Ultra-Low-Power Links for Brain Probing

Vaibhav Karkare and Dejan Marković
Electrical Engineering Department
University of California, Los Angeles
Los Angeles, CA 90095, USA
{vaibhav, dejan}@ee.ucla.edu

Abstract—This paper discusses signal processing and circuit techniques for ultra-low-power many-channel wireless telemetry. Signal processing has to be performed online to reduce transmit data rate. The choice of signal compression algorithm can greatly reduce energy, much beyond levels attainable by implementation. Since real-time performance required from biomedical implants is much lower than delay of logic gates, computations are done in near-\(V_T\) regime. Voltage scaling alone does not guarantee high energy efficiency. Architecture design has to position path delay onto pareto-optimal energy-delay curve. Power density levels below 50\(\mu\)W/mm\(^2\) can be achieved. With high signal compression, analog front-end begins to dominate energy consumption.

Index Terms—Spike sorting, Subthreshold operation, Energy-delay tradeoff

I. INTRODUCTION

Electrophysiological recordings have become essential for several clinical and scientific applications. Among the various biosignals neural action potentials present the most stringent constraints because they have the highest bandwidth and lowest amplitudes. Typical data rates for a 24 kS/s neural recording system are shown in Fig. 1.

![Fig. 1. Many-channel neural wireless telemetry. DSP compresses data to meet functional requirements and/or reduce transmit (Tx) power. Brain-machine interface (BMI) and neuroscience (NS) need different types of processing.](image)

Neural recording systems acquire action potentials using high-density implanted electrode arrays [1]. To avoid risk of injury and infection it is desirable to wirelessly transmit the recorded signals. However, transmission of many channels of raw data can exceed power levels required by implantable medical electronics [2]. This makes it necessary to implement local processing (DSP, Fig. 1) to reduce the output data rates.

II. NEURAL SPIKE SORTING

Each electrode in an implanted electrode array records signals from multiple neurons. Many applications require classification of the recorded action potentials according to their source neurons by a process known as spike sorting. In addition to being a functional requirement spike sorting provides data-rate reduction that allows wireless transmission for multiple channels. In this paper, spike sorting DSPs are used to illustrate the design methods for biosignal processors.

A. Minimum Energy per Operation

Biomedical implants present an attractive opportunity for subthreshold computations. Sample rates required for BMI go up to ~30 kHz for neural spikes, they are much lower for ECoG or EEG. Today’s GHz technology is too fast for kHz rates. As shown in Fig. 2, there is an optimal energy-delay tradeoff that CMOS logic can achieve by voltage scaling. Gate sizing and \(V_T\) adjustment can also be applied. Minimum-delay point (MDP) and minimum-energy point (MEP) are bounds on achievable performance and energy levels. All designs above the pareto line are suboptimal.

![Fig. 2. Energy-delay tradeoff in digital logic. Today’s GHz technology is leakage-dominated at kHz sample rates found in biomedical implants. Architectural and circuit solutions for leakage reduction are needed.](image)

Subthreshold or near-threshold operation, around MEP, results in only about 10x energy reduction. Performance required from biomedical DSPs is outside the range of the pareto curve. If digital logic is forced to operate in the kHz regime, energy will be wasted. Further scaling \(V_{DD}\) doesn’t work since the rate of increase in leakage energy equals the rate of increase in delay (exponential). Keeping the original \(V_{DD}\) creates timing slack, but the slack time is spent leaking and there is no reward for early computation. Architecture and circuit solutions are needed to overcome leakage energy that is implicit to kHz-rate computations (dashed line in Fig. 2).

B. Design Environment

Algorithmic impact on energy could be many orders of magnitude, much higher than that from implementation. Close interaction between algorithm and hardware can be managed...
using Simulink design model, Fig. 3. Feedback about hardware cost is accounted during algorithm design, and algorithmic features are leveraged during hardware mapping.

A high-level Simulink model, built on top of MATLAB, quickly captures initial algorithm design. Simulink also uses hardware libraries and outputs RTL for FPGA or ASIC implementation. This provides a quick estimation of hardware resources. The bit-true, cycle-accurate model also allows the exploration of many architectural realizations in the energy-area space.

