Balancing Latency and Energy Efficiency in mmWave 5G NR Systems with Multiconnectivity

Vitalii Beschastnyi, Darya Ostrikova, Dmitri Moltchanov, Yuliya Gaidamaka,
Yevgeni Koucheryavy, and Konstantin Samouylov

***Abstract*—Multiconnectivity is a vital option in 5G New Radio (NR) systems allowing user equipment (UE) to maintain multiple links to nearby base stations (BS) improving service reliability. However, this functionality is very power-hungry prohibiting its application in practice. To alleviate this shortcoming discontinuous reception (DRX) mechanism can be utilized. The latter requires careful parameterization as it may drastically increase latency at the air interface. In this paper, we develop a mathematical model to characterize the trade-off between energy efficiency and latency in millimeter wave (mmWave) 5G NR systems under micromobility and blockage impairments. We then utilize it to determine the optimal type of DRX timers scaling. We show that micromobility has a positive impact on energy efficiency. For low micromobility speeds ($< 0.2^\circ$/s) proportional DRX scaling scheme with the scaling coefficient $k < 1.0$, provides the best performance in terms of considered metrics while for higher speeds it leads to a compromise between them. The optimal value of k depends on design preferences.**

***Index Terms*—5G, New Radio, millimeter wave, energy efficiency, latency, micromobility, multiconnectivity.**

I. Introduction

Nowadays, as the standardization of 5G New Radio (NR) technology is over, operators start to deploy these systems worldwide. However, their roll-out in the millimeter-wave (mmWave) band, promising to achieve the target bitrates set by ITU for IMT-2020 systems, is hampered by challenging propagation phenomena including line-of-sight (LoS) blockage [1] and micromobility [2]. These effects lead to frequent outages [3] and, as a result, to degraded service quality.

To alleviate the abovementioned challenges 3GPP has proposed a multiconnectivity functionality [4]. According to it, user equipment (UE) is allowed to maintain multiple links to nearby base stations (BS) and switch between them to avoid outages. While this option is shown to improve outage performance of mmWave 5G NR systems [3], [5], it is extremely power-hungry as it is required to spend energy for maintaining both active and backup links.

5G NR defines several techniques to conserve energy, such as bandwidth adaptation [6], cross-slot scheduling [7], discontinuous reception (DRX) [8] and others. Out of these options, DRX is the most comprehensive UE side mechanism that has been tested in 4G LTE and appropriately extended for NR. DRX defines a number of sleep timers that need to be optimized to improve the energy efficiency of UEs. In case of multiconnectivity, there is an option to specify timers separately for active and backup links to further improve UE energy efficiency. However, when these timers are not properly configured UE may experience excessive latency at the air interface, especially, in case of blockage and micromobility that result in a change of the currently active link. Thus, there exists an inherent trade-off between latency and energy efficiency in mmWave 5G systems.

Performance improvements of multiconnectivity in terms of outage and capacity under dynamic blockage and micromobility in mmWave 5G NR have been well documented [3], [9]. There are also link-level studies addressing performance of DRX scheme [8], [10]. However, there is a lack of system-level models capturing DRX performance with multiconnectivity in presence of mmWave-specific propagation phenomena. Furthermore, latency in presence of multiconnectivity and DRX has not been addressed so far leaving the question of the trade-off between energy efficiency and latency open.

In this paper, we formulate a mathematical model capturing DRX operation in 5G NR systems with multiconnectivity operation under both dynamic blockage and micromobility impairments. We utilize this model to investigate trade-offs between latency and energy efficiency as a function of the system and environmental parameters. Finally, we study the effect of DRX timer scaling. The main contributions are:

- a mathematical model capturing specifics of DRX and multiconnectivity schemes under both dynamic blockage and micromobility impairments;
- for applications with low micromobility speed ($< 0.2\ ^\circ$/s) proportional DRX scaling schemes maximizes both considered metrics, while for those characterized by higher speeds this scheme provides a trade-off between them.

This paper is organized as follows. In Section II we introduce our system model. The mathematical model is formulated and solved in Section III. Numerical results are provided in Section IV. Conclusions are drawn in the last section.

II. System Model

1) Radio Specifics: We consider mmWave 5G NR BSs deployed according to the Poisson point process (PPP) in $\Re^2$ with density λ_A and concentrate on a randomly located UE, see Fig. 1. The height of UE and BS are h_A and h_U, respectively.

