ENERGY-EFFICIENT DIGITAL DESIGN OF RELIABLE, LOW-THROUGHPUT WIRELESS BIOMEDICAL SYSTEMS

A Dissertation
Presented to
The Academic Faculty

by

Jeremy Reynard Tolbert

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
December 2012

ENERGY-EFFICIENT DIGITAL DESIGN OF RELIABLE, LOW-THROUGHPUT WIRELESS BIOMEDICAL SYSTEMS

Approved by:

Dr. Saibal Mukhopadhyay, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Sung Kyu Lim
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Hsien-Hsin Lee
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Maysam Ghovanloo
School of Electrial and Computer Engineering
Georgia Institute of Technology

Dr. Hyesoon Kim
College of Computing
Georgia Institute of Technology

Date Approved: [July 30th, 2012]

ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge God the Father, the Son, and the Holy Spirit, through him all things are possible. Without him I am nothing, and because of him I can achieve something. My relationship with Jesus Christ has calmed me in the storm, when doors were closed, the Lord made a way out of no way.

I would like to extend the highest level of gratitude to my Advisor, Saibal Mukhopadhyay. Five years from this writing, we both just began our journey at Georgia Tech, and we each took a leap of faith to work together. Based on you guidance and your concerned mentorship, I am the first student to graduate from the GREEN Lab. It is an extreme honor to say I have learned from you, and I am encouraged by the future direction of the lab.

I would also like to extend special thanks to several professors who have guided me on my path: Dr. Sung Kyu Lim, for your research insights and collaborative efforts; Dr. Maysam Ghovanloo and Dr. Hyesoon Kim, for your valuable input and critical eye with my thesis; Dr. Hsien-Hsin Lee, for your guidance on my thesis and our regular talks about career opportunities; Dr. Raheem Beyah and Dr. Thomas Conte, for your candor, mentorship and open-door policy; and Dean Gary May for developing a sound infrastructure such that underrepresented minorities can achieve in higher education.

Several fellow colleagues were critical to my sanity during this process including Lonnie Parker, Douglas Brooks, J. Chris Ford, Adilson Cardoso, Chad Rosier, George Baah, Amin Rida, Jordan Greenlee, Dr. Damon Williams, Subho Chatterjee, Minki Cho, Boris Alexandrov, and the rest of GREEN Lab.

On various research avenues, my work would be incomplete without the aid of Xin Zhao, Dr. Dae Hyun Kim, Denny Lie, Kwanyeob Chae, Pratik Kabali, Simeranjit Brar and Andrew Burks. It is very easy to get caught up in one's ego, but you all left yours at the door to help me. Your input was greatly appreciated.

I have been fortunate to have been sponsored by many programs that have directly or indirectly funded my graduate education: The National Science Foundation's Graduate Research Fellowship, The Georgia Tech President's Fellowship, The GEM Fellowship sponsored by Intel Corporation, The FACES (Facilitating Academic Careers in Engineering and Sciences) Fellowship, and the SURE (Summer Undergraduate Research for Engineers/Scientists) Program.

In my lifetime, I have been preparing for this journey based on my life experiences. This would not be possible without the positive influence and encouraging nature of my Mother, Sonia Tolbert. My Father, Randolph Tolbert, has shown me the correct model of being a man simply by the actions he performed over his lifetime. Without my only sibling, Jason Tolbert, I would not have the competitive nature and desire to never give up when I feel defeated. My family has and always means so much to me, and this degree is also a testament to their hard work in my life as well.

Several additional family members and friends have contributed a constant stream of encouraging words and actions throughout this five year journey: Venola Tolbert (R.I.P), Sherece Tolbert, Desmond Tolbert, Mai Mends-Cole, Arthur Nickerson, Harvey Freeman, Jr. and Brandon C. Boles. During the times when the moments felt overwhelming, something was said or did to inspire me to continue on.

The most important person in my life is my wife, Cianna Freeman-Tolbert. Over the last nine years, our relationship has flourished on a level that excites me to know what the future holds. Your comic relief, unwarranted honesty, and drive for success have influenced me the most, on a day-to-day basis. You were more than understanding when we decided to move to Atlanta, and your constant support is an impeccable quality known to very few. I love you more than words can express, and through this experience and the birth of our first child, I am indebted to you for life

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The main objective of this research is to improve the energy efficiency of low throughput wireless biomedical systems by employing digital design techniques. The power consumed in conventional wireless EEG (biomedical) systems is dominated by digital microcontroller and the radio frequency (RF) transceiver. To reduce the power associated with the digital processor, data compression can reduce the volume of data transmitted. An adaptive data compression algorithm has been proposed to ensure accurate representations of critical epileptic signals, while also preserving the overall power. Further advances in power reduction are also presented by designing a custom baseband processor for data compression. A functional system has been hardware verified and ASIC optimized to reduce the power by over 9X compared to existing methods. The optimized processor can operate at 32MHz with a near threshold supply of 0.5V in a conventional 45nm technology. While attempting to reach high frequencies in the near threshold regime, the probability of timing violations can reduce the robustness of the system. To further optimize the implementation, a low voltage clock tree design has been investigated to improve the reliability of the digital processor. By implementing the proposed clock tree design methodology, the digital processor can improve its robustness (by reducing the probability of timing violations) while reducing the overall power by more than 5 percent. Future work suggests examining new architectures for low-throughput processing and investigating the proposed systems' potential for a multi-channel EEG implementation.

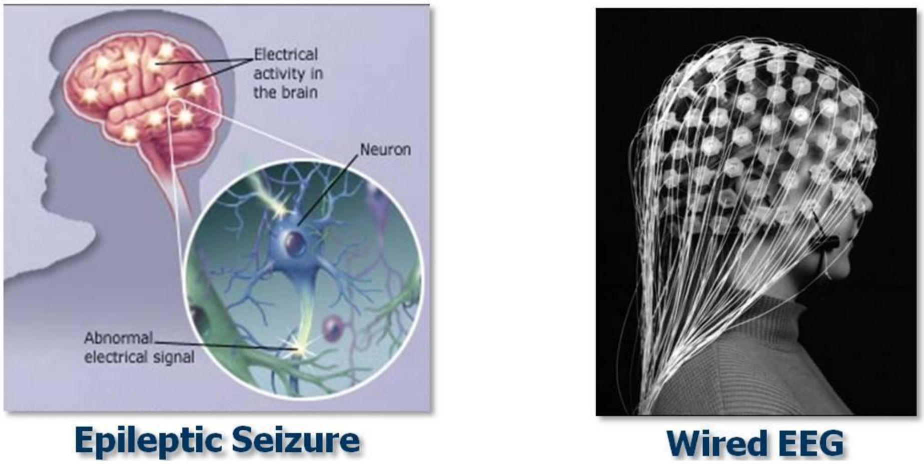
CHAPTER 1

INTRODUCTION

1.1 Background

Epilepsy is one of the most common neurological disorders and is characterized by recurrent, unprovoked seizures. Seizures result from abnormal electrical signals at neural sites in the brain (See Figure 1 (a)). About 1 in 100 people have epilepsy, and this translates into over 50 million people in the world who are hampered by this disease. With a proper diagnosis, medication can be prescribed or in serious cases surgery is an option. Electroencephalography (EEG) is the study of electrical activity produced by the brain, and can be used to diagnose epilepsy, sleep related disorders, and brain tumors. Routine (or stationary) EEGs can provide a 20 to 40 minute sample of data for the analysis of brain activity. To better localize seizure regions of the patient's brain, Intracranial EEGs (IEEG) or Ambulatory EEGs (AEEG) can be used as a method to record more results.

Intracranial EEG is an invasive technique where signals are recorded directly from the human cortex, as opposed to the surface recordings of traditional EEGS. While IEEGs provide the most accurate signal acquisition, the complexity arises in the implanting of the electrodes. Throughout this complexity, researchers feel they can benefit from accurate data acquisition when the means require such [1]. Ambulatory EEG is an outpatient procedure that allows for prolonged EEG recording in the home or work setting. For up to 72 hours, a patient's brain activity can be monitored in their natural



(a) (b)

Figure 1: (a) Epileptic seizures result from abnormal electrical signals at neural sites in the brain. An Electroencephalogram (EEG) is used to measure the electrical activity of the brain to diagnose seizures. (b) Traditional EEGs are wired, which causes patients to be tethered to a local station, limiting the patient's movement.

environment, which increases the chance of recording an important event. While AEEG systems provide the advantage of portability, a number of issues have been identified [2]:

1. Systems can weigh up to 1kg, limiting their portability;
2. Up to 32 channels of activity are typically recorded;
3. Each of these channels requires a wired connection from the patient's head to the recording unit, limiting the patient's movement (See Figure 1 (b));
4. Long term recordings generate large amounts of data for storage, approximately 1GB every 24 hours;
5. This data is time consuming to analyze, taking approximately two hours per 24 hour recording.

To overcome these issues, the design of wireless AEEG systems has been of great interest, because of the additional comfort of no wires or bulky devices to carry. In wireless EEG systems, patients wear a headset, which samples the EEG signal and wirelessly transmits it to a base station (e.g. a mobile device such as laptop or cellphone). The EEG reader can obtain the data by accessing the local host. The pictorial view of an integrated system for remote monitoring of EEG signals using a wireless headset is shown in Figure 2.

By using wireless EEG systems, patients will be able to return to their typical day-to-day schedule, instead of being required to schedule appointments for recordings. Additionally, systems can also allow for real-time feedback, when doctor monitors and diagnoses a patient remotely. Based on the improved patient portability, and real-time feedback, remote wireless monitoring systems can significantly improve the quality of the healthcare delivered to the epileptic patients.

A major concern in wireless AEEG systems is the battery lifetime due to the wireless transmission of high-quality data. A wireless EEG unit is composed of an amplifier, analog-to-digital converter (ADC), and a wireless transceiver [3]. Yates et. al. has shown that the primary source of power dissipation in wireless EEG systems is the wireless transmission, that directly depends on the volume of the transmitted data. Consequently, methods have been proposed to reduce the volume of data transmitted [2].

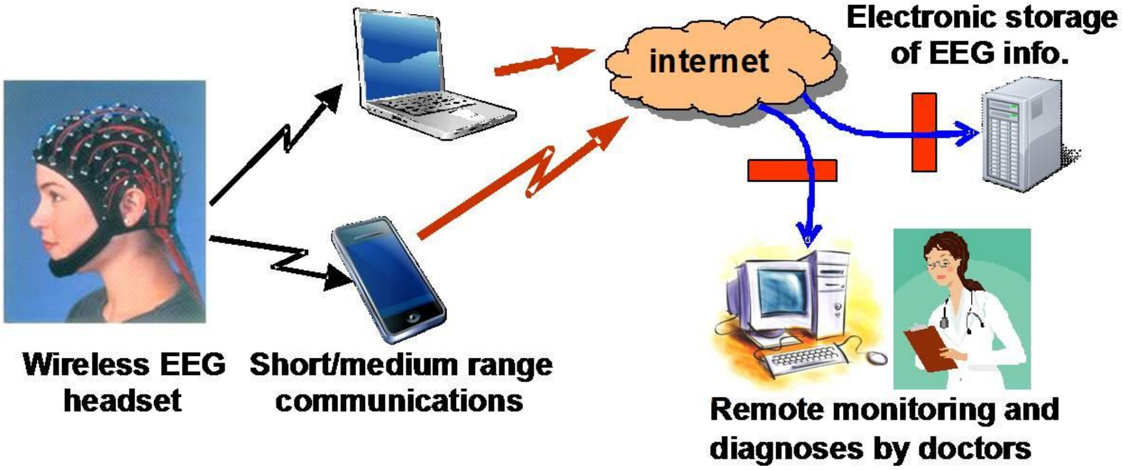

internet
Electronic storage of EEG info.
Wireless EEG headset
Short/medium range communications
Remote monitoring and diagnoses by doctors

Figure 2: Pictorial view of remote wireless electroencephalography (EEG) system monitoring

1.2 Prior Work in Wireless Electroencephalography Systems

Prior work in wireless EEG systems date back to 1994, where a 16-lead infrared based wireless system was developed [4]. The infrared system reported a maximum transmission distance of 5 meters, and used 12 bits per channel for signal representation. The main disadvantage to infrared transmission is the requirement of an uninterrupted light path between the transmitter and receiver. By 1996, experiments and design methods focused on improving wireless transmission and exploring Bluetooth frequencies (2.45 GHz) had begun [5]. While the previous two works were the first to demonstrate the feasibility of transmitting and receiving brainwave signals, they were not a true mobile wireless system; the transmitters were not battery powered, and thus immobile. In 1997, the Cleveland Medical Clinic first introduced a mobile EEG headset prototype powered by a 9 Volt battery, with an 8 channel transmitter unit [6]. This was the first of many works to successfully develop a mobile wireless electroencephalography system [1, 7-16].

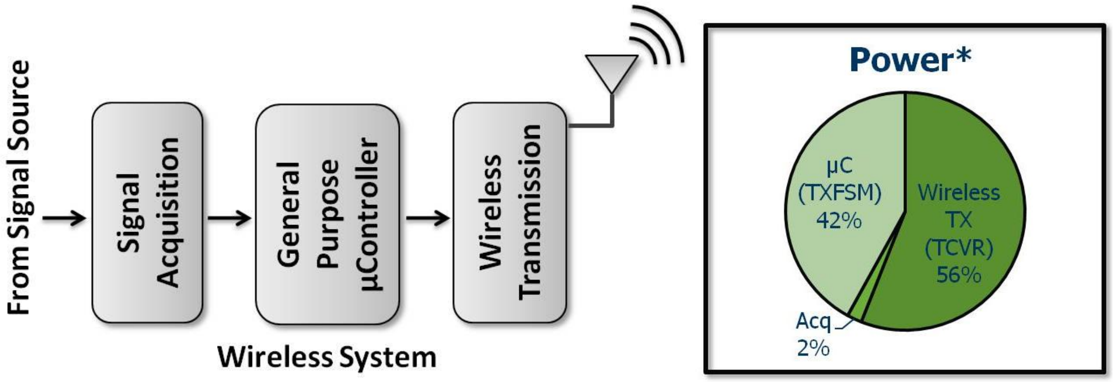
Commercial wireless EEG systems are available, but have some limitations [13, 14]. Neurosky's Mindset claims to have a research grade EEG, but only has one electrode. Many of Neurosky's target applications are also developed for personal entertainment usage. Emotiv's neuroheadset provides 14 sensors for EEG neuro-signal acquisition, and also includes a software developer kit for signal display and FFT processing. The previous have similar issues: they are bulky implementations with the primary goal of providing a means of entertainment for the consumer.

Research grade wireless EEG systems have also been presented, with the primary focus on assisting in the way healthcare information is delivered [10, 15, 16]. Advanced

Brain Monitoring, Inc developed the B-Alert platform to address sleep apnea, memory dysfunction and alert monitoring. The B-Alert X10 provides up to seven channels of biomedical signal monitoring, and is marketed towards researchers, physicians and psychologists. Filipe et. al. presented a wireless multichannel EEG platform that includes a digital controller (i.e. TI-MSP430), power management block and 8 channels for signal acquisition. IMEC has been on the forefront of research in the wireless EEG domain, developing a low-power 8-channel ASIC for signal acquisition, a TI-MSP430 to control the digital settings, and an NRF24L01 radio for low power wireless transmission. The research grade implementations demonstrate the functionality of the wireless EEG platform, but it leaves much to be desired in regards to designing the systems for optimal energy efficiency.

A generic platform for processing wireless EEG signals can be seen in Figure 3 (a). This platform is typical for the existing research grade implementations. From the signal source, there is signal acquisition which encompasses amplification and analog-to-digital conversion (ADC). After the signal has been acquired, there is a general purpose microcontroller to facilitate the data and control the ADC and wireless transmitter. After wireless transmission, a local host machine performs data processing and advanced analysis.

Based on the survey from [17], the power breakdown of a wireless EEG system can be seen in Figure 3 (b). It can be seen that the wireless transmitter is the most power intensive component in a wireless system [18]. As a result, circuit and system designers have attempted to reduce the energy by looking at the wireless transceiver [2, 19-21]. Other works have looked and designing custom individual components: the EEG sensors



(a) (b)

Figure 3: (a) Wireless system framework for existing EEG systems. (b) Power breakdown of components in a wireless EEG system**.**

[10, 22], analog-to-digital converter [23]. The large power of the microcontroller can be attributed to using a general purpose microprocessor (TI-MSP430) for an application specific task. The operation of the general purpose microcontroller is required: it directs the data to the wireless transmitter, and controls the wireless transmitter. Overall, the combined components of the microcontroller and wireless transmitter contribute 98 percent of the overall system power.

1.3 Proposed Solution

Based on existing EEG system implementations, the microcontroller and wireless transmitter contribute a significant portion of the overall power. To achieve the maximum power reduction, it's essential to focus on reducing the power of these two components. Two design principles will be implemented to reduce the overall power: data compression and a custom microcontroller.

The wireless transmitter contributes to 56 percent of the overall power and it is directly related to the volume of data that is being transmitted. By using data compression, the volume of data will be reduced, which will allow the transmitter to operate for a shorter amount of time. The microcontroller contributes to roughly 42 percent of the overall power when a general purpose microcontroller is used. Traditional systems have suggested using a Texas Instruments TI-MSP430, but it has not been customized for the specific low power application. By implementing a custom microcontroller, the power associate with controlling the wireless transmitter can be reduced as well. Figure 4 depicts the proposed design principles for power reduction.

An improved, proposed wireless system for EEG processing is shown in Figure 5. At the top level, there are three operations that are performed: signal acquisition, data processing, and wireless transmission. Compared to existing systems, the proposed system aims to perform data processing on-chip, which will reduce the energy consumed in transmission. Based on supply voltage scaling, it is possible to design a data processing block with minimum energy. However, designing an energy efficient wireless system will require knowledge based on the interactions with the neighboring blocks (acquisition, transmission).

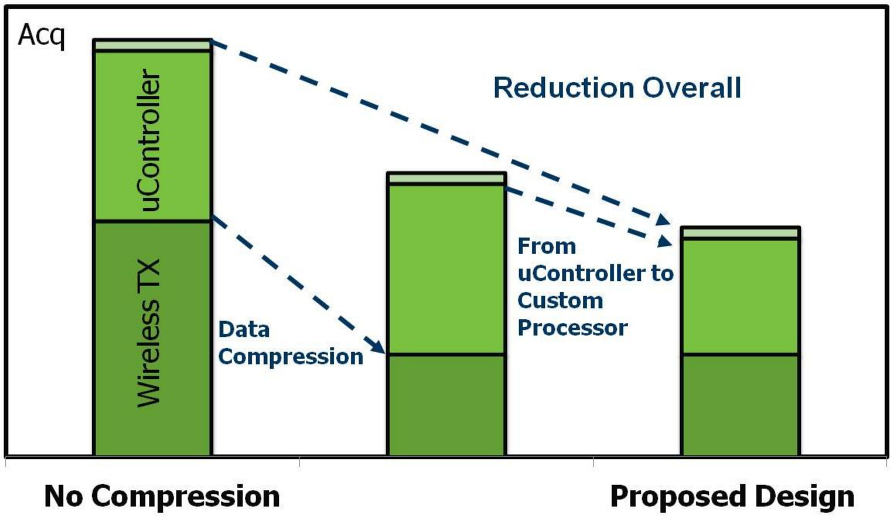


Figure 4: The proposed design principles for power reduction include data compression and developing a custom microcontroller.

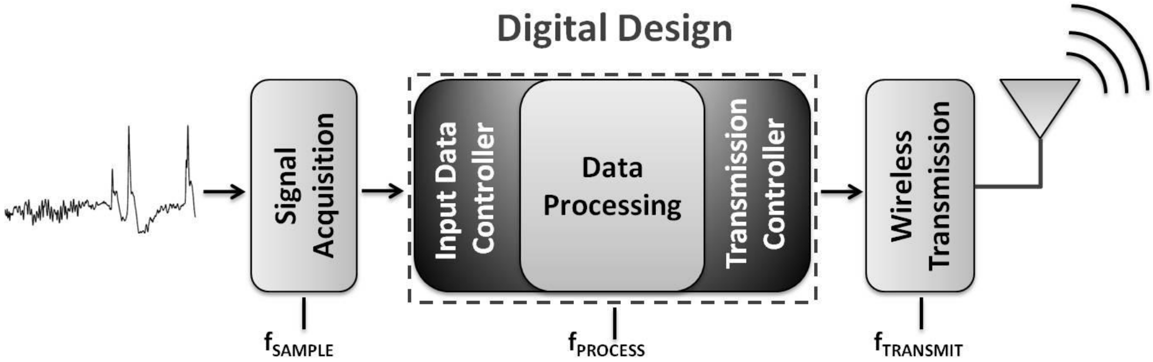


Figure 5: Proposed wireless system model. The focus of this thesis is using digital design to improve the energy efficiency of the wireless system.

In other words, independently designing each block for minimum power will not result in minimum power for the system. To consider the entire system interactions, this thesis will focus on using digital design techniques to perform a system level energy optimization.

The system level energy optimization problem is further complicated by the timing requirements that are involved in the interactions of the blocks. For example, given a long acquisition time (small f_{SAMPLE}), combined with a small processing time (large $f_{PROCESS}$) and small transmission time (large $f_{TRANSMIT}$), it is unclear how the energy will be distributed between the blocks. Based on the different frequencies, blocks will have different timing requirements. Some blocks will be dominated by idle energy (energy consumed while waiting to process), while others will be dominated by dynamic energy (energy consumed while processing). The ideal combinations of idle and dynamic energy will result in optimal energy efficiency (minimum power).

1.4 Thesis Contributions

The objective of the proposed research is to improve the energy efficiency (and thus battery life) of low-throughput wireless biomedical systems by employing digital design techniques. By employing data compression techniques and designing a custom baseband processor, the power of the wireless biomedical system will be significantly reduced. The focus of this work is on the modeling and design of a wireless electroencephalography (EEG) system that requires analog acquisition, digital processing and wireless transmission. This dissertation makes the following contributions:

- **An Accuracy and Energy Aware System for Adaptive Data Compression -** Data Compression is well known technique that can be used to reduce the amount of data that is required to represent the signal. For the proposed wireless system, reducing the volume of data with reduce the operating time (and power) of the wireless transmitter. Recognizing that accuracy is desired in epileptic regions and low power is desired overall, the first contribution presents an adaptive data compression algorithm to detect, compress and transmit wireless EEG signals.

- **Chameleon: A Content-Aware Adaptive Compression Architecture for Wireless Electroencephalography** – The general purpose microcontroller in existing EEG systems is a dominant portion of the system power as well. To reduce this, a custom digital baseband processor, Chameleon, will implement the proposed adaptive data compression algorithm, while facilitating the original purpose of the microcontroller. The Chameleon processor will be hardware verified and ASIC optimized to reduce the overall system power.

- **Analysis and Design of Energy and Slew-Aware Subthreshold Clock Systems** – The digital Chameleon processor was optimized to operate at near threshold voltages with a maximum frequency of 32 MHz. When designing low power systems, reliable clock networks need be investigate to make certain that the timing requirements are met.

1.5 Chapter-by-Chapter Summary

The rest of this dissertation is organized as follows:

Chapter 2 introduces the concept of adaptive compression while also introducing the reader to energy efficient system level design for wireless electroencephalography. The basis for this chapter involves modeling a 32-channel wireless system, and examining how subthreshold design can be used to achieve energy efficiency.

Chapter 3 develops a custom digital processor for adaptive data compression of electroencephalography signals. The processor has been optimized with low-power techniques to improve the energy efficiency overall. By introducing the custom digital processor, the system level energy-efficiency is improved, compared to existing approaches.

Chapter 4 further examines the development of the digital processor, by examining the impact of high frequency clocks in low voltage domains. A new approach to subthreshold clock tree design will be presented that reduces the clock tree energy and improves the robustness.

Chapter 5 presents the final conclusions, summarizes the thesis contributions, and proposes suggestions for follow-up research topics.

CHAPTER 2

AN ACCURACY AND ENERGY AWARE SYSTEM FOR ADAPTIVE DATA COMPRESSION

2.1 Introduction

In this chapter, a digital system is proposed to adaptively compress EEG signals, by determining the compression rate based on the real-time information content. The digital system detects the information content of EEG signals of each channel independently and performs adaptive compression. The system aims to preserve the generic behavior of the signal by transmitting compressed data for background EEG and uncompressed data during regions of epileptic (or spike related) activity. The continuously transmitted EEG signal is available for EEG interpretation, which provides the EEG interpreter a higher probability for correct diagnosis. By transmitting low power, low quality data during background regions, and high power, high quality data during epileptic regions, the processing method is best described as adaptive compression. In general, the method provides a dynamic energy and accuracy tradeoff.

2.2 Design Challenges

The primary challenge involved in data compression for EEG transmission is the conflicting energy and accuracy requirements. Traditional methods for data compression have proposed a full compression, where all data presented is compressed with the same compression ratio. While traditionally popular in information theory, these schemes for full compression have their limitations when used to process biomedical signals. For full compression of data, increasing the compression ratio reduces the accuracy of the represented signal. A higher compression also reduces the data volume. By reducing the volume of data, the wireless transceiver will require less time to transmit the data, thus reducing the energy in transmission. For EEG systems, the epileptic activity needs to be accurately transmitted, because it contains the critical signal behavior to diagnose and detect epileptic seizures and abnormal behavior. As a result of this requirement, the compression ratio needs to be chosen to reduce the errors for epileptic regions. Unfortunately, the occurrence of epileptic events is rare; therefore significant energy is lost by transmitting background activity a higher accuracy than necessary. On the other hand, if a higher compression ratio is chosen to reduce energy during background activity, the epileptic activity will be transmitted with a lower accuracy than desired. The reduced accuracy of epileptic activity will make the data more difficult for an EEG interpretation. Therefore, a real-time dynamic trade-off of energy and accuracy is not possible in a full compression system.

To eliminate the static nature of a full compression system, discontinuous compression was proposed. Described as an event related solution, discontinuous compression will present the neurologists with only events that include epileptic activity [2]. During this time, the acquired signal is directly transmitted as decompressed (high quality) data via the transceiver. However, when background activity occurs, the channel is cut off and no data is transmitted. Based on the rarity of epileptic activity, this method provides very low energy solution but does not consider accuracy in all regions. It has been noted that the analysis of EEG records are subjective, often requiring background information. Additionally, accurate definitions of what is epileptic activity can be ambiguous [24]. Moreover, algorithms to select epileptic activity are imperfect and as a result false detections exist. By removing data points for analysis, this will further increase the difficulty to characterize EEG signals and spike related events. As a result, in the presence of false-negative detections, the error will be infinite as no information will be presented to the EEG interpreter.

2.3 System Design and Methodology

To overcome the challenges related to the full and discontinuous compression methods, a new data compression method is proposed to transmit EEG data with the best accuracy and energy tradeoff. During spike related events, it is desirable to transmit EEG signals with the highest quality to preserve accuracy. At other times, accuracy is not the primary goal, and low energy is desired. Therefore, during background activity, the data can be compressed to reduce the transmission energy, while still maintaining the recognizable signal behavior. This section introduces the proposed architecture for adaptive compression. The proposed system acts as a digital baseband between EEG frontend acquisition and RF transceiver. Although the architecture is presented in the context of EEG, the system is applicable to generic wireless biomedical signal processing.

2.3.1 Overall System Objective

Single channel EEG implementations exist, but for a thorough system that is applicable in Epilepsy outpatient procedures, 16-32 channels are necessary. Based on this requirement, the proceeding results will focus on a multi-channel EEG transmission. The overall architecture of the proposed system is depicted in Figure 6. Thirty-two

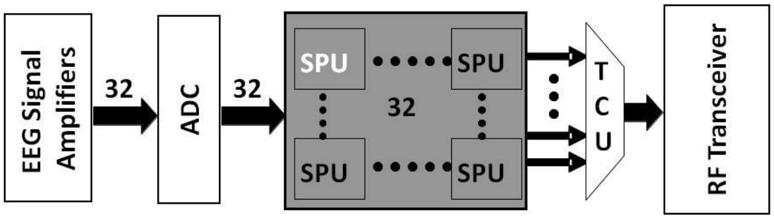


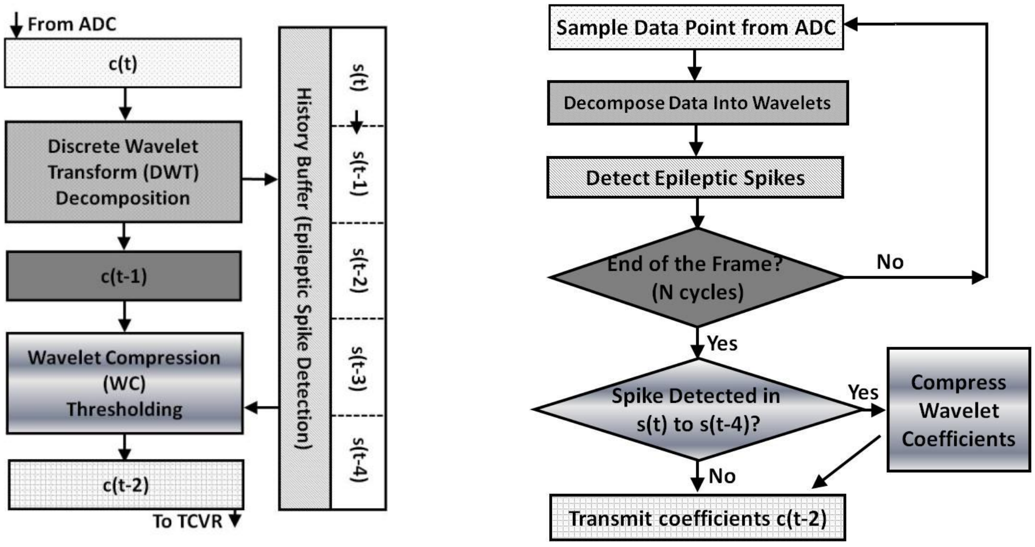
Figure 6: The proposed digital implementation of a wireless 32-channel EEG system. A multi-core system was chosen for the potential to exploit parallelism in multi-channel detection and analysis.

acquisition amplifiers are used to obtain data at different locations and they are then converted to digital signals with the Analog-to-Digital Converter (ADC). In the digital domain, a multi-channel algorithm is implemented to detect and process the EEG signals. The gray block is emphasized to denote the design focus of the proposed system architecture. In order to detect epileptic activity, an on-chip implementation of a complex algorithm is required. To improve the detection methods, multi-channel analysis can be performed. In response to the demand for complex and parallel algorithms, it is possible to use a digital system that exploits multi-core processing to complete this task. Previous works have attempted analog implementations of wireless EEG systems, but a thoroughly investigated, custom digital system has the potential provide performance and power benefits [7, 25, 26]. Additionally, only generic studies of digital implementations have been done, showing skewed results that support analog over digital systems [27]. The network of these signal processing units (SPUs) receives its input from the ADC. Once the processing is complete, the transmission control unit (TCU) initiates the RF transceiver to transmit the data to a wireless host.

