Citation	N. Reynders, W. Dehaene, (2012), Variation-Resilient Sub-Threshold Circuit Solutions for Ultra-Low-Power Digital Signal Processors with 10MHz Clock Frequency Proceedings of IEEE European Solid-State Circuits Conference (ESSCIRC), 38, 474-477.
Archived version	Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher
Published version	http://dx.doi.org/10.1109/ESSCIRC.2012.6341358
Journal homepage	
Author contact	nele.reynders@esat.kuleuven.be + 32 (0)16 321104

(article begins on next page)

Variation-Resilient Sub-Threshold Circuit Solutions for Ultra-Low-Power Digital Signal Processors with 10MHz Clock Frequency

Nele Reynders and Wim Dehaene
Katholieke Universiteit Leuven, ESAT-MICAS
Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium
Email: nele.reynders@esat.kuleuven.be

Abstract—This paper presents a variation-resilient, complete design strategy for sub-threshold Digital Signal Processors (DSP) based on a novel combination of circuit and microarchitectural techniques of which a new differential Transmission Gate logic family is the most prominent. The strategy is successfully validated by a 16 bit, 90 nm CMOS Multiply-Accumulate (MAC) unit which operates down to a supply of 150 mV at a clock frequency of 5 MHz and 0.96 pJ energy consumption per operation. Minimum energy per operation of 0.87 pJ occurs at a supply of 190 mV and a 10 MHz clock.

I. INTRODUCTION

Low energy consumption per operation is crucial for many applications. As the switching energy of digital logic decreases with V_{dd}^2 , reduction of the supply is the most effective solution to achieve this objective. Lowering V_{dd} further into the sub-threshold region results in even more energy savings. However, designing circuits to operate in the sub-threshold region poses many challenges. First, an exponential increase of delay results in a low performance. Second, sub-threshold circuits exhibit a high susceptibility to variations which leads to high variations in delay, decreased robustness and even failing gates. Consequently, sub-threshold systems often require much overhead which compromises the low-power benefit of working in sub-threshold. Therefore, sub-threshold operation is not often used in commercial applications, although applications such as [1] require ultra-low energy consumption at MHz-frequency. This paper provides a highly efficient, robust design strategy that can be applied to all units of DSP designs operating in the sub-threshold region. The Multiply-Accumulate unit is considered as the basic building block of many DSP algorithms. To illustrate the efficacy of the proposed strategy, this paper presents a MAC with better than state-of-the-art sub-threshold performance, both in terms of energy and in terms of speed.

Section II discusses low-level logic gate design in the sub-threshold region, while section III focuses on the high-level architectural design. Section IV describes the design of the Multiply-Accumulate unit. Finally, the measurements results and state-of-the-art comparison are discussed in section V.

II. LOGIC GATE DESIGN

The sub-threshold design strategy provides a solution from microarchitecture down to layout level. Starting with logic gate

Fig. 1. The single generic TG logic block with which 8 different logic gates can be constructed, including non-inverting gates. Transistor stacking is employed to reduce nMOS leakage and increase robustness. The speed loss thereby is negligible because the slowest edge, i.e. the rising edge, is dominated by the weak pMOS and will only be influenced slightly by the slower stacked nMOS transistors.

design, Transmission Gate (TG) logic is the most robust option for sub-threshold operation, as discussed by [2]. The variation-resilience arises from the inclusion of both nMOS and pMOS transistors in each conducting path. Statistically, the effect of variations on both transistors tends to be compensated by the presence of the complementary transistor. Therefore, logic gate design is minimized to the design of a single generic TG logic block (Fig. 1) out of which 8 different logic gates can be constructed, including non-inverting gates.

The design of this block has to be optimized only once for the specific technology at hand, using techniques such as sizing for optimal noise margins and transistor stacking for leakage reduction. In this 90 nm technology e.g., nMOS stacking is employed to decrease $I_{off,n}$ and thus increase the $I_{on,p}/I_{off,n}$ -ratio in order to improve robustness and to ensure functionality under all possible variations. The stacking introduces only a slight speed degradation because the critical timing specification, i.e. the rise time, is dominated by the weak pMOS and thus the increased delay of the stacked nMOS transistors has a negligible effect on the overall speed.

