MP-DPD: Low-Complexity Mixed-Precision Neural Networks for Energy-Efficient Digital Predistortion of Wideband Power Amplifiers

Wu, Yizhou; Li, Ang; Beikmirza, Mohammad; Singh, Gagan Deep; de Vreede, Leo C.N.; Alavi, Morteza ; Gao, Chang; Chen, Qinyu

DOI
10.1109/LMWT.2024.3386330

Publication date
2024

Document Version
Final published version

Published in
IEEE Microwave and Wireless Technology Letters

MP-DPD: Low-Complexity Mixed-Precision Neural Networks for Energy-Efficient Digital Predistortion of Wideband Power Amplifiers

Yizhuo Wu, Ang Li, Mohammadreza Beikmirza, *Member, IEEE*, Gagan Deep Singh, Qinyu Chen, *Member, IEEE*, Leo C. N. de Vreede, *Senior Member, IEEE*, Morteza Alavi, *Member, IEEE*, and Chang Gao, *Member, IEEE*

***Abstract*— Digital predistortion (DPD) enhances signal quality in wideband radio frequency (RF) power amplifiers (PAs). As signal bandwidths expand in modern radio systems, DPD's energy consumption increasingly impacts overall system efficiency. Deep neural networks (DNNs) offer promising advancements in DPD, yet their high complexity hinders their practical deployment. This article introduces open-source mixed-precision (MP) neural networks that employ quantized low-precision fixed-point parameters for energy-efficient DPD. This approach reduces computational complexity and memory footprint, thereby lowering power consumption without compromising linearization efficacy. Applied to a 160-MHz-BW 1024-QAM OFDM signal from a digital RF PA, MP-DPD gives no performance loss against 32-bit floating-point precision DPDs, while achieving −43.75 (L)/−45.27 (R) dBc in the adjacent channel power ratio (ACPR) and −38.72 dB in error vector magnitude (EVM). A 16-bit fixed-point-precision MP-DPD enables a 2.8× reduction in estimated inference power. The DPD code in PyTorch is publicly available on GitHub.**

***Index Terms*— Deep neural network (DNN), digital predistortion (DPD), digital transmitter (DTX), power amplifier (PA), quantization.**

I. Introduction

THE rapid evolution of wireless communication technologies has spurred an increased demand for higher data rates, improved spectral efficiency, and reduced error rates. Nonlinear distortions, predominantly caused by wideband radio frequency (RF) power amplifiers (PAs), significantly compromise signal integrity, affecting both communication reliability and energy efficiency. Digital predistortion (DPD) has emerged as a crucial technique to mitigate these issues, enhancing signal integrity. In contemporary radio digital front-ends, the DPD module is a major contributor to power consumption [1]. This challenge might be further exacerbated by the potential integration of machine-learning (ML) algorithms, such as deep neural networks (DNNs), which, despite their potential, add to the power demands.

Manuscript received 27 February 2024; accepted 18 March 2024. Date of publication 17 April 2024; date of current version 7 June 2024. *(Yizhuo Wu and Ang Li contributed equally to this work.) (Corresponding author: Chang Gao.)*

Yizhuo Wu, Ang Li, Mohammadreza Beikmirza, Gagan Deep Singh, Leo C. N. de Vreede, Morteza Alavi, and Chang Gao are with the Department of Microelectronics, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: chang.gao@tudelft.nl).

Qinyu Chen is with the Leiden Institute of Advanced Computer Science (LIACS), Leiden University, 2311 EZ Leiden, The Netherlands.

This article was presented at the IEEE MTT-S International Microwave Symposium (IMS 2024), Washington, DC, USA, June 16–21, 2024.

Color versions of one or more figures in this letter are available at https://doi.org/10.1109/LMWT.2024.3386330.

Data is available on-line at https://github.com/lab-emi/OpenDPD.

Digital Object Identifier 10.1109/LMWT.2024.3386330

Recent advancements of ML-based long-term DPD in state-of-the-art RF system-on-chip (SoC) products are given in [2]. Nevertheless, the substantial computational complexity and memory requirements of ML-based DPD systems, especially those using DNNs, pose significant obstacles to their efficient deployment in wideband transmitters, particularly in the context of future 5.5G/6G base stations or Wi-Fi 7 routers, where limited power resources constrain real-time DPD model computation.

