Final Report (July 2011)

Title: Design of Low-Power Wireless Electroencephalography (EEG) System

Sponsor: Southeastern Center for Electrical Engineering Education (SCEEE)

PI: Dr. Saibal Mukhopadhyay, School of ECE, Georgia Institute of Technology

Students worked in the project: Denny Lie and Jeremy Tolbert (NSF Fellow).

1. Introduction

Wireless Electroencephalography (EEG) systems are emerging as an efficient replacement for long-term monitoring of epileptic activity, because it allows the comfort of no wires and eliminates the need for bulky devices to store data. In these systems, signals are acquired via a headset, processed and wirelessly transmitted to a remote host (Fig. 1). Wireless systems can provide the efficiency of performing real-time signal detection (of epileptic activity) and compression of signals to extend the battery life [1]. The challenge in designing a wireless EEG system is in reducing the energy consumption involved in processing and accurately transmitting these signals.

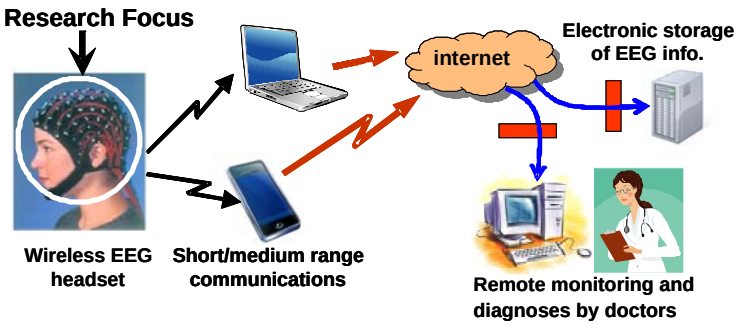

Fig. 1: A pictorial view of wireless EEG scheme

2. Summary of Performed Work

This project has developed an algorithmic and architectural design of an embedded hardware system for energy-efficient wireless EEG monitoring that accurately transmits epileptic signals by employing adaptive compression methods. Our specific goal was to address the issues of system energy. Our project involved following key steps:

1. ***Development of Algorithms***: Our first step was to develop algorithms for signal detection, compression and transmission. The majority of the system power in a wireless system is due to the power required to operate the transceiver [2]. By compressing (lossy) the data when no epileptic potentials exist, we can use the transceiver for a shorter amount of time and save power [3]. The algorithm proposed for detection will characterize the signal in real-time and determine whether compression is necessary.
2. ***Architectural Design and FPGA based Hardware Prototype***: We have developed embedded hardware architecture for wireless EEG monitoring scheme. The architecture was based on a three stage pipeline system. The embedded digital hardware receives the EEG signal from an Analog-to Digital Converter, performs the accuracy and energy-aware compression, and controls the data flow to the RF transceiver. The full system is designed using a Xilinx Virtex 5 FPGA, and the resulting implications show that

methodology is a viable option to design an ultra-low power ASIC for wireless EEG processing. The system measurements were performed that consider accuracy and energy, supporting the use of adaptive compression.

3. ***Implications for ASIC Design****:* We have performed theoretical analysis to understand the implications of implementing this design as an ASIC system. We have also tape-out a test-chip to validate the Wavelet Filter and its energy-efficiency. The test-chips were recently back. We are currently planning the necessary steps required to perform the measurements.

Related Publications

1. Jeremy R. Tolbert, Denny Lie, Akshath Rampur, Andrew Burks and Saibal Mukhopadhyay, "A Content-Aware Adaptive Compression System for Energy-Efficient Wireless EEG," submitted to IEEE BioCAS 2011.

