Body Bias Control on a CGRA based on Convex Optimization

Takuya Kojima [1], Hayate Okuhara [2], Masaaki Kondo [3], and Hideharu Amano [3]

[1]The University of Tokyo, Tokyo, Japan, tkojima@hal.ipc.i.u-tokyo.ac.jp
[2]National University of Singapore, Singapore, hayate01@nus.edu.sg
[3]Keio University, Yokohama, Japan, kondo@acsl.ics.keio.ac.jp,hunga@am.ics.keio.ac.jp

Abstract Body biasing is one of the critical techniques to realize more energy-efficient computing with reconfigurable devices, such as Coarse-Grained Reconfigurable Architectures (CGRAs). Its benefit depends on the control granularity, whereas fine-grained control makes it challenging to find the best body bias voltage for each domain due to the complexity of the optimization problem. This work reformulates the optimization problem and introduces continuous relaxation to solve it faster than previous work. Experimental result shows the proposed method can solve the problem within 0.5 sec for all benchmarks in any conditions and demonstrates up to 5.65x speed-up compared to the previous method with negligible loss of accuracy.

1 Introduction A CGRA consists of Processing Elements (PEs) communicated with the interconnection network, as illustrated in Fig. 1. Each PE contains an Arithmetic Logic Unit (ALU). The operations for each ALU and interconnection are changed depending on the target application kernel, which is generally a compute-intensive loop. It can archive high energy efficiency close to the Application-Specific Integrated Circuits (ASICs). The CGRA provides word-level reconfigurability in this respect. Thus, the reconfiguration overhead is much smaller compared to Field-Programmable Gate Arrays (FPGAs).

Body biasing is a promising technique to reduce unnecessary leakage power consumption, as shown in Fig. 2. However, it is challenging to find optimal body bias voltage for the CGRAs based on the target application. A previous work [1] proposes a solution based on Integer Linear Programming (ILP) in order to find an optimal solution. Nevertheless, its effectiveness is demonstrated only for a PE array divided only into eight control domains. This paper proposes a more scalable and faster optimization method by introducing continuous relaxation. Then, it shows that the relaxed optimization problem can be solved as a convex optimization.

2 CGRAs with Body Bias Control This paper chooses VPCMA2 as a target architecture, a low-power CGRA proposed for edge computing [2]. It has a PE array whose size is 8×12, and the PE is composed of an ALU and a Switch Element (SE), as shown in Fig. 3(a). The intermediate results can pass through the configurable pipeline registers, as specified by the configuration. Thereby, an appropriate pipeline structure can be used depending on the application, as shown in Fig. 3(b). A fine-grained body bias domain increases the possibility of the reverse bias for leakage reduction, as illustrated in Fig. 4. This figure assumes the same application mapping composed of three operations. If one domain includes 2x3 of PEs like Fig. 4(a), all the data paths belong to the same domain. Therefore, the zero bias has to be applied to satisfy a timing constraint in this example. On the other hand, in the case of finer granularity (per PE) as Fig. 4(b), a strong reverse bias can be used for the unused domains. Besides, even if operations are mapped to the PEs, the weak reverse bias (e.g., - 0.2 V) can be available for some domains as far as no timing violation occurs.

3 New Formulation Leakage power consumption should be minimized, satisfying constraints associated with the timing constraint. The previous work [1] formulates the problem as an ILP, focussing on the discreteness of the body bias voltage by practical body bias generators. However, the ILP formulation becomes challenging to be solved as the number of domains and voltages increase due to the NP-completeness of ILP. In contrast, this work reformulates the problem as a convex optimization problem, which can be solved in a polynomial time even for the non-linear functions. It allows the body bias voltages to be continuous values with upper and lower bounds. Then, the objective function can be expressed as a convex function by using the approximation model of the subthreshold leakage presented in [3] as follows:

$$\min P_{\text{leak}} = \sum_{i=0}^{N_{\text{dom}-1}} I_{\text{leak0},i} \exp(AV_{DD} + BV_{b,i} + CT) \quad (1) \qquad I_{\text{leak0},i} = N_{\text{PE},i} \times I_{\text{leak0,PE}} \quad (2)$$

where $I_{\text{leak0,PE}}$ is the leakage current parameter corresponding to a PE. $I_{\text{leak0},i}$ is simply calculated by the product of the number of PE in the i-th domain and $I_{\text{leak0,PE}}$. The objective function is a monotonic increase with the voltages $V_{b,i}$, which are independent of each other. Therefore, it satisfies the convexity of the function. Constraints regarding the delay time for each path are also represented as convex functions by using α-Power law [4]. Our preliminary analysis confirmed that the approximation error is less than 2% compared to the SPICE simulation based on a layout of the PE.

