Proceedings of the 3rd Annual International IEEE EMBS Special Topic Conference on Microtechnologies in Medicine and Biology
Kahuku, Oahu, Hawaii ú 12 - 15 May 2005

Augmenting Real-Time DSP in Implantable High-Density Neuroprosthetic Devices

K. Oweiss, Y. Suhail, K. Thomson, J. Li, A. Mason

Electrical and Computer Engineering Department, Michigan State University, East Lansing, MI

Abstract – **Developments in microfabrication of high-density electronic interfaces to the central nervous system rapidly evolved in recent years. Nonetheless, there is a lack of higher digital signal processing capability to cope with the larger data throughput. Multiresolution analysis by means of the wavelet transform has shown to yield substantial processing and compression capabilities of high volumes of neural signals while preserving the necessary information needed to understand the neural coding mechanism. We describe a systems approach for reducing the complexity of *on-chip* discrete wavelet transform (DWT) computation for multiple data channels. The reduction is achieved by exploiting regularity in the filtering steps of the lifting-based DWT algorithm associated with negligible degradation in signal fidelity with integer fixed point arithmetic representation. The main advantages of the proposed architecture lie in the scalability to an arbitrary number of channels and decomposition levels. The results demonstrate that on-chip computation is feasible prior to data transmission, permitting large savings in bandwidth requirements and communication costs.**

Keywords – **Implantable Neuroprosthetics; wavelet transform; neural signal processing; electrode arrays**

I. Introduction

Neuroprosthetic devices play a vital role in helping patients with severe motor disorders achieve a better lifestyle by enabling direct interface to the central nervous system at various levels. These devices generally consist of an array of microelectrodes implanted subcutaneously to record electrical signals and selectively stimulate neuronal populations in brain structures of interest.

For implantable neuroprosthetic applications, advances in microfabrication technology have greatly accelerated the integration of high-density microelectrode arrays on a single device [1]. A typical state-of-the-art microelectrode array can have as many as 1000 electrodes integrated on a single device [2]. This is considered a modest number considering that approximately 30,000 neurons and 2.4×10^8 synapses (assuming 8,000 synapse/neuron) may exist in a cubic millimeter of cortex tissue [3].

In a typical recording experiment (uplink) with a 96-electrode array sampled at 25kHz per channel, the aggregate rate is 2.4 Msamples/sec. At 12 bits/sample, the bit rate is nearly 29 Mbps. Equally important, a typical stimulation experiment (downlink) using a retinal prosthesis with 100 electrodes would require 10 Mbps to allow a 625-100 pixel image to be effortlessly read by the patient [4]. Existing biotelemetry developments for extracutaneous transmission quickly erode in the face of such data volume [5][6]. It is becoming clear that the onus is on the signal processing community to accommodate quantum advances in neural data processing and compression *on chip* to cope with these rapid advances in microprobe technology.

A viable alternative is to extract the useful information from the raw neural data prior to extracutaneous transmission making use of the sparsity of the neural signals on the time base as illustrated in Figure 1. In such case, more advanced signal processing needs to be performed early in the data stream. Such alternative may reduce the data throughput by approximately 75% in bursting activity and 85-90% in spontaneous neural activity.

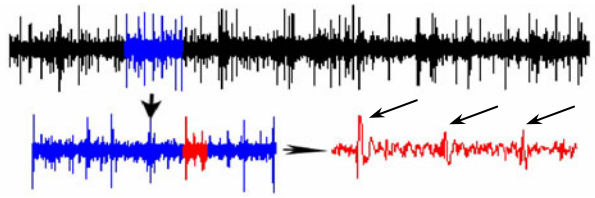

Figure 1: Traces from a typical train of neural spikes from multiple neurons recorded by a single electrode. The top trace contains 4 seconds (at 20kHz sampling rate), while the bottom left contains 400 msec, and the bottom right is 40 msec. The information is contained for the most part in the outliers from the baseline noise (indicated by arrows).

New methods for processing neural data [7][8] have shown that multiresolution analysis by means of the discrete wavelet transform (DWT) representation of the neural signals due to its high compression capabilities, and its ability to preserve the spike information of the neural signals [9]. Moreover, the development of the *lifting* scheme for computing the DWT coefficients yielded substantial reduction in hardware requirements [10]. Some recent efforts have been devoted to implement the lifting scheme for single data sequences [11][12]. However, these approaches were focused on optimizing processing speed and are neither suitable for implantable neuroprosthetic applications (with severe limitations on both power and area), nor suitable for "streaming" neural data in real-time because it is assumed that an entire data frame is available in the memory.