2.3.2 Signal Processing Units (SPU)

The microarchitecture of the SPU for a single channel is depicted in Figure 7 (a). Each SPU processes data on frame-by-frame basis, where a frame is a collection of N samples. If a spike is not detected within the frame (or neighboring frames), the data is compressed and transmitted at a later stage. When a spike is detected in the frame (or neighboring frames) the data is not compressed and later transmitted, maintaining a high quality signal. Neighboring frames are relevant for epileptic activity, because they can provide the EEG interpreter insight into behavior before and after spikes occur. In each cycle, data is shifted into the data register c(t). The register c(t) holds the last frame (N samples) of data. Additionally, spike detection is performed in every cycle and data compression is performed once at the end of every frame. Both of these processes were performed in the discrete wavelet transform domain, with a Daubechie-2 characteristic wavelet. EEG detection is performed every cycle (as each sample is acquired), yielding N detections for a frame size of N samples.

If a spike is detected, the entire frame is classified as epileptic activity. The shift-and-detect scheme for sampling and epileptic spike detection is portrayed in Figure 8.


From ADC
c(t)
Discrete Wavelet Transform (DWT) Decomposition
c(t-1)
Wavelet Compression (WC) Thresholding
c(t-2)
To TCVR
History Buffer (Epileptic Spike Detection)
s(t)
s(t-1)
s(t-2)
s(t-3)
s(t-4)
Sample Data Point from ADC
Decompose Data Into Wavelets
Detect Epileptic Spikes
End of the Frame? (N cycles)
No
Yes
Spike Detected in s(t) to s(t-4)?
Yes
Compress Wavelet Coefficients
No
Transmit coefficients c(t-2)

(a) **(b)**

Figure 7: (a) Signal Processing Unit (SPU) unit to detect and compress data so that is can be sent to the Transmission Control Unit (TCU) for time division multiplexing. (b) Algorithm of the SPU used to process EEG signal for one channel. Shaded boxes are used to denote complementary steps/blocks.

This scheme results in an inherent redundancy in detection that increases the number of computations, but reduces the number of false-negative detections. A false negative is an incorrect classification of a feature as a non-feature, in our case epileptic spikes. This means epileptic spikes will be represented as background activity. Although the shift-and-detect scheme increases the energy required for computation, it improves the overall system accuracy. The algorithmic flow chart of the adaptive compression system can be seen side by side with the SPU unit in Figure 7 (b). At the completion of the discrete wavelet transform stage (detection), values are latched into the register c(t-1).

After detection, compression is performed on the incoming EEG signal on frame-by-frame basis (i.e. compression is performed on consecutive non-overlapping frames). The compressed data is represented in the wavelet domain, and the coefficients are then latched in another register c(t-2). At the end of the frame, the spike detection result for the entire EEG frame is stored in the history buffer (HB). When the frame at time (t) has been sampled, the frame in (t-2) is ready for transmission. This delay allows for optimal transmission, by considering the epileptic activity in neighboring frames: (t), (t-1), (t-3),

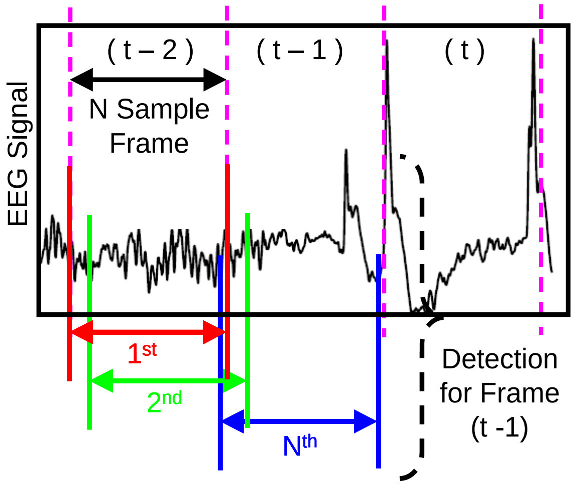


Figure 8: Shift-and-detect scheme for high resolution spike detection. The sliding window will make N detections on a frame of N samples.

and (t-4). Based on the result of the detection, the data c(t-2) will be compressed or decompressed and then sent to the transmission control unit (TCU) to facilitate wireless transmission.

2.3.3 Spike Detection and Compression with Wavelets

The method of spike detection and data compression are performed in wavelet domain to isolate the frequency bands associated with multi-level resolutions [28]. Both the method of compression and detection are based on a simple thresholding technique. If the magnitude of the wavelet coefficients in a frequency band has surpassed the spike threshold (ST), epileptic activity (or spike) is said to be detected (spike detection). For data compression, all the wavelet coefficients less than the compression threshold (CT) are removed (wavelet compression). The packet format for compressed coefficients requires a signature and corresponding data packet. The signature packet is composed of a digital (1 or 0) bit representation for each sample position, where a ‘1’ denotes a kept sample and a ‘0’ denotes a compressed sample. The data packet is composed of all decompressed values, eliminating the compressed values (now zeros) in between. With the signature and data packet, the compressed signal can be reconstructed at the receiver. The resulting method of compression is data dependent. As an added note, the results of the system are determined by the selections of ST and CT. To show the feasibility and advantage of adaptive data compression, simple techniques for detection and compression have been used. The results shown later encourage a more thorough investigation of complex detection and compression algorithms as possible future work [29-31]. The original and reconstructed signals are portrayed in the top of Figure 9. As each sample is received, the wavelets coefficients are compared against the spike threshold (ST) and the spike detection signal is enabled (bottom of Figure 9). As we will see in section 3.4.1, the reconstructed signal and the original sample are determined to be equivalent, when compared numerically.

2.3.4 Synchronization of Acquisition and Transmission

The data from an EEG sensor enters the proposed system through the ADC, and processing is performed on a frame-by-frame basis. For a 32-channel operation, the input frames are received and processed by their corresponding SPU in parallel. After processing, each SPU will generate the transmission frame corresponding to its channel. As noted before, the frames will be compressed or decompressed versions of the EEG data, depending on the result of spike detection. A transmission control unit (TCU) sequentially collects the transmission frames of each channel and creates a data payload for the RF transceiver.

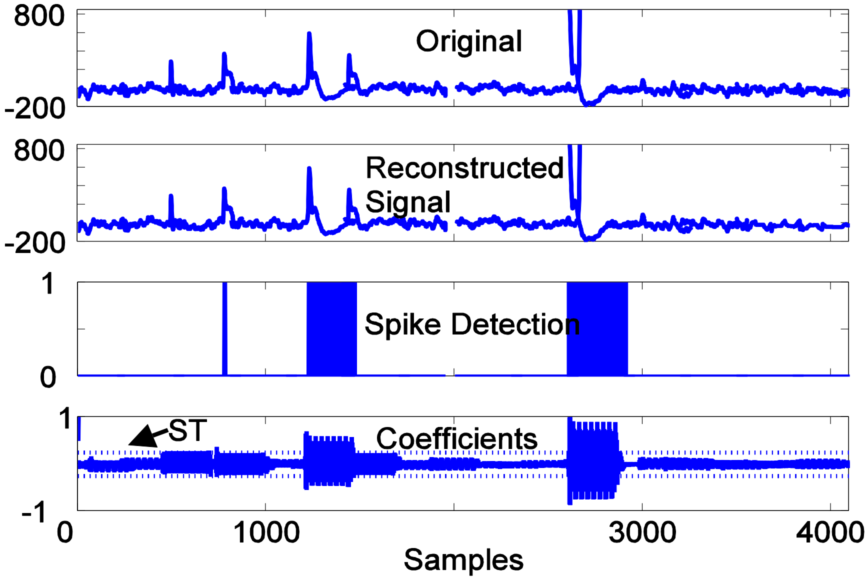


Figure 9: Digital adaptive compression system functionality with simple spike detection. Spikes are detected when the wavelet coefficients surpass the Spike Threshold (ST).

The details of parallel data sampling and serial data transmission are shown in Figure 10. In general, a data payload for RF transmission is collection of 32 transmission frames. The transmission of each data payload is synchronized with each new EEG frame acquisition. This is possible considering the low sampling rate of EEG signals (150-250Hz) and high data rate of transceivers (2 Mbps). For example, considering a frame size of 1024, 32 EEG channels, 16-bit data, and sampling frequency of 500Hz, results in acquisition time of 2.048 seconds. For an RF transceiver with 2 Mbps data rate, the transmission of the data payload will require less than 0.25 seconds. Based on this timing example, a parallel EEG acquisition unit with serial data transmission is a completely viable approach.

2.3.5 The Digital System Components

The system components incorporated with this design make it simple to envision a custom processor for adaptive data compression. The registers (i.e. c(t-n), for n = 0,1,2) and history buffer are memories that can be created by using latches and flip-flops. The discrete wavelet decomposition unit can be designed using a set of FIR filter banks and latches. Spike detection and wavelet compression is performed by using digitally

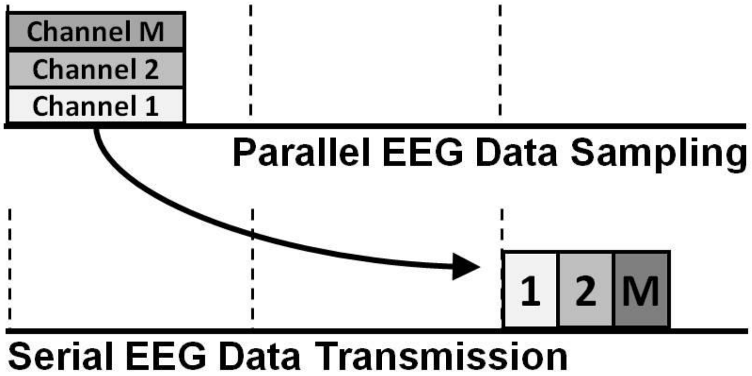


Figure 10: The 32-channel EEG signals will be sampled in parallel and (after processing) time multiplexed before transmission.

comparing the wavelet coefficients with ST and CT.

2.4 Results on Accuracy and Compression Ratio

In this section, experiments have been performed the compare the accuracy and compression ratio of the adaptive compression scheme against other methods. The other compression schemes for comparison are: 1) discontinuous compression, 2) full compression for energy and 3) full compression for accuracy. Recall that with discontinuous compression, the data is either directly presented or removed entirely. When the primary goal is to save energy, the full compression method uses a large compression ratio. When the primary goal is accuracy, the full compression method uses a smaller compression ratio. The results of both of these methods are shown, because the full compression method can only be designed to achieve one of the two goals.

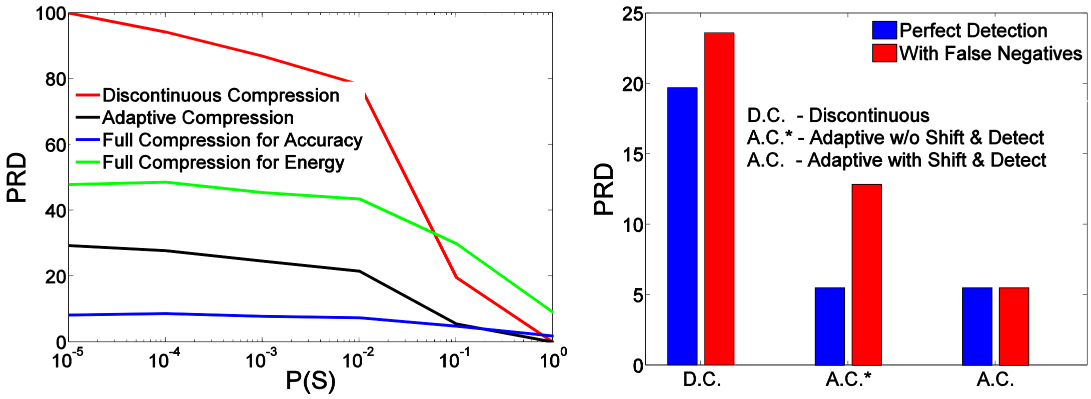
The above metrics will vary based on the data presented, and the data used was taken from actual EEG data from an online database [32]. Furthermore, the data was sampled to generate unique signals based on the probability of spike occurring within a frame, P(S). As P(S) approaches one, the entire frame becomes epileptic data. As the P(S) approaches zero, the entire frame becomes background EEG. The background EEG patterns were selected based on a normal random distribution, and epileptic signatures were inserted in time. The analysis of the compression methods over the wide range of P(S) will show the expected behavior in all scenarios.

2.4.1 Accuracy and Compression Ratio

To compare the accuracy of the reconstructed and original signal, the Percent Root Mean Square Difference (PRD) is defined by:

$$PRD = 100 \times \sqrt{\frac{\sum_{i=1}^{n}[x_{ori}(i) - x_{rec}(i)]^2}{\sum_{i=1}^{n} x_{ori}(i)^2}} \qquad \text{Eq. (2.1)}$$

A smaller value of PRD means that the signal reconstruction is more accurate. That is, perfect reconstruction occurs with a PRD equal to zero. The PRDs of the four compression methods (for the reconstructed and original signals) is shown in Figure 11 (a). The first thing to note is that the full compression methods (blue and green), have a limited dynamic range of accuracy once the system is designed. On the contrary, the discontinuous and adaptive compression methods can alter their accuracy over a wide range to achieve ideal reconstruction in the presence of spike activity. Adaptive data compression has a significant advantage over discontinuous compression; when spike activity is not detected, the general signal behavior is still transmitted. This is particularly important when discontinuous compression infers a false negative (i.e. spike activity is not detected even though a spike is present). Once this information is cutoff from transmission, it is lost, providing less useful data for the EEG interpreter. The response of both methods in the presence and absence of false negatives is shown in



(a) **(b)**

Figure 11: (a) Accuracy comparisons show that discontinuous does not provide great accuracy in all regions. The adaptive method is advantageous for its wide range of accuracies. (b) False Negative scenarios show that the discontinuous method can reduce accuracy, while the shift-and-detect scheme with adaptive compression virtually eliminates the effect. Results are taken with P(S) = 0.1.

Figure 11 (b).

As it is expected, in the presence of false negatives, discontinuous compression is less accurate (larger PRD) than without false negatives. Adaptive compression without the shift-and-detect scheme reduces the PRD both with and without false negatives. On the other hand, with the shift-and-detect scheme, the accuracy is virtually unaffected by false negatives. This occurs because shift-and-detect scheme provides multiple opportunities to detect a single spike.

The compression ratios (CR) are also directly affected by P(S) and these results are plotted in Figure 12. The compression ratio is inversely proportional to the energy of the transmitter. A larger (smaller) compression ratio translates into less (more) data being transmitted, requiring a smaller (larger) operating time and energy for the transceiver.

As before, the full compression (blue/green) methods have a limited dynamic range, but this time resulting in a small variation in compression ratios. This result means that care must be taken when designing the system with full compression methods.

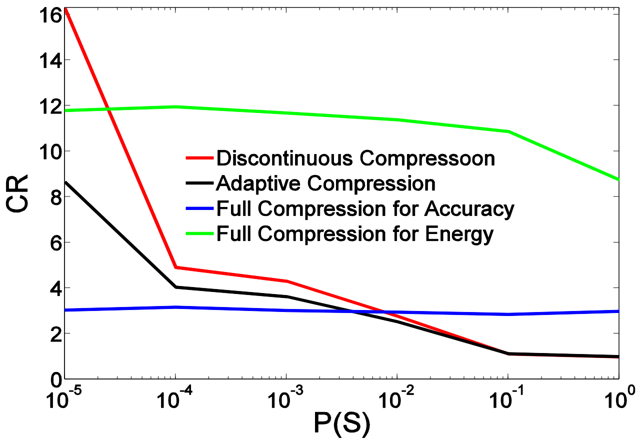


Figure 12: Compression ratios for the four compression schemes. The compression ratios are inversely proportional to the transmitter power, which is assumed to dominate the system power.

If the definition of spike dynamically, the original design choice restricts the energy-accuracy trade-off. The proposed adaptive compression system has the potential to provide a larger compression ratio during background EEGs, and the most accurate signals during epileptic activity.

2.5 System Parameters and Design Aspects

In this section system parameters will be examined and their affect on the design of the adaptive compression system will be analyzed. In most cases, there are no direct equations to determine how each parameter impacts the accuracy or compression ratio. As a result, the use of computer-aided system simulations is performed to allow a wide range of analysis.

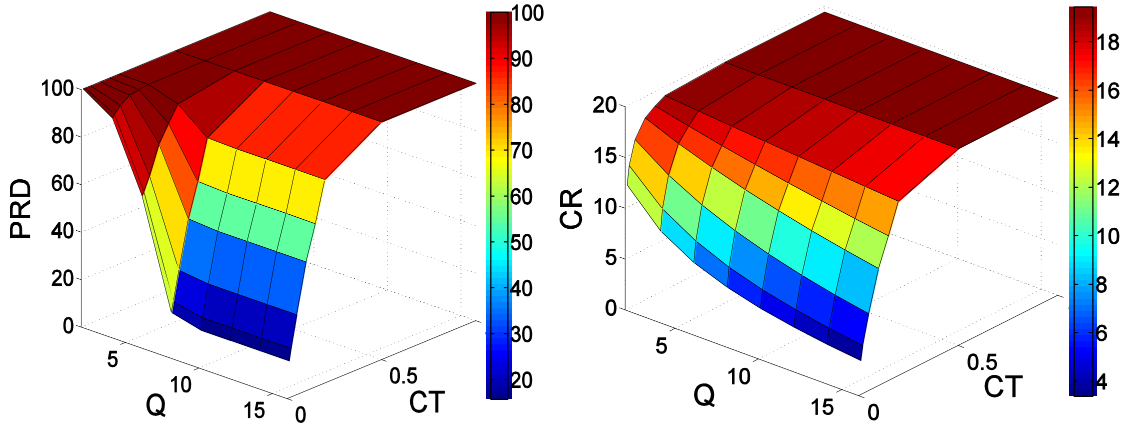
2.5.1 Quantization

As data is sampled by the EEG headset, it can be digitally represented by using as many as 16 bits. The quantization (Q) is an important parameter of the adaptive compression system that is related to how many bits are used to represent the data. This factor is only important during background EEG’s because it provides an additional level of compression. One level of compression (as described above) is achieved by thresholding the wavelet coefficients. By reducing the Q to a factor less than 16, a secondary source of data compression added.

The affect that Q has on the accuracy (PRD) and compression ratio (CR) for EEG samples (with no epileptic activity) is shown in Figure 13. The normalized compression threshold (CT), has also been plotted, and will be discussed in the next subsection. For a given CT, increasing Q will make the signal reconstruction more accurate, because using more bits will create a smaller level of precision. Since more bits are being used, the compression ratio will reduce with increasing Q. To select a quantization factor that will work optimally in the system, the smallest Q (for a higher compression) is desired, that will provide enough accuracy. From the Figure 13, selecting Q = 8 is the best candidate that will provide the same accuracy of Q = 16, but with more compression. For Q < 8, the PRD is much worse and unacceptable for accurate representation.

2.5.2 Compression Threshold

The normalized compression threshold (CT) is another system parameter that can impact system performance. A larger CT means that more data is compressed, while a smaller CT means that less data is compressed. Similar to the quantization factor, the CT is only pertinent during background EEGs, when compression is performed. Referring back to Figure 13, the normalized compression threshold alters the PRD and CR. As expected, a higher CT will increase the compression ratio, while at the same time reducing the quality of reconstructed signal. When the CT = 0, the PRD $\neq$ 0 due to



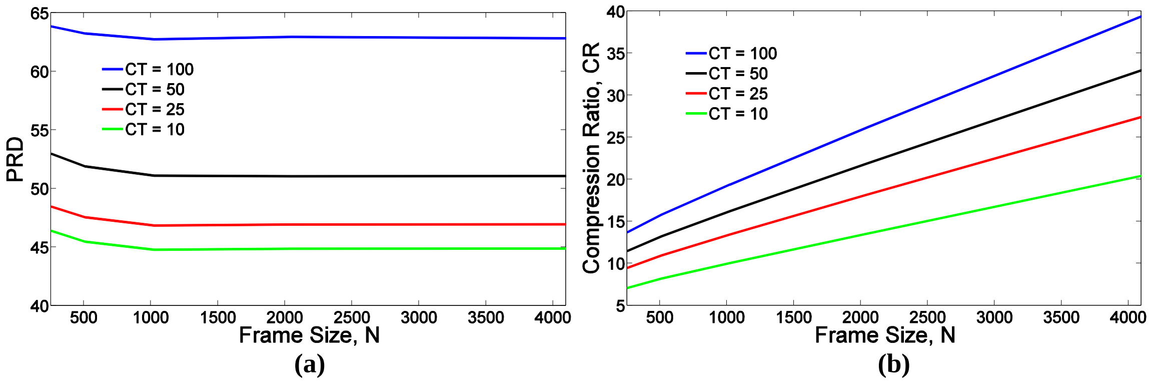
(a) **(b)**

Figure 13: Computer aided simulation used to determine the quantization factor, Q, and compression threshold, CT. The PRD is a measure of the reconstruction accuracy with PRD = 0 denoting ideal reconstruction. The compression ratio (CR) is inversely related to the accuracy.

quantization errors. The CT can be tweaked to achieve a desired compression and accuracy in background region.

2.5.3 Frame Size

The frame size (N) of the adaptive compression system is directly correlated with the size of the hardware, and also affects the accuracy and compression ratio of the reconstructed signal. A larger frame size will translate into more storage (latches and flip-flops) which naturally increases the hardware power. By increasing N, the computational energy will increase as well. To understand how the frame size can alter the reconstruction accuracy, the PRD is compared for different compression thresholds in Figure 14 (a). In this figure, the analysis was performed on background EEG data. Each curve represents a different compression ratio that is directly proportional to CT. The horizontal lines denote that accuracy has little or no dependence on frame size. This occurs because the background EEG may have periodic harmonics, which can be represented as a fixed amount of wavelet coefficients, independent of the frame size. This means that as N increases, the number of bits to represent the background data



(a) **(b)**

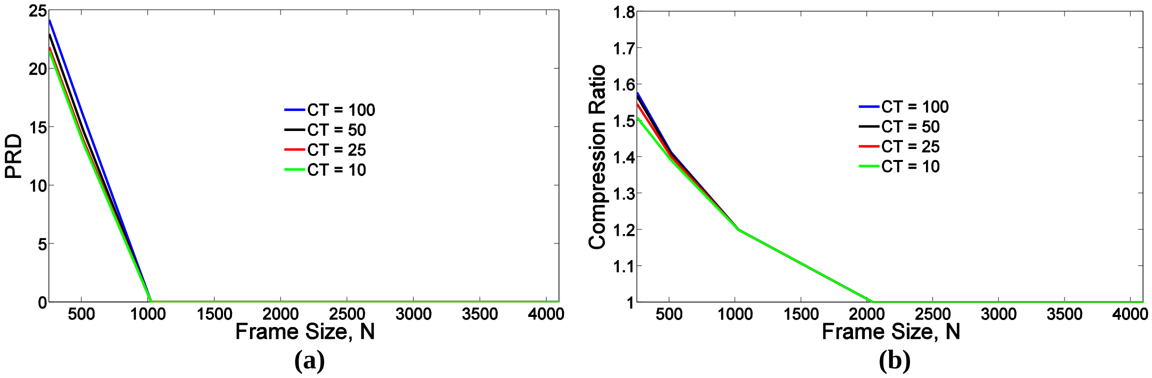
Figure 14: Results of frame size dependence for a background EEG Signal. The accuracy has little or no dependence on the frame size, N. The compression ratio increases because a fixed number of co-efficients can represent the background activity, independent of the frame size.

remains the same or increases at a much slower rate. This is why there is almost no change in PRD. For the same background EEG data, the compression ratio is compared for a varying frame size in Figure 14 (b). The trend seen is that a larger frame size will induce a large compression ratio. This behavior is consistent with a constant number (or marginal increase) of wavelet co-efficients being transmitted, even as the frame size increases.

The PRD (a) and compression ratio (b) dependence for a signal with epileptic activity is shown in Figure 15. As the frame size increases, the accuracy increases as well. In the limit that N goes to infinity, the entire signal would be sent as uncompressed data, because a spike is detected in the frame. Additionally, the compression ratio would decrease and approach one based on the same reasoning. When N is small, some frames will be compressed, causing the PRD to be large and compression ratio greater than one.

2.5.4 Spike Threshold

The spike threshold (ST) is a determining factor for spikes and can be used in all three compression methods previously described. In general, the ST should accurately



(a) **(b)**

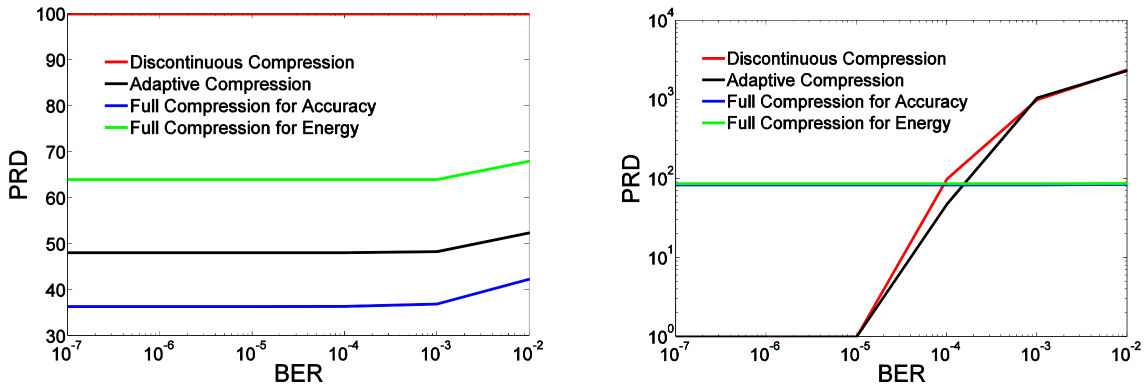
Figure 15: Results of frame size dependence for an EEG signal with epileptic activity. The accuracy improves in the limit that N goes to infinity, the entire signal would be sent as real data points. As a result, the compression ratio decreases to 1.

predict the spikes, but should be medically assigned by what a doctor defines as epileptic activity for the patient. It is not easy to determine what the spike threshold is without a given EEG history. These values can change from patient to patient, and there are ways to adaptively determine these based on a patients' history [33]. The adaptive compression system presented assumes that the spike threshold can be correctly assigned, showing ideal operating conditions. Incorrectly assigning the spike threshold can result in false detections where spikes go undetected (false negatives), or background data is detected as a spike (false positives).

2.5.5 Channel Noise Considerations

To obtain a realistic accuracy scenario, the effect of channel noise on the accuracy of the transmitted signal is considered. For simplicity, the four data compression schemes were simulated using a binary symmetric channel (BSC) model. In a BSC model, the transmitter wishes to send a bit (zero or one), and the receiver receives a bit. It is assumed that the bit is usually transmitted correctly, but that it will be flipped with a small probability. When the corresponding bits are transmitted, the bit error rate (BER) determines whether the received bits are correctly or incorrectly recovered.

The accuracy (PRD) of a background EEG is depicted in Figure 16 (a), when the error in the channel is varied. For all schemes, the accuracy is nearly constant even for severe BER of up to 10^{-2}. The constant behavior results because the data being transmitted is compressed, and the smaller packets have little or no effect on the error bits. The BER impact on the accuracy of an epileptic signal is plotted in Figure 16 (b). The accuracy degrades significantly worse during the discontinuous and adaptive



(a) **(b)**

Figure 16: (a) Channel condition simulation for background EEG sample. As the bit error rate in the channel increases, the accuracy is unaffected, because compressed data is being transmitted. (b) Channel condition simulation for EEG samples containing epileptic activity. The adaptive compression method is the best choice as long as the channel error is less than 0.0001.

compression schemes. This occurs because coefficients are being directly transmitted, instead of compressed. Since the data is uncompressed, the amount of data transmitted is larger and therefore, the chance of having an error in the transmitted packet also increases. As long as the channel condition is insured to have a BER < 10^{-4}, the adaptive compression method will always be more accurate.

In an effort to alleviate the BER restriction, a method has been proposed for further investigation: switch the transmission modes when the channel becomes severe. From the above results, when the BER approaches 10^{-4}, the adaptive system should switch and transmit in full compression mode.

2.6 Modeling and Designing for Energy Efficiency

2.6.1 Design Goals and Constraints

The purpose of data compression is to reduce the volume of data presented to the transceiver, which will reduce the operating time and energy of the transceiver. The system energy is composed of the energy in the core processor (EEG Amplifiers, ADC, and SPU) and the transceiver. For moderate data rates (~2 Mbps) the system energy is dominated by the transceiver components. Better transceiver energy efficiency is achieved by transmitting data at the highest possible data rate and putting the system into idle (sleep) mode for longer duration (duty cycle control) [18].

