Energy and Reliability Improvement of Voltage-Based, Clustered, Coarse-Grain Reconfigurable Architectures by Employing Quality-Aware Mapping

Hassan Afzali-Kusha, Omid Akbari, Mehdi Kamal, and Massoud Pedram

***Abstract*— An energy–quality scalable coarse grain reconfigurable architecture (CGRA) based on the voltage overscaling (VOS) technique is presented. The approximation level of each processing element (PE) in the CGRA is determined by the applied VOS-determined voltage level. By employing the technique, the architecture may be configured for accurate or approximate modes of computation depending on a user-specified output quality-of-service target for a given application. More precisely, operating voltages used for performing various operations in the application dataflow graph are minimized subject to the output quality constraint by using an energy–quality tradeoff algorithm. To make the hardware implementation of the scheme more efficient, PEs are clustered into groups of (e.g., 3 × 1 and 2 × 1) voltage islands. To assess the efficacy of the proposed method in improving the power (energy) consumption and reliability of CGRAs, different combinations of minimum output quality constraints, voltage levels, and cluster sizes for several benchmarks are studied. Simulation results indicate considerable reductions in energy consumption (up to 43%) and aging rate (up to 73%) when compared with the conventional CGRA with perfect output quality (i.e., with no approximate computations).**

***Index Terms*— Coarse grain reconfigurable architecture (CGRA), approximate computing (AC), voltage clustering, voltage overscaling (VOS), aging rate reduction, energy consumption lowering.**

I. Introduction

THE Coarse Grain Reconfigurable Architecture (CGRA) consists of an array of a large number of functional units (FUs) or processing elements (PEs) interconnected by a mesh network. It is a hardware platform which performance may be optimized (reconfigured) for different applications. Using careful scheduling and mapping, CGRA can achieve a computational efficiency close to that of a custom hardware architecture [1]. While lacking the fine reconfigurability of FPGAs, their incorporation of arithmetic units enables CGRAs to run compute-intensive applications efficiently both in terms of computation speed and energy consumption. CGRAs have short reconfiguration times, high performance, and low power consumption since they admit standard cell implementations [2].

The efficiency of the CGRA fabrics may be further improved by using approximate computing (AC) [3], which is an effective technique of trading off energy/power consumption or processing speed for computational accuracy [4]. The computation approximation can be made at different levels of abstractions including algorithm, data, or hardware. AC is possible for applications where some computational inaccuracy, tends to lower the output quality, can be tolerated by the end user of the application. Fortunately, there are many applications where some degree of output quality degradation may be tolerated. Examples image and video applications, recognition, and classification [4].

There are different ways of achieving AC, such as, simplifying the target hardware, reducing width of the data path, ignoring lower significant bits of operands, and performing voltage overscaling (VOS). In the VOS technique, the operating voltage is scaled down without lowering the corresponding operating frequency. Since the critical path delay of the circuit with this over-scaled voltage becomes larger than the operating clock period (including its guard-band), some setup time violation errors may occur at the output of the circuit. The number of output errors is a function of the difference between the overscaled voltage and the nominal voltage (assumed to be a safe voltage level for the desired computational speed). An erroneous output may be perceived as an approximate result and hence, the circuit level of accuracy (alternatively, the approximation level) may be dynamically adapted at runtime [5]. This technique for AC has the additional advantage of not requiring the redesign of the circuit hardware to achieve perfect computation, that is, a 100% output quality (errorless output) may be achieved by simply applying the specified nominal voltage level. Yet, another significant advantage of the VOS technique is its considerable improvement of the circuit lifetime/reliability, which is a serious concern particularly in highly scaled state-of-the-art technologies [6].

Manuscript received January 8, 2018; revised May 29, 2018; accepted June 29, 2018. Date of publication July 17, 2018; date of current version September 11, 2018. This work was supported in part by a grant from the Software and Hardware Foundations of the National Science Foundation. Computation for the work described in this paper was supported by the University of Southern California's Center for High-Performance Computing. This paper was recommended by Guest Editor V. De. *(Corresponding author: Hassan Afzali-Kusha.)*

H. Afzali-Kusha and M. Pedram are with the Electrical Engineering Department, University of Southern California, Los Angeles, CA 90007 USA (e-mail: afzaliku@usc.edu; pedram@usc.edu).

O. Akbari and M. Kamal are with the School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran 14399-57131, Iran (e-mail: akbari.o@ut.ac.ir; mehdikamal@ut.ac.ir).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JETCAS.2018.2856838

There are several aging mechanisms which among them Biased Temperature Instability (BTI) is one of the important mechanisms in the current CMOS technologies [7]. The BTI causes a threshold voltage (V_{TH}) shift, adversely affecting the transistors' ON currents, which in turn result in path delay increases over time. The rate of the V_{TH} degradation is a strong function of the supply voltage applied to transistors. This strong dependence motivates the use of VOS technique for the AC with the additional benefit of improving the circuit lifetime/reliability.

In this work, we utilize the VOS technique in the CGRA platform. Whenever no quality deterioration can be tolerated, all PEs of the CGRAs operate at the nominal voltage level whereas when some output quality reduction can be endured, lower voltages may be applied to some of the PEs. For a given output quality, the required accuracy (and hence required operating voltage) for different PEs are determined. To reduce the cost of realizing the VOS technique in hardware, only a few discrete voltage levels are assumed. To further optimize the implementation of the technique, we suggest the use of voltage-based clustering of the PEs. A careful mapping of the data flow graph (DFG) nodes (operations) to PE clusters (voltage islands) leads us to higher energy reductions. Therefore, we present an integer linear programming (ILP) formulation for quality-aware mapping to improve the energy/lifetime of the CGRA by accepting quality loss.

The rest of the paper is organized as follows. In Section II, some background concepts are reviewed. In Section III, the related work is briefly reviewed. Section IV, which includes a motivational example, presents the proposed architecture. The proposed mapping and quality aware VOS assigning scheme determining the voltage level of each PE are explained in Section V. The results are discussed in Section VI while the paper is concluded in Section VII.

II. Background

In this section, first, we briefly review the CGRA, voltage dependences of the power/energy consumption, and then discuss the bias temperature instability as one of the main aging mechanisms in the state-of-the-art digital CMOS design technologies.

A. Coarse Grain Reconfigurable Architecture (CGRA)

A typical CGRA consists of a 2-D mesh array of processing elements (PEs), host controller, context memory, and data memory [8]. The processing elements (PEs) or function units (FUs) (used interchangeably here) can execute common word-level operations, including addition, subtraction, and multiplication. Every PE is formed by an arithmetic logic unit (ALU), a local register file, and an output register. The ALUs can provide multiple logic and arithmetic operations determined by the configuration context data. Here, for the sake of simplicity, we have assumed that each ALU component only consists of an adder and a multiplier. For each specific application, the algorithm (computations) changes, and hence, the context data configures the CGRA accordingly. The configuration may be optimized for power, speed, and/or lifetime.

The CGRA is connected to the host CPU through the host controller which is connected to the data and context memory used for storing the configuration information. The architecture is symmetric where each PE is connected to its neighbors. Also, the register files are employed to hold temporary data. Schedulers assign an FU and time to every operation in the program DFG where the operand values should be routed between producing and consuming FUs. Since dedicated routing resources are not provided, an FU either serves as a compute resource or as a routing resource at a given time. A compiler scheduler manages the computation and flow of operands across the array to effectively map applications onto CGRAs [2].

B. Voltage Dependences of Power/Energy Components

To remind the dependence of power (energy) consumptions, here, we present analytical expressions for the power consumptions of the circuit. The expression containing the dynamic (switching) and leakage power components is given as [9]

$$P = P_{switching} + P_{leak} = \alpha \cdot C_{eff} \upsilon^2 f + \eta\theta^2 e^{\left(\frac{-qV_{th}}{nk\theta}\right)} \quad (1)$$

where the α is the activity of the system, f is the operating frequency, υ is the operating voltage, θ is the temperature, C_{eff}, η, n, q, and k_θ are technology- and circuit-dependent parameters. To be complete, the short circuit power also should be added to the above expression. The expression is obtained from [10]

$$P_{shortcircuit} = \frac{\beta}{12}\left(\upsilon - 2V_{th}\right)^3 \frac{\tau}{T} \quad (2)$$

where β denotes the conductivity of the transistor per power voltage in the linear region, T is the input rise/fall time, and τ is the gate delay. These equations show the strong dependency of the power (energy) consumption to the operating voltage level.

C. Bias Temperature Instability

The bias temperature instability affects both NMOS and PMOS transistors by generating the interface traps at the Si/SiO$_2$ interface [11]. Based on the wafer-level extended Measure-Stress-Measure (eMSM) measurement on sub-20nm FinFET technology nodes [12] the long-term aging-induced shift of the threshold voltage incorporating both stress and relaxation phases, is fitted by a power law given by [13]

$$\Delta V_{th,NBTI} \cong A e^{-\frac{\kappa}{\theta}} t^{\alpha} E_{OX}^{\gamma} df^{\beta} \quad (3)$$

where A, κ, α, β, and γ are technology and power-law fitting parameters (whose values are given in Table I), t is the total stress time (s), θ is the temperature in degree Kelvin (°K), *df*is the duty factor of the stress signal, and $E_{OX} = (V_{DD} - V_{th})/T_{INV}$ is the electric field across the gate oxide (where T_{INV} is the thickness of gate inversion layer) [12]. The power-law fitting is consistent with the existing aging Bias Temperature Instability (BTI) models, such as the Reaction-Diffusion mode [12]. This relation demonstrates the strong voltage dependence of the threshold voltage shift (aging) on the supply voltage.

TABLE I
THE VALUES OF PARAMETERS USED IN II-C FOR CALCULATING THE NBTI THRESHOLD VOLTAGE DRIFT [12]

Parameter	Description	PMOS Value
γ	Power-law exp.	3
α	Power-law exp.	0.173
A	Fitting constant	2.02e-2
β	Fitting constant	1/6
κ	Fitting constant	50
V_{th}	High performance V_{th}	0.25V
T_{inv} (nm)	Inversion layer	1.4

III. RELATED WORK

In this section, some of the work related to use of the voltage over-scaling (VOS) (and voltage islanding) for reducing the power consumption, Approximate CGRAs, and CGRA lifetime improvement are briefly reviewed.

A. Power Reduction Techniques Based on VOS

In [14], by shaping the quality-energy trade off, the VOS technique was used to significantly improve the energy-efficiency. The efficiency of the technique was evaluated by applying it to the adder component in the architecture used for running the inverse discrete cosine transform (IDCT) benchmark.