III. ARCHITECTURAL DESIGN

A. Latch-based Pipelining

From an architectural point of view, the design uses a pipelined architecture to increase throughput in order to achieve a

Fig. 2. Schematic of the differential-in, differential-out latch inserted between all pipeline stages.

high clock frequency. More specifically, a latch-based pipeline is utilized since latches tolerate time borrowing. This is a very beneficial concept for sub-threshold designs due to the highly variable gate delays. In addition, the used pipelined architecture solves the 2 main disadvantages of TGs: differential input signals are available at each stage, and the small output signal loss of the TGs is compensated by the latch.

Fig. 2 shows the schematic of the latch. Its differential nature adds to the variation-resilience of the total design. The latch is controlled by TG switches. It is important to note that it contains only 2 inverters to minimize leakage, because the inverters leak significantly more than the TG switches. To achieve equal noise margins in this 90 nm technology, the pMOS transistors have to be sized excessively large compared to the nMOS transistors [2]. Therefore, the standard CMOS inverter is enhanced with nMOS stacking to decrease $I_{on,n}$ in order to relax pMOS sizing. Moreover, nMOS stacking also reduces leakage. The speed penalty of introducing nMOS stacking is very small because the rise time is again the critical timing specification and is only determined by the pMOS transistor in the pull-up path. Simulations prove that registers designed with the same approach exhibit correct and robust sub-threshold behavior, thus enabling application of the design strategy in FSMs and register-based pipelines.

B. Design of Pipelined Systems

The main design considerations when pipelining systems are the length of 1 pipeline stage and the depth of the total pipeline. This paper presents a totally new approach to drastically increase variation-resilience, both of robustness and of timing. All TG logic gates in this design are constructed differentially to improve gate reliability. As a result, since TG logic gates need differential inputs, it is possible to put them in series in order to increase stage length.

In a longer pipeline stage, assuming Gaussian distributions, the mean of the timing variation increases with the number of

Fig. 3. Cumulative Distribution Function (CDF) of the propagation delay of pipeline stages with different stage length at $V_{dd} = 150$ mV, obtained with 500 Monte Carlo simulations and fitted with normal distributions. The standard deviation of stage length 1 is 3.29, while the σ of stage length 2 is 5.28, which is 1.6x higher. Introducing averaging of timing variations by increasing stage length in sub-threshold pipelined systems thus proves to be a valuable technique.

consecutive gates ($\mu' = n \cdot \mu$), whereas the standard deviation only increases with the square root of the number of gates ($\sigma' = \sqrt{n} \cdot \sigma$). By increasing the stage length, averaging of timing variations of the different consecutive gates is thus obtained. Fig. 3 shows the Cumulative Distribution Function of the propagation delay of pipeline stages with stage length 1 and 2, obtained with extensive Monte Carlo simulations. The standard deviation σ_1 of stage length 1 is 3.29, whereas σ_2 of stage length 2 is 5.28. The standard deviation is thus increased with a factor of 1.6, slightly higher than the square root of 2. However, since the system is operating in the sub-threshold region, it is very probable that the propagation delay is not normally distributed. Nonetheless, increasing the stage length in sub-threshold designs proves to be valuable because an important degree of averaging is still obtained. This averaging, combined with the latches' time borrowing capability, is very advantageous considering the large variations in sub-threshold delay.

However, increasing stage length can compromise the robustness due to a significant signal loss caused by too many consecutive TG logic gates. In the proposed design, each pipeline stage thus consists of 2 TG logic gates in series. Constructing logic gates differentially does not increase total area or energy consumption, because it allows increasing stage length and therefore decreasing the number of latches. Since the area, leakage power and dynamic energy of a latch are significantly higher than of a TG logic gate, using differential gates actually improves area and energy figures. To conclude, this paper presents a pipelined MAC architecture that makes a trade-off between stage length and pipeline depth and hence increases variation-resilience and reduces the number of latches, and thus also chip area and energy consumption. An automated but very dense layout is acquired by using the dedicated Datapath Generator [3] to perform place and route.

Fig. 4. Block diagram of the 16 bit Multiply-Accumulate unit.

