UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

TRAITEMENT ÉCO-ÉNERGÉTIQUE POUR DES SYSTÈMES MIMO MASSIFS: UNE APPROCHE DE CALCUL APPROXIMATIF

ENERGY EFFICIENT MASSIVE MIMO PROCESSING FOR NEXT GENERATION WIRELESS COMMUNICATION SYSTEMS: AN APPROXIMATE COMPUTING APPROACH.

THÈSE PRÉSENTÉ(E)
COMME EXIGENCE PARTIELLE DU
DOCTORAT EN GÉNIE ÉLECTRIQUE

PAR
ABHINAV KULKARNI

DECEMBER 12, 2024

UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

Doctorat en génie électrique (3739)

Direction de recherche :

Messaoud AHMED OUAMEUR	Directeur
Daniel MASSICOTTE	Co-directeur

Jury d'évaluation

François Nougarou	Président
Sébastien Roy	Externe
Michel Lemaire	Interne UQTR

Thèse soutenue le 22 11 2024

Résumé

Avec l'avancée des technologies, les limitations de la norme 5G en matière de débit, connectivité et faible latence exigent le développement de normes pour au-delà de la 5G (B5G - Beyonf 5G). L'émergence de la 6G sous B5G pose des défis concernant l'efficacité énergétique (EE) du traitement des stations de bases (EE - Energy Efficiency). La technologie Multiple Input Multiple Output extrêmement large (XL-MIMO - Extremely Large MIMO), qui améliore l'efficacité spectrale en utilisant des antennes multiples à grande échelle, est essentielle pour la 6G. Toutefois, les systèmes MIMO traditionnels rencontrent des difficultés, notamment une bande passante d'interconnexion élevée et une complexité computationnelle accrue, entraînant des hausses de latence et de consommation d'énergie. La technologie MIMO décentralisée (D-MIMO - Decentralized MIMO) résout ces problèmes en décentralisant les tâches de traitement, réduisant ainsi la bande passante et la complexité, et diminuant la latence et la consommation d'énergie. Le calcul approximatif peut encore améliorer l'Energy Efficiency (EE) en exploitant la résilience aux erreurs du système.

L'objectif principal de cette recherche est d'améliorer l'EE dans la détection des signaux sans fil au sein de Multiple Input Multiple Output (MIMO) grâce à une approche en deux niveaux : la décentralisation de l'algorithme et l'intégration du calcul approximatif dans la détection des signaux. La décentralisation de l'algorithme consiste à développer un nouvel algorithme de détection Decentralized MIMO (D-MIMO), en caractérisant différentes topologies matérielles pour l'accumulation et le traitement des gradients nécessaires. Une méthode heuristique pour les circuits de multiplication signée approximative basés sur Field Programmable Gate Array (FPGA) est également proposée, offrant des compromis entre précision et performance matérielle. L'impact des niveaux d'approximation sur la qualité de service (Quality of Service (QoS)) de la détection

des signaux est étudié avec le modèle Bruit de multiplication approximatif (Approximate Multiplication Noise (AMN)), qui capture les irrégularités causées par la multiplication approximative.

Un algorithme de détection D-MIMO, Newton décentralisé (Decentralized Newton (DN)), utilisant l'optimisation de Newton pour le matériel FPGA, est proposé et évalué pour la bande passante d'interconnexion, la complexité computationnelle et d'autres indicateurs de performance clés (Key Performance Indicator (KPI)) de Beyond 5G (B5G). Les circuits de multiplication signée approximative pour FPGA, développés avec la méthodologie clonage logique (Logic Cloning (LC)), sont optimisés pour la précision, la consommation de ressources et l'énergie. La performance des multiplications approximatives est évaluée pour la détection des signaux forçage zéro (Zero Forcing (ZF)) MIMO. Une analyse de l'impact de l'approximation sur le QoS est réalisée en calculant une expression analytique pour le taux d'erreur de symbole (Symbol Error Rate (SER)) à l'aide du modèle AMN. Les analyses de fidélité du signal et de résilience fournissent des informations sur les configurations fiables de multiplication approximative et les seuils de dégradation de la fiabilité du système.

En termes de SER, l'algorithme DN est comparable à la détection ZF en bruit Gaussien blanc additif (Additive White Gaussian Noise (AWGN)), maintenant une complexité linéaire constante à travers les clusters d'antennes Base Station (BS). La configuration en étoile de DN utilise moins de ressources, consommant environ 78,4× et 169× moins d'énergie que les méthodes GPU. Pour des opérandes de 16 bits, la méthode LC réduit la consommation des tables de correspondances (Look-Up Table (LUT)) de 31,05% pour Booth et de 36,85% pour Baugh Wooley (BW), tout en réduisant le Power Delay Product (PDP). L'analyse de l'impact de l'approximation sur le QoS dans Single Input Single

Output (SISO) montre que des niveaux d'approximation plus élevés améliorent l'EE dans des conditions de faible rapport signal sur bruit (Signal-to-Noise Ratio (SNR)), tandis que des niveaux plus faibles sont plus efficaces dans des conditions de SNR élevé. Cette recherche jette les bases pour l'amélioration de l'EE de la détection MIMO en utilisant la décentralisation algorithmique et le calcul approximatif.

Summary

As technology advances, the limitations of the 5G standard in meeting demands for throughput, connectivity, and low latency necessitate the development of B5G communication standards. With 6G emerging as a key development under B5G, a major concern is the EE of BS processing. Extremely Large MIMO (XL-MIMO) technology, which enhances spectral efficiency through multiple antennas at a extremely large scale, is an infrastructure enabler for 6G. However, traditional MIMO systems face challenges such as high interconnect bandwidth and computational complexity when scaled to a large scale, leading to increased latency and energy consumption. D-MIMO addresses these issues by decentralizing processing tasks to reduce interconnection bandwidth and computational complexity, thereby lowering latency and energy consumption. Approximate computing has potential for further enhancing EE by utilizing error-resiliency of the system.

The primary objective of this research is to improve EE in wireless signal detection within MIMO through a two-level approach: algorithm decentralization and incorporating approximate computing techniques in signal detection. Algorithm decentralization involves developing a novel D-MIMO detection algorithm characterizing different hardware topologies for accumulation and processing of gradients required for the algorithm. A novel heuristic methodology for FPGA based approximate signed multiplication circuits is also introduced, which provides controlled trade-offs between multiplication accuracy and hardware implementation performance. In systematic analysis of the impact of approximation levels on the QoS of signal detection, novel AMN model is evaluated to capture the irregularities caused due to approximate multiplication used for wireless signal detection.

A novel D-MIMO detection algorithm, DN, using Newton optimization for FPGA hardware, is proposed and assessed for interconnect bandwidth, computational complexity, and other B5G KPIs. Approximate signed multiplication circuits for FPGA, developed using LC methodology, are optimized for accuracy, resource consumption and power consumption. The performance of approximate multiplications is evaluated for ZF MIMO signal detection. A systematic analysis of the impact of approximation on QoS is conducted by computing an analytical expression for SER using the AMN model. Signal fidelity, approximation, and resiliency analyses provide insights into reliable approximate multiplication configurations and the thresholds at which system reliability degrades.

In terms of SER, the DN algorithm performs comparably to ZF detection in AWGN, maintaining consistent linear complexity across BS antenna clusters. The DN star configuration uses approximately 1.05× DSP48E, 0.84× FF, and 0.71× LUT compared to the DN ring, with both configurations being more power efficient, consuming about 78.4× and 169× less power than GPU based methods. For 16-bit operands, LC method effectively reduces LUT resource consumption by 31.05% for Booth and 36.85% for BW, and lower the PDP by 34% for Booth and 35% for BW. The systematic analysis of the impact of approximation level of multiplication on QoS of signal detection in SISO indicates that higher approximation levels improve EE across all bit-width values, proving advantageous in low SNR conditions, while lower approximation levels are more effective under high SNR conditions. A guideline to extend the systematic analysis of SISO to MIMO enables to gain insights on approximate multiplication configuration. The research work builds foundations using algorithmic decentralization and approximate computing for strategically approaching the research problem of improving EE for MIMO detection.

Acknowledgment

I am grateful to my supervisor, Prof. Messaoud, for his mentorship and guidance. His expertise in research, along with his innovative approach with research methodology, has been instrumental in shaping my research journey. I would also like to extend my sincere thanks to my co-supervisor, Prof. Daniel, whose expert eye, mentorship and valuable feedback have been crucial to refining my research.

I would also like to acknowledge my colleagues, Michel and Rabiul, for their support and guidance. Their assistance has been a source of strength throughout my academic journey. My deepest gratitude goes to my parents and brother, whose affection and encouragement has helped me navigate the challenges of this academic pursuit.

I believe that knowledge can be advanced through mindfulness and commitment. I also thank all the researchers whose foundational work has allowed me to develop my own research and present this thesis work. Finally, I would like to express my thanks to UQTR for providing me with the opportunity and conducive environment to develop my skills and develop as a researcher.

Table of contents

List of figures

List of tables

List of acronyms

3GPP 3rd Generation Partnership Project. 32, 33, 36

ADMM Alternating Direction Method of Multipliers. 18, 22

ADMM-GS ADMM with Gauss-Seidel iteration. 18

AMN Approximate Multiplication Noise. iv, vi, vii, 13, 103, 105, 114, 116

ANRR Average Normalized Resiliency Ratio. 78, 79, 81, 82

ASIC Application Specific Integrated Circuit. 53, 58, 103

AWGN Additive White Gaussian Noise. iv, vii, 22, 32, 36

B5G Beyond 5G. iv, vi, vii

BER Bit Error Rate. 28, 30, 76, 79, 114

BS Base Station. iv, vi, vii, xiii, 3, 6–9, 12, 19–22, 28, 31, 32, 35, 36, 103

BW Baugh Wooley. iv, vii, 12, 54–56, 59, 60, 105

CC Carry Chain. 24–26

CCU Carry Chain Unit. 24, 26

CDMA Code Divison Multiple Access. 1

CPD Critical Path Delay. 55, 60

CSI Channel State Information. 31, 35, 105

D-MIMO Decentralized MIMO. iii, iv, vi, vii, 9, 20, 21, 105

DCD Decentralized Coordinate Descent. 18, 33, 34, 36–38

DCG Decentralized Conjugate Gradient. 19, 22

DN Decentralized Newton. iv, vii, xiv, 12, 22, 32–38, 103, 105

EE Energy Efficiency. iii, v–vii, 2, 6–11, 13, 22, 25, 27, 29, 32, 36, 53, 58, 76, 78, 79,

81, 104, 106

EP Expectation Propagation. 19, 21–23, 28, 30, 32, 33, 36

FEC Forward Error Correction. 27, 29

FER Frame Error Rate. 76, 79, 114

FF Flip Flop. 33, 36

FFT Fast Fourier Transform. 27–29

FIR Finite Impulse Response. 28, 30

FPGA Field Programmable Gate Array. iii, iv, vi, vii, x, xiv, 10, 12–16, 18, 22, 24–26, 32, 33, 36, 37, 53, 54, 58, 103–105

GPRS General Packet Radio Service. 1

GPU Graphics Processing Unit. 33, 34, 36–38

GSM Global System for Mobile. 1

IRS Intelligent Reflecting Surfaces. 106

KPI Key Performance Indicator. iv, vii, xiii, 2, 11, 13, 27, 34, 37, 38, 57, 62, 104

LAMA Large-MIMO Approximate Message Passing. 19, 22, 33, 36, 37

LC Logic Cloning. iv, vii, xiv, 12, 26, 54–60, 62, 103, 105

LLR Log Likelihood Ratio. 22, 28, 32, 36

LSSI Laboratory of Signal and System Integration. 15, 16

LUT Look-Up Table. iv, vii, 12, 24–26, 33, 36, 54–56, 59, 60

MAC Multiply And Accumulate. 25, 26

MAP Maximum A Posteriori. 19, 103

MED Mean Error Distance. 55, 59, 60

MF Matched Filter. 115

MIMO Multiple Input Multiple Output. iii–vii, ix, x, xiii, xiv, 3, 6, 8–15, 18–24, 28, 31, 32, 35, 36, 54, 56, 59–61, 103–106, 114

ML Machine Learning. 4, 20, 25, 27, 106

MMSE Minimum Mean Square Error. 13, 18, 19, 31, 35, 77, 80, 105

MRED Mean Relative Error Distance. 55, 59, 60

MSE Mean Square Error. 114, 117

NMED Normalized Mean Error Distance. 55, 59, 60

NRR Normalized Resiliency Ratio. 78, 79, 81, 82

OTFS Orthogonal Time Frequency Space. 20, 23

PDP Power Delay Product. iv, vii, 54, 55, 59, 60

PP Partial Product. 24–26

QAM Quadrature Amplitude Modulation. 28–30, 32, 36, 56, 61

QoS Quality of Service. iii, iv, vi, vii, 8–11, 14, 29, 76, 79, 82, 104, 114

QPSK Quadrature Phase Shift Keying. 13, 77, 80, 105

RRC Root Raised Cosine. 28, 30

SC Successive Cancellation. 27, 29

SCMA Sparse Code Multiple Access. 28, 30

SER Symbol Error Rate. iv, vii, 12, 19, 27, 32, 34, 36, 37, 56, 60, 76–82, 103, 114

SGD Stochastic Gradient Descent. 19, 22

SISO Single Input Single Output. iv, vii, 14, 15, 28, 104, 114

SNR Signal-to-Noise Ratio. v, vii, 12, 19, 22, 28, 30, 77–79, 81, 82

TM Truncated Multiplication. 13, 77, 80, 105, 117

TP Truth Probability. 54, 59, 103

UE User Equipment. xiii, 3, 6, 18, 19, 21, 22, 28, 31, 32, 35, 36

UQTR Université du Québec à Trois-Rivières. 16

VAR Virtual and Augmented Reality. 1, 2

XL-MIMO Extremely Large MIMO. vi, 6, 10, 15, 18

ZF Zero Forcing. iv, vii, 12, 14, 31, 32, 34–37, 54, 56, 59–61, 105

Chapter 1 - Introduction

1.1 Background

Wireless communication systems have undergone revolutionary advancements since inception, driven primarily by the escalating demands for higher data rates in evolving applications. Consequently, various wireless standards have evolved to meet the required data rates of these diverse application needs [1,2]. Initially, 1G and 2G were characterized by Global System for Mobile (GSM), which facilitated communication at speeds of up to 30-35 Kbps. This was later enhanced with General Packet Radio Service (GPRS), achieving data rates of up to 110 Kbps. The advent of 3G significantly increased speeds up to 2 Mbps, enabling smartphones to handle faster communication, transfer enormous data, reduced latency, and incorporate enhanced security features. Code Divison Multiple Access (CDMA) technology in 3G facilitated communication across multiple channels simultaneously, thereby enhancing network speed and connectivity. Subsequently, 4G was introduced to elevate data rates up to 100 Mbps. It entailed a complete overhaul and simplification of the 3G architecture, resulting in substantial reductions in transfer latency, and enhancing overall network efficiency and speed. The 5G standard has evolved to provide peak speeds of up to 10 Gbps, latency as nearing 1 millisecond, improved security, extensive coverage, and increased user handling capacity.

As societal communication data needs continue to grow, communication capabilities must be able to keep pace with them. Future communication standards need to support applications such as extended Virtual and Augmented Reality (VAR), multi-sensory holographic teleportation, real-time remote healthcare, autonomous cyber-physical systems, industrial automation, and precision agriculture. Some of these applications cannot be adequately supported by the 5G standard. For instance, next-generation

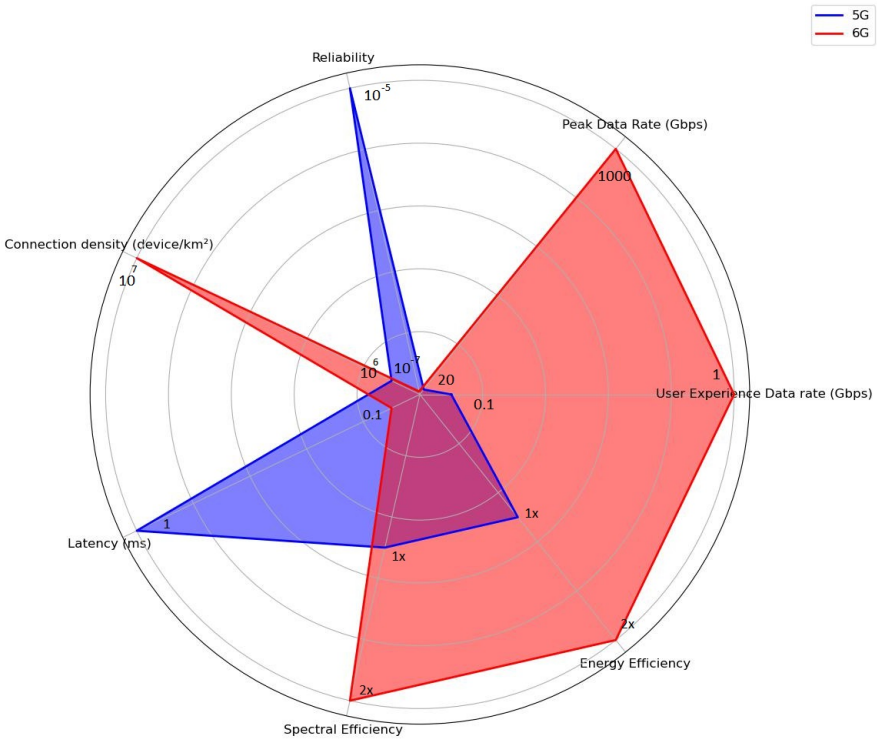

Figure 1.1 Major KPI improvement projected with 6G in comparison with 5G [1].

applications like extended VAR and holographic teleportation require data rates in Tbps range and latency below microsecond level, capabilities that exceed those specified in the frequency bands being utilized for 5G. Increasing industrial automation and the transition towards Industry 4.0 will significantly raise connectivity density, surpassing the servicing capability of 1 million devices per km^2 that 5G was designed to handle. Heightened connection density will drive demands for improved EE, an aspect that 5G was not specifically designed to address. This has led to the emergence of a new set of requirements and KPIs for evolving B5G communication standards as shown in Figure. 1.1. Spectral and energy efficiency is doubled for 6G, while user data rate and peak data rate is increased by $10\times$ and $50\times$ respectively. Latency is reduced by $10\times$, while connection density is increased by $10\times$. Reliability is increased by about $100\times$.

For B5G, the development of 6G standard encompasses advancements in three key areas: spectrum, protocol, and infrastructure [1–3] as shown in Table. 1.1. At the spectrum level, incorporation of higher frequency bands is being researched to expand available bandwidth. Protocol level enhancements aim to optimize data packet organization and transmission methodologies. At the infrastructure level, efficient hardware implementation techniques are crucial for realizing B5G networks. A particular focus is placed on reducing energy consumption to improve the overall scalability and environmental sustainability [4].

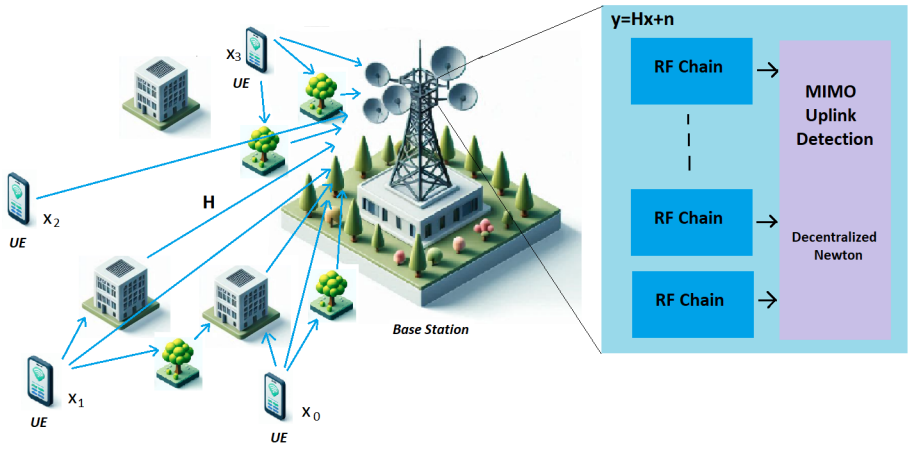

Figure 1.2 MIMO uplink signal detection with BS and UEs with single antenna and channel matrix **H**.

Table 1-1 6G enabling technologies.

Technology	**Features**	**Benefits**
Spectrum Level Enablers		
Terahertz Communications	Uses frequencies above 100 GHz in THz range for wider bandwidths and higher data rates.	Enables ultra high speed data transmission suitable for future applications like virtual reality and high definition streaming.
Visible Light Communication	Uses visible light spectrum for data transmission, integrating with illumination systems.	Offers high speed and energy efficient communication with minimal interference, ideal for indoor applications.

Technology	Features	Benefits
Dynamic Spectrum Sharing	Allows flexible allocation of spectrum resources among different operators and services.	Optimizes spectrum usage and increases network capacity.
Cognitive Radio	Utilizes spectrum sensing and interference management techniques.	Enhances spectrum efficiency by dynamically sharing resources.
Free Space Optics	Uses light propagation in free space for high capacity, point-to-point communication links.	Offers high data rates, low latency, and immunity to electromagnetic interference.
Protocol Level Enablers		
Machine Learning	Application of Machine Learning (ML) algorithms for intelligent and adaptive communication configuration.	Improves network performance, reduces latency, and supports dynamic network management.
Orbital Angular Momentum	Utilizes helical phase wavefronts for wireless modulation, enabling multiplexing with perpendicular orbital angular momentum modes.	Increases channel capacity and spectrum efficiency.
Full Duplex Communication	Enables simultaneous signal transmission and reception.	Increases spectral efficiency, bolsters network capacity, and supports real time bidirectional communication.
Non Orthogonal Multiple Access	Allows multiple users to share the same resources through power and code domain assignments.	Improves user connectivity and spectral efficiency.
Network Slicing	Provides customized virtual networks on a shared physical infrastructure.	Optimizes resource allocation, and supports diverse application requirements.
Blockchain	Implements decentralized security mechanisms for authentication and data integrity.	Enhances network resilience, protects against cyber threats, and ensures trustworthiness in communication.