In the instruction set architecture (ISA)-extended design of [5], a processor had a dual-voltage register file and two arithmetic logic units (ALUs), one was supplied by a high supply voltage level for exact computations and the other with a low supply voltage for approximate computations. Chippa *et al.* [15] proposed a new design methodology with the name of scalable effort hardware design. The notion of scalable effort was embodied in to the design process at different levels of abstractions. It involved identifying mechanisms at each level that can be used to change the computational effort expended to generate accurate results and using these mechanisms as the control knobs to trade-off between the accuracy and energy-efficiency.

A micro-architecture named Lazy Pipeline was suggested in [16]. In this architecture, the VOS approximate technique was utilized to improve the power efficiency. To reduce the timing errors of the approximated Functional Units (FUs), this microarchitecture employed vacant cycles in a VOS functional unit to extend execution and reduce the error rate. Ragavan *et al.* [17] proposed approximated operators based on the VOS technique for error-resilient applications. The energy efficiency and accuracy of the approximated operators were characterized using three knobs of the supply voltage, body biasing, and clock frequency to create models for the approximate operators (different adder designs). The statistical behavior models of the adders were used to obtain the optimum point in the energy efficiency and error margin characteristic.

In [18], techniques for designing the kernel of the error-resilient application, which can tolerate more scaling under the VOS technique, were discussed. Because this explanation has too many details which can be found in its reference and we have the space limitation, we can omit this.

There are several works on detecting and correcting the timing errors originating from VOS FUs. As an example, in [19], Razor was suggested as a technique for detecting and correcting timing errors in a VOS circuit. Lazy Pipeline micro-architecture is orthogonal to Razor and could be combined with it to reduce the number of cases where an operation needs to be repeated due to a timing error.

Recently, the use of an accuracy-aware operating voltage management unit for improving the lifetime using aggressive voltage scaling during the runtime of error-resilient applications was suggested [6]. The unit determines the operating voltage of the processor based on the type of the running application and the predefined minimum acceptable quality resulting in lifetime/reliability and power improvement at the cost of tolerable accuracy loss. There are several other works which have used the VOS technique using the similar of error detection/correction or error reduction hardware solution to reduce the quality degradation of the output. The review of these works is omitted for the sake of space.

B. Operation Level Approximation and Voltage Islanding

A different approximation-based approach works on simplifying the computations need to be performed for an operation leading to shortening the critical path delay of the function. This provides the possibility of voltage downscaling without violating the simplified circuit critical path delay. The downscaled voltage would have violated the critical path delay of the original circuit for the exact operation. For example, Lee *et al.* [20] achieved this type of operation-level approximation by bit rounding and more aggressive operation elimination. In [21], the use of approximate computing techniques at different abstraction levels was discussed. At the software level, pruning the source code based on the profiling was suggested. At the register transfer level, the authors proposed internal signal substitution (variable to variable (V2V) and variable to constant (V2C)) as well as bit-level optimization. At the high-level synthesis (HLS)-level, employing different approximated adders and multipliers was considered. Lee *et al.* [22] suggested the similar idea of using different approximate types for each operation DFG while making sure that the output quality constraint of the application is met. Downscaled voltages were applied to the operations for approximate operations without causing any timing error.

The use of voltage islands for improving the efficiency of multi-level voltages in the approximate computing also has been considered in the literature. For example, Zervakis *et al.* [23] introduced an approximate accelerators synthesis framework that enabled the usage of approximate techniques at different levels (multi-level) of algorithm, circuit, and logic level.

C. CGRA Lifetime Improvement

The only two works which have focused on improving the lifetime of CGRAs are those of [8] and [24]. In [24], the initial

TABLE II

A COMPARISON BETWEEN THE PROPOSED WORK AND SOME OTHERS IN THE AREAS OF CGRA, APPROXIMATE COMPUTING, AND VOLTAGE OVERSCALING

	CGRA Structure	**Approximate Computing**	**Lifetime Improvement**	**Energy Reduction**	**Accuracy Configurability**	**Runtime Accuracy Configurability**	**Runtime Accuracy Configurability Resolution**
[14]	×	✓	✓	✓	✓	×	-
[5]	×	✓	✓	✓	✓	×	-
[15]	×	✓	✓	✓	✓	✓	Low
[16]	×	✓	✓	✓	✓	×	-
[17]	×	✓	✓	✓	✓	×	-
[18]	×	✓	✓	✓	✓	×	-
[19]	×	×	✓	✓	✓	×	-
[6]	×	✓	✓	✓	✓	✓	Medium
[20]	×	✓	✓	✓	×	×	-
[21]	×	✓	×	✓	✓	×	-
[22]	×	✓	✓	✓	✓	×	-
[8]	✓	×	✓	×	×	×	-
[3]	✓	✓	×	✓	✓	✓	Low
This work	✓	✓	✓	✓	✓	✓	High

idea of using VOS in CGRAs considering five voltage levels was proposed by our group. Only two benchmarks of 4th order polynomial evaluation and 8-tap FIR filter were studied. In [8], a joint stress-aware loop mapping method for helping designers to select the optimized mapping with the minimal stress on the PEs was suggested. This is performed in the early phase of the CGRA mapping. Their stress effort optimization had two objectives of reducing the maximum accumulated stress on the PEs and providing more balanced stress distribution on the PEs of the CGRAs. For their first aim, they introduced a stress-aware Force-Directed scheduling method (sFDS) to schedule operations at different time slots. A rapid MCC (Maximal Compatibility Classes) search method [25] was used to find the optimal maps which had the lowest maximum stresses and distributing more operations on more PEs. A multi-map scheduling technique which used the dynamic reconfigurability feature of the CGRAs was used in their method.

D. Summary

Table II compares our work to some prior ones in the areas of CGRA, approximate computing, and voltage overscaling. It includes the parameters/features of ability to support approximate computing, lifetime improvement, energy reduction, accuracy configurability, runtime accuracy configurability, and runtime accuracy configurability resolution. In this work, the use of VOS approximate technique for improving the lifetime/reliability as well as reducing the energy/power consumption of CGRAs based on voltage islanding is proposed for the first time. This provides a platform in which a trade-off between output accuracy (quality) level from one side and energy/power consumption and lifetime/reliability from the other side may be performed while an exact computation (100% accuracy level) is also an option.

Among the previous works, only [3] and [24] focus on presenting accuracy configurable (approximate and exact mode computing) CGRA structures. Compared to the work of [3], more specifically, our proposed method provides lifetime improvement while the structure suggested in [3] mainly concentrates on energy/power reduction. In addition, our method supports different accuracy qualities with more resolution levels. Also, in comparison to [24], this work proposes the use of VOS in a CGRA considering the actual mapping, different voltage island sizes, and different voltage levels where four benchmarks are studied.

IV. PROPOSED ARCHITECTURE

In the conventional CGRA architectures, full supply voltage (V_{DD}) is used for all the PEs (as well as other blocks) and hardware modules perform exact computations. As discussed above, one may use approximate FUs to make the CGRA an approximate one (see, *e.g.*, [3]) or even use accuracy-configurable PEs to provide both approximate and exact computing modes (see, *e.g.*, [3]). In this work, we suggest an accuracy-configurable CGRA technique, which makes use of exact hardware modules while selectively using the VOS technique to switch to approximate computations. Obviously since the accuracy is a function of the PE operating voltage, to minimize the overhead of using the dynamic scaling of PE voltages (dynamic accuracy configuration), we suggest using clustering of the PEs in voltage islands. Before we explain the proposed architecture further, we discuss a motivational example which demonstrates the use of the VOS technique for reducing the energy consumption and lowering the aging rate (improving the lifetime) of the CGRA when running an application. For the rest of the section, first, we provide a motivational example and then present the details of the proposed CGRA architecture.

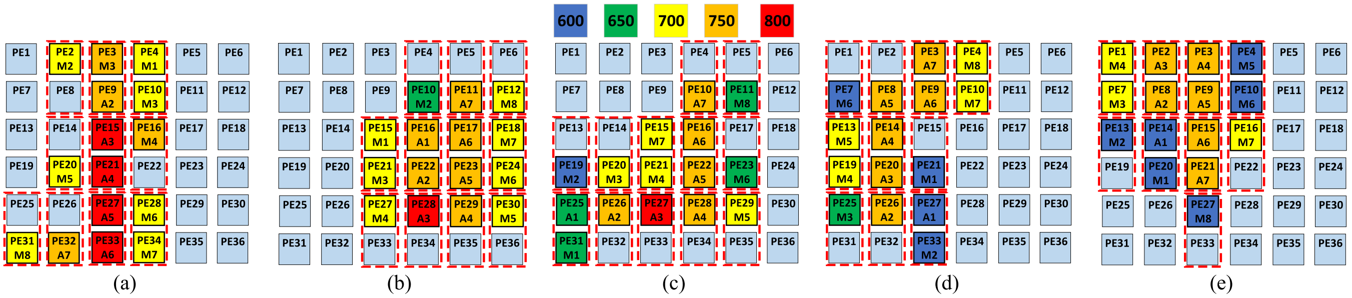

600
650
700
750
800
PE1 | PE2 M2 | PE3 M3 | PE4 M1 | PE5 | PE6
PE7 | PE8 | PE9 A2 | PE10 M3 | PE11 | PE12
PE13 | PE14 | PE15 A3 | PE16 M4 | PE17 | PE18
PE19 | PE20 M5 | PE21 A4 | PE22 | PE23 | PE24
PE25 | PE26 | PE27 A5 | PE28 M6 | PE29 | PE30
PE31 M8 | PE32 A7 | PE33 A6 | PE34 M7 | PE35 | PE36
(a)
PE1 | PE2 | PE3 | PE4 | PE5 | PE6
PE7 | PE8 | PE9 | PE10 M2 | PE11 A7 | PE12 M8
PE13 | PE14 | PE15 M1 | PE16 A1 | PE17 A6 | PE18 M7
PE19 | PE20 | PE21 M3 | PE22 A2 | PE23 A5 | PE24 M6
PE25 | PE26 | PE27 M4 | PE28 A3 | PE29 A4 | PE30 M5
PE31 | PE32 | PE33 | PE34 | PE35 | PE36
(b)
PE1 | PE2 | PE3 | PE4 | PE5 | PE6
PE7 | PE8 | PE9 | PE10 A7 | PE11 M8 | PE12
PE13 | PE14 | PE15 M7 | PE16 A6 | PE17 | PE18
PE19 M2 | PE20 M3 | PE21 M4 | PE22 A5 | PE23 M6 | PE24
PE25 A1 | PE26 A2 | PE27 A3 | PE28 A4 | PE29 M5 | PE30
PE31 M1 | PE32 | PE33 | PE34 | PE35 | PE36
(c)
PE1 | PE2 | PE3 A7 | PE4 M8 | PE5 | PE6
PE7 M6 | PE8 A5 | PE9 A6 | PE10 M7 | PE11 | PE12
PE13 M5 | PE14 A4 | PE15 | PE16 | PE17 | PE18
PE19 M4 | PE20 A3 | PE21 M1 | PE22 | PE23 | PE24
PE25 M3 | PE26 A2 | PE27 A1 | PE28 | PE29 | PE30
PE31 | PE32 | PE33 M2 | PE34 | PE35 | PE36
(d)
PE1 M4 | PE2 A3 | PE3 A4 | PE4 M5 | PE5 | PE6
PE7 M3 | PE8 A2 | PE9 A5 | PE10 M6 | PE11 | PE12
PE13 M2 | PE14 A1 | PE15 A6 | PE16 M7 | PE17 | PE18
PE19 | PE20 M1 | PE21 A7 | PE22 | PE23 | PE24
PE25 | PE26 | PE27 M8 | PE28 | PE29 | PE30
PE31 | PE32 | PE33 | PE34 | PE35 | PE36
(e)

Fig. 1. Mapping of an eight-tap FIR filter DFG on a 6 × 6 CGRA for a) 90%, b) 80%, c) 70%, d) 60%, and e) 50% quality.

