DESIGN METHODOLOGY FOR LOW POWER 3D-INTEGRATED IMAGE SENSING SYSTEM FOR NETWORK BASED APPLICATIONS

A Dissertation
Presented to
The Academic Faculty

By

Denny Lie

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

School of Electrical and Computer Engineering
Georgia Institute of Technology
May 2015

DESIGN METHODOLOGY FOR LOW POWER 3D-INTEGRATED IMAGE SENSING SYSTEM FOR NETWORK BASED APPLICATIONS

Approved by:

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

Dr. Sudhakar Yalamanchili
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Arijit Raychowdhury
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. Peng Qiu
School of Biomedical Engineering
Georgia Institute of Technology

Date Approved: January 2015

ACKNOWLEDGEMENTS

First, I would like to acknowledge God, his Son, and the Holly Spirit. I would like to thank for all the blessings, the hope, and the grace that have been poured down in life.

I would like to express my highest gratitude to my advisor, Professor Saibal Mukhopadhyay, for his guidance and support during my PhD program. He taught me to keep pushing the boundary. And I am grateful for his patience and encouragement through my difficult times. These five years have been a tremendous journey of learning, I believe I am prepared to strive through challenges on my career path.

I would like to extend special thanks to Professor Sudhakar Yalamanchili and Professor Arijit Raychowdhury for their critical advice and insight in shaping my thesis. In addition, I would like to acknowledge Professor Sung Kyu Lim and Professor Peng Qiu for investing their valuable time and effort in serving as my PhD Committee Members.

I would like to thank GREEN Lab current and former members Amit Trivedi, Boris Alexandrov, Wen Yueh, Zakir Khondker, Sergio Carlo, Jae Ha Kung, Monodeep Kar, Duckhwan Kim, Jong Hwan Ko, Mohammad Faisal Amir, Arvind Singh, Krishnamurty Yeleswarapu, Swarrnna Karthik Parthasarathy, Dr. Jeremy Tolbert, Dr. Minki Cho, Dr. Subho Chatterjee, and Dr. Kwanyeob Chae for the solidarity, friendship and support that they have offered me. I am surrounded by some of the most brilliant people, and I am honored to have them as my friends. I would also like to extend my special thanks to Dr. Minki Cho who had mentored me through my early years of PhD. I have been fortunate to be part of GREEN Lab.

On several research occasions, I would like to acknowledge Dr. Jeremy Tolbert, Dr. Kwanyeob Chae, Wen Yueh, Amit Trivedi, and Wan-Ning Lee who have contributed to the completion of my thesis.

At last, my greatest thanks to my parents Loenhian Lie and Piefung Wong, my siblings Meriyana Lie and Kevin Lie, and Stella Huang for their continuous love, care, and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Energy-efficient processing of sensor information has emerged as a key challenge for the next generation system. The need for real-time processing while maintaining the quality of the processed information is critical in various applications ranging from ultra-low-power wireless sensor node to high performance mobile systems. To improve image throughput with limited power, a 3D-integration of the image compression module is proposed as a solution in this work. A methodology in designing such system is investigated considering dynamically changing requirements (varying channel condition, and throughput demand), and low power operation through system architecture and device technology choices. First, a multi-segment/multi-core image compression approach is presented as a combined solution with 3D-stacking to reduce the workload of the compression module, effectively increasing power efficiency of the system. Second, vertical stacking reduces the rate of heat removal from the compression module and ADC. This dissertation analyzes the impact of thermal coupling to the noise level of the photosensor and quality of the compressed image. The analysis observes that image quality is strongly influenced by the desired image throughput, architectural configuration of the system, and outside environment factors, as a result of die-to-die thermal coupling in the stack. Third, a heterogeneous integration of the photosensor module and compression module, each designed in different technology nodes, is presented. In particular, the opportunity of scaling the compression engine including image buffers to deep sub-micron technology is analyzed while keeping the CMOS image sensor at less advanced 180nm process. As a result, power dissipation and die area reduce with decreasing channel length, however image quality also degrades due to increased power density and reduced heat spreading. Although 3D-integration is a concrete solution to increase power/performance efficiency of the image sensing system, die-to-die thermal coupling may provide challenges in managing the quality of the compressed images.

CHAPTER 1

INTRODUCTION

1.1 Motivation

Image sensing is an essential feature used in many fields such as consumer electronics, media, scientific research, security, medicine, and transportation. A typical digital image sensing unit consists of an image sensor, a signal processing unit (SPU), and an off-chip module for storage or transmitting the sensed image. Historically, advancement in the image sensing technology have been motivated by demands for highly functional (i.e. low noise, high dynamic range), high performance (i.e. high speed, high resolution), and low cost digital cameras and video cameras [1]. Recent advancement in portable electronics, such as smart phones and smart appliances, have fueled demands for fully integrated, small sized, and low power image sensors, while still pushing for maximum functionality and performance [1, 2].

In the last fifty years, advancement in system integration and performance has been achieved mainly through technology scaling. However, as the complexity of the system grows, floor planning constraints post new challenges in efficiently routing the interconnects and minimizing the communication bottlenecks. These issues may be minimized with 3D-integration technology [3]. For example, in a traditional 2D-integrated image sensing system, the image sensor and the SPU are placed separately with limited interconnect ports to read out the image data from the sensor to the SPU. With the 3D-integration, these two components can be stacked on top of each other, with many vertical interconnects for high bandwidth read out. In addition, the two tiers can potentially be build using different process technologies, where each component can be highly optimized [4]. This technology has the potential benefits in achieving not only high performance, but also low power and highly portable image sensing unit.

Along with improving the device technology, image processing is also an essential step

in increasing the usability of the image data. Typical examples of image processing includes data compression, noise reduction, smoothing filter, dynamic range enhancement, and features detection. Among the above, image compression is critical in reducing redundancy in the image data. For example, a JPEG2000 compression algorithm [5] typically has a compression gain of 20 percent with a 1920x1080 image resolution. The significant reduction in the image size can potentially minimize power in transferring data across a wireless channel [6], as well as store the image efficiently. However, image compression requires heavy computation, and the computing requirement grows with increasing image resolution. Therefore, it is necessary to develop techniques to maximize the efficiency of the system integration.

3D-integration may restrict heat flow generated by the SPU and other components in the system [4, 7, 8]. A variation in the temperature has direct impact to the device parameters of the photodiode arrays, the pixel circuits, and the column circuits, thus changing the CIS characteristics and affecting image quality. Essentially, there is a complex interactions between the performance, power, and thermal coupling in the integrated system. Understanding and taking the interactions into account can help avoid harmful effects on the image quality.

1.2 Existing Works

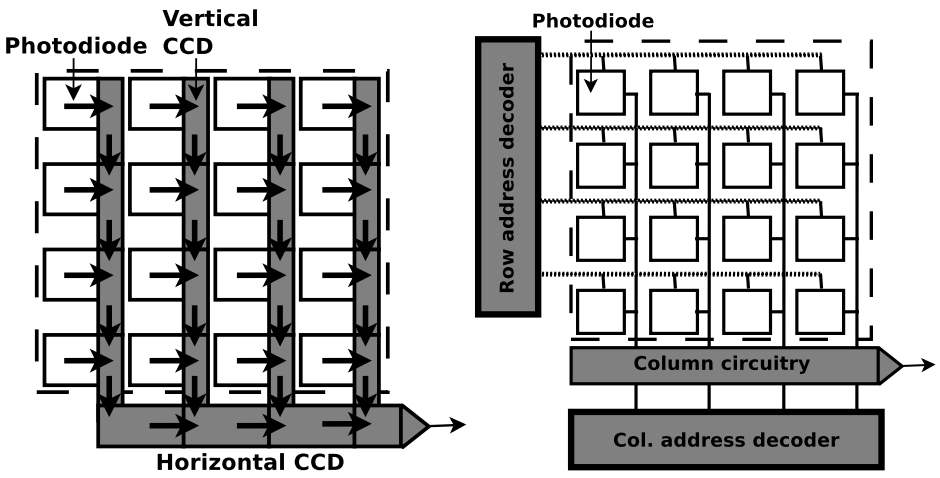
Prior works in digital image sensors date back to 1964, with the invention of the first CCD and CMOS based image sensor. Since then, developments in image sensing and compression technology have continued to progress and mature. In addition, recent developments in 3D integrated circuit (3DIC) technology have set a new trend in designing complex digital systems, and may prove beneficial for image sensing application. The following subsections briefly observe existing works in image sensor, compression, and 3DIC.

1.2.1 CMOS Image Sensor

Early development in digital image sensor technology have resulted in the creation of charge coupled device (CCD) and complementary metal oxide semiconductor (CMOS) image sensor. Nevertheless, CCD image sensor had been the primary device of choice since the beginning of its adoption in the early 1970s [9]. At that time, the CCD image sensor was proven to have extremely low noise compared to the CMOS image sensor, producing a superior image quality [1, 10]. For the next 20 years, majority of the research and development efforts had focused primarily on improving performance of the CCD image sensor. Throughout the first half of the 1990s, most of the digital still cameras and video capturing cameras used CCD technology, while only low end imaging functionality, such as an optical mouse, would consider CMOS based implementation. By the early 1990s, the CMOS process technology matured and was extensively use in digital electronics and microprocessors. Research and development efforts to improve image quality in the CMOS image sensor technology restarted, motivated by the possibility for low power operation, miniaturization, and high integration with a vision for camera-on-a-chip devices [11]. Since then, the performance of the CMOS image sensor has been steadily improving to match and potentially surpass its CCD counterpart [12].

Unlike CCD image sensor that use sequential charge transfer mechanism to read out an image (Figure 1a) [13, 14, 15], the CMOS image sensor use an X-Y address mechanism similar to the read out scheme in a digital memory (Figure 1b) [13, 16, 17]. The X-Y address scheme offers ability to access a specific pixel location, allowing for a flexible readout pattern. In addition, the CMOS image sensor has significantly lower power consumption during the readout sequence [10, 13]. In the X-Y address scheme, only activated pixels are switching. Meanwhile, in the charge transfer scheme, all the pixels have to be activated to transfer photo charges from one pixel to the next.

The active pixel sensor (APS) architecture is the prominent choice for the CMOS based pixel design. In the APS, a photo generated charge is amplified before it is read out to help



(a) (b)

Figure 1: Read out mechanism in image sensor. (a) CCD image sensor. (b) CMOS image sensor.

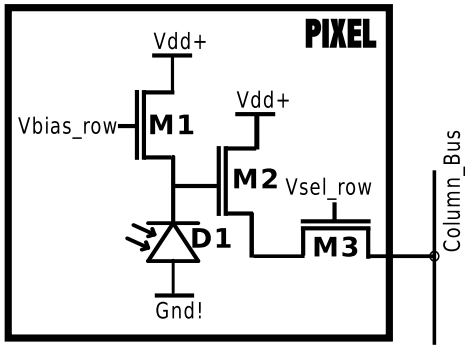


Figure 2: Pixel schematic of a typical logarithmic CIS.

suppress noise in the readout path, as opposed to the original MOS passive pixel sensor (PPS) that doesn't have a gain stage in its pixel. During photo conversion, incidental light is linearly converted to photocurrent by the photodiode. Then this photocurrent is integrated into charge during the exposure. Finally, the collected charge is converted into voltage by the readout path. The above method is commonly referred as direct integration, and is used in the CCD image sensor as well as majority of the CMOS based pixel. In direct integration, photo conversion occurs in a highly linear fashion, and the response can be treated as an output of a linear system, hence it is often referred as linear pixel.

Figure 3: Flow diagram of a typical wavelet based image compression.

Although the current linear CMOS image sensors have surpassed their predecessors in many aspects, its dynamic range is still inferior as compared to the human eyes [18, 19, 20]. One possible option to improve dynamic range is by using a logarithmic photo conversion technique. Figure 2 shows a typical logarithmic CMOS image sensor (logarithmic CIS). Lights incidental to the photodiode generate photocurrent. The photocurrent has to be small enough to keep transistor M1 in the weak inversion region, hence the voltage at the source of transistor M1 has a logarithmic relationship with the photocurrent. The logarithmic CIS has a continuous pixel response, which leads to faster image capture time than direct integration sensor. The main drawback of the logarithmic CIS is its output noise is difficult to eliminate as compared to linear pixel [21, 18, 22].

1.2.2 Wavelet Based Image Compression Method

Image compression is essential in reducing redundancy in the image data so that data can be transmitted in an efficient way. Image compression can be divided into lossless compression and lossy compression. In a lossless compression, every single bit in the image data can be fully reconstructed to its original strings after the image is uncompressed. Huffman Coding [23], Arithmetic Coding [24], and Lempel-Ziv Coding [25, 26] are examples of famous traditional lossless compression algorithm that are still widely used today. Typical applications for lossless compression include medical imaging and space explorations [27, 28]. In a lossy compression, the image can not be reconstructed to its original image, because part of the detail in the image is discarded during compression. The lossy compression are generally capable of achieving high compression ratio with the expense of adding compression artifacts into the image. Typical applications for lossy compression include multimedia broadcasting and video streaming.

Over the years, the use of discrete wavelet transforms (DWT) as a transform stage in the compression process has gained popularity. Figure 3 shows a flowchart for typical wavelet based image compression. First, the image is transformed to its frequency-spatial representation (wavelet coefficients) using the two dimensional discrete wavelet transform (2D-DWT). Next, the wavelet coefficients are quantized, and a coding technique is applied to reduce data representation of the compressed image. A popular wavelet filters pair is the biorthogonal Cohen-Daubechies-Feauveau (CDF) wavelet, which is used in JPEG2000 [29]. After the transform stage, different coding techniques such as Embedded Zerotree Wavelet (EZW) [30, 31], Set Partitioning In Hierarchical Trees (SPIHT) [30, 32], Embedded Bock Coding with Optimized Truncation (EBCOT) [33], or a simple lossy thresholding method can be applied to compress the image.

However, 2D-DWT requires heavy computation [34]. Although it is easy to compute the 2D-DWT in software, it is difficult to achieve a real-time hardware based implementation especially when the desired image size and frame rate are high. In traditional digital signal processing theory, the DWT is computed using a convolution method. Research efforts in the area had developed the lifting method which significantly reduces the number of adders and multipliers [35, 36]. The 2D-DWT is effectively an extension of a 1D-DWT for transforming a two-dimensional data array. In general, the 2D-DWT is a separate two-steps process that involves applying the 1D-DWT horizontally and vertically across the image. After the first DWT step (i.e. the horizontal DWT), a storage module is required to temporarily store the intermediate data, before it is sent to the second DWT step (i.e. the vertical DWT). The size of the intermediate storage is proportional to the length and width of the image. Since then, a line based scheme was developed as a simple and effective solution to significantly reduce the intermediate storage size [37, 38]. In the line based 2D-DWT, the intermediate storage size is proportional to the length of the image and the length of the finite impulse response (FIR) filter. On the other hand, numerous non-separable techniques that compute the transform in a single step had also been presented

[39, 40, 41]. Although, the non-separable techniques offer faster performance than the separable techniques, system implementation is relatively complex and requires significant hardware overhead.

1.2.3 3D Integrated Circuits

Vertically stacked circuits can be integrated together using the combination of through silicon vias (TSVs) and bonding structures [42, 43, 44, 45]. The die-to-die integration can be organized using a face-to-face stacking, face-to-back stacking, or back-to-back stacking.

The main benefits of 3D-integrated circuits for image sensing application include wide bandwidth interconnects, and reduced footprint of the system [3, 45]. In addition, the technology offers reduction in interconnect length, which leads to reduced interconnect delay and switching power. When an image is captured, electrical signal from each pixel has to be converted to digital signal and transferred to a storage element. In a conventional planar integration, it is a pixel by pixel sequential process. 3D-integration technology provides multiple parallel connections to the pixel array, which can reduce the capture time of the image. This technology also allows the photodiode to be build on a separate layer from the rest of the circuits to increase fill factor. In addition, the reduced footprint can be used to improve the resolution of the sensor.

However, 3D-integration technology have thermal management issues, which may hurt the image sensing application. Components such as analog to digital converter (ADC) and SPU consumes power and generates heat [46, 47]. In a dense 3D-integrated image sensor, heat flow is limited and the photo receptor circuit is thermally coupled with the rest of the system. Temperature variation affects noise behavior and pixel response characteristics of the photo sensor [10, 48].

1.2.4 Existing Work in 3D Integrated Image Sensor

3D integrations open new opportunities for high-speed and high-density image sensor design [49, 50, 51, 52, 53, 54]. Kiyoyama et. al. studied a block-parallel image sensing and

processing architecture to allow analog to digital conversion and other image processing algorithm tasks be distributed across multiple blocks and performed in parallel [51, 52]. However, the focus of the work is only the design of the ADC and its integration to the pixels block. One of the early 3D stacked pixel sensor circuits was demonstrated by Suntharalingam et. al. [53], and later by Zhang et. al. [54]. Both works demonstrated 3D stacked CIS with high fill factor by separating the photodiode and the rest of the pixel circuitry on different layers. A simple pre-processing module integration with the sensor pixel has also been demonstrated by Zhang. The above research efforts provide initial concepts of an integrated photo sensor and image processor system in a 3D-stack, and demonstrations of the key components.

1.3 Thesis Objective and Organization

Although a full integration of photo sensor module and image processing engine has been conceptualized, a thorough analysis and optimization on such system is still yet to be performed. This research builds on the prior works on the 3D image sensors, analysis of temperature effects on CIS, and wavelet based image compressions. The objective of the proposed research is *to investigate a methodology in designing energy efficient 3D-integrated image sensor for network based applications considering dynamically changing requirements (varying channel condition, and throughput demand) and low power operation through system architecture and device technology choices*. This research examines the design integration of a CMOS image sensor (CIS) with image processing unit on a 3D stack. Specifically, a wireless image sensor node is considered in majority of the analysis. The potential of highly parallel image compression using a multicore SPU is explored as the main image processing engine. The effect of thermal coupling between the image sensor and the SPU, due to the power and performance of the system, on the noise characteristics of the CIS pixels is analyzed. Finally, a higher degree of optimization method for different performance range of applications through heterogeneous integration is explored.

The rest of this dissertation is organized as follow:

Chapter 2 proposes a multicore system architecture that is suitable for a 3D integration of a memory (image buffer) and a signal processing unit (SPU) in performing heavy compression task. The implication of the architecture and design parameters to the power, performance, and image quality of the system is discussed.

Chapter 3 investigates the effect of thermal coupling inside the 3D stack of the system. The analysis includes discussions on the power and performance of the system, thermal coupling in the 3D stack, and noise characteristics of the sensor. The quality of the compressed image depends on not only the noise level of the sensor but also the lighting condition of the captured object.

Chapter 4 examines the system level benefits of heterogeneous integration for the system. The idea is to optimize the system by integrating the image sensor elements and image processing elements in different technology nodes.

Chapter 5 presents the final conclusion, thesis contributions, and suggestions for follow up research topics.

CHAPTER 2

MULTI-SEGMENT IMAGE COMPRESSION FOR THE 3D INTEGRATED IMAGE SENSOR

2.1 Introduction

Image sensing is a common functionality in various applications like smart camera, and smart phone. A typical image sensing system includes an image sensor, a signal-processing unit (SPU), and an off-chip module for storage or transmission of the sensed image. Advancing high-speed imaging has been an important objective in many imaging applications. To achieve the goals, a 3D integration of photodiode array, multiple Analog-to-Digital Converters (ADC), and frame memory (image buffer) has been proposed to help fetch the image from the photodiode array faster [49, 50, 52, 53, 54].

Image compression is essential to reduce redundancy in the image data so that the data can be stored or transmitted in an efficient way. Different algorithms had been introduced to maximize compression ratio while maintaining an acceptable image quality. Among the many algorithms, the wavelet based compression has gained popularity. Namely the two-dimensional discrete wavelet transform (2D-DWT) is set as the standard transform function for JPEG2000 [5], MPEG4, and many compression algorithms. However, 2D-DWT

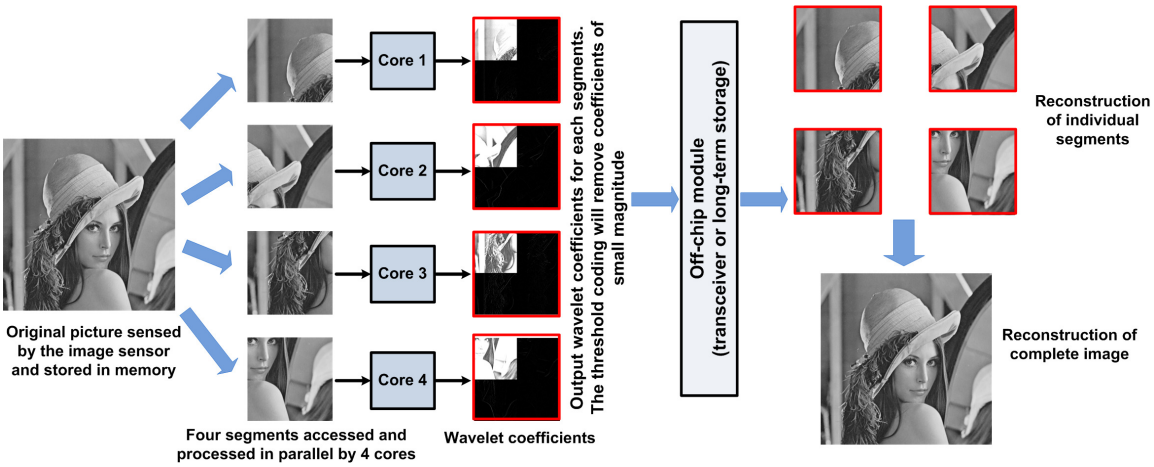


Figure 4: A flow diagram of the multi-segment image compression.

requires heavy computations, and hence various VLSI architectures had been proposed for the benefit of real-time image processing [34, 39, 36, 37, 38]. A key challenge in the DWT is the need for high-bandwidth connection between the SPU and the image buffer. Hence, it is expected that 3D stacking of the SPU and the image buffer can provide major performance gain for the processing unit [3].

In this chapter, a multi-segment image compression method is proposed as a high throughput signal processing architecture for a 3D-integrated image sensing system. In this method, the concept of parallel processing is applied by using a multi-core signal processing unit (multi-core SPU). Essentially, a captured image is divided into multiple segments (Figure 4). Each segment is treated as an independent image, and locally assigned to a signal processing core for compression. By maximizing the spatial locality between the photodiode array and the signal processing unit, the total compression latency of the image is significantly reduced. However, dividing the image into many segments reduces the compression ratio. In the case where off-chip data transmit rate (channel rate) is limited, the reduced compression ratio significantly increases the transmission latency. It also increases the power dissipation of the system. This section describes the design space of the multi-segment (number of cores) and the operating clock speed of the system considering specific target image throughput (frames per second), varying channel rate, and power dissipation of the system.

2.2 Background

This section provides background and reasoning for choosing the multi-segment image compression as a suitable signal processing architecture for the 3D-integrated image sensing system. As mentioned in the previous chapter, image compression is crucial in reducing image data volume for efficient transmission and storage. In this work, the wavelet based image compression is presented as the algorithm of choice. Different image coding algorithms such as the threshold coding, the Huffman coding, and a mixed threshold and

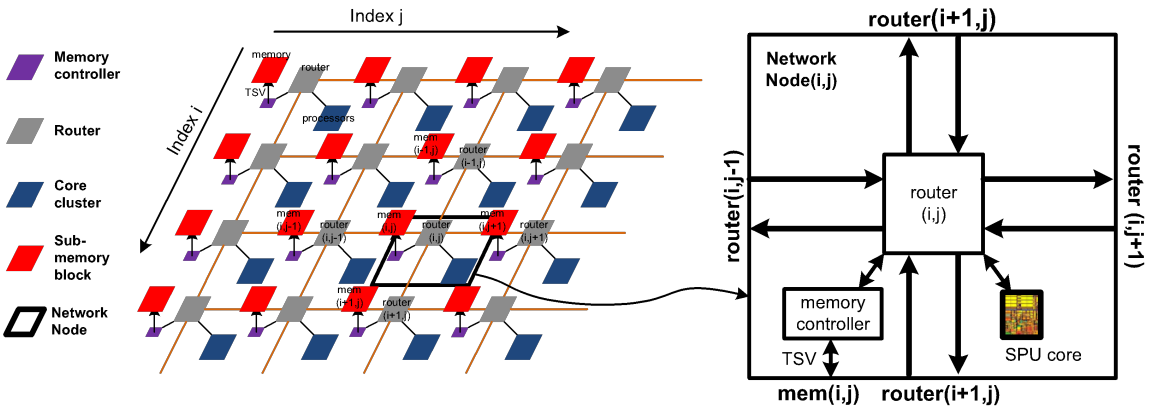


Figure 5: An illustration of the X-Y routing scheme for the core-to-buffer network.

Huffman coding are presented.

2.2.1 Core-to-Buffer Network for 3D-Integrated System

To understand the algorithms that can successfully benefit from 3D stacking, the core-to-buffer network in the 3D-stack needs to be carefully analyzed. In this work, a mesh type network topology with an X-Y routing scheme is considered. The entire network consists of multiple nodes (Figure 5). Each node comprises of a processor, a memory controller, and a network router. The router has four gates, i.e. north, west, south, and east, which connect it to neighboring nodes. The queue lines inside the router are implemented using FIFOs. In the case of multiple accesses, the arbitration is decided using the round robin algorithm. To motivate the need for MuSIC, we first study the generic behavior of the network considering randomly generated memory traffic from individual cores. The distance distribution of the memory accesses of each core is generated from a Poisson's process. The rate parameter (λ) of the Poisson's process is determined by the memory access rate (r). A higher spatial locality implies a core is more likely to access the memory banks, which are physically close to the node to which the core is connected. A lower spatial locality implies that a core is more likely to access memory banks connected to the far nodes. We model the spatial locality behavior of the program using the Gaussian random process. Lower σ value indicates higher spatial locality i.e. transactions will be

consumed at near memories and less likely to traverse to the far node. As σ value increases, cores tend to access far memories. As shown in Figure 6, a higher access locality leads to lower core-to-memory latency (assuming one cycle latency for each link) due to closer signal routing that minimizes latency through the 2D network and maximally exploit the 3D bandwidth.

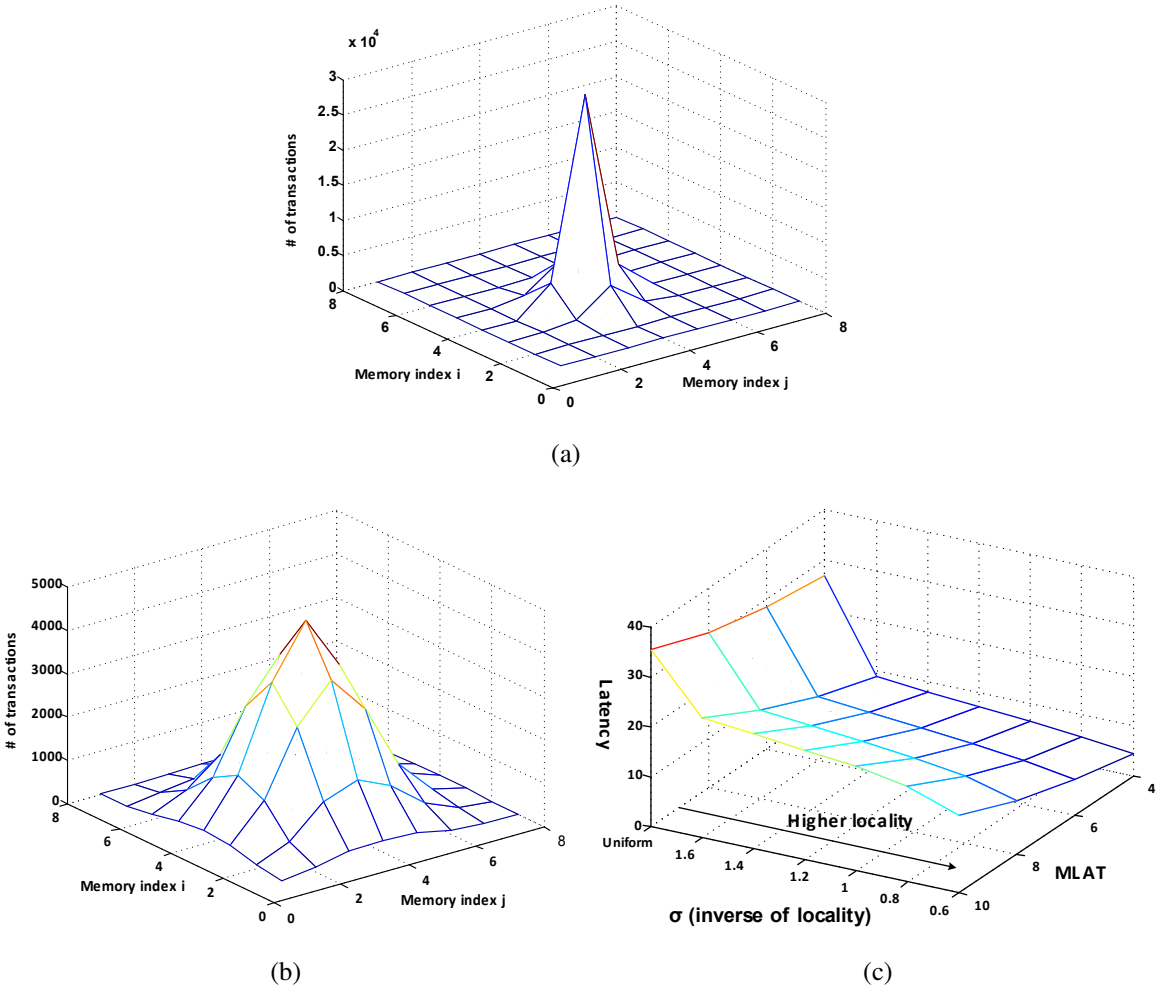


Figure 6: Network simulation for varying spatial locality to illustrate the need for MuSIC: (a and b) target access memory((i,j) node) location for (a) higher (σ = 0.6), (b) lower (σ = 1.6) spatial locality; (c) Sensitivity of latency to the address locality and the memory latency (8x8 nodes).

