Advanced Design And Implementation Of Reducing Power Reduction In Multiplier Through Dynamic Fine Control

M. Zahir Ahmed, Lecturer in Physics, Osmania College (Autonomous),Kurnool-518001

Andhra Pradesh, India.

Abstract:

In many applications such as artificial intelligence, computer vision, multitasking, image recognition and digital signal processing, the connections are simple arithmetic operations. The use of heavy equipment leads to large power consumption, which is very difficult, especially in weak environments such as mobile devices. We offer a unique way to solve this problem by reducing energy consumption by controlling accuracy. Our solutions are based on the concept of flexible switching, which can reduce energy consumption and increase performance at a certain cost. We present a new model that enables adaptive changes in accuracy based on application errors. The basis of our idea is to create a 4-2 compressor with extraordinary precision, which will work as the basis of a dynamic precision control system. We also include powerful error compensation to improve the accuracy of the addition process. In addition, by integrating parallel assertions into a convolutional neural network (CNN), we demonstrate the simplicity and flexibility of the real-time control system. As this integration shows, the multiplier is an excellent solution for deep learning applications, demonstrating the ability of each layer to meet different and energy needs. As a result, our scheme will be a good idea to reduce electricity consumption. Enabled by groundbreaking dynamic accuracy control, on-the-fly adjustments intelligently reduce power consumption while maintaining control's received learning levels. Our company's design offers an attractive next-generation energy efficiency solution because it reduces energy consumption and can be optimized to meet different needs of the application.

Index terms: 4-2 Compressor, CNN, Power consumption, accuracy control

1. Introduction:

Multiplier is one of the most important mathematical functions in many applications such as signal processing (DSP), computer vision, multiprocessing, image recognition, and wise man. These applications usually require a lot of parallelism, which results in large power consumption[1]. High power consumption poses a challenge to these applications, especially in mobile devices. Therefore, many studies propose strategies to reduce energy consumption of connected products. One solution to reduce more power consumption is to use approximate equations if the target of the application is to allow violation or affect the hearts of the same people, and vice versa[2]. Due to the limitations of the human mind, such as limited vision and hearing, it is not necessary to calculate the correct results. Approximate 2*2 bit multiplier is prepared by updating the Karnaugh map (K-Map) once. 3 bits (including 4 bits) are used for the output of

2*2 integers[3]; The design reduces the power consumption from 31.78% to 45.4% with an average error of only 1.39% to 3.32%. Different ways to generate approximations include the use of racks or pumps as used in Wallace wood or Dada wood; for approximating 4*4 bit Wallace multipliers, erroneous 4:2 racks are requested. A large multiplier is designed to use approximating 4*4 bit multipliers to achieve short latency, low power consumption, and more. Two approximating 4*2 compressors are created by changing a few values in the truth table; four samples are then requested using the approximating 4*2 compressor in a Dada tree with 8*8 multipliers. Power consumption and latency are reduced[4]. A new approximating multiplier is prepared using an approximating adder that ignores the carry propagation of some of the product. 8 8-bit multipliers provide 20% lower latency and 69% lower power consumption. Most of the existing equation models consider the partial division and integration stage, and almost all of them are for the operation of unsigned numbers.

2.Existing system:

The current system in the current system (Vos) is not strong enough to mitigate many attacks without compromising the main power supply. However, VOS can cause significant signal-to-noise ratio (SNR) degradation. The new Algorithmic Noise Tolerance[5] (ANT) technology combined with the VOS main block with reduced repeatability (RPR) to effectively prevent false positives while saving energy is important in terms of electricity. Some ANT deformation designs have been proposed and the ANT concept design has been extended to the system level. Combining ANT with VOS enables the creation of a low-power signal that operates with more energy than existing systems[6].

Important Procedures A new ANT procedure called Reduced Precision Redundancy (RPR) aims to combat the light-weight problems while providing significant energy savings. RPR uses a scaled-down version of the DSP system [called the main DSP (MDSP)] to identify and correct errors occurring at the output of the MDSP system[7]. Existing RPR-based ANT technology differs from predictive error control (PEC) or adaptive error cancellation (AEC) schemes.

The RPR design in ANT design is a structure and is not easy to adopt and copy. The RPR built into ANT models can run very fast, but their hardware is very complicated. Therefore, the RPR design of ANT designs is still the most popular model because of its simplicity. However, using RPR still requires additional space overhead and power consumption[8]. There is no need to pay because the payment will be insufficient.

