donor's every link in high traffic load

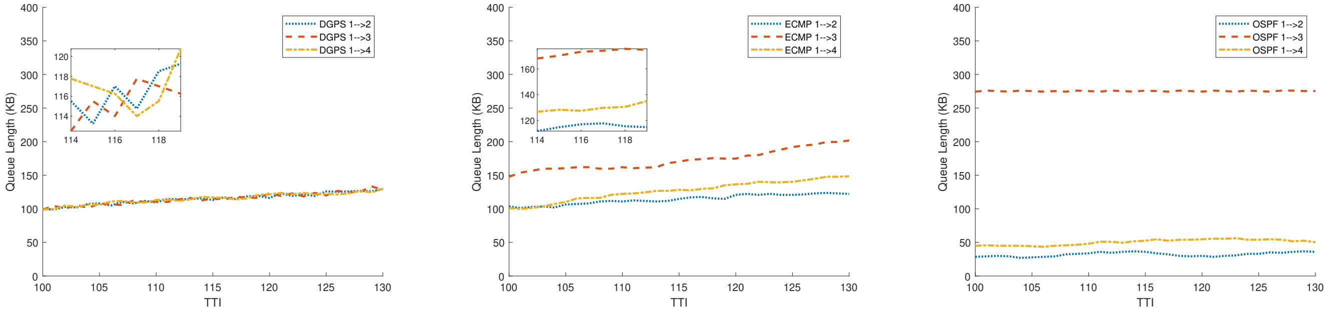
We examine our proposed solution across various traffic load scenarios outlined in Table III. In this regard, we considered two types of applications supported by our network for downlink transmission: (1) the time-sensitive application generates packets of size 1500 B and requires a guaranteed E2E delay of 10ms (DS packets), and (2) the non-time-sensitive packets with size of 1500 B and delay budget of 1 s (NDS packets). The considered buffer size and WRR weights for DGQ and BEQ are 12 KB and 600 KB, and 0.8 and 0.2,

TABLE II: Buffers and packets configuration

Parameters	DGQ	BEQ
Buffer Size b^q_i (KB)	12	600
WRR weight w^q_i	0.8	0.2
	DS	NDS
Packet Size $pktsz$ (B)	1500	1500
Delay budget t_{DB}(TTI)	40	4000

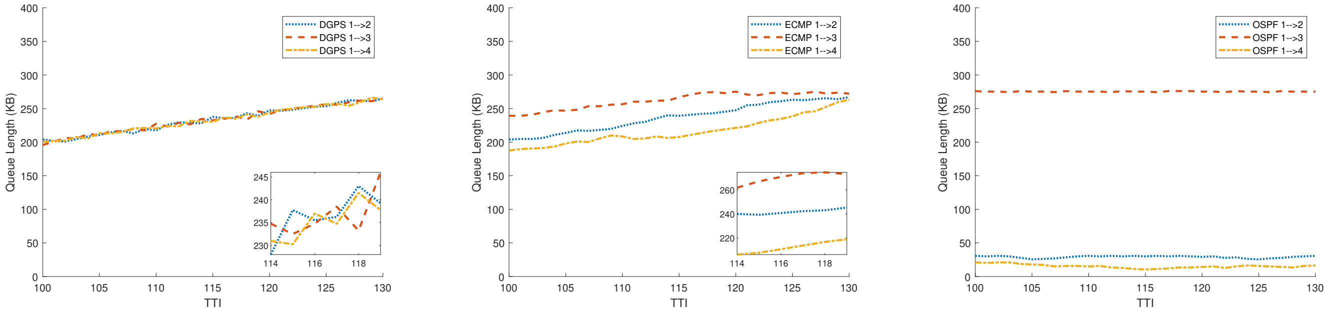
TABLE III: Simulation Parameters

Traffic Load (on/off)	High	Medium	Low	
$\text{Pr}_{b,\text{DS}}$, $\text{Pr}_{b,\text{NDS}}$	0.5,1	0.5,0.75	0.5,0.5	
ρ	0.70	0.59	0.48	
Modulation	QPSK	16QAM	64QAM	64QAM
Coding rate×1024	602	490	466	873
μ	2			
t_I(ms)	0.25			



