UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM MICROELETRÔNICA

LEONARDO BANDEIRA SOARES

A Simulation-Based Methodology focused on Energy-efficient Approximate Hardware Accelerators Design

Thesis presented in partial fulfillment of the requirements for the degree of Doctor in Microelectronics

Prof. Dr. Sergio Bampi
Advisor
Prof. Dr. Eduardo Antonio César da Costa
Co-advisor

Porto Alegre
2018

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Soares, Leonardo Bandeira

A Simulation-Based Methodology focused on Energy-efficient Approximate Hardware Accelerators Design / Leonardo Bandeira Soares. – 2018.

128 f.:il.

Orientador: Sergio Bampi; Co-orientador: Eduardo Antonio César da Costa.

Tese (Doutorado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Microeletrônica. Porto Alegre, BR – RS, 2018.

1.Computação Aproximada. 2.Eficiência energética 3.Concepção de circuitos CMOS. I. Bampi, Sergio. II. Costa, Eduardo Antonio Cesar da . III. Título.

“*On ne voit bien qu’avec le coeur. L’essentiel est invisible pour les yeux.*”

Antoine de Saint-Exupéry – Le petit prince

AGRADECIMENTOS

Agradeço a Deus pelas sucessivas bênçãos em minha vida.

À minha mãe, Mariza Bandeira Soares, por ser uma mulher batalhadora que não mediu esforços para que eu tivesse uma educação básica de qualidade e pudesse conquistar meus objetivos mesmo nas situações em que enfrentamos dificuldades. Obrigado, mãe, por todo incentivo, conselhos, amor e carinho. Te amo!

À minha esposa, Andréia Peres de Oliveira, e minha filha, Antônia de Oliveira Soares, por serem meu porto seguro e a minha fonte de motivação, cumplicidade, amor, e por tornarem tudo mais fácil durante esta etapa do Doutorado. Andréia, a tua dedicação para com tua profissão e família, bem como tua humanidade me motivam a ser uma pessoa cada vez melhor. Obrigado pelo teu incentivo e compreensão em todas as etapas do Doutorado. Te amo! Antônia, desde que você nasceu meu mundo se encheu de alegrias, amor e motivação. És minha filha amada! Te amo!

Às minhas tias, Marilda Correa Bandeira e Mara Ione Correa Bandeira, por todo apoio, carinho e amor. Vocês certamente representam muito mais que tias em minha vida! Amo vocês!

À minha sogra e mãe, Maria Bernardina de Oliveira, por todo apoio e carinho durante este período.

In Memoriam para meu pai, Jorge Luis Ferreira Soares, meu avô, Pedro Bandeira, e minha avó, Olívia Antônia Correa Bandeira. Pai, tu representaste um referencial importante em minha vida, e tu, da tua maneira, soube demonstrar o amor de pai para um filho. Obrigado! Te amo! Vô, obrigado por ter sido um pai para mim. Você me ensinou valores importantíssimos. Te amo! Vó, até hoje sinto saudades de tanta doçura, carinho e amor. Obrigado por ter me ensinado lições que levarei por toda a vida. Te amo!

Ao meu orientador e amigo, Sergio Bampi, pelas orientações e contribuições durante o Doutorado. Obrigado por todas as oportunidades que me foram dadas e também pela confiança que tiveste em mim e em meu trabalho.

Ao meu co-orientador e amigo, Eduardo Costa, pelas contribuições técnicas e por todas incansáveis revisões feitas aos trabalhos escritos feitos durante o Doutorado.

A todos os colegas e amigos feitos durante todo o tempo de permanência no laboratório 215 do Instituto de Informática da UFRGS. Ao amigo, André Luís Rodeghiero

Rosa, pela parceria e apoio incondicional. Ao amigo, Cláudio Diniz, por toda ajuda neste período de Doutorado. Aos amigos Mateus Grellert, Dieison Soares, Eduarda Monteiro, Felipe Sampaio, Daniel Palomino, Kleber Stangherlin, Cristiano Thiele, Marcos Hervé, Bruno Zatt, Bruno Vizzotto, Leandro Ávila, Guilherme Paim, Leandro Rocha, Ana Mativi, Brunno Abreu e Giovanni, por ajudas na pesquisa, por compartilhamento de ideias, pelas conversas filosóficas, futebolísticas e, também, pelos momentos de lazer.

Aos colegas da Universidade Católica de Pelotas, Matheus Stigger, Júlio Oliveira e André Sapper, pelo suporte e compartilhamento de ideias e por toda a colaboração com o bom andamento da pesquisa.

A todos os demais colegas e amigos da UFRGS que se esforçam diariamente para desenvolverem suas pesquisas.

A todos os servidores e docentes da UFRGS que prestam seus serviços com excelência.

À sociedade brasileira pelo custeio de universidades públicas de qualidade, na esperança por dias melhores para nosso país.

ABSTRACT

The increasing power density and the pervasive use of compute-intensive and power-hungry applications demand energy-efficient CMOS design. This work proposes a systematic simulation-based design flow to explore the integration of state-of-the-art approximate adders inside hardware accelerator architectures regarding approximation-tolerant applications. The approximate computing concept emerged as a promising technique to drive energy efficiency for CMOS technologies. In this context, the proposed techniques are focused on the tradeoff between accuracy and energy efficiency. Most of the state-of-the-art methodologies for approximate computing exploration are analytical or concentrated in the arithmetic and logic layers of abstraction and they do not consider real input data distributions. Another characteristic found in related works is the weak capability to connect layers, when performing quality-power-performance profiles from the arithmetic up to the application layer. Differently from the state-of-the-art, in this thesis the proposed methodology takes into account the cross-layer integration challenge and presents different quality-power-performance results by considering real test-cases. Three distinct case studies are evaluated in approximation-tolerant applications scope: i) FIR filters for audio processing; ii) Canny edge detection for computer vision algorithms; iii) Motion estimation computation for video coding application. Results show that the proposed design flow is suitable for exploring cross-layer approximate computing integration by considering both the energy efficiency analysis and the application quality. In terms of energy efficiency evaluation, the proposed approach plus the search heuristics are able to seek for suboptimal approximation during design-time which resulted in an energy reduction of up to 57.4%. In addition, the accuracy-configurable approach is proposed in architectural level by exploring coarse grain pruning. In this context, the proposed schemes are designed to accomplish run-time capabilities for distinct power-performance-accuracy profiles. The proposed accuracy-configurable accelerators present dynamic power reduction of up to 64% for the case where most of the operational blocks are clock gated. For quality analysis, realistic objective metrics were systematically explored by considering a large set of real test cases. Results indicate that the proposed methodology contributes with an in-depth characterization for quality-power-performance profiles.

Keywords: Approximate Computing, Accelerator architectures, Low power CMOS design, Digital signal processing applications

Concepção de uma Metodologia baseada em Simulações Focada no Projeto de Aceleradores de Hardware Energeticamente Eficientes e Aproximados

RESUMO

O aumento da densidade de potência e do uso pervasivo de aplicações com alto custo em esforço computacional e potência exigem eficiência energética no projeto CMOS. Este trabalho propõe um fluxo de projeto baseado em simulações para explorar a integração entre somadores aproximados do estado da arte e aceleradores de *hardware* para aplicações tolerantes a erros. O conceito de computação aproximada emergiu como uma técnica promissora para fomentar eficiência energética em tecnologias CMOS recentes. Neste contexto, as técnicas propostas são focadas no balanço de compromisso entre exatidão e eficiência energética. A maioria das metodologias do estado da arte é analítica ou concentrada na camada de abstração aritmética sem considerar casos de teste reais. Outra característica encontrada nos trabalhos relacionados refere-se ao baixo acoplamento quando considerados perfis de qualidade-potência-desempenho computacional, desde a camada aritmética até a camada da aplicação. Diferentemente do estado da arte, a metodologia proposta neste trabalho leva em consideração o desafio de integração entre camadas de abstração e apresenta diferentes perfis de qualidade, potência e desempenho computacional, quando são utilizados casos de teste reais. Três estudos de caso são avaliados no escopo de aplicações tolerantes a erros: i) filtros FIR no processamento de áudio; ii) detector de bordas Canny; e iii) métricas para a estimativa de movimento em aplicações de codificação de vídeo. Os resultados indicam que o fluxo de projeto proposto é adequado para explorar integração entre camadas de abstração no contexto de computação aproximada quando considerados os critérios de eficiência energética, bem como a qualidade da aplicação. Em termos de eficiência energética, a proposta deste trabalho resultou em redução no consumo energético em até 57,4%. Em adição, este trabalho propõe aproximação com granularidade grossa em aceleradores de *hardware* com o objetivo de obter uma solução configurável. Neste contexto, os esquemas propostos foram projetados para atender diferentes perfis de qualidade-potência-desempenho computacional em tempo de execução. As arquiteturas configuráveis apresentam redução na dissipação de potência dinâmica de até 64%. Para a análise de qualidade, métricas objetivas e realísticas foram sistematicamente exploradas considerando um conjunto maior de casos de teste reais. Resultados indicam que a solução proposta contribui com uma caracterização abrangente em termos de qualidade, potência dissipada e desempenho computacional.

Palavras-chave: Computação aproximada, aceleradores de hardware, projeto CMOS de baixa potência, aplicações de processamento digital de sinais.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABREVIATURES AND ACRONYMS

ACA	Almost Correct Adder
ARM	Advanced RISC Machine
ASIC	Application Specific Integrated Circuit
BD-BR	Bjøntegaard-Delta Bit-Rate
BD-PSNR	Bjøntegaard-Delta PSNR
BSD	Berkeley Segmentation Dataset
CIF	Common Intermediate Format
CLA	Carry Look-ahead Adder
CMOS	Complementary Metal-Oxide-Semiconductor
CORDIC	COordinate Rotation DIgital Computer
CPU	Central Processing Unit
CSA	Carry Select Adder
CTC	Common Test Conditions
DA	Distributed Arithmetic
dBFS	decibel level relative to Full Scale
DCT	Discrete Cosine Transform
DSP	Digital Signal Processing
DVFS	Dynamic Voltage Frequency Scaling
ETAI	Error Tolerant Adder I
ETAII	Error Tolerant Adder II
ETAIV	Error Tolerant Adder IV
FFT	Fast Fourier Transform
FINFET	Fin Field-Effect Transistor

FIR	Finite Impulse Response
FHD	Full High Definition
FPGA	Field Gate Programmable Array
fps	frames per second
GDA	Gracefully-degrading accuracy-configurable adder
GeAr	Generic accuracy-configurable adder
GF	Gaussian Filter
GPGPU	General Purpose Graphic Processor Unit
GPU	Graphic Processor Unit
HD	High Definition
HEVC	High-Efficiency Video Coding
HLS	High-Level Synthesis
HM	HEVC Model
HT	Hadamard Transform
IIR	Infinite Impulse Response
IoT	Internet of Things
IP	Internet Protocol
IPs	Intellectual Properties
JPEG	Joint Photographic Experts Group
kbps	kilobits per second
KSA	Kogge-Stone Adder
LOA	Lower-part-OR adder
LSB	Least Significant Bit
ME	Motion Estimation
MEF	Mean Energy per Frame

MEOp	Mean Energy per Operation
MMCM	Multiplier-less Multiple Constant Multiplication
MPBR	Maximum Pass Band Ripple
MPEG	Moving Picture Experts Group
MSB	Most Significant Bit
MSBR	Maximum Stop Band Ripple
MSE	Mean Squared Error
NPA	Nearest Pixel Approximation
NP-Complete	Nondeterministic polynomial time - Complete
PDK	Process Design Kit
PSNR	Peak Signal to Noise Ratio
QHD	Quad High Definition
QP	Quantization Parameter
RCA	Ripple Carry Adder
RMSE	Root Mean Squared Error
ROM	Read Only Memory
RTL	Register Transfer Level
SAD	Sum of Absolute Differences
SATD	Sum of Absolute Transformed Differences
SIMD	Single Instruction Multiple Data
SNR	Signal to Noise Ratio
SOCs	System-on-chips
SRAM	Static Random Access Memory
SSIM	Structural Similarity Index
TCF	Toggle Count Format

TDP	Thermal Design Power
THD+N	Total Harmonic Distortion plus Noise
TV	Television
UHD	Ultra High Definition
VCD	Value Change Dump
VHDL	Very High-Speed Integrated Circuits Hardware Description Language
VLSA	Variable Latency Speculative Adder
VLSI	Very Large Scale Integration
VOS	Voltage over-scaling

TABLE OF CONTENTS

1 INTRODUCTION

The semiconductor industry faces challenges at each new CMOS (Complementary Metal-Oxide-Semiconductor) technology node. One of them is the power density increase which according to (DENNARD, 2015) is related to the mismatch between the CMOS transistor channel length and nominal power supply voltage scaling factors. The latter scales down slower than the former and indicates that the Dennard scaling stated in (DENNARD et al., 1974) is not feasible anymore. Dennard scaling showed that the power density would remain constant with transistor channel length scaling. While this observation was valid, efficiency was enabled through smaller transistors operating at higher frequencies with no substantial increase in power consumption. On the other hand, according to (SHAFIQUE et al., 2014), in recent deep sub-micron planar CMOS technologies, the reduction of transistor threshold voltage results in exponential growth in static power due to leakage current. This undesired growth in static power imposes severe limitation in the transistor threshold voltage scaling which, according to (DENNARD, 2015), can no longer be scaled down. At this point, the progressive scaling applied to the nominal power supply voltage incurs substantial clock frequency reduction. Since in deep sub-micron technologies the nominal power supply voltage scaling factor is typically higher than the one practiced for the transistor channel length, the power per silicon area dramatically increases at each new planar CMOS technology node. Power density is one of the most concerning issues which imposes difficulties to continuously leverage the benefits provided by the Moore's Law (MOORE, 1965).

The previously mentioned power wall in deep sub-micron technologies made many CPU (Central Processing Unit) manufacturers started to design multi- and many-cores CPUs. The multi- and many-cores manufacturing trend was the solution to accomplish the growing workload requirements and to overcome the limitation imposed by the end of Dennard Scaling. However, even this core scaling is constrained to power budgets due to the significant transistor integration capability. Recent works state that the semiconductor

industry is facing the so-called “Dark Silicon era” (SHAFIQUE et al., 2014) (ESMAEILZADEH et al., 2011). In other words, operational blocks in a chip must be powered off to manage excessive heat dissipation. The heat may be controlled through the use of a TDP (Thermal Design Power) constraint which limits the maximum number of transistors simultaneously powered on and working at full performance (SHAFIQUE et al., 2014). When the TDP is violated, the cooling system of a given chip cannot handle the heat dissipation. As a consequence, undesirable and harmful events such as the acceleration of the aging effects are experienced (RAHMANI et al., 2017). The alarming point is that, with the progressive planar CMOS technology scaling, more and more percentage area of operational blocks in chips must be powered off to accomplish TDP requirements. For instance, (SHAFIQUE et al., 2014) presents an analysis grounded on the model provided by (ESMAEILZADEH et al., 2011) showing that, for 8nm CMOS node and CPU-based design, more than 80% of the chip is forecasted to be “dark”. In (HENKEL et al., 2015) it is shown that the analysis performed by (ESMAEILZADEH et al., 2011) did not consider emerging technologies such as FinFETs (Fin Field-Effect Transistor) and the use of consolidated low power techniques like DVFS (Dynamic Voltage Frequency Scaling). Therefore, any analysis based on this previous model results in an over-estimated dark silicon projection. According to the new model presented in (HENKEL et al., 2015), for 11nm FinFET-based technology and DVFS enabled, the worst scenario shows that approximately 50% of the chip area is forecasted to be “dark”.

Another CMOS power concern is associated with the pervasive use of portable and battery-powered devices which are associated with compute-intensive and power-hungry applications. Multimedia applications (*i.e.*, audio, image, and video processing) are at the top of the most used applications in recent years. For instance, in (CISCO, 2016) it is shown a forecast indicating that by 2020 the IP (Internet Protocol) video traffic will represent 82% of the global consumer Internet traffic. This significant amount of video content flowing across the Internet forces the use of video compression. In (VANNE et al., 2012) the state-of-the-art HEVC (High-Efficiency Video Coding) standard is presented. When compared to the previous H.264 standard, the HEVC was proposed to compress twice for the same perceptual video quality. On the other hand, the bottleneck in HEVC turns out to be the computational effort to achieve such high compression capabilities. According to (VANNE et al., 2012), the HM (HEVC model) reference software has computational effort up to 3.2 X higher than the previous H.264 reference software. This number indicates the challenging scenario to reduce

power dissipation while providing higher demanded computational effort when designing SoCs (System-on-a-chips) for mobile devices.

This scenario tends to become more dramatic when considered the emerging ubiquitous concept of IoT (Internet of Things) as defined in (ATZORI; IERA; MORABITO, 2010) (KAMILARIS; PITSILLIDES, 2016). In 2011, the forecast presented by (CISCO, 2011) performed a prediction of 50 billion devices connected to the Internet by 2020. Although recent observations indicate that the 50 billion connected devices prediction is outdated, (NORDRUM, 2016) shows that experts from different companies projected that this number would remain high reaching approximately 30 billion by 2020. From this large amount of devices, many of them refer to sensors which need to implement energy harvesting schemes in order to provide sustainability. Though the large proportion of autonomous sensor systems used in ubiquitous applications, the processing elements on mobile or datacenter devices need to handle much more amount of data provided by the IoT scenario. For instance, according to (ATZORI; IERA; MORABITO, 2010), (NALBANDIAN, 2015), and (KAMILARIS; PITSILLIDES, 2016), data from many fields such as agriculture, logistics, surveillance, health, sports, gaming and so forth arise from the IoT applications and sensors. Hence, to process and extract useful information from this huge amount of data, computational effort plus power concern raises as limiting factors in digital CMOS design scope.

From these previously mentioned issues but not limited to those, one can conclude that is mandatory to shift from performance-driven to energy-efficient oriented CMOS digital circuit design in all abstraction levels. Energy efficiency is defined in (MARKOVIC et al., 2004) as being the maximum number of operations (*i.e.*, instruction fetching/decoding, arithmetic) per energy budget or the minimum consumed energy per operation.

1.1 Motivation and Problem Definition

Due to the challenges previously mentioned considering the CMOS digital design, one can conclude that the use of many energy-efficient techniques is necessary for recent deep sub-micron planar CMOS technologies. This observation is also applicable to emerging CMOS technologies such as FinFETs. Although these technologies substantially attenuate the short-channel effects of deep sub-micron planar bulk, the integration capability and the previously mentioned Dark Silicon projections indicate that power density also affects these technologies. The same observation is also provided in (BAILEY, 2016) which shows an

increase in power density (*i.e.,* W/mm²) for the 22 nm Intel's FinFET-based processor when compared to the 32 nm planar-based one.

The work in (SHAFIQUE et al., 2014) cites three key energy-efficient approaches to cope with Dark silicon era: i) architectural heterogeneity and specific hardware accelerators design; ii) power management by using near-threshold computing; iii) the use of approximate computing.

Architectural heterogeneity and the use of ASIC (Application Specific Integrated Circuit) accelerators are energy-efficient techniques to execute the most compute-intensive kernels of an application (SHAFIQUE et al., 2014). On the other hand, the remaining tasks which demand less energy consumption can be scheduled to general-purpose processors. As a result, general-purpose processors' workload is alleviated due to the use of energy-efficient specific processing cores. Power-management schemes can be implemented to power off accelerators or general-purpose cores when not in use, thus respecting the TDP constraint. The works in (IYER, 2012) and (CONG et al., 2014) show that despite the challenges of architectural integration introduced by this new design paradigm, these accelerator-rich architectures play an essential role in energy efficiency for recent applications. For example, (HAMEED et al., 2010) shows that an ASIC solution is 500 X more energy-efficient than a four core general-purpose processor when considered H.264 video coding application. In the same context of video coding, (DINIZ, 2015) proposed dedicated and reconfigurable hardware accelerators showing solid energy-efficient results.

Near-threshold computing and ultra-low voltage operation leverage the strong quadratic relationship between the power supply voltage and the dynamic power dissipation in CMOS digital circuits. Therefore, when lowering the supply voltage one can reduce the dynamic power dissipation at the expense of decreased computational performance. According to (PINCKNEY; BLAAUW; SYLVESTER, 2015), the near-threshold region is typically located close and above the threshold voltage of the CMOS transistors, while minimum energy point is in the ultra-low voltage or sub-threshold voltage region. The difference between the former and the latter is that operating at near-threshold voltage results in moderate performance and energy reductions, while at the ultra-low voltage the minimum energy point is achieved in detriment of poor computational performance. For instance, (PINCKNEY; BLAAUW; SYLVESTER, 2015) shows an example where operating at near-threshold and ultra-low voltages presents energy reductions of 75% and 87.5% in comparison with the energy consumption at nominal power supply voltage, respectively. On the other

hand, the clock frequency reductions of 80% and 96% are experienced when operating in near-threshold and minimum-energy point, respectively. These numbers are ratified in (STANGHERLIN; BAMPI, 2013) and (ROSA et al., 2015), where substantial energy reduction and poor computational performance are achieved at minimum energy per operation point. According to (PINCKNEY; BLAAUW; SYLVESTER, 2015), different classes of applications can leverage the advantages of using near-threshold and ultra-low voltage computing. For example, autonomous sensor systems required for IoT environment could operate at very low clock frequencies while general-purpose processors or other ASIC accelerators could sustain moderate energy reduction due to the limitations in clock frequency target. The key point is that the supply voltage can also be adjusted during run-time by using DVFS technique aiming at different power-performance profiles. Even considering the variability challenges of operating at lower voltages (SHAFIQUE et al., 2014), near-threshold and ultra-low voltage computing are promising paradigms to overcome the power concern imposed by the progressive CMOS technology scaling and recent computational workloads.

The approximate computing paradigm emerged to increase performance and to reduce power dissipation (HAN; ORSHANSKY, 2013). The key approach in approximate hardware is to reduce the computation accuracy in favor of energy-efficiency. In circuit level design, this is performed by designing simpler circuits to speed up the critical path timing and / or to consume less power. Approximate computing techniques take advantage of approximation-tolerant applications which do not need high accuracy all the time but only “good enough” or “sufficiently good” results for output perceptual quality. In (VENKATARAMANI et al., 2015) is stated the following properties to define an approximation-resilient application: i) there is no a golden or accurate result, but a range of acceptable ones and ii) robustness to input noisy data. For example, multimedia applications (*e.g.*, video coding, audio filtering, image processing, and so on), highly demanded by current portable devices, are intrinsically related with human senses. Since in (ZHU et al., 2010a) is stated that human senses process analog information and have difficulty to realize digital approximations, the multimedia signals are in fact approximation-tolerant applications. In other words, it is possible to adopt approximate computing techniques to improve energy efficiency in multimedia applications by adequately exploring the user experience at different profiles of quality. The scope of approximation-tolerant applications is more substantial than multimedia one. Recent industry researchers in (MISHRA; BARIK; PAUL, 2014) and (ESMAEILZADEH et al., 2012) indicate that the scope of approximation-tolerant applications covers many others cutting-

edge applications such like data mining, machine learning, computer vision, mobile computing, and so on. Based on that, the use of approximate computing may enable energy-efficiency for non-critic quality tasks of IoT scenario.

The excellent point of approximate computing is that this paradigm can be adopted at any abstraction level from transistor-level up to software application (XU; MYTKOWICZ; KIM, 2016). Furthermore, approximate computing can be an additive design component for both the accelerator-rich architectures and near-threshold computing. One can consider that the use of approximate hardware accelerators brings further improvements in terms of energy efficiency (XU; MYTKOWICZ; KIM, 2016). In addition, this promising paradigm can be developed to rescue performance of circuits operating at the near-threshold supply voltage (SOARES et al., 2015), since the approximate hardware is designed to reduce power dissipation and increase computational performance.

On the other hand, approximate computing is in its infancy, and many challenges arise in the design of approximate approaches. According to (XU; MYTKOWICZ; KIM, 2016), one of these challenges is the limited exploration of integration among approximate computing techniques at different levels of abstraction. This integration capability or cross-layer approximate computing is mentioned in (SHAFIQUE et al., 2016) as being a crucial design technique to bridge the gap between approximate approaches from different layers. Recent research is often focused on a single layer of the hardware/software stack. For example, related works have proposed approximate arithmetic operations (*e.g.*, adders and multipliers) with discreet architecture exploration of these key components in approximation-tolerant applications. Also, one of the challenges which arise with cross-layer approximate computing is the difficulty in determining the application quality due to approximation in many abstraction levels. Although approximate computing research is grounded on leveraging error resiliency condition of some applications, an uncontrolled error may substantially degrade the application quality. From these observations emerges the following research question: how to propose cross-layer integration among state-of-the-art approximate computing techniques to improve energy efficiency by respecting quality constraint in the applications under evaluation?

This thesis foresees that the synergy among distinct approximate computing techniques from different abstraction layers brings substantial energy efficiency when executing approximation-tolerant compute-intensive applications. These approaches can be performed during design- or run-time, achieving one of the possible acceptable quality

responses for applications amenable to adopt approximation. The hypothesis to be proofed is guided by the use of a simulation-based methodology to deal with cross-layer integration and to provide energy efficiency in CMOS design. Furthermore, integration between approximate computing, heterogeneous architectures, and near-threshold computing is a promising trend in CMOS digital design.

1.2 Thesis Objectives

The global objective of this thesis is to reveal the implications of integrating approximate computing techniques from different layers of hardware/software stack by considering application quality as well as energy efficiency evaluation. Furthermore, this thesis is focused on proposing integration between approximate computing techniques for both design- and run-time scopes.

The secondary objectives are listed as follow:

- Compute-intensive kernels identification in approximation-tolerant applications.
- Analysis and balance of tradeoff between different levels of quality and energy efficiency for the approximation-tolerant applications under analysis.
- Systematic exploration of state-of-the-art approximate adders to design energy-efficient ASIC accelerator architectures.
- Exploration of run-time power-performance profiles through quality configurable hardware accelerators with management interface for system stack level.
- Investigation of possible integration among approximate computing and accelerator-rich architectures to bring additional energy efficiency.

1.3 Contributions of this work

This work presents the following novel contributions:

- The design of a simulation-based methodology for cross-layer approximate computing integration. In this scope, the state-of-the-art approximate adders are explored in architectural and application layers. The proposed methodology considers real test cases and three different approximation-tolerant applications. Therefore, different power-performance-quality profiles are provided in this work.

- Since simulation-based exploration is time-consuming and prohibitive, new heuristics are proposed to reduce the search for suboptimal solutions.
- Accuracy configurable solutions are proposed in architectural level which can be managed during run-time. This coarse grain technique is capable of producing dynamic response according to the output quality or energy efficiency requirements.

1.4 Outline

The remaining of this thesis is organized as follows: approximate computing related works are reviewed in **Chapter 2**. The approximation-tolerant applications under evaluation are presented in **Chapter 3**. **Chapter 4** presents the proposed methodology to explore state-of-the-art approximate adder parameterization and approximate architectures for ASIC hardware accelerator design. In this same chapter, results regarding application quality and energy efficiency are shown. Accuracy configurable ASIC architectures plus results are presented in **Chapter 5**. Finally, conclusions and future works are drawn in **Chapter 6**.

2 RELATED WORKS ON APPROXIMATE COMPUTING

The previous chapter defined the basic principles of approximate computing paradigm and enumerated the broad set of cutting-edge applications which is amenable to adopt approximation. In this chapter, a review of related works regarding approximate computing is performed.

2.1 Approximate computing in the arithmetic layer

Adders are the basic building block of any computing application. Due to the massive presence of this operator in many compute-intensive tasks, one can conclude that proposing approximation in adders is a key approach to drive energy efficiency in digital CMOS design. According to (ERCEGOVAC; LANG, 2004), the fundamental module of an *n*-bit adder is the full-adder. One possible implementation for the full-adder cell is represented in Figure 2.1.

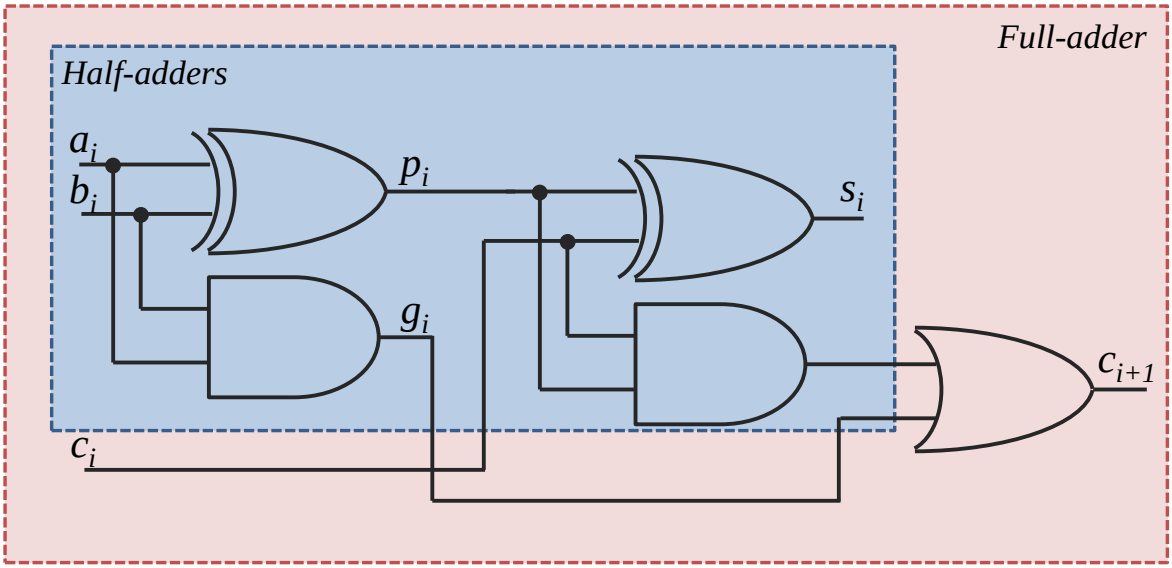

Figure 2.1 – Full-adder circuit composed of half-adders. Adapted from (ERCEGOVAC; LANG, 2004).

In Figure 2.1, the full-adder is composed of two cascaded half-adder sub-modules. These two half-adders are identified in the illustration. The inputs a_i, b_i, and c_i denote the two

operands and carry-in at the ith bit position of a given *n*-bit adder. The terms p_i and g_i refer to the carry-propagate and carry-generate operation outputs which are implemented by XOR (*i.e.*, exclusive OR) and AND logic gates, respectively. The output sum s_i is calculated through an XOR operation between p_i and c_i. The carry-out c_{i+1} is implemented by using an OR operation between two AND operations results: i) the carry-generate of the a_i and b_i input operands; ii) the AND operation between the carry-propagate p_i and the carry-in c_i. The full-adder outputs s_i and c_{i+1} can be represented by the following equations in (1) to (4).

$$p_i = a_i \oplus b_i \tag{1}$$

$$g_i = a_i \wedge b_i \tag{2}$$

$$s_i = p_i \oplus c_i \tag{3}$$

$$c_{i+1} = (p_i \wedge c_i) \vee g_i \tag{4}$$

Based on the fundamental full-adder component, the circuit of an *n*-bit adder may be designed by adopting different topologies. This design may be driven by low power and area or high computational performance requirements. The worst timing path in an adder is the carry chain propagation from the LSB (Least Significant Bit) up to MSB (Most Significant Bit) (ERCEGOVAC; LANG, 2004). In general, low area and power adder circuits tend to present low computational performance. For instance, the RCA (Ripple Carry Adder) presents low power dissipation and area. On the other hand, the cascading topology of *n* full-adders incurs high critical path delay (ERCEGOVAC; LANG, 2004). To cope with low computational performance, carry chain propagation can be accelerated at the expense of both higher logic count and power dissipation. The example of this class is the parallel-prefix adders, whose taxonomy is presented in (HARRIS, 2003).

The previously mentioned tradeoff between computational performance vs. power dissipation (or circuit area) can be alleviated when proposing approximate adders. This is because approximation simplifies the circuit so that the critical path can be reduced as well as the power dissipation. For example, the classical truncation in LSBs of a given adder circuit topology may be the most intuitive source of approximation. In truncation technique, LSBs are pruned or statically set to zero, while the remaining MSBs are implemented as an accurate and conventional adder topology. Following this idea of approximating the LSBs, (GUPTA et al., 2011) and (GUPTA et al., 2013) proposed many versions of approximate adders, in transistor level, based on the pruning of some series connected transistors regarding the mirror full-adder schematic cell. This approach is used to facilitate faster charging/discharging of

node capacitances. Among the different approximate versions, the most energy-efficient practice is to replace the entire full-adder cell by buffers. One of the input operands (a_i or b_i) is copied to the sum result bit s_i without considering carry-propagation between these approximate LSBs. This approach has 50% probability of correct sum result. Therefore, the n-bit adder is broken into two parts: i) the $(n - k)$-bit precise part which contains the MSBs; and ii) the k-bit approximate part which contains the LSBs. Since the adder is divided into two parts, the precise one needs carry-in estimation. In classical truncation, the carry-in is statically set to "0" logic value. Thus, the probability of a correct estimation is 50%. In (GUPTA et al., 2013) this estimation is performed by copying one of the operand bits at the MSB in the approximate part (*i.e.*, the $(k\text{-}1)^{th}$ bit position) as shown in the example of Figure 2.2. This assignment has 75% probability of correct carry-in estimation according to the full-adder truth table. For convenience, this approximate adder is called "copy adder" along this thesis.

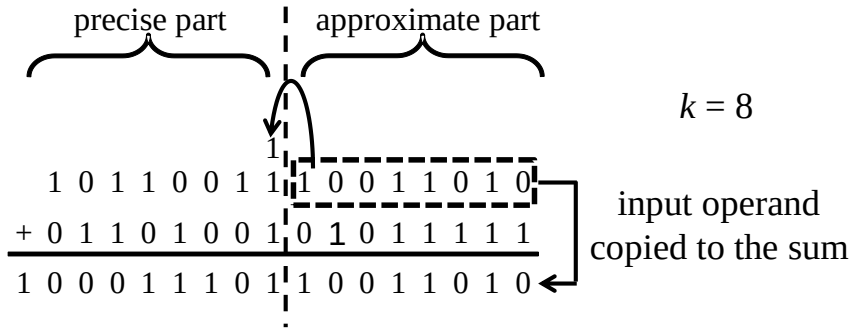

Figure 2.2 – Copy adder example.

The Error Tolerant Adder I (ETAI) proposed in (ZHU et al., 2010a) follows the same technique of dividing a n-bit adder into k-bit approximate and $(n - k)$-bit precise parts. There is no carry chain implemented in the approximate block. It is composed of half adders at each bit position. In the approximate block, the sum bits are computed from the MSB to the LSB direction as follow: each bit is computed by half adders until the first carry-generate operation equal to "1" is found. After that, all the remaining least significant sum bits are set to "1". The precise block can be implemented by any conventional adder topology and computes the sum considering carry-in statically set to "0". As previously mentioned, this technique has a lower probability of correct carry-in estimation than the proposed in (GUPTA et al., 2013). Figure 2.3 shows an example of how the ETAI works.

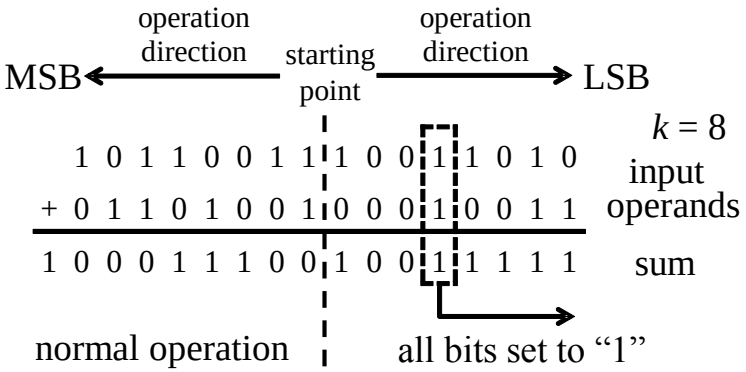

Figure 2.3 – ETAI example. Adapted from (ZHU et al., 2010a).

