A Multi-Connectivity Architecture with Data Replication for XR Traffic in mmWave Networks

Muhammad Affan Javed, Pei Liu, and Shivendra S. Panwar
Department of Electrical and Computer Engineering, NYU Tandon School of Engineering,
Brooklyn, New York, USA.
Email: {maj407, peiliu, panwar}@nyu.edu

***Abstract*—mmWave communications are paving the way for next generation cellular networks due to their inherent ability to provide high data rates and mitigate interference. However, the intermittency of mmWave links and the advent of low-latency eXtended Reality(XR) applications presents a conundrum for satisfying strict QoS constraints. We propose a multi-tiered multi-connectivity network architecture which exploits mmWave macro-diversity to allow users (UEs) to connect to multiple base stations (gNBs) simultaneously and rapidly switch between them. The power of our proposed architecture lies in multiple tiers of multi-connectivity, wherein we selectively replicate UE data at a subset of associated gNBs in order to improve the response time to blockages. We evaluate the performance of XR traffic with standard scheduling algorithms. Our results show that connecting to multiple gNBs and enabling the scheduler to rapidly switch between them shields the UEs from higher handover delays and minimizes data plane interruptions. Although we show that our network architecture allows for much better performance even with conventional scheduling algorithms, we also highlight the need for better scheduling algorithms optimized for the multi-connectivity paradigm.**

***Index Terms*—multi-connectivity, millimeter wave, mmWave, handover, blockages, low latency, XR applications, quality of service**

I. Introduction

The promise of eXtended Reality (XR) applications, which include Virtual Reality (VR), Augmented Reality (AR), and Cloud Gaming (CG), has taken the world by storm [1]. These services are the cornerstone of next-generation wireless networks and fundamental changes in network architecture and protocols are needed in order to meet their requirements of high bandwidths, low latencies, and strict deadlines [2]. Fortunately, as we move into the millimeter wave (mmWave) spectrum and beyond, the fundamental characteristics of these wireless systems offer us potential solutions for catering to XR traffic demands.

Fifth generation (5G) cellular networks have already led the charge into mmWave technology, which operates at frequencies above 24 GHz, thereby utilizing the enormous amount of spectrum available in these frequency bands [3]. At these frequencies, the radio propagation characteristics are starkly different from their microwave counterparts. First, according to the Friis transmission equation [4], the path loss can easily exhibit 30-40 dB more attenuation. This higher path loss necessitates focusing power into fairly narrow and very directional beams, that can be realized through phased antenna arrays, whose implementation is made possible thanks to the smaller wavelengths that correspond to these frequencies. Furthermore, due to the exacerbated blockage and shadowing effects [5], the wireless links exhibit rapid variations in quality, thereby leading to severe intermittency in link connectivity between the user (UE) and the base station (gNB).

To address these challenges, and to maintain an acceptable level of service despite this intermittency, the density of gNBs in mmWave cellular networks is expected to be significantly higher than in sub-6 GHz systems [6]. It will be greatly beneficial for the UEs to harness macro-diversity from the nearby gNBs in sixth generation (6G) and future cellular networks. As XR applications are becoming an increasingly vital component of the ecosystem, it is essential to study the impact of the degree of multi-connectivity and handover delays on the performance of XR applications that operate under strict Quality of Service (QoS) constraints.

In this paper, we propose a network architecture that utilizes mmWave multi-connectivity in order to reduce the handover delays experienced by the UEs and minimize data plane interruptions. We evaluate the performance of XR applications over such a network and validate that the higher level of connectivity offered by such a network does in fact translate to better performance for XR applications.

The key contributions of this paper are as follows:

- We propose a multi-tiered network architecture for mmWave multi-connectivity in the access network that provides better performance even with conventional scheduling algorithms. We show that our architecture allows us to shield the UEs from high handover latencies in case of blockages, minimizes data plane interruptions and enables fast switching between multiple gNBs.
- We present system-level performance evaluation results for XR applications in a multi-cell mmWave network using the statistical traffic model given in 3GPP standards. Our results show that our multi-tiered multi-connectivity architecture significantly reduces the response time to blockages and leads to better performance for XR applications with strict deadlines.

This work was supported in part by NYU Wireless, the NY State Center for Advanced Technology in Telecommunications (CATT), and NYU IT High-Performance Computing resources, services, and staff expertise.

ISBN 978-3-903176-57-7

The rest of the paper is organized as follows. Section II presents related work. We propose our multi-connectivity architecture in Section III and describe the system models in Section IV. In Section V we describe our simulation setup, present results obtained by our simulations, and discuss the key takeaways. Finally, Section VI concludes our paper and highlights possible avenues for future research.

II. Related Work

Exploiting multi-connectivity in the access network to gain better performance is not a new concept, nor is it unique to mmWave networks. In fact, multi-connectivity was first proposed for sub-6 GHz networks with the introduction of Dual Connectivity (DC) in heterogeneous Long Term Evolution (LTE) networks in 3GPP Release 12 [7]. DC refers to the most basic multi-connectivity where the UE is connected to only two base stations. Although DC contributed to throughput gains, it did not gain much traction in sub-6 GHz networks because the overhead involved in maintaining dual connectivity far outweighed any performance improvements to be had. With the move towards mmWave networks in 5G, multi-connectivity has received renewed interest due to several reasons. First, it is easier for a UE to be within range of multiple gNBs due to the high densification of gNBs required to provide adequate coverage at mmWave frequencies. Second, directional beams in mmWave networks offer an opportunity to provide multi-connectivity without creating excessive interference between neighboring gNBs. Last, meeting the strict QoS constraints of next-generation applications such as XR provides further incentives that makes the high overhead cost of multi-connectivity tolerable from a cost-benefit tradeoff perspective.

Multi-connectivity in mmWave networks has been studied in [8], where the impact of gNB discovery time, handover execution times and degree of multi-connectivity was studied with respect to QoS criteria such as out-of-service probability, outage duration and radio link failure (RLF) probability. However, the weakness of the proposed architecture was that data would either have to be replicated at all connected base stations, which would be prohibitively expensive in practice, or would have to be redirected from the Master base station to the Secondary base stations, which would incur additional delays. In [9], a new transport network architecture was proposed that would enable fast control signalling and leverage multi-connectivity via a fiber ring to improve QoS for different applications. Petrov et al. [10] considered different multi-connectivity scenarios to study the impact of the degree of connectivity, and showed that a high degree of multi-connectivity would enhance the reliability of the system at the cost of significant signaling and computation overhead. On a similar note, Gapayenko et al. [11] showed that increasing the degree of multi-connectivity up to 4 could provide benefits in terms of lower outage probability and higher spectral efficiency.