(a) DGPS (b) ECMP (c) OSPF

Fig. 5: Queue length of IAB-donor's every link in high traffic load in lower SNR state



(a) DGPS (b) ECMP (c) OSPF

Fig. 6: Queue length of IAB-donor's every link in bursty mode with high traffic load in lower SNR state

TABLE IV: Transition probability matrices

IAB-Doner	IAB-node
$\begin{bmatrix} 0.88 & 0.12 & 0 & 0 \\ 0.8 & 0.18 & 0.02 & 0 \\ 0 & 0.9 & 0.07 & 0.03 \\ 0 & 0 & 0.99 & 0.01 \end{bmatrix}$	$\begin{bmatrix} 0.35 & 0.65 & 0 & 0 \\ 0.15 & 0.7 & 0.15 & 0 \\ 0 & 0.25 & 0.7 & 0.05 \\ 0 & 0 & 0.95 & 0.05 \end{bmatrix}$

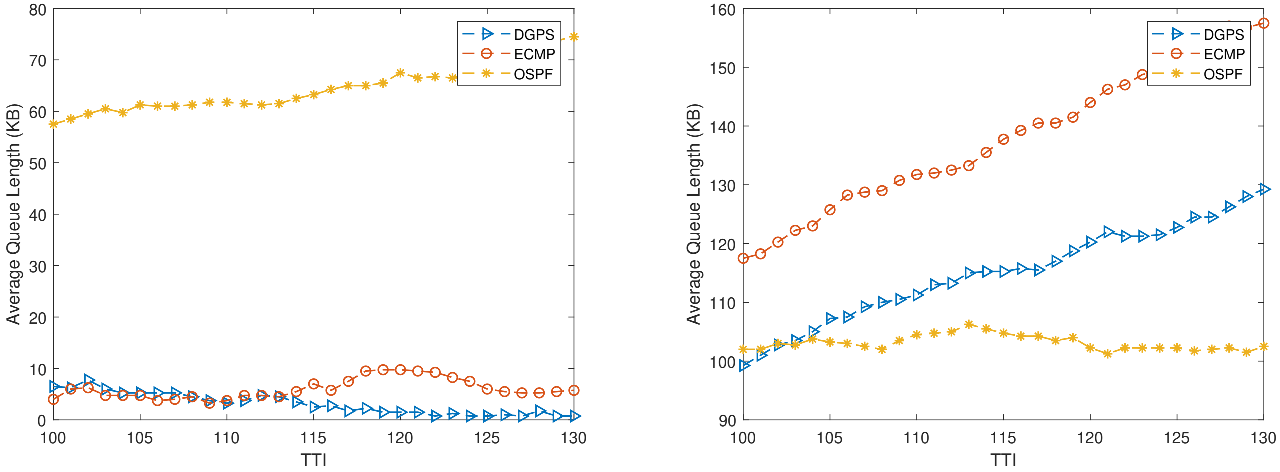
respectively.

We utilize an on/off packet generation model, originating from the core network and intended for each IAB-node, including IAB-donor. These packets are then distributed via an IAB-donor to different IAB-nodes throughout our network. At every TTI, within the core network, there exists one NDS packet generator and one DS packet generator, each designated for a specific IAB-node as the destination. The probability $\mathrm{Pr}_{b,\mathrm{DS}}$ and $\mathrm{Pr}_{b,\mathrm{NDS}}$ indicate the likelihood of the DS and NDS packet generators being active for each IAB-node b as their destination. The probability of each traffic scenario and its average arrival rate divided by the departure rate (ρ) is summarized in Table III.