A. Motivational Example

Consider an 8-tap finite impulse response (FIR) filter benchmark, which DFG is shown in Fig. 3. The results of mapping this DFG using the proposed algorithm (which will be discussed in a next section) on a 6 × 6 CGRA for different output qualities are depicted in Fig. 1. For the perfect output quality, all PEs are connected to full V_{DD}. Let us suppose that an output quality of 50% may be tolerated by the application using this filter. For this level of quality reduction, one may apply the five voltage levels of 600, 650, 700, 750, and 800 mV to the PEs as shown in Fig. 1e. Also, one voltage switch box is used for each 2 × 1 voltage island.

Given the allowed reduced output quality, we can use lower operating voltages for the FUs by utilizing an optimization approach which, for example, has as its objective the minimization of the sum of all PE operating voltages. This application requires that 15 PEs become involved in the calculation. For the accurate calculation where full operating voltage of 800mV is applied, the average V_{DD} of PEs is 800mV. In the case of tolerable reduced quality of 50%, the average V_{DD} of the PEs becomes 680mV which is 15% lower than that of the exact case. Now, given the fact that the dynamic and leakage power components as well as lifetime/reliability degradation mechanisms depend super-linearly (even exponentially in some cases) on the operating voltage, the reduction would lead to a significant improvement for the lifetime/reliability. Also, with this voltage reduction, the CGRA dissipates 27% less energy.

B. Proposed CGRA Architecture

The general architecture of the proposed CGRA, which has two major differences with the conventional CGRAs, is depicted in Fig. 2. The first major difference is availability of two separate operating voltages, one for the I/O and the other for the core of the PEs. The former could be the nominal voltage of the technology. The input for the switch power boxes, realized by MOSFET switches, are different VOS voltage levels and the output of these switch power boxes are the core voltages of the voltage islands. Using a single power switch box for each PE provides us with the flexibility of assigning any voltage level to any PE. This obviously increases the efficiency of the proposed CGRA core architecture in terms of energy consumption and lifetime/reliability improvements.

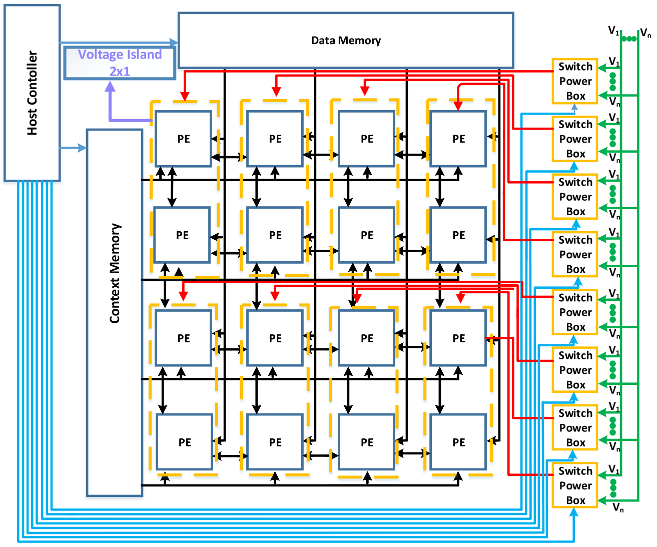


Fig. 2. The overall architecture of proposed CGRA.

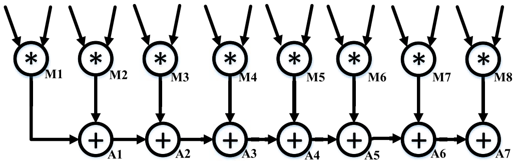


Fig. 3. Data flow graph of 8-tap FIR filter [26].

It, however, has the disadvantage of increasing area and power overheads and voltage level routing complexity.

In our suggested technique, by clustering the PEs to, *e.g.*, 2 × 1 and 1 × 3, the overheads and complexity are reduced. While, in this work, we considered the same sizes for the islands, in general the sizes could be different. The importance of using as few number of voltage levels as possible in multiple supply voltage systems for having a less complex (overhead) power network has been emphasized in some other works (e.g., [27]–[29]). Moreover, specifically in [28], it is stated that placing blocks with the same voltage together saves power routing resources, simplifies power planning, and reduce IR drop. Also, in [29], a layout plan where modules with same voltage are placed together for reducing the complexity of the power network is recommended. The power supply rails are routed using the top interconnect layers, where the thickness, width and pitch of these wires are larger than the wires in the lower layers. These voltages are connected to the

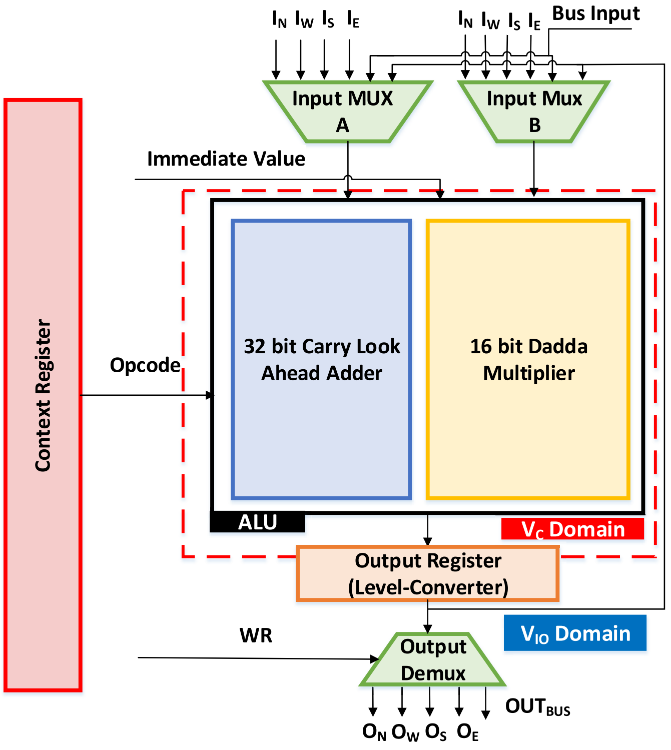


Fig. 4. The proposed architecture of each process element (PE).

supply contacts of the transistors using several via electrical connections which brings down the voltage level from the higher layers down to the contact. The vias, however, consume areas lowering the chip wiring efficiency which may cause some chip area increase. As an example, in [30], it has been shown that employing dual supply voltage leads to 15% area increase. Obviously, the larger is the number of the vias, the higher is the via blockage. [31] states that the via blockage can be up to 50% on metal 1 layer. The model in [31] has been used in the literature for calculating the impact of the via blockage on lowering the wiring efficiency of the chip (see, e.g., [32]). Since in the proposed structure, each island requires only one supply voltage level, instead of one set of vias for each PE, one set is required for connecting the proper voltage level to all the PEs in the same island. Therefore, it may be concluded that the number of via sets and, hence, via blockage are inversely proportional to the sizes of the islands.

The proposed CGRA structure works based on determining the required minimum accuracy of each DFG node for satisfying the output quality constraint. Then, the node is mapped on a PE whose voltage level is determined by the accuracy of the node. In mapping a node on a PE, the clustering of the nodes based on voltage islands are performed such that the objective function is optimized. In this work, four voltage island sizes of 2 × 1 (two rows by one column), 1 × 3 (one row by three columns), 2 × 2 (two rows by two columns) 3 × 2 (three rows by two columns) are considered.

The internal structure of each PE and the ALU and I/O voltage domains (V_C and $V_{I/O}$, respectively) are shown in Fig. 4. For this work, PEs are considered to comprise of a 32-bit Carry Look Ahead adder and a 16-bit Dadda multiplier. The operating voltage of the said ALU (consists of one adder and one multiplier) was determined by the core voltage. The output of the ALU is connected to a 32-bit register level-converter (consisting of level-converter flip-flops [33]). The inputs of the register level converter are in the V_C voltage domain whereas the output of this register is in the I/O voltage domain. For connecting the PE to its neighboring PEs, there are two 5 to 1 multiplexers for the input of the ALUs of the PEs and one 1 to 5 demultiplexer for sending the output of the PEs to its four (up, down, left, right) neighbors or the output bus. All of these multiplexers and demultiplexers and the bus connections work based on $V_{I/O}$. The core voltage level for each voltage island is determined by the host processor from a LUT which has the mapping for the DFG operations and the corresponding voltage levels for each island. Due to possibility of long idle times (*e.g.*, no operation mapped to the PE), to alleviate the energy dissipation, power gating switches on V_c and V_{IO} of each PE is considered.

V. Mapping Formulation

For the optimum voltage level setting and mapping an application to the accuracy-configurable CGRA, we formulate the determination of the accuracy level of each DFG node (and the voltage level of the PE that will be performing that operation) and the physical mapping of the DFG nodes to specific PEs on the CGRA fabric as an optimization problem. In the proposed formulation, all nodes mapped on a set of PEs in an island have the same operating voltage level. The scheduling and binding are two NP-complete problems [34]. For this purpose, a set of linear constraints, which will be explained in the following subsections, are included in the optimization framework.

