Two Stage Downlink Scheduling for Balancing QoS in Multihop IAB Networks

Shashi Ranjan*, Pranav Jha*, Abhay Karandikar* † and Prasanna Chaporkar*
*Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, India
Email: {shashi.svk, pranavjha, karandi, chaporakar}@ee.iitb.ac.in
†Director, Indian Institute of Technology, Kanpur, India
Email: karandi@iitk.ac.in

***Abstract*—The 3GPP has envisioned IAB as a key enabler to support the flexible and dense deployment of 5G cells with significantly reduced deployment cost. However, IAB introduces new research challenges, especially when studying multihop topology. This paper considers the radio resource scheduling problem and presents a QoS-based downlink scheduler designed explicitly for IAB networks. The scheduler is devised after considering multihop relaying topology, QoS requirements and backhaul constraints. We investigate its performance using system-level simulations and show that it fulfills QoS requirements for different types of services, even at heavy network load. The scheduler also maintains excellent fairness among QoS flows belonging to the same service.**

***Index Terms*—IAB, integrated access and backhaul, wireless backhaul, multihop relaying, resource scheduling, QoS aware**

I. Introduction

Ultra-dense deployment combined with millimeter-wave (mmWave) communication has emerged as an effective solution to realize the vision of Fifth Generation (5G) mobile networks [1], [2]. The mmWave spectrum offers enormous bandwidth to achieve a high data rate and low delay, supporting diverse services. Although mmWave communications also introduce severe path and penetration losses, it compensates for them by using high directivity antennas or beamforming [1]. On the other hand, ultra-dense deployment achieves better frequency reuse, energy efficiency and more line-of-sight links due to reduced inter-site distance. Having said that, ultra-dense deployment may not be an economical choice for operators as they need to connect each cell to 5G Core Network (5GC) using a wired backhaul (e.g., optical fiber or digital subscriber line), which may become a significant part of total capital and operational expenditures. Furthermore, the deployment of wired backhaul requires considerable installation time and might not be possible/allowed in certain areas.

Integrated Access and Backhaul (IAB) [3], [4] has been envisaged as a scalable and cost-effective solution to overcome geographical constraints in ultra-dense deployment. In IAB, a Base Station (BS) shares the same spectrum to serve User Equipments (UEs) in access links and communicate with other BSs in wireless backhaul links. Sharing the spectrum or radio resources can be done in either time, frequency or spatial domain. In the IAB network, wired backhaul to the 5GC is only available at a few specific BSs and these BSs act as access gateways for other BSs having wireless backhauls. Thus IAB enables a network to form a multihop backhaul topology where a UE can communicate with the 5GC over any number of wireless backhaul links. An ultra-dense network typically produces massive backhaul traffic in the 5GC, which means a wireless backhaul would require high bandwidth and reliability similar to those of fiber connection. This issue is attempted to be addressed in IAB by utilizing mmWave spectrum for backhaul as well as access. Therefore, IAB reduces reliance on wired backhaul availability at each BS, enabling an operator to provide a faster and more flexible rollout of a 5G network with significantly reduced deployment cost.

Keeping in mind these merits, the 3rd Generation Partnership Project (3GPP) has recently standardized an IAB architecture in Release 16 [3]. The IAB feature, however, poses many open research challenges that are still not addressed. In particular, radio resource scheduling is widely believed to be a valuable tool to use available spectrum and improve overall system throughput efficiently in a wireless network. In literature, many channel-aware radio resource schedulers have been proposed. However, gains of these channel-aware schedulers may be limited in IAB networks due to several reasons, namely IAB network topology, the multihop distance between UEs and the 5GC and sharing of spectrum by access and backhaul links. Specifically, multihop relaying makes resource scheduling more challenging as access traffic passes through multiple backhaul links impacting network and UE performance. In fact, UEs that are far away from IAB-donors may not have sufficient throughput for basic services. The issue worsens if UEs need services with different Quality of Service (QoS) requirements and the network is expected to guarantee them. Therefore, the resource scheduler must be aware of IAB topology QoS requirements and backhaul constraints for effective implementation to improve network performance.

Motivated by the aforementioned scheduling issues, we propose a 2-stage QoS-based and channel-aware downlink scheduler tailored for IAB networks and show that it conforms to UEs QoS requirements. The scheduler also aims for acceptable fairness across UEs belonging to the service type. We describe both stages of the scheduler in detail with its underlying principle for QoS provisioning. The first stage schedules access and backhaul links at a BS. The second stage distributes traffic volume allocated for a backhaul link to its underlying downstream links. Thus the scheduler at a

BS considers QoS requirements of every downstream (direct or indirect) UEs. In other words, the scheduler fulfills the QoS requirements of each UE irrespective of its topological location within the network. The main contributions of this paper are summarized as follows.

- Design scheduling frameworks for delay-constrained services based on their QoS requirements. For example, a service that necessitates a minimum bitrate has a different scheduling framework than a service that does not have such constraints.
- Devise delay frameworks for different types of services. These frameworks are probabilistic and in accordance with the 3GPP definition for delay.
- Propose a simple rate adaptation technique that periodically adapts sending rate based on minimum bitrate and congestion level in the network to improve resource efficiency and control any unwarranted increase in delays.

II. Related Works

Both academia and industry researchers have shown major interest in IAB in recent times. These research activities have demonstrated the feasibility, potentials and challenges of mmWave-based IAB networks using either end-to-end simulations [5], field trials [6] or optimal association and routing [4], [7]. Specifically, on the resource scheduling problem, many existing solutions are based on network throughput maximization [2], [8], [9]. Others have formulated resource scheduling as a joint problem with power control [10], [11], routing [4], [7], [12] and/or network delay [13]–[15]. Particularly, the authors of [15] propose a delay-aware heat diffusion algorithm to provide proportional prioritization of heterogeneous traffic and minimizes time average network delay. Although the algorithm considers heterogeneous traffic, it has not been explored for different traffic bitrates, per-flow end-to-end (E2E) delay and other QoS requirements that we discuss in Section IV.

All these works (and references therein) have shortcomings in some respects. Except for [15], they are designed for a two-hop IAB network, homogeneous and same priority service and without delay differentiation. In practice, however, there may be a mix of services having different priorities and QoS requirements. This is especially true for 5G networks where a wide gamut of services would be available, some of which are not there in previous generations of mobile networks. Therefore, researchers have been developing new techniques to tackle QoS differentiation in mobile networks over the past few years. For example, authors of [16]–[19] have designed QoS-based scheduling algorithms for Long Term Evolution (LTE) or LTE-relay networks based on 3GPP specifications. However, these works may be inefficient for multihop mmWave 5G networks.

Based on the studies above, we observe that literature is still exploring practical solutions for resource scheduling that enable QoS provisioning in IAB networks. Further, it is necessary to maintain a reasonable tradeoff of the QoS characteristics for different services to improve network performance. Our proposed QoS-aware scheduler aims to provide a simple per-flow QoS provisioning in a multihop IAB network according to the 3GPP specifications with no restrictions on the number of hops. The scheduler also provides a balance of QoS performance between higher and lower priority services.

