Multi-level Optimization of an Ultra-Low Power BrainWave System for Non-Convulsive Seizure Detection

Multi-level Optimization of an Ultra-Low Power BrainWave System for Non-Convulsive Seizure Detection

Barry de Bruin, *Student Member, IEEE*, Kamlesh Singh, *Student Member, IEEE*, Ying Wang, Jos Huisken, José Pineda de Gyvez, *Fellow, IEEE*, and Henk Corporaal, *Member, IEEE*

***Abstract*—We present a systematic evaluation and optimization of a complex bio-medical signal processing application on the BrainWave prototype system, targeted towards ambulatory EEG monitoring within a tiny power budget of $<1\,\mathrm{mW}$. The considered BrainWave processor is completely programmable, while maintaining energy-efficiency by means of a Coarse-Grained Reconfigurable Array (CGRA). This is demonstrated through the mapping and evaluation of a state-of-the-art non-convulsive epileptic seizure detection algorithm, while ensuring real-time operation and seizure detection accuracy. Exploiting the CGRA leads to an energy reduction of 73.1%, compared to a highly tuned software implementation (SW-only). A total of 9 complex kernels were benchmarked on the CGRA, resulting in an *average* $4.7\times$ speedup and *average* $4.4\times$ energy savings over highly tuned SW-only implementations. The BrainWave processor is implemented in 28-nm FDSOI technology with $80\,\mathrm{kB}$ of Foundry-provided SRAM. By exploiting near-threshold computing for the logic and voltage-stacking to minimize on-chip voltage-conversion overhead, additional 15.2% and 19.5% energy savings are obtained, respectively. At the Minimum-Energy-Point (MEP) ($223\,\mu\mathrm{W}$, $8\,\mathrm{MHz}$) we report a measured state-of-the-art 90.6% system conversion efficiency, while executing the epileptic seizure detection in real-time.**

***Index Terms*—Ultra-low power architectures, Coarse-Grained Reconfigurable Arrays, Bio-medical Signal Processing, Non-Convulsive Epileptic Seizure Detection, Voltage-Stacking**

I. Introduction

In the last decade, ambulatory or remote health monitoring of common chronic neurological diseases, such as Epilepsy and Parkinson's Disease (PD), has become of increasing importance. This can be attributed to the tremendous cost-savings opportunities of preventive health-care and non-hospital diagnosis and treatment, as well as the large quality-of-life improvements for patients of chronic diseases to remain mobile and autonomous. Commercial devices for wearable ambulatory monitoring do exist; however, these devices generally have limited battery lifetime [1], or use non-EEG sensors that are insufficient to reliably detect more complex brain-related seizure types [2], [3]. While non-EEG sensors can be sufficient to reliably detect convulsive epileptic seizures [4], many patients suffer from non-convulsive epileptic seizures or Non-Convulsive Status Epilepticus (NCSE), which lack obvious visual and motor symptoms and are therefore challenging to detect without real-time EEG signal analysis [5]. Existing monitoring devices with non-EEG sensors are therefore not suitable for long-term ambulatory monitoring of patients with non-convulsive epilepsy [3].

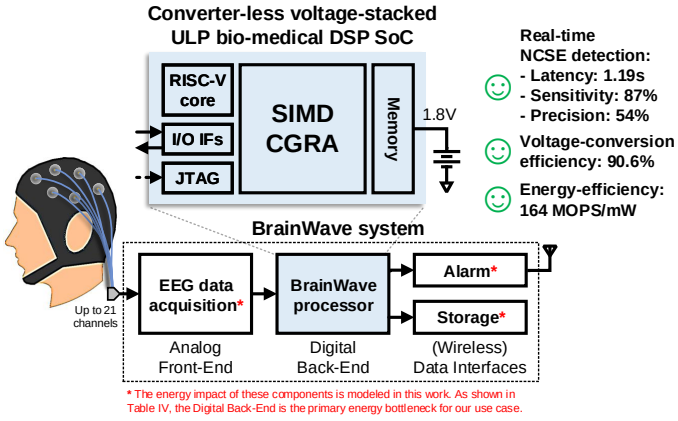


Fig. 1. Overview of BrainWave system for EEG-based seizure detection. BrainWave aims to enable ambulatory EEG monitoring in a portable battery-operated device with an ultra-low-power BrainWave processor.

An ongoing trend for EEG-based ambulatory monitoring devices is to move the signal analysis from the cloud to the edge, where devices are equipped with processing capabilities to perform bio-medical signal analysis and data reduction. State-of-the-art bio-medical signal processing platforms consist of one or multiple processor cores and are typically coupled with hardware accelerators [6]–[12]. Unfortunately, these architectures either lack energy-efficiency if the architecture is fully programmable, or are specialized towards a limited set of kernels and applications. Prior art on wearable EEG systems focuses primarily on the optimization of traditional seizure detection algorithms by providing dedicated hardware acceleration for spectral, time-frequency and entropy features i.e. [6]–[10], [13]. However, these algorithms are insufficient for reliable non-convulsive seizure detection, which demands more complex algorithms [5], [14]. For detection of these complex brain-related seizures i.e. non-convulsive epileptic seizures, but also PD Freezing-of-Gait, research is ongoing for the optimal algorithms or sensors. As such, there is a clear need for processing platforms that are both energy-efficient and flexible to account for future improvements in algorithms.

Hardware acceleration using a Coarse-Grained Reconfigurable Array (CGRA) is being promoted as a good compromise between

Manuscript received XX, 2021; revised XX, 202X and XX, 202X; accepted XX, 202X. Date of publication XX, 202X; date of current version XX, 202X. This work was supported in part by the Dutch NWO Project 14714 BrainWave. (*Corresponding author: Barry de Bruin.*)

Barry de Bruin, Kamlesh Singh, Jos Huisken, and Henk Corporaal are with the Department of Electrical Engineering, Eindhoven University of Technology, 5612 AZ Eindhoven, The Netherlands (e-mail: e.d.bruin@tue.nl; k.k.singh@tue.nl; j.a.huisken@tue.nl; h.corporaal@tue.nl).

Ying Wang is with the Department of Biophysics, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands, and also with the Department of Electrical Engineering, Eindhoven University of Technology, 5612 AZ Eindhoven, The Netherlands, and also with the Biomedical Signals and Systems Group, Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS), University of Twente, Enschede, The Netherlands, and also with the ZGT Academy, Ziekenhuisgroep Twente, Almelo, The Netherlands (e-mail: ying.wang@utwente.nl).

José Pineda de Gyvez is with the Department of Electrical Engineering, Eindhoven University of Technology, 5612 AZ Eindhoven, The Netherlands, and also with NXP Semiconductors, San Jose, CA 95134 USA (e-mail: j.pineda.de.gyvez@tue.nl).

Color versions of one or more figures in this article are available at https://ieeexplore.ieee.org.

Digital Object Identifier XX.XXXX/TBCAS.2021.XXXXXXX.

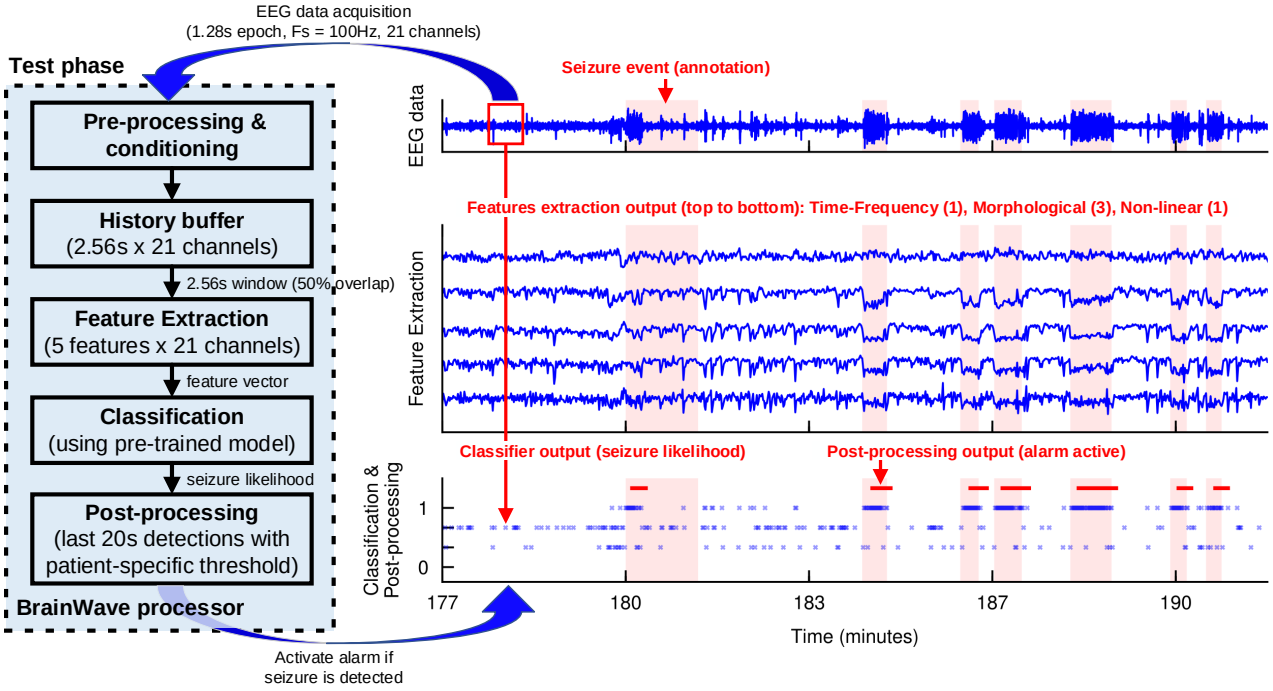


Fig. 2. (left) Overview of complete non-convulsive epileptic seizure classification pipeline running on the BrainWave processor. (right) Illustrative example of system behavior for several minutes of EEG data (one channel shown) with corresponding feature values, and classification and post-processing output.

flexibility and energy-efficiency for bio-medical signal processing on ultra-low power (ULP) systems ($<1\,\mathrm{mW}$) [11], [15]–[19]. Unfortunately, the efficiency of these platforms is typically evaluated on small and regular kernels or the system-level energy savings are sub-optimal due to under-utilization of the CGRA. Often, only part of the system is considered in the efficiency evaluation, and important system aspects like the voltage-conversion efficiency and area overhead of on-chip power delivery circuitry are ignored. In contrast to these works, we present a system-level energy evaluation where the CGRA is heavily utilized (91.41% duty cycle).

