A cost-error tunable round-off method: Finite-length absorption

Hamid Reza Mahdiani[1,2a] **and Sied Mehdi Fakhraie**[1b]

[1] *School of ECE, University of Tehran,*
North Kargar Ave., Tehran 14395–515, Iran

[2] *Computer and Electronics Department, Sh. Abbaspour University of Technology,*
Tehranpars, Hakimie, Tehran, Iran

a) *mahdiany@ut.ac.ir*

b) *fakhraie@ut.ac.ir*

Abstract: Although the round-off noise is normally small, it might accumulate and significantly degrade the output quality in most computationally intensive emerging applications such as FFT calculation in data communication protocols or neural networks. A new hardware-friendly error-tunable round-off method is introduced. The most important feature of this new method is that it can provide acceptable and arbitrary selectable accuracies along with different hardware implementation costs.

Keywords: round-off noise, truncation, rounding

Classification: Integrated circuits

References

[1] J. Cong, Y. Fan, G. Han, Y. Lin, J. Xu, Z. Zhang, and X. Cheng, “Bitwidth-aware scheduling and binding in high-level synthesis,” *ASP-DAC*, vol. 2, pp. 856–861, Jan. 2005.

[2] G. A. Constantinidis, “Word-length optimization for differentiable nonlinear systems,” *ACM Trans. Design Autom. Electron. Syst.*, vol. 11, no. 1, pp. 26–43, Jan. 2006.

[3] D. U. Lee, A. A. Gaffar, O. Mencer, and W. Luk, “MiniBit: bit-width optimization via affine arithmetic,” *DAC*, June 2005.

[4] N. Doi, T. Horiyama, M. Nakanishi, and S. Kimura, “Minimization of fractional wordlength on fixed-point conversion for high-level synthesis,” *ASP-DAC*, pp. 80–85, Jan. 2004.

[5] G. A. Constantinides, P. Y. K. Cheung, and W. Luk, “Synthesis of saturation arithmetic architectures,” *ACM Trans. Design Autom. Electron. Syst.*, vol. 8, no. 3, pp. 334–354, July 2003.

1 Introduction

The cost/accuracy trade-off is the most important factor to decide about when choosing the hardware to implement any computational system and

DOI: 10.1587/elex.6.1312
Received June 28, 2009
Accepted August 24, 2009
Published September 25, 2009

the Word Length (WL) is the dominant factor that determines this trade-off [1]. Increasing the WL significantly increases the accuracy as well as the system implementation costs [1, 2]. WL optimization problem is performed in two parts [3]. The first part is range analysis that optimizes the 'integer length' to prevent most of the overflow situations that produce rare but large errors. On the other side, the precision analysis is the next step that tries to find the best trade-off between the round-off errors and the 'fraction length'.

Although the round-off errors are normally very small values, need special care since they frequently occur in every system and might accumulate and highly degrade the output accuracy. Increasing the fraction length results in lower round-off errors as well as larger and slower arithmetic components or higher implementation costs. However; there are some lower cost round-off handling techniques that improve the output accuracy of the system without increasing the system WL. A new round-off technique is introduced in the next section. The other sections compare the accuracy and hardware implementation costs of this method with respect to other common round-off techniques.

2 The "Finite-Length Absorption" (FLA) round-off method

Truncation (TR), biased-rounding (BR), and Unbiased-Rounding (UBR) are the most important and popular round-off strategies in hardware [4]. The TR only discards the extra bits which are out of the WL range. The BR discards the extra bits and then rounds up or down the remaining bits while the most significant discarded bit is '1' or '0' respectively. The UBR or banker's rounding rounds-up or down the numbers similar to BR. The only difference arises when the discarded bits are just equal to the half range (i.e. all discarded bits are '0' except for the most significant bit). In this case, UBR rounds up or down the remaining bits with 50% probability. There are some other round-off strategies such as floor, ceiling or round-away from zero however; they are not explicitly included in this paper. Because they are either similar to BR/UBR or they are not commonly used in hardware realizations.

The common base of all round-off strategies (other than TR) is that they define some situations in which the remaining bits should be roundup to compensate the round down errors and provide a moderate overall error. The roundup process implies adding one unit or incrementing the remaining bits. This might introduce a carry that can propagate toward the more significant remaining bits according to the values of the low order bits string. As the worst case, this carry may even result in an overflow situation when all the remaining bits are '1'. So the rounding process output should be saturated to prevent the overflows [5].

