Fast Inter-Base Station Ring (FIBR): A New Millimeter Wave Cellular Network Architecture

Athanasios Koutsaftis, *Student Member, IEEE*, Rajeev Kumar, *Student Member, IEEE*, Pei Liu, *Member, IEEE*, and Shivendra S. Panwar, *Fellow, IEEE*

***Abstract*— Fifth Generation (5G) Millimeter Wave (mmWave) cellular networks are expected to serve a large set of throughput-intensive, ultra-reliable, and ultra-low latency applications. To meet these stringent requirements, while minimizing the network cost, the 3rd Generation Partnership Project has proposed a new transport architecture, where certain functional blocks can be placed closer to the network edge. In this architecture, however, blockages and shadowing in 5G mmWave cellular networks may lead to frequent handovers (HOs) causing significant performance degradation. To meet the ultra-reliable and low-latency requirements of applications and services in an environment with frequent HOs, we propose the Fast Inter-Base Station Ring (FIBR) architecture, where Base Stations (BSs) that are in close proximity are grouped together, interconnected by a bi-directional counter-rotating buffer insertion ring network. FIBR enables high-speed control signaling and fast-switching among BSs during HOs, while allowing the user equipment to maintain a high degree of connectivity. We demonstrate that the FIBR architecture efficiently handles frequent HO events in mmWave cellular systems, and thus more effectively satisfies the QoS requirements of 5G applications.**

***Index Terms*— Ring, 5G, millimeter wave, latency, multi-connectivity, blockages, handover, URLLC, fast switching.**

I. Introduction

FIFTH Generation (5G) cellular networks are expected to serve a variety of new applications and services including eHealth, Augmented Reality (AR), Virtual Reality (VR), and the tactile Internet. The 3rd Generation Partnership Project (3GPP) categorizes them in three different classes of services, namely, massive Machine Type Communication (mMTC), enhanced Mobile BroadBand (eMBB), and Ultra-Reliable Low Latency Communication (URLLC) depending on the throughput, latency, and reliability requirements. A comprehensive set of requirements for these services and applications is presented in Table I. The high throughput requirement of eMBB services and the high traffic density required by URLLC applications [1] cannot be satisfied by the legacy sub-6 GHz band alone due to spectrum scarcity [2]. Thus, the 5G Next Generation Radio Access Network (NGRAN) will also use Millimeter Wave (mmWave) frequencies up to 52.6 GHz [3], where abundant bandwidth is available to support the demands of these applications and services [4].

While mmWave systems are capable of transmitting at speeds of multiple gigabits-per-second on the air interface, they are quite vulnerable to blockages and shadowing [5]. Even the human body can cause up to 35 dB attenuation in the signal strength [6]. Thus, mmWave links are inherently intermittent due to blockage and user mobility. Our work on the frequency of blockage events in [7], [8] suggests that a dense deployment of base stations (BSs) will be necessary to overcome blockages and satisfy the reliability requirements of URLLC applications in mmWave cellular networks. However, frequent handovers (HOs) $(0.1 - 1$ HO/sec) to maintain connectivity will be unavoidable. Our previous work [8] further suggests that in some conditions, UEs need to have either simultaneous connections with up to 12 BSs or an efficient HO mechanism to achieve the URLLC QoS requirements.

To satisfy the diverse requirements of these applications and to provide network flexibility and controllability, the 3GPP has proposed centralization of a few functions for the next Generation NodeB (gNB). The centralized part of a BS is called the gNB-CU and the decentralized part of a BS is called the gNB-DU. Furthermore, in the 3GPP transport architecture, a gNB is connected to the 5G Core Network (5G-CN) via the NG interface, gNBs are inter-connected using the Xn interface, and gNB-CU and gNB-DU are connected through the F1 logical interface (see Fig. 1(a)). The 3GPP further considers optimal placement of different functional blocks in the transport network to meet the diverse QoS requirements of applications and services. In particular, to satisfy the latency requirements of URLLC applications, the gNB-CU, the gNB-DU, and the mobile cloud can be pushed closer to the network edge. However, we argue that moving these closer to the network edge may not be practical [9], since during each HO, data and cloud computation needs to be forwarded. Moreover, as the 3GPP transport architecture is connection-oriented, i.e., a UE has to establish a connection to BSs before receiving/transmitting a packet, the frequent HOs due to blockages present a severe challenge to mmWave cellular systems [10]–[12]. Furthermore, as the mmWave channel is sporadic in nature, many HO procedures may result in Radio Link Failure (RLF) if the signal quality to the source or

Manuscript received April 30, 2019; revised September 9, 2019; accepted October 10, 2019. Date of publication October 23, 2019; date of current version November 27, 2019. This work was supported in part by the U.S. National Science Foundation under Grant 1527750, NYU Wireless, and in part by the NY State Center for Advanced Technology in Telecommunications (CATT). *(Corresponding author: Athanasios Koutsaftis.)*

The authors are with the Department of Electrical and Computer Engineering, Tandon School of Engineering, New York University, Brooklyn, NY 11210 USA (e-mail: tkoutsaftis@nyu.edu; rajeevkr@nyu.edu; peiliu@nyu.edu; panwar@nyu.edu).

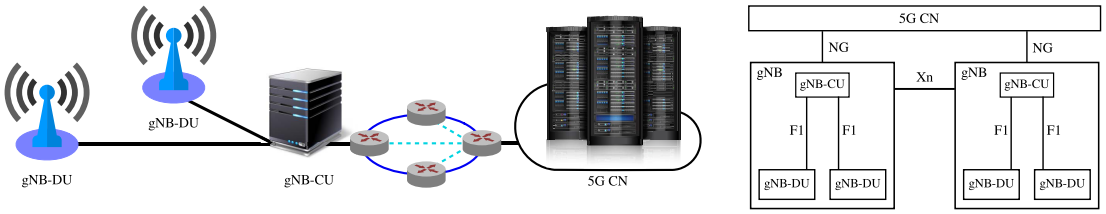
Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSAC.2019.2947940

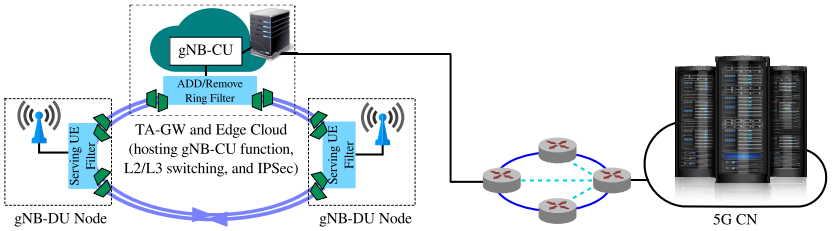
TABLE I

QoS REQUIREMENTS AND EXAMPLE APPLICATIONS FOR mMTC, URLLC AND eMBB SERVICES [14]–[16]

Services/Application Category	Throughput	Air-link Latency	Reliability	Example Applications
mMTC	1-100 Kbps	10 ms - 1 hr	90%	smart city, smart home
URLLC	1-10 Mbps	1 ms	99.9 % - 99.9999 %	eHealth, factory automation, robotics
eMBB	0.1 - 10 Gbps	4 ms	99.9 %	AR/VR, tactile Internet, 360 degrees video



(a) Physical and logical transport network architectures proposed by 3GPP for 5G



(b) Our proposed 5G transport network architecture

Fig. 1. Transport network architecture for 5G: In the 3GPP transport architecture gNB-DU, gNB-CU and the mobile cloud can be pushed to the network edge to satisfy ultra-low latency application requirements, but frequent HOs in mmWave systems will degrade network performance. To address this, we envision a transport network architecture, where the gNB-DU, the gNB-CU, and the mobile cloud are connected via a high-speed ring.

the target BS deteriorates during the HO procedure, due to blockage or UE mobility. For example, if a UE is moving at 120 km/s and the user plane service interruption time is 1 ms, 99.999% service reliability cannot be satisfied [13]. Thus, maintaining a high application QoS during frequent HOs and UE mobility is a major challenge for mmWave networks.

To alleviate the performance degradation of applications due to frequent HOs in 5G mmWave networks, we introduce a ring-based 5G transport network architecture, called the Fast Inter-Base Station Ring (FIBR) (see Fig. 1(b)). In FIBR, a number of BSs (gNB-DUs) in close proximity are grouped together to form a bidirectional buffer insertion ring network. Rather than being associated with a single BS, a UE in the FIBR architecture is associated with the *Target Area Gateway (TA-GW)*. To meet the QoS requirements of URLLC applications, the TA-GW hosts the gNB-CU, Layer 2/Layer 3 (L2/L3) switching, and the edge cloud. The TA-GW connects the user to the core network without regard to which BS on the ring the UE is served by. This provides FIBR with a framework for fast signaling among gNB entities, which helps in overcoming blockages and frequent HOs. Even when a UE has a low degree of connectivity, FIBR can provide reliability that would otherwise require a much higher degree of connectivity in the 3GPP transport network, thanks to the high speed signaling among gNB-DUs. The key contributions of this paper are as follows:

- We present a new transport network design for mmWave cellular systems, which connects a group of BSs in close proximity (target area) with high speed links to form a logical ring topology. In the proposed architecture, each individual UE gets associated with the TA-GW instead of a single BS, which will significantly reduce the HO signaling overhead due to frequent HOs in mmWave cellular networks.
- Next, we describe the proposed 3GPP architecture in detail and analyze the recent advances in 3GPP HO procedures. We particularly focus on single-connectivity and multi-connectivity HO schemes, and compute the associated control and data plane delays. Finally, we illustrate that FIBR can significantly reduce the HO latency for eMBB services and URLLC applications by enabling fast switching between BSs. Using the random waypoint mobility model, we showed that FIBR can achieve significantly lower blockage and RLF probabilities, as compared to the 3GPP transport network. Our results also show that FIBR can achieve high throughput and low user plane latency, and significantly smaller signaling overhead as compared to the 3GPP architecture. In essence, FIBR enables opportunistic utilization of intermittent mmWave links.

The rest of the paper is organized as follows. Related work is presented in Section II. A study of the new 3GPP transport architecture and our proposed FIBR transport architecture is presented in Section III. A detailed description of the 3GPP and FIBR HO procedures is highlighted in Section IV. Section V presents a performance comparison of the 3GPP and FIBR architectures using simulation results. Finally, Section VI concludes the paper.

II. Related Work

A. Background on Ring Architectures

Ring local area networks, such as token rings, attracted a lot of interest in the 1980's [17]. In token rings, a node is allowed to transmit only when it receives a free token. Then, the node removes the free token and replaces it with a busy one, indicating that the ring is currently occupied by the node. Major disadvantages such as fairness among nodes and node failures were addressed in [18] and [19].

In buffer insertion rings, when a packet arrives at a ring node, its destination address is examined and, if it is the current node, the packet is removed from the ring and placed in a reception buffer, otherwise, it is passed to the next node. The performance analysis of single-channel and multi-channel buffer insertion rings is presented in [20].

The *Resilient Packet Ring (RPR)* network was introduced in IEEE 802.17 [21]. It consists of two counter-rotating rings, which improves the reliability of the ring topology. While RPR allows packet-based access to the ring, the traffic scheduling policy is still flow-based with the aim of achieving a fair bandwidth sharing policy among all RPR stations.

All of these prior ring technologies were focused on exploiting the broadcast nature of rings, and the cost efficient shared access to high bandwidth for multiple stations that it offers. For FIBR, by contrast, the main consideration of a ring architecture is the ring's ultra-fast capability to accommodate UEs whose point of attachment to the network can change frequently, but yet the connections it carries cannot be interrupted or delayed in order to meet 5G's URLLC objectives. *We therefore claim that this is a unique application of ring technology*.