2. Jeremy R. Tolbert, Pratik Kabali, Simeranjit Brar, and Saibal Mukhopadhyay, "Modeling and Designing for Accuracy and Energy Efficiency in Wireless Electroencephalography Systems," accepted for publication in *ACM Journal on Emerging Technology in Computing (JETC).*

3. Detail Description of Different Components

3.1. EEG Specific Algorithm

The algorithm implemented on chip to detect, compress and transmit EEG data is based on an adaptive method. The overall goal is to transmit the data with maximum energy-efficiency while preserving the accuracy of relevant signals, (epileptic spikes). Figure 2 shows a human EEG signal taken from a patient who suffers from epileptic seizures [9]. EEG signals are low frequency signals (0-100Hz) and their amplitudes range from 10 to 100 micro volts when sampled from the scalp. We can classify the data in (at least) two distinct regions: 1.) Background activity and 2.) Epileptic (spike) activity. All humans' exhibit background EEGs, but it is the epileptic activity that is most important for neurological diagnosis

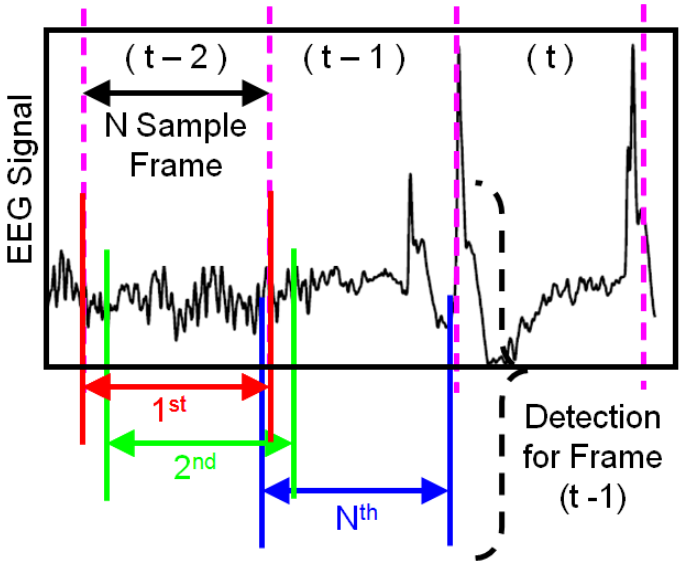

Fig. 2: An EEG Signal can be defined by at least two distinct regions: background and epileptic (spike) activity. In our algorithm for spike detection, a frame size of N-samples will involve N overlapping detections.

and classifications. In epileptic patients, background EEGs occur more often than not, making it difficult to understand and isolate epileptic behavior.

To detect and compress EEG signals, we have used Mallats' theory of multi-level resolution to decompose the EEG into separate frequency bands [4]. The wavelet transform was selected to perform this function, because it is appropriate for non-stationary signals like EEG. Once the signal has been divided into separate bands, the magnitude of the wavelet coefficients can provide insight into epileptic activity (detection) and background activity (coefficients eligible for compression). For spike detection, a set of wavelet coefficients were compared across a pre-determined *spike* threshold. If any coefficients surpass this threshold, an epileptic action potential is said to be detected. Similar methods of detection have been investigated and shown to produce greater than 90.5% accuracy, 91.7 % sensitivity and 89.3 % specificity [5]. To compress the data, the coefficients are compared with a different pre-determined *compression* threshold. The coefficients that do not surpass this threshold are discarded. This method can provide appreciable compression with reasonable signal reconstruction [6].

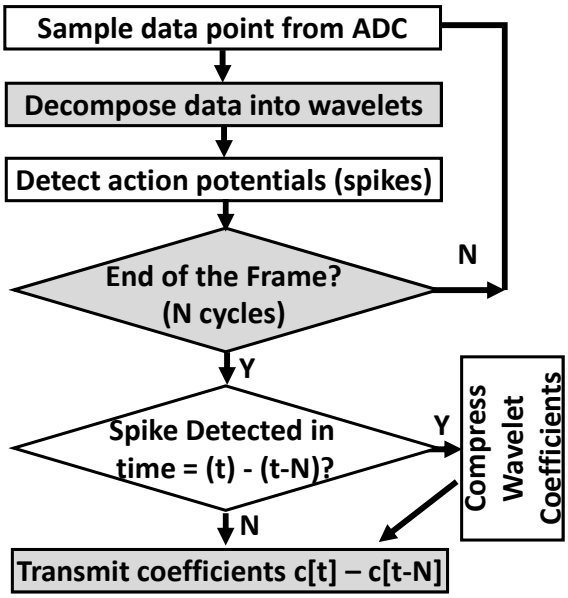

Fig. 3: Adaptive algorithm for spike detection and data compression of EEG signals.

