BioCare: An Energy-Efficient CGRA for Bio-Signal Processing at the Edge

Zahra Ebrahimi, Akash Kumar
Center for Advancing Electronics Dresden (Cfaed), Technische Universität Dresden
Corresponding Author's Email: zahra.ebrahimi_mamaghani@tu-dresden.de

***Abstract*—*Coarse Grained Reconfigurable Architectures* (CGRAs) have proved to be viable platforms for health monitoring applications. Targeting energy-efficiency, *state-of-the-art* (SoA) CGRAs are augmented with approximation techniques, while still maintain acceptable accuracy at final *Quality of Result* (QoR). However, such CGRAs suffer from overheads of collecting separate Add/Mul/Div units. We propose *BioCare* as an area- and energy-efficient CGRA for health- monitoring edge devices, which exploits the synergistic effects of multiple approximations across HW/SW stack. BioCare offers different levels of energy-accuracy trade-off through the plasticity of its small PEs, each can support precision-adaptability with a *Single Instruction, Multiple Data* (SIMD) manner. BioCare demonstrates its superiority over SoAs, by achieving up to 32% and 67% area- and energy-savings, with 3.6× higher throughput. In addition to analysis on multiple kernels, evaluations on a multi-kernel ECG application shows that BioCare speed-ups the QRS detection latency by 61%, with 0% loss in accuracy. Our implementations will be available at *https://cfaed.tu-dresden.de/pd-downloads*.**

***Index Terms*—Bio-signal, ECG, EEG, CGRA, Approximate Computing, SIMD, Energy-Efficiency, Edge Computing.**

I. Introduction

The number of wearable devices – having one of the fastest-growing industries – is projected to surpass one billion and their market value is expected to grow triple and worth over $54 billion by 2023 [1], [2]. Enabling a remote and smart health (s-health) monitoring through these wearable nodes entails high-processing speed, secure and swift data transportation and storage, for which relying merely on conventional cloud computing not only will impose a drastic traffic on the network, but also cannot guarantee patients' privacy protection [3]. A promising solution to address these concerns is to efficiently *process* bio-signals at the edge, referred to as *Multi-access Edge Computing* (MEC) [3]. MEC can provide many advantages for s-health e.g., short response time through extracting necessary features and transporting them rather than whole sampled data. Thus, energy-efficiency is of utmost importance for such a battery-operated node.

To accelerate compute-intensive bio-signal analysis on personalized gadgets, SoA studies have shown merits of CGRAs [4], [5] (commercialized in e.g., Samsung Galaxy smartphones/smartwatches [6], [7]): ① post fabrication datapath flexibility with near-ASIC energy-efficiency, ② higher computation speed and smaller area/power vs. FPGAs [8]. Still, to enable a real-time processing in 24/7 portable gadgets, higher throughput should be achieved in stringent power-budget. To cope with such design constraints, several approximation techniques have emerged, however, they suffer from inevitable routing-power and die-area overhead, due to collecting accuracy-configurable Add/Mul/Div units, with an inter- or intra-PE heterogeneity[9]–[11].

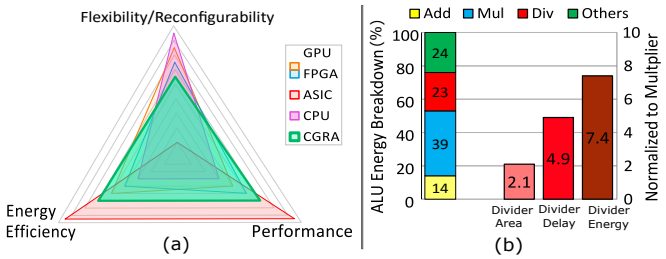

Fig. 1: a) Performance comparison of CGRA with other architectures[8] and b) operations energy in bio-signal kernels/Pan Tompkins application (16-bit ALU)

Approximation potentials are even more pronounced in bio-signal analysis, as 1) *processing* stage dissipates up to 70% of total energy in wearable nodes [12]. 2) These signals are inherently subjected to noise and their processing algorithms exhibit high parallelism and error-resiliency. Such computations can thus benefit greatly from an approximate CGRA, easily adjustable with patient's changing condition/ activity/physiology and application upgradability which usually outpace hardware updates. Yet, merely one cutting-edge work [13] has shown notable gains by deploying inexact Add/Mul in kernels of an ECG program, albeit in a fully-customized ASIC implementation, without investigating application or architecture-level techniques.

