Quantization Effects of twiddle factors in FFT for Beamforming Using Planar Arrays

Juan A. Vásquez-Peralvo
SnT, Uni. of Luxembourg
Luxembourg, Luxembourg
0000-0001-7304-095X

Juan Carlos Merlano Duncan
SnT, Uni. of Luxembourg
Luxembourg, Luxembourg
0000-0002-9652-679X

Vu Nguyen Ha
SnT, Uni. of Luxembourg
Luxembourg, Luxembourg
0000-0003-1325-3480

Rakesh Palisetty
Shiv Nadar Institution of Eminence
Delhi, India
0000-0003-3222-6576

Geoffrey Eappen
SnT, Uni. of Luxembourg
Luxembourg, Luxembourg
0000-0002-4065-3626

Luis Manuel Garcés-Socarrás
SnT, Uni. of Luxembourg
Luxembourg, Luxembourg
0000-0002-4164-3673

Jorge Luis González Rios
SnT, Uni. of Luxembourg
Luxembourg, Luxembourg
0000-0003-4415-9649

Paul H. Aldás Ponce
Departamento Ciencias, de la Ingeniería
Universidad Israel
Quito, Ecuador
0009-0009-9890-7105

Symeon Chatzinotas
SnT, Uni. of Luxembourg
Luxembourg, Luxembourg
0000-0001-5122-0001

***Abstract*—This paper examines the impact of quantization on the twiddle factors in the Fast Fourier Transform (FFT) used for beamforming applications. A comprehensive scenario utilizing a Medium Earth Orbit (MEO) satellite system is developed to analyze these effects. The setup comprises a 10×10 phased array antenna, where each antenna element consists of a subarray of 5×5 elements spaced at half-wavelength intervals, all distributed in a circular aperture. We investigate the influence of 1, 2, 3, and 4-bit quantization on various aspects of system performance, including the overall radiation pattern, the grid of beams generated by a 16-point FFT, and the Signal-to-Interference Ratio (SIR) in a multibeam scenario. Our findings provide insight into the trade-offs between quantization resolution and beamforming efficacy, highlighting the critical role of twiddle factor precision in optimizing satellite communication systems. Finally a set of recommendations is given for selecting the quantization that is most suitable for different cases.**

***Index Terms*—Phase arrays, FFT, Digital Beamforming.**

I. Introduction

Digital beamforming has become a focal point in recent years, driven by the commercial viability of fully controllable phased arrays. This development enables the practical simulation, design, and application of numerous beamforming algorithms that were previously theoretical. These algorithms, however, present challenges in terms of power consumption, multi-beam efficacy, and flexibility. The impact of these challenges on various beamforming techniques is summarized in Table I, highlighting the computational complexities associated with each technique [1].

This work was supported by European Space Agency under the project number 4000134678/21/UK/AL "EFFICIENT DIGITAL BEAMFORMING TECHNIQUES FOR ON-BOARD DIGITAL PROCESSORS (EGERTON)" (Opinions, interpretations, recommendations and conclusions presented in this paper are those of the authors and are not necessarily endorsed by the European Space Agency). This work was supported by the Luxembourg National Research Fund (FNR), through the CORE Project (C^3): Cosmic Communication Construction under Grant C23/IS/18116142/C3.

TABLE I: Computational Requirements, Multi-Beam Efficacy, and Flexibility of Digital Beamforming Algorithms

Beamforming Algorithm	Comp. Req.	Multi-Beam Efficacy	Flexibility
Delay-and-Sum Beamforming	Very Low	Low	Low
FFT-Based Beamforming	Low	High	Medium
Zero-Forcing Beamforming	Medium	High	Low
MVDR Beamforming	Medium	Medium	High
LCMV Beamforming	High	High	High
Eigenbeamforming	High	Medium	Medium
Maximum Likelihood Beamforming	Very High	Low	Medium

Fast Fourier Transform (FFT)-based beamforming is highlighted in Table I as a particularly efficient technique, characterized by low power consumption and capability to support multi-beam operations with both regular [2] and potentially irregular lattice structures [3]. Moreover, its implementation becames even more efficient by replacing the FFT by the Discrite Furier Transform (DFT) to obtain the beamforming matrix [4]. Despite the advantages, a significant challenge persists in implementing these digital beamforming algorithms in phased array antennas, particularly due to the use of phase shifters.

