

Optimization of DSP Applications Using Parameterized Error Models for Low Power Approximate Adders

Celia Dharmaraj, *Student Member, IEEE*, Vinita Vasudevan, *Member, IEEE*,
and Nitin Chandrathoodan, *Member, IEEE*

Abstract—Approximate circuit design has gained significance in recent years targeting error tolerant applications. In this paper, we consider the problem of minimizing the power for a given accuracy, in a signal processing application with accurate adders replaced by low-power approximate adders. We first demonstrate that the commonly used assumption that the inputs to the adder are uniformly distributed results in an inaccurate prediction of error statistics for multi-level circuits. To overcome this problem, we propose the use of parameterized error models for adders, with input static probabilities as parameters. The static probability computation in our work considers not just the functionality of the adder but also its position in the circuit, functionality of its parents and the number of approximate bits in the parent blocks. This parameterized error model can be incorporated in any optimization framework. We demonstrate up to 6.5 dB improvement in the accuracy of noise power prediction when the proposed model is used to optimize an 8×8 DCT.

Index Terms—Accuracy, approximate adder, error model, low power, noise power, optimization, static probability.

I. INTRODUCTION

Approximate computing is widely used in signal and image processing applications to obtain improvements in power and/or speed while maintaining the required accuracy. Adders are the basic building blocks in these applications and a typical implementation has a large number of adders. A variety of approximate adders have been proposed in the literature, with different levels of trade-offs between accuracy and performance. These adders can be classified as low-latency [1] (and references therein) and low-power approximate adders [2]–[7]. In this paper, our focus is power optimized implementations of signal processing algorithms using low power approximate adders (LPAA). This requires an optimization routine to find the maximum number of approximate bits possible in each adder for a given accuracy. This in turn requires accurate error models for the approximate adders.

In the literature, multiple approaches have been proposed to find optimal approximation levels for adders used in low power implementations. An approximate Finite Impulse Response (FIR) filter is designed by fixing the level of approximation of the adders using Monte-Carlo simulations in [8]. Approximate mirror adder-5 (AMA-5) [4] modeled assuming uniformly distributed inputs are used in a 2D Discrete Cosine Transform (DCT) module constructed using 1D DCT blocks in [9] and

the optimization problem is solved using a mixed integer non-linear problem solver. Cartesian Genetic Programming (CGP) is used to design various approximate implementations of four point 1D DCT in [10]. An expression for variance of error of AMA 1-5 adders [4] and Lower part OR adder (LOA) [5] is obtained in [11] empirically by regression assuming uniform inputs, and heuristics are used to solve the approximation-level optimization problem. In [12], AMA 1-5 adders and Transmission Gate based Approximate adders TGA I-II [13] are considered. An expression for mean square error (MSE) is obtained assuming that the distribution of inputs and error are uniform. This is then used in a Lagrange multiplier based optimization approach.

All of the above previous works on optimization use error metrics based on uniformly distributed inputs. Moreover, the same error model is used for all adders in the circuit. To verify the validity of these assumptions, we analytically compute the noise power (NP_1) at the output of an approximate 8×8 DCT module [14] and compare it with that obtained using Monte-Carlo simulations (NP_{sim}) in Table I. For Monte-Carlo simulations, we considered 10^5 uniformly distributed random inputs. The DCT module consists of 288 adders spread over 6 levels denoted by $L_1 - L_6$ in the table. The accurate adders are replaced with LOAs. The inputs to the DCT module are assumed to have 15 bits of precision, i.e. Q1.15 in fixed point format. Analytical and simulated noise power are computed for various combinations of approximate bit assignments for adders in different levels as mentioned in the table. The absolute error in noise power prediction is given by $|e_1| = |NP_{sim} - NP_1|$. From Table I, we see that the analytical noise power differs from the simulated value by as much as 11 dB.

While the assumption of uniformly distributed lower order bits may be justified for the primary inputs, neither the output nor the error is uniformly distributed at the output of most LPAAs. A more accurate method of obtaining the probability mass function (PMF) of error is proposed in [15]. However, including this method within an optimization routine would require extensive computations. Moreover, in most applications, an accurate estimate of the mean error and MSE is sufficient and we do not need the PMF of the error.

In this paper, we use parameterized error models based on static probabilities of the inputs to each adder. A uniform distribution implies a static probability of 0.5, which is what

The authors are with the Department of Electrical Engineering, Indian Institute of Technology Madras, India.
E-mail: ee13d003,vinita,nitin@ee.iitm.ac.in.