The concept of SIMD is recently exploited by e.g., Xilinx [16] and Intel [17], to provide support for precision-variability *in a single unit*. Such efforts, however, are restricted to two Muls with a common operand, or Mul/Div [15], in FPGAs. This highlights the demand for a generic approximate SIMD ALU, that can also be utilized in health-monitoring platforms. Moreover, while Add/Mul are frequent functions in bio-signal processing workloads, [9], [15], [18] and our evaluations show that long latency/high power of division, not only limit application speed, but also consumes considerable portion of ALU area/energy (Fig. 1b). These obstacles have hitherto prevented offloading of division operations to most of CGRA accelerators and relieving host processor from context-switching on such occasional interrupts.

To address above-mentioned concerns, *BioCare* sets out as an area- and energy-efficient CGRA which supports runtime function-versatility and precision-adaptability through the plasticity of its PEs; each can perform Add, Mul, or Div, on different bit-width. Particularly, featuring SIMD at intra- rather than conventional inter-PE granularity mitigates stress on global routing. Such a light-weight SIMD architecture can perfectly fit for Mul-exhaustive bio-signal processing, while also allows better utilization of PEs when division is required as well[1]. In short, our **novel contributions** are:

[1]Division is unavoidable in bio-medical programs, e.g. feature extraction in arrhythmia detection/classification by transformations/kmeans/NN (softmax layer).

TABLE I: Summary of SoA studies in the literature from the perspectives of approximation and parallelization

SIMD	Approximate Add/Mul/Div	Optimimization-Layer	Description of Work	Platform	Performance Improvement	Accuracy
✗	✓/ ✓/ ✓	Circuit	Power-gate configurable Add/Mul [10], [11] & Div [9] in PEs	CGRA	{Delay, Energy} +	PSNR 26, SSIM 0.9 in images
✗	✓/ ✓/ ✗	Circuit/Application	Quantization with inexact Muls in NN ([14] and its references)	ASIC	Energy ++	$<$ 10% classification loss
✗	✓/ ✓/ ✗	Circuit/Application	Inexact Add/Mul in ECG analysis (fixed precision kernels) [13]	ASIC	{Area, Energy} ++	PSNR/SSIM/QRS: 11, 0.3, 100
✓	Accurate	✗	Process bio-signals by multi-datapath PEs (ALU+Reg+Mux) [4], [5]	CGRA	{Energy, Throughput} +	-
✓	✗/ ✓/ ✓	Circuit/Architecture	First hybrid Mul/Div (LUT-based, customized for FPGAs) [15]	FPGA	{Energy, Throughput} +	ARE 0.8%, PSNR 45
✓	✗/ ✓/ ✓	Circuit/Architecture /Application	Enable a chain of HW/SW approximations (Mul/Div & Precision scaling) efficiently, in a cross-layer hierarchy	CGRA	{Energy, Power, Throughput} ++	Final PSNR/SSIM/QRS: $>$29.3, 0.82, 100

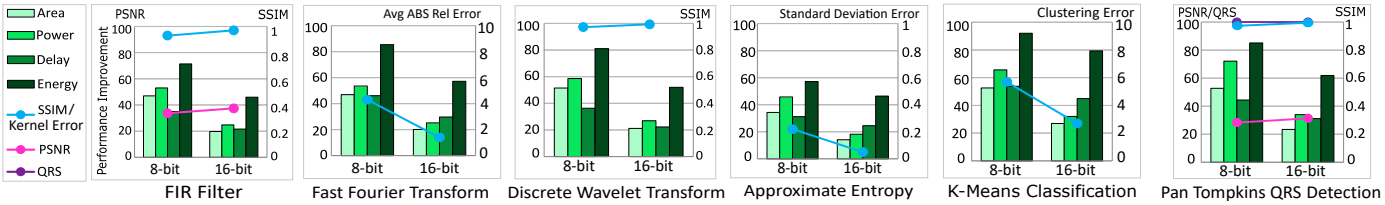

Fig. 2: Intra-kernel sensitivity analysis: performance gain & QoR changes by using SIMDive Mul/Div & lowering precision (Ref: 16-bit kernel, accurate Mul/Div)