An ideal transition cycle for the adaptive compression system is depicted in Figure 17. During idle times, the contributions of power are only from the core processors (P_{CORE}). When the signal processing for a frame has completed, the transceiver switches on and transmits the packets. Once transmission is complete, the system returns to the sleep state. Since the SPU will operate at a low frequency (~ 100's of Hz) and the transceiver will operate at a maximum data rate (~ 2Mbps), the transceiver will be in the idle state for a majority of the cycle. With the assumption that the transmission power, P_{TX}, is much larger than the core processor power, P_{CORE}, it is the designers goal to have $t_{SLEEP} >> t_{TX}$. In reality, there are also components of wakeup power (sleep to transmit transition) that need to be considered. For this analysis, the transmit power has been overestimated to compensate for this neglect. Lastly, by

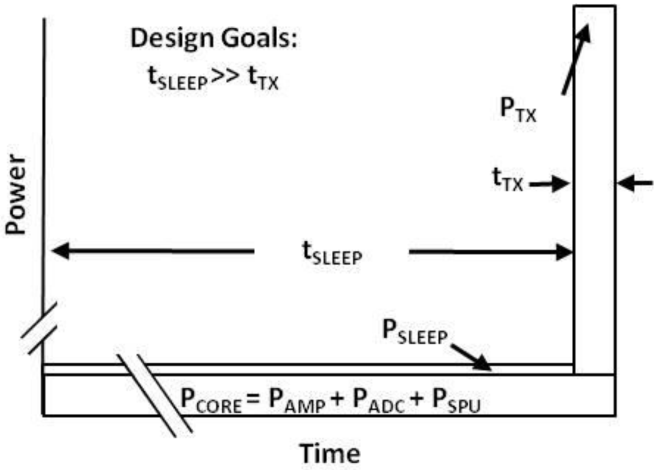


Figure 17: Power and Time Design Goals to Minimize Energy

selecting an existing transmitter/receiver pair (NRF24L01), the issues of synchronization can be avoided resulting in a continuous signal stream.

To design for energy efficiency, it is essential to develop a metric for comparison that is applicable across the various system design parameters: frame size (N), compression ratio (CR) and quantization (Q). These parameters are knobs that the designer can use to achieve a desired accuracy requirement. The impact of these system parameters on the system energy needs to be realized. Before the metric of energy efficiency can be addressed, the system energy components must be examined. The total energy per frame computed, E_{TOTAL} is defined as

$$E_{TOTAL} = E_{CORE} + E_{TCVR}. \qquad \text{Eq. (2.2)}$$

The energy of the core processor (E_{CORE}) is formed from the energy of the EEG amplifiers (E_{AMP}), analog-to-digital conversions (E_{ADC}), and signal processing units (E_{SPU}).

$$E_{CORE} = E_{AMP} + E_{ADC} + E_{SPU} \qquad \text{Eq. (2.3)}$$

The transceiver energy, E_{TCVR}, was previously described as having a dependence on the sleep time (t_{SLEEP}) and the transmission time (t_{TX}). The transmission time is also defined based on the data rate of the transmitter (r), the length of the input packet (L_{PACKET}) and the compression ratio (CR).

$$E_{TCVR} = P_{TX}t_{TX} + P_{SLEEP}t_{SLEEP} \qquad \text{Eq. (2.4)}$$

$$t_{TX} = L_{PACKET}/(r \cdot CR) \qquad \text{Eq. (2.5)}$$

A higher CR reduces the transmission time and allows the system to be in the sleep mode for longer time thereby reducing the energy consumed. The L_{PACKET} is essentially the number of bits that will be processed during an entire frame and is a function of the quantization (Q) and the frame size (N).

$$L_{PACKET} = N \cdot Q \qquad \text{Eq. (2.6)}$$

Designers can understand the entire energy associated with EEG signal acquisition, data compression and wireless transmission by referring to (2.2) to (2.6). Using (2.2) to obtain energy comparisons for a system with different compression ratios is valid, because the energy will scale accordingly. It is not valid to use (2.2) for comparisons with different input packet lengths. As L_{PACKET} increases, the system energy increases, because there is a larger volume of data being operated on. To normalize this effect, the energy per bit will be used to determine the energy efficiency in this system. By using the results of (2.2) and (2.6) the energy per bit is

$$\frac{Energy}{bit} = \frac{E_{TOTAL}}{L_{PACKET}}. \qquad \text{Eq. (2.7)}$$

This energy efficiency metric is the energy associated with the acquisition, conversion, data compression, and transmission of one bit. In an energy efficient system, (2.7) will be minimized. Similar equations that relate the energy per bit of acquisition, conversion, compression and transmission can all be independently found. One cannot assume that maximizing the energy efficiency (i.e. minimizing equation (2.7)) of the four individual components will result in overall energy efficiency, due to the complexity of this system.

2.6.2 Modeling Energy Components

The focus of this chapter has now shifted to designing an energy efficient system for a wireless EEG system that employs adaptive compression techniques. To achieve this, a SPU has been designed that can operate with commercial-off-the-shelf (COTS) components. In this subsection, a method to estimate the energy in the SPU will be derived. Additionally, the impact of various system variables will be analyzed to achieve energy efficiency. The energy of the EEG amplifiers, analog-to-digital converter and wireless transceiver are modeled using reported (i.e. data from measurements) results.

2.6.2.1 EEG Signal Amplifiers, ADC and Transceiver

The EEG signal amplifiers and analog-to-digital conversion models were based on an ultra low power implementation based on the work of Verma [8]. The amplifier was able to operate at a 1.0V supply voltage while delivering 3.5uW of power. The analog-to-digital converter operates at the same supply voltage, but can achieve 250pJ per sample conversion. To stay with the mindset of an ultra low power scheme, the transceiver energy model is based on the Nordic NRF24L01+ 2.4 GHz wireless transceiver. The Nordic transceiver was selected because it has been designed for ultra low power sensor networks [34]. The NRF24L01+ transceiver can achieve 1.71 uW of power in the sleep state while delivering 21.5 mW of power at a supply voltage of 1.9V. Due to the low processing frequency of the proposed system, an optimal transceiver with a smaller carrier frequency could be custom designed to further reduce the overall power. For the purposes of this study, the Nordic transceiver will suffice. The estimated power and energy of the EEG system components is summarized in Table 1.

Table 1: Power and Energy Estimates for EEG System Components

System Component	Supply Voltage	Power	Energy
EEG Signal Amplifier [8]	1.0 V	3.5 uW	-
ADC Conversion [8]	1.0 V	-	250pJ per conversion
NRF24L01+ Transmission [34]	1.9-3.3 V	21.5 - 37.3 mW	-
NRF24L01+ Sleep [34]	1.9-3.3 V	1.71 - 2.97 uW	-
Signal Processing Unit [This Work]	1.0-0.3V	-	$n \cdot \alpha \cdot C_L \cdot V_{DD}^2 + n \cdot I_{OFF} \cdot V_{DD} \cdot t_{COMPUTE}$

2.6.2.2 Signal Processing Unit

The signal processing unit (SPU) is the computational workhorse of the proposed adaptive compression system. The SPU is the only digital component of this system, which allows one to model the energy based on the works of Calhoun and Zhai. [35, 36]:

$$E_{SPU} = E_{DYN} + E_{STA} \qquad \text{Eq. (2.8)}$$

The sum of the energy in a digital system is composed of the individual contributions of dynamic energy and static energy, as stated in (2.8). The dynamic energy is the energy associated with charging and discharging load capacitances (C_L) on the logic paths and has a square dependence on the supply voltage (V_{DD}). As one can imagine, there is also a dependence on the number of logic gates (n) as well as how frequently the logic gates switch (α).

$$E_{DYN} = n \cdot \alpha \cdot C_L \cdot {V_{DD}}^2 \qquad \text{Eq. (2.9)}$$

As the name denotes it, the dynamic energy of a digital circuit is the energy that is dissipated when circuit experiences logical transitions.

The common mathematical model for dynamic energy in any digital circuit is shown in (2.9). The activity factor (α) denotes a measure of how often the gates switch and is usually on the order of 0.1-0.2. Static energy results from small leakage currents that occur because it is impossible to completely turn off the transistors. The static energy is defined as:

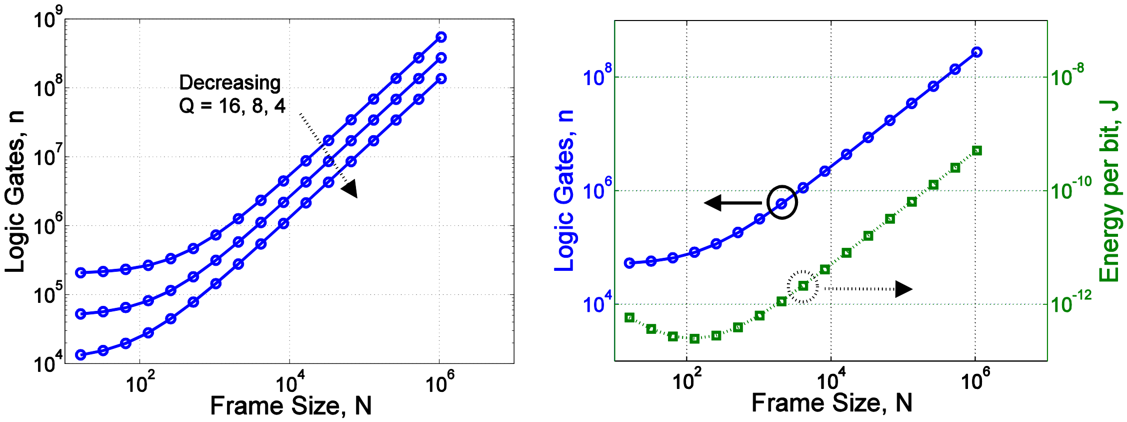
$$E_{STA} = n \cdot I_{OFF} \cdot V_{DD} \cdot t_{COMPUTE} \qquad \text{Eq. (2.10)}$$

The first three terms on the right of (2.10) determine the leakage power associated with all of the logic gates, while $t_{COMPUTE}$ converts this leakage power into the leakage energy. All logic gates exhibit static leakage energy. For a logic gate that is switching, the dynamic energy will dominate over static leakage energy. If a logic gate does not switch, the leakage energy contributes to a majority of the total energy. In order to accurately estimate the total energy associated with our SPUs, the specific technology parameters (C_L, I_{OFF}) and the number of logic gates in the design must be known. The capacitance load per logic gate (C_L) and leakage current (I_{OFF}) have been defined for a standard CMOS NAND gate in 180nm technology [37]. The approximate number of logic gates for the signal processing unit is a function of the quantization (Q) and sample frame size (N):

$$n(SPU) = 768 \cdot Q^2 - 336 \cdot Q + 33 \cdot N \cdot Q \qquad \text{Eq. (2.11)}$$

With the previous analysis, one can now examine how the system parameters impact the number of logic gates in the SPU, and energy efficiency. The models are created based on simulations in the 180nm CMOS Predictive Technology Model (PTM) with a $V_{DD} = 1.0V$, $V_{TN} = 0.39$ V and $V_{TP} = -0.42$ V. Load capacitances (C_L) and leakage currents (I_{OFF}) were also determined by estimates in this technology. The number of logic gates and the dependence on frame size is plotted in Figure 18 (a). As expected, increasing either the number of quantization (Q) bits or the frame size (N) will increase the total number of logic gates (n). The lowest power could be achieved with a Q equal to four, but based on the earlier accuracy analysis a minimum Q of eight is required to fulfill the accuracy requirement.

The accompanying energy per bit curve for Q = 8 is plotted in Figure 18 (b). There is a direct correlation with the number of logic gates and energy per bit when the frame size is large (i.e. $N > 10^3$). When $N < 10^3$, the energy per bit approaches a minimum and then increases again. This minimum denotes an energy-efficient operating



(a) **(b)**

Figure 18: (a) Estimate number of logic gates for the SPU with frame size dependence and vary quantization. (b) Comparison between estimate number of gates and energy per bit for SPU.

point in terms of processing data per bit. Recall that the sampling rate of the system is on the order of 500 Hz, meaning that within the sample time the processing on the new data point has to be performed. Since this sampling occurs at such a small frequency, one can attempt to reduce the energy by scaling the supply voltage to subthreshold voltages.

A plot of how the energy per bit changes with a decreasing supply voltage from 1.0 V to 0.3V is shown in Figure 19 (a). There is nearly a 9X reduction in energy can be saved by reducing the supply voltage to subthreshold voltages (~ 300mV). Additionally, the energy-efficient point has shifted to a smaller frame size as the supply voltage is reduced. This shift in the curve is best described by looking at Figure 19 (b), which plots the static and leakage contributions of the SPU energy per bit. As noted from Calhoun and Zhai [35, 36], when devices are scaled to subthreshold voltages, the energy contributions of the static leakage components begin to dominate over the dynamic energy. Essentially, as the voltages are scaled down the leakage curve in Figure 19 (b) will shift upwards, moving the energy-efficient point towards a smaller frame size. Stated in another fashion, as the voltages are reduced, it takes a longer time to compute,

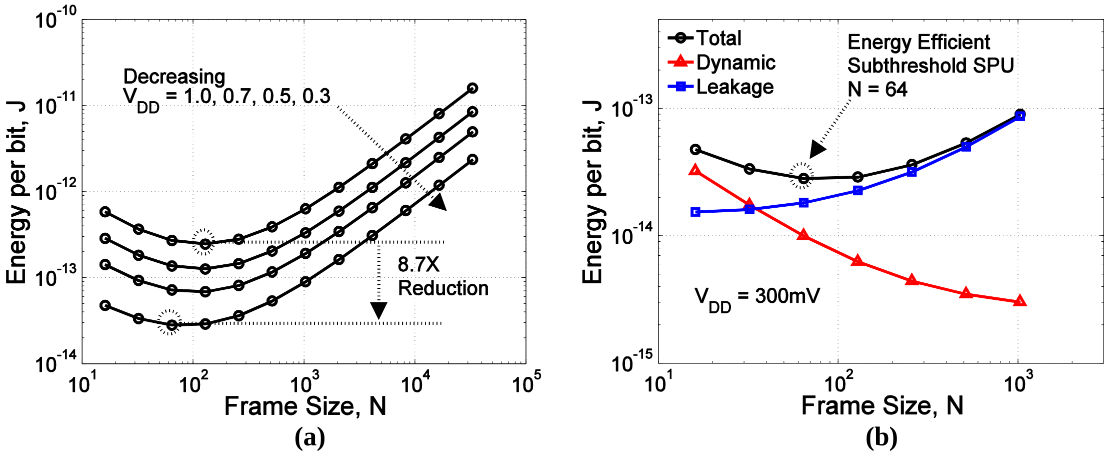


Figure 19: (a) Energy savings in the SPU achieved by reducing supply Voltage (b) Dynamic and Static leakage contributions of the SPU and the Energy Efficient operating point for the SPU.

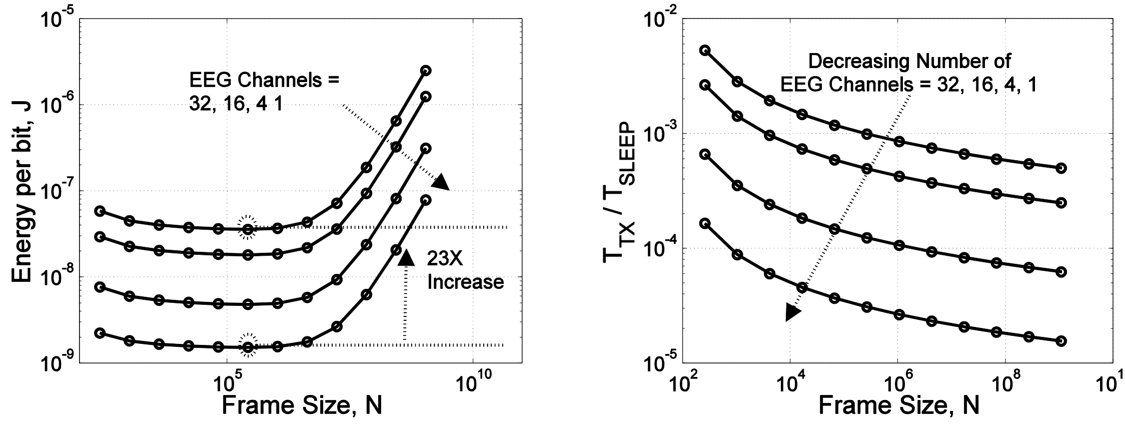
creating more time for circuits to generate static leakage energy. With this thought, it would make sense to reduce the frame size, in order to reduce the number of leaking devices. In summary, it is advantageous to scale the supply voltage because of the energy savings and low sampling frequency that allows it. Lastly, for a quantization of 8 bits, and supply voltage of 300mV, the energy-efficient frame size occurs at N = 64.

2.6.3 An Energy-Efficient, Multi-Channel, Wireless EEG System

The previous analysis of a single SPU can be extended to approach the goal of designing an energy efficient, multi-channel, wireless EEG system. In the following investigation several processing cores (EEG Amplifier, ADC, and SPU) are combined to model and understand the energy associated with a multi-channel EEG system. Since the system is operating at a very low sampling frequency (~ 500 Hz) and transmitting data at a high frequency (~2 Mbps), parallel signal acquisition and serial data transmission can be used as denoted in Figure 10. This means that as the number of cores increases, all cores will be able to share a single transceiver. For the remaining analysis the following parameters are assumed: Q = 8 and V_{DD} = 300mV. A compression threshold of 50 was selected, which creates a compression ratio that is modeled by the curve of Figure 14 (b) (CT = 50).

The wireless EEG system energy per bit is compared across our design parameter, N, and plotted in Figure 20 (a). By increasing the number of EEG channels from 1 to 32, the energy consumed increases by a factor of 23X. This appears reasonable, as one would expect at least a 32X increase in energy if parallel acquisition and parallel transmission (i.e. using multiple transceivers) were performed. The system design goal of transmission and sleep times is analyzed in Figure 20 (b). For a smaller number of channels, $T_{TX} << T_{SLEEP}$, as desired, but as the number of channels increases, the ratio increases. This occurs because for a given frame size, the time to compute is constant and independent of the number of channels. However, as the number of channels is increased, the transmission time increases because one transceiver is being used. For the 32 channel implementation, the design goal is continually met (i.e. $T_{TX}/T_{SLEEP} \gg 0.1$), but this may be of concern for systems with larger channels.

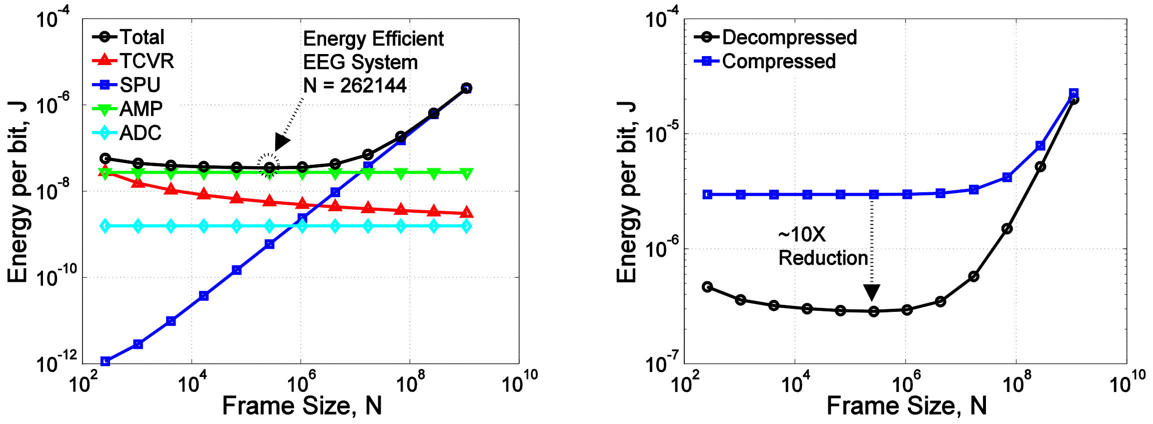
The remaining results and analysis focuses on a design that includes 32 channels.



(a) **(b)**

Figure 20: (a) Energy increase for the complete Wireless EEG System as the number of EEG Channels is increased. (b) Design goals met across all channels to target an Energy Efficient Wireless EEG System.

The system components are analyzed and the energy per bit is compared in Figure 21 (a). The first thing to notice is that there exists a minimum energy per bit point which is defined as the energy-efficient point for the 32 channel wireless EEG system. In this case, the energy-efficient condition occurs at N = 262,144. Recall earlier that the energy-efficient point for the SPU occurred at N = 64. From the results, this shows that to determine the minimum energy per bit point for an entire system, it is not valid to simply determine minimum energy per bit for the individual components. Another thing to note is the EEG signal amplifiers (AMP) and the ADC energies are constant. The constant energy occurs because these components have no dependence on the frame size. The transceiver, however, reduces its energy per bit as the frame size increases. This reduction in energy per bit is aligned with Wang [21], which states that the maximum energy efficiency occurs at a maximum data rate. As the frame size increases, more compression cab be achieved based on the curves in Figure 14 (b). Additionally, for a large N, the SPU energy dominates, and this is a result of the large volumes of data that require processing to be performed on them.



(a) **(b)**

Figure 21: (a) Energy component breakdowns for EEG Signal Amplifier, ADC, Signal Processing Units and Transceiver. (b) Potential energy savings when using an energy-efficient design methodology that employs adaptive compression techniques.

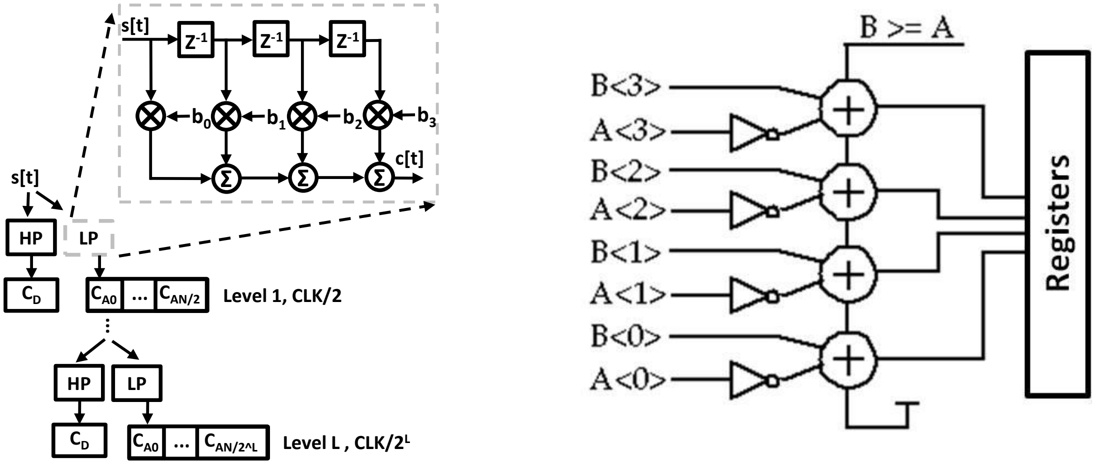
In the final experiment (Figure 21 (b)), the potential energy savings are determined by using adaptive compression and designing for an energy-efficient system. At the energy-efficient point, one can achieve a 10X savings in power when the data is compressed, compared to when the data is decompressed. In the presence of background activity, there is a 10X reduction in power when compared epileptic activity.

2.7 Discussion

2.7.1 Discrete Wavelet Transform and Wavelet Compression Blocks

The discrete wavelet transform is implemented as a filter bank of low-pass and high-pass filters. The implementation of the DWT is shown in Figure 22 (a), 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) co-efficients. 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 downsampling, the co-efficients are latched at half the frequency of the sampled signal s[t]. This downsampling effectively reduces the number of latches needed to represent the co-efficients. At each DWT level, the low pass co-efficients are further decomposed until the desired L^{TH} level is reached. From Indiradevi [31], the optimal number of levels for EEG decomposition to facilitate spike detection is six. In the end, an N sample signal will produce N co-efficients, resulting from all high-pass coefficients and the last level low-pass co-efficients.

Wavelet compression (WC) is performed based on a thresholding technique, where the wavelet coefficients are compared against a compression threshold (CT). If the



(a) **(b)**

Figure 22: (a) The Discrete Wavelet Transform (DWT) logical schematic. (b) The Wavelet Compression (WC) system level representation.

co-efficients are greater than the threshold, they are preserved; otherwise the coefficients are zeroed out and will not be transmitted. This logic can be performed by using an adder in subtract mode. If the difference between the wavelet coefficient and the CT is positive, the coefficient is discarded. If the result is negative or zero, we can preserve the wavelet coefficients. A four bit comparator is shown in Figure 22 (b).

2.7.2 Estimating the Number of Logic Gates

This section derives the model used to estimate the number of logic gates associated with the signal processing units (SPUs). From the design in Figure 7 (a), the majority of the logic is performed in the discrete wavelet transform (DWT), which is composed of several FIR Filters. For simplicity, each SPU is assumed composed of the discrete wavelet transform, wavelet compression, and the resulting registers that will transfer the logic. Additionally, all logic blocks are formed using two-input NAND gates, and all system blocks (e.g. multipliers, adders, etc.) are created from basic implementations described in Rabaey [38]. In general, this approximate number of logic gates is an overestimate, since circuit, logic and system level optimizations have not been

Table 2: Summary of NAND Logic Gate Models

CMOS Gate	**n(X)**	**Equivalent Logic**	**# Gates**
Nand	n(NAND)	---------------	= 1
Inverter	n(INV)	= n(NAND)	= 1
And	n(AND)	= n(NAND) + n(INV)	= 2
Or	n(OR)	= n(NAND) +2· n(INV)	= 3
Xor	n(XOR)	= 4·n(NAND)	= 4
D-Flip Flop	n(FF)	= 6·n(NAND)	= 6
Q-bit Register	n(Q-FF)	= Q·n(FF)	= 6·Q
Half Adder	n(HA)	= n(AND) + n(XOR)	= 6
Full Adder	n(FA)	= 2·n(XOR) + 2·n(AND) + n(OR)	= 15
Q-bit Full Adder	n(Q-FA)	= Q·n(FA)	= 15·Q
Q-bit Multiplier	n(MU)	= Q·n(HA) + $((Q-1)^2-1)$·n(FA) +Q^2·n(AND)	= 16·Q^2 - 22·Q
Q-bit FIR Filter	n(FIR)	= 3·n(Q-FF) + 4·n(MU) + 3·n(Q-FA)	= 64·Q^2 -28·Q
DWT	n(DWT)	= 12·n(FIR) + 2·N·n(Q-FF)	=768·Q^2-336·Q +12·N·Q
WC	n(WC)	= N·(n(Q-FA) +Q· n(INV) + n(Q-FF))	= 22·N·Q
SPU	n(SPU)	= n(WC) + n(DWT)	=768·Q^2-336·Q +34·N·Q

performed. This worst case number of logic gates will provide an overestimate of energy consumed by each SPU Core, which are still significantly less than the transmission power.

Based on the above description and Figure 22, the DWT is composed of flip-flops, adders and multipliers. The approximate number of gates is derived with a bottom up approach, beginning with a standard Flip Flop. In general, the logic gate equivalent translations between standard CMOS gates and NAND Logic are shown in Table 2. The DWT is formed by creating 12 FIR Filters and a 2N (2*Frame Size) Q-bit registers for all of the filter banks. The resulting estimate number of gates is n(DWT). Similarly, to estimate the number of logic gates in the wavelet compression (WC) a Q-bit Full Adders, Q inverters and Q flip flops will be needed to compare for one sample. Since N samples are used, the multiplicative factor has been introduced and the resulting equation n(WC) is shown in Table 2. The most important result is the estimate number of the SPU which is composed of the DWT and WC gates:

$$n(SPU) = n(DWT) + n(WC) = 768 \cdot Q^2 - 336 \cdot Q + 33 \cdot N \cdot Q \tag{2.12}$$

2.8 Summary

A data dependent adaptive compression approach and associated system design for power reduction in a wireless EEG system has been presented. The adaptive compression scheme provides very accurate data transmission during epileptic activity and low power (but less accurate) transmission during background EEGs. A shift-and-detect scheme for spike detection was introduced, and it increased the detection accuracy by reducing the impact of false negatives in spike detection. By presenting a custom digital system approach to EEG processing, one can take advantage of parallelism and high throughput when complex algorithms are implemented [29-31]. The effect of important system parameters were also analyzed, which will aid in the design of an optimal adaptive compression system. An energy model was also developed to determine the energy efficient point, defined by selecting the appropriate frame size. When combining the adaptive compression method and the energy efficient design methodology, a prediction of 10X energy savings can be achieved. The preceding analysis gives encouraging results into digital EEG systems and the use of adaptive compression as well. The proposed system could form a generic basis for remote wireless monitoring of other physiological signals.

CHAPTER 3

CHAMELEON: A CONTENT-AWARE AWARE ADAPTIVE ARCHITECTURE FOR LOW POWER WIRELESS ELECTROENCEPHALOGRAPHY

3.1 Introduction

In this chapter, an algorithmic and architectural design of an embedded hardware system for low power wireless EEG monitoring is presented. Based on the algorithm presented in Chapter 2, the proposed architecture accurately transmits epileptic spikes by employing adaptive compression methods. A custom architecture has been optimized for low power in the low-throughput domain that performs signal detection, compression and transmission. To reduce the overall power, the digital processor has been designed to consider the system level impact of other devices, such as the wireless transceiver.

It is well known that the majority of the system power in a wireless system is due to the power required to operate the transceiver [18]. By compressing (lossy) the data when no epileptic potentials exist, the transceiver can be operated for a shorter amount of time and save power [21]. The algorithm implemented for detection characterizes the signal in real-time and determines whether compression is necessary. By utilizing an algorithm that considers power and accuracy, both energy-efficiency and signal quality are maintained. The architectural three stage pipeline is intended to be used in an embedded hardware based wireless electroencephalography monitoring scheme. The embedded digital hardware receives the EEG signal from an analog-to-digital converter

(ADC), performs the accuracy and energy-aware compression, and controls the data flow to the RF transceiver.