B. Instant queue length

We first analyzed the queue lengths across each link within the IAB-donor, as illustrated in Figs. 4 through 6. The x-axis represents the TTI, while the y-axis indicates the queue length measured in KB. In the DGPS algorithm, each link is associated with two queues. Therefore, the queue length displayed in the figures is the sum of the lengths of these two queues corresponding to each link. Fig. 4 illustrates the instant queue length on each link for all algorithms. In the initial 200 TTIs, the wireless channel quality is poor, leading the IAB-donor to employ lower MCS for transmission, resulting in prolonged queues. However, as the SNR improves and the IAB-donor adopts a higher MCS, DGPS effectively reduces the queue length close to zero, which is not achieved by ECMP and OSPF. Also, We observe that in OSPF the queue length for the link from IAB-donor to IAB node 3 remains relatively constant over a period of time. This consistency is due to our assumption that OSPF and ECMP algorithms can discard packets in their queue when they reach their delay budget. This strategy helps enhance the efficiency of their queues since we have DS packets which need to be delivered in time. Next, we are going to focus on TTIs where the wireless channel is not in a good situation, posing a greater challenge. Two distinct configurations are employed to illustrate packet distribution in each queue for three different algorithms.

For the high traffic load scenario, Fig. 5a illustrates that DGPS effectively distributes packets across various links, maintaining a balanced load and mitigating congestion on specific links. Similarly, Fig. 5b depicts ECMP attempting a similar distribution, albeit less efficiently than DGPS. Additionally, Fig. 5c reveals poor traffic distribution, resulting in prolonged queues for the link from IAB-donor to IAB-node 3. The superior performance of DGPS is attributed to its utilization of global network knowledge, including the delay upper bound of subsequent IAB-nodes, coupled with



(a) Medium traffic load

(b) High traffic load

Fig. 7: Average queue length of IAB-donor in medium and high traffic load

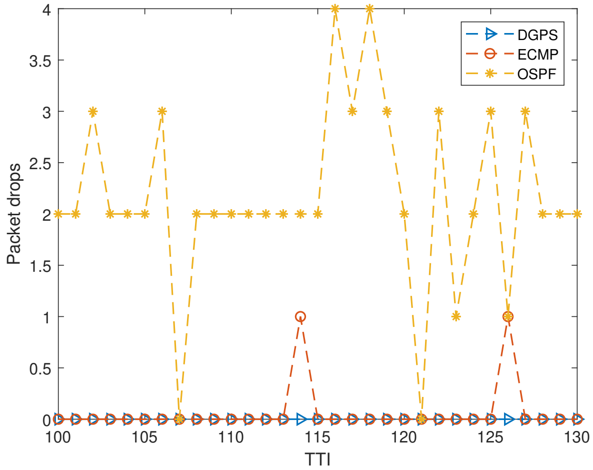


Fig. 8: Number of dropped packets at IAB-donor in high traffic load

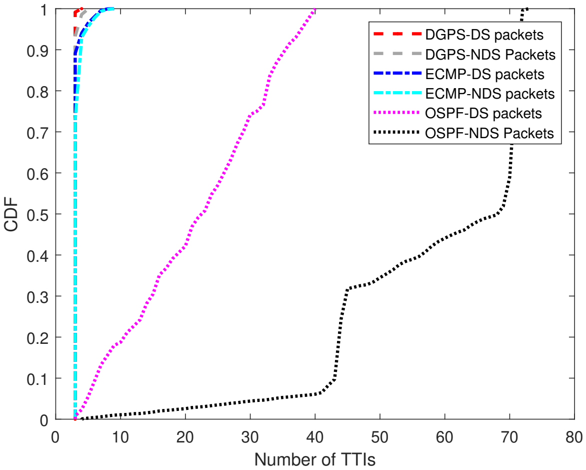


Fig. 9: Comparison of E2E latency in Medium Traffic Load

local knowledge of its own TB size and queueing delay. This comprehensive information allows DGPS to make informed decisions regarding path selection and queue utilization. Similarly, a consistent pattern is observed in Fig. 6 for bursty mode under high traffic load. DGPS demonstrates a well-balanced distribution of queue lengths, outperforming other algorithms. In Fig. 6b, it is evident that the queue lengths for the links from IAB-donor to IAB-node 3 and IAB-donor to IAB-node 4 at TTI 116 are 271 KB and 211 KB, respectively. This difference of 60 KB between the two links can adversely impact network performance.