This paper extends our previous work [19] as follows: First, this is the first work that performs a system-level optimization of a real-time non-convulsive seizure detection algorithm, which requires more complex features than the traditional spectral and time-frequency features found in low-power seizure detection SoCs for convulsive seizures [6]–[10], [12]. Second, in contrast to [19], which was based on simulations, all results in this paper are based on chip measurements on our BrainWave processor, which was fabricated in 28-nm FDSOI. This paper aims to tackle the energy problem of a wearable EEG-based monitoring system for non-convulsive seizure detection. We present a prototype system consisting of an ULP programmable and re-configurable System-on-Chip (SoC) for wearable and on-device (digitized) signal acquisition and analysis. The proposed system is fully programmable, while improving the energy-efficiency by 2.7–6.1× compared to highly tuned SW-only implementations of a diverse set of complex EEG feature extraction algorithms on a RISC-V core with DSP extensions. To the best of our knowledge, this is the first implementation and optimization of a complete non-convulsive epileptic seizure detection algorithm on a ULP micro-controller, i.e. BrainWave processor, targeted towards real-time seizure detection in a wearable form-factor. The novel contributions are at four system design levels:

1) **System-level:** evaluation of first system prototype for real-time non-convulsive seizure detection. This includes an energy trade-off analysis between cloud and edge processing (Section III) and an investigation on the impact of EEG data precision and algorithm fixed-point quantization on event detection accuracy (Section VI-A).
2) **Algorithm-level:** extensive evaluation of the SW optimizations of complex EEG features. For non-linear entropy and morphological visibility graph (VG) features we report a speedup of 5.3× and 155.6×, respectively, compared to reference implementations on the RISC-V core (Section IV). This is the first work on the acceleration of VG features for time-series on an embedded platform with limited on-chip memory.
3) **Architecture-level:** off-loading of performance and power costly features to the hardware accelerator (CGRA) results in average energy-savings of 4.4× at the kernel-level compared to a highly tuned SW-only implementation (Section VI-B), and a 73.1% energy reduction at the system-level (Section VI-C1), with a CGRA duty-cycle of 91.41%.
4) **Circuit-level:** we explore near-threshold computing combined with voltage-stacking and obtain additional energy savings of 15.2% and 19.5%, respectively, with a state-of-the-art system voltage-conversion efficiency of 90.6%, while executing the seizure detection application in real-time (Section VI-C2).

The remainder of this paper is organized as follows: Section II introduces the non-convulsive seizure detection algorithm and BrainWave processor architecture. Section III provides a brief summary of important system design aspects. In Section IV the implementation and optimization of the algorithm on the BrainWave processor is described, to enable real-time and energy-efficient operation. Section V details the BrainWave processor implementation and measurement setup. Measurement results and discussion are provided in Section

VI, followed by concluding remarks in Section VIII.

II. Preliminaries and background

Non-convulsive epileptic seizures or NCSE are challenging to detect reliably without real-time monitoring of EEG signals by medical professionals, due to the lack of physical convulsions or motor symptoms. Several automated NCSE detection algorithms were proposed in literature, e.g. [5], [14], but minimizing the number of false alarms while obtaining an acceptable seizure event detection rate ($>80\,\%$) remains a major challenge [5]. Additionally, none of the recent works on automated NCSE detection focus on the design of a real-time implementation on an embedded processor platform with tight resource constraints. The algorithm and dataset that are used in this work are based on a state-of-the-art non-convulsive epileptic seizure detection pipeline, which is previously published by Y. Wang *et al.* [5]. We port this algorithm to the resource-contrained BrainWave processor with only $80\,\mathrm{kB}$ of on-chip memory, while minimizing energy consumption at multiple design levels and ensuring real-time operation and algorithmic robustness against signal quantization. The structure of the complete seizure detection pipeline is illustrated in Fig. 2 (left).

A. Real-time non-convulsive epileptic seizure detection

1) Dataset properties: The considered clinical EEG dataset [5] contains $126.7\,\mathrm{h}$ of scalp EEG recording and consists of $13.9\,\mathrm{h}$ of seizure data (316 events) with an average event duration of $158\,\mathrm{s}$. The dataset was captured and annotated by experts in a clinical environment, digitized using three different systems (BrainRT, Micromed, and EEG stellate) at different sampling frequencies ($256\,\mathrm{Hz}$, $200\,\mathrm{Hz}$, and $100\,\mathrm{Hz}$). To standardize the recordings between different EEG acquisition systems all subject data was down-sampled to $100\,\mathrm{Hz}$. Twenty-one common electrodes between the systems were used.

2) Pre-processing and conditioning: Every $1.28\,\mathrm{s}$ the on-chip EEG data is segmented in a $2.56\,\mathrm{s}$ window with 50% overlap with the previous window. The 256-sample window ($2.56\,\mathrm{s} \times 100\,\mathrm{Hz}$) was chosen to limit the on-chip memory requirements. An on-line 10th order $1\,\mathrm{Hz}$–$45\,\mathrm{Hz}$ Butterworth band-pass filter (BPF) is applied to every EEG channel to eliminate low-frequency artifacts and remove $50\,\mathrm{Hz}$ mains interference. Furthermore, the EEG channels are re-referenced to the average of all channels to reduce movement and muscle artifacts. Finally, a Hann window is applied to every channel.

3) Feature extraction and selection: Feature extraction is performed on the individual EEG channels. A feature importance analysis was conducted on a total of 45 different features. A minimum feature set was derived using a correlation-based feature selection method. More information can be found in [5]. The 5 resulting features that were used are from the non-linear, time-frequency and morphological domain. For non-linear features the Sample Entropy (SampEn) of the EEG time-series is used. For time-frequency features we compute the standard deviation of the Detail coefficients of the 4th level ($3.125\,\mathrm{Hz}$ to $6.25\,\mathrm{Hz}$) Discrete Wavelet Decomposition (DWT). The DWT filters are derived from the Daubechies 4 (db4) wavelet. The morphological features are based on the Visibility Graphs (VG) of the EEG time-series. We compute the Node Degree (ND) of the Normal VG (NVG), Horizontal VG (HVG) and Difference VG (DVG), and extract the Approximate Entropy (ApEn) of all ND vectors as a feature.

4) Classification and post-processing: We use the classifier and post-processing stage as presented in [5]. A RUSBoost tree ensemble classifier [20] is used to deal with the imbalanced nature of the dataset. Alternative classifiers that are commonly used, especially in low-power systems, are SVMs with non-linear kernels [8], [9]. However, their computational and memory requirements can be quite high [8]. First, the RUSBoost algorithm is trained using 2 classes (seizure, no seizure). Based on the misclassifications of this 2-class model, a synthetic 4-class dataset is constructed and a 4-class model is trained (false negatives are labeled as 'suspected no seizure', false positives are labeled as 'suspected seizure'). To reduce the number of false alarms, the post-processing stage uses the classifier outputs of the last $20\,\mathrm{s}$ with a patient-specific threshold to determine whether the alarm should be activated, which corresponds to sending a small notification message to a wireless end-point that is observed by clinical experts. The run-time behavior of the system is illustrated in Fig. 2 (right).

5) Accuracy metrics: The performance of the algorithm on the labeled dataset is evaluated in terms of event detection performance, as non-convulsive seizure events span multiple epochs. The event detection accuracy is counted as follows: Every 2 minutes of correctly classified non-seizure data counts as a true negative (tn) event. Aggregating epochs into 2-minute events simulates a realistic scenario, where a clinician or caregiver checks whether the patient is experiencing a non-convulsive seizure in regular intervals of 2 minutes. If a seizure is missed or detected too late ($>$2 minutes after onset), the event is counted as a false negative (fn) event. A seizure event that is detected within 2 minutes before or after the annotated onset is counted as a true positive (tp). If the system incorrectly detects a seizure for $20\,\mathrm{s}$, it is counted as a false positive (fp) and the detector is muted for 2 minutes, to simulate actions taken by clinicians in case of a false positive. The muting system of 2 minutes is designed to simulate the actions taken by clinicians against false

TABLE I
NCSE dataset summary [5].

Subject	Recording duration (h)	Seizure duration (h)	Seizure events	Avg. event duration (h)
#1	0.8	0.7	2	0.36 ± 0.02
#4	8.8	0.8	15	0.06 ± 0.03
#5	2.9	0.4	14	0.03 ± 0.01
#6	22.3	0.6	5	0.13 ± 0.11
#7	21.7	0.8	118	0.01 ± 0.00
#8	18.2	1.3	13	0.10 ± 0.27
#11	10.1	1.1	38	0.03 ± 0.02
#12	16.7	0.6	38	0.01 ± 0.01
#14	18.0	5.6	35	0.16 ± 0.13
#15	7.2	1.9	38	0.05 ± 0.06
Total	126.7	13.9	316	0.04 ± 0.09

TABLE II
Average run-time per epoch of baseline seizure detection pipeline on RISC-V core (21 channels $\times$ 256 samples epoch size).

Stage	Calls (#)	Cycles ($\times 10^6$)	Total (%)
Data acquisition	128	0.09	0.01
Band-pass filter	21	0.89	0.13
Epoch conditioning	21	0.33	0.05
Non-linear features	21	40.61	6.09
Morphological features	21	624.05	93.58
Time-Frequency features	21	0.52	0.08
Construct feature vector	1	<0.01	<0.01
Classification + post-proc.	1	<0.01	<0.01
Other	1	0.36	0.05
Total	-	666.85	100.00

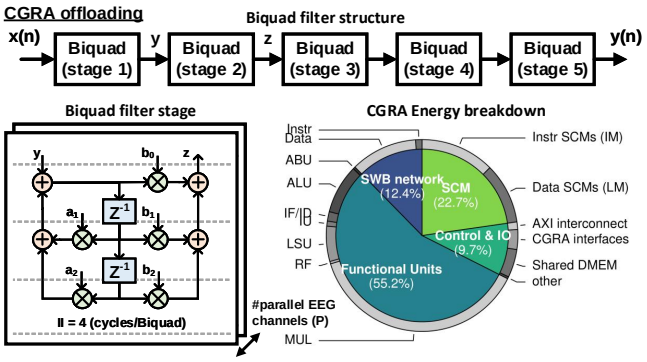


Fig. 3. Mapping and energy breakdown of BPF application on Blocks CGRA (back-annotated netlist simulation at 0.7 V, 50 MHz, 25 °C).

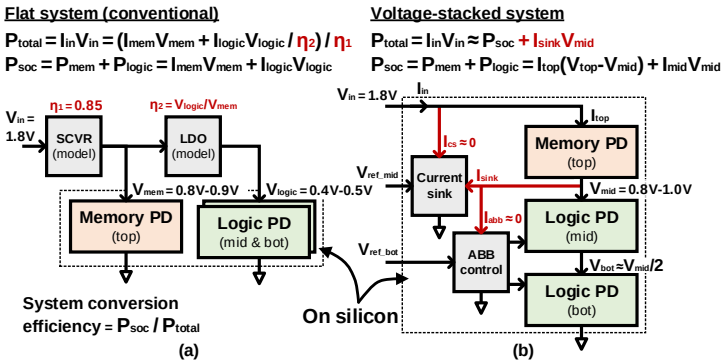


Fig. 4. (a) Conventional ULP System-on-Chip (SoC) requires (on-chip) voltage-conversion which have a large area penalty and sub-optimal system efficiency. (b) In voltage-stacked mode all power domains are stacked in series, thereby eliminating expensive explicit voltage-conversion circuitry.

alarms. A system that generates many false alarms could lead to alarm fatigue for the clinicians, which can lead to dangerous situations for the patient. Muting the system for a short duration will not affect the normal clinical practice.

B. Energy-efficient signal processing on the BrainWave processor

The BrainWave processor (overview in Fig. 5) is based on the open-source Pulpino micro-controller[1], which consists of a single-issue RISC-V core with DSP extensions [21], coupled with 80 kB of tightly-coupled program (PMEM) and data memory (DMEM). The performance of this core is comparable to the ARM Cortex-M4 core, a popular choice for low-power embedded signal processing. For hardware acceleration an instantiation of the Blocks CGRA [17] with 21 functional units (FUs) is chosen, which acts as a co-processor.

1) Baseline SW-only implementation and bottleneck analysis: The complete algorithm is implemented in C and functionally verified against a high-precision Matlab reference implementation. For efficient deployment on the BrainWave processor, the pipeline is quantized to 16-bit fixed-point (with 32-bit intermediate results). Commonly used complex functions such as fixed-point ln(), log2() are implemented using a small 16-bit 32-segment piece-wise linear approximation lookup table (≈ 60 cycles/call). Less frequently used functions such as an integer sqrt() are implemented using iterative approximations (≈ 57/90 cycles/call for 16/32-bit values). Table II depicts the average run-time breakdown of the baseline implementation of the non-convulsive seizure detection algorithm on the RISC-V processor (measured using on-chip timers). It follows that the pre-processing and feature extraction stages (especially the non-linear entropy and the Morphological Visibility Graph features) consume >99% of the run-time, and are therefore primary candidates for further optimization. To enable real-time classification on the BrainWave processor at near-threshold operation conditions (i.e. <10 MHz) we need to improve the throughput by >66×.