TABLE I: Noise power computation (in dB) using analysis and simulation of an 8×8 DCT module that uses Lower part OR adders.

NP_1	NP_{sim}	$ e_1 $	No. of approximate bits
-53.45	-51.20	2.25	$L_1 - L_4$: 5; $L_5 - L_6$: 6
-51.76	-44.67	7.09	$L_1 - L_2$: 5; $L_3 - L_6$: 6
-51.32	-43.00	8.32	$L_1 - L_2$: 5; $L_3 - L_5$: 6; L_6 : 7
-50.55	-41.21	9.34	$L_1 - L_2$: 5; $L_3 - L_4$: 6; $L_5 - L_6$: 7
-47.59	-36.50	11.09	L_1 : 5; $L_2 - L_3$: 6; $L_4 - L_5$: 7; L_6 : 8

is used in most error models. Since the inputs to most of the adders come from other approximate adders, we derive the value for the static probability at the output of each approximate adder. The mean error and MSE contributed by an adder is computed based on the static probabilities of the output of its parent blocks. Therefore, in our framework, the optimizer is not just aware of which approximate adder is used, but also its parents and the number of approximate bits used in the parent blocks. This significantly improves the accuracy prediction, as will be seen in the results.

Each approximate adder requires the static probabilities of its input bits, which eventually traces back to the primary inputs. As mentioned, in the literature, the distribution of the lower order bits is assumed to be uniform. Some justification for this assumption on primary inputs is included in [7]. In this paper, we derive an exact condition under which the lower order bits of a signal is uniformly distributed and check the validity of the assumption on primary input of an image processing application, where the intensity distribution of image pixels is very non-uniform.

To summarize, our main contributions are as follows:

- 1) We develop an optimization framework using parameterized error models for LPAA's to maximize the number of approximate bits in each adder of a signal processing application for a given accuracy constraint.
- 2) We show that the Discrete Fourier Transform (DFT) of a signal satisfies a certain condition if the distribution of the lower order bits is uniform.
- 3) We obtain power-optimized implementations of an FIR filter and a 2D 8×8 DCT module using the optimization framework and demonstrate significant improvement in accuracy prediction due to the use of the parameterized error models.

We have organized this paper as follows: In Section II, we present the methodology to derive the parameterized error model and to assign input probabilities to approximate adders in a multi-level circuit. In Section II-A, we derive the conditions under which the lower order bits of the primary inputs will have uniform distribution. The details of the optimizer that is developed to obtain the optimal number of bits that can be approximated for a given accuracy constraint are presented in Section III. The results are shown in Section IV and finally Section V concludes the paper.

II. PARAMETERIZED ERROR MODELS FOR LOW POWER APPROXIMATE ADDERS

The general assumption in deriving various error metrics for approximate adders is that the k LSBs of the N -bit input A are uniformly distributed, resulting in static probabilities $P_{a_i} = 0.5$ for $i = 0, 1, \dots, k - 1$. In the literature, the expressions

for mean error and MSE are derived using this. However, the static probability at the output of approximate adders is not 0.5 (for example, for LOA it is 0.75 and for AMA-1 it is 0.25), which means that the output PMF is not uniform. If this forms the input to a subsequent adder, the error models are inaccurate. Instead, we can parameterize the error metrics in terms of static probabilities and use the right values for each adder.

Generally an N -bit LPAA is constructed using k approximate full adders to compute the lower part sum and $N - k$ accurate full adders to compute the upper part sum. Let $\hat{s}_i = f(a_i, b_i, \hat{c}_{i-1})$ and $\hat{c}_i = g(a_i, b_i, \hat{c}_{i-1})$ be the sum and output carry of an approximate full adder. Therefore, $P_{\hat{s}_i}$ and $P_{\hat{c}_i}$ are functions of P_{a_i} , P_{b_i} and $P_{\hat{c}_{i-1}}$. To find $P_{\hat{s}_i}$ and $P_{\hat{c}_i}$, we assume that (a) the inputs are independent of each other and (b) the probability of getting a carry in each bit is the same, i.e., $P_{\hat{c}_i} = P_{\hat{c}_{i-1}}$. The first assumption is an approximation when the circuit has reconvergent fanouts. However, it is a reasonable approximation in many cases as correlations are diluted as the logic depth increases, as argued in [15]. We do not have a rigorous justification for the second assumption, but estimates for $P_{\hat{c}_i}$ are close to what is obtained using $c_{-1} = 0$ and working out the statistics for each bit location as in [4]. Also, the error expression for AMA-1 adder obtained using this assumption is what is used in [12]. Simulations also indicate that this is a good assumption. With these assumptions and using the truth table of the approximate full adder, it is possible to derive expressions for $P_{\hat{s}_i}$ and $P_{\hat{c}_i}$.