The FLA method rounds up and down the numbers similar to BR. The main difference is that FLA modifies the roundup process and does not let the carry to freely propagate through any number of remaining bits. In case of roundup, it tries to absorb the carry only within the first 'A' low order

DOI: 10.1587/elex.6.1312
Received June 28, 2009
Accepted August 24, 2009
Published September 25, 2009

remaining bits which are called ‘Absorption Bits’. The ‘A’ or the number of absorption bits is called ‘Absorption Length’ (AL). If it fails to absorb the carry, cancels the roundup process and simply discards the carry that corresponds to a round down. The absorption fail probability exponentially decreases as ‘A’ increases and is equal to $1/2^A$ (for when all the absorption bits are equal to ‘1’). In the following, the FLA_M abbreviation stands for the FLA method with an absorption length of ‘M’.

3 Accuracy simulation results of round-off methods

The mean squared error (MSE) or noise power of different round-off methods are measured by exhaustive simulations. The importance of MSE is due to its direct relation with the overall system signal to noise ratio (SNR) [3]: $SNR \approx 1/MSE$. Fig. 1 illustrates the logarithmic MSE values of new FLA method (with different ALs) vs TR, BR and UBR. The MSE values are shown for different WLs from 4 to 14 while the length of the discarded bits is equal to WL. In other words, the length of the number before rounding is twice as its length after rounding that really occurs at the output of the hardware multipliers. The figure shows that TR method always provides the worst accuracy while BR and UBR methods provide the best and MSE values. As to FLA method, it is interesting to note that it provides gradually-improving accuracy levels between the worst case MSE of TR and the best case MSE of BR\UBR according to AL value.

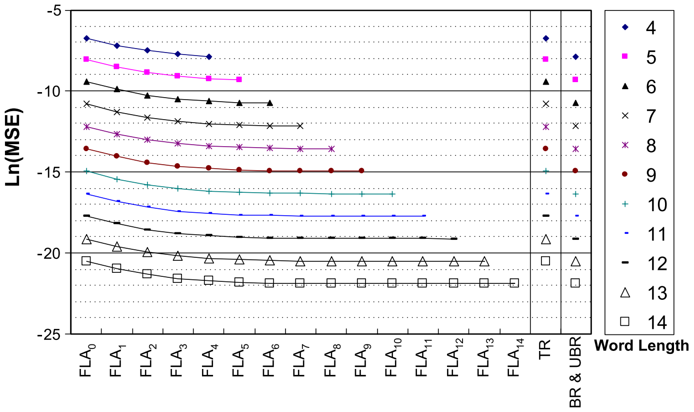


Fig. 1. Logarithmic Mean Squared Error of round-off methods for different WLs.

The results of Fig. 1 that are verified by analytic studies show that FLA method provides the same error values of the TR method when the AL is zero. As the AL increases, the error decreases and when the AL is equal to N, the FLA method behaves like the BR and UBR methods. The below equations describe the important relations

$$\lim_{A\to 0} MSE(FLA_A) \to MSE(TR) \tag{1}$$

$$\lim_{A\to N} MSE(FLA_A) \to MSE(BR) \tag{2}$$

$$\lim_{A\to N} MSE(FLA_A) \to MSE(UBR) \tag{3}$$

DOI: 10.1587/elex.6.1312
Received June 28, 2009
Accepted August 24, 2009
Published September 25, 2009

Other simulation results show that Equations (1), (2) and (3) are valid between "Average Error" values of those methods. Therefore, the first important feature of the FLA method is that it can be fine-tuned to achieve different either "MSE" or "Average Error" values on an accuracy spectrum with the TR and BR\UBR methods on its two extremes.

The logarithmic curves also demonstrate another important property of FLA. They show that for all WLs, the accuracy improves significantly with the increase of AL when AL < 5. As the AL increases more than 5, the slopes of all curves are tend to zero and so, there is no important improvement. The simulations show that FLA_4 and FLA_5 round-off methods provide maximum 19% and 9% worse MSE in comparison with the best existing methods (BR/UBR) for all WLs. The results also show that when AL is equal to 10, the difference between MSE values of FLA_{10} and BR/UBR is less than 0.44% regardless of WL, which is negligible. These results imply that the BR/UBR round-off methods might be replaced with FLA_{10} which has less implementation cost and higher performance (as shown in the next section) while the overall system SNR remains unchanged. Also the results propose that some other versions of FLA (such as FLA_5), might be considered as other good replacements for BR/UBR methods in some applications that can tolerate a small SNR degradation.

4 Hardware implementation results of round-off methods

Fig. 2 (a) shows the hardware structure of UBR method. The structure of BR method is similar to UBR while it does not include the half range detector circuit and the glue logic. The N-bit half adder is used to roundup the remaining bits when necessary. The outputs of the half adder are than connected to the saturate logic to prevent the potential overflow situations that might be caused by roundup process. Fig. 2 (b) demonstrates the internal structure of the saturate logic. It is a simple block that consists of 'N' parallel OR gates to saturate all outputs of the half adder to '1' when a carry out is produced by N-bit half adder.