- *BioCare*, specialized for bio-signal processing, implements a chain of customizable approximations, seamlessly across HW/SW stack. By capitalizing on *and optimizing* SoA Mul/Div[15], each proposed PE of BioCare supports function- and precision-adaptability in a small area-footprint.
- An application-level sensitivity analysis on prevalent bio-signal kernels & a multi-kernel ECG program, highlighting over-provisioning of fixed 16-bit precision. Hence, adapting precision via SIMD at runtime extends battery lifespan.
- Show $\frac{\Delta\text{Throughput}}{\Delta\text{QoR}}$ is a deciding metric that optimally reflects intensity of performance-gain over a possible accuracy loss. A gradient descent based heuristic based on this metric has also been presented that *maximizes performance gain* for a *user-defined accuracy threshold* within kernels.

II. Related Work

SIMD CGRAs: SIMD execution model has already been employed for *accurate* computations [4], [5]. However, such CGRAs suffer from high area, stems from multiple datapath in PEs. Amortizing this penalty, few works have narrowed their focus to design *approximate* SIMD Mul and/or Div [15], [19].

Approximate CGRAs/Bio-signal Processing: As mentioned, Multi-access Edge Computing, MEC, is the solution to energy-efficient processing of bio-signals [3]. Works in this cutting-edge track are, however, limited to power-gating PEs (having separate inexact Add/Mul/Div) [9]–[11]. Targeting bio-signal approximation, XBioSiP [13] has applied LSB simplification on kernels of an ECG program, *albeit all fixed to 16-bit precision*.

We adopt Mul/Div approximation of SIMDive [15], the rationale behind which is multi-fold: ① it has proved having a superior resource-accuracy trade-off over SoAs [20], plus its error-strategy is adjustable and independent from Mul/Div size. ② Its errors are unbiased: they are centered nearly symmetrical around zero, therefore, can cancel out each others in consecutive bio-signal kernels (having mostly Add/Mul) and prevent PSNR degradation. ③ It elegantly fits into i) ALU function transformation: 2D Mul/Div is converted to 1D shift & Add/Sub in logarithmic representation. ii) SIMD design: higher-precision units are achieved with modest overhead by connecting smaller instances. We distinguish our additional contributions over SIMDive: further reduce its error-coefficients, specialize its LUT-customized structure for a close-ASIC implementation to embed it in our SIMD CGRA.

III. Proposed Architecture

A. Motivation and Sensitivity Analysis

Bio-signals are usually sampled at 8 to 16-bit, which satisfies accuracy requirement of battery-powered remote monitoring gadgets. To design a CGRA that enables runtime accuracy-energy trade-off[2], we target approximation techniques at both HW & SW level (i.e. inexact operations and precision scaling). In this regard, we have conducted an analysis on prevalent kernels shared by bio-signal applications, from pre-processing (band- pass FIR filter for noise removal), to feature extraction (FFT, Discrete Wavelet Transform, and approximate entropy), and classification (K-Means). We also have analysed a widely-used multi-kernel ECG program, QRS detection: it not only is the atomic task in heart diseases diagnosis, but also employed in epilepsy/sleep apnea analysis, biometric authentication, etc. We have used Pan Tompkins QRS detection algorithm, serves as the main standard for wearable devices. This application has five kernels: low-pass followed by high-pass filter, differentiator, squarer, and moving average window. Following points summarize key observations (Fig. 1b and Fig. 2):