2) Kernel offloading to the Blocks CGRA: Complex EEG features are offloaded to the Blocks CGRA, which is characterised by its grid of programmable FUs and reconfigurable instruction/data switchbox (SWB) network. The FU grid consists of 5 different FU types that each support a subset of instructions to reduce the instruction width (e.g. a Load-Store-Unit (LSU) for memory operations, a Multiplier Unit (MUL) for multiply/shift operations). More information on the Blocks ISA can be found in [17]. Units can share the same program counter to operate as a very-large-instruction-word (VLIW) processor, and SIMD-parallelism is naturally supported since instructions can be broadcasted to multiple FUs over the reconfigurable instruction network. The reconfigurable data network allows the CGRA to bypass the register file, and pipeline multiple FUs together. Every LSU has access to the shared DMEM to access the EEG data and can store intermediate results in small 1 kB private standard-cell memories (SCMs) for local processing (LM). A CGRA kernel is pre-loaded in private 256-word instruction memories, and is reused by consecutive acceleration requests to reduce reconfiguration overhead.

3) CGRA mapping example: Fig. 3 illustrates how a band-pass filter (BPF) kernel is mapped to the CGRA. The BPF is implemented as a cascade of Biquad sections. Each Biquad filter stage consists of 2 parallel multiply-accumulate paths, which are mapped with an initiation interval (II) of 4 cycles to the CGRA. During every cycle, operations (with the same color) are mapped on FUs that share the same instruction (fine-grained SIMD). Intermediate results are explicitly bypassed between FUs to reduce register writes and reads. To further reduce instruction overhead, SIMD-parallelism is exploited by executing multiple EEG channels in parallel. From the CGRA energy breakdown it follows that over half of the total energy is spent in FUs, due to efficient data-movement and instruction sharing. For reference, it should be noted that many off-the-shelf processors typically spend less than 10% of all energy on processing [22].

C. Near-threshold operation and voltage-stacking

The BrainWave processor is designed assuming an always-on battery-operated system. Voltage scaling to the near/sub-threshold (V_{th}) region is a commonly used strategy to maximize energy-efficiency of integrated circuits by operating at the Minimal-Energy-Point (MEP). Unfortunately, near/sub-V_{th} design is challenging from an implementation and performance perspective.

1) Design for near/sub-V_{th} operation: The operating voltage for commercial (foundry-provided) high-density SRAM memories does not scale to the near/sub-V_{th} region. As such, multiple external or on-chip voltage converters are required to provide distinct voltages for SRAM blocks and logic circuits, which greatly increases design complexity and area overhead [23]. Additionally, achieving high system voltage-conversion efficiency (>85%) for ULP (<1 mW) systems is challenging [6], [15], [24]–[28]. Alternative solutions like custom low-voltage SRAM memories come with significant design and area overhead, and in most cases an industry standard 1.8 V supply is still required for the chip IO-pads. In the BrainWave processor we address these issues using voltage-stacking. Voltage-stacking is a charge-recycling technique to eliminate the need for

[1]https://github.com/pulp-platform/pulpino

explicit voltage-conversion, by connecting multiple power domains (PD) in series. The BrainWave processor is implemented in a 28-nm FDSOI technology and can be configured to run in a 3-level voltage-stacked mode, such that the logic is running in the near-V_{th} range and the SRAM blocks are running at a higher voltage, following the foundry-provided specification.

2) Conventional and voltage-stacked operation: The operation principle of the conventional (flat) and voltage-stacked system is illustrated in Fig. 4. In conventional mode multiple (external or on-chip) voltage converters are required to supply the memory and logic PDs. In voltage-stacked mode, a single external $1.8\,\mathrm{V}$ supply voltage is used. The foundry-provided SRAMs in the memory PD (top) operate in the range of $0.8\,\mathrm{V}$ to $0.9\,\mathrm{V}$ ($V_{in} - V_{mid}$), and both logic PDs (mid and bot) operate in the voltage range of $0.4\,\mathrm{V}$ to $0.5\,\mathrm{V}$, which is a typical range for near-threshold operation, and is close to the MEP for our use case. To balance the intermediate top-to-mid voltage rail, a current sink controller is used to sink excess current from the memory domain. Similarly, the intermediate voltages rails of the logic domains are balanced using adaptive-body-bias (ABB) controllers. The system conversion efficiency is maximized when the sink current equals zero, as the energy overhead of the current sink controller and the ABB controllers is negligible. The design and implementation of the 3-level voltage-stacked chip and balancing circuitry was presented in K. Singh *et al.* [24]. In this work we evaluate this at the system-level, thereby, for the first time, demonstrating the energy savings of a near-threshold voltage-stacked system on a real-world use case.

III. BrainWave system design

To design a wearable EEG monitoring system with a battery life of >1 week, without compromising signal quality, different components in an EEG monitoring system need to be carefully tuned. A representative system for wearable EEG monitoring is presented in Fig. 1. The system is divided in (analog) Front-End (FE) and (digital) Back-End (BE). The FE is responsible for data acquisition and analog-to-digital conversion. The BE performs the signal conditioning and seizure classification and utilizes a wireless link to notify medical experts or to store data in the cloud for post-analysis.

State-of-the-art ULP EEG-based seizure detection systems utilize 10–12 bit ADCs, low noise amplifiers and advanced filtering in the Analog Front-End (AFE) to maximize battery life [9], [13], [29]. Using over-designed ADCs with higher precision than necessary leads to up to 2 orders of magnitude more energy per sample [30]. For battery-constrained systems it is therefore critical to design the ADC around the minimal application dynamic range requirements. The importance of properly sizing the ADC is also illustrated in Table III, which provides a brief overview of 3 representative EEG data acquisition solutions. The TMSi Mobita is a medical-grade system for wearable ambulatory EEG logging. The 24-bit ADC makes it unable to operate >1 day on a single battery charge. When we compare the TI ADS1299, which has comparable specifications to the TMSi Mobita, with a power-optimized AFE with 12-bit ADC [29], the energy consumption is tremendously reduced from $20\,\mu\mathrm{J/Sample}$ to $52\,\mathrm{nJ/Sample}$. Another important system design decision to further enhance the system battery lifetime is whether the seizure detection pipeline is executed in the cloud or at the edge (on-chip). In the former case there is no on-chip signal analysis; the raw EEG data is transmitted to the end-point where the seizure detection is performed. In the latter case seizure detection is performed on-chip, and only an alarm is sent to a wireless end-point. For medical devices, autonomous operation without depending on wireless connectivity is strongly preferred from a system reliability and data security aspect.

TABLE III
Overview of medical-grade ambulatory EEG data logger and two state-of-the-art EEG data acquisition solutions.

	TMSi Mobita	**TI ADS1299**	**Xu *et al.* [29]**
Solution	Medical-grade EEG logger	General-purpose EEG AFE+ADC	Power-optimized EEG AFE+ADC
Channels	32	8	16
ADC (sample rate)	24-bit $\Sigma\Delta$ (2kS/s/chn)	24-bit $\Sigma\Delta$ (0.25kS/s/chn)	12-bit SAR (2kS/s/chn)
Power diss.	$<1.5\,\mathrm{W}$	$39\,\mathrm{mW}$	$1.7\,\mathrm{mW}$
Energy/sample, Battery life	Wireless: $\approx 8\,\mathrm{h}$ Flash: $\approx 18\,\mathrm{h}$	$20\,\mu\mathrm{J/Sample}$	$52\,\mathrm{nJ/Sample}$

TABLE IV
Wearable EEG processing system energy breakdown per epoch (21 channels $\times$ 256 samples epoch size). Numbers between parentheses: energy consumption of the optimized seizure detection pipeline on the BrainWave processor.

	BrainWave system energy (mJ/epoch)	
System component	**Cloud processing**	**Edge processing**
AFE + 12-bit ADC [29]	0.14	0.14
RISC-V MCU [31]	not used*	25.76† (0.264)
Wireless Radio - Tx [32]	2.96	≈ 0
Data logger - SD card	0.34	0.34
Total	3.44	≈ 26.26 (0.76)

* Energy cost of baseline seizure detection pipeline in cloud ignored.
† Energy cost of baseline seizure detection pipeline on RISC-V core.

Table IV lists an estimated energy breakdown based on a representative system for these two operation scenarios. This system consists of a power-optimized 12-bit EEG analog front-end with an energy efficiency of $52\,\mathrm{nJ/sample}$ and a Dialog DA14580 SoC [32] with integrated Bluetooth Low Energy (BLE) transceiver for wireless communication. This SoC consumes approximately $4.7\,\mathrm{mA} \cdot 3\,\mathrm{V} / 128\,\mathrm{kbit/s} = 110\,\mathrm{nJ/bit}$ (payload) in transmission mode. The energy consumption of the baseline seizure detection pipeline on a RISC-V micro-controller (MCU) is approximately $25.76\,\mathrm{mJ/epoch}$[2]. Finally, for data logging we consider an external SPI flash memory. For our application a $4\,\mathrm{GB}$ SD card should be able to record the digitized EEG data up to a week. We estimate the SD write energy consumption to be approximately $12\,\mathrm{nJ/bit}$[3] (assuming typical current and latency values and power-down modes).

It can be observed that on-chip processing leads to a large reduction in wireless traffic energy. This is in line with other works that generally performs digital signal processing and data reduction on-chip to minimize wireless communication [9], [13], [33]. However, if we compare both cases in terms of energy per classification, the Table suggests that edge processing is less attractive than cloud processing. Also, the baseline classification pipeline (Table II) prohibits real-time operation on the RISC-V MCU, which causes the system to potentially miss seizure events. In the remainder of this article we will focus on reducing the energy and throughput bottleneck in the digital BE, and demonstrate that edge processing can greatly outperform cloud processing. More specifically, the energy cost per epoch will be

[2]Based on chip measurements; $V_{logic} = 0.6\,\mathrm{V}$, $V_{mem} = 0.9\,\mathrm{V}$, $38.6\,\mathrm{pJ/cycle}$, $32\,\mathrm{MHz}$, $666.85\,\mathrm{Mcycles/epoch}$.

[3]Estimate based on 'SanDisk SD Card OEM Product Manual (April 2011)'.

TABLE V
IMPACT OF ALGORITHMIC AND SW OPTIMIZATIONS ON AVERAGE RUN-TIME AND MEMORY USAGE ON RISC-V PROCESSOR ($N = 256$).

Algorithm + optimizations	**Cycles** ($\times 10^6$)	**Memory** (B)	**Speedup**
BPF baseline (direct-form 1)	0.084	0	1.0×
BPF cascade of Biquads	0.025	0	3.4×
VG ND baseline (Alg. 1)	26.88	0	1.0×
+ min. slope tracking [34]	1.57	0	17.1×
+ divide-and-conquer [35]	stack overflow	-	-
+ eliminate recursion	0.30	$4N$	91.1×
+ limit partition stack depth	0.30	$4\log_2(N)$	91.2×
+ eliminate divisions	0.18	$4\log_2(N)$	156.6×
SampEn* baseline (Eqn. 3)	1.93	0	1.0×
+ fusion/symmetry opt. [36]	0.54	0	3.6×
+ early stopping [37]	0.34	$4N(+4N)^\dagger$	5.8×
+ inline early stopping	0.37	$2N(+4N)^\dagger$	5.3×

* Comparable run-time improvements were obtained for ApEn. Subvector length $m = 2$ and threshold $r = 0.2 \cdot \sigma(\vec{x})$ was used.
† Additional memory is required for ApEn calculation.

reduced from $26.26\,\mathrm{mJ/epoch}$ to $0.76\,\mathrm{mJ/epoch}$, while executing the application in real-time. This gives a $4.5\times$ energy reduction compared to the cloud solution.