For a time reference, the signal is divide into frames, which are define a N cycles (see Figure 2). If a spike is not detected within the frame, the compressed wavelet coefficients will be transmitted. If a spike is detected, all the wavelet coefficients will be transmitted. The adaptive compression algorithm for EEG detection, compression and transmission is depicted by the decision chart in Figure 3.

3.2. Design of the Embedded Hardware Architecture

We have developed an embedded hardware architecture for processing of EEG signals requires in three primary steps: Signal Detection, Data Compression, and Transmission. Figure 4 shows the three stage pipelined architecture with one stage each for detection and compression, and the final stage for transmission and control.

In the first pipeline stage; detection occurs as the signal sampled from the analog-to-digital converter (ADC) is decomposed into wavelet coefficients. Detection is performed on the 4^{th} and 5^{th} level coefficients as noted in [5]. This result is stored in the History Buffer, which is later used to determine if data compression is necessary. Compression thresholding occurs in the second stage of the process, by way of a generated signature. Essentially, every coefficient is compared to a compression threshold (CT), and if it is greater than CT, a 1 is stored. The resulting signature is a binary pattern that corresponds to the sampled data points in the frame.

During the second pipeline stage, compression occurs via the collapse buffer. The collapse buffer will scan the History Buffer to determine if a spike has been detected in the frame (or neighboring frames). If a spike has been detected, no compression will occur. If no spike has been detected, the wavelet coefficients that are below the CT are removed, and the coefficients are "collapsed" to ensure there is a continuous stream of relevant data. This lossy compression results in a buffer of wavelet coefficients, with zeros in the higher order spaces. With the collapsed coefficients and signature, a receiver system can reconstruct the EEG signal.

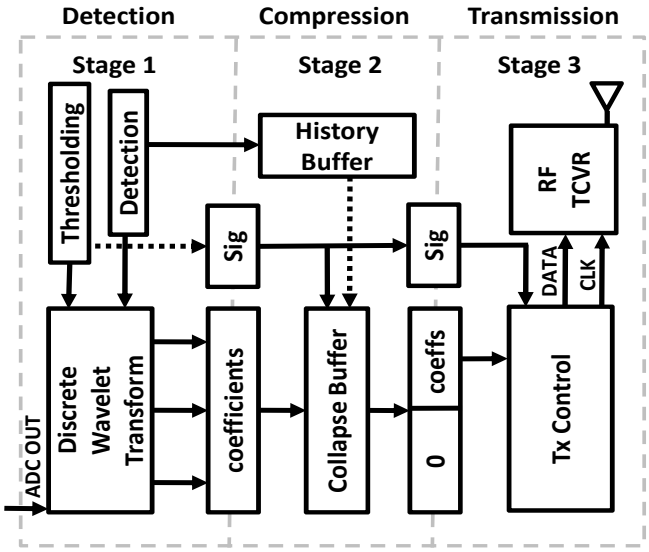

Fig. 4: Architectural of the three-stage pipeline (Detection, Compression, Transmission) for adaptive EEG processing.

In the last stage, the operations to facilitate transmission occur. After the coefficients have been compressed, data payloads have to be generated based on the packet size for a given transceiver. The transmission control logic also sends the data into the transceiver, and enables transmission when ready.