III. An Overview of 3GPP IAB

An IAB network has two Radio Access Network (RAN) elements, namely IAB-donor and IAB-node. An IAB-donor is similar to a next-generation Node B (gNB) and is connected to the 5GC using a wired fiber connection. In contrast, an IAB-node can only wirelessly backhauls UE or access traffic, possibly through multiple hops to IAB-donor. The IAB-donor then forwards access traffic to the 5GC. In the downlink, the IAB-donor first backhaul access traffic to an appropriate child IAB-node. Then IAB-node either forwards it to an intermediate IAB-node or delivers it directly to the UE. An illustration of a three-hop IAB network is shown in Figure 1 with a data path between a UE (UE8) and the 5GC. It should be noted that any data path always has a single access link and any number of backhaul links.

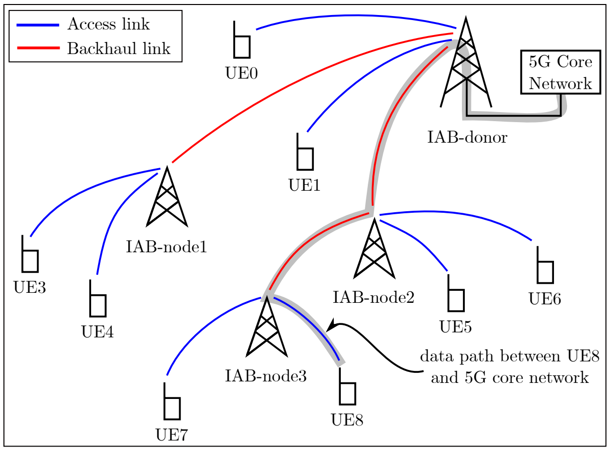


Fig. 1: An illustration of multihop IAB network.

In an IAB network, an IAB-node plays a dual role as a UE from the perspective of upstream IAB-nodes or IAB-donors and as a BS from the perspective of downstream IAB-nodes and UEs. The UE part of a IAB-node is termed as Mobile Termination (MT), which, like a UE, performs cell (re)selection and transmits periodic measurements of backhaul link to its parent node.

To forward access traffic through backhaul links, a new protocol stack known as Backhaul Adaptation Protocol (BAP) [4] is introduced in intermediate hops between the IAB-donor and access IAB-node (the IAB-node serving UE). The IAB-donor configures each IAB-node with a unique BAP ID and as a result, each IAB-node maintains a routing table to forward packets to appropriate child nodes. The IAB-donor inserts header to packets containing BAP ID of the access IAB-node and path identifier before forwarding to the next hop in the downlink. The BAP protocol also performs mapping between ingress and egress backhaul QoS flows to ensure that packets are treated with proper QoS provisioning.

IV. 5G QoS Model

The 3GPP 5G QoS model supports a diverse range of services with different QoS requirements [20]. A QoS flow is primarily defined with eight characteristics – 5G QoS Identifier (5QI), Priority Level, Resource Type Guaranteed Flow Bit Rate (GFBR), Maximum Flow Bit Rate (MFBR), Maximum Data Burst Volume (MDBV), Averaging Time Window (ATW) and Packet Delay Budget (PDB).

The 5QI determines the packet forwarding treatment that gNB has to apply to its packets. Each characteristic helps gNB perform scheduling properly. Priority Level is used to differentiate between QoS flows and prioritize QoS flow over another QoS flow during a scheduling process. A lower Priority Level indicates higher priority.

A QoS flow can be Guaranteed Bit Rate (GBR), Non-GBR or Delay-critical GBR flow. Non-GBR flows are the most generic resource type that most applications use and typically have services like Transport Control Protocol (TCP)-based streaming or File Transfer Protocol (FTP) service. GBR flows are of delay-sensitive resource type and thus are restrictive than Non-GBR flows. They include services like conversational video or real-time gaming. Delay-critical GBR flows are of the most restrictive characteristics. The network should use these flows with caution as they can easily starve other types of flows and do not respect any fairness. Typical services within this category are use cases like robotics and discrete automation.

GFBR specifies the bitrate that is guaranteed to be provided by the network to GBR or Delay-critical GBR flow. For Non-GBR flow, GFBR is considered zero. MFBR is the maximum bitrate for GBR or Delay-critical GBR flow, and excess traffic may get discarded at UE, BS or User Plane Function (UPF) at the 5GC. ATW is the sliding time duration over which GFBR and MFBR are calculated at BS, UE or UPF. The network guarantees bitrates up to GFBR, whereas it tries to fulfill bitrates from GFBR up to MFBR. Consequently, GFBR and MFBR can be mapped to the lowest acceptable and highest required bitrate of service.

PDB defines the upper bound for time that a packet can be delayed between a UE and the UPF. If a packet is delayed beyond PDB, corresponding data is no longer relevant. As we are interested in the budget for RAN, we assume that the delay between the UPF and IAB-donor is negligible.

Typically PDB for GBR and Non-GBR flows indicates "soft upper bound" in the sense that expired packet (i.e., the packet that has exceeded PDB) does not need to be discarded. For GBR flows, it is assumed that congestion-related packet drops do not occur and a packet should not experience delay exceeding PDB with the confidence of 98%. Although Non-GBR flows also have some PDB, their packets are first to be dropped in a situation of scarce resources. For Non-GBR flows, 98% of packets that have not been dropped due to congestion should not experience delay exceeding PDB.

MDBV is only applicable to Delay-critical GBR flows and denotes the maximum amount of data that RAN can deliver within the window of PDB. MDBV should be small to assure delivery without degrading the bandwidth of the whole network. For Delay-critical GBR flows, packet delayed more than PDB is counted as lost if transmitted data burst is not exceeding MDBV and bitrate is not exceeding GFBR.

V. Resource Scheduling in IAB

One of the prime challenges of the IAB resource scheduler is to maintain fairness across different QoS flows irrespective of their topological locations within a network. Enforcing fairness is challenging at IAB-node because its parent node aggregates traffic for multiple UEs into a single backhaul flow before forwarding it. This causes it to be aware of only its backhaul flow and not of downstream QoS flows. BS scheduler may solve this fairness issue if it is provided information (e.g., QoS requirements and buffer status) about downstream QoS flows carried on each backhaul flow. BS also needs to maintain an individual queue for every downstream flow. In effect, the scheduler has to be updated whenever attributes of a QoS flow change on any of its backhaul flows. For example, if the current 5QI is no longer suitable for downstream QoS flow, a new 5QI needs to be assigned to flow and upstream schedulers are updated accordingly. For this reason, the 3GPP has extended Logical Channel ID (LCID) space [3] in Release 16 to identify and maintain each downstream QoS flow within backhaul flows to help in optimal scheduling and resource coordination between parent nodes and their downstream child nodes.

Another challenge in the IAB network is to maintain E2E delay for each QoS flow. Each hop in RAN between sender and receiver takes a slice of PDB, making "effective" PDB at the hop much smaller. It means that with an increasing number of hops, the delay constraint at each hop becomes tighter. Therefore, we need a scheduler that considers delay constraints for QoS flows at each hop, especially for services with strict delay requirements.

VI. System Model

We consider an IAB network having a single IAB-donor and multiple IAB-nodes located at different hop levels from the 5GC. Suppose that there are $I+1$ BSs in the network that are represented by $\mathrm{BS}_i,\ i=0,1,\cdots,I$. We represent the IAB-donor as BS_0 and IAB-nodes as $\mathrm{BS}_i,\ i>0$. We also assume that except for BS_0, each BS_i (specifically its MT) is connected to its parent node.

