Cloud-RAN Fronthaul Rate Reduction via IBM-based Quantization for Multicarrier Systems

Johannes Demel, Tobias Monsees, Carsten Bockelmann, Dirk Wübben, Armin Dekorsy
Department of Communications Engineering
University of Bremen, Bremen, Germany
Email: {demel, tmonsees, bockelmann, wuebben, dekorsy}@ant.uni-bremen.de

***Abstract*—Industrial radio communication is identified as a new use case in the Industry 4.0 (I4.0) initiative as well as in the 3rd Generation Partnership Project (3GPP). 5th Generation (5G) Ultra Reliable Low Latency Communication (URLLC) requirements comprise high reliability and burst error resilience for short packets as well as low latency for I4.0 communication systems. We consider a Cloud Radio Access Network (Cloud RAN) architecture with distributed Radio Access Points (RAPs) that are connected via a rate limited fronthaul to a General Purpose Processor (GPP) cloud-platform. Thus, we can flexibly balance fronthaul data rates and joint processing gains to fully leverage spatial diversity. Here, we conduct an investigation on a functional split within the PHYsical layer (PHY) to harvest these benefits in the uplink while maintaining moderate data rates on the fronthaul for joint decoding. We investigate how data compression according to the Information Bottleneck Method (IBM) on the fronthaul link affects system performance for Generalized Frequency Division Multiplexing (GFDM) as well as OFDM. We show that 3 bit IBM quantization already achieves close to floating point performance in frequency-selective channels.**

***Index Terms*—Industry 4.0, 5G, NR, URLLC, multicarrier, polar code, IBM, functional split, fronthaul, quantization**

I. Introduction

Industrial radio communication is identified as a new use case in the Industry 4.0 (I4.0) initiative as well as in the 3rd Generation Partnership Project (3GPP) [1]. 5th Generation (5G) Ultra Reliable Low Latency Communication (URLLC) spawns a new set of requirements with highly reliable short packets, low latency and resilience to burst errors for I4.0 communication systems.

To achieve low latency as well as high reliability, we exploit spatial diversity in the uplink. We achieve spatial diversity in an Ultra Dense Network (UDN) where multiple Radio Access Points (RAPs) observe the same message through different channels. Further, we consider these distributed RAPs in a Cloud Radio Access Network (Cloud RAN) architecture [2], [3]. A Cloud RAN offers the flexibility to execute preprocessing on distributed RAPs that are connected via fronthauls to a centralized General Purpose Processor (GPP) cloud-platform. Joint signal processing on a cloud-platform promises to leverage spatial diversity and thus improve reliability. While full IQ sample forwarding drastically increases fronthaul data rates, we propose a functional split within the PHYsical layer (PHY) after the demapper step. With this architecture we can leverage joint processing gains in the upper PHY on a cloud-platform while distributed preprocessing in the lower PHY ensures achievable data rates on rate limited fronthauls [2].

Recently polar codes were adopted for 5G mobile communication systems [4]. Especially control channels and URLLC are anticipated use cases for polar codes as they promise good error correction performance for short packets. Further, low latency decoders for polar codes are available [5], e.g., Fast Simplified Successive Cancellation (Fast-SSC). We consider 5G Orthogonal Frequency Division Multiplexing (OFDM) multicarrier modulation. Further, we consider Generalized Frequency Division Multiplexing (GFDM) because it promises to reduce latency with only one Cyclic Prefix (CP) per frame and offers greater flexibility than OFDM [6]. Also, low latency GFDM implementations are known in literature [7].

Without further measures, we would need to forward Log-Likelihood Ratios (LLRs) with high resolution and thus high data rate over a fronthaul. Here, we investigate how to reduce this data rate burden by coarser quantization. Our approach is an offline design of scalar quantizers for the RAPs via Information Bottleneck Method (IBM) [8] to reduce the rate on the fronthaul link and to have a minimum loss in the e2e performance. In common quantizer design based on rate-distortion theory the objective is to obtain a compressed representation of the observation which does not exceed a predefined distortion between the input and the output of the quantizer such as Mean Square Error (MSE) . In contrast, IBM based quantizer design is a pertinent design approach which considers a suitable distortion measure for the communication setup, namely, the mutual information between the quantizer output and the original transmitted source signal. The IBM has been successfully utilized in different communication areas, such as the design of channel quantizers [9], polar code construction [10], relay networks [11], and advanced discrete decoder design [12]–[14].

Our main contribution is the investigation of IBM based quantization of multicarrier signals in order to reduce the fronthaul data rate in a Cloud RAN architecture. We demonstrated that the application of 3 bit IBM quantizers for eight different SNRs are sufficient to leverage most diversity in a frequency selective Rayleigh fading scenario.

This work was partly funded by the German ministry of education and research (BMBF) under grant 16KIS0720 (TACNET 4.0).