The work we present herein constitutes a novel approach to design efficient architectures for lifting-based DWT computation for multichannel neural data processing such that an appropriate compromise among power, required bandwidth, computational load, memory, and delay time is achieved within the constraints imposed by implantability requirements. We primarily focus in this paper on recording array microsystems since the applications are much more widespread among the neuroscience and bioengineering communities.

II. Theory

A coarse layout of our approach for augmenting the DSP components in a neural interface system front-end is

illustrated in Figure 2. Our focus in this paper is on the implementation of the DWT block. Other blocks are reported elsewhere [13].

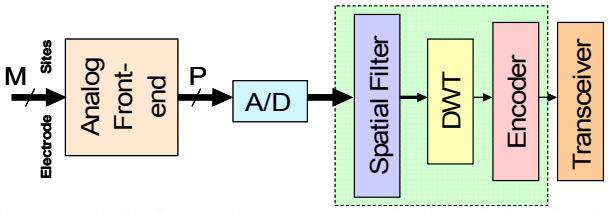

Figure 2: Block diagram for an augmented electronic interface in neuroprosthetic devices. The highlighted area represents the new DSP components needed to reduce the bandwidth (indicated by the relative width of the arrows at intermediate stages) for extra-cutaneous transmission of recorded neural data.

A. Hardware Design Constraints

For implantable neuroprosthetic applications, both power and area for the wavelet transform hardware are severely limited. For reliable hardware design, the following considerations can relax some of the design constraints:

- Real-time decoding of neural signals for closed-loop control applications has been demonstrated from specific brain regions provided that adequate telemetry and real-time processing takes place [17][18].
- A neuroprosthetic system can tolerate delays of *several* milliseconds. Therefore, the DWT can be simultaneously computed for multiple data channels and multiple decomposition levels without compromising real-time operation.
- From a packaging viewpoint, 32 data channels from a closely-spaced device are a manageable number of interconnects and signal conditioning components in the volume over the tissue being considered [1].
- A 14mm burr hole used by neurosurgeons is sufficient for a package of several hundred thousand transistors.
- 25kHz quantized to 10 bits is a usual and logical sampling scheme if scaling is controlled.

On the other hand, the following considerations restrict the available design space:

- The implant power dissipation is severely constrained because brain tissue temperature should not rise above 1-2°C in order not to damage surrounding neurons and cause brain trauma.
- The implant size is constrained because any rise above the cortical surface should not be more than 1mm in order not to compromise the stability of the implant by making it independent of the skull as the brain moves inside the cranial cavity.

B. Lifting DWT Overview

From an algorithmic point of view, the integer lifting wavelet transform is an effective method to reduce computational load, reducing the number of required transistors. However, through custom design of hardware, circuit area and power consumption can be further reduced. The lifting method is based on 3 steps: First, *splitting* the data $\underline{a}^j$ at any given decomposition level *j* into *even* and *odd* samples $\underline{f}_0$ and $\underline{g}_0$, respectively; Second, *predicting* the odd samples from the even samples such that the prediction error becomes the high pass coefficients $\underline{g}_1$; and Third, *updating* the even samples with the $\underline{g}_1$ coefficients to obtain the approximation coefficients $\underline{f}_1$. This process is repeated *N* times. The data at each step, after applying the new filters are labeled as $\underline{f}_0, \underline{f}_1, ..., \underline{f}_N$ and $\underline{g}_0, \underline{g}_1, ..., \underline{g}_N$, respectively At an arbitrary prediction and update step *n* within level *j*, the prediction and update filters $T_n(z)$ and $S_n(z)$ are obtained by factorizing the wavelet low pass and high pass filters into *N* lifting steps [10] as illustrated in Figure 3. The last step is a multiplication by a scaling factor G and G^{-1} to obtain the approximation $\underline{a}^{j+1}$ and details $\underline{d}^{j+1}$ of the next level.

