Discrete & Continuous Models
& Applied Computational Science
ISSN 2658-7149 (Online), 2658-4670 (Print)

2024, 32 (1) 86–98
http://journals.rudn.ru/miph

Research article

UDC 004.2, 004.7
PACS 07.05.Tp
DOI: 10.22363/2658-4670-2024-32-1-86-98

EDN: CCHVBS

A new link activation policy for latency reduction in 5G integrated access and backhaul systems

Anna A. Zhivtsova, Vitaly A. Beschastnyy

RUDN University, 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation

(received: February 22, 2024; revised: March 12, 2024; accepted: March 25, 2024)

Abstract. The blockage of the propagation path is one of the major challenges preventing the deployment of fifth-generation New Radio systems in the millimeter-wave band. To address this issue, the Integrated Access and Backhaul technology has been proposed as a cost-effective solution for increasing the density of access networks. These systems are designed with the goal of avoiding blockages, leaving the question of providing quality-of-service guarantees aside. However, the use of multi-hop transmission negatively impacts the end-to-end packet latency. In this work, motivated by the need for latency reduction, we design a new link activation policy for self-backhauled Integrated Access and Backhaul systems operating in half-duplex mode. The proposed approach utilizes dynamic queue prioritization based on the number of packets that can be transmitted within a single time slot, enabling more efficient use of resources. Our numerical results show that the proposed priority-based algorithm performs better than existing link scheduling methods for typical system parameter values.

Key words and phrases: 5G, IAB, millimeter wave, half-duplex, link scheduling, network control

1. Introduction

The digitalization of many areas of human activity relies upon a communication system capable of providing a wide range of services. The 5th generation (5G) mobile networks enable the provision of different services including Enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communications (URLLC), and Massive Machine-Type Communications (mMTC).

The services provided by 5G networks require improvements in various performance indicators. For example, eMBB needs to offer high throughput (up to 10 Gbps) and support high mobility devices (up to 500 km/h). URLLC requires delay reduction down to one millisecond. Finally, for mMTC services, the number of connected devices must be increased to up to 10 million per square kilometer, while also improving their energy efficiency [1].

In order to provision the required performance indicators in 5G, significant changes have been made to the architecture and operations of the 5G core (5GC) and radio access networks (RAN). For example, flexibility and adaptability in synchronization procedures, as well as the allocation and splitting of bands into subcarriers, have been increased. Additionally, modulation, coding, and error correction have been improved [2].

In addition to enhancing the RAN functionality, an important technical innovation of 5G is its substantially expanded frequency range. This allows for higher throughput by allocating vast bandwidth at high frequencies (greater than 24 GHz), while maintaining wide coverage through the utilization of lower frequencies. It is worth noting though that communications in the new high-frequency spectrum suffer from high propagation losses and require significant capital expenditures for upgrading and expanding network hardware infrastructure. In particular, as the coverage area of a base station is reduced due to propagation issues, network densification is necessary, which involves increasing the number of access points (APs) per unit area.

One way to densify 5G networks is to utilize the Integrated Access and Backhaul (IAB) technology. It employs relay nodes that are not wired connected to the core network as additional APs. The

interference issues in the resulting multi-hop wireless network call for the half-duplex transmission, meaning that no network node can receive and transmit data at the same time. In turn, a half-duplex system requires an efficient link activation policy, which determines over which links data can be transmitted at any given time.

In this paper, we aim to design a new link activation policy for 5G IAB networks that allows for packet delay reduction and throughput maximization and can be employed in both centralized and distributed manners. The rest of the paper is structured as follows. First, in Section 2, we discuss the IAB technology and briefly overview the related work. Then, we formalize the model of an IAB network in Section 3 and propose a new link activation policy in Section 4. Next, in Section 5, we obtain realistic simulation parameters and numerically evaluate performance of the proposed policy in comparison with well-known link activation algorithms. Conclusions are drawn in the last section.

2. Background and related work

To minimize capital expenditures in deploying dense 5G networks, the 3GPP (3rd Generation Partnership Project) standardization body has proposed the IAB [3]. IAB allows to use relay nodes that are not directly connected to the core network as relaying APs. As depicted in figure 1, there are two types of APs in an IAB network: an IAB donor directly connected to the core network by a wired link, and one or more IAB nodes which transmit traffic from or to the core network through the IAB donor. The wireless links in the IAB network are divided into two types: access links between an AP and a User Equipment (UE), and backhaul links between APs. Both types of links use a shared time-frequency resource, as the name of the technology implies.

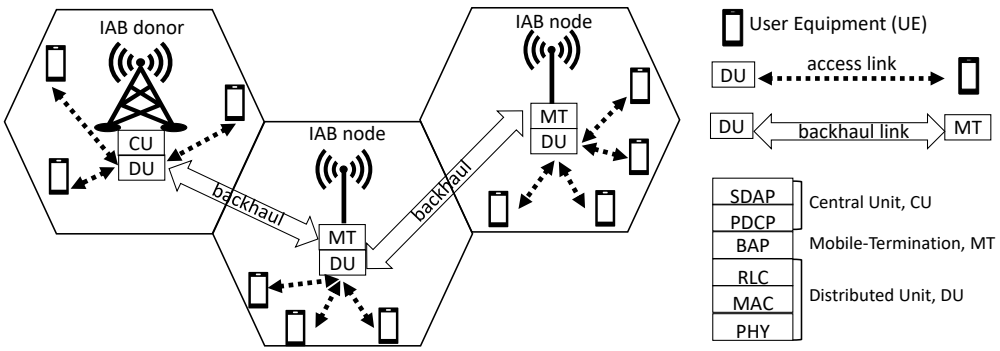


Figure 1. The main components of the IAB network

The IAB technology is based on the distributed architecture of 5G networks. This architecture separates the layers of the data transfer protocol stack between central and distributed units, as shown in figure 1. A Distributed Unit (DU) implements Radio Link Control (RLC), Medium Access Control (MAC), and Physical Layer (PHY). The DU is present at each AP and ensures the establishment, maintenance, and termination of radio connections. The Central Unit (CU) implements Service Data Adaptation Protocol (SDP) and Packet Data Convergence Protocol (PDCP). The CU is only present in the donor and provides connection with the core network. Each IAB node contains a Mobile Termination (MT). This component supports the Backhaul Adaptation Protocol (BAP), which forwards data streams that travel through multiple IAB nodes to and from the IAB donor.

