Techniques for the Design of High Speed and Low Power MAC Unit: A State-of-the-art Review

Anu
M. Tech. Scholars- ECE, DCRUST, Murthal

Prachi Chaudhary
Assistant Professor- ECE, DCRUST, Murthal

Pawan Kumar Dahiya
Assistant Professor- ECE, DCRUST, Murthal

ABSTRACT
The multiplication operation is used in many parts of a digital system or digital computer, usually in signal processing, video/graphics and scientific computation. With advances in technology, various techniques have been developed to design multipliers, which offer high speed, low power consumption and lesser area. Thus making them suitable for various high speeds, low power compact VLSI implementations. These three parameters i.e. power, area and speed are always traded off. In this paper, different techniques used for efficient operations resulting in high speed and low power consumption are discussed. Such as parallelism, pipelining, modified booth algorithm (MBA), spurious power suppression technique (SPST), block enabling technique.

Keywords
Multiply and Accumulate (MAC), Modified Booth Algorithm (MBA), parallel modified booth multiplier, Spurious Power Suppression Technique (SPST), block enabling technique.

1. INTRODUCTION
The core of every microprocessor, Digital Signal Processor (DSP), and data processing Application Specific Integrated Circuit (ASIC) is its data path. The statistics shows that more than 70% of the instructions perform additions and multiplications in the data path of RISC machines [20]. At the heart of data-path and addressing units in turn are arithmetic units, such as adders, and multipliers. Multiplication operations are among some of the frequently used computation-intensive arithmetic functions, currently implemented in many DSP applications such as convolution, fast Fourier transform, FIR filters, discrete wavelet transform and in microprocessors in its arithmetic and logic unit. Since multiplication process has large amount of delay in most of the DSP algorithms, so there is a need of high speed multiplier. The demand for high speed processing has been increasing as a result of increasing computer and signal processing applications [8]. Low power consumption is also an important parameter in multiplier design. To reduce significant power consumption, it is needed to reduce the number of operation thereby reducing dynamic power which is a major part of total power consumption. Therefore, the need for high speed and low power multiplier has increased. The designers, mainly, concentrate on high speed and low power efficient circuit design. The objective of a good multiplier is to provide a physically packed together, high speed and low power consumption unit. In this paper we will discuss different techniques used in MAC for efficient operations resulting in high speed and low power consumption [7].

The paper is organized as follows:

Section 2 presents the basic operation and design of MAC unit. Section 3 describes different techniques used for improving the performance of MAC unit. Section 4 discusses the performance of various techniques mentioned in previous section. Section 5 concludes the paper and in the last section references are listed.

2. OVERVIEW OF MAC UNIT
In digital signal processing, the basic operation is multiplication and accumulation. The MAC unit provides the operations such as high speed multiplication with accumulation. Hence, if a MAC is working under high speed operation, it must support multiple operations and parallel MAC comprises of three important sections:

1. Adder
2. Multiplier
3. Accumulator

![Fig.1: Basic MAC unit [1].](image)

If an operation to multiply two N-bit numbers and accumulates into a 2N-bit number is considered, 2N-bit accumulation operation determines the critical path [1].

3. TECHNIQUES FOR MAC DESIGN: A STATE-OF-THE-ART
Some of the techniques used for designing high speed and low power MAC unit are as follows:

2.1 Parallel Multiplier Using MBA
Modified Booth Algorithm (MBA) is the commonly used method for achieving high speed by reducing the number of partial products. By using radix-4, radix-8, radix-16 and radix-32 booth encoding schemes, the partial products are further reduced, improving the performance but complexity is increased [1]. To enhance the speed and performance, many parallel MAC architectures have been designed. There
are two ways that make use of parallelism for improving the performance of MAC unit. The first one is to reduce the number of partial product rows and second one is to use the carry-save-tree technique to reduce multiple partial product rows into two “carry-save” redundant forms [2]. In parallel MAC implementation, accumulator stage that provides the largest critical path delay in MAC is combined with multiplication stage to enhance speed and decrease the hardware architecture. By combining multiplication with accumulation in a hybrid type of carry save adder (CSA) tree, the performance was improved. Since the accumulator that has the largest delay in MAC was merged into CSA, the overall performance was elevated. To further improve the performance of final adder, the no. of input bits should be decreased. In order to reduce the no. of input bits, the multiple partial products are reduced into a sum and carry by CSA tree. The number of bits of sums and carries to be transferred to the final adder is reduced by adding lower bits of sums and carries in advance. A 2-bit CLA is used to add the lower bits in CSA.

Fig.2: Hardware architecture of parallel MAC [14].