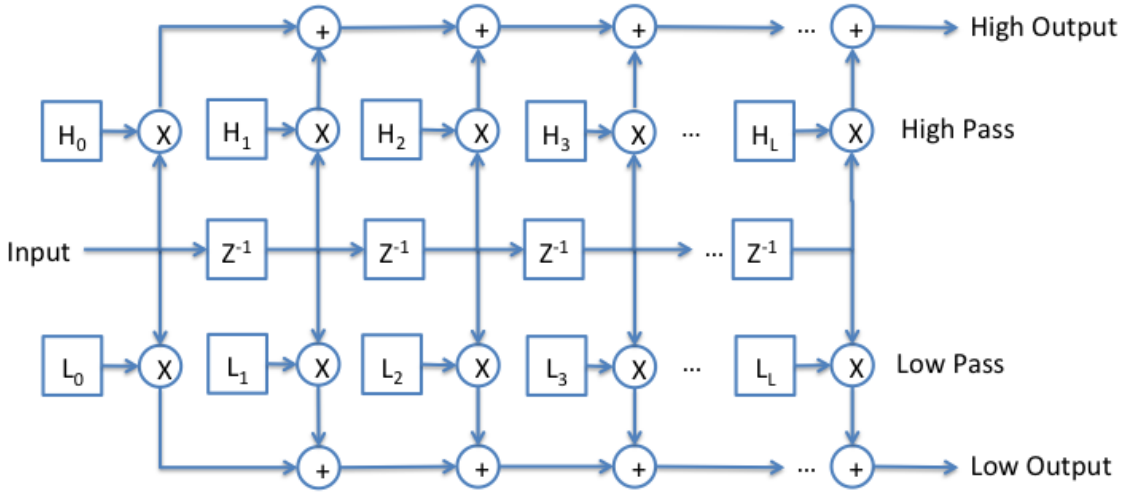


Figure 7: Block diagram of DWT convolution.

2.2.2 Implementation of the Coding Algorithm

In this work, we implement the line-based 2D DWT scanning architecture with convolution-based 1D DWT FIR filters [37]. First, let us consider the convolution-based DWT FIR filter as shown in Figure 7. Given that our FIR filters has a length L, a shift register of the length L-1 is needed for the input data. The input data is shifted into the register one at a time. Let us now consider that the maximum length of our FIR filter is L = 2S when the filter length is even, or L = 2S + 1 if the filter length is odd, where S is an arbitrary positive number. Then, the minimum length of input needed to perform the convolution is S if the filter length is even, or S + 1 if the filter length is odd. This is because symmetric extension is applied to the input data to create a 2S or 2S + 1 input data for the even or odd length FIR filters. Figure 7 also shows that the input data can be fed into both the high-pass filter (HPF) and low-pass filter (LPF) in parallel.

A 2D DWT is an extension of a 1D DWT. Considering a RAM based 2D DWT architecture, a temporary storage element is needed to store intermediate results. The conventional and simple way to perform 2D DWT is to first do row-wise 1D DWT filtering to all the rows, save the intermediate results in the temporary storage element, and then perform

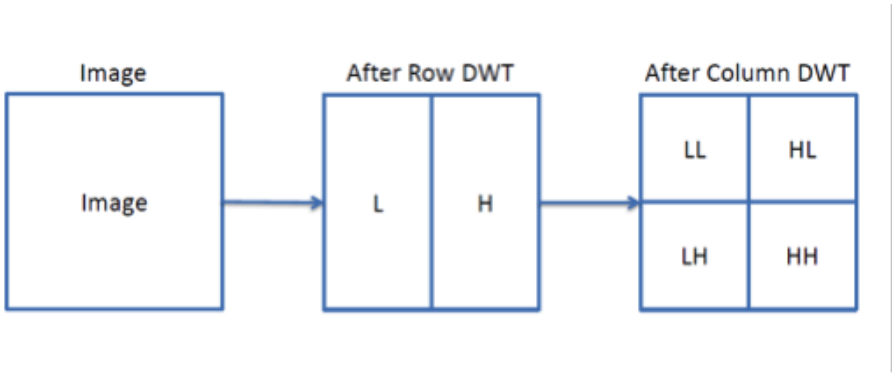
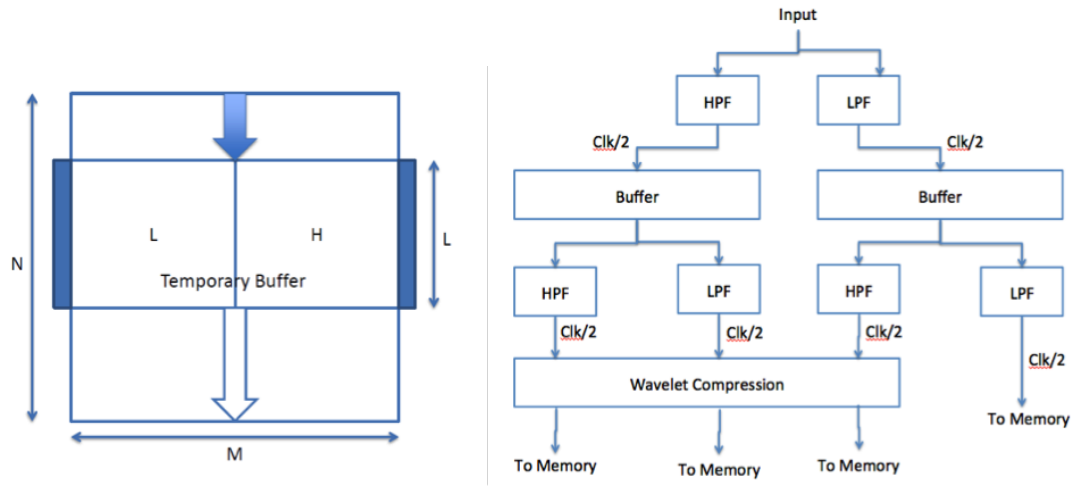


Figure 8: Block diagram of memory access for direct implementation of 2D DWT.

column-wise 1D DWT filtering, as shown in Figure 8. Assuming an image of size N by M, a temporary storage element of size NxM is needed to store the intermediate results. These coefficients are fed in as inputs for the column-wise 1D DWT filtering. A more efficient method to perform a 2D DWT is the line-based method, shown in Figure 9a. We first start by doing the row-wise filtering. But then, the difference is the column-wise filtering starts



(a) (b)

Figure 9: Efficient line based 2D discrete wavelet transform (2D-DWT). (a) Temporary buffer to contain the intermediate coefficients during the 2D transform. (b) A flow diagram of the 2D transform.

Table 1: Discrete Wavelet Transform Filter Pair

Low Pass Decomposition Filter	0	-1/8	1/4	3/4	1/4	-1/8
High Pass Decomposition Filter	0	-1/2	1	-1/2	0	0
Low Pass Reconstruction Filter	0	1/2	1	1/2	0	0
High Pass Reconstruction Filter	0	-1/8	-2/8	6/8	-2/8	-1/8

as soon as sufficient rows of the intermediate values are present. With line-based architecture, the size of temporary storage element is reduced to NL, where L is the length of the FIR filters. Normally the filter length (L) is significantly lower than the image length (M) itself. Once the temporary storage is filled up, the column-wise DWT is executed. After that, the next row of input data is taken in for another set of row-wise and column-wise filtering. In the case of a multilevel 2D DWT, the lowpass-lowpass (LL) sub-band, referred as the approximate coefficients, is used as the input for the next level of decomposition, however in this work we present a one-level 2D DWT. Figure 9b shows a flow diagram of the DWT. Table 1 shows the coefficients of the 5/3 LeGall filter pair used in this work [55].

2.3 System Description

The discussion in Section 2.2.1 shows that minimum latency in the memory-core stack can be achieved when each core only accesses the nearest memory port, and provides the reason for choosing the multi-segment image compression scheme. A conceptual diagram of the 3D-integrated image sensor with multi-segment compression is shown in Figure 10. The system consists of five tiers: 1) the photodiode tier; 2) the column circuitry tier; 3) the ADC and control logic tier; 4) the image buffer (image storage) tier; 5) the SPU tier. The tiers are connected vertically by through-silicon vias (TSVs). The imaging procedure consists of three stages: 1) sensing: the image is captured and converted into digital signal; 2) compression: the digital data is compressed using the discrete wavelet transform method

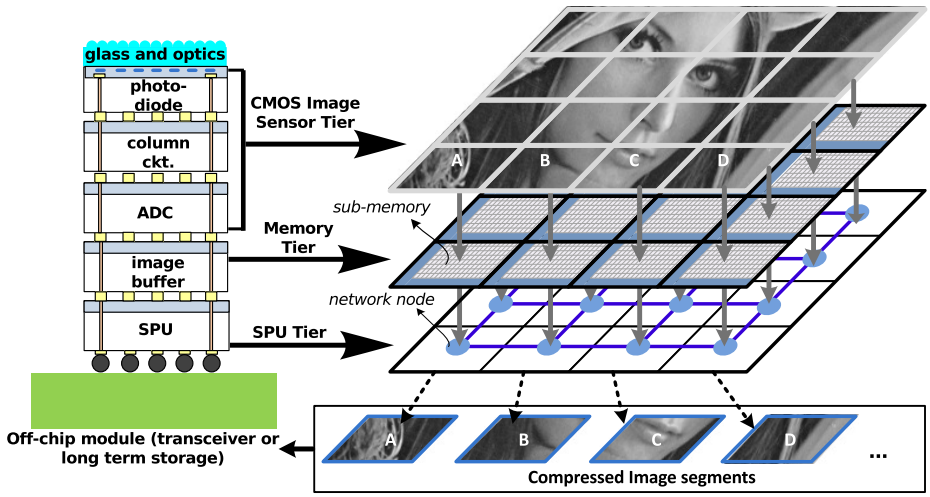


Figure 10: An illustration of the 3D-integrated multi-segment image compression system.

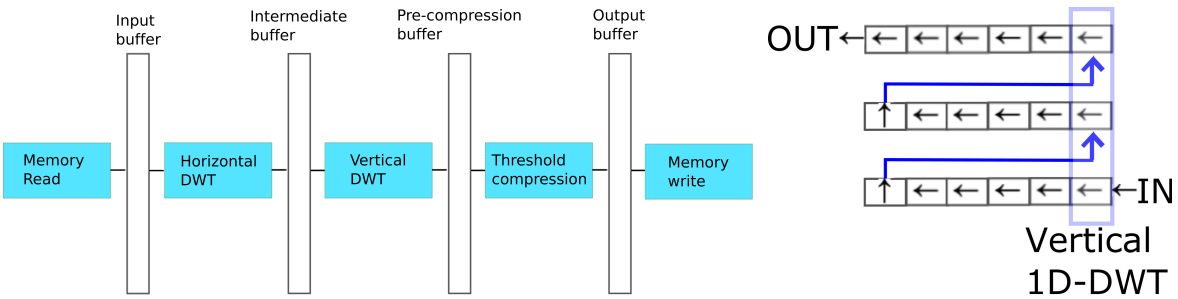
followed by threshold or/and huffman coding; 3) transmission: the compressed digital data is then transmitted to an off-chip module. This module can be a wireless transceiver. The next image is captured (sensed) only after the previous image is completely transmitted. The SPU and the image buffer is connected by a mesh type network topology with X-Y routing scheme. Detail explanation of the network is presented in Section 2.2.1. Essentially, the simulation framework can be divided into two main parts:

1. Architectural timing analysis - the architectural performance of the compression algorithm is estimated by simulating the data patterns in the network nodes. The generated traffic pattern is discussed below.

2. Power and performance analysis - the power dissipation of the SPU, network router, and image buffer is extracted from a synthesized design, ORION [56], and CACTI [57], respectively, considering a 180nm CMOS process technology.

In generating the traffic pattern, we consider a pipelined line based architecture (Figure 11a) with a finite impulse response (FIR) filter of length 6 (L = 6) that fits with 5/3 LeGall filter pair. The DWT starts with a memory read operation to fetch input data from the frame memory (image buffer). The input data is then used to perform horizontal 1D-DWT, and

the result is stored in the intermediate buffer in the next clock cycle. The intermediate data is not immediately used for vertical 1D-DWT. Sufficient amount of intermediate data points need to be collected before the vertical 1D-DWT can be computed. Once the vertical DWT is computed, threshold based compression is performed only on the detailed coefficients (namely LH, HL, and HH sub-bands), before they are written back to the frame memory (image buffer). To generate the pattern, at the beginning, three consecutive memory-read operations are needed to prepare the first input set. After the third read operation, horizontal 1D-DWT calculation is performed to generate the first intermediate coefficient. The generated coefficient is then stored in the temporary storage buffer for column filtering. The temporary storage buffer is implemented as two separate buffer chains of moving registers, one chain for L sub-bands, other for H sub-bands. Figure 11b illustrates data movement in one of the buffers chain. Data from horizontal 1D-DWT gets into the buffer chain from the bottom right corner, and move to the next neighboring register every clock cycle. Six registers of the intermediate buffer are connected to the input ports of the vertical 1D-DWT filters, as shown in Figure 11b. Vertical 1D-DWT calculation starts as soon as three coefficients are available to the vertical filter. The output of the vertical calculation is then stored back to the memory. If we assume a 512X512 image, that means there is 512X3 memory read operations before the first memory write operation starts. After this point, the traffic



(a) (b)

Figure 11: DWT architectural consideration. (a) SPU pipeline stage. (b) Registers for the intermediate buffer.

pattern becomes an alternating write and read access up to the end of an image frame. The remaining part of the pattern consists of write accesses until the last pixel is calculated.

2.4 Analysis for Threshold Coding

The first part of the analysis focuses on the performance of the multi-segment image compression with threshold coding algorithm. The key benefit of the threshold coding is that the algorithm is relatively simpler to implement compare to the Huffman coding, thus requiring a smaller hardware and die area.

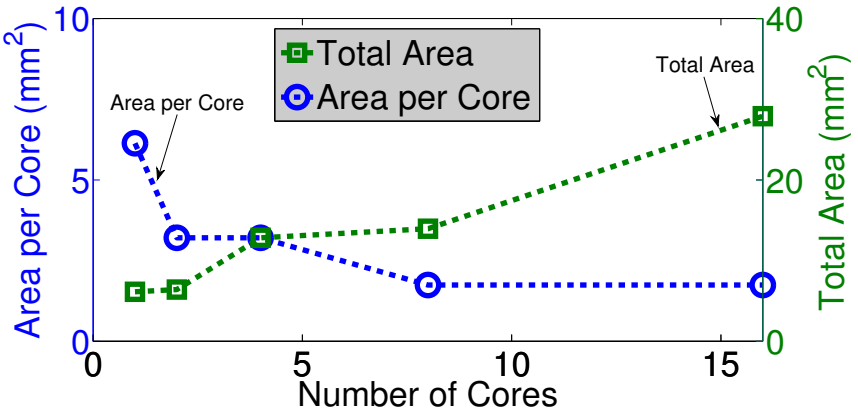
2.4.1 Die Area of the Multi-Segment Image Compression

Figure 12a illustrates the increase in die area of the signal processing unit (SPU) with increasing number of cores for the DWT with Threshold coding scheme. The die area of the signal processing unit increases with increasing number of cores (image segments). However, the increase in the die area is not linearly dependent with the number of segments. Majority of the signal processing unit core die area is occupied by a set of registers to store intermediate data during the wavelet transform computation. The size of the register file is determined by the width, but not by the length, of the image (Figure 12b). Therefore, the die area of the signal processing unit depends on the method in which the image is divided into segments. In this research, the image is first divided along the image width direction to create the 2-cores system, and then it is divided along the image length direction to create the 4-cores system. This alternating division method is also used to create the 8-, and 16-cores system.

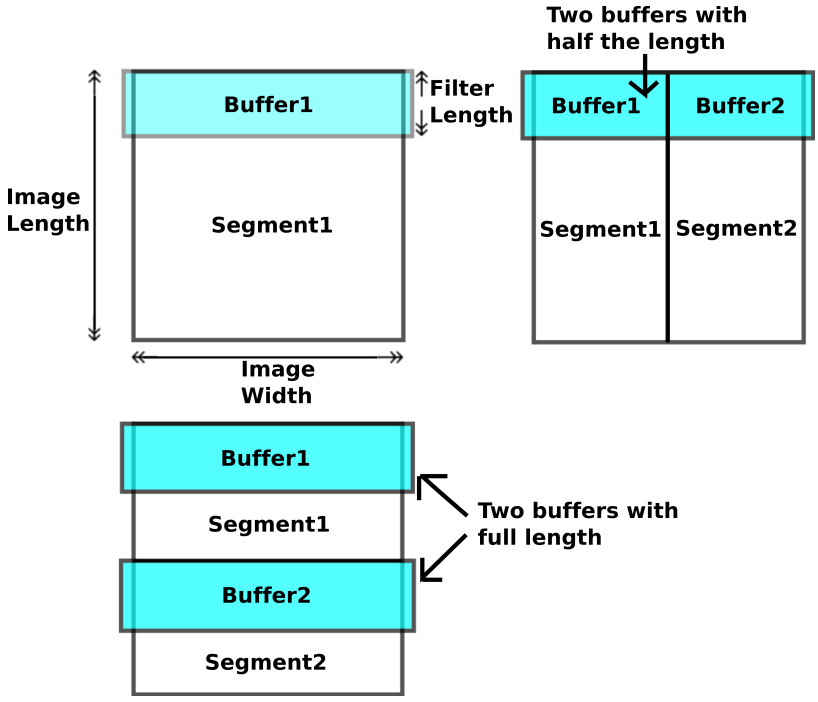
2.4.2 Data Volume Overhead and Image Quality of the Multi-Segment Image Compression

Dividing the image into many segments reduces the compression ratio. As the number of segments increases, the image data associated with the borders of the image segments also increases. Thus, output data volume expands as a result of the wavelet transform calculation. Figure 13 shows the expansion of output data packets with increasing number

of segments for the threshold coding. The black solid line represents the average data



(a)



(b)

Figure 12: Effect of increasing number of segments to the SPU die area. (a) Area versus number of cores in Threshold coding scheme. (b) Intermediate buffer size expansion with an increase in number of segments along the x-direction (top right) and y-direction (bottom left).

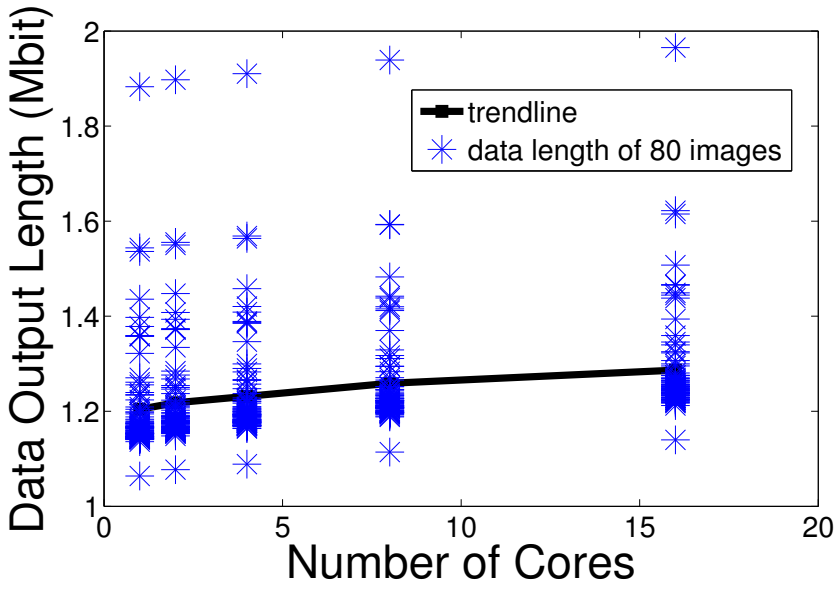


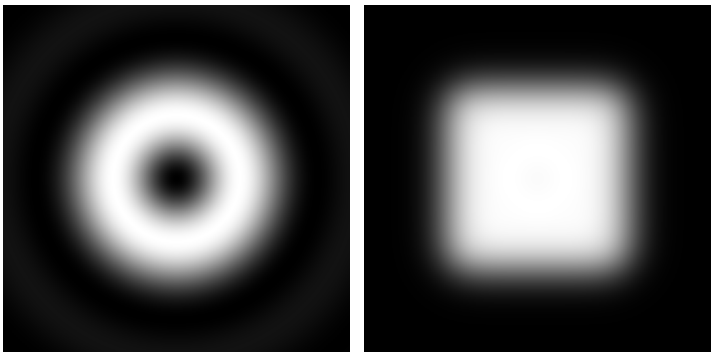
Figure 13: Effect of increasing number of segments to the SPU data volume for threshold coding.

volume for the 80 test images.

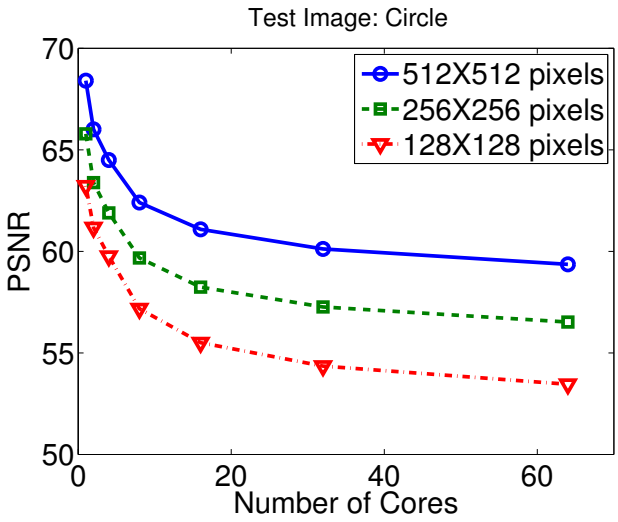
To test for effect of multi-segment image compression to the image quality, a test image (Figure 14a and Figure 14b) is generated in different sizes: 512X512, 256X256, 128X128 [58]. The image quality is measured using the peak signal to noise ratio (PSNR), in which a low distortion level yields a high PSNR value [59]. Figure 14c - 14e show the rising distortion level with increasing number of segments. This increased distortions come from limitation in the convolution of the image data along the edges of a segment, and the increased amount of edges with increasing number of segments. On the other hand, PSNR reduces as the image size is reduced because the edges to non-edges ratio increases.

2.4.3 Performance and Power Analysis of the Multi-Segment Image Compression

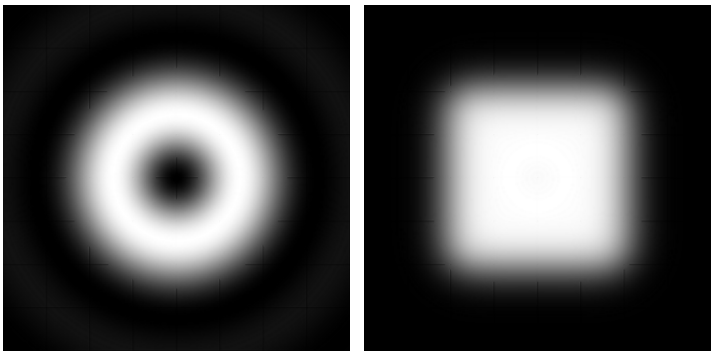
It is possible to integrate the image buffer and the SPU in a 2D fashion for multi-segment compression, although the 3D-integrated system (Figure 15a) yields better performance. In the 2D-integrated system the image buffer is stacked vertically with the ADCs, the column



(a) (b)

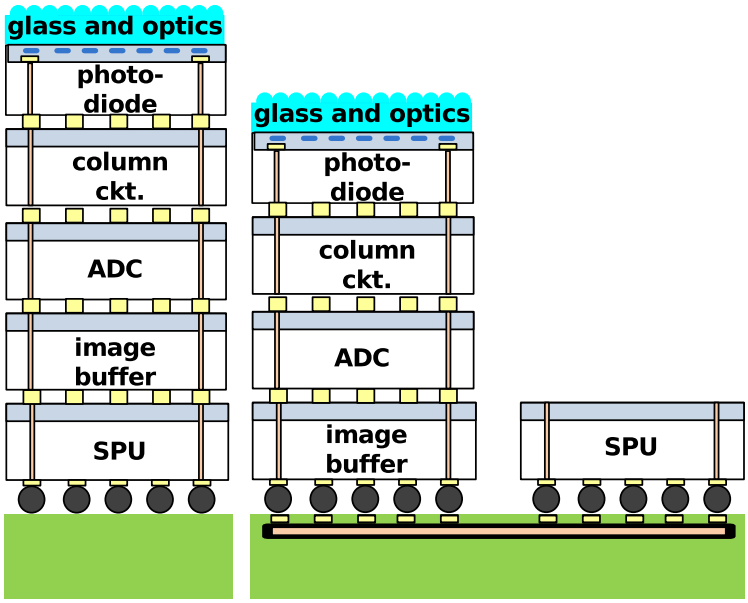


(c)

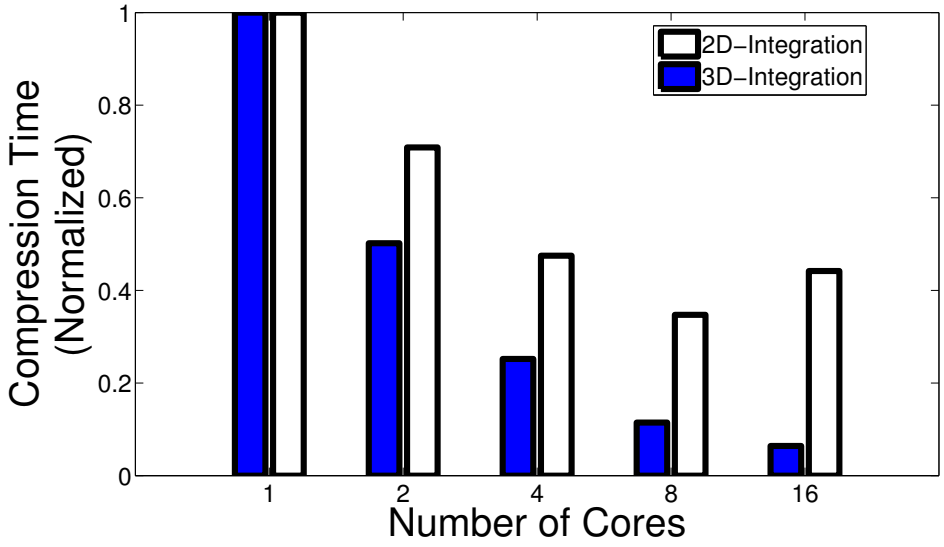


(d) (e)

Figure 14: Effect of multi-segment compression to the image quality. (a) Circle test image. (b) Square test image. (c) Image degradation of the circle test image introduced by MuSIC. (d) After compression circle image compressed with 64 cores. (e) After compression square image compressed with 64 cores.

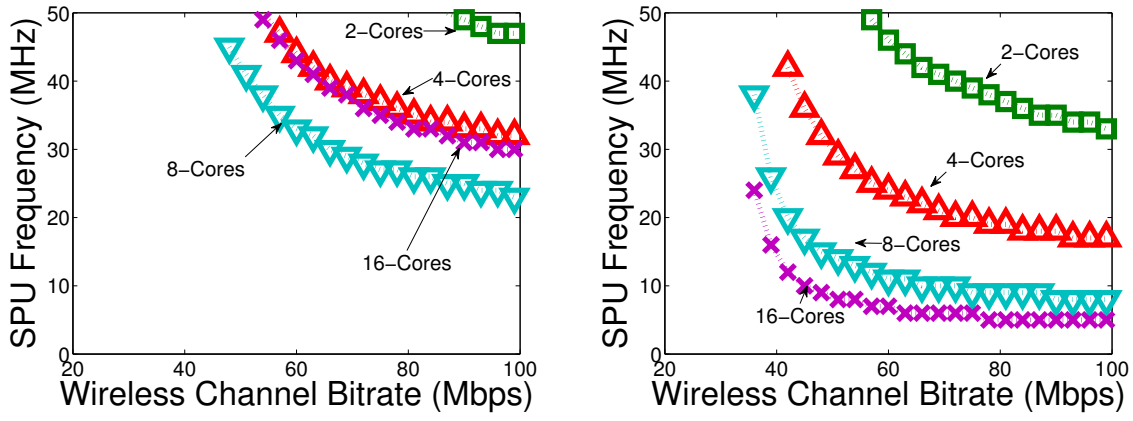


(a) (b)



(c)

Figure 15: Effect of increasing number of segments to the architectural performance in 2D- and 3D-integration. (a) 3D-integration of image buffer and SPU. (b) 2D-integration of image buffer and SPU. (c) Compression time versus number of cores.



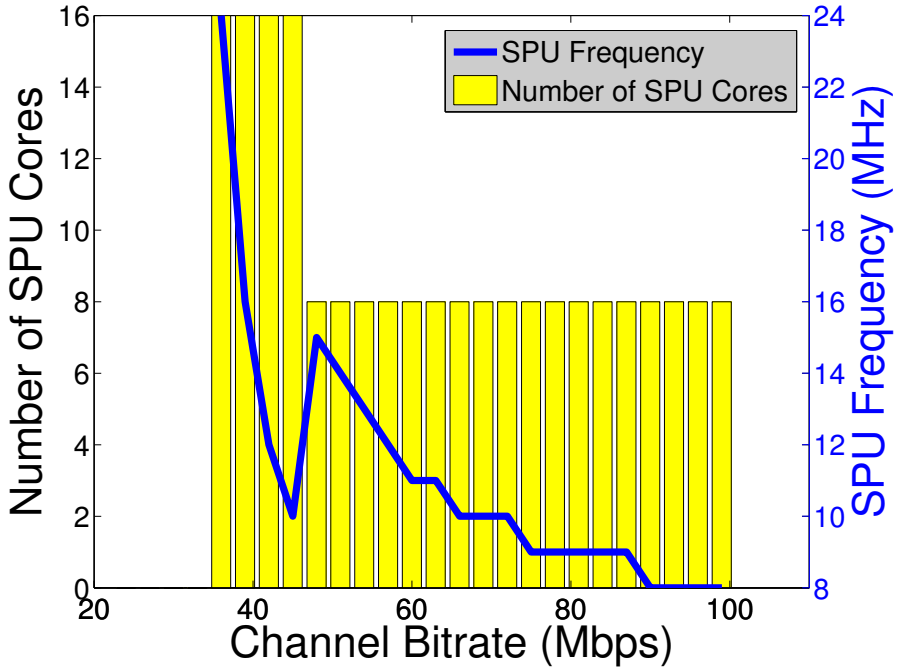
(a) (b)

Figure 16: System performance of the image compression system in 2D- and 3D-integration, considering variations in wireless channel condition, SPU clock speed, and number of SPU cores. (a) System in 2D-Integration. (b) System in 3D-Integration.

