

Full-stack Hybrid Beamforming in mmWave 5G Networks

Felipe Gómez-Cuba¹, Tommaso Zugno², Junseok Kim³, Michele Polese⁴, Saewoong Bahk⁵, Michele Zorzi²

¹AtlantTIC, University of Vigo, Spain. Email: gomezcuba@gts.uvigo.es

²Department of Information Engineering, University of Padova, Padova, Italy.

Email: {zugnotom, zorzi}@dei.unipd.it

³System LSI, Samsung Electronics, Gyeonggi-do, South Korea. Email: junseok.kim@samsung.com

⁴Institute for the Wireless Internet of Things, Northeastern University, Boston, MA USA.

Email: m.polese@northeastern.edu

⁵Department of ECE and INMC, Seoul National University, Seoul, South Korea. Email: sbahk@snu.ac.kr.

Abstract—This paper analyzes Hybrid Beamforming (HBF) and Multi-User Multiple-Input Multiple-Output (MU-MIMO) in millimeter wave (mmWave) 5th generation (5G) cellular networks considering the full protocol stack with TCP/IP traffic and MAC scheduling. Prior work on HBF and MU-MIMO has assumed full-buffer transmissions and studied link-level performance. We report non-trivial interactions between the HBF technique, the front-loaded channel estimation pilot scheme in NR, and the constraints of MU-MIMO scheduling. We also report that joint multi-user beamforming design is imperative, in the sense that the MU-MIMO system cannot be fully exploited when implemented as a mere collection of single-user analog beams working in parallel. By addressing these issues, throughput can be dramatically increased in mmWave 5G networks by means of Spatial Division Multiple Access (SDMA).

Index Terms—mmWave, hybrid beamforming, 3GPP NR, end-to-end, system-level simulations

I. INTRODUCTION

The next generations of mobile networks will need to support a wide range of applications that demand significantly higher rates, such as video and content delivery; lower latency, such as eHealth and Connected Autonomous Vehicless (CAVs); and increased reliability, such as Internet of Things (IoT) in smart factories and smart cities [1]. The 3rd Generation Partnership Project (3GPP) specification for 5G cellular networks [2] is positioned to address these challenges, introducing a new Radio Access Network (RAN) design (i.e., 3GPP NR). It features a flexible Orthogonal Frequency Division Multiplexing (OFDM) frame structure with a flexible *numerology*, and, for the first time in the RAN, the support for mmWave communications in the 24–53 GHz band, referred to as Frequency Range 2 (FR2).

NR in the mmWave spectrum supports much larger bandwidths with respect to legacy 3GPP Radio Access Technologies (RATs) in sub-6 GHz bands [3], with up to 400 MHz for each carrier [2]. The small wavelength permits to integrate a large number of antenna elements even in a small mobile device. By means of Beamforming (BF) techniques it is possible to concentrate the transmitted power in a single direction and to offset the higher path loss at higher carrier frequencies. Since the use of fully digital large antenna arrays

with high-resolution Analog to Digital Converters (ADCs) consumes too much power, energy efficient BF architectures are considered. In Analog Beamforming (ABF) and HBF architectures, one or $K > 1$ Radio Frequency (RF) chains with analog circuits feed an array of $N \gg K$ antennas, respectively. In ABF, a single frequency-flat analog beam can be employed at a time, whereas in hybrid BF more sophisticated beam design is possible at the expense of some more power consumption [4]. Other proposals consider the usage of low-resolution ADCs in fully-digital BF architectures [4], which we leave for consideration in future work.

HBF is capable of using digital low-dimensional linear operations to modify the effective beams seen by different OFDM subcarriers in a frequency-selective manner. In addition, HBF is capable of steering multiple SDMA *layers* at the same time, i.e., different signal beams delivering independent data streams. It is typical that mmWave channels are sparse in the sense that sending multiple SDMA layers to the same user is ineffective, making MU-MIMO imperative to achieve spatial multiplexing gains in mmWaves [5].

Abundant physical layer literature has studied beam design, transceiver circuits, and spatial multiplexing for mmWave, such as [4]–[7]. Nowadays the 3GPP NR specifications support HBF and MU-MIMO [2], [8]. However, the state of the art currently lacks an analysis of how a *physical layer* based on HBF interacts with the *full protocol stack*.