Technology	Features	Benefits
Proactive Caching	Caching of frequent utilized content.	Reduces latency by reduced access delay and efficient traffic offloading.
Infrastructure Level Enablers		
Edge Intelligence	Integrates ML algorithms at the network edge for real time data analytics and decision making.	Reduces latency and power consumption. Supports autonomous applications in distributed environments.
Holographic Radio	Creates a continuous electromagnetic aperture using electromagnetic wave interference, akin to optical holography.	Enables ultra-high density and resolution spatial multiplexing, facilitating accurate electromagnetic field reconstruction.
Wireless Energy Transfer	Enables wireless charging and power delivery to IoT devices and sensors.	Enhances device autonomy, reduces maintenance costs, and supports sustainable network operations.
Satellite Networks	Integrates satellite communication systems for global coverage and ubiquitous connectivity.	Extends network reach, supports remote and rural areas.
Integrated Access and Backhaul	Combines wireless access and backhaul/fronthaul infrastructure.	Optimizes spectrum usage, reduces deployment costs, and enhances data transfer efficiency in dense and heterogeneous 6G environments.
Integrated Broadcast and Multicast Networks	Enables large-scale content delivery to multiple devices simultaneously via broadcast and multicast.	Improves network flexibility to support diverse service needs and utilizes existing infrastructure effectively.
Backscatter Communication	Utilizes ambient RF signals for interaction between devices for short range communication.	Optimizes resource deployment and enhances network services by minimizing power consumption.
Intelligent Reflective Surfaces	Manipulates electromagnetic waves using metamaterials.	Improves spectral efficiency and reliability encountered in multipath signal propagation.

Technology	Features	Benefits
Tactile Internet	Enables real time human to machine interactions using haptic sensors.	Enhances user experience by enabling real-time and low latency tactile feedback.
Quantum Technologies	Computing and communication through quantum principles.	Boosts computational power for complex algorithms. Provides massively reduced latency and very high throughput.
MIMO	Uses multiple antennas for simultaneous transmission of data streams. XL-MIMO utilizes these antennas at a very large level.	Increases data throughput and spectral efficiency by transmitting multiple streams concurrently. Enhances link reliability and extends coverage area.

XL-MIMO is one of the infrastructure enabler for 6G as given in Table. 1.1, MIMO with extremely large number of antennas. In the downlink scenario, the BS employs its multiple antennas to transmit data to UEs. The downlink channel matrix represents the spatial relationships between BS and UE antennas. Conversely, the uplink process for signal detection at BS, as illustrated in Figure. 1.2, involves UEs transmitting data back to the BS. The uplink channel matrix captures the spatial characteristics between UE and BS antennas. As MIMO, XL-MIMO largely improves spectral efficiency and reliability in multipath signal propagation using spatial diversity and spatial multiplexing.

Meeting simultaneous requirements for ultra high speed, high capacity, and low latency connectivity poses a significant challenge. While traditional communication networks have primarily focused on optimizing metrics such as throughput and latency, EE has recently gained critical importance due to growing environmental concerns regarding electricity consumption. This shift is driven by the substantial power demands of BS operations. BS alone account for about 60% to 80% of overall network energy

consumption [5, 6]. This consumption primarily stems from baseband processing, RF processing, and signal power amplification during active modes. For example, typical baseband energy consumption for BS is approximately 150 Watts for 4G LTE and 220 Watts for 5G NR, constituting 5-15% of their total energy use [7]. The efficiency of RF and power amplification is also dependent on baseband processing efficiency [8].

The EE of algorithms implemented for BS baseband processing is influenced by their computational complexity and hardware resource utilization. Algorithms with higher computational complexity demand more processing power and longer execution times, leading to increased energy consumption as hardware components operate for extended periods. Efficient management of hardware resources, such as memory and bandwidth, optimizes data movements and storage operations, thereby reducing overall power consumption. The goal of EE algorithm design is to achieve computational objectives while minimizing energy expenditure. Consequently, prioritizing algorithms with lower computational complexity and resource usage can substantially increase the EE of baseband processing.

Communication systems require extensive data processing and substantial computing resources, yet the growth of computing capabilities lags behind demand. As transistors shrink to smaller sizes, they encounter physical limits like quantum tunneling, short-channel effects, and increased leakage currents. These factors restrict further performance gains by causing reduced gate control and challenges in heat dissipation, thus hindering the ability to increase the number of transistors on integrated circuits.

To address this challenge, there is need for techniques that optimize existing hardware capabilities to deliver necessary performance while minimizing resource and power consumption. One such approach is approximate computing [9], [10]. The principle

underlying approximate computing is that systems do not always require the highest level of accuracy for all applications [11]. By assessing the error resiliency of such applications, the surplus reliability can be managed, creating opportunities for approximate computing implementation. The deployment of an approximate computing technique is tailored to a specific application. Subsequently, the error resilience of various components within the communication system's processing algorithms have to be assessed based on their criticality. This evaluation is crucial for determining where and how approximate computing can be effectively utilized to optimize performance while maintaining acceptable QoS.

1.2 Rationale

Enhancing the EE of baseband processing, integral to BS operations, directly influences the overall energy consumption of the BS. The work [12] emphasizes the incorporation of energy-efficient methods into the design of baseband processing for BS, aiming to reduce operational expenses and environmental impact. Similarly, improving the EE of wireless signal detection, essential for baseband processing, contributes to overall EE in baseband processing.

The transition to next-generation communication systems demands higher data rates, often necessitating an increase in carrier frequency to achieve desired throughput. As frequencies advance into the THz range, conventional communication approaches encounter challenges such as heightened signal attenuation and propagation loss. To address these issues and improve reliability and spectral efficiency, XL-MIMO systems offer a promising solution. Higher carrier frequencies enable the use of smaller antennas, facilitating effective utilization of a large number of antennas for XL-MIMO infrastructure.

However, integrating XL-MIMO with numerous antennas brings about a few technical challenges. A major problem is the immense interconnect bandwidth necessary between the increased antennas and single BS processing unit. In XL-MIMO infrastructure, the data transfer rates and computational complexity can burden the network infrastructure and consume substantial power because of increased resource consumption. To tackle this issue, effective architectures for XL-MIMO uplink processing are vital. D-MIMO stands out as a potential solution by allocating computational tasks and signal processing across several smaller processing units instead of centralizing them at the BS, where every processing unit is allocated a set of smaller number of dedicated antennas. This method lessens the requirement for high bandwidth interconnects and minimizes computational complexity and interconnect bandwidth. By employing D-MIMO, the system can more effectively handle the computational workload and thus reduce energy consumption.

Approximate computing in the context of wireless signal detection operations can improve EE by strategically compromising on computational precision by using the inherent error resiliency. By relaxing the requirement for high computational precision in signal processing, computational resource usage can be reduced. For instance, algorithms can be designed to operate with reduced bit-widths or employ simplified computation methods that trade-off minor accuracy losses for substantial energy savings. This approach can be particularly advantageous in D-MIMO systems where numerous antennas and distributed processing nodes collaborate.

In situations where rigorous computational accuracy is not essential for attaining the desired QOS, algorithms dynamically modify their precision levels according to real-time performance indicators. This adaptive ability can enable D-MIMO uplink systems to customize their signal processing methods to the current operating conditions, reducing

energy consumption without sacrificing overall system reliability. To further reinforce this procedure, hardware implementation on a FPGA platform can offer benefits due to its reconfigurability. This reconfigurability is especially valuable in scenarios requiring rapid prototyping, iterative design enhancements, or adjustments to evolving standards.

1.3 Research Problem

As personal computing devices become increasingly interconnected on a large scale, the data processing requirements for next-generation communication systems are escalating. The emergence of new enterprise applications is also driving the need for increased capabilities in wireless signal detection, making these systems more sophisticated and increasing the demand for computing resources.

However, the traditional trajectory of transistor scaling is encountering diminishing returns as it is becoming increasingly difficult to accommodate more transistors on integrated circuits. This phenomenon, coupled with power dissipation challenges, acts as a barrier to further increase computing capabilities across hardware platforms for signal detection systems. Consequently, there is a need for efficient computing approaches to harness gains in EE.

The research work addresses the challenge of improving the EE of wireless signal detection for B5G with emphasis on XL-MIMO. The thesis is composed of three research articles. The first research article expounds on a novel D-MIMO detection algorithm. The second article details novel signed approximate multiplication circuits. The work proposed in these articles is evaluated for FPGA implementation. The third research article proposes a systematic analysis linking the QoS of signal detection with the level of approximation in the arithmetic multiplication circuit.

1.4 Objectives

The goal of this research work is to find novel techniques using two level approaches for efficient signal detection in context of XL-MIMO processing. The first level approach uses D-MIMO signal detection. The second level approach incorporates approximate arithmetic multiplication operations for signal detection. The goal is achieved through the following objectives:

1. Algorithm decentralization in centralized XL-MIMO uplink detection.

2. Proposition of a heuristic methodology for inducing approximation in signed arithmetic multiplication circuits.

3. Systematic analysis for the application of approximate multiplication to wireless signal detection to link the level of approximation in multiplication operation and QoS of the system.

1.5 Contributions

The research outlines techniques aimed at achieving its objectives, with each contribution yielding explicit gains in specific KPIs. This facilitates the adoption of these techniques that require homomorphic improvements in KPIs necessary for wireless communication applications. Specifically, the emphasis on EE highlighted by these contributions would reduce electricity consumption associated with wireless signal detection. This reduction not only saves operational costs but also decreases the carbon footprint of these systems, thereby improving their sustainability. For each of the objectives, the contributions are explicitly listed in respective research article. A summary of these contributions is as follows:

- **Objective 1:**

  The article introduces a novel algorithm that adapts the centralized Newton optimization algorithm for D-MIMO detection, focusing on developing a local objective function across decentralized antenna clusters of BS. The proposed DN algorithm demonstrates performance comparable to ZF SER, especially under low SNR conditions. The work further presents ring and star topological architectures designed for FPGA hardware implementation of DN, enabling efficient gradient and Hessian sampling while leveraging decentralized antenna cluster configurations. Further, the work conducts an analytical comparison of interconnect bandwidth, throughput, latency, computational complexity, hardware resource consumption, and energy consumption between ring and star topologies and other contemporary D-MIMO uplink detection techniques. The ring topology maintains a consistent interconnect bandwidth despite increasing cluster numbers, contrasting with the star topology which exhibits lower latency but the ring topology excels in throughput capabilities.

- **Objective 2:** A heuristic methodology LC is introduced which utilizes the probability of logic '1' of the LUT-LS output to induce approximation in the accurate multiplication circuit for FPGA implementation. Novel approximate multiplication circuits are designed from BW and Booth multiplication schemes for signed operation using LC methodology. LC-BW circuits are characterized by high multiplication accuracy, their approximation being controlled with finer granularity, while LC-Booth circuits are characterized by low LUT resource consumption. Also, impact of approximate multiplication circuits at the application level is evaluated for signal fidelity analysis of ZF MIMO uplink signal detection.

- **Objective 3:** A systematic analysis of the application of approximate multiplication to wireless signal detection is conducted in this work. For such assessment, Truncated Multiplication (TM) operation on Quadrature Phase Shift Keying (QPSK) Minimum Mean Square Error (MMSE) signal detection is explored. AMN model is introduced as a constant noise model for signal fidelity analysis. The EE gains are analysed through approximation analysis. Insights into reliable configurations of TM for QPSK MMSE signal detection are provided using the resiliency analysis.

1.6 Research Methodology

This abstract research methodology primarily aims to improve EE in wireless signal detection for B5G using MIMO technique. The research work begins with an examination of existing MIMO techniques to assess its capabilities and limitations in scaling signal detection for XL-MIMO. This investigation identifies bottlenecks that currently restrict XL-MIMO's potential as a key enabler for B5G. Subsequently, D-MIMO approach to uplink signal detection is identified as a primary strategy to overcome these bottlenecks, aligning with B5G KPIs. The work involves evaluating efficient D-MIMO architectures that prioritize EE, minimize latency, maximize throughput, and optimize interconnect bandwidth to meet the demands of B5G.

The research introduces a novel technique for D-MIMO uplink detection based on Newton optimization. The proposed D-MIMO technique is implemented and evaluated for FPGA, employing multiple topologies to assess their performance according to KPIs. The research explores the critical role of arithmetic multiplication operations in MIMO signal detection and investigates the potential of approximate computing to improve EE. Novel approximate multiplication circuits are developed using a heuristic methodology

that introduces controlled approximation to accurate multiplication operations, evaluated for FPGA, aiming to balance accuracy with power and resource consumption. The approximate multiplication circuits are evaluated for QoS at the system level for ZF MIMO signal detection. The research identifies incomprehension between the level of approximation in approximate multiplication circuits and their impact on signal detection performance. A systematic analysis is presented to remove this incomprehension, providing insights into regulating approximation level with intended QoS for SISO. Further, guideline is outlined for extending this analysis to MIMO systems. Research methodology is outlined in Figure. 1.3. A detailed methodology for each article is provided in its respective chapter.

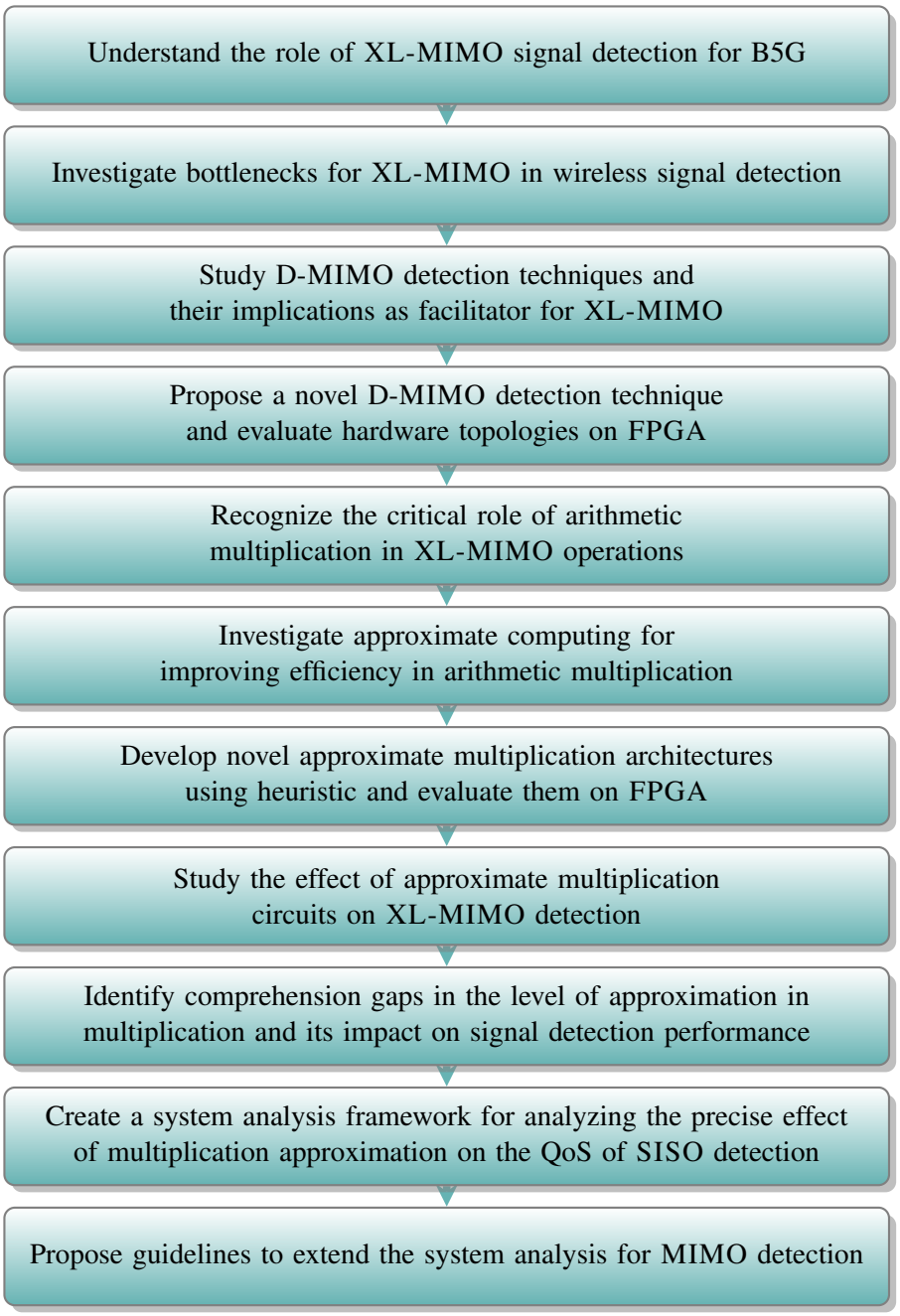

Figure 1.3 Research Methodology

1.7 Research Infrastructure

The research infrastructure and available funding had major contribution in the success of the work. The primary research was conducted at the Laboratory of Signal and System

Integration (LSSI) at Université du Québec à Trois-Rivières (UQTR), which provided access to essential computing resources, including a workstation and an IBM System x3850, necessary for heavy computational tasks. In addition to computing facilities, the UQTR Bibliothèque played a vital role by offering necessary stationary supplies and dedicated study space.

The research utilized various simulation tools to achieve its objectives. Python and C/C++ were employed for numerical simulations, enabling modeling and analysis required for the work. Furthermore, the Vivado Design Suite 2021.1 was utilized for hardware implementation, specifically targeting Xilinx Virtex 7 and Virtex Ultrascale FPGA devices.

The research was funded through multiple organizations. The contributors are Natural Sciences and Engineering Research Council of Canada, Prompt, Canadian Foundation for Innovation, and CMC Microsystems. Additional funding support came from Opal-RT Technologies Inc., Hydro-Québec, NUTAQ Innovation, and LSSI Laboratoire des Signaux et Systèmes Intégrés, as well as Chaire de recherche sur les signaux et l'intelligence des systèmes haute performance.

1.8 Thesis Outline

The thesis is organised in the following manner. Chapter 1 provides an introduction and rationale for the work, outlining the research objectives, achieved contributions, research methodology and infrastructure utilized. Chapter 2 details the essential literature study conducted for the work. Chapter 3 delves into the research article addressing algorithm decentralization. Chapter 4 discusses the research article focusing on a novel methodology for signed approximate multiplication circuits. Chapter 5 covers the research article

dealing with the systematic analysis of the application of approximate multiplication operations for wireless signal detection. Finally, Chapter 6 presents the conclusion of the research and outlays future avenues for research.

Chapter 2 - Literature Review

This chapter presents contemporary research in the domain of each of the research objectives, categorically segregating them into separate sections according to their objectives. The section on algorithm decentralization explores various decentralized approaches under different configuration scenarios and addresses challenges in XL-MIMO. The further section investigates FPGA softcore arithmetic circuits and explores approximate circuits for signed multiplication operations. The next section examines research on systematically applying approximate computing for wireless signal detection, discussing its relevance in communication systems. Each section conducts a thorough literature review to evaluate state-of-the-art techniques, discussing their characteristics and challenges. These research works lay the groundwork for the contributions of the proposed research, aimed at fulfilling the research objectives.

2.1 Algorithm decentralization for MIMO uplink detection

The work [13] comprehensively examines novel paradigms critical for XL-MIMO, focusing on the near-field region visibility and leveraging signal sparsity, aspects often overlooked in traditional MIMO systems. D-MIMO is an important advancement for XL-MIMO. In the work [14], Decentralized Coordinate Descent (DCD) based MIMO detection computes partial uplink signals at each distributed BS antenna cluster using the coordinate descent method. These partial signal estimates are adjusted by the BS antenna cluster variance and merged to create the final uplink signal at the BS. In the work [15], Decentralized Alternating Direction Method of Multipliers (ADMM) is a high computational complexity method based on consensus exchange, providing near MMSE performance with minimal iterations for low UE. In the work [16], ADMM with Gauss-Seidel iteration (ADMM-GS) integrates Gauss-Seidel iteration in ADMM

for better performance in SER. This method is appropriate for high SNR and high UE situations and is robust against channel estimation errors. The work [15] utilizes Decentralized Conjugate Gradient (DCG) method and yields near MMSE performance with minimal iterations in high UE scenarios per BS antenna cluster. The work [17] utilizes a fully decentralized architecture in daisy chain topology for D-MIMO detection with Stochastic Gradient Descent (SGD), eliminating the need to reconfigure the central cluster when adding new BS antenna clusters and maintaining constant interconnect bandwidth between two clusters for a specific number of UEs.

The work [18] explores a Maximum A Posteriori (MAP) estimation based decentralized algorithm known as Large-MIMO Approximate Message Passing (LAMA), and presents two architectures for hardware implementation: Partially Decentralized as LAMA-PD and Fully Decentralized as LAMA-FD. D-MIMO detection using LAMA-PD and LAMA-FD equalization achieves optimal performance under the assumption that the channel matrix $\mathbf{H}$ follows an i.i.d distribution and is characterized by a variance of $1/B$, where B denotes the number of BS antennas [18]. However, LAMA's robustness is questioned in realistic channel environments [19] like 3GPP spatial channel model. The work [20] proposes an Expectation Propagation (EP) based method, which rivals LAMA in error rate performance but involves explicit matrix inversion, thus escalating its computational complexity. The work [21] improves upon LAMA's performance, especially at high SNR, although it necessitates high interconnect bandwidth. MAP methods incur computational overhead to enhance numerical stability for variance computation from noise statistics. Also, in MAP methods, partial local estimates must be fused and processed using a soft detector for signal estimation, with proposed enhancements discussed in the work [22]. The work [23] discusses a tree-based D-MIMO detection architecture, outlining a decentralized scalable BS system where

interconnect links grow logarithmically with the addition of BS antenna clusters.

The work in [24] introduces a new D-MIMO detection algorithm based on approximate message passing. This algorithm incorporates a configurable additional inner loop into the decentralized version of GEC-SR (Generalized Expectation Consistent for Signal Recovery) and optimizes the utilization of computing resources in local decentralized computing units. In [25], a D-MIMO detection algorithm is presented that addresses spatial non-stationarities arising from the deployment of XL-MIMO, further proposing approaches to accelerate the convergence of the algorithm. The work in [26] explores an approximated version of the expectation propagation algorithm using MIMO-Orthogonal Time Frequency Space (OTFS), where OTFS is a joint time-frequency space modulation of symbols designed to mitigate high inter-symbol interference in scenarios with high mobility, affected by Doppler effects.

The work [27] applies reinforcement learning techniques within ML to mitigate issues arising from spatial non-stationarities. These non-stationarities stem from modeling systems in the near-field region, as opposed to the far-field region typical in traditional MIMO systems. The work in [28] introduces a D-MIMO detection method that addresses Out-of-system noise, characterized by unknown statistics. In the work [29], the focus is on integrating local means and variances from decentralized antenna clusters for symbol demodulation. The variance values significantly influence algorithm convergence, with smaller variances leading to faster convergence rates. The algorithm exhibits linear complexity with regards to the number of users. The work highlights that non-uniform antenna clusters outperform uniform configurations, stressing on the need for optimal antenna allocation methods.