C. Average queue length and Dropped packets

Furthermore, the average queue length at each TTI for both medium and high traffic loads is depicted in Fig. 7. Notably, DGPS maintains a lower average queue length at every TTI compared to ECMP, due to its better link utilization and traffic distribution. In Fig. 7b, OSPF exhibits the lowest average queue length. This result is because of the extended queue latency at one of the links, causing a significant number of DS packets to be dropped as they fail to meet their delay budget. In our design, packets are dropped by each IAB-node upon reaching the delay budget threshold to save the resource for better utilization.

The count of dropped packets from the IAB-donor at each TTI under high traffic load is illustrated in Fig. 8. Notably, OSPF exhibits the highest number of dropped packets at each TTI due to its prolonged queues, while DGPS consistently outperforms the other algorithms in minimizing packet drops.

D. E2E delay comparison

We assess our algorithm's E2E delay performance, presented in Figs. 9 to 11, across varying traffic loads for received packets. We examine three traffic load scenarios outlined in Table III. For moderate traffic loads, Fig. 9 illustrates DGPS's

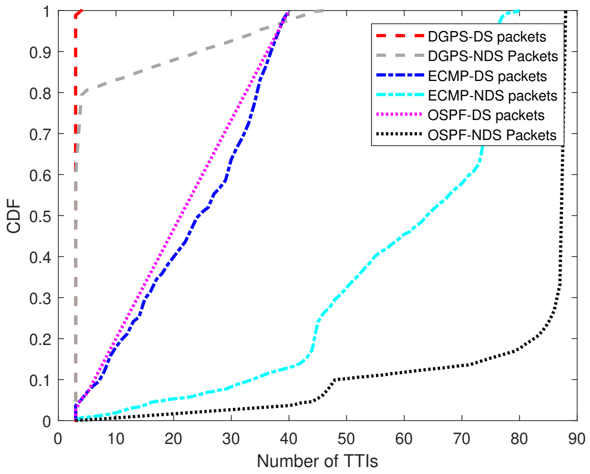


Fig. 10: Comparison of E2E latency in High Traffic Load

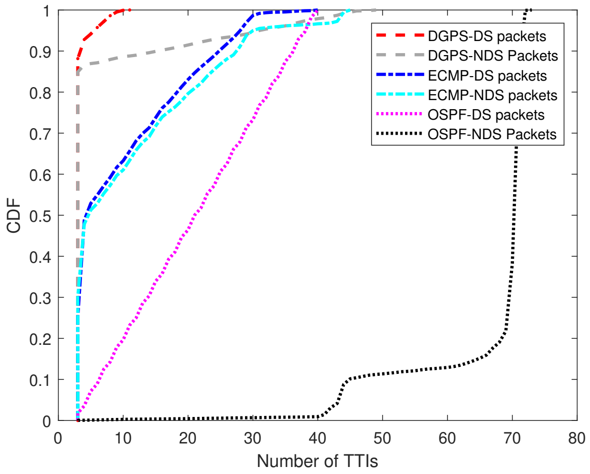


Fig. 11: Comparison of E2E latency in Bursty mode in Medium Load

superior E2E latency performance for both DS and NDS packets compared to ECMP and OSPF. DGPS and ECMP demonstrate similar performance for DS and NDS packets. The reason behind the better performance of DPGS and ECMP is that they distribute the traffic accordingly and improve the congestion situation in the network. On the other hand, OSPF is not capable of using alternative paths.

