Investigating CMOS Process Technology with a Multiplier-Accumulator

B. Perry, K. Altendor
ECE 471- Oregon State University

Abstract—A pipelined 4x4-bit multiplier with 12-bit accumulator was designed using CMOS technology. Several new and smaller process IC technology processes have developed in the past decade. This document investigates the performance of the MAC design across several sub-micron processes ranging from .25um to 22nm.

I. INTRODUCTION

The DSP multiplier-accumulator was initially designed on a 250nm static CMOS technology. The pipelined architecture on the MAC can improve the throughput of the design at the cost of latency, due to reliance on the slowest logic function when clocking pipeline stages. However, the focus of this document is on the advantages of improving deep sub-micron technologies down to 22nm. The inherent disadvantages of decreased technology sizes are IR drop, electromigration, and leakage currents due to small channel widths. This phenomenon, however, are beyond the scope of this paper. Sections 1 through 3 focuses on the MAC design and section 4 focuses HSpice simulation results. The simulation results were based on standard nominal process technology simulation parameters. The software used for the 250nm design was Cadence Schematic-composer and Virtuoso on Linux OS.

II. MAJOR PIPELINE CELLS

The two major pipeline cells are the full-adder and the tri-state inverters with inverter that we designed. The layout for the full-adder and accumulator can be seen in Fig.1 and Fig. 2 respectively. Both these layout designs were compact for optimum cell space. In both, metal1, metal2 and poly were used for internal cell signals, metal1 was used for positive and negative voltage rails, metal4 was used for clock signals (overlaying the voltage rails), and metal3 was used for clock signals. All cell components including inverters and NAND gates were designed to be the same height, to simplify the later design of the pipeline architecture.

A. Full-Adder

The full-adder we implemented [1] was a conventional static CMOS structure. It is also used in the MAC architecture to implement the half-adder by shorting one of the inputs to ground or logical zero. The approximate layout size is 13.2 x 11.1 microns.

B. Clocked Inverter

This cell is the primary component of our pipeline architecture. It consists of two tri-state inverters and a F01 inverter. Each tri-state inverter component has clock and inverse clock inputs. These components are not edge triggered. The inverter simply turns on or passes the inversion of the input value when the clock signals toggle either the PMOS or NMOS into saturation mode. The third non-clocks inverter was added for ease of design of our inversion paths to the output and to help drive the signal output by providing the net between tri-state inverters with a path to the supply rails. The cell size is approximately 6 microns by one side than the full-adder at 19.9 x 11.1 microns.
III. MAJOR DESIGN BLOCKS

In our MAC design of course the multiplier and accumulator are our two major block components. Both the multiplier and accumulator have some triangular-shaped pre-skewing and post-skewing circuits are added for timing purposes. When implementing the MAC we remove the multiplier post-skewing circuit and the accumulator pre-skewing circuit for higher throughput.

The multiplier and accumulator schematics of Fig. 3 were positioned with the clocks of each stage running horizontal and the signals propagating vertically. The pre- and post-skewing portion of the circuits can clearly be seen. To determine the maximum clock frequency of the system we tested all of the logic clock delays as driven by our tri-state inverters.

A. Pipelined Multiplier

The multiplier [3] we used was a unsigned carry-add type design. Though multiple test circuits it was found the full-adder had maximum possible delay when the sum output depended on the worst case transition of the carry output. The latency is approximately 8 or 9 clock periods, before results are propagated to the output.

B. Pipelined Accumulator

The accumulator [1] effectively adds the input of a stage to the output of the last output of that stage. For example, if we focused on the first stage with a constant input of ‘1’ (bit-0 high), the first clock cycle you would see one at the output of that the first full-adder. The ‘1’ would then begin to propagate down through the post-skew path. However, also within that cycle the ‘1’ would also feedback into the first full-full adder resulting in an output of ‘2’ (bit-1 high). With an input of ‘1’ into the accumulator this would in effect act like a counter, with slight latency. The layout of the accumulator not including the post-skew circuit can be seen in Fig. 4.
C. Clock Tree

The clock tree consists of a series of inverters that drive the clock devices and transitions in the circuit. It consisted of 7 inverters ranging from a fan-out of 729 to a fan-out of 1 by multiples of 3. This number was primarily based on the number of gates the clock was driving in both the accumulator and multiplier.