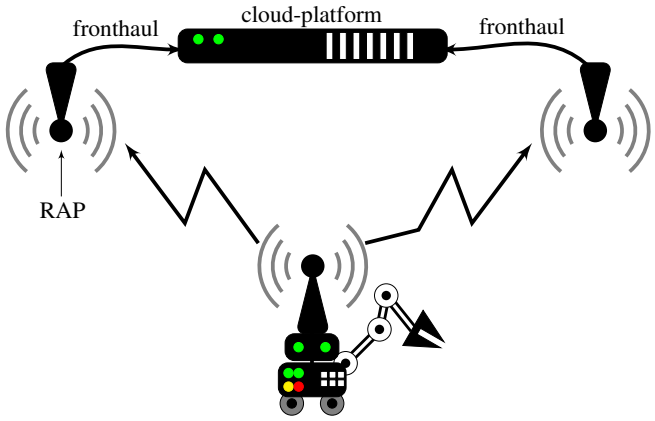

Fig. 1. Cloud RAN setup with fronthaul

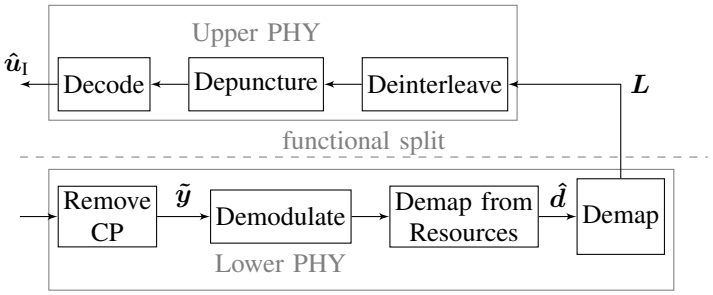

Fig. 2. PHY receiver processing where the demapper location is quantizer dependent

II. Cloud RAN

A. Architecture

In order to fulfill URLLC requirements in the uplink a Cloud RAN setup is considered to enable high reliability. In Fig. 1 we outline our architecture with distributed RAPs that are connected to a cloud-platform via a fronthauls to leverage spatial channel diversity from mobile units [2]. In case all RAPs forward their IQ samples as they are digitized, the fronthaul data rate would be very high and possibly infeasible for rate-limited fronthauls. In order to leverage joint processing gains to improve reliability while maintaining a moderate fronthaul data rate, we focus on a functional split within the PHY as indicated in Fig. 2. In each RAP we perform lower PHY processing distributedly, including synchronization and multicarrier demodulation. Next, a symbol demapper produces LLRs with high resolution that we want to forward to our cloud-platform which are then used for decoding. In Fig. 2 we indicate this functional split. We propose to use IBM in the RAPs to drastically reduce resolution and thus fronthaul data rates while preserving relevant information for joint decoding. On the cloud-platform we use Look-Up-Tables (LUTs) to obtain representative LLRs for joint upper PHY processing, mainly polar decoding. As a first step we consider one RAP and show that 3 bit IBM quantization is sufficient to almost achieve high resolution floating point decoding performance.

B. Polar Codes

Polar codes are first presented in [15]. Here, we want to summarize prior research that is relevant to our work. We consider a bit vector $\boldsymbol{u} \in \mathbb{F}_2^{N_\mathrm{c}}$ of size N_c and obtain the codeword

$$\boldsymbol{c} = \boldsymbol{u} \cdot \boldsymbol{B}_{N_\mathrm{c}} \cdot \boldsymbol{G} \quad \text{with} \quad \boldsymbol{G} = \boldsymbol{F}^{\otimes \log_2(N_\mathrm{c})} \tag{1}$$

where $\boldsymbol{B}_{N_\mathrm{c}}$ is a bit reversal matrix and $\boldsymbol{F}^{\otimes \log_2(N_\mathrm{c})}$ is the $\log_2(N_\mathrm{c})$th Kronecker product [15], [16]. The bit vector $\boldsymbol{u}$ consists of N_u information bits $\boldsymbol{u}_\mathrm{I} \in \mathbb{F}_2^{N_\mathrm{u}}$ and so called frozen bits $\boldsymbol{u}_\mathrm{Fr} \in \boldsymbol{0}^{N_\mathrm{c}-N_\mathrm{u}}$. The set of frozen bit positions $\mathcal{A}_\mathrm{Fr}$, with $|\mathcal{A}_\mathrm{Fr}| = N_\mathrm{c} - N_\mathrm{u}$, in a bit vector $\boldsymbol{u}$ is determined via polar channel construction. We refer to [17] and references therein for an in-depth discussion on possible strategies to obtain frozen bit positions $\mathcal{A}_\mathrm{Fr}$. We opt for the Bhattacharyya Bounds method because of its simplicity and good performance. Polar codes are then defined as $(N_\mathrm{c}, N_\mathrm{u}, \mathcal{A}_\mathrm{Fr})$ codes. Further, we consider systematic polar codes in our work [16].