2.Proposed System:

Approximate mode it reduces the power consumption and area at the cost of low accuracy .when it operates at exact mode it produces the exact results and has high power consumption and area. In this paper, we propose four 4:2 compressors, which have the flexibility of switching between the exact and approximate operating modes.

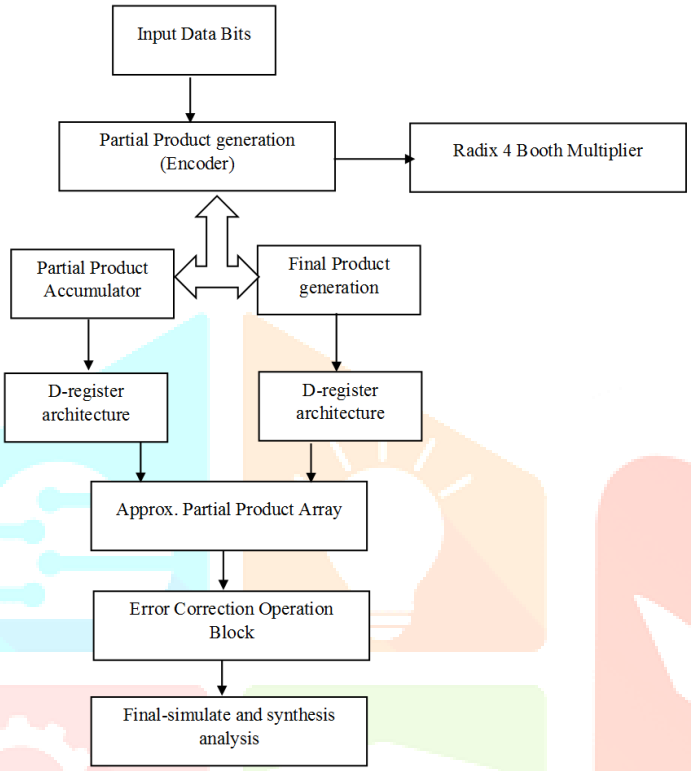

Fig.1. Proposed system block diagram

1. In the approximate mode, these dual-quality compressors provide higher speeds and lower power consumptions at the cost of lower accuracy[9]. Using these compressors in the structures of parallel multipliers provides configurable multipliers whose accuracies may change dynamically during the runtime.
2. ***Data splitting:*** In this step, the pre-processed data's are split into train set and test set for decision
   - ✓ ***Train data*** *is used for evaluate the model (80%).*
   - ✓ ***Test data*** *is used for predict the model (20%).*
3. ***Classification:*** In this step, we can implement the Deep learning algorithms such as,
   - ✓ *CNN*
4. It is based on Test bench using Verilog Code
5. ***Output/objective****:* The objective is ***to classify or predict the Type*** based on dataset attributes by using the ***classification*** algorithms.

6. ***Performance Estimation:*** In this step, we can analyse some performance metrics such as,

- ✓ *Accuracy control*
- ✓ *Precision*
- ✓ *Recall*
- ✓ *Area*
- ✓ *Power*
- ✓ *Delay*

We propose a high accuracy approximate 4-2 compressor that can be used to construct the proposed approximate multiplier. We design a simple error compensation circuit to further reduce the error distance. We propose a dynamic input truncation technique that can be used to adjust accuracy and power required for a multiplication[10]. The proposed technique is suitable for CNNs as power consumption can be easily adjusted depending on the different requirements for each layer. Based on the proposed 4-2 compressor, error compensation circuit and the dynamic input truncation technique, we propose a high-accuracy and reconfigurable approximate multiplier.