D. Pipelined MAC

The MAC schematic we used for our process technology simulations is the multiplier and accumulator schematics meshed together without the multiplier post-skew and accumulator pre-skew circuits. As stated earlier, this in effect reduces the latency ad increases throughput by allowing the accumulator to begin tabulating results and propagating to the output as soon as the first bit (bit-0) is received. By the next clock cycle when the accumulator needs the bit-2 it should just be propagating out of the multiplier.

IV. SIMULATION RESULTS

<table>
<thead>
<tr>
<th>Process (nm)</th>
<th>Latency (ns)</th>
<th>Chip Area (um²)</th>
<th>Clock Power (mW)</th>
<th>Logic Power (mW)</th>
<th>Total Power (mW)</th>
<th>Frequency (GHz)</th>
<th>GHz mW</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>10.0</td>
<td>43,300</td>
<td>222</td>
<td>35.0</td>
<td>277</td>
<td>1.4</td>
<td>0.00502</td>
</tr>
<tr>
<td>130</td>
<td>3.73</td>
<td>11,700</td>
<td>54.4</td>
<td>12.4</td>
<td>66.8</td>
<td>3.75</td>
<td>0.0561</td>
</tr>
<tr>
<td>90</td>
<td>3.50</td>
<td>5,610</td>
<td>23.3</td>
<td>5.77</td>
<td>29.1</td>
<td>4.60</td>
<td>0.137</td>
</tr>
<tr>
<td>65</td>
<td>2.80</td>
<td>2,930</td>
<td>19.4</td>
<td>5.00</td>
<td>24.4</td>
<td>5.60</td>
<td>0.205</td>
</tr>
<tr>
<td>45</td>
<td>2.95</td>
<td>1,400</td>
<td>9.21</td>
<td>2.45</td>
<td>11.7</td>
<td>4.75</td>
<td>0.406</td>
</tr>
<tr>
<td>32</td>
<td>2.95</td>
<td>709</td>
<td>4.59</td>
<td>1.18</td>
<td>5.77</td>
<td>4.75</td>
<td>0.823</td>
</tr>
<tr>
<td>22</td>
<td>2.21</td>
<td>335</td>
<td>4.15</td>
<td>1.20</td>
<td>5.35</td>
<td>6.33</td>
<td>1.18</td>
</tr>
</tbody>
</table>

Table 1. MAC simulation results comparing each technology process.

Our simulations of Table 1 were carried out in HSpice based on the schematic netlists. As one would these results show that the by scaling technology by approximately a factor of about 1, from 250nm to 22nm, decreases chip area and power consumption, while increasing the maximum frequency. Over the full range GHz per mW increases by a factor 3. However, at 90nm and below though we see a proportional power decline with technology decrease, the rates benefits to frequency fade.

Fig. 5 shows the linear relationship between operating power and chip area. We also see that approximately 80% of power is dissipated in the clock.

Fig. 6. Latency and frequency variations between process technologies on the MAC.

Fig. 6 shows the relationship between latency and frequency. The decrease is latency at smaller processes can be owed to smaller clock periods and added precision.

V. CONCLUSION

From our results the best design for speed, total power, and chip size would be the 22nm model. However, in terms of cost for R&D, the larger processes could suffice depending on your preferred application. If I were worried about power consumption I might reduce the supply voltage and frequency of the chip rather than design a new one. Based on Hspice model parameters the circuit was robust down to 22nm.

In retrospect, it is interesting to note about this design is that with minimal inputs to the multiplier resulting in a counter, the 12-bit accumulator will overflow 1.6 million times per second (assuming the 6.5 GHz clock). The design is clearly not practical as multipliers commonly operate at much slower frequencies and can often include a 40-bit accumulator.

Further research I would like to see done includes: an investigation into the pros and cons of higher or lower supply voltages on deep sub-micron chip processes and an investigation of optimization techniques.

ACKNOWLEDGMENTS

The author would like to thank the K. Altendor, P. Chiang and Oregon State University for support and guidance during this project.
REFERENCES