Polar code codeword sizes are inherently restricted to a power of 2. We circumvent this obstacle with Frozen Quasi Uniform Puncturing (Frozen-QUP) [18] to obtain a punctured codeword $\boldsymbol{c}_\mathrm{P} \in \mathbb{B}^{N_\mathrm{P}}$. We employ a standard random interleaver to leverage diversity in order to obtain an interleaved bit vector $\boldsymbol{b} \in \mathbb{B}^{N_\mathrm{P}}$ from the punctured codeword $\boldsymbol{c}_\mathrm{P}$ [19]. In this paper, we consider QPSK where $D = 2$ bits are mapped to a symbol d from the symbol alphabet $\mathcal{D}$, $|\mathcal{D}| = 2^D$. Here, Gray labeling is assumed in order to optimize the Hamming distance between neighboring symbols. For the sake of simplicity we assume the mean signal power to be $\sigma_\mathrm{d}^2 = E\{|\mathrm{d}|^2\} = 1$.

In our work, we consider a Successive Cancellation (SC) polar decoder [15]. This SC decoder may be optimized to the Fast-SSC decoder in order to improve throughput and reduce latency [5].

C. Multicarrier Modulation

We consider Generalized Frequency Division Multiplexing (GFDM) and Orthogonal Frequency Division Multiplexing (OFDM) as multicarrier modulation schemes. We discuss the transmitter and detail the receiver as depicted in Fig. 2.

A frame with $N = MK$ elements spans over M timeslots and K subcarriers with K_on active subcarriers used for transmission. The resource grid for a multicarrier frame is represented by the matrix $\boldsymbol{D} \in \mathcal{D}^{M \times K}$ where each element $d_{m,k}$ corresponds to a symbol in the mth timeslot on the kth subcarrier [6]. With $\boldsymbol{d}_k \in \mathcal{D}^{M \times 1}$ we denote all elements on subcarrier k in a frame. In case $K_\mathrm{on} < K$, subcarriers which are unused correspond to columns in $\boldsymbol{D}$ which are filled with zeros and, consequently, the occupied bandwidth is reduced. Thus, the resource mapper groups $N_\mathrm{d} = MK_\mathrm{on}$ transmit symbols together with $M(K - K_\mathrm{on})$ zeros into $\boldsymbol{D}$.

The GFDM modulator computes

$$\tilde{\boldsymbol{x}} = \boldsymbol{A}_{N \times N} \boldsymbol{d} \tag{2}$$

where $\boldsymbol{d}$ is obtained by stacking $\boldsymbol{D}$'s columns and $\boldsymbol{A}_{N \times N}$ is the modulation matrix [6]. Matrix multiplication is an

expensive operation, especially when N tends to be large. We perform efficient frequency domain GFDM modulation and demodulation to drastically reduce complexity [20] and achieve low latency processing [7]. Therefore, we rewrite (2) to

$$\boldsymbol{x} = \sum_{k=0}^{K-1} \boldsymbol{P}^{(k)}_{N\times MN_{\text{ov}}} \boldsymbol{G}_{MN_{\text{ov}}\times MN_{\text{ov}}} \boldsymbol{R}_{MN_{\text{ov}}\times M} \boldsymbol{\mathcal{F}}_M \boldsymbol{d}_k \tag{3}$$

and further transform this frame $\boldsymbol{x}$ into time domain

$$\tilde{\boldsymbol{x}} = \boldsymbol{\mathcal{F}}_N^{-1} \boldsymbol{x} \, . \tag{4}$$

First, the symbols $\boldsymbol{d}_k$ are transformed to frequency domain with an M-point Discrete Fourier Transform (DFT) matrix $\boldsymbol{\mathcal{F}}_M$. Next, upsampling in frequency domain is performed by means of a repetition matrix $\boldsymbol{R}_{MN_{\text{ov}}\times M}$ with overlap factor $N_{\text{ov}} \leq K$. $\boldsymbol{G}_{MN_{\text{ov}}\times MN_{\text{ov}}}$ is a diagonal filter matrix with the MN_{ov} prototype filter taps $\boldsymbol{g} \in \mathbb{C}^{MN_{\text{ov}}\times 1}$ on its diagonal. $\boldsymbol{P}^{(k)}_{N\times MN_{\text{ov}}}$ performs subcarrier modulation by shifting the samples into a vector of size N at the corresponding position of the kth subcarrier. For $K = N_{\text{ov}}$, (3) is an alternative representation of (2). If the prototype filter is chosen such that its Out-Of-Band (OOB) leakage decays outside its subcarrier bandwidth, N_{ov} can become smaller than K. We use Root-Raised-Cosine (RRC) filters with roll-off factor $\alpha = 0.5$ and thus $N_{\text{ov}} = 2$ is sufficient. In this case, GFDM is a non-orthogonal modulation scheme and thus we must expect self-interference. OFDM is a special case of GFDM with $M = 1$ timeslot and the prototype filter is fixed to a rectangular filter in time domain. With GFDM we shorten the frame duration and thus latency because we only use one CP per frame in contrast to OFDM where we use one CP per timeslot.