The Lower-part-OR adder (LOA) proposed in (MAHDIANI et al., 2010) also considers approximating the *k* LSBs of a given *n*-bit adder. This approximate adder replaces full-adder cells by OR gates. The carry-in estimation for the precise part in the k^{th} bit position is performed by considering the carry-generate operation between the input operand bits in the $(k\text{-}1)^{th}$ position. Following the same response of the copy adder, this technique still has 75% probability of correct carry-in estimation for the carry to the LSB of the precise part. On the other hand, the computational cost of an AND gate is higher than the use of a buffer. An example of the LOA is shown in Figure 2.4.

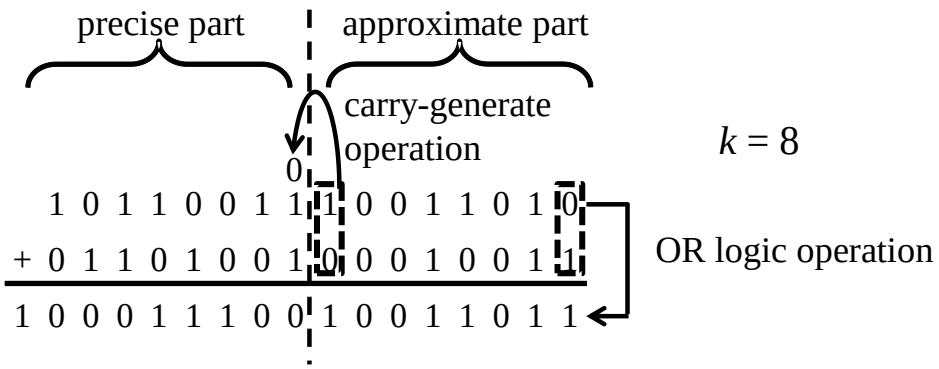

Figure 2.4 – LOA example.

The four approximate adders previously presented are classified as power-oriented approximations. This is because there is logic complexity reduction when compared to the corresponding non-approximate conventional adder topology. Another advantage of this class of approximate adders is that all of them can be implemented in any conventional adder topology. This is because these adders are divided into approximate and precise parts. Therefore, it is expected to reduce energy consumption for any adder topology when compared to the case where there are no approximate LSBs. All of these adders do not

consider error recovery at run-time and try to attenuate error occurrence by using simpler techniques. For example, the copy adder in (GUPTA et al., 2013) presents a power-efficient technique with 75% probability of correct carry-in estimation. According to (HUANG; LACH; ROBINS, 2012), approximate adders which propose simpler cells in the LSBs present highly-frequent small-magnitude error characteristics. Since *k* bits are all approximated with 50% chance of correct sum result per bit, the frequency of errors turns out to be higher. On the other hand, the error magnitude can be controlled by the correct parameterization of the approximate block bit-length (*i.e.*, the *k* parameters).

In a different point of view, there is a class of works which proposed approximate adders to improve the computational performance. The claim is that, for random uniformly distributed pairs of operands, more extended carry propagation rarely occurs (VERMA; BRISK; IENNE, 2008). Therefore, it is possible to divide the *n*-bit adder into *m* independent blocks, where each of them contains *k* bits. This may improve performance if $k << n$, because these shorter blocks can be computed in parallel. The work in (VERMA; BRISK; IENNE, 2008) proposed the block-based Almost Correct Adder (ACA). This approximate adder is divided into *k*-bit overlapping blocks as shown in Figure 2.5.

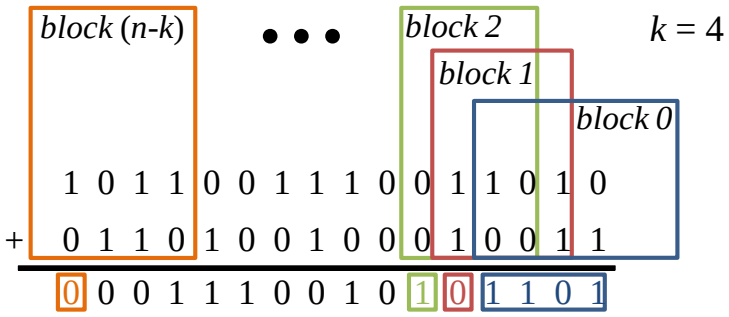

Figure 2.5 – Almost Correct Adder example.

One can observe that the first block (*i.e.*, *block 0*) is responsible for the computation of the first four LSBs sum result. On the other hand, the next blocks (*i.e.*, from 1 up to *n-k*) compute only 1-bit sum result. In other words, one can conclude that each 1-bit sum result is computed considering carry chain speculation of *k*-1 bits. Depending on the conventional adder topology, this adder may incur substantial cost of area and power dissipation. When considered an RCA topology, one can conclude that *n-k*+1 independent *k*-bit adders must be implemented. To cope with this limitation, the authors in (VERMA; BRISK; IENNE, 2008), show that it is possible to reuse logic operators without additional cost regarding area when

adopting Kogge-Stone Adder (KSA) parallel prefix topology. Therefore, for better hardware utilization, this approximate technique is restricted to KSA.

In (ZHU; GOH; YEO, 2009) the Error Tolerant Adder II (ETAII) is proposed. Unlike the ACA, this approximate adder is composed of non-overlapping independent blocks. The adder is divided into *m* blocks of *k* bits as shown in Figure 2.6. Since there is no block overlapping, the proposed method to attenuate error in sum result is the use of *k*-bit carry speculation scheme based on the Carry Look-ahead Adder (CLA). The CLA block estimates carry-in as being "0" and provides carry-in speculation of *k* bits to its next most significant *k*-bit RCA block. The sum generation is performed by the *k*-bit RCA block in which the carry-in is fed by the previous *k*-bit CLA carry speculation block. Therefore, the critical path timing of this approximate adder is composed of *k*-bit CLA plus *k*-bit RCA. Since the ETAII may result in higher magnitude errors due to the non-overlapping condition, in the same work a modified version of the ETAII is proposed. The modified version allows the non-uniform length of carry-in speculation. Therefore, for the most significant parts, the length of carry-in speculation is allowed to be larger than the least significant part. This procedure increases the critical path timing of the adder.

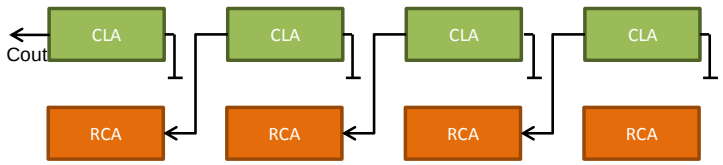

Figure 2.6 – Error Tolerant adder II topology.

The Error Tolerant Adder IV (ETAIV) was further proposed in (ZHU et al., 2010b). This adder can be seen in Figure 2.7. The difference is that the carry speculation has a more extended chain than the ETAII approach. To do that, the ETAIV interleaves CLA and carry select topology (CSA). Each CLA block statically estimates carry-in as being "0" and provides carry-in speculation for its next *k*-bit RCA block. Also, the CLA also selects the multiplexer of the next CSA block which provides carry-in speculation for its next RCA block.

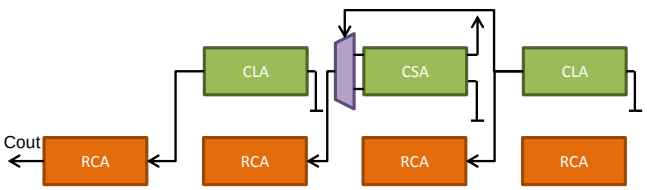

Figure 2.7 – Error Tolerant Adder IV topology.

Hence, if the block bit-length is determined by the *k* parameter, then the critical path is *k*-bit CLA plus *k*-bit CSA plus *k*-bit RCA. This results in higher accuracy than ETAII, at the cost of higher circuit delay.

In sum, one can observe that all the previously mentioned block-based approximate adders are restricted to a specific conventional adder topology such as CLA, KSA, RCA, CSA, and so on. According to (HUANG; LACH; ROBINS, 2012) these block-based adders may have lowly-frequent high-magnitude error characteristic. In other words, the correct choice of carry-in speculation bit-length may result in a low frequency of errors. On the other hand, the presence of many divisions may incur higher magnitude errors when they occur. As a counterpart of the previously mentioned power-oriented adders, these block-based approximate adders are designed to be faster than conventional ones and are not power-efficient. This is because additional logic is implemented to speculate carry-in for each sum-generation block. The exception is the ACA adder that reuses logic from KSA topology to implement the block-based carry speculation without additional complexity. The authors claim that power-efficiency could be achieved by further exploring supply voltage scaling since the block-based adders are designed to improve critical path delay.

Due to the possibility of occasional high-magnitude errors in block-based approximate adders, related works proposed accuracy configurable block-based adders or even error detection and further correction. In the ACA proposed by (VERMA; BRISK; IENNE, 2008), there is a scheme to detect and correct error occurrence. The proposed ACA plus the error detection and correction is called Variable Latency Speculative Adder (VLSA). The VLSA may take up to two clock cycles for error recovery. In (KAHNG; KANG, 2012) is proposed the Accuracy-Configurable Adder that enables run-time accuracy configuration. The proposed adder structure follows the same idea of the ACA. This is because the proposed structure considers overlapping blocks. The only difference is in the overlapping granularity, which presents a coarser grain scheme than the ACA. According to the author, the use of overlapping adder blocks substantially reduces the probability of error induction when compared to ETA II and ETAIV. Therefore, the use of simpler logic approaches is facilitated for error detection and correction, when entirely accurate results are required. The accuracy-configurable adder employs 4-stage pipelined adder architecture to enable accuracy-configuration. The Gracefully-degrading accuracy-configurable adder (GDA) proposed in (YE et al., 2013) has similar structure when compared to ETAII. The difference is that there are multiplexers between pairs of blocks. This scheme controls if the chosen carry-in for the

next sum-generator block (*i.e.,* the RCA block in Figure 2.6) will be provided either by the current sum-generator or carry speculation block (*i.e.,* the CLA block in Figure 2.6). For an entirely accurate result, one can control the GDA by connecting all the sum-generator blocks. If all the multiplexers select carry-out from the corresponding carry-speculative blocks, then the adder response will be equal to the ETAII one. Observing that there is a relation between all the previously mentioned block-based adders, the work in (SHAFIQUE et al., 2015) proposed a clever generic accuracy-configurable adder (GeAr) in which the user can determine the block configurations as well as the overlapping conditions. Also, the author developed probability model of error occurrence and proposed an error correction scheme for the generic approximate adder.

The accuracy-configurable adder, the VLSA, and the GeAr approaches contain additional logic count when compared to the block-based ones. Therefore, these adders are designed to improve computational performance and also to support applications which may demand entirely accurate results during a specified period. This period of accurate computing must be short. Otherwise, the penalty in energy efficiency will be aggressive due to many error detections and corrections. In this class of accuracy-configurable adders, the exception is for the GDA in which the use of multiplexers does not result in substantial area increase. All the main characteristics of the classes of approximate adders reviewed in this subsection are shown in Table 2.1. The characterization was performed according to the results and comparison with corresponding conventional topologies provided in each related work.

Regarding power, only the first class of adders plus the ACA present lower dissipation than its corresponding conventional adder topology. The accuracy-configurable adder, GeAr, and ACA do not provide power results. The power characterization in Table 2.1 does not consider any additional low power techniques such as supply voltage scaling.

Regarding computational performance, all the classes, except the VLSA, are faster than its conventional adder topology. The accuracy-configurable adder, GDA, and GeAr may vary computational performance due to accuracy configuration capability. More accurate results are obtained in detriment of moderate or no performance improvement, while less accurate ones substantially reduce the worst path delay. Although in this thesis many state-of-the-art approximate adders are reviewed, this is not a closed research subject. There are many other aspects which can be addressed in this scope such as the accuracy-configurable capability for adders divided into LSBs approximate and MSBs precise parts.

Table 2.1- State-of-the-art approximate adders characterization.

approximate adder	power dissipation[1]	computational performance[2]	error characteristic	adder topology restriction
truncation copy adder **ETAI** **LOA**	≈ 40% to 50% reduction	≈ 1.1 to 1.7 X faster	frequent low-magnitude error	no
ETAII **ETAIV**	≈ 10% to 14% increase	≈ 3.9 to 4.7 X faster	infrequent high-magnitude error	yes
Accuracy-configurable adder **GeAr**	N/A	≈ 1.1 to 1.4 X faster	accuracy-configurable or correction	no
ACA	N/A	≈ 1.5 to 2.5 X faster	infrequent high-magnitude error	yes
VLSA	N/A	≈ 2% performance reduction	correction capability	yes
GDA	≈ 64% increase	≈ 5.4 faster	accuracy-configurable or correction	no

[1] Power dissipation characteristic in comparison with the corresponding non approximate adder topology

[2] Computational performance characteristic in comparison with the corresponding non approximate adder topology

Based on that, new content in this abstraction level is expected to be published for the next years. Approximate multipliers have also been proposed in recent years, but they are not within the scope of this thesis.

2.2 Approximate adders validation in application level

Evaluation in approximation-tolerant applications is demonstrated to validate the reviewed approximate adders. In (GUPTA et al., 2013), three key applications are explored: i) the Sum of Absolute Differences (SAD) used in the motion estimation module regarding the MPEG (Moving Picture Experts Group) video standard; ii) JPEG (Joint Photographic Experts Group) image compression by approximating the Discrete Cosine Transform (DCT); iii) low-pass FIR (Finite Impulse Response) filter. According to (GUPTA et al., 2013), SAD is a metric used for block matching in Motion Estimation (ME) video context. To motivate approximation in ME, the authors state that this block is responsible for 70% of the total power dissipation in MPEG application. The SAD architecture was implemented by a tree of adders. These adders are approximated by using the same parameterization for all of them (*i.e.*, uniform approximation). Based on that, 50 frames from the Akiyo video sequence in CIF resolution (*i.e.*, 352 x 258) are simulated to evaluate application quality. Results indicated that the average PSNR (Peak Signal to Noise Ratio) ranges from 35 dB up to 37.5 dB for all the adders being approximated from 4 LSBs to 1 LSB. The maximum power saving of 42% is achieved by k = 4 LSBs being approximated. The DCT architecture was implemented by using shifts and additions. A significance driven methodology is adopted to explore approximate adders in the DCT architecture. The authors state that low-frequency coefficients of the DCT have more significance than the high-frequency ones regarding image quality. In other words, adders which compute low-frequency coefficients may be less approximated than the ones responsible for the high-frequency components. They classify the adders by significance considering three different groups for approximate exploration and show PSNR evaluation only for the “Lena” image through an exhaustive search. In sum, the copy adder exploration in DCT results in PSNR level of 25.3 dB and power savings of 69%. The low-pass FIR filter is also implemented by using shift and addition approach. The proposed grouping scheme to approximate the different regions of the filter is driven by the significance of each coefficient to the quality result. To do this, the authors set each coefficient to zero and evaluate the FIR filter quality by considering the maximum pass-band and stop-band ripple (*i.e.*, MPBR and MSBR). The FIR filter under evaluation is a 25-tap filter which process 16-bit input samples. Results indicate percentage change in MPBR and MSBR of 0.27% and

1.92%, respectively. This approximate solution reaches 61% of power reduction. In the last two applications (*i.e.*, the DCT and the 25-tap FIR filter), the power dissipation result is estimated by considering further use of power supply voltage over-scaling (VOS). This is performed since the approximate solution results in a timing slack that can be leveraged. According to the authors, VOS is a technique which differs from the traditional voltage scaling because the clock frequency is not accordingly scaled. Therefore, there is no performance penalty, while the power is reduced. Although approximation for three applications is presented in (GUPTA et al., 2013), the drawback of this work is that the proposed analysis remains far from a more systematic and precise exploration in the application domain. For example, only one image and video are considered for quality and power results, while for the FIR filter there is no application being evaluated. The approximation of adders in these applications considers few configurations. Therefore, this results in the discreet exploration of power-performance-quality profiles.

In (ZHU et al., 2010a) the ETAI utilization is demonstrated by considering the FFT (Fast Fourier Transform) operational block for image processing. The method adopted to approximate the adders from the transform is not explained. The authors follow the same practice presented in (GUPTA et al., 2013): only the Lena image is considered in application scope. There is no objective quality analysis such as PSNR. The subjective analysis is performed by plotting the output images after forward plus inverse transforms when implemented by conventional and approximate adders. The FFT application is also evaluated in (ZHU et al., 2010b) regarding the ETAIV. There are no details on how the approximation is explored in the FFT structure and the end-user application is audio processing. The only audio signal adopted is the pronunciation of “One-Two” sentence. The authors do not consider objective metrics and only plot the audio signal in the time domain to enable subjective comparison. The drawback of the previously mentioned works which show examples for FFT application is the absence of objective quality metrics and a more robust set of real signals to evaluate the average response. The work in (ZHU; GOH; YEO, 2009) which proposed the ETAII does not present any application analysis.

In (SHAFIQUE et al., 2015) only one Full-HD image (*i.e.*, 1920 x 1080) is considered to evaluate the GeAr response regarding critical path timing for the image integral, SAD, and low-pass filter applications. There is no quality analysis or power estimation for the applications under evaluation. In (KAHNG; KANG, 2012) the accuracy-configurable adder is validated by considering the Gaussian image filter application. Following the same trend in

previously mentioned works, only the Lena image is used. Among all the possible accuracy configurations the least accurate one may produce power reduction of 51.6% with a PSNR response of 24.5 dB. The power results also consider additional technique of VOS.

The LOA approach proposed in (MAHDIANI et al., 2010) consider two applications: i) neural networks for face recognition, and ii) fuzzy logic. The use of objective metrics such as Mean Squared Error (MSE) and percentage error are exercised. There is no explanation about how the adders are approximated in the application architectures, but in this case, a broader set of 100 training epochs are used. The synthesis results are shown only regarding the area and computational performance. For the VLSA and ACA proposed in (VERMA; BRISK; IENNE, 2008), there is no application validation, while in (YE et al., 2013) the DCT is used to validate the GDA. The transform is implemented by using the concept of distributed arithmetic (DA) which consists of ROM (Read-Only Memory) and Accumulator components. This is a serial architectural approach where the adder in the accumulator can be configured to perform more or less accurate sum. According to (YE et al., 2013), the configuration is performed to increase and decrease accuracy for low and high-frequency components, respectively. The highest PSNR result for one processed image is of 25.83 dB. Synthesis results are not presented in the application scope.

Based on the aforementioned related works, Table 2.2 summarizes all the aspects of approximate adders exploration towards application scope. One can conclude that all the works present a discreet exploration when considered application-specific context. Only the works in (GUPTA et al., 2013) and (YE et al., 2013) give information about how the adders are approximated in the applications. The set of applications are concentrated in Digital Signal Processing (DSP) domain including DCT, FFT, SAD for video coding, and so forth. The exception is in (MAHDIANI et al., 2010), where artificial intelligence algorithms are explored. Some works adopt objective metrics such as PSNR, MPBR, MSBR, and MSE. On the other hand, other works do not provide any metric for quality evaluation. None of the related works summarized in Table 2.2 consider the evaluation of multi-level quality for a given application. Therefore, only one specific level of objective metric response is analyzed. For instance, in (GUPTA et al., 2013) the copy adder exploration in DCT domain results in a single PSNR response. This specific analysis imposes severe limitations in the power-performance-quality analysis for approximate computing concept validation. As earlier mentioned in Chapter 1, one of the challenges in approximate computing is the absence of

direct and trivial tradeoff balance between energy efficiency and application quality. Hence, a more systematic exploration of this scope is needed.

Table 2.2 – Comparison among approximate adders exploration for application-specific scope

Related Work	Technique to approximate adders inside the application	Applications under evaluation	Objective metric	Levels of quality analysis	Use of additional low power technique	Benchmark used to evaluate the results
(GUPTA et al., 2013)	Uniform and significance-driven	SAD, DCT, and FIR filter	PSNR (35 dB up to 37.5 dB), MPBR (0.27%), and MSBR (1.92%)	Single level	VOS	Limited to one image
(ZHU et al., 2010a)	N.A.	FFT and IFFT	N.A.	N.A.	N.A.	Limited to one image
(ZHU et al., 2010b)	N.A.	FFT and IFFT	N.A.	N.A.	N.A.	Limited to one audio signal
(SHAFIQUE et al., 2015)	N.A.	Image integral, SAD and FIR filter	N.A.	N.A.	N.A.	Limited to one full-HD image
(KAHNG; KANG, 2012)	N.A.	Gaussian Image filter	PSNR (24.5 dB)	Single level	VOS	Limited to one image
(MAHDIANI et al., 2010)	N.A.	Neural Networks and Fuzzy Logic	MSE (0.0006)	Single level	N.A.	Up to 100 training epochs
(YE et al., 2013)	Run-time configuration of one adder in DA approach	DCT	PSNR (25.8 dB)	Single level	VOS	Limited to one image

The works in (GUPTA et al., 2013), (KAHNG; KANG, 2012), and (YE et al., 2013) adopt a combination of approximate computing and VOS technique. This is due to the improvement provided in critical path timing of approximate adders when compared to conventional accurate ones. Regarding the benchmark, most of the works use only one input

signal to validate the application quality. The exception is the work in (MAHDIANI et al., 2010) which consider a broader set of training epochs for neural network application. The limitation in the number of input signals, levels of quality metric analysis, and integration of approximate adders with higher layers of abstraction indicates that the application-specific domain in the approximate computing context is an open research subject.

2.3 Proposed methodologies for cross-layer approximate computing integration

Some works have proposed more systematic methodologies, to cope with the gap between approximate adders exploration inside application layer. In (LIU; HAN; LOMBARDI, 2015) an analytical methodology is proposed to characterize the error response of different block-based state-of-the-art approximate adders. All the probabilistic models are analytically developed by considering random uniformly distributed pairs of input operands. Based on error metrics for approximate adders such as mean error distance and error rate, the authors present an estimation method to show the correlation between those metrics and the PSNR for image processing applications. This is motivated to avoid computational and time-consuming application-specific numeric simulations. The applications under evaluation are: i) image sharpening, ii) point detector, and iii) mean filter. Six images are used as a benchmark to validate the proposed method. Although the authors defend the point of view that there is a correlation between PSNR and mean error distance, only one specific case presents the best match between estimated and simulated PSNR. For all the other cases, there are substantial discrepancies in dB for the estimated and simulated PSNR. For instance, the ETAII with block bit-length $k = 4$, applied to point detector, results in 10 dB for simulated PSNR and 5 dB for estimated one. For all the cases, the proposed estimation produces pessimistic numbers. This indicates that their estimation method based on uniform input distribution may not be the best approach. This is because when performing power-performance-quality profiles for hardware accelerators, the pessimistic quality estimation will negatively restrict reductions in power dissipation and computational performance improvements. Since the power-performance profile is not addressed in (LIU; HAN; LOMBARDI, 2015), this observation does not emerge in that work. In (SHAFIQUE et al., 2016) it is affirmed that data-driven resilience must be investigated and explored for cross-layer approximate computing techniques.

In (VENKATARAMANI et al., 2012) an approximate synthesis methodology is developed where the user can determine the threshold error magnitude. The synthesis flow considers logic elements that may be simplified in the approximate version by respecting the threshold error magnitude value. Power-performance-quality profiles are shown regarding percentage error magnitude for different topologies of adders, multipliers, DCT, FIR filter, and so on. On the other hand, no benchmark is adopted, so that the quality analysis is performed only at the low logic level of abstraction. This is the drawback of the proposed synthesis approach since error magnitude is rarely adopted as metric in the application-specific scenario.

Following the same practice proposed in (LIU; HAN; LOMBARDI, 2015), the probabilistic error model for approximate adders proposed in (MAZAHIR et al., 2017) is developed to characterize error, area, and critical path timing of block-based state-of-the-art approximate adders. Regarding application-specific analysis, the authors show different quality profiles for Gaussian image filters and the SAD metric for video coding application. On the other hand, the estimated metrics are kept at a low level of abstraction instead of mapping to application level metrics. The exercised approximate adder is the GeAr, where different configurations are selected to provide results regarding error probability.

One can conclude that all the presented methodologies are focused on an analytical solution with richer analysis on low abstraction layer. This is because those approaches infer statistical properties of the state-of-the-art approximate adders or logic functions to characterize error metrics for arithmetic or logic operations. The work in (VENKATARAMANI et al., 2012) also evaluates low-level error metrics. On the other hand, their approach is focused on approximate logic synthesis instead of adopting approximate adders in application level. Those analytical approaches are useful regarding the low computational time taken to model and characterize error metrics but are limited in the following aspects: i) the adders of a given application usually are uniformly approximated by the same parameter; ii) present poor analysis at the application-specific level. In architectures with many levels of adders, if heterogeneous parameterization is adopted for the approximate adders, then the error characterization turns out to be time-consuming and expensive regarding computational cost. Most of the analytical methodologies assume uniformly distributed input data. Since different applications present diversified input data distribution, this may produce a discrepancy between low-level error metrics and application quality

metrics. Therefore, the only way of performing a detailed application quality profiling is to use simulation-based approach.

Table 2.3 – Methodologies summarization for approximate computing in application level scope

Related work	Characteristic	Power-performance-quality profiling	State-of-the-art approximate adders	Objective quality metric	Benchmark used
(LIU; HAN; LOMBARDI, 2015)	Analytical, uniform parameter for approximate adders	Only quality	Block-based approximate adders	PSNR	6 images
(VENKATARAMANI et al., 2012)	Approximate logic synthesis	Yes	No use of approximate adders	Error magnitude	N.A.
(MAZAHIR et al., 2017)	Analytical, uniform parameter for approximate adders	N.A.	Block-based approximate adders	Error probability	N.A.
(KANG; KIM; KANG, 2016)	Simulation, heterogeneous parameter for approximate adders	Yes	ETAI	Accuracy	Random inputs
(SHAFIQUE et al., 2016)	Simulation, uniform parameter for approximate adders	N.A.	Copy adder	N.A.	"football" video sequence

Based on that, the work in (KANG; KIM; KANG, 2016) proposed a simulation-based approach to synthesize FIR filters. The filters are implemented by using shifts and additions, and a search heuristic based on the filter adder step is introduced. The filter adder step refers

to the number of adders in the critical path timing. Therefore, each adder which belongs to a specific level is configured with the same approximation parameter. The approximate adder under evaluation is the ETAI, and 5 FIR filters with a number of taps ranging from 15 up to 49 are used to validate the methodology. The quality metric is the accuracy of the FIR filter, and random inputs are used as a benchmark to determine power-performance-quality curves. Only one simulation with real images is performed, where for the 65.4% and 45.6% accuracy targets the filtered image has PSNR equal to 14.7 and -9.5 dB, respectively. Therefore, one can conclude that the accuracy metric cannot be directly mapped to PSNR due to the substantial quality degradation in PSNR when accuracy target is changed from 65.4% to 45.6%. One possible hypothesis for that may be the use of accuracy metric plus random input data to validate the FIR filters. In sum, (KANG; KIM; KANG, 2016) proposed a simulation-based scheme which does not leverage the possibility of adopting an application level quality metric and real signals to validate the methodology.

In (SHAFIQUE et al., 2016), a cross-layer methodology is presented where the study case on SAD is developed by using approximate adder versions proposed by (GUPTA et al., 2013). The SAD metric for block matching in ME scope is exercised by considering all the adders being uniformly parameterized. For instance, all the adders from the SAD are set to 1, approximate LSB, 2, 3, and so forth. The authors present results for one video sequence called "football" where is shown that the candidate block does not tend to change with increasing approximation. According to (SHAFIQUE et al., 2016), 4-bit LSB approximation would be the most reasonable solution for this video sequence. This is because there is no substantial video quality degradation and it does not significantly increase the video bitrate. On the other hand, no more details are given regarding power reduction and the study does not consider a broader set of videos.

Table 2.3 summarizes all the reviewed methodologies proposed in approximate computing context for the application-level scope. The main characteristics are shown for each methodology. Most of the works propose the use of approximate adders in application level. The exception is the work in (VENKATARAMANI et al., 2012) where approximate logic synthesis is introduced. As can be seen, most of the works do not provide power-performance-quality profiles. Only the work in (LIU; HAN; LOMBARDI, 2015) presents an estimation of PSNR as being the objective metric, while the others are focused on low-level error metrics. Regarding the benchmark, the proposed methodologies are limited to few input signals or random inputs. Based on that, one can conclude that the cross-layer approximate

computing integration through methodology is an open research subject. Also, none of these proposed methodologies consider the use of approximation in higher layers such as the architectural one. These previous works only consider approximations at the arithmetic level or logic gates.

2.4 Approximate computing in the architectural layer for application-specific scope

Since this thesis is also focused on bridging the gap of approximation among different abstraction levels, some works which proposed approximate techniques in architectural layer instead of considering the use of approximate adders is reviewed.

Works which propose approximate computing in architectural level are generally focused on a specific application. In (PARK; CHOI; ROY, 2010) a dynamic bit-length DCT is proposed. Following the example in (GUPTA et al., 2013), the work in (PARK; CHOI; ROY, 2010) ratifies the difference of significance between low and high-frequency components in DCT. Therefore, an iterative analysis is previously performed to explore power-performance-quality analysis when reducing the number of bits to represent least significant coefficients (*i.e.,* high frequency). After this evaluation, three profiles are determined and used for reconfiguration during run-time. According to the author, the reconfigurable approach results in power reduction of 36% with PSNR degradation of 0.61 dB. This work considers seven images as a benchmark. The work in (JAISWAL et al., 2015) explores the intrinsic characteristic of neighbor pixels in images. According to the authors, for most of the cases, neighbor pixels are nearly equal with 1% to 4% variation. This condition is defined as Nearest Pixel Approximation (NPA) and can be used to eliminate an excessive number of convolutions when filtering images. Therefore, the proposed work convolves a subset of pixels with the respective Gaussian coefficients to estimate the output filtered sample. This reduces the number of convolutions, and the hardware design can be simplified at the expense of application quality. Two images are used to validate the proposed technique. Results show substantial energy reduction in detriment of low degradation regarding PSNR and SSIM (Structural Similarity Index) quality metrics. In the same application of the Gaussian filter, the work in (CABELLO et al., 2015) proposed the approximation of this filter by adopting fixed-point representation. The fixed-point solution is implemented in FPGA platform and is compared to the floating-point representation. According to the authors, the representation 8.11 (*i.e.,* 8-bit integer part and 11-bit fractional part) is sufficient to maintain

the Root Mean Squared Error (RMSE) below 5% when running a database of 20 images. In (KAUSHIK; KUMAR, 2015) a mirror short pixel approximation is proposed for energy-efficient Gaussian filtering. The authors proposed their solution observing the same condition of NPA shown in (JAISWAL et al., 2015). Based on that, the authors in (KAUSHIK; KUMAR, 2015) proposed to select one pixel by row and copy these pixels for the entire rows to approximate the input image block. The rounding procedure is further implemented. PSNR results are shown for the Lena image where maximum quality is of 28.32 dB. The hardware implementation is synthesized for FPGA, and results indicate power, area and delay reductions of 8%, 85%, and 58%, respectively.

In (HE; GERSTLAUER; ORSHANSKY, 2011) the approximate Inverse DCT (IDCT) is addressed by considering the same significance concept of low and high-frequency IDCT coefficients. On the other hand, the authors propose modifications in the architecture to allow VOS exploration. Therefore, instead of only reducing the bit-length of least significant coefficients, reordering scheme in arithmetic level is proposed to avoid addition between low magnitude and opposite sign operands. This is because, according to the author, in 2's complement notation this specific condition results in worst path delay due to carry propagation. After those proposed modification schemes, the authors explore energy reduction through VOS. The Lena image is used to validate the proposed technique and the highest energy reduction results are found for a PSNR of 32.3 dB. Following the same concept of VOS, in (MOHAPATRA et al., 2011) some compute intensive kernels mostly used in approximation-tolerant applications are identified as follows: i) L1 norm, ii) dot product, and iii) L2 norm. If VOS is applied to the conventional versions of these computational kernels, then uncontrolled timing errors will occur. On the other hand, the authors propose techniques to dynamically configure accuracy inside those kernels to attenuate error due to timing violations. One of the techniques is called dynamic segmentation with error compensation. In dynamic segmentation, the critical path of these circuits is divided into smaller paths and multiplexers are inserted among those sub-modules. Therefore, for aggressive VOS, one can conclude that the control logic must disable most of the parts of the critical path to maintain approximation in sustainable level. The other design technique is called delay budgeting which is based on the insertion of transparent latches to freeze a given operational block output. Thus, increasing the time budget of a given combinational path and reducing the error due to timing violation. Those proposed techniques are exercised considering the SAD for ME in video coding application. The validation is performed by

using 12 CIF resolution video sequences. According to the authors, at iso-quality analysis, their techniques provide energy reduction of 17%.

Table 2.4 – Architectural exploration in approximate computing scope

Related Work	Technique	Additional low power exploration	Quality metric	Benchmark used
(PARK; CHOI; ROY, 2010)	Significance-driven in DCT	N.A.	PSNR	7 images
(JAISWAL et al., 2015)	NPA exploration in Gaussian image filter	N.A.	PSNR and SSIM	2 images
(HE; GERSTLAUER; ORSHANSKY, 2011)	Significance-driven in IDCT	VOS	PSNR	1 image
(MOHAPATRA et al., 2011)	Dynamic segmentation and time budgeting for compute-intensive kernels	VOS	PSNR	12 CIF resolution video sequences
(CABELLO et al., 2015)	Fixed point exploration in Gaussian image filter	N.A.	RMSE	20 images
(KAUSHIK; KUMAR, 2015)	Mirror short pixel approximation in Gaussian image filter	N.A.	PSNR	1 image

Architectural approximate exploration is not a new concept, but it gained substantial attention with the emerging paradigm of approximate computing. There is a challenging gap between simplifications performed in architectures and logic/arithmetic levels which can be addressed in a cross-layer approximate computing context. As can be seen in Table 2.4, the trending directions are as follows: i) to group the operational blocks by their output quality significance in order to explore more or less approximation, ii) to apply VOS in different operational blocks to reduce dynamic power, and iii) to provide run-time configurations in order to dynamically balance the trade-off between accuracy and energy efficiency for these architectures. All these related works consider these techniques as being approximate computing approaches because they are focused on relaxing accuracy to improve energy efficiency. Despite all these substantial efforts to explore approximation in the architectural layer, one can conclude that most of these works still present a limited number of evaluated input signals even considering that this abstraction level is nearer to the application layer. Therefore, bridging the gap between architectural and application layers is an open research subject.

2.5 Summary of the chapter

In this chapter, a systematic review was performed regarding different approaches for approximate computing scope. Approximate computing at arithmetic level was firstly reviewed, where three classes of approximate adders were evidenced: i) power-oriented adders, ii) performance oriented adders, and iii) accuracy-configurable adders. Next, the primitive exploration of these state-of-the-art approximate adders was observed in subsection 2.2, where most of the works did not evaluate heterogeneous approximate parameterization. Furthermore, these related works did not systematically explore the high-level applications. After that, a perspective scenario was observed regarding all the proposed methodologies, where most of them are focused only on a specific abstraction layer. In addition, most of these related works did not evaluate multiple power-performance-quality profiles. Finally, the chapter presented approximate techniques and trends for architectural level, where one of them is related to accuracy-configurable approaches. In the next chapter, the case studies under evaluation are presented.

3 CASE STUDIES ON APPROXIMATION-TOLERANT APPLICATIONS

In this chapter, a brief introduction is presented to the three case studies which are addressed in this work. Motivations, approximation-resiliency characteristics, and compute-intensive bottlenecks will be exercised in the next sections. Since multimedia applications are amenable to adopt approximations, in this thesis audio, image, and video processing applications are addressed.

3.1 FIR filters in audio processing scope

Audio is one type of signal which is processed by human senses. As earlier mentioned, approximations performed in digital signals may not be perceptible in the analog domain for the human sensorial system. The exploration of psychoacoustic characteristics is not a new concept and is adopted as a standardized practice for lossy audio compression as shown in (ISO/IEC, 1993).

