# Resource Constrained VLSI Architecture for Implantable Neural Data Compression Systems

Awais M. Kamboh, Karim G. Oweiss and Andrew J. Mason
Department of Electrical and Computer Engineering
Michigan State University, East Lansing, Michigan, USA
{kambohaw, koweiss, mason}@msu.edu

***Abstract*— Neural recordings from high-density microelectrode arrays implanted in the cortex require time-frequency domain processing to alleviate the data telemetry bottlenecks of bandwidth and power. Our previous work has shown that the energy compaction capability of the Discrete Wavelet Transform (DWT) offers a practical data compression solution that faithfully preserves the information in the neural signals. This paper presents a complete compression system including both lossy and lossless compression schemes, namely the DWT and Run Length Encoding. Performance tradeoffs and key design decisions for implantable applications are analyzed. A 32-channel, 4-level version of the circuit is presented. Custom designed in 0.5µm CMOS, occupying only 5.75mm$^2$ and consuming 3mW of power (95µW per channel at 25Ks/sec), the implantable compression circuit is well suited for intra-cortical neural interface applications.**

## I. INTRODUCTION

Brain machine interfaces have come to be recognized as some of the most powerful tools in helping patients with neural disorders. Especially, individuals with severe motor limitations can benefit greatly with the advancement in neuroprosthetic devices. Control of artificial limbs is dependent on the accurate decoding of the neural signals, which contain preset movement parameters. In order to enable extraction of these parameters the activity of cortical neurons needs to be recorded [1], using microelectrode arrays of hundreds of elements. Since the algorithms required to extract any useful information from these neural signals are computationally complex and resource hungry, the signals need to be transmitted out of the body to powerful external processing units (EPU).

Transmission of neural recordings to the EPU can be done with wired connections, but this limits patient mobility to a few feet and introduces various surgical complications. One of the challenges facing neural engineers is achieving the wireless transmission of neural recordings, or the information they contain, to overcome wiring limitations. Wireless transmission of potentially hundreds of signals must address three major limitations: bandwidth, implant area, and power consumption. For example, without compression, a 32 channel system with a sampling rate of 25KHz per channel and 10 bits of data precision generates data at 8Mbps. Even state-of-the-art wireless transceivers for biomedical applications are not capable of providing the required data bandwidth, necessitating signal compression before transmission. The hardware required for signal compression within the implant must be, firstly, area efficient, to enable minimally invasive surgical procedures, and secondly, power efficient, to avoid any temperature-induced damage to surrounding tissues. High energy efficiency is also required to enable longer periods of operation with little available power.

To date, most efforts to address neural signal compression have been directed to data in time-domain [2-4]. In contrast, our approach is based on the Discrete Wavelet Transform (DWT) which has not only been shown to be a very effective signal compression and denoising tool, but is also inherently well suited for real-time spike sorting [5-6]. In earlier work [7], we have presented a compression scheme that is tailored to suit neural signals and based on the DWT. In this paper we expand on the theory and design details of the DWT-based system and introduce performance evaluation criteria for circuit blocks that complete the neural signal compression system. We show that this system can be implemented in highly area-power efficient hardware and results in large compression ratios while extending flexibility to the neuroscientists to choose the desired degree of perfection in spike trains' reconstruction from a multitude of neural signals.

## II. COMPRESSION

Fig. 1 provides a system level view of an integrated neural data compression circuit fabricated on the back of a microelectrode array that is implanted in the cortex and communicates wirelessly to an EPU. Neural signals from multiple channels are amplified, digitized and fed to the DWT block, which generates a sparse representation of these

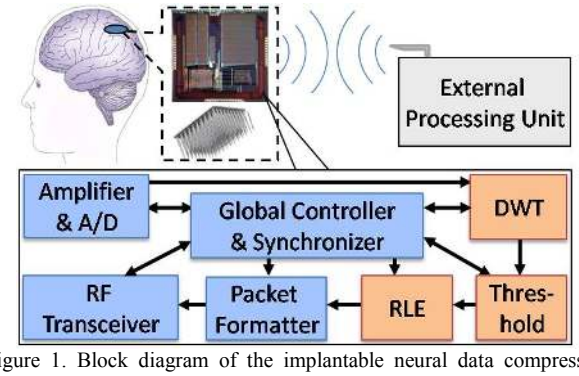


Figure 1. Block diagram of the implantable neural data compression system and its position within an implantable neural recording system.