D. Channel model

The signal $\tilde{\boldsymbol{x}}$ in time domain is transmitted with a CP over the channel. At the receiver, after CP removal and a DFT, we denote the received signal

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n} \tag{5}$$

in frequency domain where $\boldsymbol{H}$ is a $N \times N$ diagonal matrix. Further, $n \sim \mathcal{CN}(0, \sigma^2_{\text{n}})$ is Additive White Gaussian Noise (AWGN) and we denote our Signal-to-Noise-Ratio (SNR) as

$$\text{SNR} = \frac{E\{|\text{d}|^2\}}{E\{|\text{n}|^2\}} = \frac{\sigma^2_{\text{d}}}{\sigma^2_{\text{n}}} \, . \tag{6}$$

At the receiver we perform one-tap Zero-Forcing (ZF) frequency domain channel equalization. In case of GFDM, and thus a non-orthogonal modulation, we demodulate and employ Interference-Cancellation (IC) in order to mitigate self-interference after equalization [20]. After demodulation and a subsequent resource demapper, we obtain the soft estimate $\hat{\boldsymbol{d}}$ which is fed into a quantizer or symbol demapper, detailed in Sec. III.

We introduce the equivalent subcarrier channel

$$\hat{d} = h_k d + n \tag{7}$$

and denote the per subcarrier Carrier-to-Noise-Ratio (CNR) as

$$\text{CNR}_k = \frac{|h_k|^2 \sigma^2_{\text{d}}}{\sigma^2_{\text{n}}} \tag{8}$$

for a fixed channel realization. CNR_k characterizes the equivalent channel (7) on subcarrier k with transition probability $p_k(\hat{d}|d)$ for designing the subsequent processing steps in Sec. III.

Industrial radio measurement campaigns [21] show that all time-domain channel taps $|\tilde{\text{h}}|$, regardless if they are Line-Of-Sight (LOS) or Non-Line-Of-Sight (NLOS), are Rayleigh distributed. Also, the Power Delay Profile (PDP) $\boldsymbol{p}$ of the channel follows an exponential distribution with a maximum channel delay $\tau_{\text{max}} < \tau_{\text{CP}}$. Thus, we obtain a time domain channel realization $\tilde{\boldsymbol{h}}$ with its elements $\tilde{h}_i$ drawn from $\mathcal{CN}(0, p_i^2)$. Finally, frequency-domain channel taps are obtained as $\boldsymbol{h} = \mathcal{F}_N \tilde{\boldsymbol{h}}$ with $E\{|\text{h}_k|^2\} = 1$. We assume perfect system synchronization and a block fading channel with perfect channel knowledge, i.e. the channel is constant over the duration of a frame and thus over all frame timeslots. Thus, each subcarrier in our frequency-domain channel model is affected by only one channel tap h_k.

III. Fronthaul Signal Generation

A. Log-Likelihood Ratios

After processing the received signals in the RAP the soft estimates $\hat{\boldsymbol{d}}$ have been calculated and need to be forwarded to the cloud-platform for joint decoding. Instead of forwarding corresponding IQ samples, we may exchange the LLRs determined by symbol demapping [2]. With given statistic $p_k(\hat{d}|d)$ of the complex channel for subcarrier k and assuming equiprobable symbols d, the LLR for a corresponding codebit b_i is given by

$$L(b_i|\hat{d}) = \ln \frac{\sum_{d\in\mathcal{D}_i^0} p_k(\hat{d}|d)}{\sum_{d\in\mathcal{D}_i^1} p_k(\hat{d}|d)} \, , \tag{9}$$

where $\mathcal{D}_i^\nu$ is the set of symbols where the i-th bit is $\nu = \{0, 1\}$. Here, we assume that the complex channel $p_k(\hat{d}|d)$ is determined by the equivalent channel for each subcarrier k with fading coefficient h_k and noise variance σ^2_{n} as defined in (7) with the CNR_k in (8). After forwarding all LLRs $\boldsymbol{L}(\boldsymbol{b}|\hat{\boldsymbol{d}})$ to the cloud-platform, they are deinterleaved and depunctured to obtain the codeword LLRs $\boldsymbol{L}(\boldsymbol{c})$.

Forwarding the LLRs (9) in floating point precision to the cloud-platform would result in a very high data rate on the fronthaul. Thus, in order to reduce the fronthaul rate it is common to transmit quantized LLRs by the RAP [2]. Here, we assume uniform quantization of (9) in the intervall $[-8,\ 8]$ with $N_Z = \{3, 4\}$ bit, i.e., using 8 or 16 quanitzation levels.

B. Information Bottleneck based Quantizer Design

In order to improve the e2e performance, we propose to use the more sophisticated IBM quantization to generate the message to be forwarded by the fronthaul. To this end, we describe subsequently the offline design of CNR dependent

quantizers per subcarrier in the RAP via IBM, where we assume an AWGN model for the design with CNR = SNR. Similar to the calculation of LLRs in (9) this design approach requires knowledge about the CNR dependent joint distribution $p_k(d, \hat{d})$.