We describe our downlink scheduling algorithm for generic BS_i as each BS (IAB-donor or IAB-node) performs scheduling irrespective of its hop level. Suppose that BS_i has J_i child nodes, out of which M_i are MTs and the rest $J_i - M_i$ are direct UEs. Without loss of generality, MTs and direct UEs are represented by index $j \leq M_i$ and $j > M_i$, respectively. If MT_j (specifically its corresponding IAB-node) has U_j downstream UEs, the total number of indirect UEs within BS_i is $K_i = \sum_{j=1}^{M_i} U_j$. For the sake of simplicity, we assume that each UE has exactly one QoS flow. The scheduling algorithm that we propose in the next section, however, would also be valid for multiple QoS flows per UE. In the end, BS$_i$ has a total of $J_i - M_i + K_i$ QoS flows, out of which $J_i - M_i$ flows are from direct UEs and K_i flows are from indirect UEs.

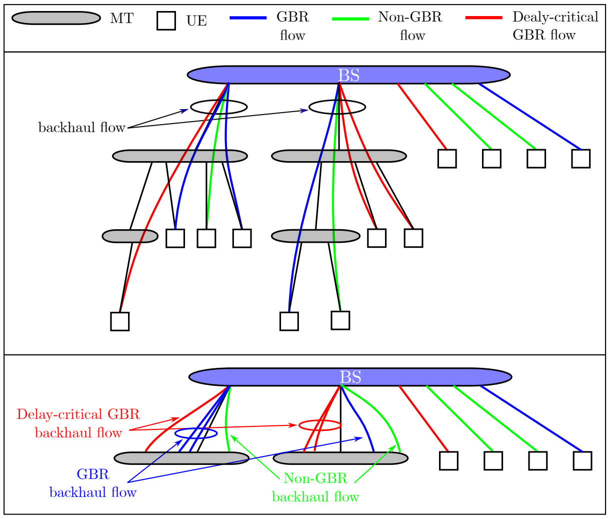


Fig. 2: An illustration depicting (a) accumulation of downstream flows within each backhaul flow and (b) segregation of GBR Non-GBR and Delay-critical GBR backhaul flows.

As MT aggregates flows for multiple indirect UEs of the same resource type into a single backhaul flow, MT$_j$ eventually has three backhaul flows, one of each resource type. The GFBR requirement for backhaul flow of resource type $x = \{\text{Non-GBR GBR Delay-critical GBR}\}$ then becomes

$$\Gamma_j^{(x)} = \sum_{k \in \mathcal{V}_j^{(x)}} \Gamma_{kj}. \tag{1}$$

Here Γ_{kj} is GFBR of downstream flow k under MT$_j$ and $\mathcal{V}_j^{(x)}$ is a set of downstream flows of resource type x under MT$_j$. Again, GFBR for Non-GBR backhaul flow becomes zero. These definitions are needed, as from the scheduling perspective of a BS, a backhaul link act as a normal UE link with QoS requirements of all underlying flows merged into a single “super flow”. BS$_i$, therefore, while scheduling sees a maximum of $J_i + 2M_i$ flows comprising of $J_i - M_i$ flows (GBR, Non-GBR and Delay-critical GBR) of direct UEs and M_i backhaul flows of each resource type. An example of this many-to-one flow mapping process is shown in Figure 2. An MT would not always have backhaul flows of all three types; when MT does not have any downstream flow of a particular resource type, say GBR, it would not have backhaul flow of that type.

VII. QoS Based Scheduling Framework For Each Resource Type

This section presents scheduling frameworks according to 3GPP guidelines and resource types. The objective of these scheduling frameworks is to satisfy the GFBR of each GBR or Delay-critical GBR flow and at the same time maximize throughput of Non-GBR flows with acceptable fairness. In that way, a balance would be reached to avoid starving Non-GBR flows while prioritizing GBR and Delay-critical GBR flows, especially in the non-saturated state.

A. Scheduling framework for Non-GBR flows

As mentioned earlier, most services are tied to Non-GBR flows and the quality of these services is typically determined by their throughput. Therefore, we use Weighted Proportional Fair (WPF) scheduling to serve Non-GBR flows. In WPF scheduling, BS$_i$ evaluates utility for flow k at timeslot t as

$$PF_k[t] = w_k \cdot \frac{R_k[t]}{\overline{R_k}[t-1]}, \tag{2}$$

where $R_k[t]$ and $\overline{R_k}[t]$ represent instantaneous throughput and moving average throughput of flow k at timeslot t, respectively. Weight factor w_k for flow k reflects its scheduling urgency and is defined as

$$w_k = \begin{cases} U_k & \text{if flow } k \text{ is backhaul flow,} \\ 1 & \text{otherwise.} \end{cases}$$

Here U_k is the number of downstream flows within backhaul flow k. Once scheduling round completes, $\overline{R_k}[t]$ is updated as

$$\overline{R_k}[t] = (1-\alpha) \cdot \overline{R_k}[t-1] + \alpha \cdot \mathbb{1}_k[t] \cdot R_k[t], \tag{3}$$

where $\mathbb{1}_k[t]$ is indicator function which equals 1 if flow k was scheduled at timeslot t and 0 otherwise. The factor α is a predetermined constant that affects the tradeoff between throughput and fairness. The default value for α is set as 0.1.

1) Delay requirement: The 3GPP specifications state that 98% of packets related to Non-GBR flows that have not been dropped due to congestion should not experience delay exceeding their PDBs. For this reason, we propose a probabilistic packet drop framework based on the sigmoid function.

Since PDB requirement is meant for UEs and not MTs, we use flow k here to represent any downstream Non-GBR flow. Suppose that packet p of flow k arrives at BS$_i$ at time τ_{pi} then we need to evaluate its effective PDB at BS$_i$ by estimating the amount of time it already has spent in RAN. The E2E delay of a packet after it enters RAN typically depends on the number of BSs along the path and parameters like traffic load, interference and link quality.

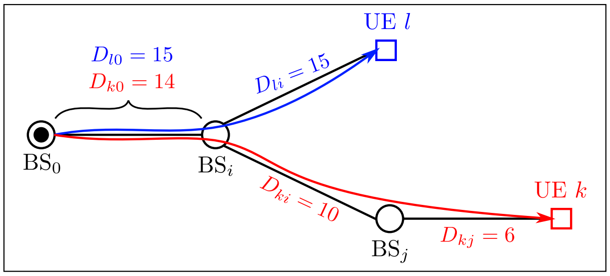


Fig. 3: An example illustrating effective PDBs at different BSs for flows k and flow l belonging to the same service.

In this work, we assume the impact of interference and link quality on E2E delay is negligible due to high directivity in mmWave communications. Hence we propose a load-based effective PDB estimator at the IAB-donor. Suppose that $\mathcal{B}_k$ denotes a set of BSs along the path from flow k to the 5GC and BS$_l \in \mathcal{B}_k$ has U_l number of downstream QoS flows. We further assume that BS$_l$ is assigned a constant fraction ρ_l of bandwidth to serve its downlink traffic. Therefore, packet transmission time at BS$_l$ is proportional to U_l/ρ_l and the IAB-donor estimates effective PDB of flow k at BS$_i$ as

$$D_{ki} = \frac{U_i}{\rho_i}\frac{D_k}{\sum\limits_{l\in\mathcal{B}_k} U_l/\rho_l}, \tag{4}$$

where D_k is the PDB for flow k.The IAB-donor then indicates each BS$_l \in \mathcal{B}_k$ to set effective PDB for flow k to D_{kl}. An example of dividing PDB among BSs (and evaluating effective PDB at each BS) along the path for two UEs with the same service is shown in Figure 3.