With regards to standardization, in Release 12 3GPP introduced the Intra-E-UTRA Dual Connectivity (DC) which is the inter-site DC between two LTE base stations (i.e., same Radio Access Technology), where both base stations are connected to the Evolved Packet Core (EPC). Since then, 3GPP has iteratively expanded on use cases and functionality of dual connectivity, and it is now a key feature of the 5G NR standard. According to the 3GPP NR Release 16 standard [12], Multi-Radio Dual Connectivity (MR-DC) is the term that is generally being used for multi-connectivity. With the introduction of 5G New Radio (NR), 3GPP introduced four configurations for MR-DC, of which only one (NR-NR Dual Connectivity or NR-DC) falls under the standalone architecture and represents the 5G equivalent of the LTE DC.

The performance of XR applications in different networks and systems has also been a keen area of interest recently. XR is characterized by both high data rate and a strict packet delay budget (PDB), thereby giving it the difficult-to-meet constraints of both 5G enhanced mobile broadband (eMBB) and ultra reliable low latency communications (URLLC) [13]. In [14], system-level performance results for XR over a 5G-NR network were presented and several enhancements, such as traffic aware scheduling, were proposed in order to boost the performance. Petrov et al. [15] also performed a case study which demonstrated that 5G NR can already support XR services, but with a limitation on the number of XR devices per cell at high data rates. A key drawback of these studies is that they fail to explicitly take into account the effect of blockages, which severely affect the performance of any mmWave network. The focus of this paper is to quantify the effect of blockages on XR traffic performance, and evaluate how a multi-connectivity architecture can enable us to circumvent data plane interruptions that arise due to blockages and satisfy strict QoS constraints.

III. Multi-Connectivity Architecture

We consider a mmWave wireless network comprising of a set of gNBs, $|\mathcal{M}| = M$, and a set of UEs, $|\mathcal{N}| = N$. Thus, there are up to $M \times N$ mmWave links in the system. The critical component of our infrastructure is the UEs' ability to connect to multiple gNBs simultaneously, a feature of emerging 3GPP standards [12]. The cornerstone of this architecture is that it further devolves multi-connectivity into two main tiers, Association and Data Replication, based on the level of connection and data availability. The bifurcation of the multi-connectivity architecture is motivated in part by the overhead costs of replicating UE data at a large number of gNBs. By choosing to associate with a larger number of gNBs, and replicating the data at only a smaller subset of them we can reap the benefits of a higher degree of connectivity while significantly reducing the overhead costs. Moreover, the two-tier architecture allows us to reduce the handover delay experienced by the UEs in the vast majority of blockage scenarios. This allows us to minimize data plane interruptions and boost QoS performance for XR applications. Fig.1 depicts our multi-connectivity network architecture.

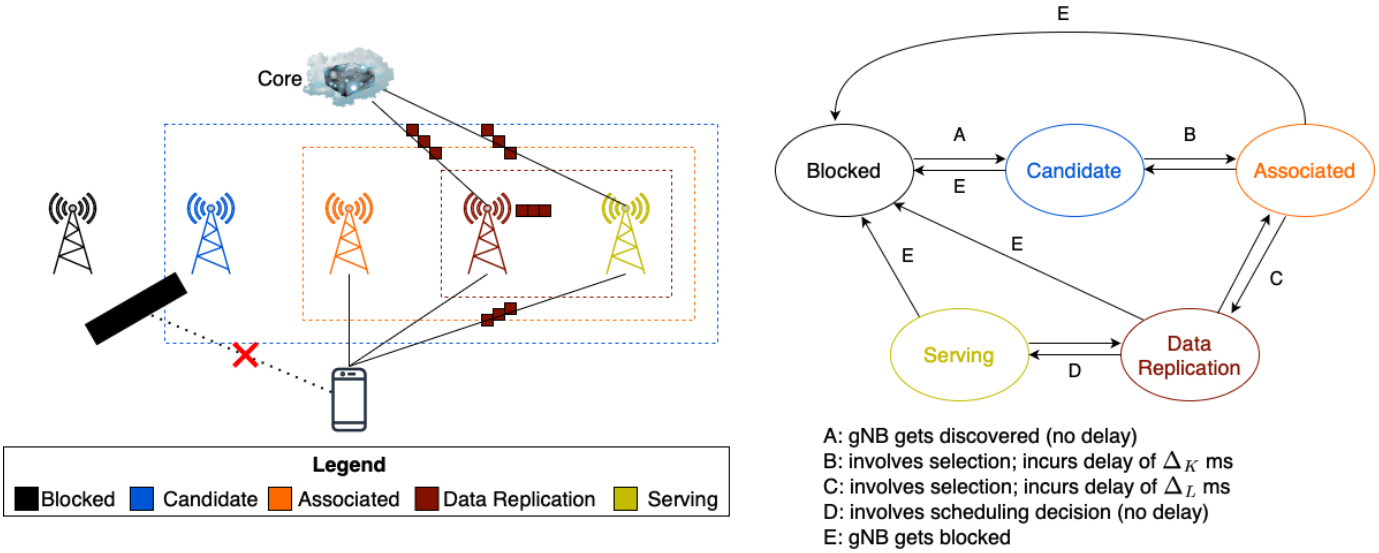


Fig. 1: Network architecture, illustrating the different tiers of multi-connectivity. Here, $K = 3$ and $L = 2$.

A. Multi-Connectivity Tiers

Since the range of mmWave links is quite short, it is possible that some gNBs are out of range of the UEs and, hence, no connection is possible. Even if a gNB is within range, it is possible that it is blocked and hence undiscovered by the UE. We define a set $\mathcal{C}_{n,t} \subset M$, which comprises of all the *candidate* base stations for user n at time t:

$$\mathcal{C}_{n,t} = \{m : \sigma_{m,n} > \sigma_{th}, |\mathcal{C}_{n,t}| \leq M \quad \forall m \in \mathcal{M}\}, \quad (1)$$

where $\sigma_{m,n}$ is the signal-to-noise ratio (SNR) of the link between the gNB m and UE n, and σ_{th} is the minimum SNR required for a successful connection between a gNB-UE pair. User n is in range of all base stations in $\mathcal{C}_{n,t}$ and can choose to connect to any of them.