B. Background on HO Techniques and Multi-Connectivity

Although HOs in the legacy LTE heterogeneous networks are well studied [22]–[24], all these HO procedures are based on the break-before-make technique, i.e., the UE breaks the connection with its source BS before the HO procedure to its target BS has been initiated. For LTE networks, that results in an around $40 - 50$ ms user plane latency or service interruption time [13]. To reduce the service interruption time during the HO procedure, 3GPP has introduced *Make-Before-Break (MBB)* and *Random Access Channel (RACH)-less* techniques [25]. In MBB, the UE breaks the connection with its source BS only when the HO procedure is completed. In RACH-less HO, the UE skips the RACH procedure to the target BS. The *MBB combined with RACH-less* HO technique can reduce the service interruption time to 6 ms. Furthermore, the service interruption time can be further decreased to 0 ms if the synchronized RACH-less technique is used, where the target BS starts sending downlink data before receiving the HO complete message [13].

The aforementioned techniques have the potential to reduce the HO delay. However, the HO process will fail if the channel conditions for both the source BS and the target BS deteriorate, due to simultaneous and sudden blockages, which may occur frequently in mmWave networks [7]. Note that HO failure in 5G mmWave cellular networks may not only occur due to blockages but also as a result of UE mobility; UE mobility is the major cause of HO failures in the legacy LTE network [26]. We anticipate that HO failures due to UE mobility will further escalate. To meet the reliability requirement of URLLC applications, the HO failure rate must be kept significantly low.

To ameliorate the intermittent connectivity of mmWave systems, multi-connectivity has been considered by the 3GPP, industry, and the research community [27], [28]. In the context of multi-connectivity, two ideas have been put forward in the literature: (i) all BSs transmit the same signals to the UE, which helps in achieving a higher reliability at the cost of significant wastage of physical resources [28], and (ii) a single BS transmits the signal, while the UE maintains connectivity to multiple BSs [29]. The latter option may result in lower reliability as compared to the former but avoids wastage of resources [28]. However, a high reliability can be attained if a UE can switch to other BSs very fast [30]. In FIBR, we have fast control signaling among gNB-DUs, thus we chose the latter option with fast signaling among BSs and re-selection of gNB-DUs in case of blockages.

Polese *et al.* [31] have proposed a multi-RAT dual connectivity (DC) framework to perform fast switching between BSs. In this work, it is assumed that a UE is connected to a single LTE BS and a single mmWave BS. During a blockage the UE switches to the LTE BS after receiving a HO command, and once a new mmWave BS is found, the UE switches to the discovered mmWave BS. However, as the mmWave BSs can be frequently blocked [8] and many of the application flows cannot be offloaded to an LTE BS due to its limited bandwidth as compared to mmWave [32], the QoS of applications can degrade significantly. Hagos and Kapitza [32] have considered offloading traffic to a WiFi network during blockages to complement limited LTE resources. WiFi systems are designed to achieve high throughput but not consistently low latency [33]. Petrov *et al.* [29] have considered different multi-connectivity scenarios to study the impact of the degree of connectivity. A high order of multi-connectivity will result in a higher reliability, however, this also results in increased signaling and computation overhead [29].

In addition to having multi-connectivity between the UE and the BSs, we believe that there is a need for a paradigm shift from a connection-oriented transport network to a more opportunistic *connection-less* transport network. The wireless links will become more intermittent with both 5G mmWave and the THz bands being proposed for 6G. As each individual link becomes less reliable, it is essential for all UEs to harness macro-diversity from all nearby BSs. Current connection-oriented transport networks require all UEs to finish a HO procedure before granting access to the new source BS. In mmWave and THz systems, the connection time for each link before an HO is at least an order of magnitude shorter than sub-6 GHz systems. As a result, the HO procedure quickly becomes very expensive in terms of signaling overhead and HO delays for such systems. With FIBR, the data connection to each UE from the transport network is anchored at the TA-GW. Between the TA-GW and the BSs, user data is transmitted in a connection-less manner. Thus, UEs can roam freely between BSs on the same ring, since the signaling overhead due to an HO procedure is minimized. Access network level

switching is handled by the FIBR network. Such a design greatly simplifies the transport network architecture, and the HO is only necessary when a UE moves out of the TA. In such cases, the UE context information can be exchanged between respective TA-GWs.

FIBR is an architecture that aims to satisfy the QoS requirements of URLLC and eMBB applications by enabling fast switching between BSs. To demonstrate this, we first present both the FIBR and 3GPP transport network architectures in Section III, and then discuss the HO procedures for the two architectures in Section IV. In Section V, we present numerical results based on simulations and compare the performance of the two architectures.

III. 3GPP and FIBR Architectures for 5G Cellular Systems

In this section, we describe the 3GPP transport network architecture and our proposed FIBR architecture.

A. 3GPP Transport Network Architecture

To satisfy a wide range of applications with diverse requirements for 5G cellular systems and to provide flexibility and efficiency while reducing the network cost, the 3GPP has proposed centralization of a few functions of the gNB. The selection of a functional split will dictate the transport network capacity and latency requirements as well as the placement of nodes in the network [34]. One possible design choice to meet the QoS requirements of URLLC applications is the functional split between Packet Data Convergence Protocol (PDCP) and Radio Link Control (RLC), where unlike the LTE eNB, PDCP and Radio Resource Control (RRC) constitute the centralized unit of gNB, while RLC and lower layers constitute the decentralized unit of gNB, defined as gNB-CU and gNB-DU respectively.

In Fig. 1(a) the 3GPP proposed functional split and the 5G transport network are shown. A gNB is connected to the 5G-CN via the NG interface, and gNBs are inter-connected via the Xn interface. For any gNB, the gNB-CU and the gNB-DU can be separated geographically [35]. The gNB-CU processes non-real time protocols and services, while the gNB-DU may process physical, medium access control, and RLC layer protocols and real-time services. The gNB-CU and the gNB-DU are connected through the F1 logical interface, which has uplink and downlink capacity requirements of 3 Gbits/s and 4 Gbits/s respectively. A gNB-DU can be connected to a single gNB-CU, while a gNB-CU can be connected to multiple gNB-DUs. This provides a framework for dual-connectivity or multi-connectivity [36], [37]. In the multi-connectivity setting, 3GPP only considers multi-Radio Access Technology (RAT) Dual Connectivity (DC) [27]. In case a UE can support multi-RAT DC, it will opt to utilize resources from two different BSs [27]. Thus, in addition to providing network flexibility and controllability, the new 3GPP transport architecture also provides a framework for both single and dual connectivity to achieve higher QoS requirements. However, as discussed earlier, even with the significant changes of the 3GPP transport architecture, due to the intermittent nature of mmWave links, meeting the QoS requirements of different applications, and in particular URLLC, is quite challenging. Therefore, we next discuss our proposed FIBR architecture and demonstrate its ability to satisfy the QoS requirements of those applications.

B. FIBR Transport Network Architecture

FIBR is a bidirectional buffer-insertion ring architecture, where a number of gNB-DUs in close proximity are grouped together with a gNB-CU and the mobile cloud (see Fig. 1(b)). The capacity of the ring is kept significantly higher than the throughput requirements of applications served by the gNB-CU. The coverage area of the FIBR ring is called *Target Area (TA)*. In FIBR, a UE is not associated with a single gNB-DU and/or gNB-CU. Instead, it is associated with the TA-GW, which is connected to the next-generation 5G-CN. Note that the TA-GW can host a gNB-CU, L2/L3 switching functions, IPSec, and the edge cloud. In FIBR, as the gNB-CU and gNB-DUs are connected through the ring and packets are not addressed to any particular gNB-DU, *the connectivity between the gNB-CU and gNB-DUs is connection-less*. Thus, it is unnecessary to make and tear down the connection between the gNB-DU and gNB-CU at every HO event.

When a UE enters the TA, it conducts the cell search procedure to find the gNB-DUs in its communication range. In a K-connectivity model, after the UE discovers all the available gNB-DUs in its coverage range, it selects the K best available gNB-DUs based on the Received Signal Strength Indicator (RSSI) values. Once a UE selects the K best gNB-DUs, it requests these gNB-DUs to serve as its proxies on the ring. The gNB-DUs accept the request and add the UE ID to their Address Filter Database (AFD), which contains all UEs served by them. Following the cell search and selection, the UE starts the RACH procedure with these K gNB-DUs. Furthermore, for the selection of transmitting gNB-DUs, the selected K gNB-DUs may use coordinated scheduling and beamforming in both uplink and downlink directions [38]. Note that FIBR is a Layer 2 scheme to achieve fast HO. Therefore, any discussion of physical layer techniques is not within the scope of this paper, but can be addressed in further work if it impacts the link layer. FIBR is capable of fast switching among BSs, thus it encourages air interfaces to utilize ephemeral and less reliable links, without decreasing the overall service reliability.

Note that there can be two scenarios of blockages: 1) if the primary serving gNB-DU gets blocked, the UE switches to one of the secondary gNB-DUs, and 2) if a secondary gNB-DU gets blocked, the UE finds a new secondary gNB-DU. The UE and the gNB-DUs maintain a periodic (e.g., 20ms) heartbeat signal to check connectivity.

Next, we discuss different aspects of the FIBR architecture, such as connectivity schemes, packet processing, and ring protection schemes.

1) Single-Connectivity: In this scenario, unless and until the channel quality between the gNB-DU and the UE degrades, the UE will be served by the same gNB-DU. Thus, at any given time, only one gNB-DU will have the UE address in its AFD. As soon as a downlink packet arrives at a gNB-DU, the gNB-DU performs a lookup action at its AFD, to check

whether it serves the UE that the packet is destined to. If there is a match, we considered two different architecture options:

(a) the gNB-DU *copies* the packet into its downlink buffer. If no other packets are being served at that instant, the source gNB-DU frames the data and transmits it over the air interface. Otherwise, the packet is kept in the gNB-DU downlink bearer buffer until there is a transmission opportunity. The packet will circulate the entire ring and return to the TA-GW, which will then remove it from the ring. Once the gNB-DU receives an acknowledgement of the transmitted packets, it circulates the acknowledgement in the ring. Upon the reception of an acknowledgement, the gNB-CU at the TA-GW removes the associated packets from its PDCP buffer. If the packet is not acknowledged after 4 slots (we consider a slot duration of 125 μs), the TA-GW will put the packet into the ring again. If it hasn't received the packet after another 4 slots, the TA-GW surmises that an RLF occurred for the UE. In case of an RLF, the serving gNB-DU deletes the UE ID from its AFD. During RLF, downlink packets cannot be transmitted over the air, thus they will travel the ring once and will finally be removed by the TA-GW. When the UE establishes a connection with a new gNB-DU, the downlink packets can be forwarded again by the TA-GW. The packets will be removed from the PDCP buffer of the gNB-CU either after being acknowledged, or upon the expiration of a timer.

(b) the gNB-DU *removes* the packet from the ring and puts it into its downlink buffer. The packet is kept in the gNB-DU downlink buffer until there is a transmission opportunity. Once the gNB-DU receives an acknowledgement for the transmitted packet, it circulates the acknowledgement in the ring and the TA-GW can then remove the packet from its PDCP buffer. During RLF, the packets are no longer removed by the gNB-DU, and circulate the whole ring until they reach at the TA-GW. In this architecture option, if the TA-GW observes the same packets returning, it concludes that the UE is not connected to any gNB-DU. Thus, the TA-GW will not retransmit the packets unless and until the UE is connected with a new gNB-DU. The packets will be removed from the PDCP buffer of the gNB-CU either after being acknowledged, or upon the expiration of a timer.

In the uplink, traffic is transmitted over the air to the gNB-DU and waits in the *gNB-DU uplink bearer buffer* before entering the ring. Once the uplink packets are injected into the ring, they will be forwarded to the 5G-CN by the TA-GW.