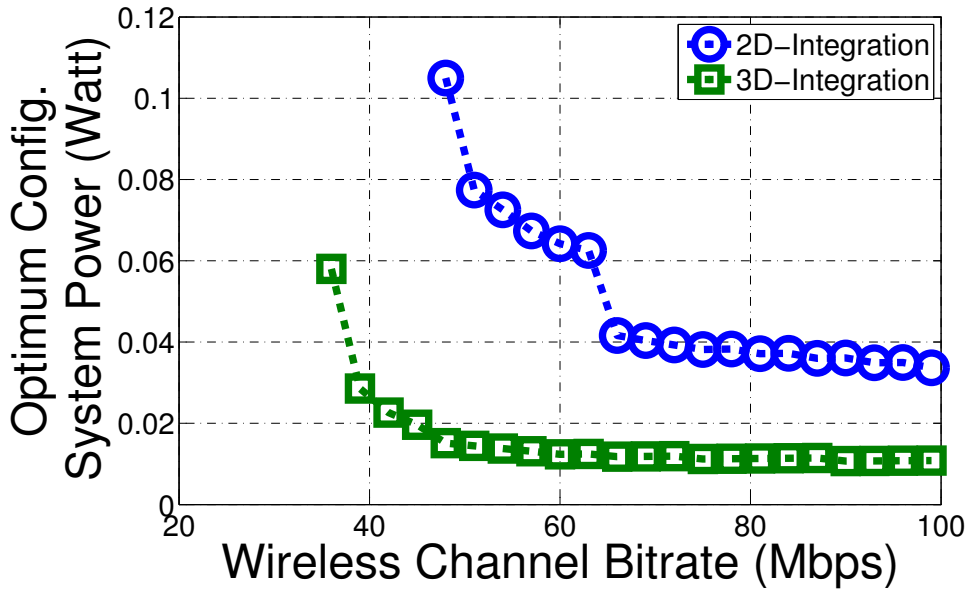
circuitry, and the photodiodes. However, the SPU is placed horizontally next to the image buffer, thus limiting the number of access path between the SPU and the image buffer (Figure 15b). Figure 15c demonstrates how 3D-integration of multi-segment is more effective than 2D-integration. As the number of segments increases, architectural compression time reduces. However, reduction in the compression time for the 2D-integration is not as high as in the 3D-integration, due to lateral traffic movement in the network router in the 2D-integration. This lateral traffic becomes more congested with increasing number of segments, and may eventually penalize the benefit of the multi-segment compression. The chart in Figure 15c shows that the 2D-integration of a 16-cores system performs worse than the 8-cores system.

The performance of the image sensor system (i.e. how many images is sensed and compressed per second) is governed by the total time required to capture, compress, and transmit the image, as shown by the following formula:

$$f_{image} = \frac{1}{t_{total}} = \frac{1}{t_{capture} + t_{compress} + t_{transmit}} \quad (1)$$



(a)



(b)

Figure 17: Effect of varying channel conditions to the SPU configurations and power dissipation. (a) SPU configurations versus channel bitrate. (b) SPU power versus channel bitrate.

The analysis framework in this chapter assumes a wireless sensor node that captures and compresses a constant number of frames per second (image throughput). In a wireless sensor network, the transmit time depends on the available bandwidth and the noise level of the wireless channel. To continuously meet the performance target (image throughput target) under varying channel condition, the SPU clock speed needs to remodulated to account for the degrading/improving wireless channel condition. The SPU clock speed can be controlled by dynamically changing the supply voltage. Figure 16 shows the minimum system clock speed that satisfies the target image throughput assuming a performance target (image throughput target) of 24 images/sec for the 2D- and the 3D-integrated system under varying wireless channel condition. The plot demonstrates that by moving to the 3D-integration of the system and increasing the number of core the performance target can be satisfied with reduced system clock speed, which leads to a potential reduction in power dissipation. Figure 16 also shows that as the wireless channel bitrate deteriorates, the SPU clock speed needs to be boost up to compensate for the slow transmit time. The power dissipation of the system is related to the number of cores and clock speed of the SPU. The optimum core configuration (number of cores and clock speed) can be determined through a design space exploration method across the varying channel condition. The goal is to pick the number of cores and clock speed that leads to the lowest system power on a given channel bitrate. Figure 17a indicates that the least number of cores or the lowest SPU clock speed does not always guarantee lowest power dissipation in meeting the performance target. Figure 17b illustrates the power dissipation for the optimally configured system for both 2D- and 3D-integration. If the channel condition degrades, the system has to spend more power to keep performance up. In addition, the multi-segment image compression is more effective in 3D-integration than in 2D-integration.

2.4.4 Multiple Clock Domain Approach

In the case where the SPU power accounts for the majority of the system power, minimizing the SPU clock speed may reduce the overall power dissipation. Since the multi-core

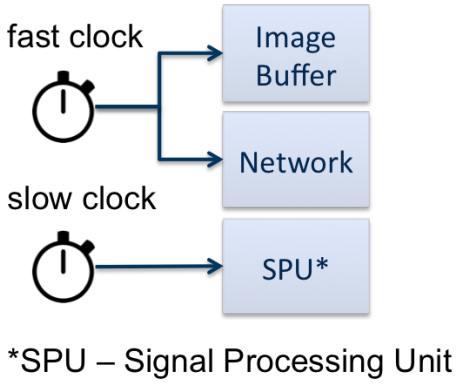


Figure 18: Multi clock domain scheme.

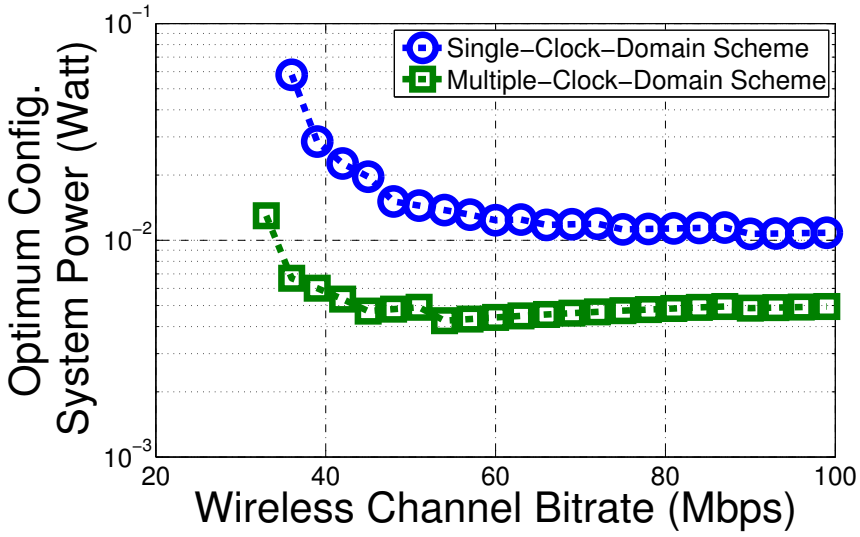


Figure 19: Comparison in the SPU power dissipation between the multi-clock-domain scheme and the single-clock-domain scheme.

SPU and the image buffer is connected by a router network, read and write access takes multiple cycles to complete. The SPU spends multiple cycles without useful load, thus reducing hardware utilization. To reduce the number of idle cycles in the SPU, the SPU clock domain can be separated from the join buffer-network clock domain such that the read and write access is completed within one clock cycle for the SPU 18. In applications such as the wavelet based compression, the read and write access pattern is fix and predictable. Therefore it is possible to estimate the wait time for each read and write access request, and determine a clock ratio between the SPU and the network. A discussion on how to estimate the wait time and clock ratio can be found in [60]. Figure 19 illustrates the potential power reduction that can be achieved from the clock separation. In this experiment, the 3D-integrated system assumes optimal core configurations across varying channel condition with 24 images/sec performance target. By implementing the multiple-clock-domain scheme, the SPU clock speed is reduced to one sixth with the cost of a slight increase in the buffer-network clock speed.

2.5 Analysis for Huffman Coding

In this section, the application of the Huffman coding to the MuSIC engine is discussed and compared with the threshold coding. The Huffman coding is a lossless compression, thus the original image can be perfectly reconstructed. However, in practice, minor distortions are introduced during the quantization stage. Several architectural implementations of the Huffman coding for the MuSIC platform are discussed, and their respective impacts towards the die area, data volume, performance, and power are compared.

2.5.1 System Description of the Huffman Coding

In this work, a zero-run-length Huffman encoder similar to the JPEG Huffman encoder is implemented. The encoder block has an input length of 64 coefficients and a predetermined Huffman table that is implemented by a look-up-table. There are two ways of integrating the Huffman encoder with the multisegment image compression (MuSIC):

(a) a parallel Huffman integration, or (b) a serial Huffman integration.

In the parallel Huffman scheme (Figure 20a), multiple Huffman encoder is implemented so that each core has exclusive access to its own Huffman block. As a result, it allows the DWT coefficients from each core to be directly encoded before they are written back to the image buffer (frame memory) and transmitted off-chip. Since the compression flow, which consists of DWT and thresholding, has already been pipelined, adding the Huffman coding as a stage to the flow does not significantly increase the overall compression time.

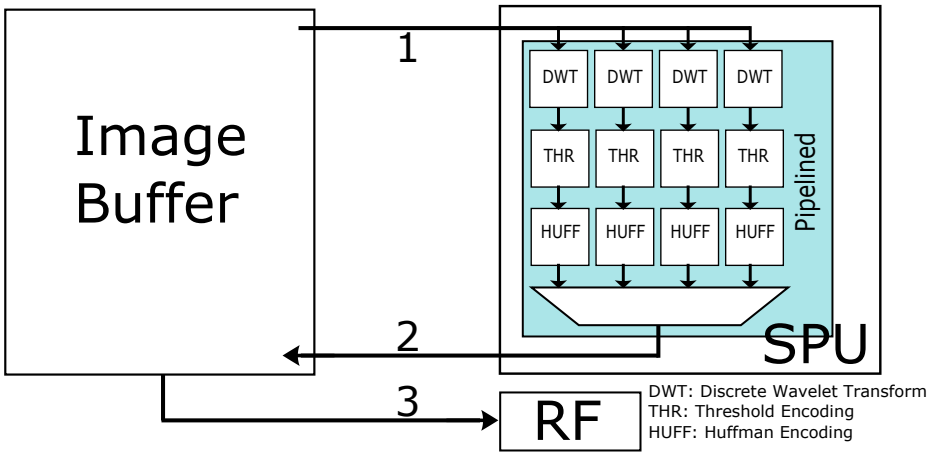
On the other hand, the serial Huffman (Figure 20b) only has one Huffman encoder block in the SPU, thus it has to be shared among the cores. In this case, the DWT coefficients from each core are serialized and written back to the image buffer. Then, they are streamed back to the Huffman block for encoding, and transmitted off-chip. One advantage of the serial Huffman is that it consumes less die area than the parallel Huffman. But, it also leads to a slower compress time than the parallel Huffman.

2.5.2 Die Area of MuSIC with Huffman Coding

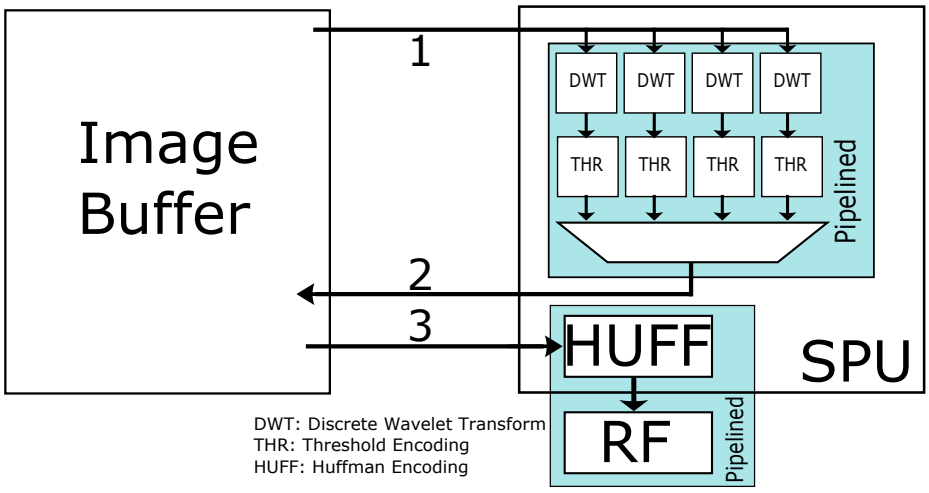
Figure 21 shows the die area of the SPU for serial and parallel Huffman schemes. In the case of a single-core SPU, the Huffman encoder takes around 22% of the SPU die area. Unlike the DWT module, the area of a single Huffman encoder block is independent from the number of cores. In the parallel Huffman schemes, however, the number of the Huffman block increases linearly with the cores. A significant portion of the Huffman encoder is consumed for storing the Huffman table and buffering the input coefficients from the DWT module during the quantization and encoding phase. Figure 22 highlights the area overhead needed to change the encoder from threshold coding to Huffman coding, especially in a many-core design.

2.5.3 Data Volume Overhead of MuSIC with Huffman Coding

In this section, data volume overheads for threshold coding and Huffman coding are compared. Figure 23 shows the impact of increasing number of segments to the data volume and



(a)



(b)

Figure 20: Flow diagram to illustrate: (a) the parallel Huffman scheme, and (b) the serial Huffman scheme.

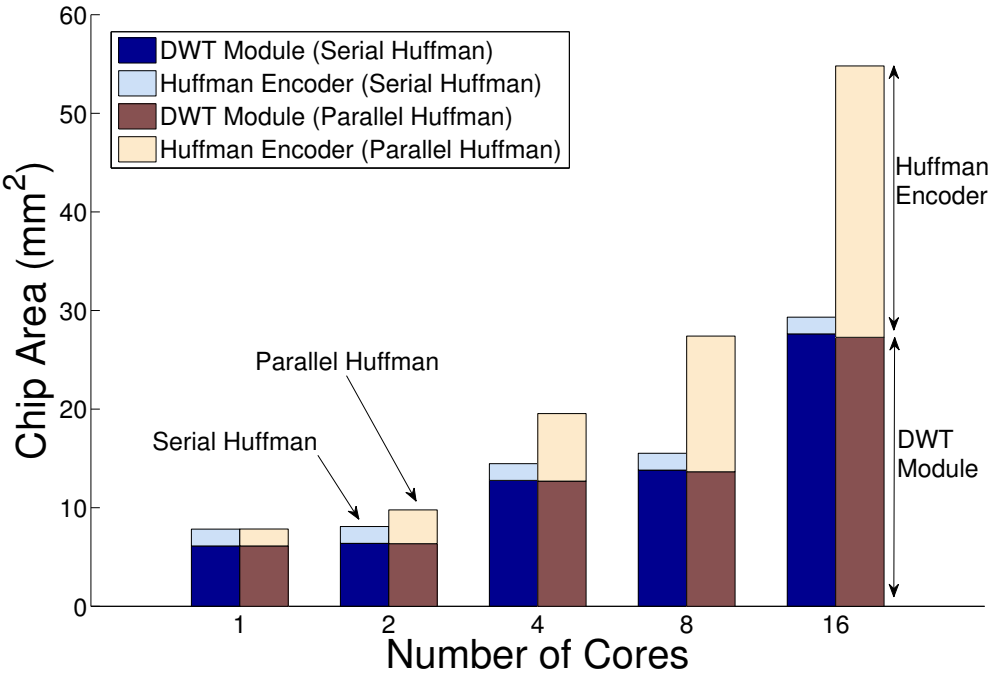


Figure 21: Die area comparison between the serial and parallel Huffman schemes.

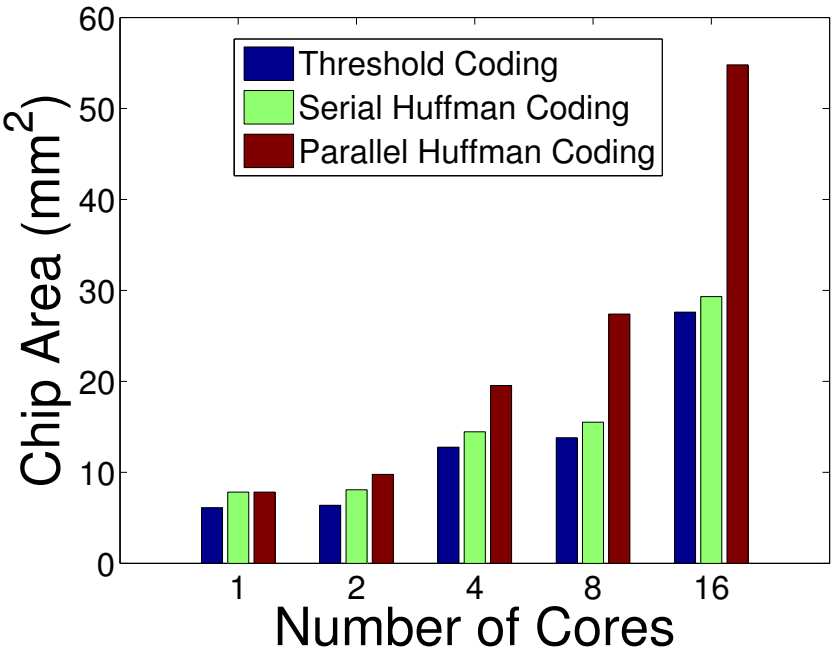
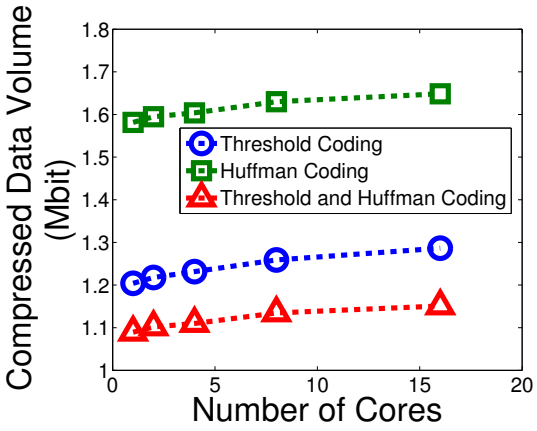


Figure 22: Die area comparison between the Threshold and the Huffman coding shcemes.

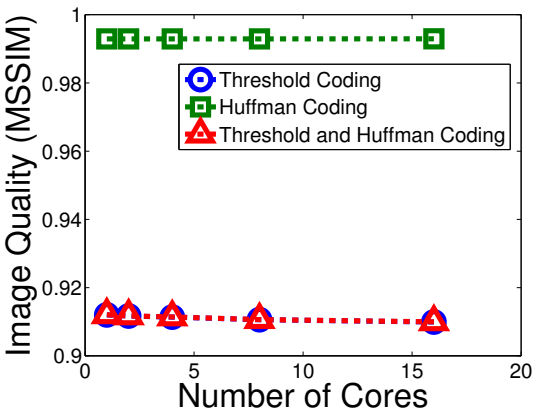
the quality of the compressed image. It is shown that threshold coding yields a lower data volume than the Huffman coding (Figure 23a). Threshold coding is capable of achieving a significantly high compression ratio by increasing the threshold level and removing more coefficients from the encoded data. However, it would also reduce the image quality (Figure 23b). The overhead due to increasing number of cores in Huffman coding is less than in threshold coding (Figure 23c). A significant portion of the overheads associated with the segment borders has zero value, which is compressed more efficiently by the Huffman encoder. On the other hand, adding a Huffman encoding to the thresholded coefficients boosts compression ratio without introducing additional distortions to the image. This is because, in theory, the Huffman encoding is a reversible process, thus image distortions only come from the threshold coding and quantization. In this experiment, a pre-determined generic Huffman table is used to encode the data. The length of the Huffman block is 64 coefficients and the maximum zero-run-length is 16 [61].

2.5.4 Performance and Power of MuSIC with Huffman Coding

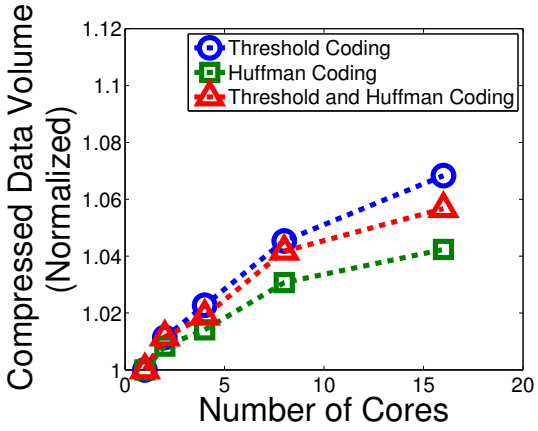
The performance of the image sensor system is determined by the total time required to capture, compress, and transmit the image. Assuming a wireless sensor node that captures and compresses frames of images with a constant throughput under a changing channel condition, Figure 24 shows the minimum system clock speed that meets a throughput target of 24 images/sec for the serial Huffman and parallel Huffman schemes in a 3D-stack. In this analysis, the system with the serial Huffman scheme (Figure 24b) needs to run in a higher clock speed, especially in a high channel bitrate region, than the system with the parallel Huffman scheme (Figure 24c) and the threshold coding scheme (Figure 24a), as shown in Figure 24d. The Huffman coding in the serial Huffman scheme is an extra stage, and it runs at the same clock speed as the DWT core. At high bitrate region, Huffman encoding in the serial Huffman scheme is significantly slower than transmitting the compressed data, thus the SPU clock speed can not be reduced to make up for the Huffman encoding time (Figure 24d).



(a)



(b)



(c)

Figure 23: Effect of increasing number of segments to: (a) the compressed data volume, (b) the image quality, and (c) the data volume normalized to the single-core case for different coding methods.

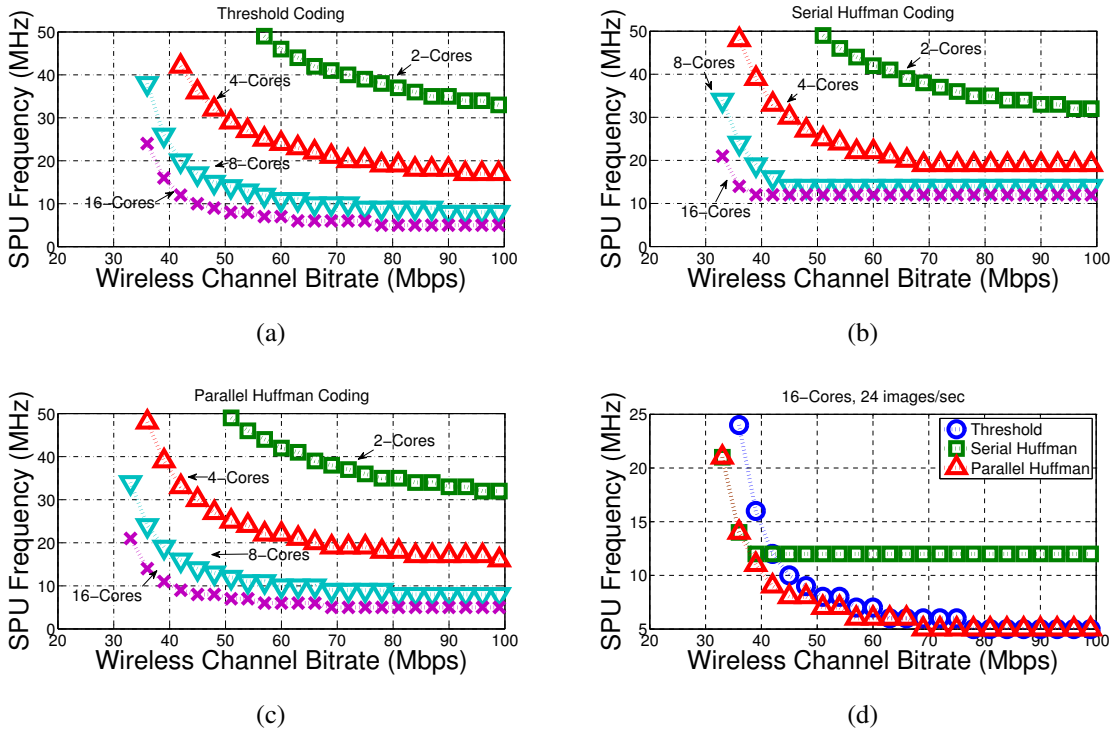


Figure 24: Performance of the image compression system considering variations in wireless channel conditions for various encoding schemes: (a) threshold coding, (b) serial Huffman coding, (c) parallel Huffman coding. (d) SPU clock speed comparison between the threshold, serial Huffman, and parallel Huffman coding for a 16-cores SPU.

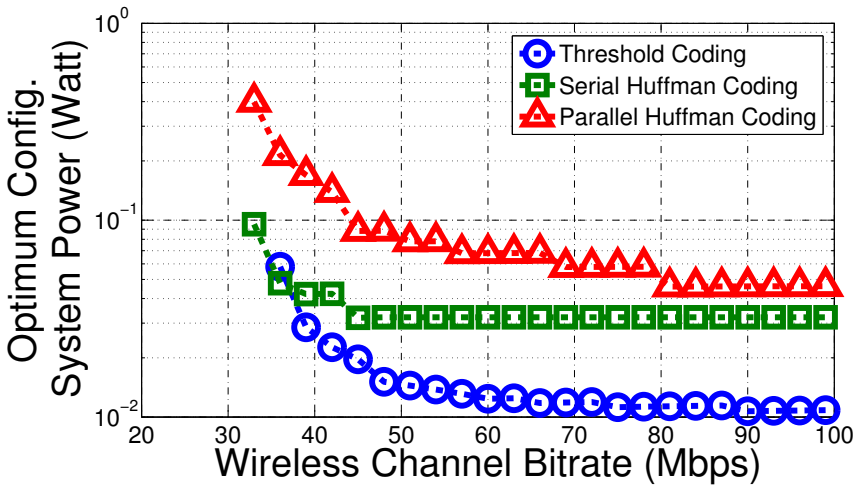


Figure 25: Comparison in the SPU power dissipation between the threshold coding, the serial Huffman coding, and the parallel Huffman coding.

Figure 25 shows the corresponding power dissipation of the systems. At low channel bitrate region, the parallel Huffman scheme consumes the most power of the three schemes due to the extra Huffman encoder blocks. However, as the channel bitrate improves, power improvement in the serial Huffman scheme starts to saturate, due to bottleneck during the Huffman encoding process.

2.6 Summary

An analysis of an image compression unit on a 3D stack for low power wireless sensor node application was presented. The proposed method exploits the high bandwidth advantage from vertical stacking to perform the computation-heavy 2D-DWT transforms using parallel processing concept, by dividing the image into multiple segments and reducing the workload of each core. We analyze the image throughput and the power behavior of the multi-segment architecture under varying off- chip data rate. A design space exploration is presented to optimize the number of cores and the SPU frequency for minimum power consumption in achieving a target IPS under varying channel bitrate condition. The effectiveness of the multi-segment architecture is analyzed for both the 3D- and 2D- integration

of the system. Our analysis shows that the multi-segment architecture benefits from the highly parallel connections of the 3D stack.

CHAPTER 3

NOISE ANALYSIS FOR THE 3D-INTEGRATED IMAGE SENSOR

3.1 Introduction

A key challenge for the 3D image sensor is that the power dissipation in the SPU, image buffer, and ADC tiers generates heat that increases the temperature within the 3D stack. The photodiode arrays on the top tier are thermally coupled with the rest of the stack. In addition, the top layer is covered with glass, which has low thermal conductivity. Hence, power dissipated in the entire 3D stack increases the temperature in the photodiode tier. Figure 26 illustrates the direction of the heat flow. A variation in temperature affects device parameters of the photodiode arrays, the pixel circuits, and the column circuits thereby changing the CMOS Image Sensor (CIS) characteristics (spatial noise and dynamic characteristics) and affecting the image quality.

This chapter analyzes the effect of thermal coupling inside a 3D image sensor with an integrated compression unit. This work builds on the parallel image processing and the multi-segment image compression (MuSIC) architecture for 2D-DWT with threshold encoding algorithm for image compression in a 3D stack. The analysis includes discussions on the power dissipation, thermal coupling, and spatial noise characteristics of a CMOS image sensor with 3D stacked multi-core SPU. The power and performance of the integrated sensor is studied through simulations in 180nm CMOS technology and a logarithmic CIS [62]. The coupled analysis shows that 3D stacking of the image compression unit results in a strong interaction between the image quality, desired image throughput, and environmental conditions.

3.2 Background

This section introduces the basic operating concept of the Logarithmic CMOS image sensor (logarithmic CIS), and the noise components associated with its operation.

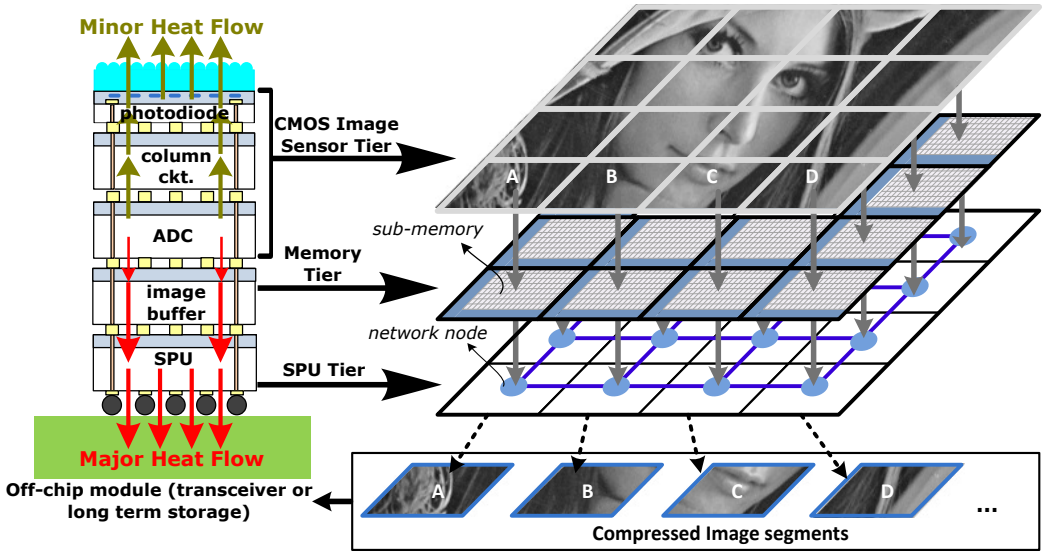


Figure 26: Conceptual diagram of the 3D integrated image sensing and compression system.