The density of pedestrians acting as blockers is λ_B. They are modeled by cylinders with height and radius h_B and r_B,

D. Moltchanov and Y. Koucheryavy are with Tampere University, Tampere, Finland. Email: {dmitri.moltchanov, evgeni.kucheryavy}@tuni.fi

V. Beschastnyi, D. Ostrikova, Yu. Gaidamaka, and K. Samouylov are with Peoples' Friendship University of Russia (RUDN University), Moscow, Russia. Email: {ostrikova-dyu, beschastnyy-va, gaydamaka-yuv, samuylov-ke}@rudn.ru

Yu. Gaidamaka and K. Samouylov are also with Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Russia.

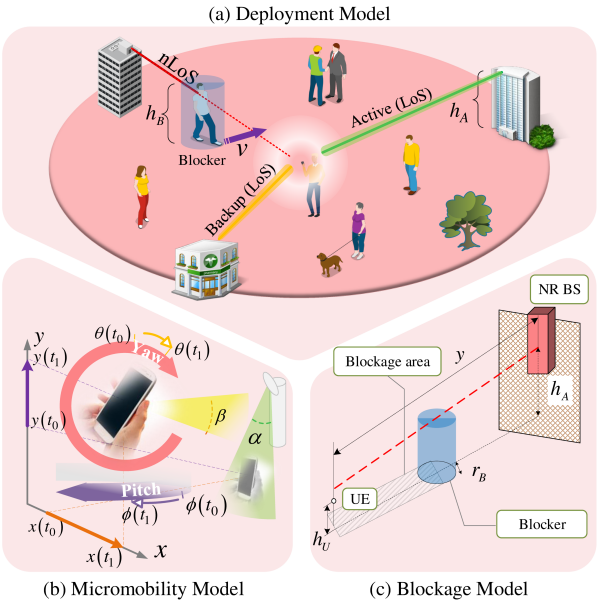


Fig. 1. The considered scenario and main components of the system model.

respectively. Pedestrians are assumed to move according to the random direction model (RDM). According to it, a direction is first chosen randomly and uniformly in $(0, 2\pi)$ and then a blocker moves in the chosen direction at constant speed v for exponentially distributed time with mean $1/\mu_B$.

To capture propagation losses in mmWave band we utilize Urban-Micro (UMi) Street-Canyon model [11]

$$L_{dB}(y) = 32.4 + \zeta 10\log y + 20\log f_c, \tag{1}$$

where f_c is the carrier frequency (GHz), y is the distance (m), ζ is the path loss coefficient. By using (1) the signal-to-interference plus noise (SINR) at UE takes the form

$$S(y) = P_T G_{T,A} G_{T,U} \left[\frac{y^{-\zeta}}{N_0 + M_I}\right], \tag{2}$$

where P_T is the emitted power, $G_{T,A}$ and $G_{T,U}$ are the BS and UE antenna gains, N_0 is the thermal noise, M_I is the shadow fading and interference margin.

We assume planar arrays at both BS and UE sides. To represent their radiation patterns, we utilize the cone model [12], where the gain over the main lobe, G, and half-power beamwidth (HPBW), α, are calculated according to [13]

$$\alpha = 2\left|\arccos\left(-\frac{1}{\pi}\right) - \arccos\left(\frac{2.782}{(N_{(\cdot)}\pi)}\right)\right|,$$
$$G = \frac{1}{\theta^{+}_{3db} - \theta^{-}_{3db}} \int_{\theta^{-}_{3db}}^{\theta^{+}_{3db}} \frac{\sin(N_{(\cdot)}\pi\cos(\theta)/2)}{\sin(\pi\cos(\theta)/2)} d\theta, \tag{3}$$

where $N_{(\cdot)}$ is the number of rows or columns in array, while $\theta^{\pm}_{3db} = \arccos[\pm 2.782/(N_{(\cdot)}\pi)]$.

We also account for micromobility impairments, i.e., loss of connection caused by small displacements and angular movements of UE in hands of a user [14], [15]. To this aim, we represent time to outage using an exponential distribution with mean $1/\lambda_{T_A}$, where the mean values are reported in [2].

For beamsearching, we assume the exhaustive search procedure. The time complexity of this procedure is $T_B = N_U N_T \delta$, where δ is the antenna array switching time, N_U and N_T are the numbers of BS and UE antenna configurations.