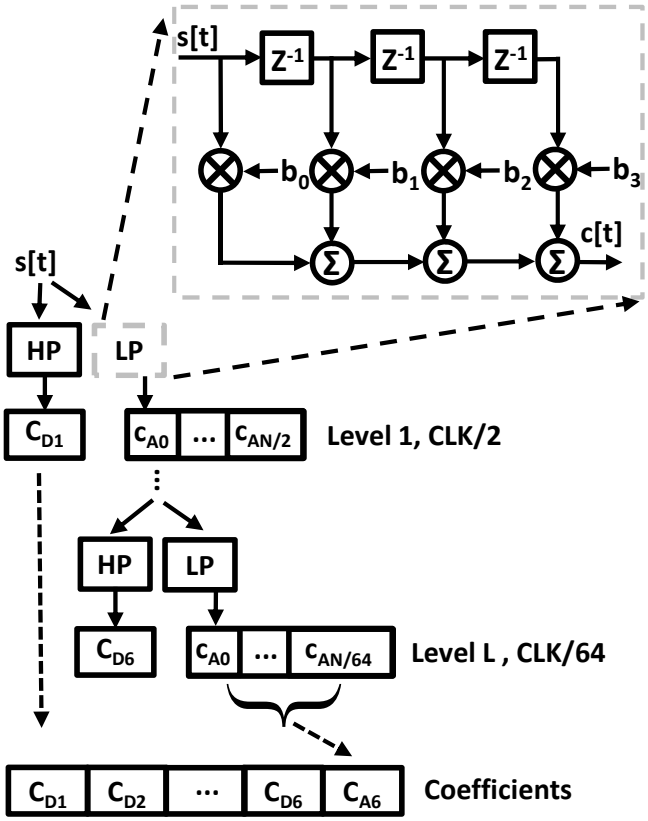

Fig. 5: Adaptive algorithm for spike detection and data compression of EEG signals.

The clocking scheme of the three stage pipeline is also a noteworthy discussion. Recall that in our earlier discussion of EEG signals, the highest frequency component is in the order of 100's of Hz. There is no need to sample and operate at a frequency much higher than Shannon sampling frequency. In the final stage, data must be clocked into the transceiver at a rate determined by the SPI protocol, in this case 2MHz. Since this streaming processor has a unidirectional flow of data, a simple asynchronous protocol can allow for the transmission controller to idle in a low power state (SLEEP) until the collapse buffer enables it to awake and begin the transmission protocol.

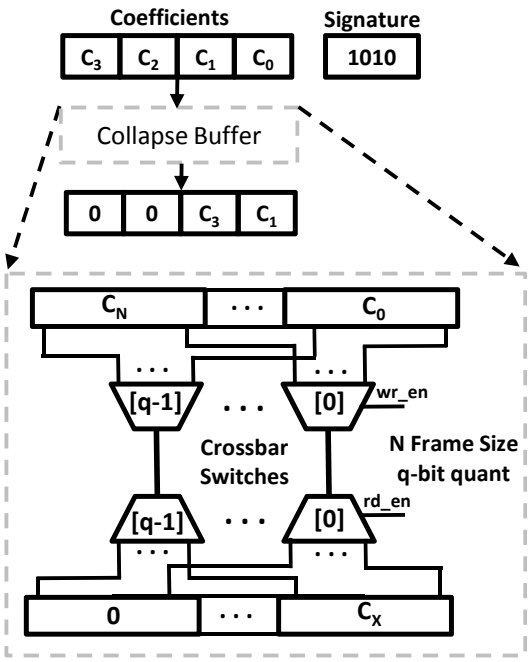

Fig. 6: The collapse buffer zeros the coefficients below the compression threshold and collapses the relevant data to the least significant bits.

The discrete wavelet transform (DWT) is implemented as a filter bank of low pass and high pass filters. Figure 5 shows the implementation of the DWT, where each filter is implemented as a 4-tap FIR Filter. An N sample signal, s[t], will produce N approximate (C_A) and N detailed (C_D) coefficients. As a result of dividing the signal into frequency bands, and making use of Shannon's Sampling Theorem, each resulting output can be represented as N/2 samples. To facilitate the effect of down sampling, the coefficients are latched at half the frequency of the sampled signal s[t]. This in effect reduces the number of latches needed to represent the coefficients. At each DWT level, the low pass coefficients are further decomposed until the desired L^{TH} level is reached. From [5], the optimal number of levels for EEG decomposition to facilitate spike detection is 6. In the end, an N sample signal will produce N coefficients, resulting from all high pass coefficients and the last level low pass coefficients.

The collapse buffer was implemented as a crossbar switch network to facilitate multiple input, multiple output connections. The overall operation can be explained with a simple example shown in Figure 6. In this case, the input to the collapse buffer is four Q-bit wavelet coefficients and a corresponding four bit signature. The signature denotes that coefficients C_1 and C_3 are relevant for accurate representation of the reconstructed signal. At the end, the negligible coefficients that correspond to '0' in the signature are discarded, and their space is taken up by the relevant coefficients.