3.Methodology:

Approximation equations can be constructed in various ways, such as measuring the benefit, halving some products, constructing an equivalent equation model, and using an approximation compressor to halve the number of parts. It controls the power supply of the logic gate, which helps reduce power consumption. If the supply voltage is lower than the desired nominal voltage, a timing violation will occur, which will lead to predictable results. However, if the timing of the violation occurs significantly, the error rate can be very large. From these, the approximation equation is constructed by truncating a portion of the multiplication column close to the least significant (LSB) column to reduce the length of the spread. The closer the partial product is to the LSB column, the smaller the weight of the partial product. Since the weight of the cut part is small, it will not cause a large error in the distance. In the planning process, a simple way is proposed to use full-width RPRs instead of full-width RPR blocks. By using wide-width RPRs[11], errors in calculations can be corrected with lower power consumption and smaller area. We use probabilistic, statistical and partial weight analysis to find estimated compensation vectors for various realistic RPR designs. In order not to increase the priority of the delay, the system limits the amount of electricity in the RPR, which should not be in the priority path[12]. Therefore, the system can realize ANT design with small area, low power consumption and less importance of electronic equipment. ANT technology includes a main digital signal processor (MDSP) and error correction (EC) blocks. In fact, when using ANT architecture, it has lower power consumption and smaller area, but also higher SNR, higher performance, lower performance and

lower power consumption[13]. This system presents our extensive RPR based on ANT design in ANT multiplier.

Advantages:

1.RPR design in ANT designs can run very fast. Support logic gate. Lower usage and lower local overhead.

2.Wallace Multiplier:

Wallace Multiplier is proposed as a fast multiplicative algorithm called Wallace tree multiplier. In Wallace tree multipliers, exactly 2-2 compressors (half adders) and 3-2 compressors (full adders) are used to reduce the number of rows of half multiplications. A transparent 4-2 compressor is also available with the Wallace Tree[14] Multiplier to create a more compact setup. The reduced product is collected between the carry propagation collectors to obtain the final product. Due to its popularity, most of the past work was designed to approximate compressors instead of the 4-2 compressor. Details of the actual and approximate compressors are explained in the next column.

3.CNN Implementation:

Our proposed approximation is suitable for quantized CNNs because we can adjust our approximation to determine the number of objects that allow the calculation to be inaccurate at runtime. In each convolution layer, we use the 4-bit truncation configuration parameter Trunc to specify which partial object rows to discard. By using the truncation parameters, we can dynamically adjust our multiplier idea according to the decimal point position of the quantized weight to reduce the power consumption of many MACs.

4.Quantization:

The weights of CNN are usually in floating-point number format, which makes the hardware cost for CNN implementation high. Therefore, in this example study, the floating-point bracket is first converted to 8-bit integer fixed-point weight and then used in CNN using the prepared quantization method. We create a scale in floating-point format representing the division of each layer. Before the results of the previous layer are transferred to the next layer, they need to be evaluated and optimized, the evaluation factor is approximately a multiple of 2, and the function is used to change the distribution in order to reduce the competition[15]. Approximate scaling to the power of 2 allows us to determine the number of elements that are changed, this is called the change count. Since the result needs to be changed before moving to the next layer, some of the result will be discarded; therefore, even if the parts to be removed are cut, the actual result will not be affected too much. Therefore, we use the transition number of each layer to determine the Trunc signal. The larger the change, the more we can cut without affecting the result too much. We try several Trunc signals using the number of variables obtained from each layer and choose the one that has a good balance between accuracy and utility[16]. In other words, we use different configuration of the desired prediction coefficient as the weight of each layer, which will be discussed in the next subsection.

Advantage:

Equal numbers should have the following properties:-

1. Accurate: - A good equation should give the correct result.
2. Speed: - The multiplier should work fast.
3. Area: - Candidates should have minimum cut-off and LUT score.
4. Power: - Multiplier should consume less power

5. Conclusion:

In this paper, approximate radix-8 Wallac multiplier of 12*12 bits with different signs is constructed. Initially, a 2-bit adder with a 3-input XOR gate was intended to multiply binary numbers. Error detection, compensation, and recovery circuits are also provided for the approximation of the 2-bit adders. The lower part of the adder is approximated using a 2-bit adder to produce a triplet match without spreading; truncation is used in the implementation of signature estimation of the Radix-8 Wallac multipliers (called ABM1 and ABM2) to further save energy and time. Then parallel processing of Wallace tree is used to speed up the partial processing. Simulation results show that the proposed adders (ARA8, ARA8-2C and ARA8-2R) are suitable for radix-8 Wallac multipliers (in terms of hardware and accuracy) compared to other approximate adders. Recoding of adders is important in terms of significant latency of multipliers. However, the error caused by recoding of adders is larger than the error caused by truncation (assuming that the truncation number of partial multiplication is less than or equal to 9 to 12*12 bit multiplier). Simulation results of FIR filter implementations show that the proposed ABM1, ABM2-C9 and ABM2-R9 perform well with only 3 dB degradation in output signal-to-noise ratio. The proposed design is better than other equivalents in FIR filter operation at similar PDP values, and therefore this design can be used because of poor performance and poor accuracy.