Prior approaches to address DPD energy consumption include reducing the sample rate [3], employing a sub-Nyquist feedback receiver in the observation path [4], dynamically adjusting model cross-terms based on input signal characteristics [5], and devising simpler computational pathways for DPD algorithms [6]. This work presents a novel approach by implementing mixed-precision (MP) arithmetic operations and model parameters in a gated recurrent neural network (RNN)-based DPD model for wideband PAs. The proposed method curtails the DPD model inference[1] power consumption by substituting most high-precision floating-point operations with low-precision fixed-point operations through quantizing neural network weights (W) and activations (A). This strategy reduces the energy of arithmetic operations and memory access and facilitates the design of energy-and-area-efficient DNN computing hardware suitable for DPD deployment in power-sensitive environments [7]. Additionally, our method is compatible with existing strategies, allowing for further power savings when combined.

II. DPD Computing's Energy Problem

To effectively correct the in-band signal and reduce out-of-band emission, DPD systems typically operate at sample rates ranging from 1.5× to 5× the baseband signal bandwidth [3]. As bandwidths in future radio systems expand, the energy demands of DPD computation intensify. The energy consumed per DPD model inference for each input I/Q sample is approximated by

$$E_{\mathrm{INF}} = E_{\mathrm{MUL}} + E_{\mathrm{ADD}} + E_{\mathrm{MEM}} \tag{1}$$

where E_{MUL}, E_{ADD}, and E_{MEM} denote the energy consumption of multiplications (MUL), additions (ADD), and memory (MEM) access per DPD model inference, respectively. Since each inference processes one I/Q data point of the input signal, the estimated dynamic power consumption of the DPD model inference is given as

$$P_{\mathrm{INF}} = E_{\mathrm{INF}} \cdot f_s \tag{2}$$

[1]Inference of a neural network model is the process of making predictions based on the learned model parameters. Learning in a model involves training the model to update the parameters with a dataset to classify patterns (classification) or to track a time-varying discrete variable (regression).

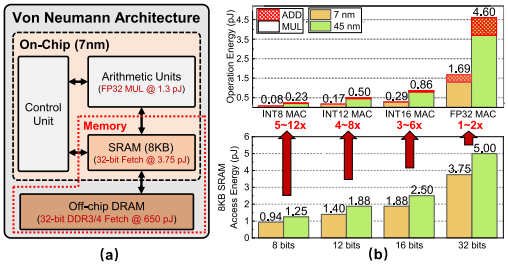

Fig. 1. (a) Von Neumann architecture with energy costs. (b) Operation and 8 KB SRAM access energy in 45 nm [8] and 7 nm [9] versus precision.

where f_s represents the DPD input I/Q data sample rate.

Utilizing 32-bit floating-point (**FP32**) arithmetic, while beneficial for accuracy, can increase model size, negatively impacting energy efficiency. Prior studies demonstrate that DNNs with low-precision, fixed-point calculations effectively minimize the memory footprint in demanding applications such as image recognition and large language models. This reduction is achieved with minimal accuracy loss, decreasing power consumption in hardware implementations. As shown in Fig. 1(b), multiply-accumulate (MAC) operations using 8-bit fixed-point integers (INT8) are up to 20× more energy-efficient than FP32 MAC operations, across both 45-nm [8] and 7-nm [9] technology nodes. Most neural network computations occur on Von Neumann architecture-based hardware, depicted in Fig. 1(a). This architecture often faces significant memory bottlenecks, as highlighted in Fig. 1(b). The energy consumption of on-chip static random access memory (SRAM) is up to 12.2× higher than that of a MAC operation. Moreover, the energy costs for off-chip memory access are roughly three orders of magnitude greater than for arithmetic operations. Therefore, the memory access demands, directly linked to the DPD model size, play a crucial role in determining overall power consumption.

III. Mixed-Precision Neural Networks DPD

Building on these insights, this section describes how to quantize weights and activations of gated recurrent neural networks (RNNs) into low precision for energy reduction.

A. Gated Recurrent Unit-Based DPD

Gated RNNs utilize gates to manage information flow through their high-dimensional hidden states according to new input stimuli. This approach effectively addresses the vanishing gradient issue in modeling long sequences and makes them widely adopted in prior research on long-term DPDs [10], [11]. In this work, the GRU-based DPD is defined as

$$\mathbf{r}_t = \sigma\left(\mathbf{W}_{ir}\boldsymbol{\phi}_t + \mathbf{b}_{ir} + \mathbf{W}_{hr}\mathbf{h}_{t-1} + \mathbf{b}_{hr}\right) \tag{3}$$