In the first IAB standardization document [3] released in 2018, the IAB network was defined as a multi-hop wireless network with static APs and the ability of path selection. Also, the standard provides a list of possible options for implementation. For example, either in-band or out-of-band backhauling can be used. The use of time, frequency, or spatial multiplexing is permitted, as is end-to-end or hop-by-hop automatic repeat request (ARQ). The resource allocation is not fully determined by the standard, and has been explored in various research projects. For an extensive review, see [4].

As previously mentioned, the IAB standard allows the simultaneous operation of access and backhaul links within the same frequency band. This reduces the downtime for the radio resources,

but also increases interference [5]. Each transmitter interferes with all other active receivers in the network, except the one it is communicating with. The high-frequency 5G spectrum allows for directional transmission, reducing interference in many channels. Nevertheless, interference that occurs during simultaneous reception and transmission remains significant [6].

To eliminate interference caused by simultaneous reception and transmission in the IAB network, the standard [3] recommends using the half-duplex mode. This mode helps to reduce interference by limiting the number of channels on which transmission occurs at any given time. More precisely, half-duplex mode prevents any AP in the IAB network from receiving and transmitting data simultaneously. Although the half-duplex mode limits the network throughput and increases delays, it is an effective and simple way to reduce interference.

To efficiently implement half-duplex, it is essential to schedule transmission over links. This can be done by dividing time into slots and marking each link with 1 (ON) if it is allowed to transmit in the slot and 0 (OFF) otherwise [7–10]. Such link scheduling permits to ensure that the half-duplex constraints are met and to optimize selected performance metrics. For example, in [9] the link scheduling algorithm maximizes minimal user throughput, in [10] it optimizes the sum of user throughputs, and in [7, 8, 11] it targets some convex function of user throughput (such as the sum of logarithms).

In [7–11] the link scheduling is performed by solving an optimization problem with the objective function of throughput. On the other hand, constructing a queuing model of the studied network allows to evaluate and optimize the delay [12, 13], as well as to prove the stability of the network under some scheduling algorithms with any acceptable rates of incoming traffic [11, 14]. This approach was used to derive a number of link scheduling algorithms for general multi-hop wireless networks with interference, and in particular several throughput optimal greedy dynamic algorithms for efficient centralized control of multi-hop networks, which choose a transmission mode based on the current system state via argmin or argmax. *Backpressure* [15] is the most recognized throughput-oriented algorithm for network control and can be utilized for link scheduling, routing or flow control problems [11, 16–18]. While *backpressure* handles queue lengths, such algorithms as the *largest weighted delay first* [19, 20], *oldest cell first* [21] and *delay-based backpressure* [22] use packet delays to specify the system state. The latter is the delay-based version of *backpressure* and allows to reduce the maximum packet delay in the original *backpressure* algorithm. The *α-algorithm* [23] is a modification of *backpressure* aimed at reducing the total delay while remaining optimal in throughput. It uses a constant $\alpha \geqslant 1$ as a per-component power in the *backpressure* algorithm to point up the longest queues. The *αβ-algorithm* [24] algorithm aims to reduce the probability of buffer overflow and thus to provide shorter queue lengths and smaller delays. The activation of a link in this algorithm depends on the lengths of all queues that packets have passed before this link and will pass after.

The introduction of the IAB technology has revived interest in existing link scheduling methods for multi-hop wireless networks, however they should be analysed and modified by taking into account the specifics of IAB and the needs of 5G services. The present paper provides a step in this direction.

3. Model formalization

We consider a half-duplex IAB network where transmission takes place over either access or backhaul links at any given time. Furthermore, a link may be activated in either the uplink or downlink direction. We assume that the throughput of each link is constant. Additionally, we assume that all data packets traversing the network have the same size, and thus, in what follows, a packet is used as a unit of data.

We represent the considered IAB network as a directed graph consisting of four vertices as shown in figure 2. The vertices represent the IAB donor (the circle), the IAB node (the square), the UEs connected to the IAB donor (modeled as a single vertex and depicted by the left triangle), and the UEs connected to the IAB node (also modeled by a single vertex, the right triangle). The edges correspond to the communication links for direct wireless transmission. In what follows we use the terms vertices and nodes, as well as edges and links interchangeably.

The links are divided into uplink, which carry packets from UEs to the IAB donor, and downlink, carrying data from the IAB donor to UEs. Furthermore, a link can be either backhaul, responsible for data transmission between the IAB donor and node, or access, connecting UE nodes to their access points, see figure 2.

Each link of the IAB network graph can be viewed as a server accompanied by a queue of unlimited size where packets awaiting transmission are stored. The system can thus be represented by a queuing network depicted in figure 3. It consists of $I = 6$ service nodes (or queues) with queues 1, 3 and 5

corresponding to the downlink links, and 2, 4 and 6 – to the uplink. Queues 1 and 2 are coupled with backhaul links, and the rest – with the access links. Packets departing queue 1 enter queue 3, and packets departing queue 4 enter queue 2, which describes the two-hop transmission. Packets departing queues 2, 3, 5, 6 leave the system. The set of all queues is denoted by $\mathcal{I}$.