A. Accuracy Levels of Nodes

As mentioned before, in the proposed architecture, the degree of inaccuracy of each PE is set based on its applied operating voltage. In the proposed formulation, a mapping between the output accuracy constraint and the operating voltage levels of the operations are established. The output error of a DFG is a linear combination of the operation errors [35]. Hence, the output error is determined based on the amount of the generated error by each approximate operation reaching to the output of the DFG. The error distribution of each operation in a DFG is independent from those of other operations unless both operations have the same input data and same hardware implementation.

The error propagations show strong structural correlation which should be used to model the output error [35]. In the work of [35], for obtaining the output error, an error sensitivity parameter (*ES*) was introduced. The error sensitivity shows the impact of the accuracy of a node on the output when the other nodes are precise. Hence, the error sensitivity of the i^{th} node (denoted by $ES_{i,o}$) was defined as $ES_{i,o} = \frac{\epsilon_{i,o}}{\epsilon_i}$ where $\epsilon_{i,o}$ and ϵ_i were the error distance of the DFG output and the i^{th} node, when only the i^{th} node was in the approximate operating mode. Based on this error sensitivity, the variance of the output (denoted by $\upsilon(\epsilon_o)$) was obtained from [35]

$$\upsilon(\epsilon_o) = \sum_{i\in\{DFG\ \ Nodes\}} ES_{i,o}^2 \cdot \upsilon(\epsilon_i) \tag{4}$$

where $\upsilon(\epsilon_i)$ was the output variance of the i^{th} node. Now, by considering the variance as the error metric and employing (4), one may formulate the problem of determining the operating voltage level of the DFG nodes under the predefined expected variance (minimum quality) denoted by υ_{EXP}.

Based on the above definitions, for formulating the accuracy level determination, we consider a binary variable ($x_{i,j}$). When the j^{th} operating voltage level (where $1 \leq j \leq L$; L is the number of the considered voltage levels) is considered for the i^{th} DFG node, $x_{i,j} = 1$. Hence, the output expected quality constraint is defined by

$$\sum_{\forall i\in\{DFG\ \ Nodes\}} \sum_{j=1}^{L} ES_{i,j,o}^2 \cdot \upsilon\left(\epsilon_{i,j}\right) \cdot x_{i,j} < \upsilon_{EXP} \tag{5}$$

where $ES_{i,j,o}(\epsilon_{i,j})$ shows the error sensitivity (error distance) of the i^{th} node in the j^{th} operating voltage level. The values of $\epsilon_{i,j}$ and $ES_{i,j,o}$ are obtained before the mapping process.

B. Mapping of the Nodes on PEs

In the considered CGRA, each PE is connected to all four neighbors (except the ones placed in the borders). Therefore, the mapping constraint should guarantee that each two adjacent nodes in DFG, are neighbors in the CGRA. By considering a binary variable ($b_{i,r}$) for showing the mapping of the i^{th} DFG node on the r^{th} PE, the mapping process may be formulated as

$$\forall_{i\in\{DFG\ \ Nodes\}, r\in\{CGRA\ \ PEs\}} \left(\sum_{i'\in\{adjecents\ \ of\ \ i^{th}\ \ node\}} \sum_{r'\in\{adjecents\ \ of\ \ r^{th}PE\}} \left(b_{i',r'}\right) \geq b_{i,r} \times E_i \right) \tag{6}$$

where E_i is the degree of the i^{th} DFG node. In (6), when $b_{i,r}$ is one, all the adjacent nodes of i^{th} DFG node must be mapped onto E_i neighbors of r^{th} PE. In addition, to map each node on only one PE, the following formula should be used.

$$\forall_{i\in\{DFG\ \ Nodes\}} \left(\sum_{r\in\{CGRA\ \ PEs\}} b_{i,r} = 1 \right) \tag{7}$$

Note that since the output of each PE in the considered CGRA structure is connected to only four neighbors, when the fanout of a DFG node is larger than four, the fanout could be reduced by inserting NOP (no operation) nodes (other approaches may be taken as well).

C. Guaranteeing the Same Voltage Level for the Nodes in an Island

In the optimization formulation, for each island, the possible mapping of Q nodes (Q is the number of the PEs in each island) on Q PEs should satisfy the condition of the same voltage level. Therefore, for each possible mapping for an island, under each operating voltage level (*e.g.*, the j^{th} voltage level), we propose to employ the set of the following inequalities:

$$\forall_{(i,r)\in\{possible\ \ mapping\ \ of\ \ Q\ \ Nodes\ \ on\ \ Q\ \ PEs\}} \big(w_j \leq 1 - b_{i,r} + x_{i,j}\big) \tag{8}$$

Here, w_j in (8) is a binary variable that is 1 if the chosen operating voltage level index for all the mapped nodes on the PEs of the island is j. The possible mapping of I nodes on Q PEs is equal to Q-permutation of I. Now, (8) should be defined for all the voltage levels and in at least one operating voltage level, the w should be one. Hence, the following inequality should be employed to meet this constraint:

$$\sum_{j=1}^{L} w_j \geq 1 \tag{9}$$

Note that the set of (8) and (9) should be described for each voltage island of the CGRA.

D. Objective Function

The goal of this optimization problem is to reduce the operating voltage levels of the nodes improving the energy consumption and lifetime/reliability of the CGRA. Here, we consider the constraint of minimizing the summation of the voltage levels of the DFG nodes as the objective function as

$$\sum_{i\in\{DGF\ \ Nodes\}} \sum_{j\in\{Voltage\ \ Levels\}} VDD_j \times \alpha \times x_{i,j} \tag{10}$$

where VDD_j indicates the corresponding operating voltage level of the j^{th} operating voltage level index and α is the weight coefficient. In this work, we consider the weight value of 0.8 (1) for the case of adder (multiplier). The reason for considering a larger value for the multiplier is its more suffering from the aging mechanisms compared to that of the adder [8].

Now, by employing VI as the objective function, and inequalities (5) to (9) as the constraints, the mapping process on the proposed CGRA can be formulated. Note that in the mapping process for an application in the CGRA, we may end up with idle nodes or the nodes operating with lower voltages experiencing different voltage threshold drifts by the PEs. This provides us with the possibility of remapping of the DFG on the CGRA during the lifetime to distribute the stress more uniformly for prolonging the lifetime even further.

VI. Results and Discussion

A. Simulation Setup

To assess the efficacy of the proposed CGRA structure, 6 × 6 CGRAs with 2 × 1, 1 × 3, 2 × 2 and 3 × 2 voltage islands have been considered. For each CGRA, we have evaluated the VOS method under two (600mV and 800mV), three (600mV, 700mV and 800mV) and five (600mV, 650mV, 700mV, 750mV, and 800mV) different operating voltage levels. The nominal voltage was 800mV.

TABLE III

THE NUMBERS OF THE TOTAL NODES AND EACH OPERATION TYPE FOR EACH BENCHMARK

Benchmark	Multiply	Add	Total
FIR	8	7	15
MMM	8	7	15
SMT	9	8	17
SHP	9	8	17

For extracting the design parameters of the PEs, the components of the PEs were described by the Verilog HDL, and synthesized using Synopsys Design Compiler using a 15nm technology file [36]. For extracting the impact of the voltage scaling on the design parameters, we have characterized the 15-nm technology file by employing the Cadence Encounter Library Characterization under the considered operating voltage levels. Therefore, the design parameters have been extracted by Synopsys Design Compiler under these characterized technology libraries.

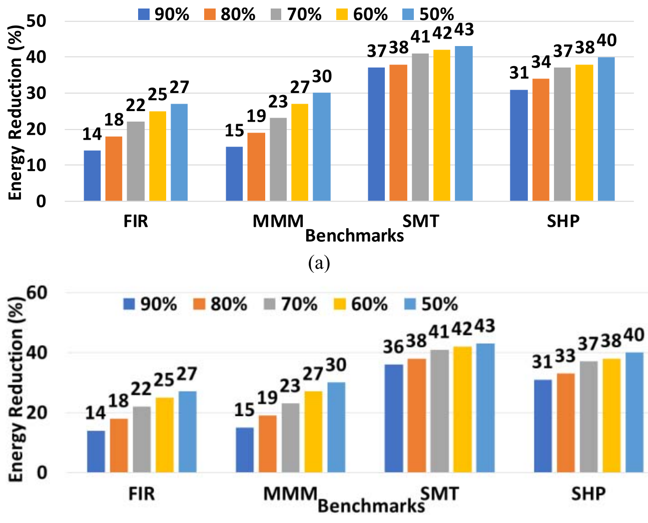
In this work, without loss of generality, it has been assumed that the arithmetic unit of each PE contained one adder (32-bit CLA) and one multiplier (16-bit Dadda). For extracting the output quality degradation due to the VOS, the post-synthesis gate-level of these two components were simulated by ModelSim HDL simulator. The timing information of the gates in these simulations for each operating voltage level was extracted by Synopsys Design Complier. The accuracy of the arithmetic operations was obtained by injecting many randomly generated input operands. Also, in the studied CGRA architectures, one and three cycles considered for the add and multiply operations, respectively, when considering the same clock frequency at the considered voltage levels. The clock frequency was obtained such that that the PEs of the CGRA did not have any output error at the nominal (800mV) operating voltage level.

To explore the effectiveness of the proposed method, four benchmarks from different application domains were considered. The benchmarks set included 8-Tap FIR filter, 8 × 8 Matrix Multiplication (MMM), Smoothing filter (with 3 × 3 filter) (SMT), and Sharpening filter (with 3 × 3 filter) (SHP). These applications contain only the add and multiply operations. The details of the number of the operations in each application have been reported in Table III.

For mapping these applications on the CGRAs, we employed the mapping approach proposed in Section V where the ILP formulations was described by the python language and solved by Gurobi LP solver [37]. In these studies, five different output quality (approximation) levels, 90%, 80%, 70%, 60%, and 50% have been studied.

B. Results

1) Energy Reduction: The results for the energy reduction of the CGRA under two cases of 2 × 1 and 1 × 3 voltage islands for different benchmarks and quality constraints compared to that of the CGRA operating at the nominal operating voltage level (no quality loss) have been presented in Fig. 5. This CGRA architecture which was used in this study is a common basic CGRA architecture utilized in some other recent works (see, *e.g.*, [8], [38], [39]). Since the main idea of this paper has been on proposing voltage overscaling (VOS) and voltage islanding schemes for improving the reliability/lifetime and energy improvements of CGRAs, we considered the basic CGRA architecture for the study. Without any limitation, the idea may be well applied to other CGRA architectures (which there exist a few of them) where the amounts of improvements may be different for different CGRA architectures.



(a)

(b)

Fig. 5. The energy reduction of different benchmarks under different quality constraints mapped on CGRA with (a) 2 × 1 and (b) 1 × 3 voltage islands in the case of considering five voltage levels has been considered.