In the multi-connectivity setting, we assume that a UE can be *associated* with multiple gNBs at the same time. Specifically, the UE maintains a control plane connection with all the gNBs in the Associated set ($\mathcal{K}_{n,t}$). We define K, the degree of association, which determines the maximum number of gNBs a UE will simultaneously associate with, i.e. $|\mathcal{K}_{n,t}| \leq K$.

The set $\mathcal{K}_{n,t} \subset \mathcal{C}_{n,t}$ comprises of the gNBs that UE n is associated with at time t. We assume that the best subset of gNBs to associate with at time t is the set of gNBs with the highest channel quality to the UE at time t.The algorithm for selecting $\mathcal{K}_{n,t}$ would start with an ordered set of SNRs and pick the gNBs corresponding to the K highest SNR values. An *associated* gNB-UE pair would have an active control channel open between them and will routinely exchange control messages and signalling required to maintain the UE state at the gNB, as well as any signalling required for beam tracking, alignment and beam switching. However, associated gNBs (except for one) do not have a data plane connection with the UE or up-to-date UE data available for delivery.

A smaller subset of $\mathcal{K}_{n,t}$ is then chosen as the Data Replication set of gNBs ($\mathcal{L}_{n,t}$). The set $\mathcal{L}_{n,t} \subset \mathcal{K}_{n,t}$ is the set of all gNBs that are associated with UE n and have copies of UE n's data ready for transmission at time t. gNBs in $\mathcal{L}_{n,t}$ pre-fetch UE data and track UE data delivery status. We also define L, where $L \leq K$, as the degree of replication - another parameter that determines the maximum number of gNBs that will replicate the UE data and have it instantaneously ready for transmission, i.e. $|\mathcal{L}_{n,t}| \leq L$. At any given instance, a UE will have a data plane connection open with only one Serving gNB, which is chosen from $\mathcal{L}_{n,t}$ by the scheduling agent. The scheduling agent's job includes selecting a Serving gNB for the UE from the gNBs in $\mathcal{L}_{n,t}$. Thus, $\mathcal{L}_{n,t}$ consists of one master/serving gNB and several other secondary gNBs. We assume zero delay in the selection of a Serving gNB from $\mathcal{L}_{n,t}$ - hence, there are no data plane interruptions until and unless all gNBs in $\mathcal{L}_{n,t}$ get blocked.

B. Handover Process

The gNB status depends upon whether the link between the gNB and the UE is blocked or unblocked. Until a gNB-UE link becomes unblocked, the gNB cannot be discovered by the UE. Even after a gNB-UE link gets unblocked, it remains undiscovered until the UE discovers the gNB through physical layer procedures, such as a cell search and measurement reports. We disregard the gNB discovery time, as the discovery procedure for new gNBs can occur in the background if a UE is still associated with other discovered gNBs. A discovered gNB is a *candidate* for association. The association procedure or the association handover delay (in case one gNB from $\mathcal{K}_{n,t}$

gets blocked, and another gNB from $\mathcal{C}_{n,t}$ is chosen to replace it) takes up to Δ_K ms.

The induction of a gNB from $\mathcal{K}_{n,t}$ to $\mathcal{L}_{n,t}$ incurs an additional handover delay of Δ_L, which is the delay incurred in fetching the UE data so that it is available for immediate delivery. This transition also involves selection, and is of particular interest to us because it determines the set of gNBs where the UE's data will be replicated. Finally, the scheduling agent picks one gNB from $\mathcal{L}_{n,t}$ to be the Serving gNB. The Serving gNB can change either due to necessity, i.e. if the current Serving gNB gets blocked and the scheduling agent is forced to switch to another gNB, or due to choice, i.e. if the scheduling agent decides that switching to another Serving gNB is the optimal action according to its scheduling policy.

Consider the following blockage scenarios, and how they translate to data plane interruptions at the UE:

- *Serving gNB gets blocked:* Instantaneous switching occurs to another gNB in $\mathcal{L}_{n,t}$. No handover delay is incurred nor is there any data plane interruption.
- *non-Serving gNB in $\mathcal{L}_{n,t}$ gets blocked:* The gNB is immediately dropped from $\mathcal{C}_{n,t}$, $\mathcal{K}_{n,t}$ and $\mathcal{L}_{n,t}$. After a handover delay of Δ_L ms, a new gNB from $\mathcal{K}_{n,t}$ is added to $\mathcal{L}_{n,t}$. Similarly, to replace the blocked gNB, a new gNB from $\mathcal{C}_{n,t}$ is added to $\mathcal{K}_{n,t}$ after a handover delay of Δ_K ms. However, these handovers occur in the background and do not interrupt the UE data plane as long as there is still one unblocked gNB available in $\mathcal{L}_{n,t}$.
- *gNB in $\mathcal{K}_{n,t}$ gets blocked:* The gNB is immediately dropped from $\mathcal{C}_{n,t}$ and $\mathcal{K}_{n,t}$. After a handover delay of Δ_K ms, a new gNB from $\mathcal{C}_{n,t}$ is added to $\mathcal{K}_{n,t}$. There is no UE data plane interruption.
- *All gNBs in $\mathcal{L}_{n,t}$ get blocked concurrently:* UE experiences a maximum data plane interruption of Δ_L ms, the time needed for gNBs from $\mathcal{K}_{n,t}$ to be added to $\mathcal{L}_{n,t}$.
- *All gNBs in $\mathcal{K}_{n,t}$ get blocked concurrently:* UE experiences a maximum data plane interruption of $(\Delta_K + \Delta_L)$ ms, while new gNBs from $\mathcal{C}_{n,t}$ are chosen for $\mathcal{K}_{n,t}$, and $\mathcal{L}_{n,t}$ is chosen from the new $\mathcal{K}_{n,t}$.

Thus, the UE will be out-of-service, and hence experience data plane interruption, in the following scenarios: 1) UE is out of coverage or completely blocked from all of the gNBs in its coverage region, i.e., $\mathcal{C}_{n,t} = \emptyset$ and 2) all the gNBs in $\mathcal{L}_{n,t}$ get blocked, and an unblocked gNB from $\mathcal{K}_{n,t}$ is not added promptly enough due to handover execution times to prevent a period of blockage.