2) Multiconnecitivity and Traffic: To avoid outages caused by blockage or micromobility, UE supports multiconnectivity operation. The choice of BSs is made upon time-averaged SINR. The number of supported links is limited by N. Only one link is utilized at a time for transmission. This link is referred to as an active link. Once the connection is lost, UE switches to the backup link currently having the highest SINR.

We consider the on/off traffic model representing source fluctuating between active and stand-by states having exponentially distributed times with parameters λ_O and λ_S.

3) DRX Operation: The DRX mode is a power-saving technique that manages the power of UE [8], [16]. Once UE detects that there are no packets for transmission during inactivity timer t_I, the DRX turns UE to a sleep cycle consisting of two periods, sleep and "On Duration". During the latter, UE wakes up regularly for t_{on} ms and checks if there are packets for transmission or reception. If not, UE proceeds to another sleep cycle. Otherwise, UE returns to the active mode to resume communications. To decrease the latency caused by the wake-up procedure, 3GPP has proposed two types of sleep cycles: (i) the short sleep cycle with duration t_{CSS} is applied first, and (ii) the long sleep cycle with a duration t_{CLS} is applied upon the expiry of the specified number of consecutive short cycles, n_{CSS}, without active communications.

In 3GPP Rel. 17, the new "RRC-Inactive" state has been introduced to allow for more flexible power-saving strategies in 5G NR systems. It has the same sleep cycles as in RRC-Connected, but with rather long durations of t_{ISS} and t_{ILS} [17] and the unlimited number of long sleeps n_{ILS}. The UE enters RRC-Inactive long sleep after consecutive n_{CLS} cycles in the RRC-Connected state. Each type of sleep cycle is characterized by corresponding power consumption denoted by the constants e_{CSS}, e_{CLS}, e_{ISS}, and e_{ILS}, while the "On Duration" period takes e_{ON} mW/slot, see Table I.

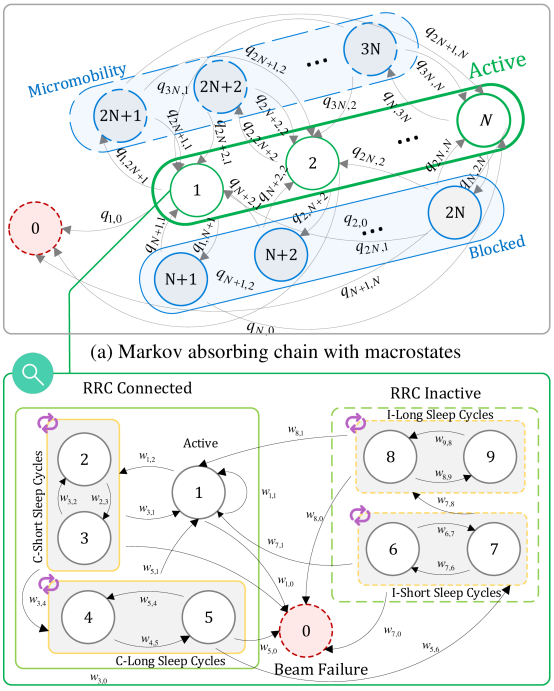
4) Metric and DRX Scaling Strategies: To characterize the trade-off between energy efficiency and UE performance we consider two metrics: (i) latency of data transmission, (ii) energy efficiency. The latter metric is measured in bit/J and expresses how well the channel is utilized.

To exploit the energy-latency trade-off, we specify and evaluate the following strategies to control DRX sleep timers:

- *Proportional.* UE uses the same timers for all N BSs supported via multiconnectivity. This is expected to minimize the latency at the expense of energy efficiency.
- *Linear scaling.* The sleep timers are scaled linearly as a function of SINR of links. The rationale is that the blockage is less likely to occur with BS located nearby implying that UE spends more time connected to them.
- *Exponential scaling.* This is a case of scaling, where timers are scaled up exponentially as a function of SINR. This is intended for aggressive power conservation.

III. Performance Evaluation Framework

The proposed energy consumption model, shown in Fig. 2, is based on the absorbing Markov macrostate model, see Fig.



(a) Markov absorbing chain with macrostates

(b) Markov absorbing chain with microstates

Fig. 2. DRX-enabled communication model.