Since the usually practiced bitrate in audio applications is in the order of hundreds of kbps (kilobits per second), at first analysis the demanded computational cost may not be a bottleneck. On the other hand, emerging applications related to audio signals are imposing a substantial increase in computational effort. According to (BLEIDT et al., 2015), new digital TV standards are considering improvements in the audio system. The new concept is to deliver an object-based audio system. From the new elements considered by the object-based audio, the following ones are highlighted: i) interactive audio elements and consumer choice; ii) immersive sound; iii) extending consumption with barrier-free accessibility. The first is related to the user capability in selecting one specific audio element. For example, users may select one language out of many options, commentaries from home or away soccer team, additional sound effects, and so on. This imposes a substantial increase in the amount of handled audio data. The immersive sound is emerging as a new requirement outside the cinemas. New configurations in loudspeakers and the increasing number of channels incur

high bitrates. For example, according to the author in (BLEIDT et al., 2015), for 24 channels configuration, a bitrate of 1200 kbps is required. This immersive paradigm has been attracted attention even for mobile devices manufacturers. The work in (QUALCOMM, 2015) shows a system technology which provides immersive surround system for mobiles without the need of headphones. In addition to the many audio objects and immersion, an important point is to also deliver additional components to break boundaries and provide accessibility, so that the user experience of blind people, can be enlarged, for instance.

To handle the increasing computational significance of audio in the recent and upcoming era of digital TV, the works in (BELLOCH et al., 2013) and (BELLOCH et al., 2016) propose accelerating approaches for the massive multi-channel immersive scenario. Those accelerating techniques are based on SIMD (Single Instruction Multiple Data) ARM processors and GPGPU (General Purpose Graphic Processor Unit) units aiming at improving computational performance of FIR filters. These filters are used to enable the correct treatment of large multi-channel application. The authors affirm that massive multi-channel sound signal processing is mainly based on combining the filtered output signals to produce a given special acoustic effect. Therefore, FIR filters become an essential building block in audio processing context for emerging scenario of object-based audio.

FIR filters can be implemented by the following convolution operator shown in (5) (PROAKIS; MANOLAKIS, 1996).

$$y(n) = \sum_{k=0}^{M-1} h(k)x(n-k) \tag{5}$$

In (5), *y* and *x* refer to the output filtered and the input signals, respectively. Also, *n* and *k* denote the current sample and tap positions. The total number of taps is indicated by *M*. The FIR filters are mostly used in many digital signal processing applications due to their stability and linear-phase properties when compared to IIR (Infinite Impulse Response) filters. On the other hand, the drawback of FIR filters is the increased computational cost due to the higher required number of taps. Therefore, one can conclude that energy reduction schemes for FIR filters are necessary.

3.2 Canny edge detection application

As earlier mentioned in Chapter 1, one of the emerging and challenging topics for new CMOS design is related to IoT scenario. Many IoT applications consider the use of cameras

as sensors to extract video or image features by adopting computer vision algorithms. Those tasks enable services in many fields such as agriculture, safety, transportation, and so on. For example, in (CHIEN et al., 2015) a System-On-a-Chip (SoC) for smart cameras is proposed and shows two applications for distributed computing of IoT: i) video surveillance to store data when critical events occur, and ii) vehicle localization for intelligent transportation. Also, others IoT applications which rely on computer vision can be checked on (ZUBAL; LOJKA; ZOLOTOVÁ, 2016) and (KIM et al., 2015), where industrial safety and liquid-level estimation are addressed, respectively. According to (ESMAEILZADEH et al., 2012), computer vision is an application amenable to energy-accuracy tradeoff exploration. As earlier mentioned, this type of task does not have a golden result, but a set of acceptable ones.

According to (CHEIKH et al., 2014), in the pre-processing step of many computer vision algorithms, edge detection technique is needed to allow object segmentation further. The first step of some edge detectors is related to the Gaussian image filter to blur the image and remove undesired noise for the next steps (CANNY, 1986). On the other hand, as will be herein demonstrated, the Gaussian convolution operator may be one of the most compute-intensive tasks among all computation steps of edge detectors. Furthermore, another challenging point is that video or image sensors produce more data than others types of sensors like humidity, infrared, temperature, and so forth. As a result, the computational effort may even be increased to accomplish real-time processing requirements. The primary consequence is the substantial power consumption increase, which is entirely undesired in IoT and semiconductor industry scope. The Canny edge detection algorithm proposed in (CANNY, 1986) can be divided into the following steps: i) image smoothing to attenuate noise level; ii) gradient filter in horizontal and vertical directions to highlight regions with high spatial derivatives; iii) relate the edge gradients to directions that can be traced; iv) tracing valid edges or non-maximum suppression; and v) hysteresis thresholding to eliminate breaking up of edge contours. Those steps are identified in the diagram of Figure 3.1.

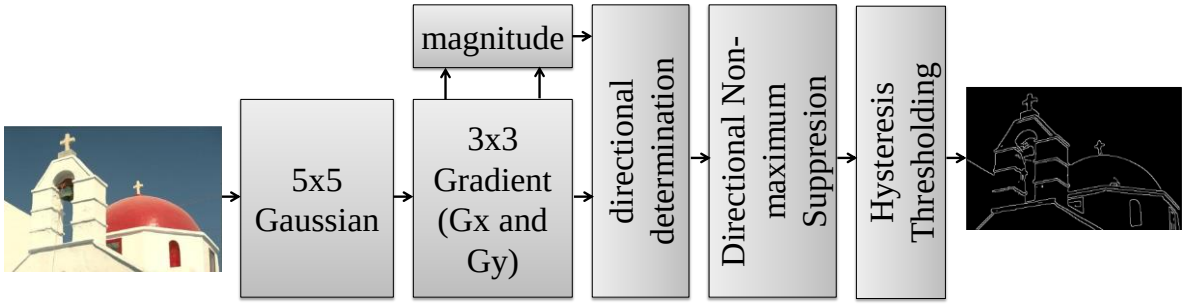

Figure 3.1 – Canny edge detection algorithm steps

The Gaussian filter is a smoothing filter to remove noise. The two-dimension (*i.e.*, *x* and *y* directions) Gaussian kernel is obtained as shown in (6).

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{6}$$

In (6), σ^2 denotes the variance, and this is one of the parameters to obtain different versions of Gaussian kernels. The other parameter is the window size which determines the number of image pixels to be convolved. Higher quality is obtained through larger window sizes (JAISWAL et al., 2015). In general, the most used window size is the 5x5 one. On the other hand, for larger window sizes, the computational cost substantially increases.

After the image has been smoothed and the Gaussian filter has reduced the noise, the next step consists in finding the intensity gradients of the image. For this task, the Sobel operator is used, whose pair of 3x3 convolution masks is presented in (7) and (8). One of the masks estimates the gradient in the y-direction (rows), while the other estimates the gradient in the x-direction (columns). The magnitude of the gradient is calculated by the square root of the sum of the squares of horizontal and vertical derivatives.

$$\frac{1}{4}\begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{7}$$

$$\frac{1}{4}\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \tag{8}$$

When considered the entire Canny edge detection algorithm, the Gaussian and Gradient filters are the most compute-intensive kernels. Table 3.1 summarizes the number of arithmetic operations which need to be performed on a 512 x 512 grey scale image. The 5x5 Gaussian filter is a convolution operator which perform twenty-five multiplications and twenty-four additions per convolution. Since the convolution must be performed 512 x 512 times, the total number of those arithmetic operations is higher than 12 million. For each 3x3 Gradient convolution, nine multiplications and eight additions are needed. Since two instances of the Gradient filter are used, for the derivatives in horizontal and vertical directions, eighteen multiplications and sixteen additions are performed per convolution. In Table 3.1, the number of arithmetic operations is considered for both the magnitude and Gradient. For the magnitude, two multiplications, one addition, and one square root are needed.

Therefore, the total number of multiplications and additions for both the Gradient filter and magnitude operator, are twenty and seventeen, respectively. Since this procedure is repeated 512 x 512 times, the total number of arithmetic operations (*i.e.*, multiplications, additions, and square root) is higher than 9.5 million. In the remaining blocks of the Canny edge detector, the sum of multiplications represents only 6.67% of the total number of multiplications for the 5x5 Gaussian plus 3x3 Gradient filters. This percentage shows that the convolution operations are more compute intensive than the remaining steps.

Table 3.1 – Number of Arithmetic Operations Per Canny Edge Detector Step Considering 512 X 512 Grey Scale Image

	5x5 Gaussian	**3x3 Gradient & magnitude**	**Directions determination**	**Non-maximum suppression**	**Hysteresis thresholding**
Multiplication	6,553,600	5,242,880	262,144	-	524,288
Addition	6,291,456	4,456,448	-	-	-
Comparison	-	-	1,048,576	786,432	1,048,576
Square root	-	262,144	-	-	-

3.3 Motion estimation for the HEVC standard

As previously mentioned in Chapter 1, video data are dominating internet traffic and mobile computing. In recent years, the increase of video resolutions is demanding higher computational effort to improve video compression. The HEVC standard (ITU-T; ISO/IEC, 2013) emerged to double the compression capability when compared to the previous H.264 (ITU-T; ISO/IEC, 2011) for the same perceptual video quality (VANNE et al., 2012). As a consequence, the computational effort substantially increases up to 3.2 X to sustain this improvement (VANNE et al., 2012). Therefore, additional low power techniques are necessary to alleviate the intrinsic power-hungry characteristic of state-of-the-art video coders.

In (VANNE et al., 2012) a profiling scheme is addressed to identify which functional blocks present the most compute-intensive kernels. The authors concluded that considering the HEVC Test Model (HM) reference software, Fractional Motion Estimation and Sum of Absolute Transformed Differences (SATD) represent up to 59% and 18% of the computational effort demanded by the entire encoding process, respectively.

Motion estimation blocks are based on a given search algorithm and a block matching scheme to find similarity between blocks in temporal frames. The goal of this key component is to improve compression by computing a motion vector instead of coding the entire current frame. The algorithms use objective metrics to measure similarity. The most common metric is known as SAD which is defined in (9).

$$SAD = \sum_{i,j} |y_{i,j}| \tag{9}$$

In (9) *y* denotes the residue between the candidate block and the current block at i^{th} row and j^{th} column pixel position. For the SAD metric higher values indicate poor similarity.

More complex metrics are also considered in block matching for video coding to drive better similarity estimation, and one of them is the SATD. According to (SILVEIRA et al., 2015), SATD has superior performance regarding both compression ratio and video quality than SAD. The SATD is defined in (10), (11), and (12).

$$W = H \cdot Y \cdot H^T \tag{10}$$

$$H = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \tag{11}$$

$$SATD = \sum_{i,j} |w_{ij}| \tag{12}$$

In (10) *W*, *H*, and *Y* refer to the transformed coefficients, Hadamard Transform matrix, and the residue pixels, respectively. The 4x4 pixels example of the Hadamard Transform is shown in (11), and in (12) the SATD is presented as the sum of the transformed coefficients *w* at the i^{th} and j^{th} pixel positions.

Table 3.2 – Comparison between SATD and SAD in terms of arithmetic operations count

	SATD			**SAD**		
block size	2x2	4x4	8x8	2x2	4x4	8x8
adders	11	79	447	3	15	63
absolute operators	4	16	64	4	16	64
total	15	95	511	7	31	127

Source: (SOARES et al., 2016)

As can be compared in (9) to (12), the SATD metric has higher computational cost than the SAD. Even considering the arithmetic simplifications that can be performed in hardware design, the SATD has substantial higher computational cost when compared to the SAD. Table 3.2 shows the required number of arithmetic operations for SATD and SAD computation (SOARES et al., 2016), for a fully parallel SATD architecture implemented by additions and shifts. One can conclude that the SATD has up to 4 X more number of arithmetic operations than the SAD. This justifies the compute-intensive nature of SATD.

As mentioned in Chapter 2, many works have proposed approximate SAD architectures. In (SHAFIQUE et al., 2016) is affirmed that depending on the level of approximation in adders, the candidate block may not be changed. Thus, in this specific case, there is no degradation regarding video bitrate and quality. Based on that, SAD and SATD metrics computation for Motion Estimation in HEVC are amenable to adopt approximation.

3.4 Summary of the chapter

In this chapter, the three analyzed case studies were shown followed by their respective background. In addition, compute-intensive kernels inside those applications were identified and motivated for approximate computing exploration. In the next chapter, the proposed methodology to bridge different abstraction levels is presented followed by an exploration of approximate adders for hardware accelerators when considered the case studies under evaluation.

4 PROPOSED METHODOLOGY FOR APPROXIMATE HARDWARE ACCELERATORS DESIGN

In this Chapter, approximate techniques are addressed to drive energy efficiency in digital CMOS design. First, a simulation-based methodology to explore state-of-the-art approximate adders in architectural level is proposed. This methodology is focused on search heuristics to speed up multi-level quality and power-performance profiling when adopting state-of-the-art approximate adders in ASIC architectures. Next, case studies are presented followed by their respective baseline architectures. Furthermore, proposed approximations in architectural level and results for evaluated case studies are presented in this Chapter.

4.1 The proposed methodology and search heuristics

One key approach to implement multiplier-less energy-efficient digital filters and transforms is to consider the use of adders and shift operators. Since coefficients are constants, the use of multipliers turns out to be dispensable. From this observation emerges the Multiple Constant Multiplication problem which is intended to find the minimum number of adders/subtractors for a given filter or transform. According to (VORONENKO; PÜSCHEL, 2007), finding the optimum configuration is known to be an NP-Complete problem, so that heuristics may be the best solution to suboptimal response. The proposed methodology considers filters and transforms implemented by trees of adders and shifts.

The proposed methodology is presented in Figure 4.1. At first, the approximate adders under evaluation are modeled in C language or MatLab. After that, the application (*i.e.*, FIR filters, Transforms, and so on) is implemented in MatLab. Next, simulation based on heuristic is performed considering the application and real test cases. After all the iterations are tested, a search method is applied to classify approximate configurations which have an average objective metric response near different levels of quality. The next step is to generate the approximate designs by using Python scripts. Once the RTL (Register Transfer Level) description is obtained, the RTL simulation and logic synthesis can be performed. The RTL

simulation also uses the real test cases and provides the VCD (Value Change Dump) or TCF (Toggle Count Format) files used for post-synthesis simulation and power estimation. The logic synthesis may be performed in the following ways: i) synthesis for a specific clock frequency target, and ii) bisection method search to evaluate maximum frequency response. Finally, with the VCD or TCF file and the synthesized netlist, post-synthesis simulation and power estimation are addressed.

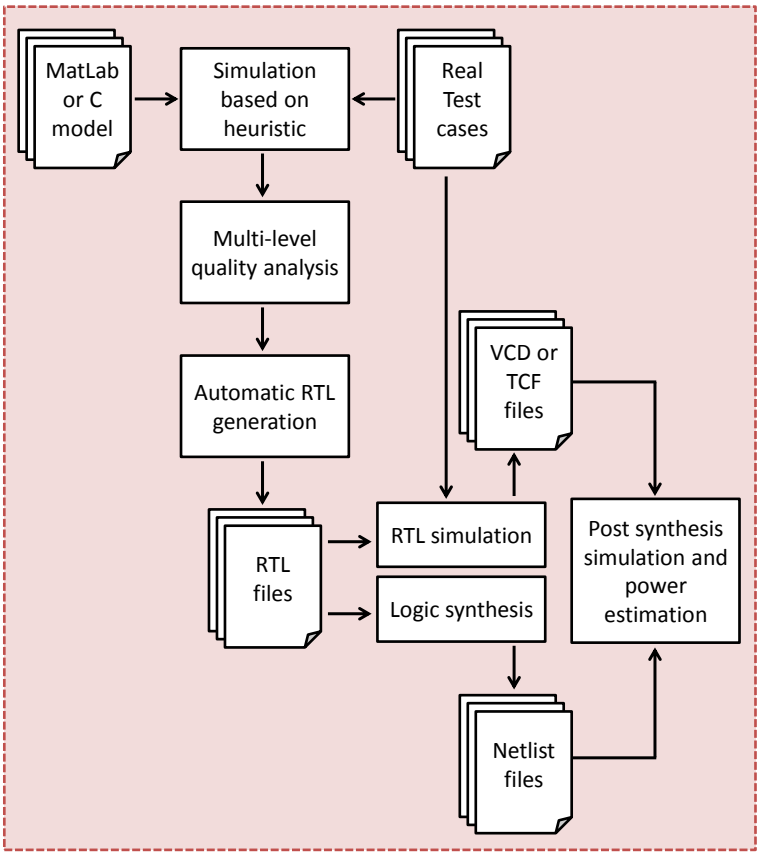

Figure 4.1 – The proposed methodology to explore the use of state-of-the-art approximate adders in approximation-tolerant applications.

When considered exploration of state-of-the-art approximate adders for trees of adders and shifts, one can observe that, depending on the number of adders, the exhaustive search may be time-consuming or a prohibitive approach. For example, consider a simpler FIR filter for image convolution. This filter is implemented by a tree containing ten 16-bit adders and the use of approximation in the LSBs ranging from 1 up to 8 approximate bits. Assuming convolution time in the order of µs (*i.e.*, measured in MatLab), the time taken to process a 512x512 image is 261.14 ms. On the other hand, to search the optimum approximation for each one of the ten adders, 10^8 image convolutions are needed. This results in approximately 304 days to identify the optimum choice regarding application quality. Based on that, the

simulation time per image is defined in (13), where, *t* refers to the unitary processing time of a given application.

$$simulationTime = t \cdot n^k \tag{13}$$

The terms *n* and *k* denote the number of adders and the range of approximation parameter being explored, respectively. The simulation time is unfeasible even considering the multi-core platform. This is because ten adders is a simpler example, and this number is usually higher. Furthermore, one image or signal indicates a limited analysis scenario. Therefore, the use of heuristics to search suboptimal solution is necessary.

4.1.1 Heuristic based on distinct functional blocks

Depending on the application, different functional blocks can be identified. For example, consider that different functional blocks may allow more or less approximation. Therefore, the proposed heuristic is focused on grouping the adders by the functional blocks they belong. The flow in Figure 4.2 depicts how the proposed heuristic works.

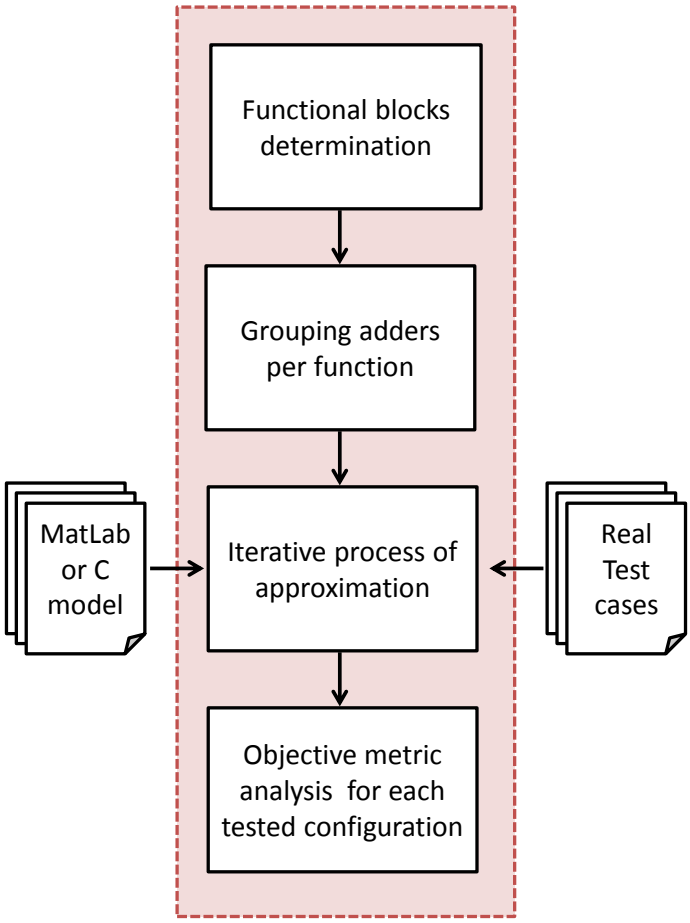

Figure 4.2 – Proposed heuristic based on different functional blocks

In Figure 4.2, the flow corresponds to the process block called “Simulation based on heuristic” in Figure 4.1. At first, for a given application the functional blocks are determined. After that, approximation parameter is determined for all the adders which belong to each functional block. For example, if two different functional blocks are identified, then two approximation parameters are determined (*e.g., k1* and *k2*). The next step is to perform an iterative search to simulate combinations among distinct approximation parameters. Based on that, different configurations between *k1* and *k2* are tested for the application under evaluation. Finally, the objective metric is measured for each configuration. One can conclude that the time taken to perform simulation based on heuristic is shorter than the exhaustive search time shown in (13).

4.1.2 Heuristic based on the estimated output sum magnitude

The previous heuristic handles approximation in functional block level. The secondly proposed heuristic is based on grouping adders which may produce the same order of magnitude at the output sum. Since this thesis evaluates different applications and those applications are data dependent, the proposed heuristic estimates output magnitude by observing how the adders are structured. All the evaluated applications in this study are implemented by shift and addition scheme as explained in (AKSOY et al., 2010) and (VORONENKO; PÜSCHEL, 2007). This type of implementation is composed of trees of shift and additions. The highest level of cascaded adders gives the worst path delay in this approach. Based on that, the proposed heuristic estimates the output magnitude for adders in the first level considering the following parameters: i) the input bit-length of the operands, and ii) the number of bits that are shifted. Right and left shifts result in lower and higher output sum magnitude estimation. The output estimation for one adder is given by considering the largest estimated input operand regarding bit-length, except when there is right-shift operation in one of the operands. For this specific case, the minimum bit-length is chosen to reduce substantial quality degradation. The adders in the remaining levels are estimated by using the same methodology. The only difference is that, for a higher level adder, the input operand may be a previously estimated magnitude sum. The overall heuristic scheme is shown in Figure 4.3.

The first step verifies if there are remaining levels to be estimated in the tree of adders. In the affirmative case, the next step is focused on determining the estimated output magnitude for all the adders in a given level. To clarify how the estimation is performed, an example is shown in Figure 4.4. Assume that the input samples are 16 bits long. Therefore,

the heuristic estimates the output sum for all the adders in the first level at the top of the illustration.

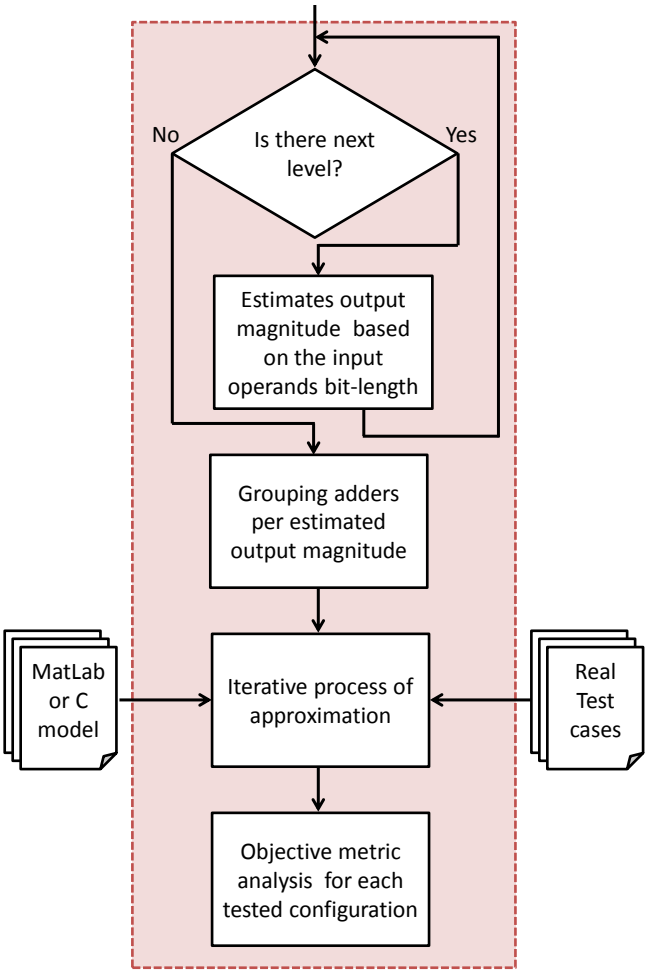

Figure 4.3 – Proposed heuristic based on the estimated output sum magnitude

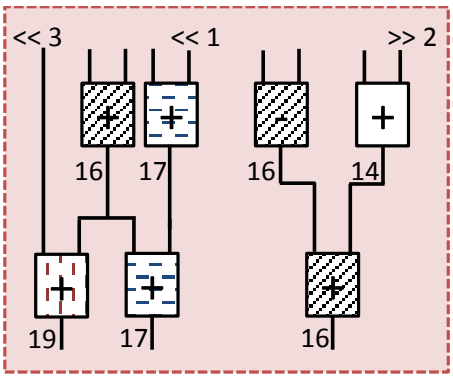

Figure 4.4 – Example of output sum magnitude estimation

The adders in the second level are evaluated considering the estimation performed in the first level and inputs from the original signal samples. In sum, for example in Figure 4.4 the four different groups are identified (*i.e.*, output sum magnitude estimation of 14, 16, 17,

and 19 bits). These groups are depicted in illustration with different types of textures. After estimation is done, the next steps are the same in Figure 4.2.

In this chapter two case studies are exercised based on the proposed heuristics. Even considering that these heuristics can be applied to any approximate adder previously mentioned in Chapter 2, only the power-oriented copy adder and ETAI were evaluated. This choice is because in trees with a substantial number of adders, the use of performance-oriented adders incurs higher power dissipation and area. This is also ratified in (REHMAN et al., 2016) which states that the adders focused on delay reduction cannot be used to explore power-efficiency in structures which demand the massive use of additions like the multipliers.

4.2 A case study on FIR filters for audio processing

The FIR filters are implemented in the transposed form topology, as presented in the example in Figure 4.5, where constant coefficients are multiplied by the same input. As mentioned in Chapter 3, the constant multiplications can be implemented by adopting the parallel structure of additions/subtractions and shift operations. Since the full implementation of multipliers is costly in hardware, the MMCM (Multiplier-less Multiple Constant Multiplication) problem is formulated as an optimization approach to find the minimum number of addition/subtraction operations which implement the constant multiplications. The FIR filters used in this study is obtained through the exact depth-first search algorithm developed and reported in (AKSOY et al., 2010). This algorithm finds the minimum solution through the use of lower and upper bound values of the search space for the MMCM problem instances.

Figure 4.5 depicts a hypothetical example of an optimized ten taps digital FIR filter regarding the MMCM operation of a transposed form filter whose taps are represented by 127x, 95x, -93x, -15x, 78x, 81x, 80x, 592x, 721x, and 129x. The input sample and output are denoted by x and y, respectively. The word registers and adders are represented as boxes. In this example, the parallel filter is designed with left shifters (<<), addition (+) and subtraction (-) operations. At the bottom of the figure, the filter delay line is depicted. This operational block realizes the sum of the partial terms and outputs one filtered sample (y) per clock cycle. The remaining part of the architecture is known as the partial terms computation or the adder tree hardware. The adder tree is previously optimized by the solver technique presented in (AKSOY et al., 2010).

In this study, five low-pass symmetric transposed FIR filters generated by the Remez algorithm are used. The number of taps ranges from 40 up to 120. Table 4.1 presents the filter specifications, where pass-band and stop-band denote the pass-band and stop-band frequencies normalized to the Nyquist frequency, respectively, and #taps column has the number of coefficients for each filter or taps.

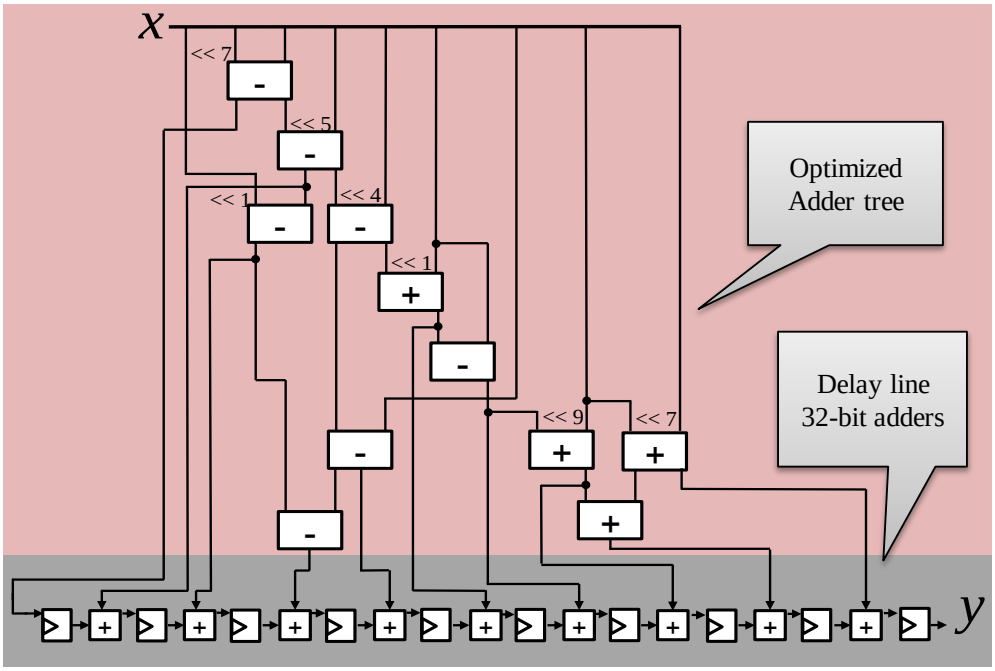

Figure 4.5 – A hypothetical example of ten taps transposed FIR filter implemented by the MMCM algorithm.

Table 4.1 – FIR filters specification

FIR Filter	*pass-band*	*stop-band*	*# taps*
1	0.10	0.20	40
2	0.10	0.15	60
3	0.12	0.18	80
4	0.10	0.12	100
5	0.24	0.25	120

For the case study of the FIR filters, the state-of-the-art copy adder and the ETAI are evaluated. The adopted heuristic is the one presented in subsection 4.1.1 which groups the

adders by considering the functionality. In Figure 4.5, two different functional blocks which contain adders are shown: i) the optimized adder tree for the partial terms computation, and ii) the delay line. The adder tree corresponds to the multiplication of the input samples by the coefficients. Higher approximation and error induction in this first block may substantially degrade the output quality when the approximation is additionally explored in the delay line. Therefore, all the adders inside this tree are grouped by the same approximation parameter *k*1, and *k*2 parameter approximates the adders from the delay line.

4.2.1 Results and discussion

Figure 4.6 and Figure 4.7 shows the average SNR (Signal to Noise Ratio) in dB per approximate FIR filter composed of the copy adder and ETAI, respectively. The SNR accounts for the noise plus distortion introduced by all approximated adders in the entire filter. The average SNR is calculated based on ten 16-bit audio signals sampled at 22.05 kHz with different genres (*e.g.*, rock, reggae, classical, jazz, blues, and so on) (TZANETAKIS; COOK, 2002). In the *x*-axis, the *k*1 and *k*2 parameters refer to the number of LSBs being approximated in the adder tree and delay line operational blocks, respectively. Both the *k*1 and *k*2 parameters are iterated from 0 to 10 to perform the simulation with the approximate FIR filters. This results in 121 combinations which were analyzed. In Figure 4.6 and Figure 4.7, the results are sorted by the approximate parameters *k1* and *k2* in ascending order. One can observe that the curves are similar, but the approximate FIR filters composed by the ETAI present a slightly lower SNR quality than the filters approximated by the copy adder.

After the evaluation of audio quality for the proposed approximate technique, different levels of SNR were selected to provide a multi-quality and energy profiling. Those target levels are: 50dB, 60dB, 70dB, and 80dB. The lower and upper bounds of 50 dB and 80 dB is related to a signal to noise plus distortion ratio which is relying on the 10^{-3} and 10^{-4} order, respectively. All these quality targets were also subjectively analyzed to observe the quality in filtered audio signals.

In Figure 4.6 and Figure 4.7 one can observe that many configurations are relying on the imposed SNR target levels. The cost function to search for the most approximated configuration of *k1* and *k2* which produces average SNR response nearest and above each target level is given as in (14).

$$\forall\, k1, k2 \in \mathbb{N}: 0 \leq k1 \leq 10 \text{ e } 0 \leq k2 \leq 10;\ \ w(k1, k2) = \max(k1 + 2k2) \qquad (14)$$

In equation (14) the term *w*(*k1*,*k2*) refers to the cost function to select the parameters *k1* and *k2* for each FIR filter under evaluation. Therefore, the chosen configuration is the one which maximizes *w*. The weights in (14) are determined by the proportion of adders in the adder tree and delay line regions. After analysis for each filter, the average proportion is twice the number of adders in delay line when compared to the adder trees.

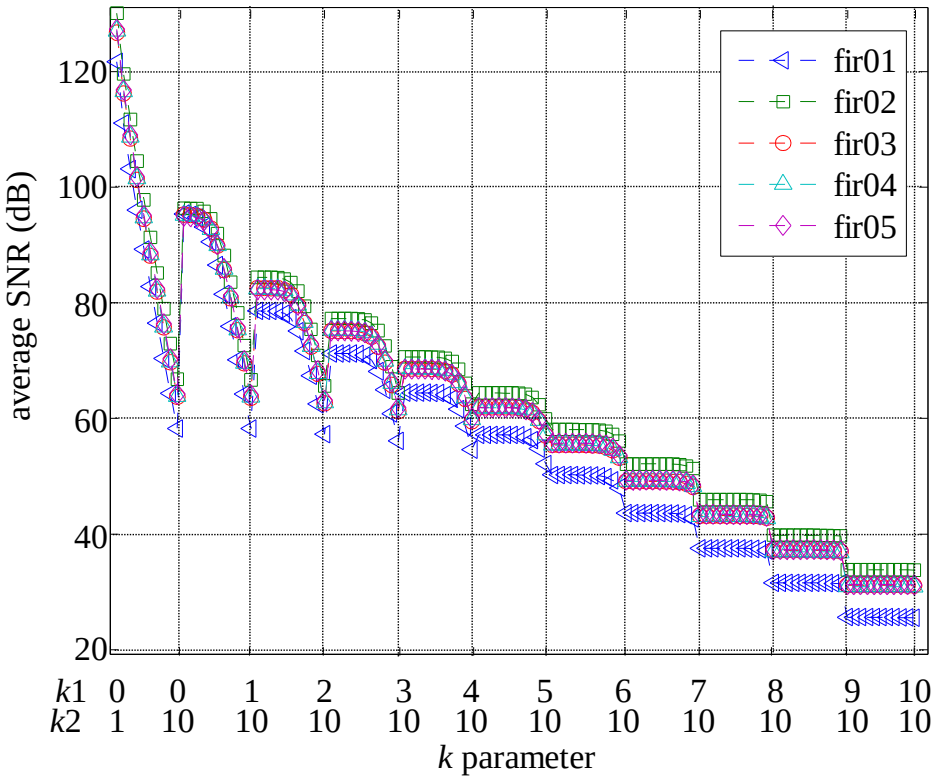

Figure 4.6 - Average SNR vs. *k* parameters combination regarding copy adder.

Table 4.2 and Table 4.3 show *k*1 and *k*2 parameters for each approximate FIR filter implemented by the copy adder and ETAI, respectively. The parameters are shown by each SNR target. As can be seen, the FIR filters implemented by the copy adder present higher *k*1 and *k2* approximation parameters than the ETAI for all the cases. This is because the copy adder may present lower degradation due to its superior performance in carry-in estimation for the precise block as mentioned in Chapter 2.

Furthermore, one can observe in Table 4.2 and Table 4.3 that the adders of the delay line are more approximated than the optimized adder trees. Two major observations can be made to explain this effect: i) the adder trees are already optimized for low power through the algorithm proposed by (AKSOY et al., 2010), and ii) the adopted cost function to search for

*k1*and *k2* gives priority to the delay line since it presents the higher number of adders for all the FIR filters under evaluation.