There is rising recent interest on the evaluation of the end-to-end performance of 5G mmWave networks, with analysis, simulations, and experiments. However, most end-to-end simulation literature has focused on single-layer analog BF [9], [10], whereas most of the work considering SDMA focused on full buffer link-level evaluation [4]–[7]. To fill this gap, in this paper we study the integration of HBF techniques for MU-MIMO SDMA into a full-stack 5G and beyond cellular network, focusing on beam design in presence of inter-beam interference and on the scheduling constraints arising in a context of multiple simultaneous transmissions.

We extend prior models for full-stack mmWave 5G networks by considering MU-MIMO transmissions. We find that the Codebook Beamforming (CBF) schemes that work well

in the single-user case lead to high inter-beam interference in a SDMA scenario, and propose a Frequency-Selective MMSE BeamForming (SMBF), which performs low-dimensional linear pre-processing for HBF beam design in order to remove this inter-beam interference. We notice a suboptimal interaction between scheduling and MU-MIMO beam design. This arises from the fact that the *effective* channel coefficient (after beamforming is applied) needs to remain unchanged during an entire allocated period of time where a single pilot is used to estimate the OFDM channel gains. As a consequence, to mitigate the inter-beam interference through the introduction of a SMBF scheme, all simultaneous transmissions must begin at the same time. To this aim, the scheduler needs to allocate some padding blank resources without a transmission, leading to resource waste.

Finally, we demonstrate our results in a novel HBF extension¹ for the ns-3 mmWave module [9]. We believe this is the first open source software to support HBF MU-MIMO SDMA at mmWaves with 3GPP-like Medium Access Control (MAC), the 3GPP mmWave channel model, and realistic TCP/IP traffic. Our results show that MU-MIMO SDMA relying on HBF can increase the capacity of a single-layer mmWave network, but HBF-aware scheduling design is fundamental to achieve the potential gains. Additional material and results can be found in the draft available in [11].

The rest of the paper is structured as follows. Sec. II discusses our HBF and scheduling model for 3GPP NR networks. Sec. III describes the performance evaluation results. Finally, Sec. IV concludes the paper.

II. FULL-STACK INTEGRATION OF HBF FOR MMWAVES

In the NR frame, complex symbols are mapped in a 3-dimensional OFDM resource grid, comprising the OFDM symbol number in time (n), the OFDM subcarrier number in frequency (k), and the *SDMA layer* number (ℓ) [12]. Some options of the waveform are suited for sub-6 GHz frequencies but not for mmWaves, for instance, since HBF features frequency-flat analog operations that cannot differ for different OFDM subcarriers, Orthogonal Frequency Division Multiple Access (OFDMA) is not employed at FR2 and all subcarriers k in a port-symbol pair are assigned to the same user. Thus, the scheduling reduces to a 2-dimension Time Division Multiple Access (TDMA) and SDMA grid (n, ℓ). Moreover, typical mmWave Multiple-Input Multiple-Output (MIMO) channel matrices are rank deficient (i.e., the second largest eigenvalue is much smaller than the first) [5]. This means that assigning two or more SDMA layers to the same user is ineffective. Finally, the layers are mapped to one or more antenna ports (p), defined as an RF input to the array that experiences the same frequency-flat analog circuitry, using low dimensional digital linear precoding that can be frequency-selective [12].

The channel matrix between the Base Station (BS) and each User Equipment (UE) u is denoted by $\mathbf{H}_u[n, k]$ in OFDM

symbol n and subcarrier k . In Downlink (DL), the BS selects a BF vector for each layer and subcarrier $\mathbf{v}_\ell[k]$ using a certain BF scheme, and the UE receives with the analog BF vector \mathbf{w}_u . Thus, the *effective* scalar complex channel between the transmit port p and the receive antenna port of the UE is given by

$$h_{eq}[u, \ell, n, k] = \mathbf{w}_u^T[k] \mathbf{H}_u[n, k] \mathbf{v}_\ell[k],$$

while the Uplink (UL) channel is computed with the transposed channel matrix and swapping transmitter and receiver beamforming vectors, resulting in the same complex scalar number.

We assume a Signal to Interference plus Noise Ratio (SINR)-based point-to-point link performance model. We compute the SINR of each link, assuming that simultaneous links are decoded separately, and map their SINR to the Block Error Rate (BLER) of the transmission. This is a simplification of real decoding hardware that makes the simulation of a large network tractable. Real NR demodulation and decoding may use sophisticated joint decoding such as sphere decoding [13], as well as LDPC and Polar channel codes [12].