3.2.1 Operations of a Basic Logarithmic CMOS Image Sensor

A typical logarithmic CMOS image sensor (CIS) circuit is shown in Figure 27 [20]. The photodiode is generating small amount of current, and makes M1 operates in subthreshold region. In subthreshold, V_{gs} of M1 has logarithmic relationship to I_{ds}. M1 is diode connected so that it stays in subthreshold saturation region. The photoresponse voltage at node V_{pd} is described by the following equation [63],

$$V_{pd} = V_{DD} - V_{th,M_1} - \frac{nkT}{q} \ln\left(\frac{I_{ph} + I_{dc}}{I_{s0}}\right) \tag{2}$$

where I_{ph} and I_{dc} are the photocurrent and dark current(reverse biased current) of the diode. In general, photocurrent (I_{ph}) is linearly proportional to light intensity sensed by the photodiode [48, 63]. The rest of the circuit works as source follower and select circuitry for the logarithmic CIS.

3.2.2 Noise Elements in the Logarithmic CMOS Image Sensor

There are two types of noise elements in the operations of the logarithmic CIS:

1. Spatial noise: the pixel-to-pixel variation in the output values across the CIS unit.

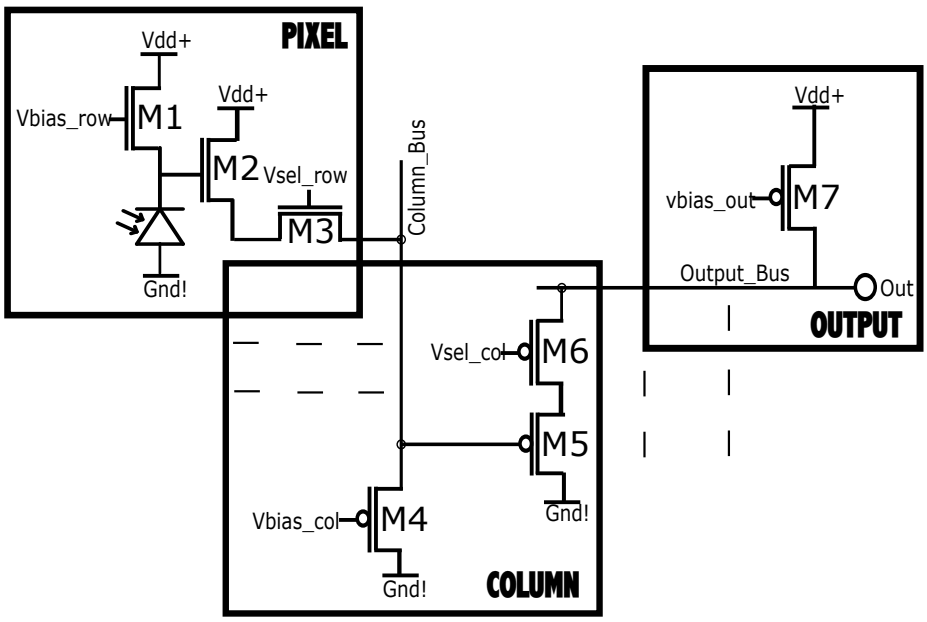


Figure 27: Logarithmic CMOS Image Sensor Circuit.

2. Temporal noise: the temporal variation in the output values of a single pixel unit under constant illumination and temperature.

Spatial noise is the main contributor of image distortion in the logarithmic CIS. It is caused by variations in device characteristics, such as doping concentrations and feature sizes, from pixel to pixel [64]. Unlike temporal noise, spatial noise does not change with time and produces certain distortion pattern on the captured image, hence the name fixed pattern noise (FPN). In a logarithmic CIS, FPN is nonlinearly coupled with illumination [20, 19, 62, 22]. In addition, the response of the logarithmic image sensor and FPN is temperature dependent since photodiode reverse bias current, and transistor parameters (i.e. threshold voltage, leakage current, sub-threshold slope) depend on temperature [48]. The dependence of the pixel response on temperature is a coupled effect of linear decrease of threshold voltage and increase of dark current (doubles with approximately 7 °C) with an increase in the temperature [63, 65].

On the other hand, temporal noise varies with time, and is caused by thermal noise and flicker noise associated with the devices in the readout path and shot noise associated

with the photodiode and transistor subthreshold operation [13, 66, 10, 67, 68]. Among the temporal noise elements, only thermal noise depends on temperature. It is caused by thermally induced agitation of charge carriers in resistive regions such as the transistor channel in strong inversion. The shot noise is caused by the fluctuations in the number of discrete charge carriers passing through the depletion region. The shot noise is the dominant noise source in the photodiode operation. It is related to variation in the photons conversion, as well as dark current generation. The flicker noise is influenced by a number of factors including contaminants in the devices. The power spectral density of this noise is dominant in the low frequency region, but it quickly drops below the thermal noise level in the high frequency spectrum.

3.2.3 Fundamentals for Modeling the Spatial Noise of the Logarithmic CIS

The relation between digital output response (Pixel), illuminance (Ev), FPN, and temperature of a typical logarithmic image sensor has been previously derived and modeled in [16]. During pixel read operation, the output voltage of the CIS (V_{out}) in Figure 27 is equal to the voltage at the source of transistor M5 ($V_{s,M5}$), the gate voltage of transistor M5 ($V_{g,M5}$) is equal to the source voltage of transistor M2 ($V_{s,M2}$). It is governed by the following equations:

$$V_{out} = V_{g,M5} - V_{th,M5} - \sqrt{\frac{\mu_{M7} C'_{ox,M7} \frac{W_{M7}}{L_{M7}}}{\mu_{M5} C'_{ox,M5} \frac{W_{M5}}{L_{M5}}}} \left(V_{gs,M7} - V_{th,M7}\right) \tag{3}$$

$$V_{g,M5} = V_{g,M2} - V_{th,M2} - \sqrt{\frac{\mu_{M4} C'_{ox,M4} \frac{W_{M4}}{L_{M4}}}{\mu_{M2} C'_{ox,M2} \frac{W_{M2}}{L_{M2}}}} \left(V_{gs,M4} - V_{th,M4}\right)$$

$$V_{g,M2} = V_{pd} = V_{DD} - V_{th,M_1} - \frac{nkT}{q} \ln\left(\frac{I_{ph} + I_{dc}}{I_{s0}}\right)$$

$$I_{ph} = G_{optics} E_v$$

where E_v corresponds to the light intensity on the surface area of the photodiode or illuminance, and G_{optics} is the gain related to the optics of the sensor. The relations between the output voltage of the CIS (V_{out}) and illuminance (E_v), shown in (3), is simplified into the

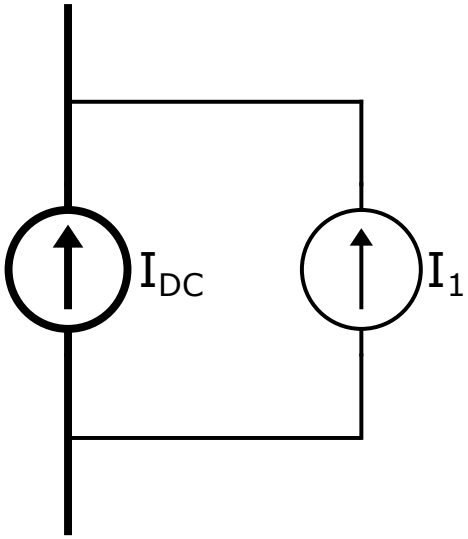


Figure 28: Circuit model for thermal and flicker noise.

following form:

$$V_{out} = a + b \ln (c + E_v) + \epsilon \tag{4}$$

where a, b, and c are the offset, gain, and bias of the logarithmic CIS. The gain and bias of the CIS vary from pixel-to-pixel due to process variations and are a function of temperature. The term ϵ corresponds to the error associated with temporal noise. For FPN analysis, ϵ can be minimized by taking multiple samples over time and averaging the response. The conversion from CIS output voltage to pixel value is modeled using ideal ADC and assumed to be independent of temperature variation as:

$$Pix = clip(F_{ADC} + G_{ADC} V_{out}) \tag{5}$$

where FADC and GADC are the offset and gain of the ADC respectively, to code the CIS output voltage to an integer value between 0-255.

3.2.4 Fundamentals for Modeling the Temporal Noise of the Logarithmic CIS

In this subsection, we recall the fundamental mathematical model of the temporal noise elements (thermal, shot, and flicker) for the photodiode, transistors, and ultimately the logarithmic CIS [69, 70, 71].

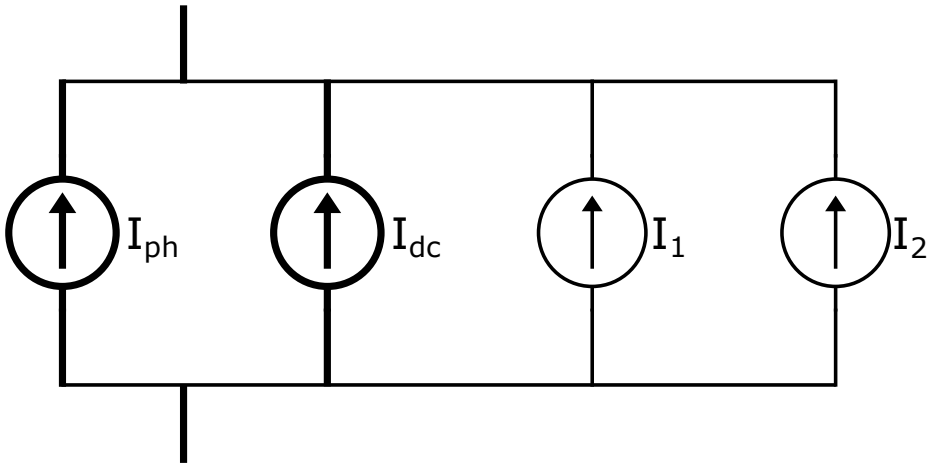


Figure 29: Circuit model for temporal noise of a photodiode.

The dominant sources of temporal noise in a photodiode are shot noise and flicker noise. The shot noise is modeled as a current source in parallel with a DC current source (I_{DC}) with a Gaussian distribution over a wide bandwidth (Figure 28). It's power spectral density is described by the following equation:

$$S_I(f) = qI_{DC}, \forall f < f_{max} \tag{6}$$

where I_{DC} is the DC current source, and q is the electron charge constant. The flicker noise is also modeled as a current source in parallel with a DC current source, however it's power spectral density drops with f. It is described by the following equation:

$$S_I(f) = \frac{k_f I_{DC}^a}{f^e} \tag{7}$$

where k_f is a process dependent constant, I_{DC} is the DC current source, e is often assumed to be one, and a is often assumed to be between 0.5 and 2.

Now, the total noise of the photodiode can be modeled as a parallel combinations of the shot noise due to photocurrent generation, shot noise due to dark current generation, flicker noise due to dark current generation, photocurrent, and dark current sources (Figure 29). The corresponding power spectrum densities of the noise sources are described by the following equations:

$$S_{I1}(f) = q\left(I_{ph} + I_{dc}\right), \forall f < f_{max} \tag{8}$$

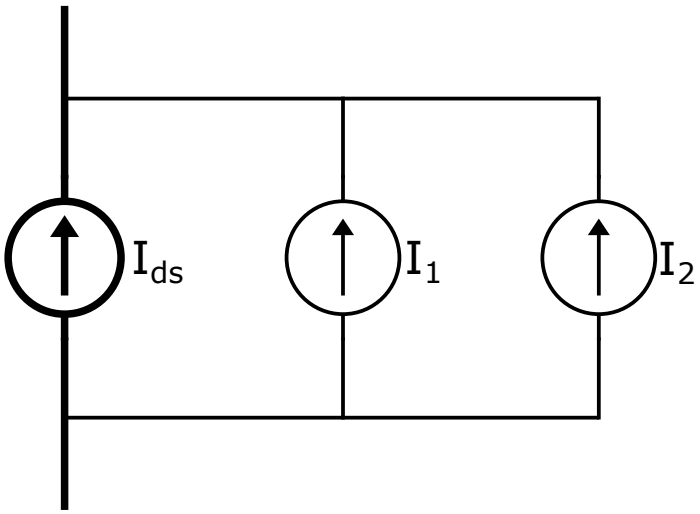


Figure 30: Circuit model for temporal noise of a CMOS transistor.

$$S_{I2}(f) = \frac{k_f I_{dc}^a}{f} \tag{9}$$

where I_{dc} and I_{ph} are the dark current and photocurrent of the photodiode.

In a CMOS transistor, the dominant sources of the temporal noise depends on the operating condition of the the device itself. In a strong inversion mode transistor channel is resistive, thus thermal noise and flicker noise are the dominant noise sources. If the transistor is operated in the subthreshold region, the dominant sources of noise are shot noise and flicker noise. Figure 30 illustrates the circuit representation of the temporal noise model in a CMOS transistor. The noise sources are independent of each other and are modeled as parallel combinations of current sources. The noise power spectral density equations for the transistor in saturation region are described as follow:

$$S_{I1}(f) = 4kT\frac{2}{3}g_m, \forall f < f_{max} \tag{10}$$

$$S_{I2}(f) = \frac{k_f I_{ds}^a}{f} \tag{11}$$

where S_{I1} and S_{I2} correspond to the thermal and flicker noise sources respectively, k is the Boltzmann constant, T is the absolute temperature in Kelvin, and g_m is the transconduction of the transistor. The noise power spectral density of the transistor in the linear region is

described by the following equations:

$$S_{I1}(f) = \frac{4kT}{R_{on}}, \forall f < f_{max} \tag{12}$$

$$S_{I2}(f) = \frac{k_f I_{ds}^a}{f} \tag{13}$$

where S_{I1} and S_{I2} correspond to the thermal and flicker noise sources respectively, and R_{on} is the channel resistance of the transistor in linear region. In the subthreshold region, the noise power spectral density of the transistor is governed by the following equations:

$$S_{I1}(f) = qI_{ds}, \forall f < f_{max} \tag{14}$$

$$S_{I2}(f) = \frac{k_f I_{ds}^a}{f} \tag{15}$$

where S_{I1} and S_{I2} correspond to the shot and flicker noise sources respectively, I_{ds} is the subthreshold drain current.

Figure 27 shows the schematic of the logarithmic CIS. During a pixel readout operation, transistor M1 is operated in the subthreshold region. Transistors M2, M4, M5 and M7 act as source followers, thus they are operated in the saturation region. Transistors M3 and M6 act as access transistors, therefore they are operated in the linear region.

3.3 System Simulation and Analysis Framework

The simulation framework is built to analyze a 3D stacked imaging system considering system power and predict FPN, response characteristics, temporal noise, and image quality considering tier-to-tier thermal coupling. The analysis considers SPU, image buffer, and ADCs as the main heat contributors that affect temperature in the pixel-array and column circuit tier. The developed framework, as shown in Figure 31, can be broken down into three separate simulation blocks: 1) power analysis of multicore SPU and image buffer for different number of cores and clock frequencies, 2) thermal coupling analysis of the 3D stack using distributed RC model, and 3) FPN, CIS response, and temporal noise analysis of the logarithmic CMOS image sensor.

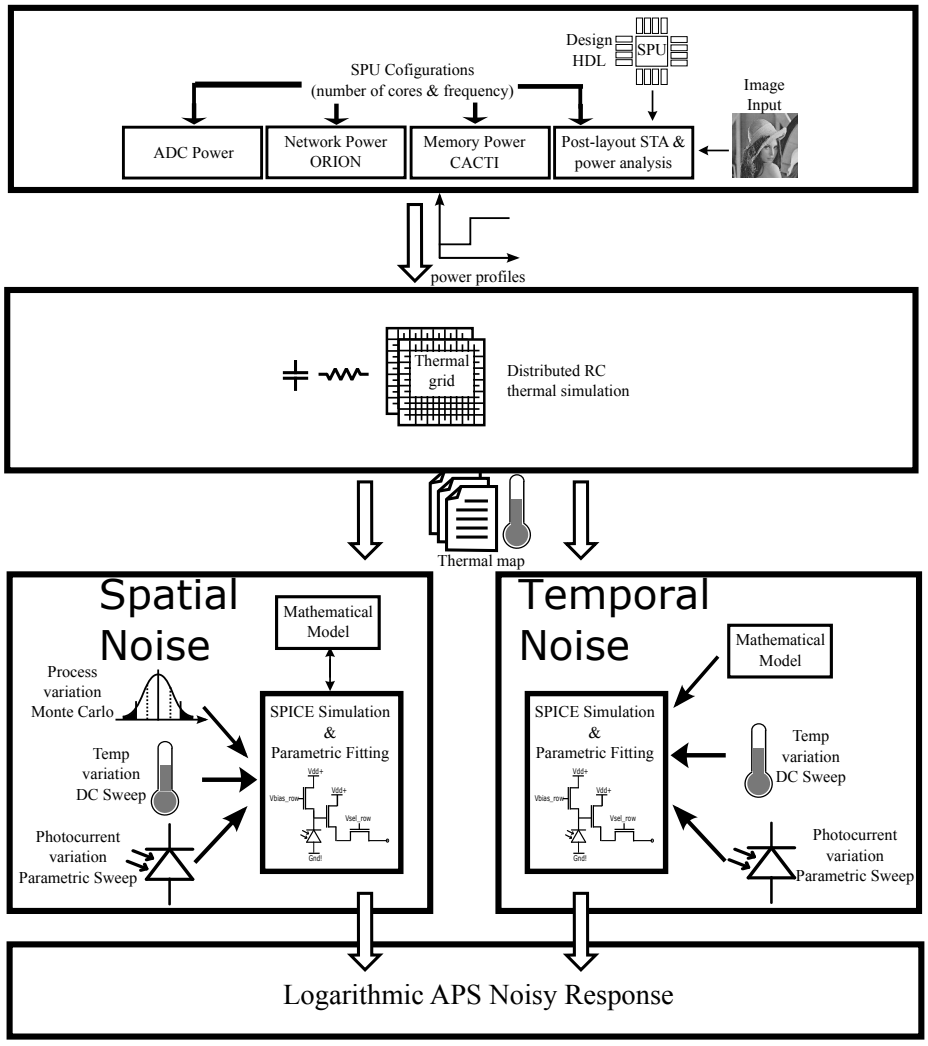


Figure 31: A diagram of the thermal simulation framework for the 3D-integrated image sensor.

3.3.1 Power Analysis of SPU, Image Buffer, and Network

The analysis framework used in this chapter for power estimation is similar to the framework used in Chapter 2. The one-level 2D DWT, with JPEG 5/3 FIR filter pairs are used for wavelet transform in this chapter. Then, the wavelet coefficients are encoded using the threshold coding. The test images are standard 512X512 gray-scale images. To study data traffic movement between image buffer, network, and SPU during read/write request, and to compute the time to compress an image, traffic patterns for the 2D-DWT are generated and injected to the model. The horizontal filtering, vertical filtering, and threshold coding computation of the compression algorithm is pipelined. The SPU was implemented in synthesizable hardware description language (HDL). The post-layout design was simulated in 180nm process, taking into account its parasitic components.

To estimate the power dissipated by the compression unit, test images are used to obtain switching activities data during processing. The SPU clock frequency is again limited to 50 MHz. A static timing analysis is considered in meeting the delay specs. The image buffer access energy and the leakage power are extracted from CACTI assuming 180nm SRAM with low leakage devices, and single read/write port [57]. The size of the image buffer module is designed only for storing the raw data and the compressed coefficients of a single image. The router access energy and the link switching energy are extracted from ORION [56] in 180nm node. In our work, each ADC is to accommodate a 32X32 pixels array block. A total of 256 ADCs are used to form the 512X512 pixels imager. The total power consumption of the ADCs is estimated to be 230.4 mW following the design by Kiyoyama et. al. [51].

The total compression power ($P_{compress}$) is estimated by adding the SPU (P_{SPU}), network ($P_{network}$), and image buffer (P_{SRAM}) power, as shown below:

$$P_{compress} = P_{SPU} + P_{network} + P_{SRAM} \tag{16}$$

Total dynamic portion of the image buffer and the network power depends on the image throughput and the image size. In addition, the network access energy primarily consists

of the energy associated with accessing each router node, and the link between two nodes. The link access energy is related to the lateral (node-to-node) movement of the data.

3.3.2 Thermal Analysis of 3D Stack

Our thermal simulation framework is built using a distributed 3D RC grid where R represents the thermal resistance and C represents the specific heat [46, 47]. The constructed grid (Figure 32) consists of a five tiers stack comprising of photo-arrays, column circuitry, ADCs, image buffer, and SPU. The stacks are assumed to be face-to-back bonded. The die thickness of each tier is assumed to be 40um, with thickness of bonding layer assumed to be 5um, is derived from the ITRS roadmap. The effective thermal conductivity is calculated using the existing weighted average method (parallel model) with a 1:3 metal to oxide ratio. For the FEOL and bulk, we assume the composite to be silicon (149 W/(m.K)) and copper (385 W/(m.K)) [72], while for the BEOL, we assume the oxide to be SiO2 (1.4 W/(m.K)) [73] and the metal to be aluminum (205 W/(m.K)). The termination resistor RFC is estimated based on the thermal conductivity of the free convection of air. Table 2 shows the parameters used for thermal simulation. The top layer of the image sensing is covered with microlens, trapped air, and optical lens. Due to their thermally non-conducting property, in calculating the thermal resistance of the top layer, we simplified the thermal conductivity of the top layer to be 0.001 W/(m.K) with 1mm thickness. In calculating the thermal resistance of the bottom layer, we use the thermal conductivity of the plastic interposer as 0.5 W/(m.K) with 1mm thickness.

Table 3 presents the dimensions of the different tiers. The image buffer consumes the highest chip area, followed by the SPU. The photodiode array, the column circuitry, and the ADC are connected and grouped together in a block parallel structure. Each parallel block carries 32X32 pixels, a set of 32 column source follower circuits with one output source follower, and one ADC. The area of the photodiode tier is estimated to be 0.16X0.16 mm^2 for each parallel block based on the 5X5μm^2 pixel size (36% fill factor). The area of the column circuitry for one parallel block is 0.033X0.001 mm^2. The area of one ADC unit is

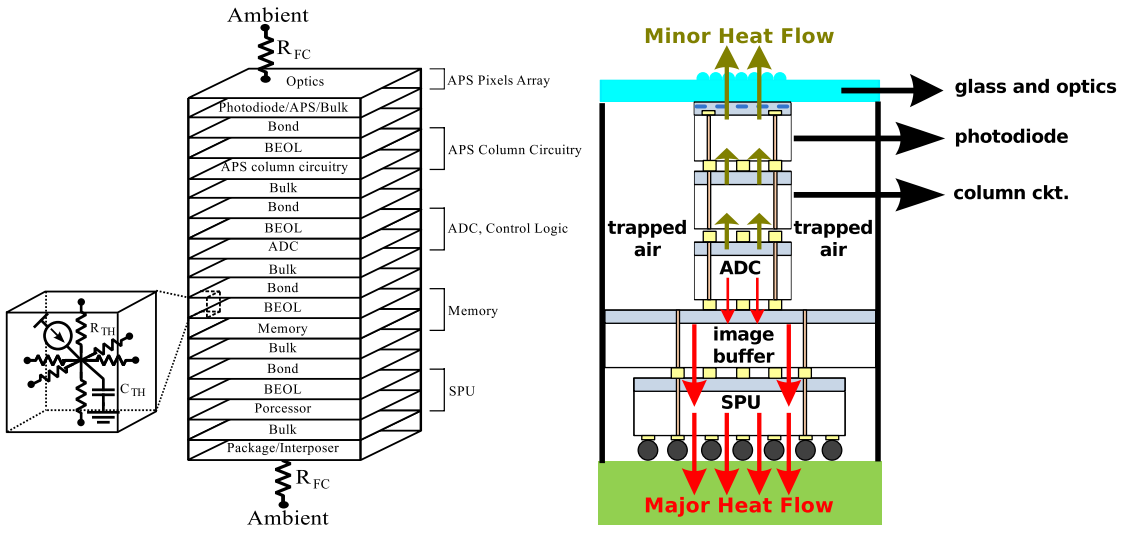


Figure 32: Thermal grid model of the 3D image sensor: (a) The grid unit cell and the stacked layers used in the thermal grid model. (b) The 3D stacking scenario of the image sensor system.

Table 2: Materials Parameters

Parameters	Thickness (m)	R (W/mK)	C (J/m^{3}K)
Optics	0.001	0.025	3.55 M
BEOL	16 μ	40	4 M
Device layer	4 μ	200	1.75 M
Bulk	20 μ	200	1.75 M
Bond	5 μ	0.1	4 M
Package	0.001	0.5	3.55 M

Table 3: Grid and Area Parameters

Number of Cores	1	2	4	8	16
Number of parallel block	256	256	256	256	256
No. of pixels per block parallel	32x32	32x32	32x32	32x32	32x32
Photodiode array area (mm^2)	2.56x2.56	2.56x2.56	2.56x2.56	2.56x2.56	2.56x2.56
Column ckt. area (mm^2)	2.56x2.56	2.56x2.56	2.56x2.56	2.56x2.56	2.56x2.56
ADC area (mm^2)	2.56x2.56	2.56x2.56	2.56x2.56	2.56x2.56	2.56x2.56
Image buffer area (mm^2)	7.68x7.68	7.68x7.68	7.68x7.68	7.68x7.68	7.68x7.68
SPU area (mm^2)	3.2x3.2	4.48x2.56	4.48x4.48	6.4x3.2	6.4x6.4

determined to be 0.1X0.1 mm^2. Thus the minimum effective areas of the photodiode array, column ckt., and ADCs for a 512X512 pixels are 2.56X2.56 mm^2, 0.528X0.016 mm^2, and 1.6X1.6 mm^2 respectively. The die size for the top three stacks can be determined to be 2.56X2.56 mm^2. They are connected together via TSVs. The image buffer consists of SRAM array. The area of the SRAM array is extracted from CACTI [57]. The SPU area changes with different number of cores. Both the metal wires and the TSVs are needed to connect the SPU tier, the image buffer tier, and the ADC tier. The photodiode array, column circuitry, and ADC have similar area dimension, which is smaller than the area of the image buffer and the SPU. The 3D stack is aligned to the center, and to match the grid dimensions between all the tiers in the thermal simulation, the grids of the photodiode array tier, the column circuitry tier, the ADC tier, and the SPU tier are padded with the resistivity of trapped air.

3.3.3 Noise Analysis of a Logarithmic CMOS Image Sensor

The image sensor circuit was simulated using commercial SPICE tools in 180nm 1.8V CMOS process technology. The base simulation methodology for estimating FPN of the

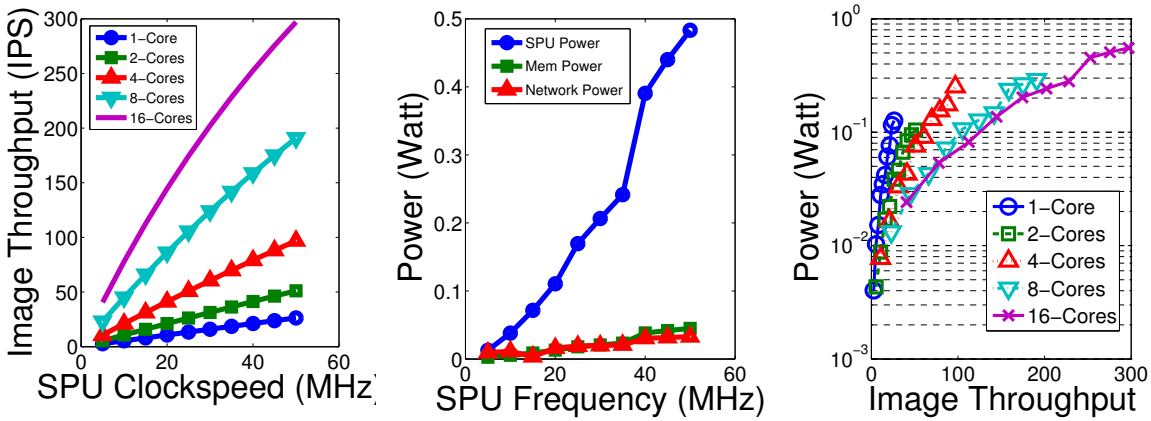
logarithmic CIS responses is adopted from [48]. A simultaneous temperature sweep, photocurrent sweep, and Monte-Carlo simulation was performed. Data points for the temperature sweep are obtained from the thermal simulation described in the previous section. The Monte-Carlo simulation was repeated 1000 times to simulate mismatch in the device. The output voltage of the logarithmic CIS values and the photodiode current values from every Monte-Carlo simulation are fitted to equation (4) to extract the offset, gain, and bias information of the mismatched logarithmic CIS. Then, each pixel for the logarithmic CIS is randomly selected from the batch of 1000 mismatched devices.

To estimate the contributions of temporal noise elements at the output node, a simultaneous temperature and photocurrent sweep with noise analysis simulation was performed. The squared output noise voltage is calculated by integrating the noise power spectral density across the noise spectrum. The noise spectrum is determined to be the minimum of the input bandwidth of the ADC and the gain bandwidth of the pixel circuit. The temporal noise level of each pixel is estimated to be a random voltage that has a gaussian probability distribution function. The standard deviation is the root mean square (RMS) value of the noise voltage.

