Seyed Mohammad Ali Zeinolabedin, Franz Marcus Schüffny, Richard George, Florian Kelber, Heiner Bauer, Stefan Scholze, Stefan Hänzsche, Marco Stolba, Andreas Dixius, Georg Ellguth, Dennis Walter, Sebastian Höppner, Christian Mayr

A 16-Channel Fully Configurable Neural SoC With 1.52 μW/Ch Signal Acquisition, 2.79 μW/Ch Real-Time Spike Classifier, and 1.79 TOPS/W Deep Neural Network Accelerator in 22 nm FDSOI

Erstveröffentlichung in / First published in:

IEEE Transactions on Biomedical Circuits and Systems. 2022, 16 (1), S. S. 94 - 107. IEEExplore. ISSN 1940-9990.

DOI: https://doi.org/10.1109/TBCAS.2022.3142987

Final edited form was published in "IEEE Transactions on Biomedical Circuits and Systems". 16 (1), S. 94-107. ISSN: 1940-9990.
https://doi.org/10.1109/TBCAS.2022.3142987

A 16-Channel Fully Configurable Neural SoC With 1.52 μW/Ch Signal Acquisition, 2.79 μW/Ch Real-Time Spike Classifier, and 1.79 TOPS/W Deep Neural Network Accelerator in 22 nm FDSOI

Seyed Mohammad Ali Zeinolabedin, *Member, IEEE*, Franz Marcus Schüffny, Richard George, Florian Kelber, Heiner Bauer, Stefan Scholze, Stefan Hänzsche, Marco Stolba, Andreas Dixius, Georg Ellguth, Dennis Walter, Sebastian Höppner, and Christian Mayr, *Member, IEEE*

***Abstract*—With the advent of high-density micro-electrodes arrays, developing neural probes satisfying the real-time and stringent power-efficiency requirements becomes more challenging. A smart neural probe is an essential device in future neuroscientific research and medical applications. To realize such devices, we present a 22 nm FDSOI SoC with complex on-chip real-time data processing and training for neural signal analysis. It consists of a digitally-assisted 16-channel analog front-end with 1.52 μW/Ch, dedicated bio-processing accelerators for spike detection and classification with 2.79 μW/Ch, and a 125 MHz RISC-V CPU, utilizing adaptive body biasing at 0.5 V with a supporting 1.79 TOPS/W MAC array. The proposed SoC shows a proof-of-concept of how to realize a high-level integration of various on-chip accelerators to satisfy the neural probe requirements for modern applications.**

***Index Terms*—Biomedical electronics, biomedical signal processing, digital integrated circuits, energy efficiency, neural recording system, implantable devices, accelerator architectures, spike sorting, unsupervised learning.**

I. Introduction

THE research of neuronal microcircuits and the creation of neuroprosthetic devices requires the analysis of neural activities recorded at high spatial and temporal resolution. Large populations of neurons can be recorded extracellularly with multi-channel arrays, however, the transferal of vast amounts of raw data off-chip in such systems prohibits real-time applications, such as Brain-Machine Interfaces (BMIs) and neural prosthetics [1], [2]. Moreover, wireless transmission of the raw data off-chip is limited by bandwidth limitations and the power requirement, and furthermore requires large off-chip memory capacities to save it.

The first step in the analysis of neuronal activity is decoding extracellular potentials to identify spikes from neurons in the vicinity of micro-electrodes. This process is called spike sorting (SS), a classification that requires training on acquired data. During the training phase, an unsupervised/semi-supervised training algorithm is executed to identify distinct spike sources by waveform shape. Training is performed intermittently and outside regular operating mode. Classifications are performed by first detecting the presence of an action potential, and subsequently passing its shape through a clustering algorithm. The classification phase disambiguates spike events from background activity and superimposed multi-unit activity [1]–[3].

Conventional neural recording SoCs include analog front-ends (AFEs) alone, to record and digitize the raw data. However, developing high-density multi-electrode arrays (MEAs) requires complex on-chip data processing to satisfy the real-time and power consumption requirements of the neural probes. Fig. 1 shows that for a 1000-Channel recording system, the power consumption of raw data transmission is about 250 mW. However, it is not supported by implant applications where it should be below 35 mW for reasons of thermal biocompatibility, as seen in the 3D Utah electrode array [4]. Besides, the data rate in such systems is estimated as high as 180 Mbps in the analysis provided. Following these considerations, the design decision to integrate on-chip digital processing is inevitable. On-chip multi-channel spike detection reduces the power and data rate by 66.72% as shown in Fig. 1. However, on-chip spike detection requires transmitting the whole spike, e.g. 64 $\times$ 9bits per spike Whereas further complex on-chip data processing like spike classification reduces the power by 96.63% and reaches the data rate of 1.8 Mbps, as given in Fig. 1. Including more on-chip digital processing causes the AFEs power consumption to become dominant and therefore new circuit-level techniques should be developed to keep the overall system power consumption low.

In an environment where electrode movement and tissue reactions are the norm, as in modern high-density MEAs [5]–[9],

Manuscript received October 18, 2021; revised December 20, 2021; accepted January 9, 2022. Date of publication January 13, 2022; date of current version May 9, 2022. This work was supported in part by the projects GEPARD and LOTUS under Grants 100215497 and 100352812 by the Free State of Saxony and the European Regional Development Fund (ERDF), and in part by the European Union's Horizon 2020 Project "SYNCH" under Grant 824162. This paper was recommended by Associate Editor Chung-Chih Hung. *(Corresponding author: Seyed Mohammad Ali Zeinolabedin.)*

The authors are with the Technische Universität, 01069 Dresden, Germany (e-mail: ali.zeinolabedin@tu-dresden.de; franz_marcus.schueffny@tu-dresden.de; richard_miru.george@tu-dresden.de; florian.kelber@tu-dresden.de; heiner.bauer@tu-dresden.de; stefan.scholze@tu-dresden.de; haenzsch@msx.tu-dresden.de; marco.stolba@tu-dresden.de; andreas.dixius@tu-dresden.de; georg.ellguth@tu-dresden.de; dennis.walter@tu-dresden.de; sebastian.hoeppner@tu-dresden.de; christian.mayr@tu-dresden.de).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TBCAS.2022.3142987.

Digital Object Identifier 10.1109/TBCAS.2022.3142987

Final edited form was published in "IEEE Transactions on Biomedical Circuits and Systems". 16 (1), S. 94-107. ISSN: 1940-9990.
https://doi.org/10.1109/TBCAS.2022.3142987

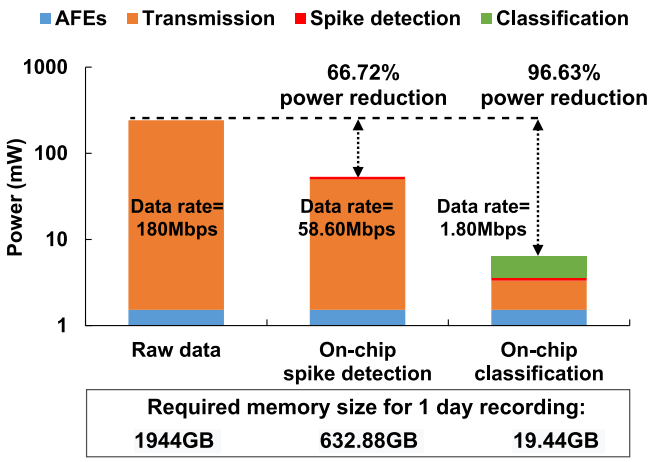


Fig. 1. Analysis of a 1000-Ch neural recording system. (P(AFE) = 1.52 μW/Ch, P(SD) = 0.29 μW/Ch, P(Classification) = 2.5 μW/Ch, maximum firing rate = 100 spikes/sec, transmission energy = 1 nJ/bit, sampling frequency = 20 kHz, and 9-bit data.) .

on-chip programmability is crucial for real-time learning, adapting the spike-sorting classifier, and sustaining high classification accuracy. For these purposes, biosignal processing SoCs require the integration of an ultra-low power CPU, aided by hardware accelerators. However, these new features are absent in state-of-the-art works [5]–[19].

In [10], [11], spike detection (SD) and spike feature extractor (SFE) are implemented for a 64-channel system. In [12], SD and SFE are implemented as a part of classification phase for 128-Channel system. Reference [13] is the first 16-channel SoC performing training and classification on-chip. Customized hardware is designed for running the OSort online training algorithm described in [20] where it is shared between all channels. This design suffers from a large memory size and high power consumption and therefore the number of channels is restricted. The average classification accuracy is about 75%. A 32-Channel chip is reported at [14] which only performs the classification and the averaged classification accuracy is between 70% to 90% and the training is performed offline. In [15], a single-channel spike sorting chip is designed to perform the training and classification. The average classification accuracy is about 84.5% and the power consumption is 148 μW/Ch which is very high for high-density MEAs. A 128-ch chip reported at [16] integrates classification and a modified K-means training. It achieves the average classification accuracy of 74% and the power of 0.175 μW/Ch. A template matching classification method is implemented in [17] to reduce the computational complexity of spike sorting and it achieves 90% accuracy. However, this design requires an extensive offline supervised training process to calculate the templates. A parallel OSort algorithm is implemented in [18] for a single channel and the average accuracy is 87%. It also requires a large memory footprint for multi-channel design, e.g around 8.6 Mb for a 16-channel implementation. An analog neural signal recorder and classifier is implemented in [19] and achieves 93.2% clustering accuracy. Reference [19] uses a fixed and limited number of features that may not be sufficient to cover varieties of the datasets.

None of these designs [10], [11], [13], [15]–[18] includes the AFEs and instead focused on various on-chip data processing/learning hardware realizations. References [13], [15], [16], [18], [19] implement dedicated hardware to perform the training phase in a power efficient mode. Although customized hardware accelerators for training add some levels of adaptivity, they do not provide trained parameters across a variety of conditions which are detrimental considering the changing recording conditions outlined above, in chronic implantation scenarios.

In this paper, we present a fully configurable neural SoC integrating the following components:

- 16-channel analog front end (AFE) containing 16 LNAs time-multiplexed to a chopped VGA and 9 b SAR ADC. Digital assisted filtering makes the AFE more robust with respect to PVT-variation.
- On-chip CPU providing the programmability feature to perform various training/evaluation algorithms in power-efficient modes.
- Dedicated bio-processing accelerators running detection and classification independent of the CPU to achieve ultra-low power performance.
- 16x4 multiply accumulate (MAC) array providing 8-bit unsigned acceleration of time-critical matrix multiplications or 2D (dimensional) convolution.

The proposed system achieves an average classification accuracy of 94.12% and a power consumption of 2.79 μW/Ch for the classification and 1.52 μW/Ch for the AFE. Various architecture- and circuit-level techniques are deployed to achieve the ultra-low power operation.

