Energy-Efficient Configurable Discrete Wavelet Transform for Neural Sensing Applications

Tang-Hsuan Wang[1], Po-Tsang Huang[1], Kuan-Neng Chen[1], Jin-Chern Chiou[2], Kuo-Hua Chen[3], Chi-Tsung Chiu[3], Ho-Ming Tong[3], Ching-Te Chuang[1] and Wei Hwang[1]
Department of Electronics Eng.[1], Department of Electrical Eng.[2], National Chiao Tung University, Hsinchu, Taiwan
Advanced Semiconductor Engineering (ASE) Group[3], kaohsiung, Taiwan

Abstract—**Highly integrated neural sensing microsystems are crucial to capture accurate signals for brain function investigations. In this paper, an energy-efficient configurable lifting-based discrete wavelet transform (DWT) is proposed for a high-density neural sensing microsystems to extract the features of neural signals by filtering the signals into different frequency bands. Based on the lifting-based DWT algorithm, the area and power consumption can be reduced by decreasing the computation circuits. Additionally, both the time window and mother wavelets can be adjusted via the configurable datapth. Moreover, the power-gating and clock-gating techniques are utilized to further reduce the energy consumption for the energy-limited bio-systems. The proposed configurable DWT is designed and implemented using TSMC 65nm CMOS low power process with total area of 0.11 mm^2 and power consumption of 26 µW. Moreover, this proposed DWT is also implemented in Lattice MachXO2-1200 FPGA and integrated in a 2.5D heterogeneously integrated high-density neural-sensing microsystem with the power consumption of 211.2 µW.**

I. Introduction

The brain computer interface (BCI) helps patients to communicate and control between human and computers [1]. The brain activity neuronal network behavior provides important physiological and psychological information. Hence, electrophysiological signals, Electroencephalography (EEG), and Electrocorticography (ECoG), are widely used to reveal dynamic activity and functional connectivity of the brain due to their excellent temporal resolution on the order of milliseconds. Moreover, heterogeneously integrated and miniaturized neural sensing microsystems for accurately capturing and classifying signals are essential for brain function investigation and neural prostheses realization [2]. In recent years, many neural sensing microsystems have been proposed to provide small form factor and biocompatible properties for neural sensing applications [3, 4]. These heterogeneous biomedical devices compose of sensors and CMOS circuits for biopotential acquisition, signal processing and transmission.

For miniaturized neural sensing microsystems, the required energy is prohibitively large to transmit raw data. Therefore, discrete wavelet transform (DWT) is a very effective method for extracting neural features and compressing neural data. Additionally, DWT can provide higher time resolutions for higher frequency bands, and the time window can be adjustable easily with both stationary and non-stationary signals. In view of these, a 2.5D heterogeneously integrated bio-sensing microsystem with µ-probes is presented for high-density multi-channel neural sensing applications as shown in Fig. 1. In this microsystem, configurable multi-channel lifting-based DWTs are designed to extract features of neural signals and to cluster

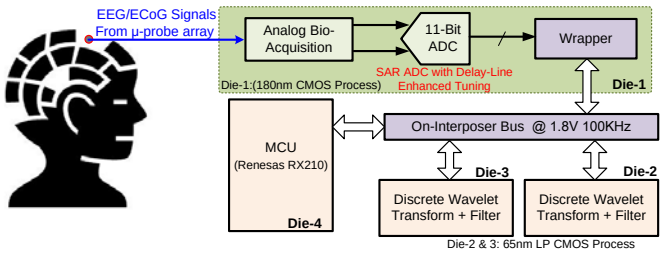


Fig. 1. 2.5D heterogeneously integrated bio-sensing microsystem.

the signals for medicine diagnosis by filtering the neural signal into different frequency bands. Both the time window and mother wavelets can be adjusted in this configurable DWT. Moreover, the power-gating and clock-gating techniques are utilized for further realizing the energy saving by decreasing both the active power and leakage power.

II. Design of Liftinh-Based Configurable DWT

The configurable DWT is realized for different applications with four types of mother wavelet, Haar, Daubechies 2 (D2), Symmlet 4 (Sym4) and Symmlet 6 (Sym6). The differences between these four types are the length of time window and the filter coefficients. Additionally, the DWT is designed based on the lifting-based algorithm to reduce the arithmetic circuits [5]. For instance, the Sym6 implemented by the conventional DWT requires 22 multiplications and additions, respectively. In the lifting-based DWT, only 12 multiplications and 12 additions are needed during an operated iteration.