IV. MULTIPLY-ACCUMULATE DESIGN

Fig. 4 shows the architecture of the 16 bit MAC with 32 bit accumulation. To efficiently pipeline the architecture, it is essential that the multiply and accumulate operations are not sequentially used, but rather that both operations are interwoven. More specifically, the proposed 2's complement MAC operation is based on the Modified Baugh-Wooley multiplier algorithm [4] and is extended with an interwoven diagonal accumulation structure. In addition, the system is expanded so that it can operate in 3 different modes: MAC, multiplier (MULT) and multiply-add (MADD) mode. Feedback, necessary for the accumulation in MAC mode, is obtained by using 32 dedicated feedback latches (Fig. 5). The feedback latch not only realizes the accumulation, but also enables operation in MULT and MADD mode. The configuration of the latch is different in every mode:

- MAC: The clock signal en_MAC is configured to be complementary to $\overline{en}$. The MAC input bits are the previously calculated product bits used for accumulation and the MADD inputs bits are cut off ($en_MADD=0$). At start-up of MAC mode, it is compulsory that the feedback latches are reset to ensure the first accumulation with zeroes.
- MULT: The feedback latches are permanently reset while MAC and MADD input bits are cut off ($en_MAC=0$, $en_MADD=0$), so that no accumulation occurs.
- MADD: Feedback is cut off ($en_MAC=0$) and a third 32bit input is provided from the right side (Fig. 4). Due to the pipelining, these MADD input bits need to be delayed by placing latches in order to ensure arrival to the feedback latches on the correct time. The signal en_MADD is permanently high, while $\overline{en}$ is always low.

Resetting the feedback latch is performed differentially by two minimal pMOS transistors in order to minimize leakage. A minimal pMOS transistor leaks 12.6 times less than a minimal nMOS transistor at $V_{dd} = 150$ mV. Moreover, using two pMOS transistors requires only a single reset signal, instead of two differential signals, which facilitates routing.

Fig. 5. Schematic of the feedback latch that enables operation in MAC, MULT and MADD mode.

Fig. 6. Boxplot of (top) measured maximum clock frequencies and (bottom) measured energy consumption per operation as function of V_{dd} , measured out of 34 dies. All values mentioned in the text are average values.

V. MEASUREMENTS

Fig. 6 shows boxplots of measurement results of 34 dies. Measurements show that the MAC is fully functional down to a supply voltage of 150 mV. Clock frequencies from 5 to 80 MHz are obtained by varying V_{dd} from 150 to 310 mV at energy consumptions of 0.87 to 1.5 pJ per operation. The minimal energy consumption per operation point occurs at a supply of 190 mV and a 10 MHz clock. This work thus succeeds to achieve MHz-performance in the sub-threshold region, while retaining an ultra-low energy consumption. The FOM used in the proposed design strategy is the Energy-Delay Product (EDP) because it combines energy consumption and clock frequency. Table I provides detailed measurement results at 4 key voltages. The key voltages are chosen for

TABLE I

DETAILED AVERAGE MEASUREMENT RESULTS AT 4 KEY VOLTAGES OF THE MAC. THE KEY VOLTAGES ARE CHOSEN FOR MINIMAL SUPPLY OPERATION, MINIMUM ENERGY OPERATION AND CLOCK FREQUENCIES OF 15 AND 50 MHz, AS INDICATED BY THE BOLD VALUES.

V_{dd} [mV]	Throughput [MHz]	Latency [μ s]	Energy [pJ]	EDP [pJ. μ s]
150	5.0	6.40	0.96	0.193
190	10.5	3.05	0.87	0.083
210	15.4	2.08	0.94	0.061
280	51.3	0.62	1.29	0.025

Fig. 7. Chip micrograph of the 90 nm CMOS, 16 bit Multiply-Accumulate.

operation at minimal supply voltage, operation at minimal energy consumption and operation at clock frequencies of 15 and 50 MHz. The active MAC area is $310 \times 225 \mu\text{m}^2$ (Fig. 7).

Table II provides a state-of-the-art-comparison of sub-threshold Multiply-Accumulate units. Making a comparison between this work and the only other published sub-threshold MAC [5] is difficult because it is not processed in the same technology and it has a different bit length. As shown in the technical specs, the MAC in [5] has been produced in a $0.14 \mu\text{m}$ CMOS technology. It consists of an 8 bit sequential MAC, i.e. an 8 bit array multiplier followed by a standard 24 bit ripple carry adder and accumulator.

