A 230MHz Half Bit Level Pipelined Multiplier using True Single Phase Clocking

Dinesh Somasekhar V. Visvanathan
Department of Electrical Communication Engineering
Indian Institute Science
Bangalore, India 560012.

Abstract

An 8 bit by 8 bit signed two's complement pipelined multiplier in $1.6\mu m$ *N well CMOS, capable of throughputs of* 230 *million multiplications per second, is described. A half bit level pipelined architecture, and the use of true single phase clocked circuitry, are the key features of this design. Simulation studies indicate that the multiplier dissipates* $540mW$ *at* $230MHz$. *The chip complexity is 5176 transistors, and the area is* $1.5mm \times 1.4mm$.

1 Introduction

A high speed multiplier core forms one of the basic building blocks for digital signal processors. Signal processing applications typically implement iterative algorithms having a large number of multiply add/accumulate operations. High throughput rates required in such applications are best satisfied by pipelining standard multiplier architectures. Recent advances in real-time filter architectures indicate that pipelined multipliers, with clock frequency much higher than the data rate, can significantly reduce the overall chip area without increasing the power consumption or the i/o bandwidth. The above reasons motivate us to explore the limits of pipelined multipliers implemented in a standard, mature CMOS process.

This work describes the implementation of a CMOS 8 bit by 8 bit signed two's complement multiplier using a very finely pipelined carry save array architecture in order to achieve a throughput rate of 230 million multiplications per second. Pipelined multiplier architectures described in the literature all show the importance of the clock skew problem at high clock rates. The current architecture uses the true single phase clocking scheme [1] which has the inherent advantage that clock skew problems are restricted to a proper distribution of a single clock phase. This scheme has been demonstrated as being very attractive for the design of very fast adder elements [2]. A full circuit simulation, although desirable, is not practical for large circuits as in the present case. Multiplier characterization has been done by SPICE simulation for smaller blocks and timing simulations for the overall multiplier. This scheme allows the design to be verified in reasonable time, without resorting to overall circuit simulation [3]. Simulation studies indicate a power dissipation of $540mW$ at a clock speed of $230MHz$. Implementation is in $1.6\mu m$ N well CMOS single polysilicon double metal process using the NELSIS IC Design System [4]. The multiplier has 5176 transistors in an area of $1.5mm$ by $1.4mm$.

2 Single Phase Clocked Pipelines

Various clocking schemes exist in CMOS for implementing pipeline stages. Two phase clocking, four phase clock, $NORA$ circuitry and C^2MOS being some. These schemes rely on multiple clock phases for obtaining race free clocking. In contrast *true single phase clocking* [1] uses a single clock phase and achieves race free clocking by using the complementary nature of N and P devices in CMOS to simulate the effect of two clock phases. At very high clock rates the need for maintaining non-overlap of clock phases in multiple phase clocking schemes becomes difficult to achieve. In contrast a true single phase clocking has to ensure proper rise and fall times, and keep skew between communicating blocks within bounds to ensure proper operation. In a single phase clock strategy pipelining is achieved by having alternate P and N blocks. Possible ways of doing this illustrating their salient features are described in [1].

The current work uses the TSPC1/TSPC2 (true single phase clocked type 1 and type 2) circuits for the full and half adder compute elements. This scheme fits in well with a finely pipelined design, and yields very high speed circuits [2]. A dynamic clocking strategy is used for these circuits. This method gives a latency of an n stage pipeline as $n/2$. Latches are integral to the compute block in these circuits. This scheme yields a utilization of 50% for a compute block. However the other 50% time is utilized for precharging, because of the dynamic nature of TSPC1/TSPC2 circuits, and is not wasted. Latch stages required for skewing multiplier bits and for deskewing multiplicand bits, use true single phase latches [1]. Pipelining is carried out within full-adder blocks, so that a full-adder has a P half and an N half. Considering a full-adder row to be two stages, $2n$ stages have a latency of n clock cycles.