The work [30] introduces an integrated communication and learning framework

utilizing MIMO for decentralized federated learning applications, specifically tailored for edge devices. The work analyzes the convergence bounds of the algorithm to assess the impact of communication errors and mixing matrices, resulting in an optimized algorithm aimed at mitigating these challenges. In the work [31], a D-MIMO uplink detection method based on EP is presented. This method groups users to suppress interference, thereby reducing latency. The paper proposes partially and fully decentralized architectures in the form of star and daisy-chain topologies. Star topology performs better when the ratio of BS to UE antennas is low. Subsequently, an approximate version of the same algorithm is introduced to lower computational complexity.

The work [32] investigates D-MIMO uplink detection, specifically addressing colored noise where the channel noise correlation matrix becomes non-diagonal. In the work [33], Gaussian elimination is utilized for recursive D-MIMO detection, and an adaptive recursion termination method is proposed to improve detection latency. The work [34] explores cell-free Massive MIMO systems and proposes decentralized optimization methods for enhancing spectral efficiency. The work [35] advocates for non-linear processing in MIMO for 6G, aiming to mitigate computational complexity through a proposed massively parallel non-linear processing framework. The work [36] proposes a decentralized algorithm without a centralized unit to address computational complexity, scalability, and non-stationarities for D-MIMO uplink detection. The work [37] explores Kalman filters with a square-root implementation for MIMO uplink signal detection.

The challenges in XL-MIMO detection include high computational complexity, significant interconnect bandwidth requirements, and scalability limitations, which D-MIMO aims to alleviate. Realistic phenomena, such as Doppler effects and colored noise correlation, further exacerbate the challenges of XL-MIMO. Additionally,

integrating XL-MIMO with advanced applications, such as federated learning and cell-free massive MIMO systems, necessitates further optimization of resource allocation. In the current work, DN using Newton optimization is proposed as a novel D-MIMO technique. Two novel hardware topologies, the DN ring and star, are introduced for FPGA implementation to optimize interconnect bandwidth, resource consumption, latency, and EE. Literature study is summarized in Table 2-1.

Table 2-1 D-MIMO detection techniques.

Work	**Research**
Han et al. (2023) [13]	Comprehensively analyzed novel paradigms for XL-MIMO.
Li et al. (2019) [14]	Addressed D-MIMO detection using the coordinate descent method with distributed BS antenna clusters.
Li et al. (2017) [15]	Implemented decentralized ADMM with high computational complexity that was advantageous for low UE scenarios.
Ouameur et al. (2019) [16]	Integrated Gauss-Seidel iteration into ADMM for improved error-rate performance in high SNR and high UE scenarios.
Li et al. (2017) [15]	Utilized the DCG method for near MMSE performance, which was advantageous in high UE scenarios.
Sanchez et al. (2019) [17]	Conducted D-MIMO detection with SGD in a fully decentralized architecture, removing the need for central cluster reconfiguration during BS cluster scaling.
Jeon et al. (2019) [18]	Implemented LAMA algorithms in both partially decentralized LAMA-PD and fully decentralized LAMA-FD architectures, offering optimal detection performance in the AWGN channel.
Wang et al. (2020) [20]	Conducted D-MIMO detection with high computational complexity by adapting EP, providing high error-rate performance.
Zhang et al. (2020) [21]	Utilized EP with Log Likelihood Ratio (LLR) for D-MIMO detection, improving the performance of the EP algorithm with log likelihood ratio.

Work	Research
Seidel et al. (2019) [22]	Implemented D-MIMO detection using a binary tree to reduce latency.
Bertilsson et al. (2016) [23]	Designed a D-MIMO detection architecture for decentralized scalable BS systems, where the growth of interlinks was logarithmic.
Yang et al. (2022) [24]	Developed a D-MIMO detection algorithm based on approximate message passing, optimizing decentralized computing resources.
Croisfelt et al. (2021) [25]	Addressed spatial non-stationarities in XL-MIMO and accelerated algorithm convergence.
Li et al. (2024) [26]	Explored an approximated EP algorithm using MIMO-OTFS for scenarios with high mobility and consideration of Doppler effects.
Liu et al. (2024) [27]	Applied reinforcement learning techniques for mitigating spatial non-stationarities in D-MIMO.
Shaik et al. (2024) [28]	Addressed Out-of-system noise with unknown statistics in D-MIMO detection.
Zhang et al. (2021) [29]	Integrated local means and variances from decentralized antenna clusters for symbol demodulation.
Zhai et al. (2024) [30]	Utilized D-MIMO for decentralized federated learning applications, analyzing convergence bounds.
Li et al. (2022) [31]	Proposed a D-MIMO detection method based on EP, suggesting star and daisy-chain topologies for optimal hardware implementation.
Zhao et al. (2021) [32]	Investigated D-MIMO detection under colored noise conditions with a non-diagonal channel noise correlation matrix.
Zheng et al. (2022) [33]	Utilized Gaussian elimination for recursive D-MIMO detection, proposing adaptive termination methods to improve latency.
Datta et al. (2021) [34]	Proposed decentralized optimization methods for enhancing spectral efficiency in cell-free Massive MIMO systems.
Ducoing et al. (2021) [35]	Advocated for non-linear MIMO processing in MIMO and a massively parallel decentralized framework to mitigate ensuing high computational complexity.

Work	Research
Amiri et al. (2021) [36]	Presented an algorithm for D-MIMO detection without any centralized unit, addressing scalability and non-stationarities.
Helmersson et al. (2022) [37]	Explored Kalman filters with square-root implementation for D-MIMO detection.

2.2 Heuristic methodology for FPGA based signed approximate multiplication circuits.

The efficient implementation of an accurate signed multiplication circuit using the Booth algorithm [38] reduces the critical path delay by 3 Carry Chain Unit (CCU)s for generating Partial Product (PP), thereby reducing hardware resource consumption. However, for PP accumulation, an external hardware entity such as an adder or a compressor is necessary, and the accuracy of the circuit depends on the methodology used for PP accumulation.

In the FPGA implementation of an accurate unsigned multiplication circuit [39], an architecture employing LUT and CCUs for accurate signed multiplication circuit based on Booth encoding [40], known as Booth-Opt, saves LUTs by integrating the logic of the rightmost and leftmost carry generating LUTs into adjacent LUTs in a PP row, thereby reducing Carry Chain (CC)s. An improvement in the critical path delay from the work in [40] is presented in subsequent work [41] through optimization of PP generation and use of a PP reduction tree; however, this approach leads to increased LUT resource consumption.

Performance improvements are achieved by incorporating approximations into architectures designed for accurate multiplication. These approximations can be applied in the accumulation of PPs using FPGA based approximate adders [42–44] or approximate compressors [45], while maintaining accurate PP generation. Configurable architectures for signed multiplication circuits, as described in [46], utilize both accurate methods and

four approximate compressors for PP accumulation. These architectures employ explicit PP accumulation methods, leading to varying levels of LUT resource consumption depending on the circuit's accuracy in performing operations of specific bit-widths.

Functional approximation introduces approximation into accurate circuits during the generation of PPs. The Booth-Opt circuit is adapted to Booth-Approx [40] by approximating the generation of PPs, specifically by approximating the carry signal generation at the rightmost LUT for all PPs except the last PP. Booth-Approx demonstrates reduced CC and LUT resource consumption. However, Booth-Approx lacks configurability to balance multiplication accuracy and EE.

AxBM circuits [47] modify the functional operation of the accurate Radix-8 Booth encoding for FPGA implementation in generation of multiples of multiplicand, particularly addressing the challenge of generating a 3×Multiplicand. AxBM1 and AxBM2 approximates 3×Multiplicand by 4×Multiplicand. Also, -3×Multiplicand is approximated by -4×Multiplicand in AxBM1 and by -2×Multiplicand in AxBM2, respectively. To generate multiples of the multiplicand in AxBM1 and AxBM2, the 4× signal is generated through implicit XNOR operations of 1× and 2× signals in the LUT of PP bit generation, configuring the Booth encoder from a 5-input 2-output LUT configuration. However, AxBM1 and AxBM2 are characterized by non Multiply And Accumulate (MAC) PP accumulation and exhibit poor accuracy in multiplication.

The work [48] proposes ASMPEC, an approximate multiplier framework employing unique summation methods for PP generation and error correction, achieving substantial area savings and improved EE. The work [49] presents AxOCS, a methodology using ML based supersampling to design approximate operators, showing heightened optimization across different bit-widths and improvements in quality metrics.

Challenges in FPGA based approximate signed multiplication include the optimization of resource usage, and increasing efficiency in PP generation and accumulation through methods like Booth encoding and compressors. In the current work, LC methodology is a proposed heuristic methodology which aids in creating modular arithmetic multiplication circuits intended for FPGA implementation with controlled approximation. Literature study is summarized in Table 2-2.

Table 2-2 FPGA based techniques for approximate signed multiplication.

Work	**Research**
Kumm et al. (2015) [39]	Developed an accurate unsigned multiplication circuit using LUT and CCU with MAC used for PP accumulation.
Ullah et al. (2020) [38]	Reduced hardware resource consumption by 3 CCUs compared to the work [39] for PP generation, decreasing LUT consumption in a signed multiplication circuit using the Booth algorithm.
Ullah et al. (2020) [40]	Optimized resource consumption with Booth-Opt architecture for signed multiplication, integrating carry generating logic to reduce LUTs and CCs. Introduced Booth-Approx by approximating PP generation to lower CC and LUT resource consumption.
Ullah et al. (2021) [41]	Optimized PP generation and used a PP reduction tree to decrease critical path delay with respect to the work [40].
Venkatachalam et al. (2017) [45]	Employed approximate compressors for PP accumulation in multiplication circuits.
Van et al. (2020) [46]	Proposed configurable architectures for multiplication circuits using both accurate methods and approximate compressors for PP accumulation.
Waris et al. (2021) [47]	Proposed AxBM multiplication circuits based on Radix-8 Booth encoding, addressing challenges in generating multiples of the multiplicand with functional approximations.

Work	Research
Aizaz et al. (2023) [48]	Proposed ASMPEC, an approximate multiplier framework with unique summation methods and error correction.
Sahoo et al. (2024) [49]	Proposed AxOCS methodology using ML based supersampling for approximate operators.

2.3 Systematic analysis of impact of approximate multiplication on wireless signal detection.

In the context of next-generation communication systems, which prioritize flexible performance targets to enhance EE [50, 51], the use of approximate computing techniques becomes relevant as another approach for improving EE. These techniques enable a controlled trade-off by degrading system accuracy, thereby contributing to improved gains in EE in communication systems. A comprehensive survey [52] on the potential of approximate computing techniques in current and future B5G posits SER as a critical KPI for channel related issues, while EE remains a primary KPI for resource allocation. By integrating approximate computing techniques within a fixed power budget for communication systems, it becomes viable to reduce overall system power consumption. This reduction in power consumption subsequently creates additional capacity for system scaling within the same power budget.

The decoder utilizing the Successive Cancellation (SC) algorithm improves the Forward Error Correction (FEC) performance of polar codes; however, it introduces constraints on the throughput of hardware implementations. To address this challenge, the work [53] introduces configurable approximation units into modified computation function blocks within the SC algorithm, aiming to boost decoder throughput. The work [54] investigates the error-resilient characteristics of the inherent Fast Fourier Transform

(FFT) operation in industrial wireless communication systems, illustrating the potential advantages of approximate computing. Exact addition and subtraction operations within the FFT's butterfly structure are substituted with approximate adders, and the impact of these modifications on system level performance is thoroughly evaluated. It also highlights the significance of establishing correlations between the attributes of approximate adders and overall system reliability and performance metrics.

The work [55] explores the application of approximate computing in the EP algorithm used for the Sparse Code Multiple Access (SCMA) receiver. By employing approximation technique, the complexity of the algorithm is reduced, which is characterized by the number of arithmetic operations. Approximations are integrated into the EP algorithm at the variable and function node updates, as well as in the calculation of LLR to decrease computational complexity. Further, parameter optimizations are proposed to strike a balance between detection performance and algorithm computational complexity.

In the work [56], exact computing units are substituted with approximate ones in Root Raised Cosine (RRC) Finite Impulse Response (FIR) filters used for pulse shaping at the BS and decoders/equalizers at the UE in SISO and MIMO 6G downlink operations. The Bit Error Rate (BER) performance of the proposed approximate computing empowered 6G SISO downlink is superior to its MIMO counterpart, where the induced approximations achieve substantial power savings. The BER performance degradation is more pronounced in the high SNR regime compared to the low/medium SNR regime. The work in [57] utilizes gradient bounds to propose a novel encoding scheme for Quadrature Amplitude Modulation (QAM) mapping in the communication system required for a federated learning model. In a fixed SNR scenario, the test accuracy

of the model deteriorates as the QAM order increases.

While approximation techniques can improve EE and reduce hardware resource consumption at the cost of reduced accuracy, their application often tend to compromise system reliability. In the work [58], strategies for testing approximate circuits are detailed, emphasizing the role of reliability in deploying approximation techniques across systems. To tackle the reliability challenge, accurate estimation of system accuracy relative to the level of approximation induced becomes essential. In the work [59], approximation techniques are explored based on deterministic accuracy and granularity control.

Although approximate computing holds potential for improving EE and resource efficiency, it presents a challenge in terms of system stability and reliability if applied without proper control. As a result, significant research has been dedicated to understanding the impact of approximation on the system's QoS. This work aims to address this challenge by conducting a systematic analysis that links the level of approximation in the multiplication operation of signal detection to the system's QoS. Literature study is summarized in Table 2-3.

Table 2-3 Approximate Computing Techniques for B5G.

Work	**Objective of Research**
Damsgaard (2023) [51, 52]	Discussed the relevance of approximate computing techniques in enhancing EE and the potential of approximate computing techniques in B5G.
Zhou et al. (2018) [53]	Introduced configurable approximation units in the SC algorithm for polar codes to enhance FEC performance without compromising throughput.
Hao et al. (2019) [54]	Explored the use of approximate adders in FFT operations in industrial wireless communication systems to improve system-level performance and reliability.

Work	**Objective of Research**
Xiao et al. (2019) [55]	Applied approximate computing techniques to reduce complexity in the EP algorithm used in SCMA receivers.
Idrees et al. (2021) [56]	Substituted exact computing units with approximate ones in RRC FIR filters for 6G downlink operations, achieving superior BER performance.
Ma et al. (2023) [57]	Proposed gradient bounds for QAM mapping for federated learning applications, balancing accuracy degradation with modulation order in fixed SNR scenarios.
Anghel et al. (2018) [58]	Detailed testing strategies for approximate circuits, focusing on reliability considerations.
Moreau et al. [59]	Explored accuracy and granularity control in approximation techniques.

Chapter 3 - Algorithm decentralization for MIMO uplink detection

3.1 Résumé Long

3.1.1 Contexte de Recherche

MIMO a été intégré dans la 5G pour augmenter de manière significative l'efficacité spectrale et la fiabilité [60, 61]. Les développements récents indiquent que XL-MIMO promet d'atteindre une efficacité spectrale supérieure par rapport aux systèmes MIMO traditionnels [62], ouvrant la voie des standards de communication actuels vers le B5G. Cette avancée permet à une BS de servir simultanément un plus grand nombre de UEs, en utilisant des milliers d'antennes et en bénéficiant potentiellement de stratégies de déploiement sans cellule.

Pour la détection de signaux MIMO, des techniques de traitement centralisé comme ZF et MMSE offrent des performances optimales mais posent des défis dans la mise en œuvre matérielle [63] et sont viables avec un nombre limité de UEs et d'antennes de BS. Cependant, l'évolutivité pour un plus grand nombre de UEs nécessite d'augmenter les antennes de BS, ce qui accroît la bande passante d'interconnexion entre les antennes et l'unité de traitement central au niveau du fronthaul [62, 64]. En outre, transférer toutes les Channel State Informations (CSIs) collectées par les antennes de BS à l'unité centrale de traitement de BS augmente la complexité du traitement et la latence, réduisant le débit système [65]. Les techniques de détection de signaux D-MIMO facilitent la réduction de la bande passante d'interconnexion ainsi que la complexité computationnelle et la consommation de ressources matérielles.

3.1.2 Méthodologie

Les techniques contemporaines de D-MIMO ont été analysées pour leur application au XL-MIMO. L'algorithme DN a été proposé dans ce travail, en tirant parti de l'optimisation de Newton. Une analyse comparative a été menée pour évaluer la fidélité des signaux de l'algorithme DN par rapport aux méthodes contemporaines via des simulations en Python utilisant les bibliothèques Numpy et Mpmath, intégrant des scénarios de canaux réalistes du modèle de canal spatial 3rd Generation Partnership Project (3GPP). De plus, deux topologies matérielles — en anneau et en étoile — ont été proposées pour une implémentation sur FPGA. Ces topologies ont été conçues pour optimiser la bande passante d'interconnexion, la latence, le débit, la consommation de ressources matérielles et l'efficacité énergétique (EE, bits par Joule). Des prototypes matériels des topologies en anneau et en étoile ont été développés en utilisant la synthèse de haut niveau dans Vivado, ciblant les dispositifs FPGA Xilinx Virtex 7 et Ultrascale.

3.1.3 Synthèse Complète

Considérons une configuration D-MIMO utilisant une modulation 16-QAM avec trois itérations. La BS dispose de 128 antennes, divisées en quatre clusters de 32 antennes chacune, desservant 8 UEs. La méthode EP-LLR présente la bande passante d'interconnexion la plus élevée. En comparaison, la bande passante d'interconnexion de l'anneau DN est 90$\times$ inférieure, tandis que celle de l'étoile DN est 17$\times$ inférieure. En évaluant les performances de SER, la méthode EP excelle sous AWGN et le modèle de canal spatial 3GPP, dépassant la détection ZF. Pendant ce temps, les performances de DN se rapprochent de celles de la détection ZF sous AWGN et avec des itérations accrues sous le modèle de canal spatial 3GPP. Cependant, la méthode EP présente la complexité computationnelle du troisième ordre la plus élevée pour les clusters non-apex, avec

une complexité linéaire pour les clusters apex impliquant des opérations exponentielles. L'approche DN maintient une complexité linéaire constante pour les clusters apex et non-apex. LAMA ne converge pas dans le modèle de canal spatial 3GPP.

Cependant, la méthode EP présente la complexité computationnelle du troisième ordre la plus élevée pour les clusters non-apex, avec une complexité linéaire pour les clusters apex impliquant des opérations exponentielles. L'approche DN maintient une complexité linéaire constante pour les clusters apex et non-apex. LAMA ne converge pas dans le modèle de canal spatial 3GPP.

L'étoile DN utilise environ 1,05× les DSP48E, environ 0,84× les Flip Flop (FF), et environ 0,71× les LUT consommés par l'anneau DN. La DCD implémentée sur un Graphics Processing Unit (GPU) atteint la latence la plus faible. L'implémentation FPGA de l'anneau DN présente une latence 2,4× plus élevée, tandis que la latence de l'étoile DN est 1,5× plus élevée que celle de la DCD. Concernant le débit maximal, LAMA-FD offre le débit le plus élevé, dépassant l'implémentation FPGA de l'anneau DN de 1,9× et l'étoile DN de 5,3×.

La consommation d'énergie est particulièrement élevée pour LAMA-FD, LAMA-PD et DCD en raison de leurs implémentations sur GPU, tandis que l'anneau DN consomme 78,4× moins d'énergie et l'étoile DN consomme 169× moins d'énergie. En termes de nombre de bits transmis par Joule, DCD présente la plus faible efficacité. Cependant, l'anneau DN permet de transmettre environ 51,5× plus de bits par Joule, tandis que l'étoile DN en transmet environ 40,5× plus par Joule.

DN démontre une faible complexité computationnelle, permettant la transmission d'un nombre plus élevé de bits par Joule grâce à l'implémentation sur FPGA tout en atteignant

des performances de détection SER comparables à celles de la détection ZF. La topologie en anneau DN offre un débit élevé avec une bande passante d'interconnexion constante, tandis que la topologie en étoile DN assure une latence plus faible avec des variations prévisibles de la consommation des ressources matérielles à mesure que le système évolue. La topologie en anneau DN peut accueillir des sous-porteuses supplémentaires avec une augmentation fractionnelle de la latence et un débit système accru. Le tableau 3-2 met en évidence une comparaison des KPIs des topologies en anneau et en étoile DN.

Table 3-1 Comparaison des topologies en Anneau et en Étoile DN.

KPI	**Anneau DN (FPGA)**	**Étoile DN (FPGA)**
Bande passante d'interconnexion	90× inférieure à EP-LLR.	17× inférieure à EP-LLR.
Performance SER	Approche la détection ZF sous AWGN et le modèle de canal spatial 3GPP avec des itérations accrues.	
Complexité computationnelle	Linéaire pour les clusters apex et non-apex.	
Latence	2,4× supérieure à DCD (GPU).	1,5× supérieure à DCD (GPU).
Utilisation matérielle	L'anneau DN consomme 1,05× DSP48E, 0,84× FF, 0,71× LUT de l'étoile DN.	
Débit	1,9× inférieur à LAMA-FD (FPGA).	5,3× inférieur à LAMA-FD (FPGA).
Consommation d'énergie	78,4× inférieure aux méthodes sur GPU.	169× inférieure aux méthodes sur GPU.
Bits par Joule	51,5× plus de bits transmis par Joule que DCD (GPU).	40,5× plus de bits transmis par Joule que DCD (GPU).

3.1.4 Droits d'Auteur

L'article suivant est publié [66]. La permission est accordée en Annexe B.

3.2 Long abstract

3.2.1 Research Context

MIMO has been integrated into 5G to significantly boost spectral efficiency and reliability [60, 61]. Recent developments indicate that XL-MIMO holds promise for achieving superior spectral efficiency compared to traditional MIMO systems [62], paving the way of current communication standard towards B5G. This advancement allows BS to simultaneously serve a larger number of UEs, employing thousands of antennas, and potentially benefiting from cell-free deployment strategies.

For MIMO signal detection, centralized processing techniques like ZF and MMSE offer optimal performance but pose challenges in hardware implementation [63] and are viable with a limited number of UEs and BS antennas. However, scaling to more UEs requires increasing BS antennas, thereby raising interconnect bandwidth between antennas and the central processing unit at the fronthaul [62, 64]. Further, transferring all CSI gathered by BS antennas to the central processing unit of BS escalates processing complexity and latency at the BS, reducing system throughput [65]. D-MIMO signal detection techniques facilitate reduction in interconnect bandwidth as well as reduce computational complexity and hardware resource consumption.