978-1-4244-3828-0/09/$25.00 ©2009 IEEE

signals. The threshold block serves a dual purpose of de-noising and spike detection. The Run Length Encoder (RLE) removes the redundancy from the streaming data which is then formatted into packets for wireless transmission and sent to the transceiver. The system employs both lossy and lossless compressions to reduce the data, where the DWT, threshold and RLE blocks form the compression engine.

### *A. Discrete Wavelet Transform*

Based on the chip real estate available to this research project, the DWT block shown in Fig. 2 was designed to support up to 32 channels of data simultaneously with 4 levels of decomposition. Multiple decomposition levels generally result in fewer significant coefficients. The relatively long intervals between samples of neural signals allow for computation hardware that prioritizes power and area efficiency over speed [7]. Derived from our prior system-level analysis [5], the power-area product can be minimized by an architecture that sequentially evaluates the DWT of multi-channel data in real time. The lifting scheme is used to compute results; it has been shown to require fewer computations than convolution based filtering [8].

The computation core performs sequential calculation of DWT coefficients. The memory blocks store temporary data and intermediate results and have been partitioned based on different access patterns. The controller manages timing and data flow among the blocks. The controller was synthesized from a library while the other blocks were custom designed to minimize power and chip area.

### *B. Threshold*

Coefficients at the output of the DWT block can be viewed as sparse packets of energy which do not, by themselves, result in any compression. However, following the DWT by a thresholding stage, which reduces the in-significant low-energy coefficients to zero and lets high-energy coefficients pass, does permit compression. The low-energy coefficients have little or no significant information and mainly contribute to noise. The high energy coefficients invariably correspond to different spikes and events in the neural signal, allowing spike sorting even without reconstruction [5]. The values to which the thresholds are set are of critical importance since they determine both the quality of reconstruction and the final rate of compression.

Methods to determine the optimal threshold is an ongoing investigation. However, it has been established that the best compression is achieved by using separate threshold values for each decomposition level of each channel [5]. Thus, results from each level of each channel must be treated separately and will form a stream of data containing information that is virtually independent of the information from other channels and levels within the data stream.

Fig. 3 shows the main functions of the threshold and RLE stages. The threshold block includes a set of memory registers that contain threshold values for each level of each channel. Since the last level of DWT produces two separate coefficients streams, an N level system would have N+1 threshold values for each channel. These values would be determined externally and then stored into the memory sequentially before the DWT operation begins. A serial peripheral interface (SPI) is used to store threshold values in each of the corresponding memory registers. Any of these values can be updated during system operation. The DWT block generates values in signed-magnitude form. The magnitude is compared against the threshold value using a magnitude comparator; if found smaller than the threshold, a 10 bit zero is generated at the output. If equal or greater, the original value is regenerated at the output.

The overall compression system is designed to operate in two modes: Monitor mode and Acquisition mode. In Monitor mode the system bypasses the DWT, threshold and RLE blocks and sends the neural signal directly to the packet formatter. Due to lack of compression, only a few channels can be monitored at a given time. Monitor mode is used by the EPU to analyze the statistical properties of individual channels and calculate the optimal threshold values for each channel and level. These threshold values are then transmitted back to the implanted system to set compression parameters for use in Acquisition mode, where all system blocks are activated to compress data.