Based on the lifting algorithm, the DWT is transferred into the lifting steps of "predict" and "update" functions, $s_n(z)$ and $t_n(z)$, respectively. First, the input data is split into the even and odd samples, then applying the $s_n(z)$ and $t_n(z)$ filters sequentially. The last step is a multiplication by a scaling factors K and 1/K. For simplifying the datapath of the configurable lifting-based DWT, the lifting steps are designed in a backbone as shown in Fig. 2. The backbone of the lifting steps is based on the Sym6 wavelet with the maximum length of time window, and the decompositions of other mother wavelets are derived according to this backbone. Therefore, the configurable datapath can realize other mother wavelet by

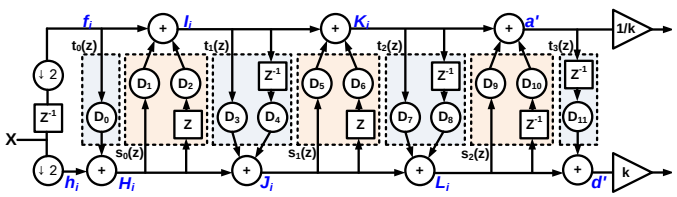


Fig. 2. Backbone of the configurable lifting-based DWT.

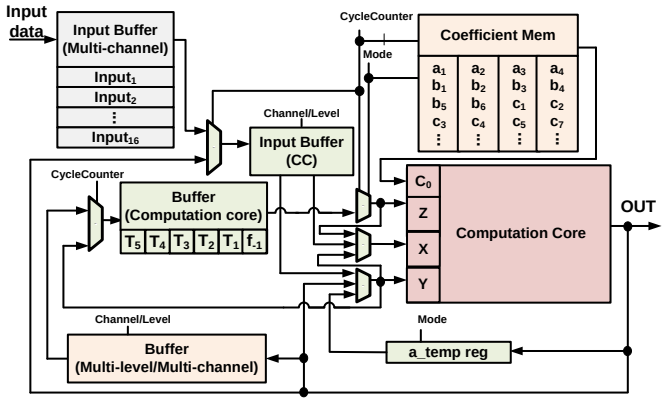
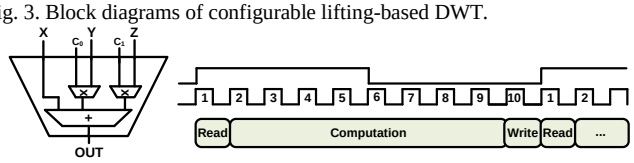


Fig. 3. Block diagrams of configurable lifting-based DWT.



(a) (b)

Figure 4. (a) Computation Core (b) Execution in 1-level iteration.

adjusting the filter coefficients only to reduce the complexity of the controller.

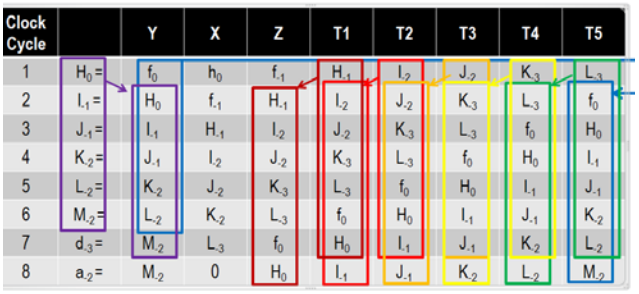
The lifting filters of Sym6 can be factorized and operated in the sequence as follows, where the h and f are the odd and even samples, respectively.

$$Step1: H_0 = h_0 + D_0 \times f_0$$
$$Step2: I_{-1} = f_{-1} + D_1 \times H_{-1} + D_2 \times H_{-1}$$
$$Step3: J_{-1} = H_{-1} + D_3 \times I_{-1} + D_4 \times I_{-2}$$
$$Step4: K_{-2} = I_{-2} + D_5 \times J_{-2} + D_6 \times J_{-1}$$
$$Step5: L_{-2} = J_{-2} + D_7 \times K_{-2} + D_8 \times K_{-1}$$
$$Step6: M_{-2} = K_{-2} + D_9 \times L_{-2} + D_{10} \times L_{-1}$$
$$Step7: d_{-3} = L_{-2} + D_{11} \times M_{-2}$$
$$Step8: a_{-2} = (1/k) \times M_{-2}$$

The outputs of H, I, J, K and L are the temporary (intermediate) value. D_i represents the filter coefficients. Finally, the a' and d' are the approximation and detail coefficients without scaling by 1/k and k, and 1/k and k are scaling factor.

