Exploiting Slack Time in Dynamically Reconfigurable Processor Architectures

Thomas Schweizer, Tobias Oppold, Julio Oliveira, Sven Eisenhardt, Kai Blocher, Wolfgang Rosenstiel

Department of Computer Engineering, University of Tuebingen
Sand 13, 72076 Tuebingen, Germany
crc@informatik.uni-tuebingen.de

***Abstract*— In dynamically reconfigurable processors, different contexts as well as different data paths within one context usually vary in their execution time. Voltage scaling offers the ability to utilize this variation to reduce power consumption. In this paper, we propose a dual-V_{DD} dynamically reconfigurable processor architecture which utilizes the varying execution time to reduce dynamic power consumption without adapting the clock frequency. Gate-level simulations reveal that the proposed dual-V_{DD} architecture reduces the power consumption of a processing element up to 22.1% and the total power consumption up to 10.5% compared to a single voltage architecture instance.**

I. Introduction

Power consumption in standard CMOS circuits can be attributed to switching power, leakage power, and short circuit power. Switching power is expressed as:

$$P \cong aC_L V_{DD}^2 f_{OP} \tag{1}$$

where a is the activity factor, C_L is the load capacitance, V_{DD} is the supply voltage, and f_{OP} is the operating frequency. According to this formula, lowering V_{DD} is the most effective way to reduce dynamic power consumption because it is proportional to the square of V_{DD}. Lowering supply voltage is generally difficult because the propagation delay in CMOS gates increases. As described in [1], delay in a CMOS gate can be approximated as:

$$T_{pd} \propto \frac{V_{DD}}{(V_{DD} - V_{th})^{\alpha}} \tag{2}$$

where α is a technology dependent parameter with values between 1 and 2, and V_{th} is the threshold voltage. However, we show that dynamically reconfigurable processor architectures can take advantage of these effects to reduce dynamic power consumption.

Power consumption is an important factor in embedded system design. To keep such systems flexible, very often reconfigurable solutions are employed. Recently, a number of dynamically reconfigurable processors have appeared on the market [2]. Some of them, for example NEC's DRP [3], can interchange their hardware configuration during runtime in less than a clock cycle to enable efficient reuse of resources. These hardware configurations, usually denoted as "hardware context", may vary in their execution time [4]. Similarly, different data paths within each context may execute in different time [5].

In this work, we propose an architectural improvement on dynamically reconfigurable processors with multiple hardware contexts that allows to exploit the slack time between different operations to reduce the power consumption by executing the faster operations on lower voltage. To realize the proposed approach we extended a model for dynamically reconfigurable processor architectures with voltage islands. For that purpose we partitioned a processing element (PE) of this model into different voltage regions.

The remainder of this paper is organized as follows: in the next section, we present related work. In Section III, we discuss the underlying architecture model and necessary architectural features to support the dual-V_{DD} approach. In Section IV, we present our experimental results. In the last section, we conclude our paper.

II. Related Work

We distinguish different voltage scaling approaches by the granularity of the circuit components. In [6], Usami *et al.* propose a dual-V_{DD} voltage approach, namely clustered voltage scaling (CVS). CVS is used to let the gates from non-critical paths run at lower voltage. This technique works well for ASIC Design. Li *et al.* designed a pre-defined dual-V_{DD} and dual-V_{th} circuit to reduce FPGA power [7]. Lackey *et al.* describe a system architecture and chip implementations methodology, that can be used to reduce power for System-on-Chip (SoC) designs [8].

In [4], Amano *et al.* apply frequency scaling to exploit the slack time on dynamically reconfigurable processors to improve the performance. Although power consumption increases, the energy consumption decreases in this approach. The Synchroscalar array uses columns of processor tiles organized into statically-assigned frequency-voltage domains [9]. The values of frequency-voltage pairs are determined according to the slack time of the task running at the frequency-voltage domain.

Our approach is similar to the CVS technique used at gate level. However, we apply that idea to the functional unit (FU) of dynamically reconfigurable processor architectures at operation level. These approaches have in common that they target power optimization by creating bounded regions for different voltages and that for each component of the system a suitable working voltage is determined. To the best of our knowledge, we are the first group applying voltage islands

to dynamically reconfigurable processor architectures. In our approach no frequency adaption is necessary. Therefore, we improve power and energy consumption.