Under high traffic load, Fig. 10 shows that DGPS substantially outperforms ECMP and OSPF. DGPS algorithm delivers 99% of DS packets and 79% of NDS packets in four TTIs. Notably, DGPS's E2E latency for DS and NDS packets differs due to its consideration of the packet's delay budget in routing and scheduling. Note that this comparison is only among the received packets in their required latency and it did not consider dropped packets. Although in Fig. 10, OSPF outperforms ECMP, it has a lower packet delivery ratio (ν) as shown in Fig. 12.

Furthermore, we examine DGPS in bursty mode. For bursty DS Mode, we assumed there are three DS packets generated for a specific destination (IAB-node 8) at each TTI. Fig. 11 shows the E2E latency comparison in Bursty DS packet mode in a medium background traffic load. With DGPS, E2E delay for received DS packets is less than 10 TTIs, and over 87% of the packets are received in four TTIs.

E. Packet delivery ratio

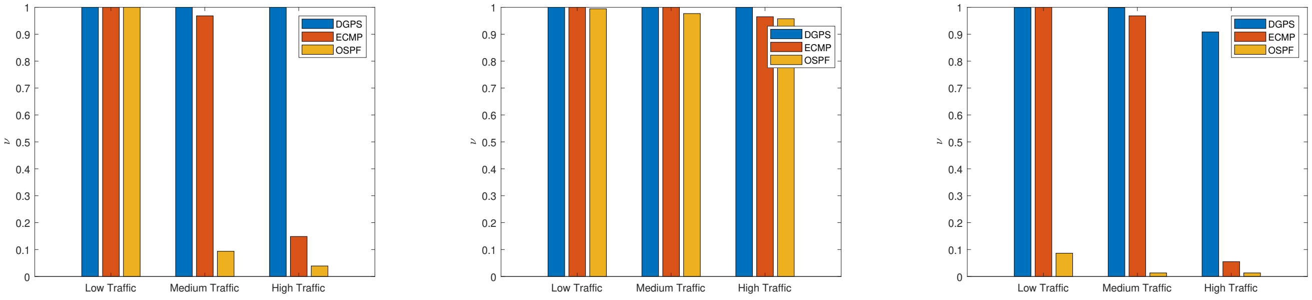
Fig. 12 depicts the comparison among packet delivery ratio of all algorithms in different traffic load situations. We defined packet delivery ratio as the number of received packets at the destination within the delay requirement over the total generated packets for the same destination. It is noteworthy to mention that packet drops only happen when coming packets exceed the buffer size or we cannot satisfy the t_{DB} of the packet. Fig. 12a displays nearly 100% packet delivery ratios for all algorithms with low traffic for DS packets. The similar performance is due to the minimal network congestion, allowing all algorithms to efficiently utilize available resources without significant queuing delays.

Nevertheless, as background traffic increases, ECMP and OSPF's performance declines, leading to significant DS packet drops in high-traffic conditions. DGPS, however, maintains its performance despite the high load. This behaviour occurs because ECMP and OSPF do not prioritize packet delivery based on delay requirements, and adapt to traffic loads and channel variation which leads to more congestion and packet drops as traffic increases. Similarly, this trend is observed for NDS packets in Fig. 12b, where ECMP and OSPF achieve a superior packet delivery ratio compared to DS packets. This difference arises because NDS packets possess a higher t_{DB}, So they have not dropped due to the delay budget. Fig. 12c shows the packet delivery ratio comparison of all approaches in bursty DS packet mode in different background traffic loads. While in low background traffic load, DGPS and ECMP perform well, they perform poorly in higher background traffic load compared to DGPS. The high amount of traffic can overwhelm the queues, resulting in a lower packet delivery ratio compared to the more controlled low-traffic or non-bursty scenario. However, DGPS still outperforms OSPF and ECMP.

We further present confidence intervals for the packet delivery ratio under two challenging traffic conditions. For DS packets in high traffic load, the 95% confidence interval for packet delivery ratio is $[0.9982, 0.9999]$, indicating very high reliability. In the bursty mode with high background traffic, the 95% confidence interval for packet delivery ratio is $[0.8941, 0.9154]$, showing the robustness of the algorithm even in more challenging scenarios. These confidence intervals provide insight into the system's reliability and demonstrate the algorithm's ability to consistently deliver packets within their delay budgets.