IV. ALGORITHMIC AND SW OPTIMIZATIONS

In this section we detail the algorithmic optimizations that were applied to improve the run-time and memory-efficiency of the main bottlenecks in the non-convulsive seizure detection algorithm that was previously introduced.

1) Band-pass filter: The baseline implementation of the 10th order (21 taps) Butterworth filter on the RISC-V core takes approximately 16 cycles per filter tap. The coefficients require at least 24 bits of integer precision to maintain filter stability, due to the sharp cut-off point near 0 Hz, which results in input-coefficient products that exceed 32-bit. As such, 64-bit multiplications and accumulations are emulated by the 32-bit data-path of the RISC-V core, which require 2–4 instructions per operation. To optimize the run-time of this filter, the filter was implemented using a cascade of 5 second-order sections, which requires only 12 bits of coefficient precision. Using this approach the throughput of the online BPF is improved by $3.4\times$ on the RISC-V core, as is depicted in Table V. The filter operates in-place and does not require additional memory while processing.

2) Visibility Graph Node Degree: The baseline VG ND algorithm for time-series is an $\mathcal{O}(N^3)$ algorithm that counts for every sample in an N-length vector $\vec{x}$ (e.g. a digitized EEG epoch) how many other samples within the same vector satisfy the visibility condition. For every pair of samples (i.e. (x_i, x_j)) it checks if no samples between indices i and j block the visibility. This implementation is slow as the same combinations of samples are checked multiple times (e.g. for pairs (x_i, x_j) and (x_i, x_{j+1}) points between i and j are checked twice). Since we are only interested in the ND vectors of the NVG, HVG and DVG, we skip the explicit adjacency matrix calculations and fuse the calculation of the ND vectors with the visibility checks to save memory. Unlike popular signal processing kernels like FFT and MatMul, there are no reference implementations of the VG ND degree kernel. Therefore we have provided a description of the considered baseline implementation in Alg. 1.

A faster approach to compute the VG of a time-series with $\mathcal{O}(N^2)$ time complexity was introduced in [34]. The algorithm tracks the minimum slope and eliminates duplicate visibility checks of points between i and j. When this approach is integrated in the VG ND computation, the run-time is improved by $17.1\times$, on average. More recently, a divide-and-conquer (D&C) algorithm with $\mathcal{O}(N \log_2 N)$ time complexity (on average) was proposed [35]. This algorithm works based on the observation that all nodes to the left of the largest node in $\vec{x}$ are not visible for nodes to the right and vice versa. This principle is then recursively applied to both partitions. We apply the proposed optimizations to the VG ND feature, and optimize the run-time and memory-efficiency on the BrainWave processor. A direct implementation with the slope-tracking and D&C optimizations leads to a stack overflow on our RISC-V processor. As such, we reduce the stack usage and call-recursion overhead by implementing the recursive algorithm as an iterative algorithm. Despite its more complex control flow, it turned out to be almost $4\times$ faster (on average) compared to the $\mathcal{O}(N^2)$ algorithm for 256-point EEG vectors, with a break-even point around 128 samples. To save memory the partition stack depth is limited from N to $\log_2(N)$ by prioritizing exploration of the smallest partition first. Finally, as the iterative hardware divider on the RISC-V processor is quite slow, thus divisions in the visibility condition are eliminated by re-arranging, which results in a total speedup of $156.6\times$ over the baseline implementation with limited memory overhead.

3) Non-linear features - SampEn: The baseline implementation to compute the SampEn feature follows from the definition. We consider a time series with N samples: $\vec{x} = \{x_1, x_2, ..., x_N\}$. From this sequence we extract $N - m + 1$ partially overlapping subvectors of length m, where $X_i^m = \{x_i, x_{i+1}, ..., x_{i+m-1}\}$. We define $P_i(m, r)$ as the likelihood of any subvector to be similar to X_i^m (excluding

Algorithm 1 Baseline normal and horizontal visibility graph node degree calculation for N-length input $\vec{x}$. Θ^N and Θ^H correspond to the node degree vectors for the normal and horizontal visibility graphs, respectively.

Initialize $\Theta^N = \Theta^H = \{1, 2, ..., 2, 1\}$. ▷ minimum visibility of any node are his neighbours
for $i = 0 \rightarrow N - 2$ **do**
  for $j = i + 2 \rightarrow N$ **do**
    $C = \{x_k \mid x_k \geq \min(x_i, x_j)$ for $i < k < j\}$ ▷ possible blocks
    if $C = \emptyset$ **then** ▷ horizontal visibility
      $\Theta_i^H = \Theta_i^H + 1;\ \Theta_j^H = \Theta_j^H + 1$
    if $\forall x_k \in C \left\{ \frac{x_k - x_i}{k - i} < \frac{x_j - x_i}{j - i} \right\}$ **then** ▷ normal visibility
      $\Theta_i^N = \Theta_i^N + 1;\ \Theta_j^N = \Theta_j^N + 1$

Algorithm 2 Improved Sample Entropy algorithm for N-length input $\vec{x}$, sub-vector length m and threshold r with optimizations of [36], [37] and inline early stopping.

Initialize similarity counters $A = B = 0$.
$ind = \{$sort indices of X_0^0 to X_{N-m+1}^0 in ascending order$\}$.
for $i = ind_0 \rightarrow ind_{n-m}$ **do**
  for $j = ind_{i+1} \rightarrow ind_{n-m+1}$ **do** ▷ skip redundant checks
    if $x_j > x_i + r$ **then** ▷ inline early stopping
      break;
    if $\max|X_i^m - X_j^m| \leq r$ **then**
      $B = B + 1$ ▷ $\phi(m, r) = B\,/\,(N - m + 1)$
    if $j + m < N$ **and** $\max|X_i^{m+1} - X_j^{m+1}| \leq r$ **then**
      $A = A + 1$ ▷ $\phi(m + 1, r) = A\,/\,(N - m)$
Compute $SampEn(m, r)$ using Eqn. 3.

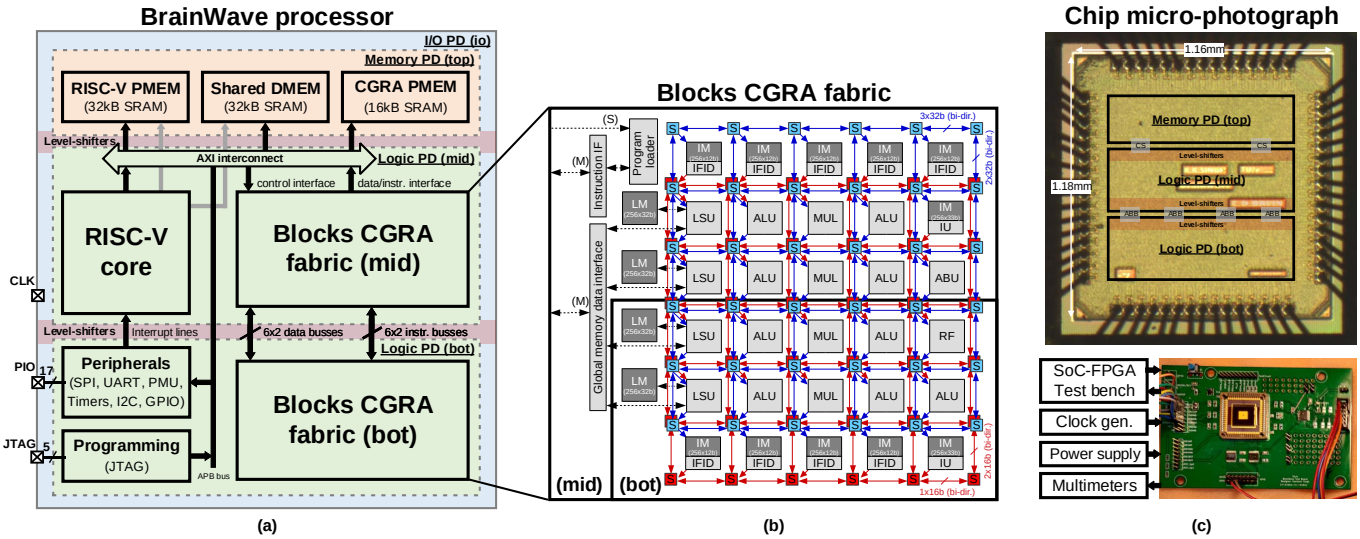


Fig. 5. (a) Architecture overview of BrainWave processor with different power domains. (b) Detailed overview of instantiated CGRA. Note that the CGRA is split between the two logic power domains to balance the activity in voltage-stacked mode. (c) Chip die photograph and test setup.

self-similarity):

$$P_i(m,r) = (N-m)^{-1} \sum_{j=1,\, j\neq i}^{N-m+1} C(X_i^m, X_j^m) \tag{1}$$

where the similarity condition for threshold r is defined as:

$$C(X_i^m, X_j^m) = \begin{cases} 1, & \text{if } \max|X_i^m - X_j^m| \leq r. \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Sample Entropy is now computed as follows [38]:

$$SampEn(m,r) = \ln \phi(m,r) - \ln \phi(m+1,r), \tag{3}$$

$$\text{where } \phi(m,r) = \frac{\sum_{i=1}^{N-m+1} P_i(m,r)}{N-m+1} \tag{4}$$

The run-time of SampEn is determined by the number of similarity condition checks. To reduce the execution time several basic optimizations are typically used [36]. More specifically, the computation of $\phi(m,r)$ and $\phi(m+1,r)$ can be fused and redundant similarity checks can be skipped due to symmetry in the distance function. These optimizations yield a $3.6\times$ speedup, over the baseline implementation.

A faster algorithm for SampEn (and ApEn) was proposed by Pan *et al.* [37]. By performing the vector similarity check in Equation 1 in ascending order, i.e. sorted based on the first element of every sub-vector, an early stopping rule can be constructed. The main idea is that if we iterate j in ascending order until X_j^m becomes dissimilar to X_i^m, then the remaining sub-vectors will also be dissimilar and can be skipped. The number of similarity comparisons is heavily reduced from $(N-m)^2/2$ to $c{\cdot}(N-m)$, where c is a data-dependent constant. Compared to the baseline, we report a speedup of $5.8\times$. The memory usage of this implementation is $2N$ bytes to store the sorted indices and $2N$ bytes to store the stopping condition for every X_i^m.

At the cost of a small performance penalty (i.e. 0.34 to 0.37 Mcycles, Table V) we can reduce the memory usage significantly, by computing the the early-stopping condition on the fly, as depicted in Algorithm 2. Note that this implementation can be further accelerated by exploiting SIMD-parallelism to compute multiple independent EEG channels in parallel, if the early stopping condition of individual EEG channels is fused in a logical-AND tree. Since threshold r scales with the standard-deviation of the input vector, the early-stopping points of different input vectors are relatively close to each other. Some preliminary experiments indicated that the instruction energy reduction due to SIMD-parallelism is mostly offset by the throughput penalty, which makes this parallelization strategy only useful for throughput-constrained applications.

All discussed optimizations for SampEn were also applied to accelerate ApEn on the RISC-V core and CGRA. ApEn is slightly more expensive to compute, since it requires more natural logarithm calls, but the run-time is still dominated by the number of vector similarity checks and the sort.