Various implementations of FFT have been proposed in the literature. For instance, Ariyarathna et al. introduce a 16-point DFT approximation method that effectively eliminates the need for multipliers, typically a major contributor to power consumption and chip area in digital beamforming systems [5]. Another application of FFT beamforming is presented by Madanayake et al., where the authors focus on developing a digital beamforming architecture optimized for size, weight,

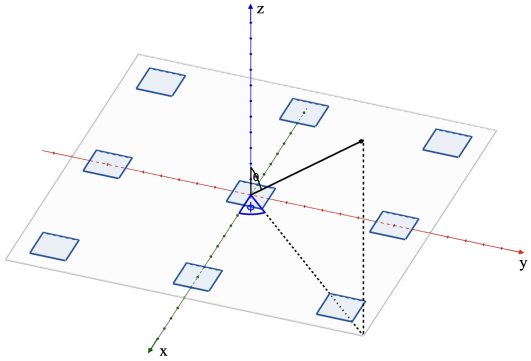

Fig. 1: Coordinate system used for the mathematical formulation.

power, and cost, crucial for modern wireless communication systems. This study introduces a multiplication-free 32-point linear transform approach that significantly reduces the arithmetic complexity of beamforming, compared to traditional methods, and decreases the chip area and power consumption by 46% and 55%, respectively [6]. The last work presented in [7] implements a beamforming network using a sparse-matrix-based user selection, a 2D DFT-based digital beam generation, which is implemented by a FFT algorithm, and a spatial windowing module for selecting the antenna pattern. However, neither of these works addresses the effects of quantization over the multi-beam scenarios they present.

This paper analyzes the impact of quantizing the twiddle factors in the FFT on beam distribution, beam pattern, and Signal-to-Interference Ratio (SIR) in multi-beam scenarios. The resolution (in bits) of these devices often limits their effectiveness, underscoring the importance of quantization analysis in multi-beam applications.

II. Mathematical Formulation

The coordinate system used in this study is illustrated in Fig. 1. Using this coordinate system, we can derive the unit vector representing the direction of incoming or outgoing electromagnetic waves as shown in the following.

$$\overrightarrow{r}_m = \sin\theta_m \cos\phi_m \overrightarrow{x} + \sin\theta_m \sin\phi_m \overrightarrow{y} + \cos\theta_m \overrightarrow{z}, \quad (1)$$

where $\overrightarrow{x} = (1,0,0)$, $\overrightarrow{y} = (0,1,0)$, and $\overrightarrow{z} = (0,0,1)$ represent the unit vectors according to $x-, y-, z-$ axes, respectively. Additionally, the positions of the antenna elements are defined accordingly to $\overrightarrow{x}, \overrightarrow{y}, \overrightarrow{z}$ as

$$\hat{p}_{[m,n,o])} = (0,0,0) + md_x \overrightarrow{x} + nd_y \overrightarrow{y} + od_z \overrightarrow{z}. \quad (2)$$

For planar arrays, we set the z-direction elements to zero. Then, $\hat{p}_{m,n,o}$ can be simplified as

$$\hat{p}_{[m,n,o]} \equiv \hat{p}_{[m,n]} = md_x \overrightarrow{x} + nd_y \overrightarrow{y}. \quad (3)$$

Based on these definitions, we compute the weight matrix for a beam pointing in the q-direction represented by $\overrightarrow{r}_q$, as

$$b_{[m,n]}(\overrightarrow{r}_q) \approx e^{+j\beta \hat{p}^T_{[m,n]} \overrightarrow{r}_q}. \quad (4)$$

Thanks to this framework, one can derive the expressions used to calculate the final frequency spectrum as

$$Y_{[o,q]}(t) = \sum_{m=0}^{M} \sum_{n=0}^{N} X_{[m,n,o,q]}(t) e^{-i2\pi\left(\frac{mo}{M} + \frac{np}{N}\right)}. \quad (5)$$

Here, X denotes the symbols to be transmitted, while o and p are indices for the desired beam. This expression represents the well-known FFT. Lastly, the radiation pattern is computed using the following array factor (AF) formula:

$$\mathrm{AF} = \sum_{m=1}^{M_x} \sum_{n=1}^{M_y} |A_{m,n}| W_{[m,n,o,q]} \mathrm{e}^{j(m)(\beta d_x \sin(\theta)\cos(\phi))} \mathrm{e}^{j(n)(\beta d_y \sin(\theta)\sin(\phi))} \quad (6)$$

where $A_{m,n}$ represents the amplitude weight matrix for each antenna element, β represents the wave-number, and $W_{[m,n,o,q]}$ is the FFT beamforming matrix. Herein, $W_{[m,n,o,q]}$ can be defined as

$$W_{[m,n,o,q]} = ZX_{[m,n,o,q]}Z^T, \quad (7)$$

where Z denotes the quantized matrix of Q-points containing the required progressive phase shift, and $X_{[m,n,o,q]}$ is the single entry matrix that specifies the active beam location. The result of this equation will be quantized for $1, 2, 3$, and 4 bits in the subsequent subsections.