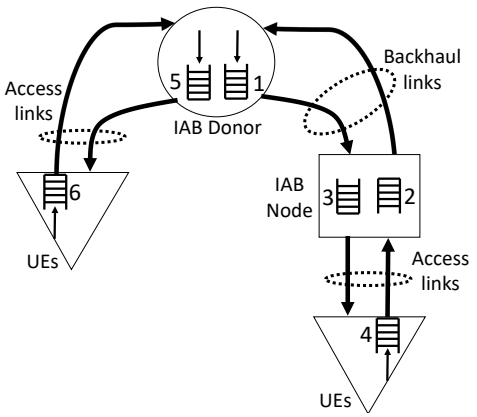


Figure 2. The considered IAB network as a directed graph

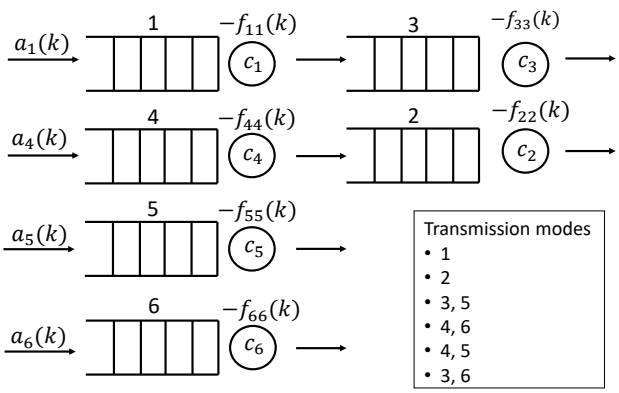


Figure 3. The queuing network corresponding to the modeled IAB network

The system is considered in discrete time indexed by $k = 0, 1, 2, \ldots$. We denote by $a_i(k)$ the number of packets exogenously arriving to the i-th queue in time slot $k \geqslant 0$. We have $a_2(k) = a_3(k) = 0$ for all $k \geqslant 0$, because packets entering queues 2 and 3 are first serviced in stations 4 and 1, respectively. For each of the remaining queues $i \in \mathcal{I}_0 = \{1, 4, 5, 6\}$ it is assumed that $a_i(k)$, $k \geqslant 0$, are independent and identically distributed (i.i.d.) random variables with finite first and second moments. We denote the row vector of arrivals in time slot k by $\mathbf{a}(k) = (a_i(k))_{i\in\mathcal{I}}$. The arrival rate to queue i is equal to the expectation of $a_i(k)$ and denoted by $\lambda_i = \mathbb{E}a_i(k)$.

We say that a packet is served when it is transmitted over a link, and that a queue is served (or active) when the packets it holds are serviced. The service duration is assumed exactly one time slot. Packets are served in batches. The maximum size of a batch that can be served in queue i in one time slot is fixed and denoted by $c_i \in \mathbb{N}$. The column vector $\mathbf{c} = (c_i)_{i\in\mathcal{I}}$ is called the link capacity vector. If the number of packets in an active queue i is fewer than c_i, then all packets in the queue are served in the time slot, otherwise packets are taken for service according to the discipline First Come First Served (FCFS), i.e., in the order of arrival.

The IAB specifics impose constraints on simultaneous activation of queues. By a transmission mode we understand a feasible combination of simultaneously active queues. Queues i and j such that $i \in \{1, 2\}$ and $j \in \{3, 4, 5, 6\}$ and the queues 1 and 2, 3 and 4, and 5 and 6, pairwise cannot be active in the same time slot due to the half-duplex constraints. Moreover, to maximize resource utilization, we do not consider transmission modes that activate fewer queues than allowed by the constraints. This results in the following transmission modes for the system: {1}, {2}, {3, 5}, {4, 6}, {4, 5}, {3, 6}. We denote the set of these transmission modes by Θ and assume they are indexed by l=1, ..., L, $L = |\Theta| = 6$, in the above order.

To specify the connectivity corresponding to the transmission modes listed above, we define, for each $\theta \in \Theta$, an $I \times I$ matrix $\mathbf{F}$ with elements

$$f_{i,j} = \begin{cases} 1, & \text{if } i \in \theta, \quad (i, j) \in \{(1, 3), (4, 2)\}, \\ -1, & \text{if } i \in \theta, \quad j = i, \quad i \in \mathcal{I}, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

We denote the set of such matrices by $\mathcal{F}$ and let them be ordered and indexed as in Θ. Since there is a one-to-one correspondence between the sets $\mathcal{F}$ and Θ, in what follows, we will specify a transmission mode by either $\theta \in \Theta$ or $\mathbf{F} \in \mathcal{F}$ interchangeably.

We assume that in each time slot only one transmission mode can be applied by a controller. Thus, in each time slot k, the system operates according to $\mathbf{F}(k) \in \mathcal{F}$.

Figure 4 shows the timing of events in a time slot $k \geqslant 0$, by which we understand the time $[t_k,\ t_{k+1})$, where $\Delta = t_{k+1} - t_k$ is a constant time slot duration. At the beginning of time slot k the system assumes a transmission mode $\mathbf{F}(k)$ for the time slot. Then, the queues activated by $\mathbf{F}(k)$ are served. Served packets from queues 2, 3, 5 and 6 depart the system, and served packets from queues 1 and 4 move, respectively, to queues 3 and 2. Then, before the end of time slot k, new packets arrive into the system and join queues 1, 4, 5, 6. Thus, no packet can join and depart a queue in one time slot.

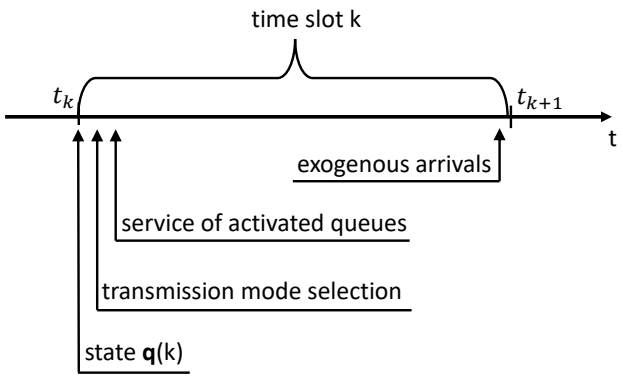


Figure 4. Timing of events in the considered model

Denote by $\mathbf{q}(k) = (q_i(k))_{i \in \mathcal{I}}$ a row vector whose entry $q_i(k)$ is the number of packets in queue i at the beginning of time slot k. Let $\mathbf{q}(0) = \mathbf{0}$ be a zero row vector of length I. Let a row vector $\mathbf{s}(k) = (s_i(k))_{i \in \mathcal{I}}$ with entries

$$s_i(k) = \min(c_i, q_i(k)), \quad i \in \mathcal{I}, \tag{2}$$

represent the number of packets that will be served in queue i in time slot k if the queue is active in this slot. Now, vector $\mathbf{q}(k+1)$ defining the system state in time slot $k+1$ relates to $\mathbf{q}(k)$ and the transmission mode $\mathbf{F}(k)$ as

$$\mathbf{q}(k+1) = \mathbf{q}(k) + \mathbf{s}(k)\mathbf{F}(k) + \mathbf{a}(k), \quad k \geqslant 0. \tag{3}$$

In what follows, we also assume that the transmission mode $\mathbf{F}(k) \in \mathcal{F}$ chosen in time slot k depends only on the system state at time k given by $\mathbf{q}(k)$. A function $\pi(\mathbf{q}(k)) = \mathbf{F}(k)$ will be referred to as the link scheduling (or control) policy.

The capacity region of the system is defined as the set of all combinations of arrival rates $(\lambda_1, \lambda_4, \lambda_5, \lambda_6)$ such that there exists a control policy that provides a finite time-average number of packets in the system operating with these rates as $k \to \infty$. Having a finite average number of packets in all queues is considered as a network stability criterion. A control policy providing network stability for all sets of arrival rates in the capacity region is called throughput optimal [25].