A comparison has been made on reported specifications at the minimal supply voltages of both designs, shown in the middle table. [5] is able to operate at a minimal V_{dd} of 175 mV, while this work is functional down to 150 mV. In order to get a meaningful comparison, the specifications have been reworked in the bottom table to acquire equal bit length and technology. First, the measurement results of this work have been transferred from a 16 bit to an 8 bit design. Second, [5] is ported from a 140 nm to a 90 nm CMOS technology in an optimistic manner, without taking into account the increased leakage. This comparison shows that while [5] reaches a smaller energy consumption, it also has a significantly smaller operating frequency. Therefore, this work has a 4.4 times better EDP than [5] for minimal V_{dd} . To conclude, this design outperforms [5] in performance and EDP.

VI. CONCLUSION

This paper presents a complete design strategy for sub-threshold DSPs. The use of differential TG logic and the design of the latch guarantees variation-resilience. The impact of timing variations in sub-threshold systems is minimized by two techniques. First, the latch-based pipelined architecture allows time borrowing between different pipeline stages. Second,

TABLE II

STATE-OF-THE-ART SUB-THRESHOLD MAC COMPARISON.

Technical specs	This work	[5]
Bit length MAC	16 bit	8 bit
Technology node	90 nm	$0.14 \mu\text{m}$
Architecture	Pipelined, interwoven	8bit Multiplier + 24bit Adder

Specifications	This work	[5]
Minimal V_{dd} [mV]	150	175
Throughput [MHz]	5.0	0.166
Power [μ W]	4.8	0.014
Energy [pJ]	0.96	0.084
EDP [pJ. μ s]	0.193	0.508

Specifications @ minimal V_{dd}	This work			[5]		
	16 bit	→	8 bit	140 nm	→	90 nm
V_{dd} [mV]	150	=	150	175	=	175
Throughput [MHz]	5	=	5	0.166	$\times S$	0.258
Energy [pJ]	0.96	$\div 4$	0.24	0.084	$\div S$	0.054
EDP [pJ. μ s]	0.193	$\div 4$	0.048	0.508	$\div S^2$	0.210

$$S = 140 \div 90 = 1.556$$

by increasing the stage length averaging of timing variations is obtained. Moreover, increasing stage length reduces the number of latches which results in area and energy savings. To summarize, the proposed design strategy proves to be very variation-resilient, both in terms of robustness and in terms of timing. This design strategy is successfully implemented in a 16 bit MAC that reaches both ultra-low energy consumption and MHz-speed. The MAC is functional down to a V_{dd} of 150 mV at a clock frequency of 5 MHz and 0.96 pJ energy consumption per operation. The minimum energy per operation point of 0.87 pJ occurs at a supply of 190 mV and a 10 MHz clock. Furthermore, the design of the MAC proves that this design strategy is applicable in systems that contain feedback. The state-of-the-art comparison demonstrates that this design achieves a large improvement in speed and a significantly lower Energy-Delay-Product.

ACKNOWLEDGMENT

This work is supported by the Research Foundation - Flanders (FWO). The authors would like to thank Tobias Noll and Oliver Weiss from RWTH Aachen for the opportunity to work with Datapath Generator and the provided support.

REFERENCES

- [1] M. Ashouei *et al.*, "A voltage-scalable biomedical signal processor running ECG using 13pJ/cycle at 1MHz and 0.4V," in *ISSCC*, Feb. 2011, pp. 332–334.
- [2] N. Reynders and W. Dehaene, "A 190mV supply, 10MHz, 90nm CMOS, pipelined sub-threshold adder using variation-resilient circuit techniques," in *A-SSCC*, Nov. 2011, pp. 113–116.
- [3] O. Weiss, M. Gansen, and T. Noll, "A flexible datapath generator for physical oriented design," in *ESSCIRC*, Sept. 2001, pp. 393–396.
- [4] M. Hatamian and G. Cash, "A 70-MHz 8-bitx8-bit parallel pipelined multiplier in $2.5\text{-}\mu\text{m}$ CMOS," *JSSC*, vol. 21, no. 4, pp. 505–513, 1986.
- [5] J. Kao, M. Miyazaki, and A. Chandrakasan, "A 175-mV multiply-accumulate unit using an adaptive supply voltage and body bias architecture," *JSSC*, vol. 37, no. 11, pp. 1545–1554, Nov. 2002.