III. Fourier Matrix Approximation by Quantized Twiddle Factors

In the FFT, the Fourier matrix Z_N, essential for efficiently computing the DFT, is decomposed into several simpler matrices. This decomposition involves repeatedly applying a block diagonal butterfly matrix B, along with multiple permutation matrices and twiddle factor matrices, tailored to a radix-R FFT algorithm.

The matrix decomposition of Z can be succinctly represented as:

$$Z_N = P_1 B T_1 P_2 B T_2 \cdots P_{\log_R N - 1} B T_{\log_R N - 1} P_{\log_R N} B$$

In this decomposition:

- P_k denotes the permutation matrices at each stage k, which reorder the data to ensure that inputs are correctly aligned for the butterfly operations that follow. These permutations are crucial as they reconfigure the input array's ordering at each stage of the transform.
- B is a block diagonal butterfly matrix applied at each stage, structured as 2×2 blocks in radix-2 FFT and 4×4 blocks in radix-4 FFT. Being each block the Fourier matrix of size 2 and size 4. Note that the butterfly operations only require four summations per output in Radix-4.
- T_k represents the twiddle factor matrices at each stage k up to $\log_R N - 1$, which are diagonal and contain elements of the form $e^{-2\pi ik/N}$. These matrices adjust the phase of the DFT outputs, aligning them correctly for subsequent combinations.

The process involves multiple stages, each of which applies a butterfly operation to combine DFT outputs followed by a phase adjustment using twiddle factors. After each stage, a permutation is necessary to rearrange the elements before the next stage begins. The systematic application of these matrices reduces the computational complexity from $O(N^2)$ in a straightforward DFT to $O(N \log_R N)$ in the FFT, significantly enhancing the efficiency for large-scale computations. This matrix-based framework not only simplifies the understanding of the FFT's operational structure but also highlights the algorithm's efficiency in hardware implementations. The twiddle factor block is the only part of the algorithm that requires a variable-by-constant multiplication. The efficiency of this block might be further improved for highly quantized twiddle factor constants. In such cases, the twidle multiplication operates by a limited number of adders, which highly improves the operation efficiency [8].

IV. Simulation

The simulation scenario, depicted in Figure 2, progresses through five distinct stages:

1) The initial stage involves designing the unit cell antenna, specifically employing a double patch antenna with an integrated cavity to diminish mutual coupling among antennas. In addition, the unit cell has been designed to have a period $p = \lambda_0/2$. A comprehensive, step-by-step guide on the antenna's design is available in [9].
2) Subsequently, the process involves assembling the unit cells into a subarray configuration. While this arrangement is known to have certain limitations, most notably, the presence of grating lobes, its application is justified for MEO satellite orbits. Employing subarrays enables a reduction in the number of RF chains, as detailed in [10], and also decreases computational time complexity After calculations, we have selected a 5×5 elements subarray, which give us an inter-subarray space $p_{\text{sub}} = 2.5\lambda_0$.
3) The following stage entails calculating the contribution of each antenna, our case 100 antennas, using the array factor with its corresponding spacial widowing and tapering, to the beam angles of interest, incorporating the W matrix, which was determined through an N-point FFT. In this scenario we have selected a circular widowing meaning that the 100 elements will be distributed in a circular aperture. Next, the quantity of beams generated correlates directly with the number of points in the FFT. Since for this example we have selected $O = P = 16$ beams, we will have a total of 256 beams.
4) A comprehensive representation of all beam contributions across the specified directions is subsequently obtained (full radiation patterns). This representation facilitates the assessment of quantization effects in three distinct domains: the beams per se, the overall beam projections, and the SIR.

For easiness of representation, we will use a u-v coordinate system, where:

$$u = \sin\theta\cos\phi; v = \sin\theta\sin\phi \tag{8}$$

The subsequent sections will detailed into the implications of quantization of the twiddle factors.

V. Quantization effects

1) Quantization effect on the radiation pattern: The effects on the radiation pattern has been assessed by plotting the the beam corresponding to the 2DFFT index (4, 4) for the case of 1, 2, 3, and 4 bit quantization and compared with the case of no quantization illustrated in Figure 3. For the 1-bit quantization case, the radiation pattern obtained is illustrated in Figure 4. Compared with the no quantization case we can see that the radiation pattern has high Side Lobe Levels (SLL) due to the leading imperfect constructive and destructive interference pattern for the low quantization. However, this destructive interference does not generate grating lobes even though the high inter-element space. The error calculated using this and the no-quantized scenario is illustrated in Figure 4b. We can appreciate that the error is acceptable even though this is the one bit quantization case.