For Quadrature Amplitude Modulation (QAM) constellations with Gray labeling, the real and imaginary part are independent of each other and can be quantized separately. Subsequently, we use the variables d_a and $\hat{d}_a$ to denote either the real or the imaginary part of d and $\hat{d}$, respectively, with the same conditioned probability $p_k(\mathrm{Re}\{\hat{d}\}|\mathrm{Re}\{d\}) = p_k(\mathrm{Im}\{\hat{d}\}|\mathrm{Im}\{d\}) = p_k(\hat{d}_a|d_a)$ with noise variance $0.5\sigma_\mathrm{n}^2$. The underlying optimization task of IBM based quantizer design is to optimize a quantizer mapping $z = Q_k^\star(\hat{d}_a)$ from the noisy observation $\hat{d}_a$ into a compressed representation $z \in \mathcal{Z}$ containing a maximum amount of relevant information $I(\mathsf{z}; \mathsf{d}_a)$ for the original source signal $d_a \in \{-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\}$, i.e.

$$Q_k^\star = \underset{Q_k}{\operatorname{argmax}}\, I(\mathsf{z}; \mathsf{d}_a) \;\text{ s.t. }\; |\mathcal{Z}| \leq Z\,. \tag{10}$$

Notice, that the value of the compressed variable z is not of importance at all as only its probability measures. Thus, we simply assume unsigned integer values, $\mathcal{Z} = \{0, 1, ..., |\mathcal{Z}|-1\}$. The optimization problem (10) is a special case of the IBM optimization problem ($\beta \to \infty$ in [8]). For the system under investigation we can utilize the optimum quantizer design for binary input signals [22] with the joint distribution $p_k(d_a, \hat{d}_a) = p_k(\hat{d}_a|d_a)p_k(d_a)$ and equiprobable $p_k(d_a)$ to obtain the quantizer mapping $Q_k^\star(\hat{d}_a)$. In this case, we limit the compression rate $I(\hat{\mathsf{d}}_a; \mathsf{z})$ and the resulting fronthaul rate by a small number of bits $N_Z = \log_2(|\mathcal{Z}|)$ with $N_Z = \{3, 4\}$.

Since the IBM based quantizer design depends on the CNR_k of the subcarrier k, we design the quantizers offline for a uniform grid of CNR_k values (e.g. $N_\mathrm{Q} = 32$ quantizer mappings from $-4.5\,\mathrm{dB}$ to $11\,\mathrm{dB}$ in $0.5\,\mathrm{dB}$ steps). The RAP has to transmit the representative z (or its binary representation) together with $\log_2(N_\mathrm{Q})$ bits of overhead information about the used quantizer mapping to the cloud-platform. We assume that the cloud-platform has a CNR_k dependent LUT for each subcarrier k with the LLRs given by

$$L(b_i|z) = \ln \frac{p_k(z|b_i = 0)}{p_k(z|b_i = 1)}. \tag{11}$$

At the cloud-platform, the discrete LLRs are deinterleaved and depunctured for further decoding. Alternatively, IBM-based discrete decoder implementations similar to [23] could be applied directly on the compressed variables z.

As already pointed out, the RAP needs to inform the cloud-platform about the used quantizer per subcarrier, i.e., about the CNR_k. In order to reduce this overhead, we investigate the performance with a reduced number of CNR_k dependent quantizer mappings (i.e. $N_Q \in \{32, 16, 8, 4, 2\}$) per subcarrier in Section IV.

IV. Numerical Results

We evaluate the proposed Cloud RAN architecture through a series of simulations. Here, we focus on the impact of coarser quantization on system performance. A lower resolution eases the data rate burden and also contributes to lower latency on rate limited fronthauls. All simulations are run with $K = 64$ subcarriers, $K_\mathrm{on} = 50$ active subcarriers, $M = 5$ timeslots, $N_\mathrm{u} = 256$ information bits, $N_\mathrm{p} = 500$ punctured code bits, and Quadrature Phase Shift Keying (QPSK) modulation under the assumption that one RAP is used.

First, we want to learn about the expected performance in an AWGN channel serving as a benchmark where the CNR is identical on all subcarriers. In Fig. 4 we observe that OFDM outperforms GFDM by $0.5\,\mathrm{dB}$ in an AWGN scenario due to non-orthogonal GFDM. Moreover, we observe that even a $3\,\mathrm{bit}$ IBM quantizer only incurs a minor $0.15\,\mathrm{dB}$

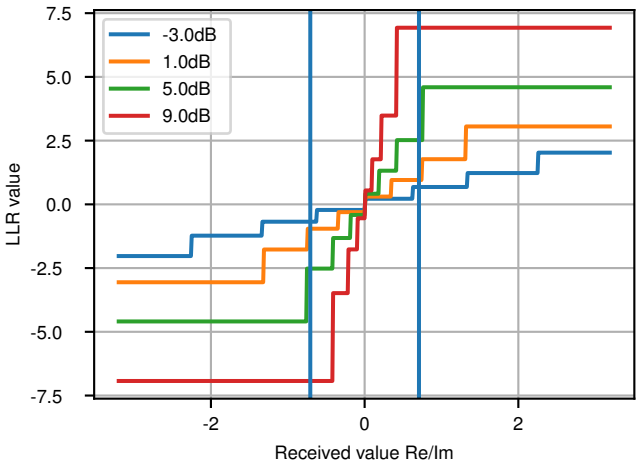

Fig. 3. Discrete LLRs based on (11) for the optimized quantizer mapping $z = Q_k^\star(\hat{d}_a)$ for different CNR values. Vertical lines represent real or imaginary part of the QPSK symbol.