The remainder of this paper is organized as follows. Section II briefly presents the basics of neural signal analysis taken into account in the design of the SoC. Section III introduces the proposed architecture with its components. Section IV describes the analog front-end and the proposed bio-processing accelerators are explained in Section V. Specifically, the MAC unit is discussed in Section VI. The chip measurement results and comparison with the state-of-the-art designs are presented in Section VII, followed by a conclusion in Section VIII.

II. Considered Fundamentals of Neural Signal Analysis

The proposed smart neural SoC aims at providing a research platform. It's capable of recording and processing electrophysiological data from a variety of different recording locations and electrode types, in an experimental setting. As such, the main motivation behind the use of an onboard processor was to gain the flexibility to accommodate state-of-the-art algorithms for a variety of real-time applications and recording configurations.

In this section, we provide a short overview of the signal domains considered, implementable algorithms, and their use-cases.

- Action Potentials (AP): AP are stereotypical signals in a frequency range of 100 Hz-10 kHz. Unsupervised/semi-supervised clustering algorithms, i.e. spike sorting algorithms, are frequently used for the analysis of such neural signals to identify the source [16], [21]. As such, real-time

Final edited form was published in "IEEE Transactions on Biomedical Circuits and Systems". 16 (1), S. 94-107. ISSN: 1940-9990.
https://doi.org/10.1109/TBCAS.2022.3142987

spike-detection and sorting is a crucial enabler of neurally inspired bio-hybrid systems [22].

- Multi-Unit Activity (MUA): MUA occurs in cases where the APs of several sources overlap. They are informative for less granular measures such as average firing frequency and time-to-first-spike, shown to be sufficient in a variety of applications of BMIs, found e.g. in [23], [24].
- Local Field Potentials (LFP): Local depolarization of the cellular membrane voltage creates superimposed electrical fields that, in sum, are observed as local field potentials within a comparably lower frequency range. Important and informative features can be extracted from LFP as biomarkers for Parkinson's symptoms and a variety of other neuropathologies [25].
- Electroencephalograms (EEG): EEG is recorded on the subjects' scalp and provides a signal that is low in spatial and temporal resolution. Traditionally, EEG power spectra are divided into arbitrary bands to study the physiological signals (e.g. in the observation of sleep patterns and sensory evoked potentials), and to diagnose neuropathologies.

Deep neural networks (DNN) pose an alternative solution to detect biomarkers and possibly perform the matching of stimuli to corresponding evoked spike patterns in situations where supervised learning is applicable. For example, the TrueNorth neuromorphic architecture was successfully used to demonstrate classification accuracy of 76% in a hand squeeze task, using EEG recordings, at a maximum peak power consumption of only 70 mW utilizing convolutional neural networks [26]. To efficiently execute DNN models in low-power hardware, it is recommended to incorporate a dedicated multiply accumulate accelerator to perform computationally expensive operations in a highly parallel and real-time fashion [27], [28]. Such accelerators can be further utilized to effectively implement classical frequency domain transforms such as Short-Time Fourier Transform (STFT), Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT) for an extended signal analysis in the spectral domain to create feature vectors [29].

The proposed system provides a power-efficient platform to analyze those above-mentioned applications in real-time thanks to the high-level system integration of 16-channel AFE, on-chip CPU, power-efficient bio-processing accelerators, and powerful MAC unit.

III. Proposed Architecture

Fig. 2 shows the proposed architecture of the neural SoC comprising a 16-Channel AFEs, an in-order 32-bit RISC-V processor (CPU) with support for IMCX-pulpv2 instructions [30] with a separate 0.50 V power domain. The proposed architecture also contains a bio-processing accelerator consisting of a spike detector (SD), and spike feature extractor/classifier (SFEC) hardware units, all within a separate 0.80 V power domain.

The CPU and the core interfaces of the dual-rail SRAM macros are implemented at an ultra-low voltage of nominally 0.50 V. It is enabled by adaptive body-biasing (ABB), following the ABB-aware implementation methodology from [31]. A forward bias solution is chosen [32] to allow robust performance

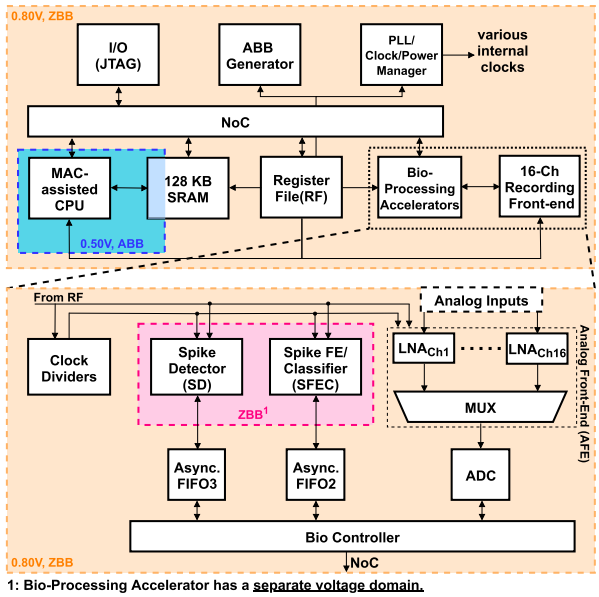


Fig. 2. Block diagram of smart neural SoC including MAC-assisted CPU, 16-Ch AFE, bio-processing accelerator, and ABB unit. All blocks are located in three power domains. CPU and SRAM periphery are in power domain 0.50 V and the SRAM bit cells are in 0.80 V domain. Bio-processing accelerator has a separate 0.80 V domain. (ZBB stands for Zero Body Bias.).

at 0.50 V over the full process, voltage and temperature range. The all-digital-phase-locked-loop (ADPLL) based clock generator from [33] is used for clock generation. Furthermore, the CPU is assisted by a dedicated accelerator specializing in fast and power-efficient execution of matrix multiplication and 2D (dimensional) convolution.

In the proposed architecture, the complex (re)training phase is initially run on the CPU to identify spike sources. CPU is also utilized to calculate the threshold values for SD and to check the quality of cluster centroids over time. Running these intermittent operations requires the flexibility provided by the CPU that is absent in other designs. In contrast, in normal operating mode, i.e. classification phase, SD and SFEC provide real-time operation at frequencies of 400 kHz (SD) and 60 MHz (SFEC) while the CPU is in sleep mode. Therefore the overall system achieves the ultra-low power operation. The mentioned sequence of the operating modes is controlled by the 'Bio Controller'.

The Bio Controller unit (refer to Fig. 2) realizes various operating/testing modes (M1-M7) which are listed below (shown in Fig. 3):

- M1: Recording data via AFEs and performing various training algorithms by means of CPU and optionally the MAC array.
- M2: Utilizing SD individually by reading the input data from a user-specified section of SRAM and store the results back in SRAM. CPU is in sleep mode.
- M3: Utilizing SFEC individually by reading the input data from a user-specified section of SRAM and store the results back in SRAM. CPU is in sleep mode.
- M4: Recording data via AFEs and storing them in a user-specified section of SRAM and transmitting it over GPIO, as well. CPU is in sleep mode.

Final edited form was published in "IEEE Transactions on Biomedical Circuits and Systems". 16 (1), S. 94-107. ISSN: 1940-9990.
https://doi.org/10.1109/TBCAS.2022.3142987

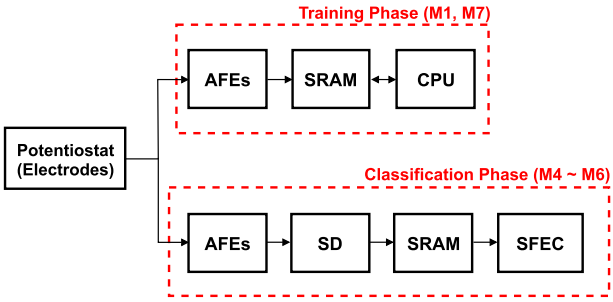


Fig. 3. Active blocks in various operating modes. Bio Controller handles the flow of data between different components. CPU is only active in M1 but not necessarily in M7.

- M5: Recording data via AFEs and performing spike detection by SD in real-time. CPU is in sleep mode.
- M6: Recording data via AFEs and performing classification by SD and SFEC in real-time. CPU is in sleep mode.
- M7: Utilizing the MAC array individually by reading the input data from a user-specific section of SRAM and storing the results back in SRAM. CPU is in sleep mode.

In M1, SRAM is partially organized as two pages whose sizes are user-defined e.g. 32 KB per page. In this mode, Bio Controller stores the recorded data on page 1 and page 2 alternatively so that once page 1 is full, an interrupt request is sent to the CPU to process page 1. Meanwhile, Bio Controller continues storing data on page 2. Because of the slow nature of the neural signal, the CPU finishes the analysis of the page 1 before page 2 gets full. In such way, the same memory region can be periodically utilized to avoid a larger memory footprint. After (re)training phase, the calculated parameters are stored either at register file (RF) or SRAM for later use in M2 to M6.

M2 and M3 opt for serving as stand-alone accelerators to perform spike detection or feature extraction and classification transmitted to the chip via I/O without interfering with the CPU. Testing modes M4-M6 are designed to perform the recording and data processing in real-time and ultra-low power fashion after (re)training is performed in M1.

As shown in Fig. 2, data is transferred between AFEs, SD, and SFEC blocks via asynchronous FIFO by Bio Controller to satisfy the various operating frequency requirement of each block. In M6, Bio Controller concurrently sends the recorded data to SD and also stores them in SRAM for the last 64 data of each channel. The new data automatically overwrites the previous ones so that there is always the latest 64 recorded data, i.e. 64-data-wide window, for any channel available in SRAM. Once a spike is detected from any channel, Bio Controllers receives a notification from SD, and then it retrieves the corresponding 64-data-wide window of that channel together with other related parameters (explained in V) and sends them to SFEC for classification. After classification, the results are sent outside the chip for real-time monitoring and also stored in SRAM. As a result, all the blocks shown in the classification phase of Fig. 3 are involved in the operation in testing mode M6.

I/O block provides a general-purpose port, serial peripheral interface (SPI), and JTAG interface to communicate off-chip. RF enables users to configure various internal blocks via the JTAG interface. A clock divider block generates the different clocks for AFE, SD, and SFEC blocks from the ADPLL.

All main blocks that are destinations or sources for processed data are connected by the network-on-chip (NoC) architecture. A 2D-mesh structure is set up with two routers. Each router supports up to 4 clients. NoC packet data rate is adopted to bandwidth requirements. There is low-speed configuration data that is distributed with 400 MByte/s and high speed neural data can use up to 2.4 GByte/s. This allows low latency communication of all information between all blocks. The NoC architecture is fully digital, scalable and the clock is gated during IDLE times. Packets are routed using a fully connected crossbar approach.

