On The Probability Distribution Of Fixed-Point Multiplication

Arash Ahmadi and Mark Zwolinski
Electronic Systems and Devices Group
School of Electronics and Computer Science, University of Southampton, UK
Email: {aa5,mz}@ecs.soton.ac.uk

Abstract— **The area, delay and power consumption of functional units all are dependent on the word-length of the processed data. Since choosing an optimum word-length is a tradeoff between design costs and accuracy, it is essential to have an accurate model of computational errors for each operation. Errors resulting from the output rounding or truncation of the functional units is generally considered as uniformly distributed over the error limits regardless to the functionality of the unit. This work presents an analysis of the probability distribution of a fixed-point multiplier and its output truncation error. Simulations show that the uniform distribution assumption is violated in the case of truncation of a multiplication result of less than 8 bits.**

I. INTRODUCTION

One of the main objectives of hardware designers is to find an optimal design in terms of area, latency, throughput, and power consumption. The Word-Length (WL) of signals is one of the parameters that designers can modify to improve these metrics. In contrast to instruction processors, customizable hardware such as Field Programmable Gate Arrays (FPGAs) and Application-Specific Integrated Circuits (ASICs) provide freedom for WL-optimization in a given application. Hardware designers, however, face increasing difficulties in choosing the best WL. The objective is to find the minimal number of bits to represent a signal, while satisfying error constraints. A naive way to optimize WL is to evaluate various combinations one by one and observe the output for each design. This technique, however, involves an enormous search space and is not practical for large designs.

Generally, it is desirable to implement computational algorithms (DSP algorithms for instance) on cheap Fixed-Point (FP) hardware, but representing real vales with finite precision FP numbers is an error-prone practice [1]. From an optimization point of view, reducing the Word-Length (WL) of functional units means a cost reduction in the final design [2], [3]. On the other hand, reducing the WL of functional units to less than some minimum value produces an error in the output, which normally is modelled as additional white noise [4].

Multiplication has received significant attention because reducing the size of multipliers has a considerable impact on area and power consumption [5]. This issue can be even more important where there are resource restrictions. WL optimization methods can be divided into categories with respect to their approach to computational error modelling. First are methods, such as [3], which consider errors as additive noise to the data. The second approach deals with errors as interval or solid symbols which propagate, contract or dilate through the computation tree of the algorithm, [6]. Symbolic Noise Analysis (SNA) is based on Probability Density Function (PDF) propagation of the error. The PDFs of the errors at each point of the computation tree are modelled as noise symbols which can be combined to calculate the computational error. Since multiplication is an important and costly unit in computational algorithms, its output PDF and truncation error are required for accurate error modelling.

The paper is organized as follows: the background to the study is briefly reviewed in section two. Section three provides an analysis of the output PDF for a fixed-point multiplier. Truncation error of the multiplier is investigated in section four to provide a more accurate model and section five presents an application and some results.

II. BACKGROUND

Urabe, [7], presented a PDF for fixed point multiplication by proving this theorem: “In fixed-point multiplication, divide the range of the roundoff error into 2^n equal intervals. Then the probability for the roundoff error to fall into any one of these intervals converges to 2^{-n} as the number of digits of the factors is increased indefinitely.” He concluded that roundoff error occurs nearly at random if the number of digits of the factors is large. Goodman and Feldstein, [8], [9], also enumerated the round-off errors in fixed-point multiplication both for rounding by chopping and for symmetric rounding and computed the mean and variance.

Bareiss and Barlow, [10], constructed probabilistic models of floating point and logarithmic arithmetic using assumptions with both theoretical and empirical justifications. These models were applied to errors from sums and inner products. A comparison was made between the error analysis properties of floating point and logarithmic computers. They concluded that the logarithmic computer has smaller error confidence intervals for roundoff errors than a floating-point computer with the same computer word size and approximately the same number range.

Tokaji and Barnes, [11], presented a general statistical analysis of the roundoff error that is generated when a discrete random multiplicand, taking only integer values, is multiplied by a real coefficient and the result rounded back to the nearest integer. Numerical results were provided for mean, variance, correlation with multiplicand, and correlation between roundoff errors, as functions of multiplier coefficient value, variance of multiplicand, and correlation of multiplicands.

