A 650 MHz Pipelined MAC for DSP Applications using a New Clocking Strategy

F. Fraternali, G. Masera, G. Piccinini, M. Zamboni
Politecnico di Torino - Dipartimento di Elettronica
Corso Duca degli Abruzzi 24 - I10129 TORINO - Italy

ABSTRACT

A 8x8 bit multiplier and accumulator unit for high speed applications is presented in this paper. The multiplier architecture is directly derived from the Baugh and Wooley algorithm, with some modifications, to reduce area and latency while the accumulator section is distributed along the multiplier structure. In this way the accumulator's latency is hidden in the multiplier's one. A new clocking strategy has been used for the design of the four stages pipelined accumulator cell, based on a full adder with partial feedback. The unit is synthesized in a 0.7μm N well CMOS technology. A one phase dynamic logic (True Single Phase Clocking - TSPC) has been adopted and the transistors widths had been sized by using an optimization algorithm achieving a clock frequency of 650 MHz with a latency of 36 clock cycles.

1 Introduction

Multiply and accumulate operations (MAC) play a central role in typical algorithms in digital signal processing applications. The high throughput required by real time processing systems demands for the development of very high speed multiplier and accumulator units. Figure 1 reports the block diagram of a 8x8 MAC unit where the accumulator is extended to 20 bits.

DSP algorithms usually require a continuous flow of MACs, without heavy test and branch conditions (small data dependency). For this reason a even large latency is acceptable as counterpart of high operation rate.

Latency is related to the number of pipeline stages introduced in the logic design to reduce combinatorial delays and consequently the period of the clock.

Different architectures for high speed multipliers are available in the literature; among them the parallel solutions [1] [2] are proved to be more suitable for high speed pipelined implementations.

The multiplier architecture designed is a CSA structure derived from the Baugh and Wooley algorithm [3]. This solution allows an easy management of two's complement number multiplications, without heavy architectural modifications.

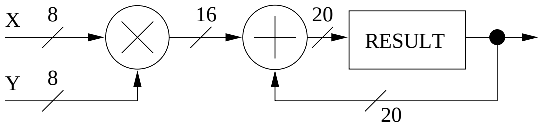

Figure 1: MAC block diagram

An example of a 4x4 bits multiplier CSA structure is shown in figure 2.

The great advantage of the Baugh and Wooley algorithm is that all partial products appear in the CSA with positive sign. In this way it is possible to perform the multiplication just with full adder blocks, for the partial sum calculation, and AND gates, for the partial products generation.

Moreover this algorithm keeps local each interconnection between internal blocks, so that the parasitics are as low as possible.

Since the CSA is pipelined along the data flowing direction, preskewing and deskewing registers are necessary to maintain the correct synchronization and to allow data propagation. These registers, as the accumulator extension bits are not indicated in figure 2 for sake of simplicity.

As previously mentioned, the accumulator section is distributed along the CSA array: each bit is calculated on the appropriate column by a full adder properly feedback.

The accumulator reset is a critical task in pipelined structures, since it is important not to flush out the pipelining queue but it should be possible to reset one bit per cycle (from the LSB to the MSB) without affecting the most significant partial results and allowing new MAC operations on the previously reset bits.

The high degree of pipelining distributed in the MAC requires an accurate choice of the logic and electric design style in order to keep complexity, transistor count, delays and power consumption at low levels.

Moreover, today's sub-micron CMOS technology allows to achieve very high speed and low power. Obviously, this kind of speed performance can be reached only with fully dynamic CMOS gate design.

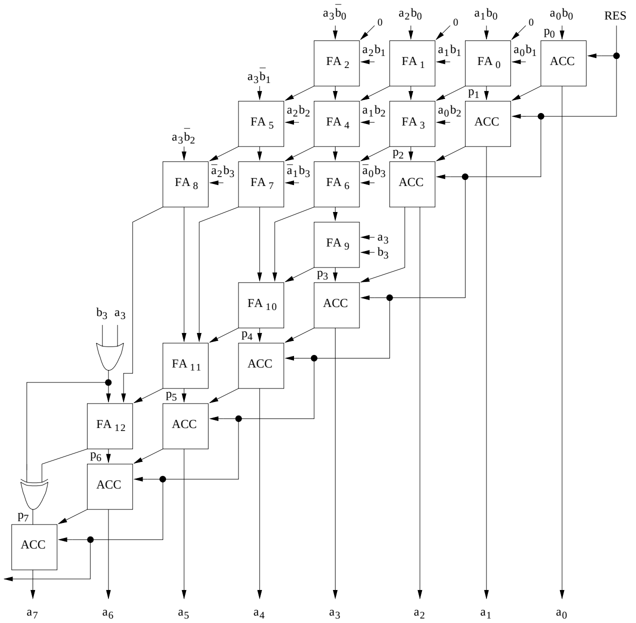

Figure 2: 4x4 bits Baugh and Wooley multiplier and accumulator

Furthermore, a sub-micron CMOS technology allows to maintain a lower power consumption than other high speed technologies, such as BiCMOS and Gallium Arsenide.