III. The Dual-V_{DD} Architecture Model

To model a dual-V_{DD} architecture, we use the CRC model (Configurable Reconfigurable Core) [10] that was developed to represent a wide range of dynamically reconfigurable processor architectures. The CRC model consists of an array of PEs as depicted in Fig. 1. The operation of the FU as well as the multiplexers and the register banks are controlled by the outputs of a context memory. An entry of the context memory is selected at the beginning of each clock cycle by a finite state machine (FSM). Therefore the hardware context can be changed in each clock cycle. We denote this special kind of multi-context reconfiguration as "processor-like reconfiguration".

Similiar to the ADRES architecture [11], the CRC model is an architecture template rather than a fixed architecture. To create instances of the CRC model, it is configured with a variety of parameters, e.g. the number of PEs or the number of contexts. For this contribution, we augmented the CRC model with dual-V_{DD} capabilities as depicted in Fig. 2.

We used a commercial tool targeting a 90 nm multi-voltage standard cell technology to synthesize and analyze the two different PE types considering the following design aspects.

The PE_H type is much like the original processing element of the CRC model. Its functional unit is designed to operate in the same voltage level (1.0 V) as its neighboring elements within the PE. Therefore, no additional components such as level shifters are necessary. PE_H is meant to execute operations in the critical path or operations that would violate our timing constraints if executed on lower voltage. The PE_H type typically consumes more power, but it is faster.

The PE_L type is a dual-voltage module where all the components are supplied with 1.0 V except the functional unit which is supplied with 0.7 V. We altered the design of the PE to accommodate a low voltage FU without violating the design timing constraints of the PE of PE_H type. First, we redesigned the FU with a new supply voltage. The multiplier module was left out because it did not meet the timing constraints when executed in 0.7 V. Multiplication must then be always executed in the PE_H type. Second, we built in a level shifter at the output of the functional unit. At this point, the signal voltage level must be converted back to 1.0 V in order to appropriately stimulate the register bank and output multiplexers. A level shifter preceding the input ports of the functional unit is not necessary because the higher voltage from the environment is already an adequate input stimulus for the FU modified circuit. As a result of these modifications, PE_L is functionally and structurally (except for the multiplier) equal to the PE_H type. However, their circuit netlists differ from each other because the synthesis tool tries to keep the timing constraints while considering the new power supply conditions.

These two PE types may now be used to compose dual voltage instances of the CRC array. The number and spatial distribution of each PE type is decided at the architecture design time. Such design decisions are highly dependent on the target application and its mapping strategy.

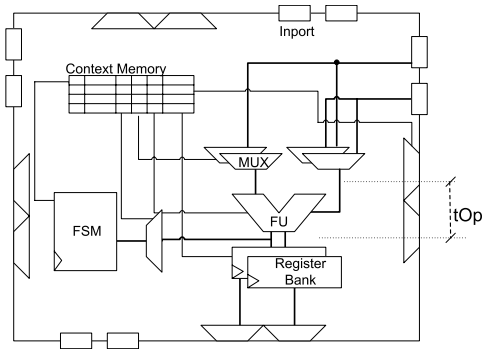

Fig. 1. Standard processing element (PE_H).

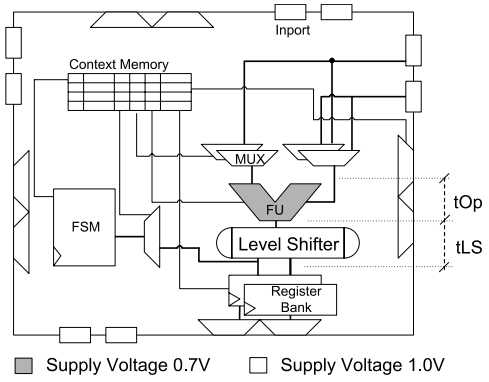

Fig. 2. Dual-V_{DD} processing element (PE_L).

IV. Experimental Results

A. Delay and Area

To present our results, we estimate the execution delay as indicated in Fig. 1 and Fig. 2. This delay consists of the time for a given operation to be executed in the FU (tOp). For PEs of PE_L type, it must be considered that the delay also includes the time spent on the voltage level shifter (tLS). Table I compares the delay of the data path section composed of operation and level shifter delay on a PE of PE_L type with the delay of operations running on the FU of PE_H type. The state-reg column denotes the time from the rising clock edge until the result is available to be stored in one of the registers. The corresponding paths, being composed of the FSM, the context memory, and the operation, are subject to a timing constraint during synthesis.