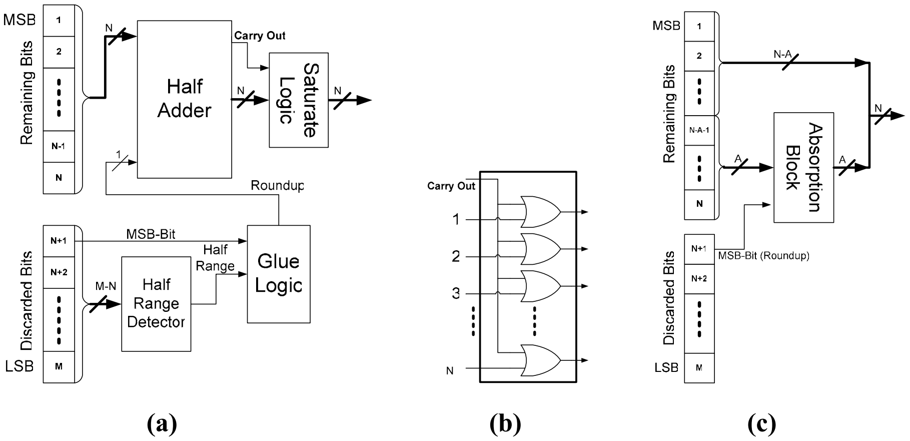


Fig. 2. Hardware structures of the (a) UBR, (b) Saturate Logic and (c) FLA.

© IEICE 2009
DOI: 10.1587/elex.6.1312
Received June 28, 2009
Accepted August 24, 2009
Published September 25, 2009

Fig. 2 (c) also shows the FLA hardware structure. The absorption block is a dedicated block whose role is to absorb the potential produced carry within the first ‘*A*’ lower remaining bits. To have a fair comparison between synthesis results of the FLA method with respect to BR and UBR, in this paper the absorption block is made in a similar way that is composed of an ‘*A*’ bit half adder followed by a saturate logic while the saturate logic contains ‘A’ parallel OR gates.

Equations (4) to (9) demonstrate the gate count and gate delay of different round-off methods as a function of the remaining bits length (WL), Absorption length (A) and Discarded bits length (D).

$$GateCount(FLA_A) = 4 \times A \tag{4}$$

$$GateDelay(FLA_A) = A + 1 \tag{5}$$

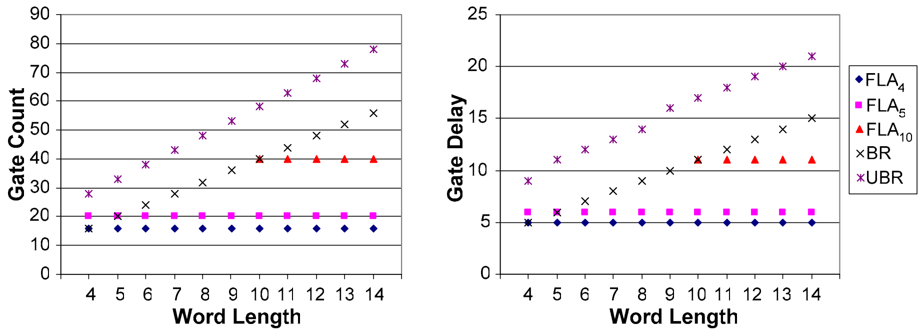
$$GateCount(BR) = 4 \times WL \tag{6}$$

$$GateDelay(BR) = WL + 1 \tag{7}$$

$$GateCount(UBR) = 4 \times WL + D + 8 \tag{8}$$

$$GateDelay(UBR) = WL + 3 + \lceil \log_2(D) \rceil \tag{9}$$

Fig. 3 shows the gate count (Fig. 3 (a)) and gate delay (Fig. 3 (b)) of different round-off methods for different WLs based on above equations. The length of discarded bits is equal to WL. It shows that FLA_4, FLA_5, and FLA_{10} provide constant and better area/delay with respect to BR and UBR methods regardless of the WL. It should be simultaneously considered that these methods provide only less than 19%, 9% and 0.44% worse MSE with respect to BR/UBR. As an instance, for WL = 16, the FLA10 provides the same MSE values of BR and UBR methods. Also it improves the area and speed of the round-off circuit about 35% and 50% with respect to BR and UBR respectively.



(a) **(b)**

Fig. 3. Gate count (a) and gate delay (b) of round-off methods.

5 Conclusion

A new round-off method with tunable precision and implementation cost is introduced in this paper. It can provide different cost/accuracy trade-offs

© IEICE 2009
DOI: 10.1587/elex.6.1312
Received June 28, 2009
Accepted August 24, 2009
Published September 25, 2009

and so can accommodate many applications with different noise sensitivities. The accuracy and cost of this new method are compared with other round-off methods to show its efficiency.

DOI: 10.1587/elex.6.1312
Received June 28, 2009
Accepted August 24, 2009
Published September 25, 2009