In the transmission control block, the coefficients are padded with an address, packet id and header to create a payload. For this case study, we have design a transmission protocol to operate with a commercial low power transceiver, the Nordic nRF24L01+ [7]. This transceiver is designed to operate at low currents (~mA) during transmission and ultra-low currents (~uA) during sleep mode. However, the proposed method can be adapted to any transceiver. The transmission control was implemented as a finite state machine, and Figure 7 depicts the state diagram. At the

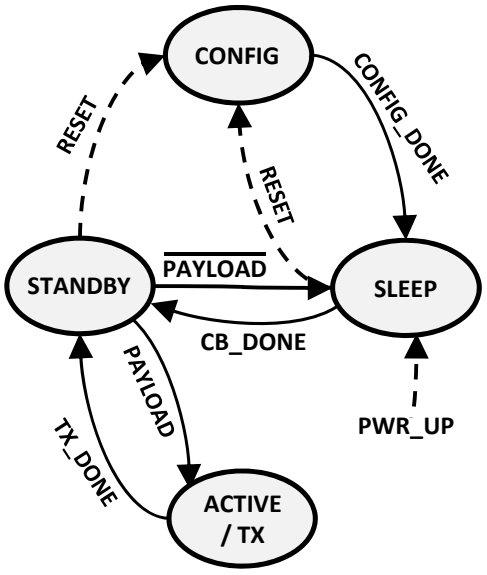

Fig. 7: The transmission controller was implemented as a finite state machine where the number of transmissions depends on the length of the payload.

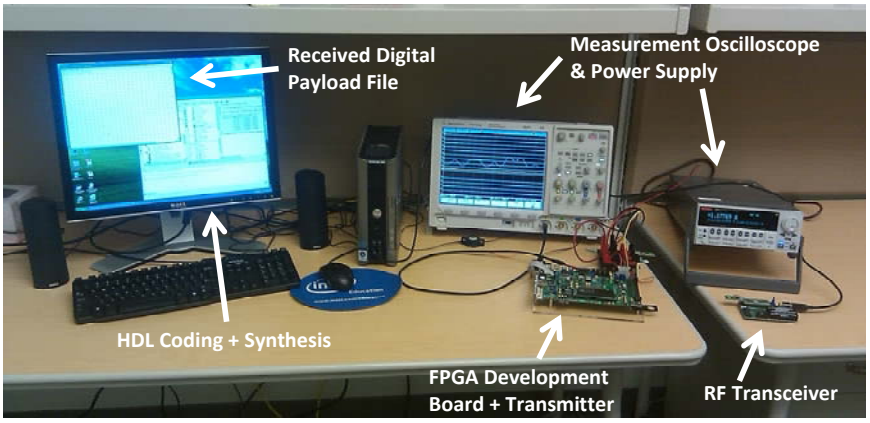

Fig. 8: FPGA based wireless setup developed in this project.

power up of the system, the controller immediately operates in the SLEEP mode, forcing the transceiver to occupy minimal power. When the system is reset, the controller begins its transceiver configuration (CONFIG) communicating information such as payload size, operating modes, and data rate. After the one time configuration is complete, the controller returns to SLEEP mode. Many cycles later, when the collapse buffer is complete (CB_DONE), the controller awakens the transceiver and waits in the STANDBY mode. After voltages have settled, the transceiver begins transmitting data in the ACTIVE/TX mode. Once the transmission is complete, the transceiver will then be placed in the STANDBY state. This STANDBY transition is used as an intermediate state needed between transmissions, since there are a maximum number of bits that can be transmitted in each packet. If there are more payloads to transmit, the transceiver will return to the ACTIVE/TX mode. Otherwise, the transceiver will SLEEP and await the next set of data to be transmitted. The transmission of each frame will begin and commence with the SLEEP state.