3.4 Simulation Results

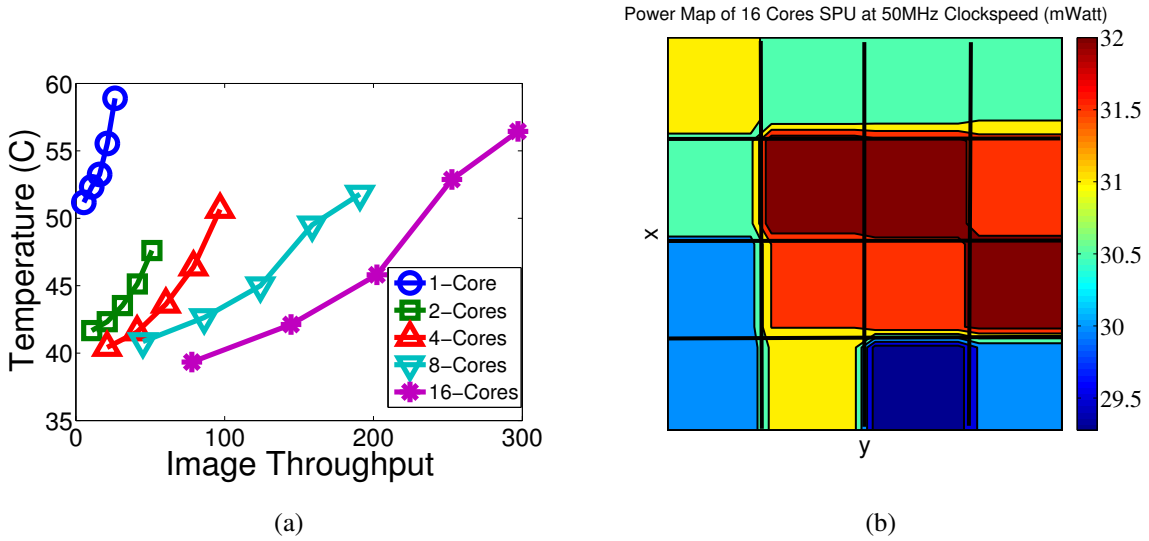
3.4.1 Power and Performance of Image Compression Unit

The first part of the investigation involves analyzing the relation between the clock frequencies, the number of cores, the image throughput, and the power dissipation of the image compression unit. The simulation method to do this is introduced in Chapter 2, and explained in Section 3.3.1. We use 1-, 2-, 4-, 8-, and 16-core SPU for our case studies. First, we demonstrate the system's sensing and processing capability in achieving high compression rate under unrestricted wireless channel output data rate. Figure 33a shows the achievable image throughput increases almost linearly with increasing number of cores as well as clock frequency. Figure 33b shows the power dissipation of a 16-core system with varying clock frequency. The SPU power dominates over the image buffer and the



(a) (b) (c)

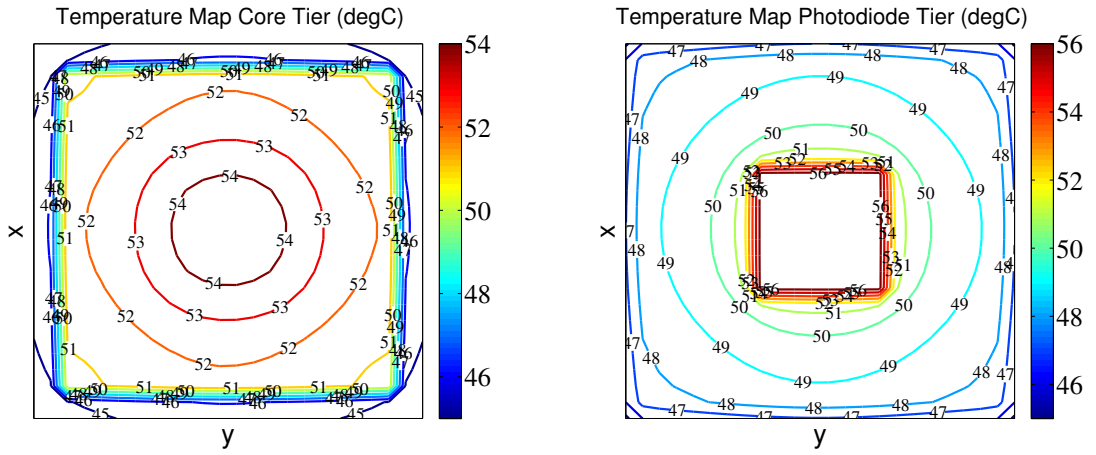
Figure 33: The analysis of the performance and power of the 3D image compression unit: (a) Image rate with varying SPU clock speed of the multi-core image compression system. (b) Power dissipation versus SPU clock speed of a 16-Cores image compression system. (c) Power versus performance of the multicore image compression system, when the SPU clock speed is varied from 5 MHz to 50 MHz.



(a) (b)

Figure 34: Thermal analysis of the 3D image sensor: (a) Temperature variation of the photo-diode tier with varying image rate throughput. The throughput corresponds to the following SPU clock speed: 10, 20, 30, 40, and 50 MHz. (b) Power map of the 16-cores SPU at 50 MHz.

network power. The clock network consumes majority of the SPU power. Majority of the power spent by the clock network is used to drive the clock signals of the registers in the intermediate registers of the 2D-DWT module. In our power simulation model, we consider voltage scaling to achieve a target throughput with minimum power. Figure 33c shows the performance versus power curves of the system with different number of cores as the SPU clock speed is varied from 5 MHz to 50 MHz. The system with 16-core SPU provides the highest image throughput for the same amount of power spent by the compression unit. Multicore system allows lower clock speed operation, which leads to lower supply voltage operation, thus reducing power dissipation.

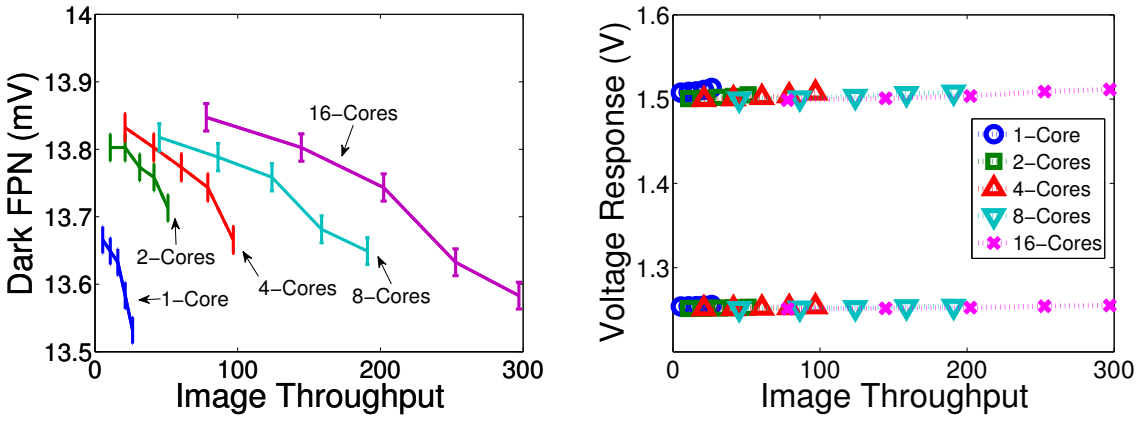


(a) (b)

Figure 35: Thermal analysis of the 3D image sensor: (a) Temperature map of the core tier, and (b) temperature map of the photodiode tier of the 16-cores system at 50 MHz.

3.4.2 Thermal Coupling, and CIS Output Response

In this section, we discuss the relations between the power dissipation, the temperature, the output responses of the CIS, and the image quality. For the simulation, we set the ambient temperature to be 25C. Figure 34a shows the projected temperature from the power profile shown in Figure 33c. The 1-core system is shown to have a high operating temperature, due to the small SPU die area thus having a high SPU power density and low heat spreading. Variations in the power dissipation from core-to-core are less than 11% at the extreme case, which results in a relatively uniform surface temperature on the photodiode tier. Figure 34b shows the core-to-core power map for the 16-core SPU running with a 50 MHz clock speed. The temperature map of the SPU tier and the photodiode tier is shown in Figure 35a and Figure 35b. The SPU and the ADC contribute most of the generated heat. Temperature difference between the photodiode tier (top tier) and the column circuitry tier (2nd tier from the top) is very small. The two temperatures are practically equal even with considerable power dissipation in the ADC. The small temperature difference between the column circuitry tier and the photodiode tier is due to the close proximity of the two layers,



(a) (b)

Figure 36: Effect of the image throughput to the sensor noise and the output voltage response considering a plain dark (black) image, and a plain bright (white) image: (a) The fixed pattern noise (FPN) at dark considering variations in 100 imagers, and (b) the output range of the logarithmic CIS with varying image throughput.

and the very low heat-flow through the glass (lens, trapped air, microlens) cover at the top. Therefore, the heat generated by the ADC, memory, and SPU flows through the package and hence, the temperature difference between the tiers above ADC is very low. In our thermal simulation, the top layer of the image sensing is covered with microlens, trapped air, and optical lens. If the thermal conductivity of the die-to-die interface is reduced or the thermal conductivity of the glass cover is increased the temperature difference between the two tiers will increase.

Figure 36 shows the standard deviation of the pixel-to-pixel output voltage variations at dark lighting condition (dark FPN). The corresponding error bars illustrate estimated variations in the dark FPN across 100 imagers. The change in the FPN at dark with temperature is shown to be marginal (Figure 36). However, Figure 36b shows the change in the output voltage range of the CIS with temperature is noticeable. The top limit of the voltage corresponds to current in the dark image condition. The bottom limit of the voltage corresponds to current in the bright image condition. The output voltage range is 240 mV at ambient

temperature. At very high temperature of 90C, the top limit and the bottom limit of the voltage could change by as much as 47 mV and 16 mV, respectively. If the ADC input voltage range is set up for operation at the ambient temperature, then as the temperature of the CIS increases, there will be a mismatch between the CIS output voltage range and the ADC input voltage range, which can contribute to error in converting a pixel voltage to the digital bit strings for the image.

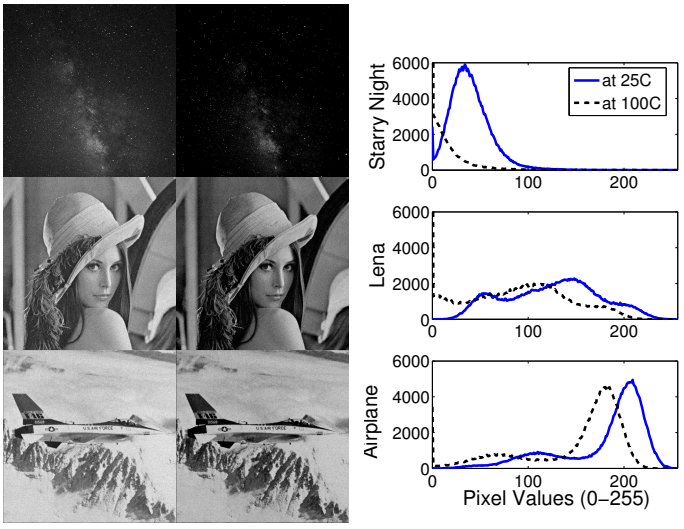
3.4.3 Spatial Noise

3.4.3.1 The Effect of Image Throughput on Image Quality

To measure the variations in the image error rate, we use the noise-induced image at ambient temperature as the baseline image. The ADC is assumed to be ideal and is optimized for use at the ambient temperature. We compare test images that have been injected with FPN profiles for different temperature points against the baseline images. Figure 37a visually shows the change/degradation in three test images due to a temperature change from 25C to 100C on the photodiode tier. Figure 37b shows that with the increase in temperature, the histogram shifts to the left. The top and the bottom limit of the CIS output voltage increases with temperature, thus every pixel will seem darker with increasing temperature.

To assess the image quality, we use a Peak Signal to Noise Ratio (PSNR), which is a metric of purely numerical comparison between two images [59], without any reference to a human perception. A high PSNR means that the two images being compared are numerically similar to each other. The effect of light intensity to the image quality under varying image throughput is shown in Figure 38. We scale the brightness of the airplane image to create a dark, and a bright version of the image. It is shown that the bright version of the airplane image generally has a higher PSNR, while the dark version of it has a lower PSNR than the normal version of it. Thus, the effect of the change in temperature in the photodiode tier to the image quality degradation is more severe in the darker version than in the brighter version of the image.

The rest of the analysis considers the effect of the different scenes as well as the lighting



(a) (b)

Figure 37: A scene of a starry night, Lena, and an airplane from top to bottom. (a) Effect of FPN distortion to the images at 25C (left) and 100C (right). (b) Pixel histogram of the images at 25C and 100C.

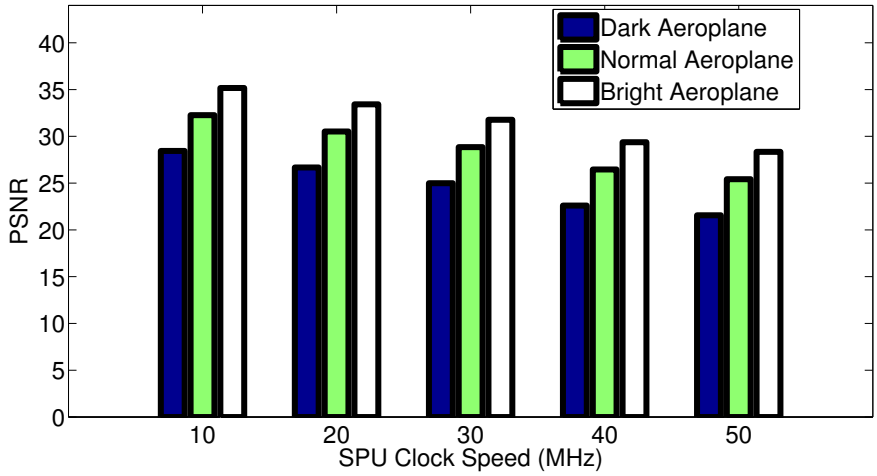


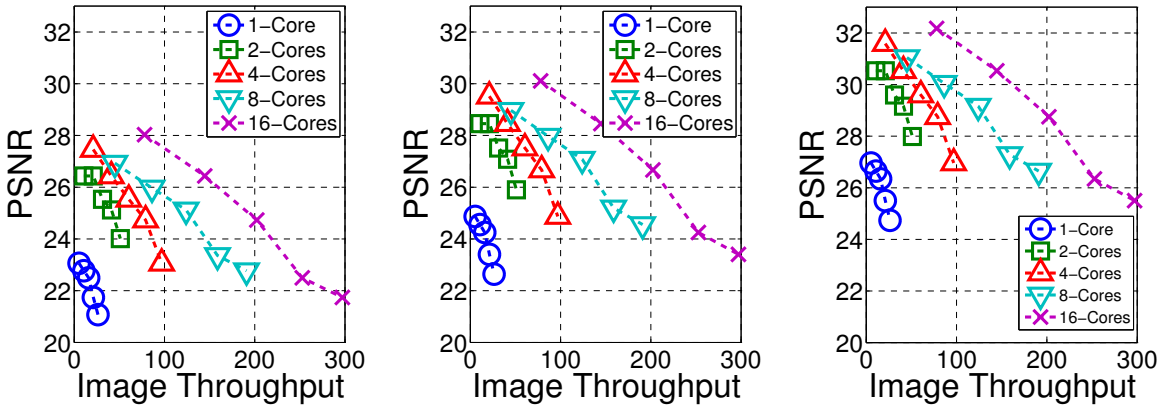
Figure 38: Effect of the lighting condition to the image quality of the airplane scene due to spatial noise: PSNR comparisons of the airplane scene in different lighting condition when the image is processed by a 16-cores system.

conditions in the test images. We define the starry night image as the dark image (i.e. the image has high distribution of low light intensity pixels), the Lena image as the normal image (i.e. the image has balanced distribution of low light intensity pixels as well as high light intensity pixels), and the airplane image as the bright image (i.e. the image has high distribution of high light intensity pixels). Figure 39 captures the achievable image quality and throughput for different number of cores as the SPU clock speed is varied from 10 MHz to 50 MHz. The figure shows a reduced PSNR with an increase in the image throughput as the DWT cores operate with a higher clock speed and generate more heat. Figure 39 shows that the PSNR of the dark image is lower than the normal image and bright image. The CIS output voltage variation in the low light condition is severe compared to in the bright light condition. Lower number of cores results in not only low image throughput, but also low PSNR due to the high temperature in the system (Figure 34a). This shows the advantage of multicore systems in achieving high image throughput while maintaining reasonable image quality. The above analysis shows that with the 3D integrated image compression unit, the CIS noise and associated quality degradation are correlated to the desired image throughput, and the correlation depends on the lighting condition.

3.4.3.2 The Effect of Output Datarate on Image Quality

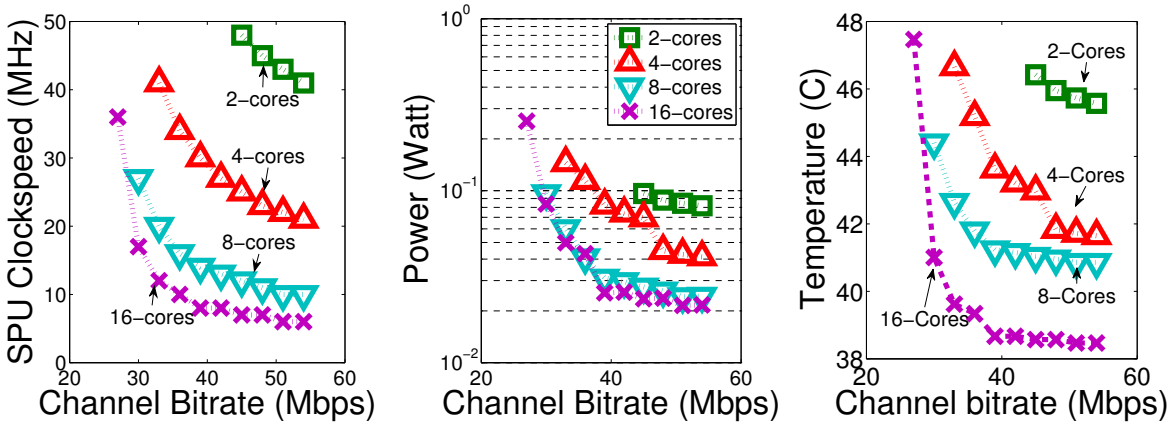
We next consider the case of a wireless image sensor where the CMOS image sensing and processing system is connected to a standard wireless transmitter. In a wireless network the output bitrate from the sensor depends on the available bandwidth and the noise levels in the channel. To meet the image throughput under variable channel data rate, the compression time ($t_{compress}$) needs to be changed (assuming constant compression ratio). The image compression unit needs to modulate the clock frequency and hence, the power dissipation. Therefore, due to the thermal coupling in the 3D image sensor, there will be an inherent correlation between the variable output bit-rate and the sensor noise.

We consider a target image throughput is 24 images/sec. A reduction in the channel bitrate implies an increase in the required SPU clock speed to satisfy performance (Figure



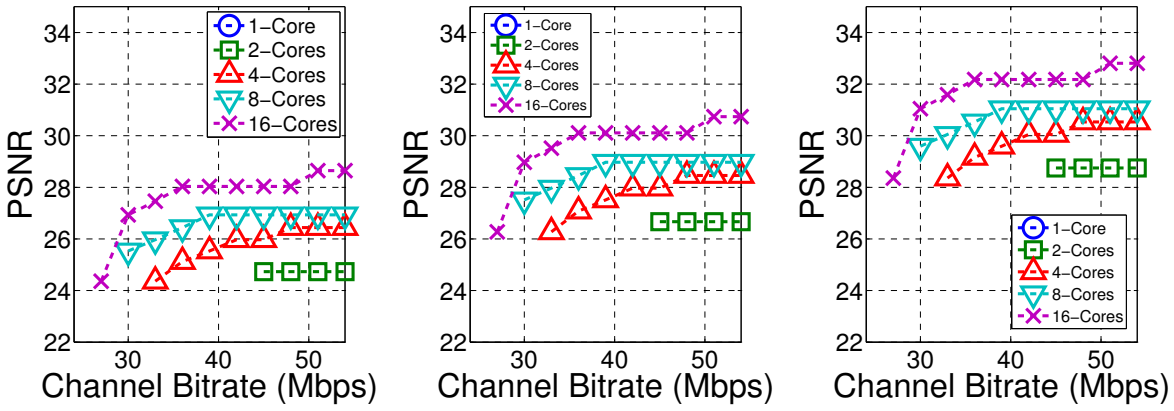
(a) (b) (c)

Figure 39: Effect of the lighting condition to the image quality due to spatial noise: (a) scene with dark lighting, (b) scene with normal lighting, and (c) scene with bright lighting. The throughput corresponds to the following SPU clock speed: 10, 20, 30, 40, and 50 MHz.



(a) (b) (c)

Figure 40: Effect of the wireless channel bitrate on the 3D image sensor, assuming a 24 images/sec throughput: (a) SPU clock speed versus channel bitrate. (b) Power (SPU, image buffer, and network power) versus channel bitrate. (c) Temperature at the photo-diode tier versus channel bitrate.

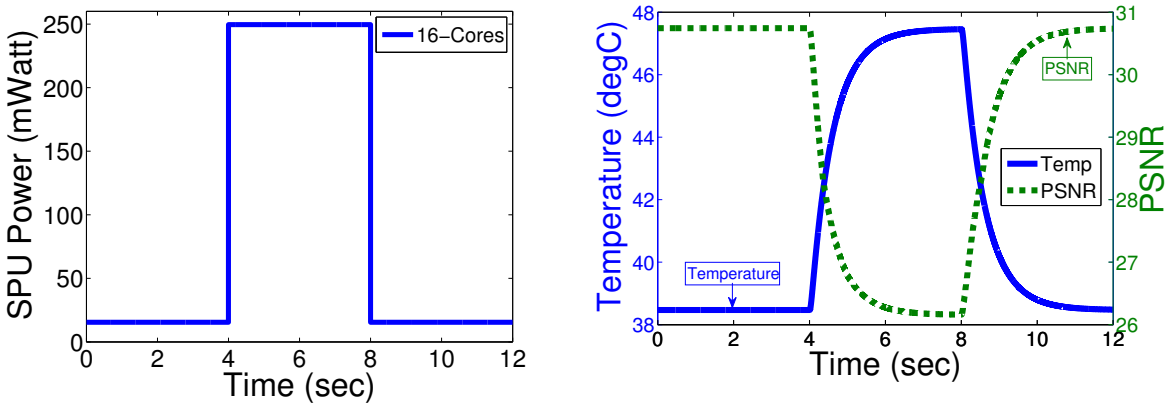


(a) (b) (c)

Figure 41: Effect of the wireless channel bitrate on the image quality due to spatial noise: (a) scene with dark lighting, (b) scene with normal lighting, and (c) scene with bright lighting. The image throughput is 24 images/sec.

40a). As the SPU consumes majority of the power, a higher SPU clock speed increases the power dissipation (Figure 40b). Due to increased power at lower channel bitrate, the die temperature increases at a reduced channel bitrate (Figure 40c). We observe that the 16-core and 8-core SPU shows similar power but the temperature of the 16-core system is lower than the 8-core system because of higher SPU die area. Figure 41 shows the relations between the image quality, and the varying channel bitrate for the different number of cores and lighting conditions. As expected, a reduced channel bitrate leads to higher temperature and hence, lower PSNR for all types of images. Also, as mentioned earlier, we observe the noise effect is most severe in the low light condition, and least severe in the bright light condition.

Let us now consider a dynamic variation in the channel bitrate. Consider a 16-core SPU and assume the channel bitrate drops from 54Mbps to 27 Mbps before it stabilize again at 54 Mbps. To maintain 24 images/second image throughput, the SPU has to increase its clock speed from 6 MHz to 36 MHz thereby resulting in a transient variation in the system power (Figure 42a). There is a corresponding time-dependent variation in the temperature



(a) (b)

Figure 42: Transient analysis of a 16-cores system compressing images at 24 images/sec throughput rate, and the channel bitrate switches from 54Mbps to 27Mbps and back to 54Mbps. (a) SPU versus time. (b) Temperature of the photodiode tier and the corresponding PSNR versus time.

of the photodiode tier and hence, the image quality (Figure 42b).

The above analysis shows that there is a correlation between the image quality, the channel data rate, the lighting conditions, core configurations, and the image throughput demand. The image quality varies with time due to the time-varying nature of the wireless channel bitrate (and hence, power dissipation) and the thermal capacity of the stack.

3.4.4 Temporal Noise

This section evaluates the effect of temporal noise to the image quality under varying lighting condition, image throughput, and channel bitrate. The analysis considers the starry night, Lena, and the airplane images as representatives for the dark, normal, and bright images respectively. First we look at the effect of the varying image throughput to the PSNR values considering multicore systems with varying clock speed up to 50 MHz, and under a very fast off chip bitrate. Then, we look at the effect of varying channel condition to the image quality considering a constant performance throughput target of 24 images/second.

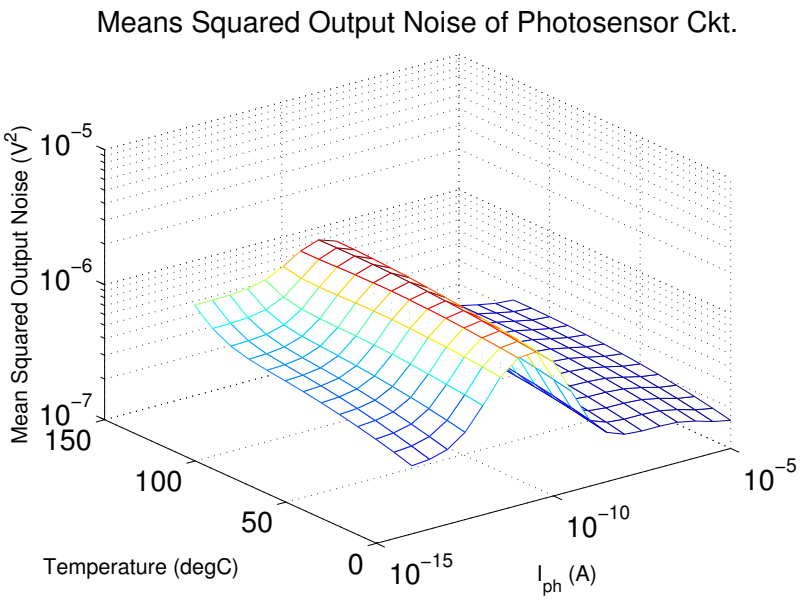


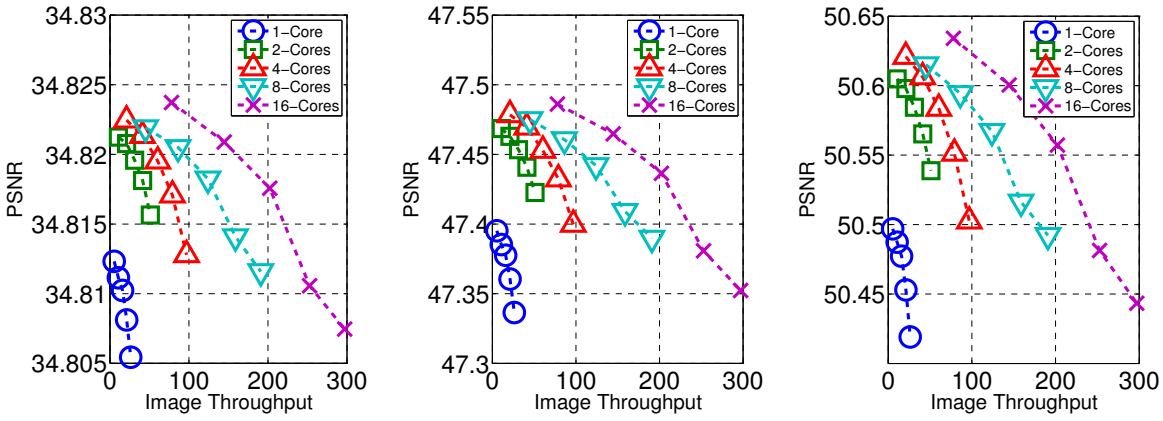
Figure 43: Temporal noise squared output voltage across temperature and illumination.

The case study for the temporal analysis in this section is similar to the corresponding analysis in the spatial noise section.

The squared output temporal noise voltage plot across variations in temperature and illumination is shown in Figure 43. The temporal noise level is significantly lower than the spatial noise level (Figure 36a). Figure 44 and Figure 45 shows the corresponding image degradation due to temporal noise for the multicore systems with varying image throughput and the multicore systems with varying channel bitrate respectively. These results display a trend that is consistent with the results from spatial noise analysis. However, the high PSNR values demonstrate that distortions due to temporal noise is less significant compared to spatial noise.

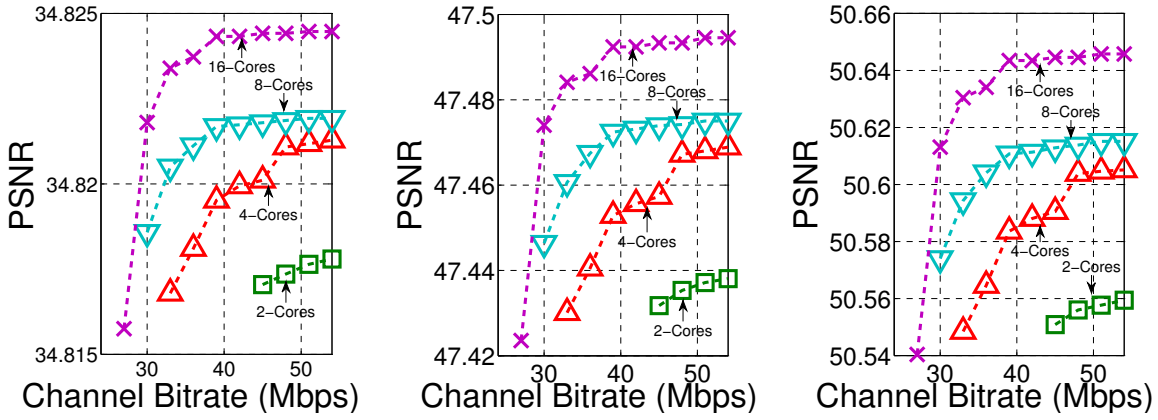
3.4.5 The Effect of Spatial and Temporal Noise to Image Quality

The effect of varying lighting condition, image throughput, and channel bitrate to the image quality due to both spatial and temporal noise is discussed in this section. The case studies for the analysis are the multicore systems with varying clock speed up to 50 MHz and a very fast off chip bitrate, and the multicore systems with varying channel condition with a



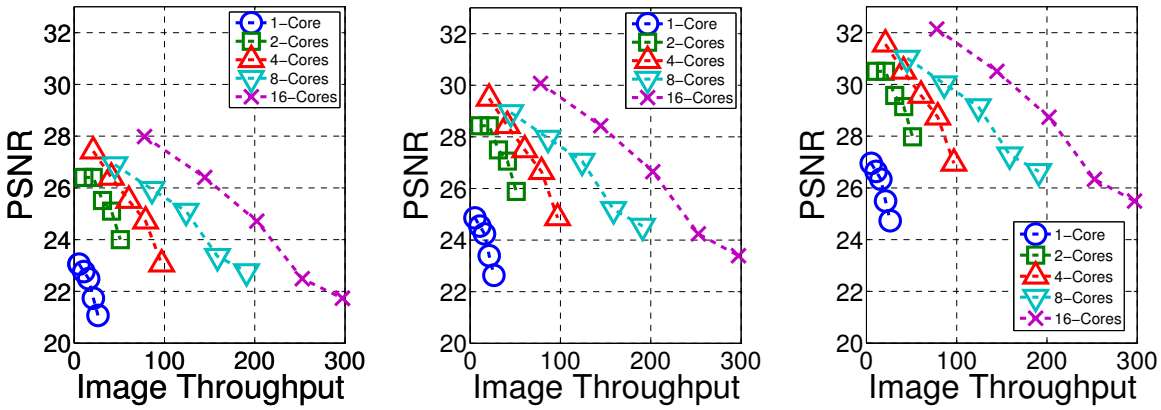
(a) (b) (c)

Figure 44: Effect of the lighting condition to the image quality due to temporal noise: (a) scene with dark lighting, (b) scene with normal lighting, and (c) scene with bright lighting. The throughput corresponds to the following SPU clock speed: 10, 20, 30, 40, and 50 MHz.