2) Multi-Connectivity: Recall that in FIBR, where we have fast control signaling among gNB-DUs, if a UE is capable of multi-connectivity, it maintains connectivity with multiple gNB-DUs but only a single gNB-DU transmits the data. The connectivity with the other gNB-DUs is maintained using heartbeat signals with configured periodicity. In the multi-connectivity setting, multiple gNB-DUs can have the UE address in their AFD. The transmitting gNB-DU is initially selected as the one with the highest RSSI value. Note that

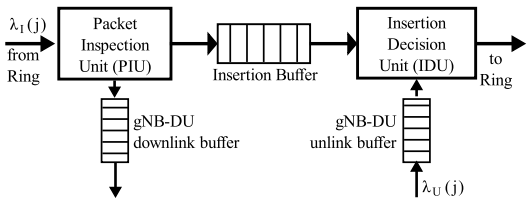


Fig. 2. FIBR gNB-DU includes functional blocks for uplink and downlink packet processing. In the downlink ring during normal operations, IDU is disabled. Similarly, in the uplink ring during normal operations, PIU is disabled. After a failure, uplink and downlink traffic are routed to the a single ring with both functional blocks enabled.

FIBR encourages the use of ephemeral links thanks to its fast switching ability. Again, we consider two architecture design options. In the *first* option, all these gNB-DUs will copy the downlink packets in their downlink bearer buffer. After a gNB-DU circulates the acknowledgement of a transmitted packet, all other gNB-DUs will flush this packet from their buffer upon reception of the acknowledgement. In this setting, gNB-DUs collectively try to achieve macro-diversity for high reliability. When the UE switches to one of the other gNB-DUs that it is connected to, the new serving gNB-DU can start transmitting the downlink packets immediately, as they were already copied in its downlink buffer. In the *second* option, the serving gNB-DU removes the packet from the ring and the other gNB-DUs do not copy the packet in their downlink buffer. In case of the UE switching serving gNB-DUs, the new serving gNB-DU will forward a control signal into the ring indicating that it now serves the UE and the TA-GW can transmit into the ring the packets that have not yet been acknowledged.

3) Packet Processing in FIBR: Downlink and uplink packet processing is conducted in the ring as follows:

- **Downlink packet processing:** At every ring node, the Packet Inspection Unit, or PIU (see Fig. 2) examines the header of every incoming packet and copies the packet to the gNB-DU downlink buffer if 1) the destination UE is found in the gNB-DU AFD, and 2) there is enough space in the gNB-DU downlink buffer. If both of these conditions are not satisfied, the packet will not be copied in the gNB-DU downlink buffer. Note that the gNB-DUs maintain separate buffers for each UE and service class depending upon the 5G Quality Indicator (5GQI) [39].
- **Uplink packet processing:** When an uplink packet is received at a gNB-DU node, it is initially stored in the gNB-DU uplink buffer. During normal operation, packets are then extracted from the gNB-DU uplink bearer buffer and put in the uplink ring. However, in the case of a single ring failure, the Insertion Decision Unit (IDU) decides whether packets from the ring (insertion buffer) or the gNB-DU uplink buffer should be prioritized. The IDU implements policies such that both uplink and downlink QoS can be satisfied in the case of a failure.

4) Ring Protection: To ensure the reliability of the ring in FIBR, we consider $1 + 1$ ring protection. In other words, in the case that one ring fails, both the uplink and downlink packets will share a single ring. During normal ring operation, i.e., when there is no failure on either ring direction, uplink

and downlink packet flows will be transmitted on separate rings. Note that the point-of-presence on the ring is only at the TA-GW, i.e., every downlink packet originates from the TA-GW and every uplink packet terminates at the TA-GW in the ring. We next consider two failure scenarios, ring node failure and fiber cut failure. After a ring node failure, the network operator will perform a wrap on the nodes adjacent to the failed one, and both uplink and downlink traffic will share the same ring. After a fiber cut failure, the network operator will perform a wrap on the two nodes adjacent to the fiber cut. In both failure scenarios, downlink and uplink traffic will eventually share the same directional ring. Thus, to handle a failure, every gNB-DU node in the ring is equipped with all the necessary functional blocks to process both uplink and downlink packets on the surviving ring. Each gNB-DU node in the ring consists of a packet inspection unit, an insertion buffer, and an insertion decision unit for both uplink and downlink rings (see Fig 2). During normal operations in the downlink ring, there is no uplink traffic, hence the IDU remains disabled. Similarly, in the uplink ring, there are no downlink packets, thus the PIU remains disabled. A queueing analysis to compute the downlink and uplink packet latency considering $1 + 1$ protection is presented in Appendix B. The analysis presents an overview on the number of gNB-DUs that can be satisfied based on the ring capacity, the QoS requirements of different applications and services, and the $1+1$ ring protection scheme.

5) Complexity of the FIBR Transport Architecture: The main source of complexity associated with the FIBR architecture is related to the processing in the TA-GW and the remaining ring nodes. Recall that the TA-GW can host a gNB-CU, thus its complexity is comparable with a gNB-CU that hosts PDCP and the layers above it [35]. As far as the other ring nodes are concerned, their complexity is associated with the hardware processing capability. This hardware complexity is comparable to the complexity of nodes in ring architectures built in the past [40].

IV. HO Procedures in 3GPP and FIBR

In this section, we first present the recent advancements in the 3GPP HO procedure and then we discuss the HO procedure in our proposed FIBR architecture.

A. 3GPP HO Procedures

3GPP has discussed different HO procedures for both single-connectivity [41] and multi-connectivity [27] settings. Note that in the multi-connectivity setting, 3GPP only considers multi-RAT DC. However, there are two major problems with multi-RAT DC. Firstly, eMBB services and some of the URLLC applications put a load on the network high enough so that a single LTE eNB cannot satisfy it [32]. Thus, data plane traffic needs to be offloaded to multiple eNBs. Secondly, due to the intermittent connectivity of the mmWave channel, connectivity to only one extra gNB-DU cannot fulfill the reliability requirement [8].

There are two types of HO procedures: (i) intra-gNB-CU HO, where UE traffic can be offloaded to a different gNB-DU, but it remains connected to the same gNB-CU, and (ii) inter-gNB-CU HO, where UE traffic is offloaded to a completely different gNB using the Xn interface or 5G core entities. Within the framework of this paper and FIBR architecture, only intra-gNB-CU HOs, which will be far more frequent, will be compared with the proposed 3GPP intra-gNB-CU HO techniques. We believe that in FIBR, inter-TA HOs will have similar complexity and performance as inter-gNB-CU HOs in the 3GPP transport architecture. Therefore, we will only discuss intra-gNB-CU HO procedures.

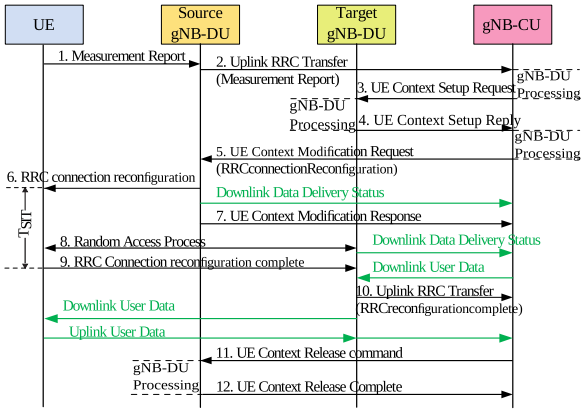


Fig. 3. The 3GPP intra-gNB-CU HO procedures [41].

1) Single Connectivity HO Procedures: As the mmWave systems are prone to blockages and the cell sizes of gNB-DUs are smaller, intra-gNB-CU HOs will be more frequent as compared to the intra-LTE HOs (source and target cells belong to the same LTE network) in legacy LTE cellular systems [8], [42]. The 3GPP intra-gNB-CU HO procedures and mobility [41] management are presented in Fig. 3. The UE sends periodic measurement reports to the source gNB-DU, which are forwarded to the gNB-CU for HO decision. If the criteria of HO procedures are met, e.g., the Received Signal Strength Indicator (RSSI) value falls below the designated threshold, the gNB-CU initiates a HO procedure by sending the UE context setup information to the target gNB-DU. After receiving the response from the target gNB-DU, the gNB-CU sends the UE context modification request, including an RRC connection reconfiguration request, to the source gNB-DU, which is ultimately sent to the UE. Following this, the source gNB-DU sends a downlink delivery status to the gNB-CU to indicate any unsuccessfully transmitted downlink data. Note that the PDCP layer in the gNB-CU keeps a copy of all packets until it receives a delivery status acknowledging successful transmission. Therefore, it is unnecessary to forward unacknowledged packets from the source gNB-DU to the target gNB-DU; the PDCP layer in the gNB-CU takes care of this. The UE follows a RACH procedure to establish a connection with the target gNB-DU. After the completion of the RRC connection reconfiguration procedures, the UE notifies the target gNB-DU. Data plane communication between the target gNB-DU and the UE can be initiated at this point and the UE context is then released from the source gNB-DU.

Using the control signaling and processing at the UE, the source gNB-DU, the target gNB-DU and the gNB-CU,

we compute the control plane latency $T_{\text{CP}}^{\text{SgNBCU}}$ as:

$$T_{\text{CP}}^{\text{SgNBCU}} = T_{\text{gNBDU-UE}} + 6T_{\text{gNBDU-gNBCU}} + T_{\text{SIT}} + 4T_{\text{PgNB}}, \quad (1)$$

where T_{SIT} is the service or HO interruption time, $T_{\text{A-B}}$ is the propagation delay between nodes A and B, and T_{PgNB} is the processing delay at the gNB. For the calculation of control plane latency, signaling and processing from steps $2-12$ are considered, excluding steps 7-9 since time associated in steps $7-9$ is considered in the service interruption time. In steps $7-9$, the user plane can also be interrupted. In the legacy HO procedures (break-before-make), as soon as the UE receives an RRC reconfiguration message, it discontinues the data plane service. Although 3GPP has introduced MBB and RACH-less procedure for reducing data plane latency, the UE still has to follow control plane procedures in HO events. Thus, in scenarios where the source gNB-DU can be suddenly blocked, the data plane latency will be at least as much as the control plane latency. Furthermore, note that due to the intermittent connectivity of mmWave links, RLFs may happen quite frequently (RLF generally happens due to HO procedures and mobility [26]). In the RLF case, the UE needs to start RLF recovery procedures by either initiating an RRC connection reestablishment procedure (if it can connect to the previous serving BS) or cell search and RRC connection procedures. Note that both of these procedures may induce significantly higher control and user plane latency. In general, RLF is declared after the expiration of T310 and N310 timers, which corresponds to a latency of 30 ms [43].

2) Multi-RAT DC HO Procedures: Many different multi-connectivity scenarios have been discussed in [27], [41], where the HO is handled by the LTE evolved packet core with the LTE Mobility Management Entity (MME) as the anchor point. The HO procedure using the 5G-CN is still under discussion in the 3GPP standard (Release 15). As of the current release of the standard, if a gNB-DU gets blocked, the connection to this gNB-DU is released and the LTE eNB starts serving the UE. Once a new gNB-DU is found, the connection to this gNB-DU is initiated (see Fig. 4). When a gNB-DU gets blocked, the 5G cellular systems may have to temporarily (until a new gNB-DU is found) stop services to high throughput applications due to limited resources.

From Fig. 4, we can see that in the case of dual connectivity under Release 15 of the 3GPP Standard, control plane functions are carried out by the LTE eNB (which makes eNB the master node and gNB the secondary node), thus increasing the connection reliability. However, this also limits the number of applications that can be offloaded to LTE eNB.