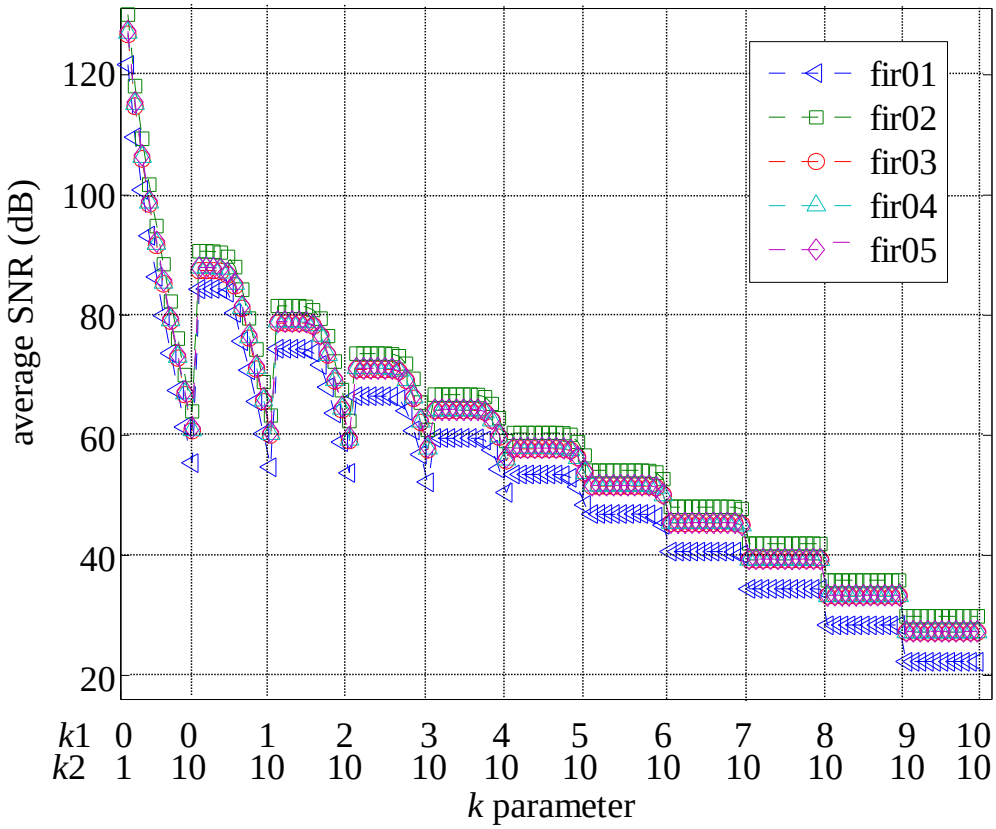

Figure 4.7 - Average SNR vs. *k* parameters combination regarding ETAI.

Table 4.2 – *k*1 and *k*2 parameters for FIR filters implemented by the copy adder.

FIR filter	**50dB**		**60dB**		**70dB**		**80dB**	
	*k*1	*k*2	*k*1	*k*2	*k*1	*k*2	*k*1	*k*2
1	5	10	3	9	1	8	1	6
2	7	10	4	10	2	9	1	7
3	6	10	3	10	2	8	1	7
4	6	10	3	10	2	8	1	7
5	6	10	4	10	1	9	1	7

Table 4.3 - *k*1 and *k*2 parameters for FIR filters implemented by the ETAI.

FIR filter	**50dB**		**60dB**		**70dB**		**80dB**	
	*k*1	*k*2	*k*1	*k*2	*k*1	*k*2	*k*1	*k*2
1	4	10	1	9	1	7	1	5
2	6	10	3	10	0	9	0	7
3	6	10	1	10	1	8	1	6
4	5	10	1	10	1	8	1	6
5	6	10	4	9	1	8	1	6

Additionally to the SNR evaluation, the THD+N (Total Harmonic Distortion plus Noise) analysis is also performed in this thesis for the precise as well as the approximate versions of the FIR filters regarding the 4 practiced SNR target levels. The evaluation is guided by the recommendations presented in the standard for digital audio measurement (AES, 1998). Therefore, 16-bit pure tone sinusoidal signals sampled at 44160 Hz were adopted to evaluate THD+N. The selected fundamental frequencies and amplitudes are shown in Table 4.4. One can observe that 12 different pure tones were exercised from 20 Hz up to 2000 Hz with steps projected in linear space. The upper boundary of 2000 Hz was selected because this is approximately the stop-band for most of the evaluated FIR filters. The amplitudes were determined considering the dB level relative to full scale (dBFS). According to the recommendations shown in (AES, 1998), the amplitudes should be -1dBFS and -20dBFS.

Table 4.4 – Experimental setup for THD+N evaluation

Criteria	**Selected values**
Fundamental frequency	20 Hz, 200 Hz, 380 Hz, 560 Hz, 740 Hz, 920 Hz, 1100 Hz, 1280 Hz, 1460 Hz, 1640 Hz, 1820 Hz, 2000 Hz
Amplitude	-1 dBFS and -20 dBFS

In Figure 4.8 to Figure 4.12 are shown the THD+N results considering the precise and all exercised approximate versions of the 5 FIR filters analyzed in this study. In these graphs,

each group of 12 columns represents the precise and four approximate FIR filters, while one column depicted in grayscale represents one tested frequency tone.

In Figure 4.8, the highest distortion in FIR filter #1 is shown in (b) for the copy adder version targeting 50dB SNR level when filtering signals with an amplitude of -20 dBFS. This maximum value is of 0.023% for the frequency of 920 Hz. All other approximate versions presented a much lower level of THD+N. However, the highest THD+N level of 0.023% is considered a low ratio between the artifacts and the tone signal. This is because the distortion plus noise amplitude is about 72.8 dB below the signal amplitude. Furthermore, one can observe that for the remaining configurations, the THD+N of the approximate versions are even lower than the degradation presented by the worst-case scenario.

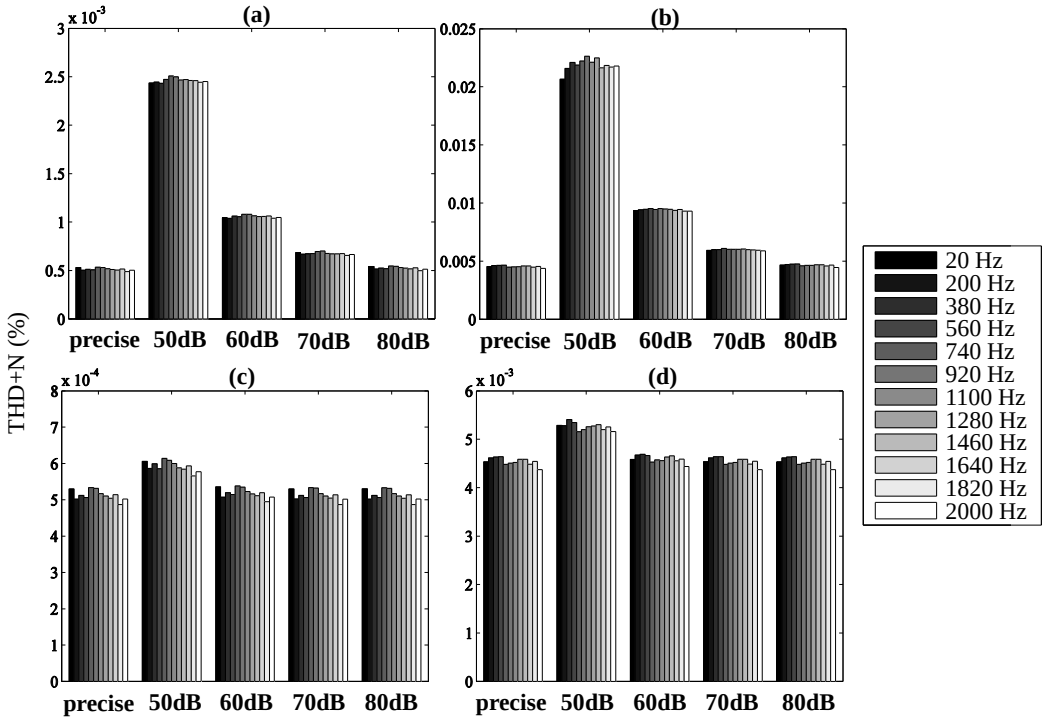

Figure 4.8 – THD+N results for FIR filter # 1. (a) precise plus Copy adder version at -1dBFS, (b) precise plus Copy adder version at -20 dBFS, (c) precise plus ETAI version at -1 dBFS, and (d) precise plus ETAI version at -20 dBFS.

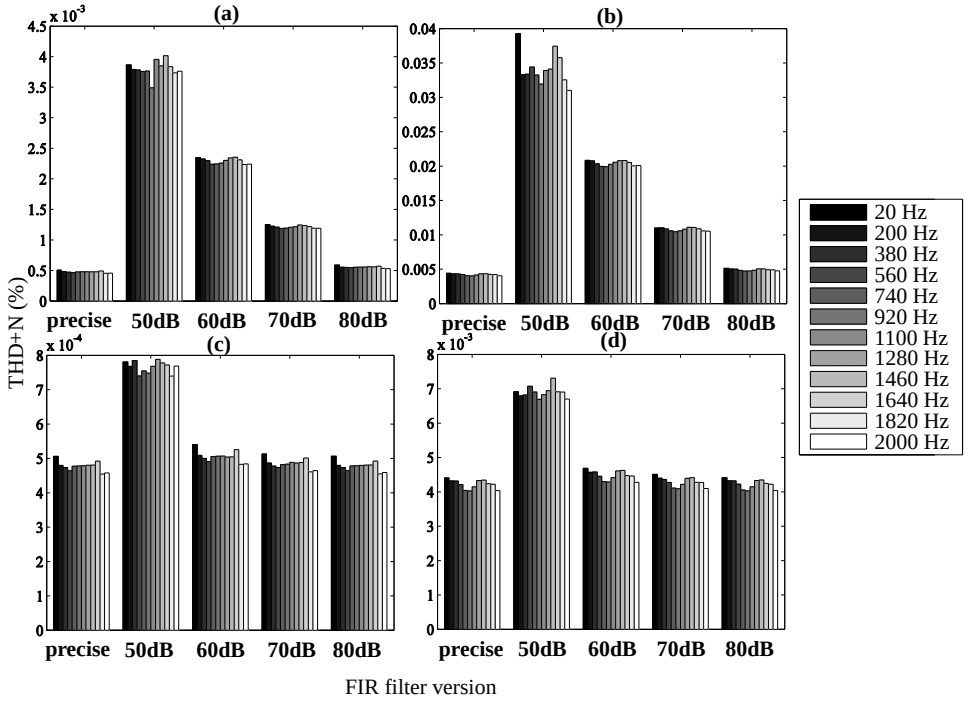

Figure 4.9 - THD+N results for FIR filter # 2. (a) Precise plus Copy adder version at -1dBFS, (b) Precise plus Copy adder version at -20 dBFS, (c) Precise plus ETAI version at -1 dBFS, and (d) Precise plus ETAI version at -20 dBFS.

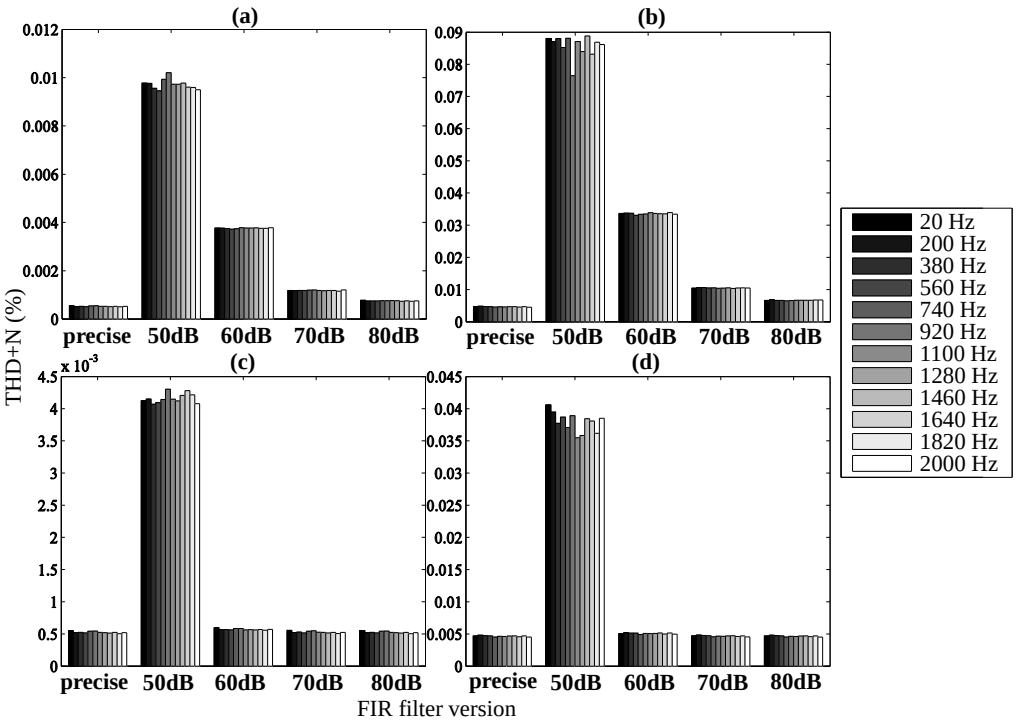

Figure 4.10 - THD+N results for FIR filter # 3. (a) Precise plus Copy adder version at -1dBFS, (b) Precise plus Copy adder version at -20 dBFS, (c) Precise plus ETAI version at -1 dBFS, and (d) Precise plus ETAI version at -20 dBFS.

In Figure 4.9 and Figure 4.10 the results for the FIR filters #2 and #3 are presented, respectively. In the former, the highest THD+N is of 0.04%, while in the latter is of about 0.09%. In other words, these numbers represent distortion plus noise levels of 67 dB and 60 dB below the signal level. For both the cases, the copy adder version with 50dB SNR response and amplitude of -20dBFS represents the worst scenario. One can notice that, for most of the approximate versions, the THD+N values are slightly higher than the precise results.

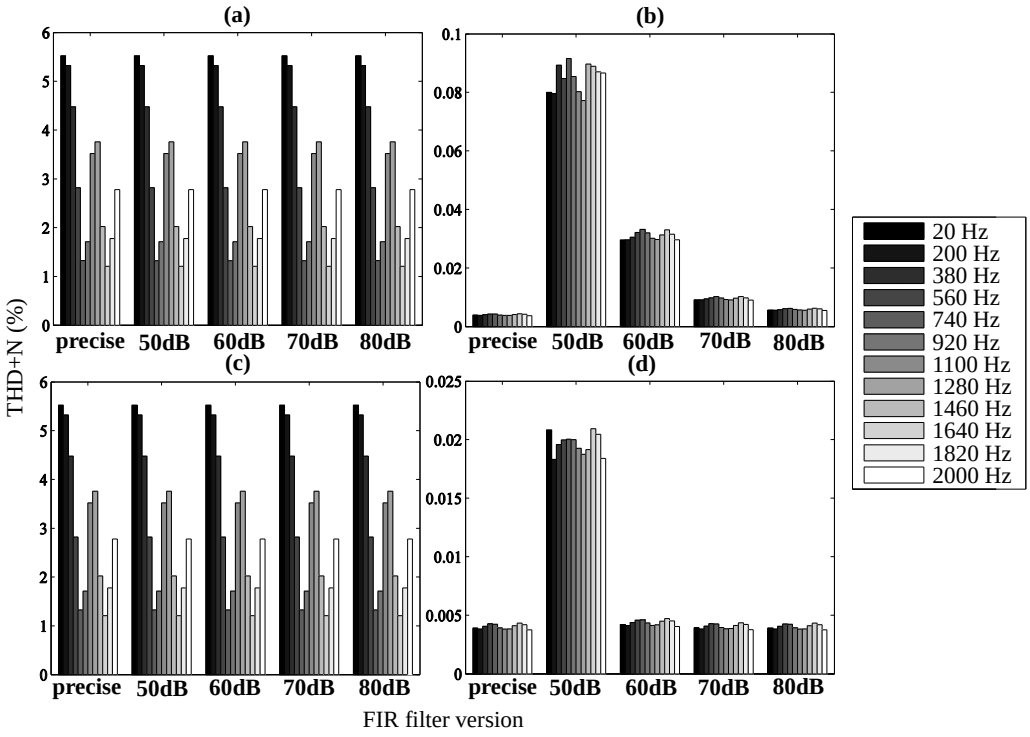

Figure 4.11 - THD+N results for FIR filter # 4. (a) Precise plus Copy adder version at -1dBFS, (b) Precise plus Copy adder version at -20 dBFS, (c) Precise plus ETAI version at -1 dBFS, and (d) Precise plus ETAI version at -20 dBFS.

In Figure 4.11 is presented the worst case when considered all the FIR filters under evaluation. One can observe that for the amplitude of -1dBFS, the output is substantially degraded for the FIR filter #4 in (a) and (c). Also, the maximum proportion of distortion is introduced by the precise filter, while the copy adder and ETAI approximate versions do not present higher considerable contribution regarding additive THD+N. After evaluation, it was observed that the FIR filter #4 presented the highest pass-band ripple when compared to the remaining FIR filters. This results in substantial THD+N of up to 5.5 % when adopting higher amplitude of -1dBFS. When considered the -20 dBFS in (b) and (d), the maximum THD+N level is about 0.9% or artifacts at 40 dB below the signal amplitude. In Figure 4.12 is shown

the second most substantial THD+N level of 0.13% for the FIR filter #5. This is experienced for the approximate filter implemented by copy adders with SNR target of 50 dB and signal amplitude of -20 dBFS. In sum, the THD+N analysis is important to observe the nonlinearities introduced by the filters' approximations and serves as an additional evaluation for the approximate FIR filters. In this scope, the ETAI versions present better THD+N results than the copy adder ones. This may be explained by the higher approximation practiced for the copy adders when compared to the ETAI. In filtered audio analysis driven by the SNR metric, it was observed that the implementation of the filter with copy adder enabled higher values for *k1* and *k2* than ETAI for the same SNR target. This superior level of approximation for the copy adders evidenced worse THD+N results when compared to the ETAI implementation. On the other hand, the THD+N values observed in the FIR filters approximated by the copy adders indicate that the degradation level is not more substantial than the pure tone signal amplitude. One can conclude that most of the approximate configurations do not present noticeable harmonic distortion plus noise when compared to the precise THD+N response.

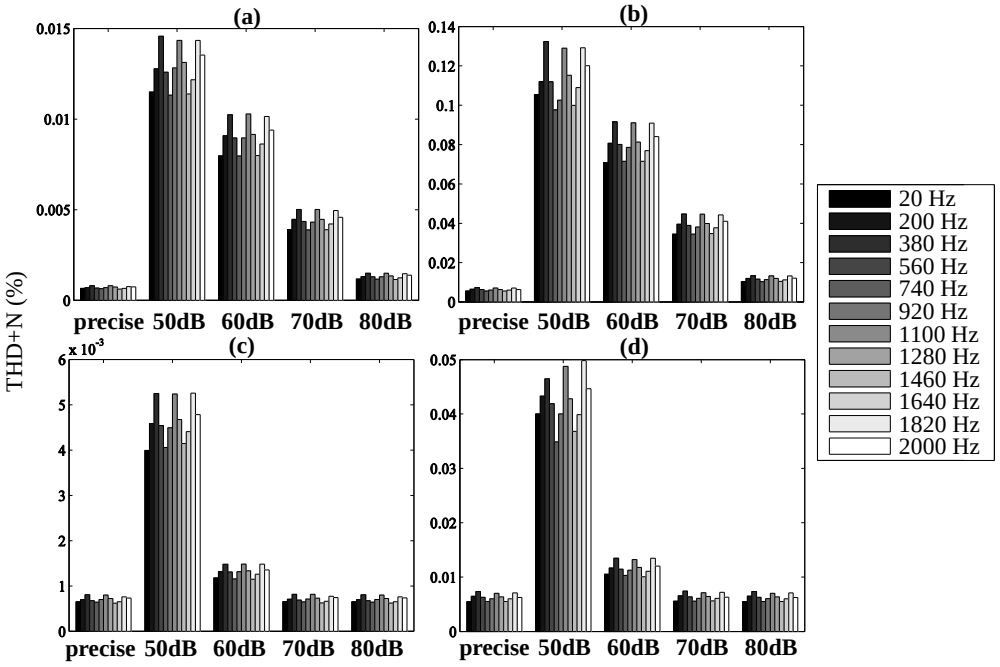

Figure 4.12 - THD+N results for FIR filter # 5. (a) Precise plus Copy adder version at -1dBFS, (b) Precise plus Copy adder version at -20 dBFS, (c) Precise plus ETAI version at -1 dBFS, and (d) Precise plus ETAI version at -20 dBFS.

Given that the proposed approximation strategy is fully evaluated regarding quality, the next step is referred to the FIR filters hardware synthesis. First, all the five FIR filters

approximated by the copy and ETAI (*i.e.*, regarding the four SNR levels for each FIR filter) plus the precise one were all described in VHDL (Very High Speed Integrated Circuits Hardware Description Language), for a total of forty-five designs. The synthesis procedure follows the method previously presented in Figure 4.1. An iterative process is used to synthesize all the designs, by executing RTL Compiler on all the VHDL sources. In the synthesis procedure, the designs are mapped onto a Nangate 45 nm Free PDK for CMOS VLSI (Very Large Scale Integration) implementation. The switching activity was extracted by simulating the designs with 10,000 samples from different audio files. This is performed to estimate power. The energy efficiency was obtained only by logic reduction, and there is no additional low power technique applied to the logic synthesis.

The energy reductions at 100 MHz for the approximate FIR filters composed by the copy adder and ETAI are shown in Figure 4.13 (a) and (b), respectively. For most of the cases in (a) and (b) the highest energy reduction is related to the FIR filters with SNR target of 50 dB. The only exception is for the ETAI version in (b), where the 60 dB FIR filter # 1 presents a slightly higher energy reduction than the 50 dB one. Another trend in the results accounts for the higher energy reduction when the number of taps increases. Based on that, the highest energy reduction among all the results is of 25.67% for the FIR filter # 5 with 120 taps, approximated by the use of copy adders, and targeting SNR response of 50 dB. On the other hand, the lowest energy reduction of 3.57% is experienced by the FIR filter #1 with 40 taps, approximated by the use of ETAI, and targeting SNR output quality of 80 dB.

The average energy reductions for each SNR target in Figure 4.13 (a) are of: i) 22.14% for the target of 50 dB, ii) 19.5% for the target of 60 dB, iii) 14.34% for the target of 70 dB, and iv) 9.52% for the target of 80 dB. The average energy reductions for each SNR target in Figure 4.13 (b) are of: i) 14.76% for the target of 50 dB, ii) 12.65% for the target of 60 dB, iii) 8.8% for the target of 70 dB and iv) 5.89% for the target of 80 dB. Based on that, one can observe that the copy adder present higher energy reduction per SNR level than the ETAI. This is because, during the SNR analysis, the copy adder achieved higher values of approximation when compared to the ETAI approach. In addition, the approximate part of a copy adder is implemented by a buffer while for the ETAI the half adder is adopted. The results for 100 MHz clock speed show that in an iso-performance scenario (*i.e.*, all at the same clock rate) the power dissipation is reduced when adopting approximate filters.

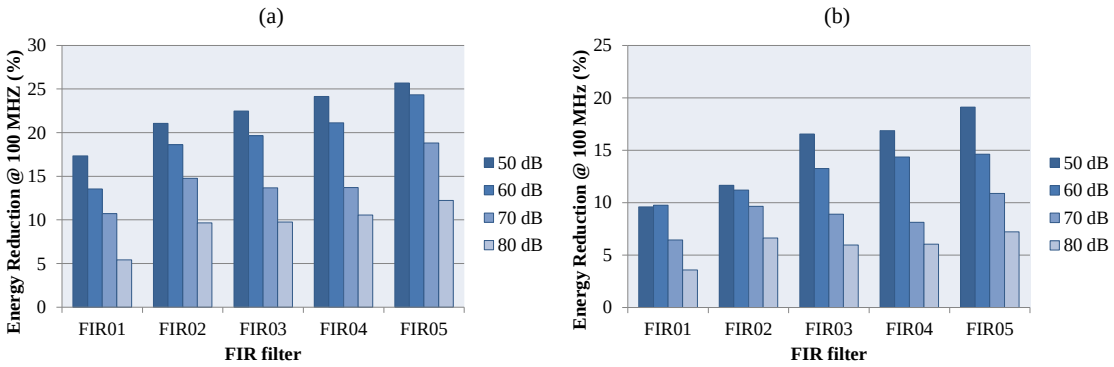

Figure 4.13 - Energy reductions at 100 MHz: (a) FIR filters approximated by the copy adders; (b) FIR filters approximated by the ETAI.

The second scenario of analysis is considering the average reductions in energy when a much lower computational performance target, at 10 MHz clock frequency, is required for the FIR filters. In Figure 4.14, the average energy per SNR response is shown for both the approximate versions of the FIR filters. In this scenario of lower clock speed, the FIR filter versions approximated by copy adders also present higher average energy reductions than the ETAI ones. One can observe that the copy adder filters achieve the highest energy reduction of 18.4% for the SNR target of 50 dB. As expected, these numbers tend to decay when objective quality increases. For the copy approximation, the lowest energy reduction of 6.94% is observed for SNR response of 80 dB. The maximum and minimum energy reductions for the ETAI version are 4.38 % and 0.69 %, respectively. These results apparently ratify that copy adder approximation for the FIR filters under evaluation are more energy efficient than the filters approximated by ETAI.

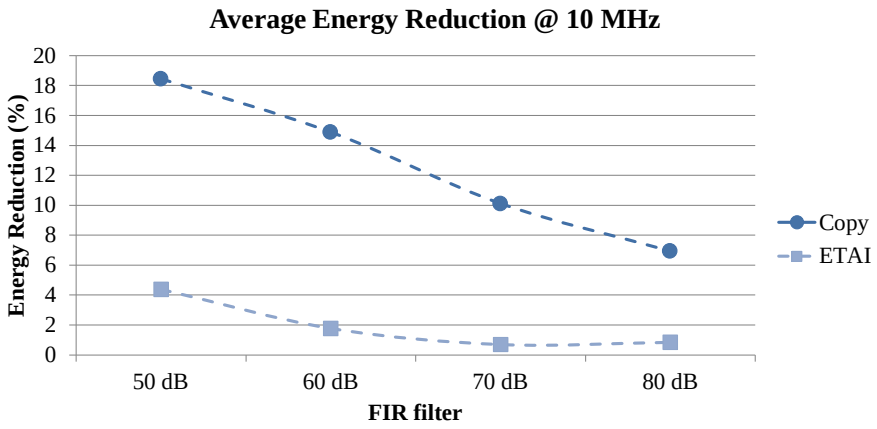

Figure 4.14 - Average energy and area reductions at 10 MHz regarding FIR filters composed of copy adder.

Based on the energy efficiency results, one can notice that those reductions in energy can be considered as additional savings to the MMCM low power optimization previously performed in the FIR filters benchmark. Therefore, the proposed heuristic enables further average energy reduction of up to 22.14 %.

4.3 A case study on Canny edge detector

In the Canny edge detection context, the heuristic based on the estimated output magnitude, presented in subsection 4.1.2, is exercised for the compute-intensive Gaussian and Gradient filters. According to the profile presented in Chapter 3, these filter modules have the largest count of arithmetic operations when compared to the remaining Canny edge steps. First, the proposed Canny edge detection architecture is presented in 4.3.1, while the heuristic evaluation is shown in 4.3.2. Finally, the results and discussion are presented in subsection 4.3.3.

4.3.1 State-of-the-art and proposed Canny edge architecture

4.3.1.1 State-of-the-art Canny edge architectures

In this thesis, a Canny edge detection architecture is proposed. Most of the related works which proposed accelerators for Canny edge detection is focused on hardware architectures implemented for FPGAs (Field Programmable Gate Arrays). In (RAO; VENKATESAN, 2004) and (NEOH; HAZANCHUK, 2004) high-level synthesis (HLS) tools are used to implement the Canny edge detection algorithm targeting FPGA synthesis. The former adopted the use of Handle-C language, while the latter used the DSP builder, which integrates Altera Quartus II® and high-level description in MatLab/Simulink®. Although HLS design brings many benefits to software designers, it still faces limitations and challenges regarding hardware synthesis as shown in (BAILEY, 2015).

The work presented in (HE; YUAN, 2008) shows a Canny edge detector implementation for FPGA, but the authors do not provide any details on how the steps of the edge detector were implemented. For example, there are no details of how the convolution operators from 5x5 Gaussian and 3x3 Gradient filters were designed. Moreover, there is no information if the implementation was done by using a Register Transfer Level (RTL) or HLS language input description. This same observation is applied to the work proposed in (GENTSOS et al., 2010), which presents details mainly about the necessary quantity of input pixels for each Canny edge detector step and the memory organization to store this content.

Moreover, there is no detailed explanation of the methodology adopted to describe the architecture (*i.e.*, if they used RTL or HLS hardware description).

On the other hand, works in (LI; JIANG; FAN, 2012; SANGEETHA; DEEPA, 2016; XU et al., 2014) present much more details about their methodology to implement the Canny edge architecture. In general, their design is focused on IPs (Intellectual Properties) provided by the FPGA vendors. The work in (POSSA et al., 2014) presents a comparison between two different implementations to accelerate the edge detection application: the use of Graphic Processing Units (GPUs), and the FPGA architecture. The authors in (POSSA et al., 2014) proposed an algorithm to provide support for multi-resolution images. They concluded that the FPGA solution provides a more energy-efficient design than the GPU implementation, but no comparison is given to ASICs (Application Specific Integrated Circuits). According to (KUON; ROSE, 2007) there is a gap between FPGA and ASIC implementations regarding area and energy efficiency. In general, the latter design flow presents better characteristic regarding the area, computational performance, and power dissipation. Based on that, logic synthesis results for the Canny edge detector targeting ASIC flow are presented in (LEE; TANG; PARK, 2016). Their proposed architecture is focused on data reuse inside the 3x3 Gaussian and Gradient filters. They consider approximation only for the magnitude operation.

4.3.1.2 The proposed architecture

This work proposes a new ASIC architecture implementation for the Canny edge detection algorithm to provide a solution which does not use IPs from FPGA vendors. The novelty is related to a more comprehensive exploration of approximate computing techniques in this application when compared to the state-of-the-art.

The proposed accelerator architecture processes row by row of any image size, and it has a throughput of one 8-bit pixel per clock cycle. In the proposed architecture there are data dependencies among the steps from Canny edge algorithm which need to be solved. Figure 4.15 shows the proposed data path for the accelerator architecture. From the output to the input direction (*i.e.*, from bottom to top in the illustration), one can observe that a 3x3 matrix containing 12-bit samples is needed for the non-maximum suppression and hysteresis thresholding block. This number of samples is needed because this operational block determines one edge or non-edge pixel by considering the central sample and its eight neighbors. Based on that, to produce the previously mentioned amount of samples for the non-maximum suppression and hysteresis thresholding block, three instances of the Gradient

and magnitude operations are necessary. Therefore, the three instances of the Gradient filter require a shift register structure with three register lines. Each register line contains five 8-bit registers. Based on that, five instances of the Gaussian filter are needed to feed this shift register structure. Finally, to produce 5 Gaussian filtered pixels per clock cycle, a shift register structure of 5 lines containing nine 8-bit registers are required.

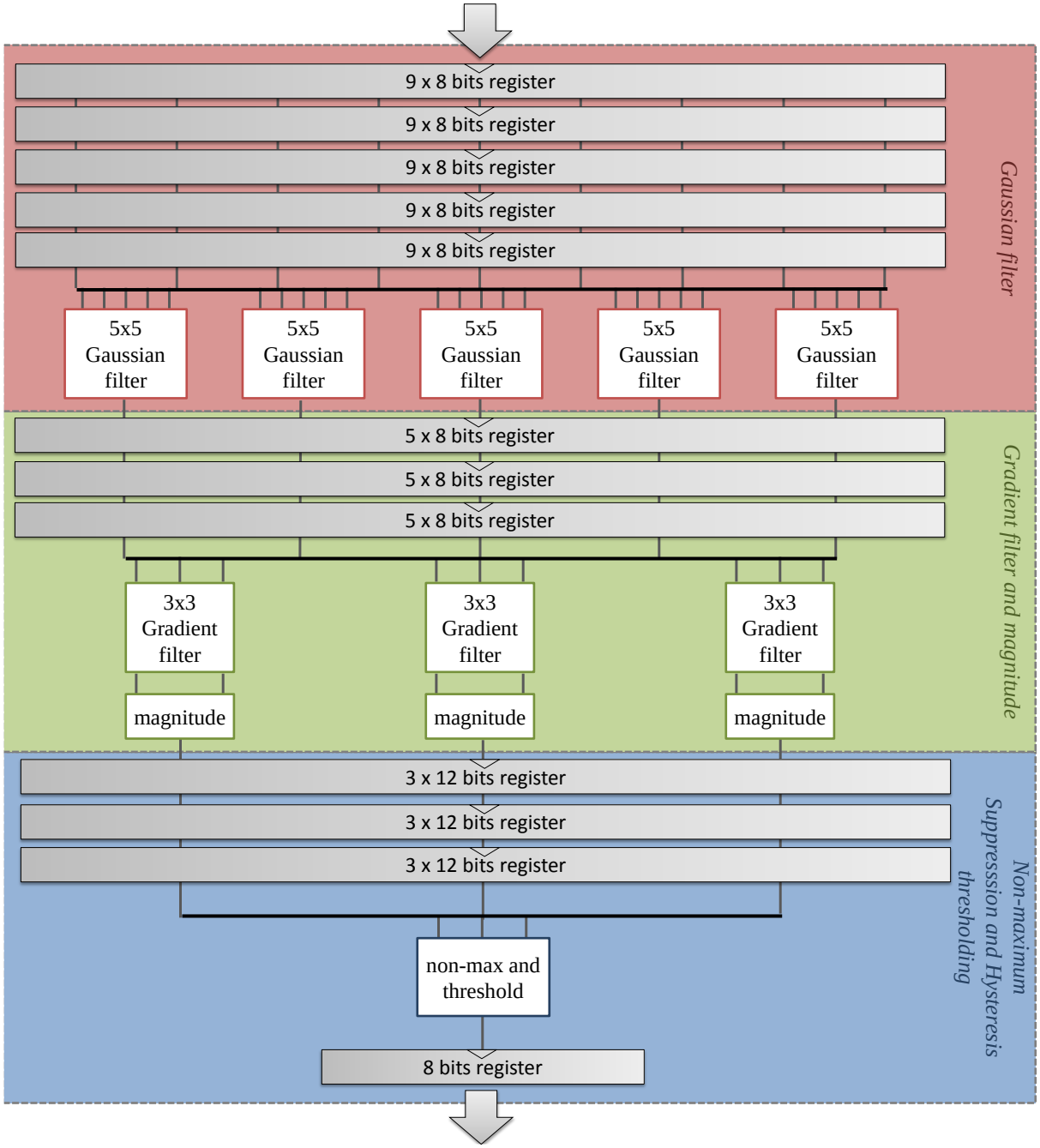

Figure 4.15 – Proposed datapath for the Canny edge detection architecture.

The 5x5 Gaussian and 3x3 Gradient design structures are shown in Figure 4.18 (a) and (b), respectively. This illustration is placed in next subsection to present the resultant

grouping scheme when considered the proposed heuristic to approximate these filters. One can observe that these filters are also implemented by shift operations and additions. The related work in (XU et al., 2014) uses two instances of 1D FIR filter IP from CoreGen in Xilinx platform. On the other hand, in (LI; JIANG; FAN, 2012) the convolution operator is designed by using parallel multipliers.

The conventional magnitude operation is shown in (15) and is adopted by the work in (XU et al., 2014). On the other hand, the hardware implementation in equation (15) is costly in terms of hardware, and most of the related works use alternative, approximate, and simpler techniques to implement the magnitude operation. Based on that, in (POSSA et al., 2014) the absolute sum of *x* and *y* derivatives, shown in (16), is adopted in their proposed architecture. In (SANGEETHA; DEEPA, 2016) an approximate implementation, shown in (17), is explored where the results tend to be better than the approach presented in (16). In (18) is shown a similar technique which was evaluated in (LEE; TANG; PARK, 2016).

$$mag = \sqrt{(dx)^2 + (dy)^2} \tag{15}$$

$$mag = |dx| + |dy| \tag{16}$$

$$\begin{gathered} mag = \max(0.875a + 0.5b, a) \\ where, a = \max(|dx|, |dy|)\ and\ b = \min(|dx|, |dy|) \end{gathered} \tag{17}$$

$$mag = \max(|dx|, |dy|, \frac{|dx|+|dy|}{\sqrt{2}}) \tag{18}$$

In this thesis, the selected magnitude operator is the one presented in (17). This is because its hardware implementation can be designed by using comparators, shift operations, and adders as presented in Figure 4.16. This implementation reduces the computational effort when compared to the baseline in (15) and does not present substantial error regarding Canny edge detection as further shown in the next subsection for the approximate Canny edge detectors results.