For the considered model, the network capacity region can be obtained as follows. Let p_l be the fraction of time when transmission mode θ_l, $l = 1, \dots, L$, is applied given some control policy π. Note that $\sum_{l=1}^{L} p_l = 1$ as only one transmission mode can be applied in each time slot. Now, the condition for the system to have a finite average number the packets can be written as

$$\begin{aligned} \lambda_6 &\leqslant c_6(p_4 + p_6), \quad \lambda_4 \leqslant c_4(p_4 + p_5), \\ \lambda_5 &\leqslant c_5(p_3 + p_5), \quad \lambda_1 \leqslant c_3(p_3 + p_6). \end{aligned} \tag{4}$$

By dividing each inequality by the capacity of the corresponding link and then summing up, we obtain the capacity region of the system in the form

$$\frac{\lambda_6}{c_6} + \frac{\lambda_4}{c_4} + \frac{\lambda_5}{c_5} + \frac{\lambda_1}{c_3} + \frac{2\lambda_1}{c_1} + \frac{2\lambda_4}{c_2} \leqslant 2. \tag{5}$$

As a key performance indicators we consider the average end-to-end delay $\overline{D}$ and the 99th percentile of the end-to-end delay probability distribution, denoted by P_{99}. The end-to-end delay is defined for each packet that has departed the system as its sojourn time in the system. We also consider such important aspects of every control policy as its throughput optimality and the control-induced overhead.

4. Centralized and distributed priority-based link scheduling

The idea behind the proposed priority-based link scheduling algorithm is as follows. To chose the transmission mode, we first prioritize the transmission modes according to whether the activated thereby queues hold more packets than can be served in one time slot. Then, among the transmission modes with the highest priority, we choose the one providing transmission of the greatest number of packets. This approach is similar to the *P-TREE* algorithm [26], which is a low-complexity scheduling algorithm designed for the multi-hop tree-shaped networks with only uplink traffic.

Recall, that for $\mathbf{F} \in \mathcal{F}$ a diagonal entry $f_{i,i}$ is -1 or 0 depending on whether queue i is activated or not in the transmission mode specified by $\mathbf{F}$. In the proposed *priority-based* algorithm, in each time slot k we obtain a set of priority transmission modes $\mathcal{F}^*(k)$ by the following procedure consisting of three steps:

1. Let the priority set $\mathcal{F}^*(k)$ include all transmission modes for which the maximum possible number of packets is served in all active queues, i.e., let

$$\mathcal{F}^*(k) := \{\mathbf{F} \in \mathcal{F} : s_i(k) f_{i,i} = c_i f_{i,i} \ \forall i \in \mathcal{I}\}. \tag{6}$$

2. If after Step 1 the set $\mathcal{F}^*(k)$ is empty, then let it include all transmission modes for which the maximum possible number of packets is served in at least one queue, i.e., let

$$\mathcal{F}^*(k) := \{\mathbf{F} \in \mathcal{F} : f_{i,i} = -1, \ s_i(k) = c_i \text{ for some } i\}. \tag{7}$$

3. If the set $\mathcal{F}^*(k)$ is still empty, then let $\mathcal{F}^*(k) := \mathcal{F}$.

Now, among the transmission modes of set $\mathcal{F}^*(k)$ we choose the one that results in serving the most packets in time slot k. Let $\mathrm{diag}(\mathbf{F}) = (f_{i,i})_{i \in \mathcal{I}}$ denote the column vector of diagonal elements of matrix $\mathbf{F}$. Since the number of packets served in time slot k under transmission mode $\mathbf{F}$ is $-\mathbf{s}(k)\mathrm{diag}(\mathbf{F})$,

the sought transmission mode is given by

$$\pi_{pb}(\mathbf{q}(k)) = \underset{\mathbf{F}\in\mathcal{F}^*(k)}{\operatorname{argmin}}\, \mathbf{s}(k)\operatorname{diag}(\mathbf{F}). \tag{8}$$

The choice of a transmission mode based on the current network state requires significant signaling. Next, in this section we propose an approach to designing a policy for distributed link scheduling whose performance is close to that of the centralized algorithm. The method is based on the use of *shadow queues* introduced in [27] and then implemented for delay reduction in multi-hop networks in [18]. We assign to each queue $i \in \mathcal{I}$ a shadow queue, which is a variable $\tilde{q}_i(k)$ such that the row vector $\tilde{\mathbf{q}}(k) = (\tilde{q}_i(k))_{i\in\mathcal{I}}$ evolves as

$$\tilde{\mathbf{q}}(k+1) = \tilde{\mathbf{q}}(k) + \min(\tilde{\mathbf{q}}(k), \mathbf{c})\mathbf{F}(k) + \tilde{\lambda}, \quad k \geq 0, \tag{9}$$

where min represents a per-component minimum. Here

$$\tilde{\lambda} = \big((1+\epsilon_1)\lambda_1, \ldots, (1+\epsilon_I)\lambda_I\big), \tag{10}$$

is a row vector in which ϵ_i, $i \in \mathcal{I}$, are positive constants such that $((1+\epsilon_i)\lambda_i)_{i\in\mathcal{I}_0}$ belongs to the system's capacity region.

The dynamics of the *shadow queues* (9) differ from that of the actual queues $\mathbf{q}(k)$ given by (3) in the use of the fixed $\tilde{\lambda}$ instead of the random disturbance $\mathbf{a}(k)$ representing the actual numbers of arrivals. Arrival rates λ_i and constants ϵ_i may not be integers, hence the components of $\tilde{\mathbf{q}}(k)$ may not be integers either, unlike the components of $\mathbf{q}(k)$.

As previously for the actual queues, we let $\tilde{\mathbf{q}}(0) = \mathbf{0}$. Then, to obtain $\tilde{\mathbf{q}}(k+1)$ by (9), its value in time slot k, $\tilde{\mathbf{q}}(k)$, is used in some given centralized control policy π_c to select a transmission mode, i.e., $\mathbf{F}(k) = \pi_c(\tilde{\mathbf{q}}(k))$. The chosen transmission mode is then substituted in (9). Thus, transmission mode selection does not depend on the actual network state and can be implemented in a distributed manner, where all nodes locally use the same policy π_c with the same fixed disturbance $\tilde{\lambda}$ and obtain the same controls, which they apply to the network. It was shown in [18] that such a control ensures a finite average number of packets in all actual queues as long as the non-zero elements of (10) are interior to the capacity region and the policy π_c is throughput optimal.