For both PE types, a timing constraint of 3.75 ns was specified for synthesis. One can see that this timing constraint can be met for all operations executed on the PE of PE_H type. As indicated in Table I, the multiply operation is the critical component and only this operation yields a violation of the timing constraint on the PE of PE_L type. It is obvious that the multiplication is the slowest operation executed on the

TABLE I

COMPARISON OF DELAY OF OPERATIONS PERFORMED ON A DUAL-V_{DD} PROCESSING ELEMENT AND A STANDARD PROCESSING ELEMENT.

	PE_L				PE_H	
Op.	FU [ns]	LS [ns]	FU+LS [ns]	state-reg [ns]	FU [ns]	state-reg [ns]
*	-	-	-	-	2.67	3.61
+	2.25	0.31	2.56	3.44	0.99	2.27
−	2.23	0.29	2.52	3.44	0.94	2.26
<<	2.09	0.31	2.40	3.48	1.36	2.49
==	1.74	0.20	1.94	2.81	0.74	1.90
! =	1.66	0.20	1.86	2.73	0.58	1.86
>	1.71	0.20	1.91	2.78	0.75	1.91
>=	1.56	0.20	1.76	2.64	0.68	1.84
<	1.33	0.20	1.53	2.63	0.53	1.84
<=	1.33	0.20	1.53	2.63	0.53	1.84
and_d	0.78	0.31	1.09	2.19	0.26	1.52
or_d	0.77	0.31	1.08	2.19	0.26	1.52
xor_d	0.78	0.31	1.09	2.19	0.26	1.56
not_d	0.34	0.33	0.67	1.67	0.26	1.39
and_s	0.58	0.20	0.78	1.64	0.57	1.39
or_s	0.98	0.20	1.18	1.97	0.57	1.39
xor_s	1.30	0.20	1.50	2.38	0.59	1.41
not_s	0.31	0.20	0.51	1.38	0.25	1.07

PE of PE_H type. This means that we can execute the other operations on lower voltage, because the timing constraint is not violated by the additional delay due to level shifters and increased delay on lower voltage.

33 level shifters (32 at the data output, 1 at the flag output) are inserted at the output of the FU in a 32-bit PE. This leads to an area increase of 4.5%.

B. Power Estimation

The mapping of applications onto dynamically reconfigurable processor architectures can be done in different ways. To validate our approach and to obtain the power estimations, we mapped the luminance calculation $(xy = (c1 * xr + c2 * xg + c3 * xb + c4) >> c5)$ of the RGB to YIQ conversion from the Embedded Microprocessor Benchmark Consortium (www.eembc.org) Consumer Benchmark under two different mapping strategies.

1) Multi-context pipelined execution: For processor-like reconfigurable architectures, the 7 operations of the example can be distributed over 3 clock cycles on 3 FUs so that in each clock cycle exactly one multiplication is executed. The states resulting in this multi-context pipelined execution [12] are depicted in Fig. 3. Applying our approach to this mapping strategie means that operations which run in parallel with the multiplication can be executed on lower voltage to reduce the power consumption. Therefore, we synthesized a 1x3 array of PEs, composed of one PE of PE_H type (PE[0][0]_H) for the multiplications and two PEs of PE_L type (PE[0][1]_L and PE[0][2]_L) for the other operations. To evaluate our approach we compared the power results at gate-level of this architecture instance with an architecture instance featuring 3 PEs of PE_H type. For the comparison only one of the 3 PEs has got a multiplier module. We set the operational frequency to 243 MHz for both instances. All results are related to dynamic power consumption. We neglect leakage power, as it only amounts to 1–4% of dynamic power in our experiments.

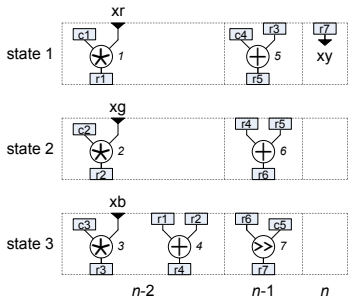

Fig. 3. States of the RGB2Y-example performing a multi-context pipelined execution.