$$\mathbf{z}_t = \sigma\left(\mathbf{W}_{iu}\boldsymbol{\phi}_t + \mathbf{b}_{iz} + \mathbf{W}_{hz}\mathbf{h}_{t-1} + \mathbf{b}_{hz}\right) \tag{4}$$

$$\mathbf{n}_t = \tanh\left(\mathbf{W}_{in}\boldsymbol{\phi}_t + \mathbf{b}_{in} + \mathbf{r}_t \odot (\mathbf{W}_{hn}\mathbf{h}_{t-1} + \mathbf{b}_{hn})\right) \tag{5}$$

$$\mathbf{h}_t = (1 - \mathbf{z}_t) \odot \mathbf{n}_t + \mathbf{z}_t \odot \mathbf{h}_{t-1} \tag{6}$$

where $\boldsymbol{\phi}_t$ is the input feature vector extracted from the I/Q modulated signal $\mathbf{X} = \{\mathbf{x}_t | \mathbf{x}_t = I_{\mathbf{x},t} + jQ_{\mathbf{x},t}, I_{\mathbf{x},t}, Q_{\mathbf{x},t} \in \mathbb{R}, t \in 0, \ldots, T-1\}$ at time t. $\mathbf{h}_t$ represents the hidden state at time t. The $\mathbf{W}$ and $\mathbf{b}$ terms are the weight matrices and bias vectors, respectively. The terms $\mathbf{r}_t$, $\mathbf{z}_t$, and $\mathbf{n}_t$ correspond to the reset gate, update gate, and new candidate state, respectively. σ represents the sigmoid activation. $\odot$ denotes the element-wise multiplication. The GRU is followed by a fully connected (FC) layer to generate the DPD output I/Q signal

$$\hat{\mathbf{y}}_t = \mathbf{W}_{\hat{y}}\boldsymbol{h}_t + \mathbf{b}_{\hat{y}} \tag{7}$$

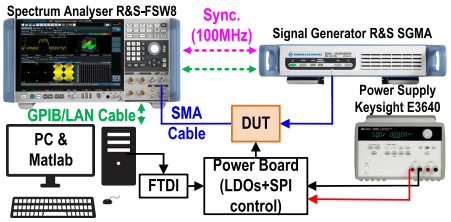

Fig. 2. Setup for dataset acquisition and DPD performance measurement.

where $\hat{\mathbf{y}}_t \in \hat{\mathbf{Y}} = \{\mathbf{y}_t | \mathbf{y}_t = I_{\hat{\mathbf{y}},t} + jQ_{\hat{\mathbf{y}},t}, I_{\hat{\mathbf{y}},t}, Q_{\hat{\mathbf{y}},t} \in \mathbb{R}, t \in 0, \ldots, T-1\}$.

B. Mixed-Precision DPD

To enhance the energy efficiency of DPD models, we adopt an MP strategy utilizing low-precision fixed-point integer arithmetic for inference. This method involves a quantization scheme that converts the model's weights and activations, including other intermediate variables, to lower precision while retaining full-precision operations for feature extraction $\boldsymbol{\phi}$ from I/Q signal $\mathbf{x}$, effectively balancing accuracy and computational complexity.

The quantization process is defined as follows: for a data point x, a quantization scale s, and a range $[Q_{\min}, Q_{\max}]$, the fixed-point representation q of x is calculated using

$$q = s \times \texttt{Round}\left(\texttt{Clip}\left(\frac{x}{s}, Q_{\texttt{min}}, Q_{\texttt{max}}\right)\right) \tag{8}$$

where `Clip` bounds the input and `Round` rounds to the nearest integer. For n-bit quantization, unsigned data ranges from $Q_{\texttt{min}} = 0$ to $Q_{\texttt{max}} = 2^n - 1$, and signed data from $Q_{\texttt{min}} = -2^{n-1}$ to $Q_{\texttt{max}} = 2^{n-1} - 1$. During training, each neural network layer's quantization scale s is optimized using gradient descent and adjusted to the nearest power-of-two, ensuring a fixed-point representation q. For precise fixed-point computations and enhanced energy efficiency, we use a quantization-aware training method [12]. This approach maintains full-precision variable copies updated during gradient descent while using quantized values for forward propagation of the DNN model. The gradient of the `Round` function is approximated using the straight-through estimator [13] for trainability.

IV. Experimental Results

A. Experimental Setup

Fig. 2 illustrates the experimental setup. The baseband I/Q data was processed by a 40-nm CMOS digital PA (DPA) [16] at a 2.4-GHz carrier frequency.