The full system was implemented using a Xilinx Virtex 5 FPGA and a Nordic RF transceiver, and the resulting implications show that methodology is a viable option to design an ultra-low power ASIC for wireless EEG processing. For accurate power estimations, a standard ASIC design flow was used to design, and simulate a post-silicon layout of the processor. All hardware experiments were performed using medically recorded EEG data made available from a publicly available online database [32]. This chapter makes the following contributions:

1. An algorithmic and architectural system design to address the signal integrity and hardware power.
2. An FPGA and transceiver hardware implementation to validate the design and measure energy efficiency.
3. An ASIC methodology that implements low power optimizations to reduce the system level power of the wireless EEG system.

Overall goal of this chapter is to present the design methodology, signal measurements and low power customizations involved in implementing an adaptive compression system, called Chameleon, for wireless EEG monitoring. As a result of this case study, it will be shown that for monitoring of signals with distinct regions, adaptive data compression represents a viable option to achieve accuracy and energy goals in all regions.

3.2 Content-Aware 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). In this section, the developed algorithm to adaptively compress the EEG signals using real-time detections will be reintroduced.

A human EEG signal taken from a patient who suffers from epileptic seizures is shown in Figure 23 [32]. Although not visible in this figure, EEG signals are low frequency signals (0-100Hz) and their amplitudes range from 10 to 100 micro volts when sampled from the scalp. The signal can be classified 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 and classifications. For patients who have epilepsy, background EEGs occur more often than not, making it difficult to understand and isolate epileptic behavior.

To detect and compress EEG signals, Mallats' theory of multi-level resolution will be

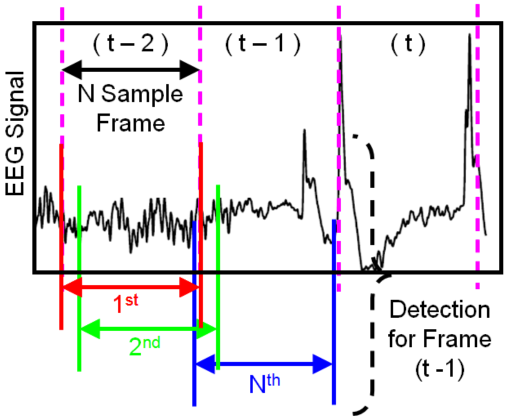


Figure 23: An EEG Signal can be defined by at least two distinct regions: background and epileptic (spike) activity. In the algorithm for spike detection, a frame size of N-samples will involve N overlapping detections.

used to decompose the EEG signal into separate frequency bands [28]. The wavelet transform was selected to perform this function, because it is appropriate for non-stationary, aperiodic signals (i.e. EEG signals). The Daubechie-2 (db-2) was used as the mother wavelet.

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). The concept of wavelet thresholding is simple to understand, and it is applied in two fashions. For spike detection, a set of wavelet co-efficients are compared across a pre-determined *spike* threshold. If any coefficients surpass this threshold, an epileptic action potential has been detected. If none of the coefficients are greater than the spike threshold, the data is assumed to be background activity. This and similar methods of detection have been investigated and shown to produce greater than 90.5% accuracy, 91.7 % sensitivity and 89.3 % specificity [31]. To compress the data, the coefficients are compared across a different pre-determined *compression* threshold. The coefficients that do not surpass this threshold are discarded. This lossy compression has historically shown appreciable compression rates with reasonable signal reconstruction accuracy [39]. For a time reference, the signal is divide into frames, which are define as N cycles (see Figure 23). 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 24.

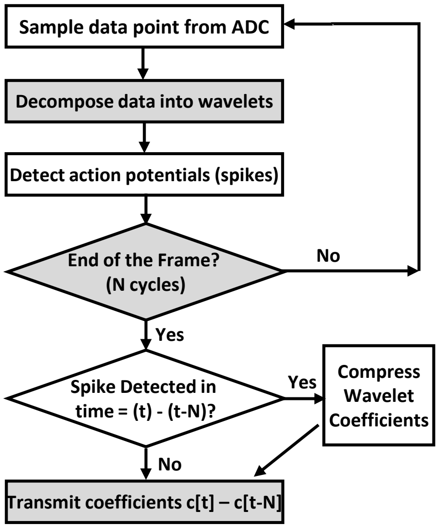

Sample data point from ADC
Decompose data into wavelets
Detect action potentials (spikes)
End of the Frame? (N cycles)
No
Yes
Spike Detected in time = (t) - (t-N)?
Yes
Compress Wavelet Coefficients
No
Transmit coefficients c[t] – c[t-N]

Figure 24: Adaptive algorithm for spike detection and data compression of EEG signals.

3.3 Low Power Methodology

In the previous section, the adaptive compression algorithm of the Chameleon system was presented as a method to reduce the volume of data transmitted. By reducing the volume of the data transmitted, the transmission energy is reduced compared to when data is not compressed.

In this section, a low power design methodology that will reduce the system energy by using SLEEP mode power cycles will be presented. To perform this analysis, the transceiver and digital processor energy will be the focus of this analysis.

The design goals of the Chameleon system are shown in Figure 25. There are several power components that need to be considered: P_{TX}, the power of the transceiver during transmission; P_{SLEEP}, the power of the transceiver during the idle times; P_{STBY}, the power of the transceiver in between transmissions; and P_{DSP}, the power of the digital embedded hardware.

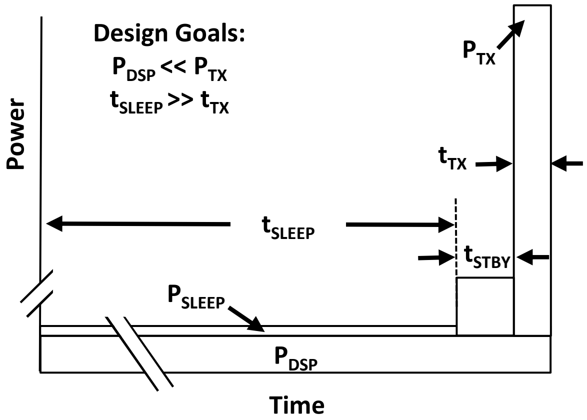


Figure 25: Power and time goals to minimize the overall system power of the Chameleon system. The transceiver power is assumed to be dominant; therefore, the main objective is to minimize its operating time.

When considering the hardware in an low power wireless system, the assumption is that the power in the transceiver, P_{TCVR}, is much greater than the power involved in the signal processing, P_{DSP}. For most ASIC wireless solutions, this assumption is fair and [18] supports this justification. This power ratio is also a function of the transceiver transmission range, and the amount of on-chip processing of the digital system. When designing the Chameleon system, the goal is to counterbalance the assumption that $P_{DSP} \ll P_{TX}$, by making $T_{SLEEP} \gg T_{TX}$. In doing this, the goal in mind is to reduce the power of the system overall. This can be achieved by operating the transceiver at the highest data rate, which also ensures maximum power savings [21]. Additionally, since the EEG signals are contained at very low frequencies (i.e. in the range of Hertz), digital sampling and processing can in that range, and much slower than the high frequency data rates of the transceiver (i.e. in the range of Megahertz). This disparity in time will also ensure that the transceiver will spend most of its time in the lowest power mode (SLEEP).

The disparity between the processing and transmitting frequencies is also dependant on the Frame Size, N. Referring back to Figure 23, the Frame Size is the data window width in which the on-chip processing will occur. For a small Frame Size, one can expect a smaller time in between transmissions. Similarly, when the Frame Size is large, there will be a longer time in between transmissions. In reference to Figure 25, the window cycle time, t_{CYCLE}, is equal to the sum of the sleep, standby and transmission times.

Based on the previous discussed relationships between the Frame Size, processing times, transmission rates, reducing the power is an intricate task. Additionally, simply minimizing the power of each individual component may not necessarily result in minimum power consumption, because of the dependence on time. For example, a minimum transceiver transmission power can be achieved, but if the transmit time is large, it can nullify the low power goal.

In general and from Figure 25, the total power to process and transmit cycle is approximately:

$$P_{TOTAL} = P_{DSP} + P_{TCVR} \qquad \text{Eq. (4.1)}$$

$$P_{TCVR} = \left(P_{TX} t_{TX} + P_{SLEEP} t_{SLEEP} + P_{STBY} t_{STBY}\right) / t_{CYCLE} \qquad \text{Eq. (4.2)}$$

$$t_{CYCLE} = t_{TX} + t_{SLEEP} + t_{STBY} \qquad \text{Eq. (4.3)}$$

$$t_{TX} = L_{PACKET} / r \cdot CR \qquad \text{Eq. (4.4)}$$

where the transmit time, t_{TX} is a function of the compression ratio (CR), the length of the data stream being processed (L_{PACKET}), and data rate (r). The standby time, t_{STBY}, occurs as a result of finite transceiver power-on times and are determined by the manufacturer. As previously stated, the cycle time, and indirectly the sleep time are both dependent on the Frame Size of the system. As the processing window increases, the cycle time is increased because of the low frequency (Hertz) sampling of the EEG signals. As a result of high frequency (Megahertz) processing, the low sampling frequency is the time bottle neck, which restricts the cycle time of the system.

3.4 Embedded Hardware Architecture

The algorithm for processing EEG signals requires three primary steps: Signal Detection, Compression, and Transmission. This section presents architectural considerations for each of these operations. The three stage pipelined architecture for the primary operations are depicted as a dotted box in Figure 26.

In the first pipeline stage, detection occurs as the signal sampled from the analog-to-digital converter (ADC) and sent to the Discrete Wavelet Transform block. Recall that the discrete wavelet transform block decomposes the signal into wavelet coefficients that can be used for detection and compression.

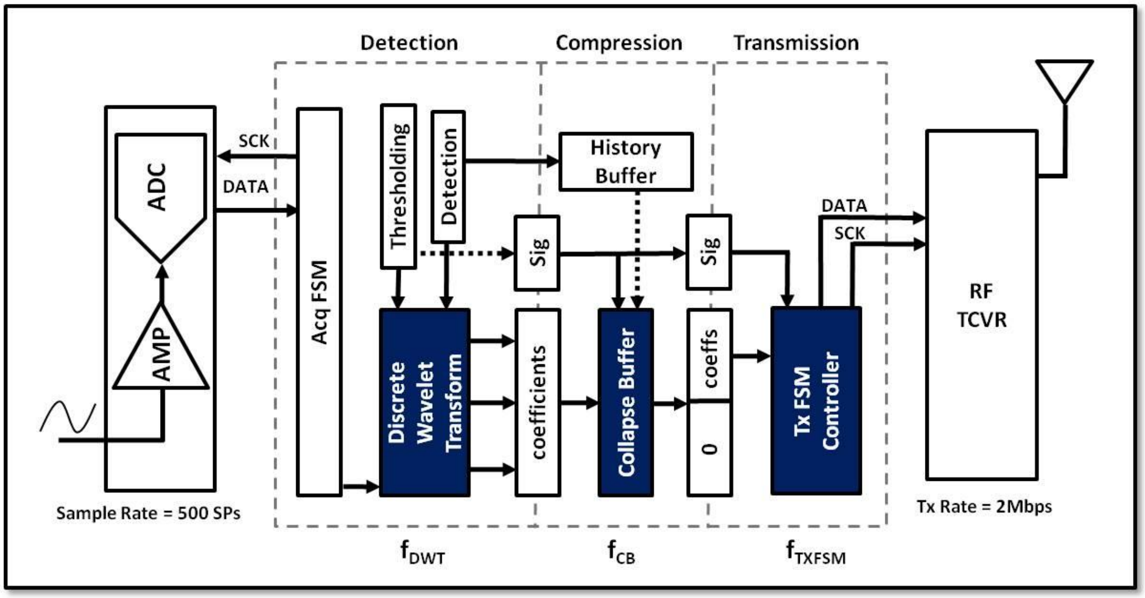


Figure 26: Architectural depiction of the three-stage pipeline (Detection, Compression, Transmission) for adaptive EEG processing.

Detection is performed on the 4th and 5th level co-efficients as noted in [31]. 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 co-efficient 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 co-efficients 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.

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.

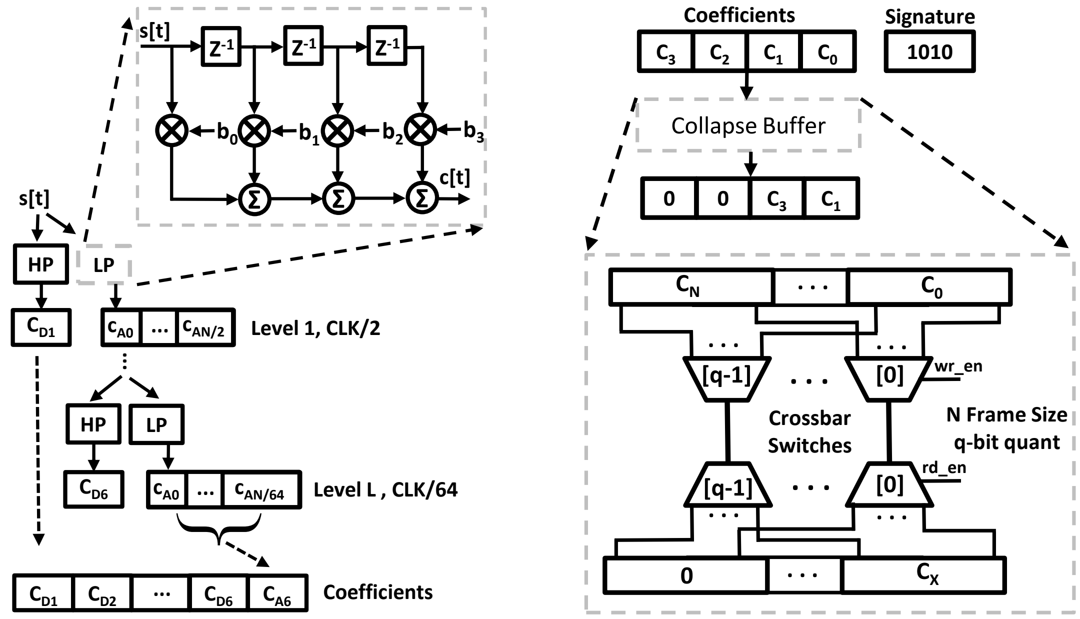
3.4.1 Acquisition Finite State Machine

The acquisition finite state machine is designed to sample data points from the analog-to-digital converter at a rate of 500 samples per second (SPs). Electroencephalography signals are well contained below 250 hertz. By sampling at 500 hertz, the integrity of the signal is preserved based on Shannon's sampling theorem. Since most commercial ADC's do not have a sampling rate this low, the finite state machine has been designed to facilitate this rate. The acquisition finite state machine has two states: ACQUIRE and SLEEP. During the ACQUIRE phase, the controller samples the analog signal and shifts in the 12-bit representation into the storage register. After the acquisition is complete, the sample is presented to the discrete wavelet transform, and the finite state machine moves to the SLEEP state. During the SLEEP state, the ADC is placed in a low power state for several cycles, until the next sample is ready to be taken.

3.4.2 Discrete Wavelet Transform

The discrete wavelet transform (DWT) is implemented as a filter bank of low pass and high pass filters. The DWT implementation is shown in Figure 27 (a), 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 [31], 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.



(a) (b)

Figure 27: (a) Spike detection is performed by using a 6-level discrete wavelet transform filter bank. (b) The collapse buffer zeros the coefficients below the compression threshold and collapses the relevant data to the least significant bits.

3.4.3 Collapse Buffer

The collapse buffer was implemented as a crossbar switch network to facilitate multiple input, multiple output connections. The write enable is controlled by the generated signature, while the read enable is only updated if a write has occurred. The serialized process performs reads and writes each cycle and is directly proportional to the length of the sample. Since the collapse buffer is serialized and the discrete wavelet transform is parallelized, the collapse buffer should operate at a frequency faster than the discrete wavelet transform.

The overall operation can be explained with a simple example shown in Figure 27 (b). 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 co-efficients 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 co-efficients.

3.4.4 Transmission Controller

In the transmission control block, the co-efficients are padded with an address, packet id and header to create a payload. For this case study, the transmission protocol has been designed to operate with a commercial low power transceiver, the Nordic nRF24L01+ [34]. 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 the state diagram is depicted in Figure 28. At the 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 (TX_DONE), 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

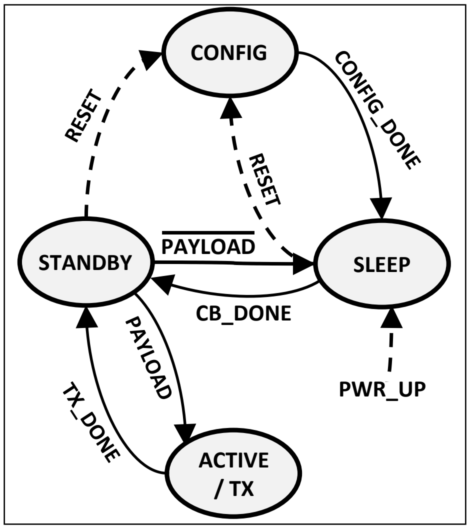


Figure 28: The transmission controller was implemented as a finite state machine where the number of transmissions depends on the length of the payload.

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.

3.4.5 Globally Asynchronous Locally Synchronous (GALS) Clocking Scheme

Recall that in the earlier discussion of EEG signals, the highest frequency component is in the order of 100's of Hz. With that being said, there is no need to sample and operate at a frequency much higher than Shannon sampling frequency. As a result, the ADC sampling frequency is 500 samples per second. The discrete wavelet transform samples a data point based on an FIR filter than depends on the previous samples. To reduce the dynamic power, the DWT operating frequency has been synchronized with the ADC sampling frequency (i.e. f_{DWT} = 500 Hz). By operating the DWT at a higher frequency, the processor would be performing redundant computations, which expend additional energy. When the DWT stage is complete, a DWT_DONE flag is enabled so that the collapse buffer can begin its computation.

At the completion of the DWT stage, the collapse buffer must perform its serialized computation. In order to ensure the processing is complete before the DWT stage ends, the collapse buffer must be set to a frequency greater than the frequency of the DWT stage (i.e. $f_{CB}>f_{DWT}$). A collapse buffer clock frequency of 1 kHz was selected to minimize the dynamic power consumed by the collapse buffer while meeting the frequency requirement. At the conclusion of the collapse buffer stage, the CB_DONE flag is enabled.

In the final stage, data must be clocked into the transceiver at a rate determined by the SPI protocol. The standard SPI protocol has frequency range from 1 to 100 MHz. For the selected transceiver, the SPI clock is 2 MHz. To allow for a simple synchronization, the clock of the transmission finite state machine has been matched with the SPI clock (i.e. f_{TXFSM} = 2 MHz). 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. A summary of the globally asynchronous locally synchronous protocol is depicted in Figure 29.

3.5 FPGA Design Verification

3.5.1 FPGA Testing Framework

The Chameleon adaptive-compression system was designed in verilog, implemented on a Xilinx Virtex 5 LXT [40] prototype board and integrated with a Nordic NRF24L01+ [34] transmitter/receiver system. After the signal has been received, a SiLabs C8051F320 microcontroller transfers the received data to a host computer for

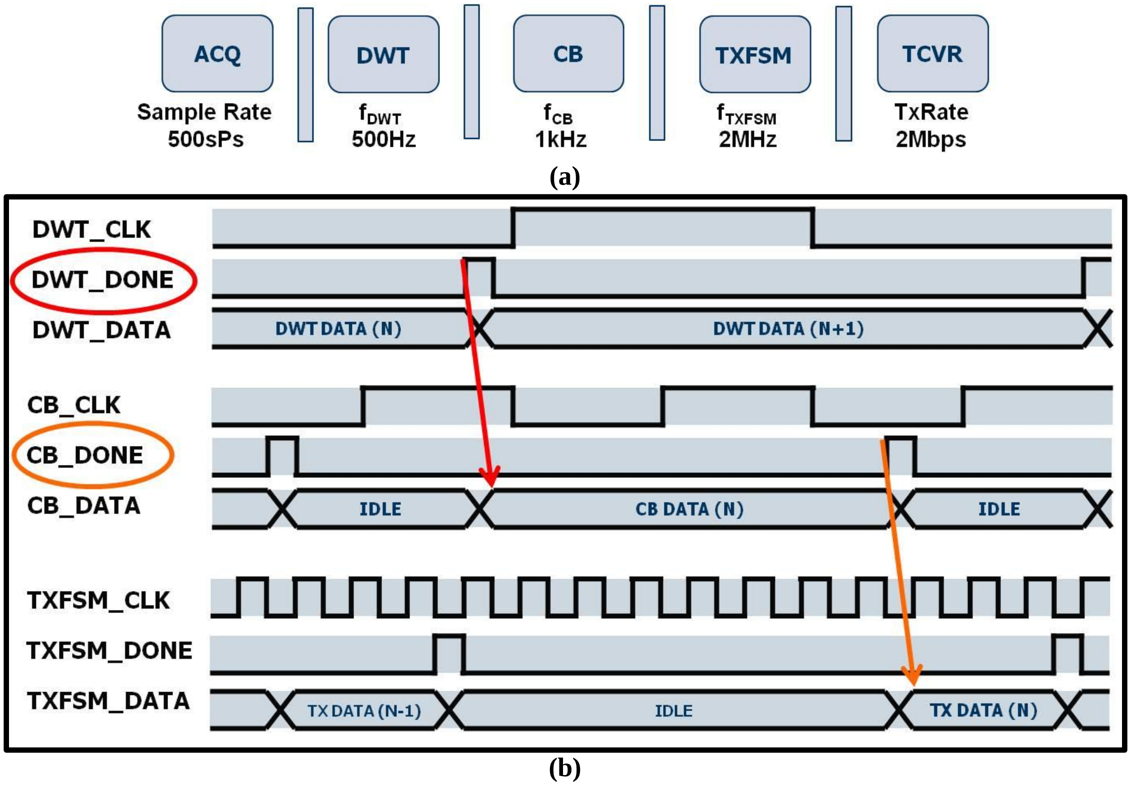


Figure 29: (a) Globally asynchronous locally synchronous (GALS) clocking scheme for the Chameleon System. (b) Asynchronous flags *_DONE denote when a stage is complete, signaling the next stage to begin.

post-processing and signal reconstruction using MATLAB. An Agilent MS07054A Oscilloscope was used to measure transmission delays by way of the transmission controllers' FSM. The testing framework can be seen in Figure 30.

3.5.2 Hardware Verification

The Chameleon system was verified at the hardware level by using a 7680-point sample, with 12-bit quantization factor (Q), and a 64 bit frame size (N). The EEG sample was taken from an online database of patients who suffer from epilepsy [32]. To verify the correctness of the system operation, the transmission controller FSM was monitored. The transitions of the 2-bit FSM controller are verified in Figure 31. Additionally, the disparity between the sleep times and transmission times are observed. As desired, the

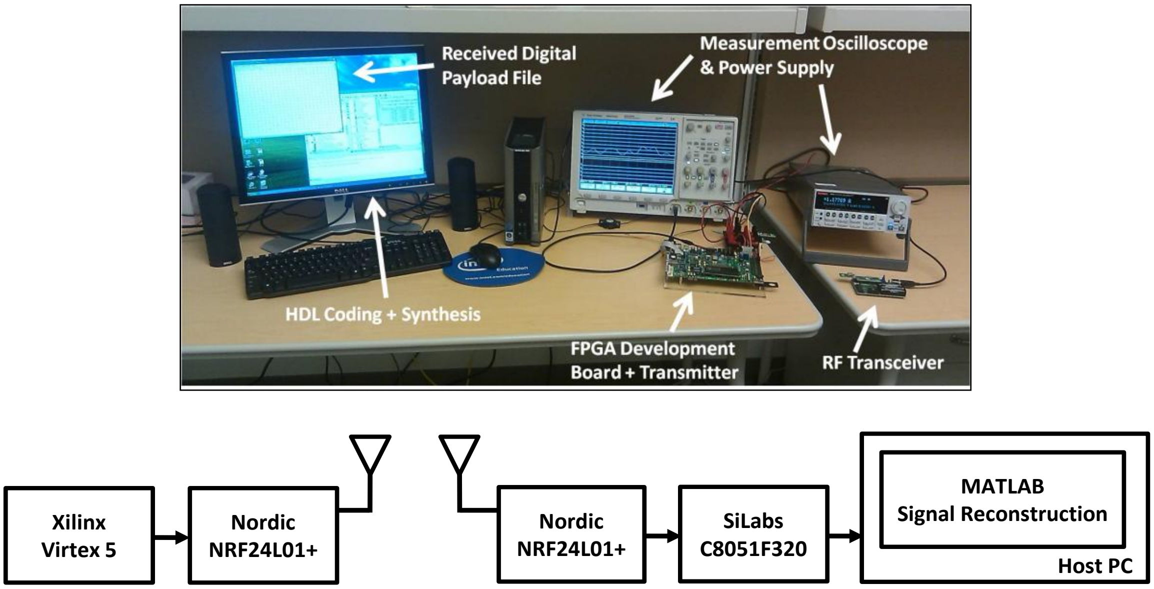

Received Digital Payload File
Measurement Oscilloscope & Power Supply
HDL Coding + Synthesis
FPGA Development Board + Transmitter
RF Transceiver
Xilinx Virtex 5
Nordic NRF24L01+
Nordic NRF24L01+
SiLabs C8051F320
MATLAB Signal Reconstruction
Host PC

Figure 30: Testing framework for the Chameleon system.

transceiver spends a minimal time transmitting and majority of the time in the ultra-low-power sleep mode.

3.5.3 Signal Processing Verification

To verify the correct signal processing operations, the received signal co-efficients were reconstructed into a time-domain signal using a custom MATLAB program. The signal processing operations were verified by using the same 7680-point sample as described above. A comparison of the original and reconstructed EEG signals is shown in Figure 32 (a). The signal of interest has two dominant spikes, namely near the 5000th and 6000th sample, respectively. Based on the naked-eye perspective, the Chameleon system should compress data that is outside these regions, leaving the epileptic spikes unaffected by the signal processing approaches. To gain a better perspective on the Chameleons’ adaptive data compression scheme, we can zoom into the dotted rectangular shown in Figure 32 (a).

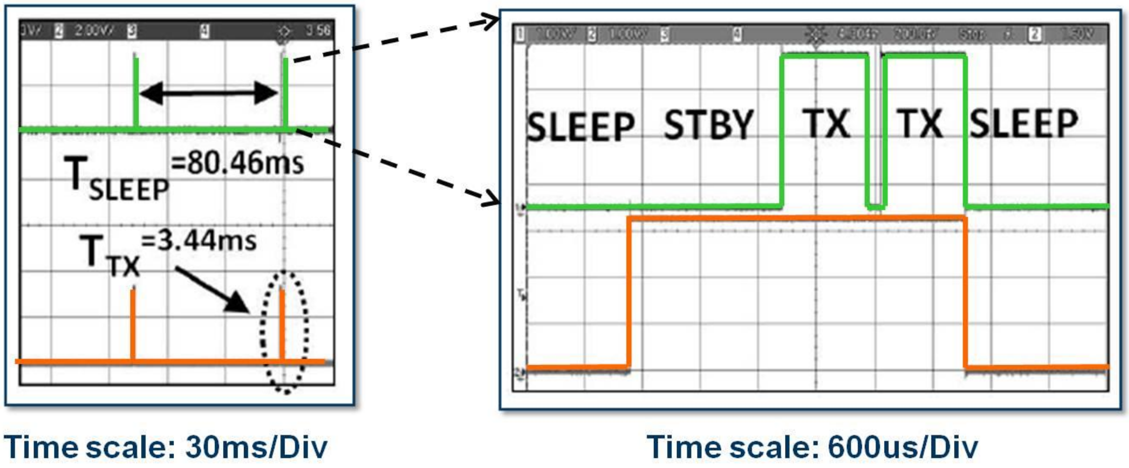
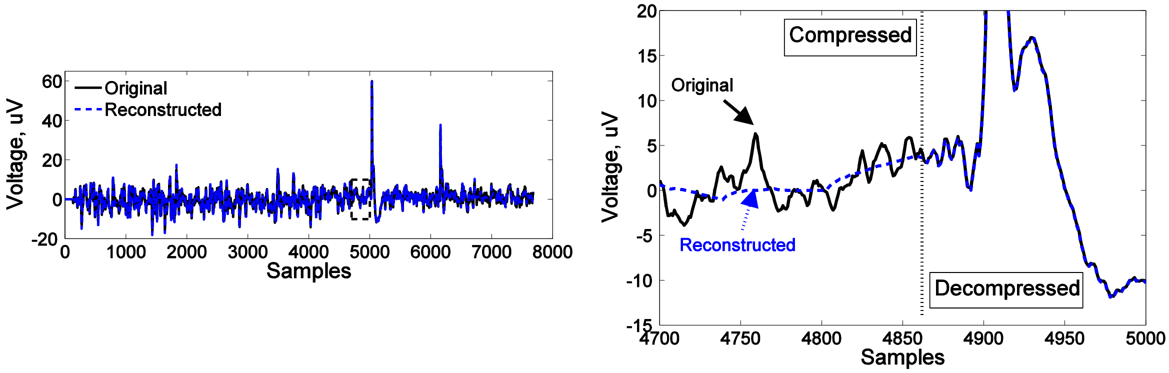


Figure 31 : Hardware verification of the Chameleon System for the 7680 point sample: (left). The FSM power cycling operation with sleep modes is depicted. (right) The oscilloscope operation of the FSM transmission controller verifies the correct states are being transitioned.

From the zoomed-in Figure 32 (b), it can be seen that prior to the emergence of the epileptic spikes, the reconstructed data is compressed, eliminating several high frequency transitions.

For all intents and purposes, this is acceptable in this background EEG region, because no important (epileptic spikes) are present. Just after sample 4850, the Chameleon system has detected an epileptic spike, causing the reconstructed signal to follow the original data more closely. In the 'Compressed' region, the fewer samples



(a) **(b)**

Figure 32: (a) The Chameleon systems' reconstructed EEG signal comparison with the original signal. (b) An expanded view of the dotted rectangular region shows how the adaptive compression is content-aware, decompressing data once epileptic behavior is detected.

transmitted allows for a low energy transmission, while in the ‘Uncompressed’ region, the highest accuracy is preserved for advanced epileptic spike analysis. The overall advantage is the adaptive nature of the Chameleon system: Regions where high accuracy is required can be designed independently of regions where low energy transmissions are desired.