2(a), and captures the basic properties of NR communications: micromobility (states $2N+1,\ldots,3N$), LoS blockage (states $N+1,\ldots,2N$), multiconnectivity (transitions between states $1,\ldots,N$), and the outage conditions (state 0). The model is a composite one with micro-models embedded into macrostates $1,\ldots,N$ corresponding to the periods of continuous association with i-th BS, see Fig. 2(b). We also assume that from a modeling perspective the beam failure detection and recovery, as well as radio link recovery procedures, are bundled together with RRC-Idle periods, though, generally, these procedures are performed in RRC-Connected/Inactive modes.

Following [18], the probability density function (pdf) of the distance between the UE and the i-th nearest NR BS is

$$f_i(x) = \frac{2(\pi\lambda_A)^i}{(i-1)!} x^{2i-1} e^{-\pi\lambda_A x^2},\ x>0,\ i=1,\ldots,N. \quad (4)$$

By combining the results of [1], [19], the intensity of blockage with i-th BSs is given by

$$\mu_{B,i} = \int_0^\infty f_i(x) \frac{2r_B\lambda_B v\left(x\left[h_B-h_U\right]+r_B\left[h_A-h_U\right]\right)}{\left(h_A-h_U\right)} dx, \quad (5)$$

while the pdf and cumulative distribution function (CDF) of the interval between LoS blockages are $f_{T_{L,i}}(t) = \mu_{B,i}e^{-\mu_{B,i}t}$ and $F_{T_{L,i}}(t) = 1-e^{-\mu_{B,i}t}$, respectively. According to [15], the pdf of the time to beam misalignment due to UE micromobility is defined by the exponential distribution

$$f_{T_A}(t) = \lambda_{T_A} e^{-\lambda_{T_A} t}. \quad (6)$$

Now, we are in the position to derive the pdf of the continuous association with i-th BS as the minimum of the two random values corresponding to time to beam misalignment and time to LoS blockage as

$$f_{C_i}(t) = f_{T_{L,i}}(t)\left[1-F_{T_A}(t)\right] + f_{T_A}(t)\left[1-F_{T_{L,i}}(t)\right]. \quad (7)$$

Then, we embed the DRX model shown in Fig. 2(b) into the macrostates $1,\ldots,N$ of the main model. Here we define $N_{DRX} = 2(n_{CSS}+n_{CLS}+n_{ISS}+1)+1$ states, where state 1 represents active communication with NR BS, other odd states correspond to short and long sleep in RRC-Connected and RRC-Inactive states, while even states stand for "On Duration" periods of the sleep cycles. State 0 reflects the end of the continuous association. These states compose another absorbing Markov chain and enable us to construct the infinitesimal generator $Q_i = \left[q_{jk}\right]$, $j,k=0,\ldots,N_{DRX}$. Denoting $N_{CSS} = n_{CSS}+1$, $N_{CLS} = n_{CSS}+n_{CLS}+1$, $N_{ISS} = n_{CSS}+n_{CLS}+n_{ISS}+1$, $g(T_E,T_C) = \int_{T_C+t_{ON}}^\infty f_{T_E}(x)\,dx$, the transition rates are given by

$$\begin{aligned}
&q_{1,1} = g(O,t_I)\,g(C_i,t_I),\\
&q_{1,2} = (1-g(O,t_I))\,g(C_i,t_I),\\
&q_{1,0} = (1-g(C_i,t_I)),\\
&q_{2j,2j+1} = 1,\ j=1,...,n_{ISS},\\
&q_{2j+1,2j+2} = g(S,t_{CSS})\,g(C_i,t_{CSS}),\ j=1,...,n_{CSS},\\
&q_{2j+1,1} = (1-g(S,t_{CSS}))\,g(C_i,t_{CSS}),\ j=1,...,n_{CSS},\\
&q_{2j+1,0} = 1-g(C_i,t_{CSS}),\ j=1,...,n_{CSS},\\
&q_{2j+1,2j+2} = g(S,t_{CLS})\,g(C_i,t_{CLS}),\ j=N_{CSS},...,N_{CLS},\\
&q_{2j+1,1} = (1-g(S,t_{CLS}))\,g(C_i,t_{CLS}),\ j=N_{CSS},...,N_{CLS},\\
&q_{2j+1,0} = 1-g(C_i,t_{CLS}),\ j=N_{CSS},...,N_{CLS},\\
&q_{2j+1,2j+2} = g(S,t_{ISS})\,g(C_i,t_{ISS}),\ j=N_{CLS},...,N_{ISS},\\
&q_{2j+1,1} = (1-g(S,t_{ISS}))\,g(C_i,t_{ISS}),\ j=N_{CLS},...,N_{ISS},\\
&q_{2j+1,0} = 1-g(C_i,t_{ISS}),\ j=N_{CLS},...,N_{ISS},\\
&q_{2N_{ISS}+1,2N_{ISS}} = g(S,t_{ILS})\,g(C_i,t_{ILS}),\\
&q_{2N_{ISS}+1,1} = (1-g(S,t_{ILS}))\,g(C_i,t_{ILS}),\\
&q_{2N_{ISS}+1,0} = 1-g(C_i,t_{ILS}),\\
&q_{i,j} = 0,\ \text{otherwise.}
\end{aligned} \quad (8)$$