IV. Analog Front-Ends

The front-end signal acquisition chain is shown in Fig. 4. It consists of 16 LNA multiplexed to a single VGA and ADC which are chopped. Its output is then processed digitally by an averager, limiting the bandwidth to 9.8 kHz and a high-pass filter. The LNA has the most power and noise contribution. Other components add up to this given power and should be well below the LNA power. To minimize noise contribution and increase PVT robustness, the digital assisted architecture in Fig. 4 is introduced in [34]. Two LNA versions are implemented on chip, whereas the first one has a gain of 100 and the other one a gain of 10. The variable gain amplifier (VGA) has a gain of either 2.6 or 4.

Small scaled technologies benefit from low area and power for digital circuits but suffer from higher process variation. This is critical for high PVT sensitive components like the sub-Hz filter in the LNAs. Therefore as shown in Fig. 5 a high-ohmic noise-contributing pseudo-resistor is replaced by a switch. If it is activated to ensure DC operating point and ADC range, it creates a step in the signal towards zero, as illustrated in Fig. 4 that is compensated digitally. In case of reset, the last sample is subtracted from the integrator giving a step to compensate for the reset pulse. Since this resetting creates a disturbance, it is executed rarely at points of time with a low signal slope. A high slope increases disturbance since the signal chain is blind during the reset phase. The missing analog high-pass (HP) filtering is implemented in the digital domain with a negative feed-back integrator giving a 0.78 Hz well-defined corner frequency [34].

The second digital assisted noise reduction technique is chopping of VGA and ADC. The multiplexed design switches between the channels. The multiplexer has two polarity options for each channel to compensate for low-frequency noise and offset in the VGA and ADC. This way, it is shifted to 20 kHz and then removed in the digital averaging filter (AVG) [34].

The third power reduction is achieved by multiplexing one ADC and VGA for 16 channels. For small-scale technologies, a faster operating speed can be achieved but leakage is relatively higher than in older technologies. That is why power reduction is achieved by sharing VGA and ADC [34].

A. Noise and Power Reduction of LNA

To design the analog front-end the main concern is the limited power resource. As shown in the noise efficiency figure (NEF) [35] current reduction increases (thermal) noise in the

Final edited form was published in "IEEE Transactions on Biomedical Circuits and Systems". 16 (1), S. 94-107. ISSN: 1940-9990.
https://doi.org/10.1109/TBCAS.2022.3142987

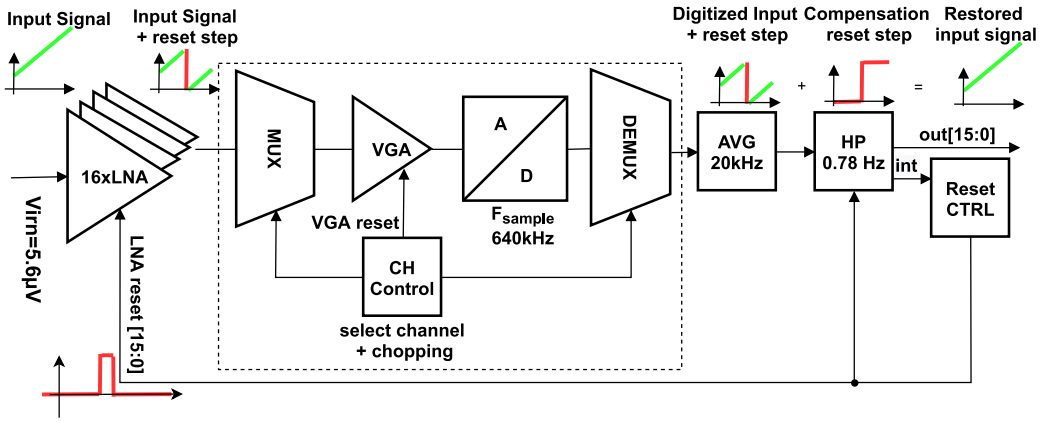


Fig. 4. Overall working principle of full-custom analog part.

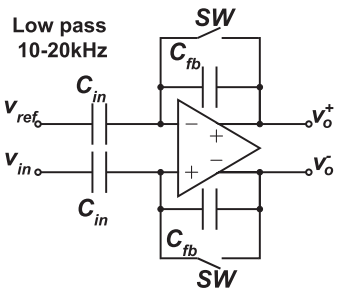


Fig. 5. Detailed LNA schematic. A high-ohmic pseudo resistor is replaced with a switch and its non-linear effect is compensated digitally at HP (refer to Fig. 4).

LNA.

$$NEF = V_{irn}\sqrt{\frac{2 \cdot I}{\pi \cdot V_T \cdot 4kT \cdot BW}} \tag{1}$$

For differential designs NEF usually above 1 is giving a minimum current I for a given total noise V_{irn}, bandwidth BW and temperature T. This results in a given power.

Since the LNA has the most power and noise contribution, its design is described more in detail here. NEF shows the relation of noise, current, temperature, and bandwidth. It is based on the thermal noise of a single transistor. It is minimized by sub-threshold stacked input transistors to get maximum gm for a given current. However, there is a slight technology dependency. In addition to thermal noise giving a constant NEF, there is flicker and igs-shot noise, degrading NEF. Increasing the gate area decreases flicker noise but increases igs-shot noise. Additional noise from the feedback resistor is avoided by replacing the feedback resistor with a reset switch [34]. Limited by the threshold of the transistors, the overall power can be reduced by the lower supply voltage.

B. Power Reduction of VGA, ADC and Digital Filter

The SAR ADC power is limited by its digital-to-analog converter (DAC) and switching power. Because of kTC noise a certain input cap is required, which then results in a required output current of the VGA to charge the ADC during the sampling phase. This is minimized by an AB output stage. The reference current is shared with as many channels as possible, to reduce power per channel. Multiplexing of more channels does not help too much because VGA needs to have more current to charge faster, which eats up power reduction per channel. The noise added by VGA is minor because LNA amplification of 100 reduces noise requirement by a factor of 100. Since it is shared with 16 channels, there is more current available anyway. ADC resolution determines quantization noise. The minimum input range is 2 mV. With an effective number of Bits (ENOB) of 7.44, we get a quantization noise of:

$$\sigma = \frac{\Delta}{\sqrt{12}} = \frac{2\,mV/2^{7.44}}{\sqrt{12}} = 3.3\,\mu V < \sigma_{LNA} \approx 5\,\mu V \tag{2}$$

The digital filter should be as short/small as possible to save power. Measurement showed that first-order high-pass (HP) filter and averager (AVG) is enough [34]. Reset control is based on a differentiation to detect low slopes to allow resetting. Sign check to do resetting towards the middle range of ADC comes cheap in hardware terms. Besides, there is a counter guaranteeing minimum time between resets. Design supply voltage reduction is feasible because the operating frequency is in the sub MHz range.

V. Bio-Processing Accelerators

This section explains Spike Detector (SD) and Spike Feature Extractor/Classifier (SFEC) shown in Fig. 2. They are two CPU-independent accelerators that can provide real-time and power-efficient operations to detect and classify action potential activities in the neural signal.

A. Spike Detector (SD)

As explained in Section III, Bio Controller sends the recorded data to SD in Modes M2, M5, and M6 via asynchronous FIFO. Because a spike features sudden changes in the waveform, a nonlinear energy operator (NEO) defined at (3) exploits it to intensify the spike activity from the background noise. Besides, NEO improves the (signal-to-noise ratio) SNR of the signal resulting in being less sensitive to a threshold value compared

Final edited form was published in "IEEE Transactions on Biomedical Circuits and Systems". 16 (1), S. 94-107. ISSN: 1940-9990.
https://doi.org/10.1109/TBCAS.2022.3142987

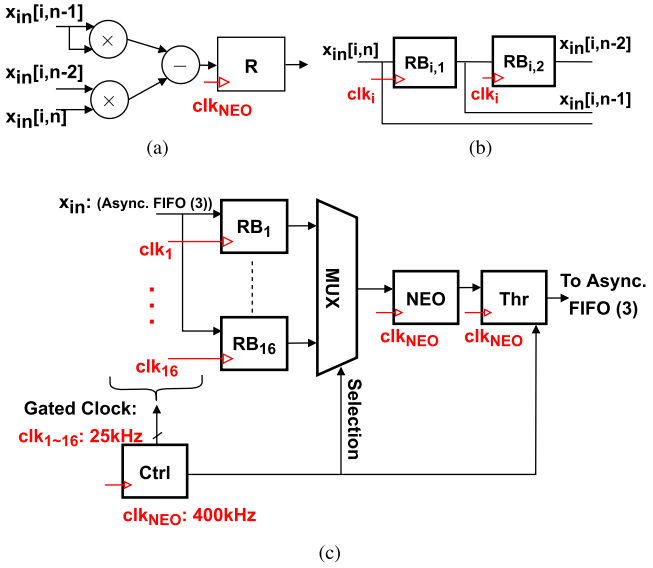


Fig. 6. (a) NEO engine. (b) Register Bank for a sample channel (RB_i). (c) Block diagram of SD.

to a case without NEO [36]–[38].

$$\Psi(x(n)) = (x(n))^2 + x(n-1)x(n+1) \tag{3}$$

where the $\Psi(x(n))$ is the NEO output of neural signal $x(n)$ at time n.

Fig. 6 shows the block diagram of SD. It proposes a time-multiplexing architecture sharing a single NEO (Fig. 6(a)) engine among 16 channels. The register bank (RB_i) in Fig. 6(b) stores the last three samples of each channel required to do the calculation. Bio Controller transfers the 16 channels' data at the frequency of 400 kHz (i.e. $=16 \times 25$ kHz) to SD. However, each channel's register bank receives the proper data every 16 clock cycles, i.e. each RB_i should clock at 25 kHz. Therefore, an extensive clock gating technique is applied to RB_i to further reduce the dynamic power. The Thr block in Fig. 6(c) compares the output of the NEO point-by-point to a fixed threshold value and generates a high signal if it is larger than the threshold value. Although NEO is much less sensitive to a threshold value, it is required to estimate the proper threshold value with the respect to recorded data. To satisfy this condition, the threshold value is calculated during the training and the result is stored at RF to be utilized by SD during the operation.

There are most likely cases where the Thr block output is activated multiple times for a given spike. To avoid that, the SD Ctrl block only allows a single activation within the last n read data, where n can be also set by the user. However, it is in general set to the absolute refractory period where no spike activities can occur [36].

B. Spike Feature Extractor/Classifier (SFEC)

SFEC can be utilized in modes M3 and M6. In M3, it performs the feature extraction and classification over the data stored in SRAM to characterize and verify the functionality of the

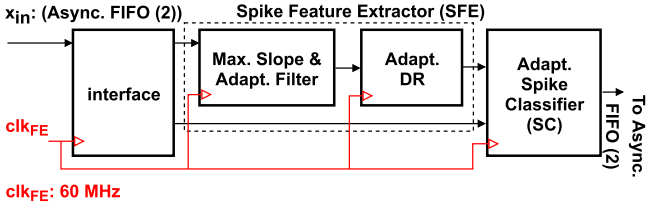


Fig. 7. Block diagram of SFEC.