3.5.4 FPGA Comparisons with Existing Methods

The adaptive compression algorithm was developed as an intermediate solution between two alternative approaches: full compression (data compression in all regions) and no compression (no data compression in all regions). 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. The adaptive compression implementation will be compared against the aforementioned methods that were presented in [2]. For future references: AC denotes adaptive compression; FC denotes full compression; and NC denotes no compression. The signal and energy measurements presented will be based on the FPGA-Transceiver system depicted in Figure 30.

The signal measurements of the three compression methods are shown in Table 3. The Percent Root mean square Difference (PRD) is a commonly used measure of reconstruction accuracy that provides a numerical measure of the residual root mean square error [39] . Perfect reconstruction occurs at a PRD = 0; therefore, a smaller PRD translates into a more accurate signal comparison. The EEG signal has been divided, providing two PRDs: One for the background EEG data and another for the epileptic

Table 3: Signal and Power Measurement Comparisons for Data Compression Schemes

Data Compression Scheme	**Signal Measurements**			**Energy**	
	Background PRD	**Epileptic Spike PRD**	**Compression Ratio (CR)**	**E_{TCVR} (uJ)**	**E_{FPGA} (mJ)**
No Compression	2.85	1.29	1.00	391.6	3.60
Full Compression	25.9	10.8	8.45	80.1	1.75
Adaptive Compression	25.9	6.78	6.42	91.7	1.92

spike data. The data has been separated into these two regions to examine how the accuracy varies based on the type of data. Additionally, with the system that has been implemented, each frame is designated as background or spike data. The compression ratio (CR) is calculated as well.

When there is no compression (NC), CR = 1.00 and the PRD = 2.85 in the best case accuracy condition. When all the data is compressed (FC) the PRD is high (worst case accuracy) but there is more compression . The adaptive compression (AC) method has accuracy in the epileptic regions as shown by the Spike PRD of 6.78, and can achieve a reasonable compression ratio of 6.42 to reduce the volume of data transmitted. The adaptive compression method can be designed to meet the users accuracy needs, based on the design parameters (i.e. ST, CT, or N).

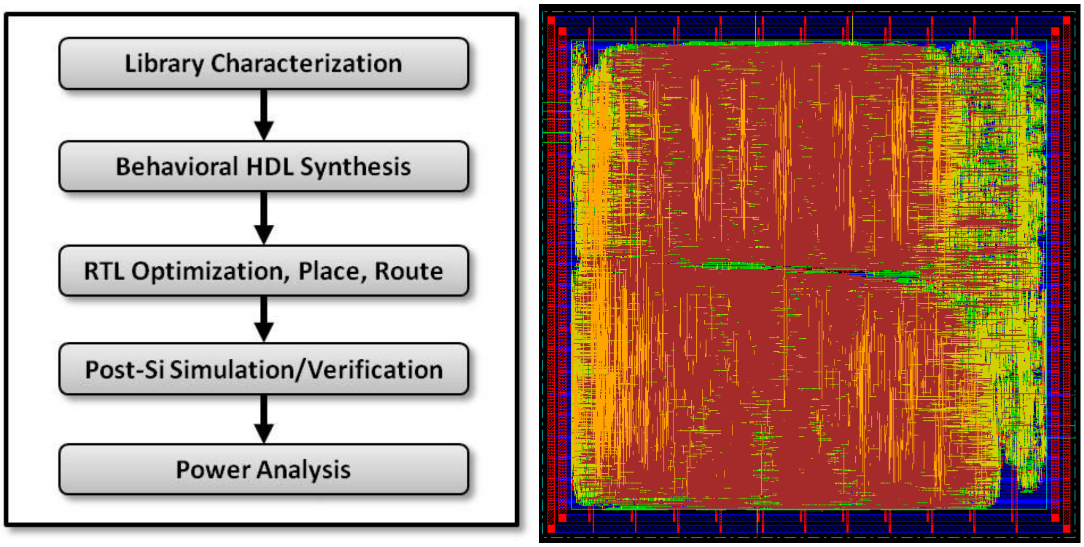
The energy measurements for three compression methods have also been presented in Table 3. In a similar fashion, the AC method is cross between the FC and NC methods. For the transceiver energy, E_{TCVR}, the AC method (91.7 uJ) is comparable to the FC method (80.1 uJ), and both provide a 4.0 X savings compared to the NC method (391.6 uJ). For the FPGA energy, E_{FPGA}, the AC and FC methods provide nearly a 50 percent reduction in energy, compared to the NC method. Overall, the energy savings that can be achieved by the AC method make it an appropriate choice for data compression, because these savings can be combined with the adaptive signal accuracy.

3.6 ASIC Design Methodology

The preceding verification ensures that the FPGA implementation of the Chameleon system is correct, however, an ASIC design is required to estimate the actual power consumption of the design. In order to accomplish these power estimations, an ASIC design flow is performed. The design flow begins with a verilog HDL implementation of the system and ends with a power analyzed, silicon based netlist.

3.6.1 ASIC Design Flow

The basic descriptions of the design methodology can be seen in the flow chart depicted in Figure 33 (a). The first step in this design flow is to select a technology and characterize the standard cells associated with the library. For the Chameleon system, the TSMC 180nm and Nangate Open Access 45nm technologies were used. Library characterization involves simulating the standard cells to develop tables with delay, power, and layout information that are used during netlist synthesis, placement and routing. The second step of the flow involves converting the behavioral verilog netlist into a structural verilog through RTL synthesis. With the structural verilog netlist, the RTL can then be optimized, placed and routed with the technology library files. After the netlist has been placed, a layout based simulation model can be created that includes switching activity, nodal capacitance tables and detailed delay. In the end, power analysis is performed on the layout based simulation model, and provides an accurate estimation for post-silicon measurements. The ASIC design flow has been completed for the proceeding results. The layout developed in TSMC 180nm technology is shown in Figure 33 (b).



(a) **(b)**

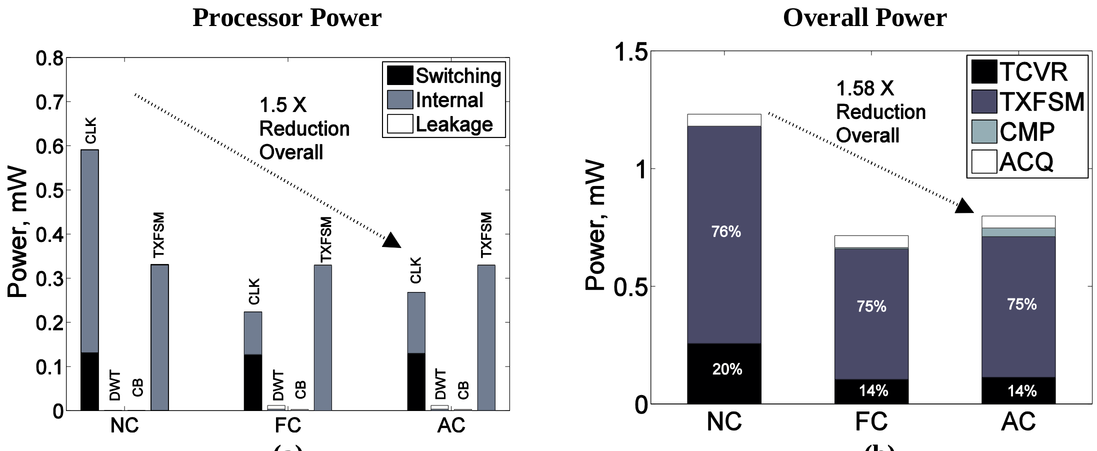
Figure 33: (a) ASIC design flow used to develop the (b) silicon based layout and perform accurate power analysis.

3.6.2 ASIC Power Comparisons with Existing Methods

To estimate the ASIC power of the three compression schemes, the proposed architecture in was used, so that a fair comparison could be made with all three designs. With no compression (NC) scheme, the same architecture depicted in Figure 26 is used; however, the collapse buffer stage has been removed to eliminate the data compression stage. When the full compression scheme is used, the same architecture as the proposed adaptive compression scheme is used, but the compression rate is not variable. All designs were created with a TSMC 180nm standard cell library which operates at a supply voltage of 1.8V a V_{TP} = -400 mV and a V_{TN} = 367 mV. In these designs, the f_{DWT} = 500Hz, f_{CB} = 1kHz and f_{TXFSM} = 2MHz (See Figure 26).

The digital processor power of the three compression methods are depicted in Figure 34 (a). The dynamic power is composed of the switching and internal power, while the static power is equivalent to the leakage components. The switching power is defined as the power at the boundary of the cell (i.e. interconnect related), while the internal power is made up of parasitic power that is contained within the cell. In Figure 34 (a), each processor is divided up into four power contributions: the power of the clock network (CLK), discrete wavelet transform stage (DWT), collapse buffer stage (CB), and transmission controller stage (TXFSM).

The first thing to note is that in the digital processor, the dynamic CLK and TXFSM power tend to dominant. The dynamic power is a large fraction of the overall power because of the high frequency clock in the TXFSM stage relative to the other stages. Additionally, the NC processor does not have a collapse buffer stage, but it still consumes the highest power compared to the other two processors. The NC processor has to process more bits of data, because compression does not occur, which increases clocking and transmission power overall. The most important metric from Figure 34 (a), is that a processor with an adaptive compression ratio (AC) has a similar power to the processor with a fixed compression ratio (FC). As stated in Chapter 3, the goal is to have


Processor Power

Overall Power

(a) (b)

Figure 34: Comparison of the (a) digital processor power and (b) the complete wireless system power of the 'No Compression' (NC), 'Full Compression' (FC) and 'Adaptive Compression' (AC) methods. The Chameleon processor was designed with the AC method and demonstrates it low power that is comparable when a flat compression rate (FC) is used.

the adaptive compression algorithm have the signal characteristics of a system that has no compression (NC), and the power characteristics of a system with a fixed compression (FC). In general, when moving from the NC scheme to the AC scheme, a 1.5 X power reduction can be achieved in the digital processing.

As a result of the implemented digital processors, the overall system power of each design is analyzed in Figure 34 (b). The power contributions of the system are divided up into four areas: the transceiver power (TCVR), the transmission controller power (TXFSM), the power introduced to performed compression (CMP) and the power associated with acquiring and amplifying the inputs (ACQ). For this and future designs, the Texas Instruments OPA333 [41] and ADS7866 [42] were selected as a low power front end pair, because of its optimal use with medical instrumentation.

Similar to the digital power trend, the NC system consumes more power (~1.3 mW) overall because it has to process and transmit more bits than a method that uses compression. The adaptive compression (AC) based system also has similar power consumption to that of a full compression method, with a power of nearly 0.8 mW. The AC system results in a 1.58 X reduction in system power compared to the NC system. By and large, the digital power is the dominant portion of the system, contributing to at least 75 percent of the overall power. Previous works have suggested that the transceiver power is dominant [17]; however, this digital processor has not been optimized to minimize power. The upcoming sections will attempt to tackle the digital processing power of the Chameleon system, such that the system operates at a lower power.

3.7 Low Power Design Optimizations

The goal of the proposed digital processor is to reduce the overall system level power associated with detecting, compressing and transmitting electroencephalography signals. From the initial implementation of the Chameleon system, the results of Figure 34 demonstrated that the adaptive compression algorithm is a viable low power candidate for the Chameleon system. The results in Table 3 demonstrated the advantage of an adaptive compression algorithm to modify its compression ratio, dependant on the type of EEG data that is being processed. Based on these two arguments, the adaptive compression algorithm will be used for the Chameleon system processor. The proceeding analysis will focus on the digital system design optimizations to achieve a low power solution.

3.7.1 Clock Gating

From the initial implementation of the Chameleon system, the dynamic power was dominant due to the high frequency clock (CLK) associated with the transmission finite state machine controller (TXFSM) (See Figure 34). Recall that the TXFSM is operated at a high frequency to improve to energy efficiency and reduce the overall power. With the clocking scheme depicted in Figure 29, the TXFSM only needs to be active during non-idle times, which is small compared to the cycle time. With these two concepts in mind, the burden of the clock and the transmission controller can be reduced by clock gating the TXFSM stage when not in use. Essentially, the clock will be active only during the time the transmission controller and wireless transceiver need to be operating. The updated clocking scheme when clock gating has been implemented is

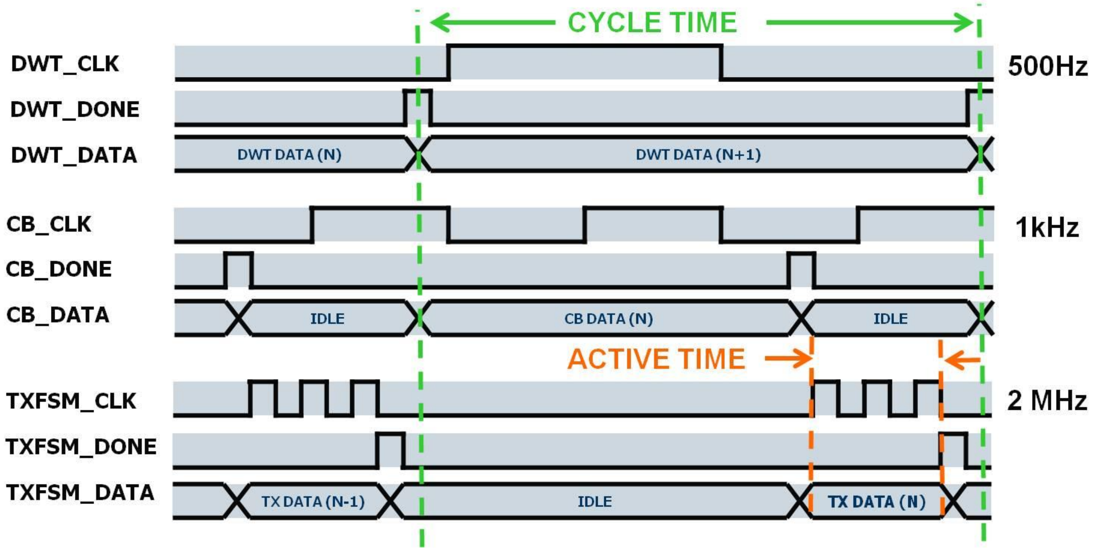
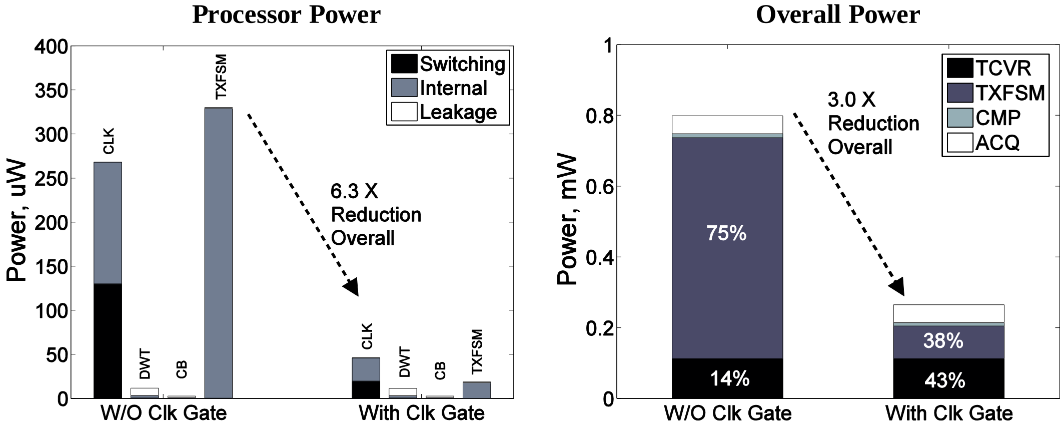


Figure 35: Updated clocking scheme when clock gating has been implemented. The active time (10.4ms) should be much smaller than the cycle time (131ms) in order to take advantage of clock gating.

presented in Figure 35. The active time of the TXFSM is 10.4ms, which is much smaller than the cycle time (131ms) that is set by the low throughput DWT stage. The small ratio of active to cycle time allows for maximum power savings, by reducing the dynamic power during non-idle times.

The impact of clock gating on the Chameleon digital processor (with adaptive compression) is depicted in Figure 36 (a). Prior to clock gating being implemented, the dynamic power of the clock and transmission controller contributes to more than 90 percent of the overall system power. After clock gating is implemented, there is a 6.3X reduction in the digital power overall. It can be seen from the figure that the clock power and transmission controller power is significantly reduced. Recall that this reduction in power was possible because of the small ratio of active to cycle times.

The impact of clock gating on the overall system power is plotted in Figure 36 (b). By implementing clock gating, the overall system power is reduced to below 0.3 mW. Prior to clock gating, the overall system power was on the order of 0.8 mW. As a



(a) (b)

Figure 36: (a) The impact of clock gating is analyzed with the Chameleon system and a 6.3 X reduction in digital power can be achieved. (b) Overall, the TXFSM power is significantly reduced, yielding a 3.0 X reduction is the system power.

result of implementing clock gating, the overall system power can be reduced by a factor of 3.0 X. Another thing to note is that the overall system power is still dominated by the digital TXFSM and the radio frequency TCVR.

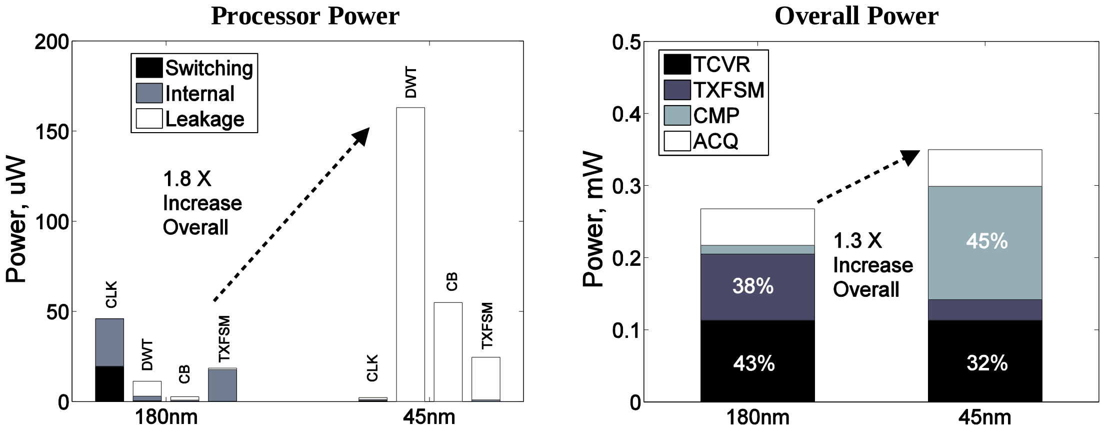
3.7.2 Technology Scaling

After clock gating was implemented, system was still dominated by the transceiver power, composing 43 percent of the overall power. The digital power, which composed 38 percent of the overall power, is mostly dynamic switching power (shown in Figure 36 (a)). To reduce the dynamic power, the supply voltage can be reduced to sub or near threshold voltages. The preceding ASIC design was performed in a TSMC 180nm technology, with a supply voltage of 1.8V. Due to the current density limitations of the 180nm technology, a more advanced technology is required to achieve the delay requirements, as the supply voltage is scaled. In this subsection, the Chameleon processor is implemented in Nangate’s 45nm Open Cell Library operating at 1.0 V [43].

The digital processor power is the technology is scaled from 180nm to 45nm is shown in Figure 37 (a). In 180nm, the power is dominated by dynamic switching energy, but it then becomes dominated by leakage energy in 45nm. In 45nm, there is a 1.8X increase in power, and the leakage energy contributes to more than 98 percent of the digital power. Now, the DWT power is now dominant, and this is because of the low frequency processing frequency, which causes the system to have more time to leak.

The overall system power in the Nangate 45nm technology is plotted in Figure 37 (b). Due to the increased leakage of the digital processor, the system power increases by 1.3X to roughly 350 uW. Additionally, 45 percent of the overall power is made up of the DWT and CB power: the power introduced into the system by adding data compression. In general, due to low throughput processing frequency, the increased leakage cause the power to increase at all levels.

To further investigate the implications of leakage on technology scaling, Figure 38 shows how the area compares when transitioning between 180nm to 45nm. There is a 23X reduction in area when scaling occurs, and the supply voltage is reduced as well.



(a) **(b)**

Figure 37: (a) The technology scaling impact on the digital processor shows that the leakage power tends to dominate at smaller technologies. (b) In 45nm, the overall system power is dominated by the digital power, more specifically the power introduced into the system by compression.

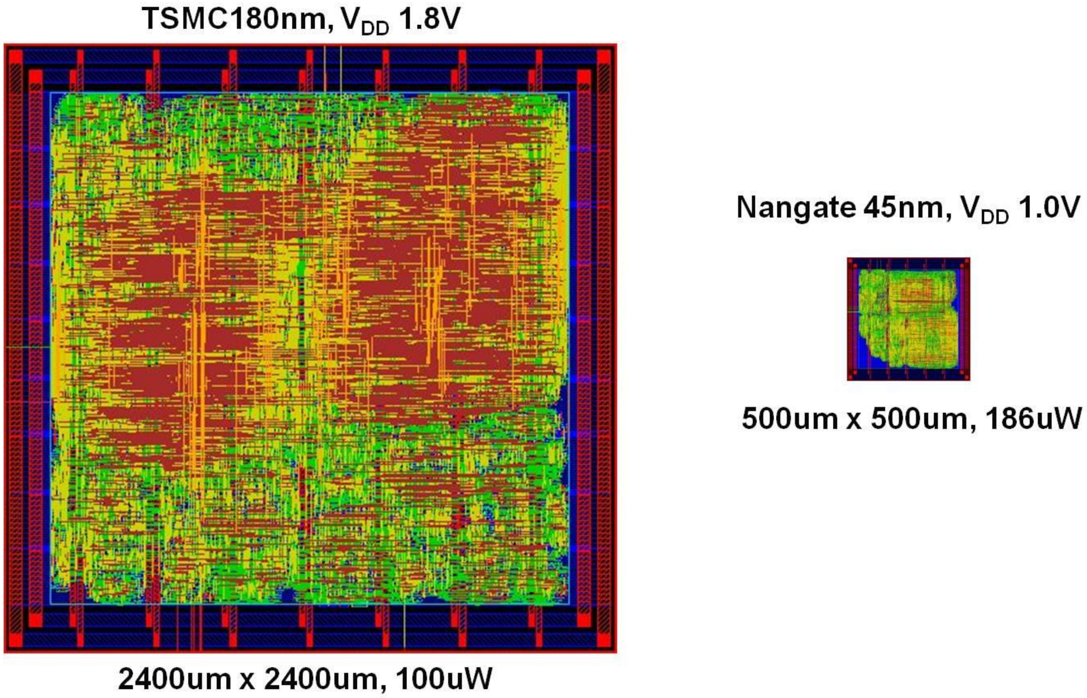


Figure 38: The area and power of the digital processor is compared as a result of technology scaling. The 23 X reduction in area and lowered supply voltage cannot overcome the increased current density of the smaller technology nodes.

The resulting evidence of the increased current density and leakage power comes from the increase in overall digital power (from 100uW to 186uW).

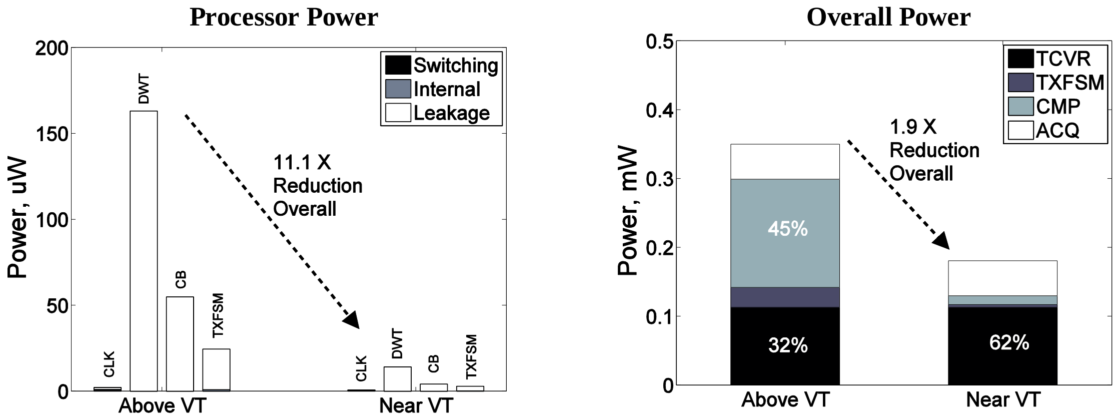
3.7.3 Near Threshold Computing

From the results of the previous subsection, technology scaling has proven to increase the leakage, which adds to the overall power. The goal of technology scaling was to move to a technology that would allow for a reduced supply voltage. In 180nm, achieving a 2 MHz frequency in the near threshold domain was not possible. Given that the Chameleon processor operates in the low-throughput domain (i.e. f_{DWT} = 500Hz), it will be possible to obtain further savings by reducing the supply voltage to near threshold voltages. To examine the impact of near threshold computing, the 45nm design was operated at a supply voltage of 0.5V, with V_{TN} = 469 mV and a V_{TP} = 492 mV.

The impact of near threshold computing on the digital processor power is depicted in Figure 39 (a). In 45nm, both the above threshold and near threshold digital processor is composed mainly of leakage power. By simply reducing the supply voltage, the largest ratio of gains can be achieved, an 11.1 X savings in power. By looking at Figure 39 (b), the overall system level power and its impact on near threshold computing is analyzed. One thing to note from Figure 39 (b) is that the digital processor power (CMP + TXFSM) is not longer dominant. As anticipated with most wireless systems, the transceiver (TCVR) power contributes to more than half (62 percent) of the system power. In the near threshold regime, the overall system power can be reduced to 180uW. Prior to reducing the supply voltage, the overall system power was approximately 350uW. Near threshold computing is a viable option because of the low frequency targets for the system, and the overall objective to reduce the system power. In the end, near threshold computing reduces the overall system power by a factor of 1.9X.

3.7.4 Burst Mode Processing and Power Gating

The combination of clock gating and near threshold computing provides a low

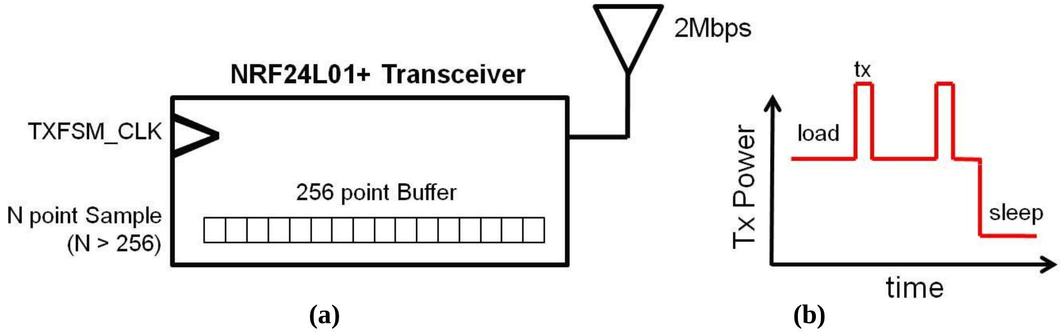


(a) **(b)**

Figure 39: (a) The impact of near threshold computing on the Chameleon system can reduce the digital processing power by 11.1 X. (b) Overall the system power can be reduced by nearly 2 X, operating at just below 200 uW.

power implementation of the digital processor, but more can be done to reduce the overall system power. From Figure 39 (b), the transceiver power is now the dominant portion of the system power. Previously, the lowest power achieved by the digital processor was dominated by leakage power, as shown in Figure 39 (a). By using burst mode processing, combined with power gating, both the digital and RF components of power can be reduced.

The basis of burst mode processing relies on understanding the loading characteristics on the NRF24L01+ transceiver. A schematic of the NRF24L01+ and the timing graph for loading the transceiver is shown in Figure 40. The left figure depicts the block diagram of the transceiver that has a loading clock frequency controlled by TXFSM_CLK (currently set at 2MHz), and a transmission data rate that is constant at 2 Mbps. The transceiver has a fixed size data buffer that can hold 256 bits. When a sample larger than 256 bits, the transceiver must cycle between the transmission and load phase, until all bits has been transmitted. For example, Figure 40 (b) depicts the power and timing graph when the sample size is greater than 256 bits, but less than 512 bits. To reduce the power consumed by overall system, the transceiver should operate in sleep



(a) **(b)**

Figure 40: Loading characteristics of the NRF24L01+ transceiver. Due to the fixed size transmission buffer, multiple loadings and transmissions are required. To reduce the loading time (and power), the TXFSM_CLK can be increased.

mode as often as possible. Since the data rate is fixed by the commercial transceiver, the rate at which the data is loaded into the transceiver is the only choice for this optimization. To accomplish this task, burst mode processing will be performed.

Burst mode processing is a method of designing a digital circuit such that the delay of the critical path is much faster than the target frequency of the system. The purpose of burst mode processing is to complete the computation as soon as possible, so that power gating can be used to further reduced system power. In the design of the Chameleon digital processor (See Figure 26), the low frequency sampling and processing (e.g. f_{DWT} = 500Hz) creates a fixed time window, that is independent of the compression (f_{CB}) and transmission (f_{TXFSM}) data rates. With that being said, it is possible to increase the frequency of the CB and TXFSM stages, without impacting the dynamic power of the system. To explain this argument, one must understand that regardless of the frequency, the dynamic energy of computation is $C_L V_{DD}^2$. Therefore the dynamic power is constant, because it is equivalent to the dynamic energy (fixed by the hardware design), divided by the cycle time (fixed by the DWT).