B. HO Procedures in FIBR

Recall that in FIBR, the connectivity between the gNB-CU and gNB-DUs is connection-less, while the connectivity between gNB-DUs and UEs is connection-oriented. In Section III, we discussed both single connectivity and multi-connectivity in FIBR depending upon the UE capabilities to support them. Here, we continue our discussion of single connectivity and multi-connectivity in the context of HO for the FIBR architecture.

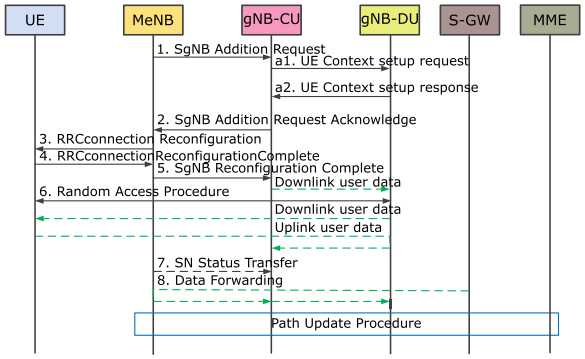


Fig. 4. 3GPP NR-DC HO: Only the master eNB (MeNB) maintains the control plane connection with the core network, thus, when the gNB-DU gets blocked, the connection to the gNB-DU is dropped and the LTE eNB starts serving the UE. Once the new gNB-DU is found, the connection to this gNB-DU is initiated [41].

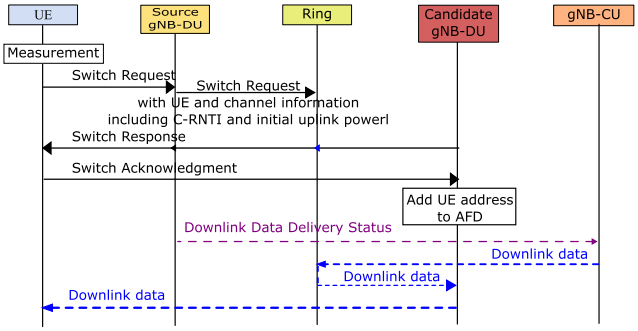


Fig. 5. Single connectivity HO processes in FIBR.

1) Single Connectivity HO Procedures in FIBR: In the FIBR architecture, we focus on user-centric networking to minimize HO latency. Based upon a measurement, a UE can send a switch request to its current serving gNB-DU. The serving gNB-DU sends this switch request onto the ring which includes all of the RRC and physical layer configuration parameters, and UE information. Note that all of the gNB-DUs, and gNB-CUs are synchronized, thus there is no timing difference between gNB-DUs, and synchronization is not needed. Based upon measurements, the UE can indicate which gNB-DUs are the best candidate BSs. Upon the reception of the switch request, the candidate gNB-DUs check whether they can provide services to the UE using the previous RRC and physical layer configuration parameters. If they can, they send a switch response to the UE. The first gNB-DU to reach the UE via the switch response is selected as the serving gNB-DU. Then, the UE replies to the gNB-DU with a switch acknowledgement. Upon the reception of the switch acknowledgment, the gNB-DU adds the UE address to its AFD. After adding the UE address to its AFD, the gNB-DU starts copying the UE downlink data from the ring.

In the case of sudden blockages, however, the UE may still need to follow similar procedures as in 3GPP transport architecture for RLF recovery. This can take a significantly long time. We therefore consider multi-connectivity in FIBR to achieve higher reliability.

2) Multi-Connectivity and HO Procedures in FIBR: In the FIBR architecture, a UE simultaneously maintains connections

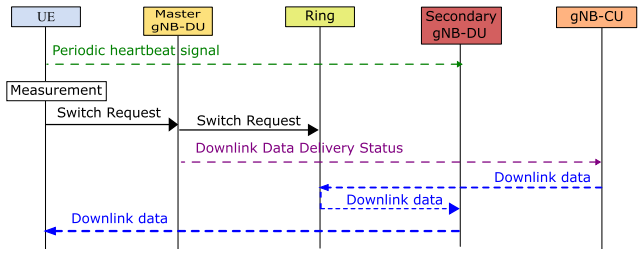


Fig. 6. Multi-connectivity HO processes in FIBR.

to multiple gNB-DUs for data transmission. However, as discussed earlier, although connectivity is maintained with multiple gNB-DUs, only a single gNB-DU sends traffic to the UE at any time. In the case of blockages, the UE switches to a secondary gNB-DU for services. The UE sends periodic heartbeat signals to the all other gNB-DUs to check connectivity. If a secondary gNB-DU gets blocked, it finds a replacement for the blocked secondary gNB-DU. The HO procedure for FIBR in the multi-connectivity case is presented in Fig. 6. In the case of a secondary gNB-DU blockage, the HO procedure is similar to Fig. 5. Note that an RLF can still happen if all gNB-DUs having connectivity to the UE get blocked. However, this probability decreases significantly due to fast control signaling and multi-connectivity in FIBR. In the case of an RLF, the UE will need to perform similar recovery processes to those discussed in the 3GPP standard.

The purpose of multi-connectivity in FIBR is to achieve high reliability instead of obtaining high throughput like in legacy LTE networks [44]. The FIBR architecture provides a framework that removes the need for setting up and tearing down connections after blockage events. Thus, even if multiple BSs suffer simultaneous blockages, FIBR can provide an alternative data path to transmit packets in uplink and/or downlink as long as at least one BS remains unblocked. This removes the control and data plane latency associated with RRC reconfiguration procedures. Occasionally, if a UE suffers blockages from all of its connecting BSs, an RLF will take place and the UE will start an RLF recovery process. However, since FIBR uses multi-connectivity to alleviate the need for frequent HOs due to blockages, the number of RLF events will be significantly reduced. Therefore, the RLF probability in FIBR is close to the simultaneous blockage probability for all connecting gNB-DUs.

V. Numerical Results

In this section, we compare the FIBR transport architecture with the 3GPP transport architecture using MATLAB simulations. For the comparison of the two architectures, we consider blockage and RLF probabilities, throughput, and data plane latency. In the simulation, the UE is considered stationary at the origin and blockers are uniformly distributed at a radius of 100 m around the UE. For the blocker mobility, we use the random waypoint mobility model [45], [46]. We also compare the theoretical upper and lower bounds on blockage and RLF probabilities with the corresponding numerical results obtained via simulation.

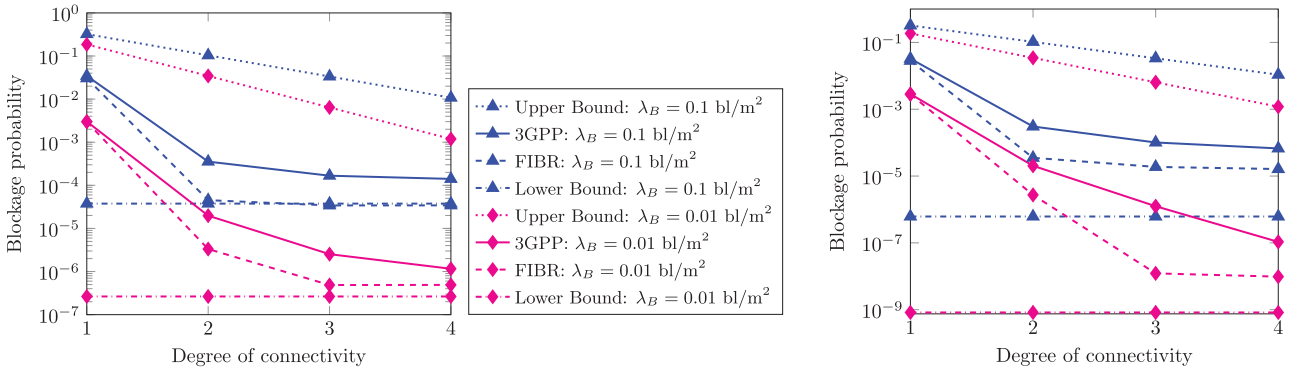
TABLE II
Simulation Parameters

Parameters	Values
LOS Radius, R	100 m
Velocity of Dynamic Blockers, V	1 m/s
Height of Dynamic Blockers, h_B	1.8 m
Height of UE, h_R	1.4 m
Height of gNB-DU, h_R	5 m
Expected blockage duration, $1/\mu$	0.5 s
Self-blockage angle, ω	60°

For the simulation, we consider a square of size 200 m $\times$ 200 m with blockers located uniformly in this area. Our area of interest is a disc of radius $R = 100$ m, which perfectly fits in the considered square area. The blockers choose a direction randomly, and move in that direction for a time-duration chosen uniformly in $[0, 60]$ seconds. For the simulation, we performed 5,000 runs and each run consisted of the equivalent of 4 hours of blockers mobility. To maintain a fixed density of blockers in the square region, we consider that once a blocker reaches the edge of the square, it gets reflected. Note that the blockage duration is exponentially distributed with parameter $1/\mu = 0.5$ seconds. We consider two values (9 and 12) as the number of gNB-DUs in the UE coverage area, which are uniformly distributed in a disc of radius $R = 100$ m. Furthermore, we consider four values (1,2,3, and 4) for the degree of connectivity. In both 3GPP and FIBR architectures, once blockage of the serving/master gNB-DU and the secondary gNB-DUs are detected, the UE performs an HO to other available unblocked gNB-DUs using MBB and synchronized RACH-less HO techniques [47]. Thus, the HO latency is considered 0 ms as long as the UE can be served by at least one BS. An HO latency will be introduced in two scenarios: 1) the UE is completely blocked from all of the BSs and 2) a new gNB-DU is not found during blockages. In such scenarios, the UE needs to start an RLF recovery procedure. Note that in the 3GPP transport architecture, blockages of secondary gNB-DUs can only be detected following periodic measurements while blockages of the serving gNB-DU blockage can be detected soon after it takes place. However, in FIBR, both serving/master and secondary gNB-DUs can be detected rapidly thanks to the fast control signaling and the periodic heartbeat monitoring (with a short period) of secondary gNB-DUs. The rest of the simulation parameters are presented in Table II.

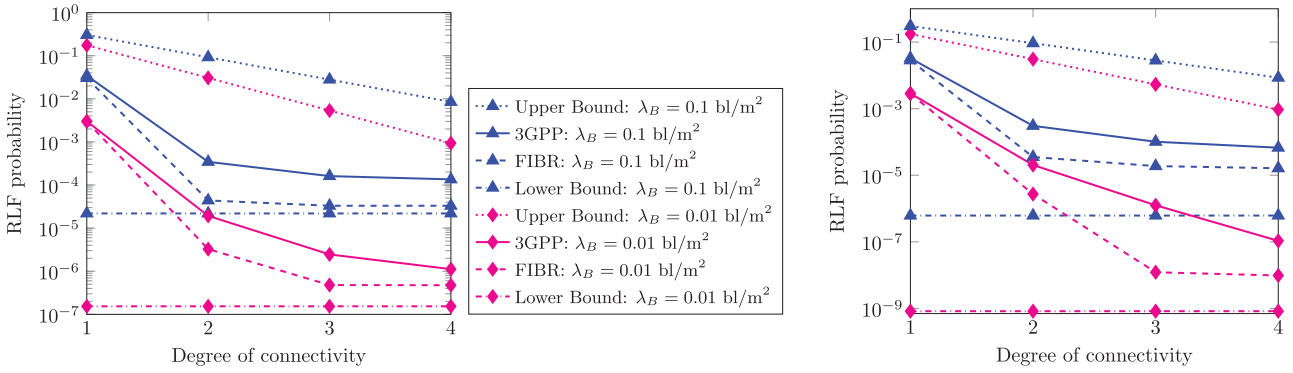
A. Blockage and RLF Probabilities

Fig. 7 and Fig. 8 plot the blockage and RLF probabilities with different degrees of connectivity and numbers of gNB-DUs in the UE coverage region together with the corresponding theoretical lower and upper bounds. As discussed in Appendix A, the lower bound on the blockage and RLF probabilities will be obtained if the UE can switch to any gNB-DU in its coverage region without any HO latency, i.e. the UE switches to an unblocked gNB-DU instantly during a blockage event. Thus, the UE has multi-connectivity effectively to all



(a) 9 gNB-DUs in UE coverage. (b) 12 gNB-DUs in UE coverage.