SFEC. In M6, it acts as a part of the real-time recording and classification of the neural signal.

In M6, there is always the latest 64-data-wide window of any channel stored in SRAM and once SD detects a spike for any channel it notifies the Bio Controller. The Bio Controller then fetches both the 64-data-wide spike window centered around the detection point, and trained parameters of that channel and feeds them into the SFEC. Fig. 7 shows the block diagram of the SFEC. It includes Spike feature Extractor (SFE) and adaptive Spike Classifier (SC). For each channel, trained parameters are 1) filter's coefficients (maximum 8^{th}-order filter), 2) dimensionality reduction (DR) scheme (2 to 7 features) [36], 3) the number of clusters, and 4) corresponding cluster means, which are all calculated during mode M1. The number of clusters is set between three to eight because [39] shows that there are most likely up to eight single-unit activities observable by each micro-electrode. The first trained parameters are used by SFE to perform the filtering and calculate features and the last three ones are used by SC to assign the detected spike to the nearest cluster.

SFE calculates the index of the maximum slope (4) of the detected spike and performs the filtering at the same time.

$$idx = \arg\max_n(x(n) - x(n-1)) \tag{4}$$

where idx is the maximum slope index.

SFE requires 69 clock cycles to finish these operations. It can be at most 8^{th}-order FIR filter and its coefficients can be specific for each channel. Before the spike data is given to SFE, the filter coefficients, DR scheme (feature_flag: a 7-bit vector indicating which feature should be included), and the number of clusters are transmitted sequentially. The Adapt. DR block is then selecting up to 7 features. This procedure is applied to every detected spike and the features are selected in alignment to the maximum slope index because the maximum slope of the spike has biological significance and results in superior clustering accuracy [36].

Once all features are ready, the cluster means are sequentially transmitted to SC as described in (5).

$$distance_j = \sum_{i=0}^{Num\_feature} |f_i - \mu_{ji}|$$

$$label = \arg\min(distance_j) \quad \forall j \in (3, \ldots, Num\_Cl) \tag{5}$$

where the `Num_feature` and `Num_Cl` are the number of features and number of clusters, correspondingly.

Each cluster mean can be up to a 7-dimensional vector. So to optimize the SC architecture, a serial architecture is proposed

Final edited form was published in "IEEE Transactions on Biomedical Circuits and Systems". 16 (1), S. 94-107. ISSN: 1940-9990.
https://doi.org/10.1109/TBCAS.2022.3142987

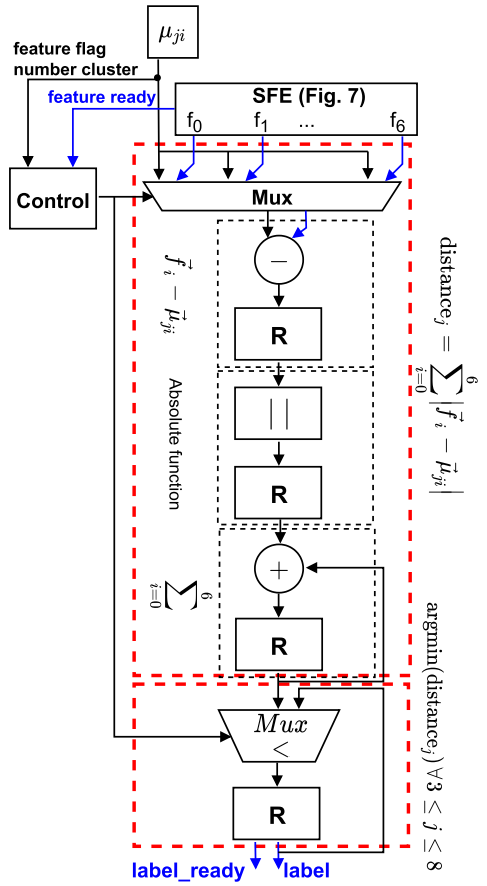


Fig. 8. Proposed architecture of adaptive spike classifier (SC).

as shown in Fig. 8. Every time one dimension, μ_i, of the current cluster arrives, the distance to its corresponding feature is calculated. This procedure is iteratively done for all dimensions of that cluster and all subsequent clusters in a sequential fashion to calculate the distance of the detected spike to all cluster means. Concurrently, whenever a distance is calculated it is compared to the previous distance value and if it is smaller, the spike label is updated. In the end, the *label* contains the nearest cluster index to the detected spike. Depending on the number of features and number of clusters, SC requires a different number of cycles to finish the task as given in (6).

$$T_{SC} = 9 + Num\_Cl \times (2.5 \times Num\_feature + 4) \quad Cycles \tag{6}$$

Fig. 9 indicates that SFE needs 69 clock cycles to calculate the features and SC requires 181 clock cycles in the worst-case scenario to perform the classification, resulting in 250 clock cycles to complete one SFEC operation as given in (7).

$$T_{SFEC} = T_{SC} + 69 \quad Cycles \tag{7}$$

Realizing the high-density neural probes requires the real-time and power/area-efficient design. To optimize the area and energy efficiency, a single SFEC is shared among all channels and it runs at 60 MHz to achieve the real-time operation.

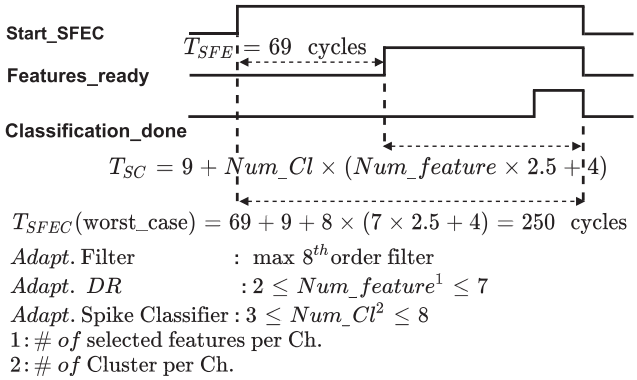


Fig. 9. SFEC running time.

Equation (8) shows that for a typical case, SFEC can process a detected spike in less than 2.5 μsec, that is, a time resolution of 16-channel (1/400 kHz). In other words, the process of the currently detected spike is finished before a new sample of the next channel is recorded. In a worst-case scenario of having a spike on all channels at the same time and assuming `Num_Cl=8`, SFEC takes 66.7 μsec to process all channels as given in (9), which is equivalent to recording 27 samples, that is, 1.7 samples per channel. During the SFEC operation, Bio Controller continues recording the data and therefore no data is lost. In the worst-case scenario, SFEC reduces the data rate to 0.19 Mbit/s for the maximum firing rate of 100 spikes per second. Therefore, single SFEC reduces the data rate significantly by 99% resulting in saving transmission power component and area.

$$\begin{aligned} &if \quad Num\_Cl = 3, Num\_feature = 7 \Rightarrow \\ &T_{SFEC}(typ) = 142.5\, Cycles \\ &operating\ time = 2.38\,\mu\text{sec} < 2.5\,\mu\text{sec} \end{aligned} \tag{8}$$

$$\begin{aligned} &if \quad Num\_Cl = 8, Num\_feature = 7 \Rightarrow \\ &T_{SFEC}(worst\_case) = 250\, Cycles \\ &operating\ time = 250 \times 16 \times \frac{1}{60\,MHz} \Rightarrow \\ &= 66.7\,\mu\text{sec} \end{aligned} \tag{9}$$

VI. MAC Unit

To enable real-time execution of time-critical signal processing, a MAC array was included to speed up algorithms relying on matrix multiplication (MM) or 2D convolution (CONV2D) e.g. classification of causes of specific spike patterns. Though the module can work independently, it functions as a support module for the RISCV CPU and can be controlled over Advanced High-performance Bus (AHB) or through specific NoC control packets. It is therefore possible to disable the CPU during any execution of the accelerator to save energy. Fig. 10 provides an overview of the structure. The module consists of 16x4 multiply-accumulate cells processing 64x2 unsigned 8-bit operations per clock cycle. For that, the accelerator fetches at maximum 2x128 bits per clock cycle over a direct connection to the SRAM and the NoC simultaneously. If the source has

Final edited form was published in "IEEE Transactions on Biomedical Circuits and Systems". 16 (1), S. 94-107. ISSN: 1940-9990.
https://doi.org/10.1109/TBCAS.2022.3142987

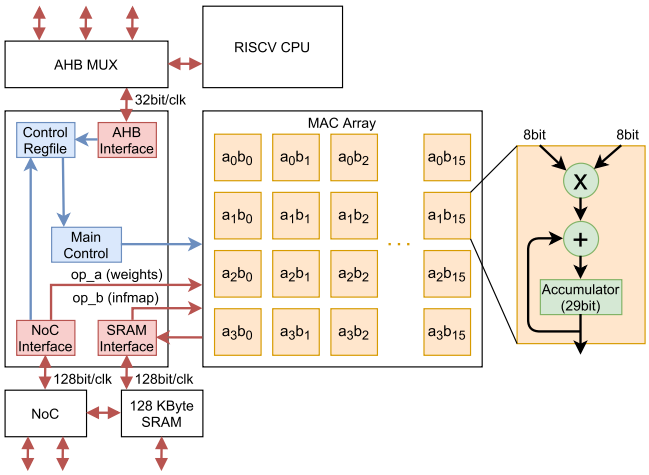


Fig. 10. Structure of the MAC array.

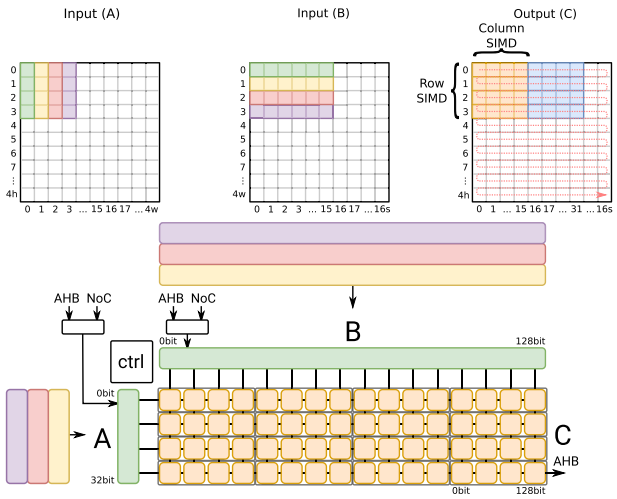


Fig. 11. Execution of matrix multiplication. The accelerator requests memory aligned and zero padded data representing the matrices A and B. Per execution step, 128-bit data arrives for matrix B (green) and 128 bits per 4 execution steps for matrix A (green to purple). Each 8-bit value from incoming data is distributed per row and column and processed by one MAC cell (green to purple matching per execution step) in a SIMD (Single Instruction Multiple Data) parallelization. After iterating through the shared matrix dimension, the content of the accumulators is written out to the nearby PE SRAM (orange). The matrix multiplication is done after going through the column dimension of matrix B and row dimension of matrix A in this order.