5. Numerical results

We now proceed illustrating the performance of the proposed approach. We assume that the capacities of the backhaul links, downlink access links, and uplink access links are all pairwise equal. That is, we let $c_1 = c_2$, $c_3 = c_5$, and $c_4 = c_6$. We also assume that IAB network is using the FR2 band with 200 MHz of bandwidth and a subcarrier spacing of 120 kHz, which corresponds to the NR numerology 3. Thus, the number of primary resource blocks, $N_{PRB}^{BW,\mu}$, is equal to 132, and the symbol duration T_s^{μ} is equal to 8.92×10^{-6}. Additionally, the uplink and downlink overheads, as defined by [28], are $OH_{UL} = 0.1$ for the uplink and $OH_{DL} = 0.1$ for the downlink.

We consider three different scenarios, each with a different set of parameter values. In the *maximum UL/DL* scenario, there are no hardware limitations for UEs in both the uplink or downlink directions. In the *limited UL* scenario, the capabilities of UEs are limited in the uplink direction only. Finally, in the *limited UL/DL*, UEs have limitations in both the uplink and downlink directions. Table 1 provides the scenario-specific values for the parameters used in our analysis.

According to 3GPP [28] the data rates of the access links can be estimated as

$$C_{X[\text{Mbps}]} = 10^{-6}\nu_{L,X}Q_{m,X}fR_X\frac{12N_{PRB}^{BW,\mu}}{T_s^{\mu}}(1-OH_X), \quad X \in \{UL, DL\}. \tag{11}$$

Let the time slot duration be 1 ms and let the packet size be 1500 bytes. Thus, to calculate, e.g., the capacity of the downlink access link, $c_3 = c_5$, we first compute C_{DL} in Mbps by (11) and then convert the value to packets per time slot as

$$c_{DL\,[\text{pkts/ms}]} = 10^{-3}(C_{DL\,[\text{Mbps}]} \times 10^6)/(8 \times 1500), \tag{12}$$

after which c_{DL} is rounded down to an integer and assigned to $c_3 = c_5$.

Table 1

Scenario-specific UE parameters

Parameter	**Notation**	**Maximum UL/DL**	**Limited UL**	**Limited UL/DL**
UL number of multiplexed layers	$\nu_{L,UL}$	4	2	1
DL number of multiplexed layers	$\nu_{L,DL}$	6	6	1
UL modulation order	$Q_{m,UL}$	6	4	4
DL modulation order	$Q_{m,DL}$	6	6	6
Scaling factor	f	1	1	0.75
UL error coding rate	R_{UL}	948/1024	490/1024	490/1024
DL error coding rate	R_{DL}	948/1024	948/1024	438/1024
UL rate, Mbps	C_{UL}	3547	611	229
DL rate, Mbps	C_{DL}	4848	4848	280

Since backhaul links are characterized by a higher transmission power and hence a high-order modulation scheme can be used, we take the backhaul link capacities one and a half times as large as the access downlink capacities, i.e., $c_B = 1.5c_{DL}$. Then c_B is also rounded down to an integer and assigned to $c_1 = c_2$. Thus, we obtain three vectors of link capacities $\mathbf{c}$: (606, 606, 404, 295, 404, 295) for *maximum UL/DL*, (606, 606, 404, 50, 404, 50) for *limited UL*, and (34, 34, 23, 19, 23, 19) for *limited UL/DL*.

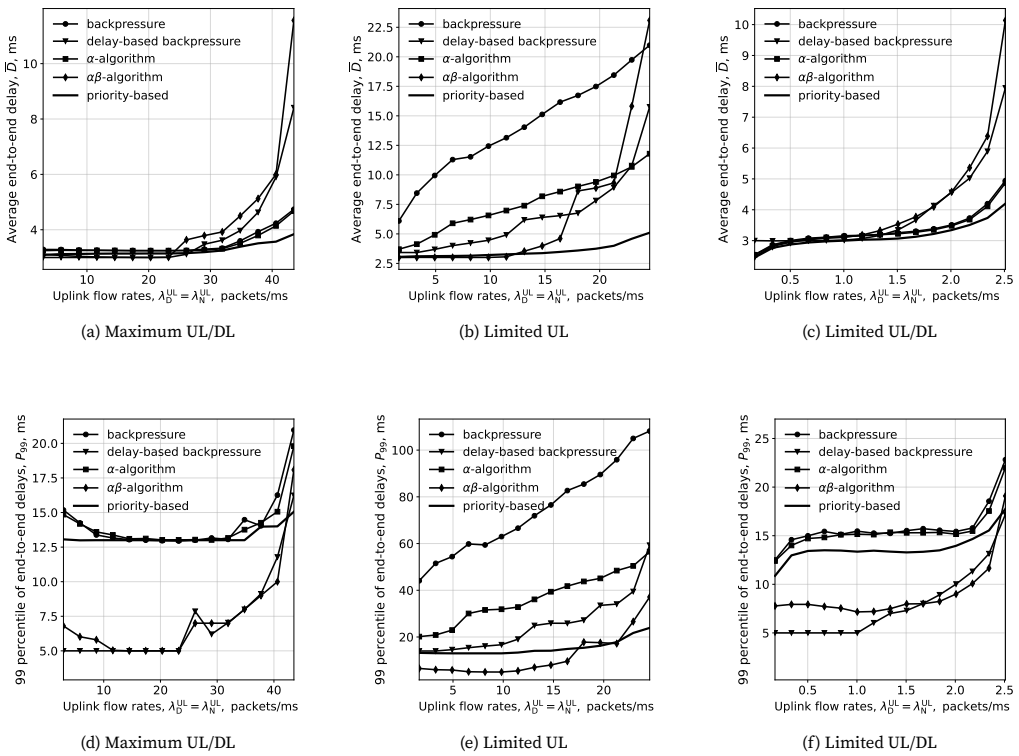
Finally, by following the recommendations for traffic modeling in the standard [3], we assume that the number of packets $a_i(k)$ arriving to queue $i \in \mathcal{I}_0$ in each time slot $k \geqslant 0$ are i.i.d. random variables distributed according to Poisson law with mean λ_i.