Fig. 7. Blockage probability: a comparison of the FIBR and 3GPP transport architectures with different numbers of gNB-DUs in the UE coverage area, blockage density values and degrees of multi-connectivity. Note that the blockage probabilities derived by simulation lie between the lower and upper theoretical bounds. The theoretical lower bound is obtained when the UE can switch to any gNB-DU instantly during a blockage event. The theoretical upper bound is obtained in a K-connectivity setting when there are only K gNB-DUs in the coverage region, i.e., the UE cannot update its K serving gNB-DUs even if they get blocked and there are unblocked gNB-DUs in UE coverage region.

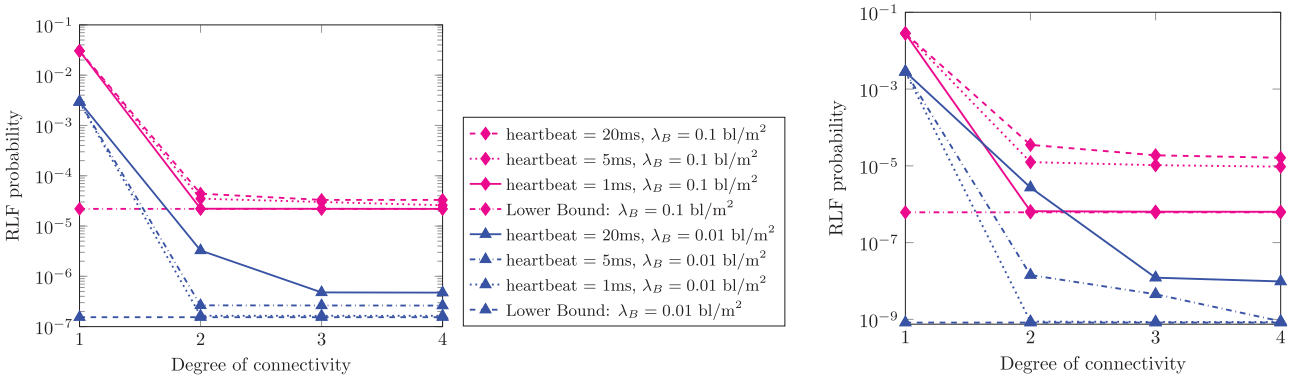


(a) 9 gNB-DUs in UE coverage. (b) 12 gNB-DUs in UE coverage.

Fig. 8. RLF probability: a comparison of the FIBR and 3GPP transport architectures with different numbers of gNB-DUs in the UE coverage area, blockage density values and degrees of multi-connectivity. Note that the RLF probabilities derived by simulation lie between the lower and upper theoretical bounds. The theoretical lower bound is obtained when the UE can switch to any gNB-DU instantly during a blockage event. The theoretical upper bound is obtained in a K-connectivity setting when there are only K gNB-DUs in the coverage region, i.e., the UE cannot update its K serving gNB-DUs even if they get blocked and there are unblocked gNB-DUs in UE coverage region.

of the gNB-DUs in its coverage region. The theoretical upper bound in the K-connectivity setting can be obtained if the UE has only K gNB-DUs in its coverage region; if K gNB-DUs get blocked, the UE cannot find another available gNB-DU. From Fig. 7(a) and Fig. 7(b), we can observe that when the UE cannot switch to a gNB-DU instantly and has a higher number of gNB-DUs in its coverage region compared to K (where K is degree of connectivity), the blockage probability lies between the two bounds. Similar observations are obtained from Fig. 8(a) and Fig. 8(b) for the RLF probability.

From Fig. 7, we can observe that the blockage probability decreases as the degree of connectivity increases. The highest improvement in the blockage probability can be observed when the degree of connectivity is increased from 1 to 2. In the single connectivity case, once the UE is blocked, it is not able to search for another gNB-DU. In that case, RLF will only be avoided if the blockage duration is shorter than the T310 and N310 timers (30 ms). However, in the dual connectivity case, if one of the gNB-DUs gets blocked, the UE will be able to search for a new gNB-DU to replace the blocked one using its active connection and as a result the reliability is greatly increased. A higher degree of connectivity provides the UE with a higher degree of freedom to find unblocked gNB-DUs in its coverage region. However, if all gNB-DUs in the UE coverage area are blocked, having a higher degree of connectivity will not help. That is why a degree of connectivity higher than two results in decreasing improvement in blockage probability. Comparing Fig. 7(a) and Fig. 7(b), we can also observe that if there is a significant number of gNB-DUs in the UE coverage region and a reasonable degree of connectivity is available, the QoS requirements URLLC applications can be met. In a scenario with a blocker density of 0.01 bl/m^2, the 3GPP architecture was able to achieve 99.9999% reliability with degree of connectivity equal to 4, when at least 9 gNB-DUs are in the UE coverage region. In FIBR, by contrast, a degree of connectivity of only 3 is required, which is a significant improvement given the additional overhead that a higher degree of connectivity imposes in the 3GPP architecture [29]. Furthermore, we can observe that FIBR achieves significantly lower blockage probability as compared to the 3GPP transport architecture due to its ability to perform fast signaling. Thus, it reduces the need for a very dense deployment of gNB-DUs,



(a) 9 gNB-DUs in UE coverage.

(b) 12 gNB-DUs in UE coverage.

Fig. 9. RLF probability in FIBR for different heartbeat signal periodicities: the RLF probability decreases with faster heartbeat signals. In the case of a periodicity of 1 ms with multi-connectivity, it converges to the theoretical lower bound.

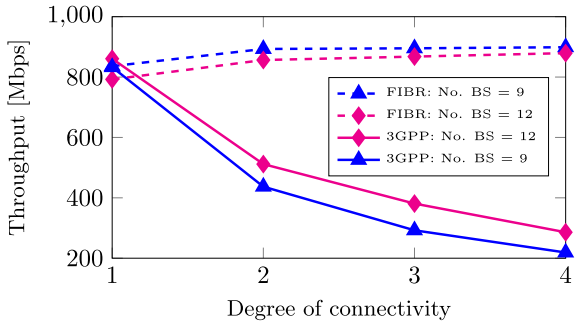


Fig. 10. Throughput: a comparison of the 3GPP and FIBR architectures with different number of gNB-DUs and degrees of connectivity. A dynamic blocker density of $0.1\text{bl}/\text{m}^2$ is considered.

particularly in densely populated areas with high blocker densities. This could lead to significant cost savings. Similar observations are obtained from Fig. 8 for the RLF probability. Note that the blockage and RLF probabilities do not vary from each-other significantly, since an RLF will occur with high probability if a UE is blocked from all its serving gNB-DUs (the RLF timer of 30 ms is relatively small compared to the average blockage duration of 500 ms).

To further investigate the effect of a faster heartbeat signal, we conducted experiments with heartbeat signal periodicities of 5ms and 1 ms. As shown in Fig. 9, the RLF probability decreases significantly with a reduction in the heartbeat signal period. For a heartbeat signal periodicity of 1 ms, we observe that the RLF probability converges to the theoretical lower bound for both blockage densities and number of gNB-DUs in the UE coverage area, for cases when the UE can support at least dual-connectivity. However, more frequent heartbeat signaling induces significant computational overhead in the UE and increased bandwidth utilization, thus the trade-off between the desired reliability and resources allocated to heartbeat signaling needs to be carefully studied.

B. Throughput

For the computation of throughput, we considered an ON-OFF process, where during the blockages (when all serving/master and secondary gNB-DUs are blocked) throughput is 0 Mbps and in the unblocked duration throughput is obtained using an empirical path loss model [48], [49]. Note that in the 3GPP transport architecture, the need for a handover to a new gNB-DU can be detected only through periodic measurement with a periodicity of 200 ms [50]. Thus, to achieve high reliability, repetition coding must be used in the 3GPP transport architecture. However, this will result in significant wastage of radio resources. In FIBR, by contrast, due to the fast control signaling among the BSs and heartbeat signaling (we assume a periodicity of 20 ms here) between the UE and the secondary/non-serving BSs, the blocked gNB-DU can be replaced with a new gNB-DU in a timely manner. This helps us achieve high reliability in the FIBR architecture without having to broadcast URLLC traffic over multiple gNB-DUs. We therefore improve the spectral efficiency and achieve a significantly higher throughput as compared to the 3GPP transport architecture (see Fig. 10). Note that in the 3GPP transport architecture, we always select the gNB-DU with highest signal-to-noise ratio to the UE, whereas in FIBR, the gNB-DU is selected randomly. Thus, in the single connectivity scenario, the 3GPP transport architecture achieves slightly higher throughput. Furthermore, in the FIBR transport architecture with dense deployment of gNB-DUs (to achieve a high reliability) and random selection of gNB-DUs to avoid blockages, the obtained throughput may degrade slightly, as the selected gNB-DUs may be far from the UE. For example, in the FIBR architecture, we achieve higher throughput when there are 9 gNB-DUs in the UE coverage region (see Fig. 10). However, note that a higher number of gNB-DUs in the UE coverage region results in higher reliability that may be a key QoS metric for many URLLC applications (see Fig. 7).

C. Data Plane Latency

As discussed earlier, in both 3GPP and FIBR, MBB and synchronized RACH-less HO process are considered. Thus, the data plane latency during these HOs remains zero. However, the data plane connection can be interrupted in the following two scenarios:

- *If all gNB-DUs connected to the UE get blocked*: RLF will be declared and an RLF recovery process will be

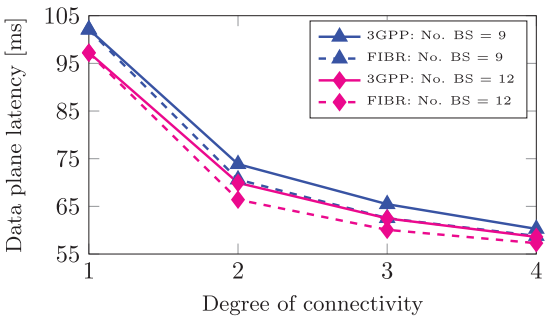


Fig. 11. Data plane latency: a comparison of the 3GPP and FIBR architectures with different number of gNB-DUs and degrees of connectivity. A dynamic blocker density of $0.1\text{bl}/\text{m}^2$ is considered. This delay should *not* associated with the handover delay, but it occurs due to outage, i.e., when the UE is blocked from all its serving gNB-DUs and the data plane is interrupted.

initiated. If the RLF recovery process succeeds, the data plane services can be re-established.

- *If all gNB-DUs in the UE coverage region get blocked*: in the event of simultaneous blockage of all connected gNB-DUs, the UE will initiate an RLF recovery process. If all other gNB-DUs in the UE coverage region are also blocked, then the RLF recovery process may fail resulting in a long data plane interruption unless and until one of the gNB-DU in UE coverage region gets unblocked.

For the computation of the data plane latency, we considered the above two blockage scenarios in our simulation. From Fig. 11, we observe that the data plane latency decreases significantly with the degree of connectivity and the number of gNB-DUs in the UE coverage region. Furthermore, the FIBR architecture only modestly outperforms the 3GPP architecture. Two important points must be noted, however. First, that the data plane latency is actually dominated by the outage duration and *not* the HO duration. i.e., even if there is a scheme with 0 ms HO delay, the expected data plane latency will remain higher than 55 ms and 42 ms when there are 9 and 12 gNB-DUs in the UE coverage range, respectively [8]. Therefore the faster HO that FIBR offers only leads to a modest improvement over the 3GPP approach. Secondly, from Fig. 8, we can observe that the corresponding RLF probabilities are much lower for FIBR, thus the data plane delay, although comparable in duration when they do occur, will occur less frequently for FIBR than 3GPP. URLLC applications may tolerate such delays if they are sufficiently infrequent, e.g., they occur with probability 10^{-6}.