### *C. Lossless Compression*

Several lossless data compressors exist in literature, with varied computational complexity and storage requirements. A few popular techniques are Huffman coders, Lempel Ziv coder, arithmetic encoders and their variants. Most of these algorithms require prior statistical knowledge of the incoming data set, and thus the rate of compression achieved is directly related to the accuracy of this information. Since the algorithms are variable length encoders, under certain conditions, they approach the theoretical limits of compression, limited by the entropy of the incoming signal. However, these dictionary-based algorithms require

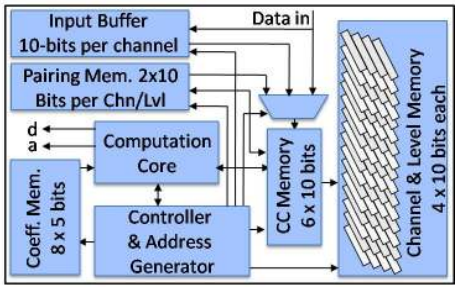


Figure 2. System diagram for sequential calculation of DWT.

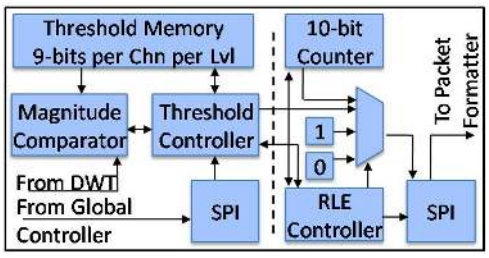


Figure 3. System diagram for the Threshold and the RLE blocks.

prohibitively large storage to maintain the dictionary of codes. Use of Huffman coding is not possible since the size and statistical properties of the source alphabet (number of possible data values) depend on the threshold value, which are ideally controllable by the neuroscientist.

The only statistical information available for our thresholding compression system is that much of the data stream consists of zeroes. Run Length Encoding is best suited for data with long repetitive strings of values; in addition, it is very conservative in required resource. Because we expect long strings of zeros at the output of the threshold block, RLE a good lossless compression choice. Though RLE is not an optimal encoding scheme in general, when given very long repetitive sequences it approaches the performance of near-optimal algorithms.

Given that, for this implementation, a byte refers to 10-bit values ranging from -511 to +511 and that a 10-bit counter can count up to 1023, our implementation of RLE can be summarized by the following rules.

1. Transmit all non-zero bytes as is.
2. Convert all negative zeros (represented by X) to positive zeros.
3. Replace a sequence of zeros (two or more) with an X (negative zero) byte and a zero-count byte.
4. If the zero count reaches 1023, send 1023 and restart a new sequence of zeros.

Following these rules, the example RLE operation of Fig. 4 shows the original 40 byte sequence has been reduced into a 20 byte sequence. Sequences from real neural recordings have been observed to yield much better compression ratios. Since we do not expect long sequences of repeating non-zero values, this implementation compresses only the sequences of zeros. This scheme results in fixed length codes, which have a computational resource advantage.

## III. Analysis and Results

A 0.5µm CMOS process was used to design all the blocks of the neural compression system. The DWT block has been fabricated and the 3mm x 3mm chip is shown in Fig. 5. The controller was synthesized using OSU's Standard Cells Library [9], and all other blocks were custom designed for low power and low area. The active components of the prototype 32-channel, 4-level DWT implementation occupy roughly 3.84mm$^2$. The layout for the threshold and RLE blocks requires about 0.95mm$^2$ of area. Thus the combined compression system is expected to require approximately 5.75mm$^2$ including global routing for a 0.5µm process. Table I lists the area consumed by each module. Empirical measurements show that this area would be reduced by roughly a factor of 15 if implemented in a 0.18µm process. However, this system is sufficiently small for implantation even in 0.5µm CMOS. The tested system consumes only 3mW of power while processing 32 channels at 25Ksamples per second, or equivalently 95µW per channel. The power consumption per channel is directly proportional to the neural data sampling frequency.

Example: Let {0, A, B, C, D} be the source alphabet. Let X represent a negative zero byte. Consider the following input sequence

**BD000A0000000A000000CB0A000000000D00000D**

Once RLE is applied, the sequence reduces to

**BDX3AX7AX6CB0AX9DX5D**

Input sequence length = 40 bytes;
Output sequence length = 20 bytes

Figure 4. Example of the system specific run length encoding.