The choice of the dynamic logic has to be carried out by considering several problems. First of all the high speed performance are strictly related to the simplicity of the clock handling and distribution; this simplifies the skew handling if a proper topology is selected to eliminate races problems. As a consequence it is mandatory to choose single phase clocking strategies.

These reasons brought us to consider TSPC (*True Single Phase Clocking*) logic [4] as the best candidate to solve the above mentioned problems.

Moreover it fits well the layout topology defined by the multiplier considering the signals and phase propagation inside the MAC unit.

2 MAC basic cells

The architecture of the multiplier is derived from the Baugh and Wooley algorithm. The algorithm applies to two's complement operands

$$\begin{aligned} A &= -a_{m-1}2^{m-1} + \sum_{i=0}^{m-2} a_i 2^i \\ B &= -b_{n-1}2^{n-1} + \sum_{i=0}^{n-2} b_i 2^i \end{aligned}$$

from which the conventional product form derives:

$$P = a_{m-1}b_{n-1}2^{m+n-2} + \sum_{i=0}^{m-2}\sum_{j=0}^{n-2} a_i b_j 2^{i+j} +$$

$$-\sum_{i=0}^{n-2} a_{m-1} b_i 2^{m-1+i} - \sum_{i=0}^{m-2} b_{n-1} a_i 2^{n-1+i}$$

Using the algorithm, the product can be expressed in the following form:

$$P = a_{m-1}b_{n-1}2^{m+n-2} + \sum_{i=0}^{m-2}\sum_{j=0}^{n-2} a_i b_j 2^{i+j} +$$

$$+2^{m-1}(-2^n + 2^{n-1} + \overline{a}_{m-1}2^{n-1} + a_{m-1} + \sum_{i=0}^{n-2} a_{m-1}\overline{b}_i 2^i) +$$

$$+2^{n-1}(-2^m + 2^{m-1} + \overline{b}_{n-1}2^{m-1} + b_{n-1} + \sum_{i=0}^{m-2} b_{n-1}\overline{a}_i 2^i)$$

Such expression contains only positive partial products that can be rearranged (in the example of m = n = 4) as in the following

2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0
				$a_3\overline{b}_0$	a_2b_0	a_1b_0	a_0b_0
			$a_3\overline{b}_1$	a_2b_1	a_1b_1	a_0b_1	
		$a_3\overline{b}_2$	a_2b_2	a_1b_2	a_0b_2		
	a_3b_3	$\overline{a}_2b_3$	$\overline{a}_1b_3$	$\overline{a}_0b_3$			
	$\overline{a}_3$			a_3			
1	$\overline{b}_3$			b_3			
p_7	p_6	p_5	p_4	p_3	p_2	p_1	p_0

From this scheme, it is possible to derive the architecture shown in figure 2 introducing the correct number of pipelining stages on the columns.

The high degree of regularity of the architecture, centered on the full adder cell that constitutes the basic block of the Carry Save Adder (CSA), reduces the number of different cells making the optimization of the design more effective.

2.1 Full Adder

In a pipelined VLSI architecture, the maximum clock rate depends on the worst delay among latched nodes; this means that higher clock frequencies can be achieved reducing the most critical propagation delays. This can be obtained limiting the complexity of each logic level by means of logic function decomposition keeping in mind the target logic family.

In the case of a full adder with inputs A,B and C the output function (SUM and CARRY) can be rewritten as follows:

$$\begin{aligned} \mathrm{M} &= \mathrm{A} \oplus \mathrm{B} = \overline{\mathrm{A}}\mathrm{B} + \mathrm{A}\overline{\mathrm{B}} \\ \mathrm{SUM} &= \mathrm{M} \oplus \mathrm{C} = \overline{\mathrm{M}}\mathrm{C} + \mathrm{M}\overline{\mathrm{C}} \\ \mathrm{CARRY} &= \overline{\mathrm{M}}\mathrm{B} + \overline{\mathrm{M}}\mathrm{C} \end{aligned}$$

This decomposition points out how the use of intermediate memory elements shorts the maximum combinatorial delays in spite of an increased latency.