One can notice that the step related to the directional determination is not represented in Figure 4.15. This is because the proposed architecture for the directional determination is integrated to the non-maximum suppression operator. The traditional Canny edge detection algorithm uses the ratio between vertical derivative and horizontal derivative to determine the angle by computing the arctan function. After the angle is computed, the direction is determined considering only 4 direction angles (0°, 45°, 90°, and 135°), as shown in Figure 4.17. For example, to determine a given computed angle α between 0° and 45°, one can check

if α is lower, greater, or equal to 22.5°. When α is greater or equal to 22.5°, α will rely on the 45° direction. Otherwise, it will rely on the 0° direction. The same verification can be performed to decide the remaining directions.

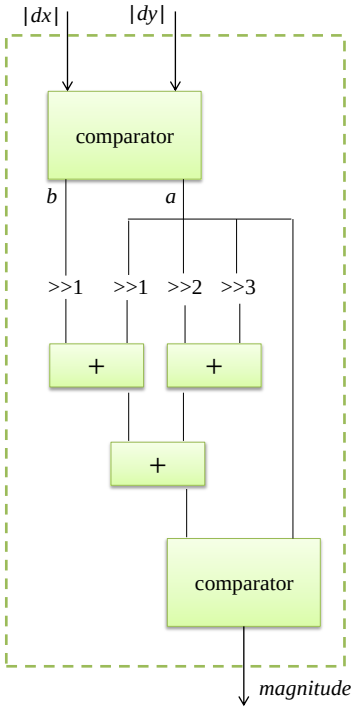

Figure 4.16 – Hardware implementation for the magnitude operator

For example, to decide between 45° and 90º, the difference is that the threshold angle is 67.5° instead of 22.5°. The proposed architecture uses only comparators, shifts, and adders to determine the direction. This is based on the following relation presented in (19).

$$\tan\theta = \frac{derivative\ y}{derivative\ x} \tag{19}$$

As earlier mentioned, considering the decision between 0° and 45° directions, the threshold angle must be 22.5°. Since the tan 22.5° is equal to 0.414214, there is a relation between the tan of 22.5° and the derivative *x* and derivative *y*, as shown in (20).

$$derivative\ y = 0.414214\ derivative\ x \tag{20}$$

Based on that, the directional determination for angles falling between 0º and 45º is reduced to the following verification in (21):

$$\begin{cases} 0° & if\ derivative\ y < 0.414214\ derivative\ x \\ 45° & otherwise \end{cases} \tag{21}$$

In other words, in (21), the determined direction is 0° if the derivative *y* is lower than 0.414214 derivative *x*. Otherwise, the direction is set to 45°. The same procedure would be performed considering the angle of 67.5º as being the threshold to determine the direction for the angle between 45º and 90º. On the other hand, it is possible to reuse the tan 22.5° instead of using the tan 67.5° to determine the 45° or 90° directions as shown in (22), (23), and (24).

$$\tan 67.5° = 2.414214 \tag{22}$$

$$derivative\ x = \frac{derivative\ y}{2.414214} \tag{23}$$

$$derivative\ x = 0.414214\ derivative\ y \tag{24}$$

Based on equations (22), (23), and (24), one can observe that the term $\frac{1}{\tan 67.5°} = \tan 22.5°$. This enables the direction determination between 45° and 90° by adopting tan 22.5° and isolating the derivative *x* instead of derivative *y*, as shown in (23) and (24). Therefore, the verification to be performed is shown in (25).

$$\begin{cases} 90° & if\ derivative\ x \leq 0.414214\ derivative\ y \\ 45° & otherwise \end{cases} \tag{25}$$

The previously mentioned verifications in (21) and (25) can be reused to determine the 135° direction. The only difference is that the signs of those derivatives have to be stored in order to decide whether the determined direction will be 135º or 45º. Equal signals (*i.e.*, ++ and --) and different ones (+- and -+) determine the 45º and 135º directions, respectively. In (LI; JIANG; FAN, 2012), the direction determination is performed by considering both the tan 22.5° and the tan 67.5°, by implementing approximations regarding the use of shifts and adders. The value 0.414214 is implemented using the term $\frac{1}{2} - \frac{1}{16}$, while the value 2.414214 is implemented using the term $2 + \frac{1}{8}$. Those approximations result in errors of 1.13 and -2.7 degrees, respectively. As a novel contribution, this thesis proposes the directional determination for the four different directions by using only the tan 22.5°. Therefore, it is needed to implement only the multiplicative constant equal to 0.414214. Instead of using the same term proposed by (LI; JIANG; FAN, 2012), this work proposes a more accurate term: $\frac{1}{2} - \frac{1}{16} - \frac{1}{64} - \frac{1}{128}$. This term results in an angle of 22.4926° with a lower error of -0.0074°. In (XU et al., 2014) the directional determination is performed by using one divider and two multipliers from Xilinx CoreGen. In (SANGEETHA; DEEPA, 2016) the directional determination is performed by a CORDIC block, which implements the tangent function. One

can observe that, in this work, the proposed solution has only three adders, four shifts, and three comparators.

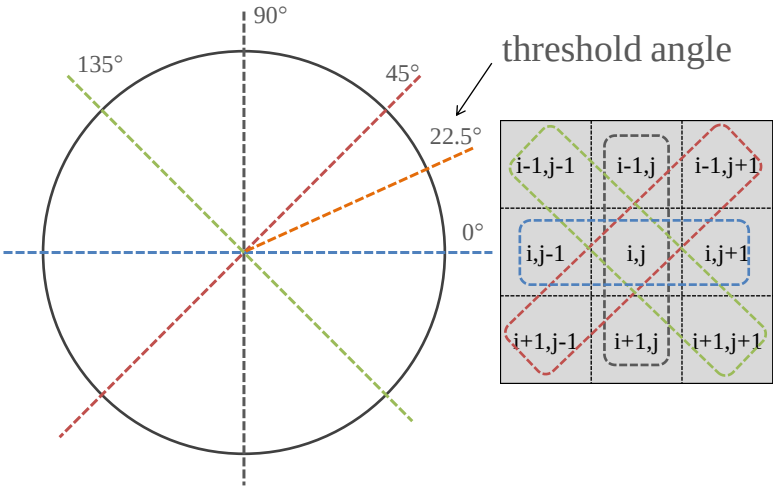

Figure 4.17 – Directional determination for non-maximum suppression.

Once the directional determination is performed, the i,j central magnitude needs to be compared with the two neighbor samples by considering the selected direction. For example, if the determined direction is 0°, then the neighbors at position i,j-1 and i,j+1 are compared to the central magnitude pixel. If this central pixel i,j is not the maximum value when compared to its neighbors, then it is suppressed or set to zero. Otherwise, this sample is maintained. The possible configurations to be tested according to the direction determination are shown in Figure 4.17.

The last step in the Canny edge detector is to perform hysteresis thresholding. As can be seen in Figure 4.15, this last operational block is integrated inside the non-maximum suppression. This can also be considered as an approximation because the hysteresis thresholding originally is performed when the non-maximum suppression block verified all the samples. On the other hand, in the next subsection is shown that this approach does not substantially degrade the Canny edge performance in the approximate configurations. This work adopts the use of comparators to determine whether the output image pixel will be “0” or “255”. The threshold operator only evaluates the magnitude samples which were not suppressed to zero. In the proposed architecture the threshold input is provided to allow dynamic threshold settings to turn the edge detection more or less sensitive.

In this operator, the threshold input value is multiplied by lower and upper constant boundaries (*i.e.*, TLow and THigh). After simulations, the values of 0.075 and 0.175 for TLow and Thigh are adopted, respectively. These are the same values adopted in the Canny

function from MatLab. The TLow and THigh parameters are approximated in hardware implementation by $\frac{th}{16}+\frac{th}{128}+\frac{th}{256}$ and $\frac{th}{4}-\frac{th}{16}+\frac{th}{128}+\frac{th}{256}$ which are equal to 0.074 and 0.176, respectively. The *th* indicates the threshold input value. The way that the thresholding operator works is shown in (26).

$$\begin{cases} 255 & if\ s > THigh \\ 0\ or\ 255 & if\ TLow \leq\ s\ \leq THigh \\ 0 & otherwise \end{cases} \quad (26)$$

In (26) *s* refers to the i,j magnitude pixel which is not suppressed in non-maximum suppression operation. Therefore, *s* is an edge when the magnitude is higher than the THigh, and a non-edge when is lower than TLow. For the case where *s* is between TLow and THigh, additional verification is performed: If one of the eight neighbor samples of *s* is higher than THigh, then *s* is set to "255". Otherwise, *s* is set to "0".

In Figure 4.15, the zero padding technique is considered for the Gaussian and Gradient filters. Therefore, only three cycles are needed to start the Gaussian filter operation, while five cycles are needed for the Gradient filters. Based on that, this work proposed a finite state machine which enables Gaussian and Gradient operation only when the required number of cycles is achieved. The initial latency in the proposed architecture is of 9 cycles to process each image row.

4.3.2 Heuristic evaluation

The heuristic proposed in the subsection 4.2.1 was adopted to explore power-performance-quality profiles for the Canny edge application. In Figure 4.18 (a) and (b), the adders at the bottom of the architectures and the subtractors are not grouped by the proposed heuristic. This is because it is possible to determine the number of adders desired to be approximated. In Figure 4.18 (a), the ungrouped adders correspond to the division by 159 and the subtractor operation. This is performed to attenuate degradation for these operations. In Figure 4.18 (b), the subtractors are not approximated to avoid or reduce the possibility of the incorrect sign for the derivatives. One can observe that in (a) and (b), the proposed heuristic results in three and two different groups of approximate adders, respectively.

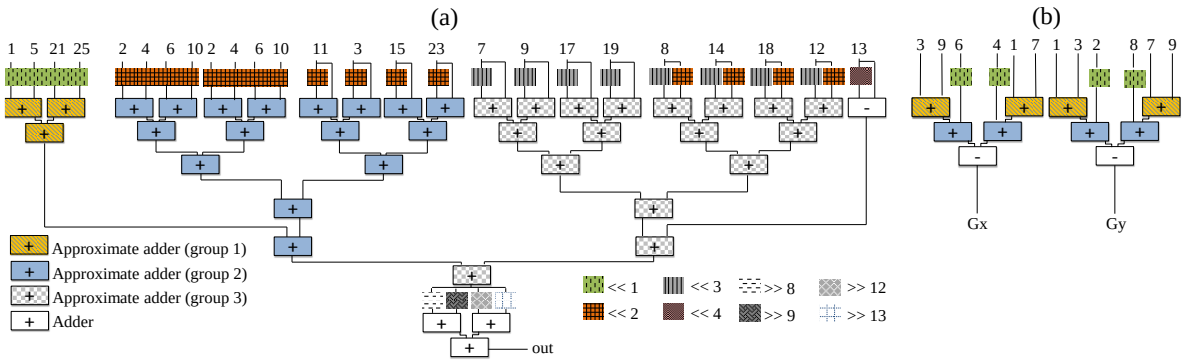

Figure 4.18 – Grouping the approximated adders. (a) 5x5 Gaussian filter. (b) 3x3 Gradient filter.

Once all the adders are grouped, the simulation can be performed with independent iteration ranges per group. For example, in Figure 4.18 (a), these iteration ranges are the following: i) 1 to 5 for the group 1, ii) 1 to 6 for the group 2, and iii) 1 to 7 for the group 3. This is feasible since lower groups are related to adders which possibly have lower input operands magnitude. In other words, there is no need to test higher *k* approximation parameters for these adders because the error magnitude would substantially degrade the application output quality.

During the iterations, all adders which belong to a specific group are uniformly approximated. This procedure substantially reduces the simulation time and makes feasible the process of searching for approximation in a set of adders. For this example, 210 different configurations are explored. Since the Gaussian filter is used to blur the image and to remove noise, the PSNR metric is used to evaluate the degradation introduced by the approximation in the filter. The PSNR is calculated for the approximations by considering the precise filtered image as the reference.

Eight grayscale images with 8-bit pixels are used to perform the simulation. From this database, seven images are genuine from MatLab, and the remaining image is the “Lena” benchmark. Figure 4.19 shows the average PSNR vs. configurations under evaluation regarding copy adder and ETAI approximations in the 5x5 Gaussian filter. Differently from the Figure 4.6 and Figure 4.7, the average PSNR results shown in Figure 4.19 are sorted in ascending order. One can observe that the copy adder presents better or similar average PSNR results than the ETAI for each approximate configuration (*i.e.*, the combination of *k1*, *k2*, and *k3*).

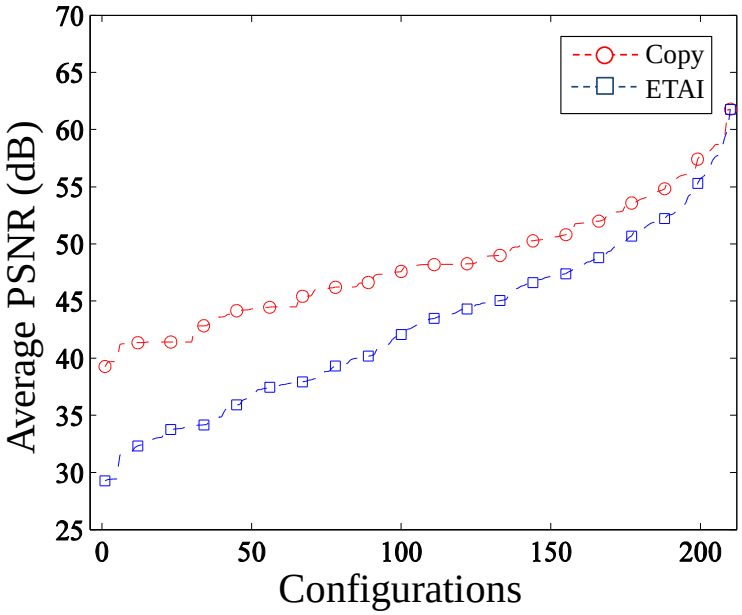

Figure 4.19 - Configurations for approximate 5x5 Gaussian filter vs. Average PSNR.

In (PARK; CHOI; ROY, 2010) and (HE; GERSTLAUER; ORSHANSKY, 2011) is stated that the 30 dB boundary may correspond to images with "good enough" output quality. On the other hand, the worst PSNR result regarding the approximation with copy adder version is about 40 dB. Therefore, in this work, the lowest selected level is 40 dB, and the highest one is of 50 dB. The lowest level of 40 dB is chosen because the 30 dB boundary may result in substantial degradation at the edge detection performance. Also, the level of 50 dB is selected to observe the impact of higher quality Gaussian filtered images on the edge detection response. The cost function in (27) is adopted to determine the three approximate parameters *k1*, *k2*, and *k3*. Following the same idea of the FIR filters, this cost function *w*(*k1*,*k2*,*k3*) is driven by the proportion in the number of adders per group.

$$\forall\, k1, k2, k3 \in \mathbb{N}: 1 \leq k1 \leq 5;\ 1 \leq k2 \leq 6; 1 \leq k3 \leq 7$$
$$w(k1, k2, k3) = \max(k1 + 5k2 + 6k3) \tag{27}$$

Therefore, the selected approximate configuration is the one which respects the PSNR targets and maximizes the cost function *w*. In Table 4.5 it is possible to observe the parameterization for the approximate Gaussian filters. One can notice that the copy adder version resulted in higher approximation than the Gaussian filter approximated by the ETAI. This can be explained due to the superior performance of copy adder for the 5x5 Gaussian filter case study as presented in Figure 4.19.

Table 4.5 – Approximate 5x5 Gaussian filter parameterization.

	40 dB	**50 dB**
Copy adder version	*k1 = 5* ; *k2* = 6 ; *k3* = 6	*k1 = 5* ; *k2* = 4 ; *k3* = 3
ETAI version	*k1 = 4* ; *k2* = 5 ; *k3* = 4	*k1 = 3* ; *k2* = 3 ; *k3* = 2

Once the *k* parameterization is selected for the Gaussian filter, the next task is to simulate the Gradient filter to evaluate the quality of edge detection for each one of the 4 selected configurations shown in Table 4.5. The iteration ranges for the two groups of adders in Figure 4.18 (b) are as follows: i) group 1 adders ranging from 1 to 5, and ii) group 2 adders ranging from 1 to 5. After the Gradient filter is performed for each approximation under evaluation, the output image containing the detected edges is obtained by running the remaining steps from Canny edge detection algorithm. The precise Canny edge detected output image is considered the reference to measure the approximate versions.

Since the edges are generally represented as white pixels (*i.e.*, 255), and the non-edge information as black pixel (*i.e.,* 0), the differences between precise and approximate images may indicate the following four possible responses: i) true positives, where the approximate version detects an edge pixel which is also detected by the precise one; ii) false negatives, where the approximate version does not detect an edge pixel which is detected by the precise one; iii) false positives, where the approximate version detects an edge pixel which is not detected by the precise one and iv) true negatives, where the approximate version does not detect an edge pixel which is also not detected by the precise one. Based on that, two indicators are used to evaluate the performance of a given edge detector: recall in (28) and precision in (29).

$$recall = \frac{tp}{tp + fn} \tag{28}$$

$$precision = \frac{tp}{tp + fp} \tag{29}$$

In (28) and (29) the terms *tp*, *fp*, and *fn* refers to true positives, false positives, and false negatives, respectively. The *recall* could be understood as the ratio between the number of edge pixels which are correctly detected by the approximate version and the number of edge pixels that should be correctly detected (*i.e.,* edges detected by the precise Canny edge detector). The *precision* could be understood as the ratio between the number of pixels

correctly detected as edges and the total number of pixels which are detected as an edge by the approximate version. The performance metric shown in (30) is mostly used to evaluate the edge detection (LEE; TANG; PARK, 2016).

$$performance(\%) = \min(recall, precision)\,100 \tag{30}$$

In (30) the minimum between the *recall* and *precision* is related to the performance. One can conclude that performance is the worst result between the *recall* and *precision*. This is the adopted metric to select the approximation parameters for the Gradient filter. The performance targets are 80% and 90%. The latter is used to compare with the maximum performance presented by the architecture proposed in (LEE; TANG; PARK, 2016), while the former is an additional evaluation for the scenario when the performance can be lower. In this scope, the cost function *w*(*k1*,*k2*) to search for the approximation parameters is presented in (31).

$$\forall\, k1, k2 \in \mathbb{N}: 1 \leq k1 \leq 5\ and\ 1 \leq k2 \leq 5;\ \ w(k1, k2) = \max(k1 + k2) \tag{31}$$

One can observe that in (31) the proportion of adders in group1 and group2 are the same. Therefore, the equal weight is adopted for both the terms *k1* and *k2* in the Gradient filter. Table 4.6 presents the selected *k1* and *k2* parameters. In two cases the maximum performance achieved is below the 90% target. This occurs for the copy adder and ETAI approximate versions with 40 dB PSNR response when considered the Gaussian filter. This ratifies that Gaussian filtering aiming at 30 dB PSNR response may degrade the edge detection. For the PSNR of 50 dB, the performance which is nearest and above 90% are 90.9% and 92.7% for the copy and ETAI approximate Canny edge detection, respectively. For the Gradient filters, the ETAI achieved higher approximation parameters and less degradation for the performance metric.

Table 4.6 – Approximate Gradient filter parameterization

	90% or the highest	**80%**
Copy with PSNR of 40 dB	*k1* = 1 and *k2* = 1 (81.9%)	*k1* = 1 and *k2* = 2
Copy with PSNR of 50 dB	*k1* = 1 and *k2* = 1 (90.9%)	*k1* = 2 and *k2* = 3
ETAI with PSNR of 40 dB	*k1* = 1 and *k2* = 1 (89.9%)	*k1* = 3 and *k2* = 3
ETAI with PSNR of 50 dB	*k1* = 1 and *k2* = 1 (92.7%)	*k1* = 3 and *k2* = 3

This may be related to the high-pass characteristic of this filter which differentiates this design from the low-pass Gaussian and FIR filters previously explored.

4.3.3 Results and discussion

This section presents application quality evaluation for the BSD (Berkeley Segmentation Dataset) (MARTIN et al., 2001) benchmark making a comparison among the precise architecture and the approximate versions. Furthermore, the comparison is also performed with the ground truth images. The ground truth is the image segmented by human observers. After the quality evaluation, energy efficiency results are shown for the approximate Canny edge detectors followed by discussion and comparison with the state-of-the-art.

4.3.3.1 Edge Detection Results

In order to evaluate the output quality of the detected edges eight different approximate configurations are considered as follows: i) The Gaussian and Gradient filters implemented with copy adders considering average PSNR and performance responses of 40 dB and 81.9%, respectively; ii) The Gaussian and Gradient filters implemented with copy adders considering average PSNR and performance responses of 50 dB and 90.9%, respectively; iii) The Gaussian and Gradient filters implemented with copy adders considering average PSNR and performance responses of 40 dB and 80%, respectively; iv) The Gaussian and Gradient filters implemented with copy adders considering average PSNR and performance responses of 50 dB and 80%, respectively; v) The Gaussian and Gradient filters implemented with ETAI considering average PSNR and performance responses of 40 dB and 89.9%, respectively; vi) The Gaussian and Gradient filters implemented with ETAI considering average PSNR and performance responses of 50 dB and 92.7%, respectively; vii) The Gaussian and Gradient filters implemented with ETAI considering average PSNR and performance responses of 40 dB and 80%, respectively; and viii) The Gaussian and Gradient filters implemented with ETAI considering average PSNR and performance responses of 50 dB and 80%, respectively. For convenience, these approximate configurations for Canny edge detection are identified by the name of the approximate adder followed by the average PSNR response of the Gaussian image filter and the average performance response after Gradient filter. For example, the Canny edge detector approximated by the copy adder targeting average PSNR response of 40dB for the Gaussian filter and average performance of 80% after Gradient filtering is identified as Copy 40dB 80%.

For all the quality evaluations, 16 images from the BSD benchmark were selected. The set of chosen images presents different levels of edges. There are images with a low count of edges and others with a higher density of edges and details. The first analysis considers the performance metric in (30), and the reference image is the one processed by the precise version of the Canny edge detector. The average performance results are shown in Figure 4.20 for each approximate configuration. In Figure 4.20 (a) is exercised the original benchmark. Both the copy and ETAI approximations reach a maximum average response of 95%. For the BSD benchmark, all the configurations achieve a higher result than the selected performance target. This may be explained due to the observation that the benchmark is different from the one used to approximate the Gaussian and Gradient filters.

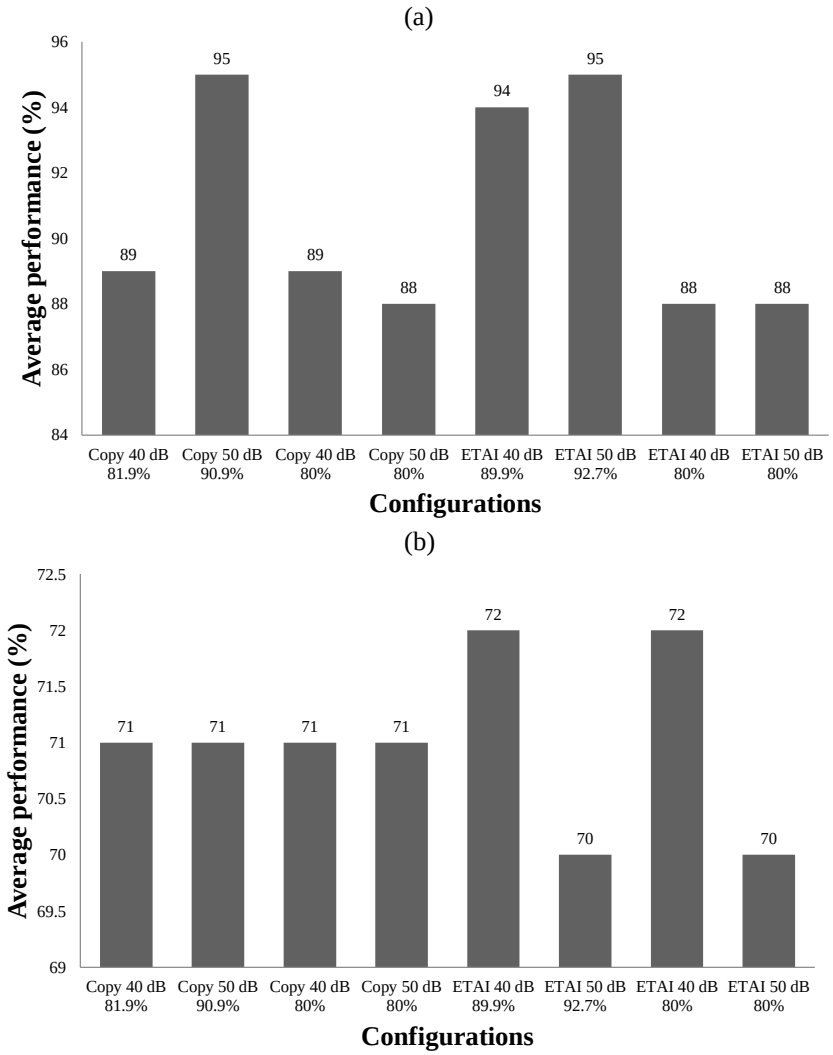

Figure 4.20 – Performance results vs. approximate configurations. (a) Normal benchmark. (b) Noisy benchmark (Gaussian noise with $\sigma^2 = 0.01$).

The average performance for the configurations is around 90.8%. The results for the noisy benchmark corrupted by additive Gaussian noise with variance $\sigma^2 = 0.01$ are shown in Figure 4.20 (b). The variance for the Gaussian noise is the same practiced in (LEE; TANG; PARK, 2016). One can observe that the noise decreases the average performance of all the configurations from 70% up to 72%. The average response for all configurations is around 71%.

An alternative analysis is considered in this thesis taking into account the ground truth images as being the reference. The ground truth database is acquired through segmentation performed by human observers, and it is used in the computer vision area to measure the performance of a given edge detection algorithm (MARTIN et al., 2001). When analyzing the performance metric in this scope, all the approximate configurations achieve the performance of 21%. For the precise Canny edge detector, this value is 22.2%. One can conclude that there is no substantial difference regarding objective metric when the ground truth image is the reference for the selected images from the BSD benchmark. When considering the simpler technique of Sobel for edge detection, the performance is of 14.5%. This clearer show that even the approximate versions of Canny edge detection present better quality results than the Sobel filter.

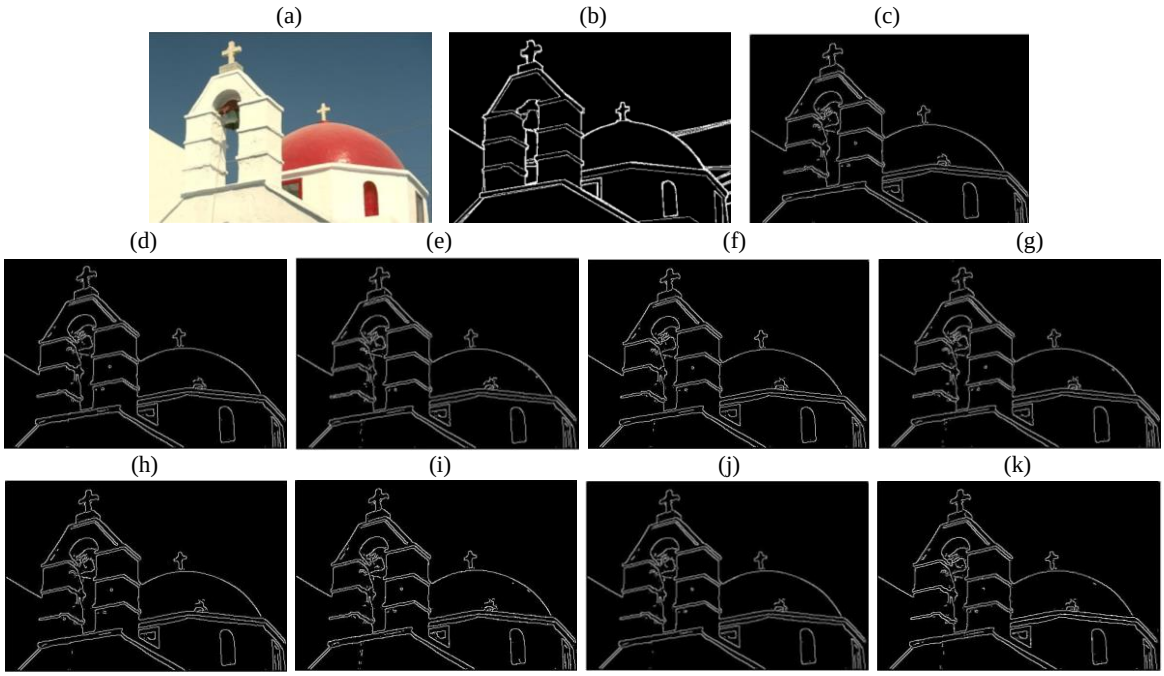

Figure 4.21 - Edge detection subjective analysis. (a) Image from BSD database. (b) Ground truth. (c) Precise Canny edge detector. (d) Copy 40 dB 81.9% (e) Copy 40 dB 80%. (f) Copy 50 dB 90.9%. (g) Copy 50 dB 80%. (h) ETAI 40 dB 89.9%. (i) ETAI 40 dB 80%. (j) ETAI 50 dB 92.7%. (k) ETAI 50 dB 80%.

Furthermore, the subjective analysis is presented in Figure 4.21. This analysis confirms the objective results. In Figure 4.21 (c) the precise Canny edge image output is similar to the approximate configurations in (d) to (k). The low proportion of artifacts is produced in the approximate versions when compared to the precise solution. This enables the use of approximate computing in edge detection when low proportion of artifacts can be accepted in application level. This confirms the findings in (KHUDIA et al., 2016), where the authors state that computer vision applications can leverage the condition of error tolerance to provide energy-efficient hardware accelerators.

4.3.3.2 Energy Efficiency Results

The precise and approximate designed architectures were fully described in VHDL. The designs were synthesized by using the Cadence RTL Compiler tool and mapped onto 45 nm Nangate Open Cell Library by considering nominal power supply of 1.1 V. To estimate the maximum frequency, many timing syntheses were performed by using the bisection search method. This analysis is essential to determine the maximum throughput for each design.

Table 4.7 – Maximum frequency and frame rates achieved by each design.

design	**Maximum frequency (MHz)**	**Supported frame rate (fps) HD image (1280 x 720)**	**Supported frame rate (fps) FHD image (1920 x 1080)**	**Supported frame rate (fps) QHD image (2560x 1440)**	**Supported frame rate (fps) UHD image (3840 x 2160)**
Precise	577.8	622.5	277.3	156.7	69.6
Copy 40 dB 81.9%	588.8	634.4	282.6	159.7	70.9
Copy 40 dB 80%	591.1	636.9	283.7	160.3	71.2
Copy 50 dB 90.9%	587.3	632.8	281.9	159.3	70.8
Copy 50 dB 80%	590.8	636.5	283.5	160.2	71.2
ETAI 40 dB 89.9%	614.5	662.1	294.9	166.6	74.0

ETAI 40 dB 80%	585.8	631.1	281.1	158.9	70.6
ETAI 50 dB 92.7%	610.3	657.5	292.9	165.5	73.5
ETAI 50 dB 80%	594.3	640.3	285.2	161.2	71.6

For power results, the switching activity was considered through simulation of 5,000 pixels from real images.

In Table 4.7 is shown the maximum frequencies and frame rates per video resolution by considering the designs under evaluation. The approximate configurations present higher speed than the precise approach. On the other hand, even the precise architecture achieves real-time operation with a frame rate higher than 60 frames per second (fps) when operating at the maximum frequency and considering different video resolutions.

Table 4.8 - Energy efficiency analysis for Canny edge detectors @ 300 MHz

	precise	**Copy 40dB 81.9%**	**Copy 40dB 80%**	**Copy 50dB 90.9%**	**Copy 50dB 80%**	**ETAI 40dB 89.9%**	**ETAI 40dB 80%**	**ETAI 50dB 92.7%**	**ETAI 50 dB 80%**
# of cells	10194	6542	6519	8361	8442	7950	8010	9422	9482
Area (EG)	21814.5	14825.8	14870.9	18162.9	18388.4	17290.7	17576.4	20184.2	20469.9
Dynamic Power (mW)	7.2	2.9	2.9	5.1	5.1	4.3	4.3	6.1	6.1
Total Power (mW)	7.5	3.2	3.2	5.3	5.3	4.5	4.6	6.4	6.4
MEF HD (µJ)	23.2	9.9	9.9	16.6	16.7	14.1	14.2	18.8	18.9
MEF FHD (µJ)	52.2	22.2	22.3	37.4	37.5	31.8	31.9	44.5	44.6
MEF QHD	92.8	39.5	39.5	66.4	66.5	56.5	56.6	79.1	79.2

(µJ)									
MEF UHD (µJ)	208.6	88.9	88.9	149.3	149.5	127.1	127.2	177.9	178

When considered the target frame rate of 30 frames per second, one can observe that the operation at maximum clock frequency may represent higher performance and energy consumption. Based on that, the iso-performance analysis is presented in Table 4.8 for the clock frequency of 300 MHz.

In Table 4.8, the term MEF refers to Mean Energy per Frame. The Copy 40 dB with performance targets of 81.9% and 80% are the most energy efficient approximate configurations. The maximum energy reduction for these configurations is of 57.4%. One can notice that the approximation in the Gaussian filter is decisive for the energy reductions. This is because the substantial difference is observed in energy consumption and area when considered the quality profiles for the Gaussian filter (*i.e.*, PSNR targets of 40dB and 50dB). For the copy configuration with PSNR target of 50 dB, the energy reduction is of 28.4%. For the ETAI configurations, the energy reductions are of 39%, and 14.6% for the 40 dB and 50 dB evaluated PSNRs, respectively. These results ratify the same findings observed in subsection 4.2 for the FIR filters case study: the copy adders are more energy efficient than the ETAI ones for similar output quality.

Based on the energy efficiency results, one can conclude that the proposed heuristic for approximations in the Canny edge detector architecture brings substantial energy savings, with real-time processing, without degrading the quality of the edge detection.

The authors in (GUPTA et al., 2013) state that the supply voltage is inversely proportional to the delay of a given circuit as shown in (32).

$$V_{DD} \propto \frac{1}{Delay} \tag{32}$$

Based on this observation, the authors in (GUPTA et al., 2013) estimate the approximate scaled voltage. This voltage is achieved due to delay reduction provided by the use of approximate techniques in the designs. The estimation is performed as in (33).

$$V_{DDAPP} = V_{DD}\left(1 - \frac{slack}{Tc}\right) \tag{33}$$

In (33), V_{DD} and V_{DDAPP} denote the nominal supply voltage and the approximate scaled voltage, respectively. The term *slack* refers to the delay difference between the precise and approximate versions, while *Tc* is the clock period of a given circuit. Therefore, one can perform a primitive estimation of additional dynamic power reduction due to the use of VOS. At the beginning of this subsection was shown that the approximate versions could achieve maximum operating frequency higher than the precise ones. Based on this observation, Table 4.9 shows the dynamic power when the approximate circuits are operating at 300 MHz and nominal supply voltage. Also, dynamic power dissipation is also shown for the approximate circuits operating at the same frequency and approximate scaled voltage. Results show that the estimated additional reduction provided by the VOS technique does not provide any substantial power reduction (*i.e.*, up to 3.7% further reduction). This is expected since the copy adder and ETAI are power-oriented design, so that they are not focused on reducing the critical path of the conventional adders.