V. BrainWave Processor Architecture

A. Architecture overview

An overview of the implemented BrainWave processor is depicted in Fig. 5a. The SoC comprises a RISC-V core and CGRA and contains 80 kB of SRAM memory to store the RISC-V core and CGRA programs, the EEG history buffers, pre-trained classification model and the necessary scratchpad space to perform computations. A detailed overview of the instantiated CGRA is depicted in Fig. 5b. It consists of 21 FUs and 11 private 256-word standard-cell instruction memories (IM) that can be connected to one or more FUs over the instruction network. Inputs and outputs of FUs can be interconnected over the data network to enable spatial computing and register file bypassing. 4 private data memories (DM) of 1 kB each are used to facilitate energy-efficient local data movement.

B. Physical implementation

The BrainWave processor was fabricated in a Foundry 28-nm FDSOI high-density 8-track low-V_{th} (LVT) standard-cell library with Foundry 6T-SRAM memories. The three power domains can externally be reconfigured for flat or voltage-stacked operation mode. The level-shifters between these power domains are designed to work in both configurations. The chip is designed to operate in near-V_{th} (0.4 V–0.5 V), but can run up to 100 MHz at nominal voltages (in flat mode). The top stack with the SRAMs operates at 0.9 V (V_{top}–V_{mid}). The middle stack with the RISC-V core and half

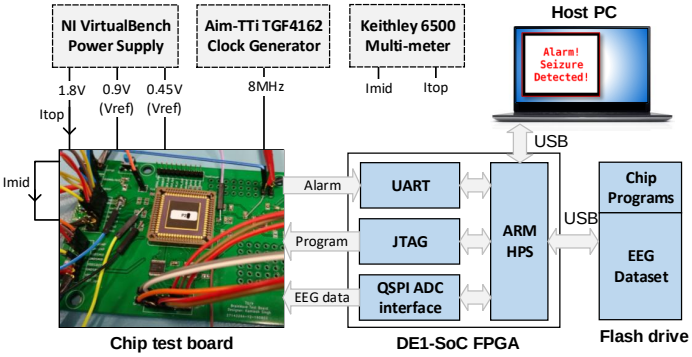


Fig. 6. Testing and measurement setup of BrainWave processor in voltage-stacked mode for real-time non-convulsive seizure detection.

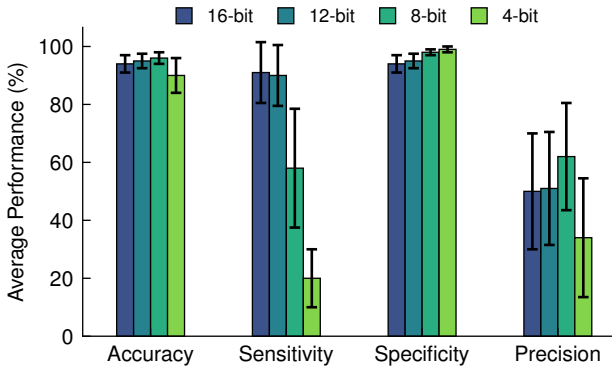


Fig. 7. Seizure detection performance (with standard error) for different EEG quantization levels (using a high-precision pre-trained classification model).

TABLE VI
OVERVIEW OF EVALUATED KERNELS.

Kernel	Description
Band-pass filter[1]	5-stage Biquad band-pass filter
DWT decomposition[1]	full db4 wavelet decomposition
Index Sort[1]	mergesort that sorts array indices
Sim. checking (ApEn)[1]	vector comp. loop with early exit
Sim. checking (SampEn)[1]	vector comp. loop with early exit
NVG/HVG Node Degree[1]	slope-following NVG/HVG algorithm
NVG/HVG Node Degree (D&C)	+ D&C optimization
FFT	256-point 16-bit fixed-point DIF cFFT
MatMul	32×32 16-bit fixed-point MatMul

[1] Tested on 2 channels × 256-point vector to account for parallel processing on the CGRA.

TABLE VII
SEIZURE DETECTION PERFORMANCE AFTER QUANTIZATION OF FEATURE EXTRACTION STAGE (QUANTIZED) AND THE CLASSIFICATION AND POST-PROCESSING STAGES (EMBEDDED).

Version	Accuracy	Sensitivity	Specificity	Precision
Baseline (float32)	94%	91%	94%	50%
Quantized[1]	95%	85%	96%	53%
Embedded[1,2]	94%	87%	95%	54%

[1] Feature extraction mapped onto 32-bit integer data path.
[2] Classification and post-processing mapped onto 32-bit integer data path.

of the CGRA, operates at 0.45 V (V_{mid}–V_{bot}). The bottom stack contains the remaining half of the CGRA, peripherals, and IO-cell control logic, operating at 0.45 V (V_{bot}). The partitioning of the logic domains is based on a post-synthesis leakage power distribution. The voltage-stacked system is designed such that the top domain supplies a stable current for the logic domains to operate, as the system cannot supply additional current (only sink excess current). The idea is that when the logic operates in near-threshold, the constant memory leakage current dominates the switching current, thereby becoming easier to stabilize, The design and evaluation of the level-shifters, current sink and adaptive body-biasing controllers is detailed in [24].

C. Measurement setup and demonstration

The chip die is packaged and put onto a test PCB. The test PCB includes decoupling capacitors and level-shifters to connect to an SoC-FPGA board. The SoC-FPGA board is used to program and control the chip. The chip is programmed via a JTAG interface, and can interface and communicate with the SoC-FPGA using the SPI and UART interfaces. The testing and measurement setup is illustrated in Fig. 6. The PCB contains jumpers to configure the power domains in flat (conventional) mode or voltage-stacked mode. In flat mode the chip requires an external clock signal and power supply which provides V_{logic}, V_{mem} and V_{io}, while in voltage-stacked mode V_{top} and two reference voltages are required to balance V_{mid} and V_{bot} using the ABB controllers and current sink. In a future design the clock generator and reference voltage generation circuitry should be embedded on-chip, but for now their energy impact is ignored.

To measure the throughput and energy-efficiency of the system, an external multi-meter monitors the power dissipation of the V_{logic} and V_{mem} rails in flat mode (and compensates for the system voltage conversion losses, as indicated in Fig. 4). The execution time of applications and kernels is measured using on-chip timers. In voltage-stacked mode the total system power dissipation of the V_{in} voltage rail is monitored, as well as the power dissipation of the memory and logic rails (V_{top} and V_{mid}). Table VI provides a brief summary of the evaluated kernels on the CGRA. Most of these kernels were used in the seizure detection application. A fixed-point FFT and MatMul kernel were added to showcase the CGRA flexibility and allow for easier comparison with other works.

We have implemented the complete non-convulsive seizure detection on the BrainWave processor. The seizure detection accuracy on the full dataset is evaluated off-line. The functional equivalence between the off-line fixed-point C-implementation and the real-time implementation that runs on the BrainWave processor was manually verified over 100 epochs. The implementation of this algorithm requires approximately 20 kB of data memory (11 kB for the EEG history buffers (2.56 s × 21 channels × 2 B/sample), 1 kB for the Hanning window, 3 kB for the BPF delay lines, 5 kB for the RUSBoost classifier model, 2 kB is reserved for the stack, and approximately 10 kB for auxiliary scratchpad memory. The RISC-V application requires approximately 24 kB of instruction memory. If we use the CGRA for offloading, an additional 16 kB is required to store the CGRA kernels.

A patient-specific classification model is trained and evaluated using a leave-one-subject-out cross-validation approach [5]. We split the EEG data chronologically, with the first 25% of the left-out subject being used to calibrate the patient-specific threshold, and the final 75% being used as a test set for evaluation of the algorithm accuracy. The test sets include 164 out of the 316 seizure events that were present in the complete dataset. Given the test set *t*rue/*f*alse *p*ositive/*n*egative detections, we report the average event detection accuracy on the test set of all subjects in terms of Accuracy, calculated

TABLE VIII
BLOCKS CGRA PERFORMANCE EVALUATION ON KERNELS FROM TABLE VI (V_{logic} = 0.7 V, V_{mem} = 0.95 V, 50 MHz).

	Execute on CGRA						Execute on RISC-V core	
Kernel	**#instr**	**Ops**	**Cycles**	**FUs active**	**Utilization[1]**	**Energy (μJ)**	**Cycles**	**Energy (μJ)**
Band-pass filter	34634	66756	9221	19/21	38%	0.61	49276	2.18
DWT decomposition	31736	49654	7671	19/21	34%	0.36	46605	2.20
Index Sort	93340	166165	26443	16/21	32%	0.95	114982	5.03
Similarity checking (ApEn)[2]	845325	1500970	162106	16/21	42%	6.68	702015	30.32
Similarity checking (SampEn)[2]	556882	946121	104438	16/21	41%	4.65	622805	27.12
NVG/HVG Node Degree	1532324	2994316	297915	18/21	48%	16.59	1357594	59.49
NVG/HVG Node Degree (D&C)[2]	608121	660603	164957	11/21	19%	5.84	355696	15.71
FFT	29220	46318	9014	17/21	28%	0.38	30943	1.46
MatMul	64872	131018	23576	16/21	34%	1.43	133193	6.10

[1] Utilization = Ops / (FUs · Cycles). FUs = 21 for the instantiated CGRA.
[2] Considering an input data vector that is close the the average run-time (as depicted in Table IX).

TABLE IX
AVERAGE APPLICATION RUN-TIME (IN CYCLES $\times 10^6$) BREAKDOWN ON THE BRAINWAVE PROCESSOR.

	Baseline	**Optimized**	
Stage	SW-only	SW-only	SW+CGRA
Band-pass filter	0.89	0.26	0.05
Non-linear features[1]	40.61	7.80	1.54
Index sort	0.00	1.22	0.29
Similarity checking (SampEn)	40.60	6.58	1.25
Feature calculation	0.01	0.01	<0.01
Morphological features[1]	624.05	28.35	7.22
NVG/HVG Node Degree	506.05	3.28	1.61
Index sort[2]	0.00	3.74	0.87
Similarity checking (ApEn)[2]	115.92	19.24	4.51
Feature calculation[2]	2.09	2.09	0.23
Time-Frequency features	0.52	0.52	0.09
DWT decomposition	0.49	0.49	0.08
Feature calculation	0.03	0.03	0.01
Remaining stages (Table II)	0.78	0.81	0.57
Total (Speedup)	666.85 (1×)	37.74 (17.7×)	9.48 (70.34×)

[1] Subvector length $m = 2$ and threshold $r = 0.2 \cdot \sigma(\vec{x})$ was used.
[2] Executed for NVG/HVG/DVG Node Degree vectors (3× per channel).

as $(tp+tn)/(tp+tn+fp+fn)$, Sensitivity (Recall), calculated as $tp/(tp+fn)$, Specificity, calculated as $tn/(tn+fn)$, and Precision, calculated as $tp/(tp+fp)$.

VI. SYSTEM MEASUREMENTS AND EVALUATION

A. EEG data precision and seizure detection accuracy

The average high-precision baseline performance of the NCSE seizure event detection algorithm on the test set of all 10 patients is summarized in Table VII. On average the baseline implementation is able to detect 91% of all seizures with an average precision of 50% which is comparable to the previously published results in [5].