Recent work has demonstrated that the *symmlet*4 wavelet function is advantageous over other wavelet basis

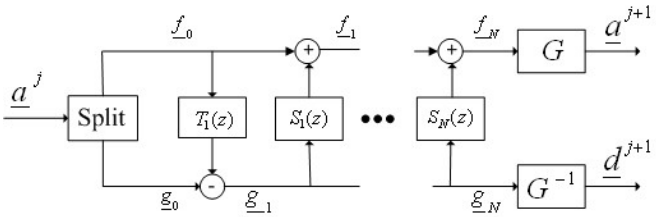

Figure 3: Lifting-scheme for computing a single level of DWT decomposition. The polynomials $T_n(z)$ and $S_n(z)$ are obtained through factorization of the wavelet low pass and high pass filters

families for processing neural signals [14]. For a lifting factorization of the *symmlet*4 wavelet, the filtering steps corresponding to Figure 3 can be written as [15]

$$\begin{aligned} &step1: g_1[i] = g_0[i] + C_1 \times f_0[i] \\ &step2: f_1[i] = f_0[i] + C_2 \times g_1[i] + C_3 \times g_1[i+1] \\ &step3: g_2[i] = g_1[i] + C_4 \times f_1[i] + C_5 \times f_1[i-1] \\ &step4: a[i] = f_1[i] + C_6 \times g_2[i] + C_7 \times g_2[i-1] \\ &step5: d[i] = g_2[i] + C_8 \times a[i+1] \end{aligned} \quad (1)$$

where the intermediate values $f_1[i]$, $g_1[i]$, and $g_2[i]$ are discarded after being used, $f_2[i] = a[i]$ is the resulting approximation coefficient, $g_3[i] = d[i]$ is the resulting detail, and C_1 through C_8 are the constant coefficients for the prediction and update filters resulting from factoring the *Symmlet*-4 wavelet basis and are given in Table I.

C. Integer Lifting DWT

Two new features in the lifting implementation can further reduce the computational load and memory requirements:

• *Integer Approximation*: Rounding-off wavelet coefficient data to be represented using fixed-point integer precision format has shown to yield negligible degradation in signal fidelity in the reconstruction phase [14]. The data is first scaled to obtain data samples within a 10-bit integer precision. The integer approximation is then computed for the scaled data. The integer-to-integer transformation involves rounding-off the result of the lifting steps T and S that are added to the odd and even data samples, respectively [16]. The last step that requires scaling by G and G^{-1} is omitted. Hence, the dynamic range of the transform at each level will now change by G.

• *Quantization of Filter Coefficients*: The filter coefficients C_1 through C_8 are further quantized to reduce multiplier complexity. The wavelet filter coefficients were computed and quantized to different resolutions ranging from 4 bits to 12 bits. The 4-bit representation is given in Table I.

TABLE I
SYMMLET-4 DWT FILTER COEFFICIENTS AND THEIR 4-BIT INTEGER APPROXIMATIONS

Coefficient	Floating point value	Scaled	Coefficient	Floating point value	Scaled
C_1	0.39114	3	C_5	0.1620	2
C_2	-0.12439	-1	C_6	0.4313	3
C_3	-0.33924	-5	C_7	0.1459	1
C_4	-1.41951	-11	C_8	-1.0492	8

D. Hardware Implementation

The arithmetic operations needed for implementing the lifting approach in (1) have a noticeable regularity that permits any arbitrary step in Fig. 3 to be defined as

$$W_{ij} = X + C_i Y + C_j Z \tag{2}$$

where the variables W, X, Y and Z take the values of f, g and, C_i and C_j are the constant coefficients given in Table I. We will refer to the hardware block that computes this equation as a computational node (CN). Observation of (2) shows that the three-term CN can be utilized to execute each of the filtering steps. Figure 4 illustrates the block diagram of such a circuit which has been custom designed to the data and coefficient bit width established by the algorithm while minimizing power and area requirements. It consists of two 6x10 Booth multipliers and a three-term adder. To accommodate integer processing, the filter coefficients are scaled by a binary factor that is easily removed at the output of each multiplier using a shifting operation. In fact, the shifting can be done without any additional hardware by properly routing the signals between the multipliers and the three-term adder to effectively divide the multiplier output by the coefficient's binary scaling factor. The resulting circuit efficiently implements steps 2-4 of (1) and can also be used for steps 1 and 5 with a zero value coefficient applied to the unused multiplier inputs to eliminate unnecessary transistor switching and power consumption.

To preclude buffering, the DWT computation has to actively take place while an entire data frame is being acquired across multiple channels. All channels must have access to a processor to execute a single level of decomposition. Highest priority is given to the lowest level of decomposition, since data is more frequently received at this level. For M channel input, they must be executed in sequential order, since each channel must have access to the CN once in a single interval.