VI. Conclusion

5G mmWave cellular networks are expected to meet the QoS requirements of different applications and services. These applications and services not only require high throughput but also impose significant challenges on the network in terms of latency and reliability. Although mmWave links can achieve data rates as high as a few Gbps, they are highly intermittent in nature causing frequent HOs. Since the 3GPP transport architecture is connection-oriented, where a connection is set up and torn down during every HO procedure, meeting the latency and reliability of URLLC applications is challenging. To satisfy the QoS requirements of different applications and services, primarily URLLC and eMBB, we proposed FIBR, a new transport network architecture for mmWave cellular networks that reduces signaling overhead and simplifies network protocols. In FIBR, a number of BSs in close proximity are grouped together to form a bi-directional buffer insertion ring network. In the FIBR transport architecture, the UEs are connected to the core network without regard as to which BS on the ring the UE is associated with at a given instant, providing an efficient framework for multi-connectivity. To evaluate the performance of FIBR, we compared our proposed architecture with the new 3GPP transport network architecture using a MATLAB simulation. We demonstrated that since FIBR achieves super-fast control signaling between BSs, it reduces the probability of UE blockage, the probability of RLF, and data plane latency. The capability of FIBR to achieve fast and reliable HOs enables the air interface to effectively utilize ephemeral and less reliable links. Thus, our proposed FIBR transport architecture improves the performance of URLLC and eMBB applications in an environment with frequent HOs.

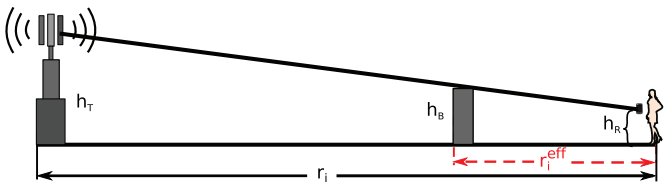


Fig. 12. Blockage of LOS path.

Appendix A

A. Blockage Probability

To compute the Line-of-Sight (LOS) blockage probability of gNB-DUs, we used the expression developed in our previous work [8]. To determine the RLF probability, we will first briefly review some the expressions for blockage probability derived in our previous paper [8]. First, let us consider blockages due to dynamic blockers. Let us also consider the link between the UE and the source gNB-DU (i^{th} gNB-DU in the UE coverage area) (see Fig. 12). Then the dynamic blockage rate α_i of this link is computed as:

$$\alpha_i = \frac{2}{\pi}\lambda_B r_i^{\text{eff}} V = \frac{2}{\pi}\lambda_B V \frac{(h_B - h_R)}{(h_T - h_R)} r_i = C r_i, \qquad (2)$$

where C is:

$$C = \frac{2}{\pi}\lambda_B V \frac{(h_B - h_R)}{(h_T - h_R)}, \qquad (3)$$

λ_B is the dynamic blockers density, V is the speed of dynamic blockers, h_B is the height of blockers, h_R is the height of the UE, and h_T is the height of the transmitter. A detailed derivation of (2) can be found in [8].

Considering an ON-OFF process with α_i (exponentially distributed blocked interval) and μ (exponentially distributed unblocked interval), the blockage probability $P(B_i^d|m, r_i)$ of the link between the source gNB-DU and the UE can be written as

$$P(B_i^d|m, r_i) = \frac{\alpha_i}{\alpha_i + \mu} = \frac{\frac{C}{\mu} r_i}{1 + \frac{C}{\mu} r_i}, \quad \forall i = 1, \cdots, m, \qquad (4)$$

Let us assume $\mathcal{K}$ denotes the set of K gNB-DUs in the K-connectivity scenario, i.e. $\mathcal{K}$ is the set of K gNB-DUs to which the UE is simultaneously connected. Note that K gNB-DUs are randomly selected among M gNB-DUs in the UE coverage area. Due to mathematical complexity, we choose to find the upper and lower bound of the blockage probability in the K-connectivity setting. Note that an upper bound on the blockage probability in the K connectivity setting will be obtained if there is a fixed K number of gNB-DUs in the UE coverage area. Furthermore, a lower bound on the blockage probability $(M > K)$ will be obtained if the UE can perform HO to other gNB-DUs in its coverage area with a zero HO duration. In the K-connectivity scenario, assuming independent blockages of gNB-DUs, the probability of simultaneous blockage $P(B_i^d|K, r_i)$ of all of the K gNB-DUs connected to a UE can be written as:

$$P(B^d|K, r_k) = \prod_{k\in\mathcal{K}} \frac{\frac{C}{\mu} r_k}{1 + \frac{C}{\mu} r_k}, \tag{5}$$

where the number of gNB-DUs m in the UE coverage follows the homogeneous Poisson Point Process BSs model in [8] given by:

$$P_M(m) = \frac{[p\lambda_T \pi R^2]^m}{m!} e^{-p\lambda_T \pi R^2}, \tag{6}$$

where for a self-blockage angle ω, the probability of self blockage $P(B^{\text{self}})$ is computed in [8] as:

$$P(B^{\text{self}}) = \frac{\omega}{2\pi}, \tag{7}$$

Assuming the independence of dynamic blockage and self blockage, the blockage probability of the link between the UE and the k^{th} gNB-DU in the K-Connectivity setting can be written as:

$$P(B_k^{LOS}|K, r_k) = 1 - \left(1 - P(B^{\text{self}})\right)\left(1 - P(B_i^d|K, r_k)\right) \tag{8}$$

Using (4), (7), and (8), the blockage probability of a link between the UE and the $k^{th}, \forall k \in \mathcal{K}$ gNB-DU can be simplified as:

$$P(B_k^{LOS}|K, r_k) = 1 - p\frac{1}{1 + \frac{C}{\mu} r_k}; \quad \forall k \in \mathcal{K}. \tag{9}$$

Thus, the upper-bound on the LOS blockage probability given K-connectivity $P(B^{LOS}|K)$ can be obtained by taking the average of $P(B^{LOS}|K, r_k)$ over the distribution of distances r_k. As K gNB-DUs are randomly selected from the M available gNB-DUs and the UE does not differentiate among the gNB-DUs in the UE coverage area, the distance distribution of gNB-DUs connected to the UE follows the same distance distribution as gNB-DUs in the UE coverage area. Thus, the blockage probability $P\left(B^{LOS}|K\right)$ in

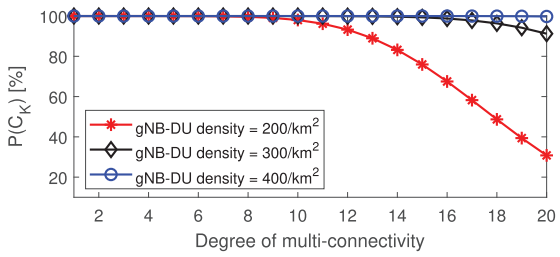


Fig. 13. The probability $P(C_K)$ of having at least K gNB-DUs in the UE coverage area. With a low gNB-DU density, it is highly unlikely to achieve a high degree of multi-connectivity. To achieve a high degree of connectivity and to satisfy the QoS requirements of URLLC applications, a high gNB-DU density is desirable.

K-connectivity setting is computed as:

$$\begin{aligned} P\left(B^{LOS}|K\right) &= \int_{r_1} \cdot \int_{r_K} \prod_{k\in\mathcal{K}} P(B_k^{LOS}|K, r_k) f(r_k) dr_1 \cdots dr_K \\ &= \int_{r_1} \cdot \int_{r_k} \prod_{i\in\mathcal{K}} \left(1 - p\frac{1}{1 + \frac{C}{\mu} r_k}\right) f(r_k) \\ &\quad \times dr_1 \cdots dr_K \\ &= \left(\int_{r=0}^{R} \left(1 - p\frac{1}{1 + \frac{C}{\mu} r}\right) \frac{2r}{R^2} dr\right)^K \\ &= \left(1 - p\int_{r=0}^{R} \frac{1}{1 + \frac{C}{\mu} r} \frac{2r}{R^2} dr\right)^K \\ &= \left(1 - \frac{2p\mu}{C^2R^2}\left(CR - \mu \ln\left(1 + \frac{C}{\mu}R\right)\right)\right)^K. \end{aligned} \tag{10}$$

Following the previous discussion about the lower-bound on the LOS blockage probability, we can write the blockage probability as:

$$P\left(B^{LOS}|k\right) = \left(1 - \frac{2p\mu}{C^2R^2}\left(CR - \mu \ln\left(1 + \frac{C}{\mu}R\right)\right)\right)^M, \quad \forall k \in K. \tag{11}$$

Furthermore, note that K-connectivity can be achieved if, and only if, there are at least K gNB-DUs in the UE coverage area. Otherwise, if the gNB-DUs density is significantly low, we simply argue that a higher degree of connectivity cannot be achieved. If there are at least K gNB-DUs in the UE coverage area, then the blockage probability in the K-connectivity setting is expressed by (10). The probability $P\left(C_K\right)$ of having at least K gNB-DUs in the UE coverage area is given by:

$$\begin{aligned} P\left(C_K\right) &= \sum_{m=K}^{\infty} P_M(m) \\ &= \sum_{m=K}^{\infty} \frac{[pq\lambda_T \pi R^2]^m}{m!} e^{-pq\lambda_T \pi R^2} \end{aligned} \tag{12}$$

Fig. 13 shows $P\left(C_K\right)$ for different gNB-DU density values. Note that for a smaller gNB-DU density, a high degree of multi-connectivity is more difficult to achieve.

B. RLF Probability

Let us now consider the scenario that a blockage event leads to an RLF. In general, an RLF is declared upon the expiration of the N310 timer and the T310 timer, together accounting for around 30 ms. On the expiration of these timers, an RLF is declared by the UE and the UE initiates an RLF recovery process. Thus, if the blockage duration is larger than 30 ms, an RLF will occur. Let us assume the blockage duration is T_B, then the probability of RLF failure can be written as:

$$\begin{aligned} P(RLF) &= P\left(B^{LOS}|K, T_B|K > 30\right) \\ &= P(B^{LOS}|K)P(T_B|K > 30). \end{aligned} \tag{13}$$

Note that for simplicity we assume that blockage duration does not depend upon the probability of the blockages.

The probability that the blockage duration is greater than 30 ms is:

$$\begin{aligned} P(T_B > 30|K) &= \int_{30}^{\infty} m\mu e^{-K\mu t} dt \\ &= e^{-30K\mu} \end{aligned} \tag{14}$$

Therefore, using (10), (13) and (14), the RLF probability can be derived as:

$$P(RLF) = (1 - ap)^K e^{-30K\mu}. \tag{15}$$

APPENDIX B

C. Queueing Analysis of FIBR

A crucial design issue is the sizing of the ring capacity to meet the bandwidth and delay QoS needs for the UEs covered by one TA-GW. The traffic on the ring will vary with the traffic on the uplink and downlink of a group of UEs. It is necessary to ensure that the ring can accommodate this varying traffic and deliver it within a tight time bound. We therefore present an analysis of the uplink and downlink queueing delay in FIBR for the worst case case when both uplink and downlink traffic share a single ring after a failure (see Fig. 2). We only consider queueing delay in our analysis. Apart from the queueing delay, other delays such as propagation (5 μs assuming a ring length of 1 km) and transmission delay (12 μs for 100 nodes in the ring assuming a packet size of 1500 bytes) remain quite low as compared to the delay requirements of URLLC applications. We therefore do not focus on them in our analysis. We model this system, which consists of these two buffers, as a prioritized non-preemptive head-of-the-line queue [51]. For a first-order evaluation of the system, we assume that the packet arrivals are Poisson distributed and that their service times are exponentially distributed, i.e., the queue discipline at the insertion buffers is M/M/1.