3.2 Pipelined Booth Multiplier:

Pipelining is a popular technique to increase throughput rate of a high speed system which divides the system into several small cascade stages and add some registers to synchronize the output of each stage. As the no. of stages increase, the power consumption and area gets increased. So, most of the times pipelining technique is introduced in Wallace tree to improve the performance. Also, when arithmetic throughput is more important than latency, pipelined multipliers are useful because the introduction of registers along the array reduces the unnecessary activity [6].

Fig.3: Block diagram of pipelined booth multiplier [7].

Modified booth encoder is used to recode the multiplicand bit in order to reduce the number of partial product. This encoder codes the three bits into single bit. It takes earlier, present and next bit into account to convert that bit into the single bit. The Wallace tree construction method is usually used to add the partial products in order to get two rows of partial products that can be added in the last stage. Wallace tree has high speed because the critical path delay is proportional to the logarithm of the number of bits in the multiplier. The prominent method considers all the bits in each column simultaneously and compresses them into two bits (a sum and a carry). To compress them into two bits many type of compressor are used such as 4:2 compressor, 3:2 compressor, 5:3 compressor. Pipelining block consists of registers. Registers consists of latches (flip-flops). Mostly, D-flip flop is used as the register. Parallel pipeline architecture is considered as more suitable for low voltage and low power systems. In a pipelining system, the maximum operating frequency is limited by the slowest stage. Final stage is also important for a multiplier because in this stage addition of large size operands is performed so in this stage fast adders like Carry-look Ahead Adder or Carry Skip Adder or Carry Select Adder and other adders such as Carry Save Adder, Kogge stone adder can be used as per requirement [7].

3.3 Spurious Power Suppression Technique:

Using SPST we can reduce power consumption in addition process. In booth multiplication, when two numbers are multiplied some portion of data may be zero in partial products, so this data can be neglected. In other words saving those computations can significantly reduce the power consumption by transient signals. The SPST technique is basically dependent on the radix-4 modified booth algorithm. It helps in the recoding of the given multiplicand and reduces the number of the intermediate stages in the multiplication operation which maintains the speed of the process at the same time the power consumed will also get reduced. The SPST uses a logic circuit (detection unit) to detect the effective data range of arithmetic units, e.g., adders or multipliers. When a portion of data does not affect the final computing results, the data controlling circuits of the SPST latch this portion of data in order to avoid useless data transitions occurring inside the arithmetic units [1].

Fig.4: Spurious power suppression technique [1].

3.4 Block Enabling Technique:

The basic building blocks for the MAC unit are multiplier, adder, and register. In block enabling technique, delay of each stage is measured and every block is enabled only after
the expected delay. So, in this technique when inputs are not enabled, the successive blocks are disabled thus saving power. Each of the block in the MAC unit has an enable signal to save power. The basic gate that is required to enable or disable the MAC is controlled using an AND gate. It is examined that the delay reduces with increase in width. As the NAND gate has delay, the blocks connected to the output of the AND gate are disabled until this time and these blocks are enabled only after the outputs are available, hence saves the power. To completely understand this technique we can take 1 bit MAC example

![Control Logic Diagram](image)

**Fig.5: Control logic for block enabling technique [13].**

Design a 1 bit MAC unit with clock gating and enable pin. Initially when the input is applied, all the blocks are enabled simultaneously, the FA block would compute the result on unknown data until AND gate delay, and the register block would be receiving unknown data for register gate delay – AND gate delay and hence there is wastage of power as these data’s are not actual ones. So, in this a control signal is incorporated that enable the blocks only after the outputs are available at their inputs. Hence we call this technique as block enable technique. Based on the delay of each block, a control signal is generated to enable the blocks [7].

4. **DISCUSSION**

The performances of the various techniques used in the previous section are tabulated in terms of the affecting parameters (such as delay, power) as follows:

<table>
<thead>
<tr>
<th>Technique</th>
<th>Delay (ns)</th>
<th>Power consumption (mW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parallel multiplier using MBA [2], [11]</td>
<td>9.48</td>
<td>0.0412</td>
</tr>
<tr>
<td>Pipelined booth multiplier [5]</td>
<td>5.4</td>
<td>1.4143</td>
</tr>
<tr>
<td>Spurious power suppression technique [13]</td>
<td>5</td>
<td>0.0121</td>
</tr>
<tr>
<td>Block enabling technique [9]</td>
<td>1.086</td>
<td>0.16386</td>
</tr>
</tbody>
</table>

5. **CONCLUSION AND FUTURE SCOPE**

As evident from the literature, the Booth multiplier has the highest operational speed and less hardware count as compared to other circuits. This algorithm is competitive with other more commonly used algorithms when used for high performance implementations. Considering different technique or design of MAC unit, parallel and pipelined booth multiplications give good performance in terms of speed and SPST and block enabling technique are better in low power consumption and area. Using higher radix MBA and partial product reduction technique by the hybrid carry save adder tree can give good results in terms of speed.

6. **REFERENCES**