3 Multiplier Architecture

Among the various multiplier architectures, the array architecture is the architecture of choice for very high throughput pipelined multipliers, mainly due to

0-8186-3180-5/92 $3.00 © 1992 IEEE

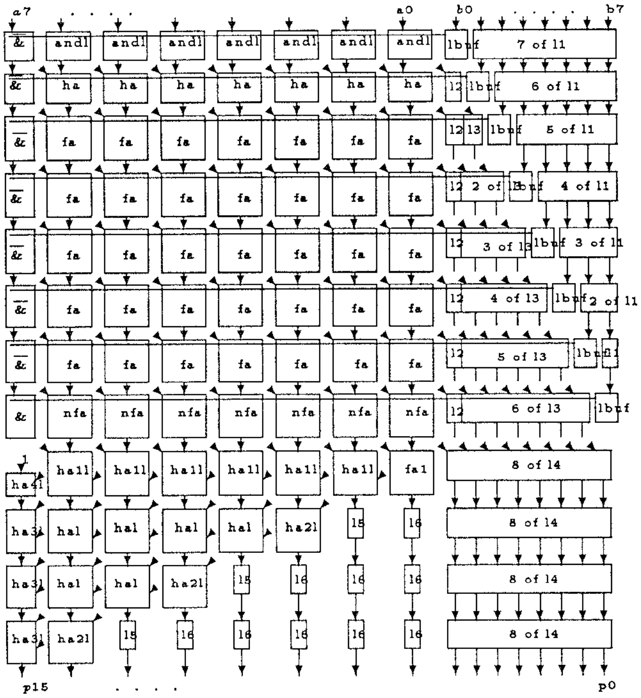

Figure 1: Multiplier Floorplan. Note, clock drivers and output buffers have been omitted.

regularity of its structure and its semi-systolic nature, wherein most of the signals propagate between local blocks. Various different forms of array architectures may be conceived of, depending on the direction of data flow, and the way in which the last row of the multiplier, the vector merge adder is implemented.

The current work uses a carry save architecture implementing the modified Baugh Wooley signed two's complement multiplication algorithm [5, 3], with the LSB partial product being evaluated first. Data flow is solely in the vertical direction. Although this architecture is semi-systolic (the multiplier bits are broadcast over an adder row), in comparison to systolic architectures using both vertical and horizontal pipelining, it is more efficient in silicon area usage [6]. The non-systolic architecture however, adds delay problems due to long data paths for multiplier lines as a design issue.

The modified Baugh Wooley algorithm used is illustrated by considering a $4bit \times 4bit$ multiplication.

				$\overline{a3.b0}$	a2.b0	a1.b0	a0.b0
			$\overline{a3.b1}$	a2.b1	a1.b1	a0.b1	
		$\overline{a3.b2}$	a2.b2	a1.b0	a2.b2		
	a3.b3	$\overline{a2.b3}$	$\overline{a1.b3}$	$\overline{a0.b3}$			
1				1			
p7	p6	p5	p4	p3	p2	p1	p0

The implementation of the above for $8bit \times 8bit$ multiplication is shown in figure 1. This reflects the actual floorplan and consists of six major blocks:

1. The partial product summing full-adder array, which consists of 5 full-adder rows (*fa*), a top row of AND gates (*andl*), a second row of half-adders (*ha*) and a final full-adder row with complemented partial product terms (*nfa*). Pipelining is at the half bit level with every full-adder having two pipeline stages. Partial products are generated within the full-adder cells using AND gates (except for the last row which uses NAND gates) and is carried out in parallel with partial product summation. This scheme is more efficient than one in which the AND array is kept separate [3]. Schemes like modified Booth recoding have not been used, because the fine level of pipelining makes a complex Booth recoder the major bottleneck.

2. The triangular vector merge adder summing up two 7 bit numbers to generate the most significant 8 bits of the product. This includes the blocks *hal*, *ha1l*, *ha2l*, *ha3l* and *ha4l*.

3. Latch stages to skew the multiplier bits (*l2*, *l1*). This includes the buffers needed to drive fairly long horizontal multiplier bit lines (*lbuf*). A latch stage is made by cascading a P latch and an N latch.