As the results show, lowering the minimum acceptable output quality causes more energy reduction. In the case of 2 × 1 voltage island (Fig. 5(a)), the FIR and MMM benchmarks enjoyed from more energy reductions (∼1.9× and 2×, respectively) by reducing the minimum acceptable quality from 90% to 50% compared to the cases of the SMT and SMP benchmarks (∼1.2× and ∼1.3×, respectively). From these results, less sensitivities of these two image processing applications to imprecise computing may be concluded. Hence, for a given specified minimum output quality, these applications may be assigned lower operating voltages for their PEs. The reductions in the energies originate from being able to assign lower operating voltages to some PEs determined by their acceptable levels of the inaccuracy. Obviously, the operating voltage reduction could be more in the case of the considered image processing applications. To demonstrate this, as an example, the operating voltage levels of the PEs for the minimum acceptance quality of 90% have been indicated for the two cases of MMM and SMT applications in Fig. 6.

As the reported energy gains in Fig. 5(a) reveal, the maximum gains belonged to the SMT benchmarks (from 37% to 43%), while the FIR benchmark had the lowest gains (from 14% to 27%). For the studied benchmarks, the average of the energy gains was from 24% to 35% when the acceptable

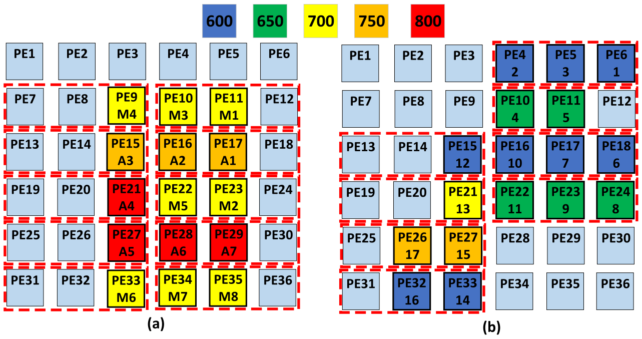


Fig. 6. Mapping of a) MMM b) SMT benchmarks on a 6 × 6 CGRA under the minimum quality of 90%.

TABLE IV

ENERGY REDUCTION (%) OF DIFFERENT BENCHMARKS UNDER DIFFERENT MINIMUM OUTPUT QUALITIES, VOLTAGE ISLAND SIZES AND OPERATING VOLTAGE RESOLUTIONS

Voltage Island	Minimum Output Quality →	90%			80%			70%			60%			50%		
	Voltage Levels→ Benchmarks↓	5	3	2	5	3	2	5	3	2	5	3	2	5	3	2
3×2	FIR	13	11	0	16	15	6	22	18	12	25	21	15	26	22	18
	MMM	14	11	0	17	15	6	22	18	12	26	24	20	30	26	26
	SMT	35	30	23	38	37	31	40	39	33	40	40	36	43	41	39
	SHP	30	21	15	33	28	26	34	34	31	38	38	33	39	39	33
2×2	FIR	13	13	0	17	15	6	21	19	12	25	21	15	27	22	18
	MMM	15	13	0	18	15	6	22	18	12	25	23	23	29	26	26
	SMT	35	31	26	38	37	31	41	39	33	41	40	36	43	41	39
	SHP	31	24	15	33	28	26	36	34	31	38	38	33	39	39	33
1×3	FIR	14	13	0	18	17	6	22	19	12	25	21	15	27	23	18
	MMM	15	13	0	19	15	6	23	18	12	27	24	23	30	28	26
	SMT	36	33	26	38	37	31	41	39	33	42	40	36	43	41	39
	SHP	31	24	15	33	29	26	37	35	31	38	38	33	40	39	33
2×1	FIR	14	13	0	18	17	6	22	19	12	25	21	15	27	23	18
	MMM	15	13	0	19	15	6	23	18	12	27	24	23	30	28	26
	SMT	37	33	26	38	37	31	41	39	33	42	40	36	43	41	39
	SHP	31	24	15	34	29	26	37	35	31	38	38	33	40	39	33

minimum output quality reduced from 90% to 50%. The results in Fig. 5(b), reveals that the energy reductions in the case of 1 × 3 voltage island are almost the same as those in the case of 2 × 1 voltage island. Since both voltage island sizes provide about the same energy reductions, the 1 × 3 islands are preferred thanks to the reduced overhead (see subsection IV.B). It should be mentioned that both power and area overheads of the power switch boxes themselves are (negligibly) small. The fact that the same power reduction gains were achieved was due to using a 6 × 6 CGRA platform which had more PEs than the number of operations required for the application DFGs. Of course, unused PEs should be power gated to prevent any leakage power consumption.

For the results presented in Fig. 5, five operating voltage levels for PEs were considered. To reduce the overhead of generating and distributing the operating voltage levels, a few number of levels may be used. Reducing the voltage levels diminishes the flexibility of using different approximation levels for the PEs limiting the opportunity to lower the energy consumption for a given output quality. To study the impact, we studied the energy reduction gain of the proposed approach by considering the cases of two and three voltage levels as well. All the energy reductions are compared to the reference case of applying the nominal voltage to all the PEs (100% output quality). The results, which are reported in Table IV show that in the two cases of the FIR and MMM benchmarks, the energy reduction gains have been affected more. This has been demonstrated further in Fig. 7 comparing the loss of the energy reduction when the number of operating voltages levels decreases from five to two. In the worst case, when the minimum tolerable output quality was 90%, no energy reduction was possible for the FIR and MMM benchmarks. This means that even if we set the voltage for one of the islands below the nominal voltage, the output quality becomes below 90%. For lower acceptable minimum output qualities, however, there is a chance for using islands with lower operating voltages. As the results show, for all benchmarks, decreasing the minimum output quality below 90%, provided us with some energy reductions. As expected, the gain is proportional to the amount of quality reduction. In addition, the results indicate that, by increasing the island size, the energy improvements for the cases of 2 × 2 and 3 × 2 compared to the those of the cases of 1 × 3 and 2 × 1 for most of the benchmarks and qualities, have reduced. For example, by increasing the island size from 2 × 1 to 3 × 2, on average (maximum) 4% (15%) reduction in the energy improvement is achieved.

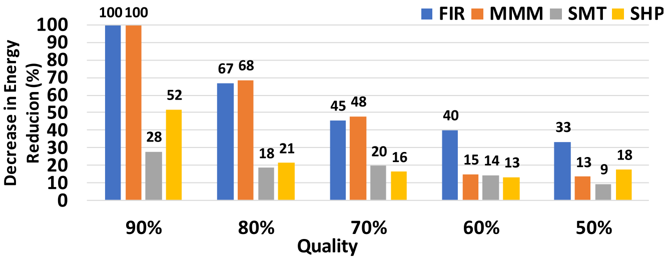
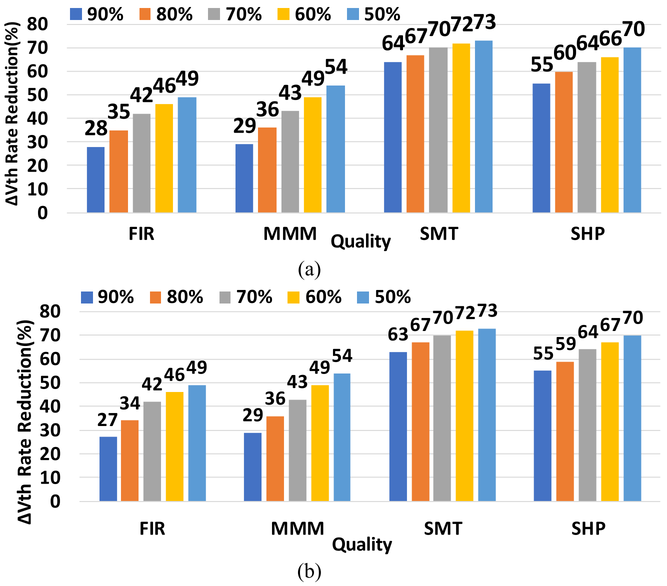


Fig. 7. The decrease of the energy reduction in the case of 1 × 3 voltage island when the number of operating voltage levels is lowered from five to two.

Finally, we notice that decreasing the voltage levels from five to three for the considered benchmarks does not yield considerable loss of the energy reduction of the VOS technique while we may reduce the associated overheads (*e.g.*, two less switches).

2) Lifetime/Reliability Improvement: First, it should be mentioned that the reliability degradation of a circuit, in a general sense, for most of associated mechanisms, is a strong function of the operating voltage (see, *e.g.*, [6]). Here, we have only concentrated on NBTI-induced threshold voltage change as a measure of the CGRA aging. The results which are for the cases of the 2 × 1 and 1 × 3 voltage islands for different minimum output qualities are plotted in Fig. 8. To obtain the threshold voltage change ($\Delta V_{th,NBTI}$), (3) was used. As the results reveal, the VOS approximate computing technique leads to reducing the aging effect when some output quality degradation is acceptable. The lower the tolerable output quality is, the lesser the threshold voltage change rate



(a)

(b)

Fig. 8. The threshold voltage change rate reduction for different benchmarks for different minimum acceptable qualities and five operating voltage levels in the case of (a) 2 × 1 (b) 1 × 3 voltage islands.

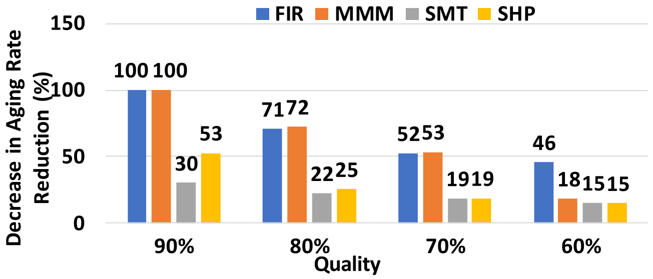


Fig. 9. Decrease of the aging rate improvement in the case of 1 × 3 voltage island when the number of operating voltage levels is lowered from five to two.