Of course, the degree of association, K, and the degree of replication, L, are two important parameters that influence the extent to which the UE is shielded from data plane interruptions in case of blockages. Associating with, and replicating the data, at a larger number of gNBs results in significantly larger overhead costs. We explore this trade-off between better performance and larger overhead to determine the optimal choice of K and L.

Scenario	PLE	Shadow Fading Std Dev (dB)
LOS	2	4.0
NLOS	3.2	7.0

TABLE I: PLEs and Shadow Fading Standard Deviations for UMi scenario

IV. System Model

The inherent randomness of the environment is captured by two important parts of the model: the channel state model which models the mmWave links, and the UE traffic model which models the statistics of the arrival processes at the UEs and the parameters of the associated XR traffic.

A. Channel Model

The mmWave channel is modeled according to the broadband statistical spatial channel model (SSCM) [16] developed by NYU and used in NYUSIM. A spatial consistency procedure developed by NYU is also implemented to provide spatially correlated LOS/NLOS probabilities [17]. The path loss exponent (PLE) and shadow fading standard deviation values for Urban Microcellular (UMi) scenario are displayed in Table I [18].

Moreover, the NYU squared model [19] for LOS probability is applied for the UMi scenario, which is given by:

$$Pr_{LOS}(d) = \left(\min\left(\frac{d_1}{d}, 1\right)\left(1 - e^{\frac{-d}{d_2}}\right)\right) + \left(e^{-\frac{d}{d_2}}\right)^2 \quad (2)$$

where $d_1 = 22m$ and $d_2 = 100m$.

1) Spatial Consistency Procedure: The close-in free space reference distance (CI) path loss model with a 1 m reference distance used in NYUSIM is a drop-based channel model. In a drop, the drop-based channel model generates a static and independent channel impulse response (CIR) at a particular transmitter-receiver (T-R) separation distance. However, there is no correlation between different drops. The shortcoming of a drop-based channel model is that it generates independent channel coefficients for different distances, even if these points are close to each other. To realize spatial consistency while calculating path loss, spatially-correlated LOS/NLOS conditions are generated [17]. By generating a map of spatially correlated LOS/NLOS conditions, similar shadow fading values are observed at closely spaced locations, which is a more accurate representation of reality than independent values for close locations used in the drop-based model.

B. Dynamic Blockage Model

Dynamic blockages in mmWave cellular networks are extensively studied in [20], [21] assuming a homogeneous Poison Point Process (PPP) with dynamic blocker density λ_B in the disc $B(o, R)$. The blocker arrival rate, or blockage rate, α_i at the i^{th} gNB-UE link is considered Poisson and was derived in [20], [21] as

$$\alpha_i = \Theta r_i, \quad i = 1, 2, ..., m, \quad (3)$$

where r_i is the 2D distance, ignoring height, between the i^{th} gNB-UE pair.

Θ is proportional to the blocker density γ_B and is given by

$$\Theta = \frac{2}{\pi}\gamma_B V \frac{h_B - h_R}{h_T - h_R}, \tag{4}$$

where V is the speed of the blocker and h_B, h_T and h_R are the heights of the blocker, the transmitter and the receiver, respectively.

We model the blocker arrival process as Poisson with parameter α_i blockers/sec. Note that there can be more than one blocker simultaneously blocking the link. Furthermore, we assume the blockage duration of a single blocker is exponentially distributed with parameter μ.The blocking event of a gNB-UE link follows an on-off process with α_i and μ as blocking and unblocking rates, respectively. In the event of a blockage, the Received Signal Strength Indicator (RSSI) of the gNB-UE link is zero, and hence the corresponding channel capacity is also zero. When there is no blockage, the NYUSIM channel model described earlier is used to calculate the path loss and, hence, the channel capacity.

C. Traffic Model

The traffic model we assume for this study is based on the 3GPP XR (Extended Reality) traffic models proposed in [2]. Specifically, we use a generic single-stream downlink model that can be used for VR, AR and CG applications. The downlink traffic is modelled as a sequence of video frames arriving periodically at the base station (gNB) according to a specified video frame rate. Random jitter, which follows a truncated Gaussian distribution, is super-imposed on the periodic arrivals to get the actual arrival time of the frames at the gNB. The size of each frame is also random according to a truncated Gaussian distribution.

Each traffic flow of a UE is assigned a specific traffic type: VR, AR or CG. The traffic type of the flow determines the underlying parameters for the distributions governing the frame size, jitter and packet delay budget of the flows. Each flow consists of a sequence of frames, and each frame is further broken up into IP packets of 1500 bytes for delivery. IP packets belonging to the same frame have the same delay budget, and arrive at the gNB simultaneously. Each UE has a separate buffer at the gNB, so traffic from different UEs do not share a buffer. This means a UE flow cannot experience head-of-line (HOL) blocking from another UE's flow.

1) Frame Size: Given R, the data rate of the flow in Mbps, and F, the frame generation rate of the flow in frames per second (fps), the frame size is modelled as a random variable following a truncated Gaussian distribution with the statistical parameters given in Table II [2].

2) Frame Arrival: The frame arrival rate is determined by the frame rate, F, which is given in frames per second. Hence, inter-arrival time for the frames is given by the inverse of the frame rate. The periodic frame arrivals implicitly assumes fixed delay contributed by the network. However, in a real system, the varying processing and transit delays introduces jitter in frame arrival times at the gNB. In this model, the jitter is modelled as a random variable which is added on top of the periodic arrivals. Thus, the jitter follows a truncated Gaussian distribution with zero mean, 2 ms standard deviation and a truncation range of $[-4, 4]$ ms [2].

Parameter	Unit	Baseline Values
Mean: M	byte	$(R \times 10^6)/(F/8)$
STD	byte	10.5% of M
Max	byte	150% of M
Min	byte	50% of M

TABLE II: Statistical Parameters for Frame Size

Parameter	VR	AR	CG
Data Rate (Mbps)	45	45	30
Frame Rate (fps)	60	60	60
Frame Delay Budget (ms)	10	10	15

TABLE III: Traffic Parameters for VR, AR and CG traffic

The given parameter values and frame generation rates ensure that the frame arrivals are always in order, i.e. the arrival time of the next frame is always later than that of the previous frame. The periodic arrival with jitter, therefore, gives the arrival time for frame with index $k(= 1, 2, 3, ...)$ as:

$$T[k|\text{with jitter}] = \frac{k \times 1000}{F} + J \text{ ms},$$

where J is a random variable capturing the jitter. Note that the actual arrival times of traffic for each UE could be shifted by a UE specific arbitrary offset.