F. Packet delivery ratio in extended network

We conducted an extended simulation by adding a fourth tier to the original network configuration to further evaluate the system's scalability and performance under more realistic conditions. This modification aimed to provide insights into



(a) DS packets

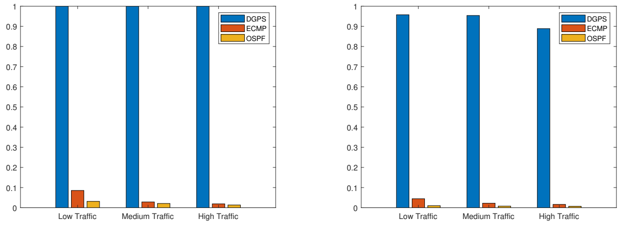
(b) NDS packets

(c) DS packets in bursty mode

Fig. 12: packet delivery ratio of different approaches for DS packets and NDS packets

the robustness and efficiency of the proposed DGPS method when applied to larger networks.

The results of this extended simulation are depicted in Fig.13. Figs.13a and 13b show trends similar to those observed in previous evaluations, indicating that the DGPS method continues to perform effectively with an increased number of nodes. The packet delivery ratio of DGPS is more than ten times higher than the benchmarks.



(a) DS packets

(b) DS packets in bursty mode

Fig. 13: DS packet delivery ratio of different approaches in four tiers scenario

VI. Conclusion

In this paper, we have proposed a distributed delay-guaranteed path selection and scheduling approach for multi-hop IAB networks using different priority queues to deliver packets. Our proposed approach has enabled each individual IAB-node to determine the optimal path and scheduling decision for each packet based on its delay requirement, knowledge of downstream IAB-nodes upper bound delay, and the current IAB-node's queue status. Each IAB-node solves a renewal optimization problem to make a decision. Through our simulation, we have demonstrated that our proposed solution substantially outperforms the existing ones in terms of latency and packet delivery ratio. There are many further research issues to extend the proposed solution such as increasing the number of DGQs and defining adaptive limits for buffers to prevent bufferbloat.

References

[1] I. Union, "IMT traffic estimates for the years 2020 to 2030," *Report ITU*, vol. 2370, 2015.

[2] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. T. Sukhavasi, C. Patel, and S. Geirhofer, "Network densification: the dominant theme for wireless evolution into 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 82–89, 2014.

[3] O. Teyeb, A. Muhammad, G. Mildh, E. Dahlman, F. Barac, and B. Makki, "Integrated access backhauled networks," in *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*. IEEE, 2019, pp. 1–5.

[4] 3rd Generation Partnership Project, ""NG-RAN; architecture description," 3GPP, TS 38.401, v16.2.0," 3GPP, Tech. Rep., 2020.

[5] R. Deng, B. Di, S. Chen, S. Sun, and L. Song, "Ultra-dense leo satellite offloading for terrestrial networks: How much to pay the satellite operator?" *IEEE Transactions on Wireless Communications*, vol. 19, no. 10, pp. 6240–6254, 2020.

[6] E. Dahlman, S. Parkvall, and J. Skold, *5G NR: The next generation wireless access technology*. Academic Press, 2020.

[7] 3rd Generation Partnership Project, "Study on integrated access and backhaul- TR 38.874," 3GPP, Tech. Rep., 2018.

[8] Y. Zhang, M. A. Kishk, and M.-S. Alouini, "A survey on integrated access and backhaul networks," *Frontiers in Communications and Networks*, vol. 2, p. 647284, 2021.

[9] V. Addanki and L. Iannone, "Moving a step forward in the quest for deterministic networks (DetNet)," in *2020 IFIP Networking Conference (Networking)*. IEEE, 2020, pp. 458–466.

