Low Power Application Specific Processing Element

Tatjana R. Nikolić[1], Goran Lj. Djordjević[1], and Mile K. Stojčev[1]

***Abstract* – Today the power consumption of integrated circuits (ICs) is considered as one of the most important problems for high performance chips. Many low power techniques have been proposed during the last 15 years [1]. In this paper the structure and instruction set of a processing element (PE), as a basic building block of complex ICs, is described. The micro-power PE's model is described writing efficient register transfer logic (RTL) code from a low- power standpoint using clock gating technique. The PE is implemented in FPGA technology. Simulation results show that power reduction of up to 10 times, in respect to standard PE design solutions, can be achieved using micro-power PE.**

***Keywords* – Low power, Clock gating, Processing element.**

I. INTRODUCTION

The ongoing progress in silicon technology fosters the transition from board level integration towards System-on-Chip (SoC) implementations of embedded systems. According to the International Technology Roadmap for Semiconductors by the end of the decade SoCs will grow to 4 billion transistors running at 10 GHz and operating at one volt [2]. In order to achieve high performance embedded systems three primary goals have to be successfully solved. The first one relates to high-speed computational PE capability, the second one is with an efficient data transfer among SoC's building blocks, and the third one is concerned with micro-power consumption. Systolic arrays, as a regular arrangement of PEs, are good candidate accelerator-architectures that are used in many SoC designs with aim to achieve high computational and communication performance. A challenging problem which should be taken into consideration by the SoC designers now relates both to increase performance and improve energy efficiency. The architectural PE design for low-power is in focus of our interest in this paper.

In particular we will consider RTL synthesis of a simple micro-power PE which represents a constituent of a complex SoC design.

The rest of the paper is organized as follows. In Section II we identify the main sources of power consumption in CMOS circuits. Section III describes the basic principles of clock gating technique. In Section IV the instruction set and hardware structure of the micro-power PE are presented. Section V deals with results obtaining during simulation and implementation of standard- and micro-power-PE on Spartan2 and Virtex4 series. Finally, Section VI gives a conclusion.

[1]Tatjana R. Nikolić, Goran Lj. Djordjević, Mile K. Stojčev are with the Faculty of Electronic Engineering, A. Medvedeva 14, 18000 Niš, Serbia, E-mail: {tatjanas,gdjordj,stojcev}@elfak.ni.ac.yu

II. POWER CONSUMPTION COMPONENTS

Total power dissipation of CMOS inverter is composed of static and dynamic consumption

$$P_{total} = P_{dynamic} + P_{static} \quad (1)$$

The static dissipation can be expressed as

$$P_{static} = I_{leakage} \cdot V_{DD} \quad (2)$$

As can be seen from (2) P_{static} depends on leakage current, $I_{leakage}$, and supply voltage V_{DD}. The leakage current depends on technology and the number of used gate.
The dynamic dissipation is given by the fallowing formula

$$P_{dynamic} = P_{cap} + P_{scc} \quad (3)$$

where P_{cap} is caused by the charging and discharging of parasitic capacitances in the circuit, while P_{scc} is short circuit dissipation.

$$P_{cap} = \alpha_{o\to 1} f_{CLK} \int_0^T i_{V_{DD}}(t) V_{DD} dt = \alpha_{0\to 1} f_{CLK} C_l V_{DD}^2 \quad (4)$$

P_{cap} is proportional to the operating frequency, f_{CLK}, switching activity, α, switched capacitor, C_l, and V_{DD} at power of two. The power dissipation when both transistors are switched on is given by

$$P_{scc} = \alpha_{0\to 1} f_{CLK} I_{sc} (\frac{t_r + t_f}{2}) V_{DD} \quad (5)$$

where t_r corresponds to the rising and t_f to falling pulse edges, respectively. According to (5) we see that P_{scc} directly proportional to the short circuit current, I_{scc}. Approximately, 20% of the total power dissipation correspond to P_{scc}, 75% to P_{cap}, and 5% to P_{static} [3]. In submicron technology dominant influence on power dissipation has $P_{dynamic}$. According to this, the main design effort related to power reduction are directed towards optimal choice of parameters α, C_l, V_{DD}, and f_{CLK}. Theoretically, power reduction can be achieved by decreasing each of the mentioned parameters. In practical realization, designers use ICs with fixed technologically defined supply voltage, so really, in this case there is no choice. The parasitic capacitance, C_l, is defined both by circuit fanout and layout. This means that this parameter is hardly to control. If we suppose now that for execution of some task n steps (operations) are needed, then there is not effect if we decrease f_{CLK}, since the number of transitions is identical in both cases. A parameter to which we can have real influence is α. In other words, α determines the number of transitions. Correct

choice of α is determined by the selected code for data presentation, and by elimination of unnecessary switching in parallel and pipeline computation. An approach based on decreasing the number of signal transitions, α, referred as clock gating, is adopted in this paper.

