A 130-μW, 64-Channel Spike-Sorting DSP Chip

Vaibhav Karkare, Sarah Gibson, and Dejan Marković
University of California, Los Angeles, USA

Abstract—Spike sorting is an important processing step in various neuroscientific and clinical studies. An on-chip spike-sorting DSP must provide data-rate reduction while maintaining a power density much less than 800 μW/mm². Most existing designs either provide only spike detection for multi-channel processing, or they provide detection and feature extraction only for a single channel. We demonstrate a chip for detection, alignment, and feature extraction simultaneously for 64 channels. Spike-sorting algorithms identified from a complexity-performance analysis are implemented on ASIC using a Matlab/Simulink-based architecture design framework. The chip has a modular architecture, which allows it to be configured to process 16, 32, 48, or 64 channels. Inactive cores are power-gated to reduce power consumption when the chip operates for less than 64 channels. The chip is implemented in a 90-nm CMOS process and has a power dissipation of 130 μW (power density of 30 μW/mm²) when processing all 64 channels. A data-rate reduction of 91.25% (11.71 Mbps to 1.02 Mbps) is achieved.

I. INTRODUCTION

In many neuroscientific studies and clinical procedures, neuronal signals are recorded from the brain using implanted electrode arrays. Each electrode records signals from multiple neurons. It is often required to classify the recorded action potentials based on their source neurons. This process of classifying the recorded action potentials according to the neurons from which they originate is called spike sorting. Spike sorting can be divided into three major steps: (1) detection and alignment, (2) feature extraction (FE), and (3) clustering, as illustrated in Fig. 1.

In the traditional, offline spike-sorting approach, the data recorded from the brain is sent outside the body using wires. Wireless transmission of data is highly desirable as it allows for freedom of movement of the subject and also reduces the risk of infection caused by the wire. However, power constraints preclude the transmission of raw data outside the body for a large number of channels. Performing spike sorting on a chip implanted inside the skull is a promising option to enable the processing and wireless transmission of information for high-channel-count systems. Spike sorting on an implantable chip also provides faster processing, thereby enabling the real-time, many-channel, closed-loop control necessary for brain-machine interfaces. An implantable spike-sorting DSP chip is subject to aggressive power density constraints. The power density needs to be much less than 800 μW/mm² which is the power density known to damage brain cells [1]. In addition, a high data rate reduction is necessary in order to support a large number of channels. Implantable ICs should also have a small area, which would allow for integration of the chip with the base of an implanted electrode array.

Most of the spike-sorting DSPs published in literature provide only detection for multiple channels [2], [3], or they provide detection and feature extraction but only support a single channel at a time [4]. In this work, we demonstrate a spike-sorting DSP chip that can perform detection, alignment, and feature extraction simultaneously for 64 channels. The power density of the chip is 30 μW/mm², which is 27 times lower than the power density limit. This low power density is achieved by careful algorithm selection and technology-driven architecture implementation. Evaluation of hardware-friendly clustering algorithms and their implementations is part of our work in progress.

II. ARCHITECTURE DESIGN

A. Algorithm selection

In our previous work [5], we analyzed the complexity-performance tradeoffs of spike-sorting algorithms using simulated neural data across a wide range of SNRs (−15 dB to 20 dB) to identify the most robust algorithms for hardware implementation. The probability of detection (P_D), probability of false alarm (P_FA), and classification accuracy were the criteria used to evaluate the performance of the algorithms. We identified the non-linear energy operator (NEO) for detection [6], maximum derivative for alignment, and discrete derivatives for FE [7] as the most practical and robust algorithms.