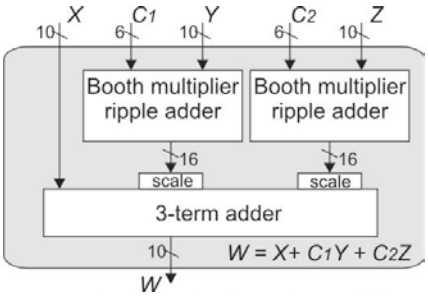

Figure 4: Customized CN to implement integer lifting wavelet transform using binary scaled filter coefficients

E. Pipeline and Sequential Processing Architectures

To execute the entire set of lifting filter equations in real time, two architectures have been explored. The first is a pipelined approach; similar to what has been reported in the literature [11][12], that is capable of very efficiently executing the filter equations with minimum storage registers. The second approach takes advantage of the limited input signal bandwidth and assumes that the propagation delay of the computational node allows the circuit to be clocked much faster than the input data sampling frequency. In this case, a single CN can be employed to sequentially process the filter steps. Both architectures have been thoroughly analyzed to determine which approach is best suited to the power and area requirements for multiple channels and multiple levels of wavelet decomposition. For the lack of space, we only report the results pertinent to the IWT and CN implementation. Hardware optimization and performance analysis is reported elsewhere [19].

III. RESULTS

We quantified the distortion in the average neural spike signals in terms of the integer approximation and quantization processes introduced in the lifting computation. The spectrum of the residual quantization and round-off noise is illustrated in Figure 5 and demonstrates that almost no degradation is observed in the signal spectrum as a result of these approximations. Another immediate observation from these results suggest that there is a negligible increase in the multiplier complexity corresponding to a dramatic increase in the SNR in which the quantization noise and the rounding-off error for integer precision are included.

IV. Conclusion

A new approach to augment the DSP capability of high-density electronic interfaces to the nervous system has been developed. The approach is based on exploiting the feasibility of computing multilevel wavelet transform using the lifting scheme for multiple channels in real-time. A computational node has been custom designed for the quantized integer lifting DWT. We have quantified the performance of the proposed approach with integer fixed-point precision arithmetic to minimize computational load and memory requirements. We have demonstrated that negligible degradation in the signal is observed with substantial reduction in circuit complexity.

Chip size and power constraints impose severe limitations on the complexity of DSP components post A/D conversion. The results we presented demonstrate that, with appropriate custom DSP design, more real-time processing capability is feasible in implantable neuroprosthetic device technology to overcome transmission bandwidth bottlenecks.

References

[1] K. Wise, D.J. Anderson, J.F. Hetke, D.R. Kipke, and K. Najafi, **"**Wireless Implantable Microsystems: High-Density Electronic Interfaces to the Nervous System," *Proc. of the IEEE*, (92)-1, 76–97, Jan. 2004

[2] M. Ghovanloo, K. D. Wise and K. Najafi, "Toward a button-sized 1024-site wireless cortical microstimulating array," *Proc. 1st Int. IEEE/EMBS Conf. Neural Engineering*, pp. 138–141, Mar. 2003

[3] V. Braitenberg and A. Schuz, *Cortex: Statistics and Geometry of Neuronal Connectivity*. New York: Springer-Verlag, 1998

[4] K. Cha, K. Horch, and R. A. Normann, "Simulation of a phosphene based visual field: Visual acuity in a pixelized vision system," *Ann. Biomed. Eng.*, vol. 20, pp. 439–449, 1992

[5] G.A. DeMichele, P.R. Troyk, **"**Integrated multichannel wireless biotelemetry system," *Proc. 25th IEEE EMBS*, (4) , 3372 – 3375, Sept. 2003

[6] C. Bossetti, J. Carmena, M. Nicolelis, and P. Wolf, "Transmission Latencies in a telemetry linked Brain Machine interface," *IEEE Trans. On Biomedical Engineering*, 51, (6), pp 919-924, June 2004

[7] K. Oweiss, D. Anderson, M. Papaefthymiou, " Optimizing Signal Coding in Neural Interface System-on-a-Chip Modules," *IEEE Conf. on Eng. in Med. & Bio.*, pp. 2016-2019, September 2003

[8] K.G. Oweiss and D.J. Anderson, " A Unified Framework for Advancing Array Signal Processing Technology of Multichannel Microprobe Neural Recording Devices, " *Proc. IEEE Int. Conf. EMBS/MMB*, pp. 245-250, 2002

[9] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 2nd edition, 1999

[10] I. Daubechies and W. Sweldens, "Factoring wavelet transforms into lifting steps," *J. Fourier Anal.* Appl.4(3), pp. 245-267, 1998

[11] H. Liao, M.Kr. Mandal and B.F. Cockburn, "Efficient architectures for 1-D and 2-D lifting-based wavelet transforms," *IEEE Trans. on Signal Processing*, volume: 52 , Issue: 5 , May 2004 pp:1315 – 1326

[12] P.Y. Chen, "VLSI Implementation for One-Dimensional Multilevel Lifting-Based Wavelet Transform," *IEEE Trans. on Computers*, vol. 53, no. 4, pp. 386-398, April 2004

[13] K. Oweiss, K. Thomson and D. Anderson, "A Systems Approach for Real-Time Data Compression in Advanced Brain-Machine Interfaces," Proc. 2nd IEEE EMBS Conf. on Neural Engineering, March 2005, to appear.