The new clocking scheme for the proposed Chameleon system is shown in Figure 41. To increase the loading rate of the transceiver, the TXFSM_CLK has been increased to 32 MHz. In addition, the collapse buffer can be operated at a higher frequency as well. By operating both at a higher frequency, once the stages complete, the power supply can cut off to reduce the leakage power. In the end, the cycle time is still defined by the frequency of the DWT_CLK (i.e. 500 Hz), which is much slower than the burst mode frequencies (i.e. 32 MHz). When the burst mode processing has completed, the idle leakage power can be reduced by using power gating.

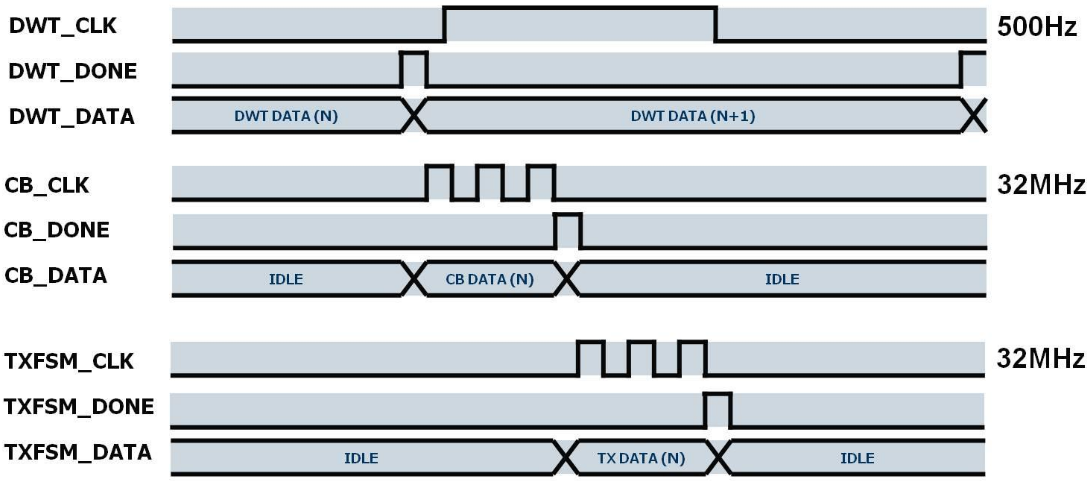
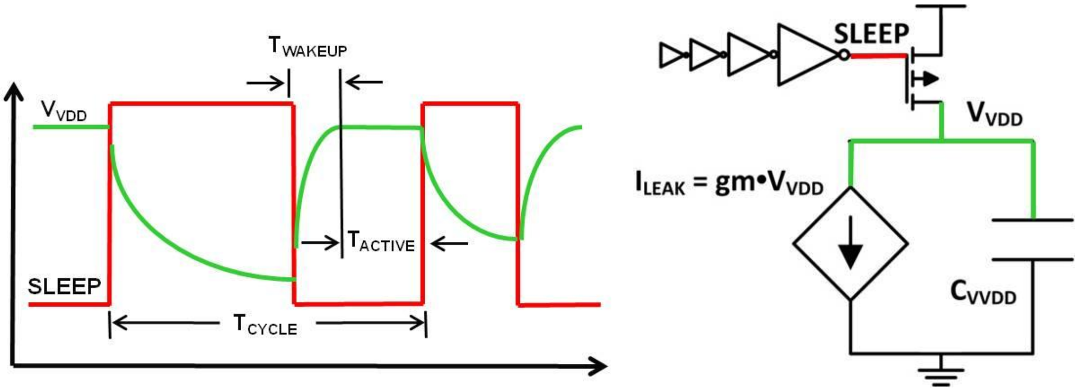


Figure 41: Clock scheme for the Chameleon system with burst mode processing. The collapse buffer and transmission controller are clocked at 32 MHz, reducing the loading time of the transceiver and offering the potential for clock gating with increased idle times.

Power gating is a low power technique that can be used to reduce the standby leakage power in digital systems. In low performance domains, processing can complete well before the required time, which allows the circuit to idle. During the idle periods, leakage energy is consumed due to the non-zero off currents in deep submicron CMOS designs. To eliminate this leakage energy, the power supply can be cut off, or gated, offering the potential for extreme power savings. In the design of the Chameleon system, there are fixed cycle times (determined by the DWT) and the power is dominated by idle leakage currents. These two aspects make power gating an ideal candidate to further reduce the system power.

The expected timing behavior of the power gating cycle is shown in Figure 42 (a). When operated correctly, the power gating cycle has three phases: active mode, the time where the circuit is computing; sleep mode, the time where the circuit is idle; and wakeup mode, the time where the circuit is transitioning from sleep to active modes. The active



(a) (b)

Figure 42: (a) Example timing behavior of the power gating cycle with sleep, wakeup and active times. (b) Power gating model used to estimate the wakeup time and wakeup power.

Table 4: Wakeup Metrics Associated with Power Gating and for a Burst Mode Frequency of 32MHz

V_{VDD} (mV)	C_{VVDD} (pF)	gm (uS)	I_{LEAK} (uA)	T_{CYCLE} (ms)	T_{ACTIVE} (ms)	T_{WAKEUP} (ns)	E_{WAKEUP} (pJ)	E_{DSP} (uJ)	E_{SYSTEM} (uJ)
491	7.84	29.0	14.5	131	3.12	34.6	1.79	2.50	10.48

and sleep mode power (times) are determined by the processing and standby operations, respectively.

The circuit in Figure 42 (b) was used to control the supply voltage, and determine the wakeup power (time). The circuit operates based on the SLEEP control signal and provides a virtual supply voltage (V_{VDD}) that is slightly reduced compared to the system supply voltage. The capacitance of the virtual supply voltage (C_{VVDD}) and the leakage current (I_{LEAK}) will be directly proportional to the size of the circuit. Initially the SLEEP signal is low, allowing the virtual supply voltage (V_{VDD}) to be supplied through the PMOS transistor (active mode). At the conclusion of the active mode, the SLEEP signal becomes high, turning off the sleep transistor (sleep mode). Ideally, no leakage current occurs during the sleep mode operation, offering the low power gains desired in design dominated by leakage current. On the falling edge of the SLEEP signal, the wakeup phase begins, as the capacitance of the virtual supply (C_{VVDD}) is charged. While the capacitance is being charged, it is also in contention with the leakage current of the devices, which has a dependence on the supply voltage (I_{LEAK} = gm·V_{VDD}). When the virtual supply voltage reaches the maximum value, the wakeup time (power) can be define

The wakeup metrics associated with power gating for a burst mode frequency of 32 MHz is shown in Table 4. The metrics in the table can be referenced from the schematic in Figure 42. The system was designed to have a virtual supply voltage of 491 mV, which introduces less than three percent of error compared to the supply voltage of 500mV. Another thing to note is that the active time ($T_{ACTIVE =}$ 3.12 ms) is much smaller than the cycle time (T_{CYCLE} = 131 ms), offering a large disparity between the sleep (T_{SLEEP} = T_{CYCLE} – T_{ACTIVE}) and active times. The large disparity shows the potential benefits of power gating and leakage savings. The wakeup energy (E_{WAKEUP}) is only a few picojoules, and is much smaller than the overall DSP energy (E_{DSP}). For this reason, there is no extra energy added into the system by implementing power gating.

Power gating with burst mode processing was implemented on the collapse buffer and transmission controller stage (Figure 26). The frequency of the collapse buffer (f_{CB}) and finite state machine (f_{TXFSM}) were increased during active mode operation and the supply voltage was cut off during sleep modes. The comparison of the Chameleon system without power gating and with power gating for a varying Frame Size is plotted in Figure 43. In the previous discussions, it was implied that the Frame Size was 64 bits long. When power gating and burst mode processing is applied, no major gains are seen when the Frame Size is held at 64. However, when the Frame Size is 128, the power is reduced nearly 20 percent to 130 uW. The power is reduced because when the Frame

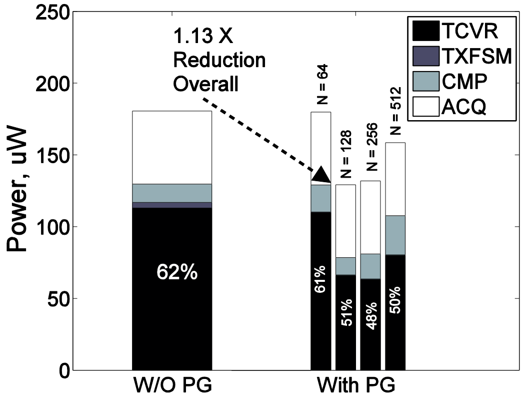


Figure 43: Comparison of the Chameleon system power without and with power gating shows that the power can be reduced by 20 percent when the Frame Size is increased as well.

Size increases, the sleep time can increases as well. Overall, power gating is most effective when the maximum frequency of the system can be achieved, in this case 32 MHz, where the disparity of active and sleep times can be maximized.

3.7.5 Power Reduction Summary

The summary of the low power design optimizations have been compared in Figure 44. The proposed system is first compared with a conventional system that does not use data compression or a custom digital processor. As adaptive date compression is implemented, the DSP and TXFSM power reduce because there is less data to process and transmit. At this point, it was discovered that the design was dominated by the high frequency clock power associated with processing the transmission controller. By implementing clock gating, the digital processing power was drastically reduced. An increase in power was then seen as the leakage power increased due to technology scaling from 180 to 45nm. To further reduce the power, the supply voltage was reduced to near threshold voltages (i.e. from 1.0V to 0.5V). Finally, the leakage and transceiver dominated design was addressed by power gating and burst mode processing. The

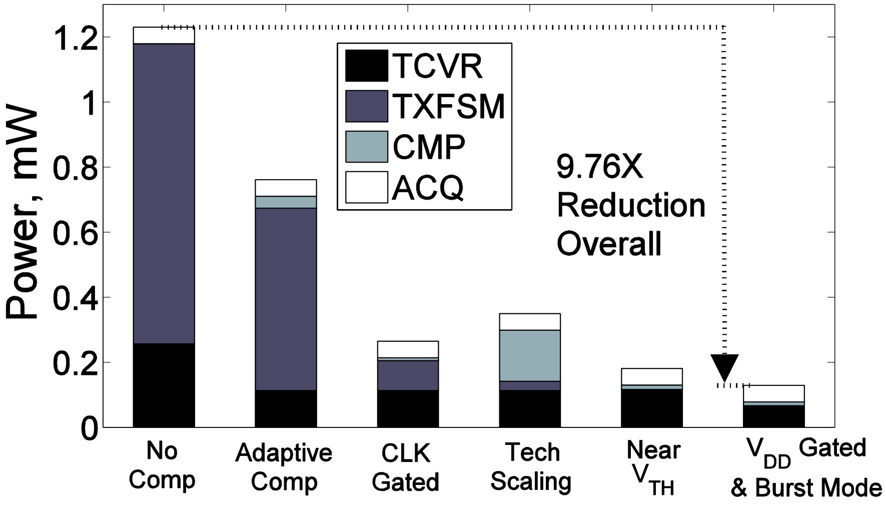


Figure 44: Summary of power reduction methods for low power Chameleon system design.

combination of these two optimizations reduced the digital power, but more importantly, minimized the transceiver power when the processing frequency is set to 32MHz. In summary, a 9.76 X reduction in overall system power can be achieve by the design and implementation of the proposed Chameleon system, compared to a conventional method.

3.8 Summary

In summary, this chapter outlined the design, implementation and analysis of a custom digital processor that would be used to reduce the power in a wireless electroencephalography system. Based on the algorithm outlined in Chapter 3, the custom processor was integrated into a proposed system called Chameleon, based on the adaptive nature. The proposed processor was implemented as a three-stage pipeline that consisted of detection, compression and transmission. The embedded hardware architecture was verified on a prototype FPGA board and an ASIC design flow was used to estimate the system power. To optimize the system, several low power design methods were used to reduce the power such as: clock gating, near threshold computing, and power gating. Overall the design of the Chameleon system has the potential to reduce the power of wireless monitoring systems by at least 9.76 times, compared to conventional approaches.

CHAPTER 4

ANALYSIS AND DESIGN OF ENERGY AND SLEW-AWARE SUBTHRESHOLD CLOCK SYSTEMS

4.1 Introduction

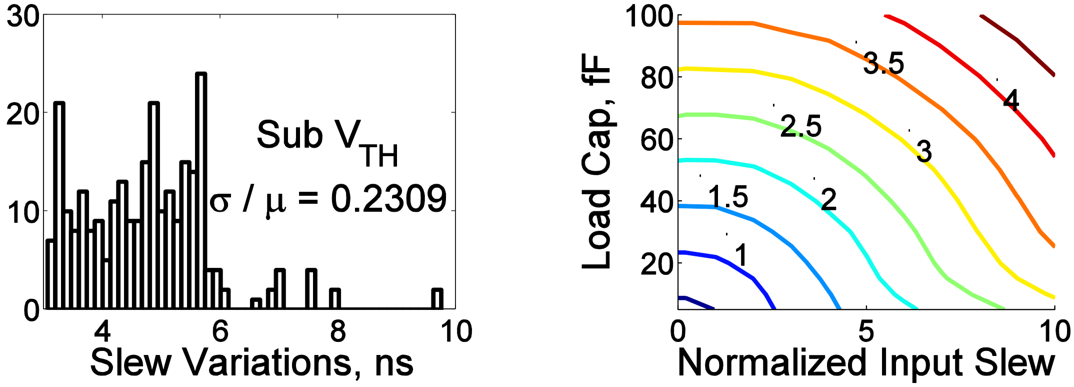
Transistors operating in the subthreshold region constitute an attractive technology for ultra-low-power mobile applications such as microsensors and biomedical devices. Subthreshold logic can significantly reduce the system energy by operating at a supply voltage lower than the threshold voltage of the devices. Even though low power is the primary goal, it is still innate for circuit designers to optimize secondary parameters such as robustness and performance. As a result of these design parameters, works have been presented to optimize energy and delay, while performing computations with minimal error [44-46]. Additionally, efforts have been made to optimize devices such that circuits can be operated at medium frequencies in the order of tens to hundreds of megahertz [47].

In addressing the design of an optimal energy-delay subthreshold system, the clock network plays a significant role. Delivering robust clock signals to hundreds (or even thousands) of flip-flops requires the clock tree to be optimally designed to handle issues of delay, skew, and jitter. In subthreshold, the signal slew (i.e. 10 to 90 percent transition time) also has the capability to affect the system performance [48, 49]. Additionally, due to its high switching activity, the clock network can contribute up to 40 percent of the total dynamic power [50]. This power trend is expected to continue when an above-threshold system is operated at subthreshold voltages. Therefore, designing a low-energy, robust clock tree is a critical challenge to implement a large-scale, subthreshold system. This challenge is increasingly difficult because subthreshold designs are always constrained by the requirement of robustness. As the supply voltage of digital integrated circuits is reduced below the device threshold, the characteristics of the transistor change. The transistor current in the subthreshold regime has an exponential dependence on gate voltage, threshold voltage, temperature, and additional process parameters. In contrast, the transistor current in the above-threshold regime has a linear or square dependence [51]. Due to the exponential trend, subthreshold devices are more sensitive to voltage and process variations, compared to above-threshold devices. To mitigate the effects of subthreshold process variations, several methods have been proposed and proven successful [52-56].

The purpose of this chapter is to analyze the impact of clock slew in subthreshold designs and propose a technique for a low-energy, slew-controlled, clock-tree design. The inherent slew variations in a clock tree will be examined and this chapter will explain that the slew variations can cause a direct increase in cycle-time computations. A systematic approach to design the clock tree will be presented. The purpose of the subthreshold clock tree is to reduce the clock-slew variations while minimizing the energy dissipation. The results will show that a smaller nodal capacitance is necessary to control the slew in a subthreshold clock tree, which can increase the energy dissipation. Recognizing that the wire resistances have a negligible effect in subthreshold circuits, proper wire sizing is necessary to reduce the clock energy. Lastly, a dynamic, nodal-capacitance-control technique will be presented. The technique will allow a larger slew at the earlier nets of the tree while controlling it more aggressively near the sink nodes.

4.2 Motivation

The purpose of this section is to demonstrate the effects of clock slew and how it can directly impact the cycle time. The input slew of a logic gate can cause the output delay to change in the range of 50 to 100 percent [48, 49]. The output slew of a logic gate has a strong dependence on the device dimensions, gate and drain voltages, as well as the load capacitance. In recent literature, this slew effect has been designated a concern for robust flip-flop design in the subthreshold regime [56]. When designed with above-threshold methods, subthreshold clock trees exhibit significant slew variations at the sink nodes (i.e. the nodes directly connected to the latches) that exacerbate timing violations. As an example, the focus of this experimental section will be based on a clock network designed using an above-threshold, zero-skew, clock-tree design algorithm. The algorithm uses a slew-control method that limits the maximum capacitance driven by the each node within the clock-tree network. The power supply of this design was then reduced to below the device threshold. In this clock tree, inverting buffers were used to reduce the number of devices, which saves energy. The clock tree was designed using a 65nm predictive technology model (PTM) with V_{TP} = -378mV and V_{TN} = 429mV [37]. The power supply voltage was 300mV and the clock tree has 267 sink nodes, each driving multiple flip-flops. The design used to define the clock-sink destinations was the IBM r1 benchmark [57].



(a) **(b)**

Figure 45: (a) Deterministic slew variations at the sink nodes of a subthreshold clock tree designed using above threshold concepts. (b) Normalized output slew contours and their dependence on input slew and total load capacitance for an inverter.

4.2.1 Deterministic Slew Variations

Deterministic (or design induced) variations occur as a function of load capacitance, routing distances, and buffer placement. Under this definition, it is unlikely that all sink nodes will have the same slew, which will result in a deterministic variation across the chip. To understand the impact of deterministic slew variations, the disparity of slew at the sink nodes and the slew dependence on are presented in Figure 45.

Ideally, every chip design would have the same spread of deterministic slew variations. The slew at the clock-sink nodes is important because they are the control for the latches and flip-flops. The coefficient of variation (CV) is a normalized measure of dispersion of a probability distribution and is defined as the ratio of standard deviation (σ) to mean (μ), or as follows:

$$CV = \sigma/\mu. \qquad \text{Eq. (4.1)}$$

In the subthreshold clock tree, the CV is 23 percent, showing there is a wide distribution of slew. Well controlled CVs are in the range of 10 to 15 percent.

The distribution of slew, corresponding to the slew variation across different sink nodes of the tree, is shown in Figure 45 (a). The output slew variation is caused by the spatial differences in the sink locations. The different sink locations require the clock signals to be routed with different paths, resistances, and capacitive loads. The inverter output slew is affected by the input slew and load capacitance, as shown in Figure 45 (b). The output slew of an inverter has a strong dependence on the input slew and load capacitance, which is supported by the curves in the figure. This dependence is important for a clock-tree path, because the input slew and capacitance of each buffer vary. To achieve a smaller output slew, a smaller input slew and load capacitance are required. The recovery of the slew (ratio of output slew to input slew) through an inverter stage is an important consideration for the clock-tree path. Ideally, the recovery is greater than one, which means the output slew is smaller than the input slew. If the slew is not controlled, it can become progressively larger as the signal progresses down the clock-tree path. This can be seen from examining Figure 45 (b). Given a load capacitance of 80 fF and input slew of 1.0 ns, the output slew is three times the input slew. Like the slew, the slew recovery is a strong function of the load capacitance. Controlling the capacitance in the clock tree is a very important tool in reducing the slew propagation and slew variations. In summary of Figure 45, if the load capacitance is selected appropriately, the inverter can recover the output slew even for a large input slew.

4.2.2 Clock-Slew Impact on Cycle Time

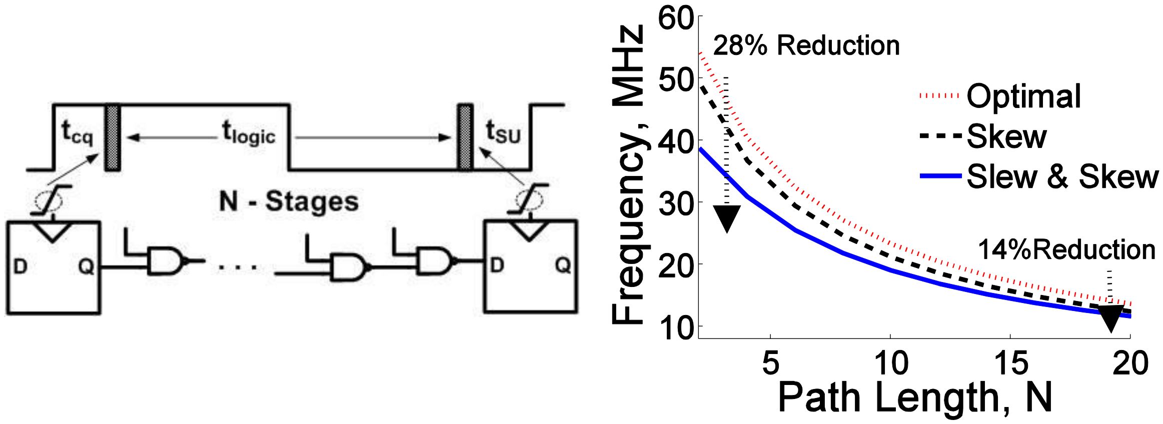
A direct motivation of this work is the impact of clock skew and slew on the cycle time. A sample logic path, with varying skew, slew and number of logic stages, is analyzed to examine its impact on frequency in Figure 46.

The schematic of a generic logic path composed of fan-out-4 (FO4) NAND gates between two registers is shown in Figure 46 (a). The minimum cycle time and the maximum clock frequency for the above system is given by

$$T_{min} \geq t_{c-q} + t_{logic} + t_{su} - \delta \text{ , and} \qquad \text{Eq. (4.2)}$$

$$F_{max} = 1/T_{min} \text{ .} \qquad \text{Eq. (4.3)}$$

The maximum propagation delay of the register (clock-to-q delay) is $t_{c\text{-}q}$. The maximum delay of the combinational logic is t_{logic}. The setup time for the registers is t_{su} and δ the clock skew. To compute the maximum clock frequency, the minimum cycle time, T_{min}, is needed. Three F_{max} cases are plotted in Figure 46 (b), as the total number of stages (N) is varied. In each case, skew and slew requirements are as follows:



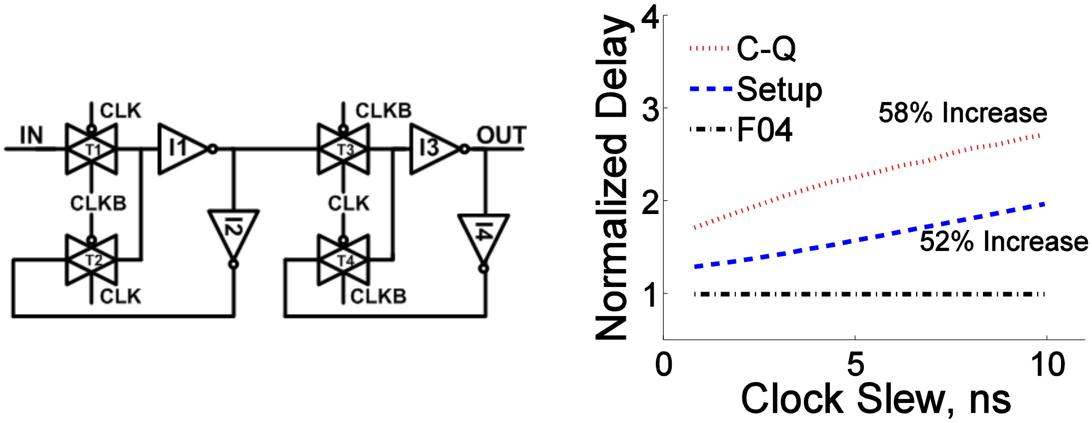
(a) **(b)**

Figure 46: (a) Sample logic path simulated with F04 NAND delays. (b) The impact of clock skew, clock slew and path length on frequency.

1. A case with 0.0 ns clock skew and 1.0 ns clock slew (Optimal);
2. A case with the clock skew that is five percent of the optimal period and 1.0 ns clock slew (Skew);
3. A case where the clock skew is five percent of the optimal period and 10.0 ns clock slew (Skew & Slew).

The 10.0 ns slew was chosen to reflect the maximum slew obtained from the clock tree that was designed in above threshold and then operated at subthreshold voltages. As expected, the addition of skew reduces the operating frequency. When both slew and skew are considered, F_{max} can be reduced from 14 to 28 percent, depending on the length of the logic path. In summary, as subthreshold systems target higher frequencies (i.e., as the logic path reduces), the effect of slew on cycle time cannot be ignored.

Since the focus of this analysis is based on clock slew, future discussion will be restricted to the impact of clock-slew variations. To understand the effect of clock slew directly, the impact that flip-flop timing metrics have on the setup and clock-to-q times has been analyzed in Figure 47. A commonly used transmission-gate flip-flop, used in subthreshold operation, is depicted in Figure 47 (a) [38]. The slew distribution directly affects the timing metrics in Equation (6.2) and these metrics can been seen in Figure 47 (b). The delays have been normalized to the delay of a FO4 gate, since the focus of this study is on the general behavior. When slew variations are applied to the clock signal, the setup time is directly proportional to it. In severe cases, the setup time can vary by 52



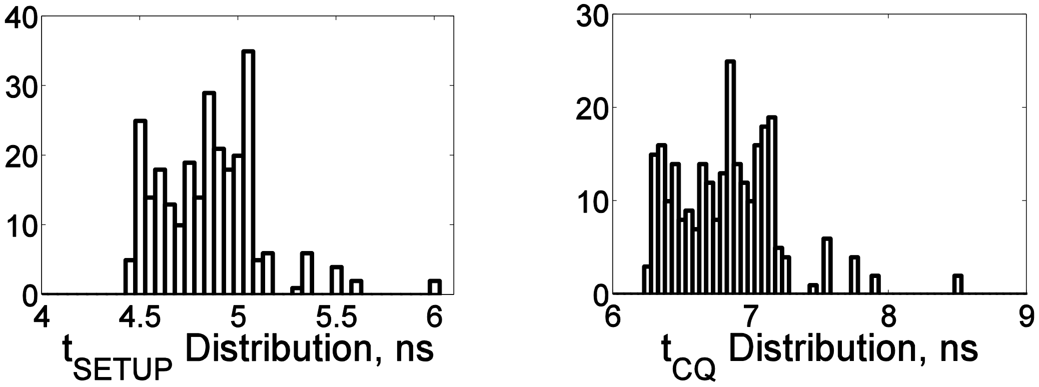
(a) **(b)**

Figure 47: (a) Transmission gate flip-flop used for subthreshold experiments. (b) Timing variations for a clock tree designed in above threshold lowered to subthreshold.

percent worse than the best case achieved. For the given slew distributions, the clock-to-q delay can be 58 percent worse than the best case value. This variation in setup and clock-to-q times reduces the time available to compute logic, and in some cases will cause errors in the logic by violating the setup-time requirements.

4.2.3 Deterministic Timing Variations

It is apparent that a clock-tree design, with inherent slew variations, has the potential to cause severe timing violations in subthreshold. In this section, it is shown that the design of the clock tree can cause a distribution of these timing variations (Figure 48). The distribution of the timing metrics that impact cycle time (i.e., t_{SETUP} and t_{CQ}) are plotted in Figure 48 (a) and (b). The setup and clock-to-q times are worsened by clock slew, as the deterministic variations show a wide scattering of values. Clock slew directly impacts timing metrics; a smaller slew variation translates to smaller setup and clock-to-q variations. By reducing the deterministic clock-slew variations, an optimal maximum frequency can be met.



(a) **(b)**

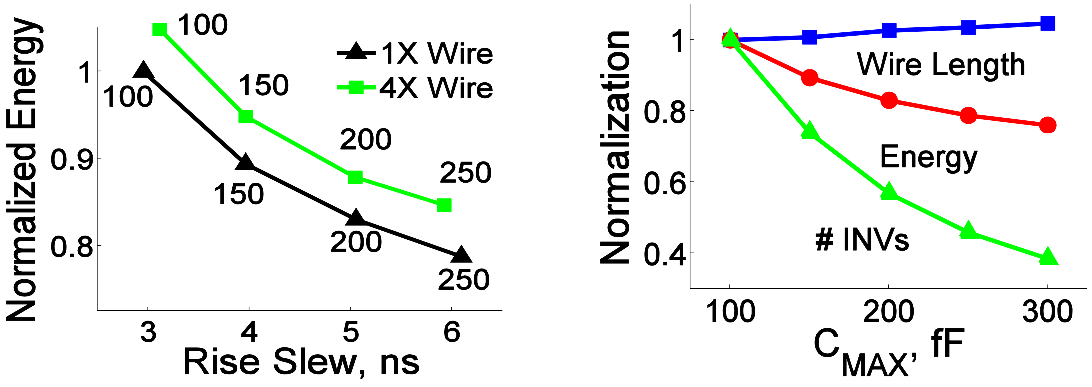
Figure 48: (a) Setup time and (b) clock-to-q distributions for a clock tree designed in above threshold and then the power supply is scaled to subthreshold voltages.

4.3 Techniques for Low-Energy, Slew-Controlled Subthreshold Design

The findings from Section 5.2 show that it is necessary to control and reduce the variations associated with clock slew to achieve robust subthreshold operation. The focus of this section is to investigate techniques for subthreshold clock-tree design that provides a smaller slew variation.

4.3.1 Smaller C_{MAX} Requirements in Subthreshold

Conventional methods for controlling clock slew currently exist and rely on limiting the maximum nodal capacitance (C_{MAX}) a buffer can drive [58-61]. In general, when a buffer reaches a node that exceeds the designated C_{MAX}, buffer insertion is required to reduce the load capacitance. In the analysis of an above-threshold clock tree operating in subthreshold, the C_{MAX} is defined by above-threshold methods. In above threshold, there is more transistor current available to drain the charge from the output node; therefore, a smaller C_{MAX} requirement should be imposed for optimal subthreshold clock trees. To examine the impact of varying C_{MAX} requirements on subthreshold clock trees, the energy, wirelength and inverter count dependence are shown in Figure 49.