In this work we model the roundoff error from the WL optimization and computational algorithms' point of view, with more emphasis on a hardware synthesis approach but also agreement with the previous mathematical model. The bit-wise distribution of fixed-point multiplication PDF is discussed in the next section.

III. AN ANALYSIS OF MULTIPLIER OUTPUT PDF

The PDF of a multiplier's output must be considered in terms of continuous and discrete distributions. Continuous distributions deal with real numbers in mathematics while discrete distribution can be construed as integer numbers or fixed point representations in digital systems. The distribution of the product of two continuous random variables can be expressed as in Equation 1, [12].

$$F_{X \cdot Y}(a) = \int_{-\infty}^{+\infty} f_{X,Y}(x, \frac{a}{x}) \cdot \frac{1}{|x|} \cdot dx, \qquad (1)$$

where X and Y are continuous random variables and $f_{X,Y}(x, y)$ is the probability density function of $(x \neq 0, y)$. Finding a closed form

978-1-4244-2182-4/08/$25.00 ©2008 IEEE.

for this distribution function is not straightforward but some solutions have been proposed, [13].

However, in the case of integer numbers, this distribution is different. To compare the continuous and discrete distributions for fixed-point numbers, a set of simulations over a large number of random numbers ($> 10^{15}$) were performed to determine the frequency of the output numbers of a 16-bit multiplier for inputs with uniform distribution. The output distribution is shown in Figure (1).

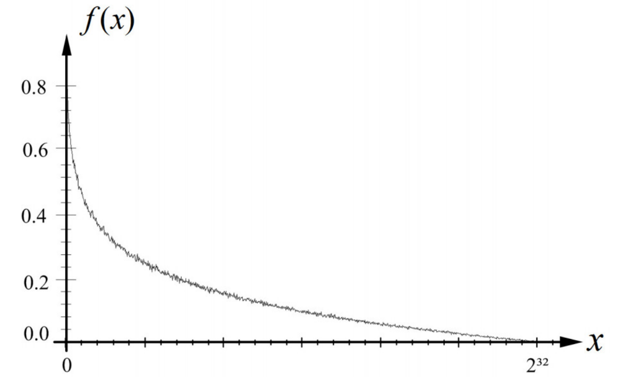

Fig. 1. PDF of 16-bit multiplier output with integer uniform distributed inputs

It can be seen from Figure 1 that multiplication results are sparser for larger numbers. In other words, to represent multiplication result of two n-bit and m-bit fixed-point numbers, we normally require an $(m+n)$-bit fixed-point number. An $(m+n)$-bit number can represent $2^{(m+n)}$ distinct numbers, where the number of distinct multiplication results of two n-bit and m-bit fixed-point numbers is far less than this value. It has been proved, and can be observed by an exhaustive search, that the multiplier's output does not cover the output range in full. The results of these investigations are in Table I where the first column shows the multiplier bit-width, the second column shows the number of distinct results, the third column indicates the total variation range of the output and the last column indicates what percentage of these possible numbers appear in the output of the multiplier. According to Table I, increasing the bit-width makes the output more sparse. The next section shows how this issue affects the truncation error PDF of the output.

TABLE I
MULTIPLIER OUTPUT COVERAGE FOR DIFFERENT WORD-LENGTHS

N×N-bit Multiplier	Number of distinct outputs	Total number of outputs	Output coverage %
1	2	4	50.00%
2	7	16	43.75%
3	26	64	40.63%
4	90	256	35.16%
5	340	1024	33.21%
6	1238	4096	30.23%
7	4647	16384	28.37%
8	17578	65536	26.83%
9	67592	262144	25.79%
10	259768	1048576	24.78%
11	1004348	4194304	23.95%
12	3902357	16777216	23.26%

IV. Truncation Error Model

The sparsity of the outputs in larger numbers suggests a non-uniform distribution for the bit pattern of the multiplication result. A set of simulations with random numbers were performed which gives a bitwise view of the matter. Figure 2 gives the probability of each bit of a 16 × 16 bit multiplier being “1”. According to this graph, the probability is not identical for all the bits of the output. Our results show that in an m-bit truncation of the multiplier with an N-bit output, the probability of a “1” occurring in the the k^{th} place of the output is:

$$P_1(k) = \frac{2^k - 1}{2^{k+1}}, \tag{2}$$

and accordingly the probability of a “0” occurrence in the the k^{th} place of the output is:

$$P_0(k) = \frac{2^k + 1}{2^{k+1}}, \tag{3}$$

where in the both formulas $k \leq \frac{m}{2}$. Practical results confirm these equations. Thus, a uniform PDF should not be expected for the output of a multiplier.