3.2.2 Methodology

Contemporary D-MIMO techniques were analyzed for their application to XL-MIMO. The DN algorithm was proposed in this work, leveraging Newton optimization. A

comparative analysis was conducted to evaluate the signal fidelity of the DN algorithm against contemporary methods through simulations in Python using Numpy and Mpmath libraries, incorporating realistic channel scenarios from the 3GPP spatial channel model. Additionally, two hardware topologies—ring and star—were proposed for FPGA implementation. These topologies were designed to optimize interconnect bandwidth, latency, throughput, hardware resource consumption and EE (bits per Joule). Hardware prototypes of the ring and star topologies were developed using high-level synthesis in Vivado, targeting Xilinx Virtex 7 and Ultrascale FPGA devices.

3.2.3 Comprehensive Synthesis

Consider a D-MIMO configuration utilizing 16-QAM modulation with three iterations. The BS has 128 antennas, which are divided into four clusters of 32 antennas each, servicing 8 UEs. EP-LLR method has the highest interconnect bandwidth. In comparison, the interconnect bandwidth of the DN ring is 90$\times$ lower, while that of the DN star is 17$\times$ lower. When evaluating SER performance, the EP method excels under both AWGN and the 3GPP spatial channel model, surpassing that of ZF detection. Meanwhile, the DN approaches performance close to ZF detection under AWGN and with increased iterations under the 3GPP spatial channel model. However, EP method exhibits the highest third-order computational complexity for non-apex clusters, with linear complexity for apex clusters involving exponential operations. The DN approach maintains consistent linear complexity for both apex and non-apex clusters. LAMA does not converge in 3GPP spatial channel model.

The DN star uses approximately 1.05$\times$ the DSP48E, about 0.84$\times$ the FF, and roughly 0.71$\times$ the LUT consumed by DN ring. DCD implemented on a GPU achieves the lowest latency. FPGA implementation of DN ring has a latency that is 2.4$\times$ higher, while the

latency of DN star is 1.5× higher than that of the DCD. Regarding maximum throughput, LAMA-FD offers the highest throughput, surpassing the FPGA implementation of the DN ring by 1.9× and the DN star by 5.3×.

Power consumption is notably high for LAMA-FD, LAMA-PD, and DCD due to their GPU implementations, while the DN ring consumes 78.4× less power and the DN star consumes 169× less power. In terms of the number of bits transmitted per Joule, DCD exhibits the lowest efficiency. However, the DN ring allows for approximately 51.5× more bits to be transmitted per Joule, while the DN star achieves about 40.5× more bits per Joule.

DN demonstrates low computational complexity, allowing for the transmission of a higher number of bits per Joule due to FPGA implementation while achieving performance in SER detection that is comparable to that of ZF detection. DN ring topology offers high throughput with a constant interconnect bandwidth, whereas DN star topology ensures lower latency with predictable variations in hardware resource consumption as the system scales. DN ring topology can accommodate extra subcarriers with fractional increase in latency and increased system throughput. Table 3-2 highlights KPI comparision of DN ring and DN star.

Table 3-2 Comparison of DN Ring and DN Star.

KPI	**DN Ring (FPGA)**	**DN Star (FPGA)**
Interconnect Bandwidth	90× lower than EP-LLR.	17× lower than EP-LLR.
SER Performance	Approaches ZF detection under AWGN and 3GPP spatial channel model with increased iterations.	
Computational Complexity	Linear for both apex and non-apex clusters.	
Latency	2.4× higher than DCD (GPU).	1.5× higher than DCD (GPU).
Hardware Utilization	DN ring consumes 1.05× DSP48E, 0.84× FF, 0.71× LUT of DN star.	
Throughput	1.9× lower than LAMA-FD (FPGA).	5.3× lower than LAMA-FD (FPGA).
Power Consumption	78.4× lower than GPU based methods.	169× lower than GPU based methods.
Bits per Joule	51.5× more bits transmitted per Joule than DCD (GPU).	40.5× more bits transmitted per Joule than DCD (GPU).

3.2.4 Copyright

Hardware Topologies for Decentralized Large-Scale MIMO Detection Using Newton Method

Abhinav Kulkarni, Messaoud Ahmed Ouameur, *Member, IEEE*, and Daniel Massicotte, *Senior Member, IEEE*

***Abstract*— Centralized Massive Multiple Input Multiple Output (MIMO) uplink detection techniques for baseband processing possess severe bottleneck in terms of interconnect bandwidth and computational complexity. This problem has been addressed in the current work by adapting the centralized Newton method for decentralized MIMO uplink detection leveraging several Base Station antenna clusters. The proposed decentralized Newton (DN) method achieves error-rate performance close to centralized Zero Forcing detector as compared to other decentralized techniques. Two hardware topologies, namely the ring and the star topologies, are proposed to assess and discuss the trade-off among interconnect bandwidth and throughput, in comparison with contemporary decentralized MIMO uplink detection techniques. As such the following findings are elaborated. On BS antenna cluster scaling for different MIMO system configurations, the ring topology provides high throughput at constant interconnect bandwidth, while the star topology provides lower latency with a deterministic variation in the hardware resource consumption. Due to strategic optimizations on the hardware implementation, additional user equipment can be allotted at a fractional increase in Field Programmable Gate Array resource consumption, improved energy efficiency, and increased transaction of bits per Joule. The ring topology can process additional subcarrier at a fractional increase in latency and improved system throughput.**

***Index Terms*— MIMO uplink detection, Newton method, FPGA, decentralized processing, hardware topology, interconnect bandwidth.**

I. Introduction

BEING a promising concept for future cellular networks, Massive Multiple Input Multiple Output (MIMO) technology has now made its way to 5G as one of the means to substantially improve both spectral and energy efficiencies [1], [2]. Future trends for 6G suggest the use of Extremely Large Aperture Array (ELAA) to provide order-of-magnitude higher area throughput compared to what massive MIMO with compact arrays can ultimately deliver [3]. It is possible for a BS to service several UEs simultaneously within the same time-frequency resources using hundreds or thousands of BS antennas. Centralized linear processing techniques for MIMO uplink signal detection like Zero Forcing (ZF), Minimum Mean Square Error (MMSE), and Maximum Ratio Combining (MRC) estimate the UE's signals by Gram matrix inversion but have caveats on hardware implementation due to high computational complexity and impose severe bottleneck in terms of interconnect bandwidth as well [1].

A. Related Work

Digital signal processing architecture design with practical system constraints for the next generation Massive MIMO uplink detection techniques is presented in [4]. For a 16-QAM MIMO system configuration with 128 BS antennas (B) and 8 UEs (U), the system parameters for centralized techniques evaluated on a FPGA are compared hereafter. MIMO uplink detection based on Neumann Series (NS) [5] achieves a throughput of 402 Mbps and high error-rate performance for large B/U ratio, however this method scales to computational complexity of $\mathcal{O}(U^3)$ for 3 series expansion terms. Conjugate Gradient (CG) based MIMO uplink detection method [6] achieves a throughput of 13 Mbps with lower FPGA resource utilization and lower error-rate performance as compared to NS method. By efficiently implementing centralized Newton method [7], the MIMO uplink detector implementation achieves a staggering throughput of 610 Mbps. Co-ordinate Descent (CD) algorithm has been adapted for MIMO uplink detection in [8], achieving a throughput of 250 Mbps at low computational complexity of $\mathcal{O}(BU)$. To alleviate the high computational complexity of NS method, Gauss Seidel (GS) algorithm has been adapted for MIMO uplink detection in [9] and achieves a throughput of 32 Mbps. An improved version of GS [10] method, that uses multiple parallel sub-carrier instances by hardware interleaving, achieves a throughput of 488 Mbps. An efficient implementation of MMSE detection has been presented in [11] and achieves a throughput of 205 Mbps. By using adaptive Successive Over Relaxation (A-SOR) to achieve fast convergence, the hardware implementation in [12] achieves a throughput of 135 Mbps with $\mathcal{O}(U^2)$ computational complexity. For high energy efficiency, ASIC based implementations [11], [13] are more advantageous over FPGA based implementations [5]–[10], [12].

Manuscript received February 17, 2021; revised May 7, 2021 and June 13, 2021; accepted June 28, 2021. Date of publication July 26, 2021; date of current version August 10, 2021. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), in part by Prompt, in part by the Canadian Foundation for Innovation (CFI), in part by CMC Microsystems, in part by NUTAQ Innovation, and in part by the Laboratoire des signaux et systèmes intégrés and the Chaire de recherche sur les signaux et l'intelligence des systèmes haute performance (www.uqtr.ca/lssi). This article was recommended by Associate Editor G. Jovanovic Dolecek. *(Corresponding author: Daniel Massicotte.)*

The authors are with the Department of Electrical and Computer Engineering, Université du Québec à Trois-Rivières, Trois-Rivières, QC G9A 5H7, Canada (e-mail: abhinav.kulkarni@uqtr.ca; messaoud.ahmed.ouameur@uqtr.ca; daniel.massicotte@uqtr.ca).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSI.2021.3097042.

Digital Object Identifier 10.1109/TCSI.2021.3097042

Centralized baseband processing techniques are feasible on hardware for a small number of UE and a low number of BS antennas for real-time processing. However, as the number of UE grows, more BS antennas are required to achieve optimal performance which increases interconnect bandwidth between BS antennas and BS central processing unit [3]. Also, all Channel State Information (CSI) has to be transferred from BS antennas to the BS central processing unit which increases computational complexity and latency at BS, thereby decreasing system throughput [14] and possess a bottleneck to ELAA implementation [3]. To address this bottleneck, several decentralized baseband processing algorithms and accompanying architectures for MIMO uplink detection have been proposed, where the baseband processing of MIMO uplink signal detection is shared by several BS antenna clusters. Decentralized Co-ordinate Descent (DCD) [15] based MIMO uplink detection computes partial uplink signal at every distributed BS antenna cluster using co-ordinate descent method. Partial signal estimates are scaled by BS antenna cluster variance and fused to produce the final uplink signal at the BS. Decentralized Alternating Direction Method of Multipliers (D-ADMM) [16] is a high computational complexity method based on consensus exchange, providing near MMSE performance with few iterations for low UE load with respect to BS antenna cluster. ADMM-GS [17] embeds Gauss-Siedel iteration in ADMM for performance enhancement in terms of error rate. This method is suited for high SNR and high UE load scenarios and is robust to channel estimation errors. The Decentralized Conjugate Gradient (D-CG) [16] method provides near MMSE performance with few iterations in high UE load scenarios per BS antenna cluster. MIMO uplink detection with Stochastic Gradient Descent (SGD) [18] uses fully decentralized architecture in Daisychain topology. In this technique, the central cluster does not have to be reconfigured when adding new BS antenna clusters, and the interconnect bandwidth between two clusters remains constant for a given number of UE.

Maximum A Posteriori (MAP) estimate based decentralized algorithms like large-MIMO approximate message passing (LAMA) with two architectures one for partially decentralized (LAMA-PD) and another one for fully decentralized (LAMA-FD) [19] and Expectation Propagation (EP) [20], [21] provide high error-rate performance at expense of increased algorithm computational complexity. MIMO uplink detection using LAMA-PD and LAMA-FD [19] equalization provides optimal performance given channel matrix $\mathbf{H}$ has i.i.d distribution and profiled with the variance of $1/B$, where B represents the number of BS antennas. However, LAMA is not robust for realistic channel environments [22]. MIMO uplink detection using EP [20] is a comparable algorithm to LAMA and involves explicit matrix inversion, which increases its computational complexity. MIMO uplink detection using EP with Log Likelihood Ratio (LLR) [21] provides improved performance than LAMA, especially at high SNR, but requires high interconnect bandwidth. MAP methods incur additional computing overheads to improve numerical stability for variance computation from noise statistics. Also, the partial local estimates have to be fused and processed using a soft-detector for uplink signal estimation in MAP methods, an improvement has been suggested by [23]. Tree K-ary based MIMO uplink detection architecture [24] discusses a decentralized scalable BS system, where interconnect links grow logarithmically on the addition of BS antenna clusters. Most of these techniques are used herein as benchmarks to discuss the error rate performance, throughput, interconnect bandwidth and the computational complexity.

B. Contributions

The choice of a MIMO uplink detection technique is based on MIMO system requirements [2] and it is a non-trivial task. Hence, there is a trade-off between error-rate performance, hardware computational complexity, latency, and system throughput based on the wireless propagation environment parameters [25]. With advancements in computing and RF technology, massive MIMO will gradually evolve into extremely large-scale MIMO systems where BS will function with thousands of antennas and in such scenarios, decentralized architectures would be more favorable. With such large MIMO antenna configurations, even the MAP methods with high computational complexity show diminishing benefits[25]. In the evolving communication standards towards 6G [26], factors of interconnect bandwidth and energy efficiency would also play a prime role along with throughput and latency for large MIMO systems. In the current work, the following contributions are presented:

- The adaptation of the centralized Newton method [27], [28] for decentralized processing of MIMO uplink detection is achieved by constructing novel local objective functions over decentralized BS antenna clusters (which we refer to as clusters for brevity). The proposed decentralized Newton (DN) algorithm provides close to the ZF symbol-error rate performance as compared to contemporary decentralized MIMO uplink detection techniques, specifically in low SNR regime and 3GPP radio channel environment.
- At the *system level*, novel proposition of ring and star topological architectures for VLSI hardware implementation to achieve gradient and Hessian sampling for the DN method, leveraging decentralized clusters at the *circuit level* to achieve trade-off among throughput, latency, energy efficiency and interconnection bandwidth.
- Analytical analysis of the interconnect bandwidth of the star and ring topologies with contemporary decentralized MIMO uplink detection techniques at the *system level*. The star topology provides lower interconnect bandwidth than EP, EP-LLR and ADMM-GS. The ring topology has lower interconnect bandwidth than the star topology and maintains constant interconnect bandwidth on MIMO configuration scaling.
- Analysis of computational complexity of the star and ring topologies with contemporary MIMO uplink detection techniques at the circuit level. Interestingly, the DN algorithm's computational complexity is in order in the number of UEs and avoids signal variance computation.

- At the *system level*, design space exploration for the hardware implementation of the star and ring topologies on FPGA and analysis of the effect of MIMO configuration scaling on system throughput, latency, energy efficiency and hardware resource consumption. The star topology provides low latency while the ring topology provides higher throughput. The implementation of the ring topology with additional sub-carrier requires a fractional increase in hardware resource consumption.
- Provide a comparative analysis of hardware implementation of the star and ring topologies with hardware architectures of contemporary MIMO uplink detection techniques at the *system level*. The star and ring topologies are feasible to implement on FPGA with high energy efficiency.

For notations, uppercase bold letter represents a matrix and lowercase bold letter represents a column vector. (t) denotes t^{th} iteration. L2 vector norm is represented as $||.||_2$. $\nabla_{\mathbf{x}}$ represents first degree gradient operator w.r.t to $\mathbf{x}$. $\nabla^2_{\mathbf{x}}$ represents second degree gradient operator w.r.t to $\mathbf{x}$. $\mathbb{E}$ represents expectation operator. For a matrix $\mathbf{A} \in \mathbb{C}^{B\times U}$, $[\mathbf{a}_1\ \mathbf{a}_2\ \mathbf{a}_3\ \dots\ \mathbf{a}_U]$ represents $\mathbf{A}$ as set of column vectors, where $\mathbf{a}_i \in \mathbb{C}^{B\times 1}$. For a matrix $\mathbf{A}$, $\mathbf{A}^H$ represents complex conjugate transpose of $\mathbf{A}$. The operation **diag**(.) extracts major diagonal of a square matrix as a column vector. The operation **diagdiag**(.) is the inverse of **diag**(.) and constructs diagonal matrix with given column vector as a major diagonal.

The paper is organized as follows; Section I.*A* discusses related work on Massive MIMO uplink detection techniques, specifically motivating the need for decentralized processing techniques. Section I.*B* presents the novel contributions of the current work. Section II lays the foundation for the decentralized Newton-based MIMO uplink detection technique and derives topological architectures for hardware implementation. Section III provides a comparative analysis against interconnect bandwidth for contemporary decentralized MIMO uplink detection techniques. Section IV analyses the computational complexity of contemporary decentralized MIMO uplink detection techniques. Section V discusses simulation and error rate performance analysis for decentralized MIMO uplink detection techniques. Section VI provides hardware implementation for the ring and star topologies and draws detailed hardware implementation analysis for both topologies, with comparative analysis with other decentralized MIMO uplink detection techniques. Section VII ends the discussion with the conclusion and future potential of ring and star topologies.

II. Proposed Technique

For the pre-processing of the DN method, local objective function f_c for $c = 1, 2, \dots, C$ at every cluster is constructed. For a generic BS model, a BS with B antennas serving U number of UEs is considered. Without loss of generality, every UE is assumed to be equipped with a single antenna. Expression **y=Hx+n** represents MIMO uplink signal at BS, where $\mathbf{y} \in \mathbb{C}^{B\times 1}$ is the vector representing receive signal over B antennas of the BS, $\mathbf{x} \in \mathbb{C}^{U\times 1}$ being signal estimate, which is transmitted from the UEs to BS. $\mathbf{H} \in \mathbb{C}^{B\times U}$ is the wireless channel model. $\mathbf{n} \in \mathbb{C}^{B\times 1}$ is the channel noise. $\mathbf{x}$ is mapped to Q bits symbol which form 2^Q-QAM modulation. As shown in Fig. 1, B antennas, $\mathbf{H}$ and $\mathbf{y}$ are equally distributed into C clusters such that every cluster c is characterized by local antennas B_c, local channel matrix $\mathbf{H}_c \in \mathbb{C}^{B_c\times U}$ and local received signal vector $\mathbf{y}_c \in \mathbb{C}^{B_c\times 1}$. The total number of antennas for the BS is represented as $B = \sum_{c=1}^{C} B_c$. $\mathbf{H}_c = [\mathbf{h}_{1,c}\ \mathbf{h}_{2,c}\ \mathbf{h}_{3,c} \dots \mathbf{h}_{u,c}]$ where $u = 1, 2, 3 \dots U$; $c = 1, 2, 3 \dots C$. $\mathbf{H}_c$ and $\mathbf{y}_c$ are known locally only to the cluster c and are not exchanged within clusters.

Lemma 1: Given $\mathbf{H}_c$ and $\mathbf{y}_c$ for $c = 1, 2, 3 \dots C$, uplink estimate at t^{th} iteration can be computed as:

$$\mathbf{x}^{(t)} = \mathbf{x}^{(t-1)} - (\mathbf{D})^{-1}\left(\sum_{c=1}^{C}(\mathbf{H}_c^H\mathbf{H}_c\mathbf{x}^{(t-1)} - \mathbf{H}_c^H\mathbf{y}_c)\right) \quad (1)$$

The detailed derivation of eq. (1) is postponed to the Appendix whereas Topologies 1 and 2 show the DN algorithm (c.f. Appendix) using two different hardware topologies as depicted in Fig. 1. The algorithm is terminated at iteration $t = T$ to obtain $\mathbf{x}^{(T)}$, which is processed using QAM decoder to obtain the uplink signal estimate. While computing eq. (1) it is important to note quantities that are static for a specific interval. In MIMO uplink signal transmission, the channel statistical characteristics remain constant during a specified interval of time. This time interval is called coherent time and $\mathbf{H}_c$ for $c = 1, 2, 3 \dots C$ remains constant during the coherent time interval. Hence, the Gram matrix $\mathbf{H}_c^H\mathbf{H}_c$ for each cluster and the approximate Hessian diagonal matrix $\mathbf{D}$ at apex cluster C have to be computed once every coherent time interval. Thus matrix multiplication of $\mathbf{D}^{-1}$ with eq. (5) involves U complex divisions, which is insignificant as compared to the total complex multiplications involved in overall algorithm.

On a single cluster, eq. (10) is implemented to obtain an uplink signal estimate. However, it is essential to design architectures that can be implemented to accumulate local computations at a single cluster.

In essence, the ring and star topologies for the DN algorithm for the MIMO uplink detection algorithm are proposed. These topological architectures enable the provision of explicit trade-offs among system latency, throughput, interconnect bandwidth, energy efficiency and hardware resource consumption. Fig. 1 shows the implementation of both topological architectures. Partial computations in a cluster are represented in eq. (6) and (5).

The ring topology is characterized by clusters organized in daisy-chain fashion. Every cluster is exactly connected to two adjacent clusters. Except the apex cluster, all the clusters are identical in functionality. Thus, every cluster receives partial computations from prior cluster, appends its local partial computations and sends resultant computations to the next cluster in the daisy-chain. All the cluster interconnections are unidirectional. As shown in Fig. 1.a, cluster C acts as an apex cluster. The apex cluster provides partial computations and also computes eq. (10) to produce $\mathbf{x}^{(t)}$ at the t^{th} iteration. To facilitate the flow of partial computations between the interconnected clusters, the interconnect variables $\mathbf{p} \in \mathbb{C}^{U\times 1}$

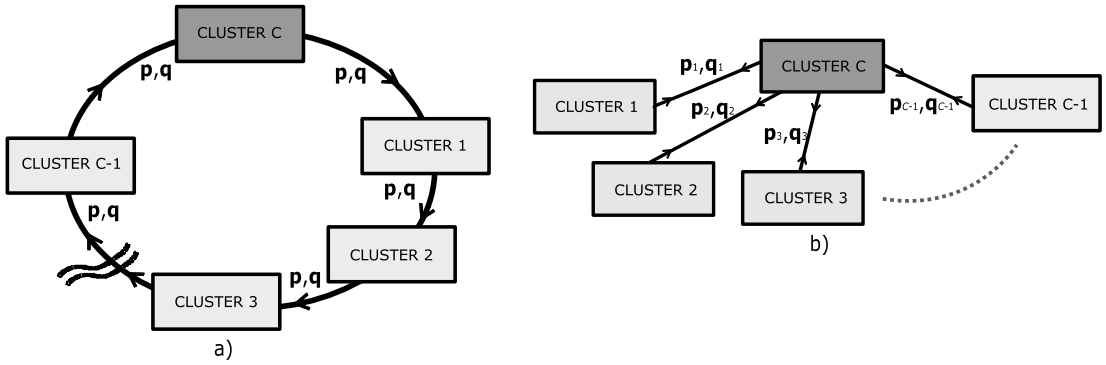

Fig. 1. DN based MIMO uplink detector implemented using a) ring topology with C clusters and interconnect variables as $\mathbf{p} \in \mathbb{C}^{U\times 1}$, $\mathbf{q} \in \mathbb{C}^{U\times 1}$ and b) star topology with C clusters and interconnect variables as $\mathbf{p}_c \in \mathbb{C}^{U\times 1}$, $\mathbf{q}_c \in \mathbb{C}^{U\times 1}$, where $c = 1, 2, 3, \ldots, C-1$. Every antenna cluster c with B_c antennas is responsible for processing local partial computations. $\mathbf{H}_c$ and $\mathbf{y}_c$ is local to every cluster and is not exchanged between clusters.