In the case of 2-bit quantization, illustrated in Figure 5a, the error obtained in this case is visually depicted in Figure 5b. The error maxima has considerable decrease to around -10 dBs.

For 3-bit quantization the result radiation pattern obtained, visually depicted in Figure 6a, shows that the biggest difference visually noticeable are the side lobes variation. This variation can be more noticeable analyzing the error illustration presented in Figure 6b.

For the last case scenario Figure 7a, the difference between the non-quantized is practicable not noticeable as it is visually depicted in the error illustration, presented in Figure 7b. Moreover, the values of error obtained in this scenario reach -30 dB which corroborates the small difference between 4-bit quantized scenario and non-quantized. Finally, it is worth mentioning that if we analyze a beam that is further away from broadside, the errors will increased due to bigger discontinuities in the phase distribution across the array.

A. Quantization Effect on Beam Projection

The beam projection is assessed by plotting the -3 dB contours of each of the 256 beams resulting from the FFT beamforming. All the results show that there are no secondary lobes that reaches values near or compared to the main beam. This can be seen by plotting the no quantization case illustrated in Figure 8 and the worst case that is the 1-bit quantization case illustrated in Figure 9.

B. Quantization Effect on the SIR

The SIR for each multi-beamforming scenario is calculated by determining the maximum power at each grid point divided by the overall interference. The baseline results without quantization are shown in Figure 10. This illustration show us that the biggest SIR will always be located at the center of each beam and it diminishes as it moves further. Moreover, all the

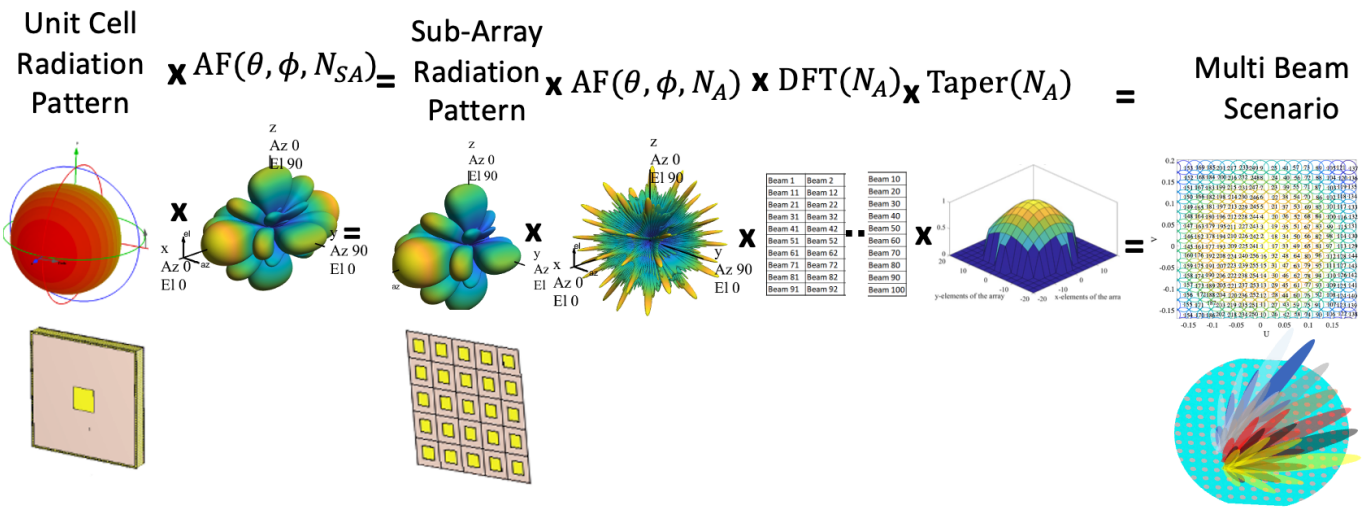

Fig. 2: Simulation scenario as a five-step process.

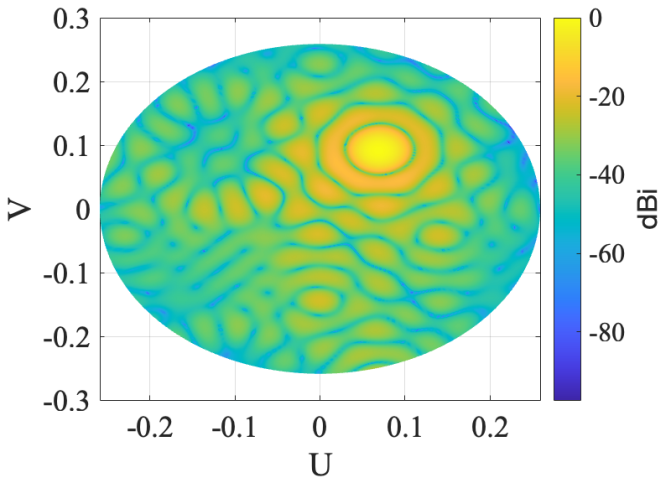

Fig. 3: No quantized radiation pattern corresponding to a 2DFFT index(4,4).