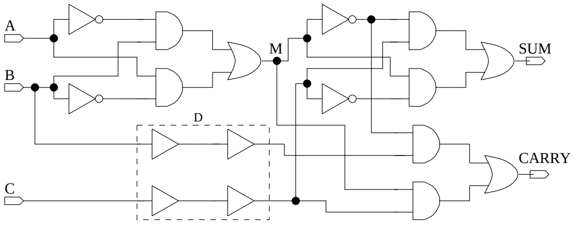

Figure 3: Four stage full adder cell

As domino and NORA logic, TSPC can only map non-inverting boolean functions. Therefore static inverters are required to complete the synthesis, so the SUM and CARRY outputs can be implemented in TSPC using AND, OR and static NOT gates.

The function AND (OR) is implemented with a precharged TSPC PC2 (NC2) cell. The output stage of each cell guarantees the self-latching function on the appropriate phase.

The mapping of SUM and CARRY by defining the intermediate variable M would provide two levels of pipe; but looking at the logic equations of M and SUM, it comes out that each variable is produced by a sum of minterms. As a consequence a pipe level can be inserted before the OR operator. This way the total number of pipe stages is four and, more important, each basic cell is a only two input function. By mapping AND gates with PC2 and OR gates with NC2, the maximum number of transistors connected in series is always two. This is the minimum transistor chain length available in TSPC.

The whole full adder is reported in figure 3, according to the four pipelining stages. It is worth noting that in TSPC two pipe stages are executed in one clock cycle, so the full adder introduces a latency of two cycles.

In order to provide proper data to the third pipe stage, two additional latches are inserted on the paths starting from B and C inputs as indicated in the D block.

2.2 The Accumulator cell

The multiplier produces a new output result on every clock cycle; this product obtained must be accumulated in the accumulator unit, i.e. it must be added to the previous clock cycle multiplication result.

The sum is carried, bit per bit, on each MAC column (P7..P0). Functionally the accumulator can be seen as a series of full adders acting in successive steps from LSB to MSB. Each bit is evaluated following a policy very similar to the additions performed on the partial products in the multiplier. Consequently the basic structure of the accumulator is again a full adder where one of the inputs is logically the output of the same block. This architectural choice does not increase significantly the latency of the MAC unit and does not limit the performance of the multiplier.

The accumulator cell is designed as reported in figure 4, where a feedback has been introduced between the output SUM and the third stage. This solution maintains the correct flow in the pipe, since input data are sampled every clock cycle and take four phases to go out of the full adder. Therefore the feedback must be closed to the third phase stage to assure synchronism.

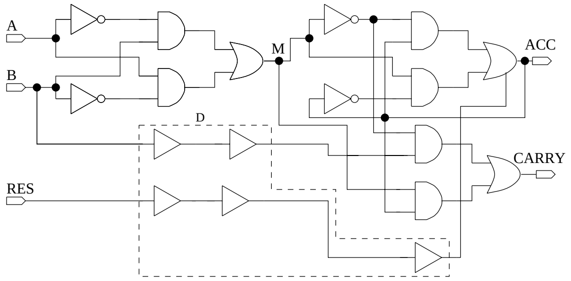

Figure 4: Four stage accumulator cell

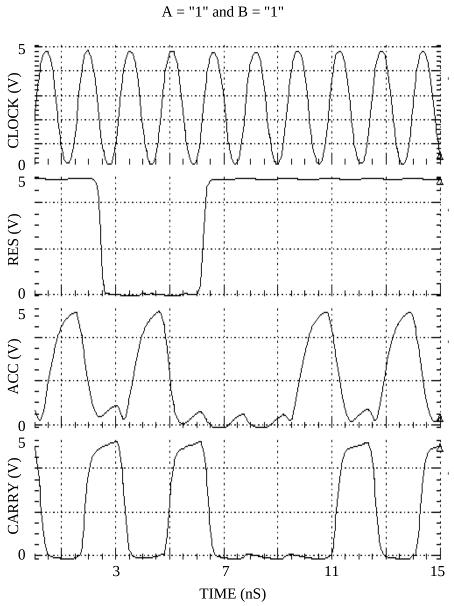

Figure 5: Accumulation cycle

The spice simulation results of the accumulator cell are indicated in figure 5.

3 Device Sizing

In a CMOS circuit, delays depend heavily on transistor widths. Thus, to reach high speed performance, it is necessary to optimize transistors widths with an appropriate sizing procedure.

An optimization tool derived from the SLOP [6] algorithm has been designed and applied on the description of the critical paths delay.

The optimization is based on an accurate delay evaluation, by using a switch level MOS model derived from the transistor dynamic resistance concept in the saturation region.

