A Low-Power Multiplication Accumulation Calculation Unit for Multimedia Applications

Oscal T.-C. Chen, Nan-Ying Shen, Chih-Chien Shen

Signal and Media Laboratories,
Dept. of Electrical Engineering,
National Chung Cheng Univ.,
Chia-Yi, 621 Taiwan
Email:oscal@ee.ccu.edu.tw

Abstract

In this work, a low-power Multiplication Accumulation Calculation (MAC) unit is proposed. In the multiplication process, the scheme of Booth encoding one of two input data with the smaller effective dynamic range is utilized to increase the probability of partial products being zero. In the addition process, the scheme is to make an adder operating at the larger effective dynamic range of two input data. The switching activities of this adder in the non-effective range can be minimized. These two schemes are integrated to design the proposed low-power MAC unit. By using the cell-based library of the TSMC 0.35um CMOS technology, the proposed and conventional MAC units based on the Farag's, Kwon's and modified Yu's architectures are implemented. In practical applications, power analysis was conducted on operations of the G.723.1 speech coder, ADPCM audio coder and wavelet transform in JPEG 2000. The proposed MAC unit using the modified Yu's architecture can save power dissipation up to 35.3%, 21.6%, and 36.3%, respectively, in these three applications. Taking into consideration of the product factor of power consumption, hardware area and critical delay, the proposed MAC units still have better performances than the conventional ones. Therefore, the proposed MAC unit can consume low power for multimedia computing at a little increasing hardware area.

1. Introduction

Nowadays, most digital signal processing in multimedia applications requires addition, multiplication, multiplication accumulation operations and so on [1]. The MAC unit becomes a basic component in most industrial digital signal processors because it can perform addition, multiplication and multiplication accumulation operations. In the literature, Kwon *et al*. proposed the MAC unit based on fast 5:2 compressors instead of 3:2 and 4:2 compressors [2]. Additionally, the modified logical decomposition of multiplication and the carry-save structure of the accumulation were designed to have speed improvement. Farag *et al*. developed a power efficient MAC unit for finite impulse response filters where their MAC unit can be programmed with different bit resolutions [3]. In the multiplier design of a MAC unit, Yu *et al*. reorganized the structure of the carry-save array multiplier using the Booth algorithm to reduce power consumption where the structure is to perform accumulation of partial products generated from the most significant bits first and then partial products from the least significant bits [4].

In CMOS circuits, power dissipation of a MAC unit mainly comes from the switching activities of its functional blocks, as represented by the following equation [5]:

$$P_{switching} = \alpha C V_{dd}^{2} f_{clk} \qquad (1)$$

where α is the switching activity parameter, C is the loading capacitor, V_{dd} is the operating voltage, and f_{clk} is the operating frequency. Here, αC could also be viewed as effective loading capacitor during switching operations. With this in mind, the best way to lower power dissipation without affecting the circuit operation performance is to minimize switching activities of a MAC unit.

The conventional MAC unit as shown in Fig. 1 includes input data latches, a multiplier, a multiplication output latch and an adder. Its operation is performed on the full word length of each input data. In this work, the MAC unit's multiplier performs on the Booth encoding one of two input data with the smaller effective dynamic range, whereas the MAC unit's adder performs on the larger effective dynamic range of two input data [6-8]. The proposed MAC unit can change its operational modes according to effective dynamic ranges of two input data. Figure 2 shows the block diagram of the proposed MAC unit that consists of input master-stage and slave-stage flip-flops, a Dynamic Range Determination (DRD) block, a multiplier, a multiplication output latch, an adder and a sign extension block. Input data are first latched in the master-stage flip-flops that provide input data for the DRD block to determine its operation mode. In the multiplication process, one of two input data with the smaller effective dynamic range is used for the Booth encoding to increase the chance of neighboring partial products being 0 that can decrease the switching activities of the multiplier. In the addition operation, one of two input data with the larger effective dynamic range is used as a current operation word length. The bits in the non-effective range are remained in their previous states to reduce switching activities of an adder. After summation, the output bits in the non-effective range must be corrected by extending the sign bit of the current operation word length. In the multiplication accumulation operation, it integrates the above multiplication and addition operations to further reduce the switching activities of the MAC unit. The proposed schemes on the multiplication and addition operations can be applied to the conventional multipliers, adders and MAC units for saving power consumption. Hence, the proposed MAC units based on the conventional architectures of the Kwon's MAC unit, Farag's MAC unit and Yu's multiplier are explored. With computing the practical input data, the proposed MAC units have lower power dissipation than the conventional ones, correspondingly, at little increasing hardware areas. However, when considering the product factor of power consumption, hardware area and critical delay, the MAC units proposed herein still show superior performances than the conventional ones.