These rates are derived based on finding the minimum of random values corresponding to duration of standby mode and continuous association with i-th BS. Defining the fundamental matrix $D_i = (I-Q_i)^{-1}$ [20], we obtain the mean number of transitions $\tau_{i,j}$, $j=1,\ldots,N_{DRX}$ before absorption as

$$\tau_{i,j} = \sum_{k=1}^{N_{DRX}} d_{kj},\ i=1,\ldots,N,\ j=1,\ldots,N_{DRX}. \quad (9)$$

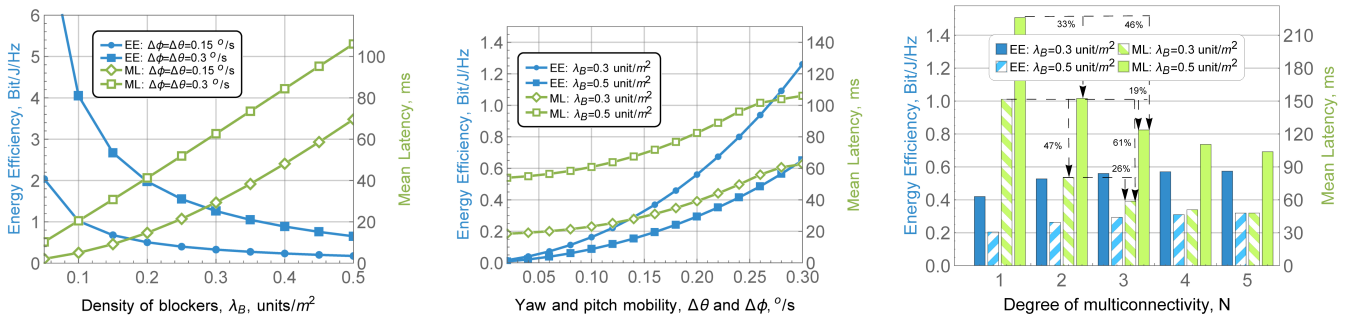
The results in (9) allow evaluating the mean power consumption within a continuous association period as

$$P_{C_i} = \vec{\tau}_i\vec{e}, \quad (10)$$

where vector $\vec{e}$ is composed of energy consumption constants in accordance with states of the chain, see Table I.

Observe that UE upon wake up may find an NR BS in DRX sleep cycle or radio problem detection phase if there is micromobility or LoS blockage. As the actual data transmission may happen in the connected mode only, the BS is prompted to wait until UE wakes up or completes the radio problem recovery inducing latency which is given by (11), where P_i is the probability of association with i-th BS, $\pi_j = \frac{\tau_{i,j}t_j}{\int_0^\infty t\cdot f_{C_i}(t)dt}$ is the probability of being in j-microstate, P_B is the probability of simultaneous blockage at all the N BSs, t_{RLF} is the time of

$$T_W = \frac{1}{2}\sum_{i=1}^{N} P_i \left(\left(1 - \frac{\tau_{i,1}}{\lambda_O \int_0^\infty t \cdot f_{C_i}(t)\, dt} \right) \sum_{j=1}^{(N_{DRX}-1)/2} \pi_j t_{2j} + \sum_{j=1}^{N_{DRX}} \pi_j q_{j,0} t_{BFM,j} \right) + \frac{1}{2} P_B t_{RLF,j}. \quad (11)$$



(a) As a function of blockers' density (b) As a function of micromobility (c) As a function of multiconnectivity

Fig. 3. Energy efficiency (EE) and mean latency (ML) as a function of blockage, micromobility, and the degree of multiconnectivity for proportional scaling

.

radio link failure (RLF) management assumed to last within LoS blockage at all N BSs, $t_{BFM} = T_{BFI}(t_{on} + t_j)$ is the time of beam management procedure, and t_j is the sleep duration in j-th state.