Great care must be taken with sub-micron technology since dynamic resistance calculation must keep into ac-

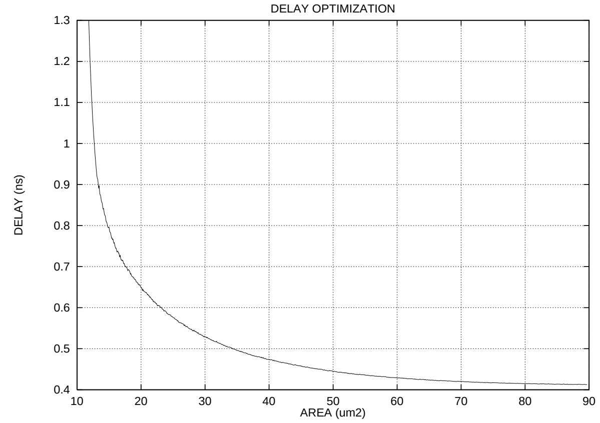

Figure 6: Accumulator cell delay (ns) vs. area (μm^2)

count the following second order effects: channel drift velocity saturation, channel length modulation and bulk effect.

Since the accumulator is the most critical element of the architecture, we consider the optimization problems applied to the basic accumulator cell.

At first, critical paths are extracted and then optimized varying the transistor widths by means of a multi-variable optimization algorithm. The algorithm converges to a global minimum, granted by the convexity of the delay functions. In this way, the designer can reach the best compromise between delay and area.

The worst delay versus the area of the accumulator cell is shown in figure 6. It can be pointed out that it is not worth to increase the transistors widths beyond the first region where significant delay reductions are gained with a reasonable increase of the total area.

4 Physical implementation

The topology of the architecture suggests great care in the design of the layout blocks in order to keep the routing overhead as low as possible. This means that only phase clock signals must be routed globally, while all the data signals are local. According to these issues a macrocell composed of three basic blocks has been designed. The blocks are the main elements of the MAC architecture: full adder, AND gate and deskewing/preskewing register.

The MAC layout, except for the first row implemented with only AND gates, is then obtained by abutments of the macro-cells, as indicated in the floorplan of figure 7.

The clock distribution is designed using a tree distribution scheme, where each vertical phase line is loaded by a column of equivalent macro-cells. Therefore all the clock lines are equally loaded and the buffers are optimized in the same way. To provide a better set-up at the input of the internal pipe stages, clock is distributed backward with respect to data flow in the MAC array (see figure 7).

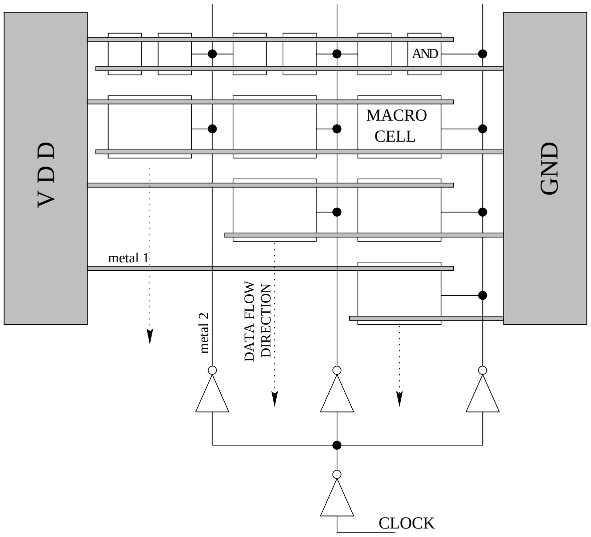

Figure 7: Chip floorplan

The layout has been designed in a full custom way, using CADENCE tools with the ES2 ECPD07 process.

References

[1] F. Lu and H. Samueli, *A 200 Mhz CMOS Pipelined Multiplier-Accumulator Using a Quasi-Domino Dynamic Full Adder Cell Design*, IEEE JSSC, vol. 28, pp. 123-132, Feb 1993.

[2] D. Somasekhar and V. Visvanathan, *A 230 Mhz Half Bit Level Pipelined Multiplier Using True Single Phase Clocking*, IEEE Trans. VLSI Systems, vol. 1, pp. 420-422, Dec 1993.

[3] C.R. Baugh and B.A. Wooley, *A Two's Complement Parallel Array Multiplication Algorithm*, IEEE Trans. Comp., vol. C-22, pp. 1045-1047, Dec 1973.

[4] J. Yuan, I. Carlsson and C. Svensson, *A True single Phase Clock Dynamic CMOS Circuit Technique*, IEEE JSSC, vol. SC-22, pp. 899-901, Oct 1987.

[5] J. Yuan and C. Svensson, *CMOS Circuit Speed Optimization based on Switch Level Simulation*, ISCAS'88, pp. 2109-2112.

[6] J. Yuan and C. Svensson, *A Simulation based Fast Algorithm for CMOS Circuit Speed Optimization*, ISCAS'89, pp. 868-871.

[7] T.G. Noll, D. Schmitt-Landsiedel, H. Klar, G. Enders, *A 300 MHz Pipelined Multiplier*, IEEE JSSC, vol. 21, pp. 411-416, June 1986.