2. The Schemes to Reduce Switching Activities

In the addition process, input data represented by the 2's complement are computed as illustrated in Fig. 3 where *X(n-1)* and *Y(n-1)* are the previous input data for addition to generate an output data of *Z(n-1)*, and *X(n)* and *Y(n)* are the current input data to derive *Z(n)*. We can see that when *X(n)* and *Y(n)* are added together, in Fig. 3(a), addition operations for the most significant 3 bits are unnecessary. To minimize the switching activity in the circuit operation, only 8 bits need be processed and the most significant 3 bits preserve the previous states. Figure 3(b) shows a diagram on how to save the switching activity. Before two input data are added together, the DRD block first detects the effective dynamic ranges of these two input data to select the larger effective dynamic range as the current operation word length. As shown in Fig. 3(b), the current sign bit is the fourth bit counting from the left. The most significant 3 bits at their previous states perform addition whereas the other 8 bits at the current states perform addition. After the addition is computed, the sign value in the fourth bit of the output result is copied to the most significant 3 bits via sign extension. By doing so, the adder can minimize the switching activities of the non-effective bits for addition operations. Figure 3 also displays a comparison of two addition operations on the numbers of input and output data switching. The comparison result clearly illustrates that the proposed addition scheme can have a lower switching activity than the conventional one.

In the multiplication process, multiplication operations of $X \times Y$ and $Y \times X$ have the same results but differ in the partial products generated. In Fig. 4(a), only the third and fifth partial products are equal to zero. However in Fig. 4(b), all partial products after the second one are zero. While both multiplication operations produce the same result, they are quite different in the partial products for accumulation. When the multiplier number is smaller than the multiplicand number, partial products generated from the most significant bits would be zero continuously; unlike the zero partial products appearing randomly as shown in Fig. 4(a). With the DRD block in the front of the multiplier, the input datum with the small effective dynamic range is selected as the multiplier number so that the chance of partial products being zeros becomes high, thus to reduce the number of switching activities. This scheme is particularly effective when used on a Booth-algorithm multiplier. Taking radix-4 Booth algorithm as an example, if the three bits for Booth encoding are either 1 or 0, the result would make the partial product become zeros. For using a Booth-algorithm multiplier, the advantage is that even if the signs are different between the previous computation and the current one, the partial products from the non-effective dynamic ranges are all zero. This condition allows the chance of partial products being zeros higher in a Booth-algorithm multiplier than a conventional 2's complement multiplier without using the Booth algorithm.

3. Proposed Low-Power MAC Unit

This work takes the multiplication and addition schemes as discussed above and integrates them into a multiplication accumulation computation unit to achieve low power dissipation, as shown in the Fig. 2. The following introduces the detailed design of each functional block in the proposed MAC unit.

3.1 DRD block:

The dynamic-range determination block detects the effective dynamic ranges of input data, from which determines whether the two input data paths should be exchanged or/and whether bits in the slave-stage flip-flop and the multiplication output latch keep the previous states. These acts would allow input data or intermediate data to remain unchanged, thus no switching activity would take place and power is saved. The DRD block as shown in Fig. 5 includes 3 partial blocks of a DRD sub-block for multiplication, a DRD sub-block for addition, and a decoder. The DRD sub-block for multiplication is used to determine one of two input data with the smaller effective dynamic range, and then to switch or pass the input data flow. In addition, the effective dynamic ranges of two input data generated from the DRD sub-block for multiplication are fed into the decoder to estimate the effective dynamic range of the product result. The DRD sub-block for addition is used to find the effective dynamic range of an input datum from the accumulation latch or other and then compare it with that from the decoder to obtain the larger effective dynamic range. After that, the non-effective bits of the multiplication output latch and a slave-stage flip-flop for addition can be determined and keep in the previous states. The control signals are then sent out to control the operations of the latch, flip-flop and sign extension block.