The error in the output is due to the approximate lower part sum and the approximate carry to accurate adder. This can be written as

$$E_s = \sum_{i=0}^{k-1} (a_i + b_i)2^i - \sum_{i=0}^{k-1} \hat{s}_i 2^i - \hat{c}_k 2^k. \quad (1)$$

The mean error of the approximate adder is given by

$$E\{E_s\} = \sum_{i=0}^{k-1} (P_{a_i} + P_{b_i})2^i - \sum_{i=0}^{k-1} P_{\hat{s}_i} 2^i - P_{\hat{c}_k} 2^k. \quad (2)$$

The MSE can be derived in a similar fashion. The expression for MSE also involves joint probabilities $P(a_i b_i) = P_{a_i} P_{b_i}$ and $P(a_i a_j)$, $i \neq j$. In addition to the assumption that the inputs are independent, we also assume that individual bits of each input are independent, which is a reasonable approximation as discussed in [15].

An exception to this method for deriving error models is ETA-I [6], where the lower part sum is not constructed using similar approximate full adders. Its error metrics are derived in our earlier work [16].

In DSP systems, the inputs to the adder are either the primary inputs or they are output of another approximate adder or (in our case, accurate) multiplier. We now consider each of these cases.

A. Static Probabilities: Primary inputs

Typical PMF of any primary input such as an image is not uniform. As an example, the PMF of Cameraman image is shown in Fig. 2a. However, we are concerned about the

PMF of the lower bits of the input signal or image, since error expression of a LPAA involves the probability of the lower k bits that are approximated. In all the previous works, the distribution of lower order bits of all primary inputs are assumed to be uniform. We now derive conditions for the k LSBs of an N bit signal to be uniform.

Let \mathcal{F}_A be the 2^N -point DFT of the PMF of N -bit signal A and \mathcal{F}_{A_L} be the 2^k -point DFT of the PMF of A_L (k LSBs of A). We have,

$$\begin{aligned} \mathcal{F}_{A_L}[m] &= \sum_{n=0}^{2^k-1} P(A_L = n) e^{-jmn2\pi/2^k} \\ &= \sum_{n=0}^{2^k-1} \sum_{l=0}^{2^{N-k}-1} P(A = l2^k + n) e^{-jmn2\pi/2^k} \\ &= \sum_{l=0}^{2^{N-k}-1} \sum_{n'=l2^k}^{l2^k+2^k-1} P(A = n') e^{-jmn'2\pi/2^k} e^{jml2^k2\pi/2^k} \\ &= \sum_{n'=0}^{2^N-1} P(A = n') e^{-jmn'2\pi/2^k} \\ &= \mathcal{F}_A[m \cdot 2^{N-k}], 0 \leq m < 2^k. \end{aligned} \quad (3)$$

If A_L is uniform, $P(A_L = n) = \frac{1}{2^k}$, $0 \leq n < 2^k$. Hence from (3), if A_L is uniform, we have

$$\mathcal{F}_{A_L}[m] = \sum_{n=0}^{2^k-1} \frac{1}{2^k} e^{-jmn2\pi/2^k} = \begin{cases} 1, & \text{if } m = 0 \\ 0, & \text{if } 0 < m < 2^k. \end{cases} \quad (5)$$

Since DFT is unique, the converse is also true. Therefore using (4), we have the following condition to be satisfied for A_L to be uniformly distributed.

$$\mathcal{F}_A[m \cdot 2^{N-k}] = \begin{cases} 1, & \text{if } m = 0 \\ 0, & \text{if } 0 < m < 2^k. \end{cases} \quad (6)$$

In [17], they have similar condition for continuous signals that are quantized, although the derivation is a little more involved.