TABLE I. Area Requirement for Hardware Modules

| Module | Area (mm$^2$) |
|---|---|
| *Complete DWT system: 32 Channel, 4 Level | 3.83 |
| *Threshold + RLE: 32 Channel, 4 Level | 0.95 |
| Expected DWT + Threshold + RLE + Routing | 5.74 |

*Area numbers do not include global routing

To test our designs and algorithms, a stream of experimentally obtained 10-bit neural data was processed through the compression system and the resulting transform coefficients were used to reconstruct the neural signal. The results were compared to the original signal to measure the quality of spike reconstruction versus the compression obtained. This analysis was performed for several different zeroing threshold values. For a fair comparison, the same threshold value was used for all channels and levels. Root Mean Squared (RMS) error and Entropy are used as the primary measures to evaluate the performance of our system. RMS error is a measure of the average difference between the original and the reconstructed signal. This difference has two major components: the error resulting from quantization into a finite word length, and the error introduced by the thresholding operation. Shannon's entropy is a measure of uncertainty associated with a random signal and can be interpreted as the minimum average length per data value, represented in bits, which must be transmitted for lossless communication. Entropy gives the theoretical limit to the achievable lossless data compression for a given data.

For a given spike train, Fig. 6 plots the number of detected spikes, RMS error, the entropy of the transmitted sequence and the RLE compression achieved with respect to

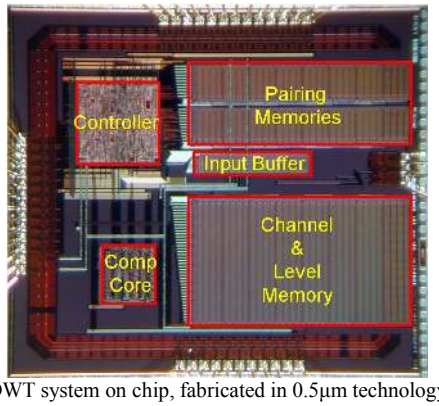


Figure 5. DWT system on chip, fabricated in 0.5µm technology.

the threshold. The plots confirm the anticipated tradeoff between compression and RMS error. As expected, when thresholding is not employed, the RLE does not result in any compression, while entropy is at its maximum and RMS error at its minimum. The RMS error never goes to zero because of quantization noise. As the threshold value (and thus the number of zeros) increases, the RLE compression approaches the theoretical limit of entropy, which proves the effectiveness of our design. A threshold increase also results in an increase in the RMS error. This region of operation removes noise from the signal while preserving all the neural spikes and their shapes. At very high threshold values, the system starts distorting neural spikes, which results in a drop in the number of spikes detected at the output. The point of operation, thus, must be selected just before this region. The optimal point of operation may vary from one application to another depending upon the quality of reconstruction desired. Because of this direct tradeoff between RMS error and compression ratio, the zeroing threshold must be chosen to match application requirements; i.e. available bandwidth, quality of signal reconstruction required, and the power available for data transmission.

Fig. 6 shows the results for a sample spike train which contains 27 spikes. Almost perfect reconstruction is achieved for a zero threshold, but at the cost of high data rate. Beyond the threshold value of 110 the system starts losing spikes for this particular data set. Thus the region of operation should be set between the values of 50 and 100 depending on the application and the desired spike reconstruction quality. The measure of quality of reconstruction depends highly on application specific spike detection and classification algorithms employed by the neuroscientist. The DWT chip has been tested and works as designed with an excellent match between simulated and experimental results. Because the RLE block has not yet been fabricated, data presented in Fig. 6 is based on a combination of measured and simulated results. Fig. 7 shows the same spike at four different de-noising thresholds and compression ratios. For our prototype 32 channel design, a conservative threshold value of 80 resulted in an output data rate of less than 370Kbps, providing a compression of more than 20 times over 8Mbps for unprocessed data. The authors are not aware of any other publications where the spike shapes have been maintained, thus we are unable to compare our results against other methods of neural signal compression.