Table 4.9 – Estimation of dynamic power reduction due to VOS technique plus approximation @ 300 MHz.

	precise	**Copy 40dB 81.9%**	**Copy 40dB 80%**	**Copy 50dB 90.9%**	**Copy 50dB 80%**	**ETAI 40dB 89.9%**	**ETAI 40dB 80%**	**ETAI 50dB 92.7%**	**ETAI 50 dB 80%**
V_{DDAPP} (V)	-	1.09	1.09	1.09	1.08	1.07	1.09	1.07	1.08
Dynamic Power @ V_{DD} (mW)	7.2	2.9	2.9	5.1	5.1	4.3	4.3	6.1	6.1
Dynamic Power @ V_{DDAPP} (mW)	-	2.9	2.9	5	4.9	4	4.3	5.8	5.9
Dynamic Power Reduction @ V_{DD} (%)	-	58.5	58.4	28.9	28.9	39.9	39.9	15	15

Dynamic Power reduction @ V_{DDAPP} (%)	-	59.3	59.5	30.2	30.5	43.6	40.7	19.7	17.5

4.3.3.3 Comparison with state-of-the-art Canny edge detectors

In this subsection, the objective is to state the contributions of this work to the emerging scope of real-time hardware accelerator design focused on edge detection for computer vision algorithms. As mentioned in the previous subsection, the proposed architectures guarantee real-time edge detection for different video or image resolutions regarding the observed frame rates. This real-time capability is also provided by most of the proposed FPGA architectures from the literature. For example, in (XU et al., 2014) the total execution time to process an 8-bit 512 x 512 image size, without considering the SRAM (Static Random Access Memory) read and write latencies, is of 0.37 ms.

Regarding quality metrics for edge detection, most of the related works also did not use any objective or subjective metric. The related works in (XU et al., 2014) and (POSSA et al., 2014) evaluated the performance of edge detection by using different metrics. In (XU et al., 2014) the authors adopted the use of objective and subjective metrics. The objective metric refers to three percentages considering the traditional Canny edge algorithm as being the reference. The percentages are the following: i) percentage of edges detected by their proposed distributed algorithm and the traditional Canny algorithm, ii) percentage of false negatives, and iii) percentage of false positives. In (POSSA et al., 2014) the SNR is adopted as metric to evaluate the edge detection.

This thesis uses the PSNR metric to evaluate the response of the 5x5 Gaussian filter while considering the performance metric to evaluate the edge detection response after the 3x3 Gradient filtering process. This is performed to evaluate an average response which is essential to determine the *k* parameters for each approximate adder in the filters. The performance metric adopted in this study is the same metric explored in (LEE; TANG; PARK, 2016). However, the work in (LEE; TANG; PARK, 2016) adopts 3x3 kernel size for the Gaussian filter in their proposed architecture. This may not be the best approach, because higher kernel sizes tend to reduce sensitivity to noise in edge detection application. Most of

the related works cited in this thesis explore the 5x5 kernel size for the Gaussian filter. For instance, a subjective comparison is shown between the 3x3 and 5x5 Gaussian filters for Canny edge detection. Both the filters are designed with variance $\sigma^2 = 1$. In Figure 4.22 is shown the comparison between the kernel sizes for a specific image from the BSD benchmark. This image has a high density of edges, and the 3x3 kernel size presents a much worse response when compared to the 5x5 one. Based on that, this work considers the use of the 5x5 Gaussian filter instead of 3x3 to allow for higher capabilities regarding noise attenuation.

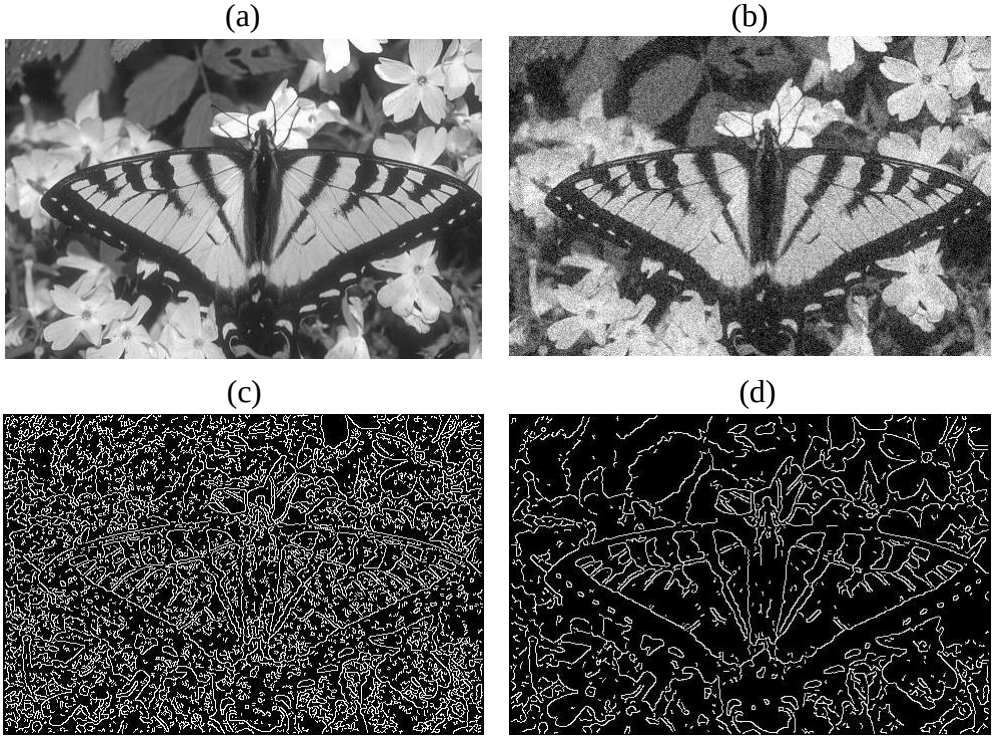

Figure 4.22 – Subjective analysis. (a) the original image, (b) noisy image, (c) edge detection result by using a 3x3 Gaussian filter, and (d) edge detection result by using a 5x5 Gaussian filter.

According to (HAN; ORSHANSKY, 2013), energy efficiency has become a paramount concern for computing systems design. On the other hand, most of the related works did not present energy efficiency analysis for their architectures making it impossible to evaluate the critical concern about power dissipation for hardware accelerators in deca-nanometer technology. Table 4.8 shows that the proposed precise approach consumes from 23.2 µJ up to 208.6 µJ for different video resolutions. The proposed approximate configurations present substantial energy reduction of 57.4%. The work in (POSSA et al., 2014) shows that for the Canny edge operator synthesized to 60 nm Cyclone IV EP4CE115 FPGA, the energy consumption is of 1.6 mJ regarding 512 x 512 image size. This energy

result presented in (POSSA et al., 2014) shows that the FPGA solution is not the most suitable platform when considered only the energy efficiency. This observation ratifies the findings in (KUON; ROSE, 2007) which concludes that ASIC approaches are more energy-efficient than FPGA ones.

The work in (LEE; TANG; PARK, 2016) presented synthesis results for their Canny edge architecture for a 65 nm library targeting ASIC design flow. The operation frequency is of 500 MHz which according to the authors are the maximum achieved frequency. The energy consumption and edge detection performance comparison are shown in Table 4.10. The same proposed architecture was also synthesized and mapped onto a 65 nm PDK. The comparison is organized as follows: i) minimum and maximum energy consumption considering only the approximate versions, and ii) minimum and maximum performance of edge detection considering only the approximate versions.

One can observe that the most energy efficient results are related to the approximations performed by this work considering the 45 nm PDK. For similar CMOS technology node, the 65 nm results of this work are about 3.5 X more energy consuming than the ones presented by the related work in (LEE; TANG; PARK, 2016).

Table 4.10 – Comparison with related work

	This work 45 nm	**This work 65 nm**	**(LEE; TANG; PARK, 2016) 65 nm**
	Mean Energy per Frame (µJ) – target of 30 fps		
HD videos	9.9 – 18.9	43.3 – 89.5	12.2
FHD videos	22.2 – 44.6	97.3 – 200.9	27.3
QHD videos	39.5 – 79.2	172.7 – 356.8	48.4
UHD videos	88.9 – 178	388.2 – 738.4	108.8
	Performance under standard conditions (%)		
	89% - 95%		91.4%

	Performance under noisy conditions (%)	
	71% - 72%	60.47%

This result makes sense, because the related work uses a 3x3 Gaussian kernel size, while in this thesis the 5x5 Gaussian size was implemented. As a counterpart, the edge detection performance results presented in this thesis shows that, at under noisy condition, the 3x3 Gaussian filter may not be an appropriate decision. This confirms the findings presented in Figure 4.22, where subjective analysis for a specific image show that the 5x5 Gaussian filter is more resilient to noise than the 3x3 one. For the 65 nm synthesis results, the Copy 40 dB 80% approximate configuration presents the maximum energy reduction of 56.1%. This indicates that the proposed heuristic is an essential approach in approximate computing scenario. In sum, this work innovates by proposing a more comprehensive approximate computing exploration without substantially degrading the application quality.

4.4 Comparison with state-of-the-art cross-layer methodologies

Table 2.3 summarized the state-of-the-art cross-layer methodologies to evaluate approximation capabilities. Most of those works do not consider a systematic evaluation regarding application quality. This information can easily be observed because those works consider a reduced number of test cases or only analytical evaluation. According to (XU; MYTKOWICZ; KIM, 2016), one of the challenges in approximate computing is just understanding what the implications of lower level approximations in the high-level application stack. Therefore, this is one of the motivations and important contribution of this thesis. This is the main reason why this work proposed a simulation-based methodology. In (LIU; HAN; LOMBARDI, 2015; MAZAHIR et al., 2017), the analytical methodologies are more focused on approximate adder error characteristics inside the application context without power-performance profiling. In other words, the authors present analysis in a low-level layer without a precise connection with the high-level application.

The simulation-based methodology proposed by (KANG; KIM; KANG, 2016) is focused only on FIR filter application, and their method may not represent the best approach regarding application quality. This is because the search for approximation is driven by random inputs in the FIR filter circuits. This adoption may also not represent the correct analysis regarding energy efficiency.

In this thesis, the novelty and contributions in this scope can be enumerated as follow:

- A cross-layer (*i.e.*, arithmetic and architectural integration) simulation-based methodology which evaluates and characterizes power-performance-accuracy profiles by considering real test cases and metrics from the application level.
- The proposal of heuristics to enable heterogeneous approximation in the state-of-the-art approximate adders. The proposed heuristics cover more configurations than only adopting uniform approximate parameterization in the adders. One can notice that the proposed heuristics test the uniform cases plus much more configurations.
- A systematic energy efficiency analysis considering real and representative input data from applications instead of considering random data. Results show energy reduction of up to 25.7% for the FIR filters processing real audio signals. For the Canny edge application, energy reduction of up to 57.4% is achieved. When considered VOS estimation, a dynamic power reduction of 59.5% is observed.
- Exploration of approximate computing techniques is provided at the architectural level. In this work, a new Canny edge detector architecture is proposed for ASIC implementation with energy reduction higher than 50%. In addition, when compared to the state-of-the-art Canny edge accelerators, this new architecture and the approximations improve application quality even considering noisy images.

4.5 Summary of the chapter

In this chapter, the proposed simulation-based methodology was shown. The methodology was exercised to evaluate the impact of the use of state-of-the-art approximate adders in different applications. One can observe that this work proposed a more comprehensive and heterogeneous exploration of approximations at the arithmetic level. Based on that, search heuristics were proposed to enlarge the design space for approximate adders regarding quality-power-performance profiles. A novel architecture was also proposed for the Canny edge exploration. In sum, this chapter explored approximate computing techniques which can be adopted during the design-time of hardware accelerators. The proposed methodology can also be used to evaluate run-time accuracy configuration. Therefore, in the next chapter accuracy-configurable architectures are proposed to enable run-time configuration for dynamic quality-power-performance profiles exploration.

5 PROPOSED ACCURACY CONFIGURABLE ARCHITECTURES

In the previous Chapter, the integration of logic and arithmetic approximation inside architectural level is evaluated. The power-performance-accuracy profiles provided by the proposed methodology bring more precious information about the cross-layer integration when compared to state-of-the-art. Based on that, the proposed methodology enables energy efficiency capability from the arithmetic level up to the application. On the other hand, in Chapter 4, the proposed approach is focused on fine grain approximation (*i.e.*, adders in a given architecture). In this Chapter, coarser grain and adaptive techniques are proposed to be further integrated with the findings from Chapter 4. All the case studies in this Chapter are based on circuit pruning. This technique is also considered approximate computing approach, since different pruning levels may result in more or less accurate responses. As a consequence, energy efficiency can be managed through the use of adaptive circuit pruning. All the evaluations in this context are performed by adopting the proposed methodology of this thesis.

5.1 A case study on SATD pruning for HEVC

As previously shown in Chapter 3, the SATD is a block matching metric mostly used in HEVC standard due to the superior performance regarding bitrate and video quality when compared to the simpler SAD metric. On the other hand, the improvement provided by the SATD results in computational effort increase as previously presented in Table 3.2. The video is a dynamic media, since at each frame the content may vary entirely or partially. Depending on the type of video, the motion estimation can be more or less accurate without significant degradation. Based on that, this thesis proposes an algorithm to prune the SATD computation, which can be further explored and configured in run-time.

The pruning-based algorithm considers fully parallel implementation transforms such as the Hadamard, DCT and so on. Figure 5.1 shows an example of the fully parallel 4x4

SATD implementation. The second and fourth steps are the horizontal and vertical 1D Hadamard Transform. The detailed implementation of the vertical or horizontal transform is shown in Figure 5.2.

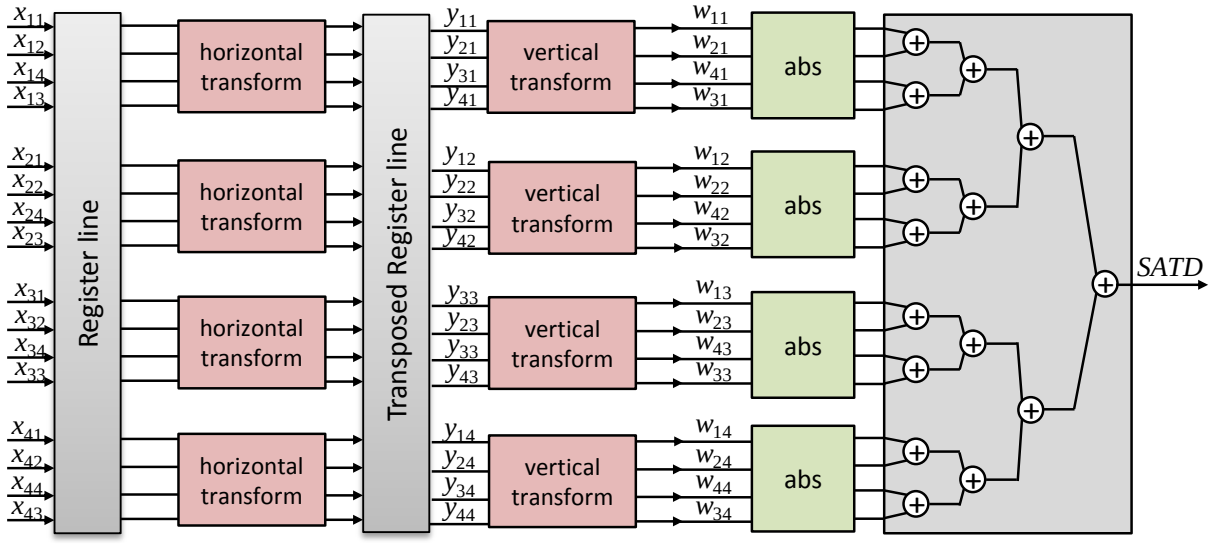

Figure 5.1 – Fully parallel 4x4 SATD implementation.

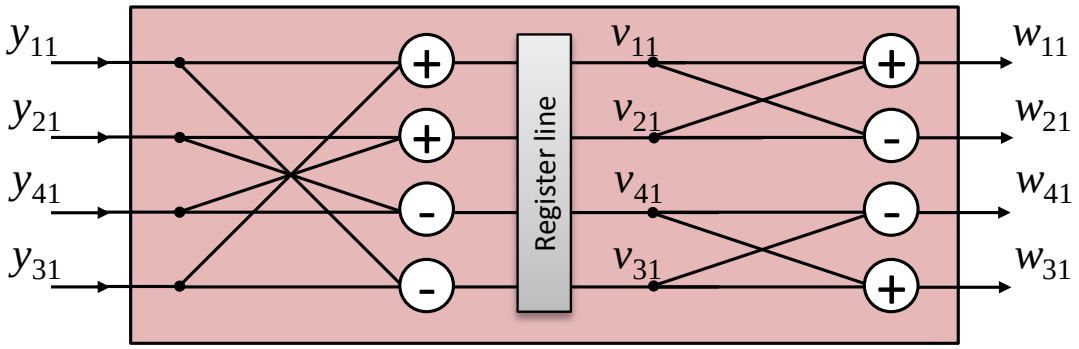

Figure 5.2 – Internal structure of horizontal and vertical transforms.

The next steps in Figure 5.1 represent the absolute operator and the SAD tree, which computes the sum of the absolute transformed differences. The proposed pruning-based algorithm shown in Algorithm 1 is independent of the block size of the SATD. The algorithm uses two data structures: i) a stack S containing the Hadamard Transform (HT) absolute coefficients w_{ij} sorted by a given policy, ii) a tree R of dependencies among the prior terms used in the HT to compute each coefficient w_{ij}. The stack is necessary to decide the order of the coefficients being discarded from the matrix W shown in (11). The policy adopted to discard the coefficients is further explained in Figure 5.4. The tree is used by the proposed algorithm to solve dependencies among the discarded coefficients w_{ij} and prior terms that can be pruned from the architecture because the remaining HT coefficients do not use them. The proposed algorithm also takes as input the number of coefficients n to be discarded. The

output is the approximate SATD RTL VHDL design generated according to the remaining terms in the pruned tree *R*.

Before explaining the algorithm, an example of the tree of dependencies is shown in Figure 5.3, since this is the central structure of the algorithm. This example refers to the sub-tree related to the vertical transform presented in Figure 5.2. All the coefficients w_{ij} are leaves in the tree. They are computed by the prior terms denoted as father nodes. For instance, the prior terms v_{31} and v_{41} are used only for the computation of the w_{31} and w_{41} coefficients. Hence, if those coefficients are discarded, the proposed algorithm must prune from the tree

---

Algorithm 1 The pruning-based algorithm for SATD

Input: *S* – a stack of coefficients sorted by any policy
R – a tree of dependencies among HT terms or any transform
n – the number of coefficients to be discarded
Output: *Q* – the approximate RTL netlist

```
1: while n > 0 do
2:     pruning := TRUE
3:     s := pop(S)
4:     l := search_leaf(s,R)
5:     while pruning do
6:         k := father(l,R)
7:         remove(l,R)
8:         if has_child(k,R) then
9:             pruning := FALSE
10:        else
11:            if is_root(k,R) then
12:                remove(k,R)
13:                pruning := FALSE
14:            else
15:                l := father(k,R)
16:                remove(k,R)
17:                if has_child(l,R) then
18:                    pruning := FALSE
19:                end if
20:            end if
21:        end if
22:    end while
23:    n := n -1
24: end while
25: Q := generate_netlist(R)
```

---

the v_{31}, v_{41}, and $y_{11} - y_{31} + y_{41} - y_{21}$ terms. One can observe that all the remaining terms of the tree must be preserved since the coefficients w_{11} and w_{21} were not discarded.

The proposed algorithm has an external loop that is performed *n* times to remove the next coefficient and its exclusive associated prior terms at each iteration. For instance, when considered the 4x4 SATD, the *n* parameter can range from 1 up to 16 (*i.e.*, no use of HT). The Boolean *pruning* variable is used to control the pruning iteration in line 5.

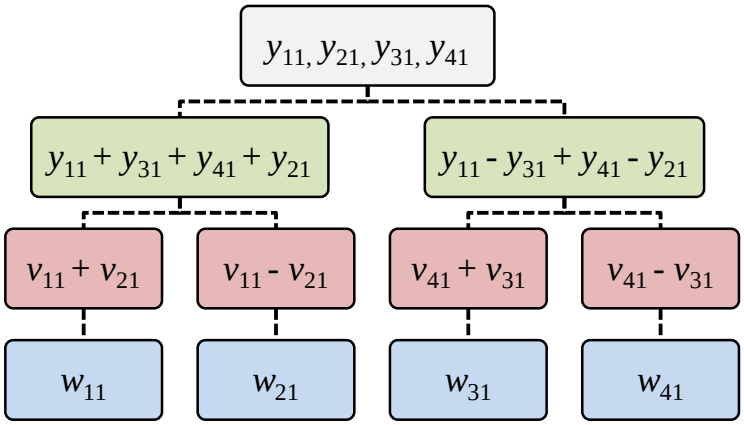

Figure 5.3 – Example of the tree of dependencies for the vertical transform.

Each coefficient is extracted from the top of the stack *S* with the *pop* function. Then, the extracted coefficient is searched in the leaves of the tree *R*. Once the coefficient is found, its location is stored in the variable *l*, and the pruning process starts. The next step is to visit its father node and store the location in variable *k*. Since the location of the coefficient to be discarded is stored in variable *l*, this information is used to remove this coefficient from the tree. The next step is to evaluate if the current father node has a remaining child. If this verification is valid, the pruning must stop, since some coefficients or terms still depend on the current father node. On the other hand, the current father node can also be removed. Before removing the current father node, it is essential to verify if this node is also the root node. If true, then this node is removed, and the pruning must be stopped. This is because there are no more coefficients or terms to be pruned. Otherwise, it is essential to store the location of the grandfather in *l* before removing the current father.

The next step is to verify if the grandfather node has a remaining child. When false, the pruning iteration (*i.e.*, line 5) must continue to remove the grandfather node and possibly more prior terms. When true, the pruning must stop. This is because the grandfather still has a child and cannot be pruned from the tree.

Until now, this work has not discussed yet which policy can be adopted to sort the coefficients during the pruning. One may consider that the significance of each coefficient has not the same weight for the output quality. Therefore, this thesis adopts a significance-driven technique to prune the coefficients from the HT. In the scope of this work, the significance is defined as the average magnitude of each Hadamard transformed coefficient considering a set of video sequences from different classes. Once the stack of most significant coefficients is determined, the proposed algorithm can perform the pruning of the SATD architecture based on the average significance of each coefficient. For example, Figure 5.4 shows the average magnitude for each coefficient in 4x4 HT considering four video sequences (*i.e.*, from class A, B, C, and D, as presented in Table 5.1).

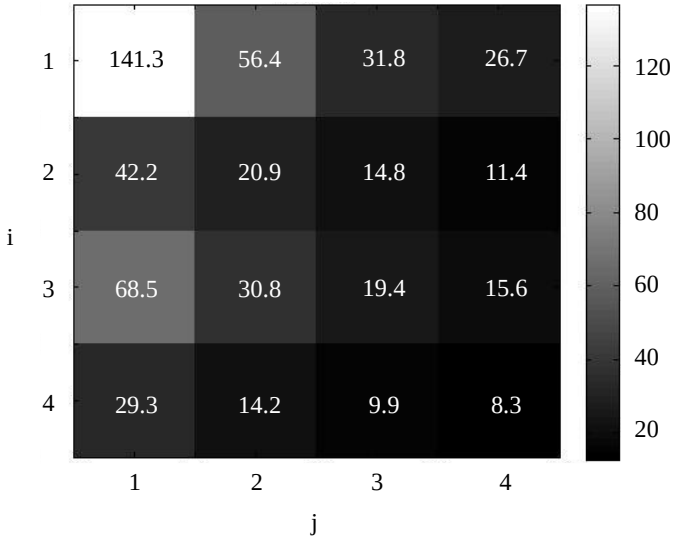

Figure 5.4 – The average magnitude of each coefficient in 4x4 HT

For instance, to discard the ten least significant coefficients from the precise example in Figure 5.1, the following coefficients will be removed: w_{44}, w_{43}, w_{24}, w_{42}, w_{23}, w_{34}, w_{33}, w_{22}, w_{14}, and w_{41}. One can observe that, based on the proposed algorithm, the tree of dependencies is pruned to reduce the number of cells in the circuit. The resulting approximate SATD architecture can be verified in Figure 5.5, where the approximate architecture has pruned horizontal or vertical transforms.

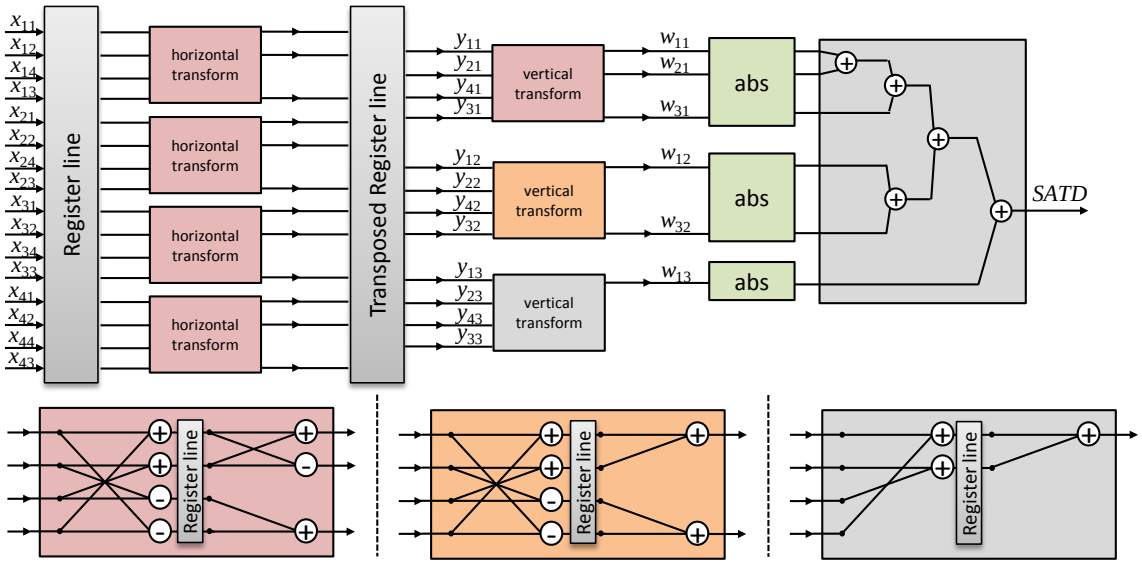

Figure 5.5 - Approximate SATD architecture with 10 discarded coefficients.

The pruned horizontal and vertical transform blocks have fewer adders than the complete ones. The reduction in the number of adders for the approximate SATD architecture in comparison with the precise one is 37.9%. In conclusion, the additional cost of the approximate SATD architecture shown in Figure 5.5 is 1.7X when compared to the 4x4 SAD architecture. Table 3.2 showed that the precise 4x4 SATD has an additional cost of 3.1X. When compared with the ten pruned coefficients approach, the precise architecture has 1.8X more computational cost.

One can notice in Figure 5.5 that some registers can be discarded when compared with the precise architecture. The key solution is to provide a configurable scheme where these registers are clock gated instead of being discarded. Based on that, dynamic power is reduced in favor of accuracy-configurable energy-efficient design. To select different levels of accuracy for the configurable approach, Table 5.1 shows the 4 videos considered to evaluate the output quality and bitrate per number of discarded coefficients. In this scope, 64 frames from these videos are encoded by considering the HM 16.0 software and the common test conditions presented in (BOSSEN, 2013). The random access configuration was selected instead of all intra or low delay. This is because it is the most complex and most generic configuration for picture ordering. Each video sequence was run for the following quantization parameters: 22, 27, 32, and 37. This is because those points are integrated to evaluate the Bjøntegaard-Delta bit-rate (BD-BR) and PSNR (BD-PSNR). For BD-BR and BD-PSNR metrics, the precise 4x4 SATD is the reference. For this analysis, all default block-sizes regarding SATD computation are allowed.

Table 5.1 – Video sequences specification

Class	Video sequence name	Resolution
A	Traffic	2560x1600
B	BQTerrace	1920x1080
C	RaceHorses	832x480
D	BlowingBubbles	416x240

Video quality (BD-PSNR) and video compression (BD-BR) results can be seen in Figure 5.6 and Figure 5.7, respectively. In the graphs, discrete points for each configuration of discarded coefficients are shown, and the dotted red line represents the 4x4 SAD (*i.e.*, when 16 coefficients are discarded in the 4x4 HT). In fact, this line represents a boundary to evaluate if there are configurations ranging from 1 to 15 discarded coefficients that have worse results than using the 4x4 SAD. Based on that, for all the cases the 4x4 SAD presents better results than some pruned configurations. For instance, when considered the BQTerrace sequence, both BD-PSNR and BD-BR 4x4 SAD results have lower degradation and higher compression than the configurations from 13 up to 15 discarded HT coefficients. Therefore, it is preferable to use the 4x4 SAD instead of those configurations, because SAD tends to be a more energy-efficient solution with better results regarding BD-BR and BD-PSNR.

Regarding BD-PSNR, higher resolution videos such as BQTerrace and Traffic have lower degradation in quality. For those video sequences, when ranging from 1 to 10 discarded coefficients, BD-PSNR results range from slightly above zero to -0.01 dB. One can realize that some configurations present better BD-PSNR results than the reference precise 4x4 SATD. The hypothesis for that is the following: some pruned HT (*e.g.*, the two discarded coefficient HT) could present slightly better properties than the complete HT regarding energy concentration. For lower resolution videos, the BD-PSNR ranges from slightly above zero to -0.05 dB when considered one discarded coefficient HT and the 4x4 SAD, respectively.

The BD-BR results are just a complimentary evaluation about the BD-PSNR. For higher resolution videos (*i.e.*, (a) and (d)) and considering the range from 1 discarded coefficient up to the 4x4 SAD, the BD-BR configuration increases up to 0.4% and 0.5%, respectively. For lower resolution videos in (b) and (c), the increase in BD-BR is not higher than 1%.

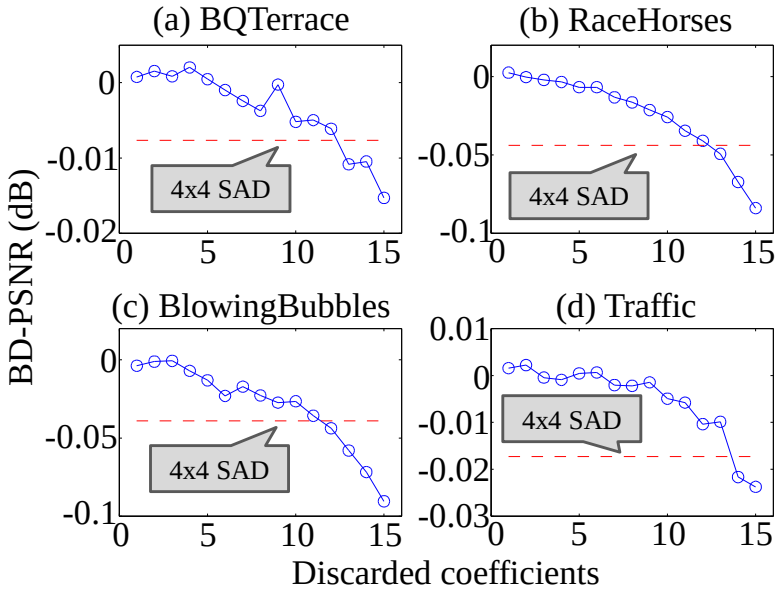

Figure 5.6 - BD-PSNR results for approximate SATD architectures.

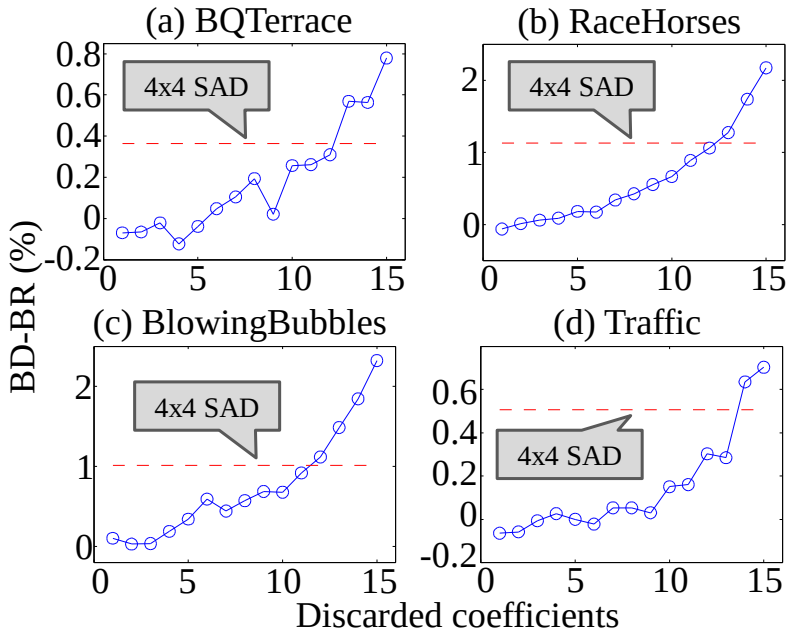

Figure 5.7 - BD-BR results for approximate SATD architectures.

Based on these findings, the selected configurations for the accuracy-configurable architecture are the following: i) four discarded coefficients, ii) eight discarded coefficients, and iii) ten discarded coefficients. These configurations were selected due to the quality results and because these configurations support an incremental configurable hardware approach as shown in Figure 5.9. In addition, four, eight, and ten discarded coefficients tend to provide the appreciable difference regarding energy reduction and quality. The configurable architecture can also sustain the precise implementation. This occurs when all

the registers are working or, in other words, when all the affected registers are not disabled through clock gating technique. The overall configurable architecture is presented in Figure 5.8.

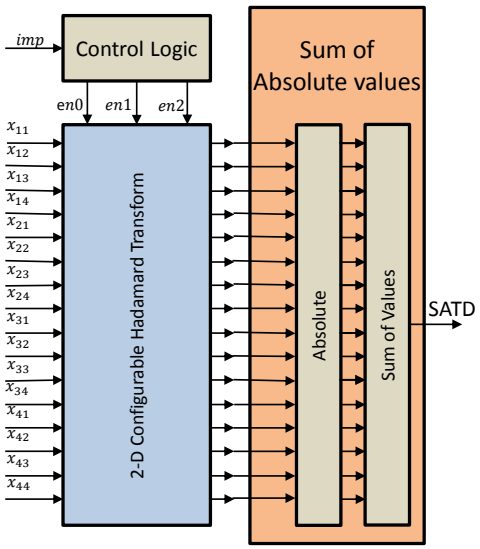

Figure 5.8 - Block diagram of the run-time quality-energy configurable 4x4 SATD architecture.

Table 5.2 – Control logic of the approximate configurations.

Configuration	**imp**	***en1***	***en2***	***en3***
Precise	00	1	1	1
Four discarded coefficients	01	0	1	1
Eight discarded coefficients	10	0	0	1
Ten discarded coefficients	11	0	0	0

The control logic works as indicated in Table 5.2. For the precise architecture, all the registers must be active. On the other hand, for the pruned versions, three distinct groups of registers are controlled through the enable signals represented by *en1*, *en2*, and *en3*. Figure 5.9 presents the internal structure of the proposed configurable HT. The distinction in colors is performed to differentiate the registers and related circuitry which should be enabled or disabled to configure the architecture.