[10] A. Nasrallah, A. S. Thyagaturu, Z. Alharbi, C. Wang, X. Shao, M. Reisslein, and H. ElBakoury, "Ultra-low latency (ULL) networks: The ieee tsn and ietf detnet standards and related 5G ULL research," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 88–145, 2018.

[11] M. Garetto and D. Towsley, "Modeling, simulation and measurements of queuing delay under long-tail internet traffic," *ACM SIGMETRICS Performance Evaluation Review*, vol. 31, no. 1, pp. 47–57, 2003.

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

[13] S. Singh, M. N. Kulkarni, A. Ghosh, and J. G. Andrews, "Tractable model for rate in self-backhauled millimeter wave cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 10, pp. 2196–2211, 2015.

[14] H. Zhang, Y. Chen, Z. Yang, and X. Zhang, "Flexible coverage for backhaul-limited ultradense heterogeneous networks: throughput analysis and η-optimal biasing," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 4161–4172, 2018.

[15] Y. Liu, A. Tang, and X. Wang, "Joint incentive and resource allocation design for user provided network under 5G integrated access and backhaul networks," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 2, pp. 673–685, 2019.

[16] M. N. Islam, N. Abedini, G. Hampel, S. Subramanian, and J. Li, "Investigation of performance in integrated access and backhaul networks," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2018, pp. 597–602.

[17] H. Alghafari and M. S. Haghighi, "Decentralized joint resource allocation and path selection in multi-hop integrated access backhaul 5G networks," *Computer Networks*, vol. 207, p. 108837, 2022.

[18] T. K. Vu, M. Bennis, M. Debbah, and M. Latva-Aho, "Joint path selection and rate allocation framework for 5G self-backhauled mm-wave networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 4, pp. 2431–2445, 2019.

[19] Y. Chang, S. Rao, and M. Tawarmalani, "Robust validation of network designs under uncertain demands and failures," in *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*, 2017, pp. 347–362.

[20] H. Yin, S. Roy, and L. Cao, "Routing and resource allocation for IAB multi-hop network in 5G advanced," *IEEE Transactions on Communications*, vol. 70, no. 10, pp. 6704–6717, 2022.

[21] R. Favraud, N. Nikaein, and C.-Y. Chang, "QoS guarantee in self-backhauled lte mesh networks," in *GLOBECOM 2017-2017 IEEE Global Communications Conference*. IEEE, 2017, pp. 1–7.

[22] M. Gupta, A. Rao, E. Visotsky, A. Ghosh, and J. G. Andrews, "Learning link schedules in self-backhauled millimeter wave cellular networks," *IEEE Transactions on Wireless Communications*, vol. 19, no. 12, pp. 8024–8038, 2020.

[23] A. A. Gargari, A. Ortiz, M. Pagin, A. Klein, M. Hollick, M. Zorzi, and A. Asadi, "Safehaul: Risk-averse learning for reliable mmwave self-backhauling in 6G networks," in *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*. IEEE, 2023, pp. 1–10.

[24] D. Triantafyllopoulou, K. Kollias, and K. Moessner, "Price of anarchy in mmwave backhaul routing and link scheduling," *IEEE Transactions on Cognitive Communications and Networking*, 2024.

[25] 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 Communications Magazine*, vol. 58, no. 3, pp. 62–68, 2020.

[26] C. Madapatha, B. Makki, C. Fang, O. Teyeb, E. Dahlman, M.-S. Alouini, and T. Svensson, "On integrated access and backhaul networks: Current status and potentials," *IEEE Open Journal of the Communications Society*, vol. 1, pp. 1374–1389, 2020.

[27] 3rd Generation Partnership Project, "Backhaul adaptation protocol (BAP) specification TS 38.340 version 16.4.0 release 16)," 3GPP, Tech. Rep., 2021.

[28] F. Sohrabi and W. Yu, "Hybrid digital and analog beamforming design for large-scale antenna arrays," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 501–513, 2016.