We start by comparing the centralized algorithms discussed in Section 2, namely *backpressure, delay-based backpressure, α-algorithm* and *αβ-algorithm*, with the centralized *priority-based* implementation in terms of the average delay $\overline{D}$ and the 99th delay percentile P_{99}. For a convenient presentation of results, we denote the uplink arrival rates from the donor- and node-associated UEs, respectively, as $\lambda_6 = \lambda_{\mathrm{D}}^{\mathrm{UL}}$ and $\lambda_4 = \lambda_{\mathrm{N}}^{\mathrm{UL}}$, and the downlink arrival rates to the donor- and node-associated UEs as $\lambda_5 = \lambda_{\mathrm{D}}^{\mathrm{DL}}$ and $\lambda_1 = \lambda_{\mathrm{N}}^{\mathrm{DL}}$. In all presented figures, at each point, 50 simulation runs, each having 1000 time slots, were generated and then averaged to obtain $\overline{D}$ and P_{99}.

The comparison of the centralized schemes is shown in figure 5, where the arrival rates at each AP are equal and the ratios of the downlink to uplink arrival rates are fixed to four, i.e., $\lambda_{\mathrm{N}}^{\mathrm{DL}} = \lambda_{\mathrm{D}}^{\mathrm{DL}} = 4\lambda_{\mathrm{N}}^{\mathrm{UL}} = 4\lambda_{\mathrm{D}}^{\mathrm{UL}}$. With such parameters, figure 5 shows $\overline{D}$ and P_{99} as functions of the uplink arrival rates for the three studied scenarios.

By analyzing the results in figure 5 we observe that the lowest average delay value is provided by the proposed *priority-based* algorithm. The closest result is demonstrated by the *backpressure* and *α-algorithm* in the *maximum UL/DL* and *limited UL/DL* scenarios. In terms of the 99th percentile P_{99}, *delay-based backpressure* and *αβ-algorithm* show the best performance in *maximum UL/DL* and *limited UL/DL*, whereas in *limited UL* the *αβ-algorithm* performs the best. Moreover, from figure 5 we can see that the *priority-based* policy provides network stability wherever the throughput optimal policies do, i.e., wherever it is possible.

We note that the qualitative behavior of all the algorithms in *maximum UL/DL* and *limited UL/DL* is similar. The rationale is that the elements of the link capacity vectors in these scenarios are closely proportional. Moreover, the ratios of the largest and smallest capacities therein are 2 and 1.8, while this ratio in *limited UL* is 12.1. The range of link capacities' values in *limited UL* is thus considerably wider, and the performance ranking of control policies it yields is different.



(a) Maximum UL/DL (b) Limited UL (c) Limited UL/DL

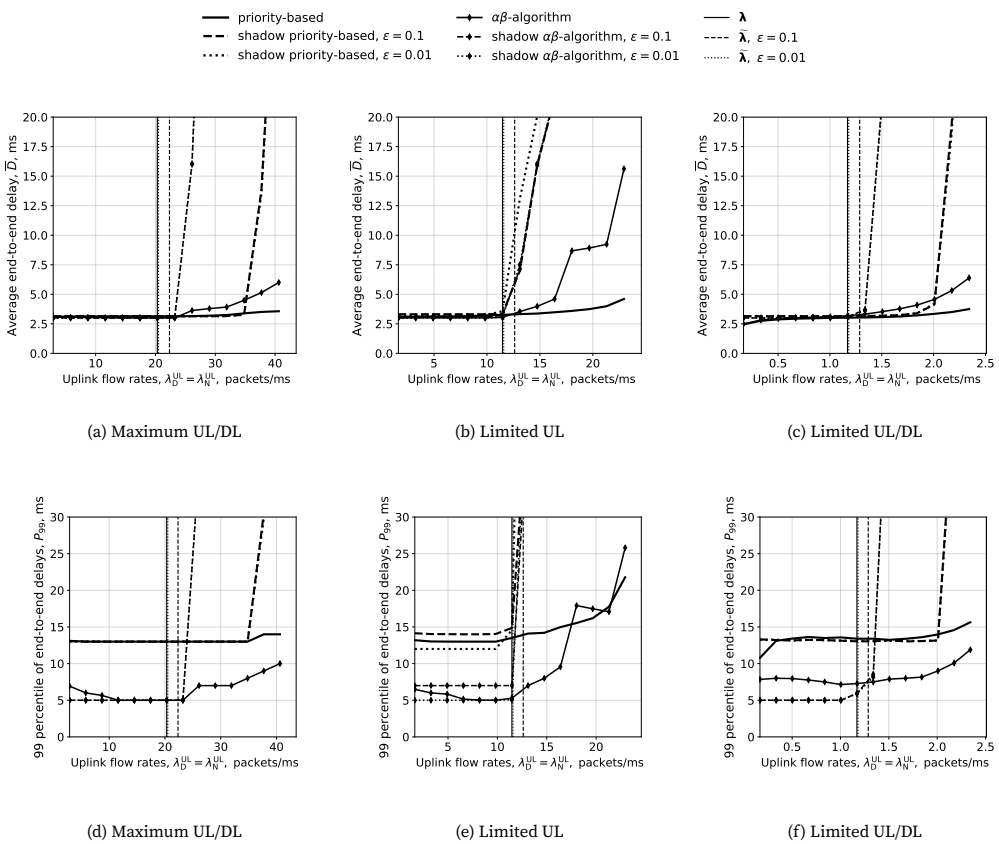
(d) Maximum UL/DL (e) Limited UL (f) Limited UL/DL

Figure 5. Performance evaluation of the centralized link activation policies in terms of the mean delay $\overline{D}$ (top) and the 99th delay percentile P_{99} (bottom) vs. the uplink arrival rates $\lambda_N^{UL} = \lambda_D^{UL}$ with $\lambda_N^{DL}/\lambda_N^{UL} = \lambda_D^{DL}/\lambda_D^{UL} = 4$.

Having identified the *αβ* and *priority-based* algorithms as performing best in terms of the 99th percentile and average end-to-end delay, respectively, we now evaluate their distributed implementations constructed using *shadow queues*. Recall, that a policy choosing the transmission mode based on the *shadow queue* lengths ensures network stability as long as $\tilde{\lambda}$ defined in (10) lies within the capacity region. This means that a larger ϵ can cause instability at high arrival rates but prevents it if the actual arrival rates increase slightly (no more than by $100\epsilon\%$) while $\tilde{\lambda}$ is fixed.

Similarly to figure 5, figure 6 shows the delay metrics $\overline{D}$ (top) and P_{99} (bottom) as functions of the arrival rates. Assuming that the system initially operates with some arrival rates λ, shown in figure 6 by the solid vertical lines, we fix two sets of shadow arrival rates: $\tilde{\lambda}_1$ defined by (10) using $\epsilon_i = 0.1$ for all i and indicated by the dashed vertical lines, and $\tilde{\lambda}_2$ defined using $\epsilon_i = 0.01$ for all i and shown by the dotted vertical lines. Then, we let the actual arrival rates vary along the horizontal and evaluate the system's performance under the centralized control (the results shown by solid lines) and using the *shadow queues* with the arrival rates $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$ fixed previously (dashed and dotted lines, respectively). Thus, to the right from the dashed and dotted vertical lines, the actual arrival rates are greater than the corresponding shadow arrival rates $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$, and to the left they are smaller.