TABLE V

AGING RATE REDUCTION (%) OF DIFFERENT BENCHMARKS FOR DIFFERENT MINIMUM OUTPUT QUALITIES, VOLTAGE ISLAND SIZES AND OPERATING VOLTAGE RESOLUTIONS

Voltage Island	Minimum Output Quality →	90%			80%			70%			60%			50%		
	Voltage Levels→ Benchmarks↓	5	3	2	5	3	2	5	3	2	5	3	2	5	3	2
3×2	FIR	26	21	0	30	29	10	41	33	20	46	38	25	47	40	30
	MMM	28	21	0	31	29	10	43	33	20	48	41	35	54	45	45
	SMT	61	53	39	66	64	52	70	67	57	70	69	61	73	71	65
	SHP	54	38	26	58	49	44	60	59	52	66	66	57	68	67	57
2×2	FIR	26	24	0	32	28	10	40	32	20	46	40	25	49	45	30
	MMM	29	24	0	35	28	10	39	32	20	44	40	40	51	45	45
	SMT	62	55	44	67	64	52	70	67	57	71	69	61	73	71	65
	SHP	55	42	26	58	49	44	62	59	52	67	66	57	67	67	57
1×3	FIR	27	24	0	34	31	10	42	35	20	46	38	25	49	41	30
	MMM	29	24	0	36	29	10	43	34	20	49	41	40	54	48	45
	SMT	63	57	44	67	64	52	70	67	57	72	69	61	73	71	65
	SHP	55	42	26	59	51	44	64	61	52	67	66	57	70	67	57
2×1	FIR	28	24	0	35	31	10	42	35	20	46	38	25	49	41	30
	MMM	29	24	0	36	29	10	43	33	20	49	41	40	54	48	45
	SMT	64	57	44	67	64	52	70	67	57	72	69	61	73	71	65
	SHP	55	42	26	60	51	44	64	61	52	66	66	57	70	67	57

will be. In the considered benchmarks, the highest improvement belonged to the SMT benchmark which the Vth change rate reduction was improved from 63% to 73% in the case of 1 × 3 voltage island. This is a direct consequence of using lower operating voltage levels for the PEs in the case of this benchmark. Also, the aging rate reductions for FIR, MMM, and SHP were from 27% to 49%, 29% to 54%, and 55% to 70%, respectively. The average of the aging rate reductions was about 62% (44%) when the minimum output quality constraint was 50% (90%). This is the one of the key advantages of using the proposed VOS approximate computing technique.

As was observed in the previous subsection, having fewer operating voltage levels, lowers the opportunity for using overscaled operating voltages for the PEs for a given minimum output quality. Similar to the case of the energy reduction gain, this lowers the aging rate improvement opportunity. As an example, Fig. 9 shows the decrease in the aging rate reduction trend versus the minimum output quality when the number of operating voltage levels changes from 5 to 2.

Table V shows the aging rate reduction of different benchmarks for different minimum output qualities, voltage island sizes, and operating voltage resolutions. As the results show, for most of the benchmarks and qualities, the 3 × 2 voltage island has the lowest aging rate reduction. For qualities below 70%, however, in some benchmarks, the 2 × 2 island case has the minimum aging rate reduction. By increasing the island size from 2 × 1 to 3 × 2, a maximum of 14% decrease in the aging rate was obtained. In addition, the figures indicate that decreasing the minimum output quality improves the aging rate reduction.

3) Folding Approach for Reliability Improvement: In addition to reducing the operating voltage, here, we consider two cases for further improving the lifetime by redistributing the impact of aging mechanisms on the PEs. These cases deal with exchanging the operations (stresses) on PEs. Before explaining these cases, consider the CGRA with 2 × 1 islands shown in Fig. 10(a) with a folding line in the middle. If some PEs in one side of the folding line have, *e.g.*, lower voltages (or power-gated), every once a while, their operations (state) may be exchanged with the PEs in the other side which suffer from (more) aging rate. This would allow for more uniform aging rate distributions of the PEs reducing the aging rate of the CGRA. For instance, in this CGRA, the pairs of PE1 and PE4 (PE5 and PE8) are candidates for the exchange process. Of course, their neighbors also should be exchanged such that the overall optimized mapping is not affected considerably (see the figure).

Now, let us describe the cases in a general form. In the first case, there is an active PE in one side while there is another idle (power-gated) PE placed symmetrically with respect to the folding line in the other side. By moving the operation of the active PE (with the assigned operating voltage level) to the idle PE, the delay drift of the active PE is reduced. In this case, both PEs will be under stress for the half of the lifetime duration. In the second case, there is two active PEs with two different operating voltages. Exchanging them as well as their corresponding neighbors with their related roles with the other

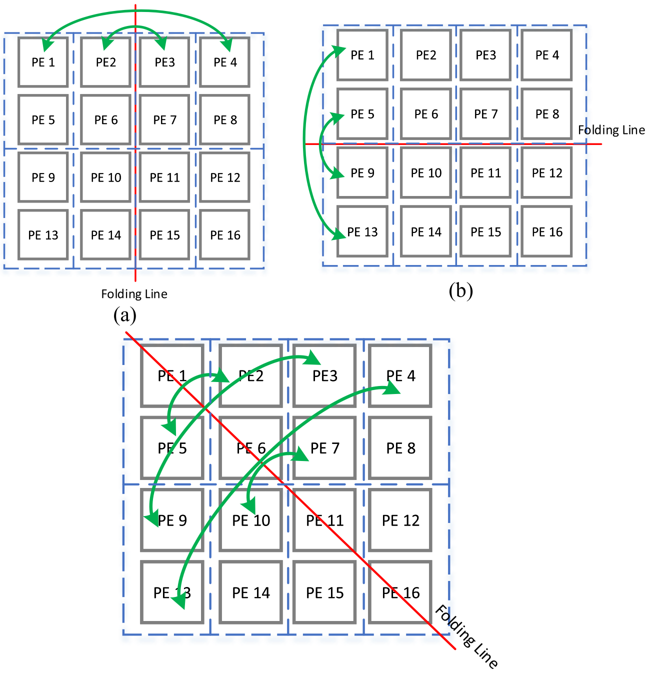


Fig. 10. Folding a 4 × 4 array of PEs with respect to a (a) vertical, (b) horizontal, and (c) diagonal folding lines passing through the middle of the array.

side lead to a more balanced aging rate for both sides which improves the CGRA lifetime.

The exchange cases (remapping) described above has some energy and runtime overheads, and hence, should be performed every once a while depending on the application in hand. Also, the efficacy strongly depends on the number of used and idle PEs and the assigned voltage levels to the islands. In this work, the mapping/re-mapping and scheduling processes for every application and given minimum output quality are performed in the offline phase (statically). The results of the mapping/re-mapping process are the context words which are utilized for programming the CGRA in different runtime slots (invocations). Proper context words, which are stored in the context memory, are loaded in the context registers used for configuring the PEs. On the other hand, since the folding process is simple having an overhead similar to the reconfiguration of the PEs, it is performed online.

The place of the folding line and the exchange direction should not change the adjacency of the PEs to considerably lower the overhead of the mapping process. This way only the direction of the neighbors in the context word of the PEs is changed. Also, it should be mentioned that there might be cases such as high minimum acceptable qualities and the use of most of the PEs of the CGRA where no room will be left for optimizing the lifetime further using this technique. As shown in Fig. 10, three possible folding lines could be considered including vertical (Fig. 10(a)), horizontal (Fig. 10(b)), and diagonal (Fig. 10(c)). In the case of the diagonal folding, due to passing of the line through the middle of some the PEs, their operations are not exchanged with other PEs. Also, there is another possible diagonal folding line passing from top-right to bottom-left which is not shown here.

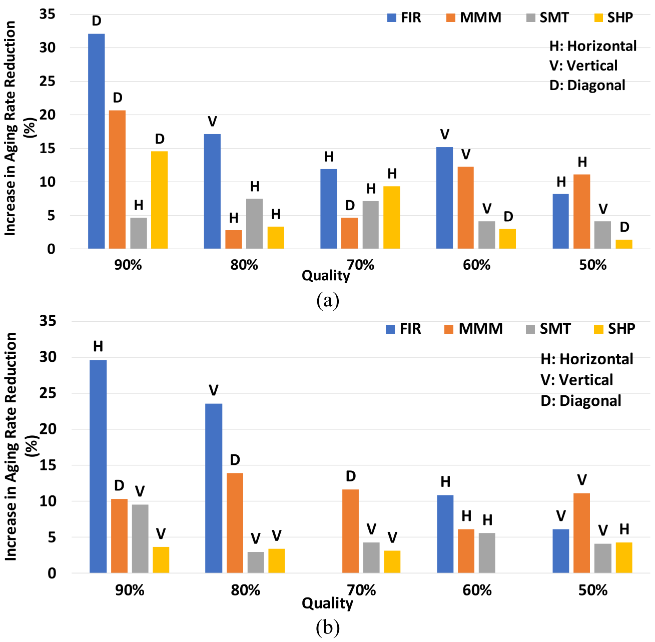


Fig. 11. The increase in aging rate reduction (%) for all the benchmarks under different minimum output qualities in the case of five voltage levels and (a) 2 × 1, (b) 1 × 3 voltage islands.

Since each PE uses more than one operating voltage, for predicting the BTI impact, (3) might not be used. In this case, for modeling the voltage threshold drift of the PMOS transistor, we have used the model proposed in [40] given by

$$\Delta V_{th} = \left(A_1^{\frac{1}{a}}\Delta t_1 + \cdots + A_n^{\frac{1}{a}}\Delta t_n\right)^{a} \tag{11}$$

where Δt_k shows the amount time that the transistor in the k^{th} voltage level, a is equal to 0.173 in the case of NBTI, and A_k is the technology parameter, which extracted from (1) and defined by

$$A_k \cong A e^{-\frac{\kappa}{\theta}} E_{OX}^{\gamma} d f^{\beta} \tag{12}$$

Here, the E_{OX} is an operating voltage dependent parameter, used for considering the operating voltage reduction impact on the lifetime of the system.