3) Frame Delay Budget (FDB): The latency requirement of XR traffic in the air interface is modeled as a limited time budget for a frame to be transmitted over the air from a gNB to a UE. The delay a frame incurs in the air interface is measured from the time that the frame arrives at the gNB to the time that it is successfully, *fully* transferred to the UE.

If a frame exceeds its FDB, it is considered to have *expired* and is no longer useful owing to the time-sensitive nature of XR applications. Hence, expired frames are immediately dropped and counted as a failed delivery. A partially delivered frame which expires is also considered a failure. If a frame is fully delivered within its FDB, it is is said to be successfully delivered. The value of the FDB varies for different applications (see Table III).

4) Traffic Type Parameters: XR traffic can be broadly classified into three main categories, each with its own set of parameters governing the data rate, frame rate and FDB: VR, AR and CG. The parameters for these various XR applications, according to 3GPP specifications [2], are specified in Table III.

Parameters	Values
Carrier Frequency, f	73 GHz
Max Spectral Efficiency, ρ_{max}	4.8 bps/Hz [24]
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, h_T	5 m
Expected Blockage Duration	500 ms [25]

TABLE IV: Simulation Parameters

D. Mobility Model

The user mobility is modeled by a Random Waypoint model [22]. The UEs are initially dropped uniformly into an area around the gNBs. Each UE then randomly selects a destination within the grid and moves towards it with a constant velocity uniformly distributed between 0 and 3 kmph [23]. Upon reaching its destination, a UE selects a new destination.

V. Simulation Results and Discussion

We do comprehensive performance evaluation by simulating the mmWave network using Python. 11 gNBs are deployed in a hexagonal grid with an inter-site distance of 100 m and 35 UEs are dropped randomly into the area. We use a connectivity threshold of 300 m, i.e. if a UE is within 300 m of a gNB and not blocked, the gNB is considered to be a *candidate* gNB. The gNB density is sufficiently high, such that in case of blockages, a UE always has other candidate gNBs to switch to. An outage is defined as an event when all gNBs in $\mathcal{L}_{n,t}$ are concurrently blocked - this will lead to an interruption of the data plane while the UE initiates a switch to other available gNBs. In order to mimic a system that is not capacity-limited, we use a per-gNB bandwidth of 400 MHz. Additionally, the system operates in discrete time slots of $125\mu s$, which is equivalent to an OFDM slot that can be used for transmitting downlink or uplink data [26]. Traffic arrivals, scheduling decisions, and blockages operate at this granularity. However, channel state updates are done at a larger time scale, once every second, because the path-loss is only affected by large-scale shadow fading, a change in which occurs on the order of seconds [4]. We simulate downlink XR traffic for the UEs and evaluate the performance for varying degrees of association (K), degrees of data replication (L), dynamic blocker densities (γ_B), and handover delays (Δ_K and Δ_L). Since XR traffic requires low latency and expires after a strict deadline, we use the percentage of frames delivered within the deadline as our primary performance metric. This captures the system performance better than other metrics such as average throughput because it explicitly takes into account only the successful traffic which was delivered within the deadline. We perform our simulation over a mobility period of 15 minutes. The rest of the simulation parameters are presented in Table IV.

For selection of $\mathcal{L}_{n,t}$ from $\mathcal{K}_{n,t}$ for each UE n, we use the Best Channel Quality Indicator (BEST-CQI) algorithm where $\mathcal{L}_{n,t}$ is selected based on channel quality alone. BEST-CQI is an algorithm which selects $\mathcal{L}_{n,t}$ by starting with an ordered set of SNRs and picking the gNBs corresponding to the L highest SNR values. It can certainly be argued that a more intelligent selection of $\mathcal{L}_{n,t}$ can be made by taking into account other factors including traffic information, gNB loads, and UE connectivity. However, designing a better selection algorithm for $\mathcal{L}_{n,t}$ is beyond the scope of this paper and is an open question that is left for future research.

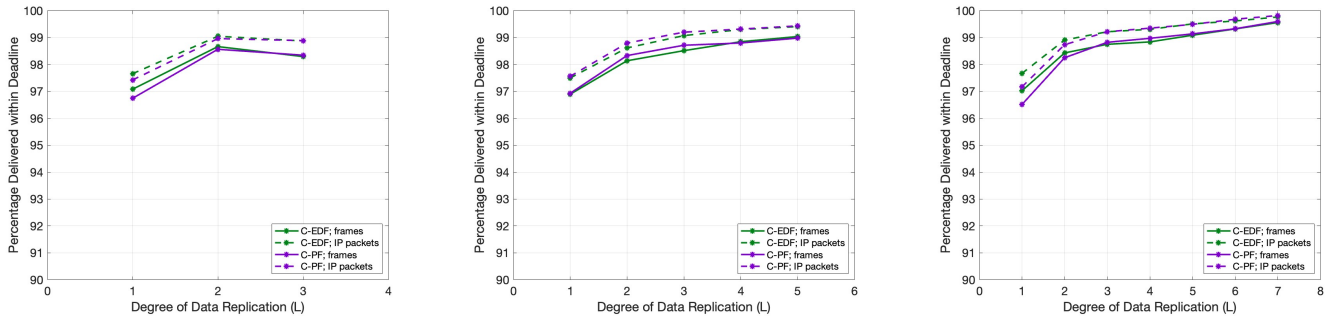
In a multi-connectivity setting, it is not sufficient to just select UEs for scheduling based on some priority value. Once a UE is selected, another selection decision needs to be made to match it to a gNB because multiple gNBs are available to each UE for data transmission. A centralized scheduler would enhance the system performance, at the cost of much higher overhead in terms of information exchange and delays in relaying the control decision. For the purpose of our simulation, we assume an omniscient, centralized scheduler that is able to operate with zero delay. We compare the performance of two centralized schedulers:

- Centralized Earliest Deadline First (C-EDF) : The UE which has the HOL frame with the earliest deadline in the network is matched to the best available gNB in $\mathcal{L}_{n,t}$.
- Centralized Proportional Fair (C-PF): The UE priority function is given by [27]:

$$P = \frac{T}{R}$$

where T is the current channel capacity of the UE-gNB link, and R is the historical average data rate of the UE. The UE with highest priority is matched to the best available gNB in $\mathcal{L}_{n,t}$.