The spectral efficiency, $E[C_{T_L}]$, is calculated as UE capacity multiplied by the fraction of the active communications time. The latter is given by the ratio between $E[T_L]$ and time before outage, $E[T_{NL}] + E[T_L]$. As the wake-up time introduces additional delay, we redefine the active communication time as $E[T_L] - T_W$ allowing to derive the energy efficiency as

$$P_E = \frac{E[T_L] - T_W}{E[T_{NL}] + E[T_L]} P_E[C_{T_L}], \quad (12)$$

where $T_{L,i}$ is the mean continuous association time with i-th NR BS, $P_E[C_{T_L}]$ is the energy efficiency when UE is associated with i-th NR BS given by

$$P_E[C_{T_L}] = \frac{1}{E[T_L]} \sum_{i=1}^{N} \nu_i T_{L,i} E[C_i] P_{C_i}^{-1}, \quad (13)$$

with the mean UE capacity provided by

$$E[C_i] = \int_0^\infty \log_2(1 + S(y)) f_i(y)\, dy. \quad (14)$$

IV. Numerical Analysis

We now elaborate our numerical results by first assessing the effect of blockage and micromobility on latency and energy efficiency and then comparing the performance of the proposed DRX scaling schemes. The system parameters are provided in Table I. To provide the numerical assessment of the model we utilized MATLAB tool with Symbolic Math Toolbox.

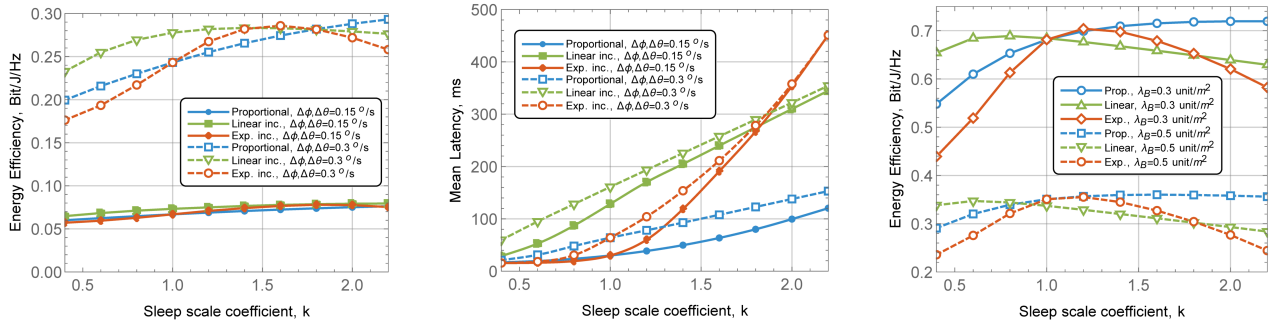
We start with the effects of blockage on latency and energy efficiency trade-off for proportional DRX scheme, shown in Fig. 3(a), where $N = 3$. Here, the latency increases with the blockers' density. The reason is that the increase in λ_B makes the outage events more frequent and longer [1]. The latter impacts energy efficiency as the system now spends more time without active connection.

The effect of micromobility is principally different, see Fig. 3(b), where $N = 3$. Here, energy efficiency still increases with higher micromobility speed as UE experiences more connection losses as a result of antenna misalignment and, thus, needs to spend time searching for a beam. However, the energy efficiency increases as well. The reason is that these misalignments force UE to spend more time at BSs located nearby and characterized by higher spectral efficiency as they are less affected by blockage [21].