(a) (b)

Figure 49: The summary of experimental methods to reduce the slew variations in subthreshold. (a) The increasing C_{MAX} increases slew and reduces power. The numbers indicate the fixed C_{MAX} in fF. (b) The impact of C_{MAX} constraint on wire length, energy and buffer count.

The results of a clock tree designed in subthreshold, with varying C_{MAX} from 100 to 250 fF, is shown in Figure 49 (a). As a reminder, C_{MAX} represents the maximum nodal capacitance a node can reach and in most cases the actual load capacitance is less than the C_{MAX}. The results of the figure show that it is possible to control the slew in subthreshold reducing the average rise slew from 6.0 ns to 3.0 ns. At the same time the slew is being reduced, the energy is increased by nearly 20 percent. This energy increase occurs because the sum of the interconnect capacitance remains constant and C_{MAX} is reduced. As a result, more buffers are added to compensate for the reduced C_{MAX}. The addition of buffers directly translates into more energy.

This same trend of increasing C_{MAX} and reducing energy and buffer count is shown in Figure 49 (b). Additionally, the total wire length of the design has small changes, because whenever an inverter is removed a small wire segment takes the place of the removed inverter. Note that going from a C_{MAX} of 100fF to 300fF, the number of inverters decreases by 60 percent, but the energy only decreases by nearly 20 percent. A large portion of this is unaffected energy comes from the large interconnect capacitance associated with the wire. In summary, to design a subthreshold clock tree, a smaller C_{MAX} is required in subthreshold compared to above threshold.

4.3.2 Minimum Wire Width in Subthreshold

The wire interconnect contributes to a significant portion of the energy in a clock network. Reducing the interconnect capacitance without sacrificing delay in the clock path can help address this energy component. When modeling the interconnect as a distributed resistance-capacitance (RC) line, the design rule of thumb is that RC wire delays should only be considered when the line being modeled has reached a critical length, L_{CRIT} [38]. The critical length is defined as

$$L_{CRIT} \gg \sqrt{\frac{t_d}{0.38rc}}, \qquad \text{Eq. (4.4)}$$

where t_d is the gate delay, r is the resistance per unit length and c is the capacitance per unit length. Based on the Equation (6.4), the L_{CRIT} for a 1.0 ns gate delay in 65nm PTM technology should be much greater than 4.0 mm. This critical length also assumes a minimum wire width to reduce the interconnect capacitance. A recent work showed a 65nm subthreshold chip with a 2.29mm x 1.86mm area [62]. Using these dimensions as a benchmark, the worst-case wire length is confirmed to be larger than 4.0 mm. Recent products in 65nm technology from Intel show similar results in regards to maximum wire lengths [63]. Since the length of the wire is expected to be less than the critical length, it is possible to neglect the distributed RC behavior of the wire. When $L << L_{crit}$, the wire can be modeled primarily as a lumped capacitance. This is possible in subthreshold because the device resistance is much higher than the wire resistance and the operating frequencies are low. Since RC wire delays are negligible, the wire resistance does not play a critical role in determining the delay and all wires can be designed with minimum width. In above threshold it is not possible to neglect RC wire delays because the operating frequencies are higher and the inverter delays are much smaller. The gate and wire delays are on the same magnitude, so larger wire widths are more appropriate in above-threshold designs.

The major advantage of reducing the wire width is the corresponding decrease in the wire capacitance. Referring back to Figure 49 (a), reducing the wire width in a subthreshold clock tree reduces the energy without sacrificing slew. The above-threshold clock tree used a wire size that was four times the minimum width, to ensure the RC delays were small. When this tree is lowered to subthreshold voltages, the large wire width only adds capacitance and energy to the system. This result can allow users to design a subthreshold clock tree with a minimum width, to reduce wire capacitance. Additionally, the wire resistance can be neglected in comparison to the driving inverters resistance

4.4 Optimizing Subthreshold Clock Trees with Dynamic C_{MAX} Methods

The previous section has provided several independent methods to reduce the clock slew (and clock-slew variation), while also considering the energy. While each method has their advantages, together they still cannot provide an optimal subthreshold clock tree for slew control. As a result, a new technique for optimal subthreshold clock-tree design will be proposed, based on a regressive C_{MAX} assignment. Since the proposed technique allows the C_{MAX} to vary dynamically from one node to another, the proposed method is called dynamic $C_{MAX.}$ The conventional approach, where the C_{MAX} is constant across all levels, will be referred to as fixed C_{MAX}. The dynamic C_{MAX} and fixed C_{MAX} clock trees were designed and compared across important design parameters: clock slew, clock skew, wirelength, and energy. The fixed C_{MAX} trees were designed in the above-threshold regime and the supply voltage was scaled to study their performance and energy in the subthreshold domain.

The following subsection will first explain the basic concept of dynamic C_{MAX}. Next, the simulated behavior of a broad selection of dynamic C_{MAX} clock trees will be presented and analyzed. The purpose of the analysis is to understand what traits the desired trees have in common, and use that information as a basis for future designs. Finally, the summarized results will be presented to show how a smaller C_{MAX}, wire width, and the use of dynamic C_{MAX} are best subthreshold clock-tree design. In summary, new design concepts are required for subthreshold clock-tree design and scaling the voltage of an above-threshold tree is not sufficient.

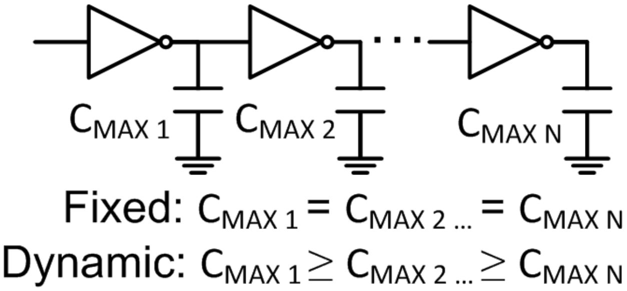
4.4.1 Principles of Dynamic C_{MAX}

From the results of Figure 49 (a), the clock-tree energy can be reduced by increasing the C_{MAX} that each node drives. In contrast, the clock-tree slew can be well controlled by reducing the C_{MAX} at each node. For the purposes of controlling slew, it is only important to reduce the slew variations at the sink nodes, which are directly attached to the flip-flops. The slew still needs to be controlled at other nodes, but the C_{MAX} constraint can be relaxed. By relaxing the C_{MAX} constraint, fewer inverters will be used, which will reduce the energy. To reduce energy while achieving a proper slew at the sink nodes, the C_{MAX} should vary at each level of the clock tree, becoming smaller as the nodes approach the sink. Additionally, if the wire width is assumed to be the minimum size, more energy savings can be achieved. The preceding methodology is the basis for dynamic $C_{MAX,}$ which can be seen in Figure 50.

The proposed technique (dynamic C_{MAX}) and a previous method (fixed C_{MAX}) for clock-tree design are compared in Figure 50. In the proposed technique of dynamic C_{MAX}, the C_{MAX} values selected are reduced from the source (buffer level '1') to the sink (buffer level 'N'). It is not guaranteed that the C_{MAX} values selected are the *exact* values that a level will drive. However, it is guaranteed that the C_{MAX} values will limit the *maximum capacitance* a level can drive.

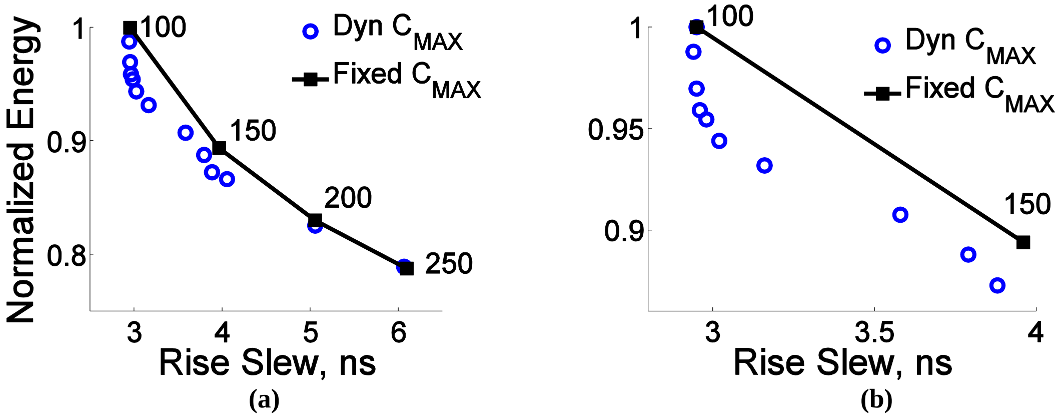
The potential advantage that a dynamic C_{MAX} method has compared to a fixed C_{MAX} method is shown in Figure 51. It is apparent that the figure represents an energy-robustness plane. In an optimal case, there is zero slew and energy, with data points near the origin.

Based on this, curves that are closer to the origin represent a better energy-robustness tradeoff. A set of dynamic C_{MAX} points is shown in Figure 51 (a), and a region where slew is small is magnified in Figure 51 (b). It is possible to reduce the



$C_{MAX\ 1}$
$C_{MAX\ 2}$
$C_{MAX\ N}$
Fixed: $C_{MAX\ 1} = C_{MAX\ 2} \ldots = C_{MAX\ N}$
Dynamic: $C_{MAX\ 1} \geq C_{MAX\ 2} \ldots \geq C_{MAX\ N}$

Figure 50: Dynamic C_{MAX} selection is the proposed technique to reduce slew variations at the sink nodes, while ensuring the energy does not increase.



(a) **(b)**

Figure 51: (a) Preliminary results of slew control and energy savings of Dynamic C_{MAX} Selection. (b) Results zoomed in for a rise slew of 3-5ns. The numbers indicate the fixed C_{MAX} in fF. By traversing the Dynamic C_{MAX} path, we can achieve better slew control at targeted power.

energy in the clock tree by nearly six percent while maintaining the same slew by employing dynamic C_{MAX} methods (Figure 51 (b)).

4.4.2 C_{MAX} Selection Trends of Clock Trees

To understand the trends of dynamic C_{MAX} clock trees, 79 trees were simulated and designed using the rules of the dynamic C_{MAX}. The trees were implemented to deliver clock to the r1 IBM benchmark, which has 267 sinks. Additionally, there are definitions for ten buffer levels and thus ten dynamic C_{MAX} selections required. The clock trees were generated using conventional low-skew, clock-tree generation methods [57, 64]. The details of the algorithm are presented in Section 6.5.2. The dynamic C_{MAX} values ranged from 100fF to 400fF, in 50fF intervals. All 79 trees were uniquely selected to provide a broad range of dynamic clock trees, so the results could provide insight into general trends. This is the best scenario for investigation, in lieu of analyzing all 7^{10} possible combinations.

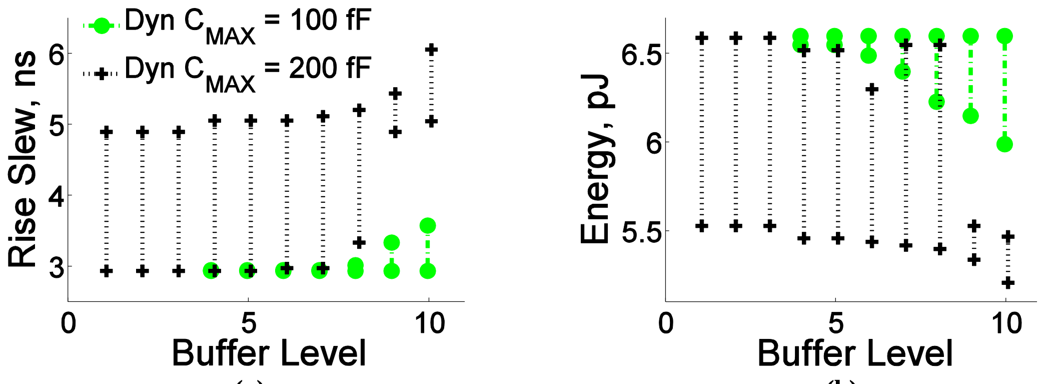
4.4.2.1 C_{MAX} Value Selection and Level Placement

The C_{MAX} value selected and the buffer level placement are critical factors in determining the performance of a dynamic clock tree. These critical factors were analyzed and their impact on clock slew and clock energy is investigated in Figure 52. Dynamic C_{MAX} values for each clock-tree level can affect the slew delivered to the latches as shown in Figure 52 (a). The endpoints (for a given level) denote the minimum and maximum ranges for slew when the dynamic C_{MAX} value has been placed. For example, of the 79 dynamic C_{MAX} trees selected, when level five has a C_{MAX} of 200fF, the final clock slew is in the range of 3.0 to 5.0 ns.

Since the results are reported on a per-level basis, it is not directly known how the remaining levels in the designated tree have been selected. That is, given that level five has a C_{MAX} value of 200fF (based on the Figure 52 (a)), the C_{MAX} values of levels one to four, and levels six to ten are unknown. Indirectly, it is known that the following holds true:

$$C_{MAX(1\ldots4)} \geq C_{MAX(5)} \geq C_{MAX(6\ldots10)}. \qquad \text{Eq. (4.5)}$$

The interesting feature to note for Figure 52 (a) is that when level five has a C_{MAX} value of 100fF, the slew is well controlled.



(a) (b)

Figure 52: The effect of independent dynamic C_{MAX} values and their impact on the (a) Final Clock Slew and (b) Total Clock Tree Energy.

Since 100fF is the smallest cap value selected, when level five is 100fF, levels six to ten are 100fF as well. This fact results in a well controlled slew delivered to the latches. Another interesting point to make requires looking at buffer level ten. When the C_{MAX} value selected at level ten is 100fF, there is a smaller range of values compared to when the C_{MAX} value is 200fF. This reinforces the concept that a smaller C_{MAX} value near the sink (buffer level ten) has the best chance to reduce the slew.

An inverse trend exists between the dynamic C_{MAX} values and energy in the clock tree (Figure 52 (b)). Although a smaller C_{MAX} at level ten can provide a lower slew at the latches, it can also increase the energy in the clock tree. The latter point is denoted by the range at level ten of 6.0 to 6.6 pJ, when the C_{MAX} is 100fF. In summary, a larger C_{MAX} value at level ten can yield a lower energy, by sacrificing the slew delivered to the latches.

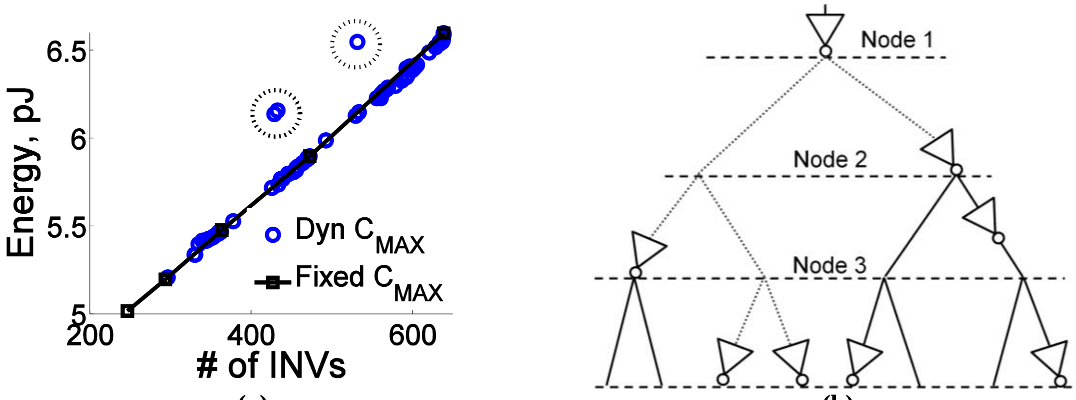
There is one last thing to note (from Figure 52): The further away from the sink level ten, the more disparity exists between minimum and maximum values for slew and energy. This occurs when the C_{MAX} value is 200fF, because there is more flexibility in the C_{MAX} values for the levels near the sink. Ultimately a C_{MAX} value closer to the sink (level ten) has more restrictions in the slew and energy compared to a C_{MAX} value near the source (levels one to three). In general, selecting a C_{MAX} value for each node can turn designing an optimal clock tree into a lesson in pure combinatorics. In summary, it should be possible to design an optimal tree using combinatorics and the previously described relationships between C_{MAX}, slew and energy.

4.4.2.2 Energy is directly related to Inverter Count

By changing the C_{MAX} values at different levels, the number of inverters in the clock tree is also being changed. The inverter count impact on the clock tree is examined in Figure 53.

The relationship between the inverter count and clock-tree energy is plotted in Figure 53 (a). For a fixed C_{MAX}, a constant C_{MAX} at each level was used. The direct trend hints that much of the energy associated in the clock trees are attributed to the inverters. In reality, this is not true via the earlier discussion of energy due to interconnect capacitance. As a reiteration, to control the number of inverters (and thus energy) in a clock tree, the C_{MAX} values can be selected at each level. This can be seen from the previously described Figure 52 (b).

To further explain the direct correlation of energy and inverter count, an example binary clock tree will be used (Figure 53 (b)). Based on the clock-tree network shown in Figure 53 (b), a merging node is defined as the junction where the clock tree splits into two directions. Buffer levels are at the output of a buffer, where C_{MAX} values need to be



(a) **(b)**

Figure 53: (a) The correlation of inverter count and clock tree energy (b) Example binary clock tree network with merging node definitions. With Dynamic C_{MAX} selection, different C_{MAX} values are selected at each buffer level, denoted by the buffer output.

defined. Recall from Figure 52 (b), when the C_{MAX} value at the clock-tree sink is 100fF, it has a higher energy than when the value is 200fF. There are two factors that contribute to this increase in energy. First, based on the nature of the binary-tree model, the clock nodes near the sink are guaranteed to have more inverters compared to the clock nodes near the source. Secondly, by reducing the CMAX values, more inverters are naturally introduced near the sink. Essentially, a large fraction of inverters are introduced by reducing C_{MAX} at the lowest level. Therefore, energy is strongly controlled by the C_{MAX} values, and more importantly, C_{MAX} values selected nearest to the sink.

4.4.2.3 Dynamic C_{MAX} can Reduce Slew, Skew and Wirelength

Knowing that inverter count is directly related to the clock-tree energy, there are also additional indirect trends. The clock slew (delivered to the latches) and inverter count are correlated to important metrics in clock-tree design: clock skew, slew, and wirelength. These important metrics are examined in Figure 54. The different points of the fixed C_{MAX} lines represent the trees generated by different fixed C_{MAX} values. When designing with a fixed C_{MAX}, there is an inverse relationship between clock skew and the clock slew (Figure 54 (a)). With the fixed C_{MAX} method, smaller C_{MAX} values are used at all levels to achieve smaller slew. For a given buffer size a smaller C_{MAX} at all levels requires more inverters in the tree. Additionally, a smaller C_{MAX} constraint results in larger deterministic skew (for a fixed C_{MAX}). If random process variations are considered, a larger number of inverters imply more sources of variations, which could further increase the random variation in skew. In summary, when designed using a fixed

C_{MAX}, there exists an inverse relationship between slew and skew: when the C_{MAX} is less, slew is larger and skew is smaller.

The observed inverse trend of skew and slew for fixed C_{MAX} trees is supported by Figure 54 (b) and (c). From a design perspective, a tradeoff is required when using the fixed C_{MAX} method: it is only possible to achieve low slew or low skew. The observed trends for a dynamic C_{MAX} tree are different. Similar to the fixed C_{MAX} method, the slew reduces with increasing the number of inverters (assuming same buffer sizes for a single design). By employing various dynamic C_{MAX} topologies, the number of inverters will increase and a reduced skew design can be achieved. Essentially, using dynamic C_{MAX} allows more degrees of freedom in the placement of inverters because different levels

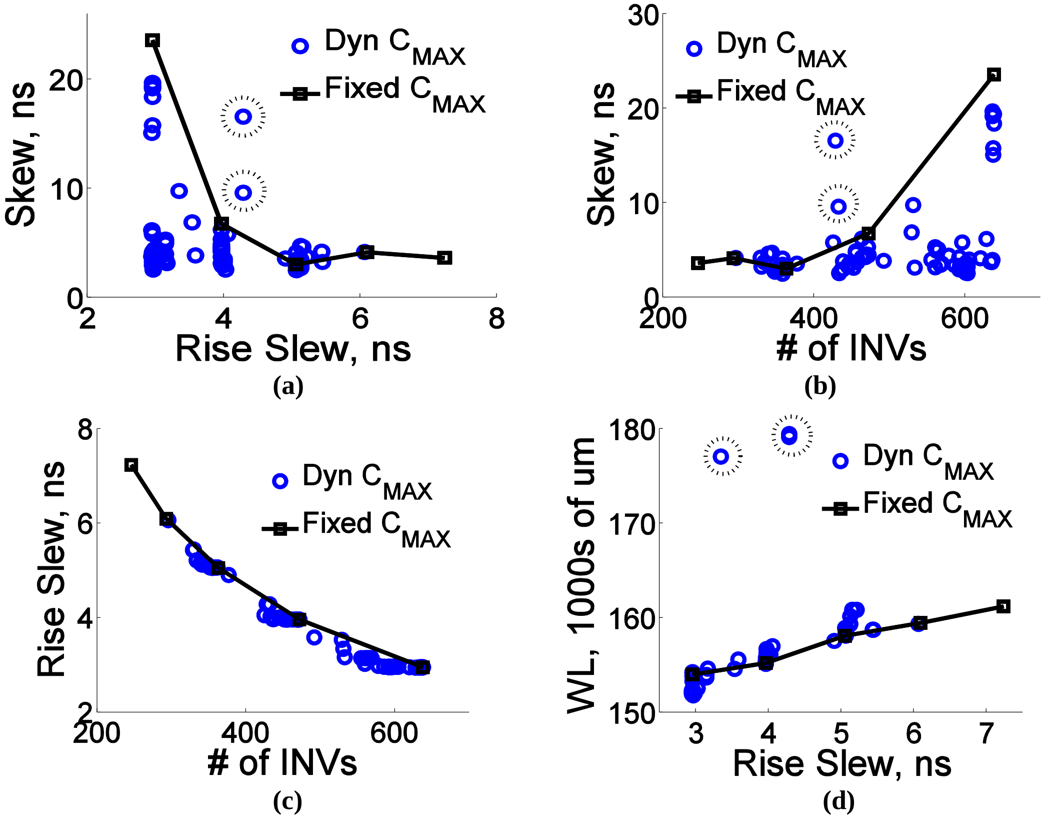


Figure 54: The indirect relationship between the of clock slew (delivered to the latches) with (a) clock skew and (d) wirelength. The relationship between inverter count with (b) skew and (c) slew.

have different C_{MAX} constraints. Consequently, there also exist several instances of clock trees (for the same design) with varying skew. As an example from Figure 54 (b), when the number of inverters is nearly 600, a dynamic C_{MAX} tree can achieve a slew in the range of 2.0 to 20 ns. With an optimal selection of dynamic C_{MAX} values, it is possible to achieve low slew and low skew by reducing the C_{MAX} near the sink, and increasing the C_{MAX} near the source (Figure 54 (a)).

The wirelength is correlated with the slew delivered to the latches as shown in Figure 54 (d). The increase in the wirelength with an increase in the slew can be explained by the absence of inverters. As stated before, when the C_{MAX} values are large, there are fewer inverters since each inverter has a larger limit of capacitance it can drive. With fewer inverters, the slew will tend to increase, because there is less control. Additionally, with fewer inverters there will be more wirelength to compensate for the removed inverters. If the fixed C_{MAX} and dynamic C_{MAX} methods are compared, there are a few points to consider. First, both methods follow the trend that a larger slew is correlated with a longer wirelength. Secondly, for smaller slew targets, it is possible achieve a smaller wirelength using dynamic C_{MAX} methods compared to fixed C_{MAX}. This can be seen by viewing all the figures where the points are located around 3.0 ns of slew.

4.4.2.4 Explanation of Outliers

During the preceding discussion of dynamic clock-tree trends, there were outliers that require explanation. These points can be seen from the dotted circles in Figure 54.

Since dynamic C_{MAX} can only limit the maximum capacitance a level can drive, it has no bearing on the minimum capacitance. For example, if

$$C_{MAX(6\ldots10)} \geq 300fF, \qquad \text{Eq. (4.6)}$$

it is not guaranteed that

$$C_6 \geq C_7 \geq C_8 \geq C_9 \geq C_{10}, \qquad \text{Eq. (4.7)}$$

where C_i denotes the actual capacitance at the level (after the clock tree has been routed). When the actual capacitance does not follow the trend in Equation (4.7), the slew is not well controlled and can become unpredictable. An algorithm to control the upper and lower capacitance values would assure optimal design.

4.4.2.5 Limitations of Buffer Reduction

In Section 5.4.2.2, the direct correlation between energy and inverter count was investigated. While it is true that shallower trees may be better for skew as described in [65], there may be secondary effects of slew depending on the size of the circuit. In an attempt to reduce the overall power by further reducing the number of inverters, a one-buffer h-tree was designed to investigate the limitations. A comparison of this design and the previously described designs can be seen in Table 5. The skew is improved using a one-buffer h-tree compared to the previous options, but the slew has increased drastically. In the clock-tree designs, the large size of the IBM r1 benchmark circuit has contributed to a long wire length and thus extra capacitance. Additionally, a large buffer has been

Table 5: Methodology Comparison for IBM r1 Benchmark

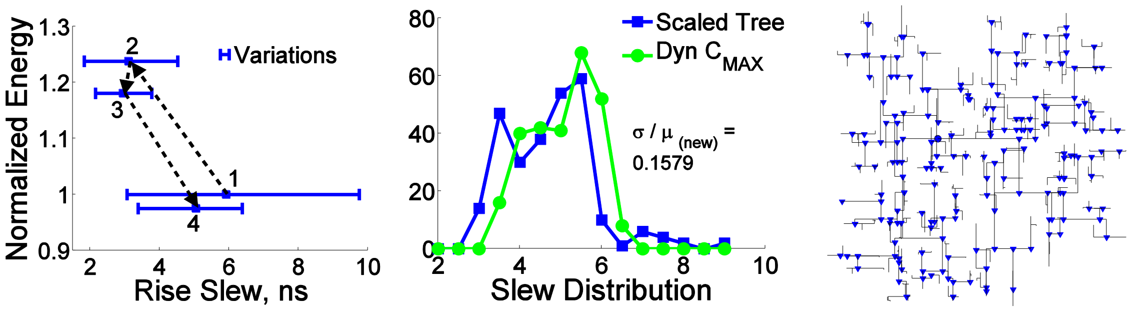
Sub-V_T Design Method	Energy (pJ)	Max Slew (ns)	Mean Slew (ns)	Skew (ns)	WL (um)
Fixed C_{MAX} [58-61]	6.60	3.77	2.95	25.9	154 048
One-Buffer H-Tree [65]	5.89	39.64	38.15	2.77	151 727
Dynamic C_{MAX} [This Work]	5.47	6.38	5.05	3.39	158 954

introduced in the one-buffer h-tree to accommodate the large design and the energy is still larger compared to dynamic C_{MAX} tree. In essence, the one-buffer h-tree is best for minimum skew and is appropriately used for small scale designs. However, for large scale designs, using a clock-tree with a dynamic C_{MAX} design, it is possible to achieve a near optimal tree with reduced slew, skew, and minimum energy.

4.4.3 Summary and Results of Dynamic C_{MAX}

In this section, the effects of using a smaller C_{MAX}, a reduced wire width, and dynamic C_{MAX} are summarized. Several subthreshold clock trees were implemented in 65nm PTM CMOS (V_{DD} = 300mV), and the best were selected. The summary of the slew reduction methods, comparison with the original design and clock-tree path are shown in Figure 55. A summary of the combined approaches to reduce the deterministic slew variations is shown in Figure 55 (a). The point '1' represents the aforementioned clock tree that was designed in above threshold and then scaled to subthreshold voltages. The details of C_{MAX} values used for these trees are shown in Table 6. The results are depicted for rise slew information, but a similar trend exists for fall slews. The techniques to reduce energy and slew were applied as follows:

1. From '1' to '2', the C_{MAX} was reduced from 250 to 100fF and the tree was redesigned using a fixed C_{MAX} method;
2. From '2' to '3', the wire width was reduced from four to one times the minimum size, and then redesigned with a fixed C_{MAX} method; and
3. From '3' to '4', dynamic C_{MAX} methods were employed with a minimum wire width.



(a) **(b)** **(c)**

Figure 55: (a) Summary of combined approaches to reduce deterministic slew variations in subthreshold. (b) Dynamic C_{MAX} deterministic slew distribution compared with the original clock tree. The new slew distribution has a coefficient of variation of 0.1579, better than the original 0.2309 of Figure 45. (c) Clock tree '4' routed using Dynamic C_{MAX} and combined methods.

With a starting point of '1' and ending point of '4', this proves that it is possible to reduce the slew without increasing power in a subthreshold clock tree. The final clock tree ('4') has smaller slew variations compared to the original above-threshold design (Figure 55 (a)).

The dynamic C_{MAX} tree slew distribution is compared with the original tree in Figure 55 (b). The dynamic C_{MAX} tree has slew variations in the range of 3.0 to 6.0 ns. In the above-threshold design, the slew variations were in the range of 3.0 to 10.0 ns. Additionally, the coefficient of variation for the new clock tree is 0.1579 compared to the larger dispersion of 0.2309 (Figure 45 (a)). The final dynamic clock tree, implemented for the r1 IBM testbench, was placed and routed in Figure 55 (c).