4 Voltage Rounding Given that the available body bias voltages are practically discrete, the optimal body bias voltages obtained by convex programming have to be rounded to the available voltages. For instance, if the optimal voltage of the convex optimization problem is 0.15 V while the voltage resolution is 0.2 V step. It has to be rounded to either 0.0 or 0.2 V. When all voltages are ceiled, the timing violation never occurs. However, it might result in a prohibitive increase in the leakage power consumption compared to the optimal case. Hence, this paper proposes two rounding methods, a scalable heuristic with $\mathcal{O}(N_{\text{dom}})$ complexity and an exact method based on an ILP. Although it is also based on ILP, the solution space is pruned efficiently by convex

programming. Therefore, it is expected to be more scalable than the entire problem is solved by ILP. However, despite the exact rounding, the result is not always the same as the method fully formulated as an ILP because the voltage assignment and the rounding are separated.

The heuristic floors all of the voltage to minimize the leakage current at the beginning. Then, voltages of some domains are ceiled step-by-step until the timing constraint is met. The order of the domains to be ceiled is ascending one of the leakage increase penalty due to ceiling. For example, let us assume domain 1 and domain 2 have the same number of PEs and are initially floored from 0.15 V to 0.0 V and from -0.15 V to -0.2 V, respectively. In this case, ceiling the voltage of domain 2 (i.e., -0.15 V to 0.0 V) brings about a larger leakage increase compared to domain 1 (i.e., 0.15 V to 0.2 V). Therefore, domain 1 is firstly selected for the ceiling in advance of domain 2. In this way, the heuristic tries to restrict a penalty for the leakage increase due to incorrect ceiling. On the other hand, the ILP for the exact rounding is formulated by introducing binary decision variables. The formulation is similar to the one in [1] so that it is omitted in this paper.

5 Experiments Several application mappings are prepared by using a mapping algorithm as listed in Table 1. For each application, different sizes of mappings are generated. The PE utilization of each mapping is well optimized. The application mappings are duplicated horizontally on the 12x8 PE array (i.e., loop unrolling) as many times as possible. For instance, two replicas of the 4x6 *dct* mapping can be placed on the rest eight PE columns, as shown in Fig. 5. The PE was designed with Verilog-HDL. The logic synthesis and place-and-route were carried out with USJC DDC (Deeply Depleted Channel) 55 nm process by using Synopsys Design Compiler and Cadence Innovus, respectively. SPICE simulations were conducted to fit the parameter used in the formulation, where the range of body bias control is between -1.0 and +0.4 V. First, the leakage current was simulated for the laid-out design, and the parameters: $A, B, C, I_{\mathrm{leak0},i}$ in Eq. (1) were estimated by a least-squares method. Next, the threshold voltage V_{t0} was estimated by a traditional constant-current-method [4] with another SPICE simulation. Lastly, α was obtained by fitting with the simulation result of the ring oscillator shown in Fig. 2. Each solver is executed with AMD Ryzen Threadripper 3960X processor.

To analyze the accuracy of the proposed formulation and the rounding methods, the optimization problems are solved while changing the domain size granularity, application mappings, and timing constraints. Four types of body bias resolution are considered: i) 0.2, ii) 0.1, iii) 0.05, and iv) 0.01 V. For example, the finest grain of the domains (i.e., 1x1) contributes to 81.8 % leakage reduction compared to the coarsest case (i.e., 12x8) for *aes* 12x6 mapping under 20MHz constraint, which is the maximum frequency without pipelining and the body biassing. Figure 6 compares the leakage power optimized with the proposed method (the new formulation with the rounding heuristic) and the previous one [1]. The leakage power varies in a wide range depending on the timing constraint. Therefore, it uses the normalized values by the cases of the previous method, which guarantees optimality. The sizes in this figure are the studied mapping sizes for each application. The coarser voltage resolution brings about the larger penalty of leakage increase in the case of a false ceiling. However, the difference is less than 5 % for 0.1 V or the finer resolutions. It is an acceptable error, considering the scalability of the method. Even though the exact rounding method can also contain a slight error, it is only around 0.1 %.