The DRD sub-block for multiplication is shown in Fig. 5. This DRD sub-block is based on detection performed on 3 bits per group, since radix-4 is used for Booth encoding. Data detection starts from the most significant bits with the comparators examining each 3-bit group. In the diagram, a bit is overlapped between each comparator, which is considered to ensure continuous comparison between the two neighboring groups. Thus effective dynamic ranges of input data are detected based on the resolution of 2 bits. When all three bits are either all zero or 1, the control signals would be 1, or otherwise 0. The control signals would pass through logic gates to obtain the effective dynamic ranges of two input data. These two effective dynamic ranges are compared to see which one is smaller. After controlling the input data paths, the input datum with a smaller effective dynamic range is used for the Booth encoding. In Fig. 5, the decoder is fed with 6 signals of m_x_1, m_x_2, m_x_3, m_y_1, m_y_2 and m_y_3. These six signals can represent the effective dynamic ranges of two input data. The decoder converts these 6 signals to generate the estimated effective dynamic range of the product result. The DRD sub-block for addition determines the effective dynamic range of an input datum based on the resolution of 4 bits where the 5-bit comparators are utilized. The effective dynamic range of the other input datum is generated from the decoder. Which one of two input data has the larger effective dynamic range is then determined in this DRD sub-block. Additionally, the control signals are produced to control the operations of a slave-stage flip-flop, a multiplication output latch and a sign extension block for low-power and correct addition.

3.2 Multiplier and Adder:

The proposed scheme is applied on the Kwon's MAC unit [2], Farag's MAC unit [3], and Yu's multiplier [4] with an adder for power saving. Especially, the Yu's multiplier is based on the partial products generated by the Booth decoders during multiplication. Partial products from the most significant bits are accumulated first and then the partial products from the least significant bits are added. This action would not affect the speed of the multiplier. Instead, when two input data are multiplied, the chance of partial products from the most significant bits being zero becomes high, which reduces the number of switching activities and that in turn lower the power dissipation. The adoption of Yu's multiplier on the MAC unit design makes it called the modified Yu's MAC.

3.3 Sign extension block

With the inclusion of the DRD sub-block for addition, some non-effective bits in the input buffers for an adder are kept in previous states, so to minimize the switching activities and reduce power dissipation. However with these bits for addition in the previous states, their added results must be corrected. The sign extension block, realized by multiplexors, has the responsibility of restoring the correct result from the output of the adder. Figure 6 shows the sign extension block that is composed of multiplexors based on the control signals of ctl1, ctl2, …, and ctl6 generated from the DRD block where these multiplexors determine if their output values are taken from the sign or value of the adder's output result.

4. Performance Analysis

In our experiment, the TSMC 0.35um cell-based library was utilized to built the Farag's, Kwon's, and modified Yu's MAC units. Additionally, the proposed MAC units based on the Farag's, Kwon's, and modified Yu's architectures are implemented. The Silicon Ensemble provided by the Cadence was used to do the placement and routing of the proposed and conventional MAC units. Lastly, Power-Mill and Time-Mill tools from the Synopsys helped to accomplish the simulation and estimation for power and speed. Here, the clock frequency of the MAC units is set at 50MHz. Input data are taken from the practical signals going through the G.723.1 speech coder, ADPCM audio coder, and wavelet transform in the JPEG 2000. A 1-minute speech is performed on the G.723.1, from which data from the first frame, 30ms, is taken. Input data for addition, multiplication and multiplication accumulation are extracted for the MAC unit. There are 9,605 records of input data of which the distribution of effective dynamic ranges is shown in Fig. 7(a). Here, the multiplication accumulation operation is represented by equation of $X \times Y+C$. From the ADPCM audio coder, the audio signals are sampled from a piano performance at a rate of 16Kbits/sec. A 1-minute audio signal is performed, from which data from the first 0.01 second is taken. 12,650 records of input data for addition, multiplication and multiplication accumulation are extracted for the MAC unit where the distribution of their effective dynamic ranges is shown in Fig. 7(b). From the wavelet transform, a 256x256-pixel Lena image is selected for analysis. Only the first 12,500 records of input data for addition, multiplication and multiplication accumulation are selected for the MAC unit. The distribution of effective dynamic range of these input data is shown in Fig. 7(c).