The waiting delay for packet p at BS$_i$ at timeslot t is derived as $d_{pi}[t] = t - \tau_{pi}$, which is the time the packet has spent in BS$_i$ buffer. Finally, we drop this packet with probability

$$P_{drop}(d_{pi}[t]) = \begin{cases} 0 & \text{if } d_{pi}[t] < D_{ki} \\ \dfrac{1}{1+e^{-\beta(d_{pi}[t]/D_{ki}+3.896/\beta)}} & \text{otherwise.} \end{cases} \tag{5}$$

Figure 4 depicts the relation of drop probability of packet with its waiting delay and parameter β that controls how fast packet would be dropped once its waiting delay exceeds the effective PDB. We consider $\beta = 10$ as its default value.

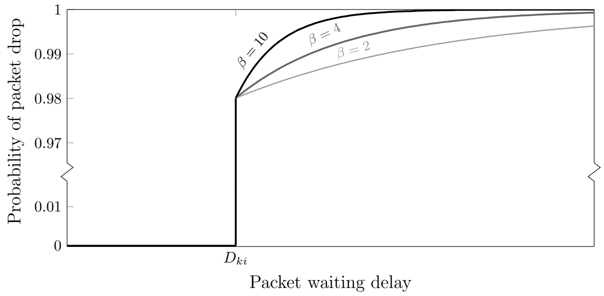


Fig. 4: Packet drop probability of flow k at BS$_i$ as a function of waiting delay. Note that the probability is zero until waiting delay does not violate the effective PDB. Also higher value of β drop packets more aggressively.

B. Scheduling framework for GBR flows

To guarantee GFBR, we propose implementing Modified Deficit Round Robin (MDRR) scheduling to serve GBR flows. In each timeslot, BS$_i$ serves its GBR flows in a round-robin fashion. Each GBR flow k is assigned quantum value Q_k and deficit counter $DC_k[t]$. The quantity Q_k is based on GFBR for flow and specifies the target amount of data that flow expects in each timeslot. If the timeslot duration is σ seconds and flow k has GFBR as Γ_k, then we define

$$Q_k = \sigma \cdot \Gamma_k.$$

The variable $DC_k[t]$ represents the difference between the amount of data actually sent and that should have been sent in timeslot t. Therefore, in the next timeslot, BS should try to serve flow k an amount of data equals to $Q_k + DC_k[t]$. By doing so, flows that received fewer data in the current slot are given an opportunity to receive more data in the next slot.

Our MDRR algorithm is different from the traditional DRR algorithm [21] (and its variants) because we allow packet fragmentation and negative deficit counters. These two modifications enable more granular control upon the amount of traffic through the link over a short-term and long-term horizon.

1) Delay requirement: For GBR flows, packets shall not suffer delays exceeding their PDB with the confidence of 98% even in congestion. This requires tight control on packet delays at each BS. For this reason, we introduce a virtual queue for GBR flow at BS$_i$ called "GBR deadline queue" to store packets whose effective PDB (evaluated from equation (4)) is to be violated in the next few timeslots. These deadline queues are managed virtually as GBR flow queues typically have these 'deadline' packets at the front. Further, GBR deadline queues have higher priority than GBR flow queues.

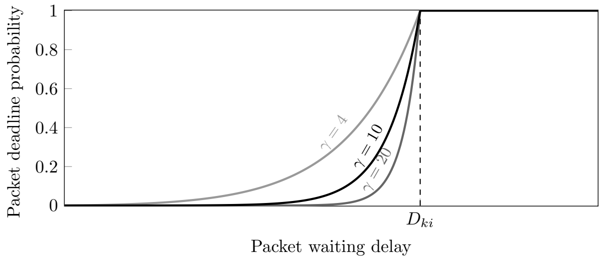


Fig. 5: Probability to move a packet to deadline queue of flow k at BS$_i$ as a function of its waiting delay. A smaller value of γ moves packets to deadline queue more aggressively even when waiting delay is much lower than the effective PDB.

The scheduler first evaluates 'packet deadline probability' of packet p for flow k at BS$_i$ as

$$P_{deadline}(d_{pi}[t]) = \min\left\{1,\ \frac{d_{pi}[t]}{D_{ki}}e^{-\gamma(d_{pi}[t]/D_{ki}-1)}\right\}, \tag{6}$$

where $d_{pi}[t]$ is the waiting delay of packet p. The scheduler then moves packet p to the deadline queue of flow k with this probability only when preceding packets of flow k are also moved to the deadline queue. Figure 5 illustrates relation of packet deadline probability with waiting delay and parameter γ. The default value of γ is 10 for our work.

The scheduling of deadline queues is performed using the proposed Earliest Average Deadline First (EADF) algorithm. Suppose $\chi_i[t]$ is a set of GBR flows at BS$_i$ whose GBR deadline queues are non-empty. We also assume that flow k has a P_k set of packets in its deadline queue. First, we evaluate deadline indicator $\delta_k[t]$ for each flow $k \in \chi_i[t]$ as

$$\delta_k[t] = \frac{1}{|P_k|}\sum_{p\in P_k}\frac{d_{pi}[t]}{D_{ki}}, \tag{7}$$

where $|P_k|$ is the cardinality of set P_k. Then we sort flows in $\chi_i[t]$ based on their deadline indicators and BS$_i$ schedules them according to this sorted list. The EADF+MDRR algorithm for GBR flows is presented in Algorithm 1.

```
Algorithm 1: EADF + MDRR algorithm
Input: $n$ – number of GBR flows
       $d_k[t]$ – data available in normal queue of flow
              $k$ at timeslot $t$
       $\hat{d}_k[t]$ – data available in deadline queue of flow
              $k$ at timeslot $t$
Initialization: $DC_k[0] \leftarrow 0, \quad \forall k = 1, 2, \cdots, n$
       $\hat{L}_i[t] \leftarrow$ sorted list of flows in $\chi_i$ based
       on their deadline indicators
       $L_i[t] \leftarrow$ sorted list of flows based on
       round robin pointer
/* Updating deficit counters */
for $k \leftarrow 1$ to $n$ do
    $DC_k[t] \leftarrow DC_k[t-1] + Q_k$
end
/* EADF algorithm */
for $i \leftarrow 1$ to $|\hat{L}_i[t]|$ do
    if Bandwidth not available then
        break
    end
    $k \leftarrow \hat{L}_i[t](i)$
    if $\hat{d}_k[t] > 0$ then
        $d \leftarrow$ amount of data that can be served with available
          bandwidth
        $d \leftarrow \min\{\hat{d}_k[t],\ d\}$
        allocate sufficient bandwidth to flow $k$ for transmitting $d$
          amount of data
        $DC_k[t] \leftarrow DC_k[t] - d$
    end
end
/* MDRR algorithm */
for $i \leftarrow 1$ to $|L_i[t]|$ do
    if Bandwidth not available then
        break
    end
    $k \leftarrow L_i[t](i)$
    if $DC_k[t] > 0$ & $d_k[t] > 0$ then
        $d \leftarrow$ amount of data that can be served with available
          bandwidth
        $d \leftarrow \min\{d_k[t],\ DC_k[t],\ d\}$
        allocate sufficient bandwidth to flow $k$ for transmitting $d$
          amount of data
        $DC_k[t] \leftarrow DC_k[t] - d$
    end
end
```

We want to emphasize that $DC_k[t]$ is also updated during scheduling of deadline queue; however, BS is not restricted to send up to $DC_k[t]$ in such cases. It means that $DC_k[t]$ can become negative and thus, flow k would not be able to send any data from its GBR flow queue in timeslot t. This may happen when total data for flow k in its deadline queue exceeds $DC_k[t]$.