B. Functional description

Figure 2 illustrates the architecture of the spike-sorting DSP. The NEO block calculates \(\psi(n)\) defined by:

\[
\psi(n) = x^2(n) - x(n+1) \cdot x(n-1) \quad (1)
\]

for each input sample \(x(n)\). During the training period, the threshold-calculation module calculates the threshold as:

\[
Thr = \frac{1}{N} \sum_{n=1}^{N} \psi(n). \quad (2)
\]

We chose the number of samples in the training period \(N\) to be a power of 2, so that the division by \(N\) is reduced...
to a simple shift operation. \( C \) in Eq. 2 is a constant empirically chosen to be 8 in our implementation. The detector signals a spike-detection event when \( \psi(n) \) exceeds the calculated threshold. When a spike is detected, 72 samples (24 pre-threshold-crossing or “preamble” samples and 48 post-threshold-crossing samples) of raw data are saved in the register-bank memory as the recorded action potential. The maximum derivative block calculates the offset for a 48-register-bank memory as the recorded action potential. When a spike is detected, 72 samples \( \psi \) typically chosen to be 8 in our implementation. The detection and alignment modules occupy 33% and 47% of the area of a single-channel DSP core occupied by registers (including the register-bank memory).

C. Architecture selection

We assume a sampling frequency of 24 kHz per channel and 8-bit resolution per sample for the input neural data. The delay requirement for single-channel processing is \( 3 \times 10^3 \)-times longer than the delay at the minimum-delay point (MDP) for a chain of 466 FO-4 inverters, corresponding to the critical path of a single-channel architecture, in a 90-nm CMOS process. The switching activity of the DSP is estimated to be around 0.05, assuming a spike firing rate of 100 Hz. The switching activity is low due to two reasons: (a) The sampling rate is around 4 times the Nyquist rate which causes high temporal correlation at the input, (b) The detection and alignment modules switch only upon a spike-detection event. The low switching activity factor and sampling frequency hint at a leakage-power-dominated design.

Given the large difference in the application delay requirement and the operating frequency that the design can support, we chose to reduce the power consumption by supply-voltage scaling. Figure 3 shows the energy-delay tradeoff for the design. The curve shows that a direct-mapped single-channel implementation of a spike-sorting DSP (with voltage scaling to a point that meets the required delay) results in a leakage-dominated design beyond the minimum-energy point (MEP). In order to bring the operating point closer to the minimum-energy point, it is desirable to interleave the design to support multiple channels. Data-stream interleaving saves the leakage power consumption of the logic, while increasing the switching power of the registers. The total power of a 64-channel design, therefore, has a minimum around 4- to 8-channel interleaving, as shown in Fig. 4. Interleaving also reduces the area by sharing the logic for multiple channels. Figure 4 shows the percentage savings in area with the number of channels interleaved. It can be seen that the savings in area saturate at 16-channel interleaving. In favor of lower area, we chose 16-channel interleaving, which gives us area savings of around 47%. The 64-channel spike-sorting DSP is thus implemented using a modular architecture of four cores, each of which is interleaved to support 16 channels.

Since the spike-sorting DSP is a leakage-dominated design,
power gating is an attractive circuit design technique to lower the energy consumption of the design (as indicated in Fig. 3). The modular architecture allows us to power-gate the inactive cores, thus reducing the total power consumption. Power gating provided a 70% reduction in the leakage and clock-distribution power of the inactive cores. Power gating also increased the “stack effect” in an active core and provided a three-times reduction in its leakage current. The low core voltage for the chip allowed us to implement power gating with NMOS-only devices on both VDD and GND lines, which implies a narrower sleep transistor for a given degradation in performance. The first DSP core on the chip was not power-gated for debug purposes.

D. Power and area minimization

In addition to data-stream interleaving and power gating, we used logic restructuring and wordlength optimization to reduce the power and area consumption. The logic is structured to reduce switching-activity propagation in the core. For instance, since the alignment and FE blocks receive valid inputs only when a spike is detected, switching their inputs only upon detection reduces their switching power by 5 times. Opportunities for hardware sharing among blocks were also exploited. For example, the hardware for calculation of the NEO $\psi(t)$ is shared between the threshold calculation and detection operations. An automated wordlength (WL) optimization tool was used to determine the WL of the signals in the design [9]. WL optimization was performed by specifying an iterative MSE constraint on the signals to avoid any detection errors due to signal quantization. Optimizing the WL allowed us to achieve a 15% area reduction compared to a design with a fixed MSE specification of $5 \times 10^{-6}$ (equal to input MSE).