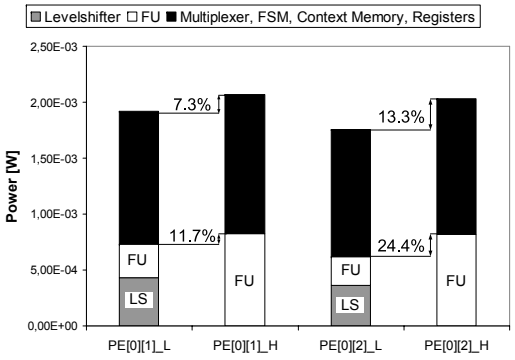

Fig. 4. Dynamic power consumption of level shifter, FU, and surrounding components of PEs of PE_L and PE_H type.

Fig. 4 presents the power results for PEs of PE_L type compared to PEs of PE_H type, executing the same operations. One can see that the power consumption of a FU of PE_L type is reduced significantly compared to a FU of PE_H type. As described before, the dual-V_{DD} PEs need level shifters at the output of the FU. Thus, we have to add the power consumption of the level shifter to that of the FU. Nevertheless, the sum of the power consumption of these two components is reduced up to 24.4%. The power consumption of a PE of PE_L type compared to a PE of PE_H type is reduced up to 13.3% in this example.

Table II summarizes the power results for the two architecture instances. The power consumption for the dual-V_{DD} architecture instance is reduced by 5.6% compared to the single-V_{DD} architecture instance executing the luminance calculation.

2) Chained execution: Similiar to NEC's DRP the CRC model allows to chain two or more data-dependent operations within one clock cycle. As depicted in Fig. 5, in a chained execution all operations are mapped to one state on 7 FUs. To realize this mapping strategie we synthesized a 3x3 array of PEs composed of 3 PEs of PE_H type performing the multiplications and 6 PEs of PE_L type for the other operations. One PE of the PEs of PE_L type is used for routing purposes

TABLE II

DYNAMIC POWER CONSUMPTION AT 243 MHZ FOR THE RGB2Y EXAMPLE PERFORMING A MULTI-CONTEXT PIPELINED EXECUTION.

Processing Element	Dual-V_{DD} Power [W]	Single-V_{DD} Power [W]	Reduction %
PE[0][0]	3.74E-03	3.76E-03	−0.5
PE[0][1]	1.92E-03	2.07E-03	−7.3
PE[0][2]	1.76E-03	2.03E-03	−13.3
Total Power	7.42E-03	7.86E-03	−5.6

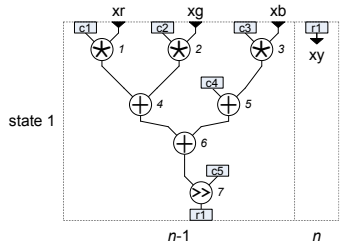

Fig. 5. States of the RGB2Y-example performing a chained execution.

only. Hence, 5 of 6 PEs of PE_L type are doing meaningful work.

In this example, we assume that the critical path is given by another context and therefore, we can use the slack time between contexts for the reduction of the power consumption. As mentioned above Amano has shown already that contexts vary in their execution time.

Our experiments show that for a chained execution, the power consumption of a FU plus level shifter power reduces up to 39% compared to a FU on high voltage level and thereby the power consumption of a PE of PE_L type reduces up to 22.1% compared to a PE of PE_H type. Table III summarizes the power results compared to an architecture instance based on 9 PEs of PE_H type. The total power consumption of the dual-V_{DD} architecture instance decreases by 10.5% compared to a single voltage architecture instance performing a chained execution of the luminance calculation.

Since we simulate only one context, no reconfiguration occurs and no power consumption can be regarded for this step. However, we consider the power dissipation of the registers in the power analysis although most of them are not needed in a chained execution. Power dissipation in registers is larger than power dissipation of the components taking part in the reconfiguration step. This means that if we switch off the registers not required but take into account the power consumption of the reconfiguration step, a further improvement can be expected.