[29] H. Shokri-Ghadikolaei, C. Fischione, G. Fodor, P. Popovski, and M. Zorzi, "Millimeter wave cellular networks: A MAC layer perspective," *IEEE Transactions on Communications*, vol. 63, no. 10, pp. 3437–3458, 2015.

[30] V. Va, J. Choi, and R. W. Heath, "The impact of beamwidth on temporal channel variation in vehicular channels and its implications," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 5014–5029, 2016.

[31] Y. Zhong, T. Q. Quek, and X. Ge, "Heterogeneous cellular networks with spatio-temporal traffic: Delay analysis and scheduling," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1373–1386, 2017.

[32] L. Cai, X. Shen, and J. W. Mark, *Multimedia services in wireless internet: modeling and analysis*. John Wiley & Sons, 2009.

[33] A. Goldsmith, *Wireless communications*. Cambridge university press, 2005.

[34] 3rd Generation Partnership Project, "Physical layer procedures for data TS 38.214 version 16.2.0," 3GPP, Tech. Rep., 2020.

[35] 3GPP, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on New Radio Access Technology: Physical Layer Aspects (Release 14)," 3rd Generation Partnership Project (3GPP), Technical Report TR 38.802 V14.2.0, Sep. 2017, release 14.

[36] 3rd Generation Partnership Project, "Physical channels and modulation-TS 38.211 version 17.4.0," 3GPP, Tech. Rep., 2022.

[37] M. J. Neely, "Fast learning for renewal optimization in online task scheduling," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 12 785–12 828, 2021.

[38] ——, "Dynamic optimization and learning for renewal systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 1, pp. 32–46, 2012.

[39] X. Ren, L. Cai, P. Yang, and J. Ji, "Congestion-aware delay-guaranteed scheduling and routing with renewal optimization," *Computer Networks*, p. 109863, 2023.

[40] J. T. Moy, *OSPF: anatomy of an Internet routing protocol*. Addison-Wesley Professional, 1998.

[41] C. Hopps, "Analysis of an equal-cost multi-path algorithm," Tech. Rep., 2000.

[42] N. S. Saatchi, H.-C. Yang, and Y.-C. Liang, "Novel adaptive transmission scheme for effective urllc support in 5G NR: A model-based reinforcement learning solution," *IEEE Wireless Communications Letters*, 2022.

[43] 3rd Generation Partnership Project (3GPP), "Base Station (BS) radio transmission and reception (3GPP TS 38.104 v15.3.0 Release 15)," 3GPP, Tech. Rep., 2018.

Pooria Seyed Eftetahi (Student Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from Isfahan University of Technology, Isfahan, Iran, in 2015 and 2018, respectively. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC, Canada. He was the recipient of the Graduate Student Fellowship Award from the University of Victoria in 2021 and 2022. His research interests include satellite communication, wireless networks, resource optimization, and machine learning applications in communication systems.

Lin Cai (S'00-M'06-SM'10-F'20) has been with the Department of Electrical & Computer Engineering at the University of Victoria since 2005, and she is currently a Professor. She is an NSERC E.W.R. Steacie Memorial Fellow, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, and an IEEE Fellow. Her research interests span several areas in communications and networking, with a focus on network protocol and architecture design supporting emerging multimedia traffic and the Internet of Things. She has been elected to serve the board of the IEEE Vehicular Technology Society, 2019 - 2024, and as its VP in Mobile Radio. She has been a Board Member of IEEE Women in Engineering (2022-24) and IEEE Communications Society (2024-2026). She has served as an Associate Editor-in-Chief for IEEE Transactions on Vehicular Technology, and as a Distinguished Lecturer of the IEEE VTS Society and the IEEE Communications Society.

Xiangyu Ren (Student Member, IEEE), deceased, received the BSc degree from the Department of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, China, in 2019. He completed his Ph.D. degree in electrical engineering with the Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC, Canada, in 2024. He was the recipient of the Graduate Student Fellowship Award from the University of Victoria in 2021. His research interests included deterministic networks, software-defined networks, vehicular networks, machine learning, and optimization with applications in networking.