2) MFBR constraint: We assume that leaky bucket traffic shaper is used at the UPF to regulate the traffic of each GBR flow so that the bitrate does not exceed its MFBR at any time.

3) Rate adaptation: We use Explicit Congestion Notification (ECN) [22] to indicate congestion in the IAB network. When BS experiences congestion, it changes ECN bits within the IP header of all concerned packets. Therefore, when packets reach their respective receivers and if ECN bits show that congestion has happened along the path, receivers notify their senders about congestion. The senders then reduce their sending rate. Conversely, when a receiver reports no congestion, its sender increases sending rate.

Although many service-specific rate adaptation techniques are available, we design simple rate adaptation to show ECN's benefits. We assume that UEs send ECN to their senders at the end of each frame. Any flow k that has experienced congestion in the last frame period (i.e., 10 ms) reports fraction ε_k of packets that have ECN bits. Suppose that sender is associated with set $\mathcal{G}$ of GBR flows and $\mathcal{G}_c$ is a set of GBR flows that reported congestion to the sender. Then the sender reduces sending rate of flow $k \in \mathcal{G}_c$ for frame $T+1$ as

$$r_k[T+1] = \max\{\Gamma_k,\ r_k[T](1-\varepsilon_k)\}, \tag{8}$$

where Γ_k is GFBR of the flow. For flows that did not report congestion, the sender updates sending rate of flow $k \in \mathcal{G} - \mathcal{G}_c$ for frame $T+1$ as

$$r_k[T+1] = \min\{\varUpsilon_k,\ r_k[T](1+\hat{\varepsilon})\}. \tag{9}$$

$\varUpsilon_k$ is MFBR of flow k and $\hat{\varepsilon}$ is defined as

$$\hat{\varepsilon} = \begin{cases} 0.05 & \text{if } |\mathcal{G}_c| = 0 \\ \dfrac{1}{|\mathcal{G}_c|} \displaystyle\sum_{k \in \mathcal{G}_c} \varepsilon_k & \text{otherwise.} \end{cases}$$

Our rate adaptation ensures that the sender keeps sending rate for the flow between its GFBR and MFBR. It also helps in achieving lower packet delays and higher resources utilization.

C. Scheduling framework for Delay-critical GBR flows

Delay-critical GBR flows are similar to GBR flows but with higher priority and stricter delay constraints. Therefore, the scheduling framework for Delay-critical GBR flows is the same as that of GBR flows. The only additional QoS attribute for Delay-critical GBR flows is MBDV.

For satisfying MDBV, we propose to use a token bucket traffic shaper. It allows the desired level of burstiness within flow by limiting its average rate and maximum burst size.

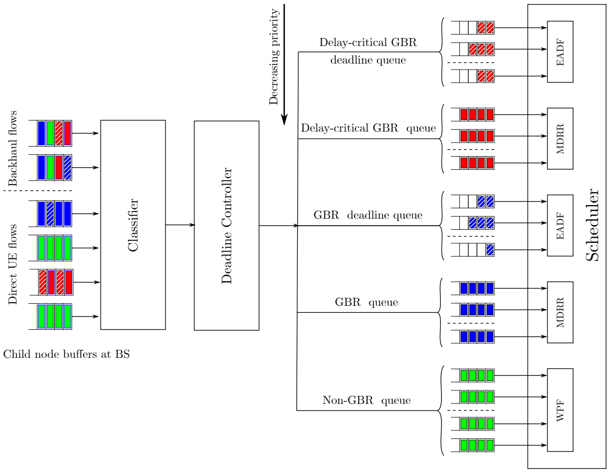


Fig. 6: Resource type classifier and deadline controller at a BS during Stage I. A similar packet processing is performed for flows of indirect UEs during Stage II.

Accordingly, we configure the token rate as MFBR and token bucket size as MBDV. These token buckets are placed before Delay-critical GBR flow queues at the IAB-donor.

VIII. Proposed 2−Stage QoS Flow Based Downlink Scheduler For IAB

In this section, we provide details of the proposed 2-stage QoS-based downlink scheduler. The scheduler combines scheduling frameworks for different resource types described in the last section. The scheduling algorithm is comprised of two stages; the operation of each stage is outlined below.

Algorithm 2: Scheduling GBR and Delay-critical GBR flows on residual bandwidth

```
Input: $L_i$ – set of GBR and Delay-critical GBR flows at BS$_i$
       $d_k[t]$ – data available for flow $k$ at timeslot $t$
while Bandwidth available do
    randomly select $k \in L_i$
    if $d_k[t] > 0$ then
        $d \leftarrow$ amount of data that can be served with available
          bandwidth
        $d \leftarrow \min\{d_k[t],\ Q_k,\ d\}$
        allocate sufficient bandwidth to flow $k$ for transmitting $d$
          amount of data
        $DC_k[t] \leftarrow DC_k[t] - d$
    end
end
```

A. Stage I*: Scheduling flows of child nodes*

In the first stage, BS$_i$ schedules $J_i + 2M_i$ flows of its child nodes. The flows are served based on their priority levels and resource types as summarized in Figure 6. After stage I, the BS allocates any residual bandwidth to GBR and Delay-critical GBR flows and schedules them randomly with the same priority as shown in Algorithm 2.

B. Stage II*: Scheduling flows of indirect UEs*

With the first stage, the BS knows how much data it has to send to each backhaul flow. In the second stage, the BS decides how much data for underlying downstream flows is to be sent within each backhaul flow. Suppose that the BS at timeslot t is to serve $S_j[t]$ amount of data to backhaul flow j after the first stage and backhaul flow j has Ψ_j set of underlying downstream flows. Then the BS schedules these indirect flows using the EADF+MDRR algorithm for GBR and Delay-critical GBR resource types, or PF algorithm for Non-GBR resource type. The stage II scheduling is presented in Algorithm 3. Therefore, scheduling in stage II is similar to the first stage except that the BS distributes $S_j[t]$ data among Ψ_j instead of bandwidth among its child flows. The flow chart of proposed 2-stage scheduling is presented in Figure 7.