We assume that there are L ring nodes in the TA and index them according to closeness to the TA-GW (assume that the index of TA-GW is $j = 1$), i.e., downlink traffic will first reach the ring node with index $j = 2$, and completes a full circle with the ring node with index $j = L$. Let us define the packet arrival rates at the gNB-DU uplink and insertion buffer of the ring node j by $\lambda_{\mathrm{U}}(j)$ and $\lambda_{\mathrm{I}}(j)$ respectively, $1/\mu$ is the mean packet size (bits/packet) and C is the FIBR capacity (bits/sec). The utilization factors $\rho_U(j)$ and $\rho_I(j)$ of the gNB-DU uplink and insertion buffer are calculated as follows:

$$\rho_U(j) = \lambda_{\mathrm{U}}(j)/(\mu C), \tag{16}$$

$$\rho_I(j) = \lambda_{\mathrm{I}}(j)/(\mu C). \tag{17}$$

The packet priorities can heavily affect the delays in the two queues. Let us consider two different priority options: a) *Ring priority*, where packets in the insertion buffer are transmitted before those in the gNB-DU uplink buffer, and b) *gNB-DU priority*, where packets in the gNB-DU uplink buffer are prioritized over those in the insertion buffer. To prevent overflow, we simplify the analysis by assuming that both buffers are sufficiently large.

The waiting time in the insertion buffer of ring node j for the ring and gNB-DU priorities are, respectively [51],

$$W_{\mathrm{I}}^{\mathrm{Ring}}(j) = \frac{R}{1 - \rho_{\mathrm{I}}(j)}, \tag{18}$$

$$W_{\mathrm{I}}^{\text{gNB-DU}}(j) = \frac{R}{(1 - \rho_{\mathrm{U}}(j))\,(1 - \rho_{\mathrm{I}}(j) - \rho_{\mathrm{U}}(j))}, \tag{19}$$

where R is the mean residual service time of packets being serviced upon arrival, and is given by [51]:

$$R = (\rho_{\mathrm{I}}(j) + \rho_{\mathrm{U}}(j))/(\mu C). \tag{20}$$

Similarly, the waiting time in the gNB-DU uplink buffer of ring node j for ring and gNB-DU priority are, respectively,

$$W_{\mathrm{U}}^{\mathrm{Ring}}(j) = \frac{R}{(1 - \rho_{\mathrm{I}}(j))(1 - \rho_{\mathrm{I}}(j) - \rho_{\mathrm{U}}(j))}, \text{and} \tag{21}$$

$$W_{\mathrm{U}}^{\text{gNB-DU}}(j) = \frac{R}{1 - \rho_{\mathrm{U}}(j)}. \tag{22}$$

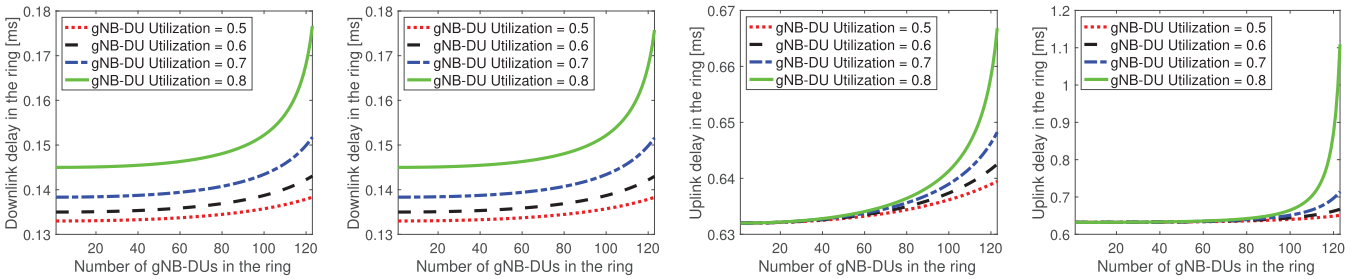
1) Downlink Traffic: We model the queue in the gNB-DU downlink bearer buffer as an M/M/1/N queue. We select the gNB-DU downlink buffer length N (in packets) to be equal to a fixed multiple of the product of the 5G slot duration (in seconds) and the gNB-DU downlink bandwidth. We select a buffer size several times higher than the product of the slot duration and the gNB-DU downlink bandwidth to ensure that minimum latency can be achieved without link starvation. In the 5G cellular systems, the slot duration is defined as $125\ \mu s$ for URLLC applications [52].

$$N = T_{\text{frame}} \times BW_{\text{gNB-DU}}. \tag{23}$$

The waiting time $W_{\text{gNB-DU}}^{\text{DL}}$ for the aforementioned queueing system is obtained in [53]. Thus, the downlink delay $W_{\mathrm{DL}}(j)$ for a packet destined to a UE associated with gNB-DU j is computed as the sum of the delays in the insertion buffers, until the previous ring node, and the current gNB-DU downlink queueing delay:

$$W_{\mathrm{DL}}^{(P)}(j) = \sum_{k=0}^{j-1} W_{\mathrm{I}}^{(P)}(k) + W_{\text{gNB-DU}}^{\text{DL}}(j), \tag{24}$$

where P is the priority used at a ring node. Note that in the bidirectional downlink ring, no packet enters the ring from the gNB-DU, thus only ring priority is considered during normal operation. However, one of the discussed ring priorities can be considered in case that a ring fails. Note that the waiting time



(a) Downlink: gNB-DU priority (b) Downlink: ring priority (c) Uplink: gNB-DU priority (d) Uplink: ring priority

Fig. 14. FIBR uplink and downlink queueing delay after the failure of one ring: the maximum number of gNB-DUs supported in the FIBR for 1+1 ring protection is bounded by the ring failure scenario. It is further bounded by the gNB-DU utilization that satisfies QoS requirements of URLLC applications. After considering both 1+1 ring protection and QoS requirements, the maximum number of gNB-DUs in FIBR is evaluated as approximately 120. The ring capacity is 400 Gbps and the service capacity at each gNB-DUs is 3 Gbps for the analysis.

$W^{\mathrm{DL}}_{\text{gNB-DU}}$ in the downlink bearer buffer includes both head-of-line processing delay and scheduling delay.

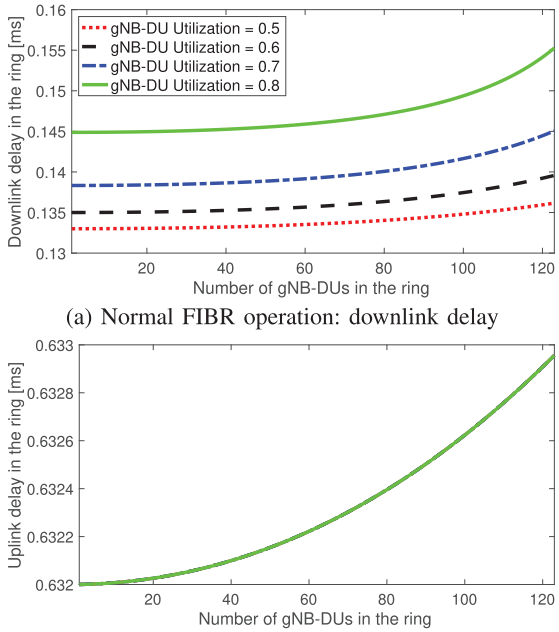
D. Uplink Traffic

The total queueing delay $W_{\mathrm{UL}}(j)$ for an uplink packet in the ring node entering the gNB-DU j is:

$$W^{(P)}_{\mathrm{UL}}(j) = \sum_{k=j+1}^{L} W^{(P)}_{\mathrm{I}}(k) + W^{(P)}_{\mathrm{U}}(j), \tag{25}$$

where P is the priority used at a ring node and j is the gNB-DU node from which uplink traffic is inserted into the ring. Similar to previous discussion, the waiting time $W^{(P)}_{\mathrm{U}}(j)$ in the uplink bearer buffer includes both head-of-line processing delay and uplink scheduling delay. The uplink scheduling delay for URLLC applications is computed as 632 μs by considering delay associated with the uplink transmission grant and its processing [54].

Note that the performance of FIBR is limited by two determining factors, (i) the protection mechanism: we consider $1+1$ protection of the ring, and (ii) the QoS agreement: maximize the utilization at the gNB-DUs while satisfying the QoS requirements of URLLC applications. In our analysis, we consider a ring capacity of 400 Gbps [55] and the service rate at each gNB-DU to be 3 Gbps [56]. Furthermore, we assume uplink traffic to be one fourth of the downlink traffic, as predicted by the International Telecommunication Union [57], thus the average uplink load is 750 Mbps at every gNB-DU. We first examine the failure scenario to evaluate the maximum number of gNB-DUs that can be supported in FIBR considering $1+1$ protection, the QoS requirements of URLLC applications, stability of the FIBR, and utilization of gNB-DUs and ring. Fig. 14 represents the uplink and downlink queueing delay in FIBR when one of the rings fail. For the considered parameters of ring capacity and service rate at gNB-DUs, we compute that around 120 gNB-DUs can be supported by the ring. From Fig. 14, we observe that the uplink and downlink delay increases with the number of gNB-DUs and their utilization. We can observe from Fig. 14(c) and Fig. 14(d) that uplink delay may be significantly higher for ring priority as compared to gNB-DU priority in FIBR. This happens as insertion queues at each gNB-DU are in general multiple times larger than uplink queues, thus prioritizing packets of insertion queues leads to additional delay for uplink packets. By contrast, prioritizing uplink packets induces additional delay to downlink packets, but this delay increment is marginal (see Fig. 14(a) and Fig. 14(b)). Thus, when the ring fails, gNB-DU priority is the natural choice. During the normal operation of FIBR, as downlink and uplink traffic are separated in different rings, we have ring priority in the ring handling the downlink traffic and gNB-DU priority in the ring handling the uplink traffic. Fig. 15 plots the downlink and uplink delay in the FIBR during normal operation. As shown in Fig. 15(a) and Fig. 15(b)), we obtain lower uplink and downlink delays as they are carried over two separate rings.



(a) Normal FIBR operation: downlink delay

(b) Normal FIBR operation: uplink delay

Fig. 15. Downlink and uplink queueing delay during the normal operations of FIBR. Downlink delay increases with the utilization of the gNB-DUs and the number of gNB-DUs in the downlink ring.

REFERENCES

[1] *5G; Service Requirements for Next Generation New Services and Markets*, document 3GPP TS 22.261, 3GPP Standard v15.5.0, Jul. 2018.

[2] M. Rybakowski *et al.*, "Challenges & solutions for above 6 GHz radio access network integration for future mobile communication systems," in *Proc. IEEE ICC*, May 2016, pp. 614–619.

[3] *5G; NR; User Equipment (UE) Radio Transmission and Reception; Part 2: Range 2 Standalone*, document 3GPP TS 38.101-2, 3GPP Standard v15.2.0, Jul. 2018.

[4] T. S. Rappaport *et al.*, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[5] K. C. Allen, N. DeMinco, J. R. Hoffman, Y. Lo, and P. B. Papazian, "Building penetration loss measurements at 900 MHz, 11.4 GHz, and 28.8 GHz," U.S. Dept. Commerce, Nat. Telecommun. Inf. Admin., Boulder, CO, USA, Tech. Rep. 94-306, May 1994, pp. 94–306.

[6] J. S. Lu, D. Steinbach, P. Cabrol, and P. Pietraski "Modeling human blockers in millimeter wave radio links," *ZTE Commun.*, vol. 10, pp. 23–28, Dec. 2012.