We model each OFDM symbol independently, so for the SINR calculation we will omit the OFDM symbol index n . For multiple simultaneous DL transmissions sent by the BS on multiple layers, we write the DL SINR of user u at subcarrier k as a function of the effective channel gains as

$$SINR_u^{DL}[k] = \frac{L_u |h_{eq}[u, \ell(u), k]|^2 P_{sc}}{\sum_{u' \neq u} L_u |h_{eq}[u, \ell(u'), k]|^2 P_{sc} + \Delta f N_o} \quad (1)$$

where $\ell(u)$ indicates the layer assigned to the UE u , L_u is the pathloss of u , P_{sc} is the transmitter power per layer and per subcarrier, N_o is the noise Power Spectral Density (PSD) and Δf is the inter-carrier spacing. We focus on the sum of interference over UEs connected to the same BS, producing inter-beam interference. We assume there is no cooperation and if there are other BSs, their interference may be modeled in bulk by increasing N_o . We note that the inter-beam interference depends on the “mismatched” vectors of u and u' . Even if this reduces the BF gain experienced by the interference, this term may be significant, making the link SINR much lower than the Signal to Noise Ratio (SNR).

The UL SINR follows the expression

$$SINR_u^{UL}[k] = \frac{L_u |h_{eq}[u, \ell(u), k]|^2 P_{sc}}{\sum_{u' \neq u} L_{u'} |h_{eq}[u', \ell(u), k]|^2 P_{sc} + \Delta f N_o}, \quad (2)$$

where the interference of user u' is received through the channel of user u' , with pathloss gain $L_{u'}$. This can make the UL interference even more severe (e.g., if $L_{u'} \gg L_u$).

This SINR point-to-point link model can support any arbitrary BF vector design. In this paper we consider the following BF techniques to achieve good link SNR or SINR values.

A. CodeBook Beamforming (CBF)

We define a *BF codebook* \mathcal{B} as a small collection of possible frequency-flat analog BF vectors. The transmitter

¹ Available at <https://github.com/signetlabdei/ns3-mmwave-hbf>

sends reference signals using all the vectors in the set \mathcal{B}_T , and the receiver tests decoding the reference signals with all vectors in \mathcal{B}_R . Finally, the receiver selects the best pair and feeds back the decision to the transmitter. As an approximation of a real maximum received reference signal power criterion, in our implementation we assume the ideal max-SNR criterion:

$$\mathbf{v}_{\ell(u)}, \mathbf{w}_u = \arg \max_{\mathbf{v} \in \mathcal{B}_T, \mathbf{w} \in \mathcal{B}_R} |\mathbf{w}_u^T \mathbf{H}_u[n, k_{ref}] \mathbf{v}_{\ell}|^2, \quad (3)$$

where k_{ref} is a single subcarrier index where we assume a narrowband reference signal was sent.

We denote the antenna array response as a function $\mathbf{a}(\theta, \phi)$ that depends on the angles of azimuth and elevation (θ, ϕ) . In our simulations we adopt the Uniform Planar Array (UPA) model with $N_1 \times N_2$ antennas separated half a wavelength, where the i -th element of the vector is

$$\mathbf{a}_i(\theta, \phi) = e^{-j\frac{\pi}{2}((i \bmod N_1) \sin(\theta) + \lfloor i/N_1 \rfloor \sin(\phi))}.$$

By generating a codebook of $N_1 N_2$ vectors using $\mathbf{a}()$ at the special angles $\{\theta = \sin^{-1}(\frac{2n}{N_1} - 1) : n \in \{0 \dots N_1 - 1\}\}$ and $\{\phi = \sin^{-1}(\frac{2n}{N_2} - 1) : n \in \{0 \dots N_2 - 1\}\}$, then the codebook conveniently becomes the set of columns of a $N_1 N_2 \times N_1 N_2$ double Discrete Fourier Transform (DFT) matrix. The DFT codebook facilitates implementation with analog phase-arrays or with lensed arrays, and is similar to precoding in [12].