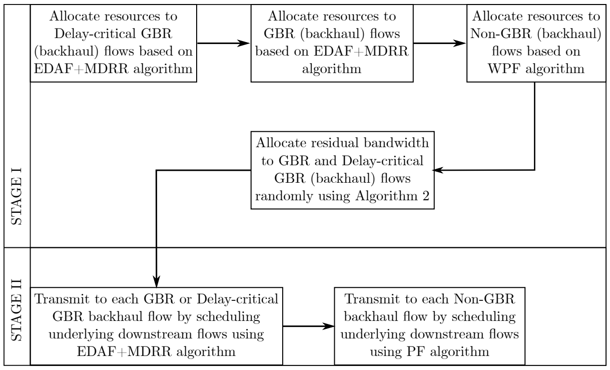

Allocate resources to Delay-critical GBR (backhaul) flows based on EDAF+MDRR algorithm
Allocate resources to GBR (backhaul) flows based on EDAF+MDRR algorithm
Allocate resources to Non-GBR (backhaul) flows based on WPF algorithm
Allocate residual bandwidth to GBR and Delay-critical GBR (backhaul) flows randomly using Algorithm 2
STAGE I
STAGE II
Transmit to each GBR or Delay-critical GBR backhaul flow by scheduling underlying downstream flows using EDAF+MDRR algorithm
Transmit to each Non-GBR (backhaul) flow by scheduling underlying downstream flows using PF algorithm

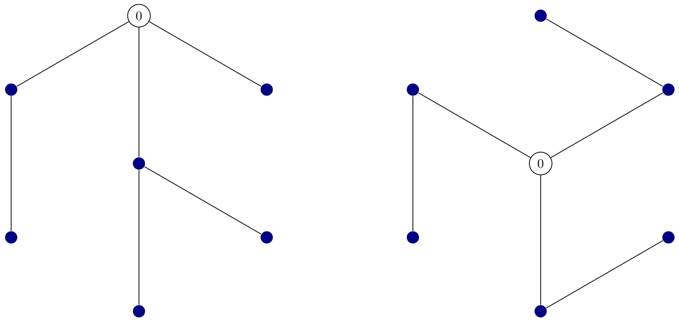
Fig. 7: Flowchart detailing each step within the proposed 2-stage QoS flow based scheduling algorithm.

Algorithm 3: Scheduling of GBR or Delay-critical GBR indirect flows

```
Input: $n^{(x)}$ – number of indirect flows of resource type
         $x = \{\text{GBR},\ \text{Delay-critical GBR}\}$
       $d_k[t]$ – data available in normal queue of flow
         $k$ at timeslot $t$
       $\hat{d}_k[t]$ – data available in deadline queue of flow
         $k$ at timeslot $t$
Initialization: $DC_k[0] \leftarrow 0, \quad \forall k = 1, 2, \cdots, n^{(x)}$
                $Q_k \leftarrow$ GFBR requirement of flow $k$
                $\hat{L}_i[t] \leftarrow$ sorted list of non-empty deadline queues
  based on their deadline indicators
                $L_i[t] \leftarrow$ sorted list of flows based on
                round robin pointer
/* Updating deficit counters                        */
for $k \leftarrow 1$ to $n^{(x)}$ do
    $DC_k[t] \leftarrow DC_k[t-1] + Q_k$
end
/* Iterate through each backhaul flow $j$              */
for $j \leftarrow 1$ to $M_i$ do
    /* EADF algorithm                                */
    $A = \hat{L}_i[t] \cap \Psi_j^{(x)}$
    for $i \leftarrow 1$ to $|A|$ do
        if $S_j[t] \to 0$ then
            break
        end
        $k \leftarrow A(i)$
        if $\hat{d}_k[t] > 0$ then
            $d \leftarrow \min\{\hat{d}_k[t],\ S_j[t]\}$
            allocate $d$ amount of data to flow $k$
            $DC_k[t] \leftarrow DC_k[t] - d$
            $S_j[t] \leftarrow S_j[t] - d$
        end
    end
    /* MDRR algorithm                                */
    flag = True
    $A = L_i[t] \cap \Psi_j^{(x)}$
    for $i \leftarrow 1$ to $|A|$ do
        while $S_j[t] > 0$ do
            $k \leftarrow A(i)$
            if $DC_k[t] > 0$ & $d_k[t] > 0$ then
                if flag is True then
                    $d \leftarrow \min\{d_k[t],\ DC_k[t],\ S_j[t]\}$
                else
                    $d \leftarrow \min\{d_k[t],\ Q_k,\ S_j[t]\}$
                end
                allocate $d$ amount of data to flow $k$
                $DC_k[t] \leftarrow DC_k[t] - d$
                $S_j[t] \leftarrow S_j[t] - d$
            end
        end
        flag = False
    end
end
```

IX. Performance Evaluation

We consider urban micro deployment [3] of IAB network with 7 hexagonal cell sites for one IAB-donor and 6 IAB-nodes. There are N UEs dropped uniformly and randomly within each cell. The IAB network topology that is randomly generated using a minimum spanning tree based on RSRP values is shown in Figure 8. The UEs associate with BSs providing maximum RSRP value in access links.



(a) Topology 1 (b) Topology 2

Fig. 8: Distribution of BSs over a plane in the representative IAB topologies. BS_0 is the IAB-donor and the other BSs are IAB-nodes.

The BSs and UEs are equipped with 16×16 and 4×4 uniform planar antenna arrays, respectively, at both transmitter and receiver sides. For physical layer aspects of mmWave frequencies, we use NYU channel model [1]. We assume that control signalings are instantaneous and don't occupy radio resources. The system-level parameters are derived from [3], [23] and important ones are summarized in Table I. We perform Monte Carlo simulations with 10 runs, where each run lasts for 10000 timeslots.

Parameters	**Urban macro scenario**
Inter-site distance	200 m
Carrier frequency	28 GHz
System bandwidth	100 MHz
Subcarrier spacing	15 KHz
Timeslot length	1 ms
Thermal noise density	-174 dBm/Hz
Pathloss model	UMi Street Canyon
Antenna height	BS: 10 m, UE: 1.5 m
Transmit power	BS: 33 dBm, UE: 23 dBm
Noise margin	BS: 5 dB, UE: 13 dB
MCS index table	based on [23]

TABLE I: Simulation parameters for IAB evaluations.

A. Downlink Resource Allocation

As a highly directional mmWave link act as pseudo-wired connection, interference between any two BSs is assumed to be negligible. BS_0 is assigned fraction ρ_0 of bandwidth and each BS that is directly connected to BS_0 is assigned fraction $\rho_1 = 1 - \rho_0$ of bandwidth. We consider load-based resource coordination between child IAB-node and its parent node. Suppose that BS_0 has M_0 child IAB-nodes and U_0 downstream UEs, and its child BS_i has U_i downstream UEs then we evaluate ρ_1 as

$$\rho_1 = \max_{1 \leq i \leq M_0} U_i / U_0.$$

The resource assignments for other downstream IAB-nodes depend on their hop levels and are illustrated in Figure 9.

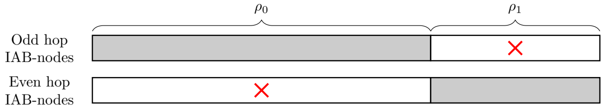


Fig. 9: Example of resource coordination during a timeslot with resources available at the IAB-donor and IAB-nodes.

B. Traffic Model

In this work, we consider three services (with QoS flows belonging to different resource types) listed in Table II. These three services encompass a wide gamut of applications and would help us study the reliability of our downlink scheduler. We use a constant bitrate for the traffic model. For services F_1 and F_2, initial sending rates are set to 1.5 times of their GFBRs, whereas it is 0.5 Mbps for service F_2. Packet size during transmission is fixed at 500 bytes for each service. For assessing reliability aspects, we also analyze a mix of services. The default configuration is a $1:1:1$ service mix, i.e., equal number of QoS flows belong to each service type.