[7] I. K. Jain, R. Kumar, and S. Panwar, "Driven by capacity or blockage? A millimeter wave blockage analysis," in *Proc. 30th Int. Teletraffic Congr. (ITC)*, Sep. 2018, pp. 153–159.

[8] I. K. Jain, R. Kumar, and S. S. Panwar, "The impact of mobile blockers on millimeter wave cellular systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 854–868, Apr. 2019.

[9] A. Sutton, "5G network architecture, design and optimisation," British Telecom, London, U.K., Tech. Rep., Jan. 2018. [Online]. Available: https://bit.ly/2E8J0m3

[10] *Technical Specification Group Services and System Aspects; System Architecture for the 5G System*, document 3GPP TS 23.501, 3GPP Standard v15.0.0, Nov. 2017.

[11] R. Trivisonno, M. Condoluci, X. An, and T. Mahmoodi, "mIoT slice for 5G systems: Design and performance evaluation," *Sensors*, vol. 18, no. 2, p. 635, Feb. 2018.

[12] *5G; NR; Medium Access Control (MAC) Protocol Specification*, document 3GPP TS 38.231, 3GPP Standard v15.3.0, Sep. 2018.

[13] H. S. Park, Y. Lee, T. J. Kim, B. C. Kim, and J. Y. Lee, "Handover mechanism in NR for ultra-reliable low-latency communications," *IEEE Netw.*, vol. 32, no. 2, pp. 41–47, Mar. 2018.

[14] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55765–55779, 2018.

[15] Y. Rao *et al.*, "New services & applications with 5G ultra-reliable low latency communication," 5G Americas, Tech. Rep., Nov. 2018.

[16] N. Alliance, "NGMN 5G initiative white paper," NGMN, Tech. Rep., Feb. 2015. [Online]. Available: https://bit.ly/2HuXhdK

[17] W. Bux, F. Closs, K. Kuemmerle, H. Keller, and H. Mueller, "Architecture and design of a reliable token-ring network," *IEEE J. Sel. Areas Commun.*, vol. 1, no. 5, pp. 756–765, Nov. 1983.

[18] W. Dobosiewicz and P. Gburzynski, "On token protocols for high-speed multiple-ring networks," in *Proc. IEEE ICNP*, Oct. 1993, pp. 300–315.

[19] R. Cohen and A. Segall, "Multiple logical token-rings in a single high-speed ring," *IEEE Trans. Commun.*, vol. 42, no. 234, pp. 1712–1721, Feb. 1994.

[20] W. Bux and M. Schlatter, "An approximate method for the performance analysis of buffer insertion rings," *IEEE Trans. Commun.*, vol. COM-31, no. 1, pp. 50–55, Jan. 1983.

[21] F. Davik, M. Yilmaz, S. Gjessing, and N. Uzun, "IEEE 802.17 resilient packet ring tutorial," *IEEE Commun. Mag.*, vol. 42, no. 3, pp. 112–118, Mar. 2004.

[22] Y. Li, B. Cao, and C. Wang, "Handover schemes in heterogeneous LTE networks: Challenges and opportunities," *IEEE Wireless Commun.*, vol. 23, no. 2, pp. 112–117, Apr. 2016.

[23] X. Yan, Y. A. Şekercioğlu, and S. Narayanan, "A survey of vertical handover decision algorithms in fourth generation heterogeneous wireless networks," *Comput. Netw.*, vol. 54, pp. 1848–1863, Feb. 2010.

[24] D. Xenakis, N. Passas, L. Merakos, and C. Verikoukis, "Mobility management for femtocells in LTE-advanced: Key aspects and survey of handover decision algorithms," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 64–91, 1st Quart., 2014.

[25] *Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN)*, document 3GPP TS 36.300, 3GPP Standard v15.5.0, Mar. 2019.

[26] H.-D. Bae, B. Ryu, and N.-H. Park, "Analysis of handover failures in LTE femtocell systems," in *Proc. IEEE ATNAC*, Nov. 2011, pp. 1–5.

[27] *Universal Mobile Telecommunications System (UMTS); LTE; 5G; NR; Multi-Connectivity; Overall Description; Stage-2*, document 3GPP TS 37.340, 3GPP Std. v15.3.0, Sep. 2018.

[28] A. Ravanshid *et al.*, "Multi-connectivity functional architectures in 5G," in *Proc. IEEE ICC*, May 2016, pp. 187–192.

[29] V. Petrov *et al.*, "Dynamic multi-connectivity performance in ultra-dense urban mmWave deployments," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 2038–2055, Sep. 2017.

[30] C. Tatino, I. Malanchini, N. Pappas, and D. Yuan, "Maximum throughput scheduling for multi-connectivity in millimeter-wave networks," in *Proc. IEEE WiOpt*, May 2018, pp. 1–6.

[31] M. Polese, M. Giordani, M. Mezzavilla, S. Rangan, and M. Zorzi, "Improved handover through dual connectivity in 5G mmWave mobile networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 2069–2084, Sep. 2017.

[32] D. H. Hagos and R. Kapitza, "Study on performance-centric offload strategies for LTE networks," in *Proc. IFIP WMNC*, Apr. 2013, pp. 1–10.

[33] C. Pei *et al.*, "WiFi can be the weakest link of round trip network latency in the wild," in *Proc. IEEE INFOCOM*, Apr. 2016, pp. 1–9.

[34] N. Alliance, "NGMN overview on 5G RAN functional decomposition," NGMN, Tech. Rep., Feb. 2018. [Online]. Available: https://bit.ly/2HuXhdK

[35] *Transport Network Support of IMT-2020/5G*, ITU-T, document CSTR-TN5G, Feb. 2018. [Online]. Available: https://bit.ly/2HnxQM9

[36] B. Bertenyi, R. Burbidge, G. Masini, S. Sirotkin, and Y. Gao, "NG radio access network (NG-RAN)," *J. ICT Standardization*, vol. 6, no. 1, pp. 59–76, May 2018.

[37] D. S. Michalopoulos, A. Maeder, and N. Kolehmainen, "5G multi-connectivity with non-ideal backhaul: Distributed vs cloud-based architecture," in *Proc. IEEE Globecom Wkshps*, Dec. 2018, pp. 1–6.

[38] D. Lee *et al.*, "Coordinated multipoint transmission and reception in LTE-advanced: Deployment scenarios and operational challenges," *IEEE Commun. Mag.*, vol. 50, no. 2, pp. 148–155, Feb. 2012.

[39] K. Pedersen, G. Pocovi, J. Steiner, and A. Maeder, "Agile 5G scheduler for improved E2E performance and flexibility for different network implementations," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 210–217, Mar. 2018.

[40] F. Alharbi and N. Ansari, "SSA: Simple scheduling algorithm for resilient packet ring networks," *IEE Proc.—Commun.*, vol. 153, no. 2, pp. 183–188, Apr. 2006.

[41] *5G; NG-RAN; Architecture Description*, document 3GPP TS 38.401, 3GPP Standard v15.2.0, Jul. 2018.

[42] H. Zhang, C. Jiang, and J. Cheng, "Cooperative interference mitigation and handover management for heterogeneous cloud small cell networks," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 92–99, Jun. 2015.

[43] LTE Quick Reference. *Radio Link Failure (RLF)*. Accessed: May 4, 2019. [Online]. Available: https://bit.ly/2lzjNLb

[44] A. Khlass, S. E. Elayoubi, and T. Bonald, "Multi-flow transmission and carrier aggregation inter-operation in HSPA+ advanced," in *Proc. IEEE VTC Fall*, Sep. 2014, pp. 1–5.

[45] D. B. Johnson and D. A. Maltz, *Dynamic Source Routing in Ad Hoc Wireless Networks*. Boston, MA, USA: Springer, 1996, pp. 153–181.

[46] M. Boutin. *Random Waypoint Mobility Model*. Accessed: Mar. 18, 2019. [Online]. Available: https://www.mathworks.com

[47] S. Barbera *et al.*, "Synchronized RACH-less handover solution for LTE heterogeneous networks," in *Proc. IEEE ISWCS*, Aug. 2015, pp. 755–759.

[48] T. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.

[49] G. R. MacCartney and T. S. Rappaport, "Rural macrocell path loss models for millimeter wave wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1663–1677, Jul. 2017.

[50] *LTE; Evolved Universal Terrestrial Radio Access (EUTRA); Requirements for Support of Radio Resource Management*, document 3GPP TS 36.133, 3GPP Standard v14.3.0, Apr. 2017.

[51] L. Kleinrock, *Queueing Systems*, vol. 1. New York, NY, USA: Wiley, 1975.

[52] *5G; NR; Physical Channels and Modulation*, document 3GPP TS 38.211, 3GPP Standard v15.2.0, Jul. 2018.

[53] J. L. van den Berg and O. J. Boxma, "The*M/G/*1 queue with processor sharing and its relation to a feedback queue," *Queueing Syst.*, vol. 9, no. 4, pp. 365–401, Dec. 1991.

[54] *Facilitating eMBB/URLLC UL Multiplexing With the Zero-Wait-Time Scheduling Request Underlay Channel*, document 3GPP R1-1701612, 3GPP RAN1 #88, 3GPP Standard, Feb. 2017.

[55] Accton Making Partnership Work. *The New World of 400 Gbps Ethernet*. Accessed: Apr. 2, 2019. [Online]. Available: https://bit.ly/2QabYqH

[56] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, Jun. 2011.

[57] *IMT Traffic Estimates for the Years 2020 to 2030*, document ITU-R, M.2370-0, ITU, Geneva, Switzerland, Jul. 2015.

Athanasios Koutsaftis received the B.Tech. degree in electrical and computer engineering from the National Technical University of Athens in 2015. He is currently pursuing the Ph.D. degree in electrical engineering with the Tandon School of Engineering, New York University, New York City, NY, USA. In Summer 2018, he worked at Nokia Bell Labs. His research interests include wireless communications and networks. He was awarded the Dean's Fellowship to pursue the Ph.D. degree at NYU.

Rajeev Kumar received the B.Tech. and M.Tech. degrees in electrical engineering from the Indian Institute of Technology Madras, Chennai, India, in 2013. He is currently pursuing the Ph.D. degree in electrical engineering with the Tandon School of Engineering, New York University, New York City, NY, USA. In the Summers of 2017 and 2018, he worked at Nokia Bell Labs. His research interests focus on latency issues related to 5G cellular systems.

Pei Liu received the B.S. and M.S. degrees in electrical engineering from Xi'an Jiaotong University, China, in 1997 and 2000, respectively, and the Ph.D. degree in electrical and computer engineering from Polytechnic University in 2007. He is currently a Research Assistant Professor with the Electrical and Computer Engineering Department, NYU Tandon School of Engineering. His research interests include designing and analyzing wireless network protocols with an emphasis on cross-layer optimizations. His current research topics include next-gen communication networks and software-defined radios.

Shivendra S. Panwar (S'82–M'85–SM'00–F'11) received the Ph.D. degree in electrical and computer engineering from the University of Massachusetts, Amherst, MA, USA, in 1986. He is currently a Professor with the Electrical and Computer Engineering Department, NYU Tandon School of Engineering. He is also the Director of the New York State Center for Advanced Technology in Telecommunications (CATT), the Faculty Director and co-founder of the New York City Media Lab, and a member of NYU Wireless. His research interests include the performance analysis and design of networks. His current research focused on cross-layer research issues in wireless networks, and multimedia transport over networks. He has coauthored a textbook titled *TCP/IP Essentials: A Lab based Approach* (Cambridge University Press). He was a winner of the IEEE Communication Society's Leonard Abraham Prize for 2004, the ICC Best Paper Award in 2016, and the Sony Research Award. He was also co-awarded the Best Paper in 2011 Multimedia Communications Award. He has served as the Secretary for the Technical Affairs Council of the IEEE Communications Society.