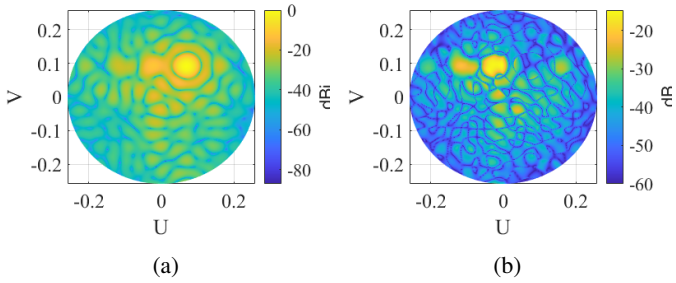

Fig. 4: Results corresponding to 1-bit case. a) Radiation pattern. b) Error between 1-bit results and no quantization.

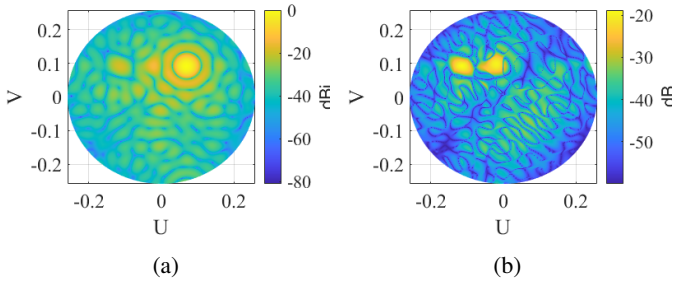

Fig. 5: Results corresponding to 2-bit case. a) Radiation pattern. b) Error between 2-bit results and no quantization.

beams at all positions will have a comparable value of SIR with a slightly higher value in the beams located near and in the broadside direction.

a) 1-bit Quantization: For the 1-bit quantization scenario, as depicted in Figure 11a, the SIR peaks at approximately -2.2 dB at broadside, while reaching a minimum of -6 dB for other beams positions. This higher SIR at broadside is particularly beneficial as it maintains power without suffering from big side lobes as other beams are affected.

b) 2, 3, 4-bit Quantization: In the 2, 3, 4-bit quantization scenario, shown in Figure 11b, 11c, 11d, there are improvements compared with 1-bit scenario, specially for 2-bit scenario but for 3 and 4-bit scenario the improvement is barely noticeable.

As previously observed, the scenario described is significantly affected by interference, primarily due to the high number of beams (256) relative to the fewer antenna elements (100). This disparity leads to considerable beam overlap, which significantly degrades the SIR. To provide a clearer understanding of the quantization effects on the SIR, we have deactivated a quarter of the beams, resulting in a total of 64 active beams at any given moment. Figure 12 illustrates this adjustment by showing the indices of the beams that remain activated. Following, we present the SIR of the no quantized beams, 1-bit quantization, 2 bit quantization, 3 -bit quantization, and 4-bit quantization in Figures 13, 14a, 14b,

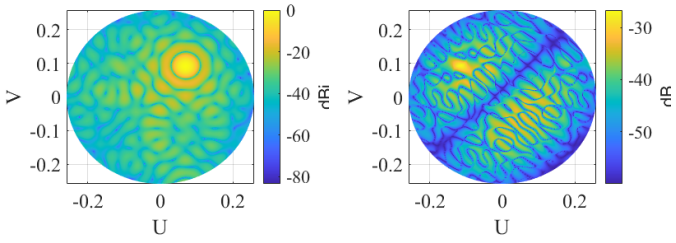

Fig. 6: Results corresponding to 3-bit case. a) Radiation pattern. b) Error between 3-bit results and no quantization.

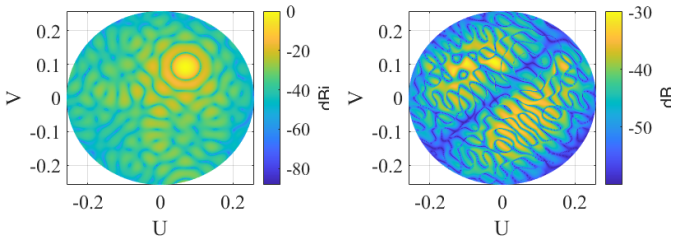

Fig. 7: Results corresponding to 4-bit case. a) Radiation pattern. b) Error between 4-bit results and no quantization.