Topology 1 DN Ring Topology

Input: $\mathbf{H}_c, \mathbf{y}_c \quad c = 1, 2, 3 \ldots C$
Output: $\mathbf{x}^{(T)}$
Initialization:
Calculate $\mathbf{D}_c$ from $\mathbf{H}_c$ using eq.(8) for $c = 1, 2, 3 \ldots C$
Initial iteration $t = 1$
for $c = 1$ **to** C **do**
  $\mathbf{x}_c \leftarrow \mathbf{D}_c^{-1}(\mathbf{H}_c^H \mathbf{y}_c)$
  $\mathbf{p} \leftarrow \mathbf{p} + \mathbf{diag}(\mathbf{D}_c)$ {Accumulate: eq.(9)}
  $\mathbf{q} \leftarrow \mathbf{q} + (\mathbf{H}_c^H \mathbf{H}_c \mathbf{x}_c - \mathbf{H}_c^H \mathbf{y}_c)$ {Accumulate: eq.(5)}
  if $c = C$ **then**
    $\mathbf{D} = \mathbf{diagdiag}(\mathbf{p})$ {Local store $\mathbf{D}$ at cluster C}
    $\mathbf{x}^{(1)} \leftarrow \mathbf{x}_c - \mathbf{D}^{-1}\mathbf{q}$ {Evaluate: eq.(10)}
    $\mathbf{p} \leftarrow \mathbf{x}^{(1)}$ {Broadcast $\mathbf{x}^{(1)}$}
    $\mathbf{q} \leftarrow 0$ {Flush}
  end if
end for
for $t = 2$ **to** T **do**
  for $c = 1$ **to** C **do**
    $\mathbf{q} \leftarrow \mathbf{q} + (\mathbf{H}_c^H \mathbf{H}_c \mathbf{p} - \mathbf{H}_c^H \mathbf{y}_c)$ {Accumulate: eq.(5)}
    if $c = C$ **then**
      $\mathbf{x}^{(t)} \leftarrow \mathbf{x}^{(t-1)} - \mathbf{D}^{-1}\mathbf{q}$ {Evaluate: eq.(10)}
      $\mathbf{p} \leftarrow \mathbf{x}^{(t)}$ {Broadcast $\mathbf{x}^{(t)}$}
      $\mathbf{q} \leftarrow 0$ {Flush}
    end if
  end for
end for

Topology 2 DN Star Topology

Input: $\mathbf{H}_c, \mathbf{y}_c \quad c = 1, 2, 3 \ldots C$
Output: $\mathbf{x}^{(T)}$
Initialization:
Calculate $\mathbf{D}_c$ from $\mathbf{H}_c$ using eq.(8) for $c = 1, 2, 3 \ldots C$
Initial iteration $t = 1$
for $c = 1$ **to** C **do**
  $\mathbf{x}_c \leftarrow \mathbf{D}_c^{-1}(\mathbf{H}_c^H \mathbf{y}_c)$
  $\mathbf{p}_c \leftarrow \mathbf{diag}(\mathbf{D}_c)$
  $\mathbf{q}_c \leftarrow (\mathbf{H}_c^H \mathbf{H}_c \mathbf{x}_c - \mathbf{H}_c^H \mathbf{y}_c)$
  if $c = C$ **then**
    $\mathbf{p} = \sum_{c=1}^{C} (\mathbf{p_c})$ {Accumulate: eq.(9)}
    $\mathbf{q} = \sum_{c=1}^{C} (\mathbf{q_c})$ {Accumulate: eq.(5)}
    $\mathbf{D} = \mathbf{diagdiag}(\mathbf{p})$ {Local store $\mathbf{D}$ at cluster C}
    $\mathbf{x}^{(1)} \leftarrow \mathbf{x}_c - \mathbf{D}^{-1}\mathbf{q}$ {Evaluate: eq.(10)}
    $\mathbf{p}_c \leftarrow \mathbf{x}^{(1)}$ {Broadcast $\mathbf{x}^{(1)}$}
    $\mathbf{q}_c \leftarrow 0$ {Flush}
  end if
end for
for $t = 2$ **to** T **do**
  for $c = 1$ **to** C **do**
    $\mathbf{q}_c \leftarrow (\mathbf{H}_c^H \mathbf{H}_c \mathbf{p}_c - \mathbf{H}_c^H \mathbf{y}_c)$
    if $c = C$ **then**
      $\mathbf{q} = \sum_{c=1}^{C} (\mathbf{q_c})$ {Accumulate: eq.(5)}
      $\mathbf{x}^{(t)} \leftarrow \mathbf{x}^{(t-1)} - \mathbf{D}^{-1}\mathbf{q}$ {Evaluate: eq.(10)}
      $\mathbf{p}_c \leftarrow \mathbf{x}^{(t)}$ {Broadcast $\mathbf{x}^{(t)}$}
      $\mathbf{q}_c \leftarrow 0$ {Flush}
    end if
  end for
end for

and $\mathbf{q} \in \mathbb{C}^{U\times 1}$ are considered. For the initial iteration $t = 0$, the variable $\mathbf{p}$ accumulates diagonal vector of $\mathbf{D}_c$ from non-apex clusters to the apex cluster C. The aggregate of $\mathbf{D}_c$ for $c = 1, 2, 3 \ldots C$ is available as $\mathbf{D}$ at the apex cluster and does not need to be computed until the next coherence time interval (since $\mathbf{H}_c$ remains constant during the coherent time interval.) For the next subsequent iterations $t = 2, 3, \ldots T$, the variable $\mathbf{p}$ is set with $\mathbf{x}^{(t)}$ at the apex cluster to be broadcasted and utilized for $(t+1)^{th}$ iteration in the computation of eq. (5). Since $\mathbf{x}^{(t)}$ is available at the end of iteration t, $\mathbf{x}_c$ is the local estimate used by cluster c in the computation of eq. (5) for $t = 0$. Initial estimate $\mathbf{x}_c$ is computed from Matched Filter $\mathbf{H}_c^H \mathbf{y}_c$ and the approximate Hessian $\mathbf{D}_c$. For all the iterations, variable $\mathbf{q}$ accumulates first gradient as partial computations of eq. (5). The DN algorithm for MIMO uplink detection mapped onto the ring topology is outlined in Topology 1.

The star topology is characterized by clusters connected to a single central processing cluster. The central processing cluster is the apex cluster denoted by C. The apex cluster is connected to other $C-1$ non-apex clusters. While every non-apex cluster

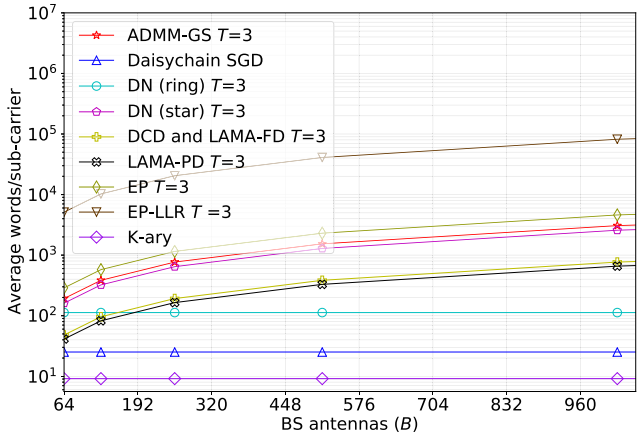

Fig. 2. Comparison of interconnect bandwidth for different decentralized MIMO uplink detection techniques with $U = 8$, $B_c = 32$ and $T = 3$.

is only connected to the apex cluster, the partial computations from all non-apex clusters are parallelly accumulated at the apex cluster. All the cluster interconnections are bidirectional. Similar to the ring topology, apex cluster provides partial computations along with eq. (10) to compute $\mathbf{x}^{(t)}$ at the t^{th} iteration. For interconnect transfers, the variables $\mathbf{p}_c \in \mathbb{C}^{U\times 1}$ and $\mathbf{q}_c \in \mathbb{C}^{U\times 1}$ for clusters $c = 1, 2, 3 \ldots C$ are considered, which handle interconnection transfer between the apex cluster and non-apex clusters. Partial computations for $\mathbf{p}_c$ and $\mathbf{q}_c$ for apex cluster $c = C$ are done internally. For initial iteration $t = 0$, the variable $\mathbf{p}_c$ accumulates diagonal vector of $\mathbf{D}_c$ from all non-apex clusters at the apex cluster to form $\mathbf{D}$, which remains constant for the coherent time interval. For the next subsequent iterations, $\mathbf{p}_c$ broadcasts $\mathbf{x}^{(t)}$ to all non-apex clusters $c = 1, 2, 3, \ldots C-1$ at the $(t+1)^{th}$ iteration. Similar to the ring topology, eq. (5) is computed using the local estimate of $\mathbf{x}_c$ for the initial iteration $t = 0$. For all the iterations, the variable $\mathbf{q}_c$ for $c = 1, 2, 3 \ldots C$ accumulates partial computations of eq. (5) to the apex cluster. The proposed DN algorithm for MIMO uplink detection mapped onto the star topology is outlined in Topology 2.

III. Interconnect Bandwidth

In centralized MIMO detection techniques, data from B antennas have to be transferred to the computing circuit of the apex cluster, which becomes a bottleneck when the detection technique is scaled to very large B as the bandwidth between antennas and computing circuit is dependent on B. So, decentralized MIMO detection techniques are employed, where B antennas are distributed into C clusters, and every cluster performs local partial computations. Local partial computations are aggregated over to the apex cluster to produce uplink signal estimation. For the apex cluster, the interconnection bandwidth is independent of B in decentralized MIMO uplink detection techniques, thereby mitigating for high data transfer between clusters. Fig.2 evaluates average interconnect transaction occurring during the coherent time interval of 1.0 millisecond (mapped to $N_{coh} = 14$ symbols) at the apex cluster.

TABLE I
Comparison of Interconnect Bandwidth

Technique	Words	Words ($B = 512$)
DCD [15]	$3CU$	384
ADMM-GS [17]	$4TCU$	1536
Daisychain SGD [18][1]	$\frac{2U^2}{N_{coh}} + 2U$	25
LAMA-PD [19]	$\frac{C\times(U^2+2N_{coh}U)}{N_{coh}}$	329
LAMA-FD [19]	$3CU$	384
EP [20]	$6TCU$	2304
EP-LLR [21][2]	$2(2T-1)UCB_c$	40960
K-ary [24][3]	$\frac{U(U+1)}{N_{coh}} + 4$	9
DN ring topology	$\frac{(2N_{coh}+1)U}{N_{coh}} + 6U(T-1)$	113
DN star topology	$\frac{(2N_{coh}+1)CU}{N_{coh}} + 4CU(T-1)$	1289

[1] Total words for formulation and filtering phase.
[2] T is the number of outer iteration.
[3] Wordlength is $2U$ for transmitting N_{coh} symbols.

For interconnect bandwidth analysis, every real entity is denoted as a word and a complex number is comprised of 2 words[14]. Interconnect bandwidth is measured by average words transferred during a coherence interval. On considering a prominent scenario for which the estimated channel in the uplink is static across a coherent time interval of N_{coh} contiguous symbols, T be the number of total iterations and C be the total number of clusters. For calculating words transacted by apex cluster for a decentralized algorithm, input and output signals are taken into account for every iteration. Interconnect transfer for LAMA-PD and LAMA-FD is given in [14]. For DN method, approximate local Hessian needs to be transmitted to the apex cluster once every coherent interval, which comprises of U words (real-valued diagonal elements). Accordingly, the apex cluster in the ring topology receives an aggregate of $(2N_{coh} + 1)U$ words during the first iteration considering all symbols in coherence interval. During subsequent iterations, the apex cluster transmits an aggregate of $4N_{coh}(I-1)U$ words and receives an aggregate of $2N_{coh}(I-1)U$ words for all symbols in coherence interval. For the star topology, the apex cluster receives an aggregate of $(2N_{coh}+1)CU$ words during the first iteration considering all symbols in coherence interval. For subsequent iterations, the apex cluster transmits an aggregate of $2N_{coh}(I-1)U$ words and receives an aggregate of $2N_{coh}(I-1)U$ words for all symbols in coherence interval. The average interconnect transfer for the ring and star topologies for coherence time interval is the average of total words transmitted and received for N_{coh} symbols for all iterations T, which is outlined in Table. I.

EP-LLR has the highest interconnect bandwidth. The interconnect bandwidth of the star topology is lower than ADMM-GS, EP and EP-LLR. For low number of BS antennas, DCD has a lower interconnect bandwidth than the ring topology. However, as the number of BS antennas increase, bandwidth of DCD also increases and surpasses that of constant bandwidth

of ring topology. DCD has a lower symbol-error rate performance with higher apex cluster computational complexity as compared to DN, which can be used to trade-off with DN even at lower number of BS antennas. LAMA-PD and LAMA-FD also have lower interconnect bandwidth than star topology, however, comparatively they have higher computational complexity and are less robust in practical wireless channel environments as investigated in [22]. The interconnect bandwidth for the star topology, EP, EP-LLR, LAMA-PD, LAMA-FD and DCD depends on the number of the clusters.

To improve upon the interconnect bandwidth performance for the DN algorithm, the ring topology exhibits lower interconnect bandwidth than LAMA-PD. The interconnect bandwidth of Daisychain SGD is lower than the ring topology, however it has a lower symbol-error rate performance by at least 3 dB and has a limitation of a single antenna per cluster. K-ary is a generic topology and has the lowest interconnect bandwidth which is provided for reference. The ring topology, Daisychain SGD and K-ary maintain constant interconnect word transfer on scaling the BS for large B antennas and the apex cluster does not have to be hardware reconfigured while varying B. Adapting the proposed DN algorithm for K-ary topology to further reduce the interconnect bandwidth is a non-trivial task and is part of ongoing research.

IV. Complexity Analysis

The computational complexity of an algorithm is mainly characterized by the number of complex multiplication and division operations. It is important to evaluate the computational complexity of MIMO uplink detection algorithms when the parameters of U, B, C, B_c, and T are varied for different MIMO system configurations. As B_c and T are fixed in scaling MIMO configuration, the critical parameters to be considered are U, B and C for analyzing the computational complexity. The MAP based MIMO uplink detection techniques of LAMA-FD, LAMA-PD, EP and EP-LLR involve exponential operation to compute the signal variance. Hence, for deriving VLSI architectures based on these algorithms, it is critical to explicitly account for the computational complexity for the implementation of exponential operation to ensure numerical stability. The performance of MAP based MIMO uplink detection algorithms depends on numerical stability, specifically at high SNR when variance becomes infinitesimally small.

The computational complexity of decentralized MIMO uplink detection algorithms is evaluated in Table. II. The order of computational complexity for the proposed DN algorithm is not affected by the choice of the topology. The EP algorithm has the highest computational complexity of third order at the non-apex clusters due to explicit matrix inversion. The LAMA-FD, LAMA-PD and ADMM-GS exhibit computational complexity of second order, however ADMM-GS does not involve the computation of the exponential operations. The Daisychain SGD exhibits second order computational complexity in terms of U and depends on the number of BS antennas B, uniform across all clusters. The proposed DN, DCD and EP-LLR exhibit linear computational complexity across the apex cluster. The DN algorithm's computational complexity is dominantly affected by computation of Gram matrix $\mathbf{H}_c^H\mathbf{H}_c$ at the non-apex clusters. However, as compared to DCD and EP-LLR, the computational complexity of the DN algorithm is lower at apex cluster.

TABLE II
Complexity Comparison

Technique	Non-apex cluster	Apex cluster	#Exponential
DCD [15]	$\mathcal{O}(U)$	$\mathcal{O}(CU)$	-
ADMM-GS [17]	$\mathcal{O}(U^2)$	$\mathcal{O}(CU)$	-
Daisychain SGD [18]	$\mathcal{O}(U^2B)$	$\mathcal{O}(U^2B)$	-
LAMA-PD [19]	$\mathcal{O}(U^2)$	$\mathcal{O}(U^2)$	$\mathcal{O}(2^Q)$
LAMA-FD [19]	$\mathcal{O}(U^2)$	$\mathcal{O}(U^2)$	$\mathcal{O}(C\times 2^Q)$
EP [20]	$\mathcal{O}(U^3)$	$\mathcal{O}(CU)$	$\mathcal{O}(2^Q)$
EP-LLR [21]	$\mathcal{O}(U)$	$\mathcal{O}(BU)$	$\mathcal{O}(CU\times 2^Q)$
DN	$\mathcal{O}(U)$	$\mathcal{O}(U)$	-

V. Error Rate Performance Analysis

Decentralized MIMO detection techniques are compared by performing simulation of BS with 128 antennas servicing 8 UEs in Gaussian i.i.d and 3GPP SCM as channel models. Also, each of these system configurations is simulated with 16-QAM to analyze the effect of modulation scheme over symbol error-rate performance. DN method is simulated with the floating-point as well as fixed-point (inline with HLS analysis). For the simulation, 32-bit data type with 16-bit for the real part and 16-bit for the imaginary part of complex number representation is used, both for floating and fixed-point analysis. Since, the ring and the star topologies are architectures for VLSI hardware implementations, they do not affect the symbol error-rate performance of the proposed DN algorithm since both equate eq. (1) using the algorithm outlined in Appendix X. All arithmetic operations for the simulation are performed using Python Numpy [31] and Python Mpmath [32].

Fig. 3.a and 3.b compares the symbol-error rate performance of the MIMO detection techniques with i.i.d Gaussian channel model. For a realistic channel model, the statistical model of a correlated fading channel model[33] for the channel matrix $\mathbf{H}_c$ is represented by $\mathbf{H}_c = \Theta_{BS}^{1/2}\mathbf{A}_{i.i.d}\Theta_{UE}^{1/2}$, where $\mathbf{A}_{i.i.d} \in \mathbb{C}^{B\times U}$ represents i.i.d Rayleigh fading channel, while $\Theta_{BS} \in \mathbb{C}^{B\times B}$ and $\Theta_{UE} \in \mathbb{C}^{U\times U}$ are correlation matrices for BS and UE respectively. Using the correlated fading model, 3GPP SCM channel correlation matrices are generated [34], [35] for BS and UE. An urban scenario with micro cell distribution is assumed for channel matrix generation, where the users are randomly distributed within a cell radius of 500m. The carrier frequency is set to 3.5GHz while the BS antenna elements spacing is half the wave length. Fig. 3.c and 3.d compares symbol-error rate performance of MIMO detection techniques for 3GPP Spatial Channel Model.

Overall, EP algorithm provides the best symbol-error rate performance, while Daisychain SGD requires the highest SNR to converge. EP algorithm provides optimal performance at cost of high interconnect bandwidth at low SNR. Output

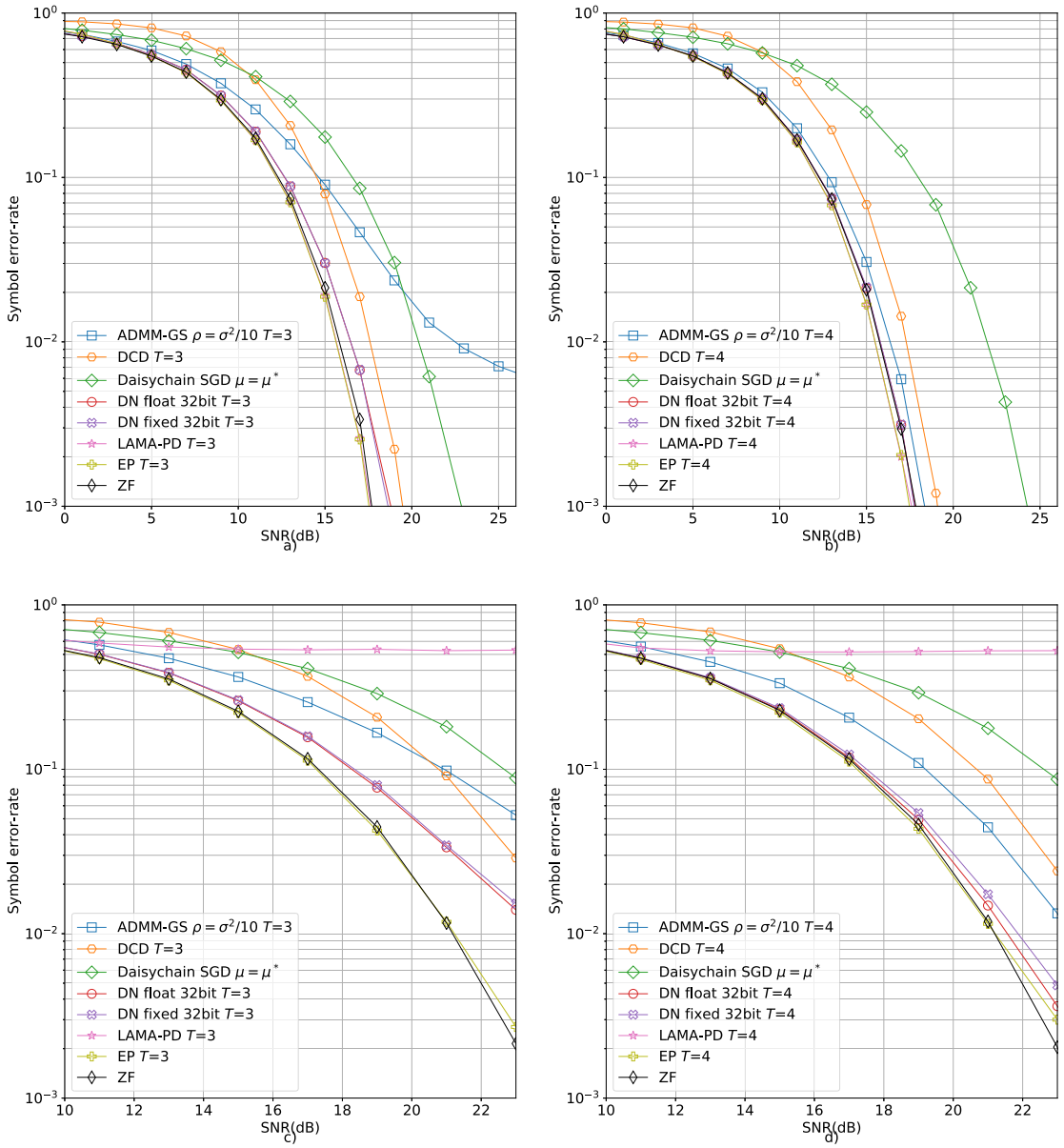

Fig. 3. Performance comparison of MIMO uplink detection algorithms for MIMO system configuration of $B = 128$, $U = 8$, $C = 4$ for 16-QAM modulation in i.i.d Gaussian channel model with 3 iterations (a) and 4 iterations (b) and in 3GPP Spatial Channel Model with 3 iterations (c) and 4 iterations (d).

equalization by EP is followed by soft-output detection [25], involving noise statistics computation. ADMM-GS is a second order algorithm which achieves fast convergence with increase in T. LAMA-PD and LAMA-FD do not converge for realistic 3GPP SCM[22]. DCD has linear computational complexity and achieves slower convergence as compared to ADMM-GS with increase in T. For the proposed DN method, the Hessian is approximated using diagonal dominance characteristics of the matrix $\mathbf{H}_c^H \mathbf{H}_c$ for $c = 1, 2, 3, \ldots C$, which saves interconnect bandwidth and provides close to ZF performance using hard-output detection.