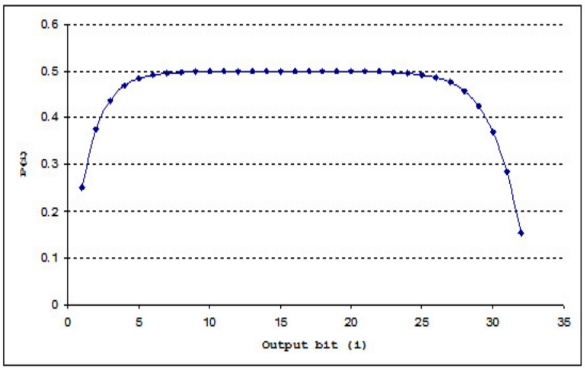

Fig. 2. Probability of each bit being “1” in the output of a 32-bit multiplier.

Figure 3 shows the PDF of the output error for an 8-bit truncation from 16-bits, derived from an exhaustive simulation of the truncated output of a multiplier for random uniform distributed inputs. It can be observed that this distribution is not uniform. Applying Equations 2 and 3 and according to the symmetry of the graph, the probability function can be formulated as:

$$\begin{aligned} P(x) &= \frac{P_1(k+1)}{2^{m-k}} \prod_{i=1}^{k} P_0(i), \qquad (4) \\ &= \frac{2^{k+1} - 1}{2^{m+2} \, . \, 2^{\frac{k(k+3)}{2}}} \prod_{i=1}^{k} \left(1 + 2^i\right), \end{aligned}$$

where k is the biggest integer number in the range $0 < k \leq m$ that 2^k divides x, which formally means:

$$k = \max \left\{ i \in \mathbf{N}, 0 \leq i \leq m, 2^i \middle| x \right\},$$

where $\mathbf{N}$ represents the set of positive integer numbers. In other words, the probability of the number x occurring in the output depends on the place of the first “1”, from the right hand side, in the base-2 representation of x, this PDF is depicted in Figure 3.

According to [14], quantization noise for a wide range of number distributions can be approximated by a random uniform distributed input. So in the case of multiplication, as shown in Figure 4, the m-bit truncation error of the output is only dependent on the least

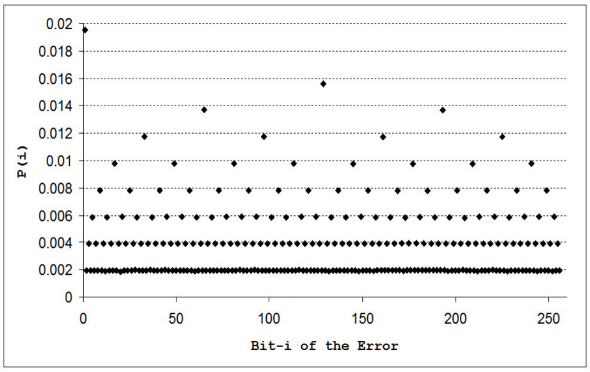

Fig. 3. PDF of 8-bit truncation error for 16-bit multiplier output with uniform distributed input.

significant m-bits of the inputs, and since m-bits of the inputs can be approximated by uniform distributed inputs, the multiplication truncation noise for a wide range of the inputs can be modelled as in Figure 3.

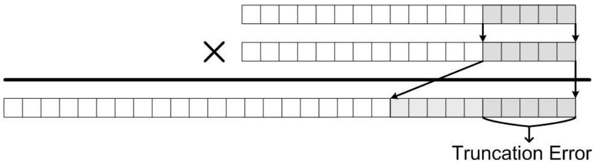

Fig. 4. Truncation error depends only on the LSB part of the inputs.