- Analyzing real-world ECG and EEG signals from MIT-BIH and CHB-MIT databases [22] shows samples are unevenly distributed in the range of 16-bit: 94.8% of ECG and 92.2% of EEG can be trimmed, efficiently to be fitted into 8-bit, respectively. Therefore, such high resolution can be relaxed while still accommodate most of information.
- Owing to its near-zero biased error, applying 4- or 16-coeff. SIMDive [15] marginally affected PSNR (still $>$ 29) and SSIM (still $>$ 0.8) of kernels/Pan Tompkins application. This tolerable degradation allows homogeneous structure of light-weight Mul/Div in all CGRA PEs. In contrast, approximation of Add could lead to high error, as it is also used in SIMDive algorithm (integer/fractional Add steps, Fig. 3d). Considering small contribution of Add in total PE area, we judiciously opted to keep this small unit accurate and instead focus on adaptable precision scaling.
- After sorting operations in the kernel by their significance on the output QoR, we lowered the precision of each operation and measured the gains in performance-metrics. Fig. 2 shows: 1) operations show different behavior w.r.t changes in QoR and performance: while precision scaling results in considerable critical path reduction in some

[2]Especially as many abnormality symptoms exist 24h prior to deterioration[21]

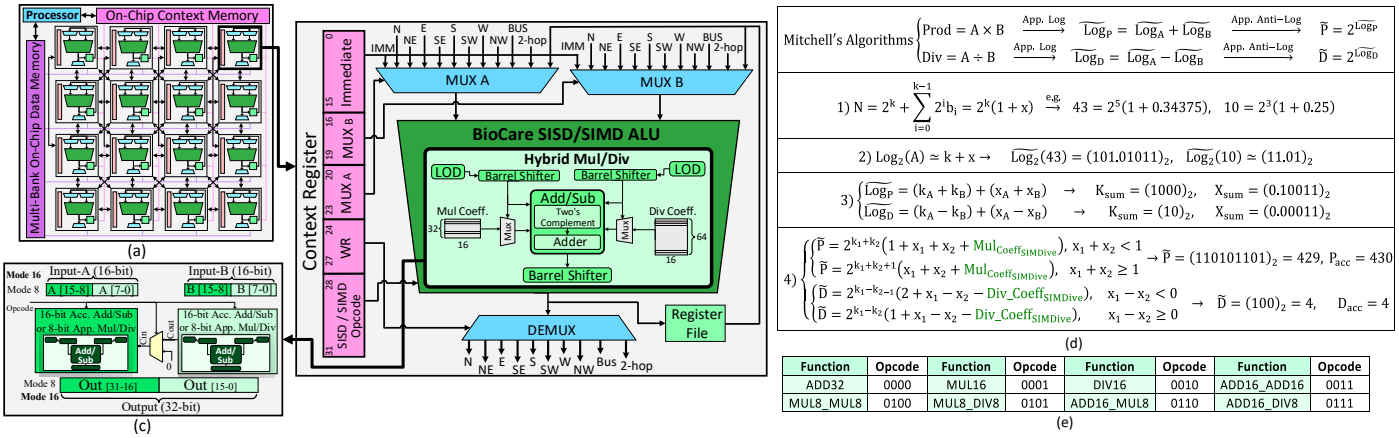

Fig. 3: (a) BioCare CGRA, (b), (c) Proposed SISD/SIMD PE/ALU (d) Mitchell's Mul/Div with SIMDive [15] Coefficients, (e) Supported SISD/SIMD opcodes.

kernels – mostly because of their division, acting as speed bottleneck operation – others might benefit more from area/power savings, 2) Even uniform 8-bit precision with accurate units provides acceptable QoR: e.g. ~100% QRS detection with PSNR of 30.3 is higher than of XBioSip which is based on *fixed* 16-bit operations (aggressively up to 12-bit at first stages, final PSNR of 11.6). Hence, a mixed-precision strategy (adjusted based on importance of operations) can be beneficial as it enables different energy-accuracy trade-offs while reduces energy and extend the battery life-time. This is achievable by our approximation heuristic and on-the-fly configurable SIMD CGRA.