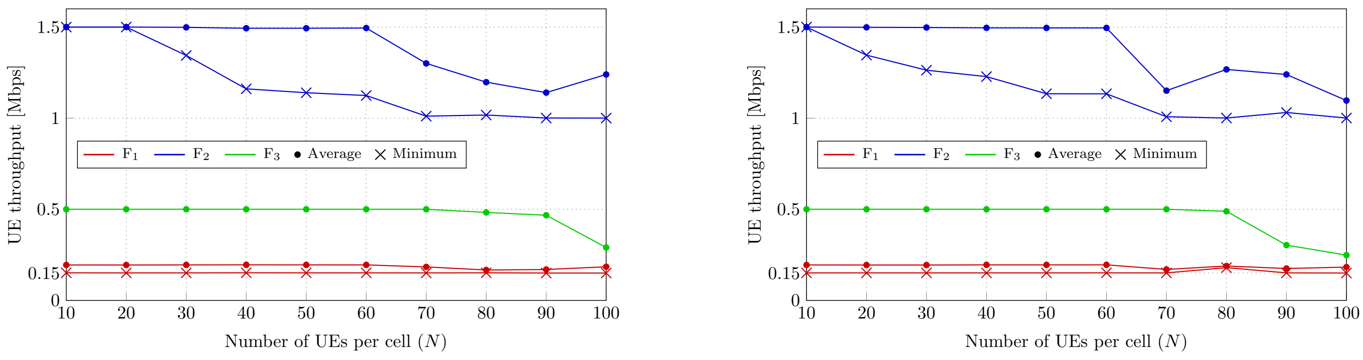
C. Simulation Results

Figures 10 and 11 show the performance of the proposed scheduler for both topologies under the default setting. We observe that each service has fulfilled its QoS requirements even at a heavy network load. For example, Figure 10 illustrates that services F_1 and F_2 have satisfied their GFBRs. We also observe that service F_3 almost maintained a throughput of 0.5 Mbps up to $N = 80$ (or until network is non-saturated). On the other hand, throughputs of services F_1 and F_2 are close to initial sending rates until $N = 60$ and then decrease with increasing N until saturation. After that, service F_3 starts suffering as network load increases. This is expected as when the network enters the saturated region, the scheduler tries to maintain GFBRs of services F_1 and F_2 at the cost of lowering throughput of service F_3. Further, the rate adaption dynamically controls sending rates for services F_1 and F_2 predominantly in the saturated region to combat congestion and reduce queueing delays.

Similarly, Figure 11 depicts that all three services have 98 percentile packet delays within their PDBs under the non-saturated region. Only at excessive network load, service F_2 fails to satisfy its PDB for Topology 1 maybe because, unlike service F_3, we cannot drop packets for service F_2. Further, inadequate bandwidth at a BS can cause deadline queues of services F_1 and F_2 to grow faster, resulting in a higher E2E congestion for service F_2. The figure also demonstrates that average packet delays for services F_2 and F_3 although increase with N are markedly lower than their PDBs. The average

Service	5QI	Priority level	Resource type	GFBR	MFBR	MBDV	PDB	ATW	Service examples
F_1	84	24	Delay-critical GBR	0.15 Mbps	0.3 Mbps	1354 bytes	30 ms	2000 ms	Intelligent transport systems
F_2	2	40	GBR	1 Mbps	2 Mbps	–	150 ms	2000 ms	Conversational video (live streaming)
F_3	6	60	Non-GBR	–	–	–	300 ms	–	Video (buffered streaming), FTP

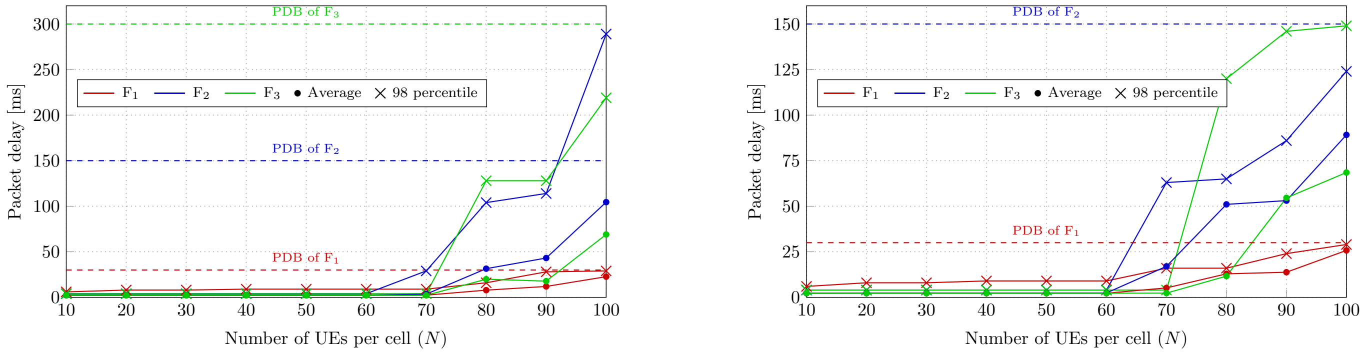
TABLE II: Services considered for performance evaluation of proposed scheduler with their corresponding QoS requirements.



(a) Topology 1

(b) Topology 2

Fig. 10: Minimum and Average throughput for each service with increasing network load.



(a) Topology 1

(b) Topology 2

Fig. 11: 98 percentile and Average E2E delay for each service with increasing network load.

packet delay for service F_1, however, does reach close to its stringent PDB criteria at extreme network loads.

Figure 12 shows fairness (using Jain's formula [24]) of resource allocation among flows belonging to the same service. As evident from the figure, fairness among QoS flows belonging to the same service is above 0.97. It is only after the network enters the saturated state, service F_3 exhibits any noticeable degradation in fairness.

Figure 13 illustrates that service F_1 is almost unaffected by different service mix configurations, whereas throughputs of services F_2 and F_3 depend on these configurations. For example, $1:2:1$ service mix is the most resource-demanding configuration, pushing the network into saturation. As a result, services F_2 and F_3 have the lowest throughputs compared with the other configurations. In contrast, the network performs the best under the $2:1:1$ service mix when resource requirement is the least. Finally, we observe that the scheduler achieves GFBRs for services F_1 and F_2 under each configuration.

X. Concluding Remarks

Many researchers believe that deploying ultra-dense networks is the primary way to realize capacity goals envisaged for the forthcoming 5G technology. However, connecting every cell to 5GC through a wired backhaul is not an economical option for operators. In this regard, IAB may prove to be a promising solution to enable faster and economical deployment of ultra-dense 5G cells. IAB provides wireless backhaul links to BSs and facilitates them to relay access traffic across the network. Further, by sharing the same spectrum between access and backhaul links, IAB also reduces equipment costs.