In Fig. 11, the improvement of the aging rate reduction for all the benchmarks for different minimum output qualities in the case of five voltage levels and 1 × 3 and 2 × 1 voltage islands have been presented. In these figure, H, V, and D stand for horizontal, vertical, and diagonal folding lines, respectively. For each benchmark and quality constraint pair, the gain for the case of the folding line resulting in highest improvement was selected. As the results indicate, the proposed folding approach could lead to about 30% improvement in the aging rate reduction in the case of the considered benchmarks. Also, by reducing the minimum output quality, the efficiency of the proposed enhancement approach in some cases decreases. This may be attributed to the fact that when the minimum output quality decreases, the assigned voltages to more PEs are reduced limiting the gain of the exchange process. The results show that the proposed folding approach

TABLE VI

AGING RATE REDUCTION IMPROVEMENT (%) USING THE FOLDING TECHNIQUE FOR DIFFERENT BENCHMARKS, DIFFERENT MINIMUM OUTPUT QUALITIES, VOLTAGE ISLAND SIZES, AND OPERATING VOLTAGE RESOLUTIONS (IMP: IMPROVEMENT, DIR: FOLDING DIRECTION)

Voltage island	Minimum Output Quality →	90%						80%						70%						60%						50%					
	Voltage Levels→ Benchmarks↓	5		3		2		5		3		2		5		3		2		5		3		2		5		3		2	
		Imp	Dir	Imp	Dir	Imp	Dir	Imp	Dir	Imp	Dir	Imp	Dir	Imp	Dir	Imp	Dir	Imp	Dir	Imp	Dir	Imp	Dir	Imp	Dir	Imp	Dir	Imp	Dir	Imp	Dir
3×2	FIR	35	D	48	H	NA	D	7	H	31	H	110	H	15	D	9	D	30	H	7	V	11	V	60	V	15	D	10	D	47	V
	MMM	32	V	48	H	NA	D	29	H	31	H	190	V	16	H	27	H	60	H	10	H	10	H	20	H	11	H	0	-	0	-
	SMT	3	D	11	V	18	H	6	D	9	H	0	-	7	H	6	V	11	V	9	H	6	H	8	D	4	D	3	V	6	D
	SHP	7	V	16	H	4	D	5	H	6	H	20	V	7	H	5	V	10	H	5	D	3	D	2	V	7	H	1	V	0	-
2×2	FIR	19	V	21	V	NA	H	19	H	21	V	110	H	15	D	0	-	45	H	0	-	0	-	60	V	10	D	0	-	17	V
	MMM	28	H	33	V	NA	H	14	D	14	H	100	D	15	V	9	H	80	D	9	H	5	V	8	D	8	V	7	H	7	H
	SMT	8	H	11	V	7	H	7	V	5	V	8	H	6	V	9	D	14	D	4	H	7	D	7	H	4	V	4	V	8	V
	SHP	7	H	10	H	23	H	7	H	10	H	18	D	5	V	7	H	8	H	4	H	5	H	5	H	4	H	7	H	11	V
1×3	FIR	30	H	8	V	NA	H	24	V	45	D	140	H	0	-	9	V	60	H	11	H	11	V	64	H	6	V	24	H	30	H
	MMM	10	D	54	H	NA	H	14	D	28	D	140	H	12	D	24	D	50	V	6	H	12	V	18	H	11	V	0	-	22	D
	SMT	10	V	0	-	7	V	3	V	5	V	10	V	4	V	4	V	11	D	6	H	4	H	8	D	4	V	4	D	5	H
	SHP	4	V	17	D	23	D	3	V	4	D	16	H	3	V	5	V	0	-	0	V	8	D	11	D	4	H	4	V	12	D
2×1	FIR	32	D	21	H	NA	V	17	V	16	H	140	V	12	H	6	D	85	V	15	V	13	H	64	V	8	H	2	H	30	V
	MMM	21	D	13	H	NA	V	3	H	31	D	190	V	5	D	15	H	30	V	12	V	17	H	13	H	11	H	8	H	16	H
	SMT	5	H	2	D	11	H	7	H	6	H	2	H	7	H	4	D	9	H	4	V	10	V	10	D	4	V	4	H	5	D
	SHP	15	D	12	H	50	D	3	H	6	V	5	H	9	H	11	H	12	V	3	D	2	H	16	V	1	D	4	V	9	H

provided, on average, 49.5% (15.5%) and 65% (6.2%) aging rate reduction (aging rate reduction improvement) in the cases of 90% and 50% minimum output quality constraint. In the case of the FIR benchmark for the quality of 70% when the voltage island size of 1 × 3 was used, the improvement of the folding approach was zero. Since the folding impact strongly depends on the bindings of the nodes to the PEs, there is no clear dependence between the aging rate reduction improvement and the voltage island size increase.

The impact of having different number of operating voltage levels on the efficiency of the folding process also was studied. The results which have been reported in Table VI contain the aging rate reduction improvement for different minimum output qualities and voltage island sizes with and without using the folding scheme. As the results show, by exploiting the folding approach, in almost all the cases, the aging rate reduction of the proposed structure has been improved. It should be mentioned that even in the cases where there was not any aging rate reduction when the folding approach was not used (*e.g.*, the FIR benchmark for the minimum output quality of 90% when two voltage levels were used), the folding reduces the aging rate. Given the fact that without folding, improvement was zero yielding infinity improvement (meaningless) and hence NA notation has been used for these cases in the table. Also, the proposed folding approach resulted in larger improvement in the case of two operating voltages. This may be attributed to the large difference between the operating voltage levels in the case of two operating voltage levels. On average, improvements of 12% and 34% were achieved in the cases of three and two operating voltage levels, respectively.

In Table VII, we have compared the improvements of the works suggested here and the work of [3] compared to those of the conventional exact CGRA for the benchmarks. In the case of the [3], we have considered only the minimum output quality level of 90% while for the proposed structure in this work, the results for an additional minimum accuracy level of 50% have been included. In addition to the power reduction, the table contains the lifetime improvement due to the reduction of the electric field (supply voltage) for the proposed approach. The results were obtained considering five voltage levels. For the case of 90%, in our approach, we cannot lower the voltages of most of the PEs much due to the fact that the MSBs are also affected by reducing the voltage considerably. This lowers the power improvement compared to that of [3]. Finally, the table contains the area overheads of both approaches where for the proposed method in this paper is negligibly low.

TABLE VII

THE IMPROVEMENTS OF THE PROPOSED METHOD AND THE ONE SUGGESTED IN [3] COMPARED TO THOSE OF THE CONVENTIONAL EXACT CGRA

	Benchmarks	Quality Constraint	Power Improvement	Lifetime Improvement	Area Overhead
[3]	FIR	90%	34%	Not Applicable	~3.3×
	MMM	90%	34%		
	SHP	90%	33%		
	SMT	90%	33%		
Proposed (2×1 Voltage Island)	FIR	90%	14%	28%	~0.2%
		50%	27%	49%	
	MMM	90%	15%	29%	
		50%	30%	54%	
	SHP	90%	37%	64%	
		50%	43%	73%	
	SMT	90%	31%	55%	
		50%	40%	70%	

4) Overheads of the Proposed Structure: In our structure, there are, *e.g.*, up to three additional control signals (for up to five voltage levels) which selects from the voltages that should be applied to each island. The values of the voltage control signals (determined beforehand for a given accuracy level) along with other controlling signals are stored in the context memory and are routed to the CGRA structure through proper interconnect layers. Obviously, the signal values are different in different invocations of the CGRA structure. Since the additional signals form a small fraction of all the signals, the overhead would be negligible. Similarly, the area and power overheads for the additional switch power boxes in the case of 2 × 1 voltage island supporting five operating voltage levels (800mV, 750mV, 700mV, 650mV, and 600mV) in the 15nm technology were estimated to be about 0.2% and 0.0003%, respectively. These overheads were extracted based on the power consumption and the delay of PEs obtained by synthesizing the PE using Synopsys Design Complier. By using the dynamic power consumption of the PE, the overall capacitance of the PE was estimated. In addition, by employing the obtained total power consumption of the PE for each VOS voltage level, the current flowing through the PE was determined. Using this current and the saturation current of a PMOS switch for one fin, the number of the fins required for each PMOS switch was determined. Next, to reduce the timing overhead of the switching of the VOS voltage, we assumed a larger number of the fins such that the switching would occur in one cycle ($\sim$490 ps). Then, the number of the fins was increased proportional to the number of the PEs inside the island. The power overhead also included the leakage powers of the PMOS switches in the OFF state.

In the mapping process, which is performed offline, the voltage levels of the islands for a given constraint as well as the mapping of the DFG nodes on the PEs are determined. The information is loaded in the context memory for the use in the runtime. The island for each PE is predetermined which along with its context word (used for configuring the PE) would configure the PE fully. The information would be loaded in the context memory when the system starts. In the runtime, based on the required configuration of the PE, the corresponding context words is loaded from the context memory to the context register. The latency of loading the context words, depends on the bandwidth between the context memory and the CGRA. The process means that no additional energy or latency for the mapping process is induced in the proposed approximate CGRA structure compared to that of the conventional one. The only latency that one might possibly conceive is the time that after one utilization of the CGRA for a given accuracy levels, for the next use of the structure, a higher level of accuracy is required. In this case, the voltage levels of some islands need to be increased. In these cases, we should give some time for the voltage of these PEs rise from the lower level to the higher level. Our estimation showed that, in the worst case, the latency for the considered switch-box was about 500ps which was equal to around one clock period of the system. In the case that the CGRA is not in use for a while, the power gating scheme could be used for the PEs. In the case of using this scheme, setting the voltages of the islands requires some time for all the PEs to reach to their final supply voltage. Thus, the proposed scheme does not impose additional latency compared to the conventional structure when similar power gating scheme is used.

VII. Conclusion

In this paper, an energy–quality scalable CGRA based on voltage overscaling (VOS) technique was proposed. In the proposed CGRA, the output quality of each Process Element (PE) was determined by the applied VOS voltage level. To reduce the overhead of the hardware implementation of the proposed structure, the PEs were clustered in a group of voltage islands. Also, an ILP based mapping algorithm for determining the operating voltage level of each PE and binding of the operations of the input application on the PEs of the CGRA was proposed. In the proposed structure, by applying the lowest possible supply voltages for the PEs, both the energy consumption and lifetime of the CGRA were improved. The efficacy of the proposed architecture was evaluated using different benchmarks, minimum output qualities, and four voltage island sizes. Furthermore, the impact of the number of the VOS voltage levels on the energy saving and lifetime improvement was investigated. Finally, a folding technique for further improvement of the proposed structure was suggested. The results indicated considerable energy savings and lifetime improvements for the VOS based approximate CGRA when some output quality degradation could be tolerated. The results showed that the proposed CGRA could lead to 43% and 73% lower energy consumption and aging rate, respectively.

References

[1] S. Oh, H. Lee, and J. Lee, "Efficient execution of stream graphs on coarse-grained reconfigurable architectures," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 36, no. 12, pp. 1978–1988, Dec. 2017.

[2] *Coarse-Grained Reconfigurable Architecture*. Accessed: Aug. 1, 2018. [Online]. Available: http://cccp.eecs.umich.edu/research/cgra.php

[3] O. Akbari, M. Kamal, A. Afzali-Kusha, M. Pedram, and M. Shafique, "PX-CGRA: Polymorphic approximate coarse-grained reconfigurable architecture," in *Proc. DATE*, Dresden, Germany, 2018, pp. 413–418.

[4] C. Li, D. Sengupta, F. S. Snigdha, W. Xu, J. Hu, and S. S. Sapatnekar, "Special session: A quantifiable approach to approximate computing," in *Proc. CASES*, Seoul, South Korea, 2017, pp. 1–2.