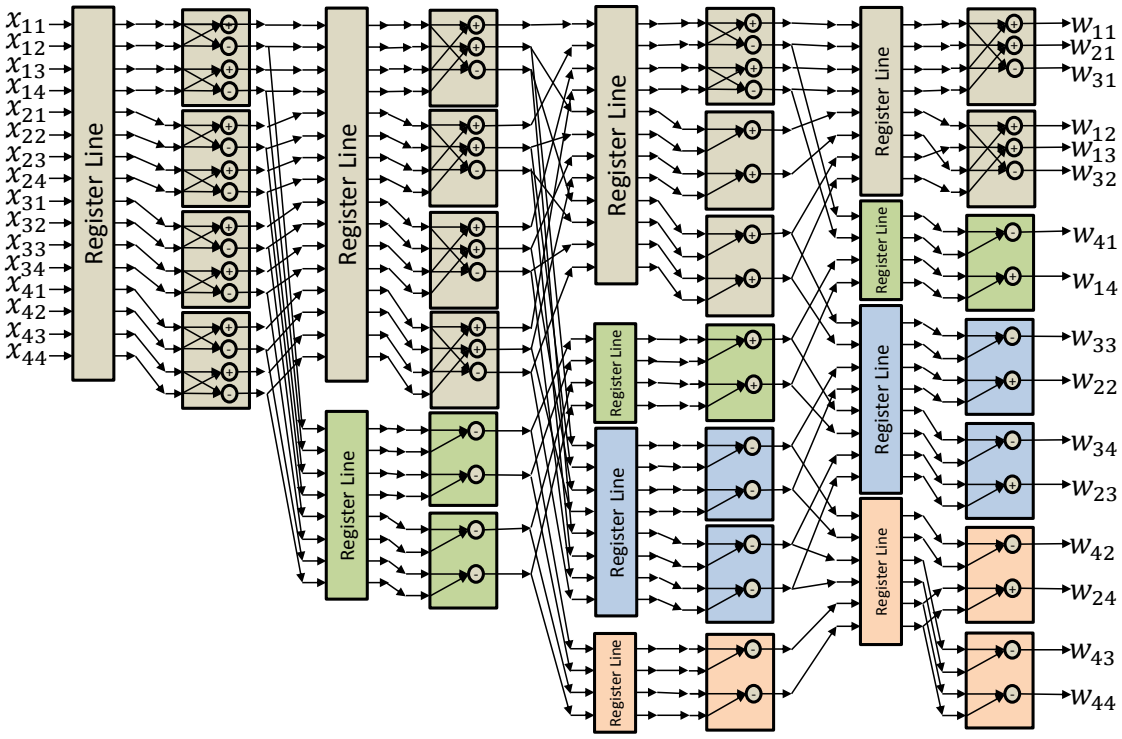

Figure 5.9 – Block diagram of 2D configurable Hadamard Transform architecture

From bottom-up direction in the block diagram, the first, second, and third colored groups refer to the clock enabling signals *en1*, *en2*, and *en3*, respectively.

5.1.1 Results and discussion

This section shows results for the case study on configurable 4x4 SATD. First, the experimental setup is presented followed by the results regarding video quality, compression, and energy efficiency. Finally, a projection is presented for the 8x8 SATD study case considering energy efficiency.

5.1.1.1 Experimental Setup

As shown in Table 5.3, fifteen video sequences are considered from the HEVC benchmark to evaluate the video quality and compression analysis. For each video sequence, 64 frames are encoded considering random access configuration due to its higher complexity when compared to low delay. All these videos are encoded with four different QP values (22, 27, 32, and 37) as recommended in CTC by (BOSSEN, 2013). Differently from the analysis previously mentioned, a more constrained quality and compression evaluation are implemented to validate the configurable architecture.

Table 5.3 – Benchmark adopted for vídeo quality and compression analysis

Video sequence	Frame rate (fps)	Video resolution
BlowingBubbles	50	416x240
BQSquare	60	416x240
RaceHorses	30	416x240
BasketballPass	50	416x240
BQMall	60	832x480
BaskeballDrill	50	832x480
PartyScene	50	832x480
RaceHorsesC	30	832x480
Kimono	24	1920x1080
Cactus	50	1920x1080
BasketballDrive	50	1920x1080
ParkScene	24	1920x1080
BQTerrace	60	1920x1080
PeopleOnStreet	30	2560x1600
Traffic	30	2560x1600

Since only the 4x4 SATD is implemented, the original version of the HM-16.0 software was modified to use only this block-size for SATD operation. This is performed to evaluate the precise baseline 4x4 SATD response without considering the implications of any other block-sizes. The pruning algorithm was also implemented inside the HM source code to evaluate the video quality and compression when the pruned versions are running.

The configurable and the baseline 4x4 SATD architectures were described in VHDL followed by synthesis. The synthesis was performed by the RTL compiler tool, where the designs were mapped onto the 45nm Nangate FreePDK, and the netlists were generated. The netlists were simulated to extract switching activity. The input stimuli are 40,000 4x4 residual blocks extracted from the HM-16.0 from different video sequences. After switching activity is

annotated, it is used to estimate the power dissipation in the circuits. Six different profiles were adopted to manage the configurations for power dissipation evaluation. These profiles can be seen in Table 5.4. Since 40,000 4x4 blocks are considered during simulation, this number is divided into four sets, where each set contains 10,000 4x4 blocks. Therefore, at each set, the configuration may be changed.

Table 5.4 – Run-time configuration profiles (number of discarded coefficients)

profile	first set	second set	third set	fourth set
#1	zero	zero	zero	zero
#2	four	four	four	four
#3	eight	eigth	eight	eight
#4	ten	ten	ten	ten
#5	zero	four	eight	ten
#6	four	eight	ten	ten

5.1.1.2 Video Quality and Compression Results

The video quality and compression results were grouped by video resolution and are shown in Figure 5.10 and Figure 5.11, respectively. One can observe that the pruning configuration which considers zero discarded coefficients is not shown in these results. This is because this configuration refers to the complete 4x4 SATD version and is used as a reference to calculate the BD-PSNR and the Bjontegaard-Delta Bit Rate (BD-BR) metric.

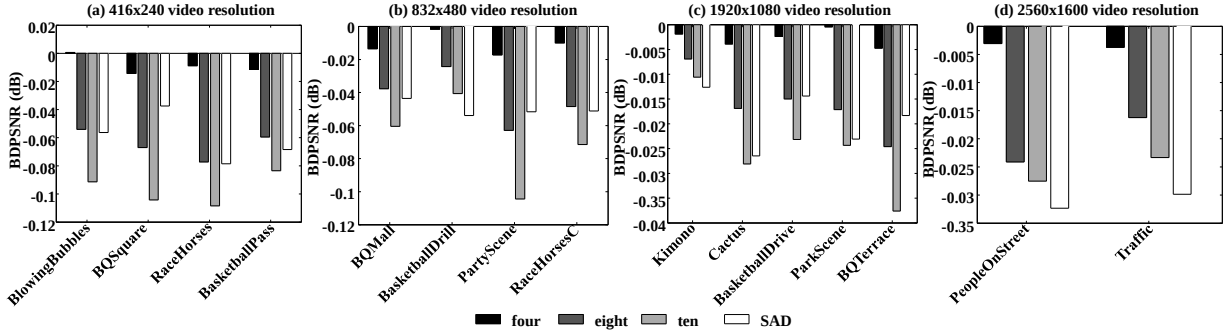

Figure 5.10 – BD-PSNR results.

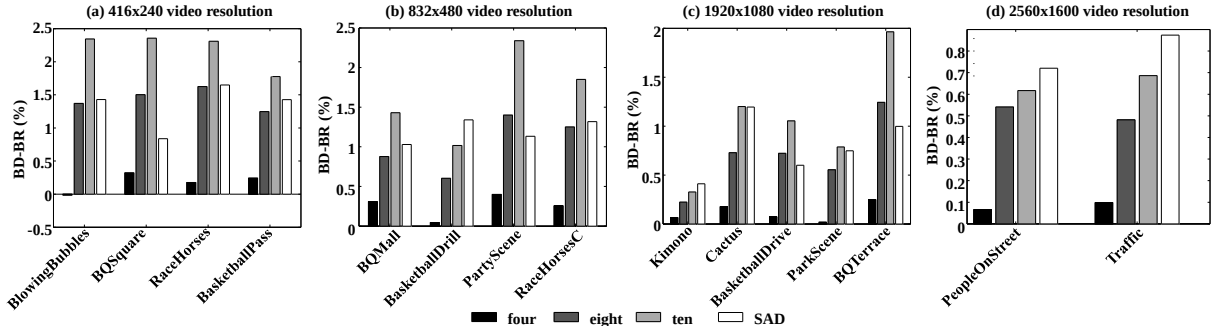

Figure 5.11 – BD-BR results.

The former metric should be interpreted as the decrease/increase of quality regarding PSNR for the same level of compression, while the latter represents the increase/decrease in terms of bit rate (compression) for the same video quality regarding PSNR. In general, when considered the 416x240, 832x480, and 1920x1080 video resolutions the four and eight discarded coefficients configurations present lower BD-PSNR degradation and lower BD-BR than the simpler SAD metric for most of the cases. The exception is for the ten discarded configuration which presents higher BD-PSNR and BD-BR than the SAD. This can be explained as follows: i) This set of results is obtained by adopting only the 4x4 SATD block-size, while in the first quality analysis any block size is allowed; and ii) In this current analysis, a broader set of videos with different motion speed characteristics are adopted. The exception is for the 2560x1600 video resolution, which all the pruned configurations present better results than adopting the SAD metric regarding video quality and compression.

Table 5.5 and Table 5.6 show the average BD-PSNR and BD-BR results per video resolution, respectively. One can observe that the previously mentioned behavior is confirmed in the average analysis. The four discarded coefficients configuration presents the best results for all the video resolutions regarding BD-PSNR and BD-BR metrics when compared to the other configurations and the SAD operation. The eight discarded coefficients present better BD-PSNR and BD-BR results than the ten discarded coefficients and the SAD. On the other hand, for the 416x240, 832x480, and 1920x1080 video resolutions the ten discarded coefficients configuration presents higher video quality degradation and bit rate than the SAD. The exception is for the highest analyzed video resolution, where the ten discarded coefficients present slightly better average BD-PSNR and lower BD-BR than the SAD operation. This indicates that for a more extensive benchmark, ten discarded coefficients in the 4x4 SATD may present worse quality results than the SAD depending on the video sequence and its resolution.

Table 5.5 – Average BD-PSNR

Configuration	416x240	832x480	1920x1080	2560x1600
Four	-0.008 dB	-0.01 dB	-0.003 dB	-0.003 dB
Eight	-0.065 dB	-0.043 dB	-0.016 dB	-0.02 dB
Ten	-0.097 dB	-0.069 dB	-0.025 dB	-0.025 dB
SAD	-0.06 dB	-0.05 dB	-0.019 dB	-0.031 dB

Table 5.6 – Average BD-BR

Configuration	416x240	832x480	1920x1080	2560x1600
Four	0.18%	0.25%	0.12%	0.08%
Eight	1.44%	1.03%	0.69%	0.51%
Ten	2.2%	1.65%	1.07%	0.65%
SAD	1.34%	1.2%	0.79%	0.8%

These results indicate that, for the benchmark under analysis, ten discarded coefficients should be adopted for higher video resolutions, while for lower ones this configuration must be avoided. This is because the simpler SAD computation would result in better results regarding video quality and energy efficiency.

5.1.1.3 Energy Efficiency Results

As previously mentioned in Chapter 3, the SATD operation is a compute-intensive kernel which demands high throughput when considered the HEVC standard. This is the motivation to analyze the maximum achievable frequency. Based on that, the proposed architecture achieved a maximum frequency of 790 MHz. This is the operating frequency adopted for energy efficiency analysis shown in Table 5.7, where the term MEOp refers to Mean Energy per Operation. The results are organized per observed configurable profile. The non-configurable baseline 4x4 SATD architecture is also implemented to analyze the energy overhead when complete SATD operation is required.

Table 5.7 – Energy efficiency analysis for configurable SATD

SATD profile	MEOp (pJ) @ 790 MHz	MEOp (pJ) x 4 @ 790 MHz
#1	56.64	226.56
#2	47.39	189.56
#3	36.19	144.76
#4	28.52	114.08
#5	42.34	169.36
#6	35.26	141.04
baseline 4x4 architecture	50.60	202.4

One can observe that there is energy overhead of 11.9 % only for the configurable profile #1. This is because this profile considers the complete 4x4 SATD is operating all the time. On the other hand, the remaining profiles present energy reductions when compared to the baseline architecture. The maximum energy reduction is of 43.65% for the profile #4. This profile considers that the ten discarded coefficients configuration is operating all the time. This is a valuable scenario for the 2560x1600 video resolution when observed the video quality and compression results for this profile in Table 5.5 and Table 5.6. When considered the profile #3, this is the one which presents energy reduction of 28.48% and has superior video quality and compression response for most of the video resolutions when compared to the SAD architecture. An additional column is provided in Table 5.7 to enable a projection for 8x8 SATD operation. This is a valid estimation since 8x8 SATD architecture can be composed of 4 instances of 4x4 SATDs.

Table 5.8 shows the area results, (*i.e.*, area and number of cells) for the configurable 4x4 SATD and the baseline architectures. One can observe that the overhead in terms of area is of 14.7%, due to the additional clock gating and the control logic to enable the architecture configurations.

Based on the previously shown results, one can conclude that the proposed accuracy configurable SATD architecture is capable of providing different profiles of application quality and energy efficiency.

Table 5.8 – Area results @ 790 MHz

Architecture	Area (μm²)	# of cells
baseline	69530	39940
configurable	79750	44490

The video quality and compression are scaled at each approximate profile, while the minimum and maximum energy reduction are of 6.8% and 43.65%, respectively. The energy and area overhead experienced for the complete 2D HT profile is of 11.9% and 14.7%, respectively. On the other hand, this additional cost could be compensated by the energy-efficient profiles during video coding process.

5.2 A case study on Gaussian filter pruning for Canny edge detection

A coarse grain approach is also provided for the Gaussian filter regarding Canny edge application. As observed in Chapter 3, the 5x5 Gaussian filter is the most compute-intensive operator followed by the Gradient filter. Therefore, one may consider that accuracy configurable solution for this task may represent energy reduction and acceptable output quality for the proposed Canny edge architecture shown in the previous chapter.

The proposed configurable architecture is focused on dynamically pruning the 5x5 Gaussian kernel according to a given system characteristic such as the battery level, the need for a higher or lower edge detection quality, and so forth.

The first shift register shown in Figure 4.15 is modified to provide a configurable solution for the Canny edge detection architecture regarding the 5x5 Gaussian filter. The proposed solution considers the use of clock gating technique in this shift register to determine the kernel size for the Gaussian computation. Based on that, the following configurations are considered: i) the precise 5x5 Gaussian filter; ii) the approximate 5x3 Gaussian filter; iii) the approximate 3x3 Gaussian filter; iv) no use of the Gaussian filter. The full 5x5 Gaussian filter is adopted every time all the registers are in activity. The approximate configurations consist of adopting clock gate technique to disable activity in the registers according to the Figure 5.12.

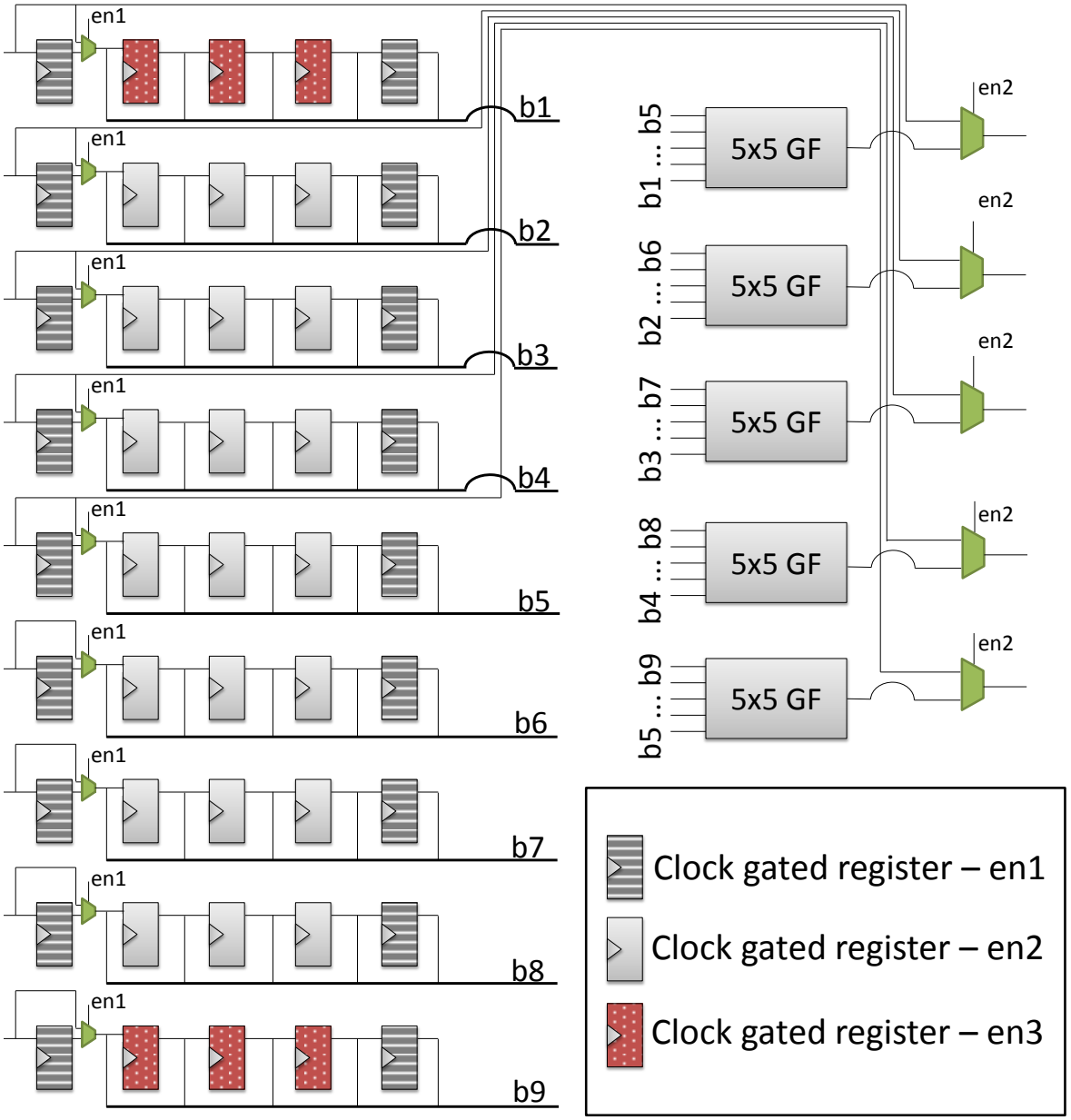

Figure 5.12 - The pruning configurable 5x5 Gaussian image filter in the Canny edge detector.

One can observe that three different patterns are adopted in Figure 5.12 to differentiate the registers which are related to the kernel sizes exercised in the configurable architecture. The striped, uniform and dotted patterns refer to the registers which are enabled/disabled by the control signals *en1*, *en2*, and *en3*, respectively. The enable signals are organized as follows: i) 5x5 Gaussian filter with *en1* = 1, *en2* = 1, and *en3* = 1; ii) 5x3 Gaussian filter with *en1* = 0, *en2* = 1, and *en3* = 1; iii) 3x3 Gaussian filter with *en1* = 0, *en2* = 1, and *en3* = 0; and iv) no use of Gaussian filter with *en1* = 0, *en2* = 0, and *en3* = 0. When 5x3 kernel size is configured, the first column of registers is disabled, and the input data should bypass this first column. This is implemented by adopting multiplexers as indicated in Figure 5.12. A similar bypass technique should be implemented when no use of the Gaussian filter is chosen. This

can be checked at the output of the Gaussian filter in Figure 5.12, where there are multiplexers to decide the data flow to the input of the Gradient filters. The configurations can be set per image. Therefore, when one image is being processed, the selected configuration must remain the same. Also, before configuration change, one cycle is taken to reset all the registers. This is performed to produce “zeros” in disabled regions. Thus, the disabled regions do not affect the computation result.

The coefficients which compose the 5x5 kernel are the same adopted for the pruned versions. Therefore, this procedure may induce error in Gaussian filtering for the pruned versions. This is because the coefficients should be different when kernel size changes for a given variance as shown in (6). Hence, it is correct to state that each pruned version is related to approximate computing technique.

5.2.1 Results and discussion

In this subsection, the Canny edge quality results through dynamic configuration are shown. In addition, energy efficiency analysis is provided for the Gaussian filter by considering different kernel size profiles.

5.2.1.1 Canny Edge Detection Analysis

The proposed simulation-based methodology was adopted to validate the Canny edge response regarding the different configurable profiles for the Gaussian image filter. In the scope of edge detection, the same BSD benchmark adopted in Chapter 4 is used. In this analysis, the use of the performance metric shown in (30) is also considered.

Table 5.9 shows the average performance metric considering the 16 images from the data set and the following profiles of 5x5, 5x3, and 3x3 Gaussian filters as well as no use of this filter. In this table, the reference image is the precise 5x5 Gaussian filter. The 5x3 and 3x3 versions experienced the edge detection performance of 79.2% and 72.3%, respectively. When the Gaussian filter is not adopted, the performance is of 34.8%. For benchmark corrupted by Gaussian noise with $\sigma^2 = 0.01$, the results show that all the configurable profiles present lower performance than processing the original benchmark. The 5x3 and 3x3 versions experienced the performance of 72.3% and 67.5%, respectively. When the Gaussian filter is not adopted, the performance is of 11.6%.

Table 5.9 – Average performance regarding the 5x5 Gaussian filtered image as a reference

	5x3	**3x3**	**No GF**
Average performance for BSD benchmark (%)	79.2%	72.3%	34.8%
Average performance for noisy BSD benchmark (%)	72.3%	67.5%	11.6%

In addition, the average performance is also performed when considered the reference as being the ground truth images. These results are shown in Table 5.10. One can observe that the 5x5 configuration presents slightly higher performance when compared to the 5x5 and 3x3 configurations. The configuration which disables the Gaussian filter experienced the lower results.

Table 5.10 – Average performance regarding the ground truth image as a reference

	5x5	**5x3**	**3x3**	**No GF**
Average performance for BSD benchmark (%)	22.2%	22%	21.8%	19.6%
Average performance for noisy BSD benchmark (%)	21%	20%	20.4%	8.8%

The subjective analysis is also provided for two images from the BSD benchmark. The images were selected by considering different characteristics as follows: i) the image in Figure 5.13 has a lower density of edges where most of the content is homogeneous pixels; ii) the image in Figure 5.14 has a higher density of edges.

In Figure 5.13 is shown in (a) the original image followed by its edge detection for 5x5, 5x3, 3x3, and no use of a Gaussian filter in Figure 5.13 (b) to (e), respectively. In this context, there is no substantial difference in the detected edges when observed all the

configurations. Only the configuration in Figure 5.13 (e) presents more artifacts in the image. This is because the Gaussian filter is responsible for attenuating the presence of these high-frequency components. Since in Figure 5.13 (e) the Gaussian filter is not used, the edge detector is prone to detect these components. When Gaussian noise in Figure 5.13 (f) corrupts the image, the respective configurations in Figure 5.13 (g) to (j) present much more artifacts than observed in Figure 5.13 (b) to (e). This is because the Gaussian filter does not totally attenuate the presence of noise, and the edge detector has its quality degraded by noise. On the other hand, one can observe that the 5x5 version is the one which is less affected by noise, while in Figure 5.13 (h) and (i), the 5x3 and 3x3 filtered images present more artifacts. In Figure 5.13 (j) the image is totally corrupted by noise and no visible edge information is observed.

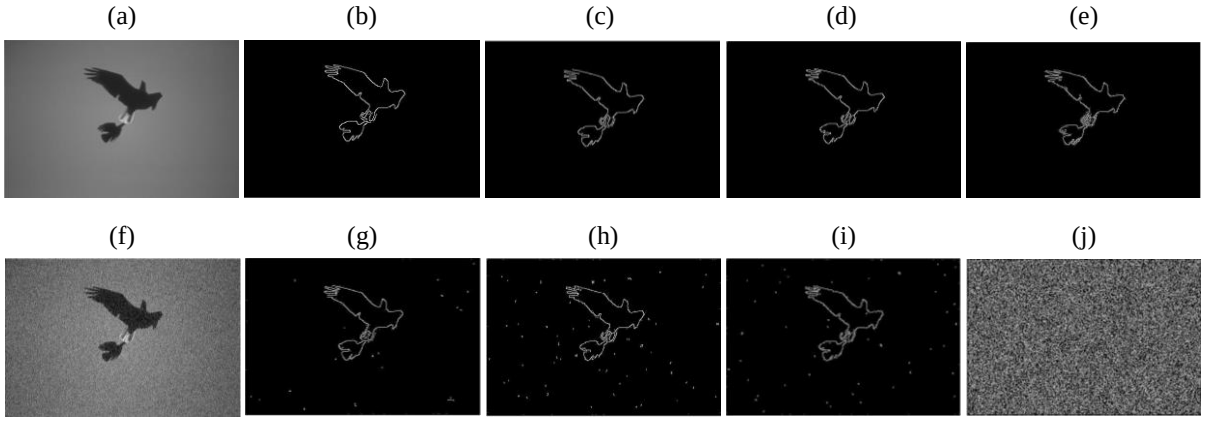

Figure 5.13 - Subjective edge detection analysis for image “135069.jpg”. (a) original image, (b) 5x5 configurable architecture detecting edges in the original image, (c) 5x3 configurable architecture detecting edges in the original image, (e) 3x3 configurable architecture detecting edges in the original image, and (f) Canny edge detector detecting edges in the original image without the use of Gaussian filter, (f) image corrupted by Gaussian noise with $\mu = 0$, $\sigma^2 = 0.01$, (g) 5x5 configurable architecture detecting edges in the noisy image, (h) 5x3 configurable architecture detecting edges in the noisy image, (i) 3x3 configurable architecture detecting edges in the noisy image, and (j) Canny edge detector detecting edges in the noisy image without the use of Gaussian filter.

A similar observation can be performed in Figure 5.14. For the original image in Figure 5.14 (a), the 5x5 configurable architecture presents the lowest level of artifacts when compared to the remaining configurations in Figure 5.14 (c) to (e).

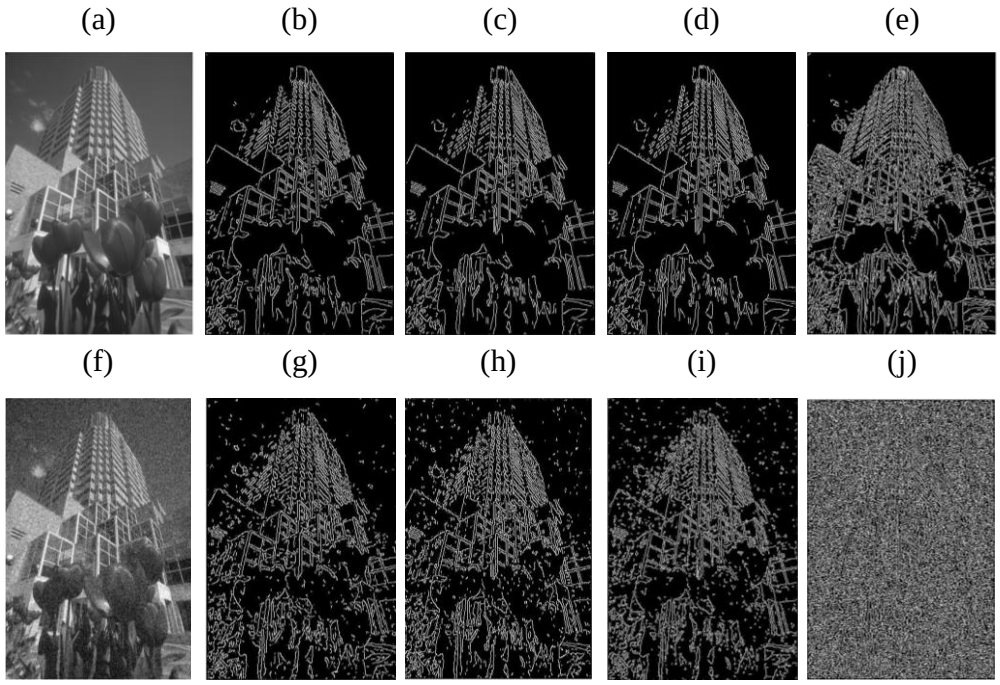

Figure 5.14 - Subjective edge detection analysis for image “86000.jpg”. (a) original image, (b) 5x5 configurable architecture detecting edges in the original image, (c) 5x3 configurable architecture detecting edges in the original image, (e) 3x3 configurable architecture detecting edges in the original image, and (f) Canny edge detector detecting edges in the original image without the use of Gaussian filter, (f) image corrupted by Gaussian noise with $\mu = 0$, $\sigma^2 = 0.01$, (g) 5x5 configurable architecture detecting edges in the noisy image, (h) 5x3 configurable architecture detecting edges in the noisy image, (i) 3x3 configurable architecture detecting edges in the noisy image, and (j) Canny edge detector detecting edges in the noisy image without the use of Gaussian filter.

On the other hand, in Figure 5.14 (e) there are more edges detected than in Figure 5.14 (b). In other words, depending on the image, the absence of Gaussian filter may increase the level of details in detected edges. When the image is corrupted by Gaussian noise as shown in Figure 5.14 (f), the configurable responses are presented in Figure 5.14 (g) to (j). As previously mentioned, the 5x5 configuration in Figure 5.14 (g) is the one which presents the lowest level of artifacts. Following the same idea of the previous subjective analysis, when the Gaussian filter is disabled in Figure 5.14 (j), the edge detection is substantially corrupted by noise. One can conclude that, depending on the application, more or fewer artifacts production can be managed to alleviate energy consumption.

5.2.1.2Energy Efficiency Analysis

The proposed configurable architecture is described in VHDL and synthesized by using the RTL Compiler for the 45 nm Nangate FreePDK and considering the same frequency

target of 300 MHz exercised in Chapter 4. For the energy efficiency analysis, the power is estimated through the toggle count extraction by simulating the pre-layout Verilog netlist with clock gate logic inserted. The simulation considers 10,000 convolution operations per image or a total of 40,000 convolution operations. Based on that, six different profiles are considered. i)5x5, 5x5, 5x5, and 5x5; ii) 5x3, 5x3, 5x3, and 5x3; iii) 3x3, 3x3, 3x3, and 3x3; iv) no use of Gaussian filter for all the convolution operations; v) 5x5, 5x3, 3x3, and no use of Gaussian filter; and vi) 5x3, 5x3, 3x3, and 3x3. Those profiles consider 10,000 pixels from 4 different images being filtered by each enumerated window sizes. One can observe that the non-configurable profile is the baseline since it is the most accurate architecture without overhead in terms of power and area. Therefore, the power reduction results for the remaining profiles consider the baseline architecture as being the reference.

Synthesis results for 300 MHz are shown in Table 5.11. The power dissipation results are shown for each one of the evaluated profiles. In comparison with the baseline non-configurable architecture, the maximum power reductions are for the profile iv), that is the case where all the images are processed without a Gaussian filter. The maximum power reduction is of 64%. The minimum power reduction is for the profile ii), that is the case where the 5x3 Gaussian filter processes all the images. The minimum power reduction is of 29.1%.

Table 5.11 – Power dissipation and Area results @ 300 MHz

Power Analysis (mW)	*Total*	*Dynamic*
i)	8.2	7.8
ii)	5.4	5.1
iii)	5.2	4.9
iv)	2.9	2.6
v)	5.3	5.0
vi)	5.3	5.0
Non- configurable Canny edge detector	7.5	7.2
Area Analysis (μm^2)	# of cells	Area
Pruning Configurable Canny edge detector	10645	18272
Non-configurable Canny edge detector	10194	17408

The number of cells and area in μm^2 for each frequency operation are shown in Table 5.11. Since the configurable architecture netlist is the same for all the profiles, the area results are independent of the profile under evaluation.

The overhead regarding area and power is also presented for the proposed configurable architecture in comparison with the non-configurable solution. The dissipated power of the profile i) results in an overhead of 8.5%. On the other hand, for the other profiles, which consider pruning, there is no overhead in terms of power dissipation. The area overhead for the configurable architecture is of 5%. In sum, low overhead for both power dissipation and area are experienced, in favor of significant power reductions provided by the configurable architecture. When comparing with state-of-the-art related works in (CABELLO et al., 2015; JAISWAL et al., 2015; KAUSHIK; KUMAR, 2015), one can conclude that this work provides a configurable kernel which can be set according to the IoT system requirements instead of proposing a static solution. The work in (HSIAO et al., 2006) also proposes a configurable kernel solution, but that work is only focused on increasing the application quality without energy efficiency analysis.

5.3 Summary of the chapter

In this chapter accuracy-configurable architectures were proposed for video coding and Canny edge detection. As previously mentioned in Chapter 2, this is a trending approach in approximate computing scope when considered run-time techniques. These accuracy-configurable approaches were proposed in architectural level, and the impact in application quality was evaluated by using the same proposed methodology. One can conclude that the proposed methodology enabled quality-power-performance profiling for accuracy-configurable approaches.

6 CONCLUSIONS AND FUTURE WORK

As previously mentioned, many challenges have arisen over recent years regarding power issue and energy efficiency for digital CMOS design. Furthermore, the advances in semiconductor industry enabled more and more complexity in end-user applications. These points were fundamental to promote investigation of emerging energy efficient techniques for CMOS design. From all the recent researches, approximate computing is one trending approach which can be used when end-user applications can sustain distinct levels of approximation in favor of energy efficiency. On the other hand, the approximate computing concept still faces its intrinsic difficulties. One of these challenges is the low integration of approximate computing techniques from low-level up to end-user application stack. In this context, state-of-the-art researchers have difficulty in stating the impact of a low-level approximation in the application stack.

This thesis is focused on this recent approximate computing challenge. Due to this inherent difficulty in approximate computing scope, this thesis presented novel contributions through the proposition of a simulation-based design flow. The simulation-based methodology is the key approach to drive integration between distinct abstraction layers. Furthermore, this technique enables comprehensive characterization of multiple quality-power-performance profiles from the arithmetic level up to application stack. This study introduced an important and desired set of findings which can be of scientific and industrial interest when considering approximate computing for digital CMOS design.

In sum, one can observe that this thesis provides substantial new findings when compared to the state-of-the-art methodologies. None of the related works exercised a solid integration among the approximate techniques and real-world applications. Based on that, this thesis provided analysis of three different approximation-tolerant applications: i) FIR filters for audio processing, ii) Canny edge detection, and iii) block matching operation for state-of-the-art video coding. For all these high-level applications, different approximations in arithmetic and architectural levels were exercised by adopting the proposed methodology.

One can conclude that the systematic exploration of design- and run-time techniques provided multiple quality-power-performance profiles.

In Chapter 4 is demonstrated the proposed methodology for the evaluation of approximate computing techniques in arithmetic and architectural levels during design-time. A more comprehensive observation was possible through the use of the proposed methodology to validate state-of-the-art approximate adders for FIR filters and Canny edge applications. Differently from the limited exploration of these adders in literature, this work enabled a broader evaluation of approximate configurations by considering the proposition of search heuristics. These proposed heuristics are a valid approach to sustain the simulation-based design flow since its essence is driven by simulating the real-world applications. Results in Chapter 4 evidenced that multiple levels of quality can be achieved followed by respective energy reductions. One can conclude that maximum energy reductions of 25.7% and 57.4% were achieved when regarding FIR filters for audio processing and Canny edge detection, respectively. For the applications under evaluation, the copy adder approximation is more energy efficient than the ETAI approximated designs. For the same quality level, the copy adders presented lower energy consumption than the ETAI designs. This can be explained by the presence of a clever technique to speculate carry-in in copy adders followed by the simpler technique of adopting buffers.

In Chapter 5 a different trend in approximate computing is exercised, where the run-time configurable solution is adopted to manage multiple quality levels and energy consumption. Video applications can leverage this run-time configuration capability since this media is intrinsically dynamic. The approximate SATD block matching technique for HEVC and the Canny edge detection for images or videos were explored in this context. The same proposed methodology was adopted where the main objective is to observe the impact of adaptive circuit pruning regarding application quality and energy consumption. Results in Chapter 5 show energy reductions ranging from 6.8% up to 43.65% for the SATD application with BD-PSNR ranging from -0.008 dB to -0.097 dB. The energy overhead is of 11.9% when the entire SATD is enabled. On the other hand, this is compensated every time a pruning-based configuration is adopted. For the Canny edge detection case study energy reductions ranging from 29.1% up to 64% regarding different Gaussian kernel sizes configuration. The energy overhead is of 8.5%. Objective and subjective quality analysis is provided for this application. One can conclude that, depending on the application, the accuracy-configurable

Canny edge detection is amenable to adopt, because energy efficiency is achieved with a low level of artifacts.