VI. Hardware Implementation Performance Analysis

An FPGA is a reconfigurable computing technology for VLSI implementation, the design flow being different than ASIC. For the ring and star topologies, the XILINX VIRTEX-7 FPGA device is used for VLSI hardware implementation analysis. The fundamental pre-verified resource elements of an FPGA for VLSI implementation are Flip-Flops (FF), Look-up Tables (LUT), Digital Signal Processor slices (DSP48E), Block RAM of 18kB (BRAM_18K). Analysis of system parameters of throughput, resource consumption, latency, and energy efficiency for the ring and star topologies is performed on FPGA. As the ring topology with additional sub-carrier processing demands more FPGA resources, XILINX VIRTEX ULTRASCALE+ FPGA device is used for this analysis. Vivado HLS [36] is a high-level synthesis (HLS) tool used for VLSI hardware prototyping. In the implementation, HLS datatype x_complex [36] is used, which performs arithmetic bit alignment operations implicitly for complex arithmetic operations. Implementing an algorithm on FPGA and

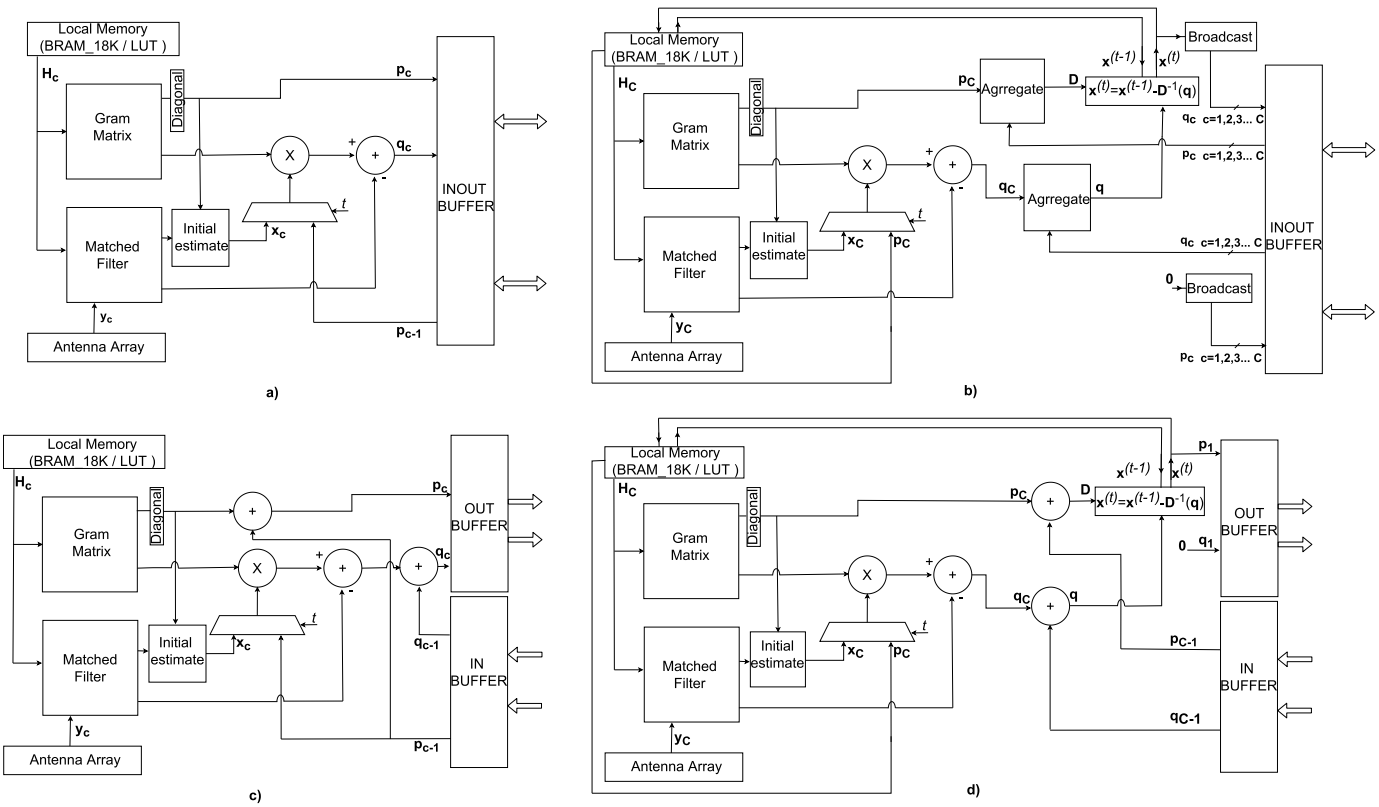

Fig. 4. The architecture diagram of MIMO uplink detection technique using the DN method. The star topology's apex cluster is shown in b) inherits the functionality of the star topology non-apex cluster shown in a). Similarly, the apex cluster for ring topology shown in d) inherits the functionality of the ring topology's non-apex cluster shown in c). All the clusters for a particular topology are implemented in a single FPGA fabric for evaluation.

TABLE III

FPGA HARDWARE RESOURCE ESTIMATES FOR DN **RING** TOPOLOGY, SINGLE SUB-CARRIER WITH 16-QAM AND $B_c = 32$ IMPLEMENTED ON XILINX VIRTEX-7 (XC7VX690T)

	Configuration	$B = 64$, $C = 2$			$B = 128$, $C = 4$		
	Number of users	$U = 2$	$U = 4$	$U = 8$	$U = 2$	$U = 4$	$U = 8$
$T = 3$	DSP48E	800 (22.2 %)	832 (23.1 %)	896 (24.9 %)	1608 (44.7 %)	1680 (46.7 %)	1824 (50.7 %)
	FF	63307 (7.3 %)	83710 (9.7 %)	122207 (14.1 %)	109849 (12.7 %)	134353 (15.5 %)	177299 (20.5 %)
	LUT	35103 (8.1 %)	51880 (12.0 %)	83162 (19.2 %)	55499 (12.8 %)	75120 (17.3 %)	111277 (25.7 %)
	BRAM_18K	4	4	8	8	8	16
	Estimated clock (MHz)	340	340	340	340	340	340
	Latency (clock cycles)	331	405	413	523	661	701
	Maximum Throughput (Mbps)	88	165	279	88	165	279
	Worst-case On-chip power (Watts)	3.573	4.095	5.080	5.342	5.992	7.177
	Power/UE (Watts)	1.786	1.024	0.635	2.671	1.498	0.897
	Mb/Joule	24.63	40.21	54.98	16.74	27.48	38.92
$T = 4$	DSP48E	1072 (29.7 %)	1120 (31.1 %)	1216 (33.7 %)	2152 (59.7 %)	2256 (62.7 %)	2464 (68.4 %)
	FF	84764 (9.8 %)	112320 (13.0 %)	164009 (18.9 %)	146832 (16.9 %)	179963 (20.8 %)	237605 (27.4 %)
	LUT	46371 (10.7 %)	68701 (15.9 %)	110398 (25.5 %)	72553 (16.7 %)	98595 (22.8 %)	146577 (33.9 %)
	BRAM_18K	4	4	8	8	8	16
	Estimated clock (MHz)	340	340	340	340	340	340
	Latency (clock cycles)	427	533	541	683	885	925
	Maximum Throughput (Mbps)	88	165	279	88	165	279
	Worst-case On-chip power (Watts)	4.302	5.004	6.325	6.643	7.519	9.108
	Power/UE (Watts)	2.151	1.251	0.790	3.321	1.880	1.138
	Mb/Joule	20.45	32.91	44.17	13.24	21.90	30.67

optimizing for performance using Vivado HLS is a non-trivial task.

HLS optimizations are applied strategically on specific arithmetic operations in the hardware architecture [37] to achieve trade-off in system latency, throughput and FPGA hardware resources utilization. HLS optimizations are applied on algorithm loop iteration or functional units. Hence, the architecture diagram for a cluster implementation of the ring and star topologies is provided in Fig. 4 for behavioral analysis, where critical arithmetic operations involving loop iterations and functional units are identified. The Gram matrix computation is the key operation with significant

computational complexity. Also, the key interconnect variables are identified between different functional units in the architecture to analyse the dependency on loop iterations.

For every cluster c, local memory registers are synthesized by BRAM_18K or LUT on the FPGA. Every cluster involves Gram Matrix ($\mathbf{H}_c^H\mathbf{H}_c$) and Matched Filter ($\mathbf{H}_c^H\mathbf{y}_c$) computations, which are computed parallelly. $\mathbf{H}_c$ is stored in local memory and utilized for every iteration. The Gram Matrix is computed at every channel coherence time interval, while Matched Filter is computed for every uplink symbol detection. Every cluster caches data from interconnect variables ($\mathbf{p}$ and $\mathbf{q}$ for the ring topology, and $\mathbf{p}_c$ and $\mathbf{q}_c$, where $c = 1, 2, 3 \ldots C$ for the star topology) using local buffers. The buffers are synthesized using BRAM_18K/LUT and operate in FIFO fashion. For the star topology, the input and the output buffers are routed to a single interconnect link, since the data flow in the interconnect variables are bidirectional. For the ring topology, the input and the output buffers are routed to separate input and output interconnect link as the data flow in interconnect variables is unidirectional.

Vivado HLS provides a pragma directive for optimizing hardware implementation [37] on FPGA to achieve a trade-off between system latency and FPGA hardware resource consumption.

A. Hardware Implementation Strategies

In Register Transfer Level (RTL) implementation, a non-apex cluster unit is built as a sub-function. For the first gradient calculation, the matrix multiplier IP core from HLS linear algebra library [36] is optimized for x_complex data-type with a fully unrolled outer row loop using pragma directive HLS UNROLL. In both topologies, every cluster (apex and non-apex) is allocated with matrix multiplier IP core with inline optimization using pragma directive HLS INLINE [37]. Inline optimization reduces processing latency of the matrix multiplier IP core at the expense of an increase in FPGA hardware resource consumption, as it constructs dedicated RTL implementation for every instance of the cluster. Dual port RAM resource implementation is used to store local channel matrix $\mathbf{H}_c$ and local receive signal $\mathbf{y}_c$. Read access to $\mathbf{H}_c^H\mathbf{H}_c$ and $\mathbf{H}_c$ are completely array partitioned in the first dimension using pragma directive HLS ARRAY_PARTITION [37]. Array partition optimization allows parallel access for every row vector of $\mathbf{H}_c^H\mathbf{H}_c$ and $\mathbf{H}_c$, which is unrolled with factor of U using pragma directive HLS UNROLL. Also, every cluster instance is pipelined using pragma directive HLS_PIPELINE [37], which reduces cluster initiation interval [36] and critical path of the cluster. In the star topology, the non-apex cluster unit only transmits $\mathbf{p}_c$ and $\mathbf{q}_c$ to the apex cluster unit. In the ring topology, every cluster unit also consists of an accumulator processing for $\mathbf{p}$ and $\mathbf{q}$ for gradient processing.

The apex cluster is built as a separate sub function and embeds functionality of non-apex cluster to calculate first gradient and the Hessian approximation. Additionally, the apex cluster performs computation of eq. (10) for every iteration t. In the star topology, the apex cluster also accumulates $\mathbf{p}_c$ and $\mathbf{q}_c$ for $c = 1, 2, 3, \ldots C$ available from the non-apex cluster, before evaluating $\mathbf{x}^{(t)}$ for t^{th} iteration. Thus, the apex cluster in the star topology has C links for each $\mathbf{p}_c$ and $\mathbf{q}_c$ where $c = 1, 2, 3, \ldots C$. Complex division for eq. (10) performed at apex cluster C for both topologies is fully unrolled using pragma directive HLS UNROLL with factor of U, which creates dedicated RTL division logic to handle each user computation of elements of $\mathbf{x}$ parallelly. All variables are implemented using dual port RAM optimized by pragma HLS RESOURCE with RAM_2P. RTL logic is realized using Configurable Logic Blocks (CLB) in FPGA [38], [39]. For the current work, the ring and star topologies are implemented on single FPGA fabric, which uses programmable interconnects between Configurable Logic Blocks (CLB) for routing algorithm.

In the ring topology, the top-level HLS synthesis function instantiates apex cluster and non-apex clusters and creates dedicated RTL implementation for every iteration of cluster instantiation. This is achieved by completely unrolling the top-level HLS synthesis function by using pragma directive HLS UNROLL [37]. Variables $\mathbf{p}$ and $\mathbf{q}$ are updated after every cluster processing for every iteration as outputs and become inputs to the next cluster in ring order. This dependence is explicitly enforced on the interconnect variables $\mathbf{p}$ and $\mathbf{q}$ using pragma HLS DEPENDENCE with Read-After-Write (RAW) option, which ensures these variables are read by the next cluster only after the write operation is performed by the current cluster. The sequential nature of ring topology enables complete unrolling of ring topology implementation to process additional sub-carrier in parallel.

In the star topology, non-apex clusters are connected to the apex cluster using dedicated variables $\mathbf{p}_c$ and $\mathbf{q}_c$ where $c = 1, 2, 3, \ldots C$. In the top-level HLS synthesis function, all non-apex clusters are instantiated with dedicated RTL logic using pragma HLS UNROLL. Every non-apex cluster c updates associated $\mathbf{p}_c$ and $\mathbf{q}_c$ parallelly and is conveyed back to apex cluster for computing $\mathbf{p}$ and $\mathbf{q}$ and thereby $\mathbf{x}^{(t)}$ for the t^{th} iteration. The assembly of the non-apex clusters and the apex cluster is unrolled for every iteration using pragma HLS UNROLL directive. Interconnect variables $\mathbf{p}_c$ and $\mathbf{q}_c$ for non-apex clusters are enforced with RAW dependence using pragma HLS DEPENDENCE for subsequent iterations.

After running the behavioral simulation for the ring and the star topologies, the respective architecture is synthesized taking account of the HLS optimizations in the RTL. After resolving critical paths, the cluster computing is time scheduled as given in Fig. 5 for MIMO configuration of $U = 8,\ B = 128,\ C = 4$ and $t = 3$. For ring topology time scheduling as shown in Fig. 5.a, owing to inter-dependency among the interconnect variables, the clusters are scheduled sequentially for every iteration $t = 1, 2, 3 \ldots T$. For the star topology, the time schedule as shown in Fig. 5.b, the non-apex clusters are scheduled parallelly since every non-apex cluster has dedicated interconnect variables with no inter-dependency. After every time interval of the non-apex cluster computation, the apex cluster computes $\mathbf{x}^{(t)}$ for every iteration t. The channel matrix $\mathbf{H}_c$ is accessed from the local memory and $\mathbf{y}_c$ is accessed from the RF frontend at initial iteration

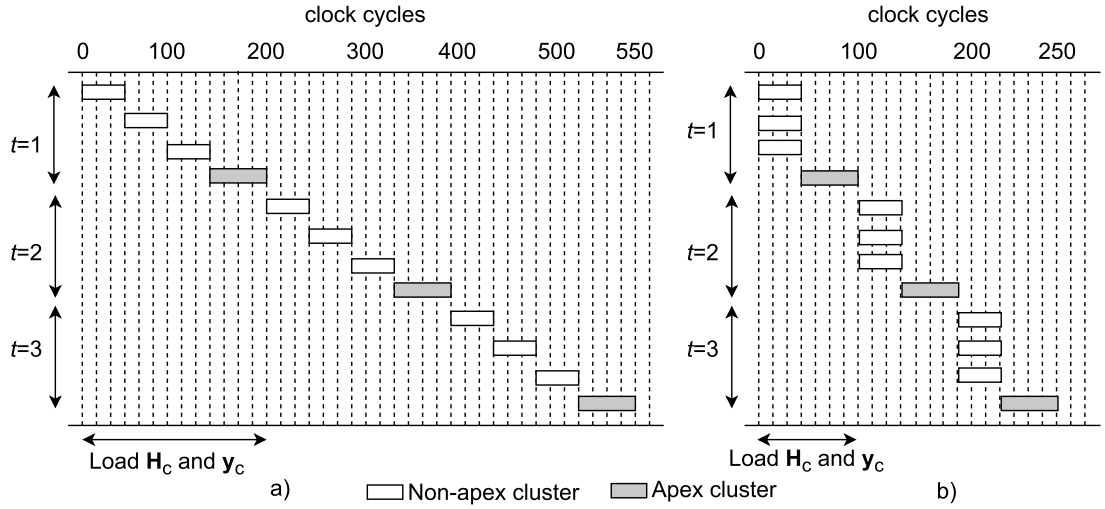

Fig. 5. The scheduling diagram from Vivado HLS Schedule viewer for the ring topology as shown in a) and the star topology as shown in b) for MIMO system configuration of $B = 128$, $U = 2$, $C = 4$, $B_c = 32$ and $T = 3$. For both topologies, $\mathbf{H}_c$ and $\mathbf{y}_c$ are accessed by each cluster during the initial iteration. The clusters are unrolled using pragma HLS UNROLL for each iteration t.

$t = 0$. After time scheduling, the maximum clock frequency for the architecture is estimated by the static timing analysis.

B. Hardware Implementation Performance Analysis

In hardware analysis, evaluation of the system parameters of latency, throughput, and on-chip power consumption for various MIMO system configurations for star and ring topologies is done. Specifically, a comparison between MIMO BS with $B = 64$ and $B = 128$ is drawn, which provides insights into change in system parameters with a change in clusters. Within particular MIMO BS, insights are provided into MIMO system configuration with UEs as $U = 2, 4, 6$ and 8 to evaluate the change in system parameters with the change in the number of UEs serviced by the system. Fig. 5 describes cluster scheduling for the star and ring topology for specific MIMO system configuration. Latency in terms of the clock cycles changes with MIMO system configuration, however, the characteristic scheduling order remains constant. For power profiling, Xilinx Power Estimator [41] is used to estimate the worst-case on-chip power consumption for different MIMO configurations presented here. For profiling, ambient temperature (25°C) and 250 Linear Feet per Minute (LFM) air supply with heat sinking as environment variables are configured. Clock toggle rate of 12.5% and enable rate of 50% is used for clock simulation. Hardware implementation for ring topology and star topology presented in Table. III and Table. IV, respectively.

1) Ring Topology Resource Analysis: Table. III gives comparative analysis for the ring topology for the system parameters. When the system with $B = 64$ ($T = 3$) is scaled from $U = 2$ configuration to support $U = 8$ configuration, overall system throughput increases by 2.2×, however throughput per UE suffers a decrease of 21%. On scaling from $B = 64$ to $B = 128$ ($T = 3$), throughput variation remain similar for system to that of $B = 64$ ($T = 3$) from $U = 2$ to $U = 8$. When star topology with $B = 64$ ($T = 3$) is scaled from $U = 2$ configuration to support $U = 8$ configuration, overall system throughput increases by 2.4×, however throughput per UE suffers a decrease of 15.4%. On scaling from $B = 64$ to $B = 128$ ($T = 3$), throughput variation remain similar to that of $B = 64$ ($T = 3$) from $U = 2$ to $U = 8$. There is no throughput variation for both topologies on increasing the number of iteration.

For the ring topology, by scaling the BS station to support additional UEs, FPGA hardware resources per UE increase fractionally. When the system with $B = 64$ ($T = 3$) is scaled from $U = 2$ configuration to support $U = 8$ configuration, there is drastic change in FPGA resource consumption per UE from $U = 2$ to $U = 8$ as 72% decrease in DSP48E, 51.7% decrease in FF, 40.7% decrease in LUT but 50% increase in BRAM is observed. On comparing FPGA resource consumption per UE for specific U, DSP48E and BRAM usage gets doubled for $B = 128$ than that for $B = 64$. However, for FF and LUT consumption per UE for specific U, a higher number of UE requires less increase in FF and LUT as compared to the lower number of UE, when the system is scaled from $B = 64$ to $B = 128$. For example, from $B = 64$ to $B = 128$, FF and LUT increases by 45.4% and 33.8% respectively for $U = 8$ as compared to 73.9% and 58% increase respectively for $U = 2$. For $B = 64$ and $B = 128$, adding an iteration with the same throughput increases DSP48E, FF, and LUT by an average of 33% with no additional BRAM requirement, for all UE cases. When system is scaled from $B = 64$ to $B = 128$ with $T = 3$ for specific number of UE, latency increase is more for higher number of UE (58% for $U = 2$ as compared to 69.7% for $U = 8$). For $B = 64$ ($T = 3$), addition of 2 UE to $U = 2$ increases latency by 22.3% as compared to addition of 4 UE to $U = 4$ with just 2% increase. Whereas for $B = 128$, addition of 2 UE to $U = 2$ and 4 UE to $U = 4$ costs 26.3% and 6% increase in latency respectively. Thus, as the number of UE increase for a particular BS, additional UE can be added at a lower increase in latency. Implementing additional iteration causes an 30% average increase in latency across $B = 64$ and $B = 128$ for all UE cases.