B. Frequency-Selective MMSE BeamForming (SMBF)

This BF scheme introduces a low-complexity, low-dimensional, frequency-selective linear matrix mapping between layers and ports, in combination with an auxiliary analog frequency-flat CBF underlying scheme. Let us denote the BF vectors selected using CBF by \mathbf{w}_u^{CB} and \mathbf{v}_{ℓ}^{CB} . We assume that first the system conducts a codebook exploration as in CBF and loads the best codebook BF vector for each user u to different antenna ports denoted $p(u)$. In addition, we assume that after the codebook exploration the BS transmits pilot signals in all subcarriers and the receivers can report back a set of effective channel coefficients $\{\sqrt{L_u} h_{eq}^{CB}[u, p(u'), n, k]\}$ for all pairs $(u, p(u'))$ and subcarrier indices k . The same pilots used for data decoding are suitable, but to report these auxiliary effective channel coefficients back to the BS would incur some overhead. Nonetheless, in this paper we assume that ideal Minimum Mean Squared Error (MMSE) precoding is possible, and we leave overhead reduction and other precoding constraints for future work.

To simplify notation, we assume that the users are numbered sequentially $u \in \{0 \dots N_u\}$ and that their assigned layer and port numbers are equally sequential $\ell(u) = p(u) = u$. Using the auxiliary scalar channel coefficients, the BS builds the MU-MIMO reference equivalent channel matrix

$$\mathbf{H}_{eq}[k] = \begin{pmatrix} \sqrt{L_1} h_{eq}^{CB}[1, 1, k] & \dots & \sqrt{L_1} h_{eq}^{CB}[1, N_p, k] \\ \vdots & \ddots & \vdots \\ \sqrt{L_{N_u}} h_{eq}^{CB}[N_u, 1, k] & \dots & \sqrt{L_{N_u}} h_{eq}^{CB}[N_u, N_p, k] \end{pmatrix}, \quad (4)$$

where $N_p = N_u$ is the number of analog BF ports, each associated to a single user. Moreover since $\ell(u) = p(u) = u$, the desired channels are in the main diagonal of this matrix.

Finally, for DL, on the receiver side, the receiving BF vectors would remain those of CBF, while on the transmitter side the BS designs a set of precoding matrices for each subcarrier k , matching layers to ports using the following MMSE DL precoding expression:

$$\mathbf{V}_{MMSE}[k] = \mathbf{H}_{eq}^H[k] (\mathbf{H}_{eq}[k] \mathbf{H}_{eq}^H[k] + \frac{N_o \Delta f}{P} \mathbf{I})^{-1}.$$

We adopt the MMSE technique because, when the noise is weak compared to the transmitted power, then $\frac{N_o \Delta f}{P_{sc}} \rightarrow 0$, and the expression converges to the pseudoinverse (zero-forcing precoder), i.e., $\mathbf{H}_{eq}[k] \mathbf{V}_{MMSE}[k] = \mathbf{I}$, thus suppressing the interference. In addition when the noise is strong, in the limit $\frac{N_o \Delta f}{P_{sc}} \rightarrow \infty$, the MMSE expression converges to the hermitian (matched filter) which maximizes the received SNR. Thus, MMSE offers a balance between interference suppression and noise reduction.

Finally, the final *effective* transmit BF vectors at the BS for DL are obtained by first computing

$$(\tilde{\mathbf{v}}_1^{MMSE}[k] \dots \tilde{\mathbf{v}}_{N_u}^{MMSE}[k]) = (\mathbf{v}_1^{CB} \dots \mathbf{v}_{N_u}^{CB}) \mathbf{V}_{MMSE}[k],$$

and then introducing the following normalization to preserve the transmitted power constraint in each layer:

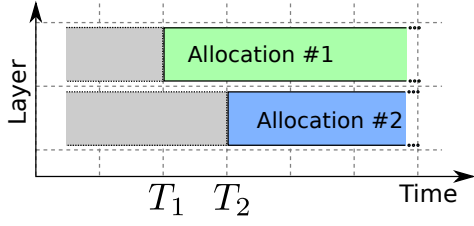
$$\mathbf{v}_u^{MMSE}[k] = \tilde{\mathbf{v}}_u^{MMSE}[k] / \|\tilde{\mathbf{v}}_u^{MMSE}[k]\|.$$

Introducing these effective vectors into (1), instead of the auxiliary CBF vectors we discussed earlier, results in the new SINR values of the MMSE technique. For UL, an equivalent hybrid combining at the BS receiver can be formulated by trasposing the matrices described in this section.