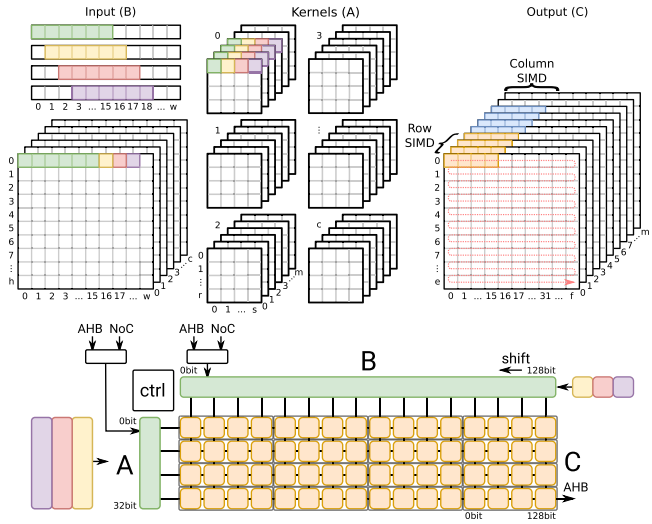


Fig. 12. Execution of 2D cross correlation. At the start, the input feature map and kernel is fetched 128 bits each (green) and distributed to the MAC array by row (kernel) and column (feature map). Per row another output channel is allocated. Each execution step shifts another 8-bit value through register B while register A iterates through kernel column and row in that order (yellow to purple). During that process, the accelerator fetches 32 bits and 128 bits per 4 execution steps and retains and reuses the previously fetched data as much as possible. After the output values are calculated, they get written back to the SRAM (orange) while the NoC interface prefetches for the next output batch (blue) and stores the values into the NoC FIFOs. The accelerator iterates over output channel (batches of 4 due to array rows), output columns (batches of 16 due to array columns) and output rows until the cross correlation is done.

increased clock delay, the NoC access can prefetch the data before it is needed. An output stationary [40] dataflow was chosen to mitigate partial-sum memory fetches for area/power efficiency. Weights are reused with a horizontal broadcast while input feature maps are distributed vertically through the array. Figs. 11 and 12 are presenting the matrix multiplication dataflow and cross-correlation dataflow respectively. To further scale down memory requests, the CONV2D controller reuses incoming rows of input feature maps by shifting them per clock cycle and requesting only the necessary data. Furthermore, instead of parallelizing the output feature map row dimension we chose the output channel dimension to fully utilize the array for scenarios with 1x1 kernel layers. For the accumulator a bit size of 29 bits was chosen to ensure high accuracy for state of the art worst-case sized models. The results are written out into the SRAM. The Main Control block tracks the state of the accelerator and synchronizes the memory fetches for the correct execution while the dimensions and features can be adapted by accessing the Control Regfile block. The module consumes 0.032mm^2 in area.

For the general use-case, the array has NoC access and therefore can either be supplied by SRAM memory or in an online fashion directly from other sources like the ADCs or Bio-Accelerators. The general dataflow for matrix multiplication and 2D convolution inside the array is illustrated in Figs. 11 and 12 respectively.

VII. Measurement and Comparison

This section provides the power and area measurements and the comparison results for the proposed system. The chip is implemented in a 22 nm FDSOI technology and its microphotograph is depicted in Fig. 13.

A. Testing Setup

The various operating modes and the components were tested by applying real datasets [41]. Fig. 14 shows the laboratory testing setup. To allow an evaluation that is both reproducible and yet approximates realistic operating conditions, the IC was connected to platinum recording electrodes inserted into Ringer Solution (Na+ 147 mmol/l; K+ 4,0 mmol/l; Ca2+ 2,3 mmol/l; Cl- 156 mmol/l). In this electrolyte similar to the extracellular medium, an additional set of electrodes was used to provide a

Final edited form was published in "IEEE Transactions on Biomedical Circuits and Systems". 16 (1), S. 94-107. ISSN: 1940-9990.
https://doi.org/10.1109/TBCAS.2022.3142987

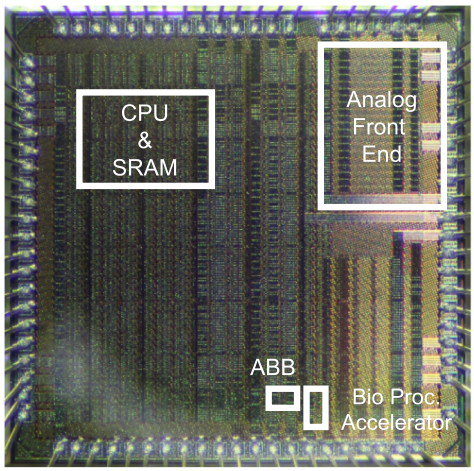


Fig. 13. 22 nm implemented smart neural chip microphotograph.

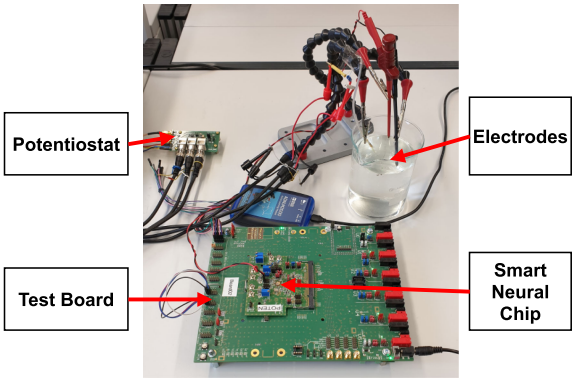


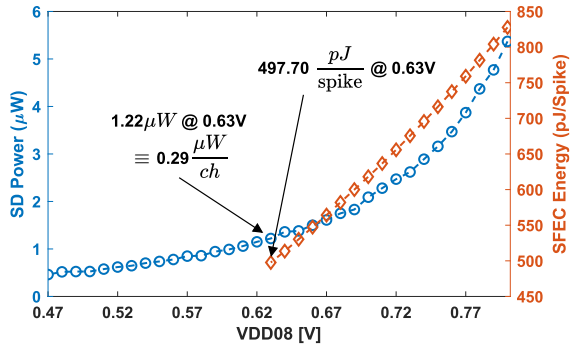
Fig. 14. Testing setup for laboratory verification of the developed front-end. Measurement electrodes are directly connected to the smart neural chip, and in contact with the Ringer solution. The Potentiostat provides a test-signal in a reproducible manner using real datasets [41].

test signal, consisting of an electrophysiological recording, with white noise added in the spectrum up to the frontends sample rate, to compensate for the filtering performed in the original acquisition process. In the application of the signal onto the recording electrodes, a custom-built potentiostat was used, to compensate for the effect of electrode/electrolyte interactions on the signal-providing electrodes.

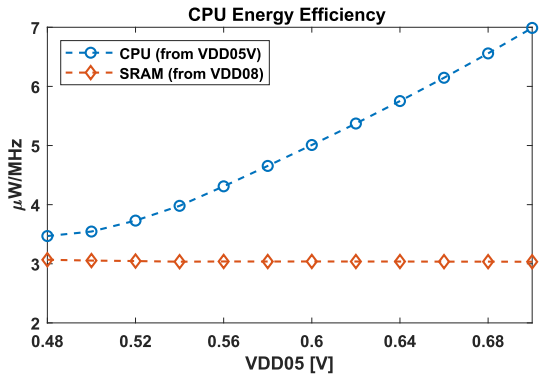
B. Power and Area

SD and SFEC are implemented using high-Vt devices that result in optimum leakage reduction at their low target frequencies of 400 kHz (SD) and 60 MHz (SFEC). In total, 0.94 μW leakage and 1.22 μW (SD) active power and 497.40 pJ/spike (SFEC) are measured at 0.63 V (Fig. 15(a)).

The CPU is implemented with a target frequency of 125 MHz at 0.50 V using low-Vt and super-low-Vt cells. As explained, the CPU is only active during the (re)training phase and otherwise remains in sleep mode. A modified training algorithm is developed based on [16] to calculate the required parameters. It takes on average 220 cycles to process an input sample. Thus, the CPU frequency must be more than 110 MHz (Fig. 16) to (re)train 16

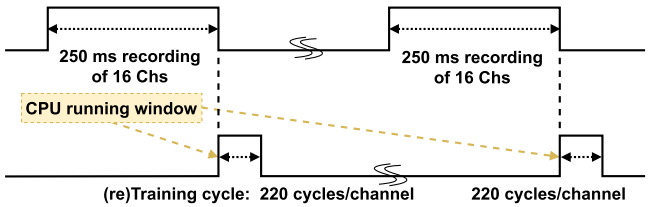


(a)



(b)

Fig. 15. (a) Measurement result for SD and SFEC. (b) Measurement result for CPU and SRAM.



Average sample processing cycle[1] : 220 (cycles)

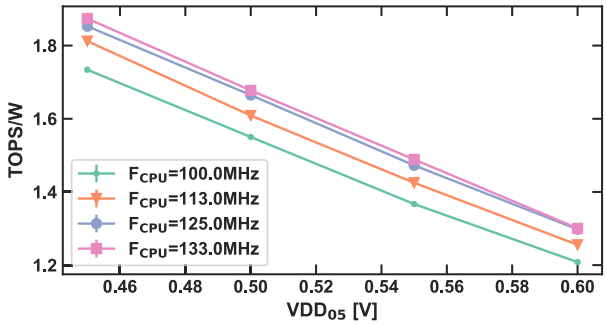
- For 16 channels: $\left(220 \times N \times F_{CPU}^{-1} \times 16\right) < 200\ (ms)$
- For at least $N = 6250$ (samples/Ch) $\Rightarrow F_{CPU} > 110\ MHz$

1: for a given training algorithm

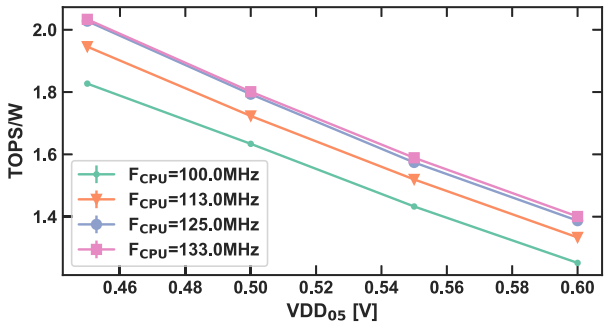
Fig. 16. Training time analysis and CPU frequency calculation.

channels in less than 200 ms for a 250 ms recording window. If a longer training period is required for any separate channel, recording is only devoted to that channel to accumulate more samples for training. With the on-chip programmability of the CPU, transmitting a huge amount of multi-channel data off-chip for training/evaluation (TE) can be avoided. Because TE is performed infrequently, the CPU can be put into sleep mode most of the time. The CPU operates from 0.50 V with 21 μW leakage and 3.5 μW/MHz dynamic power. It uses dual-rail SRAM with 0.50 V logic and 0.80 V bit cell supply. The SRAM macros used for CPU operation consume 3.0 μW/MHz. The complete on-chip SRAM (16 $\times$ 8 Kbyte) consumes 16 μW leakage at room temperature (Fig. 15(b)).