B. Proposed Approximation Heuristic

Sensitivity analysis has revealed that contribution of operations in QoR sensitivity and performance gain, varies in each kernel. Hence, the goal of our proposed intra-kernel approximation strategy is to maximize gains while also guarantees a user-defined accuracy. To this end, we use the gradient of $\frac{\Delta\text{Throughput}}{\Delta\text{QoR}}$ as a deciding metric that perfectly reflects performance gain over a possible accuracy loss[3]: first, all operations are uniformly set to 16-bit precision. Then heuristic search is navigated toward reducing precision of the *operation* having highest gradient descent in the kernel (found based on the sensitivity analysis). If accuracy drops below a threshold, heuristic backtracks and continues with the next operation having highest gradient descent. An interesting point observed is combining approximate operations sometimes results in slight increase in PSNR. Our profiling has shown the *near-zero biased* errors of SIMDive Mul/Div have canceled each other in consecutive operations, thereby increasing approximation opportunities and allowing smaller precision. Arbitrary Pareto-optimal points at different levels of accuracy, generated by our heuristic, can be stored in context memory RAM enabling runtime accuracy-performance trade-offs. In this manuscript, we adhere to final 100% QRS detection (for evaluation in the next section in Fig. 5) in Pan Tompkins, with marginal PSNR degradation 29.3 (Table I).

C. Proposed SISD/SIMD CGRA Architecture

Each PE in BioCare CGRA (Fig. 3), encompass an exclusive context memory that configures the ALU for SISD/SIMD modes. ALU is assembled by a shifter, leading-one detector (LOD), adder, and two's complement unit (handle signed numbers or Sub operation which used also in SIMDive Div). Their unifying via Mitchell's algorithm, has also facilitated implementing a light-weight approximate Mul/Div [15]. This integration 1) prevents the overhead of collecting separate Add/Mul/Div units; 2) circumvents long latency of SoA dividers (with <1% error) which act as the speed-bottleneck operation of ALU; and 3) enables an approximate SIMD Div by decomposing a larger one to smaller instances (which is not mathematically practical in *accurate* mode [15]). Fig. 3e, details proposed ALU opcodes in BioCare that supports different levels of precision on-the-fly (note, addition of Mul outputs is 32-bit). PEs also include small ROMs for Mul/Div constant error-reduction coefficients of [15]. The delicate optimization that was not noticed and utilized in [15] – that we applied in our BioCare CGRA – is that Mul-coefficients can be halved as swapping operands does not affect Mitchell's Mul error. This also has reduced the complexity of the associated Mux that selects the coefficient (though these do not impose a significant overhead in CGRA, especially as Mux complexity increases logarithmically, see Fig. 4a). As CGRAs have variation of interconnect structures depending on application-domain, we have augmented BioCare with diagonal and 2-hop links (for third and sixth rows) which enable energy-efficient processing of bio-signals [24]. Although these domain-specific requirements has increased interconnection area/power, ultimately Instruction Per Cycle (IPC) and execution time/energy of the total application has improved: prolonged schedules are avoided when PEs are not used as routing nodes. Similar observations are also drawn in [25]–[27].

IV. Results and Discussion

We have assessed the collective performance metrics for different CGRA sizes, in both SISD and SIMD modes with 16-bit (subword of 8-bit) inputs. BioCare is evaluated against approximate 16-bit X-CGRA [11], GP-CGRA [9]. We analysed Add/Muls candidates in X-CGRA and GP-CGRA and selected those with lowest *resource×error bias*. In addition, we have evaluated XBioSip approximation approach [13] by implementing it in CGRA with a minimal-overhead to make 2-16 Add/Mul LSB configurable. For a fair comparison, we kept same routing structure in all CGRAs and refrained

[3]The gradient descent based heuristic used in *layer-wise* quantization of NNs [23] targets resource, while in a *fixed-area* SIMD PE, higher throughput gained by reducing the precision implies multiple operations can be executed.