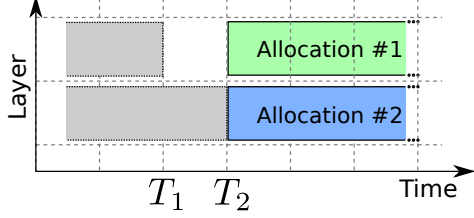
C. Scheduling

We assume that the scheduler produces allocation decisions assigning transmissions to the TDMA+SDMA grid (n, ℓ) periodically [12]. We assume that the first and the last symbol of each NR 14-symbol slot are reserved for the transmission of the Physical Downlink Control Channel (PDCCH) and Physical Uplink Control Channel (PUCCH), respectively. Data transmissions are assigned to the symbols 2 to 13 of the slot, and all symbols are “flexible,” meaning that they can be employed for DL or UL at the scheduler’s decision [12]. We assume perfect channel estimation and do not model DeModulation Reference Signals (DMRSs) explicitly. We assume that the minimum data allocation unit is 1 OFDM symbol of data transmission.

Since each allocation has only one front-loaded DMRS, the BF vector selected at the start of the transmission may not change until the allocation ends. This means that, when two transmissions overlap with different starting instants, they may not observe each other’s references DMRS and employ SMBF to reduce their mutual interference (Fig. 1(a)). For this reason we design a scheduler that introduces blank symbols at the



(a) At T_1 allocation #1 does not observe DMRS from other layers and cannot design a MMSE precoding. When Allocation #2 starts at T_2 , it does not observe other DMRSs either. After T_2 , both allocations experience interference as in CBF even though SMBF is supported.



(b) After the transmission in the top layer ends at T_1 , the scheduler leaves a padding symbol without signal and Allocation #1 starts at T_2 . Both layers start a new allocation simultaneously, they observe the other's DMRS and implement SMBF successfully. However, the resource region $T_2 - T_1$ in the top layer is wasted.

Fig. 1. SMBF conflict for parallel allocations with different start time.

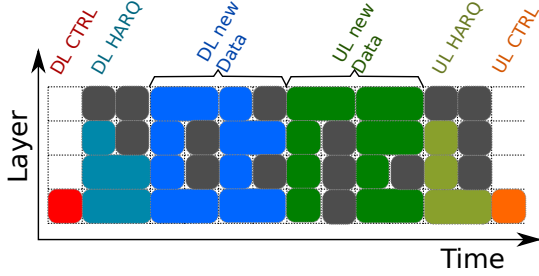


Fig. 2. PMRS example. The resources in gray are wasted as padding to guarantee equal start times for all allocations.

end of some transmissions and enforces equal start times for all simultaneous transmissions of different layers (Fig. 1(b)).

Given N_ℓ layers, N_s symbols and N_u total UEs, our Padded mmWave RR Scheduler (PMRS) equally divides the subframe in $N_b = \lceil N_u / N_\ell \rceil$ “SDMA bundles.” Each bundle consists in N_ℓ concurrent transmissions with the same start time allocated to different layers. Bundles are further time-multiplexed over the full subframe, where each bundle duration in time is exactly $N_a = \lfloor N_s / N_b \rfloor$ symbols. Layers that have fewer symbols to transmit start at the same time as the rest of their bundle, but end transmitting sooner. The interval between the end of any transmission and the start of the next bundle is a “padding” of wasted symbols (Fig. 2). Within each bundle and for each layer, a different UE is selected. If $N_u > N_s \times N_\ell$, then some UEs are left unserved and become the first UEs in the list for the next subframe in a Round Robin (RR) fashion.

III. PERFORMANCE EVALUATION

We implemented an MU-MIMO HBF extension for the ns-3 mmWave module introduced in [9]. Besides the implementation of the HBF and MU-MIMO features, we made adjustments for a more realistic simulation of 5G networks. Instead of the NYU channel model [9], we adopt the 3GPP channel model [14]. In addition, the OFDM resource grid parameters (bandwidth, subcarrier spacing, symbol duration, and number of slots per frame) reflect those of NR, as described in Sec. II and [12]. Notice that the ns-3 mmWave module assumes that control signaling is ideal and messages are never lost or corrupted.

In our implementation we have introduced modifications to numerous C++ classes in the ns-3 mmwave module. Notably, the antenna array module now supports multiple antenna ports, with different BF configurations. Moreover, the 3GPP channel model implementation has been extended to account for the multi-layer interference of Eq. (1) and Eq. (2), while the channel abstraction code and the physical layer implementation have been refactored to support multiple SDMA asynchronous layers (i.e., transmissions from a single entity). The BF strategies described in Sec. II have been implemented in a plug-and-play fashion, leveraging a novel, flexible BF module. Finally, we updated the ns-3 mmwave module MAC layer to support multiple asynchronous layers, by properly accounting for the mapping of upper layer PDUs to mmwave Transport Blocks on different antenna ports, the management of Hybrid Automatic Repeat reQuest (HARQ) retransmissions, the CQI estimation, and the control signaling. The MAC layer also features a plug-and-play implementation of the scheduler introduced in Sec. II, which is backward compatible and allows comparison with the other scheduling strategies implemented in the ns-3 mmWave module [9]. We refer the reader to the publicly available Github repository with the HBF extension for additional details.