(a) (b) (c)

Figure 45: Effect of the wireless channel bitrate on the image quality due to temporal noise: (a) scene with dark lighting, (b) scene with normal lighting, and (c) scene with bright lighting. The image throughput is 24 images/sec.



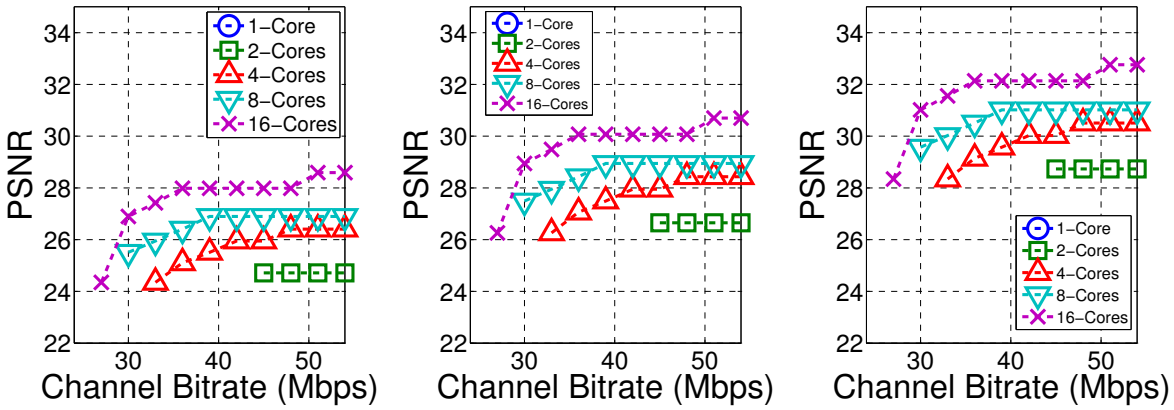
(a) (b) (c)

Figure 46: Effect of the lighting condition to the image quality due to spatial and temporal noise: (a) scene with dark lighting, (b) scene with normal lighting, and (c) scene with bright lighting. The throughput corresponds to the following SPU clock speed: 10, 20, 30, 40, and 50 MHz.

constant performance throughput target of 24 images/second similar to the case studies in the temporal analysis section. Figure 46 and Figure 47 show that the PSNR level of results from the total noise analysis is almost at the same level as the results from the spatial noise analysis shown in Figure 39 and Figure 41, which conclude that the effect of temporal noise to the image distortion is marginal compared to spatial noise.

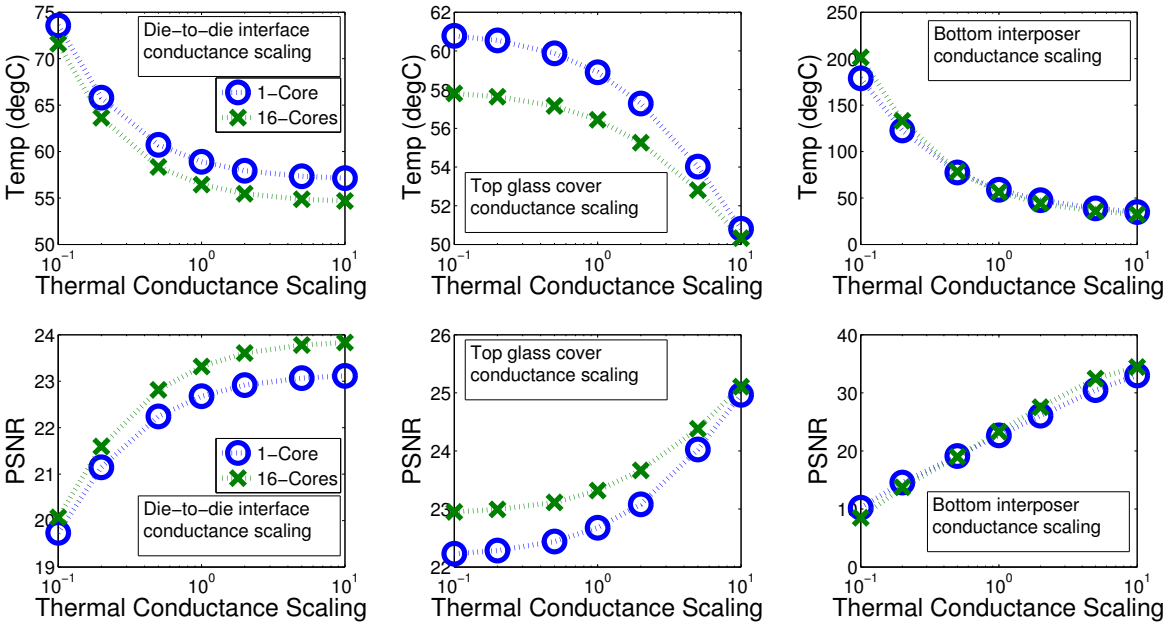
3.4.6 The Effect of Varying Thermal Conductance

The rest of the analysis considers only spatial noise, since previous sections have shown that the effect of temporal noise is marginal. In this section, we analyze how temperature of the photodiode tier changes with a change in the average conductivity between the tiers (i.e. thermal conductivity of the die-to-die interface in the 3D stack which includes the BEOL and the bonding layer in Figure 32). We have also considered variations in the thermal conductance of the top cover (glass cover), and the bottom cover (interposer). The simulations are performed considering the 1-core and the 16-cores cases running at 50



(a) (b) (c)

Figure 47: Effect of the wireless channel bitrate on the image quality due to spatial and temporal noise: (a) scene with dark lighting, (b) scene with normal lighting, and (c) scene with bright lighting. The image throughput is 24 images/sec.



(a) (b) (c)

Figure 48: Effect of varying thermal conductance in the (a) die-to-die interface, (b) top glass cover, and (c) bottom interposer to the thermal coupling and image quality.

MHz clock speed, and assuming 250C ambient. We use the gray-scale 512X512 pixels Lena image for our experiment. The thermal conductances are scaled with respect to their values in Table I. A decrease in thermal conductance of the die-to-die interface results in higher overall temperature and hence, degrades image quality (Figure 48a). As the thermal conductance of the glass cover is increased, more heat is allowed to flow and escape through the top cover. Consequently, the temperature of the photodiode tier is reduced, and the image quality improved (Figure 48b). Reducing the thermal conductance of the interposer prevents the heat from escaping through the bottom layer. As a result, the temperature across all tiers increases and the image quality degrades (Figure 48c).

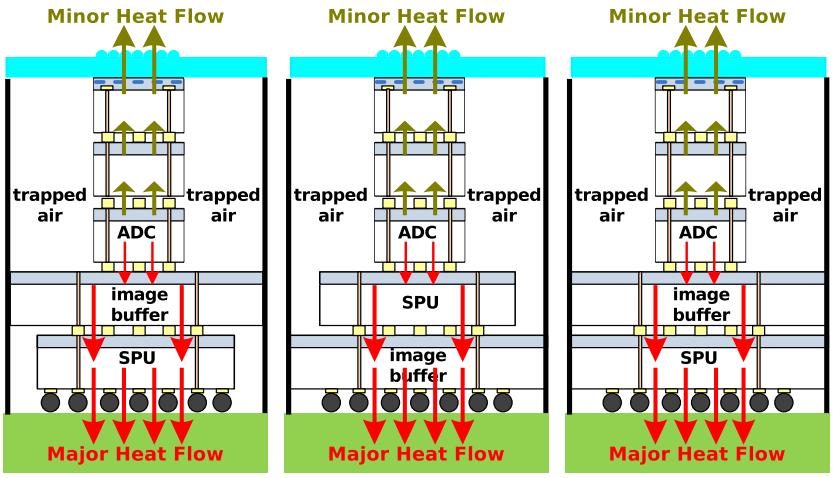
3.4.7 Alternate 3D-Stacking Scenarios

The different sizes of the SPU and memory die, raises the multiple possibilities of stack organizations. We perform analysis on three possible scenarios to vertically stack the image sensor system, as shown by Figure 49. We consider 24 images/sec compression throughput, 2-cores and 16-cores system configurations and a channel bitrate of 54Mbps. Figure 50 shows the temperature of the photodiode tier and the image quality for various stack scenarios.

Case 1 The SPU is placed at the bottom tier, below the image buffer tier. This organization has been used for the prior results in this paper. This results in the maximum temperature and the minimum image quality.

Case2 The image buffer with the largest die area is placed at the bottom tier. TSV connections between the ADC and the image buffer have to pass through the SPU tier. The surface power density of the SPU tier and the volumetric power density of the stack remains the same, but placing the largest die at the bottom improves the heat spreading. Hence, the heat outflow improves compared to case 1, thereby reducing photodiode temperature and improving image quality.

Case3 The SPU is placed at the bottom tier (as in case 1), and the area of the SPU tier is expanded to match the area of the image buffer tier. The SPU core circuit is spread out



(a) (b) (c)

Figure 49: Alternative stacking scenarios for the image sensing system. (a) Case 1: SPU at the bottom of the stack. (b) Case 2: image buffer at the bottom of the stack. (c) Case 3: the SPU die area is widened to match the image buffer die area.

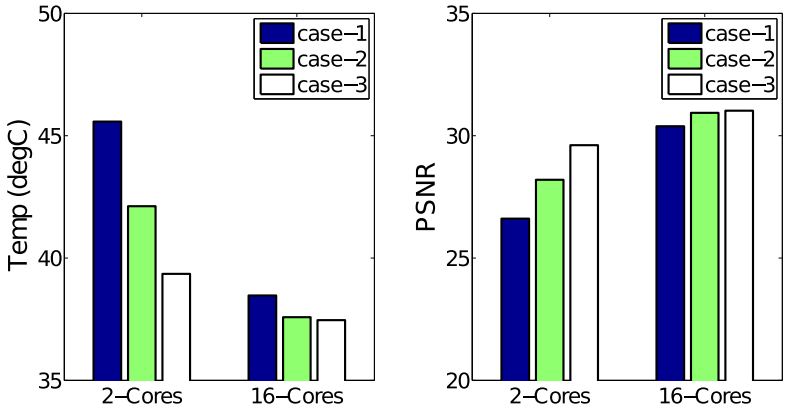


Figure 50: Effect of the alternative stacking scenarios on the temperature of the photodiode tier and the image quality.

across the entire die area. This reduces the overall power density of the SPU tier. Moreover, increasing area of the SPU tier provides more spreading opportunity for the heat generated from the ADC as well. Therefore, both the surface power density of the SPU as well as the volumetric power density of the 3D stack reduces, resulting in lowest temperature for the photodiode tier among all the cases and hence, better image quality. However, this configuration uses much more silicon than required, and hence, incur higher cost.

3.5 Summary

We have analyzed the relations between the power/performance of image compression unit and the noise characteristics of logarithmic image sensor when integrated in a 3D stack. A coupled power, thermal, and noise (fixed-pattern-noise, temporal noise, and CIS output voltage range) analysis framework is developed to correlate the architectural design choices, environmental conditions, power dissipation, and image quality in a 3D image sensors. The above analysis shows that due to the die-to-die thermal coupling, 3D integration of the image compression unit with the photo-diode array introduces new challenges to the noise management and the image quality control in an image sensor. The spatial (fixed-pattern) noise due to process variability, and the temporal noise elements (i.e. thermal, shot, flicker noise) have a weak dependence on the power dissipation in the compression unit, but the output voltage range of the sensor is significantly affected by the compression power. The time-varying nature of the power dissipation of the compression unit due to factors like different image throughput demands, lighting conditions, or wireless channel bandwidth, results in a dynamic variation in the quality of the image captured by the CMOS image sensor. As the 3D integration has a strong potential to significantly improve energy-efficiency of the image sensor, controlling the effect of die-to-die thermal coupling on the characteristics of the image sensor is crucial.

CHAPTER 4

ANALYSIS OF HETEROGENEOUS INTEGRATION FOR THE 3D STACKED IMAGE SENSOR SYSTEM

4.1 Introduction

Three Dimensional Integrated Circuit provides a vehicle for heterogeneous integration, i.e. circuit tiers are built with different process technology nodes. In the previous chapter, the 3D IC has been presented as a solution to improve power efficiency for a complete image sensor/compression system. However, the design has only considered synthesizing with a homogeneous technology node for both photo sensor and compression modules. Although 180nm process is highly optimized for designing photodiodes and pixel circuitries, its performance as a digital system lags behind the more advanced deep sub-micron processes. Unfortunately, access to the newest process technology does not improve CMOS image sensors. For example, a diode in standard CMOS process has low photo-sensitivity due to material characteristic that is opaque to the visible light spectrum [74]. Another potential disadvantage of technology scaling is the reduction in V_{DD}, which reduces the voltage swing of the analog signal representation in the CIS. With 3D integration technology the best of the two houses can be joined together. Since each component of the system (i.e. photo sensors, column circuits, ADCs, image buffer, and SPU) is placed on a separate layer, they can be designed on different wafers using different processes.

This chapter aims to assess the potential benefits that can be obtained through heterogeneous integration. The photosensor module is simulated with 180nm process, while the image compression module is scaled down to 90nm and 45nm processes. The structure of the 3D stack is shown in Figure 49a, in which the SPU is placed at the bottom tier below the image buffer. The SPU is synthesized using commercial tools. The performance of the system is evaluated considering multiple factors such as die area, compression throughput,

power, thermal coupling, and sensor noise. The effect of thermal coupling to the photosensor noise is modeled as discussed in Chapter 3. This noise introduces distortion to the compressed image, and the image quality is measured using peak-signal-to-noise-ratio (PSNR) scale. Heterogeneous integration and multi-clock domain scheme is considered to further reduce the power consumption of the system. Although numerous work have qualitatively described the benefits of heterogeneous integration, no work has presented a quantitative analysis on the issues. Technology scaling provides reduction in the digital system power and die area. However the resulting increment in power density increases temperature at the photosensor, thus reducing the image quality. At the end of the analysis, a design based on low power process technology is presented for low performance application.

4.2 Motivation for Technology Scaling

Traditionally, technology scaling in digital circuit is aimed to achieve the following goals:

1. Increase transistor density.
2. Improve performance.
3. Reduces power dissipation.

In theory, scaling reduces the channel dimension by approximately 30% at each new generation, which translates to 50% reduction in total area, 30% reduction in capacitance, and reduction in gate delay. In addition, scaling reduces the supply voltage (V_{DD}) which reduces the switching power of the digital circuits. It is an attractive technique to significantly boost system performance.

However, as the channel length is reduced below 100nm, leakage power becomes an important source of power dissipator in the system. Figure 51 shows the percentage contributions of the leakage power of a single core SPU in 90nm and 45nm process [75]. When the SPU is in a low activity state, leakage power becomes the dominant contributor. These

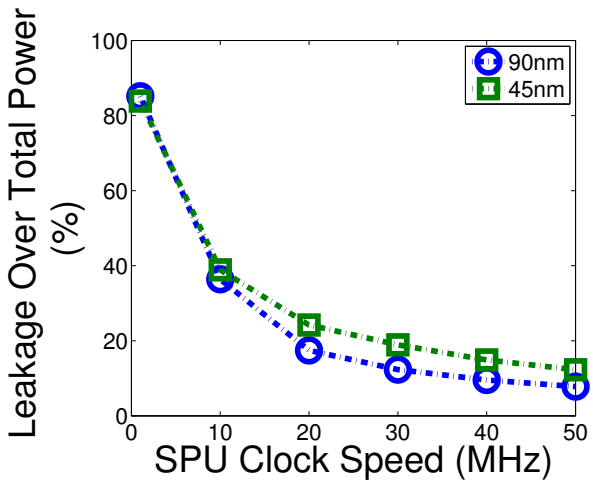


Figure 51: Percentage contributions of leakage power of the SPU in 90nm and 45nm process.

percentage values are significantly higher than the leakage power in 180nm, which is constantly less than 1% of the total power. On the other hand, the power benefits from scaling down bulk CMOS is rapidly saturating. The channel length is comparable to the depletion width of the drain-body junction, giving rise to short channel effect. This results in an increased leakage current, Vth roll-off, and Vth reduction due to drain induced barrier lowering (DIBL) [76].

To address the leakage current problem, researchers and industries have explored non planar transistor structures called FinFET. In FinFET, the gate is wrapped around its channel to give better electrostatic performance and better leakage control [77, 78, 79]. It is currently the state-of-the-art FET structure, which provides superior characteristics compared to planar CMOS device.

4.3 Experiment Setup

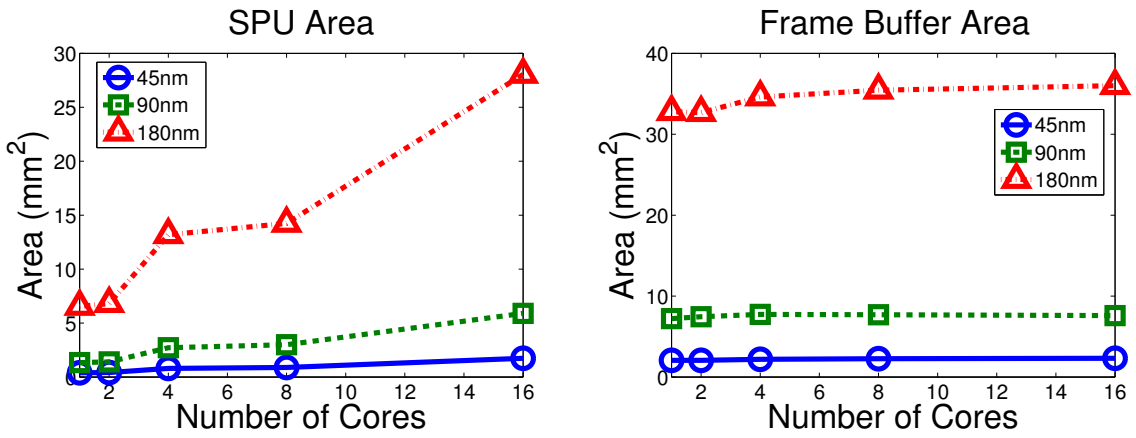
For this analysis, we consider the multisegment image compression architecture with 1-,2-,4-,8-,and 16-cores configurations. The SPU, image buffer and network are connected to a single supply voltage source. The DWT with serial Huffman compression algorithm is coded in Verilog hardware descriptive language (HDL) as an RTL to provide a basis for the

Table 4: List of Logic Gates for the Libraries

Logic gates	180nm	90nm	45nm	22nm FinFET	22nm TFET
AOI21X1	✓	✓	✓	X	X
DFFX1	X	X	X	✓	✓
DFFSRX1	✓	✓	✓	X	X
INVX1	✓	✓	✓	✓	✓
INVX2	✓	✓	✓	✓	✓
INVX4	✓	✓	✓	✓	✓
INVX8	✓	✓	✓	✓	✓
NAND2X1	✓	✓	✓	✓	✓
NAND3X1	✓	✓	✓	X	X
NOR2X1	✓	✓	✓	✓	✓
NOR3X1	✓	✓	✓	X	X
OAI21X1	✓	✓	✓	X	X
XOR2X1	✓	✓	✓	X	X

SPU core. Digital logic libraries for various technology nodes are utilized to estimate power and performance and simulate technology scaling effect of the SPU. The image buffer is simulated as an SRAM design, in which its performance and power data is estimated using CACTI [57]. The network power and performance data is extracted from ORION [56]. The network power values are calibrated based on findings presented in [80], and its trend is crosschecked with DSENT [81].

A digital logic library consists of power-delay table and physical data of a list of logic gates that can be used for RTL synthesis and automated place and route. The RTL of the SPU, which is written in behavioral hardware descriptive language (HDL), is synthesized



(a) (b)

Figure 52: Effect of technology scaling to die area of the 3D-stack.

into a structural HDL netlist using Synopsys Design Compiler tools. Next, switching activities of the SPU is simulated using Mentor Modelsim functional simulation tools by taking one of the test images as input vectors. Lastly, the power-delay table is used to perform a static timing analysis and power estimation of the design using Synopsys Primetime. For this chapter, we use 180nm, 90nm, 45nm, and 22nm logic libraries. Device models for the 90nm, 45nm, and 22nm are taken from predictive technology model developed by Arizona State University. Physical and area data is provided by TSMC180, Synopsys-90, and NCSU FreePDK45. A minimum set of logic gates are used to maintain simplicity and consistency across process technologies. The libraries are recharacterized across different V_{DD} levels to obtain power-delay tables for V_{DD} sweep analysis. Table 4 shows the list of gates used for our logic libraries. A device model that is suitable for low performance operation based on Tunnel FET (TFET) is developed in house and discussed in the subsequent section. No physical and area data is available for the FinFET and TFET logic library. These libraries maintain a smaller list than the bulk CMOS.

Table 5: Effect of Heterogeneous Integration to the Die Area of the 3D-Stack

Number of Cores	1	2	4	8	16
Photosensor area (mm^2)					
180nm	6.56	6.56	6.56	6.56	6.56
Image buffer area (mm^2)					
180nm	32.79	32.67	34.61	35.44	36.01
90nm	7.22	7.45	7.74	7.70	7.59
45nm	2.06	2.06	2.19	2.27	2.32
SPU area (mm^2)					
180nm	6.55	6.81	13.19	14.24	28.05
90nm	1.34	1.41	2.72	3.00	5.90
45nm	0.40	0.42	0.80	0.89	1.74

4.4 Effect of Heterogeneous Integration to Die Area of the 3D-Stack

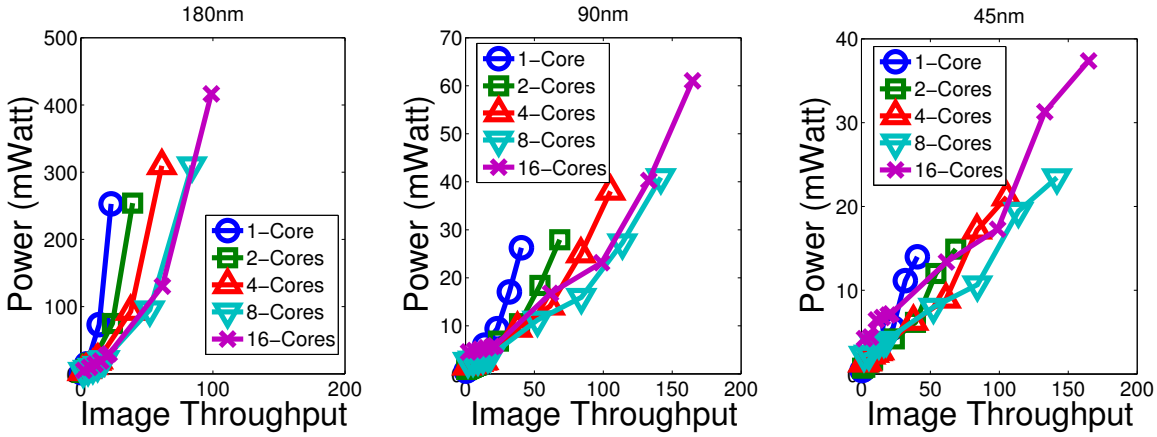
The impact of technology scaling to the area of the SPU and the image buffer is shown in Figure 52a and Figure 52b respectively. The area reduces by a factor of 4 as the device is scaled from 180nm to 90nm, and from 90nm to 45nm. This trend follows the traditional area reduction of 50% per generation. The SPU area increases non-linearly with increasing number of cores, meanwhile the frame area is relatively constant because the image size does not change. Table 5 shows the tier-by-tier area comparison between the image sensor and the image processing modules in the 3D stack.

4.5 Power

4.5.1 Effect of Technology Scaling to Power

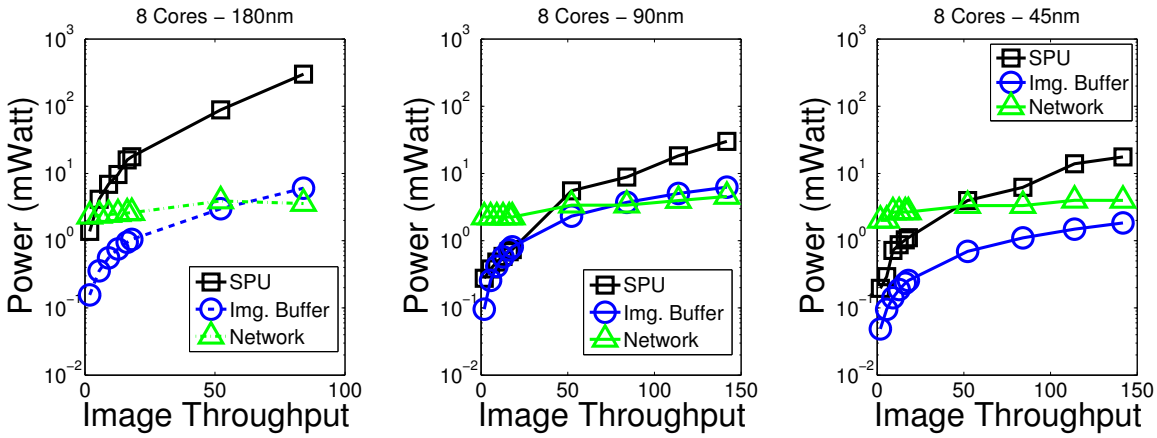
In the following investigation, the image processing module (SPU, image buffer, and network) is simulated using 180nm, 90nm, and 45nm process libraries to model and understand the change in power dissipation associated with technology scaling. First, we examine the system's power usage and pure processing capability with unrestricted data output bitrate. The SPU is set to have 1-, 2-, 4-, 8-, and 16-cores configurations with a single clock domain to control SPU, image buffer, and network clock signal. As the system clock is increased, compression throughput and power dissipation go up. Figure 53 shows the power curve of the multicore system as a function of image throughput, simulated using different process libraries. By scaling down the feature size, the same amount of throughput rate is achieved with reduced power dissipation. However, the network and image buffer power is increasingly dominant as feature size scales down. Although dynamic power significantly reduces, leakage power reduction has been marginal as compared to dynamic power. In addition, the leakage power of the network goes up with increasing parallelism, especially at low throughput region. Figure 54 shows power curves of the SPU, image buffer, and network in a 8-cores configuration. At 90nm and 45nm, leakage power accounts up to 50% and 70% of the network power respectively. At low throughput region where SPU activity is low, the network is the dominant power dissipator. This leakage power increases with increasing number of cores, which reduces power efficiency of extensive parallelism in low throughput operation. Dynamic power of the SPU and image buffer increases with image throughput because of the increased activity as the system compresses more images per second. But dynamic power of the network remains relatively constant because increasing the clock speed of the network reduces the dynamic energy of the router.

Next, we consider a multi-clock domain implementation of the system. In this scheme, the SPU is operated in a reduced clock speed, while the network and image buffer is operated in an increased clock speed, such that each memory access is completed virtually in



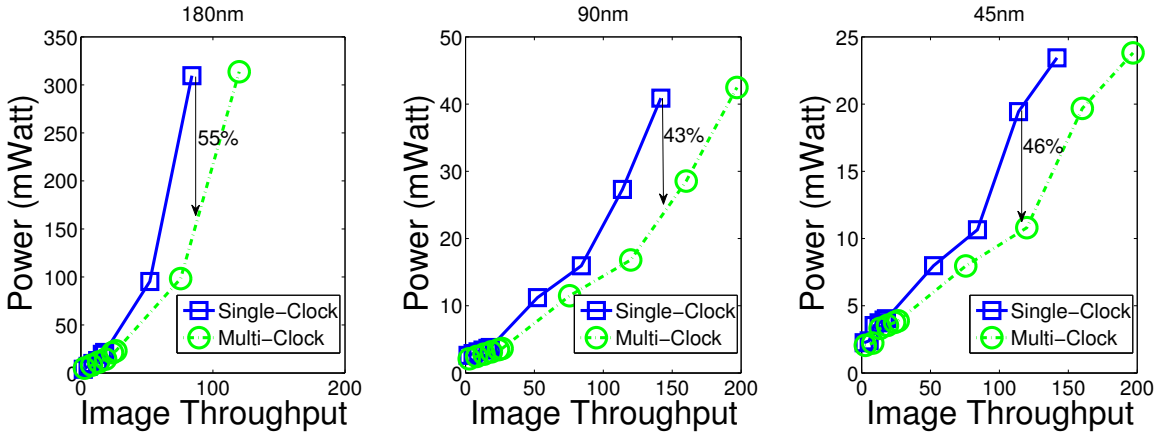
(a) (b) (c)

Figure 53: Power curves of the multicore system as a function of image throughput simulated in (a) 180nm, (b) 90nm, and (c) 45nm process libraries.



(a) (b) (c)

Figure 54: Power curves of the 16-cores SPU, image buffer, and network as a function of image throughput simulated in (a) 180nm, (b) 90nm, and (c) 45nm process libraries.