To illustrate this condition (6), let us consider the Cameraman image with $N = 8$. For the image pixel distribution, $\mathcal{F}_A[m \cdot 2^{N-k}]$ for different values of k , m varying from 0 to $2^k - 1$, is plotted in Fig. 1. It is seen that for lower values of k , the value of the transform is very close to zero for $0 < m < 2^k$. As k increases, the value of transform also increases and for $k = 5$, the values are high. This is confirmed from the actual PMF of the lower order bits of the image shown in Fig. 2. From the figure, it is seen that distribution can be considered uniform even if half the bits are

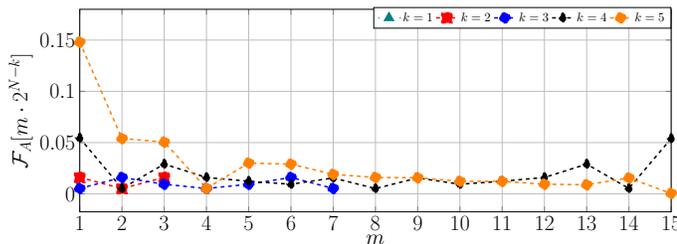


Fig. 1: Illustration of condition for k lower-order bits of Cameraman image to be uniform.

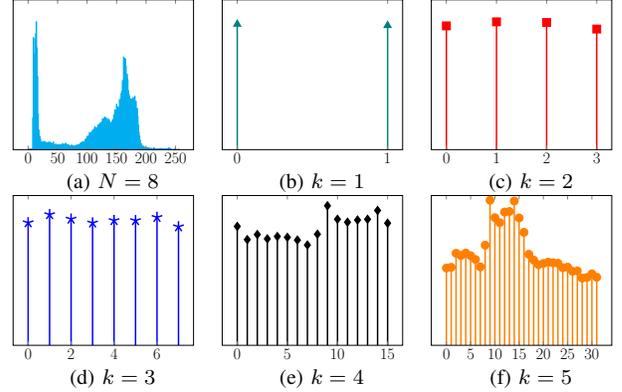


Fig. 2: (a) PMF of Cameraman image; (b)-(f) PMF of the lower k bits of the image.

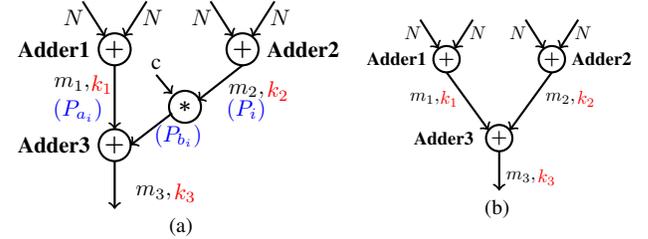


Fig. 3: Adder tree in a circuit with N bits at primary inputs.

approximated. This turns out to be true for all the standard images we have looked at. Hence, we assume that primary inputs to the approximate adder are uniformly distributed.

B. Static Probabilities: Adders in the higher levels

If the inputs to the adder are the output of another adder as in Fig. 3b, the mean and MSE are derived using $P_{\hat{s}_i}$ and $P_{\hat{c}_k}$ as discussed previously.

The other possibility is that input is the output of a multiplier. In this work, we are only optimizing adders and all multipliers are accurate, with the output truncated to the standard precision used in the circuit. Also, we only consider linear systems, so that one of the inputs to the multiplier is a constant coefficient. In Fig. 3a, consider Adder3 which has an input from the output of the multiplier. Depending on the value of the constant coefficient c , the probability of the LSBs at the output of the multiplier (P_{b_i}) will vary. Let P_i denote the probability that the i^{th} bit of the k_2 LSBs of the multiplicand (output of Adder2) is 1. Consider the following cases.

- 1) When $c = 2^l$ and $l \geq 0$, the product is the logical left shift of the multiplicand. So $P_{b_i} = 0$ for the first l LSBs and $P_{b_i} = P_{i-l}$ for the next $k_2 - l$ LSBs.
- 2) When $c = -2^l$ and $l \geq 0$, $P_{b_i} = 0$ for the first l LSBs and $P_{b_i} = 1 - P_{i-l}$ for the next $k_2 - l$ LSBs is a good approximation, accounting for flipping involved in the two's complement representation for negative numbers.
- 3) When $c = 2^l$ and $l < 0$, the product is the right shift of the multiplicand. So $P_{b_i} = P_{i+|l|}$ for $k_2 - l$ LSBs.
- 4) When $c = -2^l$ and $l < 0$, $P_{b_i} = 1 - P_{i+|l|}$ for $k_2 - l$ LSBs.
- 5) For c chosen uniformly at random, Monte Carlo simulations indicate that the average static probability of the output bits is 0.5 ± 0.03 for each of the LSBs. Therefore,

when c is not a power of 2, we assume that $P_{b_i} = 0.5$ for k_2 LSBs.