As it could be expected, the *shadow-queues*-controlled network is stable when the actual arrival rates are less than the shadow arrival rates. Additionally, we note that the delay performance is very close to that in a network with centralized control. Interestingly, in the *maximum UL/DL* and *limited UL/DL* scenarios the network is stable even if the actual arrival rates are slightly higher, than the shadow arrival rates. We note that the *priority-based* algorithm maintains the system stable over a wider range of real arrival rates than the *αβ-algorithm*.



(a) Maximum UL/DL (b) Limited UL (c) Limited UL/DL

(d) Maximum UL/DL (e) Limited UL (f) Limited UL/DL

Figure 6. Performance evaluation of the distributed link activation policies in terms of the mean delay $\overline{D}$ (top) and the 99th delay percentile P_{99} (bottom) vs. the uplink arrival rates $\lambda_N^{UL} = \lambda_D^{UL}$ with $\lambda_N^{DL}/\lambda_N^{UL} = \lambda_D^{DL}/\lambda_D^{UL} = 4$.

6. Conclusions

In this paper, we considered the IAB technology enabling cost-effective deployment of dense 5G networks operating in high frequency bands. Specifically, we concentrated on the half-duplex regime and focused on link activation as a critical task for this type of networks. By identifying throughput and delay as relevant performance criteria, we designed a *priority-based* link activation policy for 5G IAB networks, which allows for packet delay reduction and throughput maximization. The proposed policy can be implemented either by the network controller in a centralized way or in a distributed manner by each network node using the proposed *shadow queue* mechanism.

By using a model of an IAB network with a basic topology consisting of one IAB donor, one IAB node, and two groups of UEs we evaluated link activation policies for three scenarios, each with different hardware capabilities. Performance of the proposed policy was compared numerically to the well-known *backpressure* policy and its delay-oriented modifications. We have shown that the centralized *priority-based* policy provides the lowest average end-to-end delay in the considered simulation setup. It also outperforms some of the other studied policies in terms of the 99th delay percentile and achieves stability in the entire capacity region.

Finally, our results also demonstrate that a distributed implementation using *shadow queues* leads to approximately the same delays as the centralized implementation.

Funding: This paper has been supported by the Russian Science Foundation, project no. 23-79-10084, `https://rscf.ru/project/23-79-10084`.

References

1. Molchanov, D. A., Begishev, V. O., Samuilov, K. E. & Kucheryavy, E. A. *5G/6G networks: architecture, technologies, methods of analysis and calculation* 516 pp. (PFUR, 2022).
2. Holma, H., Toskala, A. & Nakamura, T. *5G Technology: 3GPP New Radio* (Wiley, 2020).
3. 3GPP. *Study on Integrated Access and Backhaul* Technical Report (TR) 38.874. Version 16.0.0 (3GPP, Dec. 2018).
4. Monteiro, V., Lima, F., Moreira, D., Sousa, D., Maciel, T., Behrooz, M. & Hannu, H. Paving the Way Toward Mobile IAB: Problems, Solutions and Challenges. *IEEE Open Journal of the Communications Society* **PP,** 1–1. doi:10.1109/OJCOMS.2022.3224576 (Jan. 2022).
5. Sadovaya, Y., Moltchanov, D., Mao, W., Orhan, O., Yeh, S.-p., Nikopour, H., Talwar, S. & Andreev, S. Integrated Access and Backhaul in Millimeter-Wave Cellular: Benefits and Challenges. *IEEE Communications Magazine* **60,** 81–86. doi:10.1109/MCOM.004.2101082 (2022).
6. Hong, S., Brand, J., Choi, J. I., Jain, M., Mehlman, J., Katti, S. & Levis, P. Applications of self-interference cancellation in 5G and beyond. *IEEE Communications Magazine* **52,** 114–121. doi:10.1109/MCOM.2014.6736751 (2014).
7. Ford, R., Gómez-Cuba, F., Mezzavilla, M. & Rangan, S. *Dynamic time-domain duplexing for self-backhauled millimeter wave cellular networks* in *2015 IEEE International Conference on Communication Workshop (ICCW)* (2015), 13–18. doi:10.1109/ICCW.2015.7247068.
8. Ahmed, I. & Mohamed, A. *On the joint scheduling and intra-cell interference coordination in multi-relay LTE uplink* in *2012 IEEE Globecom Workshops* (2012), 111–115. doi:10.1109/GLOCOMW.2012.6477554.
9. Wang, L., Ai, B., Niu, Y., Jiang, H., Mao, S., Zhong, Z. & Wang, N. Joint User Association and Transmission Scheduling in Integrated mmWave Access and Terahertz Backhaul Networks. *IEEE Transactions on Vehicular Technology,* 1–11. doi:10.1109/TVT.2023.3293788 (2023).
10. Qiao, J., Cai, L. X., Shen, X. & Mark, J. W. *STDMA-based scheduling algorithm for concurrent transmissions in directional millimeter wave networks* in *2012 IEEE International Conference on Communications (ICC)* (2012), 5221–5225. doi:10.1109/ICC.2012.6364219.
11. Gómez-Cuba, F. & Zorzi, M. Optimal Link Scheduling in Millimeter Wave Multi-Hop Networks With MU-MIMO Radios. *IEEE Transactions on Wireless Communications* **19,** 1839–1854. doi:10.1109/TWC.2019.2959295 (2020).
12. Yarkina, N., Moltchanov, D. & Koucheryavy, Y. Counter Waves Link Activation Policy for Latency Control in In-Band IAB Systems. *IEEE Communications Letters* **27,** 3108–3112. doi:10.1109/LCOMM.2023.3313233 (2023).
13. Gupta, M., Rao, A., Visotsky, E., Ghosh, A. & Andrews, J. G. Learning Link Schedules in Self-Backhauled Millimeter Wave Cellular Networks. *IEEE Transactions on Wireless Communications* **19,** 8024–8038. doi:10.1109/TWC.2020.3018955 (2020).
14. Gopalam, S., Hanly, S. V. & Whiting, P. Distributed and Local Scheduling Algorithms for mmWave Integrated Access and Backhaul. *IEEE/ACM Transactions on Networking* **30,** 1749–1764. doi:10.1109/TNET.2022.3154367 (2022).
15. Tassiulas, L. & Ephremides, A. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Transactions on Automatic Control* **37,** 1936–1948. doi:10.1109/9.182479 (1992).
16. Neely, M., Modiano, E. & Li, C.-P. *Fairness and optimal stochastic control for heterogeneous networks* in *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.* **3** (2005), 1723–1734 vol. 3. doi:10.1109/INFCOM.2005.1498453.
17. Li, Q. & Negi, R. *Scheduling in Wireless Networks under Uncertainties: A Greedy Primal-Dual Approach* in *2011 IEEE International Conference on Communications (ICC)* (2011), 1–5. doi:10.1109/icc.2011.5963357.
18. Bui, L., Srikant, R. & Stolyar, A. *Novel Architectures and Algorithms for Delay Reduction in Back-Pressure Scheduling and Routing* in *IEEE INFOCOM 2009* (2009), 2936–2940. doi:10.1109/INFCOM.2009.5062262.
19. Stolyar, A. & Ramanan, K. Largest Weighted Delay First Scheduling: Large Deviations and Optimality. *The Annals of Applied Probability* **11.** doi:10.1214/aoap/998926986 (Feb. 2001).
20. Andrews, M., Kumaran, K., Ramanan, K., Stolyar, A., Vijayakumar, R. & Whiting, P. Scheduling in a queuing system with asynchronously varying service rates. *Probability in the Engineering and Informational Sciences* **18,** 191–217. doi:10.1017/S0269964804182041 (Apr. 2004).