2) Star Topology Resource Analysis: Star topology comparative analysis is presented in Table. IV. By scaling BS

TABLE IV

FPGA HARDWARE RESOURCE ESTIMATES FOR DN **STAR** TOPOLOGY, SINGLE SUB-CARRIER WITH 16-QAM AND $B_c = 32$ IMPLEMENTED ON XILINX VIRTEX-7 (XC7VX690T)

	Configuration	$B = 64$, $C = 2$			$B = 128$, $C = 4$		
	Number of users	$U = 2$	$U = 4$	$U = 8$	$U = 2$	$U = 4$	$U = 8$
$T = 3$	DSP48E	816 (22.7 %)	864 (24.0 %)	960 (26.7 %)	1632 (45.3 %)	1728 (48.0 %)	1920 (53.3 %)
	FF	60335 (7.0 %)	70488 (8.1 %)	86581 (10.0 %)	115374 (13.3 %)	130215 (15.0 %)	150236 (17.3 %)
	LUT	27713 (6.4 %)	34072 (7.9 %)	46227 (10.7 %)	50747 (11.7 %)	60054 (13.9 %)	78607 (18.1 %)
	BRAM_18K	0	0	0	0	0	0
	Estimated clock (MHz)	340	340	340	379	379	379
	Latency (clock cycles)	242	256	304	248	262	310
	Maximum Throughput (Mbps)	70	130	237	74	138	253
	Worst-case On-chip power (Watts)	3.462	3.736	4.208	5.859	6.329	7.097
	Power/UE (Watts)	1.731	0.934	0.526	2.930	1.582	0.876
	Mb/Joule	20.22	34.62	56.24	12.63	21.81	35.61
$T = 4$	DSP48E	1088 (30.2 %)	1152 (32.0 %)	1280 (35.6 %)	2176 (60.4 %)	2304 (64.0 %)	2560 (71.1 %)
	FF	78730 (9.0 %)	92031 (10.6 %)	113023 (13.0 %)	150494 (17.4 %)	169875 (19.6 %)	195763 (22.6 %)
	LUT	35683 (8.2 %)	43801 (10.1 %)	59445 (13.7 %)	65263 (15.0 %)	77061 (17.8 %)	100733 (23.2 %)
	BRAM_18K	0	0	0	0	0	0
	Estimated clock (MHz)	340	340	340	379	379	379
	Latency (clock cycles)	313	330	398	321	338	406
	Maximum Throughput (Mbps)	70	130	237	74	138	253
	Worst-case On-chip power (Watts)	4.111	4.466	5.081	7.248	7.858	8.855
	Power/UE (Watts)	2.056	1.116	0.635	3.624	1.964	1.107
	Mb/Joule	17.03	29.11	46.64	10.21	17.56	28.57

station to support additional UE, FPGA hardware resources per UE increase fractionally. On the contrary, there is a drastic change in FPGA resource consumption per UE from $U = 2$ to $U = 8$ as a 70.6% decrease in DSP48E, a 61.54% decrease in FF, 58.20% decrease in LUT is observed. FPGA resource consumption per UE for specific U, DSP48E usage gets doubled for $B = 128$ than that for $B = 64$. However, for FF and LUT consumption per UE for specific U, system servicing a higher number of UE requires less increase in FF and LUT as compared to system servicing a lower number of UE, when two clusters are added to the system with $B = 64$. For example, from $B = 64$ to $B = 128$, FF and LUT increases by 73% and 69.15% respectively for $U = 8$ as compared to 90% and 82.8% increase respectively for $U = 2$. For $B = 64$ and $B = 128$, adding an iteration with same throughput increases DSP48E, FF and LUT by an average of 33%, 30% and 28% respectively, for $U = 2, 4$ and 8. When system is scaled from $B = 64$ to $B = 128$ with three iteration, latency increase is constant at 6 clock cycles for all UE cases. Relatively, addition of two clusters to $B = 64$ costs 2.5%, 2.34% and 2.0% increase in latency for $U = 2, 4$ and 8 respectively. Implementing additional iteration causes an average of 30.5% increase in latency across $B = 64$ and $B = 128$ for all UE cases.

3) Power Consumption Analysis: For ring topology, although total power consumption increases with the number of UE, the power consumed per UE decreases, and bits per Joule increase with the number of UE for a particular BS. For instance, from scaling $U = 2$ to $U = 8$ ($T = 3$), power consumption per UE drops by 64% for $B = 64$ and 66.4% for $B = 128$, whereas bits per Joule increase by 1.23× for $B = 64$ and 1.32× for $B = 128$, making it more power-efficient. The addition of cluster increases power consumption fractionally. By scaling from $B = 64$ to $B = 128$ ($T = 3$), power consumption per UE increases by 49.5%, 46.3% and 41.2% for $U = 2, 4$ and 8 respectively, which causes drop in bits per Joule by 33.1%, 31.6% and 29.21% for $U = 2, 4$ and 8 respectively. Addition of iteration to three iteration system increases power consumption per UE by an average of 22.4% for $B = 64$ and 25.6% for $B = 128$, whereas decreases bits per Joule by an average of 18.2% for $B = 64$ and 20.36% for $B = 128$ for all UE cases. Ring topology with a single sub-carrier can process additional sub-carrier at approximately 30% reduced latency as shown in Table. V. While processing the second sub-carrier, FPGA resource consumption for DSP48E, FF, and BRAM has almost doubled while LUT consumption increases by 53%. Throughput is increased by 1.39× for two sub-carrier as compared to single sub-carrier processing, with a 61% increase in power consumption per UE and a 53% increase in bits per Joule.

TABLE V

FPGA HARDWARE RESOURCE ESTIMATES FOR DN RING TOPOLOGY, MULTIPLE SUB-CARRIERS WITH 16-QAM, $B_c = 32$, $B = 128$, $U = 8$ AND $C = 4$ ON XILINX VIRTEX ULTRASCALE+ (XCVU13P)

Number of sub-carriers	1	2
DSP48E	1824 (14.8 %)	3840 (31.2 %)
FF	175351 (5.0 %)	367241 (10.6 %)
LUT	108709 (6.3 %)	218087 (12.6 %)
BRAM_18K	16	32
Iterations (I)	3	3
Estimated clock (MHz)	383	381
Latency (clock cycles)	707	492
Maximum Throughput (Mbps)	292	697
Worst On-chip power (Watts)	9.8	15.3
Power/UE (Watts)	1.225	1.9125
Mb/Joule	29.80	45.55

TABLE VI

COMPARISON OF **DECENTRALIZED** MASSIVE MIMO UPLINK DETECTION TECHNIQUES

Detection method	LAMA-FD [19]	LAMA-PD [19]	DCD [15]	FD-LAMA[1][40]	DN ring topology[1]	DN star topology[1]
Fabric	GPU	GPU	GPU	FPGA	FPGA	FPGA
Precision point	float[2](n.a.)	float[2](n.a.)	float (32bit)	fixed[3](32bit)	fixed (32bit)	fixed (32bit)
Parallel sub-carrier instances	-	-	-	n.a.	2	1
Modulation scheme	16-QAM	16-QAM	16-QAM	QPSK	16-QAM	16-QAM
Configuration ($B \times U$)	128 × 8	128 × 8	128 × 8	128 × 8	128 × 8	128 × 8
Iterations (I)	3	3	3	3	3	3
FF	-	-	-	76270 (2.2 %)	367241 (10.6 %)	150236 (4.3 %)
LUT	-	-	-	44420 (2.6 %)	218087 (12.6 %)	78607 (4.5 %)
DSP48E	-	-	-	1197 (9.7 %)	3840 (31.2 %)	1920 (15.6 %)
BRAM_18K	-	-	-	10	32	0
Latency (ms)	0.854	0.929	0.540	0.900	1.291	0.818
Maximum Throughput (Gbps)	1.34	1.23	1.06	0.018	0.697	0.253
Power Consumption (Watts)	1200	1200	1200	n.a.	15.3	7.097
Mb/Joule	1.12	0.94	0.88	n.a.	45.36	35.61

[1] FF, LUT, DSP48E and BRAM_18K percent estimation is normalized for XILINX VIRTEX ULTRASCALE+ (XCVU13P) FPGA device.
[2] NVIDIA cuBLAS library .
[3] Complex number representation.

For the star topology, although power consumption increases with the number of UEs, the power consumed per UE decreases, and bits per Joule increase with the number of UE for particular BS similar to the ring topology. For instance, from scaling $U = 2$ to $U = 8$ ($T = 3$), power consumption per UE drops by approximately 70% for $B = 64$ and $B = 128$, whereas bits per Joule increase by approximately $1.8\times$ for $B = 64$ and $B = 128$. The addition of clusters increases power consumption and causes a decrease in bits per Joule. By scaling from $B = 64$ to $B = 128$ for $T = 3$, power consumption per UE increases by 69%, 69% and 66.5% for $U = 2, 4$ and 8 respectively, which causes drop in bits per Joule by 37.52%, 37.00% and 36% for $U = 2, 4$ and 8 respectively. Addition of iteration to three iteration system increases power consumption per UE by an average of 19.6% for $B = 64$ and 24.6% for $B = 128$, whereas decreases bits per Joule by an average of 16.4% for $B = 64$ and 19.4% for $B = 128$.

4) Comparative Performance Analysis: On comparing the ring and star topologies, the choice of topology is dependent on the trade-off among interconnect bandwidth, throughput, energy efficiency and latency. On comparing similar MIMO configurations between the ring and star topologies, ring topology provides more throughput at expense of increased latency as compared to the star topology for $B = 64$ and $B = 128$ for all UE cases. Ring topology maintains constant interconnect bandwidth on the addition of clusters, with no RTL reconfiguration required for the apex cluster or non-apex clusters on scaling MIMO configuration for B. For similar MIMO configurations, star topology provides low latency as compared to the ring topology at the expense of reduced throughput. Also, the star topology provides a more deterministic latency increase by scaling MIMO configuration by the number of UE or clusters as compared to a ring topology. On the contrary, the ring topology processes additional sub-carrier (Table. V) at a fractional increase in latency and power consumption per UE as compared to the star topology, providing high throughput gain at the expense of twice the FPGA resource consumption.

Table. VI compares decentralized MIMO detection techniques implemented as FPGA and GPU prototypes. FD-LAMA [40] is a variant of LAMA-FD, which implements a hyperbolic tangent function for LAMA iterations, thereby increasing algorithm computational complexity. LAMA algorithm is not robust for a realistic channel environment[22]. On the application note, the star topology is favorable for a 3GPP SCM scenario that needs to be scaled with a large number of clusters at low latency at expense of high interconnect bandwidth. On the contrary, the ring topology is favorable for a 3GPP SCM scenario that requires scaling for a large number of clusters at constant interconnect bandwidth and high throughput, at expense of increased latency.

VII. CONCLUSION AND FUTURE WORK

In this work, decentralized Newton (DN) algorithm for decentralized MIMO uplink is presented, which is a novel adaptation of the centralized Newton method. Also, two novel hardware architectures for hardware implementation are proposed. Also, a comparative analysis of scaling effects on parameters of throughput, latency, FPGA resource consumption, and on-chip power consumption for both topologies is carried out. The star topology is suited for MIMO configuration scenarios that demand low-latency, while ring topology can be implemented in MIMO configuration scenarios demanding higher throughput and lower interconnect bandwidth. Interestingly, it is possible to switch between topologies using smart routing hardware to route the resource blocks allocated to enhanced mobile broadband (eMMB) services and those dedicated to Ultra-High Reliability and Low Latency (URLLC) services accordingly. In terms of scaling system for thousand of BS antennas, the ring topology maintains low and constant interconnect bandwidth as compared to star topology. Also, the ring topology can process additional sub-carrier at a fractional increase in the latency and power consumption. The star topology can be scaled for a huge number of clusters without incurring high latency. DN is a comparatively low complexity algorithm providing close to ZF performance which can be implemented feasibly on FPGA. FPGA is

inherently power efficient as compared to GPU, which makes FPGA implementation of DN algorithm power-efficient than GPU implementation of other decentralized MIMO uplink detection algorithms [15], [19].

In the future work, there is scope to implement and analyze the ring and star topologies of DN method based MIMO uplink detection using multiple FPGA, which would make the decentralized implementation more modular. Although clusters with an equal number of antennas per cluster are considered, both topologies can be implemented with non-uniform distribution of BS antennas among clusters.

APPENDIX

A. Proof of Lemma 1

$f_c(\mathbf{x})$ with $f_c : \mathbb{C}^{U\times 1} \rightarrow \mathbb{R}$ is considered as local cost function of cluster c and define it as:

$$f_c(\mathbf{x}) = ||\mathbf{H}_c\mathbf{x} - \mathbf{y}_c||_2^2 = \mathbf{x}^H\mathbf{H}_c^H\mathbf{H}_c\mathbf{x} - 2\mathbf{y}_c^H\mathbf{H}_c\mathbf{x} + \mathbf{y}_c^H\mathbf{y}_c \quad (2)$$

With ensemble of sample function $f_c(\mathbf{x})$ with respect to c, $F(\mathbf{x})$ is constructed as system objective function given as $F : \mathbb{C}^{U\times 1} \rightarrow \mathbb{R} >= 0$ for robust stochastic optimization [29], defined such that:

$$F(\mathbf{x}) = \mathbb{E}\left(f_c(\mathbf{x})\right) = \frac{1}{C}\sum_{c=1}^{C} f_c(\mathbf{x}) \quad c = 1, 2, 3 \ldots C \quad (3)$$

By adapting Newton Method [27] to evaluate $\mathbf{x}^{(t)}$, where iteration $t = 1, 2, 3 \ldots T$:

$$\mathbf{x}^{(t)} = \mathbf{x}^{(t-1)} - \left(\nabla^2_{\mathbf{x}^{(t-1)}} F(\mathbf{x}^{(t-1)})\right)^{-1} \nabla_{\mathbf{x}^{(t-1)}} F(\mathbf{x}^{(t-1)}) \quad (4)$$

where,

$$\begin{aligned}\nabla_{\mathbf{x}^{(t-1)}} F(\mathbf{x}^{(t-1)}) &= \nabla_{\mathbf{x}^{(t-1)}} \mathbb{E}\left(f_c(\mathbf{x}^{(t-1)})\right) \\ &= \frac{1}{C}\sum_{c=1}^{C}\left(\nabla_{\mathbf{x}^{(t-1)}} f_c(\mathbf{x}^{(t-1)})\right) \\ &= \frac{2}{C}\sum_{c=1}^{C}\left(\mathbf{H}_c^H\mathbf{H}_c\mathbf{x}^{(t-1)} - \mathbf{H}_c^H\mathbf{y}_c\right) \quad (5)\end{aligned}$$

and,

$$\begin{aligned}\nabla^2_{\mathbf{x}^{(t-1)}} F(\mathbf{x}^{(t-1)}) &= \nabla^2_{\mathbf{x}^{(t-1)}} \mathbb{E}\left(f_c(\mathbf{x}^{(t-1)})\right) \\ &= \frac{1}{C}\sum_{c=1}^{C}\left(\nabla^2_{\mathbf{x}^{(t-1)}} f_c(\mathbf{x}^{(t-1)})\right) \\ &= \frac{2}{C}\sum_{c=1}^{C}\left(\mathbf{H}_c^H\mathbf{H}_c\right) \quad (6)\end{aligned}$$

$\mathbf{H}_c^H\mathbf{H}_c$ is symmetrical positive-semidefinite and is decomposed arithmetically as:

$$\mathbf{H}_c^H\mathbf{H}_c = \mathbf{D}_c + \mathbf{L}_c + \mathbf{L}_c^H \quad (7)$$

where $\mathbf{D}_c$, $\mathbf{L}_c$ and $\mathbf{L}_c^H$ are diagonal, strictly lower triangular and strictly upper triangular matrices. As $\mathbf{H}_c^H\mathbf{H}_c$ is a $U \times U$ matrix, it needs $U \times U$ dimensional interconnect between the clusters. But, as $\mathbf{H}_c^H\mathbf{H}_c$ is diagonally dominant and its column vectors being mutually orthogonal, it can be approximated as $\mathbf{H}_c^H\mathbf{H}_c \approx \mathbf{D}_c$ [25], [30]. With this approximation, column vector comprising diagonal of $\mathbf{D}_c$ can be exchanged and accumulated between clusters as $U \times 1$ dimensional column vector, thus reducing interconnect bandwidth between clusters. Accordingly, column vectors of $\mathbf{H}_c$ are used to calculate $\mathbf{D}_c$.

$$(\mathbf{D}_c)_{ij} = \begin{cases} ||\mathbf{h}_{u,c}||_2^2 & \text{when } i = j = u \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

With approximation of $\mathbf{D}_c$, second gradient is calculated as:

$$\nabla^2_{\mathbf{x}^{(t-1)}} F(\mathbf{x}^{(t-1)}) = \frac{2}{C}\sum_{c=1}^{C}\left(\mathbf{H}_c^H\mathbf{H}_c\right) \approx \frac{2}{C}\sum_{c=1}^{C}(\mathbf{D}_c) \triangleq \frac{2}{C}\mathbf{D} \quad (9)$$

While using eq. (5) and eq. (9) for evaluating eq. (4), factor $\frac{2}{C}$ gets canceled:

$$\mathbf{x}^{(t)} = \mathbf{x}^{(t-1)} - (\mathbf{D})^{-1}\left(\sum_{c=1}^{C}(\mathbf{H}_c^H\mathbf{H}_c\mathbf{x}^{(t-1)} - \mathbf{H}_c^H\mathbf{y}_c)\right) \quad (10)$$

REFERENCES

[1] M. A. Ouameur, D. Massicotte, A. M. Akhtar, and R. Girard, "Performance evaluation and implementation complexity analysis framework for ZF based linear massive MIMO detection," *Wireless Netw.*, vol. 26, pp. 1–15, Apr. 2020.

[2] M. A. Albreem, M. Juntti, and S. Shahabuddin, "Massive MIMO detection techniques: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3109–3132, 4th Quart., 2019.

[3] E. Björnson, L. Sanguinetti, H. Wymeersch, J. Hoydis, and T. L. Marzetta, "Massive MIMO is a reality—What is next?: Five promising research directions for antenna arrays," *Digit. Signal Process.*, vol. 94, pp. 3–20, Nov. 2019.

[4] L. V. der Perre, L. Liu, and E. G. Larsson, "Efficient DSP and circuit architectures for massive MIMO: State of the art and future directions," *IEEE Trans. Signal Process.*, vol. 66, no. 18, pp. 4717–4736, Sep. 2018.

[5] M. Wu, B. Yin, G. Wang, C. Dick, J. R. Cavallaro, and C. Studer, "Large-scale MIMO detection for 3GPP LTE: Algorithms and FPGA implementations," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 916–929, Oct. 2014.

[6] B. Yin, M. Wu, J. R. Cavallaro, and C. Studer, "VLSI design of large-scale soft-output MIMO detection using conjugate gradients," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Lisbon, Portugal, May 2015, pp. 1498–1501.

[7] J. Chen, Z. Zhang, H. Lu, J. Hu, and G. E. Sobelman, "An intra-iterative interference cancellation detector for large-scale MIMO communications based on convex optimization," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 63, no. 11, pp. 2062–2072, Nov. 2016.

[8] M. Wu, C. Dick, J. R. Cavallaro, and C. Studer, "High-throughput data detection for massive MU-MIMO-OFDM using coordinate descent," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 63, no. 12, pp. 2357–2367, Dec. 2016.

[9] Z. Wu, C. Zhang, Y. Xue, S. Xu, and X. You, "Efficient architecture for soft-output massive MIMO detection with Gauss–Seidel method," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2016, pp. 1886–1889.

[10] C. Zhang, Z. Wu, C. Studer, Z. Zhang, and X. You, "Efficient soft-output Gauss–Seidel data detector for massive MIMO systems," *IEEE Trans. Circuits Syst. I, Reg. Papers*, early access, Oct. 26, 2019, doi: 10.1109/TCSI.2018.2875741.

[11] G. Peng, L. Liu, S. Zhou, S. Yin, and S. Wei, "A 1.58 Gbps/W 0.40 Gbps/mm^2 ASIC implementation of MMSE detection for 128×8 64-QAM massive MIMO in 65 nm CMOS," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 65, no. 5, pp. 1717–1730, May 2018.

[12] A. Yu *et al.*, "Efficient successive over relaxation detectors for massive MIMO," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 6, pp. 2128–2139, Jun. 2020.

[13] C. Jeon, O. Castaneda, and C. Studer, "A 354 Mb/s 0.37 mm^2 151 mW 32-user 256-QAM near-MAP soft-input soft-output massive MU-MIMO data detector in 28 nm CMOS," in *Proc. IEEE 45th Eur. Solid State Circuits Conf. (ESSCIRC)*, Sep. 2019, pp. 127–130.

[14] K. Li *et al.*, "Design trade-offs for decentralized baseband processing in massive MU-MIMO systems," in *Proc. 53rd Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Nov. 2019, pp. 906–912.