A. Simulation Scenario

We consider a random mmWave cellular system with one BS located at the origin of the coordinates (0,0) with a height of 25 m, and 7 UEs located at random positions uniformly distributed in a disc of radius 100 m with a height of 1.6 m. We generate and average the results over 20 such random deployments, UE locations and channels. Due to the considerable pathloss in mmWave, we assume inter-cell interference is severely attenuated and it is sufficient to simulate one cell.

We configured the NR OFDM waveform with numerology $\mu = 2$, which corresponds to a subcarrier spacing of 60 kHz. The central frequency is 28 GHz and the bandwidth 198 MHz is divided into 275 Resource Blocks (RBs), each including 12 subcarriers. There are 4 slots per subframe with duration 250 μ s, and the OFDM symbol duration is 17.85 μ s including the CPs. We adopt the channel model described in 3GPP TR 38.901 [15] and consider the “Urban Macro” scenario. The transmission power is 30 dBm, and the receiver noise figure 5 dB. We consider a 8×8 UPA with either 1 or 4 layers in the BS and a 4×4 UPA with 1 layer for the UEs.

B. Comparison of BF Solutions

We compare the CBF scheme that focuses on improving the SNR and is sufficient in single-layer cases versus our proposed SMBF scheme. We use Radio Link Control (RLC) Unacknowledged Mode (UM) (i.e., without RLC retransmissions), disable the HARQ retransmissions, and use low-traffic application in the UEs. This makes it possible to probe the channel and BF scheme at a constant rate, and to measure the statistics of the SINR and BLER in the physical layer without disruptions by the upper layers.

The low rate application is a constant traffic generator that produces a packet of 1500 bytes every 1500 μ s in each UE. Roughly speaking, when the Modulation and Coding Scheme (MCS) coding rate is greater than 3.64 bits per subcarrier, the 3300 subcarriers can carry a full packet in a single OFDM symbol. This means that the scheduler receives a demand for 14 symbols in one out of every six slot of 250 μ s. The frame has thus plenty of RBs to satisfy the demand and the cell is lightly loaded.

We represent the received UL SINR Cumulative Distribution Function (CDF) for all transmission allocations in the simulation in Fig. 3(a). We compare 1 layer (solid) and 4 layer (dashed) cases. For the 1 layer case, we use the TDMA mmWave RR Scheduler (TMRS) without SDMA capabilities that was implemented in the previous versions of the ns-3 mmWave module, with CBF. For the 4 layer case we consider the PMRS with both CBF and SMBF. Since there is no self-interference, in the 1-layer case the SINR is the same as the SNR. Moreover, with 4 layers, adopting a single-layer BF scheme (CBF) leads to SINR degradation. Finally, for the SMBF scheme the SINR CDF is nearly identical to that of the single-layer CBF case. This suggests that with SMBF, the SINR is almost equal to the SNR and the scheme removes almost all inter-beam interference.

We represent the received DL SINR CDF in Fig. 3(b). The main difference with the UL case is that in DL the desired and interfering signals at each UE experience the same pathloss, and -20 dB SINR outages with 4-layer CBF rarely happen. On the other hand, the gap between the higher range of SINRs achieved with 4-layer CBF and 4-layer MMSE BF schemes is wider than in UL.

Finally we depict the instantaneous BLER CDF for all UL transmissions in Fig. 3(c). The instantaneous BLER is dominated by outages when the channel has changed and the Channel Quality Information (CQI) is outdated, as most transmissions experience either $\text{BLER} \leq 10^{-2}$ or $\text{BLER} = 1$. As we can see, 4-layer CBF has a much larger outage probability (lower step in the BLER CDF) and results in more severe average BLER in the system. Again, SMBF behaves almost as a 1-layer CBF situation. We do not depict the DL BLER CDF due to space constraints, as its insights were identical.