6.Future Development:

Therefore, in future development, we use the addition and modification of wires to improve the previous level and operating speed. In the process, we improve the 12*12-bit input to 16*16, 24*24 and 32*32, and also reduce the power consumption and competition in the circuit.

7. References:

[1] S. Venkataramani, S. T. Chakradhar, and K. Roy. “Computing approx imately, and efficiently,” Proc. Design, Automation & Test in Europe Con ference & Exhibition (DATE), 2015, pp.748-751.

[2] J. Han and M. Orshansky, “Approximate computing: an emerging par adigm for energy-efficient design,” Proc. 18th IEEE European Test Sym posium, 2013, pp.1-6.

[3] H. Mahdiani, A. Ahmadi, S. Fakhraie, and C. Lucas, “Bio-in spired imprecise computational blocks for efficient VLSI imple mentation of soft-computing applications,” IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 57, pp. 850-862, 2010.

[4] V. Gupta, D. Mohapatra, S. Park, A. Raghunathan, and K. Roy, "IMPACT: IMPrecise Adders for Low-Power ApproximateComputing," Proc. Int. Symp. Low Power Electronics and Design (ISLPED), pp. 1-3, 2011.

[5] W. Liu, L. Chen, C. Wang, M. O'Neill and F. Lombardi, "Design and analysis of floating-point adders", IEEE Trans. Computers, vol. 65, pp. 308-314, Jan. 2016.

[6] J. Liang, J. Han, and F. Lombardi, "New metrics for the reliability of approximate and probabilistic adders," IEEE Trans. Computers, vol. 63, pp. 1760-1771, Sep. 2013.

[7] A. Wallac, "A signed binary multiplication technique," Quarterly J. Mechanics and Applied Mathematics, vol. 4, pp/236-240, June 1951.

[8] O. MacSorley, High-speed arithmetic in binary computers, Proc. IRE, vol. 49, pp. 67-91, 1961.

[9] K.-J. Cho, K.-C. Lee, J.-G. Chung, and K. K. Parhi, "Design of low error fixed-width modified Wallac multiplier," IEEE Trans. VLSI Systems, vol. 12, no. 5, pp. 522–531, 2004.

[10] M. J. Schulte and E. E. Swartzlander Jr., "Truncated multiplication with correction constant," Proc. Workshop VLSI Signal Process. VI, 1993, pp. 388–396.

[11] E. J. King and E. E. Swartzlander Jr., "Data dependent truncated scheme for parallel multiplication," Proc. 31st Asilomar Conf. Signals, Circuits Syst., 1998, pp. 1178–1182.

[12] J.-P. Wang, S.-R. Kuang, and S.-C. Liang, "High-accuracy fixed-width modified Wallac multipliers for lossy applications," IEEE Trans. VLSI Systems, vol. 19, no. 1, pp. 52–60, 2011.

[13] C.-Y. Li, Y.-H. Chen, T.-Y. Chang, and J.-N. Chen, "A probabilistic estimation bias circuit for fixed-width Wallac multiplier and its DCT applications," IEEE Trans. Circuits and Systems II: Express Briefs, vol. 58, no. 4, pp. 215–219, 2011.

[14] Y.-H. Chen, C.-Y. Li, and T.-Y. Chang, "Area-effective and power efficient fixed-width Wallac multipliers using generalized probabilistic estimation bias," IEEE J. Emerging and Selected Topics in Circuits and Systems, vol. 1, no. 3, pp. 277–288, 2011.

[15] Y.-H. Chen and T.-Y. Chang, "A high-accuracy adaptive conditionalprobability estimator for fixed-width Wallac multipliers," IEEE Trans. Circuits Syst. I: Reg. Papers, vol. 59, no. 3, pp. 594–603, 2012.

[16] P. Kulkarni, P. Gupta, and M. Ercegovac, "Trading accuracy for power with an under designed multiplier architecture," Proc. 24th Int. Conf. VLSI Design, 2011, pp. 346–351.