For the GRU-based MP-DPDs, quantization of activations and weights is performed at 8, 12, or 16 bits, except during feature extraction, which utilizes full-precision (FP32) operations to generate I_x, Q_x, $|x|$, $|x|^3$ features. We compared the MP-DPDs' performance to FP32 models, including general memory polynomial (GMP) [14], GRU, vector decomposition LSTM (VDLSTM), and the real-valued time-delay convolution neural network (RVTDCNN). The configurations for VDLSTM and RVTDCNN followed their optimal settings in [10] and [15], with adjustments in model size through the hidden LSTM and FC layer sizes.

The test signal's peak-to-average power ratio (PAPR) is 10.38 dB, and the DPA outputs at 13.75 dBm. The dataset, comprising 491 520 samples of 160-MHz 4-Channel × 40 MHz OFDM signals sampled at 640 MHz, was split into a 60% training set for DPD learning, a 20% validation set for

TABLE I

ACPR AND EVM PERFORMANCE OF DIFFERENT DPD MODELS EVALUATED WITH 160-MHz 4-CHANNEL × 40-MHz 1024-QAM OFDM SIGNALS SAMPLED AT 640 MHz ALONGSIDE THEIR ESTIMATED INFERENCE ENERGY AND DYNAMIC POWER CONSUMPTION IN 7 AND 45 nm [9]

Classes	DPD Models[a]	ACPR (dBc, L/R)	EVM (dB)	Number of MUL/ADD/MEM	Energy/Inference (nJ)		Dynamic Power (W)		Power Reduction
					45nm	7nm	45nm	7nm	
Without DPD	-	-31.69/-32.45	-27.05	-	-	-	-	-	-
FP32-DPDs	GMP [14]	-40.79/-40.86	-29.27	2190/3668/517	11.44	6.20	7.32	3.97	-
	VDLSTM [10]	-43.38/-43.02	-36.19	538/1528/542	3.38	3.32	2.16	2.12	-
	RVTDCNN [15]	-44.27/-43.50	-36.70	500/2690/512	4.28	3.60	2.74	2.30	-
	GRU	**-43.36/-45.30**	**-38.46**	**502/1417/506**	**5.66**	**3.09**	**3.62**	**1.98**	**1×**
MP-DPDs[b] (This work)	**W16A16-GRU**	**-43.75/-45.27**	**-38.72**	**502/1417/506**	**4.02**	**1.11**	**1.93**	**0.71**	**2.8×**
	W12A16-GRU	-43.03/-44.69	-37.47	502/1417/506	2.29	0.85	1.46	0.54	3.7×
	W12A12-GRU	-42.36/-43.79	-37.45	502/1417/506	2.19	0.82	1.40	0.52	3.8×
	W8A16-GRU	-41.64/-42.80	-36.24	502/1417/506	1.56	0.74	1.00	0.47	4.2×
	W8A12-GRU	-41.78/-42.90	-36.17	502/967/506	1.49	0.72	0.95	0.46	4.3×
	W8A8-GRU	-35.84/-35.70	-28.89	502/967/506	1.42	0.69	0.90	0.44	4.5×

[a] The numbers of parameters are 495 (GMP), 502 (GRU), 538 (VDLSTM), 500 (RVTDCNN).
[b] Each MP-DPD has 14 and 17 FP32 MULs and ADDs for feature extraction, respectively.

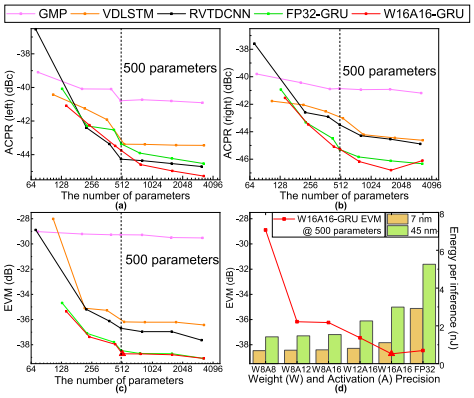

Fig. 3. Parameter scan of DPD models versus (a) ACPR (left), (b) ACPR (right), (c) EVM, (d) EVM (left Y-axis), and energy per inference (right Y-axis) versus precision.

early stopping, and a 20% test set for performance evaluation. The DPD learning process involves backpropagation through a pretrained PA model using our `OpenDPD` [17] with the collected dataset in an end-to-end approach. All PA models consist of approximately 500 parameters, except for those used in parameter scan experiments. For both PA modeling and DPD learning, the models are trained for 100 epochs using the ADAM optimizer with a learning rate of 1E-3 and a batch size of 3200 samples.