III. DATAPATH OF LIFTING-BASED CONFIGURABLE DWT

The lifting factorization algorithm is utilized for the configurable DWT, and the maximum step of operations is 8 steps for Sym6 mother wavelet, which is the basic model for the other three mother wavelets. Additionally, this sequential structure is suitable in multi-channel and multi-level [6]. Therefore, Fig. 3 presents the multi-level multi-channel lifting-based DWT architecture. Based on the correlated arithmetic functions in these eight steps, a computation core (CC) is realized for each step. Two separated input buffers are realized for efficient data accesses. The multi-channel input buffer stores the input data for different neural sensing channels, and the input buffer (CC) is utilized for the sequential samples (input data h_0 and f_0) or the output data from the last level. Additionally, the temporary data for 1-level iteration are stored in the buffer (computation core) which is composed of six shift registers. The buffer (multi-level/multi-channel) stories the intermediate data from different channels and levels. The



Clock Cycle		Y	X	Z	T1	T2	T3	T4	T5
1	H_0=	f_0	h_0	f_{-1}	H_{-1}	I_{-2}	J_{-2}	K_{-3}	L_{-3}
2	I_{-1}=	H_0	f_{-1}	H_{-1}	I_{-2}	J_{-2}	K_{-3}	L_{-3}	f_0
3	J_{-1}=	I_{-1}	H_{-1}	I_{-2}	J_{-2}	K_{-3}	L_{-3}	f_0	H_0
4	K_{-2}=	J_{-1}	I_{-2}	J_{-2}	K_{-3}	L_{-3}	f_0	H_0	I_{-1}
5	L_{-2}=	K_{-2}	J_{-2}	K_{-3}	L_{-3}	f_0	H_0	I_{-1}	J_{-1}
6	M_{-2}=	L_{-2}	K_{-2}	L_{-3}	f_0	H_0	I_{-1}	J_{-1}	K_{-2}
7	d_{-3}=	M_{-2}	L_{-3}	f_0	H_0	I_{-1}	J_{-1}	K_{-2}	L_{-2}
8	a_{-2}=	M_{-2}	0	H_0	I_{-1}	J_{-1}	K_{-2}	L_{-2}	M_{-2}

Fig. 5. Data dependency and movement of the 8 steps in 1-level iteration for the computation cure (CC).

coefficients of different wavelets are placed in the coefficient memory. Furthermore, the a_temp register is designed for the last step in 1-level iteration which scales the data to the approximation by the scaling factor.

A. Computation Core (CC)

1-level iteration is composed of 8 steps, and each step can be realized by the computation core. Obviously, the functionality of each step can be implemented via the simple form as $OUT = X + D_i \times Y + D_j \times Z$. Based on the regularity of these 8 steps, the complexity of the CC block can be reduced using two multipliers and one adder in the lifting-based DWT. Additionally, the three 10-bit inputs of X, Y and Z are defined in 2's complemented form, and D_i and D_j are the 6-bit quantized filter coefficients. Fig.4(a) shows the architecture of the CC. The outputs from the two multipliers must to be down-scale by utilizing a hardwired shifting operation for removing the x16 scaling filter coefficients. Then, transform X into 12-bit data for the 3-term adder. Finally, truncating the output from this adder is the last action during the entire CC operation.

B. Three States for 1-Level Iteration

For the multi-channel & multi-level DWT operations, 1-level iteration is the baseline via 10 execution cycles. Fig. 4(b) present 1-level DWT iteration composed of three states with the total ten execution cycles fixed by Sym6. Preparing the input data and temporary data from previous sample is in the read state by one cycle, and the following eight computation cycles are in the computation state for the 8 steps. Finally, the last one cycle is for the write state.