TABLE II: Circuit-level metrics of 16-bit Add/Mul & 8-bit Div (normalized to accurate versions)

		ARE[1]	PRE[2]	EB[3]	Area	Power	Delay	Energy
Accurate	Acc_Add (16-bit)	-	-	-	1	1	1	1
	Acc_Mul (16-bit)	-	-	-	1	1	1	1
	Acc_Div (8-bit)[4]	-	-	-	0.56	1.13	1.32	1.48
SIMDive [15]	Acc_Add (16-bit)	-	-	-	1	1	1	1
	64 Coefficient Mul/Div (16-bit)	0.8/0.7	6.9/5.2	-0.04/-0.01	0.51	0.48	0.71	0.35
XBioSip [13]	AppAdd5 (2-16)	0.01-30	100	-0.01-6.9	0.52	0.48	0.7	0.35
	AppMultV1[5] (2-16)	0.01-61	100	0.03-60	0.54	0.53	0.63	0.37
X-CGRA [11]	RAP-CLA W6	0.2	100	0.2	1.16	1.18	0.87	1.03
	Dadda DQ42C4	8.1	51	8.1	0.66	0.62	0.69	0.43
GP-CGRA [9]	Add_Design 4	0.3	100	0.01	0.57	0.45	0.66	0.35
	Mul_Version Lit	3.4	22	3.4	0.8	0.9	0.88	0.79

[1]Avg of Absolute Relative, [2]Peak Rel, [3]Error Bias (all%), [4]Norm. to 16-bit Mul

TABLE III: Architecture-level metrics of CGRAs for 100% QRS detection

		Area ($\mu m^2 \times 10^3$)	Power (mW)	Chain Latency (ns)	Peak Throughput (GOPS)	Energy (pJ)	Area ($\mu m^2 \times 10^3$)	Power (mW)	Chain Latency (ns)	Peak Throughput (GOPS)	Energy (pJ)
		PE					2×2				
Accurate		5.1	0.25	10.7	0.11	2.7	19.6	0.94	21	0.35	11.4
BioCare	SISD	3.7	0.22	4.4	0.2	0.9	13.9	0.8	8.3	0.66	3.9
BioCare	SIMD	4.3	0.23	5.2	0.17-0.34	1.2	15.8	0.85	9.7	0.62-1.24	4.9
XBioSip		7.3	0.34	17.3	0.06	5.9	27.1	1.26	33.8	0.21	25
X-CGRA		4.5	0.23	8.1	0.19	1.8	16.9	0.84	16.1	0.42	7.9
GP-CGRA		4.8	0.24	9.3	0.18	2.2	18.4	0.9	18.6	0.39	9.5
		4×4					8×8				
Accurate		72	3.5	28	1.3	48	276	13.2	77	5.3	206
BioCare	SISD	51	2.9	15.3	2.6	15.7	190	10.9	38.7	10.1	64
BioCare	SIMD	56	3.1	16.2	2.5-4.9	19.9	207	11.1	40.5	9.6-19.2	81
XBioSip		97	4.6	45	0.84	105	360	17.2	138	3.2	441
X-CGRA		61	3.1	25	1.6	32	232	11.5	73	6.4	134
GP-CGRA		68	3.4	27	1.5	39	262	12.8	76	5.9	163

[1]Peak throughput of SISD/SIMD can be up to 2× when using 8-bit operands.

from utilizing specialized scheduling/mapping optimizations. Architectures are coded from scratch in Verilog, synthesised with Nangate 45 nm using Synopsis Design Compiler, placed & routed with Cadence Innovus. We have used widely-used list scheduling algorithm [28] to traverse kernels DFG nodes with a resource-aware approach. Table II summarizes circuit-level characteristics, providing insights for selection of [15], bodes well for an ALU design, while Table III and Fig. 4 detail architecture-level metrics of CGRAs. Finally, Fig. 5 compares performance metrics of ECG Pan Tompkins application executed on CGRAs (for BioCare SIMD mode operations precision are tuned using proposed approximate heuristic). Following inferences are highlighted based on results:

- Table II justifies that SIMDive hybrid Mul/Div suits for an approximate ALU design: it achieves lowest error-bias with higher performance improvement. Analysis exhibits that even 4 coeff. SIMDive guarantees 100% QRS detection, but we have used the more accurate, 64-coeff. Mul/Div, to create opportunity for our adaptable precision-scaling.
- *BioCare outperforms other CGRAs at architecture-level*: as indicated in Table III and Fig. 4 (b) SISD *and even SIMD* BioCare are up to 32% smaller and 67% more energy-efficient, and 41% faster than the accurate counterpart. The small cost of transforming SISD to an SIMD structure is worthwhile, as it can enable up to 3.6× higher throughput vs. accurate CGRA (if all approx. PEs configured to 8-bit).
- It is also worth underlining that Mitchell-based designs are more suited than accurate counterparts for an *SIMD* ALU design. In the hierarchical-based structure of accurate SIMD Mul, resource footprint grows quadratic (x^2) when operand size is doubled (for Div is not feasible, as mentioned