V. Application

In the digital representation of data, reducing the data bit-width has a direct effect on the accuracy, which is construed as computational error or noise. From this viewpoint, WL optimization methods can be categorized as range analysis or noise analysis. The former approach considers how the maximum/minimum values of signals propagate through the system from every input to every output. Accordingly, the result of the analysis is a range in which the output falls. Several methods have been introduced in this category such as Interval Arithmetic(IA) [15], Affine Arithmetic (AA) [6] and the Taylor Model [16]. These sub-categories differ in range representation and approximation. In the noise analysis approach, on the other hand, the outcome of the accuracy reduction is represented as a random process, which also called computational noise. Different characteristics of the computational noise has been investigated and they are commonly assumed to be Wide Sense Stationary (WSS) signals [14]. Inspired by analogue signal processing, most existing work utilizes the Signal-to-Noise Ratio (SNR) error criterion as the accuracy cost.

In our method a partially known quantity x is represented in SNA form as in Equation 5, [17].

$$x = T_N(\vec{E}), \tag{5}$$

where $T(\cdot)$ is a polynomial of order N with M known coefficients $(x_1, x_2, \cdots, x_M)$; and $\vec{E}$ is an array as in Equation (6).

$$\vec{E} = [x, \varepsilon_1, \varepsilon_2, \cdots, \varepsilon_m], \tag{6}$$

where x is the right-hand side of the Equation(5) and ε_i are symbolic representation of random values.

This model, called an algebraic representation, [16], covers a wide range of nonlinear relationships which can be expressed as algebraic relations. By eliminating x from $\vec{E}$ in Equation 6, Equation 5 will be reduced to an ordinary Taylor Model. Furthermore, the AA representation can be achieved with a first order Taylor Model as in Equation 7.

$$x = x_0 + \sum_{i=1}^{m} x_i \cdot \varepsilon_i, \tag{7}$$

where the x_0 is the original value, x is the rounded value, $x_i \in R$ are constants and $-1 \leq \varepsilon_i \leq +1$ are noise symbols. As in the AA analogy, these noise symbols are unknown symbolic variables in the range $[-1, +1]$. Every noise symbol has a known Source (S) in the computation DFG and a known Probability Density Function (PDF). Accordingly, in this study, any noise symbol is defined by two other symbols $\varepsilon_i = (S, P)$, in which S represents the noise source and P indicates the PDF type. Extending symbol variables ε_i into two symbols provides more information about noise at every point of the system, however it necessitates more computational effort during the optimization process, [17].

VI. Results

There is a variety of design techniques that can be used to implement a digital multiplier. Here a truncated Booth multiplier is used to evaluate the probabilistic characteristics of the output error. Area and power consumption dependency of the multiplier and its bit-width are given in Figure 5; this data is extracted for ST 130nm technology.

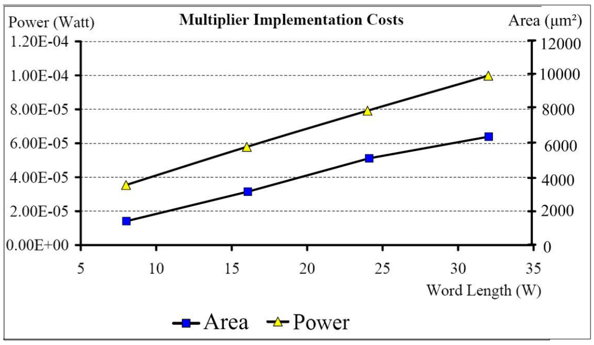

Fig. 5. Multiplier cost dependency to the word-length [2].

The mean value and variance of the truncation noise for different numbers of truncated bits are compared in Table II, these values are for positive integers when the truncation error is scaled up to $0 \leq Err < 2^m$. In Table II, the third column represents the variance of the error with a uniform PDF assumption and the fifth column gives the real value of the variance which is equal to the calculated value of our model. According to the table, by increasing the number of truncated bits, the uniform distribution model gets closer to the actual model. This result confirms the results in [10], where it is suggested that the roundoff noise of the fixed-point multiplication can be approximated by a uniform distribution for a large number of truncated bits m, and N, word-length of the multiplier output.