3.3. FPGA Based Experimental Results

3.3.1. Testing Framework and System Verification

The described adaptive compression system was designed in verilog, implemented on a Xilinx Virtex 5 LXT [8] prototype board and integrated with a Nordic nRF24L01+ transmitter/receiver system. After the signal has been received, a microcontroller transfers the received data to a host computer for post-processing and signal reconstruction using MATLAB. The system was verified by using a sample size (S) of 7680, a 12-bit quantization factor (Q), and a 128 bit frame size (N). The EEG sample was taken from an online database of patients who suffer from epilepsy [9]. Figure 8 shows the picture of the overall measurement and analysis setup. Figure 9 shows that the original and reconstructed EEG signals of the adaptive compression processor visually match. The accuracy, performance, and energy measurements are presented in Table I. Figure 10 verifies the transitions of the 2-bit FSM controller, and the disparity between sleep time and transmission time. As designed, the transceiver spends a minimal time transmitting and majority of the time in the ultra low power sleep mode. Lastly, the FPGA signal routing of the adaptive compression processor is shown in Figure 11.

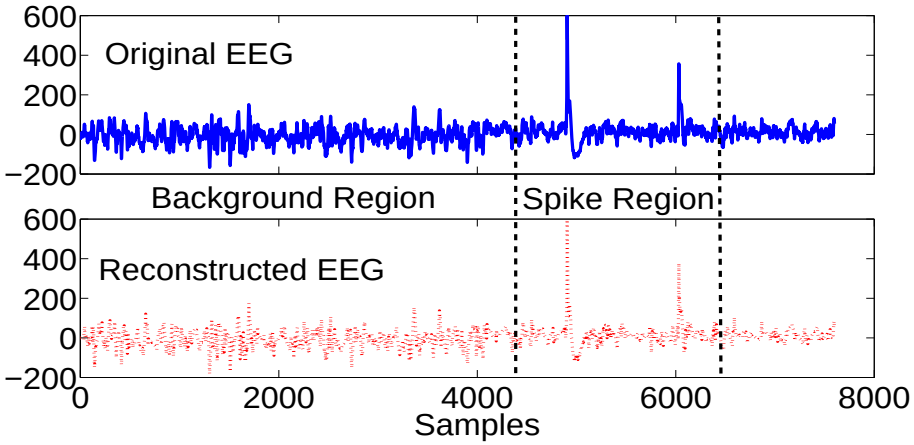

Fig. 9: The reconstructed EEG signals is visibly indistinguishable from the original EEG after passing through entire wireless data compression system.

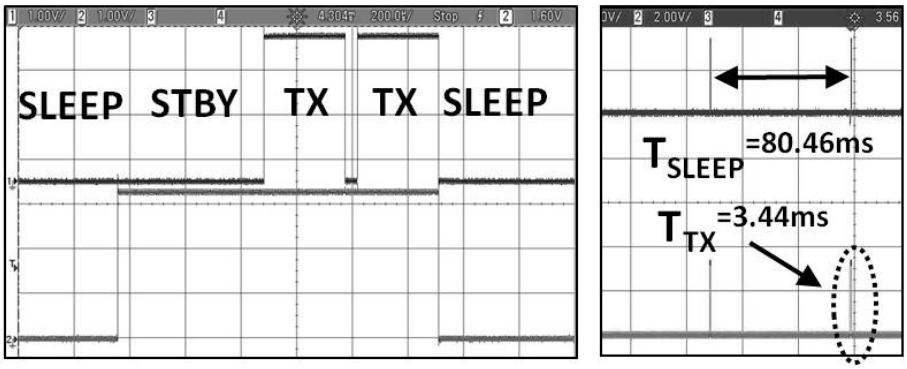

Fig. 10: Oscilloscope measurements of the 2-bit finite state machine transitions (left) and real-time measurements of the transmission and sleep times (right). The screenshots represents a transmission of compression data.

3.3.2. Signal Measurements

Our adaptive compression algorithm is an intermediate solution between two alternative approaches: full compression (at all times) and no compression. With the former case (full compression), all the data is compressed offering ultra low power system, at the expense of accuracy. With the latter (no compression), no data is compressed and an accurate transmission occurs but at the expense of high power. We have compares our adaptive compression implementation against the aforementioned methods.