Table 1 lists the power dissipation, hardware areas and critical delays of the proposed and conventional MAC units for performing the applications of G.723.1, ADPCM, and wavelet transform. The Kwon's MAC architecture yields the least critical delay but has the largest hardware area. In the conventional MAC units, the Kwon's MAC unit can have the least power consumption at these three applications. The proposed MAC unit using the modified Yu's architecture can reduce more power dissipation than the proposed MAC unit using the Kwon's architecture. This effect illustrates that the scheme of Booth encoding the one of two input data with the smaller effective dynamic range is more suitable to the Yu's architecture. Hence, the proposed MAC unit using the modified Yu's architecture can have the least power consumption at the applications of ADPCM and wavelet transform, and have the second least at the application of G.723.1. When considering the product factor of power consumption, hardware area and critical delay, the proposed MAC unit using the modified Yu's architecture can have the best performance at the three applications. Figure 8 shows the saving ratios of power consumed by the proposed versus the conventional MAC units based on the same architecture platform. Our schemes for reducing switching activities of multiplication and addition operations are the best suitable to the modified Yu's architecture, the second to the Kwon's architecture, the last to the Farag's architecture. The proposed MAC unit using the modified Yu's architecture can save power dissipation up to 35.3%, 21.6%, and 36.3% in the applications of G.723.1, ADPCM and wavelet transform, respectively.

5. Conclusion

In this work, we propose two schemes to reduce switching activities of multiplication and addition operations, and integrate them to design a low-power MAC unit. Based on the TSMC 0.35um cell-based library, the proposed and conventional MAC units using Farag's, Kwon's, and Modified Yu's architectures are implemented. Input data for addition, multiplication and multiplication accumulation are extracted from the operations of the G.723.1 speech coder, ADPCM audio coder, and wavelet transform in JPEG 2000. The proposed MAC unit using the modified Yu's architecture can achieve the least or close to the least power consumption at these three applications. Additionally, it can yield the best performance at the product factor of power consumption, hardware area and critical delay. Therefore, the proposed MAC unit can be widely used for various multimedia applications to achieve power-efficiency computing.

References

[1] A. Oppenheim and R. Schafer, *Discrete-Time Signal Processing*, Prentice Hall, New Jersey, 1993.

[2] Ohsang Kwon, K. Nowka and E. E. Swartzlander, "A 16-bitx16-bit MAC design using fast 5:2 compressors," *Proc. of IEEE International Conference on Application-Specific Systems, Architectures, and Processors*, pp. 235 –243, 2000.

[3]E. N. Farag, Ran-Hong Yan and M. I. Elmasry, "Power-efficient multiplier- accumulator design for FIR filters," *Electrical and Computer Engineering. Engineering Innovation: Voyage of Discovery*, pp. 27 -30 vol.1, 1997.

[4] Z. Yu, L. Wasserman and A. Willson, Jr., "A painless way to reduce power dissipation by over 18% in Booth-encoded carry-save array multipliers for DSP," *Proc. of IEEE Workshop on Signal Processing Systems*, pp. 571-580, 2000.

[5] A. P. Chandrakasan and R. W. Brodersen, *Low-power CMOS Design*, New York: IEEE Press (ed.), 1998.

[6] R. Sheen, S. Wang, O. T.-C. Chen and R.-L. Ma, "Power consumption of a 2's complement adder minimized by effective dynamic data ranges," *Proc. of IEEE International Symposium on Circuits and Systems*, Orlando, Florida, USA, vol. I, pp. 266-269, May 1999.

[7] Y. Wu, O. T.-C. Chen and R. Ma "A low-power digital signal processor core by minimizing inter-data switching activities," *Proc. of IEEE 44th Midwest Symposium on Circuits and Systems*, Dayton, Ohio, USA, vol. 1, pp.172-175, Aug. 2001.

[8] N. Shen and Oscal T.-C. Chen, "Low-power multipliers by minimizing switching activities of partial products," *Proc. of IEEE International Symposium on Circuits and Systems*,

Scottsdale, Arizona, USA, vol. IV, pp. 93-96, May 2002.

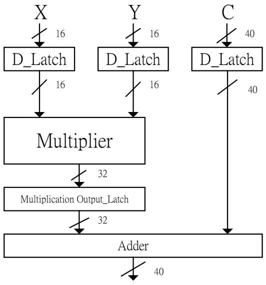


Fig. 1 A conventional MAC unit.