Final edited form was published in "IEEE Transactions on Biomedical Circuits and Systems". 16 (1), S. 94-107. ISSN: 1940-9990.
https://doi.org/10.1109/TBCAS.2022.3142987



(a)



(b)

Fig. 17. Energy Efficiency for (a) a matrix multiplication (b) a 2D convolution execution.

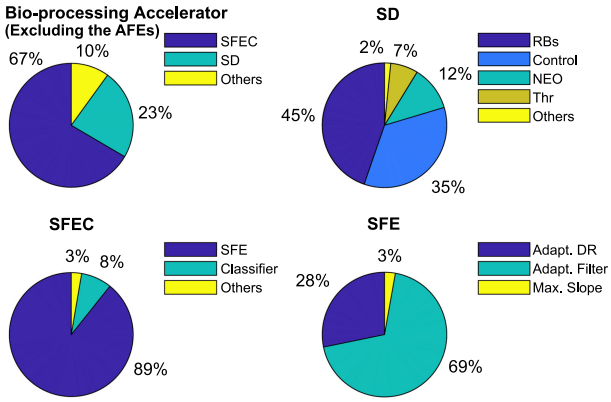


Fig. 18. Area break-down of bio-processing accelerator excluding the AFEs.

The MAC accelerator designed for dense DNN classification achieves 1.66 TOPS/W for matrix multiplication and 1.79 TOPS/W for 2D convolution with 125 MHz CPU frequency and 0.50 V supply voltage as can be seen in Fig. 17. Furthermore, it can reach up to 1.85 TOPS/W and 2.03 TOPS/W respectively if the voltage is further decreased to 0.45 V. For each execution, one example layer has been selected out of common classification models and divided to fit into 128KByte SRAM.

CPU and SRAM consume 0.22 mm^2 and the bio-processing accelerators consume 0.007 mm^2 out of which 23% is occupied by SD and 67% by SFEC. The area breakdown of the bio-processing accelerators is fully depicted in Fig. 18.

Fig. 19 shows a sample recorded and processed data by smart neural chip tested in-vitro. The neural signal is first measured by

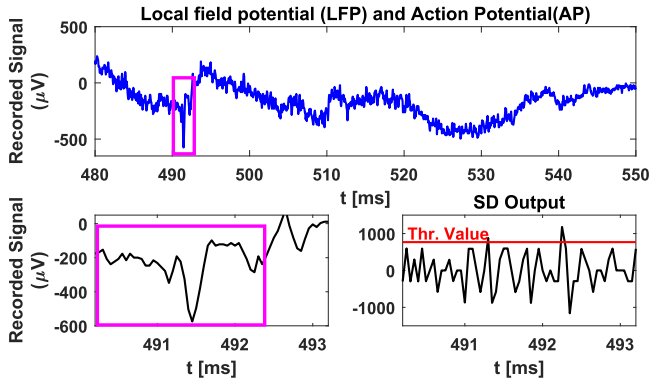


Fig. 19. A sample recorded and processed neural signal provided by the in-vitro testing setup. An exemplary spike is highlighted to be detected using the SD.

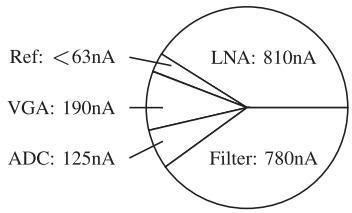


Fig. 20. AFE overall power per channel [34].

TABLE I
AFE COMPARISON WITH STATE-OF-THE-ART CHIP

	This	[42]	[43]	[44]	[45]
Process (nm)	22	65	55	65	180
Virn (μV)	5.6	3.1	6.0	8.98	3.93
Power (μW/Ch)	1.52	0.65	61.2	2.72	13.94
BW (kHz)	9.8	10	10	9.7	9.7
NEF_{sys}	3.0	0.97	17.1	6.8	2.94
Area (mm^2/Ch)	0.018	0.00656	0.0077	0.062	0.0378

the analog front-end and SD simultaneously processes the signal to detect the potential spikes. The digital-assisted AFE consumes 1.52 μW/Ch and the SD and SFEC consume 2.79 μW/Ch. The chip can run arbitrary (re)training algorithms compared to fixed algorithms at [13]–[19]. The selected algorithm consumes 28.46 μJ/Ch for 250 ms training window and training is typically run no more than once per hour.

Power details of the AFE are shown in Fig. 20. LNAs power dominates, but the digital filter is roughly the same. A design with a lower supply voltage than 0.80 V would have been beneficial and realistic due to the low speed of below 1 MHz. The ADC itself has a DNL between 0.25 and -0.4.

C. Comparison

A comparison to other state-of-the-art AFEs is presented in Table I. The proposed design has a low footprint but higher input referred noise at the same power compared to other systems.

From an area perspective, the LNAs are dominant as well. It can be seen that our design has a higher NEF than [6], but the low FOM is gained by large caps and input transistors resulting in a roughly 10x larger area per channel.

Final edited form was published in "IEEE Transactions on Biomedical Circuits and Systems". 16 (1), S. 94-107. ISSN: 1940-9990.
https://doi.org/10.1109/TBCAS.2022.3142987

TABLE II
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART CHIP

Reference	[13] JSSC'13	[14] TBCAS'17	[15] TBCAS'18	[16] TVLSI'19	[17] TBCAS'19	[18] TBCAS'19	[19] TBCAS'21	**This Work**
No. of Chs	16	32	1	128	1	1	1	**16**
AFE	N	N	N	N	N	N	Y	**Y**
Detection	Y (Absolute)	Y (NEO)	Y (NEO)	Y (ICD[1])	Y (NEO)	Y (NEO)	Y (Absolute)	**Y (NEO)**
FE	N	Y (Max-Min[2])	Y (ADDs[3])	Y (ICFE[4])	N	N	Y (FSDE Max-Min[5])	**Y (AF[6])**
Accuracy Mean	75%[7]	60-80%[7]	84.5%	74%	90%[8]	87%[8]	93.2%[9]	**94.12%**[10]
On-chip Training	Y (O-Sort)	N	Y (C-Sort)	Y (Mod. K-means)	N	Y (C-Sort)	Y (Perturbed K-means)	**Y (configurable)**
Core Voltage (V)	0.27	1.2	1.8	0.54	0.25	1.16	1.5	**CPU:0.5, Mem: 0.5, 0.8 SD, SFEC: 0.63 AFE: 0.8**
Classification Power(μW/ch)	4.68	0.75	148	0.175	0.064[11]	2.78[11]	4.35	**2.79**[12]
Area (mm^2/Ch)	0.07	0.023	2.7	0.003	0.3	2.57	1.023	**0.014**[13]
Process (nm)	65	130	180	65	45	32	180	**22 FDSOI**

[1] Integer coefficient detector (ICD). [2] Max-Min peaks of Haar DWT. [3] Adaptive discrete derivatives (ADDs).
[4] Integer coefficient feature extractor (ICFE). [5] 1st and 2nd derivative (FSDE). [6] Adaptive filter (AF).
[7] With different datasets that are not available online.
[8] It seems that three are only limited datasets from [46] utilized to calculate the accuracy.
[9] The number of cluster is set to 4 by the user resulting in high clustering accuracy.
[10] With all datasets in [46].
[11] Only post-layout simulation.
[12] Excluding AFE. P(AFE) = 1.52 μW/Ch, E(training) = 28.46 μJ/Ch running for 250 ms.
[13] Including the bio-processing accelerators, used SRAM blocks and bio controller but excluding AFE. Including AFE area is = 0.038 mm^2/Ch.

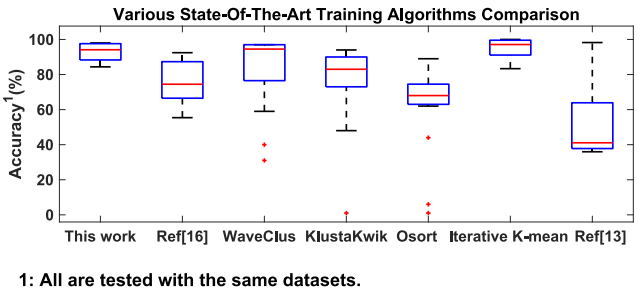


Fig. 21. Classification accuracy of smart neural chip compared to state-of-the-art spike sorting algorithm either implemented offline or on-chip.

Fig. 21 shows that the proposed design outperforms state-of-the-art online spike sorting algorithms which are publicly available using the same datasets [46] in terms of the classification accuracy. The mean classification accuracy is 94.12%. WaveClus [46] algorithm and iterative K-mean with 100-iteration still gives the better results because they are complex offline spike sorter. However, this work achieves better accuracy than KlustaKwik (mean classification accuracy: 83%) [47] and [13]–[19].

Table II compares the proposed neural chip with the latest implemented designs. This work consists of 16 analog front-ends, configurable ultra-low power bio-processing accelerators and ultra-low power CPU with dedicated DNN accelerator. The proposed SoC provides a platform to perform various (re)training methods in a real-time and power-efficient operation which is required for neural implants. This work achieves the highest level of integration for a neural implant SoC compared to recently published ones [10], [11], [13], [15]–[19]. In this work, classification accuracy is calculated with all widely-used datasets introduced in [46] and also compared to well-known off-line spike sorting algorithms [48] like WaveClus [46], KlustaKwik [49] and OSort [20].

The classification power per channel in this work is one of the lowest reported ones, considering the configurability of the design to perform various training algorithms and thanks to dedicated configurable bio-processing accelerators for the classification. References [16], [17] report 0.175 μW/Ch and 0.064 μW/Ch, respectively. However, in [16], the training algorithm is fixed and the corresponding SFEC is not configurable. Reference [17] needs also off-chip supervised training which makes it infeasible for neural implant applications.

The proposed SoC has been achieved the competitive power, area, and accuracy performance because:

- It integrates ultra-low power CPU equipped with MAC unit, 16-Ch AFEs, and dedicated power-efficient bio-processing accelerators.
- programmability feature of CPU allows performing various complex algorithms for (re)training that results in a

Final edited form was published in "IEEE Transactions on Biomedical Circuits and Systems". 16 (1), S. 94-107. ISSN: 1940-9990.
https://doi.org/10.1109/TBCAS.2022.3142987

very high accuracy. Whereas, classification is performed by the bio-processing accelerators operating in the ultra-low power mode. Similarly, MAC unit realizes the intensive arithmetic operations in an efficient way.