The signal measurements of the three compression methods are shown on the left of Table 1. The Percent-Root-Mean-Difference (PRD) is a commonly used measure of reconstruction accuracy [6]. The Spike PRD and Total PRD denote the accuracy in the epileptic regions and entire sample, respectively. As the PRD gets smaller, the reconstruction becomes more accurate, with a

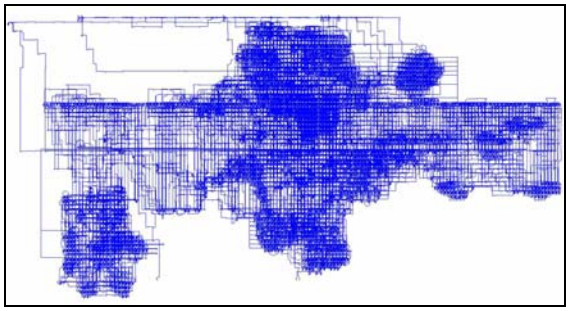

Fig. 11: FPGA signal routing of the three stage adaptive compression processor.

PRD of zero representing ideal reconstruction. Additionally, the compression ratio (CR) determined as well. When there is no compression, CR = 0.83 and the PRD = 2.65. Ideally, the CR would be 1, but due to the signature packets that accompany the coefficients, this creates an inefficient condition when there is no compression. When all the data is compressed the PRD is high (less accurate transmission) but there is more compression. The adaptive compression method has accuracy in the epileptic regions as shown by the Spike PRD of 1.29, and can achieve a reasonable compression ratio of 7.51 to reduce the volume of data transmitted.

Table I: Signal Measurement Comparison Summary of Data Compression Methods for S = 7680, N = 128, Q = 12

Compression Method	**Signal Measurements**		
	Spike PRD	**Total PRD**	**CR**
None	1.31	2.65	0.83
Full	17.1	41.3	17.2
Adaptive	1.29	40.1	7.51

3.3.2. Performance and Energy Measurements

Table II shows the results of the transceiver measurements in regards to time and energy. Depending on the type of compression method, the transmission (T_{TX}) and sleep (T_{SLEEP}) times for a frame size of data will vary. The results show that when there is no data compression maximum transmission time will be achieved. Also, when all the data is compressed, maximum sleep time can be achieved. With adaptive compression, the system can operate at the end of both spectrums dependant on the type of data that is being processed. The energy consumed by the transceiver to represent the 7680 point sample is shown in column E_{TCVR}. This measured energy can be calculated methodically by the total energy required to transmit the entire sample is $E_{TCVR} = P_{TX}T_{TX} + P_{SLEEP}T_{SLEEP} + P_{STBY}T_{STBY}$. As we would expect, the average power to transmit the sample is more than 3X larger when there is no compression, because the transmitter will have to transmit more bits. When adaptive compression is used, the total energy is slightly larger than when all the data is compressed. This result is also true because the total number of bits transmitted (TX bits) is slightly larger. In addition to the transceiver energy, the FPGA energy is also reported. This energy, E_{FPGA}, is composed of the I/O, logic and signal contributions, and it

Table II: Energy/Performance Measurement and Comparison Summary of Data Compression Methods for S = 7680, N = 128, Q = 12

Compression Method	**Performance / Energy Measurements**					
	T_{SLEEP} (ms)	**T_{TX} (ms)**	**E_{TCVR} (mJ)**	**TX bits**	**E_{FPGA} (mJ)**	**E/bit (nJ/b)**
None	71.6	12.3	8.68	135 680	4.83	146.6
Full	80.5	3.44	2.36	34 304	7.34	105.3
Adaptive	71.6 to 80.5	3.44 to 12.3	2.78	46 848	7.50	111.5

neglects the extra peripherals that dominate and do not contribute to our specific design. The minimum FPGA energy to process the sample (4.83 mJ) occurs when there is no compression, and this is because the collapse buffer stage (see Fig. 3) has been removed. Additionally, the two methods that use the collapse buffer for compression require almost 2X more energy.

The energy-efficiency of a wireless system can be quantified as the total energy required to process and transmit a sample. We can define the energy per bit as our metric as $E/bit = (E_{TCVR} + E_{FPGA})/(S*Q)$, where S is the sample size, Q the quantization, and the quantity S*Q is the total number of bits being processed. Ideally, we would like to reduce the energy per bit as much as possible to improve the energy efficiency. From Table I, we see that when the bits are full compressed, we arrive at the E/bit of 105.3 nJ/b, and the PRD is 17.1. The adaptive compression system can attack both accuracy and energy concerns simultaneously. Our results show that the adaptive compression system can achieve a low PRD (1.29) during spike behavior and relatively low energy per bit (111.5 nJ/b).