Figure 7 gives the comparison of the elapsed time in the case of *fft* 9x7 mapping, where four interesting domain sizes are extracted. The elapsed time depends on the timing constraint so that the worst-case time is selected. If the solution space is small enough, like a single domain (i.e., 12x8) with 0.2 V step, the previous method based only on ILP can solve the problem effectively and find an optimal solution faster than the other methods. Nonetheless, the domain size 1x1 (i.e., PE-by-PE control) with 0.01 V step poses the biggest optimization problem, which could not be solved by the previous method. On the contrary, the proposed method with the rounding heuristic finishes in 0.45 sec. The method based on the exact rounding can also find the solution in 4.2 sec. Figure 8 describes how much speed up the proposed methods archive compared to the previous one. This evaluation uses a middle-class problem, where the domain size is 3x2, and the voltage resolution is 0.1 V step. The proposed method with the rounding heuristic demonstrates up to 5.65x speed-up. Even though the exact rounding method reduces around 46 % of the elapsed time on average, it takes a longer time than the previous one for some cases (e.g., *gray* 3x5). That is because the separation of the voltage assignment and rounding causes nonnegligible time-overhead when the whole of the problem can be solved quickly with an ILP.

6 Conclusion This paper demonstrated that the proposed method could solve the largest-scale problem in an acceptable time. It is hoped that the scalable method will facilitate an aggressive body bias optimization for FPGAs as well as the CGRAs.

References

[1] T. Kojima, N. Ando, H. Okuhara, N. A. V. Doan, and H. Amano, "Body bias optimization for variable pipelined CGRA," in *Field Programmable Logic and Applications (FPL), 2017 27th International Conference on*. IEEE, 2017, pp. 1–4.

[2] T. Kojima and H. Amano, "Refinements in Data Manipulation Method for Coarse Grained Reconfigurable Architectures," in *2019 14th International Symposium on Reconfigurable Communication-centric Systems-on-Chip (ReCoSoC)*. IEEE, 2019, pp. 113–120.

[3] Y. Fujita, H. Okuhara, K. Masuyama, and H. Amano, "Power optimization considering the chip temperature of low power reconfigurable accelerator CMA-SOTB," in *2015 Third International Symposium on Computing and Networking (CANDAR)*. IEEE, 2015, pp. 21–29.

[4] N. H. Weste and D. Harris, *CMOS VLSI design: a circuits and systems perspective*, 4th ed. USA: Addison-Wesley Publishing Company, 2010.

Acknowledgement This work was supported in part by JSPS KAKENHI Grant 19J21493 and in part by the VLSI Design and Education Center (VDEC), the University of Tokyo in collaboration with Synopsys, Inc and Cadence Design Systems, Inc.

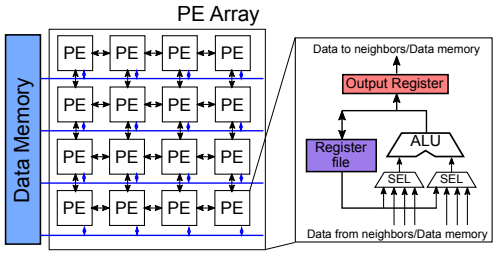


Figure 1: General description for the CGRA

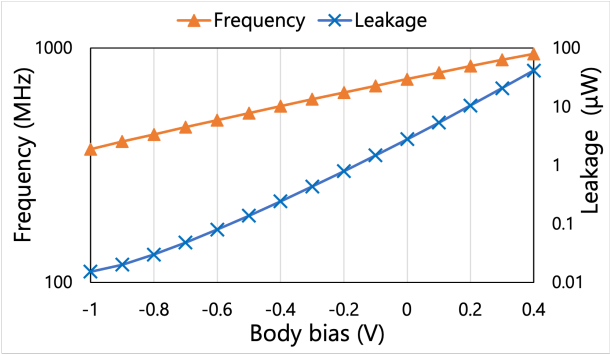
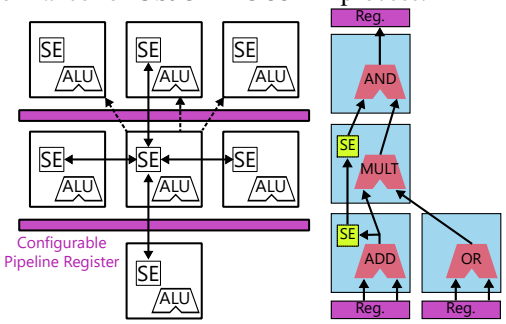
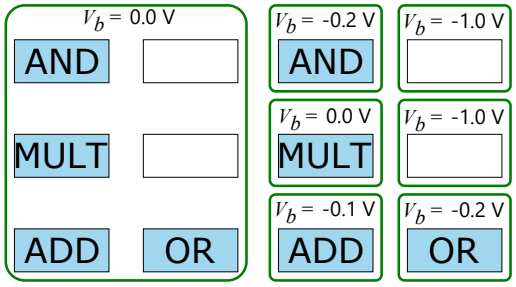


Figure 2: A trade-off possibility between leakage power and performance for USJC DDC 55 nm process.