- Various architectural techniques are applied in the bio-processing accelerators and MAC unit as explained in Sections V and VI to either reduce the power or to meet the timing.
- In AFE, avoiding a feedback resistor and replacing it by a switch, reduces noise and thus allows to reduce power for constant noise. Besides, digital filtering instead of analog filtering improves the frequency accuracy.
- Advanced 22 nm Technology with the ABB technique facilitates achieving the minimum power operation of CPU at all different operating conditions at the given frequency. It also ensures that CPU can reliably operate at the ultra-low voltage of 0.50 V.

VIII. Conclusion

In this paper, we have outlined a complete neural processing SoC with 16-Channel AFE, a low-power MAC-assisted CPU, and SD/SFEC for analysis of a wide range of neural signals in real-time. The on-chip ultra-low power CPU with its powerful MAC unit is a key element to satisfy on-chip programmability and flexibility. On-chip bio-processing accelerator facilitates the application of various types of neural signal analysis algorithms such as spike detection, feature extraction, and spike sorting independent of the CPU to achieve in ultra-low power operation. This work achieves the best classification accuracy (mean classification accuracy of 94.12%) compared to state-of-the-art online spike sorting algorithms. Moreover, it provides one of the highest levels of SoC integration for ultra-low power neuronal recording applications. The digital-assisted AFE, consuming 1.52 μW/Ch, and the classification, consuming 2.79 μW/Ch, were verified with synthetic and real datasets. The CPU operates from 0.50 V with 21 μW leakage and 3.5 μW/MHz dynamic power. For further processing, the MAC array achieves 1.66 TOPS/W for matrix multiplication and 1.79 TOPS/W for 2D convolution with 0.50 V and 125 MHz clock. Using this design, both data rate and the transmission power can be reduced by around 99% for 16-channel AFE, effectively laying the groundwork for a new class of cortical active and intelligent implants.

References

[1] R. Vetter, J. Williams, J. Hetke, E. Nunamaker, and D. Kipke, "Chronic neural recording using silicon-substrate microelectrode arrays implanted in cerebral cortex," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 896–904, Jun. 2004.

[2] J. Jun *et al.*, "Fully integrated silicon probes for high-density recording of neural activity," *Nature*, vol. 551, no. 7679, pp. 232–236, Nov. 2017.

[3] C. Rossant *et al.*, "Spike sorting for large, dense electrode arrays," *Nature Neurosci.*, vol. 19, no. 4, pp. 634–641, Apr. 2016.

[4] S. Kim, P. Tathireddy, R. A. Normann, and F. Solzbacher, "Thermal impact of an active 3-d microelectrode array implanted in the brain," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 15, no. 4, pp. 493–501, Dec. 2007.

[5] M. Ballini *et al.*, "A 1024-Channel CMOS microelectrode array with 26,400 electrodes for recording and stimulation of electrogenic cells in vitro," *IEEE J. Solid-State Circuits*, vol. 49, no. 11, pp. 2705–2719, Nov. 2014.

[6] D. Han, Y. Zheng, R. Rajkumar, G. S. Dawe, and M. Je, "A 0.45 v 100-Channel neural-recording IC with sub-uw/channel consumption in 0.18 um CMOS," *IEEE Trans. Biomed. Circuits Syst.*, vol. 7, no. 6, pp. 735–746, Dec. 2013.

[7] C. M. Lopez *et al.*, "A neural probe with up to 966 electrodes and up to 384 configurable channels in 0.13 um SOI CMOS," *IEEE Trans. Biomed. Circuits Syst.*, vol. 11, no. 3, pp. 510–522, Jun. 2017.

[8] C. M. Lopez *et al.*, "An implantable 455-Active-Electrode 52-Channel CMOS neural probe," *IEEE J. Solid-State Circuits*, vol. 49, no. 1, pp. 248–261, Jan. 2014.

[9] R. Shulyzki *et al.*, "320-Channel active probe for high-resolution neuromonitoring and responsive neurostimulation," *IEEE Trans. Biomed. Circuits Syst.*, vol. 9, no. 1, pp. 34–49, Feb. 2015.

[10] T.-T. Liu and J. M. Rabaey, "A 0.25 v 460 nW asynchronous neural signal processor with inherent leakage suppression," *IEEE J. Solid-State Circuits*, vol. 48, no. 4, pp. 897–906, 2013.

[11] V. Karkare, S. Gibson, and D. Marković, "A 130-μw, 64-Channel neural spike-sorting DSP chip," *IEEE J. Solid-State Circuits*, vol. 46, no. 5, pp. 1214–1222, May 2011.

[12] M. S. Chae, Z. Yang, M. R. Yuce, L. Hoang, and W. Liu, "A 128-Channel 6 mW wireless neural recording IC with spike feature extraction and UWB transmitter," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 17, no. 4, pp. 312–321, Aug. 2009.

[13] V. Karkare, S. Gibson, and D. Marković, "A 75-μw, 16-channel neural spike-sorting processor with unsupervised clustering," *IEEE J. Solid-State Circuits*, vol. 48, no. 9, pp. 2230–2238, Sep. 2013.

[14] Y. Yang, S. Boling, and A. J. Mason, "A hardware-efficient scalable spike sorting neural signal processor module for implantable high-channel-count brain machine interfaces," *IEEE Trans. Biomed. Circuits Syst.*, vol. 11, no. 4, pp. 743–754, Aug. 2017.

[15] M. Zamani, D. Jiang, and A. Demosthenous, "An adaptive neural spike processor with embedded active learning for improved unsupervised sorting accuracy," *IEEE Trans. Biomed. Circuits Syst.*, vol. 12, no. 3, pp. 665–676, Jun. 2018.

[16] A. T. Do, S. M. A. Zeinolabedin, D. Jeon, D. Sylvester, and T. T. Kim, "An area-efficient 128-channel spike sorting processor for real-time neural recording with 0.175 μw/channel in 65-nm CMOS," *IEEE Trans. Very Large Scale Integrat. Syst.*, vol. 27, no. 1, pp. 126–137, Jan. 2019.

[17] D. Valencia and A. Alimohammad, "An efficient hardware architecture for template matching-based spike sorting," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 3, pp. 481–492, Jun. 2019.

[18] D. Valencia and A. Alimohammad, "A real-time spike sorting system using parallel OSort clustering," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 6, pp. 1700–1713, Dec. 2019.

[19] H. Hao, J. Chen, A. G. Richardson, J. Van der Spiegel, and F. Aflatouni, "A 10.8 μw neural signal recorder and processor with unsupervised analog classifier for spike sorting," *IEEE Trans. Biomed. Circuits Syst.*, vol. 15, no. 2, pp. 351–364, Apr. 2021.

[20] U. Rutishauser, E. M. Schuman, and A. N. Mamelak, "Online detection and sorting of extracellularly recorded action potentials in human medial temporal lobe recordings, in vivo," *J. Neurosci. Methods*, vol. 154, no. 1/2, pp. 204–224, Jun. 2006.

[21] S. M. A. Zeinolabedin, A. T. Do, D. Jeon, D. Sylvester, and T. T.-H. Kim, "A 128-channel spike sorting processor featuring 0.175 μw and 0.0033 mm^2 per channel in 65-nm CMOS," in *Proc. IEEE Symp. VLSI Circuits*, 2016, pp. 1–2.

[22] R. George *et al.*, "Plasticity and adaptation in neuromorphic biohybrid systems," *iScience*, vol. 23, no. 10, Oct. 2020, Art. no. 101589.

[23] M. Velliste, S. Perel, M. C. Spalding, A. S. Whitford, and A. B. Schwartz, "Cortical control of a prosthetic arm for self-feeding," *Nature*, vol. 453, no. 7198, pp. 1098–1101, 2008.

[24] R. D. Flint, Z. A. Wright, M. R. Scheid, and M. W. Slutzky, "Long term, stable brain machine interface performance using local field potentials and multiunit spikes," *J. Neural Eng.*, vol. 10, no. 5, Aug. 2013, Art. no. 056005.

[25] N. C. Rowland *et al.*, "Task-related activity in sensorimotor cortex in parkinsons disease and essential tremor: Changes in beta and gamma bands," *Front. Hum. Neurosci.*, vol. 9, p. 512, Sep. 2015.

[26] E. Nurse, B. S. Mashford, A. J. Yepes, I. Kiral-Kornek, S. Harrer, and D. R. Freestone, "Decoding EEG and LFP signals using deep learning," in *Proc. ACM Int. Conf. Comput. Front.*, 2016, pp. 259–266.

Final edited form was published in "IEEE Transactions on Biomedical Circuits and Systems". 16 (1), S. 94-107. ISSN: 1940-9990.
https://doi.org/10.1109/TBCAS.2022.3142987