III. CHIP IMPLEMENTATION AND VERIFICATION

A. Design methodology

We used a Matlab/Simulink-based graphical design environment using the Synplify DSP blockset for architecture design [8]. The spike-sorting algorithms were modeled in Simulink, which provides a bit-true, cycle-accurate representation of the design. The design process avoids design re-entry since both the HDL and the test vectors were auto-generated from the Simulink model with the Synplify DSP blockset. This approach reduced the design time and also allowed us to easily evaluate different architectural tradeoffs based on synthesis estimates of various architecture options. Since results from technology characterization can be included in the Simulink model, an estimate of area and performance numbers was obtained early in the design phase.

B. Chip implementation

The 64-channel design was implemented as four cores that process 16 channels each. The input clock is divided by 4 and is fed to each of these cores. Interleaved input data for 64 channels is clocked at a rate of 1.6 MHz, and is split by a serial-to-parallel (S/P) converter into 4 data streams which are clocked at 400 kHz. At the output the data is combined again by a parallel-to-serial (P/S) converter to form a 64-channel interleaved stream. Figure 5 shows the die photograph of the chip. The chip operates at a reduced core voltage of 0.55 V. In order to provide a 1.2-V swing at the I/O pads, a level converter using a cross-coupled PMOS device was designed. The chip supports three modes of operation, which can be chosen separately for each of the four cores. The training period for calculation of the detection threshold can be initiated for each channel independently without affecting the processing on other channels. Further, the modularity in the architecture allows for a straightforward extension of the current design to support a higher number of channels.

C. Chip verification

ASIC verification was performed with an FPGA-aided test environment using the IBOB (FPGA) board [10]. The IBOB feeds the test patterns to the ASIC through the GPIO interface. It reads the outputs from the ASIC from the same interface and performs a cycle-accurate comparison of the outputs with the expected results. The I/O interface between the PC client and the IBOB is based on the BEE platform studio design flow [11]. The power consumed by the chip is measured with the help of an on-board current-sense circuit.

Figure 6 shows a sample output of the spike-sorting DSP for simulated neural data at an SNR of $-2.2$ dB. It was known a priori that the raw data contained spikes from two neurons. These spikes are difficult to discriminate in the time domain (Fig. 6-b), but can be easily separated in the feature domain (Fig. 6-c). Subsequent clustering of the features separated the spikes with a classification accuracy of 92%, which is
significantly better than the classification accuracy of 77% when raw spike waveforms (Fig. 6-b) were used for clustering. In addition to simulated data, the chip was used to process real neural data recorded from the hippocampus of human epilepsy patients. Examination of the clustering outputs of this data found to be consistent with manual inspection by neuroscientists.

IV. MEASUREMENT RESULTS

The power of the chip is measured to be 130 μW when operated in feature-extraction mode for 64 channels at a core voltage of 0.55 V. When a single core is operated with the remaining cores in sleep mode, the power consumed by the chip is 52 μW. Leakage and clocking power (53% and 41%, respectively) are the dominant contributors of power in the register-dominated, interleaved, 16-channel spike-sorting DSP core. Signal switching power only contributes around 6% of the power of a DSP core. The power for single-core operation is not 4-times lower than that for four-core operation because the sleep devices do not completely eliminate the power consumption and 2-times reduction in power density compared to the previous state-of-the-art spike-sorting DSPs. Our design performs simultaneous processing of 64 channels with a power consumption/power density lower than previous designs that perform only detection and feature extraction for multiple channels or detection and feature extraction for a selection and technology-driven architecture implementation to damage brain cells. We demonstrate that careful algorithm selection and technology-driven architecture implementation allow us to perform multi-channel spike sorting on chip while easily meeting the design constraints for an implantable device. Demonstration of the ability to perform digital signal processing with low power motivates the use of digital signal processing for the reduction of power consumption of analog, mixed-signal, and RF components in a neural recording system. Future work includes the incorporation of a digitally assisted analog front-end/RF and the inclusion of on-the-fly clustering.

VI. ACKNOWLEDGMENTS

The authors thank the NSF grant 0847088 for funding this work, and IBM TAPO for chip fabrication.

REFERENCES