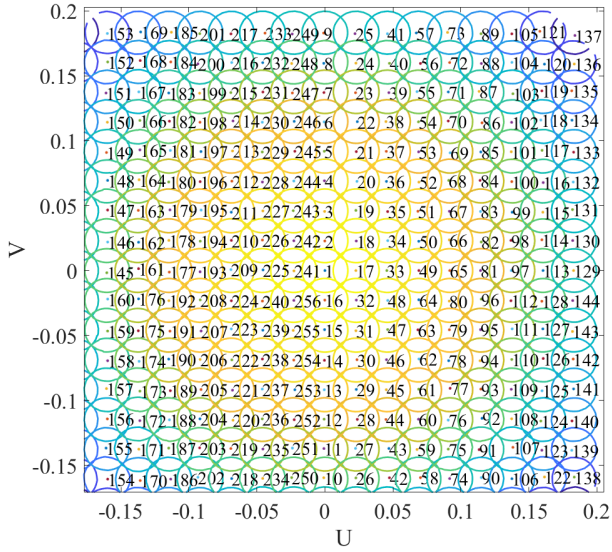

Fig. 8: Beam contour at $\theta_{-3\text{dB}}$ using a 16-point FFT, without quantization. The beam numbers are included in the figure.

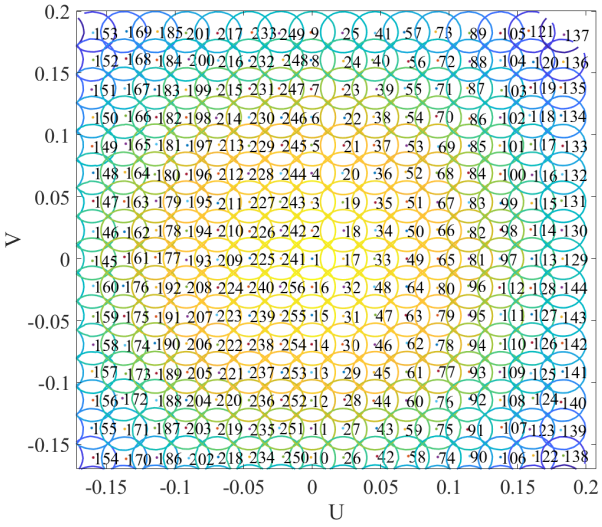

Fig. 9: Beam contour at $\theta_{-3\text{dB}}$ using a 16-point FFT, quantized at 1 bit. The beam numbers are included in the figure.

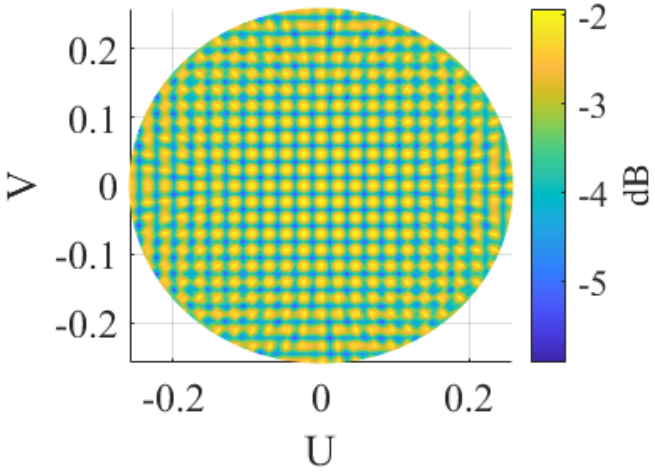

Fig. 10: SIR obtained without quantization effects.

14c and 14d, respectively.

In this analysis, we can have a more visible effect of the quantization effect compared with the previous case. The SIR improves as the quantization levels increase, with the most notable improvement occurring at 2-bit quantization, which increases the SIR by around 2 dB. For 3-bit and 4-bit quantization, the improvement is about 0.8 dB each. The difference between 3 and 4-bit quantization and no quantization is nearly imperceptible, allowing us to conclude that 3-bit or 4-bit quantization is sufficient to achieve a progressive phase shift without any significant discontinuities.