V. CONCLUSIONS

In this work, we presented a dual-V_{DD} dynamically reconfigurable processor architecture. This heterogenous architecture composed of FUs with different supply voltages, allows to exploit the slack time between different operations and different contexts to reduce power consumption without affecting performance. We execute operations with residual slack time

TABLE III

DYNAMIC POWER CONSUMPTION AT 115 MHZ FOR THE RGB2Y EXAMPLE PERFORMING A CHAINED EXECUTION.

Processing Element	Dual-V_{DD} Power [W]	Single-V_{DD} Power [W]	Reduction %
PE[0][0]	7.09E-04	7.15E-04	−0.8
PE[0][1]	7.96E-04	7.97E-04	−0.1
PE[0][2]	6.45E-04	6.27E-04	+2.9
PE[1][0]	7.81E-04	9.54E-04	−18.2
PE[1][1]	1.03E-03	1.25E-03	−17.2
PE[1][2]	5.10E-04	5.11E-04	−0.3
PE[2][0]	2.16E-04	2.17E-04	−0.5
PE[2][1]	6.54E-04	8.39E-04	−22.1
PE[2][2]	5.63E-04	6.91E-04	−18.6
Total Power	5.91E-03	6.60E-03	−10.5

on FUs supplied with lower voltage. We demonstrated that the total power reduction of a dual-V_{DD} architecture instance reduces up to 10.5% compared to a single voltage architecture instance.

ACKNOWLEDGMENT

This work is funded by DFG under RO-1030/13 within the 'Priority Program 1148' which is focused on reconfigurable computing systems.

REFERENCES

[1] T. Sakurai and R. Newton, "Alpha-power law mosfet model and its application to cmos inverter delay and other formulas," *IEEE JSSC*, vol. 15, no. 5, pp. 584–594, 1990.

[2] H. Amano, "A survey of dynamically reconfigurable processors," *IEICE Transcations on Communications*, vol. E89-B, no. 12, pp. 3179–3189, 2006.

[3] M. Motomura, "A dynamically reconfigurable processor architecture," in *Microprocessor Forum*, 2002.

[4] H. Amano, Y. Hasegawa, S. Abe, K. Ishikawa, S. Tsutumi, S. Kurotaki, T. Nakumura, and T. Nishimura, "A context dependent clock control mechanism for dynamically reconfigurable processors," in *International Conference on Field Programmable Logic and Applications (FPL)*, 2006.

[5] T. Schweizer, J. Oliveira Filho, T. Oppold, T. Kuhn, and W. Rosenstiel, "Evaluation of temporal-spatial voltage scaling for processor-like reconfigurable architectures," in *Euro DesignCon*, 2005.

[6] K. Usami and M. Horowitz, "Clustered voltage scaling technique for low power design," in *International Symposium on Low Power Electronics and Design (ISLPED)*, 1995.

[7] F. Li, Y. Lin, L. He, and J. Cong, "Low-power FPGA using pre-defined dual-vdd/dual-vt fabrics," in *International Symposium Field-Programmable Gate Array (FPGA)*, 2004.

[8] D. E. Lackey and P. S. Zuchowski, "Managing power and performance for system-on-chip designs using voltage islands," in *International Conference on Computer-Aided Design (ICCAD)*, 2002.

[9] J. Oliver, R. Rao, P. Sultana, J. Crandall, E. Czernikowski, L. W. Jones, D. Franklin, V. Akella, and F. Chong, "Synchroscalar: a multiple clock domain, power-aware, tile-based embedded processor," in *International Symposium on Computer Architecture*, 2004.

[10] T. Oppold, T. Schweizer, T. Kuhn, and W. Rosenstiel, "Cost functions for the design of dynamically reconfigurable processor architectures," in *Workshop on Synthesis and System Integration of Mixed Information Technologies (SASIMI)*, 2004.

[11] B. Mei, A. Lambrechts, D. Verkest, J.-Y. Mignolet, and R. Lauwereins, "Architecture exploration for a reconfigurable architecture template," *IEEE Design & Test of Computers*, vol. 22, no. 2, pp. 90–101, 2005.

[12] T. Oppold, T. Schweizer, J. Oliveira Filho, S. Eisenhardt, T. Kuhn, and W. Rosenstiel, "Execution schemes for dynamically reconfigurable architectures," in *Workshop on Synthesis and System Integration of Mixed Information Technologies (SASIMI)*, 2006.