[27] H. A. Gonzalez *et al.*, "Hardware acceleration of eeg-based emotion classification systems: A comprehensive survey," *IEEE Trans. Biomed. Circuits Syst.*, vol. 15, no. 3, pp. 412–442, Jun. 2021.
[28] F. Kelber *et al.*, "Mapping Deep Neural Networks on spinnaker2," in *Proc. Neuro-Inspired Comput. Elements Workshop*, New York, NY, USA: Association for Computing Machinery, Mar. 2020, pp. 1–3.
[29] S. Mallat, *A Wavelet Tour of Signal Processing*. New York, NY, USA: Academic, 1999.
[30] M. Gautschi *et al.*, "Near-threshold RISC-v core with DSP extensions for scalable IoT endpoint devices," *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 25, no. 10, pp. 2700–2713, Oct. 2017.
[31] S. Höppner *et al.*, "Adaptive body bias aware implementation for ultra-low-voltage designs in 22FDX technology," *IEEE Trans. Circuits Syst. II*, vol. 67, no. 10, pp. 2159–2163, Oct. 2019.
[32] S. Höppner *et al.*, "How to achieve world-leading energy efficiency using 22FDX with adaptive body biasing on an arm Cortex-M4 IoT SoC," in *Proc. IEEE Eur. Solid-State Device Res. Conf. (ESSDERC)*, 2019, pp. 66–69.
[33] F. Schraut, H. Eisenreich, S. Höppner, and C. Mayr, "A fast lock-in ultra low-voltage ADPLL clock generator with adaptive body biasing in 22 nm FDSOI technology," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, 2019, pp. 1–5.
[34] F. Schüffny, S. Hoeppner, S. Hänzsche, R. George, S. M. A. Zeinolabedin, and C. Mayr, "An ultra-low area digital-assisted neuro recording system in 22 nm fdsoi technology," *IEEE Trans. Circuits Syst. II: Exp. Briefs*, to be published, doi: 10.1109/TCSII.2021.3121034.
[35] M. S. J. Steyaert and W. M. C. Sansen, "A micropower low-noise monolithic instrumentation amplifier for medical purposes," *IEEE J. Solid-State Circuits*, vol. 22, no. 6, pp. 1163–1168, Dec. 1987.
[36] S. Gibson, J. W. Judy, and D. Marković, "Spike sorting: The first step in decoding the brain: The first step in decoding the brain," *IEEE Signal Process. Mag.*, vol. 29, no. 1, pp. 124–143, Jan. 2012.
[37] I. Obeid and P. D. Wolf, "Evaluation of spike-detection algorithms fora brain-machine interface application," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 905–911, Jun. 2004.
[38] S. M. A. Zeinolabedin, A. T. Do, K. S. Yeo, and T. T. H. Kim, "Design of a hybrid neural spike detection algorithm for implantable integrated brain circuits," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2015, pp. 794–797.
[39] C. Pedreira, J. Martinez, M. J. Ison, and R. Q. Quiroga, "How many neurons can we see with current spike sorting algorithms?" *J. Neurosci. Methods*, vol. 211, no. 1, pp. 58–65, Oct. 2012.
[40] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. IEEE*, vol. 105, no. 12, pp. 2295–2329, Dec. 2017.
[41] D. Henze *et al.*, "Simultaneous intracellular and extracellular recordings from hippocampus region CA1 of anesthetized rats," 2009, doi: 10.6080/K02Z13FP.
[42] A. Uran, Y. Leblebici, A. Emami, and V. Cevher, "An AC-Coupled wideband neural recording front-end with Sub-1 $mm^2 \times$fj/conv-step efficiency and 0.97 NEF," *IEEE Solid-State Circuits Lett.*, vol. 3, pp. 258–261, Aug. 2020.
[43] S. Wang *et al.*, "A 77-dB DR 16-Ch 2^{nd}-order Δ-$\Delta\Sigma$ neural recording chip with 0.0077mm^2/ch," in *Proc. Symp. VLSI Circuits*, 2021, pp. 1–2.
[44] D.-Y. Yoon, S. Pinto, S. Chung, P. Merolla, T.-W. Koh, and D. Seo, "A 1024-channel simultaneous recording neural SoC with stimulation and real-time spike detection," in *Proc. Symp. VLSI Circuits*, 2021, pp. 1–2.
[45] D. Wendler, D. D. Dorigo, M. Amayreh, A. Bleitner, M. Marx, and Y. Manoli, "28.7 a 0.00378mm^2 scalable neural recording front-end for fully immersible neural probes based on a two-step incremental delta-sigma converter with extended counting and hardware reuse," in *Proc. IEEE Int. Solid- State Circuits Conf.*, 2021, vol. 64, pp. 398–400.
[46] R. Q. Quiroga, Z. Nadasdy, and Y. Ben-Shaul, "Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering," *Neural Comput.*, vol. 16, no. 8, pp. 1661–1687, Aug. 2004.
[47] S. N. Kadir, D. F. M. Goodman, and K. D. Harris, "High-dimensional cluster analysis with the masked EM algorithm," *Neural Comput.*, vol. 26, no. 11, pp. 2379–2394, Nov. 2014.
[48] J. Wild, Z. Prekopcsak, T. Sieger, D. Novak, and R. Jech, "Performance comparison of extracellular spike sorting algorithms for single-channel recordings," *J. Neurosci. Methods*, vol. 203, no. 2, pp. 369–376, Jan. 2021.
[49] K. D. Harris, D. A. Henze, J. Csicsvari, H. Hirase, and G. Buzsáki, "Accuracy of tetrode spike separation as determined by simultaneous intracellular and extracellular measurements," *J. Neuriophysiol.*, vol. 84, no. 1, pp. 401–414, Jul. 2000.

Seyed Mohammad Ali Zeinolabedin (Member, IEEE) received the B.Sc. degree in electrical engineering from Azad University, Isfahan, Iran, in 2006, the M.Sc. degree in electrical engineering from the Isfahan University of Technology, Isfahan, Iran, in 2010, and the Ph.D. degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2017. He is currently a Postdoctoral Researcher with the Chair of Highly-Parallel VLSI Systems and Neuro-Microelectronics, Technische Universität Dresde, Dresden, Germany. His research interests include ultra-low power integrated system-on-chips with high energy efficiency for biomedical and digital signal processing applications. He was the recipient of the Low Power Design Context Award at ISLPED2016 and Best Demo Award at APCCAS2016.

Franz Marcus Schüffny received the Dipl.-Ing. (M.Sc.) degree in electrical engineering from Technische Universität Dresden, Dresden, Germany, in 2018. He is currently working toward the Ph.D. degree with the Chair of Highly-Parallel VLSI-Systems and Neuromorphic Circuits. His research interests include circuits for energy harvesting, bio implants, low-power systems-on-chip in advanced technology nodes, with special focus on data transmission and analog to digital converters. He has experience in designing full-custom circuits for bio related ADCS in academic research projects.

Richard George received the M.Sc. degree in medical engineering from HTW Saarland, Germany with a specialization Neural Engineering, in 2013, and the Ph.D. degree in computational neurosciences from the Institute of Neuroinformatics of UZH and ETH Zürich, Zürich, Switzerland, in 2018, for his work on structural plasticity in neuromorphic systems. He is currently a Postdoctoral Researcher with the Chair for Highly-Parallel VLSI-Systems and Neuromorphic circuits, Technische Universität Dresden, Dresden, Germany. His research focuses on creation of active and intelligent neuroprosthetic devices. His particular focus is in the creation of energy efficient computational architectures capable of processing electrophysiological signals and forming electrical response stimuli within biohybrid closed-loop systems.

Florian Kelber received the Dipl.-Ing. (M.Sc.) degree in information systems engineering from Technische Universität Dresden, Dresden, Germany, where he is currently working toward the Ph.D. degree with the Chair for Highly-Parallel VLSI-Systems and Neuromorphic Circuits. His research interests include the design of digital low-power accelerators to mitigate bottlenecks in multiprocessor system-on-chip systems and mapping of load-balanced high level algorithms on highly parallel architectures.

Heiner Bauer received the Dipl.-Ing. (M.Sc.) degree in electrical engineering from Technische Universität Dresden, Dresden, Germany, in 2017. He is currently working toward the Ph.D. degree with the Chair of Highly-Parallel VLSI-Systems and Neuromorphic Circuits. His research interests include the design of low-power microprocessors and programmable logic.

Final edited form was published in "IEEE Transactions on Biomedical Circuits and Systems". 16 (1), S. 94-107. ISSN: 1940-9990.
https://doi.org/10.1109/TBCAS.2022.3142987

Stefan Scholze received the Dipl.-Ing. (M.Sc.) degree in information systems engineering from Technische Universität Dresden, Dresden, Germany, in 2007. Since 2007, he has been a Research Assistant with the Chair of Highly-Parallel VLSI-Systems and Neuromorphic Circuits, Technische Universität Dresden. His research interests include design and implementation of low-latency communication channels and low-power multiprocessor system-on-chips.

Stefan Hänzsche received the Dipl.-Ing. (M.Sc.) degree in electrical engineering and the Ph.D. degree from Technische Universität Dresden, Dresden, Germany, in 2006 and 2015, respectively. His research interests include circuit design of analog to digital converters and mixed-signal simulation.

Marco Stolba received the M.Sc. degree in nanoelectronic systems from Technische Universität Dresden, Dresden, Germany, in 2016. He is currently working toward the Ph.D. degree with the Chair of Highly-Parallel VLSI-Systems and Neuromorphic Circuits. His research interests include network on chip (NoC), MPSoC architectures, digital design, and verification.

Andreas Dixius received the Dipl.-Ing. (M.Sc.) degree in information systems engineering from Technische Universität Dresden, Dresden, Germany, in 2014. He is currently a Research Associate with the Chair of Highly-Parallel VLSI-Systems and Neuromorphic Circuits, Technische Universität Dresden. His research interests include MPSoC architecture, digital design, code generation, physical implementation and design for test.

Georg Ellguth received the Dipl.-Ing. (M.Sc.) degree in electrical engineering from Technische Universität Dresden, Dresden, Germany, in 2004. Since 2004, he has been a Research Assistant with the Chair of Highly-Parallel VLSI-Systems and Neuromorphic Circuits, Technische Universität Dresden. His research focuses on low-power implementation techniques in multiprocessor system-on-chip.

Dennis Walter received the Dipl.-Ing. (M.Sc.) degree in information and system technology from Technische Universität Dresden, Dresden, Germany, in 2010. He is currently a Research Associate with the Chair of Highly-Parallel VLSI-Systems and Neuromorphic Circuits, Technische Universität Dresden. His research interests include focused on energy-efficient implementation of integrated digital circuits at ultra-low voltages and the required modeling of statistical timing variation effects. He is also responsible for physical signoff and actively involved in all advanced node tape-outs.

Sebastian Höppner received the Dipl.-Ing. (M.Sc.) degree in electrical engineering and the Ph.D. degree from Technische Universität Dresden, Dresden, Germany, in 2008 and 2013, respectively. He is currently a Research Group Leader and Lecturer with the Chair of Highly-Parallel VLSI-Systems and Neuromorphic Circuits. He is the author or co-author of more than 80 publications and 12 patents in his research field, which include circuits for low-power systems-on-chip in advanced technology nodes, with special focus on clocking, data transmission, and power management. He has experience in designing full-custom circuits for multiprocessor systems-on-chip (MPSoCs), like ADPLLs, register files and high-speed on-chip and off-chip links, in academic and industrial research projects. He is managing the full-custom circuit design and SoC integration for more than 12 MPSoC chips in 65 nm, 28 nm, and 22 nm CMOS technology. Currently, he leads the chip design of the SpiNNaker2 neuromorphic computing system within the Human Brain Project(HBP). He was the recipient of the Barkhausen Award.

Christian Mayr (Member, IEEE) received the Dipl.-Ing. (M.Sc.) degree in electrical engineering and the Ph.D. and Habilitation degrees from Technische Universitat Dresden, Dresden, Germany, in 2003, 2008, and 2012, respectively. He is currently a Professor of electrical engineering with Technische Universitat Dresden. From 2003 to 2013, he was with Technische Universitat Dresden, with a secondment to Infineon during 2004–2006. From 2013 to 2015, he did a Postdoc with the Institute of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland. Since 2015, he has been the Head of the Chair of Highly-Parallel VLSI-Systems and Neuromorphic Circuits, Technische Universitat Dresden. He is the author or co-author of more than 100 publications and holds four patents. His research interests include bio-inspired circuits, brain-machine interfaces, bio-inspired AI, and mixed-signal VLSI-design. He has acted as an editor/reviewer for various IEEE and Elsevier journals. He was was the recipient of several awards.