1) Effect of Degree of Association (K) and Replication (L): Fig. 2 shows how the percentage of frames and IP packets delivered successfully within their deadline varies with the degree of data replication (L), for different values of the degree of association (K). First, note that the percentage of IP packets delivered within the deadline is always more than the frames delivered within the deadline, which is to be expected because frame delivery is only counted as successful if the *entire* frame is delivered successfully within the deadline. This shows why the percentage of frames delivered within the deadline is a better QoS metric for deadline-driven XR applications because it only counts the useful throughput. Next, from Fig. 2 we observe that there is a huge spike in performance when we go from single connectivity ($L = 1$) to dual connectivity ($L = 2$). The availability of an extra gNB in dual connectivity ensures that the scheduler has a backup to fall back on in case of sudden service disruption due to blockages. As we further increase the degree of data replication from $L = 2$ to $L = 5$, we see diminishing returns in terms of performance improvement. This is due to the fact that the extra backup



(a) $K = 3$ (b) $K = 5$ (c) $K = 7$

Fig. 2: Effect of the degree of association (K) and the degree of data replication (L) on the percentage of frames and IP packets successfully delivered within deadline, with $L \leq K$, Δ_K = 20 ms, Δ_L = 10 ms and blocker density γ_B = 0.01 bl/m^2

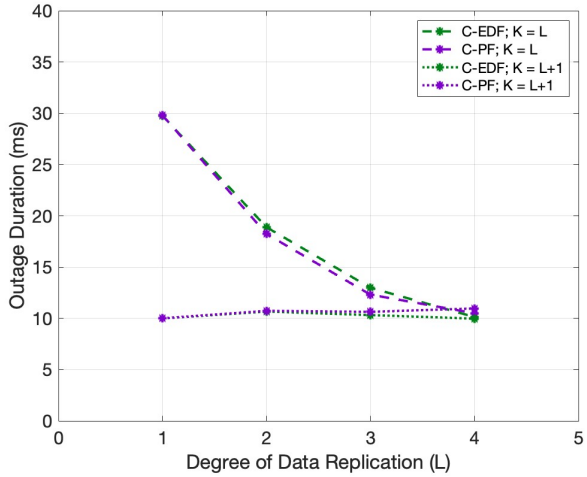


Fig. 3: Effect of the relative values of K and L on the average outage duration , with γ_B = 0.05 bl/m^2, Δ_K = 20 ms and Δ_L = 10 ms

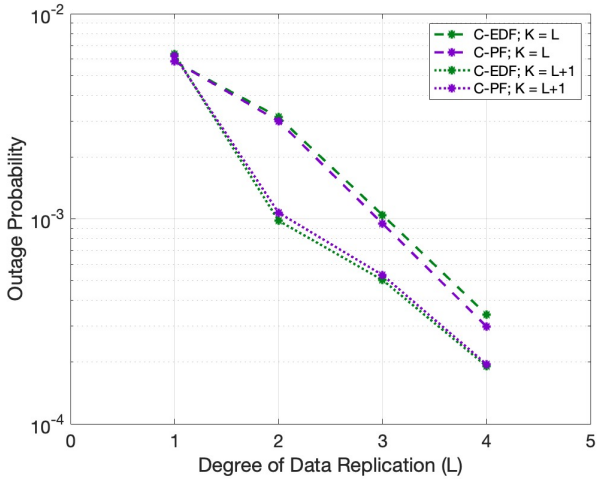


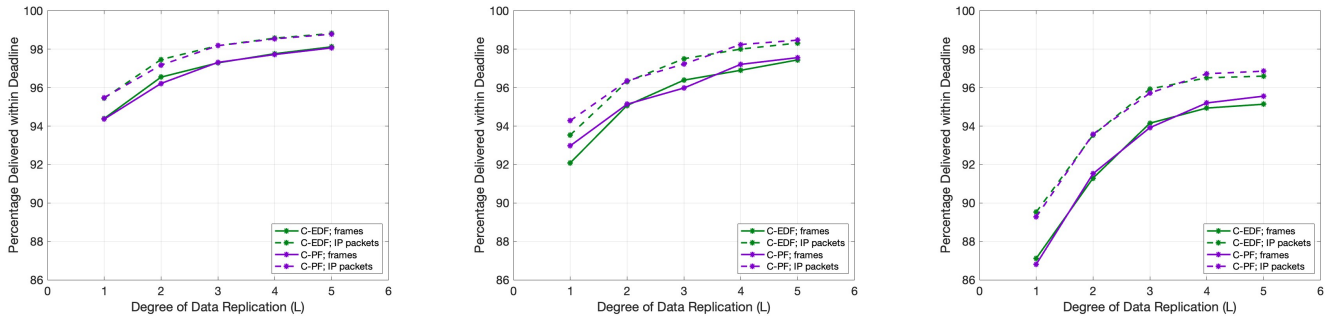
Fig. 4: Effect of the relative values of K and L on the outage probability, with γ_B = 0.05 bl/m^2, Δ_K = 20 ms and Δ_L = 10 ms

gNBs only become useful when there are several concurrent blockages. For example, when $L = 4$, the fourth gNB will only be useful in the scenario when the first three gNBs are concurrently blocked. Since the outage probability decreases exponentially with the number of gNBs, as shown in Fig.4, we see corresponding diminishing returns as L increases.

From Fig. 2, we note that with $K = 3$, there is a dip in performance going from $L = 2$ to $L = 3$. However, this is well within the confidence intervals ($\pm 0.23\%$) and the broader trend of performance increasing with multi-connectivity remains true. In fact, from Fig. 2c we can see that we boost performance from 96.5% when $L = 1$ to 99% when $L = 5$. This is a significant improvement in performance given the fact that one of the main QoS criteria for XR applications is to deliver 99% of a UE's traffic within the deadline [2]. Moreover, we see that the prime benefit of increasing K is that it allows us to potentially replicate the data at a larger number of gNBs, since $L \leq K$. However, if we fix L, there is no benefit to be gained in further increasing K beyond $K = L + 1$. For example, with $L = 2$, we see similar performance, disregarding the minor variations which are within the confidence intervals, as K is increased from 3 to 7.