C. Data Dependency of 1-Level Iteration

The lifting-based DWT can be realized by the single computation core (CC) to calculate the equations in 1-level iteration. However, the data access is a crucial issue for the design of the datapath. Therefore, five shift registers are applied to store the intermediate values for the next sample from the same channel. The data dependency and movement status of 1-level iteration in Sym6 is as shown in Fig. 5. For executing DWT at different steps, a 4-bit counter is employed with eight cycles for calculation, one for reading and one for writing. These registers X, Y and Z are the input of computation core, and the temporary registers $T1, T2, T3, T4$ and $T5$ are designed for storing temporary data and then providing these data for different equations. At the first cycle, the values of h_0 and f_0 are read from the input buffer (CC), and f_{-1}, H_0, I_{-1}, J_{-1}, K_{-2} and L_{-2} are read from buffer (CC). After the eight

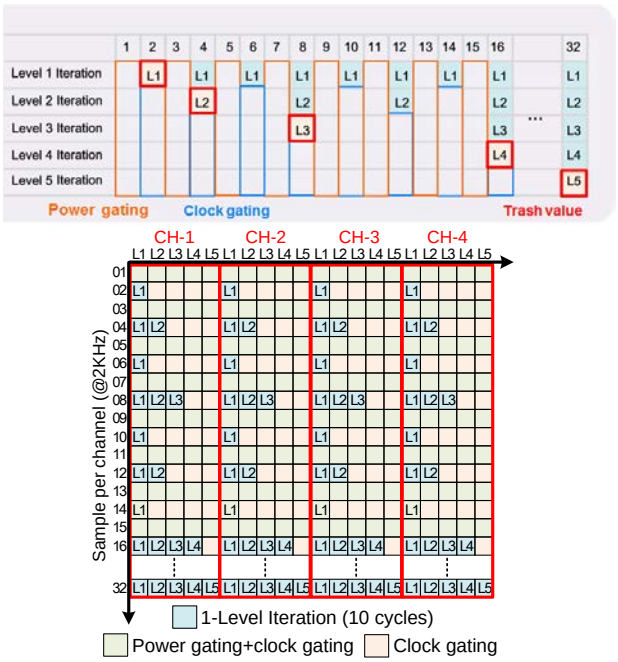


Fig. 6. Timing diagram of clock gating and power gating for 5-level single channel DWT and 5-level 4-channel DWT.

cycles of computations and data shifts, the detail coefficients d_{-3} and the approximation coefficient a_{-2} are the output. Accordingly, the approximation coefficient has to be stored for the latter processes in the next level also.

D. Buffer Implementation for Multi-Lev el and Multi-Channel

In the datapath of the multi-level and multi-channel DWT, numerous storages are required as the coefficient memory for the quantized filter coefficients, input buffers (multi-channel and CC) and other buffers (CC and multi-channel/multi-level). For 1-level DWT iteration, the DWT execution requires the input signals from the system input or the pre-level output a_i, where i is the sample time. For different levels and channels, two registers are essential for h_0 and f_0 except level-1. These registers are designed in 10-bit integer, and thus the input buffer (CC) is composed of $2(\text{register}) \times 10 \times (L-1) \times CH$ bits, where L is the number of level and CH represents the number of channels. For each level and channel, the buffer (multi-level/multi-channel) is implemented to store the intermediate values, such as f_{-1}, H_0, I_{-1}, J_{-1}, K_{-2}, L_{-2} based on Sym6 wavelet. The total size of the buffer (multi-level/multi-channel) is $6\,(\text{register}) \times 10 \times L \times CH$ bits. For different levels and channels, the corresponding data of the selected level & channel can be accessed from the buffer using two-level wire multiplexers.

IV. PowerGating & Clock Gating for DWT

Low energy dissipation is a critical challenge for miniaturized neural sensing microsystems. Therefore, the interleaving architecture for multi-level DWT has been proposed to reduce the operation frequency for active power saving [7]. However, as nano-scale technologies become more advanced, leakage currents increasingly dominate their overall power consumption, especially operated at low frequencies. A gating architecture for multi-channel and multi-level DWT is designed to reduce the active power and leakage power via clock gating and power gating, respectively. Fig. 6 presents the timing diagram of the clock gating and power gating for multi-level single channel DWT. The grey iterations are inactive and can be gated to reduce the power consumption. Therefore, both power gating and clock gating are utilized in the odd sampling periods. During the even sampling periods of the input data, 1~5 different iterations are executed within 5x10 execution cycles, and the process of different levels is executed sequentially from level-1 to level-5. The inactive time is variable and shorter than that in the odd sampling period. Hence, only the clock gating is applied within the inactive iterations during the even sampling periods.