C. Application Performance on Loaded Cell

Next, we consider the BLER and throughput with high traffic load applications. Qualitatively we wish to check that the PMRS does not introduce too much padding so as to

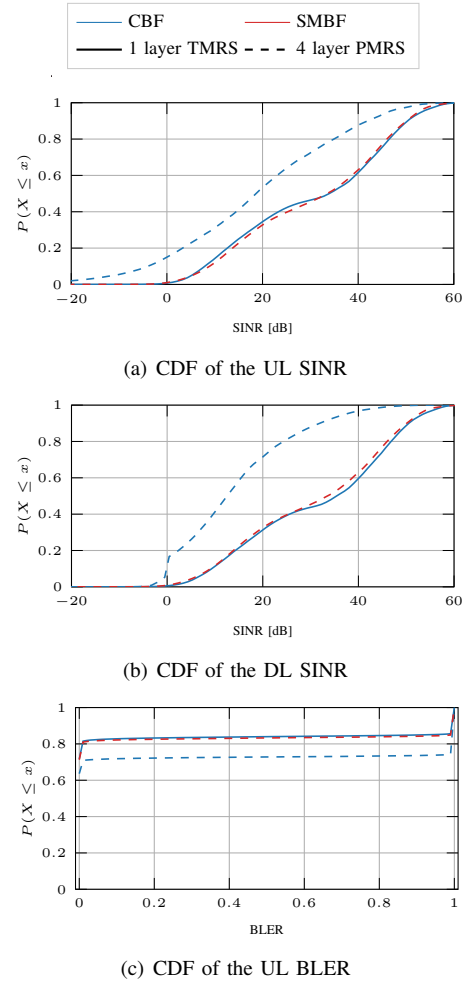


Fig. 3. Comparison of the different BF schemes.

completely cancel the increased capacity of the multi-layer SDMA capabilities of the physical layer. Once again we use RLC UM and disable the HARQ retransmissions. For space constraints we leave the impact of retransmissions in application throughput and delay for future extensions of our work.

Since we have already determined the best BF scheme for each number of layers, we consider the TMRS 1-layer scheduler with CBF, and our proposed PMRS for 4 layers with SMBF.

The high-rate application is a constant bit rate source that generates a packet of 1500 bytes every 150 μ s in each UE, with a symmetric traffic in uplink and downlink. For every slot of 250 μ s, the scheduler always receives requests for at least ~ 23 symbols. In the 1-layer case there are 12 data symbols per slot, which are not enough to allocate all the demand. In the 4-layer case, there are 12×4 available data symbols, i.e., more than enough when the users demand ~ 23 symbols.

Figure 4(a) reports the average DL and UL BLER for the two cases. The BLER of PMRS with 4 layers is comparable to that of TMRS with 1 layer. Recalling our prior remark that BLER is mostly driven by CQI outage, this is consistent with

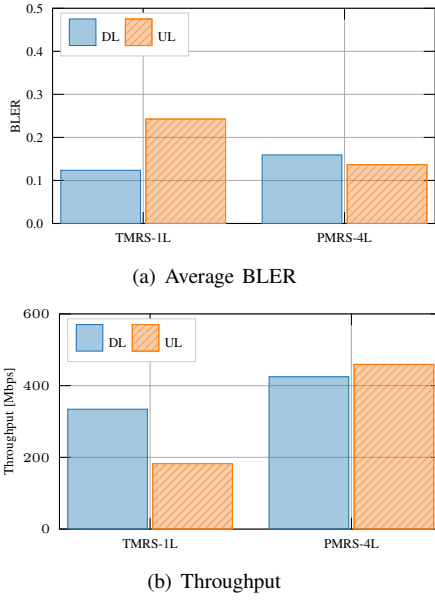


Fig. 4. Comparison of the different scheduling strategies.

the SINR plots discussed in the previous section. Since SMBF removed almost all interference, both schemes experienced comparable CQI and outage events.

Figure 4(b) depicts the average throughput in the simulations. The sum source rate is 560 Mbps. For TMRS with CBF, we see that the throughput is 330 Mbps in DL and 180 Mbps in UL, with significant asymmetry and lower value than the offered traffic. This is because the demand exceeds the number of data symbols of the 1-layer frame even with the maximum MCS rate. In the 4-layer case, the cell is not saturated, and the throughput with PMRS exceeds 420 Mbps in DL and 450 Mbps in UL. This shows the main advantage of SDMA MU-MIMO in mmWave networks, i.e., *an increase in the number of available RBs by a factor of N_{layers} allows the network to support much more traffic*. Particularly in our simulation a delivered traffic that is $2\times$ the capacity of the single-layer frame was achieved without stressing the system. The SDMA scheme could still saturate for even higher application rates, or experience issues with delay or with retransmissions either by HARQ or Transmission Control Protocol (TCP), which we leave for future extensions.