[5] H. Esmaeilzadeh, A. Sampson, L. Ceze, and D. Burger, "Architecture support for disciplined approximate programming," in *Proc. ASPLOS*, London, U.K., 2012, pp. 301–312.

[6] F. Nakhaee, M. Kamal, A. Afzali-Kusha, M. Pedram, S. M. Fakhraie, and H. Dorosti, "Lifetime improvement by exploiting aggressive voltage scaling during runtime of error-resilient applications," in *Proc. Integr.*, vol. 61, Mar. 2018, pp. 29–38.

[7] S. Ramey *et al.*, "Intrinsic transistor reliability improvements from 22 nm tri-gate technology," in *Proc. IRPS*, Anaheim, CA, USA, 2013, pp. 4C.5.1–4C.5.5.

[8] J. Gu, S. Yin, and S. Wei, "Stress-aware loops mapping on CGRAs with considering NBTI aging effect," in *Proc. DAC*, Austin, TX, USA, 2017, pp. 1–6.

[9] M. Ghasemazar, H. Goudarzi, and M. Pedram, "Robust optimization of a chip multiprocessor's performance under power and thermal constraints," in *Proc. ICCD*, Montreal, QC, Canada, 2012, pp. 108–114.

[10] C. Piguet, *Low-Power CMOS Circuits: Technology, Logic Design and CAD Tools*. Boca Raton, FL, USA: CRC Press, 2006.

[11] J. W. McPherson, *Reliability Physics and Engineering Time-to-Failure Modeling*, 2nd ed. New York, NY, USA: Springer, 2013.

[12] H. Kükner *et al.*, "Scaling of BTI reliability in presence of time-zero variability," in *Proc. IRPS*, Monterey, HI, USA, 2014, pp. CA.5.1–CA.5.7.

[13] N. Goel, P. Dubey, J. Kawa, and S. Mahapatra, "Impact of time-zero and NBTI variability on sub-20 nm FinFET based SRAM at low voltages," in *Proc. IRPS*, Monterey, CA, USA, 2015, pp. CA.5.1–CA.5.7.

[14] K. He, A. Gerstlauer, and M. Orshansky, "Controlled timing-error acceptance for low energy IDCT design," in *Proc. DATE*, Grenoble, France, 2011, pp. 1–6.

[15] V. K. Chippa, D. Mohapatra, K. Roy, S. T. Chakradhar, and A. Raghunathan, "Scalable effort hardware design," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 22, no. 9, pp. 2004–2016, Sep. 2014.

[16] G. Tziantzioulis, A. M. Gok, S. M. Faisal, N. Hardavellas, S. Ogrenci-Memik, and S. Parthsarathy, "Lazy pipelines: Enhancing quality in approximate computing," in *Proc. DATE*, Dresden, Germany, 2016, pp. 1381–1386.

[17] R. Ragavan, B. Barrois, C. Killian, and O. Sentieys, "Pushing the limits of voltage over-scaling for error-resilient applications," in *Proc. DATE*, Lausanne, Switzerland, 2017, pp. 476–481.

[18] D. Mohapatra, V. K. Chippa, A. Raghunathan, and K. Roy, "Design of voltage-scalable meta-functions for approximate computing," in *Proc. DATE*, Grenoble, France, 2011, pp. 1–6.

[19] D. Ernstet al, "Razor: A low-power pipeline based on circuit-level timing speculation," in *Proc. MICRO*, San Diego, CA, USA, 2003, pp. 7–18.

[20] S. Lee, L. K. John, and A. Gerstlauer, "High-level synthesis of approximate hardware under joint precision and voltage scaling," in *Proc. DATE*, Lausanne, Switzerland, 2017, pp. 187–192.

[21] S. Xu and B. C. Schafer, "Exposing approximate computing optimizations at different levels: From behavioral to gate-level," *IEEE Trans. VLSI Syst.*, vol. 25, no. 99, pp. 3077–3088, Nov. 2017.

[22] S. Lee, D. Lee, K. Han, E. Shriver, L. K. John, and A. Gerstlauer, "Statistical quality modeling of approximate hardware," in *Proc. ISQED*, Santa Clara, CA, USA, 2016, pp. 163–168.

[23] G. Zervakis, S. Xydis, V. Tsoutsouras, D. Soudris, and K. Pekmestzi, "Multi-level approximation for Inexact accelerator synthesis under voltage island constraints," in *Proc. Workshop Approx. Comput.*, Pittsburgh, PA, USA, 2016.

[24] H. Afzali-Kusha, O. Akbari, M. Kamal, and M. Pedram, "Energy consumption and lifetime improvement of coarse-grained reconfigurable architectures targeting low-power error-tolerant applications," in *Proc. GLSVLSI*, Chicago, IL, USA, 2018, pp. 431–434.

[25] M. Hamzeh, A. Shrivastava, and S. Vrudhula, "EPIMap: Using epimorphism to map applications on CGRAs," in *Proc. DAC*, San Francisco, CA, USA, 2012, pp. 1284–1291.

[26] P. M. Heysters and G. J. M. Smit, "Mapping of DSP algorithms on the montium architecture," in *Proc. IPDPS*, Nice, France, 2003, p. 6.

[27] W. K. Mak and J. W. Chen, "Voltage island generation under performance requirement for SoC designs," in *Proc. ASPDAC*, Yokohama, Japan, 2007, pp. 798–803.

[28] J. M. Lin and Z. X. Hung, "SKB-Tree: A fixed-outline driven representation for modern floorplanning problems," *IEEE Trans. VLSI Syst.*, vol. 20, no. 3, pp. 473–484, Mar. 2012.

[29] J. M. Lin and J. H. Wu, "F-FM: Fixed-outline floorplanning methodology for mixed-size modules considering voltage-island constraint," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 33, no. 11, pp. 1681–1692, Nov. 2014.

[30] K. Usami *et al.*, "Automated low-power technique exploiting multiple supply voltages applied to a media processor," *IEEE J. Solid-State Circuits*, vol. 33, no. 3, pp. 463–472, Mar. 1998.

[31] Q. Chen, J. A. Davis, P. Zarkesh-Ha, and J. D. Meindl, "A compact physical via blockage model," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 8, no. 6, pp. 689–692, Dec. 2000.

[32] R. Sarvari, A. Naeemi, P. Zarkesh-Ha, and J. D. Meindl, "Design and optimization for nanoscale power distribution networks in gigascale systems," in *Proc. IITC*, Burlingame, CA, USA, 2007, pp. 190–192.

[33] M. W. Chen *et al.*, "A dual-edged triggered explicit-pulsed level converting flip-flop with a wide operation range," in *Proc. SOCC*, Erlangen, Germany, 2013, pp. 92–97.

[34] T. Peyret, G. Corre, M. Thevenin, K. Martin, and P. Coussy, "Efficient application mapping on CGRAs based on backward simultaneous scheduling/binding and dynamic graph transformations," in *Proc. ASAP*, Zurich, Switzerland, 2014, pp. 169–172.

[35] C. Li, W. Luo, S. S. Sapatnekar, and J. Hu, "Joint precision optimization and highlevel synthesis for approximate computing," in *Proc. DAC*, San Francisco, CA, USA, 2015, pp. 1–6.

[36] M. Martins *et al.*, "Open cell library in 15 nm FreePDK technology," in *Proc. ISPD*, Monterey, CA, USA, 2015, pp. 171–178.

[37] *Gurobi Optimizer Reference Manual*, LLC Gurobi Optim., Beaverton, OR, USA, 2018.

[38] M. Balasubramanian, S. Dave, A. Shrivastava, and R. Jeyapaul, "LASER: A hardware/software approach to accelerate complicated loops on CGRAs," in *Proc. DATE*, Dresden, Germany, 2018, pp. 1069–1074.

[39] M. Karunaratne, A. K. Mohite, T. Mitra, and L. S. Peh, "HyCUBE: A CGRA with reconfigurable single-cycle multi-hop interconnect," in *Proc. DAC*, Austin, TX, USA, 2017, pp. 1–6.

[40] R. Zheng, J. Velamala, V. Reddy, V. Balakrishnan, E. Mintarno, and S. Mitra, "Circuit aging prediction for low-power operation," in *Proc. CICC*, Rome, Italy, 2009, pp. 427–430.

Hassan Afzali-Kusha received the B.S. and M.S. degrees in electrical engineering from the University of Tehran in 2011 and 2014, respectively. He is currently pursuing the Ph.D. degree in electrical engineering with the University of Southern California. His research interests include configurable architectures, approximate computing, and innovative SRAM cell designs.

Omid Akbari received the B.Sc. degree from the University of Guilan, Rasht, Iran, in 2011, the M.Sc. degree from the Iran University of Science and Technology, Tehran, Iran, in 2013, and the Ph.D. degree from the University of Tehran, Tehran, in 2018, in electrical engineering, electronics - digital systems sub-discipline. He was as a Visiting Researcher with the CARE-Tech Group, Vienna University of Technology (TU Wien), Austria, in 2017. He is currently a Research Assistant with the Low-Power High-Performance Nanosystems Laboratory, University of Tehran. His current research interests include low-power design, energy-efficient computing, machine learning, and fault-tolerant system design.

Mehdi Kamal received the B.Sc. degree from the Iran University of Science and Technology, Tehran, Iran, in 2005, the M.Sc. degree from the Sharif University of Technology, Tehran, in 2007, and the Ph.D. degree from the University of Tehran, Tehran, Iran, in 2013, all in computer engineering. He is currently an Assistant Professor with the School of Electrical and Computer Engineering, University of Tehran. His current research interests include reliability in nanoscale design, approximate computing, neuromorphic computing, design for manufacturability, embedded systems design, and low-power design.

Massoud Pedram received the B.S. degree in electrical engineering from the California Institute of Technology, Pasadena, CA, USA, in 1986, and the M.S. and Ph.D. degrees in electrical engineering and computer sciences from the University of California at Berkeley, Berkeley, CA, USA, in 1989 and 1991, respectively. In 1991, he joined the Ming Hsieh Department of Electrical Engineering, University of Southern California (USC), Los Angeles, CA, USA, where he is currently the Stephen and Etta Varra Professor with the USC Viterbi School of Engineering. He was a recipient of the National Science Foundation's Young Investigator Award in 1994, the Presidential Early Career Award for Scientists and Engineers in 1996, two Design Automation Conference Best Paper Awards, the Distinguished Paper Citation from the International Conference on Computer Aided Design, three Best Paper Awards from the International Conference on Computer Design, the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS Best Paper Award, and the IEEE Circuits and Systems Society Guillemin-Cauer Award.