4. Deskewing latch stages for the product bits (*l6*, *l5*, *l4*, *l3*).

5. Clock distribution circuitry (see figure 3).

6. Output buffers consisting of inverter chains to buffer product bits (these are not shown in figure 1).

In this design at a given clock tick the current multiplicand, is multiplied with the multiplier clocked in on the previous clock cycle. The actual implementation is very nearly square with dimensions of $1.5mm \times 1.4mm$ and has in all 5176 transistors. About 25% of the area is occupied by the vector merge adder and 10% is occupied by the clock drivers.

3.1 True Single Phase Clocked Full-adder

The schematic of the pipelined single phase clocked adder is shown in figure 2. Each full-adder also includes the partial product generation circuitry. A full adder is partitioned into two pipeline stages a P (left half of the schematic) and an N block (right half of the schematic). This circuit differs from that in [2], in allowing partial product evaluation in parallel to *sum* generation and in minimizing the complexity of the P half.

Evaluation of the *sum* and *carry* is a two step process. In the first step during the clock low period the P half generates: the *xor* of the *si* and *ci* inputs, the product term *aj.bi*, a latched version of the *ci* input, and a latched version of the *aj* input. Either the *xor* or *xnor* term may be implemented with the same transistor complexity. In the current design the choice of generating the *xor* term was governed by layout issues. Generation of the *xor* term is carried out by generating $\overline{si.ci}$ by using a fully complementary CMOS gate with the *xor* term being obtained from *si.ci* , *si* , *ci*.

In the high period the N half now evaluates the *sum* and *carry* terms. $sum = \overline{aj.bi}.xor + aj.bi.\overline{xor}$ and $cy = aj.bi.xor + cil.\overline{xor}$. Inverters are used to

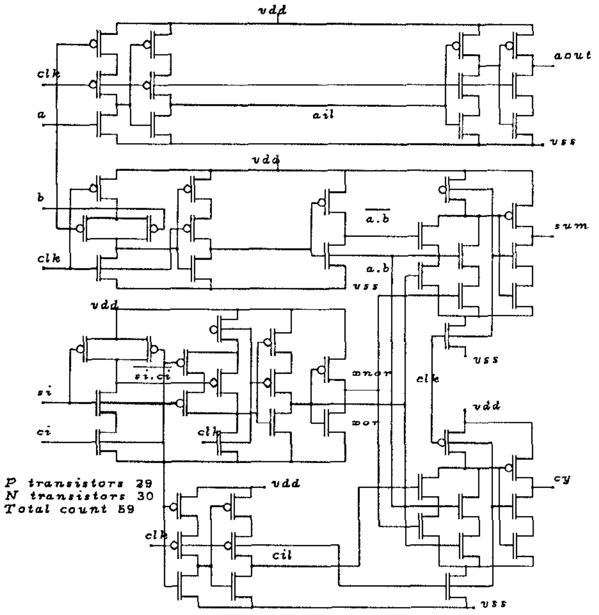

Figure 2: Full-adder Circuit Schematic

generate *xnor* and inverted partial product term from *xor* and *aj.bi* terms. further N latch is used to latch the *ail* line.

The above scheme has the advantage of evaluating the partial product in parallel with the P half evaluation of the *sum* and *cy*. It partitions the addition process into two halves with the computation being so arranged so that more of it is in the faster N half.

The *xnor* and *aj.bi* terms in the P half are generated by using the TSPC1 circuit, while the N half exclusively uses the TSPC2 circuit. Because the P half uses the TSPC1 circuit the outputs generated by it show spikes which are positive going [2]. This is minimized by proper transistor sizing, but are not totally eliminated. The outputs of the N half show far smaller spikes, since TSPC2 circuits are used.

Transistor count for the full adder block is 59. If the transistor count of latches, and the partial product generator is excluded we get a count of 40 for the actual adder block. In comparison, an FCC full adder has a transistor count of 24, while a *NORA* full-adder requires 25 transistors. Area of the full-adder is $142.8\mu m \times 147\mu m$.

Each full adder has been designed for abutment. Data flow is in the vertical direction except for the bi line which runs horizontal. Power and clock lines run horizontally in first metal, without breaks.

3.2 Buffered latch

Since each row of the multiplier evaluates its inputs on the clock low level, the buffered latch must ensure that the multiplier bit is valid during this phase. The buffered latch is built by having a P single phase latch, an inverter, a comparatively larger N latch, and a large sized inverter capable of meeting the timing constraints while driving the multiplier line.