IV. CONCLUSIONS

We have studied MU-MIMO HBF implementations for 3GPP NR mmWave end-to-end cellular systems. We have shown that SDMA greatly increases the system capacity. Moreover, by associating each frequency-flat BF vector to a separate antenna port, the MU-MIMO signal processing can be handled in a space of reduced dimensions instead of the large arrays characteristic of mmWave. It is necessary to alleviate the inter-beam interference in order to improve the SINR, as otherwise if we merely used separate analog beams the SINR would degrade significantly. We reveal a conflict between the design of MU-MIMO schedulers and MU-MIMO BF, which stems from the unique front-loaded DMRS per transmission. In

future work we intend to study feedback overhead reduction, asynchronous scheduling without padding, and the effect of retransmission schemes on throughput and the delay observed by the applications. We also leave for future work the study of non-linear MU-MIMO signal processing methods, such as joint sphere decoding or successive interference cancelling, which can theoretically outperform MMSE in terms of BLER.

REFERENCES

- [1] F. Boccardi, R. W. Heath Jr, A. Lozano, T. L. Marzetta, and P. Popovski, "Five Disruptive Technology Directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, Feb. 2014.
- [2] 3GPP, "NR and NG-RAN Overall Description," TS 38.300 (Rel. 15), 2018.
- [3] 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, Mar. 2014.
- [4] R. W. Heath, N. González-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An Overview of Signal Processing Techniques for Millimeter Wave MIMO Systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 436–453, Apr. 2016.
- [5] S. Sun, T. Rappaport, R. Heath, A. Nix, and S. Rangan, "MIMO for Millimeter-Wave Wireless Communications: Beamforming, Spatial Multiplexing, or Both?" *IEEE Communications Magazine*, vol. 52, no. 12, pp. 110–121, Dec. 2014.
- [6] F. Sohrabi and W. Yu, "Hybrid Analog and Digital Beamforming for mmWave OFDM Large-Scale Antenna Arrays," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 7, pp. 1432–1443, July 2017.
- [7] M. N. Kulkarni, A. Ghosh, and J. G. Andrews, "A Comparison of MIMO Techniques in Downlink Millimeter Wave Cellular Networks With Hybrid Beamforming," *IEEE Transactions on Communications*, vol. 64, no. 5, pp. 1952–1967, May 2016.
- [8] B. Mondal, V. Sergeev, A. Sengupta, G. Ermolaev, A. Davydov, E. Kwon, S. Han, and A. Papathanassiou, "MU-MIMO and CSI Feedback Performance of NR/LTE," in *53rd Annual Conference on Information Sciences and Systems (CISS)*, Baltimore, MD, USA, USA, 2019.
- [9] M. Mezzavilla, M. Zhang, M. Polese, R. Ford, S. Dutta, S. Rangan, and M. Zorzi, "End-to-End Simulation of 5G mmWave Networks," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 2237–2263, Apr. 2018.
- [10] S. Choi, J. Song, J. Kim, S. Lim, S. Choi, T. T. Kwon, and S. Bahk, "5G K-SimNet: End-to-End Performance Evaluation of 5G Cellular Systems," in *16th IEEE Annual Consumer Communications Networking Conference (CCNC)*, Las Vegas, NV, USA, USA, Jan. 2019.
- [11] F. Gomez-Cuba, T. Zugno, J. Kim, M. Polese, S. Bahk, and M. Zorzi, "Hybrid Beamforming in 5G mmWave Networks: a Full-stack Perspective," *arXiv preprint*. [Online]. Available: <https://arxiv.org/abs/2010.04220>
- [12] 3GPP, "NR; Physical channels and modulation, v16.1.0," TS 38.211, Apr. 2020.
- [13] U. Fincke and M. Pohst, "Improved methods for calculating vectors of short length in a lattice, including a complexity analysis," *Mathematics of computation*, vol. 44, no. 170, pp. 463–471, Apr. 1985.
- [14] T. Zugno, M. Polese, N. Patriciello, B. Bojovic, S. Lagen, and M. Zorzi, "Implementation of A Spatial Channel Model for ns-3," in *Proceedings of the 2020 Workshop on ns-3 (WNS3)*, Gaithersburg, MD, USA, Jun. 2020.
- [15] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," TR 38.901, v15.0.0, Jun. 2018.