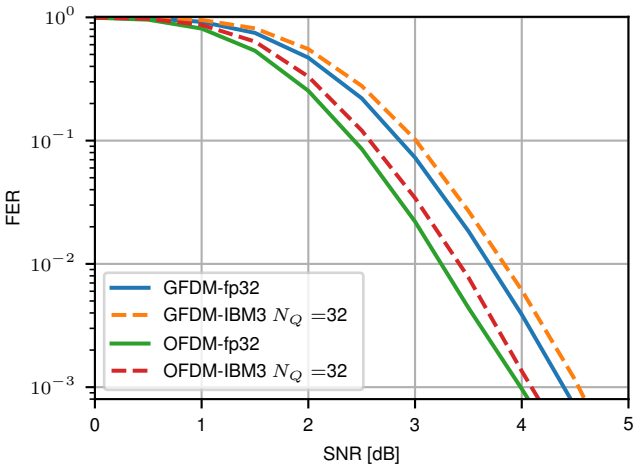

Fig. 4. AWGN FER performance for 3 bit IBM quantization and fp32 LLR forwarding

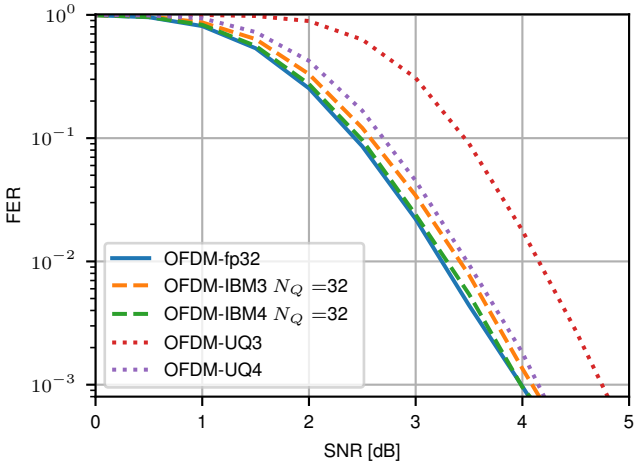

Fig. 5. AWGN OFDM FER performance for various quantization strategies

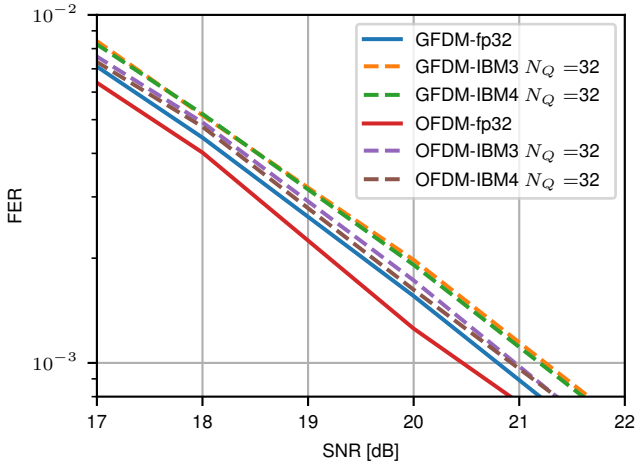

Fig. 6. FER performance for frequency-selective channels with different quantizers

performance hit compared to LLR forwarding with floating point (fp32) precision. This result indicates that a fronthaul data rate reduction is possible while maintaining performance.

Next, we evaluate the performance of different quantization strategies in an AWGN channel. In Fig. 5 we focus on OFDM and note that the results are also valid for GFDM. We observe that $4\,\mathrm{bit}$ IBM quantization only incurs a minimal performance loss compared to fp32 LLR forwarding. Further, we observe that $3\,\mathrm{bit}$ IBM quantization delivers better performance than $4\,\mathrm{bit}$ uniform quantization. With $3\,\mathrm{bit}$ uniform quantization we observe a larger $0.75\,\mathrm{dB}$ performance loss.

As a first result, we conclude that IBM clearly outperforms uniform quantization. Further, we conclude that $3\,\mathrm{bit}$ IBM quantization performs close to fp32 LLR forwarding and thus enables lower fronthaul rates. $4\,\mathrm{bit}$ IBM quantization is sufficient to close this performance gap.

In Fig. 6 we compare different IBM quantizers for GFDM and OFDM with frequency-selective Rayleigh fading. We observe that $3\,\mathrm{bit}$ and $4\,\mathrm{bit}$ IBM quantization deliver almost the same performance unlike in an AWGN channel. We reckon that the observed performance loss is caused by the used quantization tables that only go up to $11\,\mathrm{dB}$ and will be the subject of future research. Further, we observe the same minor performance loss for GFDM compared to OFDM due to self-interference. We note that GFDM offers better efficiency compared to OFDM and potentially lower latency [6].