Radio resource scheduling in IAB is an open research problem. Many QoS-based schedulers designed for single-hop cellular networks become impractical for IAB networks. There are also schedulers designed for two-hop networks, but they become inefficient for multihop topology. Lastly, these proposals are typically not according to the 3GPP specifications, rendering them useless under practical scenarios. In this paper, we present QoS flow-based downlink scheduler tailored for

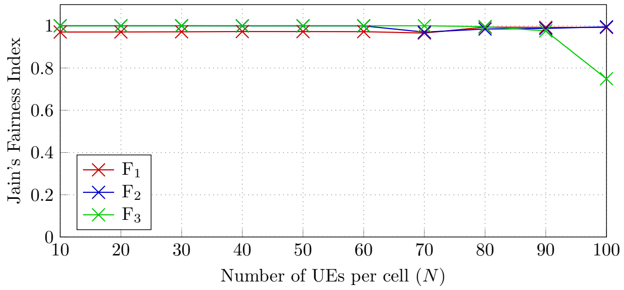


Fig. 12: Fairness of each service for Topology 1.

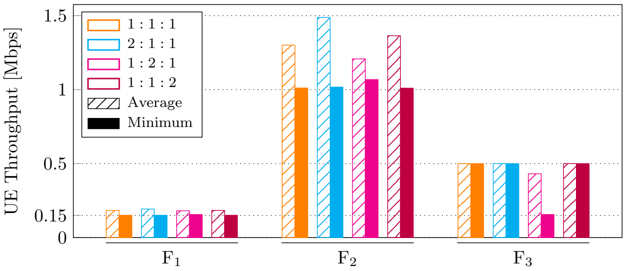


Fig. 13: Minimum and Average throughput for each service under different service mix when $N = 70$.

IAB networks. The scheduler is designed as per the 3GPP specifications and considers multihop relaying topology QoS requirements and backhaul constraints.

The proposed scheduler consists of two stages; the first stage schedules access and backhaul flows, whereas the second stage distributes traffic volume allocated for backhaul flows to their underlying downstream flows. The scheduler serves GBR and Delay-critical GBR flows using the proposed EADF+MDRR algorithm with virtual deadline queues and Non-GBR flows using WPF algorithm with probabilistic packet drop approach. These algorithms are devised by accounting for different QoS characteristics of each resource type. We also proposed a rate adaption technique to control sending rates for flows and congestion levels in the network based on their GFBRs.

We have investigated the scheduler's performance using system-level simulations with a different mix of resource types. Preliminary results show that scheduler can fulfill QoS requirements for each resource type, even at heavy network load. Additionally, our scheduler has maintained excellent fairness for each resource type and fairness degrades only for Non-GBR flows when the network enters saturated region. These are remarkable results considering that UEs at different hop levels may demand different QoS requirements under IAB relaying network. As a result, the scheduler would help facilitate a faster and more flexible rollout of 5G networks with significantly reduced deployment costs.

Acknowledgement

This work has been supported by the Department of Telecommunications, Ministry of Communications, India as part of the Indigenous 5G Testbed project.

References

[1] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE journal on selected areas in communications*, vol. 32, no. 6, pp. 1164–1179, 2014.

[2] C. Saha and H. S. Dhillon, "Millimeter Wave Integrated Access and Backhaul in 5G: Performance Analysis and Design Insights," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 12, pp. 2669–2684, 2019.

[3] 3GPP TR 38.874, "Study on integrated access and backhaul," 2019.

[4] 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.

[5] 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.

[6] T. Tian, Y. Dou, G. Ren, L. Gu, J. Chen, Y. Cui, T. Takada, M. Iwabuchi, J. Tsuboi, and Y. Kishiyama, "Field trial on Millimeter Wave Intergrated Access and Backhaul," *IEEE Vehicular Technology Conference (VTC-Spring)*, pp. 1–5, 2019.

[7] M. N. Islam, S. Subramanian, and A. Sampath, "Integrated Access Backhaul in Millimeter Wave Networks," *IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, 2017.

[8] W. Lei, Y. Ye, and M. Xiao, "Deep Reinforcement Learning-Based Spectrum Allocation in Integrated Access and Backhaul Networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 3, pp. 970–979, 2020.

[9] B. Zhang, F. Devoti, and I. Filippini, "RL-based Resource Allocation in mmWave 5G IAB Networks," *Mediterranean Communication and Computer Networking Conference (MedComNet)*, pp. 1–8, 2020.

[10] Y. Liu, X. Fang, and M. Xiao, "Discrete Power Control and Transmission Duration Allocation for Self-Backhauling Dense mmWave Cellular Networks," *IEEE Transactions on Communications*, vol. 66, no. 1, pp. 432–447, 2017.

[11] Y. Li, E. Pateromichelakis, N. Vucic, J. Luo, W. Xu, and G. Caire, "Radio Resource Management Considerations for 5G Millimeter Wave Backhaul and Access Networks," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 86–92, 2017.

[12] 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.

[13] A. Ortiz, A. Asadi, G. H. Sim, D. Steinmetzer, and M. Hollick, "SCAROS: A Scalable and Robust Self-Backhauling Solution for Highly Dynamic Millimeter-Wave Networks," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 12, pp. 2685–2698, 2019.

[14] Y. Zhang, V. Ramamurthi, Z. Huang, and D. Ghosal, "Co-Optimizing Performance And Fairness Using Weighted PF Scheduling And IAB-Aware Flow Control," *IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, 2020.

[15] J. García-Rois, R. Banirazi, F. J. González-Castaño, B. Lorenzo, and J. C. Burguillo, "Delay-Aware Optimization Framework for Proportional Flow Delay Differentiation in Millimeter-Wave Backhaul Cellular Networks," *IEEE Transactions on Communications*, vol. 66, no. 5, pp. 2037–2051, 2018.

[16] P. Ameigeiras, J. Navarro-Ortiz, P. Andres-Maldonado, J. M. Lopez-Soler, J. Lorca, Q. Perez-Tarrero, and R. Garcia-Perez, "3GPP QoS-based scheduling framework for LTE," *EURASIP Journal on Wireless Communications and Networking*, vol. 2016, no. 1, pp. 1–14, 2016.

[17] N. K. Madi, Z. M. Hanapi, M. Othman, and S. K. Subramaniam, "Delay-based and QoS-aware packet scheduling for RT and NRT multimedia services in LTE downlink systems," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, no. 1, pp. 1–21, 2018.

[18] T. M. de Moraes, A. B. Saleh, G. Bauch, and E. Seidel, "QoS-Aware Traffic Scheduling in LTE-Advanced Relay-Enhanced Networks," *IEEE Vehicular Technology Conference (VTC Spring)*, pp. 1–5, 2013.

[19] J. Hajipour, J. M. Niya, and D. W. K. Ng, "Energy-Efficient Resource

Allocation in Buffer-Aided Wireless Relay Networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 10, pp. 6648–6659, 2017.

[20] 3GPP TS 23.501, "System architecture for the 5G System (5GS)," 2020.

[21] M. H. MacGregor and W. Shi, "Deficits for bursty latency-critical flows: DRR++," *IEEE International Conference on Networks (ICON)*, pp. 287–293, 2000.

[22] K. Ramakrishnan, S. Floyd, D. Black *et al.*, "The addition of explicit congestion notification (ECN) to IP," 2001. [Online]. Available: https://tools.ietf.org/html/rfc3168.

[23] 3GPP TS 38.214, "Physical layer procedures for data," 2020.

[24] D.-M. Chiu and R. Jain, "Analysis of the Increase and Decrease Algorithms for Congestion Avoidance in Computer Networks," *Computer Networks and ISDN systems*, vol. 17, no. 1, pp. 1–14, 1989.