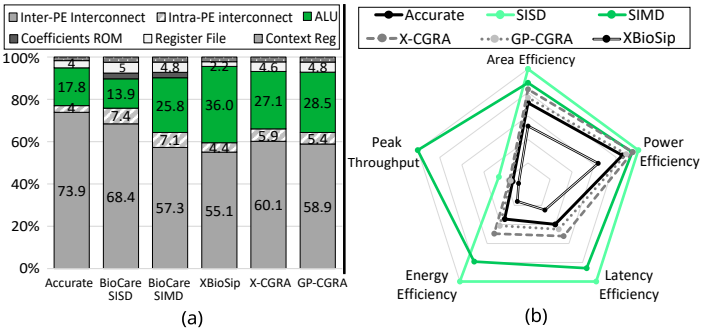

Fig. 4: a) Area breakdown after PnR, b) performance efficiency of CGRAs (for relative comparison, efficiency is translated as the inverse of resource consumption)

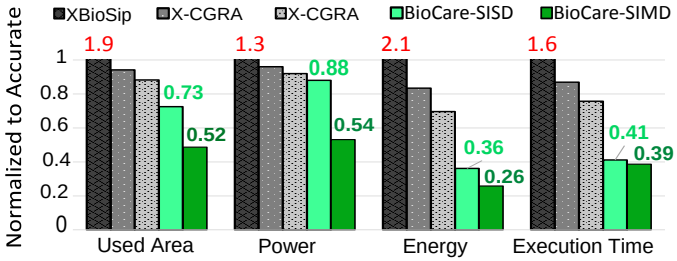

Fig. 5: Application-level comparison of CGRAs, by running ECG program (Pan Tompkins QRS Detection)

previously). This factor is smaller in BioCare ALU: for instance, a double sized Add/Sub is achieved by connecting two smaller instances. Furthermore, it is also re-used for Add/Sub step in the *logarithmic*-based structure of Mul/Div.

- Table III / Fig. 5 indicate that LSB approximation of XBioSip is rewarding, only in a customized ASIC implementation and this approach will be counter-productive when realized in a CGRA (due to the inevitable overhead of making FAs and 2×2 Muls up to 16-bit accuracy-configurable).
- *BioCare offers significant application-level savings, while maintaining acceptable accuracy*: Although architecture-level gains of other CGRAs (Table III) are also reflected at application-level (Fig. 5), ultimately SIMD BioCare provides superior savings as it accommodates kernels operations with reduced precision (achieved by the proposed approximation heuristic), in smaller PE instances. Especially 49%, 74%, and 61% reductions are gained respectively, in area, energy, and application-execution time of Pan Tompkins program, with ~0% quality loss (this shorter detection latency is appealing for real-time processing). Unutilized PEs in SIMD mode can be power-gated or leveraged for analysis of bio-signals at a higher sampling rate or accommodation of more tasks (extracting more feature at edge can alleviate overheads/challenges of communication with cloud).

V. Conclusion and Future Works

BioCare serves as a stepping stone for bio-signal processing at the edge. The proposed SISD/SIMD CGRA template benefits from high-throughput and energy-efficiency of instruction- and data-level parallelism enabled by the light-weight PE which enables adaptive precision-scaling. As future work, intend to evaluate BioCare on more multi-kernel applications such as heart arrhythmia detection through ECG processing.

Acknowledgment

This research is funded by the German Research Foundation (DFG) on project ReAp (Project Number: 380524764).