21. McKeown, N., Mekkittikul, A., Anantharam, V. & Walrand, J. Achieving 100% throughput in an input-queued switch. *IEEE Transactions on Communications* **47,** 1260–1267. doi:10.1109/26.780463 (1999).
22. Ji, B., Joo, C. & Shroff, N. B. *Delay-based Back-Pressure scheduling in multi-hop wireless networks* in *2011 Proceedings IEEE INFOCOM* (2011), 2579–2587. doi:10.1109/INFCOM.2011.5935084.
23. Venkataramanan, V. J. & Lin, X. On Wireless Scheduling Algorithms for Minimizing the Queue-Overflow Probability. *IEEE/ACM Transactions on Networking* **18,** 788–801. doi:10.1109/TNET.2009.2037896 (2010).
24. Venkataramanan, V. J., Lin, X., Ying, L. & Shakkottai, S. *On Scheduling for Minimizing End-to-End Buffer Usage over Multihop Wireless Networks* in *2010 Proceedings IEEE INFOCOM* (2010), 1–9. doi:10.1109/INFCOM.2010.5462117.
25. Neely, M. *Stochastic Network Optimization with Application to Communication and Queueing Systems* doi:10.2200/S00271ED1V01Y201006CNT007 (Jan. 2010).
26. Venkataramanan, V. J. & Lin, X. *Low-complexity scheduling algorithm for sum-queue minimization in wireless convergecast* in *2011 Proceedings IEEE INFOCOM* (2011), 2336–2344. doi:10.1109/INFCOM.2011.5935052.
27. Bui, L., Srikant, R. & Stolyar, A. *Optimal resource allocation for multicast flows in multihop wireless networks* in *2007 46th IEEE Conference on Decision and Control* (2007), 1134–1139. doi:10.1109/CDC.2007.4434451.
28. 3GPP. *User Equipment (UE) radio access capabilities* Technical Specification (TS) 38.306. Version 17.2.0 (3rd Generation Partnership Project (3GPP), Sept. 2022).

To cite: Zhivtsova A. A., Beschastnyy V. A., A new link activation policy for latency reduction in 5G integrated access and backhaul systems, Discrete and Continuous Models and Applied Computational Science 32 (1)(2024)86–98.DOI: 10.22363/2658-4670-2024-32-1-86-98.

Information about the authors

Zhivtsova, Anna A.—bachelor's degree student of Department of Probability Theory and Cyber Security of Peoples' Friendship University of Russia (RUDN University) (e-mail: aazhivtsova@sci.pfu.edu.ru, phone: +7(910)484-71-44, ORCID: https://orcid.org/0009-0007-8438-6850)

Beschastnyy, Vitaly A.—Candidate of Physical and Mathematical Sciences, assistant professor of Department of Probability Theory and Cyber Security of Peoples' Friendship University of Russia (RUDN University) (e-mail: vbeschastny@sci.pfu.edu.ru, phone: +7(905)776-38-58, ORCID: https://orcid.org/0000-0003-1373-4014, Scopus Author ID: 57192573001)

УДК 004.2, 004.7
PACS 07.05.Tp
DOI: 10.22363/2658-4670-2024-32-1-86-98 EDN: CCHVBS

Стратегия активации каналов для снижения задержки пакетов в сетях интегрированного доступа и транзита 5G

А. А. Живцова, В. А. Бесчастный

Российский университет дружбы народов, ул. Миклухо-Маклая, д. 6, Москва, 117198, Российская Федерация

Аннотация. Блокировка путей распространения радиоволн является одним из основных препятствий на пути развертывания сетей сотовой связи пятого поколения (Fifth Generation) Новое Радио (New Radio) в диапазоне миллиметровых волн (30–100 ГГц). Возможным решением данной проблемы является уплотнение сетей радиодоступа, однако оно связано высокими капитальными затратами операторов связи. Экономически эффективное уплотнение может быть достигнуто с помощью технологии интегрированного доступа и транзита (Integrated Access and Backhaul), использующей ретрансляционные узлы между абонентом и базовой станцией. Такие системы были разработаны главным образом для борьбы с блокировками без учета показателей качества обслуживания (Quality of Service). При этом использование ретрансляционных узлов отрицательно влияет на сквозную задержку пакета. В данной работе предлагается новая стратегия активации каналов направленная на сокращение задержек в системах интегрированного доступа и транзита, учитывающая органичения полудуплексной передачи. Предлагаемый подход основан на динамической приоритезации очередей на базе количества пакетов, которые могут быть переданы в одном временно́м слоте. Результаты имитационного моделирования с использованием реалистичных исходных данных показывают, что предлагаемый алгоритм обеспечивает наименьшую среднюю задержку по сравнению с известными подходами для различных значений нагрузки восходящей и нисходящей передачи.

Ключевые слова: 5G, интегрированный доступ и транзит, миллиметровые волны, полудуплекс, управление активацией каналов