Addressing the effect of multiconnectivity shown in Fig. 3(c), where $\Delta\theta = \Delta\phi = 0.1°/s$, one may notice the "cross" effect – latency decreases while the energy efficiency improves as more links are added. Here, both observations are attributed to the less time UE spends in outage conditions. However,

TABLE I
The Default System Parameters

Notation	Description	Values
λ_A	BS density	0.001 units/m^2
r_B, h_B	blocker radius and height	0.4 m, 1.7 m
v	blocker speed	1 m/s
h_U, h_A	UE and BS heights	1.5 m, 4 m
P_T	BS emitted power	2 W
ζ	path loss exponent	2.1
M_I	shadow fading and interference margin	3 dBi
δ	array switching time	2 μs
N_T, N_U	BS/UE antenna array configurations	16x4, 4x4
N	degree of multiconnectivity	3
N_0	thermal noise	-84 dBi
$\Delta x, \Delta y$	x- and y-coordinate micromobility speed	3 cm/s, 3 cm/s
λ_O	intensity of switching to standby mode	2 s^{-1}
λ_S	intensity of switching to operational mode	0.5 s^{-1}
t_I	inactivity timer	4 ms
t_{on}	"On Duration" timer	2 ms
t_{CSS}	C-DRX short sleep timer	8 ms
t_{CLS}	C-DRX long sleep timer	32 ms
t_{ISS}	I-DRX short sleep timer	64 ms
t_{ILS}	I-DRX long sleep timer	256 ms
n_{CSS}	number of short sleep cycles in C-DRX	3
n_{CLS}	number of long sleep cycles C-DRX	1
n_{ISS}	number of short sleep cycles I-DRX	3
n_{ILS}	number of long sleep cycles I-DRX	unlimited
e_{SSB}	SSB detection power consumption	100 mW/slot
e_{RACH}	RACH procedure power consumption	376.5 mW
e_{IDLE}	RRC-Idle state power consumption	0 mW/slot
e_i	power consumption in active state	350 mW
e_{ON}	"On Duration" power consumption	61 mW/slot
e_{CSS}	C-DRX short sleep power consumption	45 mW/slot
e_{CLS}	C-DRX long sleep power consumption	20 mW/slot
e_{ISS}	I-DRX short sleep power consumption	20 mW/slot
e_{ILS}	I-DRX long sleep power consumption	1 mW/slot
T_{BFI}	Beam failure indication count threshold	2



(a) Energy efficiency as a function of $\Delta\theta = \Delta\phi$ (b) Latency as a function of $\Delta\theta = \Delta\phi$ (c) Energy efficiency as a function of λ_B

Fig. 4. Performance of DRX scaling schemes as a function of micromobility speed $\Delta\theta = \Delta\phi$ and blockers density λ_B.

these improvements are different for different blockers' densities. Specifically, for high blockers density improvements are much milder which is explained by the exponentially decaying nature of the blockage probability [21]. However, the gains in terms of energy efficiency are relatively similar for considered values of λ_B. Recalling that multiconnectivity also decreases the fraction of time in outage [3], the optimal trade-off between considered metrics depends on application preferences.

We now proceed to assess the performance of DRX scaling as illustrated in Fig. 4 as a function of micromobility speed and blockers density. Analyzing the impact of the former metric, we observe that for low micromobility speed of $\Delta\theta = \Delta\phi = 0.15°$/s energy efficiency varies only slightly while the latency increases for all the schemes. The smaller impact of latency is produced by a proportional scheme that can be recommended for practical implementation. Note that there is a small gap in k values, where the exponential scaling scheme outperforms the proportional one, $k \sim (0.5 - 1.0)$. However, the associated latency gain is rather small.

For higher micromobility speeds, i.e., $\Delta\theta = \Delta\phi = 0.3°$/s, the latency behavior is preserved. Although for practical values of k corresponding to the minimal latencies in Fig. 4(b), i.e., $k \sim 0.5 - 1.0$, linear scaling scheme provides the best energy efficiency. However, even at these values of k, the latency of this scheme is significantly higher than that of the proportional scaling scheme which performs slightly worse than the linear one in terms of energy efficiency. Thus, for high micromobility speed proportional DRX scaling scheme provides the trade-off between considered metrics. Fig. 4(c) further shows that these conclusions remain valid for other blockers' densities.

V. Conclusion

In this paper, we proposed a model characterizing the trade-off between energy efficiency and latency in mmWave 5G NR systems with multiconnectivity operation. We then investigated the impact of environmental characteristics and different DRX scaling schemes on the considered metrics. Our results show that as opposed to blockage having a strictly negative impact on the considered metrics, micromobility speed has a positive effect on energy efficiency. For low micromobility speeds ($< 0.2°$/s) proportional DRX scaling scheme with small values of k, $k < 1.0$, provides the best performance in terms of considered metrics while for higher speeds it leads to a compromise between them. The optimal value of the DRX scaling coefficient k depends on the design preferences.