6.1 Future Work

Even considering the dense and deep set of analyses performed in this thesis, many aspects should be considered as future work. More approximation-tolerant applications and state-of-the-art approximate adders can be exercised in the proposed simulation-based methodology. This will help to compare upcoming contributions in the arithmetic level for approximate computing scope. Furthermore, approximate multipliers can also be explored in applications by adopting the proposed methodology. Another essential point is to consider new heuristics to approximate the architectures. In addition, this methodology can be explored for situations where the input signals bit-length is reduced. This trending and classical technique was not considered in this thesis scope because the major objective was to propose cross-layer integration among different state-of-the-art approximate computing techniques. In the configurable run-time scope, much more analyses can be performed considering the design of adaptive decider blocks for the proposed configurable architectures. Therefore, the input content could be used to determine which approximate configuration is the best. For instance, the decider block can observe the level of input noise of a given image to enable one configurable profile for the Canny edge detection. Furthermore, the evaluation of new applications can be adopted in this context.

6.2 Publications by the author

This section presents the list of publications by the author since the beginning of the Ph.D. until its conclusion.

6.2.1 Journal Paper

SOARES, L. B.; COSTA, E. A. C.; BAMPI, S. Design of area and energy-efficient digital CMOS FIR filters with approximate adder circuits. **Journal on Analog Integrated Circuits and Signal Processing**. v. 89, n. 1, p. 99 – 109, Springer. Out, 2016. (QUALIS B2)

6.2.2 Conference Papers

SOARES, L. B., ROSA, A.L.R, BAMPI, S. COSTA, E.A.C. Near-threshold computing for very wide frequency scaling: approximate adders to rescue performance. In: IEEE International NEW on Circuits and Systems Conference, 2015. p. 1–4.

SOARES, L. B., COSTA, E.A.C. BAMPI, S. Approximate adder synthesis for area- and energy-efficient FIR filters in CMOS VLSI. In: IEEE International NEW on Circuits and Systems Conference, 2015. p. 1–4.

SOARES, L.B., DINIZ, C. M., COSTA, E.A.C., BAMPI, S. A novel pruned-based algorithm for energy-efficient SATD operation in the HEVC coding. In: 29th Symposium on Integrated Circuits and Systems Design (SBCCI), 2016. P. 1-6.

ROSA, A.L.R, **SOARES, L.B.**, STANGHERLIN, K.H., BAMPI, S. Designing CMOS for near-threshold minimum-energy operation and extremely wide V-F scaling. In: 28th Symposium on Integrated Circuits and Systems Design (SBCCI), 2015. p. 1-6.

OLIVEIRA, J., **SOARES,L.B.**, COSTA, E.A.C., BAMPI, S. Exploiting approximate adder circuits for power-efficient Gaussian and Gradient filters for Canny edge detector algorithm. In: Latin American Symposium on Circuits and Systems (LASCAS), 2016. p. 1-4.

OLIVEIRA, J., **SOARES,L.B.**, COSTA, E.A.C., BAMPI, S. Energy-efficient Gaussian filter for image processing using approximate adder circuits. In: IEEE International Conference on Electronics, Circuits, and Systems (ICECS), 2015. p. 1-4.

PAIM, G., **SOARES,L.B.**, OLIVEIRA, J., COSTA, E.A.C., BAMPI, S. A power-efficient imprecise radix-4 multiplier applied to high resolution audio processing. In: IEEE International Conference on Electronics, Circuits, and Systems (ICECS), 2016. p. 1-4.

PAIM, G.; **SOARES, L.B.**; FERREIRA, R.; COSTA, E.; BAMPI, S. Pruning and approximation of coefficients for power-efficient 2-D Discrete Tchebichef Transform. In: 2017 15th IEEE International New Circuits and Systems Conference (NEWCAS), 2017, Strasbourg. p. 25.

SAPPER, A.N.; **SOARES, L.B.**; COSTA, E.; BAMPI, S. Exploring the combination of number of bits and number of iterations for a power-efficient fixed-point CORDIC implementation. In: 2017 24th IEEE International Conference on Electronics, Circuits and Systems (ICECS), 2017.

MACEDO. M.; **SOARES, L.B.**; SILVEIRA, B.; DINIZ, C.; COSTA, E.; BAMPI, S. Exploring the use of parallel prefix adder topologies into approximate adder circuits. In: 24th IEEE International Conference on Electronics, Circuits and Systems (ICECS), 2017.

SOARES, L.B.; MACEDO. M.; DINIZ, C.; COSTA, E.; BAMPI, S. Exploring Power-Performance-Quality Tradeoff of Approximate Adders for Energy Efficient Sobel Filtering. In: 2018 IEEE 9th Latin American Symposium on Circuits & Systems (LASCAS), 2018.

REFERENCES

AES. **AES standard method for digital audio engineering — Measurement of digital audio equipment**, 1998.

AKSOY, Levent et al. Optimization of Area and Delay at Gate-Level in Multiple Constant Multiplications. In: 13TH EUROMICRO CONFERENCE ON DIGITAL SYSTEM DESIGN: ARCHITECTURES, METHODS AND TOOLS 2010, **Anais**... . In: 13TH EUROMICRO CONFERENCE ON DIGITAL SYSTEM DESIGN: ARCHITECTURES, METHODS AND TOOLS. : IEEE, 2010. Disponível em: <http://ieeexplore.ieee.org/document/5615614/>. Acesso em: 22 fev. 2017.

ATZORI, Luigi; IERA, Antonio; MORABITO, Giacomo. The Internet of Things: A survey. **Computer Networks**, [s. l.], v. 54, n. 15, p. 2787–2805, 2010.

BAILEY, Brian. **FinFET Scaling Reaches Thermal Limit**. 2016. Disponível em: <https://semiengineering.com/dennards-law-and-the-finfet/>.

BAILEY, Donald G. The advantages and limitations of high level synthesis for FPGA based image processing. In: PROCEEDINGS OF THE 9TH INTERNATIONAL CONFERENCE ON DISTRIBUTED SMART CAMERAS 2015, **Anais**... . In: PROCEEDINGS OF THE 9TH INTERNATIONAL CONFERENCE ON DISTRIBUTED SMART CAMERAS. : ACM Press, 2015. Disponível em: <http://dl.acm.org/citation.cfm?doid=2789116.2789145>. Acesso em: 17 mar. 2017.

BELLOCH, Jose A. et al. Multichannel massive audio processing for a generalized crosstalk cancellation and equalization application using GPUs. **Integrated Computer-Aided Engineering**, [s. l.], v. 20, n. 2, p. 169–182, 2013.

BELLOCH, Jose A. et al. Accelerating multi-channel filtering of audio signal on ARM processors. **The Journal of Supercomputing**, [s. l.], v. 73, n. 1, p. 203–214, 2016.

BLEIDT, Robert et al. Object-Based Audio: Opportunities for Improved Listening Experience and Increased Listener Involvement. **SMPTE Motion Imaging Journal**, [s. l.], v. 124, n. 5, p. 1–13, 2015.

BOSSEN, F. **Common test conditions and software configurations**JCT-VC-L1100,JCT-VC, Geneva, , 2013.

CABELLO, Frank et al. Implementation of a fixed-point 2D Gaussian Filter for Image Processing based on FPGA. In: SIGNAL PROCESSING: ALGORITHMS, ARCHITECTURES, ARRANGEMENTS, AND APPLICATIONS (SPA), 2015 2015, **Anais**... . In: 2015 SIGNAL PROCESSING: ALGORITHMS, ARCHITECTURES, ARRANGEMENTS, AND APPLICATIONS (SPA). : IEEE, 2015. Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7365108>. Acesso em: 17 nov. 2016.

CANNY, John. A computational approach to edge detection. **IEEE Transactions on pattern analysis and machine intelligence**, [s. l.], v. PAMI-8, n. 6, p. 679–698, 1986.

CHEIKH, Taieb Lamine Ben et al. Fast and accurate implementation of Canny edge detector on embedded many-core platform. In: NEW CIRCUITS AND SYSTEMS CONFERENCE (NEWCAS), 2014 IEEE 12TH INTERNATIONAL 2014, **Anais**... . In: NEW CIRCUITS

AND SYSTEMS CONFERENCE (NEWCAS), 2014 IEEE 12TH INTERNATIONAL. : IEEE, 2014. Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6934067>. Acesso em: 17 nov. 2016.

CHIEN, Shao-Yi et al. Distributed computing in IoT: system-on-a-chip for smart cameras as an example. In: THE 20TH ASIA AND SOUTH PACIFIC DESIGN AUTOMATION CONFERENCE 2015, **Anais**... . In: THE 20TH ASIA AND SOUTH PACIFIC DESIGN AUTOMATION CONFERENCE. : IEEE, 2015. Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7058993>. Acesso em: 17 nov. 2016.

CISCO. **The Internet of Things: How the Next Evolution of the Internet Is Changing Everything**, 2011. Disponível em: <https://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf>. Acesso em: 30 jan. 2017.

CISCO. **Cisco Visual Networking Index: Forecast and Methodology, 2015–2020**, 2016. Disponível em: <http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html>. Acesso em: 30 jan. 2017.

CONG, Jason et al. Accelerator-Rich Architectures: Opportunities and Progresses. In: PROCEEDINGS OF THE 51ST ANNUAL DESIGN AUTOMATION CONFERENCE 2014, **Anais**... . In: PROCEEDINGS OF THE 51ST ANNUAL DESIGN AUTOMATION CONFERENCE. : ACM Press, 2014. Disponível em: <http://dl.acm.org/citation.cfm?doid=2593069.2596667>. Acesso em: 1 fev. 2017.

DENNARD, Robert H. et al. Design of ion-implanted MOSFET's with very small physical dimensions. **IEEE Journal of Solid-State Circuits**, [s. l.], v. 9, n. 5, p. 256–268, 1974.

DENNARD, Robert H. Past Progress and Future Challenges in LSI Technology: From DRAM and Scaling to Ultra-Low-Power CMOS. **IEEE Solid-State Circuits Magazine**, [s. l.], v. 7, n. 2, p. 29–38, 2015.

DINIZ, Claudio. **Dedicated and Reconfigurable Hardware Accelerators for High Efficiency Video Coding Standard**. 2015. Ph.D. - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2015. Disponível em: <https://www.lume.ufrgs.br/bitstream/handle/10183/118394/000969495.pdf?sequence=1>. Acesso em: 2 jan. 2017.

ERCEGOVAC, Miloš D.; LANG, Tomás. **Digital arithmetic**. San Francisco, CA: Morgan Kaufmann Publishers, 2004. Disponível em: <http://site.ebrary.com/id/10203578>. Acesso em: 17 jul. 2014.

ESMAEILZADEH, Hadi et al. Dark silicon and the end of multicore scaling. In: ACM SIGARCH COMPUTER ARCHITECTURE NEWS 2011, **Anais**... . In: ACM SIGARCH COMPUTER ARCHITECTURE NEWS. : ACM, 2011. Disponível em: <http://dl.acm.org/citation.cfm?id=2000108>. Acesso em: 28 jan. 2017.

ESMAEILZADEH, Hadi et al. Architecture support for disciplined approximate programming. In: ACM SIGARCH COMPUTER ARCHITECTURE NEWS 2012, New York. **Anais**... . In: ARCHITECTURAL SUPPORT FOR PROGRAMMING LANGUAGES AND OPERATING SYSTEMS. New York: ACM, 2012. Disponível em: <http://dl.acm.org/citation.cfm?id=2151008>. Acesso em: 30 abr. 2014.

GENTSOS, Christos et al. Real-time canny edge detection parallel implementation for FPGAs. In: 17TH IEEE INTERNATIONAL CONFERENCE ON ELECTRONICS, CIRCUITS, AND SYSTEMS (ICECS) 2010, **Anais**... . In: 17TH IEEE INTERNATIONAL CONFERENCE ON ELECTRONICS, CIRCUITS, AND SYSTEMS (ICECS). : IEEE, 2010. Disponível em: <http://ieeexplore.ieee.org/abstract/document/5724558/>. Acesso em: 17 mar. 2017.

GUPTA, Vaibhav et al. IMPACT: imprecise adders for low-power approximate computing. In: PROCEEDINGS OF THE 17TH IEEE/ACM INTERNATIONAL SYMPOSIUM ON LOW-POWER ELECTRONICS AND DESIGN 2011, Fukuoka. **Anais**... . In: INTERNATIONAL SYMPOSIUM ON LOW-POWER ELECTRONICS AND DESIGN. Fukuoka: IEEE Press, 2011. Disponível em: <http://dl.acm.org/citation.cfm?id=2016898>. Acesso em: 30 abr. 2014.

GUPTA, Vaibhav et al. Low-Power Digital Signal Processing Using Approximate Adders. **IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems**, [s. l.], v. 32, n. 1, p. 124–137, 2013.

HAMEED, Rehan et al. Understanding sources of inefficiency in general-purpose chips. In: ACM SIGARCH COMPUTER ARCHITECTURE NEWS 2010, **Anais**... . In: ACM SIGARCH COMPUTER ARCHITECTURE NEWS. : ACM, 2010. Disponível em: <http://dl.acm.org/citation.cfm?id=1815968>. Acesso em: 1 fev. 2017.

HAN, Jie; ORSHANSKY, Michael. Approximate computing: An emerging paradigm for energy-efficient design. In: TEST SYMPOSIUM (ETS), 2013 18TH IEEE EUROPEAN 2013, Avignon. **Anais**... . In: IEEE EUROPEAN TEST SYMPOSIUM. Avignon: IEEE, 2013. Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6569370>. Acesso em: 30 abr. 2014.

HARRIS, David. A taxonomy of parallel prefix networks. In: THE THRITY-SEVENTH ASILOMAR CONFERENCE ON SIGNALS, SYSTEMS & COMPUTERS, 2003 2003, **Anais**... . In: THE THRITY-SEVENTH ASILOMAR CONFERENCE ON SIGNALS, SYSTEMS & COMPUTERS, 2003. [s.l: s.n.]

HE, Ku; GERSTLAUER, Andreas; ORSHANSKY, Michael. Controlled timing-error acceptance for low energy IDCT design. In: DESIGN, AUTOMATION & TEST IN EUROPE CONFERENCE & EXHIBITION (DATE), 2011 2011, Grenoble. **Anais**... . In: DESIGN, AUTOMATION & TEST IN EUROPE CONFERENCE & EXHIBITION. Grenoble: IEEE, 2011. Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5763129>. Acesso em: 30 abr. 2014.

HE, Wenhao; YUAN, Kui. An improved Canny edge detector and its realization on FPGA. In: 7TH WORLD CONGRESS ON INTELLIGENT CONTROL AND AUTOMATION 2008, **Anais**... . In: 7TH WORLD CONGRESS ON INTELLIGENT CONTROL AND AUTOMATION. : IEEE, 2008. Disponível em: <http://ieeexplore.ieee.org/abstract/document/4594570/>. Acesso em: 17 mar. 2017.

HENKEL, Jörg et al. New trends in dark silicon. In: 2015, **Anais**... : ACM Press, 2015. Disponível em: <http://dl.acm.org/citation.cfm?doid=2744769.2747938>. Acesso em: 11 fev. 2018.

HSIAO, Pei-Yung et al. A parameterizable digital-approximated 2D Gaussian smoothing filter for edge detection in noisy image. In: 2006 IEEE INTERNATIONAL SYMPOSIUM ON CIRCUITS AND SYSTEMS 2006, **Anais**... . In: 2006 IEEE INTERNATIONAL SYMPOSIUM ON CIRCUITS AND SYSTEMS. : IEEE, 2006. Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1693303>. Acesso em: 17 nov. 2016.

HUANG, Jiawei; LACH, John; ROBINS, Gabriel. A Methodology for Energy-Quality Tradeoff Using Imprecise Hardware. In: DESIGN AUTOMATION CONFERENCE (DAC), 2012 49TH ACM/EDAC/IEEE 2012, San Francisco. **Anais**... . In: DESIGN AUTOMATION CONFERENCE. San Francisco: IEEE, 2012.

ISO/IEC. **ISO/IEC 11172-3:1993. Information technology -- Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s -- Part 3: Audio.**ISO/IEC, , 1993.

ITU-T; ISO/IEC. **Advanced video coding, ITU-T Recommendation H.264 and ISO/IEC 14496-10 (MPEG-4 AVC)**ITU-T, ISO/IEC, , 2011.

ITU-T; ISO/IEC. **High Efficiency Video Coding, ITU-T Recommendation H.265 and ISO/IEC 23008-2**ITU-T, ISO/IEC, , 2013.

IYER, Ravi. Accelerator-rich architectures: Implications, opportunities and challenges. In: DESIGN AUTOMATION CONFERENCE (ASP-DAC), 2012 17TH ASIA AND SOUTH PACIFIC 2012, **Anais**... . In: DESIGN AUTOMATION CONFERENCE (ASP-DAC), 2012 17TH ASIA AND SOUTH PACIFIC. : IEEE, 2012. Disponível em: <http://ieeexplore.ieee.org/abstract/document/6164927/>. Acesso em: 1 fev. 2017.

JAISWAL, Ankur et al. SPAA-Aware 2D Gaussian Smoothing Filter Design Using Efficient Approximation Techniques. In: 28TH INTERNATIONAL CONFERENCE ON VLSI DESIGN 2015, **Anais**... . In: 28TH INTERNATIONAL CONFERENCE ON VLSI DESIGN. : IEEE, 2015. Disponível em: <http://ieeexplore.ieee.org/document/7031756/>. Acesso em: 17 nov. 2016.

KAHNG, Andrew B.; KANG, Seokhyeong. Accuracy-configurable adder for approximate arithmetic designs. In: PROCEEDINGS OF THE 49TH ANNUAL DESIGN AUTOMATION CONFERENCE 2012, San Francisco. **Anais**... . In: DESIGN AUTOMATION CONFERENCE. San Francisco: ACM, 2012. Disponível em: <http://dl.acm.org/citation.cfm?id=2228509>. Acesso em: 30 abr. 2014.

KAMILARIS, Andreas; PITSILLIDES, Andreas. Mobile Phone Computing and the Internet of Things: A Survey. **IEEE Internet of Things Journal**, [s. l.], v. 3, n. 6, p. 885–898, 2016.

KANG, Yesung; KIM, Jaewoo; KANG, Seokhyeong. Novel approximate synthesis flow for energy-efficient FIR filter. In: 2016 IEEE 34TH INTERNATIONAL CONFERENCE ON COMPUTER DESIGN (ICCD) 2016, **Anais**... . In: 2016 IEEE 34TH INTERNATIONAL CONFERENCE ON COMPUTER DESIGN (ICCD). : IEEE, 2016. Disponível em: <http://ieeexplore.ieee.org/abstract/document/7753266/>. Acesso em: 18 fev. 2017.

KAUSHIK, Sharda; KUMAR, NVS V. Vijay. Energy-efficient approximate 2D Gaussian smoothing filter for error tolerant applications. In: ADVANCE COMPUTING CONFERENCE (IACC), 2015 IEEE INTERNATIONAL 2015, **Anais**... : IEEE, 2015.

Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7154815>. Acesso em: 17 nov. 2016.

KHUDIA, Daya Shanker et al. Quality Control for Approximate Accelerators by Error Prediction. **IEEE Design & Test**, [s. l.], v. 33, n. 1, p. 43–50, 2016.

KIM, Mikang et al. Liquid-level estimation using region-based segmentation for automatic beverage refilling service. In: 2015 INTERNATIONAL SYMPOSIUM ON CONSUMER ELECTRONICS (ISCE) 2015, Madrid. **Anais**... . In: 2015 INTERNATIONAL SYMPOSIUM ON CONSUMER ELECTRONICS (ISCE). Madrid: IEEE, 2015. Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7177811>. Acesso em: 17 nov. 2016.

KUON, Ian; ROSE, Jonathan. Measuring the Gap Between FPGAs and ASICs. **IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems**, [s. l.], v. 26, n. 2, p. 203–215, 2007.

LEE, Juseong; TANG, Hoyoung; PARK, Jongsun. Energy Efficient Canny Edge Detector for Advanced Mobile Vision Applications. **IEEE Transactions on Circuits and Systems for Video Technology**, [s. l.], p. 1–1, 2016.

LI, Xiaoyang; JIANG, Jie; FAN, Qiaoyun. An improved real-time hardware architecture for Canny edge detection based on FPGA. In: THIRD INTERNATIONAL CONFERENCE ON INTELLIGENT CONTROL AND INFORMATION PROCESSING (ICICIP) 2012, Dalian. **Anais**... . In: THIRD INTERNATIONAL CONFERENCE ON INTELLIGENT CONTROL AND INFORMATION PROCESSING (ICICIP). Dalian: IEEE, 2012. Disponível em: <http://ieeexplore.ieee.org/abstract/document/6391408/>. Acesso em: 17 mar. 2017.

LIU, Cong; HAN, Jie; LOMBARDI, Fabrizio. An Analytical Methodology for Evaluating the Error Characteristics of Approximate Adders. **IEEE Transactions on Computers**, [s. l.], v. 64, n. 5, p. 1268–1281, 2015.

MAHDIANI, H. R. et al. Bio-Inspired Imprecise Computational Blocks for Efficient VLSI Implementation of Soft-Computing Applications. **IEEE Transactions on Circuits and Systems I: Regular Papers**, [s. l.], v. 57, n. 4, p. 850–862, 2010.

MARKOVIC, D. et al. Methods for true energy-performance optimization. **IEEE Journal of Solid-State Circuits**, [s. l.], v. 39, n. 8, p. 1282–1293, 2004.

MARTIN, David et al. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: EIGHTH IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION (ICCV) 2001, **Anais**... . In: EIGHTH IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION (ICCV). : IEEE, 2001. Disponível em: <http://ieeexplore.ieee.org/abstract/document/937655/>. Acesso em: 17 mar. 2017.

MAZAHIR, Sana et al. Probabilistic Error Modeling for Approximate Adders. **IEEE Transactions on Computers**, [s. l.], v. 66, n. 3, p. 515–530, 2017.

MISHRA, A. K.; BARIK, R.; PAUL, S. iACT: A Software-Hardware Methodology for Understanding the Scope of Approximate Computing. In: WORKSHOP ON APPROXIMATE COMPUTING ACROSS THE SYSTEM STACK 2014, **Anais**... . In:

WORKSHOP ON APPROXIMATE COMPUTING ACROSS THE SYSTEM STACK. [s.l: s.n.]

MOHAPATRA, Debabrata et al. Design of voltage-scalable meta-functions for approximate computing. In: DESIGN, AUTOMATION & TEST IN EUROPE CONFERENCE & EXHIBITION (DATE), 2011 2011, Grenoble. **Anais**... . In: DESIGN, AUTOMATION & TEST IN EUROPE CONFERENCE & EXHIBITION. Grenoble: IEEE, 2011. Disponível em: <http://www.date-conference.com/proceedings/PAPERS/2011/DATE11/PDFFILES/08.3_4.PDF>. Acesso em: 30 abr. 2014.

MOORE, Gordon E. Cramming more components onto integrated circuits. **Electronics**, [s. l.], v. 38, n. 8, 1965.

NALBANDIAN, Shayan. A survey on Internet of Things: Applications and challenges. In: INTERNATIONAL CONGRESS ON TECHNOLOGY, COMMUNICATION AND KNOWLEDGE (ICTCK) 2015, **Anais**... . In: INTERNATIONAL CONGRESS ON TECHNOLOGY, COMMUNICATION AND KNOWLEDGE (ICTCK). : IEEE, 2015. Disponível em: <http://ieeexplore.ieee.org/abstract/document/7582664/>. Acesso em: 28 jan. 2017.

NEOH, Hong Shan; HAZANCHUK, Asher. Adaptive edge detection for real-time video processing using FPGAs. **Global Signal Processing**, [s. l.], v. 7, n. 3, p. 2–3, 2004.

NORDRUM, Amy. **Popular Internet of Things Forecast of 50 Billion Devices by 2020 Is Outdated**. 2016. Disponível em: <http://spectrum.ieee.org/tech-talk/telecom/internet/popular-internet-of-things-forecast-of-50-billion-devices-by-2020-is-outdated>. Acesso em: 30 jan. 2017.

PARK, Jongsun; CHOI, Jung Hwan; ROY, Kaushik. Dynamic Bit-Width Adaptation in DCT: An Approach to Trade Off Image Quality and Computation Energy. **IEEE Transactions on Very Large Scale Integration (VLSI) Systems**, [s. l.], v. 18, n. 5, p. 787–793, 2010.

PINCKNEY, Nathaniel; BLAAUW, David; SYLVESTER, Dennis. Low-Power Near-Threshold Design: Techniques to Improve Energy Efficiency Energy-efficient near-threshold design has been proposed to increase energy efficiency across a wid. **IEEE Solid-State Circuits Magazine**, [s. l.], v. 7, n. 2, p. 49–57, 2015.

POSSA, Paulo Ricardo et al. A Multi-Resolution FPGA-Based Architecture for Real-Time Edge and Corner Detection. **IEEE Transactions on Computers**, [s. l.], v. 63, n. 10, p. 2376–2388, 2014.

PROAKIS, John G.; MANOLAKIS, Dimitris G. **Digital Signal Processing: Principles, Algorithms, and Applications**. 3rd. ed. Upper Saddle River: Prentice-Hall, 1996.

QUALCOMM. **Qualcomm Immersive Audio: superior surround sound on your smartphone—without headphones**, 2015. Disponível em: <https://www.qualcomm.com/news/snapdragon/2015/04/17/qualcomm-immersive-audio-superior-surround-sound-your-smartphone-without>

RAHMANI, Amir M. et al. Reliability-Aware Runtime Power Management for Many-Core Systems in the Dark Silicon Era. **IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS**, [s. l.], v. 25, n. 2, p. 427–440, 2017.

RAO, Daggu Venkateshwar; VENKATESAN, Muthukumar. An efficient reconfigurable architecture and implementation of edge detection algorithm using Handle-C. In: INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY: CODING AND COMPUTING (ITCC) 2004, **Anais**... . In: INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY: CODING AND COMPUTING (ITCC). : IEEE, 2004. Disponível em: <http://ieeexplore.ieee.org/abstract/document/1286764/>. Acesso em: 17 mar. 2017.

REHMAN, Semeen et al. Architectural-space exploration of approximate multipliers. In: PROCEEDINGS OF THE 35TH INTERNATIONAL CONFERENCE ON COMPUTER-AIDED DESIGN 2016, Austin. **Anais**... . In: PROCEEDINGS OF THE 35TH INTERNATIONAL CONFERENCE ON COMPUTER-AIDED DESIGN. Austin: ACM Press, 2016. Disponível em: <http://dl.acm.org/citation.cfm?doid=2966986.2967005>. Acesso em: 11 fev. 2017.

ROSA, Andre Luis Rodeghiero et al. Designing CMOS for Near-Threshold Minimum-Energy Operation and Extremely Wide V-F Scaling. In: 28TH SYMPOSIUM ON INTEGRATED CIRCUITS AND SYSTEMS DESIGN (SBCCI) 2015, Salvador. **Anais**... . In: 28TH SYMPOSIUM ON INTEGRATED CIRCUITS AND SYSTEMS DESIGN (SBCCI). Salvador: ACM Press, 2015. Disponível em: <http://dl.acm.org/citation.cfm?doid=2800986.2801004>. Acesso em: 1 fev. 2017.

SANGEETHA, D.; DEEPA, P. An Efficient Hardware Implementation of Canny Edge Detection Algorithm. In: 29TH INTERNATIONAL CONFERENCE ON VLSI DESIGN AND 15TH INTERNATIONAL CONFERENCE ON EMBEDDED SYSTEMS (VLSID) 2016, Kolkata. **Anais**... . In: 29TH INTERNATIONAL CONFERENCE ON VLSI DESIGN AND 15TH INTERNATIONAL CONFERENCE ON EMBEDDED SYSTEMS (VLSID). Kolkata: IEEE, 2016. Disponível em: <http://ieeexplore.ieee.org/document/7434996/>. Acesso em: 17 mar. 2017.

SHAFIQUE, Muhammad et al. The EDA Challenges in the Dark Silicon Era: Temperature, Reliability, and Variability Perspectives. In: 51ST ANNUAL DESIGN AUTOMATION CONFERENCE 2014, San Francisco. **Anais**... . In: 51ST ANNUAL DESIGN AUTOMATION CONFERENCE. San Francisco: ACM Press, 2014. Disponível em: <http://dl.acm.org/citation.cfm?doid=2593069.2593229>. Acesso em: 17 nov. 2016.

SHAFIQUE, Muhammad et al. A low latency generic accuracy configurable adder. In: PROCEEDINGS OF THE 52ND ANNUAL DESIGN AUTOMATION CONFERENCE 2015, San Francisco. **Anais**... . In: DESIGN AUTOMATION CONFERENCE. San Francisco: ACM Press, 2015. Disponível em: <http://dl.acm.org/citation.cfm?doid=2744769.2744778>. Acesso em: 22 jun. 2015.

SHAFIQUE, Muhammad et al. INVITED: Cross-Layer Approximate Computing: From Logic to Architectures. In: 2016 53ND ACM/EDAC/IEEE DESIGN AUTOMATION CONFERENCE (DAC) 2016, Austin. **Anais**... . In: 2016 53ND ACM/EDAC/IEEE DESIGN AUTOMATION CONFERENCE (DAC). Austin

SILVEIRA, Bianca et al. SATD hardware architecture based on 8x8 Hadamard Transform for HEVC encoder. In: IEEE INTERNATIONAL CONFERENCE ON ELECTRONICS, CIRCUITS, AND SYSTEMS (ICECS) 2015, Cairo. **Anais**... . In: IEEE INTERNATIONAL CONFERENCE ON ELECTRONICS, CIRCUITS, AND SYSTEMS (ICECS). Cairo: IEEE, 2015. Disponível em: <http://ieeexplore.ieee.org/abstract/document/7440382/>. Acesso em: 22 fev. 2017.

SOARES, Leonardo Bandeira et al. Near-threshold computing for very wide frequency scaling: Approximate adders to rescue performance. In: IEEE 13TH INTERNATIONAL NEW CIRCUITS AND SYSTEMS CONFERENCE (NEWCAS) 2015, Grenoble. **Anais**... . In: IEEE 13TH INTERNATIONAL NEW CIRCUITS AND SYSTEMS CONFERENCE (NEWCAS). Grenoble: IEEE, 2015. Disponível em: <http://ieeexplore.ieee.org/abstract/document/7182030/>. Acesso em: 2 fev. 2017.

SOARES, Leonardo Bandeira et al. A novel pruned-based algorithm for energy-efficient SATD operation in the HEVC coding. In: 29TH SYMPOSIUM ON INTEGRATED CIRCUITS AND SYSTEMS DESIGN (SBCCI) 2016, Belo Horizonte. **Anais**... . In: 29TH SYMPOSIUM ON INTEGRATED CIRCUITS AND SYSTEMS DESIGN (SBCCI). Belo Horizonte: IEEE, 2016. Disponível em: <http://ieeexplore.ieee.org/abstract/document/7724049/>. Acesso em: 22 fev. 2017.

STANGHERLIN, Kleber H.; BAMPI, Sergio. Energy-speed exploration for very-wide range of dynamic VF scaling. In: 26TH SYMPOSIUM ON INTEGRATED CIRCUITS AND SYSTEMS DESIGN (SBCCI) 2013, Curitiba. **Anais**... . In: 26TH SYMPOSIUM ON INTEGRATED CIRCUITS AND SYSTEMS DESIGN (SBCCI). Curitiba: IEEE, 2013. Disponível em: <http://ieeexplore.ieee.org/abstract/document/6644884/>. Acesso em: 1 fev. 2017.

TZANETAKIS, G.; COOK, P. Musical genre classification of audio signals. **IEEE Transactions on Speech and Audio Processing**, [s. l.], v. 10, n. 5, p. 293–302, 2002.

VANNE, Jarno et al. Comparative Rate-Distortion-Complexity Analysis of HEVC and AVC Video Codecs. **IEEE Transactions on Circuits and Systems for Video Technology**, [s. l.], v. 22, n. 12, p. 1885–1898, 2012.

VENKATARAMANI, Swagath et al. SALSA: systematic logic synthesis of approximate circuits. In: PROCEEDINGS OF THE 49TH ANNUAL DESIGN AUTOMATION CONFERENCE 2012, San Francisco. **Anais**... . In: PROCEEDINGS OF THE 49TH ANNUAL DESIGN AUTOMATION CONFERENCE. San Francisco: ACM, 2012. Disponível em: <http://dl.acm.org/citation.cfm?id=2228504>. Acesso em: 29 jun. 2016.

VENKATARAMANI, Swagath et al. Approximate Computing and the Quest for Computing Efficiency. In: 2015 52ND ACM/EDAC/IEEE DESIGN AUTOMATION CONFERENCE (DAC) 2015, San Francisco. **Anais**... . In: 2015 52ND ACM/EDAC/IEEE DESIGN AUTOMATION CONFERENCE (DAC). San Francisco: ACM Press, 2015. Disponível em: <http://dl.acm.org/citation.cfm?doid=2744769.2744904>. Acesso em: 31 jan. 2017.

VERMA, Ajay K.; BRISK, Philip; IENNE, Paolo. Variable latency speculative addition: A new paradigm for arithmetic circuit design. In: PROCEEDINGS OF THE CONFERENCE ON DESIGN, AUTOMATION AND TEST IN EUROPE 2008, Munich. **Anais**... . In: PROCEEDINGS OF THE CONFERENCE ON DESIGN, AUTOMATION AND TEST IN

EUROPE. Munich: IEEE, 2008. Disponível em: <http://dl.acm.org/citation.cfm?id=1403679>. Acesso em: 30 abr. 2014.

VORONENKO, Yevgen; PÜSCHEL, Markus. Multiplierless multiple constant multiplication. **ACM Transactions on Algorithms**, [s. l.], v. 3, n. 2, p. 11- es, 2007.

XU, Qian et al. A Distributed Canny Edge Detector: Algorithm and FPGA Implementation. **IEEE Transactions on Image Processing**, [s. l.], v. 23, n. 7, p. 2944–2960, 2014.

XU, Qiang; MYTKOWICZ, Todd; KIM, Nam Sung. Approximate Computing: A Survey. **IEEE Design & Test**, [s. l.], v. 33, n. 1, p. 8–22, 2016.

YE, Rong et al. On reconfiguration-oriented approximate adder design and its application. In: PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON COMPUTER-AIDED DESIGN 2013, San Jose. **Anais**... . In: PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON COMPUTER-AIDED DESIGN. San Jose: IEEE Press, 2013. Disponível em: <http://dl.acm.org/citation.cfm?id=2561838>. Acesso em: 13 jul. 2016.

ZHU, Ning et al. Design of Low-Power High-Speed Truncation-Error-Tolerant Adder and Its Application in Digital Signal Processing. **IEEE Transactions on Very Large Scale Integration (VLSI) Systems**, [s. l.], v. 18, n. 8, p. 1225–1229, 2010. a.

ZHU, Ning et al. Enhanced Low-Power High-Speed Adder For Error-Tolerant Application. In: SOC DESIGN CONFERENCE (ISOCC), 2010 INTERNATIONAL 2010b, Seoul. **Anais**... . In: INTERNATIONAL SOC DESIGN CONFERENCE. Seoul: IEEE, 2010.

ZHU, Ning; GOH, Wang Ling; YEO, Kiat Seng. An enhanced low-power high-speed adder for error-tolerant application. In: PROCEEDINGS OF THE 2009 12TH INTERNATIONAL SYMPOSIUM ON INTEGRATED CIRCUITS, ISIC'09. 2009, Singapore. **Anais**... . In: 12TH INTERNATIONAL SYMPOSIUM ON INTEGRATED CIRCUITS. Singapore: IEEE, 2009. Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5403865>. Acesso em: 30 abr. 2014.

ZUBAL, M.; LOJKA, T.; ZOLOTOVÁ, I. IoT gateway and industrial safety with computer vision. In: IEEE 14TH INTERNATIONAL SYMPOSIUM ON APPLIED MACHINE INTELLIGENCE AND INFORMATICS 2016, Herl'any. **Anais**... . In: IEEE 14TH INTERNATIONAL SYMPOSIUM ON APPLIED MACHINE INTELLIGENCE AND INFORMATICS. Herl'any