C. Truncation and Median adder (MA) in higher levels

Both Truncation and MA have their lower part sum bits fixed to constant all 0's and 1's respectively. In these adders, since the lower part sum is known, the lower part sum of the adders in higher levels can be fixed more accurately so that the accuracy of the approximate circuit is improved. In case of Truncation adder, the approximate sum is obviously zero. In Fig. 3b with Median adders, Adder1 and Adder2 will have their lower part sum as $2^{k_1} - 1$ and $2^{k_2} - 1$, respectively. For Adder3, instead of setting the lower part sum as $2^{k_3} - 1$, we improve the accuracy of the circuit by setting the sum to $2^{k_3+1} - 1$ for the following cases:

- 1) If $k_3 \leq k_1, k_2$, the lower part sum is known exactly and is equal to $2^{k_3+1} - 2$, which is closer to $2^{k_3+1} - 1$ than $2^{k_3} - 1$.
- 2) If $k_1 \geq k_3 > k_2$, the mean of the sum is $(3 \times 2^{k_3} + 2^{k_2} - 4)/2$, which is closer to $2^{k_3+1} - 1$ than $2^{k_3} - 1$.

Using this setting, we obtain up to 6 dB improvement for the adder tree in Fig. 3b.

III. OPTIMIZER FOR POWER-ACCURACY TRADE-OFF

The goal of the optimizer is to maximize the number of approximate bits of the adders in the circuit for a given noise power constraint at the output. The primary inputs to the system are normalized to $1.N$ fixed point numbers with N fractional bits. For each functional unit in the system, we use the required number of integer bits while maintaining the number of fractional bits as N . The output noise power is computed based on the mean, MSE and the transfer function from each adder to the output.

In [18], a three-step procedure that uses Minimum Width algorithm, Mildest Greedy Ascent algorithm and Tabu search algorithm was used to minimize the word length of each signal in the circuit for a given accuracy constraint. We have adapted the procedure to minimize the number of accurate bits in each adder, for a given accuracy constraint at the output. For this purpose, several modifications were made in the algorithms in order to incorporate the parameterized error models of various approximate adders and simultaneous satisfaction of constraints at multiple outputs. The main differences when compared to [18] are as follows:

- In [18], the error introduced at each node due to quantization depended only on the number of bits quantized at that node. In our case, we need to keep track of the number of approximate bits in each adder, its parent nodes and their functionality and the type of approximate adder.
- In [18], increasing the word-length of any signal results in lower quantization noise. However, in the case of approximate adders, the approximation noise can worsen even if the number of accurate bits is increased in certain cases. Although counter-intuitive, this happens as the mean error shifts significantly for some of the adders.
- In Tabu search algorithm, we target signals with maximum number of accurate bits (instead of the most

TABLE II: Error in noise power computation (in dB), when the FIR filter using ETA-I adders is optimized using $P_i = 0.5$ in the error model (NP_1) and parameterised error models (NP_2).

NP_t	NP_1	NP_{sim}	e_1	NP_2	NP_{sim}	e_2
-40	-40.36	-33.73	6.63	-40.06	-38.26	1.8
-45	-45.15	-38.68	6.47	-45.22	-43.5	1.72
-50	-50.23	-44.17	6.06	-50.25	-48.9	1.35
-55	-55.12	-49.84	5.28	-55.16	-54.61	0.55
-60	-60.15	-56.54	3.48	-60.16	-61.18	1.02

sensitive signal) for reduction and keep decreasing the number of accurate bits as long as noise power constraint is met. We found that this heuristic provides better power-accuracy trade-off.

IV. EXPERIMENTAL RESULTS

All the approximate circuits are designed using Verilog and synthesized with relaxed timing constraints using Synopsys Design Compiler (DC) for 55nm technology to get a gate-level netlist. The synthesized netlist along with Standard Delay Format file generated by Synopsys DC is simulated with 10^5 uniform random inputs for FIR filter and standard images for DCT computation. A full adder's input pin capacitance is set as the output load capacitance. Using the value change dump (VCD) file generated after simulation, dynamic power is found using Cadence Genus.