III. CLOCK GATING TECHNIQUE

The clock gating technique is extensively used in the design of low-power circuit [4]. It involves dynamically shutting off the clock to portions of a design that are idle or are not performing useful computation. Fig. 1 depicts the concept of the clock gating using an AND gate. The basic idea is to AND the clock with an enable signal, so that the register receives a clock signal only when the enable is high.

The granularity of the circuit block at which clock gating is applied greatly affects the power savings that can be achived because gating larger blocks results in higher power savings in the "off" clock cycle, but allows fewer number of "off" clock cycles. The following three levels of granularity can be distinguished: 1) modul-level clock gating, 2) register-level clock gating, and 3) cell-level clock gating [4].

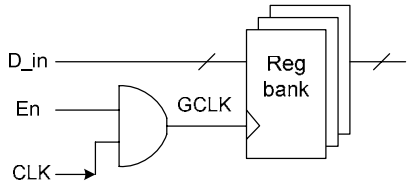

Fig. 1. Clock gating using a latch and AND gate

In this paper we will consider implementation of clock gating technique on simple PE as a constituent of 1D or 2D linear systolic array. We define systolic array as an arrangement of PEs in an array (often rectangular) where data flows synchronously across the array between neighbors, usually with different data flowing in different directions. More details concerning systolic array can be found in [5].

IV. PROCESSING ELEMENT STRUCTURE

The function of a PE, described in this paper, is defined by the following formula:

$$c_{out} = (a_i \otimes b_i) \oplus c_{in}$$

where: a_i, b_i and c_{in} are n-bit input operands; c_{out} is n-bit output operand; symbols $\otimes$ and $\oplus$ correspond to one of the arithmetic/logic operations defined in Table I.

Having in mind that in one instruction cycle the PE have to perform two operations, our design choice was to use two stage pipelining technique (see Fig. 2). The first pipeline stage (PS1) performs the operation $\otimes$ on input operands a_i and b_i, and generates output results denoted as c and time delayed input operands, a' and b'. Select input signals SEL1 define operation of PS1. The second pipeline stage (PS2) performes the operation $\oplus$ on input operands c and c_{in} and generates output results c_{out} and time delayed input operands, a_{i-1} and b_{i-1} (a_{i-1} is obtained by implementing Z^{-1} to a_i). Select input signals SEL2 define operation of PS2.

TABLE I
INSTRUCTION SET

Mnemonic	Description
`OR c,a,b`	c = a ∨ b
`AND c,a,b`	c = a ∧ b
`NOT c,a`	c = $\bar{a}$
`ADD c,a,b`	c = a + b
`SUB c,a,b`	c = a - b
`MIN c,a,b`	c = min{a,b}
`MAX c,a,b`	c = max{a,b}
`CMIN c,a,b`	$c = \begin{cases} a \text{ if } a = b \\ min(rangevalue) \end{cases}$
`CMAX c,a,b`	$c = \begin{cases} a \text{ if } a = b \\ max(rangevalue) \end{cases}$
`NEG c,a`	c = - a
`MUL c,a,b`	c = a ∗ b
`NOP`	no operation
`others operations`	reserved for future use

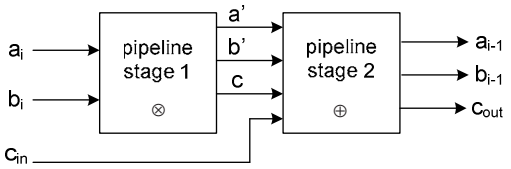

Fig. 2. Two-stage processing element

The internal PE hardware structure, composed of two PSs, is pictured in Fig. 3. As can be seen from Fig. 3, both PS are identical. Each PS is realized as a parallel hardware composed of 16 arithmetic/logic execution units, marked as EX0 to EX15. All execution units operate in parallel. The multiplexer, MUX1/MUX2, according to the control inputs, SEL1/SEL2, selects one of the 16 execution-unit-outputs at its output. Blocks CG1/CG2 marked with dashed lines correspond to the hardware structure which is needed for implementation of the clock gating technique. Let note that clock gating technique is independently realized for each PS. The block CG1/CG2 is composed of: a) 16 clocking circuits, CLKIN0 to CLKIN15; b) time delay stage for input operands, DSZ; c) opcode decoder, DEC1/DEC2; and d) clock distribution tree, CLK_TR1/CLK_TR2. The hardware block COUNT&DEC is common to both CG1 and CG2 blocks and is used as a multiphase clock generator.