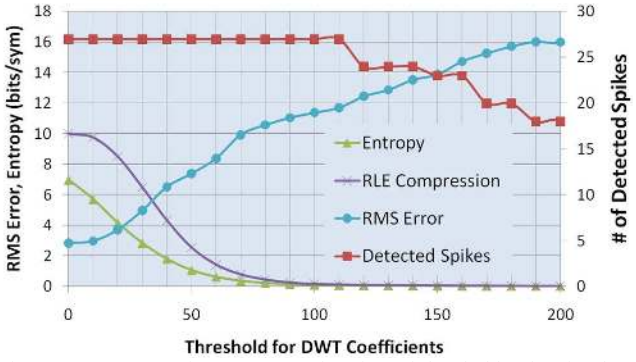

Figure 6. System performance as a function of threshold value for the neural data set used in our experiments.

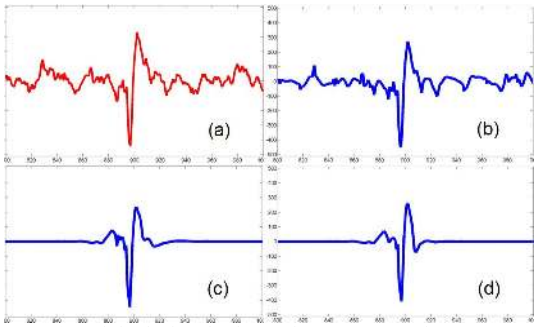

Figure 7. (a) Original signal. (b) 2 times compression. (c) 16 times compression. (d) 62 times compression.

## IV. Conclusion

A system enabling very high data compression of neural recording while maintaining high signal fidelity has been described. The system employs DWT, threshold, and RLE hardware blocks to pseudo-simultaneously processes data from multiple channels. In 0.5µm CMOS, the DWT block require only 3.84mm$^2$ of area to process 32 channels of data at 4 levels of in real time, while consuming only 3mW of power. The overall compression system is designed to fit within 5.75mm$^2$, and the small size and low power consumption of the system makes it highly suitable for implantable high-density microelectrode array devices.

## V. Acknowledgment

This work is supported by the National Institute of Health under grant number NS062031.

## References

[1] M. Nicolelis, "Actions from thoughts," *Nature*, vol. 409, pp. 403-407, Jan. 2001.

[2] A. M. Sodagar, K. D. Wise, and K. Najafi, "A fully integrated mixed-signal neural processor for implantable multichannel cortical recording," *IEEE Trans. on Biomed. Eng.*, vol. 54, no. 6, pp. 1075-1088, Jun 2007.

[3] R. R. Harrison, P. T. Watkins, R. J. Kier, R. O. Lovejoy, D. J. Black, B. Greger, and F. Solzbacher, "A low power integrated circuit for a wireless 100 electrode neural recording system," *IEEE Journal of Solid State Circuits*, vol. 42, no. 1, pp. 123-133, Jan 2007.

[4] J. G. Harris, J. C. Principe, J. C. Sanchez, D. Chen and C. She, "Pulse based signal compression for implanted neural recording systems," *IEEE Int. Symp. on Circuits and Systems*, pp. 344-347, May 2008.

[5] K. Oweiss, "A systems approach for data compression and latency reduction in cortically controlled brain machine interfaces," *IEEE Trans. on Biomed. Eng.*, vol. 53, no. 7, pp. 1364-1377, Jul. 2006.

[6] K. Oweiss, A. Mason, Y. Suhail, K. Thomson, and A. Kamboh "A scalable wavelet transform VLSI architecture for real-time neural signal processing in multichannel cortical implants," *IEEE Tran. on Circuits and Systems I*, vol. 54, no. 6, pp. 1266-1278, Jun. 2007.

[7] A. M. Kamboh, M. Raetz, K. G. Oweiss, A. Mason, "Area-power efficient VLSI implementation of multichannel DWT for data compression in implantable neuroprosthetics," *IEEE Trans. on Biomed. Circuits. and Systems*, vol. 1, no. 2, pp. 128-135, Jun. 2007.

[8] A. M. Kamboh, A. Mason, and K. G. Oweiss, "Analysis of Lifting and B-Spline DWT Implementations for Implantable Neuroprosthetics," *Journal of Signal Processing Systems,* vol. 52, no. 3, pp. 249-261, Sep. 2008.

[9] http://vcag.ecen.okstate.edu/projects/scells/, Standard cells library, Oklahoma State University.