VI. Conclusions

This study has explored the effects of quantization on the twiddle factors of the FFT in both single and multi-beam radiation patterns, and their impact on the SIR. A notable

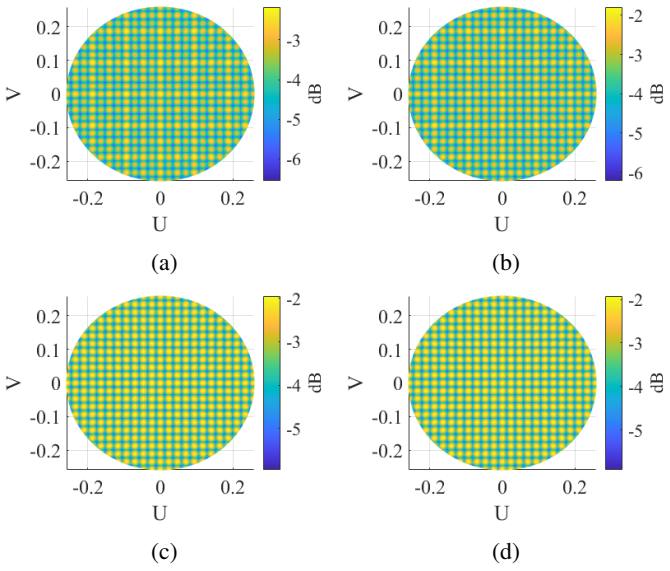

Fig. 11: SIR of the 256 beam scenario using: a) 1-bit quantization, b) 2-bit quantization, c) 3-bit quantization d) 4-bit quantization

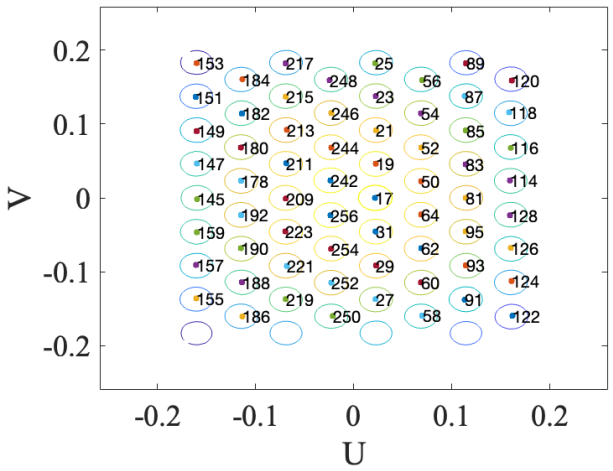

Fig. 12: 64-beams activation scenario. The beam numbers are included in the figure.

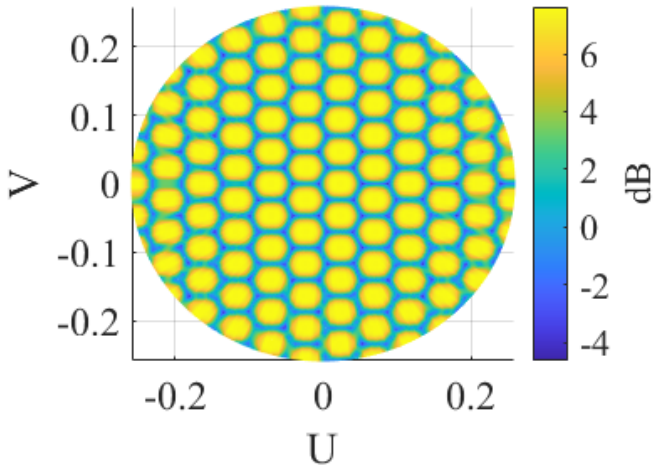

Fig. 13: SIR obtained without quantization in the 64 activated beam scenario.

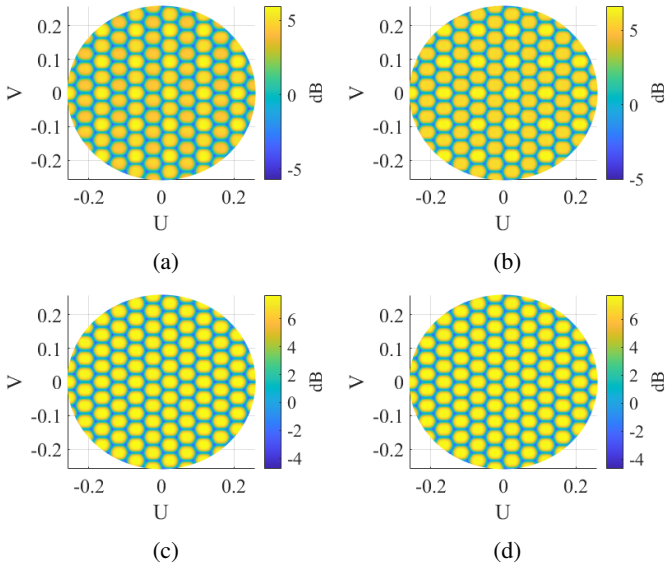

Fig. 14: SIR of the 64 beam scenario using: a) 1-bit quantization, b) 2-bit quantization, c) 3-bit quantization d) 4-bit quantization