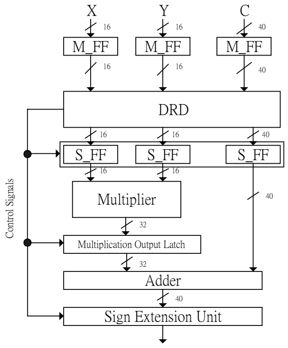
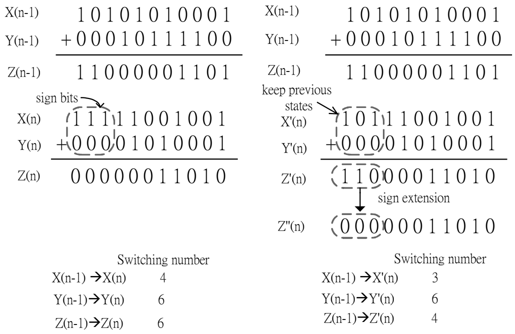


Fig. 2 The proposed MAC unit.



(a) (b)

Fig. 3 The conventional and Proposed addition operations. (a) The conventional one. (b) The proposed one.

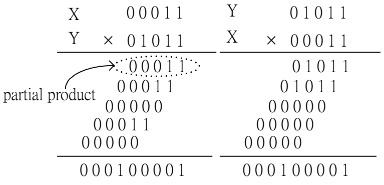


Fig. 4 The two multiplication operations. (a) X multiplied by Y. (b) Y multiplied by X.

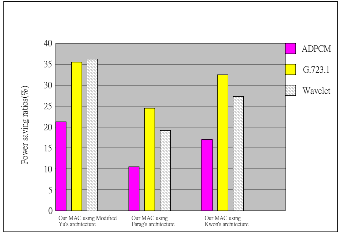


Fig. 8 The saving ratios of power consumed by the proposed versus the conventional MAC units.

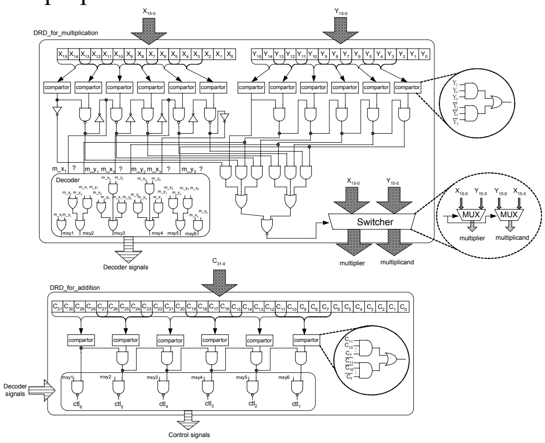


Fig. 5 The DRD block.

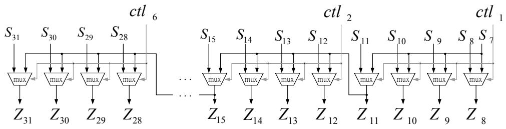
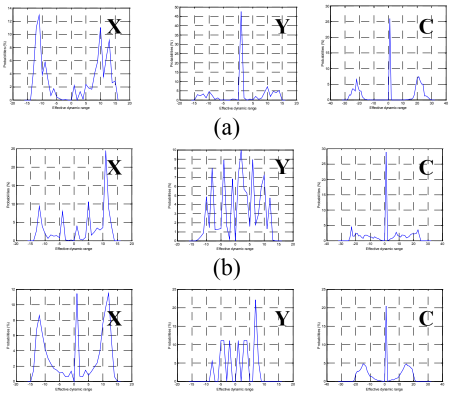


Fig. 6 The sign extension block.



(a)

(b)

(c)

Fig. 7 The distributions of input data at the three applications. (a) G.723.1 (b) ADPCM (c) Wavelet transform

Table 1 Power dissipation, hardware areas and critical delays of the proposed and conventional MAC units.

Features / MACs	Area	Delay	ADPCM		G.723.1		Wavelet	
	μm^2 (C)	Critical delay, (D) (ns)	Power (mw) Total (P)	C×D×P $(\times 10^7)$	Power (mw) Total (P)	C×D×P $(\times 10^7)$	Power (mw) Total (P)	C×D×P $(\times 10^7)$
Modified Yu's MAC [4]	447,561	14.16	31.95	20.25	32.85	20.82	27.31	17.31
Proposed MAC using modified Yu's approach	477,481	14.16	25.05	16.94	21.26	14.37	17.40	11.76
Farag's MAC [3]	443,556	15.28	34.46	23.36	34.50	23.38	28.72	19.47
Proposed MAC using Farag's approach	473,344	15.28	30.40	21.99	24.37	17.63	23.08	16.69
Kwon's MAC [2]	481,636	13.31	31.80	20.39	31.87	20.43	26.51	16.99
Proposed MAC using Kwon's approach	512,656	13.31	26.28	17.93	21.22	14.48	19.13	13.05