C. Hardware-Friendly Algorithms

Identifying hardware-friendly algorithms is a prerequisite to an efficient hardware implementation. In [3], we evaluated the complexity-accuracy tradeoffs of popular spike-sorting algorithms using simulated neural data over SNR range from \(-15\) dB to 20 dB. As an example, Fig. 4 shows the complexity-accuracy tradeoff for feature extraction algorithms. The knee point, the discrete derivatives (DD) algorithm, is the chosen feature extraction algorithm.

Feature extraction is shown to reduce data rate by an order of magnitude [6]. Highest data reduction is achieved by sending spike IDs obtained as a result of online clustering. A multi-channel implementation of the online clustering [4] would actually increase the total system power due to the high DSP power, dominated by the memory required for clustering. In [5], we introduced two-phase implementation to reduce the DSP power. Channel-specific cluster means are first calculated sequentially, followed by parallel cluster mapping, resulting in \(>5x\) memory reduction. A noise-tolerant \(L_1\)-norm distance metric is used to further simplify computations.

III. DSP ARCHITECTURE AND CIRCUITS

Key architecture- and circuit-level techniques to reduce power in spike-sorting DSPs are discussed in this section.

A. Multi-Channel Interleaving

A single-channel neural-spike DSP operating at kHz-rate is limited by leakage, Fig. 5, even when \(V_{DD}\) is scaled to meet the target delay. Interleaving of many channels allows us to reduce the impact of logic leakage and helps us bring the design closer to MEP (dots in Fig. 5). While the timing slack afforded by the design may tempt a designer to interleave a very high number of channels, the increase in register switching energy limits the number of channels that can be interleaved to about 16 [6].

![Energy-delay tradeoff for spike-sorting DSP processor with varying degree of interleaving. Channel interleaving shares logic area and increases datapath clock rate, both resulting in reduced leakage energy. Beyond 8 channels, design becomes register dominated and energy increases. The use of D flip-flops and register files is recommended for aggressive voltage reduction.](image)

B. NMOS Power Gating

Only a few channels of a multi-channel DSP acquire usable data. The unused channels should, therefore, be power gated to reduce leakage power. Conventional power gating uses PMOS header device. Gating a subthreshold design is different. As shown in Fig. 6, the \(I_{on}/I_{off}\) ratio of an NMOS device is higher than that of PMOS device for the near-threshold \(V_{DD}\) used in spike-sorting DSPs. The gate of the NMOS device is operated by the I/O voltage. A higher overdrive (on mode) and reverse \(V_{GS}\) bias (off mode) reduce the device size and leakage. NMOS headers, thus provide a more power-efficient gating option for subthreshold DSPs than their PMOS counterparts.

C. Register-Based Memory Clocking

Register-bank memories can operate in subthreshold, unlike...
SRAMs, and achieve a 2x lower power than SRAMs [5]. Also, random access to individual spike samples is not required since spike sorting algorithms deal with complete spike waveforms. Hence, the memory can be organized as spike registers (SRs), each containing one spike waveform, Fig. 7. To select a spike, the decoder toggles corresponding $\text{Clk}_{\text{en}}$ pulse for one clock cycle. This enables the clock to the first sample R(1). In the following clock cycles, the $\text{Clk}_{\text{en}}$ pulse is shifted to samples 2 to $N_{s}$. Only the clock to a single register (R) is active at any time. This reduces the clocking power of the memory by an 8x.

![Fig. 7. Selectively-clocked spike memory. Spike memory is implemented using register files to allow subthreshold operation for aggressive power reduction.](image)

IV. DISCUSSION

The total system power is significantly influenced by the choice of output data as shown in Fig. 8. Raw data can be sent for a relatively small number of channels, beyond which Tx power has significant impact. On-chip compression reduces Tx and overall power. In systems that use spike IDs (high compression), analog front-end becomes the main power consumer. Apart from balancing AFE, DSP, and Tx components of power, the system should also offer flexibility in choosing different output modes and data-compression ratios. For invasive recordings, temperature-based energy scavenging can be considered.

![Fig. 8. Systems with high amount of data compression reduce transmitter energy and dictate the need for energy-efficient analog front-end (AFE).](image)

V. CONCLUSION

Online data processing sets the relative power consumed in the analog front end, DSP, and radio subsystems. Subthreshold processing is required from kHz-rate systems such as biosignal sensors. Interleaved channel processing and NMOS power gating effectively reduce leakage power. Power-efficient analog front-end is necessary for BMI applications. Flexibility to support multiple signal bands and data processing options can be considered.

ACKNOWLEDGMENT

This work is supported by NSF grant number 0847088.

REFERENCES