outcome of quantization is the emergence of pronounced side lobes in the visible region, which intensify when the separation of sub-arrays exceeds half a wavelength. A key finding is the overall degradation of the SIR, where the presence of high side lobes introduces substantial interference, thus reducing system efficiency. Based on our analysis, we propose specific quantization strategies tailored to the computational resources available. For applications with limited resources, using a 1-bit quantized FFT algorithm can generate all required beams without producing grating lobes or high side lobes comparable to the main beam in the visible region. For scenarios with more computational capacity, 2-bit and 3-bit quantizations provide performance nearly equivalent to that of unquantized case. For the most demanding applications, where ample computational resources are available, we recommend 4-bit quantization, which achieves performance indistinguishable from the unquantized scenario. Consequently, we have opted for 4-bit quantization in our FFT beamforming network implementation, due to the previously described advantages.

REFERENCES

[1] J. Litva and T. K. Lo, *Digital beamforming in wireless communications*. Artech House, Inc., 1996.

[2] A. Capozzoli, C. Curcio, and A. Liseno, "Optimized nonuniform ffts and their application to array factor computation," *IEEE Transactions on Antennas and Propagation*, vol. 67, no. 6, pp. 3924–3938, 2019.

[3] D. Potts, G. Steidl, and M. Tasche, "Fast fourier transforms for nonequispaced data: A tutorial," *Modern Sampling Theory: Mathematics and Applications*, pp. 247–270, 2001.

[4] V. N. Ha, Z. Abdullah, G. Eappen, J. C. M. Duncan, R. Palisetty, J. L. G. Rios, W. A. Martins, H.-F. Chou, J. A. Vasquez, L. M. Garces-Socarras, H. Chaker, and S. Chatzinotas, "Joint linear precoding and dft beamforming design for massive mimo satellite communication," in *2022 IEEE Globecom Workshops (GC Wkshps)*, 2022, pp. 1121–1126.

[5] V. Ariyarathna, D. F. G. Coelho, S. Pulipati, R. J. Cintra, F. M. Bayer, V. S. Dimitrov, and A. Madanayake, "Multibeam digital array receiver using a 16-point multiplierless dft approximation," *IEEE Transactions on Antennas and Propagation*, vol. 67, no. 2, pp. 925–933, 2019.

[6] A. Madanayake, V. Ariyarathna, S. Madishetty, S. Pulipati, R. J. Cintra, D. Coelho, R. Oliviera, F. M. Bayer, L. Belostotski, S. Mandal, and T. S. Rappaport, "Towards a low-swap 1024-beam digital array: A 32-beam subsystem at 5.8 ghz," *IEEE Transactions on Antennas and Propagation*, vol. 68, no. 2, pp. 900–912, 2020.

[7] R. Palisetty, L. M. Garces Socarras, H. Chaker, V. Singh, G. Eappen, W. A. Martins, V. N. Ha, J. A. Vásquez-Peralvo, J. L. Gonzalez Rios, J. C. Merlano Duncan, S. Chatzinotas, B. Ottersten, A. Coskun, S. King, S. D'Addio, and P. Angeletti, "Fpga implementation of efficient beamformer for on-board processing in meo satellites," in *2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2023, pp. 1–7.

[8] R. Palisetty, G. Eappen, J. L. G. Rios, J. C. M. Duncan, S. Domouchtsidis, S. Chatzinotas, B. Ottersten, B. Cortazar, S. D'Addio, and P. Angeletti, "Area-power analysis of fft based digital beamforming for geo, meo, and leo scenarios," in *2022 IEEE 95th Vehicular Technology Conference: (VTC2022-Spring)*, 2022, pp. 1–5.

[9] J. A. Vásquez-Peralvo, J. Querol, F. Ortíz, J. L. González-Rios, E. Lagunas, L. M. Garcés-Socorrás, J. C. M. Duncan, M. O. Mendonça, and S. Chatzinotas, "Multibeam beamforming for direct radiating arrays in satellite communications using genetic algorithm," *IEEE Open Journal of the Communications Society*, 2024.

[10] J. A. Vásquez-Peralvo, J. Querol, F. Ortíz, J. L. G. Rios, E. Lagunas, V. M. Baeza, G. Fontanesi, L. M. Garcés-Socorrás, J. C. M. Duncan, and S. Chatzinotas, "Flexible beamforming for direct radiating arrays in satellite communications," *IEEE Access*, vol. 11, pp. 79 684–79 696, 2023.