B. Results and Discussion

Table I compares the adjacent channel power ratio (ACPR) and error vector magnitude (EVM) results for different DPD models, alongside the number of MUL and ADD operations and 8 KB SRAM accesses[2] in feature extraction and model inference [(3)–(7)]. The amplitude/phase (`arctan2`) group, `tanh`, and `sigmoid` functions can be computed using the COordinate Rotation DIgital Computer (CORDIC) algorithm over 15 iterations (30 ADDs) despite that state-of-the-art gated RNN hardware [18] uses look-up tables to approximate them with less energy and chip area overhead. The 502-parameter W16A16-GRU DPD model demonstrates the best performance among all tested models, achieving an ACPR of −43.36/−45.30 dBc and an EVM of −38.72 dB while consuming 1.13 nJ per inference in 7-nm technology and 0.72-W dynamic power at 640 MHz. Lower power can be achieved by using a smaller model size or lower precision at the price of worse accuracy, as shown in Fig. 3.

Fig. 3(a)–(c) shows the correlation between model size and ACPR/EVM, covering 100–3200 parameters. The

[2]Each input I/Q sample necessitates 2 input fetches, #parameter fetches, and 2 output write-backs between the arithmetic units and the 8 KB SRAM cache. Intermediate results are buffered locally, thus bypassing cache access.

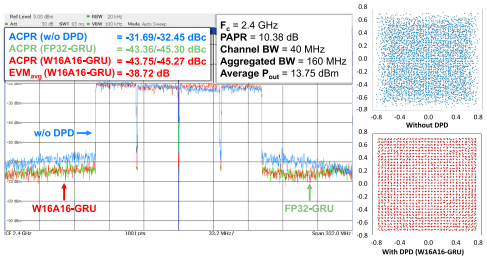

Fig. 4. Measured spectrum and constellation map on the 160-MHz signal.

W16A16-GRU DPD model notably outperforms FP32 models in many settings due to the regularization effect by training with quantization noise [12]. Fig. 3(d) presents the energy efficiency versus performance tradeoffs in MP models. The W8A8 model achieves a 4.5× power reduction over the FP32 model in 7-nm technology at the expense of linearization performance. The W12A16 and W16A16 configurations present a balanced compromise, offering 3.7× and 2.8× less power consumption than the FP32 GRU baseline DPD model while sustaining competitive EVM. Hence, W12A16 and W16A16 are optimal for power-critical applications demanding high accuracy.

Fig. 4 displays the measured spectrum and constellation map with and without DPDs. The spectrum analysis confirms that the W16A16-GRU model achieves no ACPR performance loss compared to the FP32-GRU model.

These findings underscore the effectiveness of our MP-DPD approach in reducing DPD power consumption while sustaining linearization performance.

C. Power Consumption Comparison to Prior Works

Prior hardware implementations of DPD hardly reported any power consumption numbers [19], [20]. To our best knowledge, the only work we found is a subsampling DPD field-programmable gate array (FPGA) implementation [5], which consumes 1.875 W to linearize the 100-MHz signal with a 150-MHz sampling rate and 320 parameters. For a fair comparison, we normalized it to the sample rate we used in this article, which is 640 MHz. By adopting our proposed MP A16W16-GRU DPD with 502 parameters, the power consumption can be reduced by 3.9×/10.6× to 1.93 W/0.71 W on a 45/7-nm process, respectively.

V. CONCLUSION

This work proposes the MP-DPD method for wideband RF PAs using the `OpenDPD` framework [17]. This approach reduces the computational complexity against the full-precision baseline, thereby contributing to power savings while preserving superior linearization performance for more sustainable and energy-efficient wireless communication.

REFERENCES

[1] S. Wesemann, J. Du, and H. Viswanathan, "Energy efficient extreme MIMO: Design goals and directions," *IEEE Commun. Mag.*, vol. 61, no. 10, pp. 132–138, Oct. 2023, doi: 10.1109/MCOM.004.2200958.

[2] *RF Transceiver With Dual Receivers and Transmitters, Observation Path, Integrated PLL/VCOs, and Auxiliary Converters*, document ADRV9040, Analog Devices, Wilmington, MA, USA, 2023. Accessed: Dec. 1, 2023. [Online]. Available: https://www.analog.com/en/products/adrv9040.html

[3] Y. Li, X. Wang, and A. Zhu, "Sampling rate reduction for digital predistortion of broadband RF power amplifiers," *IEEE Trans. Microw. Theory Techn.*, vol. 68, no. 3, pp. 1054–1064, Mar. 2020.