[14] Y. Suhail, K.G. Oweiss, "A Reduced Complexity Integer Lifting Wavelet Based Module for Real-Time Processing in Implantable Neural Interface Devices," *IEEE Int. Conf. on Eng. in Med. and Bio.*, pp.4552-4555, September 2004

[15] K. Thomson, Y. Suhail, and K. Oweiss "A Scalable Architecture for Streaming Neural Information from Implantable Multichannel Neuroprosthetic Devices," *Proc. IEEE Int. Conf. On Circuits & Systems*, May 2005, *to appear*

[16] R. Calderbank, I. Daubechies, W. Sweldens and B.-L. Yeo, "Wavelet transforms that map integers to integers," *Applied and Computational Harmonic Analysis* 5(3), pp. 332-369, 1998

[17] M. Nicolelis, "Actions from Thoughts, " *Nature*, vol. 409, pp.403-407, January 2001

[18] D.M Taylor, S.I. Tillery, and A. B. Schwartz, "Direct cortical control of 3D neuroprosthetic devices" *Science* 296, 1829–1832 (2002).

[19] A. Mason, J. Li, K. Thomson, Y. Suhail and K. Oweiss, "Design Optimization of Integer Lifting DWT Circuitry for Implantable Neuroprosthetics" *Proc. 3rd IEEE EMBS Conf. on Microtechnology in Med. And Bio*, May 2005, to appear.

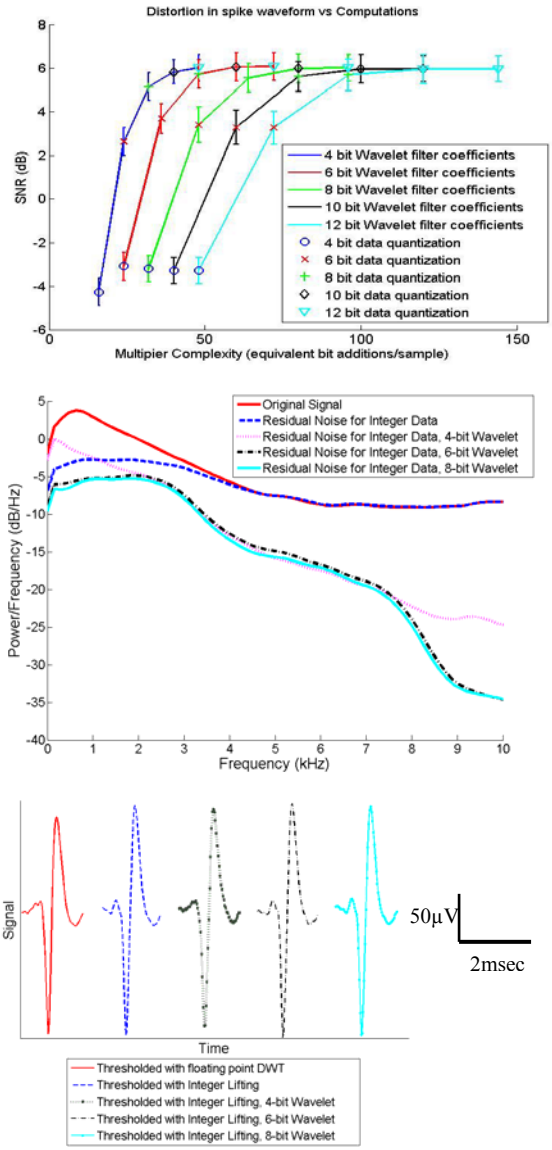

Figure 5: Multiplier Complexity versus distortion of spike waveforms for different integer precision of filter and data coefficients
(a) Effect of round-off and quantization errors on the signal fidelity as a function of multiplier complexity. (b) Power spectral density of the original data and the residual noise for integer approximated coefficient data and quantized wavelet filter coefficients for various bit widths. (c) Example spike waveform obtained in each case.