Further, we analyze the impact of different quantizers in a frequency-selective Rayleigh fading setup for GFDM. In Fig. 7 we observe a $0.4\,\mathrm{dB}$ performance loss at FER 10^{-3} for $4\,\mathrm{bit}$ IBM quantization. Further, $3\,\mathrm{bit}$ IBM compared to $4\,\mathrm{bit}$ IBM quantization only cause another $0.05\,\mathrm{dB}$ performance loss. Again, we reckon that this tiny gap will widen if we employ quantization tables that go beyond $11\,\mathrm{dB}$. Further, we observe that IBM quantization clearly outperforms uniform quantization.

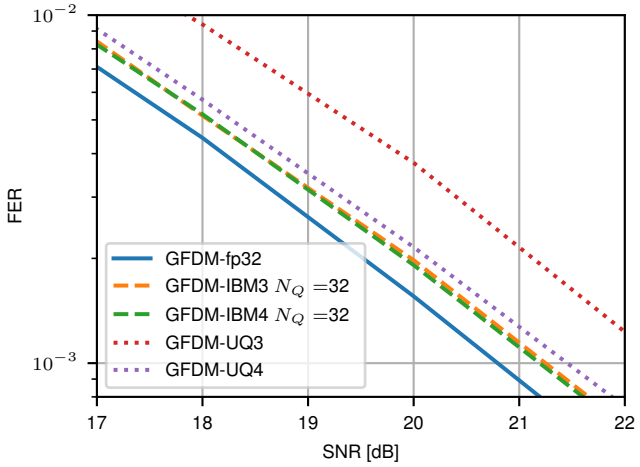

Fig. 7. GFDM FER performance for frequency-selective channels

Distributed RAPs need to signal the employed quantization table indices to the cloud platform per subcarrier which causes overhead. In our setup with $K_{\mathrm{on}} = 50$ active subcarriers and $N_{\mathrm{Q}} = 32$ quantization table indices this causes $\log_2(N_{\mathrm{Q}})K_{\mathrm{on}} = 250\,\mathrm{bit}$ overhead together with $N_{\mathrm{p}}N_{\mathrm{Z}} = 1500\,\mathrm{bit}$ for quantized symbols per frame. In case of $4\,\mathrm{bit}$ uniform quantization, a quantizer yields $2000\,\mathrm{bit}$ per frame and thus a $3\,\mathrm{bit}$ IBM quantizer is advantageous because it requires a lower data rate and shows better performance.

We want to minimize this overhead by reducing the number of indices N_{Q}. In Fig. 8 we observe that $N_{\mathrm{Q}} = 8$ indices are sufficient to achieve a similar performance as with $N_{\mathrm{Q}} = 32$ quantization tables. 2 quantization tables are insufficient but 4 quantization tables only constitute a minor performance loss. Fewer quantization tables, e.g. 8 and thus

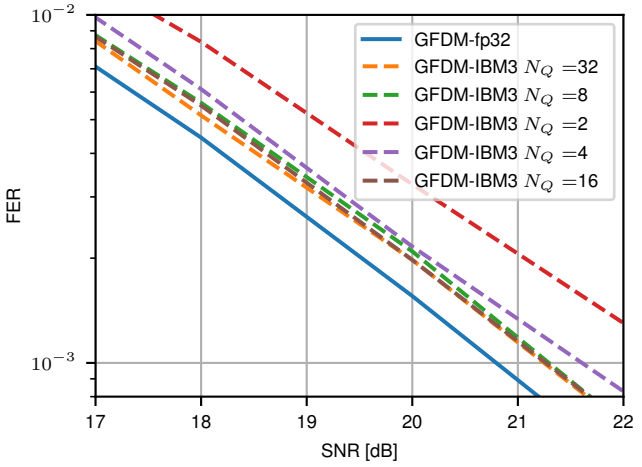

Fig. 8. GFDM FER performance for frequency-selective channels with different number of indices N_{Q}

only $\log_2(N_{\mathrm{Q}})K_{\mathrm{on}} = 150\,\mathrm{bit}$ frame overhead, further ease the fronthaul data rate burden. In a future work we will consider quantization tables for groups of subcarriers instead of per-subcarrier to further reduce this overhead.

V. Conclusion

We propose IBM quantization for multicarrier systems in a Cloud RAN architecture with rate limited fronthauls. As a first step we investigate a PHY split for Cloud RAN with a quantizer after the demapper step to reduce fronthaul data rates. We conclude that coarse $3\,\mathrm{bit}$ IBM quantization for a rate limited fronthaul is sufficient to achieve a performance close to a high resolution LLR forwarding. This observation holds for both AWGN as well as frequency-selective Rayleigh fading channels. Thus, we deduce that IBM quantization is a superior choice to lower data rates on fronthauls in a wireless communication system. Especially URLLC communication which is mainly limited by shadowing and fading can benefit from a Cloud RAN architecture with multiple spatially separated cooperating RAPs.

References

[1] A. Osseiran, J. F. Monserrat, and P. Marsch, *5G Mobile and Wireless Communications Technology*, 1st ed. New York, NY, USA: Cambridge University Press, 2016.

[2] D. Wübben, P. Rost, J. S. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis, "Benefits and impact of cloud computing on 5G signal processing: Flexible centralization through cloud-RAN," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 35–44, 2014.