[4] N. Hammler, A. Cathelin, P. Cathelin, and B. Murmann, "A spectrum-sensing DPD feedback receiver with 30× reduction in ADC acquisition bandwidth and sample rate," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 66, no. 9, pp. 3340–3351, Sep. 2019.

[5] Y. Li, X. Wang, and A. Zhu, "Reducing power consumption of digital predistortion for RF power amplifiers using real-time model switching," *IEEE Trans. Microw. Theory Techn.*, vol. 70, no. 3, pp. 1500–1508, Mar. 2022.

[6] M. Beikmirza, L. C. N. de Vreede, and M. S. Alavi, "A low-complexity digital predistortion technique for digital I/Q transmitters," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Jun. 2023, pp. 787–790.

[7] C. Gao, A. Rios-Navarro, X. Chen, S.-C. Liu, and T. Delbruck, "EdgeDRNN: Recurrent neural network accelerator for edge inference," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 10, no. 4, pp. 419–432, Dec. 2020.

[8] M. Horowitz, "1.1 computing's energy problem (and what we can do about it)," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2014, pp. 10–14.

[9] N. P. Jouppi et al., "Ten lessons from three generations shaped Google's TPUv4i: Industrial product," in *Proc. ACM/IEEE Int. Symp. Comput. Archit. (ISCA)*, Jun. 2021, pp. 1–14.

[10] H. Li, Y. Zhang, G. Li, and F. Liu, "Vector decomposed long short-term memory model for behavioral modeling and digital predistortion for wideband RF power amplifiers," *IEEE Access*, vol. 8, pp. 63780–63789, 2020.

[11] T. Kobal and A. Zhu, "Digital predistortion of RF power amplifiers with decomposed vector rotation-based recurrent neural networks," *IEEE Trans. Microw. Theory Techn.*, vol. 70, no. 11, pp. 4900–4909, Nov. 2022.

[12] M. Nagel, M. Fournarakis, R. A. Amjad, Y. Bondarenko, M. van Baalen, and T. Blankevoort, "A white paper on neural network quantization," 2021, *arXiv:2106.08295*.

[13] Y. Bengio, N. Léonard, and A. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," 2013, *arXiv:1308.3432*.

[14] D. R. Morgan, Z. Ma, J. Kim, M. G. Zierdt, and J. Pastalan, "A generalized memory polynomial model for digital predistortion of RF power amplifiers," *IEEE Trans. Signal Process.*, vol. 54, no. 10, pp. 3852–3860, Oct. 2006.

[15] X. Hu et al., "Convolutional neural network for behavioral modeling and predistortion of wideband power amplifiers," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 3923–3937, Aug. 2022.

[16] M. Beikmirza, Y. Shen, L. C. N. de Vreede, and M. S. Alavi, "A wideband energy-efficient multi-mode CMOS digital transmitter," *IEEE J. Solid-State Circuits*, vol. 58, no. 3, pp. 677–690, Mar. 2023.

[17] Y. Wu, G. D. Singh, M. Beikmirza, L. C. N. de Vreede, M. Alavi, and C. Gao, "OpenDPD: An open-source end-to-end learning & benchmarking framework for wideband power amplifier modeling and digital pre-distortion," 2024, *arXiv:2401.08318*.

[18] C. Gao, T. Delbruck, and S.-C. Liu, "Spartus: A 9.4 TOp/s FPGA-based LSTM accelerator exploiting spatio-temporal sparsity," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 1, pp. 1098–1112, Jan. 2024, doi: 10.1109/TNNLS.2022.3180209.

[19] D. Byrne, R. Farrell, S. Madhuwantha, M. Leeser, and J. Dooley, "Digital pre-distortion implemented using FPGA," in *Proc. 28th Int. Conf. Field Program. Log. Appl. (FPL)*, Aug. 2018, pp. 453–4531.

[20] C. Quindroit, N. Naraharisetti, P. Roblin, S. Gheitanchi, V. Mauer, and M. Fitton, "FPGA implementation of orthogonal 2D digital predistortion system for concurrent dual-band power amplifiers based on time-division multiplexing," *IEEE Trans. Microw. Theory Techn.*, vol. 61, no. 12, pp. 4591–4599, Dec. 2013.