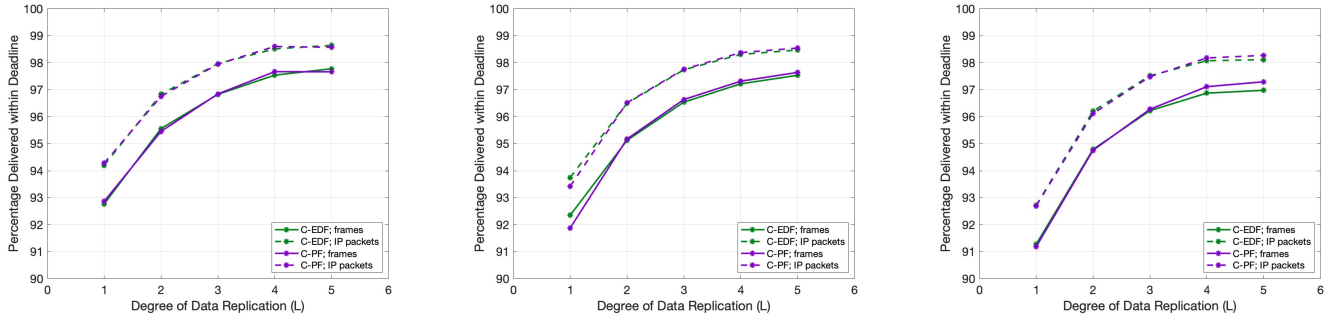
We now turn our attention towards a discussion and comparison of the performance of our two schedulers: C-EDF and C-PF. The decision to use a centralized scheduler is a deliberate one, and stems from our multi-connectivity architecture where the *selection* of a Serving gNB plays a critical role in the subsequent scheduling decision and system performance. Hence, the scheduling problem is fundamentally different from single-connectivity scenarios, where the only decision that needs to be made is the scheduling decision. Thus, it is imperative that the *selection* and *scheduling* decisions be made jointly in order to gain better performance. Even so, neither C-EDF nor C-PF is optimal. Simple examples can be crafted that show both schedulers taking sub-optimal decisions.

Moreover, we acknowledge that our schedulers operate



(a) γ_B = 0.03 bl/m^2 (b) γ_B = 0.05 bl/m^2 (c) γ_B = 0.1 bl/m^2

Fig. 5: Effect of the dynamic blocker density (γ_B) on the percentage of frames and IP packets successfully delivered within deadline, with $K = 5$, Δ_K = 20 ms and Δ_L = 10 ms



(a) Δ_K = 20 ms, Δ_L = 10 ms (b) Δ_K = 50 ms, Δ_L = 30 ms (c) Δ_K = 80 ms , Δ_L = 50 ms

Fig. 6: Effect of Association Handover Delay, Δ_K, and Data Replication Handover Delay, Δ_L, on the percentage of frames and IP packets successfully delivered within deadline, with $K = 5$ and γ_B = 0.05 bl/m^2

under ideal assumptions that will not hold in real-world scenarios, namely the availability of instantaneous channel state and traffic information at the scheduler and the instantaneous relaying and execution of the scheduling decision at the gNBs. However, our results can be used to gauge the performance of schedulers that better emulate real-world conditions and operate in a distributed manner.

From Figs. 2-6, we see that both C-EDF and C-PF have similar performance, with C-PF performing better at higher blocker densities. C-PF performs well because it jointly optimizes over the UE's historical data rate and the available channel capacities; however, its drawback is that it does not explicitly take into account the traffic deadlines nor does it attempt to do delay-aware scheduling. On the other hand, C-EDF attempts delay-aware scheduling but does not take a joint gNB selection and scheduling decision; instead, it does scheduling and selection sequentially which is sub-optimal. Hence, we can see that there is a need for new scheduling algorithms that are optimized for the multi-connectivity paradigm, i.e, which do deadline-driven scheduling in conjunction with gNB selection.

Next, we illustrate how our architecture minimizes data plane interruptions. We are interested in the average outage duration, which is the amount of time it takes a UE to recover from an outage event by resuming the data plane connection with another gNB. At 60 fps, the average frame inter-arrival time is 17 ms, so depending on the link capacity available after the interruption, at most one frame is dropped when $\Delta_K = 20$ ms and $\Delta_L = 10$ ms. From Fig. 3, we note that when $K = L$, which is the case when the Association and Data Replication tiers are collapsed into one i.e. data is replicated at all the associated gNBs, the average outage duration is upper-bounded by $(\Delta_K + \Delta_L)$ ms. However, the power of our multi-tier architecture is displayed when $K > L$. Consider the simplest case, when $K = L + 1$. With one extra gNB in $\mathcal{K}_{n,t}$, the average outage duration falls to approximately Δ_L ms. Fig. 3 also shows that this benefit does not increase with L because the response time to the outage is determined by whether an extra gNB is available in $\mathcal{K}_{n,t}$ when all gNBs in $\mathcal{L}_{n,t}$ get blocked. However, from Fig.4, we observe that increasing L decreases the outage *probability*. Thus, from Figs. 3 and 4 we can conclude that for the same value of L, $K = L + 1$ gives better performance than $K = L$, if this option is available.

2) Effect of Dynamic Blocker Density γ_B: Fig. 5 illustrates the effect of dynamic blocker density (γ_B) on the percentage of frames and IP packets delivered within the deadline. We observe that a higher blocker density results in a significant loss of performance, especially at low levels of multi-connectivity. Moreover, as the blocker density is increased the boost in

performance from a higher degree of data replication also increases. This is because a higher blocker density results in more frequent blockages, which is reflected in a higher out-of-service probability. Consequently, the benefit to be gained by having backup gNBs also increases as the density of blockers is increased.