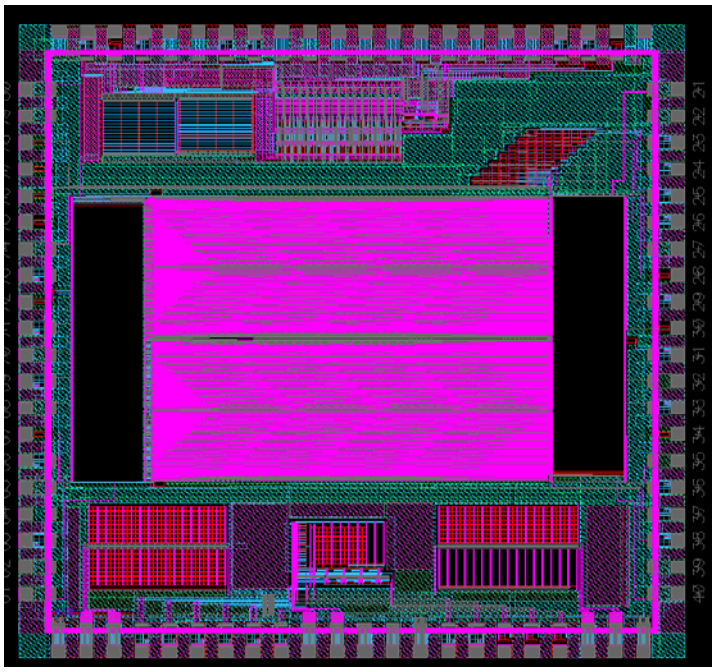

Fig. 12: Layout of the test-chip taped-out with the Wavelet filters.

3.4. Test-Chip Design of the Wavelet Filter

We have also taped-out a test-chip with design of the wavelet filter. The layout of test-chip is shown in Figure 12. The fabricated chips were recently received. We are currently planning for testing and characterization of the chips. The details of the design are given below:

- Total Design Area was 1993 um X 1100 um = 2.02 mm^2 (Total = 5.76 mm^2), # of Transistors in the FIR block = 43,388;
- We used MIT/LL 150nm FDSOI Process with nominal devices and devices optimized for subthreshold operation.
- The design consists of two DWT Filters; 1) One design with nominal devices 2) One design with subthreshold optimized devices
- The DWT Filter consists of a Low Pass and High Pass Filter Pair, each implemented as a 4-tap FIR Filter
- An 8-bit, 128 Sample FF Array was used at the input and output to scan-in and scan-out samples independently
- Level Shifters were implemented to reduce to core voltage to determine the minimum operating point and robustness tests in the new technology.

REFERENCES

[1] D. C. Yates, et. al, "A key power trade-off in wireless EEG headset design," IEEE EMBS conference on neural Engineering, May 2007.

[2] A. Wang and C. Sodini, "A simple energy model for wireless microsensor transceivers," IEEE Global Telecomm. Conf., Nov. 2005

[3] A. Wang and C. Sodini. "On the Energy Efficiency of Wireless Transceivers" Int. Conf. on Communications, 2006. ICC '06.

[4] S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," IEEE Trans. Pattern Anal. Mach. Intell. 11 (1989) 674–693.

[5] K.P. Indiradevi and et. al, "A multi-level wavelet approach for automatic detection of epileptic spikes in the electroencephalogram," Computers in Biology and Medicine, vol. 38, no. pp. 805-816, 2008

[6] Higgins, G., et al, "EEG compression using JPEG2000: How much loss is too much?" IEEE Engineering in Medicine and Biology Society (EMBC), September 2010.

[7] Nordic Semiconductor, http://www.nordicsemi.no

[8] Xillinx, http://www.xilinx.com/products/virtex5/lxt.htm

[9] Klinik für Epileptologie, www.meb.uni-bonn.de