References

[1] M. Gapeyenko *et al.*, "On the temporal effects of mobile blockers in urban millimeter-wave cellular scenarios," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 11, pp. 10124–10138, 2017.

[2] N. V. Stepanov, D. Moltchanov, V. Begishev, A. Turlikov, and Y. Koucheryavy, "Statistical analysis and modeling of user micromobility for THz cellular communications," *IEEE Trans. on Veh. Tech.*, 2021.

[3] 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, 2019.

[4] 3GPP, "NR; Multi-connectivity; stage 2 (Release 16)," 3GPP TS 37.340 V16.0.0, December 2019.

[5] M. Polese *et al.*, "Improved handover through dual connectivity in 5G mmWave mobile networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, pp. 2069–2084, September 2017.

[6] 3GPP, "5G; NR; NR and NG-RAN Overall description; stage 2 (Release 15)," 3GPP TS 38.300 V15.11.0, 3GPP, November 2020.

[7] IWPC, "5G millimeter wave frequencies and mobile networks a technology whitepaper on key features and challenges," pp. 149–150, 2019.

[8] Y.-N. R. Li *et al.*, "Power saving techniques for 5G and beyond," *IEEE Access*, vol. 8, pp. 108675–108690, 2020.

[9] D. Moltchanov, A. Ometov, S. Andreev, and Y. Koucheryavy, "Upper bound on capacity of 5G mmwave cellular with multi-connectivity capabilities," *Electronics Letters*, vol. 54, no. 11, pp. 724–726, 2018.

[10] M. Kumar Maheshwari, M. Agiwal, and A. Rashid Masud, "Analytical modeling for signaling-based DRX in 5G communication," *Trans. on Emerging Telecomm. Techn.*, vol. 32, no. 1, pp. 1–18, 2021.

[11] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz (Release 14)," 3GPP TR 38.901 V14.1.1, July 2017.

[12] V. Petrov *et al.*, "Interference and SINR in millimeter wave and terahertz communication systems with blocking and directional antennas," *IEEE Transactions on Wireless Communications*, vol. 16, pp. 1791–1808, March 2017.

[13] A. B. Constantine *et al.*, "Antenna theory: analysis and design," *Microstrip Antennas, John Wiley & Sons*, 2005.

[14] V. Petrov, D. Moltchanov, Y. Koucheryavy, and J. M. Jornet, "The effect of small-scale mobility on terahertz band communications," in *Proceedings of the 5th ACM International Conference on Nanoscale Computing and Communication*, pp. 1–2, 2018.

[15] V. Petrov, D. Moltchanov, Y. Koucheryavy, and J. M. Jornet, "Capacity and outage of terahertz communications with user micro-mobility and beam misalignment," *IEEE Transactions on Vehicular Technology*, 2020.

[16] S. Rostami *et al.*, "Wake-up radio based access in 5G under delay constraints: Modeling and optimization," *IEEE Transactions on Communications*, vol. 68, no. 2, pp. 1044–1057, 2020.

[17] L. Sharma, B. B. Kumar, and S.-L. Wu, "Performance analysis and adaptive DRX scheme for dual connectivity," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10289–10304, 2019.

[18] D. Moltchanov, "Distance distributions in random networks," *Elsevier Ad Hoc Networks*, vol. 10, pp. 1146–1166, August 2012.

[19] D. Moltchanov, Y. Gaidamaka, D. Ostrikova, V. Beschastnyi, Y. Koucheryavy, and K. Samouylov, "Ergodic outage and capacity of terahertz systems under micromobility and blockage impairments," *IEEE Transactions on Wireless Communications*, vol. 21, no. 5, pp. 3024–3039, 2022.

[20] J. G. Kemeny, J. L. Snell, *et al.*, *Finite markov chains*, vol. 356. Van Nostrand Princeton, NJ, 1960.

[21] M. Gapeyenko *et al.*, "Analysis of human-body blockage in urban millimeter-wave cellular communications," in *2016 IEEE International Conference on Communications (ICC)*, pp. 1–7, IEEE, 2016.