Another issue to be considered is the error PDF in the case of inputs with non-uniform distributions. Gaussian random number generators (GRNG) are used in a large number of computationally intensive modelling and simulation applications. A widely used method is Box-Muller, [18], in which a random number is produced with standard normal distribution from a standard uniform distributed random number by inverting the distribution function. A hardware implementation of this algorithm can be found in [19]. The truncation

TABLE II

COMPARING MEAN VALUE AND VARIANCE OF THE TRUNCATION NOISE FOR DIFFERENT NUMBERS OF TRUNCATED BITS.

m-bit	Uniform Distribution		Proposed Distribution	
Truncate	Mean	Variance	Mean	Variance
1	0.5	0.083333333	0.25	0.1875
2	1.5	0.75	0.999969	1.249939
3	3.5	4.083333333	2.750153	5.687454
4	7.5	18.75	6.501007	23.004912
5	15.5	80.08333333	14.252628	90.197059
6	31.5	330.75	30.005785	353.203323
7	63.5	1344.083333	61.769383	1392.149538
8	127.5	5418.75	125.486034	5518.875694
9	255.5	21760.08333	253.234952	21960.62686
10	511.5	87210.75	508.976108	87624.61602
11	1023.5	349184.0833	1020.686764	350021.5419
12	2047.5	1397418.75	2044.548389	1399220.573
13	4095.5	5591040.083	4092.004452	5594182.905
14	8191.5	22366890.75	8188.349327	22372861.38
15	16383.5	89473024.08	16379.13195	89482935.01

error for x and y is evaluated using extensive data inputs to find the PDF of the error. Figure 7 shows the error PDF for an 8-bit truncation of a 20-bit multiplier; error values are scaled to set 1 as the value of the LSB. It is observable that this distribution is symmetrical but not uniform.

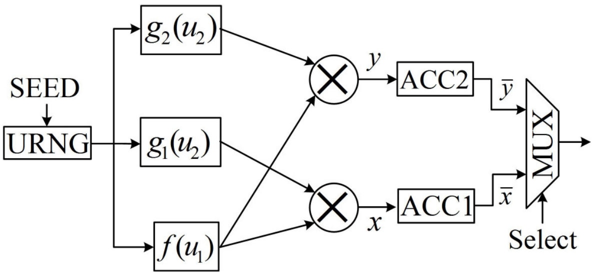

Fig. 6. Block diagram of the Box-Muller Gaussian random number generator [19].

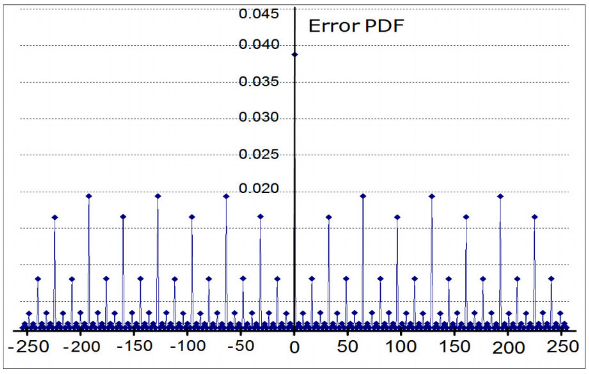

Fig. 7. Multiplier 8-bit truncation error after 20-bit multiplication.

VII. CONCLUSION

This study presents an investigation of fixed-point multiplication output and its truncation error PDF. Simulation results show deviation of the truncation error from a uniform PDF. This model is utilized in the SNA error analysis method for WL-optimization. It can be concluded that in word-length optimization, the PDF of the truncation error in N-bit fixed-point multipliers should be assumed uniform where the number of truncated bits, m, is greater than or equal to 8 and smaller than $\frac{N}{2}$. For N-bit fixed-point multipliers with m-bits ($m < \frac{N}{2}$) truncation, which $m < 8$, the PDF is different and needs to be considered for more precise error analysis.