REFERENCES

[1] Global Data, "Wearable Technology Value by 2023," 2019.
[2] Statista Global Data Platform, "Number of Wearable Devices Worldwide by 2022," 2020.
[3] A. A. Abdellatif *et al.*, "Edge computing for smart health: Context-aware approaches, opportunities, and challenges," *IEEE Network*, 2019.
[4] L. Duch *et al.*, "i-DPs CGRA: An Interleaved-Datapaths Reconfigurable Accelerator for Embedded Bio-Signal Processing," *ESL*, 2019.
[5] L. Duch *et al.*, "HEAL-WEAR: An Ultra-Low Power Heterogeneous System for Bio-Signal Analysis," *TCAS-I*, 2017.
[6] C. Kim *et al.*, "ULP-SRP: Ultra Low-Power Samsung Reconfigurable Processor for Biomedical Applications," *TRETS*, 2014.
[7] "Samsung Exynos 7420 Deep Dive - Inside A Modern 14nm SoC," 2015.
[8] L. Liu *et al.*, "A Survey of Coarse-Grained Reconfigurable Architecture and Design: Taxonomy, Challenges, and Applications," *CSUR*, 2019.
[9] M. Brandalero *et al.*, "Approximate On-The-Fly Coarse-Grained Reconfigurable Acceleration for General-Purpose Applications," in *DAC*, 2018.
[10] O. Akbari *et al.*, "Toward Approximate Computing for Coarse-Grained Reconfigurable Architectures," *MICRO*, 2018.
[11] O. Akbari *et al.*, "X-CGRA: An Energy-Efficient Approximate Coarse-Grained Reconfigurable Architecture," *TCAD*, 2019.
[12] R. Braojos *et al.*, "A Synchronization-Based Hybrid-Memory Multi-Core Architecture for Energy-Efficient Biomedical Signal Processing," *TC*, 2017.
[13] B. S. Prabakaran *et al.*, "XBioSiP: A Methodology for Approximate Bio-Signal Processing at the Edge," in *DAC*, 2019.
[14] V. Mrazek *et al.*, "ALWANN: Automatic Layer-Wise Approximation of Deep Neural Network Accelerators without Retraining," in *ICCAD*, 2019.
[15] Z. Ebrahimi *et al.*, "SIMDive: Approximate SIMD Soft Multiplier-Divider for FPGAs with Tunable Accuracy," in *GLSVLSI*, 2020.
[16] Y. Fu *et al.*, "Deep Learning with INT8 Optimization on Xilinx Devices," *White Paper*, 2017.
[17] M. Langhammer *et al.*, "High Density Pipelined 8bit Multiplier Systolic Arrays for FPGA," in *FPGA*, 2020.
[18] H. Saadat *et al.*, "Approximate Integer and Floating-Point Dividers with Near-Zero Error Bias," in *DAC*, 2019.
[19] R. R. Osorio and G. Rodríguez, "Truncated SIMD Multiplier Architecture for Approximate Computing in Low-Power Programmable Processors," *IEEE Access*, 2019.
[20] Z. Ebrahimi, S. Ullah, and A. Kumar, "LeAp: Leading-one Detection-based Softcore Approximate Multipliers with Tunable Accuracy," in *2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2020, pp. 605–610.
[21] A. Anzanpour *et al.*, "Self-awareness in remote health monitoring systems using wearable electronics," in *DATE*, 2017.
[22] A. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals." *Circulation*, 2000.
[23] P. Judd *et al.*, "Proteus: Exploiting Numerical Precision Variability in Deep Neural Networks," in *ICS*, 2016.
[24] J. Lopes *et al.*, "Evaluation of CGRA Architecture for Real-Time Processing of Biological Signals on Wearable Devices," in *ReConFig*, 2017.
[25] F. Bouwens *et al.*, "Architectural Exploration of the ADRES Coarse-Grained Reconfigurable Array," in *ARC*, 2007.
[26] J. W. Yoon *et al.*, "Architecture Customization of On-Chip Reconfigurable Accelerators," *TODAES*, 2013.
[27] S. A. Chin *et al.*, "CGRA-ME: A Unified Framework for CGRA Modelling and Exploration," in *ASAP*, 2017.
[28] G. D. Micheli, *Synthesis and Optimization of Digital Circuits*. McGraw-Hill Higher Education, 1994.