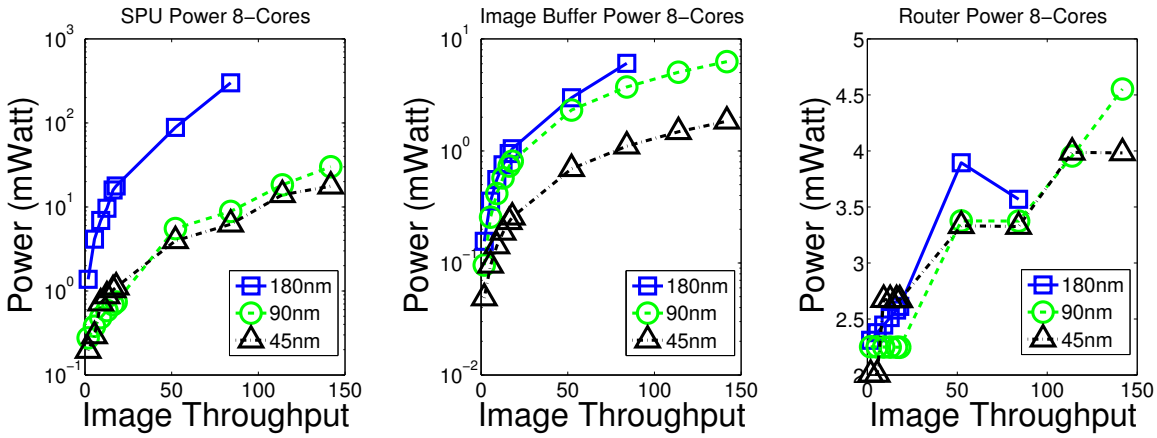
(a) (b) (c)

Figure 55: Power comparison between the multiclock-domain and singleclock-domain schemes simulated in (a) 180nm, (b) 90nm, and (c) 45nm process libraries.

one SPU clock period. A discussion related to this scheme is presented in Chapter 2. For this experiment, we take an 8-cores system, sweep the SPU clock-speed from 1 MHz to 100 MHz, and estimate the image throughput. The SPU, image buffer, and network is supplied by one V_{DD} source. Figure 55 shows the potential power reduction of the multi-clock domain scheme. At low throughput region, power reduction is relatively small because the SPU has low activity. One characteristic of the multi-clock domain is the SPU clock speed is reduced with the expense of increasing image buffer and network clock speed. At high throughput region, the SPU is the major power contributor, therefore reducing the workload of the SPU helps reduce power. The average power saving values of the multi-clock domain scheme in 180nm, 90nm, and 45nm are approximately 27%, 26%, and 23% respectively. It means that the reduction of the SPU dynamic power is greater than the increase in the image buffer and network dynamic power.

4.5.2 Second Degree Heterogeneous Integration of the Compression Module

In this section, we investigate the possibility of synthesizing the SPU and image buffer/network in different process technology in order to achieve minimum power. In deep sub-micron



(a) (b) (c)

Figure 56: Power comparisons of the (a) SPU, (b) image buffer, and (c) network router across different technology nodes.

process, leakage power dominates and may overshadow dynamic power saving from technology scaling. In addition, dynamic/leakage power balance of all digital components in the compression module shifts with changes in the activity level of the system. To minimize power consumption across various activity level, we consider synthesizing each component (i.e. SPU, network router, and image buffer) using different process library. We take an 8-cores system, scale down all the components (SPU, network, or image buffer), and compare each component separately across different technology node. Figure 56 shows the results of this comparisons. The SPU power in the 180nm is significantly higher than in 90nm and 45nm. This is because dynamic power in 180nm process is much higher than in 90nm and 45nm, and it contributes more than 70% of the SPU power. The image buffer power reduces with decreasing feature size. The router power, which is dominated by leakage power, reduces as it is scaled from 180nm to 90nm. When the router is scaled from 90nm to 45nm, power reduction only happens at regions where V_{DD} of the 45nm system is lower than the 90nm system.

In this experiment we introduce a 2nd degree of heterogeneity by synthesizing the SPU and image buffer/network in 45nm and 90nm respectively and compare them against the

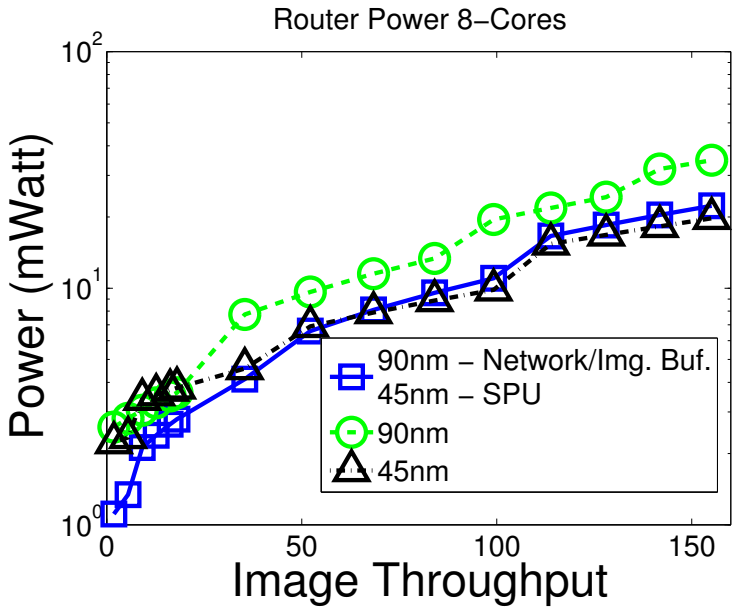


Figure 57: Power comparison between 1st degree heterogeneous and 2nd degree heterogeneous integration of the compression module. In 1st degree heterogeneous system, the compression module is implemented in one process node (90nm or 45nm). In 2nd degree heterogeneous system, the compression module is implemented using 90nm and 45nm for the image buffer/network and SPU respectively.

1st degree heterogeneous system. The network router, which uses a significant amount of power in deep sub-micron, consumes more power at 45nm than at 90nm. The network are placed at the same tier as the image buffer. Figure 57 shows that a 2nd degree heterogeneity within the compression module yields approximately 9% higher power consumption than the 45nm compression module in the region with 53 images/sec throughput or higher, but has an average of 29% lower power in the region with lower than 53 images/sec throughput. Note that the network routers, when operated at certain supply voltage and clock speed (e.g. V_{DD} at 0.4V and clock speed at 7 MHz) yields lower leakage power in 90nm than in 45nm. Although network router power in 90nm is lower than 45nm, the image buffer power in 90nm is higher than in 45nm. In the region between 9 images/sec and 20 images/sec, the 90nm compression module has 7% lower power than the 45nm compression module. At this region, V_{DD} of both 45nm and 90nm systems is at 0.4V. At region with lower than 9 images/sec throughput, V_{DD} of the 45nm compression module reduces to 0.3V while V_{DD} of the 90nm compression module saturates at 0.4V, thus leakage power of the 45nm compression module becomes lower than the leakage power of the 90nm compression module.

4.6 Power Efficiency for a Wireless Image Sensor Node

In this section we consider a low power wireless image sensor design (Figure 58). This sensor node is part of a larger Wireless Sensor Network system that is deployed to monitor a remote area. Image data that is captured is compressed and transmitted to a base station. The algorithm used for compression consists of DWT and Huffman coding. The multi segment compression with serial Huffman scheme is considered as the architecture of the compression module. The power dissipation of the ADC is taken into account in estimating the overall power of the system. This section presents two different optimization problems:

1. Given a fixed power budget, we maximize throughput of the image sensor
2. Given a fixed target throughput, we minimize power consumption

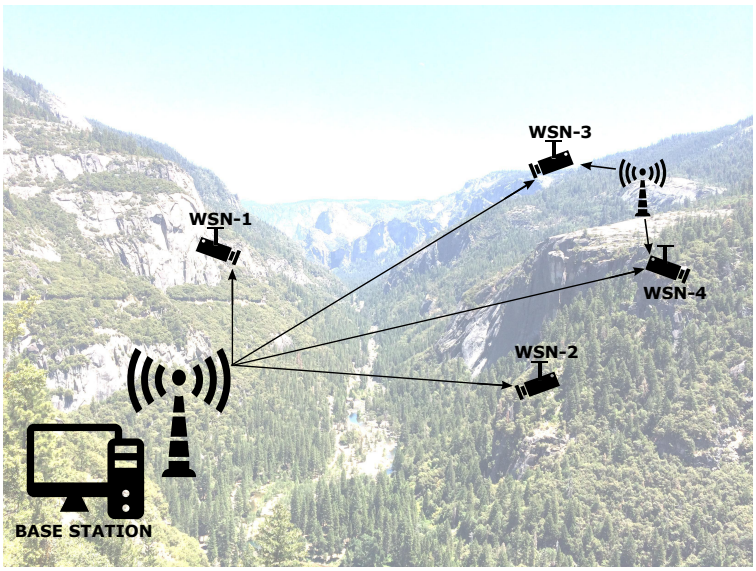


Figure 58: Wireless Image Sensor Network.

4.6.1 Maximizing Throughput for a Given Power Budget

The analysis in this section assumes a Wireless Sensor Network with a wireless channel bitrate of 54 Mbps. The system consists of a 5 tiers stack as shown in Figure 49a. The ADC tier hosts 256 ADC units to support high performance imaging. The ADCs cluster consumes 230.4 mWatt of power. A discussion related to the ADC power is presented in Chapter 3. The image size is 512X512. With a homogenous 180nm design synthesis, achieving 24 images/sec throughput requires 260 mWatt of total power consumption, in which. In the above set up, the compression module is configured with 8-cores SPU, which gives the least power consumption compared to 1-, 2-, 4-, and 16-cores configuration. Next, given a fixed power budget of 260 mWatt, we change the integration schemes of the compression module to boost the performance of the system. Several integration schemes have been discussed in this dissertation such as 3D-stacking, multiclock domain scheme, technology scaling, and heterogeneity within compression module. Figure 59a presents the improvement in the performance of the system as we change the integration scheme of the compression module. The different integration schemes are listed as follow:

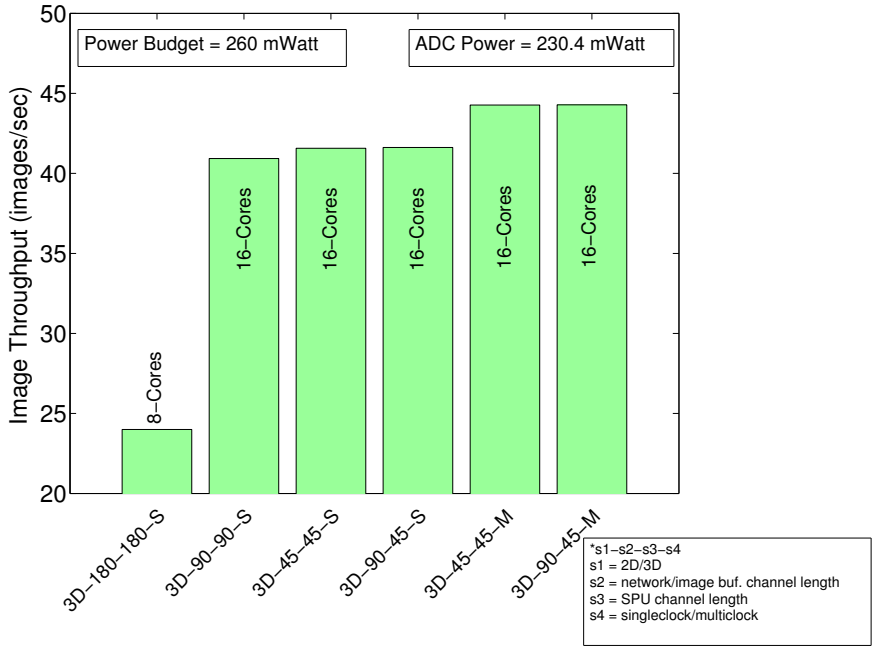
1. 3D-integration of the compression module in 180nm process with single clock scheme,

2. 3D-integration of the compression module in 90nm process with single clock scheme,
3. 3D-integration of the compression module in 45nm process with single clock scheme,
4. 3D-integration of the compression module with 2nd degree heterogenous 45nm SPU, 90nm network/image buffer, and single clock scheme,
5. 3D-integration of the compression module in 45nm with multi clock domain scheme, and
6. 3D-integration of the 2nd degree heterogeneous 45nm SPU and 90nm network/image buffer with multi-clock domain scheme.

Scaling down the compression module from 180nm to 90nm increases image throughput by 70%. As channel length is scaled down, dynamic energy of the compression system significantly reduces, which allows increased throughput for a given power budget. Scaling down the compression module from 90nm to 45nm increases image throughput by only 1.6%. The increased in throughput from technology scaling saturates due to limitation in the channel bandwidth. It takes between 5 to 39 ms for the transmitter to send an image under a 54 Mbps channel bitrate. Next, we introduce heterogeneity in the compression module by synthesizing the SPU in 45nm technology, and the network/image buffer in 90nm technology. The boost in image throughput from 45nm compression module to 2nd degree heterogeneous 90/45nm compression module is less than 1%. Although, leakage power reduces, improvement in energy efficiency is limited due to the increase in dynamic energy as network/image buffer is scaled up to 90nm. Implementing a multi clock scheme to the 45nm compression module increases the image throughput by 6.5%. Multi clock scheme yields to reduction in the dynamic energy of the SPU, which allows extra power budget to boost the clock speed of the system, thus increasing throughput. Moving from heterogeneous 90/45nm compression module with single clock scheme to heterogeneous 90/45nm compression module with multi clock scheme increases throughput by less than

1%. This is because the increase in the compression energy of the network/image buffer is almost equal to the reduction of the compression energy of the SPU.

Next, we analyze the distortion introduced in the image due to thermal coupling. The analysis considers the wireless image sensor node application discussed in the previous section. The SPU is placed at the bottom of the stack, below the image buffer tier. Technology scaling and the various power reduction scheme changes the die area as well as the power dissipation of the system, thus altering the power density of the whole system. Temperature at the photosensor tier is highly coupled with power dissipated by the ADC, SPU, image buffer, network, and other components in the system such as wireless transmitter. This changes in temperature, coupled with process variations in the photosensor die, lead to variations in the fixed pattern noise behavior and output range of the pixel circuitry. In addition temporal noise also increases with temperature, although our previous analysis shows that the temporal noise level is more than 10X lower than the fixed pattern noise. The test image considered in this analysis is the Lena image. Figure 59b presents image quality of wireless image sensor node configured with the above schemes. With the given 260 mWatt power budget, the system with 180nm compression module has the lowest distortion (highest PSNR value), and the system with 45nm compression module has the highest distortion (lowest PSNR value). The 180nm compression module has the highest die area, providing extra room for heat to spread and dissipate. Note that the image quality (PSNR value) for the single-clock 45nm module is similar to the multi-clock 45nm module, because power density of the two are the same. The 2nd degree heterogeneous 90nm network/image buffer with 45nm SPU has higher power density than the 90nm compression module but lower power density than the 45nm compression module, which results in a lower PSNR value than the 90nm compression module but higher than the 45nm compression module.



(a)

(b)

Figure 59: Performance and image quality comparison of the wireless image sensor node synthesized in various integration schemes. (a) Change in image throughput. (b) Change in image quality.

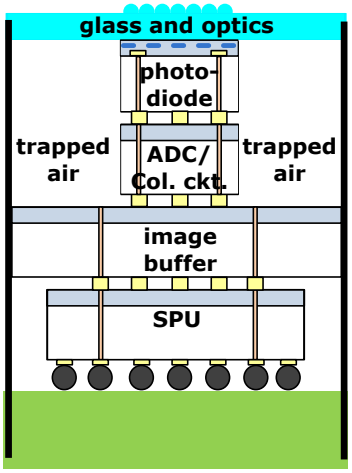


Figure 60: Four stacks structure of the 3D integrated image sensor.

4.6.2 Minimizing Power for a Given Target Throughput

The analysis in this section aims to investigate a power efficient design of the wireless image sensor node. We assume the desired compression throughput is 24 images/second and the wireless channel bitrate is 54 Mbps. The number of ADCs used in this analysis section is reduced to 16 to reduce ADC power consumption. As a result, the die area consumed by the ADC reduces significantly, thus the column circuitry and the ADCs can be placed together in a single tier, as illustrated by Figure 60. The power consumption of the ADC is 14.4 mWatt. The core configuration (i.e. number of cores, clock speed, and integration scheme) is optimized to achieve its minimum power. We compare power dissipation of the sensor node across the different integration schemes that are introduced in the previous section.

The above problem is an optimization in which the solution is provided through design space exploration. The problem is defined as follow:

optimize: n_{core}, f_{spu}, and the integration scheme (S)

to minimize: $Power$

under the constraints: $IPS \geq IPS_0$ and $f_{bitrate} = Fbit_0$

where: $IPS_0 = 24\ images/sec$ and $Fbit_0 = 54\ Mbps$

Therefore, the design space is a three-dimensional table that can be described as

$$U_{IR} \equiv \{(n_{core}, f_{spu}, S) : IPS \geq IPS_0, f_{bitrate} = Fbit_0\} \tag{17}$$

and power dissipation of the system is a function of the core configurations

$$P = f(n_{core}, f_{spu}, S) \tag{18}$$

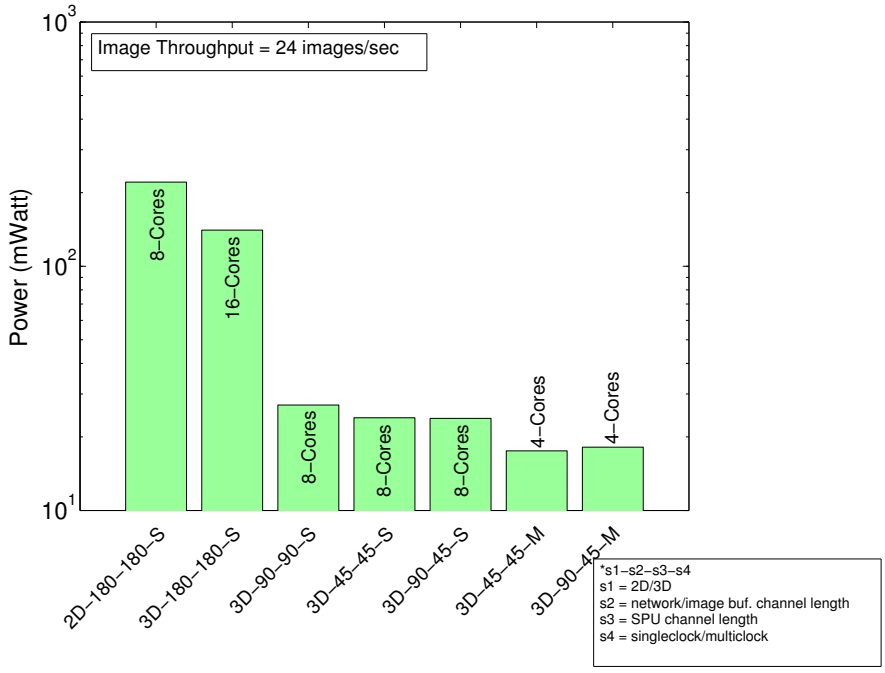
in which we select f_{spu} and n_{core} that returns the minimum power solution

$$Power = min_{n_{core}, f_{spu}, S \in U_{IR}}(P(n_{core}, f_{spu}, S)) \tag{19}$$

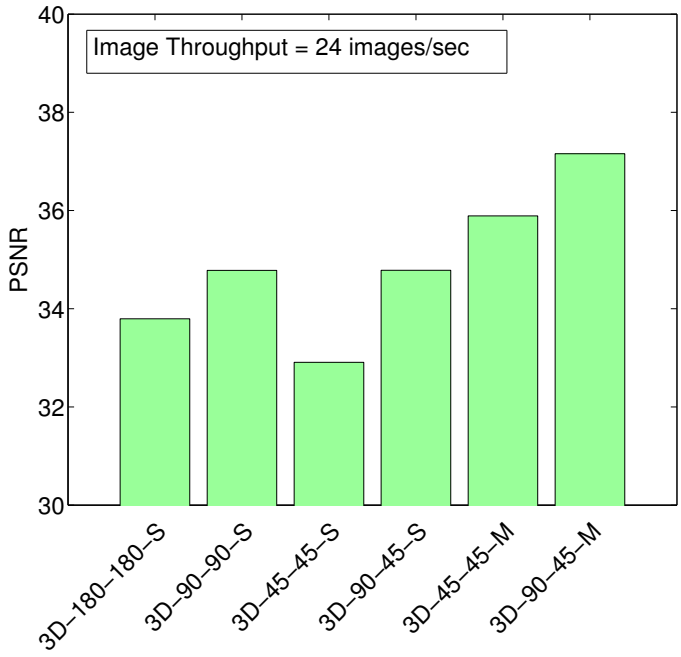
Figure 61a presents the power comparison of the system in various integration schemes. The different integration schemes (S) are listed as follow:

1. 2D-integration of SPU and image buffer in 180nm process with single clock scheme,
2. 3D-integration of the compression module in 180nm process with single clock scheme,
3. 3D-integration of the compression module in 90nm process with single clock scheme,
4. 3D-integration of the compression module in 45nm process with single clock scheme,
5. 3D-integration of the compression module with 2nd degree heterogenous 45nm SPU, 90nm network/image buffer, and single clock scheme,
6. 3D-integration of the compression module in 45nm with multi clock domain scheme, and
7. 3D-integration of the 2nd degree heterogeneous 45nm SPU and 90nm network/image buffer with multi-clock domain scheme.

Moving from 2D-stack to 3D-stack reduces power by 36%. In the 2D-stack, the SPU is placed side-by-side with the image buffer, thus limiting the number of interconnects between them. We assume there is only one memory port that connects the SPU to the



(a)



(b)

Figure 61: Power and image quality comparison of the wireless image sensor node synthesized in various integration schemes. (a) Change in power. (b) Change in image quality.

image buffer in the 2D-integration scheme. This memory port creates a bottle neck for the image data flow, significantly increasing lateral traffic movements in the network, which raises power dissipation. Scaling down the 3D-stack from 180nm to 90nm reduces the power by 81%. At 90nm, 8-cores configuration is more power efficient than 16-cores configuration by 8%. Technology scaling is shown to rapidly reduces dynamic power. However, it also increases static power, thus making parallelism increasingly expensive. As the technology is scaled down, low parallelism is more efficient than high parallelism. This trend is also observed at 45nm, in which 8-cores configuration has 11% lower power than 16-cores configuration. Next, the heterogeneous compression module is presented by synthesizing the SPU with 45nm process while the image buffer and network with 90nm process. The image buffer and the network has a high leakage power component, and scaling them up to 90nm reduces this leakage power, and ultimately it reduces total power consumption of the system. Next, the multi-clock domain scheme is applied to the 45nm compression module to reduce the clock speed of the SPU. The SPU is consuming majority of the power, thus going from the single clock 45nm compression module to multi clock 45nm compression module results in 27% reduction in power. Finally, applying multi-clock domain scheme to the 2nd degree heterogeneous 90/45nm compression module is not as beneficial as in the homogenous 45nm compression module. This is because with multi-clock domain scheme, the image buffer and the network is working in high clock speed, therefore its dynamic power is also high.

Figure 61b shows that the system in 45nm has the lowest image quality. This is because, although power dissipation of the compression module is lower in 45nm than in 180nm or 90nm, the die area of this system is also smaller than in the 180nm and 90nm processes. In addition, it has higher power dissipation than the 90/45nm compression module and the multi-clock 45nm compression module. The system with multi-clock 90/45nm compression module has the highest PSNR value. This is because multi-clock significantly reduces power consumption of the SPU, thus reducing the heat generated by the SPU.

4.7 Optimizing for Low Power Low Throughput Operation

In this section, low performance design of the wireless image sensor node is considered. Low power low performance application generally has a strict power requirement. In addition, it normally has limited capacity to store data. For example, image throughput for traffic cameras ranges between 1 to 5 images/second [82]. Monitoring devices deployed in non-busy area have lower than 1 images/second throughput. This section investigates synthesizing low throughput image compression system with low power device library to further reduce power consumption of the system.

4.7.1 Electrical Characteristics of the TFET Device Model

Tunnel FET (TFET) is a type of field-effect transistor aimed towards low energy electronics. In the TFET device, current conduction is invoked by quantum tunneling carriers from source to drain, rather than by thermionic injection typically found in conventional MOSFET device. TFET transistor is a gated p-i-n junction structure. A typical n-TFET device, shown in Figure 62, comprises of a p-type source, an n-type drain, and an intrinsic channel whose electrostatic potential is controlled by a gate terminal. The device is operated by applying a gate bias voltage to lower the conduction band of the intrinsic region, effectively narrowing the barrier width to allow band-to-band-tunneling of electrons from the valence band of the p-type source to the conduction band of the n-type drain.

A planar TFET structure with 22nm channel length is used to synthesize an ultra low power SPU. For this purpose, we consider the III-V GaSb/InAs source/channel heterojunction TFET (HTFET) which has higher performance than Si based TFET [79]. Table 6 lists the process/geometrical parameters of the TFET used in this study. A drain underlap of 4nm is utilized to suppress ambipolarity in the device. A doping gradient of 2nm/decade is considered at the source/channel junction to enhance on-current. Process to improve its doping gradient is discussed in [83]. A non-local BTBT model [84] is assumed. A non-local BTBT model adaptively searches for the steepest electric field gradient with changing bias conditions and identifies the shortest tunneling path. The BTBT parameters are

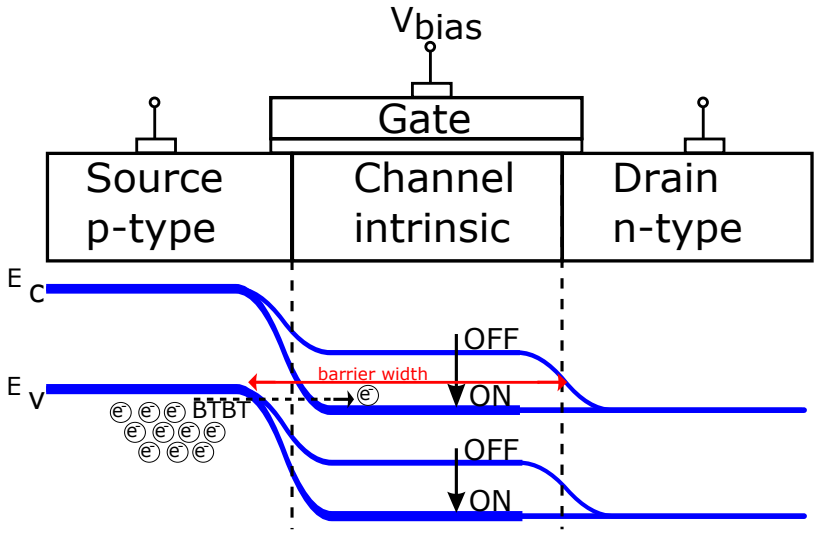


Figure 62: Schematic and band diagram of a TFET.

Table 6: TFET Process Parameters

Specification	InAs/GaSb
Source material	GaSb
Channel material	InAs
Source doping	$4x10^{19}/cm^3$
Drain doping	$4x10^{17}/cm^3$
Dielectric	Al_2O_3/HfO_2
Workfunction	4.85eV

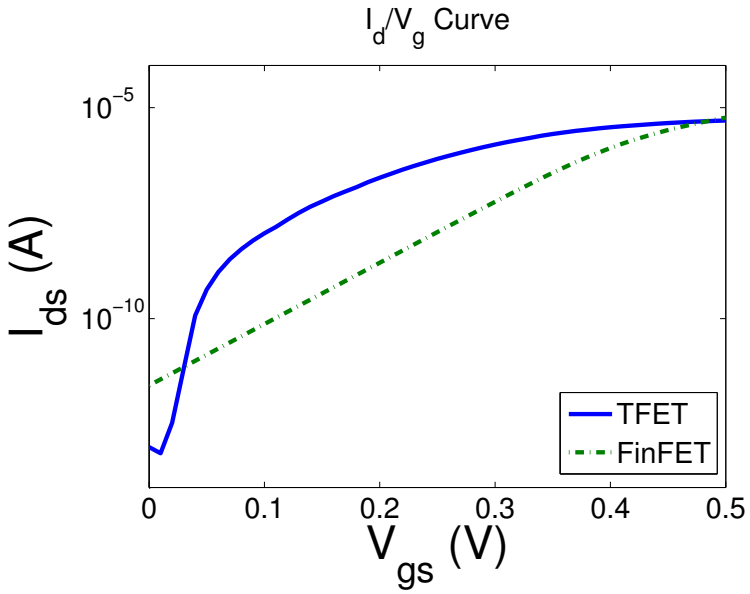


Figure 63: Transconductance of the n-TFET and n-FinFET models.

also calibrated against the characteristics reported in [85]. A Hurkx trap-assited-tunneling (TAT) model [86] is used to estimate the off-current.

Figure 63 shows the drain current (I_{DS}) comparison between the GaSb/InAs HTFET and the FinFET. The HTFET achieves I_{ON} that is comparable to the FinFET device due to the extremely narrow band-gap at the source/channel junction [85]. It has an I_{OFF} that is 50X lower than the FinFET. This gives the HTFET a higher I_{ON}/I_{OFF} ratio than the FinFET. At low V_{GS}, the HTFET has a subthreshold slope that is much steeper than the FinFET, although it degrades as V_{GS} increases. These characteristics make the HTFET a superior device in low voltage low power operation.

4.7.2 TFET Based Digital Logic Library

To synthesize a TFET based digital system, a TFET standard cell library is required. In this chapter, a simple logic library with a minimum number of standard gates are developed. Although various TFET compact models have been previously developed and presented [87, 88, 89], this work utilizes a simplified look-up-table approach for its device model. A flow diagram to generate the TFET library is shown in Figure 64. At first, electrical

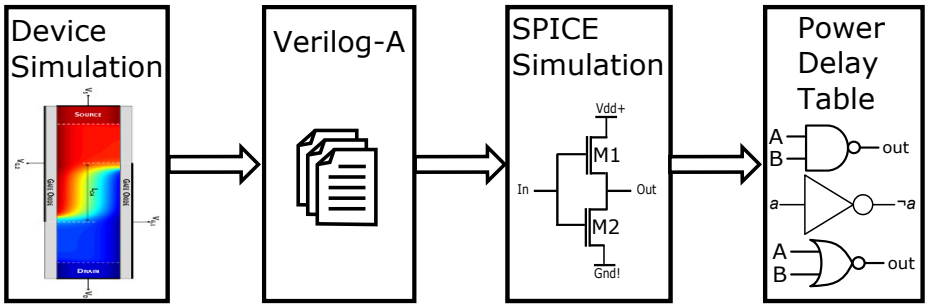
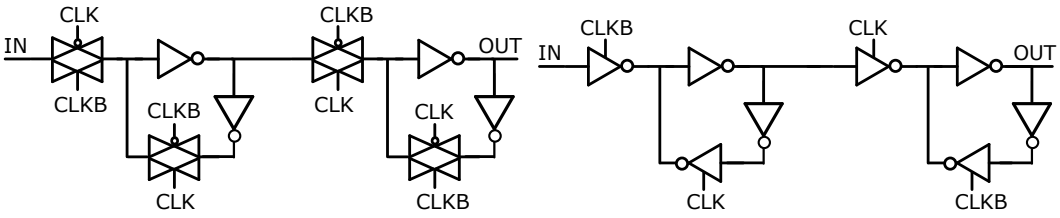


Figure 64: Flow diagram to build the TFET logic library.