A. Application: FIR filter

First, we consider the direct form I realization of an 18-tap low pass FIR filter, where the accurate adders are replaced with approximate adders. Among the LPAAs, AMA-5 [4], LOA [5], ETA-I [6], Truncation adder and MA [7] have been observed to be very promising in terms of power savings [19], [20], [4], [7]. So we use these approximate adders in our experiments. In our implementation, we have assumed that the input of the filter has 10 fractional bits and the filter coefficients and multipliers' outputs and adders' outputs have 15 fractional bits of precision.

For a given noise power at the output (NP_t), the results obtained by the optimizer when we use $P_{a,i} = P_{b,i} = 0.5$ in the error model for all the adders is given by NP_1 in Table II. As mentioned, for ETA-I adder, it is not possible to get the static probability at the output using the procedure in section II. However, based on the functionality, it is definitely greater than 0.5. We assumed a value of 0.75 and use it in the optimizer for the higher level adders and obtain NP_2 . Error in noise power computation is the difference between the actual simulated value (NP_{sim}) and the predicted one using the error model in the optimizer (NP_1 and NP_2). We see that the error drops drastically from a maximum of 6.63 dB to 1.8 dB when we use updated values of static probabilities in the error model.

The FIR filter is designed using various approximate adders, with the help of the optimizer to obtain the number of approximate bits in each adder. The dynamic power consumption is obtained using the procedure described previously. Fig. 4a shows the percentage power savings (power savings/power of accurate circuit) in the FIR filter versus output noise power. It is seen that the FIR filter implemented using MA and AMA-5 adders give maximum power savings.

TABLE III: Error in noise power computation (in dB), when the 8×8 DCT using LOA is optimized using $P_i = 0.5$ in the error model (NP_1) and parameterised error models (NP_2).

NP_t	NP_1	NP_{sim}	e_1	NP_2	NP_{sim}	e_2
-40	-40.08	-32.66	7.42	-40.0	-39.1	0.9
-45	-45.01	-37.5	7.51	-45.0	-43.72	1.28
-50	-50.05	-43.77	6.28	-50.0	-48.38	1.62
-55	-55.01	-49.92	5.09	-55.0	-53.25	1.75
-60	-60.01	-54.64	5.37	-60.0	-58.5	1.5

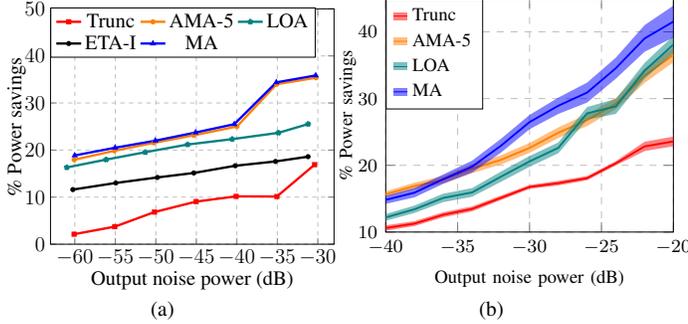


Fig. 4: (a) Percentage power savings in the FIR filter (b) Band of percentage power savings vs Noise power at the output of DCT implemented using various approximate adders.

B. Application: DCT

We consider the implementation of 8×8 DCT using the transform matrix presented in [14], which is a multiplierless transformation matrix with entries 0, 1 and -1. The circuit consists of a chain of adders along with some two's complement blocks. For a given noise power at the output (NP_t), the results obtained by the optimizer when we use $P_{a,i} = P_{b,i} = 0.5$ in the error model (NP_1) and that obtained with updated bit probabilities (NP_2) for LOA adder are shown in Table III. We see that the error drops drastically when the optimizer uses proper values of bit probabilities in the error model.

For various input images such as Cameraman, Lena, Fishing boat and Peppers, 8×8 DCT was performed with approximation optimized for various noise power values. The percentage power savings obtained is plotted as a band in Fig. 4. For a given noise power, DCT implementation using MA gives the most power savings.

V. CONCLUSION

We have proposed parameterized error models for approximate adders using input static probabilities as parameters and incorporated these error models in an optimization framework. We have shown that the parameterized error models provide better noise power prediction than the typical error models that assume uniform input distribution. We obtain power-optimized implementations of an FIR filter and a 2D 8×8 DCT of JPEG encoder using various approximate adders. In comparison to the other low power approximate adders considered in this work, Median adders are shown to provide better power-accuracy trade-off in applications.