(a) Interconnection of the PE array (b) Example of application mapping

Figure 3: Overview of VPCMA2[2]



(a) 2x3 domain (b) 1x1 domain

Figure 4: Limited reverse bias due to the large granularity of domain

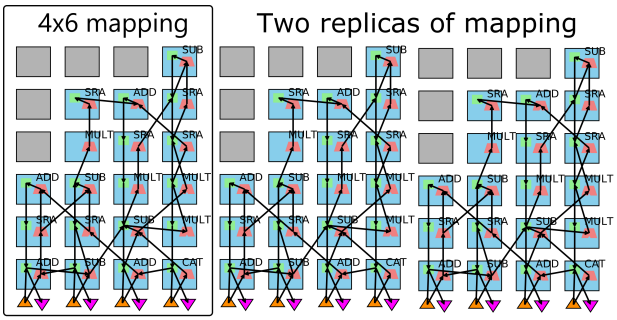


Figure 5: Mapping duplication for loop unrolling on a 12x8 PE array in the case of 4x6 *dct4* (unused upper two rows are omitted)

Table 1: Benchmark applications

Kernel	Description	map size
gray	24-bit gray scale	2x7,3x5,4x6
sepia	8-bit sepia filter	3x5,4x4,11x2
af	24-bit alpha blender	3x8,4x7,9x4
sf	24-bit sepia filter	3x8,4x7,6x6
dct	4-point discrete cosine transform	4x6,6x5,8x3
fft	Radix-4 fast Fourier transform	8x8,9x7
aes	Advanced encryption standard	11x7,12x6

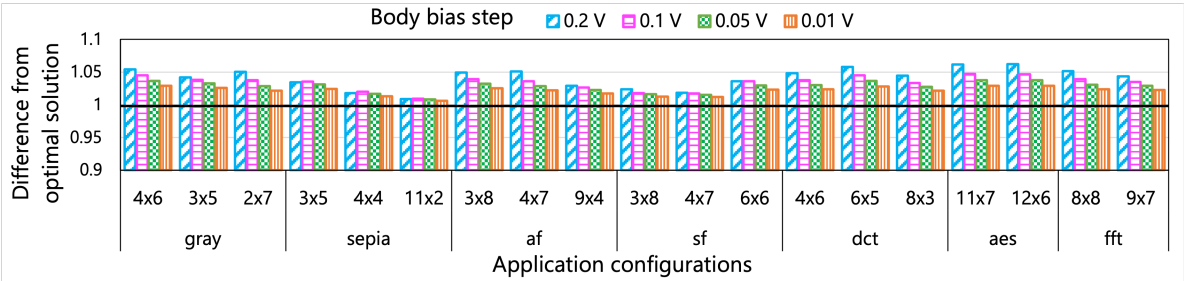


Figure 6: Difference of the optimization result by the proposed rounding heuristic from the optimal solution by the previous method [1]

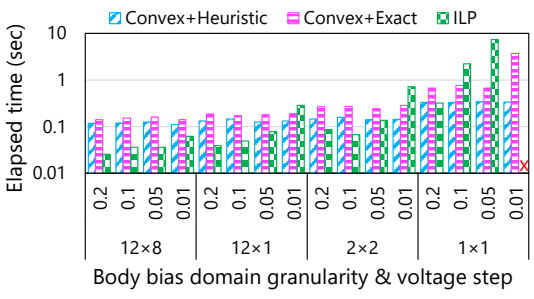


Figure 7: Elapsed time for each method (*fft* 9x7)

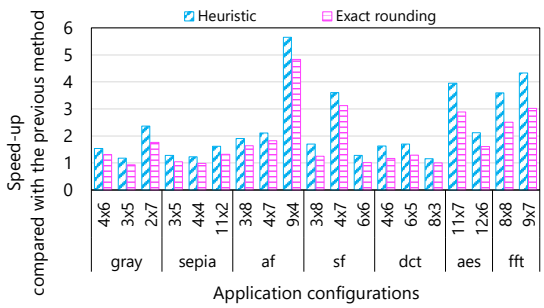


Figure 8: Speed-up of the proposed methods compared to the previous method [1]