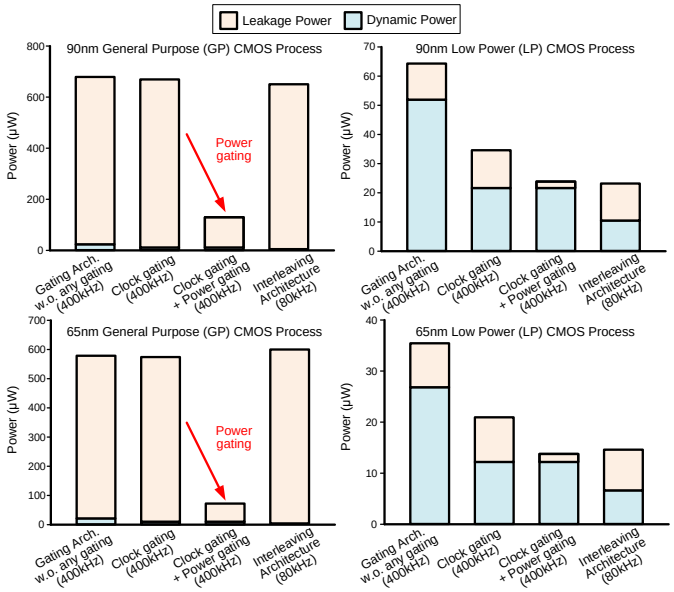


Fig. 7. Power analysis of 4-folded 16-channel DWT.

The frequency of the interleaving architecture is only related to the number of levels. However, the frequency of the gating architecture is related to both the numbers of channels and levels. For a 5-level 4-channel DWT with the sampling frequency of 2KHz, the clock frequencies of the gating and interleaving architectures are 400KHz (2k x 4 x 5 x 10) and 80KHz (2k x 4 x 10), respectively.

V. Power/Area Analyses for Multi-Channel DWT

For multi-channel DWT, the data between different channels are independent and the computation core can be shared. To obtain an optimal trade-off among energy and area, the time-multiplexing scheme is utilized. Therefore, 5-level 16-channel lifting-based DWTs are implemented on different folded numbers of time-multiplexing scheme, interleaving or gating architectures and different CMOS processes, including TSMC 90nm General Purpose (GP) CMOS process, 90nm Lowe Power (LP) CMOS process, 65nm GP CMOS process and 65nm LP CMOS process. The voltages of GP process and LP process are 1.0V and 1.2V, respectively.

Fig. 7 shows the power analysis of the four-folded time-multiplexing architecture in different CMOS processes. In both 90nm and 65nm GP processes, the clock gating technique can achieve the slight power saving. Nevertheless, the power gating

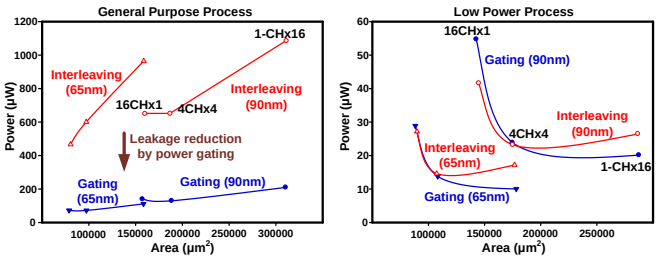


Figure 8. power & area analysis using different processes.

technique can realize power reduction drastically. Although the frequency of the gating architecture is 5 times higher than that of the interleaving architecture, the power consumption of gating architecture is much smaller than that of the interleaving architecture because the leakage power dominates the over power using GP processes. However, the leakage current of LP process is 10X~20X smaller than that of GP processes by increasing the thickness of the gate oxide. Therefore, the dynamic power reduction techniques are required using LP processes. Both clock gating and interleaving architecture can decrease the active power consumption.

Fig. 8 presents the power and area analysis of different architectures on different processes. In GP processes, both the power consumption and area are decreased with the increasing folded number using the time-multiplexing scheme. Additionally, the power of the gating architecture is much smaller than those of the interleaving one with the similar area because the leakage current is reduced significantly. In LP processes, the ratio of the dynamic power is increased. Therefore, the power consumption of 16-folded DWT increases rapidly due to the increasing operation frequency. Consequently, the power of the gating architecture is smaller than that of the interleaving one since the leakage current is reduced by power gating. Through the shrink of the gate length, both the power and area can be reduced.