In our proposal, the clock gating is realized as a two steps technique. During the first step, in order to select the specified PS's function ($\otimes / \oplus$ referred as opcode), the control signal SEL1/SEL2 is decoded and one of DEC1/DEC2 outputs Eni(0) to Eni(15), i=1,2 is activated. At the second step, the

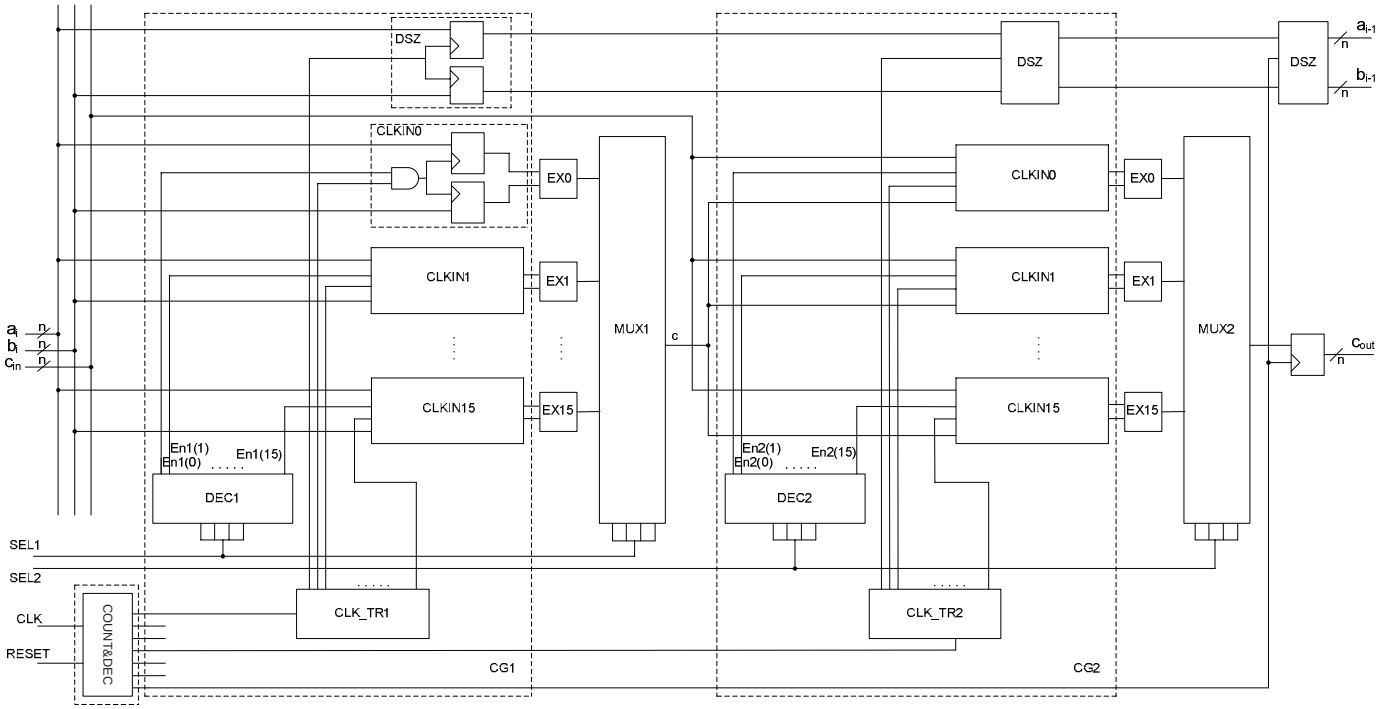

Fig. 3. Micro-power PE hardware structure

latching operation of clocking circuits CLKIN0 to CLKIN15 is enabled or disabled. Only one clocking circuit is enabled, all others are disabled.

Power reduction using this aproach is achieved thanks to the following fact: Instead to switch inputs to all execution units, only inputs to the enabled execution unit are swithed. In this way, we significantly reduce switching activity, α, defined in (4) and directly decrease energy consumption.

The PE for which we have implemented the clock gating technique is referred as a *micro-power PE*. The energy reduction is achieved at cost of decreased PE's operating speed and increased Silicon area, i.e. the number of occupied gates.

Another crucial property of the proposal relates to its possibility to implement the method at the RTL level of abstraction. In this way the proposal can be implemented independently of the target technology. In other words, standard HDL code (VHDL, Verilog) as well as standard CAD tools intended for simulation and synthesis can be used.

V. RESULTS

In order to evaluate the performance of a proposal we have implemented both the standard and the micro-power PE on Xilinx FPGA technology using a software tool ISE 9.1. We have considered and analyzed several *n*-bit PEs which can manipulate with 8-, 12-, 16-, 24- and 32-bits operands. All PEs, at behavioral level, are described using VHDL code. During the phase of HDL code-creation special care was devoted to power consumption optimization by using clock gating technique. Namely, to each execution unit a corresponding clock gating logic is appended. In this manner, significant power reduction is achieved.