(a) (b)

Figure 65: Master-slave flip flop schematics. (a) Transmission gate implementation commonly found in CMOS libraries. (b) Tristate inverter implementation used for the TFET and FinFET library.

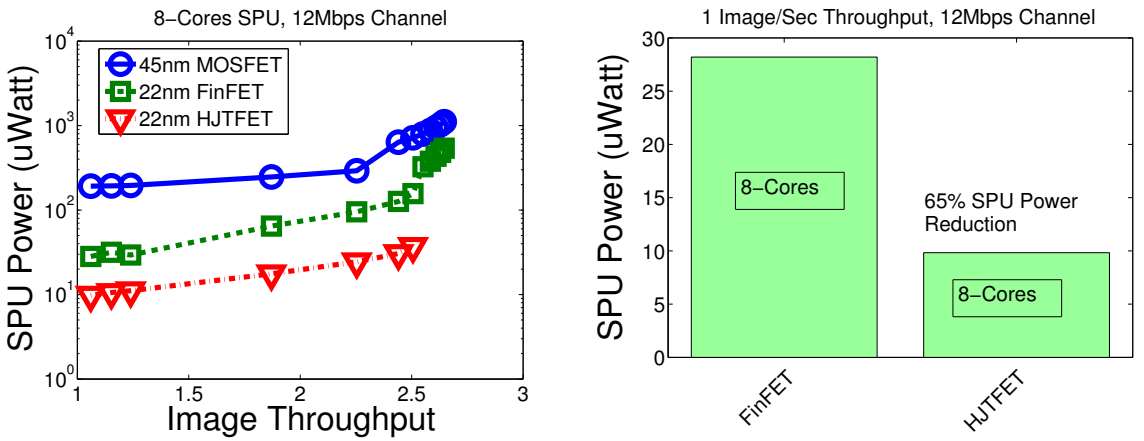
characteristics of TFET are simulated using Synopsys Sentaurus TCAD tools with bias conditions finely varying over the operating range. Next, SPICE compatible Verilog-A table models are constructed by interpolating these values with quadratic spline. Then, a set of SPICE based test vectors is generated for each cell using Cadence Encounter Library Characterizer (ELC) tools. Finally, the generated test vectors and the Verilog-A based TFET device model are simulated in HSPICE to construct the power-delay table necessary for logic synthesis. Note that a different type of flip-flop is required in TFET. Master slave flip-flop is commonly implemented using a transmission gate design. However, TFET is a unidirectional device because its source has different polarity from its drain. This issue is solved by replacing the transmission gates with tri-state inverters as shown in Figure 65.

4.7.3 Simulation Results

For this experiment we consider a low performance wireless image sensor node for monitoring a parking lot. The expected image throughput is between 1 to 3 images/sec. The image resolution is 512X512. The photosensor module has only one ADC. The wireless channel bitrate is 12 Mbps. The SPU is synthesized using 45nm MOSFET, 22nm FinFET, and 22nm HJTFET libraries. Figure 66a shows the SPU power. The power consumption of the SPU synthesized in 45nm MOSFET is higher than in 22nm FinFET because of a significantly higher leakage current. In this low throughput region, leakage power of the 45nm SPU dominates, thus the decrement of SPU power consumption with decreasing throughput saturates. In addition, Figure 66a indicates that synthesizing the SPU with HJTFET, in this low throughput region, yields approximately 60% to 70% lower SPU power than FinFET. Figure 66b shows the SPU power when compressing with a throughput rate of 1 image/sec. The power optimal configuration is determined to be an 8-cores SPU with 600 kHz clock speed. The supply voltage is set to 0.4 V for both HJTFET and FinFET design. The HJTFET is shown to be more power efficient than the FinFET, in this low throughput region, due to a higher sub-threshold slope of the HJTFET. The power of the SPU synthesized with HJTFET is approximately 65% lower than the FinFET SPU.

4.8 Summary

In this chapter, the power and die area reduction of the image processing module due to technology scaling is presented. We started with analyzing the area of the SPU and image buffer. It is shown that area reduces by 75% when the device is scaled from 180nm to 90nm and from 90nm to 45nm. Next, we analyze the power dissipation of the system in 180nm. Our results shows that the SPU is the power hungry module in the system. By scaling down the technology, dynamic power is significantly reduced, however leakage power starts to increase. At 45nm, the network and image buffer power is high, and it is at comparable level with the SPU. In the low performance operation region, network/image-buffer power



(a) (b)

Figure 66: Power comparison of the SPU of the low performance wireless image sensor synthesized in (a) 45nm MOSFET, 22nm FinFET, and 22nm HJTFET with varying throughput, (b) 22nm FinFET and 22nm HJTFET at 1 image/second throughput.

is higher than the SPU power. Leakage power dominates due to the low activity in the system, and choosing for high parallelism becomes an inefficient solution. In the high performance operation region, SPU power is higher than network/image buffer power. SPU activity is high, and dynamic power dominates. The noise level at the photosensor tiers are directly related to temperature, which is determined by the power density and thermal coupling behavior of the 3D-stack. As the technology is scaled down, total power dissipation reduces, but the area reduces with it. The results show that image quality generally reduces with technology scaling.

CHAPTER 5

CONCLUSIONS

5.1 Summary and Contribution

The purpose of this thesis is to develop a methodology for designing a 3D-integrated image processing system for low power, network based applications. Image data is generally huge in size, thus applying a compression technique as the first processing step is critical to reduce resource consumption, such as storage space and power. Many advanced compression algorithms require heavy computation, thus achieving real time compression with low power budget is challenging. This thesis thoroughly investigates, at the system level, the prospect of vertically stacking the signal processing module with the imager storage element in a 3D structure, and applying a parallel computing paradigm as an integrated solution to reduce the workload of the signal processing unit, ultimately reducing power consumption. A design methodology is developed by studying the relations between the controllable design parameters (i.e. number of cores, clock speed), uncontrollable variables from the environment (i.e. image lighting, transmission channel bitrate), and power/performance of the system. In studying the methodology, the following contributions are presented:

1. System level modeling of a 3D integrated image sensing/compression system that considers the correlation between design choices, power, performance, thermal coupling, and noise.

2. Analysis of the implication of the architecture, design parameters, external environment factors, and technology scaling to the power, performances, and image quality of the system.

3. Multi-segment image compression architecture that utilizes the benefits of 3D-stacking and multicore processing for image sensing/compression application.

The multi-segment image compression architecture is presented in Chapter 2. In designing the architecture, an implementation of the discrete wavelet transform based compression algorithm is considered. This algorithm is characterized by intensive interactions between the image storage element and the signal processing unit. The image buffer is vertically stacked on top of the signal processing unit to increase data flow. Parallel computing method is applied by dividing the image into multiple segments, which increases the locality of image buffer access and reduce access time. However, the main objective is to introduce a divide and conquer strategy to reduce the workload of each signal processing unit, scale down the supply voltage, and save power. The image throughput and the power behavior of the multi-segment architecture is analyzed under varying off-chip channel bitrate. The core configuration and clock speed optimization for minimum power consumption is achieved through design space exploration. The effectiveness of the multi-segment architecture is analyzed for bot the 3D- and 2D-integration of the system. Our analysis shows that the multi-segment architecture benefits from the highly parallel connections of the 3D stack. At 180nm, dynamic power of the signal processing unit dominates the power budget. At the end, a multi-clock domain is introduced to further reduce dynamic power of the signal processing unit, but at the cost of an increased in the network/image buffer power.

In Chapter 3, an analysis framework to investigate the relations between the power/performance of the multi-segment image compression unit and the noise characteristics of the photosensor on the top tier is presented. Noise behavior of a logarithmic CMOS image sensor is studied. Logarithmic CMOS image sensor provides faster response and wider dynamic range than linear charge CMOS image sensor, but it is also prone to noise. A coupled power, thermal, and noise (spatial noise, temporal noise, and CIS output voltage range) analysis framework is developed to correlate the architectural design choices, environmental conditions, power dissipation, and image quality of the system. The signal processing

unit has no access to a heat sink, strictly limiting the heat dissipation capability of the system. The above analysis shows that due to the die-to-die thermal coupling, 3D integration introduces new challenges to the noise management and the image quality control. The spatial noise is caused by process variability. Our analysis suggests that it has weak dependence on the power dissipation of the system. However, the output voltage range of the sensor varies with temperature, and it is significantly affected by the compression and ADC power. On the other hand, the temporal noise is caused by electronic noise sources such as thermal noise, flicker noise, and shot noise. Out of the three, only thermal noise has a direct dependence with temperature. However, the level of the temporal noise is lower than the spatial noise by approximately a factor of ten. Naturally, power dissipation of the compression unit varies with time due to factors like different image throughput demands, lighting conditions, and wireless channel bandwidth. This power behavior ultimately causes dynamic variability in the image quality captured by the photosensor. Although the 3D integration has a strong potential to improve energy-efficiency of the image sensor, managing the effect of thermal coupling on the characteristics of the image sensor is critical.

In Chapter 4, the analysis framework is extended to study the effect of heterogeneous integration of the 3D-integrated image sensor system. The image compression module, which is a digital block, is synthesized using deep sub-micron process, while the photosensor module stays with 180nm process. Scaling down the digital block results in significant reduction of die area and power consumption. The rate of power reduction with technology scaling is determined by the activity level of the compression module, which varies with image throughput target and channel bandwidth. If the compression module is designed for high performance operation with high activity level, technology scaling provides huge advantage at discounting dynamic power of the system. However, if the compression module is designed for low performance low activity application, the power budget is dominated by leakage power, which increases as the channel length is scaled down. Nevertheless, area significantly reduces as the channel length is reduced. The noise level and quality of the

compressed image is influenced by changes in the power dissipation as well as die area of the system. At ultra low performance application range, the Tunnel FET is presented as a promising alternative to further reduce power consumption.

5.2 Recommendation for Future Work

To progress the research further, this work can be extended in a number of ways. As a first step, this dissertation primarily focused on synthesizing the signal processing unit with Tunnel FET, which is more power efficient than standard CMOS for low throughput image compression. The compression module includes not only signal processing module, but also image buffer and network router. The image buffer is made of arrays of SRAM cells. An HDL implementation of a network on chip is presented in [90, 91]. To model the power and performance of the compression module, a Tunnel FET implementation of the SRAM and network on chip is needed.

As a second step, the impact of thermal coupling to the power/performance of the image compression module needs to be analyzed. A raise in temperature increases leakage current and decreases carrier mobility, which lead to high leakage power and performance degradation. This feedback effect may be amplified due to lack of heat flow to remove heat from the 3D stack.

As a third step, investigating the effect of wireless channel noise due to interference to the image quality. In cases where interference level is high, the system has to either increase the transmit power or reduce the transmission bit per symbol rate. Therefore, managing wireless channel noise requires a tradeoff between available power budget and target performance of the system.

REFERENCES

[1] E. Fossum, "Cmos image sensors: electronic camera-on-a-chip," *Electron Devices, IEEE Transactions on*, vol. 44, pp. 1689–1698, Oct 1997.

[2] A. El Gamal, "Trends in cmos image sensor technology and design," in *Electron Devices Meeting, 2002. IEDM '02. International*, pp. 805–808, Dec 2002.

[3] K. Banerjee, S. Souri, P. Kapur, and K. Saraswat, "3-d ics: a novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration," *Proceedings of the IEEE*, vol. 89, pp. 602–633, May 2001.

[4] W. Davis, J. Wilson, S. Mick, J. Xu, H. Hua, C. Mineo, A. Sule, M. Steer, and P. Franzon, "Demystifying 3d ics: the pros and cons of going vertical," *Design Test of Computers, IEEE*, vol. 22, pp. 498–510, Nov 2005.

[5] O. Al-Shaykh, I. Moccagatta, and H. Chen, "Jpeg-2000: a new still image compression standard," in *Signals, Systems amp; Computers, 1998. Conference Record of the Thirty-Second Asilomar Conference on*, vol. 1, pp. 99–103 vol.1, Nov 1998.

[6] N. Kimura and S. Latifi, "A survey on data compression in wireless sensor networks," in *Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference on*, vol. 2, pp. 8–13 Vol. 2, April 2005.

[7] A. Topol, D. Tulipe, L. Shi, D. Frank, K. Bernstein, S. Steen, A. Kumar, G. Singco, A. Young, K. Guarini, and M. Ieong, "Three-dimensional integrated circuits," *IBM Journal of Research and Development*, vol. 50, pp. 491–506, July 2006.

[8] P. Leduca, F. de Crecy, M. Fayolle, B. Charlet, T. Enot, M. Zussy, B. Jones, J.-C. Barbe, N. Kernevez, N. Sillon, S. Maitrejean, and D. Louisa, "Challenges for 3d ic integration: bonding quality and thermal management," in *International Interconnect Technology Conference, IEEE 2007*, pp. 210–212, June 2007.

[9] G. Amelio, J. Bertram, W.J., and M. Tompsett, "Charge-coupled imaging devices: Design considerations," *Electron Devices, IEEE Transactions on*, vol. 18, pp. 986–992, Nov 1971.

[10] H. Tian, *Noise analysis in CMOS image sensors*. PhD dissertation, Stanford University, August 2000.

[11] S. Mendis, S. Kemeny, and E. Fossum, "A 128/spl times/128 cmos active pixel image sensor for highly integrated imaging systems," in *Electron Devices Meeting, 1993. IEDM '93. Technical Digest., International*, pp. 583–586, Dec 1993.

[12] S. K. Mendis, S. E. Kemeny, R. C. Gee, B. Pain, Q. Kim, and E. R. Fossum, "Progress in cmos active pixel image sensors," 1994.

[13] J. Nakamura, *Image sensors and signal processing for digital still cameras*. CRC Press, 2005.

[14] Y. Matsunaga and N. Suzuki, "An interline transfer ccd imager," in *Solid-State Circuits Conference. Digest of Technical Papers. 1984 IEEE International*, vol. XXVII, pp. 32–33, Feb 1984.

[15] T. Yamada, K. Ikeda, and N. Suzuki, "A line-address ccd image sensor," in *Solid-State Circuits Conference. Digest of Technical Papers. 1987 IEEE International*, vol. XXX, pp. 106–107, Feb 1987.

[16] T. Nomoto, S. Hosokai, T. Isokawa, R. Hyuga, S. Nakajima, and T. Terada, "A 4 m-pixel cmd image sensor with block and skip access capability," *Electron Devices, IEEE Transactions on*, vol. 44, pp. 1738–1746, Oct 1997.

[17] R. Dyck and G. Weckler, "Integrated arrays of silicon photodetectors for image sensing," *Electron Devices, IEEE Transactions on*, vol. 15, pp. 196–201, Apr 1968.

[18] M. Loose, K. Meier, and J. Schemmel, "Cmos image sensor with logarithmic response and self-calibrating fixed pattern noise correction," 1998.

[19] O. Yadid-Pecht, "Wide-dynamic-range sensors," *Optical Engineering*, vol. 38, no. 10, pp. 1650–1660, 1999.

[20] D. Joseph and S. Collins, "Modeling, calibration, and correction of nonlinear illumination-dependent fixed pattern noise in logarithmic cmos image sensors," *Instrumentation and Measurement, IEEE Transactions on*, vol. 51, pp. 996–1001, Oct 2002.

[21] D. Joseph, *Modelling and calibration of logarithmic CMOS image sensors*. PhD dissertation, University of Oxford, Sept 2002.

[22] M. Loose, K. Meier, and J. Schemmel, "A self-calibrating single-chip cmos camera with logarithmic response," *Solid-State Circuits, IEEE Journal of*, vol. 36, pp. 586–596, Apr 2001.

[23] D. Huffman, "A method for the construction of minimum-redundancy codes," *Proceedings of the IRE*, vol. 40, pp. 1098–1101, Sept 1952.

[24] P. Howard and J. Vitter, "Arithmetic coding for data compression," *Proceedings of the IEEE*, vol. 82, pp. 857–865, Jun 1994.

[25] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *Information Theory, IEEE Transactions on*, vol. 24, pp. 530–536, Sep 1978.

[26] T. A. Welch, "A technique for high-performance data compression," *Computer*, vol. 17, pp. 8–19, June 1984.

[27] J. Nunez-Yanez, X. Chen, N. Canagarajah, and R. Vitulli, "Statistical lossless compression of space imagery and general data in a reconfigurable architecture," in *Adaptive Hardware and Systems, 2008. AHS '08. NASA/ESA Conference on*, pp. 172–177, June 2008.

[28] M. M. H. Chowdhury and A. Khatun, "Image compression using discrete wavelet transform," *International Journal of Computer Science Issues*, vol. 9, pp. 327–330, July 2012.

[29] E. Balster, B. Fortener, and W. Turri, "Integer computation of jpeg2000 wavelet transform and quantization for lossy compression," in *Communication Systems Networks and Digital Signal Processing (CSNDSP), 2010 7th International Symposium on*, pp. 495–500, July 2010.

[30] K. Sayood, *Introduction to data compression*. Academic Press, 2nd ed., 2000.

[31] J. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *Signal Processing, IEEE Transactions on*, vol. 41, pp. 3445–3462, Dec 1993.

[32] A. Said and W. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 6, pp. 243–250, Jun 1996.

[33] D. Taubman, "High performance scalable image compression with ebcot," *Image Processing, IEEE Transactions on*, vol. 9, pp. 1158–1170, Jul 2000.

[34] J. Kim and T. Park, "High performance vlsi architecture of 2d discrete wavelet transform with scalable lattice structure."

[35] W. Sweldens, "Lifting scheme: a new philosophy in biorthogonal wavelet constructions," 1995.

[36] C.-T. Huang, P.-C. Tseng, and L.-G. Chen, "Analysis and vlsi architecture for 1-d and 2-d discrete wavelet transform," *Signal Processing, IEEE Transactions on*, vol. 53, pp. 1575–1586, April 2005.

[37] C. Chrysafis and A. Ortega, "Line-based, reduced memory, wavelet image compression," *Image Processing, IEEE Transactions on*, vol. 9, pp. 378–389, Mar 2000.

[38] C.-T. Huang, P.-C. Tseng, and L.-G. Chen, "Generic ram-based architectures for two-dimensional discrete wavelet transform with line-based method," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 15, pp. 910–920, July 2005.

[39] C. Chakrabarti and M. Vishwanath, "Efficient realizations of the discrete and continuous wavelet transforms: from single chip implementations to mappings on simd array computers," *Signal Processing, IEEE Transactions on*, vol. 43, pp. 759–771, Mar 1995.

[40] J.-T. Kim, Y.-H. Lee, T. Isshiki, and H. Kunieda, "Scalable vlsi architectures for lattice structure-based discrete wavelet transform," *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, vol. 45, pp. 1031–1043, Aug 1998.

[41] F. Marino, "Efficient high-speed/low-power pipelined architecture for the direct 2-d discrete wavelet transform," *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, vol. 47, pp. 1476–1491, Dec 2000.

[42] L. Xue, C. Liu, and S. Tiwari, "Multi-layers with buried structures (mlbs): an approach to three-dimensional integration," in *SOI Conference, 2001 IEEE International*, pp. 117–118, 2001.

[43] M. Koyanagi, T. Nakamura, Y. Yamada, H. Kikuchi, T. Fukushima, T. Tanaka, and H. Kurino, "Three-dimensional integration technology based on wafer bonding with vertical buried interconnections," *Electron Devices, IEEE Transactions on*, vol. 53, pp. 2799–2808, Nov 2006.

[44] R. Reif, A. Fan, K.-N. Chen, and S. Das, "Fabrication technologies for three-dimensional integrated circuits," in *Quality Electronic Design, 2002. Proceedings. International Symposium on*, pp. 33–37, 2002.

[45] B. Black, M. Annavaram, N. Brekelbaum, J. DeVale, L. Jiang, G. Loh, D. McCauley, P. Morrow, D. Nelson, D. Pantuso, P. Reed, J. Rupley, S. Shankar, J. Shen, and C. Webb, "Die stacking (3d) microarchitecture," in *Microarchitecture, 2006. MICRO-39. 39th Annual IEEE/ACM International Symposium on*, pp. 469–479, Dec 2006.

[46] S. Chatterjee, M. Cho, R. Rao, and S. Mukhopadhyay, "Impact of die-to-die thermal coupling on the electrical characteristics of 3d stacked sram cache," in *Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM), 2012 28th Annual IEEE*, pp. 14–19, March 2012.

[47] W. Yueh, S. Chatterjee, A. Trivedi, and S. Mukhopadhyay, "Performance and robustness of 3-d integrated sram considering tier-to-tier thermal and supply crosstalk," *Components, Packaging and Manufacturing Technology, IEEE Transactions on*, vol. 3, pp. 943–953, June 2013.

[48] D. Joseph and S. Collins, "Temperature dependence of fixed pattern noise in logarithmic cmos image sensors," *Instrumentation and Measurement, IEEE Transactions on*, vol. 58, pp. 2503–2511, Aug 2009.

[49] M. Motoyoshi and M. Koyanagi, "3d-lsi technology for image sensor," *Journal of Instrumentation*, vol. 4, no. 03, p. P03009, 2009.

[50] K. W. Lee, T. Nakamura, K. Sakuma, K. T. Park, H. Shimazutsu, N. Miyakawa, K. Y. Kim, H. Kurino, and M. Koyanagi, "Development of three-dimensional integration technology for highly parallel image-processing chip," *Japanese Journal of Applied Physics*, vol. 39, no. 4S, p. 2473, 2000.

[51] K. Kiyoyama, Y. Ohara, K. W. Lee, Y. Yang, T. Fukushima, T. Tanaka, and M. Koyanagi, "A parallel adc for high-speed cmos image processing system with 3d structure," in *3D System Integration, 2009. 3DIC 2009. IEEE International Conference on*, pp. 1–4, Sept 2009.

[52] K. Kiyoyama, K. W. Lee, T. Fukushima, H. Naganuma, H. Kobayashi, T. Tanaka, and M. Koyanagi, "A very low area adc for 3-d stacked cmos image processing system," in *3D Systems Integration Conference (3DIC), 2011 IEEE International*, pp. 1–4, Jan 2012.

[53] V. Suntharalingam, R. Berger, S. Clark, J. Knecht, A. Messier, K. Newcomb, D. Rathman, R. Slattery, A. Soares, C. Stevenson, K. Warner, D. Young, L. P. Ang, B. Mansoorian, and D. Shaver, "A 4-side tileable back illuminated 3d-integrated mpixel cmos image sensor," in *Solid-State Circuits Conference - Digest of Technical Papers, 2009. ISSCC 2009. IEEE International*, pp. 38–39,39a, Feb 2009.

[54] X. Zhang, S. Chen, and E. Culurciello, "A second generation 3d integrated feature-extracting image sensor," in *Sensors, 2011 IEEE*, pp. 1933–1936, Oct 2011.

[55] D. Le Gall and A. Tabatabai, "Sub-band coding of digital images using symmetric short kernel filters and arithmetic coding techniques," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pp. 761–764 vol.2, Apr 1988.

[56] A. Kahng, B. Li, L.-S. Peh, and K. Samadi, "Orion 2.0: A fast and accurate noc power and area model for early-stage design space exploration," in *Design, Automation Test in Europe Conference Exhibition, 2009. DATE '09.*, pp. 423–428, April 2009.

[57] S. Li, K. Chen, J.-H. Ahn, J. Brockman, and N. Jouppi, "Cacti-p: Architecture-level modeling for sram-based structures with advanced leakage reduction techniques," in *Computer-Aided Design (ICCAD), 2011 IEEE/ACM International Conference on*, pp. 694–701, Nov 2011.

[58] T. Dertinger, R. Colyer, G. Iyer, S. Weiss, and J. Enderlein, "Fast, background-free, 3d super-resolution optical fluctuation imaging (sofi)," *Proceedings of the National Academy of Sciences*, vol. 106, no. 52, pp. 22287–22292, 2009.

[59] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, pp. 2366–2369, Aug 2010.

[60] D. Lie, K. Chae, and S. Mukhopadhyay, "On the impact of 3d integration on high-throughput sensor information processing: A case study with image sensing," in *Nanoscale Architectures (NANOARCH), 2013 IEEE/ACM International Symposium on*, pp. 128–133, July 2013.

[61] W. Pennebaker and J. Mitchell, *JPEG: Still Image Data Compression Standard*. Chapman & Hall digital multimedia standards series, Springer, 1993.

[62] L.-W. Lai, C.-H. Lai, and Y.-C. King, "A novel logarithmic response cmos image sensor with high output voltage swing and in-pixel fixed-pattern noise reduction," *Sensors Journal, IEEE*, vol. 4, pp. 122–126, Feb 2004.

[63] H. Zimouche and G. Sicard, "Temperature compensation method for logarithmic cmos vision sensor using cmos voltage reference bandgap technique," in *Electronics, Circuits, and Systems (ICECS), 2010 17th IEEE International Conference on*, pp. 910–913, Dec 2010.

[64] P. Fry, P. Noble, and R. Rycroft, "Fixed-pattern noise in photomatrices," *Solid-State Circuits, IEEE Journal of*, vol. 5, pp. 250–254, Oct 1970.

[65] A. El Gamal and H. Eltoukhy, "Cmos image sensors," *Circuits and Devices Magazine, IEEE*, vol. 21, pp. 6–20, May 2005.

[66] A. J. P. Theuwissen, "Ccd or cmos image sensors for consumer digital still photography?," in *VLSI Technology, Systems, and Applications, 2001. Proceedings of Technical Papers. 2001 International Symposium on*, pp. 168–171, 2001.

[67] H. Nyquist, "Thermal agitation of electric charge in conductors," *Phys. Rev.*, vol. 32, pp. 110–113, Jul 1928.

[68] J. B. Johnson, "Thermal agitation of electricity in conductors," *Phys. Rev.*, vol. 32, pp. 97–109, Jul 1928.

[69] H. Tian, *Noise analysis in CMOS image sensors*. PhD thesis, stanFord university, 2000.

[70] N. Kawai and S. Kawahito, "Noise analysis of high-gain, low-noise column readout circuits for cmos image sensors," *Electron Devices, IEEE Transactions on*, vol. 51, pp. 185–194, Feb 2004.

[71] R. J. Baker, *CMOS: circuit design, layout, and simulation*, vol. 18. John Wiley & Sons, 2011.

[72] A. L. F. Hugh D. Young, Roger A. Freedman, *University physics with modern physics*. Addison-Wesley, 2011.

[73] A. S. Grove, *Physics and technology of semiconductor devices*. Wiley, 1967.

[74] H.-S. Wong, "Technology and device scaling considerations for cmos imagers," *Electron Devices, IEEE Transactions on*, vol. 43, pp. 2131–2142, Dec 1996.

[75] N. I. M. O. at ASU, "http://ptm.asu.edu," 2014.

[76] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. New York, NY, USA: Cambridge University Press, 2nd ed., 2009.

[77] D. Fried, E. Nowak, B. Rainey, and D. Sadana, "Fin fet devices from bulk semiconductor and method for forming," Nov. 4 2003. US Patent 6,642,090.

[78] E. Nowak, T. Ludwig, I. Aller, J. Kedzierski, M. Leong, B. Rainey, M. Breitwisch, V. Gemhoefer, J. Keinert, and D. Fried, "Scaling beyond the 65 nm node with finfet-dgcmos," in *Custom Integrated Circuits Conference, 2003. Proceedings of the IEEE 2003*, pp. 339–342, Sept 2003.

[79] A. Trivedi, M. Amir, and S. Mukhopadhyay, "Ultra-low power electronics with si/ge tunnel fet," in *Design, Automation and Test in Europe Conference and Exhibition (DATE), 2014*, pp. 1–6, March 2014.

[80] S. Park, T. Krishna, C.-H. Chen, B. Daya, A. Chandrakasan, and L.-S. Peh, "Approaching the theoretical limits of a mesh noc with a 16-node chip prototype in 45nm soi," in *Design Automation Conference (DAC), 2012 49th ACM/EDAC/IEEE*, pp. 398–405, June 2012.

[81] C. Sun, C.-H. Chen, G. Kurian, L. Wei, J. Miller, A. Agarwal, L.-S. Peh, and V. Stojanovic, "Dsent - a tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling," in *Networks on Chip (NoCS), 2012 Sixth IEEE/ACM International Symposium on*, pp. 201–210, May 2012.

[82] J. W. King, "Cisco ip video surveillance design guide," 2009.

[83] A. Vandooren, D. Leonelli, R. Rooyackers, A. Hikavyy, K. Devriendt, M. Demand, R. Loo, G. Groeseneken, and C. Huyghebaert, "Analysis of trap-assisted tunneling in vertical si homo-junction and sige hetero-junction tunnel-fets," *Solid-State Electronics*, vol. 83, no. 0, pp. 50 – 55, 2013. Selected Papers from the 6th International SiGe Technology and Device Meeting (ISTDM 2012).

[84] Synopsys, "http://www.synopsys.com/tools/tcad/pages/default.aspx," 2014.

[85] V. Saripalli, A. Mishra, S. Datta, and V. Narayanan, "An energy-efficient heterogeneous cmp based on hybrid tfet-cmos cores," in *Design Automation Conference (DAC), 2011 48th ACM/EDAC/IEEE*, pp. 729–734, June 2011.

[86] G. Hurkx, D. Klaassen, and M. Knuvers, "A new recombination model for device simulation including tunneling," *Electron Devices, IEEE Transactions on*, vol. 39, pp. 331–338, Feb 1992.

[87] K. Bhuwalka, J. Schulze, and I. Eisele, "Scaling the vertical tunnel fet with tunnel bandgap modulation and gate workfunction engineering," *Electron Devices, IEEE Transactions on*, vol. 52, pp. 909–917, May 2005.

[88] W. Vandenberghe, A. Verhulst, G. Groeseneken, B. Soree, and W. Magnus, "Analytical model for a tunnel field-effect transistor," in *Electrotechnical Conference, 2008. MELECON 2008. The 14th IEEE Mediterranean*, pp. 923–928, May 2008.

[89] M. Bardon, H. Neves, R. Puers, and C. Van Hoof, "Pseudo-two-dimensional model for double-gate tunnel fets considering the junctions depletion regions," *Electron Devices, IEEE Transactions on*, vol. 57, pp. 827–834, April 2010.

[90] D. Becker, “http://nocs.stanford.edu/cgi-bin/trac.cgi/wiki/resources/router.”

[91] D. Becker, *Efficient Microarchitecture For Network-On-Chip Routers*. PhD thesis, Stanford University, August 2012.