VI. 16-Channel DWT in 2.5D Heterogeneously Intrgrated Bio-Sensing Microststem

A bio-sensing microsystem is implemented in 2.5D heterogeneous integration as shown in Fig. 9. In this microsystem, 4-folded 16-channel DWT are implemented via two Lattice MachXO2-1200 FPGA dies, which are fabricated using 65nm LP CMOS process. The power gating is realized by the power saving mode of these FPGAs. Fig. 10 presents the power consumptions of different DWT architectures on these 2 FPGA dies at 1.2V. The leakage power is also a critical issue in FPGA. The dynamic power of the gating architecture is slightly larger than that of the interleaving one because of the 5 times frequency (800KHz). Moreover, the power gating can reduce the leakage power substantially, which is from 223.2μW to 91.1μW. Compared to the interleaving architecture, the gating architecture is the best solution for FPGA implementations.

VII. Conclusion

Heterogeneously integrated and miniaturized neural sensing microsystems for accurately capturing and classifying signals are essential for brain function investigation. Therefore, the energy-efficient configurable lifting-based DWT design is presented to extract the features of neural signals for different

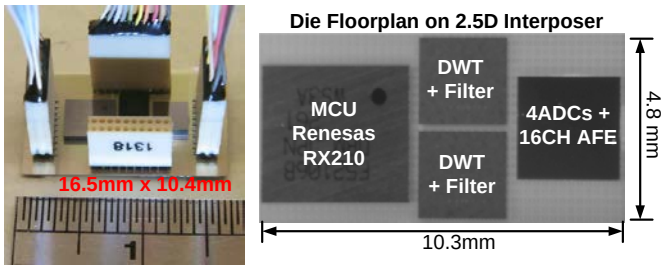


Figure 9. 2.5D heterogeneously integrated bio-sensing microsystem.

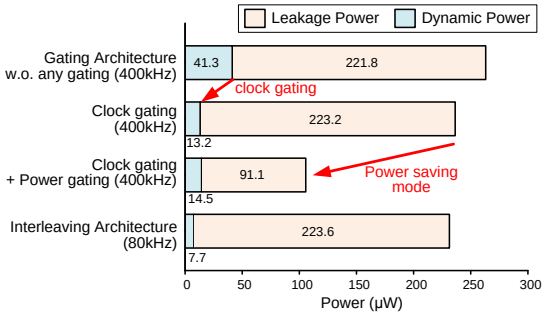


Figure 10. Power consumptions of different DWT architectures on FPGA.

neural sensing applications. The configurable DWT can select four mother wavelets using different filter coefficients. Additionally, both the time window and mother wavelets can be adjusted via the configurable datapth. Moreover, the power-gating and clock-gating techniques are adopted to reduce both the active and leakage power for the energy-limited bio-systems. This 16-channel configurable DWT is implemented using TSMC 65nm CMOS low power process with the total area of 0.11 mm^2 and power consumption of 26 μW. Moreover, this proposed DWT is also implemented in Lattice MachXO2-1200 FPGA, which is fabricated using 65nm LP CMOS process, and integrated in a 2.5D heterogeneously integrated neural-sensing microsystem with the power consumption of 211.2 μW.

References

[1] B.J. Lance, et al., "Brain–Computer Interface Technologies in the Coming Decades," *Proceedings of the IEEE*, vol.100, no. Special Centennial Issue, pp.1585-1599, May 2012.

[2] N. Verma, et al., "A Micro-Power EEG Acquisition SoC With Integrated Feature Extraction Processor for a Chronic Seizure Detection System," *IEEE Journal of Solid-State Circuits*, Vol. 45, No.4, pp.804-816, April 2010.

[3] B. Gosselin, et al., "A Mixed-Signal Multichip Neural Recording Interface With Bandwidth Reduction," IEEE Trans. on Biomedical Circuits and Systems, vol. 3, no. 3, pp. 129-141, June 2009.

[4] B.K. Thurgood, et al., "A Wireless Integrated Circuit for 100-Channel Charge-Balanced Neural Stimulation," IEEE Trans. on Biomedical Circuits and Systems, vol. 3, no. 6, pp. 405-414, Dec. 2009.

[5] I. Daubechies and W. Sweldens, "Factoring wavelet transforms into lifting steps," *J. Fourier Anal. Appl.*, vol. 4, no. 3, pp. 245–267, 1998.

[6] K.G. Oweiss, et al., "A Scalable Wavelet Transform VLSI Architecture for Real-Time Signal Processing in High-Density Intra-Cortical Implants," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol.54, no.6, pp.1266-1278, June 2007.

[7] A.M. Kamboh, et al., "Area-Power Efficient VLSI Implementation of Multichannel DWT for Data Compression in Implantable Neuroprosthetics," *IEEE Transactions on Biomedical Circuits and Systems*, vol.1, no.2, pp.128-135, June 2007.