Proper operation of the multiplier requires $t_{clkh} > t_{dN} + t_{dinv}$, where t_{clkh}, t_{dN} and t_{dinv} are the clock high period, delay of the N latch for a high to low output transition and the inverter delay respectively. The design allows driving the $500fF$ multiplier line capacitance, at a clock period of $3.8ns$. Multiplier line capacitance was kept at a minimum, since each full-adder line contributes a load equal to a single P transistor to the multiplier line (see figure 2).

3.3 Vector merge adder

The vector merge adder uses a triangular array of half adders. This structure is similar to that used in the multipliers in [6, 3]. Merging of two half-adder rows is used in the current implementation, allowing an 8 bit sum of two 7 bit numbers to be generated in 4 clock cycles [3]. A cascade of two half adders thus constitutes a basic block. This has one half adder entirely in the P half and the other entirely in the N half. Blocks on the main diagonal have one half adder in the P half and a latch in the N half. The percentage area used by the vector merge adder, along with the deskewing registers required at the output is 25% of the total area for the multiplier, not including the clock drivers and the output buffers.

The vector merge adder also takes care of the 1 addition required in the Baugh Wooley algorithm. For the $8bit \times 8bit$ case a 1 has to be added in the 9th bit position and the 16th bit position. This is done by feeding in a 1 to the top most half adder row at the MSB position. For the 9th bit position by noting that an addition of 1 implies that the *sum* is the $\overline{xor}$ of the other two inputs and the *carry* is the *or* of the inputs it is possible to design a block very close to a half adder. This is the scheme which is employed here. This block is labeled FA1 in the floorplan.

4 Clock system

A crucial issue involved in high speed clocked designs is the determination of the clock distribution method to be employed. The clock distribution used is illustrated in figure 3.

Each row of the multiplier has a common clock line – driven by a clock buffer – for all the blocks within it. The clock buffer is a CMOS inverter. In order that the clock have a 50% duty cycle a P device width 2.75 times that of the N device is required. Using parallel connected 8 N devices of width $16\mu m$ and 8 P devices of width $44\mu m$ (length $1.6\mu m$) rise and fall times as determined by SPICE simulations with a load of $5.3pf$ are $650ps$. SPICE simulations of basic blocks use rise and fall times which are much larger ($2.0ns$).

Figure 3 shows the inverter sizes, along with the load present on each row. The bottom-most row being the AND gate row. Each of the Full-adder rows has a capacitive load of $5.3pF$. Buffers used for these rows are similar. Half-adder rows however do not have a uniform capacitance. Vector-merge adder rows have a smaller capacitance than full-adder rows with the outermost row having the minimum capacitance $1.8pF$. This allows two adjacent rows to be driven by a single clock buffer. In all ten clock buffers are needed. To keep skew between adjacent rows at a minimum, buffer sizes are optimized for each row. Clock skew between

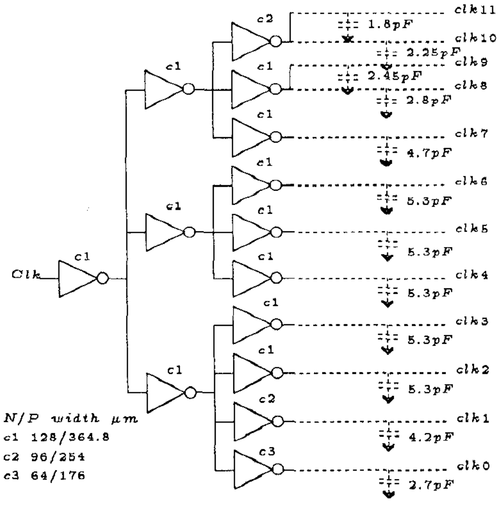

Figure 3: Clock Distribution Tree

adjacent rows has been kept within $150ps$. Routing of clock lines has been done in metal in order to reduce the propagation delay. Calculations for optimum wire width size following the method described in [6], gives an optimum width of 18λ ($\lambda = 0.2\mu m$) for a load of $5pF$. In this design a wire width of 20λ has been used.