To ensure that that the system in Section III is representative for our task, we have investigated the algorithm accuracy degradation when the dynamic range of the EEG data is reduced. It follows from the results in Fig. 7 that the dataset can be quantized down to 12 bits without a significant performance degradation. Below 12-bit quantization, the system starts to miss many seizure events, which is not acceptable. These results are in line with state-of-the-art

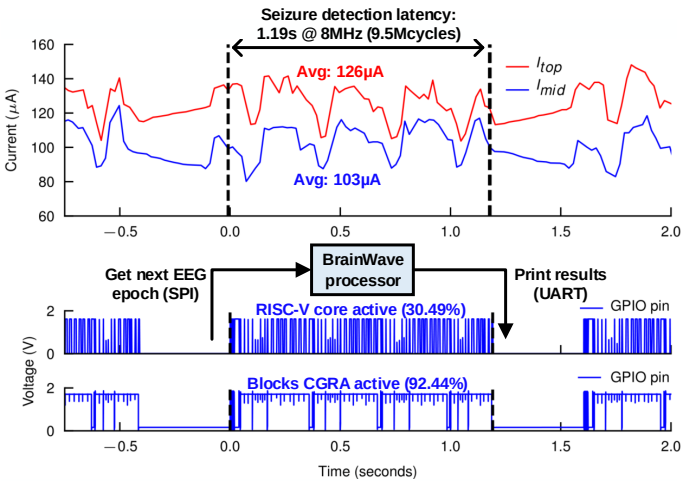


Fig. 8. Measured current (top) and processor activity (bottom) in voltage-stacked mode when running the seizure detection application on the BrainWave processor (SW+CGRA, V_{top} = 1.77 V, V_{mid} = 0.92 V, V_{bot} = 0.46 V).

EEG-based seizure detection systems [9], [13], [29], which typically consider 10 to 12 bits of dynamic range sufficient.

To enable deployment on the BrainWave processor, the classifier and post-processing stage were trained using a quantized fixed-point feature extraction stage. After training, the resulting classification model and patient-specific thresholds were quantized as well to have an end-to-end quantized NCSE seizure detector. The resulting accuracy is depicted in Table VII.

In general, the performance of the embedded algorithm is very similar to the high-precision reference. On average, the embedded version is able to detect 87% of all seizures with an average precision of 54%.

B. Kernel-level speedup and energy savings using the CGRA

We measure the speedup and energy-efficiency improvements when offloading the complex EEG kernels to the Blocks CGRA. To get a stable measurement for these individual kernels we continuously execute the kernel. All CGRA programs were manually converted from C to assembly code, as a compiler is currently not available.

It follows from the results in Table VIII that using the Blocks CGRA for kernel off-loading leads to a speedup of 2.2–6.1× (4.7× on average) and energy savings of 2.7–6.1× (4.4× on average), compared to highly tuned implementations on the RISC-V core

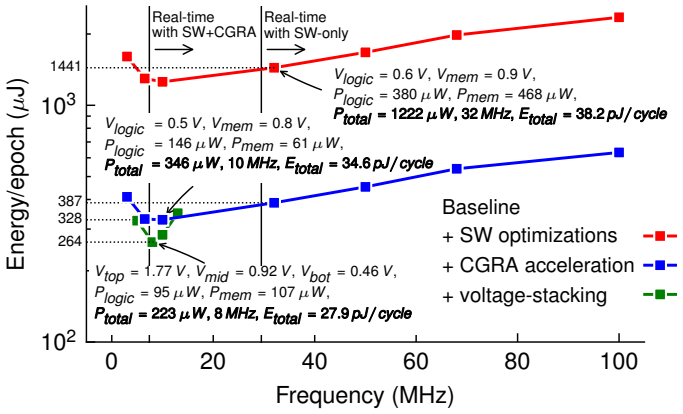


Fig. 9. Measured energy per epoch versus maximum clock frequency for different configurations when running the optimized seizure detection application on the BrainWave processor.

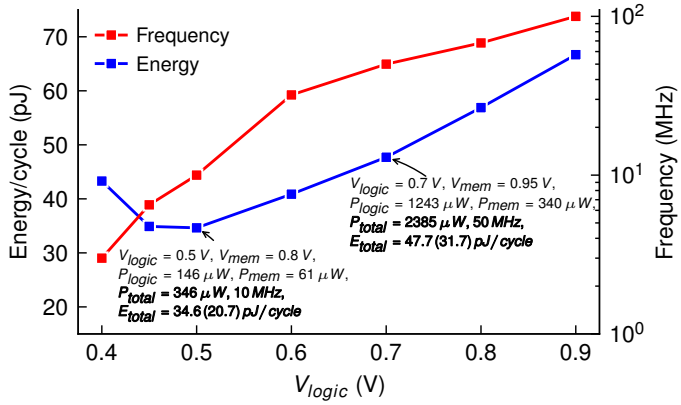


Fig. 10. Measured energy per cycle and maximum clock frequency versus logic voltage when running the optimized seizure detection application on the BrainWave processor (SW+CGRA) in flat mode. The shown power and energy numbers between brackets is without conversion losses.

[21]. A large fraction of the speedup can be attributed to parallel processing. While the RISC-V core executes close to 1 instruction per cycle (IPC), the CGRA mappings have an IPC between 4–10 (7 on average). As a consequence, the CGRA utilization ranges from 19% to 42% (35% on average), which is competitive with other CGRAs, as was recently reviewed in [39].

The CGRA code was written such that most mappings can calculate features for 2 EEG channels in parallel, thereby exploiting SIMD-parallelism, which shows from the ratio between FU operations and instructions. Most kernels are able to use most of the available FUs. One exception is the D&C algorithm for VG Node Degree calculation, which has data-dependent branches that cannot be parallelized in a meaningful way, and does not have enough instruction-level parallelism to utilize all units. However, it is still faster and more energy-efficient than the slope-following algorithm. Therefore we consider it for our performance evaluation.

In general the results indicate that the Blocks CGRA is a suitable platform for acceleration of a wide variety of signal processing and EEG processing features. A future version of the Blocks CGRA will be able to exploit more parallelism by utilizing multi-processing for data-dependent features, consist of a larger CGRA fabric with more FUs to support wider SIMD-processing, and will be able to disable unused units completely to minimize leakage.

C. Application-level speedup and energy savings

1) Real-time embedded seizure detection: A detailed run-time breakdown which shows the throughput improvements of the algorithmic optimizations and CGRA off-loading is depicted in Table IX. Using the algorithmic optimizations that were presented in Section IV leads to baseline throughput improvements of 17.7×, without compromising on seizure detection accuracy. Compared to the optimized SW-only version, the active time of the RISC-V core is reduced by approximately 92.25% by off-loading work to the CGRA (SW+CGRA version). To further optimize the seizure detection latency, the RISC-V core operates in parallel with the CGRA. This reduction in run-time through latency hiding is visible in Table IX for the 'Feature calculation' and 'Remaining stages' stages, which are performed on the RISC-V core. In this way 73.75% of the remaining cycles on the RISC-V core can be hidden by means of parallel processing. Also CGRA reconfiguration overhead (approximately 1.89% of the epoch time) can be hidden.

By off-loading computationally intensive parts of the pipeline to the CGRA, an additional 3.98× speedup is obtained. This is expected, given that the primary bottleneck kernels have slightly more than 4× speedup on the CGRA (see Table VIII). These efforts result in a final speedup of 70.34× from the initial baseline SW-only implementation. For real-time operation of the algorithm a minimum clock rate of 37.74 Mcycles / 1.28 s = 29.48 MHz and 7.41 MHz are required, for the optimized SW-only and SW+CGRA versions, respectively.

The measured run-time behavior of the SW+CGRA version is depicted in Fig. 8 (bottom). Two GPIO pins were toggled when the RISC-V core and CGRA were active. To ensure correctness, results are printed over UART after every epoch, which causes some delay that will not be present in a real system. It follows that the total active time of the RISC-V core and the CGRA is approximately 30.49% and 92.44% of the epoch time, respectively. Idle components are explicitly clock-gated to save dynamic power.

2) Near-threshold operation and voltage-stacking: We execute the seizure detection application on the BrainWave processor and measured the average power dissipation over several epochs (i.e. the total current supplied to V_{logic} and V_{mem}), as depicted in Fig. 8. We sweep the logic voltage from 0.4 V to 0.9 V and find the maximum operation frequency while the application is still operating correctly (i.e. EEG data is correctly sampled over QSPI interface and feature values and classifier output match reference). Since the system is designed for near-threshold operation, the system frequency is constrained by the logic, as the voltage of the SRAMs can only be scaled down to 0.8 V.

Fig. 10 shows the measured energy per cycle and clock frequency in flat mode. It follows from the figure that the MEP is at 0.5 V and corresponds to 20.7 pJ/cycle before or 34.6 pJ/cycle after taking conversion losses into account (computed as in Fig. 3a). Compared to the highlighted nominal operation point (V_{logic} = 0.7 V, V_{mem} = 0.95 V), an energy per cycle improvement of 35% is observed before, and 27% after conversion losses were taken into account. These numbers indicate that a significant fraction of the energy is spent in the SRAM memories and that voltage-conversion overhead reduces the energy-efficiency significantly, as can be seen from the relative difference between the sum of P_{logic} and P_{mem} over the total system power P_{total} (excluding IO).

We further investigate the energy impact of the considered architectural and circuit-level optimizations. Fig. 9 shows the measured

TABLE X
OVERVIEW OF STATE-OF-THE-ART PROGRAMMABLE ULP BIO-MEDICAL/IOT PROCESSORS AND SYSTEMS.

Metric	This work	ULP-SRP [15], [25]	PULPv1 [12], [40]	Blackghost [26]	S. Song *et al.* [28]	Catena [27]
Year	2021	2013	2016	2018	2019	2020
Technology	28-nm FDSOI	40-nm LP	28-nm FDSOI	28-nm FDSOI	55-nm LP	65-nm LP
Chip area	$1.21\,\mathrm{mm}^2$	$6\,\mathrm{mm}^2$	$3\,\mathrm{mm}^2$	$9\,\mathrm{mm}^2$	$18.49\,\mathrm{mm}^2$	$6.5\,\mathrm{mm}^2$
SoC arch.	RISC-V MCU + CGRA	VLIW MCU + CGRA	4-core RISC-V MCU	ARM M0 MCU + DSP + Vision acc.	ARM M4F MCU + FFT/MP/SRC acc.	16-core spatial processor
On-chip SRAM	**80 kB (foundry 6T)**	512 kB (foundry 6T)	36 kB	320 kB (custom 8T)	192 kB	16 kB (custom 6T)
Algorithms	**Non-convulsive seizure detection**	Heart-beat detection	Convulsive seizure detection	Face recognition application	E-health monitoring application	Fixed-point MatMul micro-benchmark
Voltages	logic: 0.46 V, memory: 0.85 V	logic: 0.5 V, memory: 0.7 V	logic: 0.5 V, memory: -	logic: 0.55 V, memory: 0.7 V	logic: 0.8 V, memory: 1.2 V	logic: 0.54 V, memory: 0.77 V
Frequency	8 MHz	7 MHz	6.4 MHz	50 MHz	10 MHz	1.56 MHz
Power diss.	223 µW	213.1 µW	419 µW†	4 mW	331 µW†§	356.8 µW†
Energy/cycle	27.9 pJ	30.4 pJ	65.5 pJ†	80 pJ	33.1 pJ†§	228 pJ†
System efficiency	**90.6 %**	70–86 %	N/A	70 %	80 %	N/A
FFT throughput	26 MOPS @ 5 MHz	33 MOPS	25 MOPS‡	800 MOPS*	N/A	7.8 MOPS @ 1 MHz
FFT efficiency	**164.9 MOPS/mW**	59.4 MOPS/mW	60 MOPS/mW‡	200 MOPS/mW*	N/A	46.2 MOPS/mW†
Energy/FFT	0.28 µJ	0.66 µJ	N/A	N/A	N/A	2.31 µJ†

† Power dissipation and energy per cycle is excluding voltage-conversion losses.
§ Power dissipation and energy per cycle is excluding AFE, BLE radio and PLL power contribution (only processor, accelerators and memory).
* Throughput and efficiency calculated using best-case DSP throughput of 16 operations per cycle (and 4 mW @ 50 MHz).
‡ Throughput and efficiency values taken from [40] while executing MatMul benchmark on all 4 cores, excluding voltage-conversion losses (optimistic estimate).