ACKNOWLEDGEMENTS

The authors would like to thank Professor Jeffrey O. Shallit, School of Computer Science, University of Waterloo, for his comments on the distribution of multiplication results.

REFERENCES

[1] G. J. A. Bioul, J. P. Deschamps, and G. D. Sutter, *Synthesis of Arithmetic Circuits: FPGA, ASIC and Embedded Systems*. John Wiley & Sons Inc, March 2006.

[2] A. Ahmadi and M. Zwolinski, "Word-length oriented multiobjective optimization of area and power consumption in DSP algorithm implementation," in *The International Conference on Microelectronics*, May 2006, pp. 614–617.

[3] G. A. Constantinides, P. Y. K. Cheung, and W. Luk, *Synthesis and Optimization of DSP Algorithms (Fundamental Theories of Physics S.)*. Kluwer Academic Publishers, 2004.

[4] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*. Pearson US Imports and PHIPEs, 1999.

[5] I. Koren, *Computer Arithmetic Algorithms*. Natick, MA, USA: A. K. Peters Ltd., 2001.

[6] D.-U. Lee, A. Abdul-Gaffar, R. C. C. Cheung, O. Mencer, W. Luk, and G. A. Constantinides, "Accuracy-guaranteed bit-width optimization," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 25, no. 10, pp. 1990–2000, October 2006.

[7] M. Urabe, "Roundoff error distribution in fixed-point multiplication and a remark about the rounding rule," *SIAM Journal of Numercal Analalysis 5, 202 (1968)*, vol. 5, no. 2, pp. 202–210, June 1968.

[8] R. Goodman and A. Feldstein, "Round-off error in products," *Computing*, vol. 15, no. 3, pp. 263–273, September 1975.

[9] R. Goodman, "On round-off error in fixed-point multiplication," *BIT Numerical Mathematics*, vol. 16, no. 1, pp. 41–51, March 1976.

[10] E. H. Bareiss and J. L. Barlow, "Roundoff error distribution in fixed point multiplication," *BIT Numerical Mathematics*, vol. 20, no. 2, pp. 247–250, July 1980.

[11] I. Tokaji and C. W. Barnes, "Roundoff error statistics for a continuous range of multiplier coefficients," *IEEE Transactions on Circuits and Systems*, vol. 34, no. 1, pp. 52–59, January 1987.

[12] J. L. Hodges, E. L. Lehmann, and J. L. Hodges, *Basic Concepts Of Probability And Statistics*, 2nd ed. Society for Industrial and Applied Mathematic, January 2005.

[13] A. G. Glen, L. M. Leemis, and J. H. Drew, "Computing the distribution of the product of two continuous random variables," *Computational Statistics and Data Analysis*, vol. 4, no. 3, pp. 451–464, January 2004.

[14] A. B. Sripad and D. L. Snyder, "A necessary and sufficient condition for quantization errors to be uniform and white," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 5, pp. 442–448, October 1977.

[15] H. Keding, M. Willems, M. Coors, and H. Meyr, "FRIDGE: a fixed-point design and simulation environment," in *DATE '98: Proceedings of the conference on Design, automation and test in Europe*. Washington, DC, USA: IEEE Computer Society, 1998, pp. 429–435.

[16] N. S. Nedialkov, V. Kreinovich, and S. A. Starks, "Interval arithmetic, affine arithmetic, taylor series methods: Why, what next?" *Numerical Algorithms*, vol. 37, no. 1-4, pp. 325–336, December 2004.

[17] A. Ahmadi and M. Zwolinski, "A symbolic noise analysis approach to word-length optimization in DSP hardware," in *ISIC'07: International Symposium on Integrated Circuits*, September 2007, pp. 497–500.

[18] G. E. P. Box and M. E. Muller, "A note on the generation of random normal deviates," *The Annals of Mathematical Statistics*, vol. 29, no. 2, pp. 610–611, January 1958.

[19] D.-U. Lee, S. J. D. Villasenor, W. Luk, and P. H. W. Leong, "A hardware Gaussian noise generator using the Box-Muller method and its error analysis," *IEEE Trans. Comput.*, vol. 55, no. 6, pp. 659–671, 2006.