3) Effect of Handover Delays, Δ_K and Δ_L: Handover Delays, Δ_K and Δ_L, are vital for performance evaluation because they affect the response time to blockages and determine the duration of data plane interruptions. Recall that, given the gNB density is high enough to ensure that there are always candidate gNBs available, a UE experiences a maximum data plane interruption of Δ_L ms if all the gNBs in $\mathcal{L}_{n,t}$ are blocked concurrently, and a maximum data plane interruption of $(\Delta_K + \Delta_L)$ ms if all the gNBs in $\mathcal{K}_{n,t}$ are blocked concurrently. From Fig. 6, we see that the system performance decreases as Δ_K and Δ_L are increased, which is due to the higher out-of-service durations as a result of higher handover delays. However, this decrease in performance is less at higher values of L, because the out-of-service probability decreases exponentially as L is increased. For example, from Fig.6a and 6c we observe that at $L = 1$, performance decreases from 92.8% to 91.2% - a decrease of 1.6% - when handover delays increase. However, at $L = 2$, the performance decreases from 95.5% to 94.8% - a smaller decrease of 0.7%.

VI. Conclusion

XR applications are expected to become a significant feature of next-generation wireless networks, and existing network architectures and protocols need to evolve to support their strict QoS requirements. mmWave studies often focus on providing connectivity and coverage analysis, and minimizing the effect of blockages from a network connectivity perspective. In this paper, we tackle both issues by first formulating a network architecture designed to minimize data plane interruptions due to blockages and then explicitly evaluating the performance of XR applications over this multi-connectivity enabled network. Our work also highlights several core modules of the network architecture which can be optimized to further improve performance, including intelligent selection of Data Replication gNBs that takes into account UE mobility predictions and traffic demands, and a delay-aware scheduler that is specifically optimized for the multi-connectivity setting.

References

[1] G. Minopoulos and K. E. Psannis, "Opportunities and challenges of tangible XR applications for 5G networks and beyond," *IEEE Consumer Electronics Magazine*, 2022.

[2] 3GPP, "Study on XR (Extended Reality) Evaluations for NR," Technical Specification Group Radio Access Network, Technical Specification (TS) 38.838, December 2021, version 17.0.0.

[3] R. Dangi, P. Lalwani, G. Choudhary, I. You, and G. Pau, "Study and investigation on 5G technology: A systematic review," *Sensors*, vol. 22, no. 1, p. 26, 2022.

[4] T. S. Rappaport, *Wireless Communications: Principles and Practice, 2/E*. Pearson Education India, 2010.

[5] S. Sun and T. S. Rappaport, "Wideband mmWave channels: Implications for design and implementation of adaptive beam antennas," in *2014 IEEE MTT-S International Microwave Symposium (IMS2014)*. IEEE, 2014, pp. 1–4.

[6] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proceedings of the IEEE*, vol. 102, no. 3, pp. 366–385, 2014.

[7] 3GPP, "Overall Description," 3GPP Std., Technical Specification (TS) 36.300, March 2015, v12.5.0.

[8] M. Özkoç, A. Koutsaftis, R. Kumar, P. Liu, and S. Panwar, "The impact of multi-connectivity and handover constraints on millimeter wave and terahertz cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 6, pp. 1833–1853, 2021.

[9] A. Koutsaftis, R. Kumar, P. Liu, and S. Panwar, "Fast inter-base station ring (FIBR): A new millimeter wave cellular network architecture," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 12, pp. 2699–2714, 2019.

[10] V. Petrov *et al.*, "Dynamic multi-connectivity performance in ultra-dense urban mmwave deployments," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 2038–2055, 2017.

[11] M. Gapeyenko *et al.*, "On the degree of multi-connectivity in 5G millimeter-wave cellular urban deployments," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 1973–1978, 2018.

[12] 3GPP, "Multi-connectivity Stage 2," TSG RAN, Technical Specification (TS) 37.340, March 2020, version 16.1.0.

[13] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Network*, vol. 34, no. 3, pp. 134–142, 2019.

[14] J. Sundararajan *et al.*, "Performance evaluation of extended reality applications in 5G NR system," in *2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 2021, pp. 1–7.

[15] V. Petrov *et al.*, "Extended reality (XR) over 5G and 5G-advanced new radio: Standardization, applications, and trends," *arXiv preprint arXiv:2203.02242*, 2022.

[16] S. Sun, G. R. MacCartney, and T. S. Rappaport, "A novel millimeter-wave channel simulator and applications for 5G wireless communications," in *2017 IEEE International Conference on Communications (ICC)*. IEEE, 2017, pp. 1–7.

[17] S. Ju, O. Kanhere, Y. Xing, and T. S. Rappaport, "A millimeter-wave channel simulator NYUSIM with spatial consistency and human blockage," in *2019 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2019, pp. 1–6.

[18] S. Sun *et al.*, "Investigation of prediction accuracy, sensitivity, and parameter stability of large-scale propagation path loss models for 5G wireless communications," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 5, pp. 2843–2860, 2016.

[19] T. S. Rappaport *et al.*, "Overview of millimeter wave communications for fifth-generation (5G) wireless networks—with a focus on propagation models," *IEEE Transactions on Antennas and Propagation*, vol. 65, no. 12, pp. 6213–6230, 2017.

[20] I. K. Jain, R. Kumar, and S. Panwar, "Driven by capacity or blockage? a millimeter wave blockage analysis," in *2018 30th International Teletraffic Congress (ITC 30)*, vol. 1. IEEE, 2018, pp. 153–159.

[21] I. K. Jain, R. Kumar, and S. S. Panwar, "The impact of mobile blockers on millimeter wave cellular systems," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 4, pp. 854–868, 2019.

[22] C. Bettstetter, H. Hartenstein, and X. Pérez-Costa, "Stochastic properties of the random waypoint mobility model," *Wireless Networks*, vol. 10, pp. 555–567, 2004.

[23] 3GPP, "5G - Study on channel model for frequencies from 0.5 to 100 GHz," 3GPP Std., Technical Report (TR) 38.901, January 2018, version 14.3.0.

[24] M. Akdeniz *et al.*, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1164–1179, 2014.

[25] G. MacCartney, T. Rappaport, and S. Rangan, "Rapid fading due to human blockage in pedestrian crowds at 5G millimeter-wave frequencies," in *GLOBECOM 2017-2017 IEEE Global Communications Conference*. IEEE, 2017, pp. 1–7.

[26] M.Mezzavilla *et al.*, "5G mmWave module for the ns-3 network simulator," in *Proceedings of the 18th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, 2015, pp. 283–290.

[27] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.