TABLE XI
SUMMARY OF SYSTEM-LEVEL ENERGY SAVINGS BASED ON BRAINWAVE PROCESSOR CHIP MEASUREMENTS.

Version	Energy per epoch	Energy savings
Baseline - SW-only[1] (Section III)	25.76 mJ	-
+ SW optimizations	1441 µJ	94.4% (1×)
+ CGRA acceleration	387 µJ	73.1% (3.7×)
+ Near-threshold operation	328 µJ	15.2% (4.4×)
+ Voltage-stacking	264 µJ	19.5% (5.4×)

[1] This version cannot run in real-time on the BrainWave processor.

energy per epoch while running the seizure detection application. The optimized SW-only version requires at least 29.48 MHz to operate in real-time and can therefore not run at the MEP. Off-loading work to the CGRA improves the average energy per epoch from 1441 µJ to 387 µJ. The speedup that the CGRA provides allows us to operate at the MEP, thereby lowering the energy per epoch further to 328 µJ. When the BrainWave processor is configured in voltage-stacked mode, its average energy per epoch is reduced to 264 µJ at the MEP of 27.9 pJ/cycle, 8 MHz. In this operating condition the BrainWave processor obtains a state-of-the-art system efficiency of $(95\,\mu\mathrm{W} + 107\,\mu\mathrm{W})/223\,\mu\mathrm{W} = 90.6\%$, while executing the seizure detection application in real-time. For completeness, the corresponding current measurements are also depicted in Fig. 8.

Finally, Table XI summarizes the energy reduction of the individual optimizations. The software optimizations on the RISC-V processor resulted in a 17.7× speedup compared to the baseline SW-only implementation, which translates in an energy-reduction of 94.4%. This result stresses the importance of multi-level optimization, as the algorithmic optimizations have a tremendous impact on energy consumption. The use of the Blocks CGRA for energy-efficient off-loading improves the efficiency by another 73.2% at the application-level, primarily caused by the reduction in instruction overhead, SIMD-processing and explicit register file bypassing. This result indicates that the Blocks CGRA is an effective solution for bio-medical signal processing, due to its ability to execute a wide diversity of workloads efficiently Using the CGRA leads to an 3.98× speedup which allows for voltage-frequency scaling of the logic from 0.7 V to 0.5 V, which saves another 15.2%. Finally, voltage-stacking is exploited to minimize voltage-conversion overhead, resulting in an additional 19.5% savings at the system level.

D. System-level evaluation

In the wearable EEG processing system that was presented in Section III, the digital signal processing was a significant energy bottleneck. We have performed a multi-level energy optimization at the system-level, algorithm-level, architecture-level and circuit-level. As such, the measured real-time seizure detection energy of the BrainWave processor is reduced from 25.76 mJ/epoch to 264 µJ/epoch, without compromising on seizure detection accuracy. Compared to the baseline SW-only edge processing case in Table IV, the system-level energy is reduced from approximately 26.26 mJ/epoch to 0.76 mJ/epoch, a 35 × energy reduction. Compared to the cloud processing case, which consumed approximately 3.44 mJ/epoch, a 4.5 × energy reduction is obtained at the system-level. Using a typical R2032 coin-cell battery with a capacity of 600 mWh, the BrainWave system could potentially perform real-time non-convulsive seizure detection and data logging on 21 EEG channels for up to 6 weeks on a single charge.

VII. COMPARISON WITH PRIOR ART

Many integrated systems towards wearable automated detection of epileptic seizures have been proposed over the last decade. [41] provides a recent overview. Most works focus on the optimization of traditional algorithms for detection of convulsive epileptic seizures, based on spectral, time-frequency and entropy features [6]–[10], [12], [13]. Systems optimized for non-convulsive seizure detection require more computationally demanding algorithms [5], [14]. The differences in applications prohibit a fair comparison in terms of seizure detection accuracy and energy per classification.

In contrast to prior EEG monitoring systems, we keep our system flexible with a programmable CGRA accelerator to account for future algorithmic improvements. To the best of our knowledge, we are the first to present accuracy and energy numbers for a real-time and embedded implementation of a non-convulsive epileptic seizure detection algorithm. On average the algorithm is able to detect 87% of all seizures in the test set, with an average precision of 54%.

Table X provides an overview of recent state-of-the-art works. For a fair comparison we consider programmable ULP processors and systems with support for energy-efficient signal processing in the bio-medical/IoT domain. We compare the systems in terms of system efficiency and energy-efficiency on a common FFT benchmark. In contrast to the works in the Table, the current BrainWave processor is not optimized for duty-cycling. The focus of this work was on energy-efficient 24/7 EEG monitoring, which requires always-on processing.

Having an integrated power delivery solution with multiple voltages for logic, memory, and IO pins comes with a large area penalty and significant voltage-conversion losses. A state-of-the-art work obtains $85\,\%$ system efficiency using on-chip voltage-converters with $36\,\%$ area overhead on a MatMul micro-benchmark [23]. However, for more dynamic real-world use cases, a system efficiency of $80\,\%$ is already challenging, as follows from the table. We measure a state-of-the-art average system efficiency and $90.6\,\%$ at $223\,\mu\mathrm{W}$, while executing the seizure detection application, as depicted in Fig. 8.

In terms energy-efficiency, our system is competitive on a representative 256-point 16-bit fixed-point FFT benchmark, compared to [15], [25], [27]. We report the efficiency in $\mathrm{MOPS/mW}$ (Million Operations Per Second per MilliWatt). At the MEP, the FFT benchmark runs at $26\,\mathrm{MOPS}$ (RISC-like operations) on the Blocks CGRA, which is approximately $25\,\%$ of its peak throughput. The authors of [26] report a higher energy-efficiency, but only report the power consumption while executing an FFT benchmark. As such, the table lists the theoretical peak energy-efficiency while assuming $100\,\%$ DSP utilization. We expect the actual utilization to be somewhere in the range of $25\,\%$–$50\,\%$. [12], [40] presents energy numbers of a real-time convulsive seizure detection algorithm that is mapped onto a 4-core MCU. Overall, we can conclude that the combination of a CGRA with near-threshold computing and voltage-stacking leads to a competitive system for ULP bio-medical signal processing.

VIII. Conclusions

Ambulatory or remote health monitoring of common chronic diseases demands an energy-efficient and programmable processing platform, capable of real-time monitoring in a tiny power budget of $<1\,\mathrm{mW}$. In this work we have provided the first systematic evaluation and optimization of a complex non-convulsive seizure detection algorithm, running in real-time with only $80\,\mathrm{kB}$ of Foundry-provided on-chip memory on the BrainWave processor, while maintaining an *average* $90.6\,\%$ system voltage-conversion efficiency. A total of 9 complex kernels were benchmarked on the CGRA, resulting in an *average* $4.7\times$ speedup and *average* $4.4\times$ energy savings over highly tuned SW-only implementations, demonstrating the Blocks CGRA to be both flexible and energy-efficient. At the system-level a $73.1\,\%$ energy reduction was achieved by utilizing the Blocks CGRA, $15.2\,\%$ by operating in the near-V_{th} region, and up to $19.5\,\%$ by exploiting charge recycling using a 3-level voltage-stacked configuration. This results in a total energy reduction of $5.4\times$ over a highly tuned SW-only implementation. The system presented in this work opens opportunities for future development of battery-operated wearable monitoring systems and algorithms for emerging applications, such as non-convulsive epileptic seizure detection and PD Freezing-of-Gait prediction.

References

[1] M. R. Carneiro, A. T. de Almeida, and M. Tavakoli, "Wearable and comfortable e-textile headband for long-term acquisition of forehead eeg signals," *IEEE Sensors Journal*, vol. 20, no. 24, pp. 15 107–15 116, 2020.

[2] A. Ulate-Campos, F. Coughlin, M. Gainza-Lein, I. S. Fernández, P. Pearl, and T. Loddenkemper, "Automated seizure detection systems and their effectiveness for each type of seizure," *Seizure*, vol. 40, pp. 88–101, 2016.

[3] E. Bruno, P. F. Viana, M. R. Sperling, and M. P. Richardson, "Seizure detection at home: Do devices on the market match the needs of people living with epilepsy and their caregivers?" *Epilepsia*, vol. 61, pp. S11–S24, 2020.

[4] J. Arends, R. D. Thijs, T. Gutter, C. Ungureanu, P. Cluitmans, J. Van Dijk, J. van Andel, F. Tan, A. de Weerd, B. Vledder, W. Hofstra, R. Lazeron, G. van Thiel, K. C. Roes, F. Leijten, and the Dutch Tele-Epilepsy Consortium, "Multimodal nocturnal seizure detection in a residential care setting," *Neurology*, vol. 91, no. 21, pp. e2010–e2019, 2018.

[5] Y. Wang, X. Long, J. P. van Dijk, R. M. Aarts, L. Wang, and J. B. A. M. Arends, "False alarms reduction in non-convulsive status epilepticus detection via continuous EEG analysis," *Physiological Measurement*, vol. 41, no. 5, pp. 1–14, jun 2020.

[6] S. R. Sridhara, M. DiRenzo, S. Lingam, S.-J. Lee, R. Blazquez, J. Maxey, S. Ghanem, Y.-H. Lee, R. Abdallah, P. Singh, and M. Goel, "Microwatt embedded processor platform for medical system-on-chip applications," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 4, pp. 721–730, 2011.

[7] J. Kwong and A. P. Chandrakasan, "An energy-efficient biomedical signal processing platform," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 7, pp. 1742–1753, July 2011.

[8] K. H. Lee and N. Verma, "A low-power processor with configurable embedded machine-learning accelerators for high-order and adaptive analysis of medical-sensor signals," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 7, pp. 1625–1637, 2013.

[9] M. A. Bin Altaf and J. Yoo, "A 1.83 µj/classification, 8-channel, patient-specific epileptic seizure classification soc using a non-linear support vector machine," *IEEE transactions on biomedical circuits and systems*, vol. 10, no. 1, p. 49—60, February 2016.

[10] S.-K. Lin, L.-C. Wang, C.-Y. Lin, H. Chiueh *et al.*, "An ultra-low power smart headband for real-time epileptic seizure detection," *IEEE journal of translational engineering in health and medicine*, vol. 6, pp. 1–10, 2018.

[11] E. De Giovanni, F. Montagna, B. W. Denkinger, S. Machetti, M. Peón-Quirós, S. Benatti, D. Rossi, L. Benini, and D. Atienza, "Modular design and optimization of biomedical applications for ultralow power heterogeneous platforms," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 39, no. 11, pp. 3821–3832, 2020.

[12] S. Benatti, F. Montagna, D. Rossi, and L. Benini, "Scalable eeg seizure detection on an ultra low power multi-core architecture," in *2016 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2016, pp. 86–89.

[13] J. Yoo, L. Yan, D. El-damak, M. Awais, B. Altaf, A. H. Shoeb, and A. P. Chandrakasan, "An 8-Channel Scalable EEG Acquisition SoC With Patient-Specific Seizure Classification and Recording Processor," *Solid-State Circuits, IEEE Journal of*, vol. 48, no. 1, pp. 214–228, 2013.