SPICE simulations indicate fairly high peak currents of $200mA$. Hence in the power net design the power lines of the clock system are distinct from the actual multiplier power net.

5 Results

SPICE simulations of the full-adder indicates operation at clock periods of $3.8ns$ at $27°C$. Outputs of the full-adder were assumed to drive inputs of a similar cell. The robustness of the full-adder design, with respect to high temperature, is indicated by its capability to sustain clock rates of $180MHz$ at $125°C$. The full-adder can accept fairly slow clock transition edged (a triangular clock waveform being used for SPICE simulations). The average power dissipation was $6mW$. Overall timing simulations to confirm multiplier operation was done using the NELSIS simulator SLS. Multiplier power dissipation as estimated by SLS is $540mW$ at $230MHz$. We expect a slightly higher dissipation, since the simulation estimate is not very accurate.

6 Conclusion

The work shows the feasibility of implementing very high speed clocked systems, by using single phase clocked circuitry. To the knowledge of the authors this is the fastest reported multiplier in a mature standard CMOS technology. Single phase clocked circuits though more complex than standard logic styles, allow performance superior to that obtained by circuits using multiple clock phases and pipelining at much finer levels, than currently employed. A comparison of the performance obtained by this multiplier, with respect to other high performance multipliers is shown in table 1.

	Noll [6] 8 X 8 multiplier	Hatamian [3] 8 X 8 multiplier	Lu [7] 12 X 12 multiply accumulate	Present Work 8 X 8 signed multiplier
Clock MHz	330	75	140	230
Latency	18	16	13	12
Transistor no.	5480	–	10616	5176
Area (mm^2)	1.5 X 0.4	–	2.5 X 3.7	1.5 X 1.4
Power diss. (W)	1.5	0.25	1.5	0.540
Technology	NMOS $1\mu m$	CMOS $2.5\mu m$	CMOS $1\mu m$	CMOS $1.6\mu m$

Table 1: Comparison of Pipelined Synchronous Multiplier architectures.

For multipliers with longer word lengths the current scheme is not suitable, because the length of the multiplier line and its capacitance is a direct function of the word length. A truly systolic architecture [8] would be more suitable. The aim of this work is to highlight the superiority of true single phase clocking for finely pipelined architectures. The advantages would be more apparent for larger and less uniform architectures where clock skew issues constitute the main design problem.

7 Acknowledgements

This work was funded in part by grants from the Department of Electronics, Government of India. The authors would also like to thank Tanay Krishna for his design effort [8] that laid the foundation for this work.

References

[1] Y. Jiren, I. Karlsson and G. Svensson, "A true single phase clocked dynamic CMOS circuit technique", *IEEE Journal of Solid State Circuits*, vol. SC-22 1987.

[2] Y. Jiren and C. Svensson, "High Speed CMOS Circuit Technique", *IEEE Journal of Solid State Circuits*, vol. SC-24 February 1989.

[3] M. Hatamian and G. Cash "70 MHz 8 x 8 Parallel Pipelined multiplier in 2.5 um CMOS", *IEEE Journal of Solid State Circuits*, vol SC-21, no. 4, August 1986.

[4] P. van der Wolf *et. al.*, "An Introduction to the NELSIS IC Design System", Delft University of Technology, The Netherlands, August 1990.

[5] C. R. Baugh and B. A. Wooley, "A two's complement Array Multiplier algorithm", *IEEE Trans. Computers*, vol C-22, no. 12 pp. 1045-1047, Dec. 1973.

[6] T.G. Noll, D.S. Landsiedel, H. Klar, and G. Enders, "A Pipelined 330 MHz Multiplier", *IEEE Journal of Solid State Circuits*, vol SC-21, no. 3, June 1986.

[7] Fang Lu and Henry Samueli, "A 140-MHz CMOS Bit-Level Pipelined Multiplier-Accumulator using a new dynamic Full-Adder Cell Design", *IEEE Symposium on VLSI Circuits*, pp 123-124, 1990.

[8] Tanay Krishna, "Bit Systolic Design for High Speed VLSI Architectures", M.E. Thesis, Indian Institute of Science, 1991.