Table 6: Summary of Fixed and Dynamic C_{MAX} Values with Slew Reduction Techniques

Tree	1 Source	2 to 5	6	7	8	9	10 Sink	Norm. Energy	μ Slew (ns)
#1 (4x WL)	250 fF	250 fF	250 fF	250 fF	250 fF	250 fF	250 fF	1.00	5.92
#2 (4x WL)	100 fF	100 fF	100 fF	100 fF	100 fF	100 fF	100 fF	1.23	3.11
#3 (1x WL)	100 fF	100 fF	100 fF	100 fF	100 fF	100 fF	100 fF	1.18	2.95
#4 (1x WL)	300 fF	300 fF	200 fF	200 fF	200 fF	200 fF	200 fF	0.97	5.05

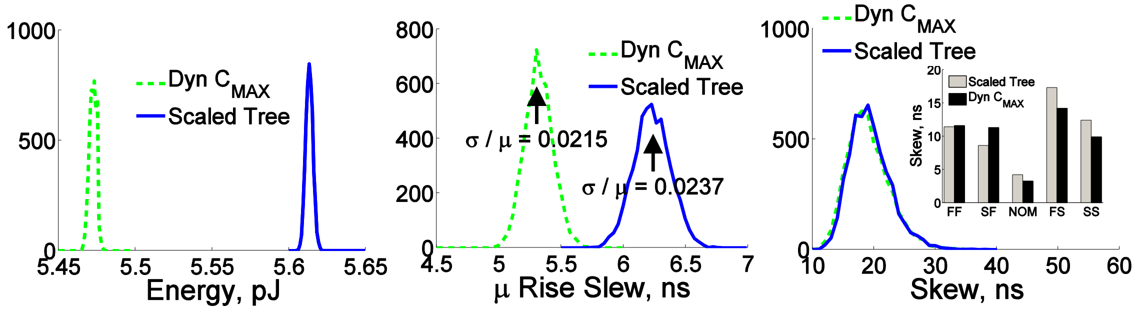
4.5 Discussions

4.5.1 Process, Voltage, and Temperature (PVT) Variations

Random slew variations are induced by process variations and have the potential to be more severe in designs because they add to deterministic slew variations. In the subthreshold regime, the local variability due to random dopant fluctuations (RDF) effects can dominate the device threshold (V_{TH}) variability [52]. Therefore, local variability in the clock buffers will be considered while simulating the random slew variations.

Using the learned techniques to design a dynamic clock tree, it is possible to reduce the clock-tree slew variations without an energy penalty. While this has remained a focus, maintaining a stable design under process, voltage and temperature variations are extremely important in subthreshold designs. In this section dynamic C_{MAX} clock trees and the impact of PVT variations are studied. The metrics to measure the impact are the clock slew, clock skew and clock-tree energy.

To model the effects of process variations in subthreshold, an independent variation was applied to the threshold voltage of each transistor in the clock tree. A 5000 point Monte Carlo (MC) simulation was performed using a Gaussian distribution. The 3σ value used was +/- 10 percent of the nominal V_{TH}. The impact of the independent transistor V_{TH} variations are plotted in Figure 56. The results for a dynamic clock tree (using dynamic C_{MAX}) and a scaled clock tree (using fixed C_{MAX}) are shown for comparison. The clock-tree energy and average slew as a result of the MC simulation under process variations is depicted in Figure 56 (a) and (b). There is a clear advantage to

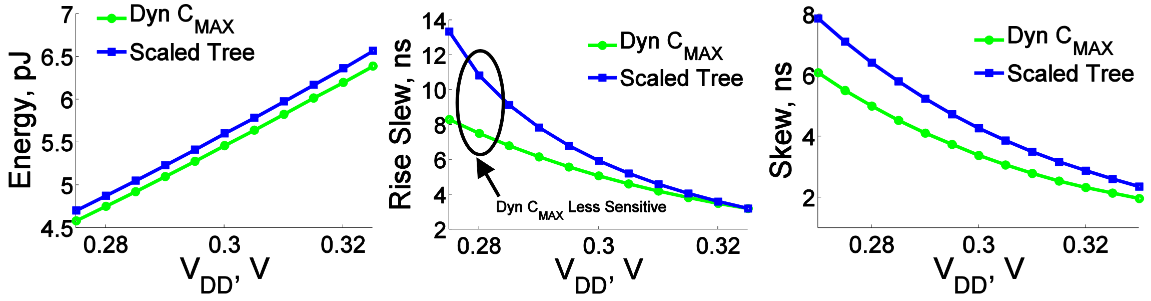


(a) (b) (c)

Figure 56: Monte Carlo simulation of varying threshold voltages for clock tree (a) Energy, (b) Slew and (c) Skew. The slew variations are reported as an average and the skew variations are reported as a worst case for each Monte Carlo simulation.

designing a tree using a dynamic C_{MAX} method, because it maintains a reduced energy and slew under variations. The MC worst-case skew and average skew variations at all process corners is shown in Figure 56 (c). From the results of Figure 56, compared to the scaled clock tree, a dynamic clock tree with smaller energy, slew and skew can be designed.

A supply voltage sweep of +/- 20 percent of the nominal V_{DD} provides a broad range of values to examine the clock-tree response. The results of the voltage sweep around a 300 mV supply are shown in Figure 57. The resulting supply voltage sweep impacts the clock tree energy, with the dynamic clock tree always more energy efficient than the scaled tree (Figure 57 (a)). The average clock slew, shown in Figure 57 (b), is more interested because it has a non-linear dependence on voltage. At lower voltages the slope of the dynamic clock tree is smaller than the slope of the scaled tree. This means that in subthreshold, the clock slew is less sensitive to supply voltage variations when designed with dynamic C_{MAX} methods. The clock skew dependence on voltage is seen in Figure 57 (c). In general, dynamic C_{MAX} is more robust across subthreshold supply variations.



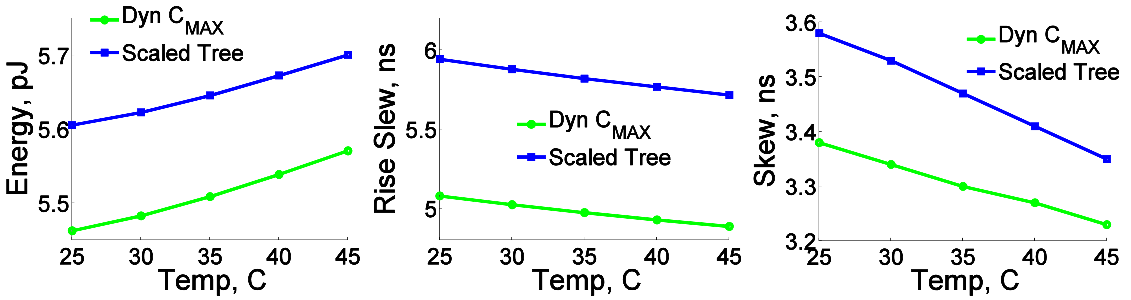
(a) (b) (c)

Figure 57: Supply voltage sweep and its impact on clock tree (a) Energy, (b) Slew and (c) Skew. At low voltages, the Dynamic C_{MAX} tree is less sensitive to supply voltage variations.

The temperature variations affect the clock energy, slew and skew as seen in Figure 58. The temperature was changed from 25 to 45 degrees Celsius. As expected, increasing the temperature reduces the skew and slew while also increasing the energy. Overall the dynamic C_{MAX} tree has lower energy, slew and skew even under different temperature conditions.

4.5.2 Algorithm for Clock-Tree Generation

The clock-routing algorithm defined in this chapter includes two major steps: abstract tree generation, and slew-aware buffering with embedding. Given a set of clock sinks, an abstract tree is generated based on the method of means and medians algorithm [66]. The objective of abstract tree generation is to decide the connection among the sink



(a) (b) (c)

Figure 58: The (a) Energy, (b) Slew and (c) Skew response to a temperature sweep shows that a dynamic clock tree is the preferable design.

nodes, internal nodes and clock source, while minimizing the wirelength. The routing topology and geometric locations of all the nodes are determined by a two-phase, slew-aware buffering and embedding method.

The clock-routing algorithm follows the classic deferred-merging and embedding flow in the above-threshold clock-network design [64]. The major difference is that buffers are inserted during the clock routing as well. First, the abstract tree is traversed by a bottom-up manner. For a pair of nodes, a set of feasible candidate solutions are created for their parent node, including the merging distances and merging styles. This bottom-up phase aims at generating zero-skew solutions, and inserting buffers. The buffer insertion is so that loading capacitance of each buffer does not exceed the user-specified maximum value (C_{MAX}). The second phase is to choose the optimum solution among the candidates by visiting the abstract tree in a top-down order. The outcomes are the entire clock-routing topology with the exact locations of the internal nodes, buffers and the clock source. The steps of the clock tree algorithmn are summarized in Figure 59.

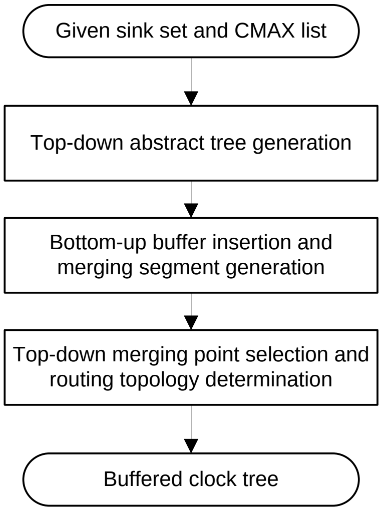

Given sink set and CMAX list
Top-down abstract tree generation
Bottom-up buffer insertion and merging segment generation
Top-down merging point selection and routing topology determination
Buffered clock tree

Figure 59: Algorithm adapted to design dynamic clock trees. This method is specifically used as a zero skew method for above threshold trees. However, with the approaches outlined in this text, a custom subthreshold algorithm should be implemented in the future.

4.6 Summary

In this chapter, the impact of slew variations on a subthreshold clock tree has been explained. By ignoring the impact the slew variations, the flow of data in a logic path can become corrupt. This notion provided the motive to design an optimal subthreshold clock tree with slew control. The following guidelines should be used when designing an optimal clock tree in subthreshold with slew control:

1. The maximum allowable nodal capacitance should be small in subthreshold;
2. Minimum wire sizes should be used at all times; and
3. The maximum nodal capacitance can be controlled dynamically to allow more slew propagation near the root of the tree while saving power.

On the other hand, near the sink nodes the maximum nodal capacitance should be reduced to better control the slew. A systematic approach has been presented, combining the above three guidelines for subthreshold clock tree design that has the potential to reduce the timing metric variations. Additionally, the guidelines will also retain the power advantage of using subthreshold design.

CHAPTER 5

CONCLUSIONS

The main objective of this thesis was to improve the energy efficiency of wireless biomedical systems by employing digital design techniques. Specifically, the development of a wireless electroencephalography (EEG) system was studied. Previous works and methodologies have been presented to optimize the individual components such as the low noise amplifiers, analog-to-digital converters and radio frequency transceivers. In designing the individual components, literature has shown that the high volume of data transmitted contributes to majority of the system level power. This work took a system level approach to reducing the power, by designing a digital processor. In designing the digital processor, three major contributions were made:

- An Accuracy and Energy Aware System for Adaptive Data Compression
- Chameleon: A Content-Aware Adaptive Compression Architecture for Wireless Electroencephalography
- Analysis and Design of Energy and Slew-Aware Subthreshold Clock Tree Systems

The first step in reducing the overall system power involved designing an adaptive data compression algorithm specific to EEG signals. The algorithm alters the compression rate of the EEG signals by detecting epileptic behavior in real-time. By using the proposed algorithm, EEG signals can be compressed by up to 8 X during background behavior, while delivering a perfectly reconstructed signal during epileptic spikes. A MATLAB system level model was developed to ensure the correctness, verify the potential data savings. The proposed algorithm was also extrapolated to a 32-channel system and a methodology was presented to solving for optimal energy efficiency.

The second step in reducing the wireless EEG system power involved designing a fully functional digital processor for detection, compression and transmission. The proposed processors works with commercial-off-the-shelf components, and its functionality was verified using an FPGA-Transceiver system. Using digital design techniques, a 9 X reduction in *system* power was achieved (compared to conventional methods) by implementing several low power optimizations: sleep mode power cycling, clock gating, and technology scaling, near threshold computing, burst mode processing and power gating. The results serve as a basis moving forward for digital design in the low power wireless system domain.

The last step in reducing the wireless EEG system involves designing a robust clock network to facilitate the digital processing in the low voltage domains. As the supply voltage is reduced, significant power gains can be made within the digital processor (up to 11.1 X savings). As the frequency of operation reaches in the MHz realm, it is important that the clock signals delivered are reliable. The reliability of the low power digital processor is directly correlated to the slew rate of the clock signals. To reduce the probability of timing failures, low voltage clock tree systems were analyzed, and a methodology to reduce the power while improving the robustness was presented. With the proposed clock tree methodology, the digital clock power can be reduced by 5 percent while the slew variations are improved by 40 percent.

To progress the research forward, future work suggests investigating serialized processing to further reduce the digital processing power. This thesis presented the design of a digital processor for the processing of low-throughput wireless electroencephalography signals. EEG signals are acquired and processed at low-throughput frequencies because of the small frequencies contained within EEG signals (i.e. 100's of Hertz). The proposed processor was parallel and pipelined, allowing for the flexibility of frequency ranges with various biomedical systems. To further reduce the digital power, one could investigate serialized architectures that are specific for low-throughput domains. This could perhaps improve the energy efficiency and reduce the leakage power by maximizing the potential of power gating and sleep modes.

A long term goal of this research is to implement a multi-channel EEG system prototype in hardware. The proposed digital processor cannot be efficiently scaled for a many channel system mainly because of the deep FIR data path required to process one channel. As a result, ADCs, low noise amplifiers and the proposed processor must be added, as each new channel is added to the EEG system. It would be beneficial to investigate a generic architecture for a single digital processor that can be adapted to operate in when multiple acquisition nodes are added to the system. The feasibility of a mobile network of processors also invites issues regarding the communication protocol.

In general, several research efforts have advanced the knowledge of how low power, wireless biomedical systems can be efficiently and effectively designed. This thesis marks one of many contributions to the field, supporting the overarching goal of providing reliable tools for the wireless monitoring of biomedical signals. With the constant effort to reduce power, while increasing the processing capabilities, it is reasonable to expect wireless biomedical systems to break through into the consumer markets within the next ten years.

REFERENCES

[1] P. Irazoqui-Pastor, I. Mody, and J. W. Judy, "In-vivo EEG recording using a wireless implantable neural transceiver," in *Neural Engineering, 2003. Conference Proceedings. First International IEEE EMBS Conference on*, 2003, pp. 622-625.

[2] A. J. Casson and E. Rodriguez-Villegas, "Data reduction techniques to facilitate wireless and long term AEEG epilepsy monitoring," in *EMBS Conference on Neural Engineering*, 2007, pp. 298-301.

[3] D. C. Yates and E. Rodriguez-Villegas, "A key power trade-off in wireless EEG headset design," in *EMBS Conference on Neural Engineering*, 2007, pp. 453-456.

[4] A. C. Metting Van Rijn, A. Peper, and C. A. Grimbergen, "A wireless infrared link for a 16-channel EEG telemetry system," in *Engineering in Medicine and Biology Society, 1994. Engineering Advances: New Opportunities for Biomedical Engineers. Proceedings of the 16th Annual International Conference of the IEEE*, 1994, pp. 906-907 vol.2.

[5] N. Raoult, J. F. Diouris, A. Sharahia, and L. Senhadji, "Experimental measurements for the design of wireless transmission in a hospital for a continuous monitoring of epileptic patients," in *Engineering in Medicine and Biology Society, 1996. Bridging Disciplines for Biomedicine. Proceedings of the 18th Annual International Conference of the IEEE*, 1996, pp. 286-287 vol.1.

[6] M. Modarreszadeh and R. N. Schmidt, "Wireless, 32-channel, EEG and epilepsy monitoring system," in *Engineering in Medicine and Biology Society, 1997. Proceedings of the 19th Annual International Conference of the IEEE*, 1997, pp. 1157-1160 vol.3.

[7] L. Chin-Teng, C. Yu-Chieh, H. Teng-Yi, C. Tien-Ting, K. Li-Wei, L. Sheng-Fu, H. Hung-Yi, H. Shang-Hwa, and D. Jeng-Ren, "Development of Wireless Brain Computer Interface With Embedded Multitask Scheduling and its Application on Real-Time Driver's Drowsiness Detection and Warning," *Biomedical Engineering, IEEE Transactions on,* vol. 55, pp. 1582-1591, 2008.

[8] N. Verma, A. Shoeb, J. Bohorquez, J. Dawson, J. Guttag, and A. P. Chandrakasan, "A Micro-Power EEG Acquisition SoC With Integrated Feature Extraction Processor for a Chronic Seizure Detection System," *Solid-State Circuits, IEEE Journal of*, vol. 45, pp. 804-816, 2010.

[9] C. Chiu-Kuo, E. Chua, T. Shao-Yen, F. Chih-Chung, and F. Wai-Chi, "Implementation of a hardware-efficient EEG processor for brain monitoring systems," in *SOC Conference (SOCC), 2010 IEEE International*, 2010, pp. 164-168.

[10] L. Brown, J. van de Molengraft, R. F. Yazicioglu, T. Torfs, J. Penders, and C. Van Hoof, "A low-power, wireless, 8-channel EEG monitoring headset," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, 2010, pp. 4197-4200.

[11] S. R. Sridhara, "Ultra-low power microcontrollers for portable, wearable, and implantable medical electronics," in *Design Automation Conference (ASP-DAC), 2011 16th Asia and South Pacific*, 2011, pp. 556-560.

[12] Y. Chung-Ping, L. Sheng-Fu, C. Da-Wei, L. Yi-Cheng, S. Fu-Zen, and H. Chao-Hsien, "A Portable Wireless Online Closed-Loop Seizure Controller in Freely Moving Rats," *Instrumentation and Measurement, IEEE Transactions on,* vol. 60, pp. 513-521, 2011.

[13] Neurosky Mindset Product Guide [Online]. Available: http://www.neurosky.com/Products/MindSet.aspx

[14] Emotiv EEG Neuroheadset [Online]. Available: http://www.emotiv.com/store/hardware/epoc-bci-eeg/developer-neuroheadset/

[15] B-Alert X10 Product Specification [Online]. Available: http://www.biopac.com/Manuals/b-alert_product_sheet.pdf

[16] S. Filipe, G. Charvet, M. Foerster, J. Porcherot, J. F. Beche, S. Bonnet, P. Audebert, G. Regis, B. Zongo, S. Robinet, C. Condemine, C. Mestais, and R. Guillemaud, "A wireless multichannel EEG recording platform," in *Engineering in Medicine and Biology Society,EMBC, 2011 Annual International Conference of the IEEE*, 2011, pp. 6319-6322.

[17] J. Penders, R. F. Yazicioglu, J. van de Molengraft, S. Patki, T. Torfs, L. Brown, and C. Van Hoof, "Wireless EEG systems: Increasing functionality, decreasing power," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, 2010, pp. 3441-3441.

[18] A. Y. Wang and C. G. Sodini, "A simple energy model for wireless microsensor transceivers," in *IEEE Global Telecommunications Conference*, 2004.

[19] A. J. Casson, D. C. Yates, S. Patel, and E. Rodriguez-Villegas, "Algorithm for AEEG data selection leading to wireless and long term epilepsy monitoring," in *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, 2007, pp. 2456-2459.

[20] A. J. Casson and E. Rodriguez-Villegas, "On data reduction in EEG monitoring: Comparison between ambulatory and non-ambulatory recordings," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, 2008, pp. 5885-5888.

[21] A. Y. Wang and C. G. Sodini, "On the Energy Efficiency of Wireless Transceivers," in *Communications, 2006. ICC '06. IEEE International Conference on*, 2006, pp. 3783-3788.

[22] R. F. Yazicioglu, P. Merken, R. Puers, and C. Van Hoof, "A 200 uW Eight-Channel EEG Acquisition ASIC for Ambulatory EEG Systems," *Solid-State Circuits, IEEE Journal of,* vol. 43, pp. 3025-3038, 2008.

[23] N. Verma and A. P. Chandrakasan, "An Ultra Low Energy 12-bit Rate-Resolution Scalable SAR ADC for Wireless Sensor Nodes," *Solid-State Circuits, IEEE Journal of,* vol. 42, pp. 1196-1205, 2007.

[24] G. W. Williams, H. O. Luders, A. Brickner, M. Goormastic, and D. W. Klass, "Interobserver variability in EEG interpretation," *Neurology,* vol. 35, pp. 1714-1719, 1985.

[25] B. Baas, Y. Zhiyi, M. Meeuwsen, O. Sattari, R. Apperson, E. Work, J. Webb, M. Lai, T. Mohsenin, D. Truong, and J. Cheung, "AsAP: A Fine-Grained Many-Core Platform for DSP Applications," *Micro, IEEE*, vol. 27, pp. 34-45, 2007.
[26] R. Sarpeshkar, "Analog Versus Digital: Extrapolating from Electronics to Neurobiology," *Neural Computation*, vol. 10, pp. 1601-1638, 1998.
[27] A. J. Casson and E. Rodriguez-Villegas, "Generic vs custom; analogue vs digital: On the implementation of an online EEG signal processing algorithm," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, 2008, pp. 5876-5880.
[28] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 11, pp. 674-693, 1989.
[29] G. Antoniol and P. Tonella, "EEG data compression techniques," *Biomedical Engineering, IEEE Transactions on*, vol. 44, pp. 105-114, 1997.
[30] K. Cheng-Wen, L. Yue-Der, C. Hsiao-Wen, and J. Gwo-Jen, "An EEG spike detection algorithm using artificial neural network with multi-channel correlation," in *Engineering in Medicine and Biology Society, 1998. Proceedings of the 20th Annual International Conference of the IEEE*, 1998, pp. 2070-2073 vol.4.
[31] K. P. Indiradevi, E. Elias, P. S. Sathidevi, S. Dinesh Nayak, and K. Radhakrishnan, "A multi-level wavelet approach for automatic detection of epileptic spikes in the electroencephalogram," *Computers in Biology and Medicine*, vol. 38, pp. 805-816, 2008.
[32] *Klinik Für Epileptologie [Online]*. Available: www.meb.uni-bonn.de
[33] H.-L. Chan, M.-A. Lin, T. Wu, S.-T. Lee, Y.-T. Tsai, and P.-K. Chao, "Detection of neuronal spikes using an adaptive threshold based on the max-min spread sorting method," *Journal of Neuroscience Methods*, vol. 172, pp. 112-121, 2008.
[34] NRF24L01 Product Specification [Online]. Available: http://www.nordicsemi.com/eng/nordic/download_resource/8041/1/40475824
[35] B. Zhai, D. Blaauw, D. Sylvester, and K. Flautner, "Theoretical and practical limits of dynamic voltage scaling," in *Design Automation Conference*, San Diego, CA, 2004, pp. 868-873.
[36] B. H. Calhoun, A. Wang, and A. Chandrakasan, "Modeling and sizing for minimum energy operation in subthreshold circuits," *IEEE Journal of Solid-State Circuits*, vol. 40, pp. 1778-1786, 2005.
[37] *Predictive Technology Model [Online]*. Available: http://www.eas.asu.edu/~ptm
[38] J. M. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits: A Design Perspective*, Second ed.: Prentice Hall, 2003.
[39] G. Higgins, S. Faul, R. P. McEvoy, B. McGinley, M. Glavin, W. P. Marnane, and E. Jones, "EEG compression using JPEG2000: How much loss is too much?," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, 2010, pp. 614-617.
[40] Virtex-5 Family Overview [Online]. Available: http://www.xilinx.com/support/documentation/data_sheets/ds100.pdf

[41] Texas Instruments OPA333 Precision Amplifier [Online]. Available: http://www.ti.com/lit/ds/symlink/opa333.pdf

[42] Texas Instruments ADC 7866 Product Specification [Online]. Available: http://www.ti.com/lit/ds/symlink/ads7866.pdf

[43] Nangate 45nm Open Cell Library [Online]. Available: http://www.nangate.com

[44] A. Wang and A. Chandrakasan, "A 180-mV subthreshold FFT processor using a minimum energy design methodology," *Solid-State Circuits, IEEE Journal of,* vol. 40, pp. 310-319, 2005.

[45] A. Wang, A. Chandrakasan, and S. Kosonocky, "Optimal supply and threshold scaling for subthreshold CMOS circuits," 2002, pp. 5-9.

[46] B. H. Calhoun and A. Chandrakasan, "Characterizing and Modeling Minimum Energy Operation for Subthreshold Circuits," in *Low Power Electronics and Design, 2004. ISLPED '04. Proceedings of the 2004 International Symposium on*, 2004, pp. 90-95.

[47] B. C. Paul, A. Raychowdhury, and K. Roy, "Device optimization for digital subthreshold logic operation," *Electron Devices, IEEE Transactions on,* vol. 52, pp. 237-247, 2005.

[48] N. Hedenstierna and K. O. Jeppson, "CMOS Circuit Speed and Buffer Optimization," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on,* vol. 6, pp. 270-281, 1987.

[49] J. R. Tolbert and S. Mukhopadhyay, "Accurate buffer modeling with slew propagation in subthreshold circuits," in *Quality of Electronic Design, 2009. ISQED 2009. Quality of Electronic Design*, 2009, pp. 91-96.

[50] N. Magen, A. Kolodny, U. Weiser, and N. Shamir, "Interconnect-power Dissipation in a Microprocessor," in *Proceedings of the 2004 international workshop on System Level Interconnect Prediction*, Paris, France, 2004, pp. 7-13.

[51] T. Sakurai and A. R. Newton, "Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas," *Solid-State Circuits, IEEE Journal of,* vol. 25, pp. 584-594, 1990.

[52] B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, "Analysis and mitigation of variability in subthreshold design," in *Low Power Electronics and Design, 2005. ISLPED '05. Proceedings of the 2005 International Symposium on*, 2005, pp. 20-25.

[53] J. Kwong and A. P. Chandrakasan, "Variation-driven device sizing for minimum energy sub-threshold circuits," in *Proceedings of the 2006 International Symposium on Low Power Electronics and Design*, 2006, p. 13.

[54] N. Jayakumar and S. P. Khatri, "A variation-tolerant sub-threshold design approach," in *Design Automation Conference, 2005. Proceedings. 42nd*, 2005, pp. 716-719.

[55] N. Lotze, M. Ortmanns, and Y. Manoli, "Variability of flip-flop timing at sub-threshold voltages," in *Low Power Electronics and Design (ISLPED), 2008 ACM/IEEE International Symposium on*, 2008, pp. 221-224.

[56] N. Verma, J. Kwong, and A. P. Chandrakasan, "Nanometer MOSFET Variation in Minimum Energy Subthreshold Circuits," *Electron Devices, IEEE Transactions on,* vol. 55, pp. 163-174, 2008.

[57] R. S. Tsay, "Exact zero skew," in *Computer-Aided Design, 1991. ICCAD-91. Digest of Technical Papers., 1991 IEEE International Conference on*, 1991, pp. 336-339.

[58] G. E. Tellez and M. Sarrafzadeh, "Minimal buffer insertion in clock trees with skew and slew rate constraints," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on,* vol. 16, pp. 333-342, 1997.

[59] C. J. Alpert, A. B. Kahng, L. Bao, I. I. Mandoiu, and A. Z. Zelikovsky, "Minimum buffered routing with bounded capacitive load for slew rate and reliability control," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on,* vol. 22, pp. 241-253, 2003.

[60] C. Albrecht, A. B. Kahng, L. Bao, I. I. Mandoiu, and A. Z. Zelikovsky, "On the skew-bounded minimum-buffer routing tree problem," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on,* vol. 22, pp. 937-945, 2003.

[61] H. Shiyan, C. J. Alpert, H. Jiang, S. K. Karandikar, L. Zhuo, S. Weiping, and C. N. Sze, "Fast Algorithms for Slew-Constrained Minimum Cost Buffering," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on,* vol. 26, pp. 2009-2022, 2007.

[62] J. Kwong, Y. Ramadass, N. Verma, M. Koesler, K. Huber, H. Moormann, and A. Chandrakasan, "A 65nm Sub-Vt Microcontroller with Integrated SRAM and Switched-Capacitor DC-DC Converter," in *Solid-State Circuits Conference, 2008. ISSCC 2008. Digest of Technical Papers. IEEE International*, 2008, pp. 318-616.

[63] *Intel Products [Online]*. Available: http://ark.intel.com

[64] K. D. Boese and A. B. Kahng, "Zero-skew clock routing trees with minimum wirelength," in *ASIC Conference and Exhibit, 1992., Proceedings of Fifth Annual IEEE International*, 1992, pp. 17-21.

[65] M. Seok, D. Blaauw, and D. Sylvester, "Clock network design for ultra-low power applications," in *Low-Power Electronics and Design (ISLPED), 2010 ACM/IEEE International Symposium on*, 2010, pp. 271-276.

[66] M. A. B. Jackson, A. Srinivasan, and E. S. Kuh, "Clock routing for high-performance ICs," in *Design Automation Conference, 1990. Proceedings., 27th ACM/IEEE*, 1990, pp. 573-579.

VITA

JEREMY REYNARD TOLBERT

Jeremy Reynard Tolbert was born on August 14th, 1984 in Flint, Michigan. He graduated from Grand Blanc High School in Grand Blanc, Michigan in 2002. He attended the University of Michigan in Ann Arbor, Michigan where he was awarded the scholar recognition award. As an undergraduate, he was active in the National Society of Black Engineer's where he served as the Pre-College Initiative chair for the university's chapter. In April 2007, he graduated summa cum laude with a Bachelor of Science in Engineering from the department of Electrical Engineering and Computer Science. After finishing his undergraduate degree, he was accepted into the doctorate program in the School of Electrical and Computer Engineering and the Georgia Institute of Technology, where he conducted research under the supervision of Dr. Saibal Mukhopadhyay. As a graduate student, he was the recipient of the GEM Ph.D. Fellowship, the Georgia Tech President's Fellowship and the National Science Foundation's Graduate Research Fellowship. His doctoral thesis was focused on energy-efficient digital design of wireless biomedical systems. More specifically, he designed and verified a wireless electroencephalography system that performs adaptive data compression to enhance the accuracy and improve the battery lifetime. In August 2012, Dr. Tolbert began work for the Samsung Austin Research and Design Center (SARC) in Austin, Texas.