The PEs are implemented on two different types of FPGA circuits, the first one to Spartan2, and the second one to Virtex4 series. Parameters that have influence on power consumption were selected to be identical in both cases. For evaluation of power consumption the Xilinx Xpower tool was used. As performance metrics the area overhead, AO, and power reduction factor, PRF, were used. The metrics AO and PRF are defined as

$$AO = \frac{\mu\_power\_PE\_gate\_count - standard\_PE\_gate\_count}{standard\_PE\_gate\_count}$$

$$PRF = \frac{standard\_PE\_power - \mu\_power\_PE\_power}{standard\_PE\_power}$$

where: μ_power - micro-power.

The obtained results which correspond to standard- and micro-power-PE, for Spartan2 and Virtex4 families, are given in Table II and III, respectively. PE's hardware complexity is presented in columns 2 and 4, and is specified as an equivalent gate count. In column 3 and 5 PE's power consumption in mW is given. The last two columns include AO and PRF metrics specified in %.

By analyzing the obtained results we can conclude the following:

1) The micro-power-PE has significantly lower power consumption in respect to standard-PE.

2) The proposed method is more efficient for larger *n*. Namely, with increasing *n* both metrices improve, i.e. the PRF increases (from 87.6% for 8-bit PE up to 97.4% for 32-bit PE, for Spartan2 serie; and from -9% for 8-bit PE up to 91.6% for 32-bit PE, for Virtex4 serie), while AO decreases (from 99.6% for 8-bit PE down to 36.7% for 32-bit PE, for Spartan2 serie; and from 172.6% for 8-bit PE down to 38.6% for 32-bit PE, for Virtex4 serie).

TABLE II
IMPLEMENTATION RESULTS FOR XILINX SPARTAN2E FPGA

number of input operands bits	standard PE		micro-power PE		area ovrhead [%]	power reduction [%]
	gate count	power [mW]	gate count	power [mW]		
8 bit	3636	104.70	7256	13.02	99.6	87.6
12 bit	7068	290.78	12270	16.22	73.6	94.4
16 bit	11236	1399.28	18032	50.24	60.5	96.4
24 bit	22065	2451.18	32106	125.98	45.5	94.9
32 bit	36240	11729.89	49550	299.85	36.7	97.4

Note: Target Device: xc2s300e-6fg456

TABLE III
IMPLEMENTATION RESULTS FOR XILINX VIRTEX4 FPGA

number of input operands bits	standard PE		micro-power PE		area ovrhead [%]	power reduction [%]
	gate count	power [mW]	gate count	power [mW]		
8 bit	2688	6,98	7328	7.61	172.6	-9.0
12 bit	6966	19,25	12252	5.19	75.9	73.0
16 bit	11020	62,30	18026	8.44	63.6	86.4
24 bit	21702	128,18	32100	15.06	47.9	88.2
32 bit	35730	462,35	49520	28.38	38.6	91.6

Note: Target Device: xc4vlx15-12sf363

3) Area overhead for both series are comparable, while power reduction is significantly larger for Spartan2 FPGA. Lower power reduction factor for Virtex4 is consequence of allready factory implemented advanced low power techniques such as, for example, the reduced power supply voltage, decreased leakage current, decreased load parasitic switching capacitors, etc.

VI. CONCLUSION

Clock gating technique is one of the most successful and widely used techniques for power reduction. In this paper the hardware structure, instruction set, and implementation of a simple micro-power processing element based on FPGA technology are described. For power reduction the clock gating technique was used. The obtained results show that power reduction up to 97.4%, with area overhead of 36.7% can be achieved. The processing element can be used as a basic building block of 1D or 2D linear systolic array.

REFERENCES

[1] C. Piguet, ed, *Low-Power Electronics Design*, Boca Raton, CRC Press, 2005.
[2] T. Kogel, R. Leupers and H. Meyr, *Integrated System-Level Modeling*, Dordercth, Springer, 2006.
[3] L. Benini, A. Bogliolo, G. De Micheli, "A Survey of Design Techniques for System-Level Dynamic Power Management", IEEE Trans. on VLSI, Vol. 8, No. 3, pp. 299-316, 2000.
[4] R. Mehra, B. Pangrle, *Synopsys Low-Power Design Flow*, pp. 40.1-40.20 in C. Piguet, ed, *Low-Power Electronics Design*, Boca Raton, CRC Press, 2005.
[5] M.O. Esonu, A.J. Al-Khalili, S. Hariri, D. Al-Khalili, "Systolic Arrays: How to Choose them", IEE Proceedings-E, Vol. 139, No. 3, pp. 179-188, 1992.