[3] P. Rost, I. Berberana, A. Maeder, H. Paul, V. Suryaprakash, M. Valenti, D. Wübben, A. Dekorsy, and G. Fettweis, "Benefits and challenges of virtualization in 5G radio access networks," in *IEEE Communications Magazine*, vol. 53, no. 12. IEEE, 2015, pp. 75–82.

[4] V. Bioglio, C. Condo, and I. Land, "Design of Polar Codes in 5G New Radio," *arXiv*, no. 1804.04389, pp. 1–9, 2018.

[5] P. Giard, "High-Speed Decoders for Polar Codes," Ph.D. dissertation, McGill University, Montreal, Canada, 2016. [Online]. Available: http://digitool.library.mcgill.ca/R/-?func=dbin-jump-full{&}object{_}id=145447

[6] N. Michailow, M. Matthe, I. S. Gaspar, A. N. Caldevilla, L. L. Mendes, A. Festag, and G. Fettweis, "Generalized frequency division multiplexing for 5th generation cellular networks," *IEEE Transactions on Communications*, vol. 62, no. 9, pp. 3045 – 3061, 2014.

[7] J. Demel, C. Bockelmann, and A. Dekorsy, "Evaluation of a Software Defined GFDM Implementation for Industry 4.0 Applications," in *Proceedings of the 2017 IEEE International Conference on Industrial Technology*, Toronto, Canada, 2017.

[8] N. Tishby, F. C. Pereira, and W. Bialek, "The Information Bottleneck Method," in *37th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, Sep. 1999, p. 368–377.

[9] S. Hassanpour, D. Wübben, and A. Dekorsy, "Overview and Investigation of Algorithms for the Information Bottleneck Method," in *11th Int. Conference on Systems, Communications and Coding (SCC)*, Hamburg, Germany, Feb. 2017.

[10] M. Stark, A. Shah, and G. Bauch, "Polar code construction using the information bottleneck method," in *2018 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, April 2018.

[11] G. Caire, S. Shamai, A. Tulino, S. Verdu, and C. Yapar, "Information bottleneck for an oblivious relay with channel state information: the scalar case," in *2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE)*, Dec 2018, pp. 1–5.

[12] J. Lewandowsky and G. Bauch, "Information-Optimum LDPC Decoders Based on the Information Bottleneck Method," *IEEE Access*, vol. 6, pp. 4054–4071, 2018.

[13] B. M. Kurkoski, K. Yamaguchi, and K. Kobayashi, "Noise Thresholds for Discrete LDPC Decoding Mappings," in *IEEE GLOBECOM 2008 - 2008 IEEE Global Telecommunications Conference*, Nov 2008, pp. 1–5.

[14] A. Balatsoukas-Stimming, P. Giard, and A. Burg, "Comparison of Polar Decoders with Existing Low-Density Parity-Check and Turbo Decoders," in *2017 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, 2017.

[15] E. Arıkan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Transactions on Information Theory*, 2009.

[16] G. Sarkis, I. Tal, P. Giard, A. Vardy, C. Thibeault, and W. J. Gross, "Flexible and low-complexity encoding and decoding of systematic polar codes," *IEEE Transactions on Communications*, vol. 64, no. 7, pp. 2732 – 2745, 2016.

[17] H. Vangala, E. Viterbo, and Y. Hong, "A Comparative Study of Polar Code Constructions for the AWGN Channel," *arXiv*, 2015. [Online]. Available: https://arxiv.org/pdf/1501.02473.pdf

[18] J. Demel, C. Bockelmann, and A. Dekorsy, "Industrial Radio Link Abstraction Models for Short Packet Communication with Polar Codes," in *International ITG Conference on Systems, Communications and Coding (SCC)*, Rostock, 2019, pp. 257–262.

[19] G. Caire, G. Taricco, and E. Biglieri, "Bit-Interleaved Coded Modulation," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 927–946, 1998.

[20] I. Gaspar, N. Michailow, A. Navarro, E. Ohlmer, S. Krone, and G. Fettweis, "Low complexity GFDM receiver based on sparse frequency domain processing," in *IEEE Vehicular Technology Conference (VTC Spring)*, Dresden, Germany, 2013.

[21] M. Düngen, T. Hansen, R. Croonenbroeck, R. Kays, B. Holfeld, D. Wieruch, P. W. Berenguer, V. Jungnickel, D. Block, and U. Meier, "Channel measurement campaigns for wireless industrial automation," *at - Automatisierungstechnik*, vol. 67, no. 1, pp. 7–28, 2019.

[22] B. M. Kurkoski and H. Yagi, "Quantization of Binary-Input Discrete Memoryless Channels," *IEEE Transactions on Information Theory*, vol. 60, no. 8, pp. 4544–4552, Aug 2014.

[23] T. Monsees, D. Wübben, and A. Dekorsy, "Channel-Optimized Information Bottleneck Design for Signal Forwarding and Discrete Decoding in Cloud-RAN," in *12th International ITG Conference on Systems, Communications and Coding (SCC)*, Rostock, Germany, Feb 2019.