REFERENCES

[1] M. Shafique, W. Ahmad, R. Hafiz, and J. Henkel, "A low latency generic accuracy configurable adder," in *Proceedings of the 52nd Annual Design Automation Conference, DAC '15*, (New York, NY, USA), ACM, 2015.

[2] H. A. F. Almurib, T. N. Kumar, and F. Lombardi, "Inexact designs for approximate low power addition by cell replacement," in *Design, Automation and Test in Europe (DATE)*, 2016.

[3] Z. Yang, A. Jain, J. Liang, J. Han, and F. Lombardi, "Approximate xor/xnor-based adders for inexact computing," *2013 13th IEEE International Conference on Nanotechnology (IEEE-NANO 2013)*, 2013.

[4] V. Gupta, D. Mohapatra, A. Raghunathan, and K. Roy, "Low-power digital signal processing using approximate adders," *IEEE Trans. on Comp.-Aided Design of Integrated Circuits and Systems*, vol. 32, 1 2013.

[5] H. R. Mahdiani, A. Ahmadi, S. M. Fakhraie, and C. Lucas, "Bio-inspired imprecise computational blocks for efficient VLSI implementation of soft-computing applications," *IEEE Trans. on Circuits and Systems I: Regular Papers*, vol. 57, 4 2010.

[6] N. Zhu, W. L. Goh, W. Zhang, K. S. Yeo, and Z. H. Kong, "Design of low-power high-speed truncation-error-tolerant adder and its application in digital signal processing," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 18, no. 8, 2010.

[7] D. Celia, V. Vasudevan, and N. Chandrathoodan, "Optimizing power-accuracy trade-off in approximate adders," in *2018 Design, Automation Test in Europe Conference Exhibition (DATE)*, March 2018.

[8] L. B. Soares, S. Bampi, and E. Costa, "Approximate adder synthesis for area- and energy-efficient fir filters in cmos vlsi," in *2015 IEEE 13th International New Circuits and Systems Conference*, June 2015.

[9] F. S. Snigdha, D. Sengupta, J. Hu, and S. S. Sapatnekar, "Optimal design of jpeg hardware under the approximate computing paradigm," in *2016 53rd ACM/EDAC/IEEE DAC*, June 2016.

[10] Z. Vasicek, V. Mrazek, and L. S. Brno, "Towards low power approximate dct architecture for hevc standard," in *Design, Automation Test in Europe Conference Exhibition (DATE), 2017*, March 2017.

[11] D. Sengupta, F. S. Snigdha, Jiang Hu, and S. S. Sapatnekar, "Saber: Selection of approximate bits for the design of error tolerant circuits," in *2017 54th ACM/EDAC/IEEE DAC*, June 2017.

[12] M. Pashaeifar, M. Kamal, A. Afzali-Kusha, and M. Pedram, "A theoretical framework for quality estimation and optimization of dsp applications using low-power approximate adders," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, Jan 2019.

[13] Z. Yang, J. Han, and F. Lombardi, "Transmission gate-based approximate adders for inexact computing," in *Proceedings of the 2015 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH15)*, July 2015.

[14] S. Bouguzel, M. O. Ahmad, and M. N. S. Swamy, "A fast 88 transform for image compression," in *2009 International Conference on Microelectronics - ICM*, Dec 2009.

[15] D. Sengupta, F. S. Snigdha, J. Hu, and S. S. Sapatnekar, "An analytical approach for error pmf characterization in approximate circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 38, Jan 2019.

[16] D. Celia, V. Vasudevan, and N. Chandrathoodan, "Probabilistic error modeling for two-part segmented approximate adders," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2018.

[17] A. Sripad and D. Snyder, "A necessary and sufficient condition for quantization errors to be uniform and white," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, October 1977.

[18] D. Menard, N. Herve, O. Sentieys, and H.-N. Nguyen, "High-level synthesis under fixed-point accuracy constraint," *Journal of Electrical and Computer Engineering*, vol. 2012, Jan. 2012.

[19] J. Liang, J. Han, and F. Lombardi, "New metrics for the reliability of approximate and probabilistic adders," *IEEE Transactions on Computers*, vol. 62, Sep. 2013.

[20] H. Jiang, J. Han, and F. Lombardi, "A comparative review and evaluation of approximate adders," in *Proc. of the Great Lakes Symposium on VLSI (GLSVLSI)*, 2015.