[14] Y. R. Aldana, B. Hunyadi, E. J. M. Reyes, V. R. Rodríguez, and S. Van Huffel, "Nonconvulsive epileptic seizure detection in scalp eeg using multiway data analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 660–671, 2019.

[15] C. Kim, M. Chung, Y. Cho, M. Konijnenburg, S. Ryu, and J. Kim, "Ulp-srp: Ultra low-power samsung reconfigurable processor for biomedical applications," *ACM Trans. Reconfigurable Technol. Syst.*, vol. 7, no. 3, Sep. 2014.

[16] L. Duch, S. Basu, R. Braojos, G. Ansaloni, L. Pozzi, and D. Atienza, "Heal-wear: An ultra-low power heterogeneous system for bio-signal analysis," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 64, no. 9, pp. 2448–2461, Sep. 2017.

[17] M. Wijtvliet, J. Huisken, L. Waeijen, and H. Corporaal, "Blocks: Redesigning coarse grained reconfigurable architectures for energy efficiency," in *2019 29th International Conference on Field Programmable Logic and Applications (FPL)*, Sep. 2019, pp. 17–23.

[18] Z. Ebrahimi and A. Kumar, "Biocare: An energy-efficient cgra for bio-signal processing at the edge," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2021, pp. 1–5.

[19] B. de Bruin, K. Singh, J. Huisken, and H. Corporaal, "Brainwave: An energy-efficient eeg monitoring system - evaluation and trade-offs," in *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*, ser. ISLPED '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 181–186.

[20] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "Rusboost: A hybrid approach to alleviating class imbalance," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 40, no. 1, pp. 185–197, 2010.

[21] M. Gautschi, P. D. Schiavone, A. Traber, I. Loi, A. Pullini, D. Rossi, E. Flamand, F. K. Gürkaynak, and L. Benini, "Near-threshold risc-v core with dsp extensions for scalable iot endpoint devices," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 10, pp. 2700–2713, 2017.

[22] R. Hameed, W. Qadeer, M. Wachs, O. Azizi, A. Solomatnikov, B. C. Lee, S. Richardson, C. Kozyrakis, and M. Horowitz, "Understanding sources of inefficiency in general-purpose chips," *SIGARCH Comput. Archit. News*, vol. 38, no. 3, p. 37–47, Jun. 2010.

[23] B. Keller, M. Cochet, B. Zimmer, J. Kwak, A. Puggelli, Y. Lee, M. Blagojević, S. Bailey, P.-F. Chiu, P. Dabbelt, C. Schmidt, E. Alon, K. Asanović, and B. Nikolić, "A risc-v processor soc with integrated power management at submicrosecond timescales in 28 nm fd-soi," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 7, pp. 1863–1875, 2017.

[24] K. Singh, B. de Bruin, H. Jiao, J. Huisken, H. Corporaal, and J. P. de Gyvez, "Converter-free power delivery using voltage stacking for near/subthreshold operation," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, pp. 1–13, 2021.

[25] M. Konijnenburg, Y. Cho, M. Ashouei, T. Gemmeke, C. Kim, J. Hulzink, J. Stuyt, M. Jung, J. Huisken, S. Ryu, J. Kim, and H. de Groot, "Reliable and energy-efficient 1mhz 0.4v dynamically reconfigurable soc for exg applications in 40nm lp cmos," in *2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, 2013, pp. 430–431.

[26] Y. Pu, C. Shi, G. Samson, D. Park, K. Easton, R. Beraha, A. Newham, M. Lin, V. Rangan, K. Chatha, D. Butterfield, and R. Attar, "A 9-mm2 ultra-low-power highly integrated 28-nm cmos soc for internet of things," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 3, pp. 936–948, 2018.

[27] J. P. Cerqueira, T. J. Repetti, Y. Pu, S. Priyadarshi, M. A. Kim, and M. Seok, "Catena: A near-threshold, sub-0.4-mw, 16-core programmable spatial array accelerator for the ultralow-power mobile and embedded internet of things," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 8, pp. 2270–2284, 2020.

[28] S. Song, M. Konijnenburg, R. van Wegberg, J. Xu, H. Ha, W. Sijbers, S. Stanzione, D. Biswas, A. Breeschoten, P. Vis, C. van Liempd, C. van Hoof, and N. van Helleputte, "A 769 µw battery-powered single-chip soc with ble for multi-modal vital sign monitoring health patches," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 13, no. 6, pp. 1506–1517, 2019.

[29] J. Xu, B. Büsze, C. Van Hoof, K. A. A. Makinwa, and R. F. Yazicioglu, "A 15-channel digital active electrode system for multi-parameter biopotential measurement," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 9, pp. 2090–2100, 2015.

[30] B. Murmann, "Adc performance survey 1997-2018," http://web.stanford.edu/~murmann/adcsurvey.html, 2018, [Online].

[31] A. Traber and M. Gautschi, "Pulpino: Datasheet," *ETH Zurich, University of Bologna*, 2017.

[32] D. Semiconductor, "Da14580 datasheet, rev. 3.4," https://www.dialog-semiconductor.com/sites/default/files/da14580_fs_3v4.pdf, 2016, [Online].

[33] R. Braojos, D. Atienza, M. M. Sabry Aly, T. F. Wu, H. . P. Wong, S. Mitra, and G. Ansaloni, "Nano-engineered architectures for ultra-low power wireless body sensor nodes," in *2016 International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS)*, 2016, pp. 1–10.

[34] B. Luque, L. Lacasa, F. Ballesteros, and J. Luque, "Horizontal visibility graphs: Exact results for random time series," *Phys. Rev. E*, vol. 80, p. 046103, Oct 2009.

[35] X. Lan, H. Mo, S. Chen, Q. Liu, and Y. Deng, "Fast transformation from time series to visibility graphs," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 25, no. 8, p. 083105, 2015.

[36] G. Manis, M. Aktaruzzaman, and R. Sassi, "Low computational cost for sample entropy," *Entropy*, vol. 20, no. 1, 2018.

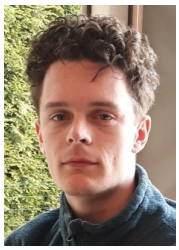
[37] Y.-H. Pan, Y.-H. Wang, S.-F. Liang, and K.-T. Lee, "Fast computation of sample entropy and approximate entropy in biomedicine," *Computer methods and programs in biomedicine*, vol. 104, no. 3, pp. 382–396, 2011.

[38] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American Journal of Physiology-Heart and Circulatory Physiology*, 2000.

[39] A. Podobas, K. Sano, and S. Matsuoka, "A survey on coarse-grained reconfigurable architectures from a performance perspective," *IEEE Access*, vol. 8, pp. 146 719–146 743, 2020.

[40] D. Rossi, A. Pullini, I. Loi, M. Gautschi, F. K. Gürkaynak, A. Bartolini, P. Flatresse, and L. Benini, "A 60 gops/w, -1.8v to 0.9v body bias ulp cluster in 28nm utbb fd-soi technology," *Solid-State Electronics*, vol. 117, pp. 170–184, 2016.

[41] J. Yang and M. Sawan, "From seizure detection to smart and fully embedded seizure prediction engine: A review," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 14, no. 5, pp. 1008–1023, 2020.

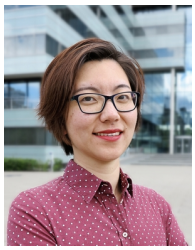


Barry de Bruin received the M.Sc. degree in embedded systems, with specialization in cyber-physical systems (CPS), from the Eindhoven University of Technology, Eindhoven, The Netherlands, in 2017, where he is currently working toward the Ph.D. degree at the Electronic Systems Group, Department of Electrical Engineering.

He continued to pursue his research interests in the areas of energy-efficient biomedical signal processing systems and architectures.

Kamlesh Singh received the master's degree in electronic systems from IIT Bombay, Mumbai, India, in 2015. He is currently working toward the Ph.D. degree at the Electronic Systems Group, Department of Electrical Engineering, Eindhoven University of Technology (TU/e), Eindhoven, The Netherlands.

He worked at Taiwan Semiconductor and Manufacturing Company (TSMC), Hsinchu, Taiwan. His current area of research is ultralow-power and energy-efficient digital integrated circuit design



Ying Wang received her Bachelor degree, Electronics and Information Engineering, at School of Information Engineering, Minzu University of China, and her Master degree, Computational Science in Engineering, at Technische Universitaet Braunschweig, Germany. From 2016 to 2021, she worked at three organizations: Donders Institute of Radboud University, Signal Processing System group of Eindhoven University of Technology, and Kempenhaeghe Research Center for obtaining her PhD degree. Currently, Ying Wang is an assistant professor at Biomedical Signals and Systems group, University of Twente, the Netherlands. Her current research interests include wearable sensing technology, biomedical signal analysis, and patient daily monitoring. In addition, she is an ethical committee member of Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS), University of Twente and a committee member of Dutch Electrical Engineering Platform (EE-NL).

Jos Huisken graduated from the University of Twente, Enschede, The Netherlands.

He joined Philips Research, Eindhoven, The Netherlands, to design their first digital signal processor in CMOS. He has been involved in automated architectural synthesis for digital signal processors and digital audio broadcast ICs in the 1990s. Since then, he has been driving low-power VLSI design from an architectural point of view. After being involved in creating a spin-off company Silicon Hive, Eindhoven, from Philips (now Intel), Eindhoven, on digital signal processing and compilation, he joined Holst Centre, imec, Leuven, Belgium, in 2008, to work on ultralow-power DSP for wireless sensor nodes for body area networks, with a strong focus on low-voltage and low-power circuit design. After joining RWTH Aachen University, Aachen, Germany, in 2013, his research shifted to reliable VLSI design and design for error resilience. In 2017, he joined the Eindhoven University of Technology (TU/e), Eindhoven, to teach VLSI circuit design and carry out research in the field of VLSI design for EEG signal processing, ultrasound, and baseband processing and continue his research in design for error resilience. He is also a researcher in energy-efficient VLSI design.

José Pineda de Gyvez received the Ph.D. degree from the Eindhoven University of Technology, Eindhoven, The Netherlands, in 1991.

He was a Faculty Member with the Department of Electrical Engineering, Texas A&M University, College Station, TX, USA. He is currently a Fellow with NXP Semiconductors, San Jose, CA, USA. He also holds the professorship ”Resilient Nanoelectronics” in the Department of Electrical Engineering, Eindhoven University of Technology. He has authored numerous articles in the fields of testing, nonlinear circuits, and low-power design and coauthored several books. He holds a number of U.S. granted patents.

Dr. Pineda de Gyvez has served as an Associate Editor for several IEEE Transactions. He is continuously involved in technical program committees of scientific symposia.

Henk Corporaal is Professor in Embedded System Architectures at the Eindhoven University of Technology (TU/e) in The Netherlands. He has gained a MSc in Theoretical Physics from the University of Groningen, and a PhD in Electrical Engineering, in the area of Computer Architecture, from Delft University of Technology.

Corporaal has co-authored over 500 journal and conference papers. Furthermore he invented a new class of VLIW architectures, the Transport Triggered Architectures, which is used in several commercial products, and by many research groups.

His research is on low power multi-processor, heterogenous processing architectures, their programmability, and the predictable design of soft- and hard real-time systems. This includes research and design of embedded system architectures, including CGRAs, SIMD, VLIW and GPUs, on accelerators, the exploitation of all kinds of parallelism, fault-tolerance, approximate computing, architectures for machine and deep learning, optimizations and mapping of deep learning networks, and the (semi-)automated mapping of applications to these architectures. For further details see corporaal.org.