[15] K. Li, O. Castaneda, C. Jeon, J. R. Cavallaro, and C. Studer, "Decentralized coordinate-descent data detection and precoding for massive MU-MIMO," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2019, pp. 1–5.
[16] K. Li, R. R. Sharan, Y. Chen, T. Goldstein, J. R. Cavallaro, and C. Studer, "Decentralized baseband processing for massive MU-MIMO systems," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 7, no. 4, pp. 491–507, Dec. 2017.
[17] M. A. Ouameur and D. Massicotte, "Efficient distributed processing for large scale MIMO detection," in *Proc. 27th Eur. Signal Process. Conf. (EUSIPCO)*, A Coruna, Spain, Sep. 2019, pp. 1–5.
[18] J. R. Sánchez, F. Rusek, O. Edfors, M. Sarajlić, and L. Liu, "Decentralized massive MIMO processing exploring daisy-chain architecture and recursive algorithms," *IEEE Trans. Signal Process.*, vol. 68, pp. 687–700, Jan. 2020.
[19] C. Jeon, K. Li, J. R. Cavallaro, and C. Studer, "Decentralized equalization with feedforward architectures for massive MU-MIMO," *IEEE Trans. Signal Process.*, vol. 67, no. 17, pp. 4418–4432, Sep. 2019.
[20] H. Wang, A. Kosasih, C.-K. Wen, S. Jin, and W. Hardjawana, "Expectation propagation detector for extra-large scale massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2036–2051, Mar. 2020.
[21] Z. Zhang, H. Li, Y. Dong, X. Wang, and X. Dai, "Decentralized signal detection via expectation propagation algorithm for uplink massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 11233–11240, Oct. 2020.
[22] C. Jeon, R. Ghods, A. Maleki, and C. Studer, "Optimal data detection in large MIMO," Nov. 2018, *arXiv:1811.01917*. [Online]. Available: http://arxiv.org/abs/1811.01917
[23] P. Seidel, S. Paul, and J. Rust, "Decentralized massive MIMO uplink signal estimation by binary multistep synthesis," in *Proc. 53rd Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Nov. 2019, pp. 1967–1971.
[24] E. Bertilsson, O. Gustafsson, and E. G. Larsson, "A scalable architecture for massive MIMO base stations using distributed processing," in *Proc. 50th Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, Nov. 2016, pp. 864–868.
[25] S. Yang and L. Hanzo, "Fifty years of MIMO detection: The road to large-scale MIMOs," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 1941–1988, 4th Quart., 2015.
[26] N. Rajatheva *et al.*, "White paper on broadband connectivity in 6G," Apr. 2020, *arXiv:2004.14247*. [Online]. Available: http://arxiv.org/abs/2004.14247
[27] J. Nocedal and S. Wright, *Numerical Optimization* (Springer Series in Operations Research and Financial Engineering). New York, NY, USA: Springer, 2006.
[28] S. P. Karimireddy, S. U. Stich, and M. Jaggi, "Global linear convergence of Newton's method without strong-convexity or Lipschitz gradients," Jun. 2018, *arXiv:1806.00413*. [Online]. Available: http://arxiv.org/abs/1806.00413
[29] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[30] T. L. Marzetta and E. G. Larsson, *Fundamentals of Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.
[31] T. E. Oliphant, *Guide to NumPy*, 2nd ed. North Charleston, SC, USA: CreateSpace Independent Publishing Platform, 2015.
[32] F. Johansson *et al.* (Dec. 2013). *Mpmath: A Python Library for Arbitrary precision Floating-Point Arithmetic (Version 0.18)*. [Online]. Available: http://mpmath.org/
[33] Y. S. Cho, J. Kim, W. Y. Yang, and C. G. Kang, "MIMO channel models," in *MIMO-OFDM Wireless Communications With MATLAB*. Singapore: Wiley, 2010, pp. 71–109.
[34] J. Salo *et al.*, *MATLAB Implementation of the 3GPP Spatial Channel Model*, document TR 25.996, 3GPP, Jan. 2005. [Online]. Available:http://www.tkk.fi/Units/Radio/scm/
[35] *Spatial Channel Model for Multiple Input Multiple Output (MIMO) Simulations*, document TR 25.996, 3GPP, Release 16, Jul. 2020.
[36] *Vivado Design Suite User Guide High-Level Synthesis UG902 (V2019.1)*, Xilinx, San Jose, CA, USA, Jul. 2019.
[37] *SDx Pragma Reference Guide UG1253 (V2019.1)*, Xilinx, San Jose, CA, USA, Jun. 2019.
[38] *Xilinx 7 Series FPGAs Configurable Logic Block User Guide UG474 (V1.8)*, Xilinx, San Jose, CA, USA, Sep. 2016.
[39] *Xilinx 7 Series FPGAs Data Sheet: Overview DS180 (V2.6.1)*, Xilinx, San Jose, CA, USA, Sep. 2020.
[40] K. Li, C. Jeon, J. R. Cavallaro, and C. Studer, "Decentralized equalization for massive MU-MIMO on FPGA," in *Proc. 51st Asilomar Conf. Signals, Syst., Comput.*, Oct. 2017, pp. 1532–1536.
[41] *Xilinx Power Estimator User Guide UG440 (V2019.2)*, Xilinx, San Jose, CA, USA, Oct. 2019.

Abhinav Kulkarni received the bachelor's degree in electronics and communication engineering from VNIT, India, in 2016, and the master's degree in embedded systems from Nanyang Technological University, Singapore, in 2017. He is currently pursuing the Ph.D. degree in electrical engineering with the Université du Québec à Trois-Rivières (UQTR), QC, Canada. From 2017 to 2019, he was with Addvalue Innovation Private Ltd., Singapore, as a Satellite Communication Engineer. He worked on Linux-based software defined radio (SDR) platform development, where he was involved with board bring-up, FPGA prototyping and troubleshooting. His varied current research interests are baseband signal processing, approximate circuits, cyber-physical systems, computer vision, and machine learning.

Messaoud Ahmed Ouameur (Member, IEEE) received the bachelor's degree in electrical engineering from the Institute national d'électronique et d'électricité (INELEC), Boumerdes, Algeria, in 1998, the M.B.A. degree from the Graduate School of International Studies, Ajou University, Suwan, South Korea, in 2000, and the master's and Ph.D. degrees (Hons.) in electrical engineering from the Université du Québec à Trois-Rivières (UQTR), QC, Canada, in 2002 and 2006, respectively. From 2001 to 2006, he worked with Axiocom Inc., as the Director of research and development, where his research activities include wireless communications, spread-spectrum systems, iterative (turbo) detection, channel estimation, smart antennas, Monte Carlo techniques for signal processing, and real-time very-large-scale integration (VLSI). He joined Nutaq Innovation in November 2006. As a radio system technical leader, his tasks involved radio system design and performance analysis of wireless communication systems including GSM, WCDMA, LTE, and 5G, embedded signal processing algorithm design and implementation, and radio transceiver prototyping from the antenna to baseband (PHY) processing. He then joined UQTR as a Regular Professor in 2018. His research interests include the field of embedded real-time systems, parallel and distributed processing with applications to distributed Massive MIMO, deep learning and machine learning for communication system design, and the Internet of Things with the emphasis on end-to-end systems prototyping and edge computing.

Daniel Massicotte (Senior Member, IEEE) received the B.Sc.A. and M.Sc.A. degrees in electrical engineering and industrial electronics from the Université du Québec à Trois-Rivières (UQTR), QC, Canada, in 1987 and 1990, respectively, and the Ph.D. degree in electrical engineering from the École Polytechnique de Montréal, QC, Canada, in 1995. In 1994, he joined the Department of Electrical and Computer Engineering, Université du Québec à Trois-Rivières, where he is currently a Full Professor. He has been the Founder and the Head of the Laboratory of Signal and Systems Integration since 1998. Since 2001, he has been the Founding President and the Chief Technology Officer of Axiocom Inc. He was the Head of the Industrial Electronic Research Group from 2011 to 2018, the Head of the Department of Electrical and Computer Engineering from 2014 to 2020, and has been the Head of the Research Chair in Signals and Intelligence of High-Performance Systems since 2018. He has proposed many methods based on modern signal and biosignal processing, such as machine learning, transform domain, and metaheuristics. He has authored/coauthored more than 200 technical papers in international conferences and journals, and 9 inventions. His research interests include advanced VLSI implementation, digital signal processing for wireless communications, measurement, and medical and control problems for linear/nonlinear complex systems. He is a member of the "Ordre des Ingénieurs du Québec," "Groupe de Recherche en Électronique Industrielle (GREI)," and "Microsystems Strategic Alliance of Quebec (ReSMiQ)." He received the Douglas R. Colton Medal for research excellence awarded by Canadian Microelectronics Corporation, the PMC-Sierra High Speed Networking and Communication Award, and the Second place at the Complex Multimedia/Telecom IP Design Contest from Europractice. He was the General Chair of IEEE NEWCAS 2014 and a Guest Editor of *Analog Integrated Circuits and Signal Processing* (Springer) for the Special Issues of NEWCAS 2013.

Chapter 4 - Heuristic methodology for FPGA based signed approximate multiplication circuits.

4.1 Résumé Long

4.1.1 Contexte de la Recherche

Les opérations de multiplication jouent un rôle crucial dans la détection des signaux sans fil, influençant la capacité de traitement et l'efficacité dans la gestion des transformations mathématiques complexes. La précision dans ces opérations est essentielle pour maintenir l'exactitude du système, impactant directement la fidélité du signal lors de la réception. Par ailleurs, les opérations de multiplication consomment des ressources computationnelles substantielles. Par conséquent, l'introduction d'approximations dans les opérations de multiplication offre une opportunité d'améliorer l'efficacité énergétique (EE) globale de la détection des signaux.

À mesure que les circuits matériels deviennent plus grands et plus complexes, il y a une quête croissante de méthodes novatrices pour optimiser la consommation d'énergie et des ressources matérielles associées. Le calcul approximatif a émergé comme un paradigme attrayant, offrant un compromis où un sacrifice limité de précision peut conduire à une amélioration de l'EE. Des techniques dédiées au calcul approximatif sont essentielles pour l'implémentation sur FPGA, car les avantages dérivés de techniques croisées entre les implémentations FPGA et Application Specific Integrated Circuit (ASIC) sont asymétriques [67].

4.1.2 Méthodologie

Cet article introduit la méthodologie LC pour introduire systématiquement une approximation contrôlée dans les circuits précis pour les opérations de multiplication signée au stade pré-synthèse, destinées à une implémentation sur FPGA. La méthodologie LC réduit efficacement l'utilisation des LUT et le PDP pour les circuits précis Booth et BW. Le travail différencie la structure logique des entités matérielles et analyse la logique du circuit. La méthodologie repose sur des valeurs Truth Probability (TP), permettant la création de circuits modulaires de multiplication approximative basés sur un seul paramètre configurable, contrairement aux circuits approximatifs statiques.

La méthodologie LC permet une gestion précise des niveaux d'approximation dans les circuits de multiplication à l'aide d'un réglage paramétrique unique. Pour évaluer l'impact global de ces circuits approximatifs, leur effet sur la fidélité du signal dans la détection montante ZF MIMO est évalué comme étude de cas, en remplaçant des multiplications précises par des opérations de multiplication approximatives.

Les circuits de multiplication conçus pour des calculs approximatifs dans l'environnement C sont compilés dans une bibliothèque dynamique pour accélérer la simulation et fournir une précision numérique élevée. Cette bibliothèque est ensuite connectée à une implémentation Python de la détection MIMO, où la matrice de canal est générée aléatoirement pour la simulation. L'implémentation en C interagit avec l'environnement Python via la bibliothèque ctypes. La bibliothèque NumPy est utilisée pour les calculs numériques et Matplotlib pour la visualisation des données en Python.

4.1.3 Synthèse Complète

En comparant les opérations de multiplication contemporaines sur 16 bits, Booth-Approx obtient les meilleures valeurs pour Mean Error Distance (MED), Mean Relative Error Distance (MRED), et Normalized Mean Error Distance (NMED), tandis que AxBM2 excelle en Max Rel. Les circuits de multiplication AxBM et MUL_xy_k présentent des performances inférieures en termes de MED, MRED, et NMED. Les variantes LC montrent des améliorations variées. LC-BW-1 améliore le MED de 1,5×, le NMED de 1,5×, et Max Rel d'environ 2×, mais dégrade le MRED d'environ 1,1×. LC-BW-2 offre des améliorations encore plus importantes, augmentant le MED, NMED, et Max Rel d'environ 3×, 3×, et 5×, respectivement, bien qu'il dégrade le MRED d'environ 6×. LC-Booth-1 présente une dégradation du MED, MRED, et NMED d'environ 1,5×, 1,6×, et 1,5×, respectivement. Cependant, il améliore significativement Max Rel par un ordre de 10^3. LC-Booth-2 améliore le MED, MRED, et NMED d'environ 1,4×, 1,1×, et 1,4×, respectivement, et Max Rel par un ordre de 10^3. LC-Booth-2 surpasse les circuits de multiplication approximative contemporains dans toutes les métriques d'erreur.

Pour l'implémentation matérielle, Booth-Approx est le meilleur en termes de #LUT, AxBM2 excelle pour le Critical Path Delay (CPD), et MUL_xy_k (x=1, y=3, k=4) est optimal pour la puissance et le PDP. LC-Booth-1 et LC-Booth-2 sont respectivement environ 1,4× et 1,3× plus efficaces en termes de #LUT, tandis que LC-BW-1 et LC-BW-2 utilisent environ 1,2× et 1,3× plus de LUT. Concernant le CPD, LC-BW-1 et LC-BW-2 présentent des CPD environ 3× plus élevés, alors que LC-Booth-1 et LC-Booth-2 sont environ 2,3× plus élevés. La consommation énergétique est environ 2× plus élevée pour LC-Booth-1 et LC-Booth-2, et environ 2,4× plus élevée pour LC-BW-1 et LC-BW-2. Les valeurs de PDP pour LC-BW-1 et LC-BW-2 sont environ 4,6× plus élevées, tandis

que pour LC-Booth-1 et LC-Booth-2, elles sont environ 2,7× plus élevées.

Pour l'analyse SER dans la détection montante MIMO à l'aide de ZF, les circuits de multiplication AxBM ne convergent pas, tandis que les circuits MUL_xy_k nécessitent une opération de 16 bits plus élevée pour converger. En revanche, les circuits LC-BW montrent les meilleures performances, avec LC-BW-2 surpassant LC-BW-1. Pour un SER de 1%, les circuits LC divergent d'environ 3 dB pour une opération sur 8 bits et d'environ 1 dB pour une opération sur 16 bits. Les performances des circuits de multiplication basés sur LC sont présentées dans le tableau 4-1.

LC démontre les meilleures performances d'erreur, tandis que les circuits LC-Booth réduisent la consommation de ressources en termes d'utilisation de LUT. Les caractéristiques d'erreur des circuits de multiplication approximative deviennent cruciales pour atteindre une détection ZF robuste, en particulier dans des configurations à faible largeur de bits et avec des modulations QAM élevées, soulignant la nécessité de prendre en compte attentivement les métriques d'erreur dans le choix du circuit.

Table 4-1 Performances des circuits de multiplication LC.

KPI	**État de l'art (Meilleur)**	**Circuits de multiplication LC.**
MED	Booth-Approx	LC-BW-1 : Amélioration de 1,5× LC-BW-2 : Amélioration de 3× LC-Booth-1 : Dégradation de 1,5× LC-Booth-2 : Amélioration de 1,4×
MRED	Booth-Approx	LC-BW-1 : Dégradation de 1,1× LC-BW-2 : Dégradation de 6× LC-Booth-1 : Dégradation de 1,6× LC-Booth-2 : Amélioration de 1,1×
NMED	Booth-Approx	LC-BW-1 : Amélioration de 1,5× LC-BW-2 : Amélioration de 3× LC-Booth-1 : Dégradation de 1,5× LC-Booth-2 : Amélioration de 1,4×
Max Rel	AxBM2	LC-BW-1 : Amélioration de 2× LC-BW-2 : Amélioration de 5× LC-Booth-1 : Amélioration d'un ordre de 10^3 LC-Booth-2 : Amélioration d'un ordre de 10^3
#LUT	Booth-Approx	LC-BW-1 : 1,2× plus élevé LC-BW-2 : 1,3× plus élevé LC-Booth-1 : 1,4× plus efficace LC-Booth-2 : 1,3× plus efficace
CPD	AxBM2	LC-BW-1 : 3× plus élevé LC-BW-2 : 3× plus élevé LC-Booth-1 : 2,3× plus élevé LC-Booth-2 : 2,3× plus élevé
Puissance	MUL_xy_k (x=1, y=3, k=4)	LC-BW-1 : 2,4× plus élevé LC-BW-2 : 2,4× plus élevé LC-Booth-1 : 2× plus élevé LC-Booth-2 : 2× plus élevé
PDP	MUL_xy_k (x=1, y=3, k=4)	LC-BW-1 : 4,6× plus élevé LC-BW-2 : 4,6× plus élevé LC-Booth-1 : 2,7× plus élevé LC-Booth-2 : 2,7× plus élevé
SER en MIMO	Booth-Approx	LC-BW-1 : Surpasse Booth-Approx LC-BW-2 : Surpasse LC-BW-1

4.1.4 Droits d'Auteur

4.2 Long abstract

4.2.1 Research Context

Multiplication operations play a heavy role in wireless signal detection, influencing processing capability and the efficiency of handling complex mathematical transformations. Precision in these operations is essential for maintaining system accuracy, directly impacting signal fidelity during reception. Also, multiplication operations consume substantial computational resources. Therefore, introducing approximations in multiplication operations offers an opportunity to enhance the overall EE of signal detection.

As hardware circuits grow larger and more complex, there is a growing pursuit of novel methods to optimize energy and hardware resource consumption associated with the circuits. Approximate computing has emerged as a compelling paradigm, offering a trade-off where sacrificing some degree of system accuracy can lead to EE. Dedicated techniques for approximate computing are essential for FPGA implementation, as the benefits derived from cross-porting techniques between FPGA and ASIC implementations are asymmetric [67].

4.2.2 Methodology

This article introduces the LC methodology to systematically introduce controlled approximation into accurate circuits for signed multiplication operations at the pre-synthesis stage intended for FPGA implementation. The LC methodology effectively

reduces LUT utilization and PDP for accurate Booth and BW circuits. The work differentiates the logic structure from hardware entities and analyzes the circuit logic. The methodology relies on TP values, enabling the creation of modular approximate multiplication circuits based on a single configurable parameter, contrasting with static approximate circuits.

The LC methodology facilitates precise management of approximation levels in multiplication circuits using single parameter tuning. To assess the broader impact of these approximate circuits, their effect on signal fidelity in ZF MIMO uplink detection is evaluated as a case study, by substituting accurate multiplication with approximate multiplication operations.

Multiplication circuits designed for approximate computations in the C environment are compiled into a dynamic library to accelerate simulation and provide high numerical precision. This library is then connected to a Python implementation of MIMO detection, where the channel matrix is randomly generated for simulation. The C based circuit implementation interacts with the Python environment through the ctypes library. The NumPy library is used for numerical computations and Matplotlib for data visualization in Python.

4.2.3 Comprehensive Synthesis

Comparing contemporary 16-bit multiplication operations, Booth-Approx achieves the best values for MED, MRED, and NMED, while AxBM2 excels in Max Rel. AxBM and MUL_xy_k multiplication circuits exhibit poorer performance in MED, MRED, and NMED. The LC variants show varied improvements. LC-BW-1 improves MED by 1.5×, NMED by 1.5×, and Max Rel by about 2× each, but suffers degradation in MRED by

about 1.1×. LC-BW-2 offers even greater improvements, boosting MED, NMED, and Max Rel by approximately 3×, 3×, and 5×, respectively, although it degrades MRED by about 6×. LC-Booth-1 shows a degradation in MED, MRED, and NMED by about 1.5×, 1.6×, and 1.5×, respectively. However, it significantly improves Max Rel by an order of 10^3. LC-Booth-2 improves MED, MRED, and NMED by about 1.4×, 1.1×, and 1.4×, respectively, and Max Rel by an order of 10^3. LC-Booth-2 outperforms contemporary approximate multiplication circuits in all error metrics.

For hardware implementation, Booth-Approx has the best #LUT, AxBM2 has the best CPD, and MUL_xy_k(x=1,y=3,k=4) excels in both Power and PDP. LC-Booth-1 and LC-Booth-2 are about 1.4× and 1.3× more efficient in terms of #LUT, respectively, while LC-BW-1 and LC-BW-2 use about 1.2× and 1.3× more LUTs. In terms of CPD, LC-BW-1 and LC-BW-2 have approximately 3× higher CPD, whereas LC-Booth-1 and LC-Booth-2 exhibit about 2.3× higher CPD. Power consumption is about 2× higher for LC-Booth-1 and LC-Booth-2, and about 2.4× higher for LC-BW-1 and LC-BW-2. PDP values for LC-BW-1 and LC-BW-2 are about 4.6× higher, while for LC-Booth-1 and LC-Booth-2, they are about 2.7× higher.

For SER analysis in MIMO uplink detection using ZF, AxBM multiplication circuits do not converge, while MUL_xy_k circuits require a higher 16-bit operation to converge. In contrast, LC-BW circuits show the best performance, with LC-BW-2 outperforming LC-BW-1. For a 1% SER, the LC circuits diverge by approximately 3 dB for 8-bit and by 1 dB for 16-bit operations. The performance of LC based multiplication circuits is shown in Table. 4-2.

LC approximate circuits demonstrate best error performance, while LC-Booth circuits exhibit reduced resource consumption in terms of LUT utilization. The error

characteristics of approximate multiplication circuits become significant in achieving robust ZF detection, especially under conditions of high QAM and low bit-width configurations, emphasizing the need for careful consideration of error metrics in circuit selection. It can also be inferred that no single error metric can be used to optimally choose the approximate multiplication circuit for ZF MIMO signal detection.

Table 4-2 Performance of LC multiplication circuits.

KPI	**State-of-art (Best)**	**LC multiplication circuits.**
MED	Booth-Approx	LC-BW-1: Improves 1.5× LC-BW-2: Improves 3× LC-Booth-1: Degrades 1.5× LC-Booth-2: Improves 1.4×
MRED	Booth-Approx	LC-BW-1: Degrades 1.1× LC-BW-2: Degrades 6× LC-Booth-1: Degrades 1.6× LC-Booth-2: Improves 1.1×
NMED	Booth-Approx	LC-BW-1: Improves 1.5× LC-BW-2: Improves 3× LC-Booth-1: Degrades 1.5× LC-Booth-2: Improves 1.4×
Max Rel	AxBM2	LC-BW-1: Improves 2× LC-BW-2: Improves 5× LC-Booth-1: Improves order of 10^3 LC-Booth-2: Improves order of 10^3
#LUT	Booth-Approx	LC-BW-1: 1.2× higher LC-BW-2: 1.3× higher LC-Booth-1: 1.4× efficient LC-Booth-2: 1.3× efficient
CPD	AxBM2	LC-BW-1: 3× higher LC-BW-2: 3× higher LC-Booth-1: 2.3× higher LC-Booth-2: 2.3× higher
Power	MUL_xy_k (x=1, y=3, k=4)	LC-BW-1: 2.4× higher LC-BW-2: 2.4× higher LC-Booth-1: 2× higher LC-Booth-2: 2× higher
PDP	MUL_xy_k (x=1, y=3, k=4)	LC-BW-1: 4.6× higher LC-BW-2: 4.6× higher LC-Booth-1: 2.7× higher LC-Booth-2: 2.7× higher
SER in MIMO	Booth-Approx	LC-BW-1: Outperforms Booth-Approx LC-BW-2: Outperforms LC-BW-1

4.2.4 Copyright

Microelectronics Journal 145 (2024) 106135

Contents lists available at ScienceDirect

Microelectronics Journal

journal homepage: www.elsevier.com/locate/mejo

Logic cloning based approximate signed multiplication circuits for FPGA ☆

Abhinav Kulkarni *, Messaoud Ahmed Ouameur, Daniel Massicotte

Electrical and Computer Engineering Department, Université du Québec à Trois-Rivières, Trois-Rivières, QC G9A 5H7, Canada

ARTICLE INFO

Dataset link: https://github.com/abhinav333/Logic-Cloning

Keywords:
Approximate computing
Reconfigurable computing
Logic cloning
Arithmetic multiplication
Massive MIMO

ABSTRACT

As hardware circuits become larger and more intricate, there is a growing need for approximate circuit techniques. These approaches offer a trade-off, sacrificing some system accuracy in exchange for greater hardware resource efficiency and energy conservation. In the context of FPGA-based computation-intensive arithmetic multiplication, Logic Cloning (LC) is introduced to systematically induce controlled approximation. LC-Baugh Wooley (BW) circuits deliver exceptional error performance with precise approximation, while LC-Booth circuits are characterized by reduced Look-Up Table (LUT) resource consumption. In the case of 16-bit operands, LC methods effectively reduce LUT resource consumption by 31.05% for Booth and 36.85% for BW. Additionally, compared to their accurate counterparts, they lower the Power Delay Product (PDP) by 34% for Booth and 35% for BW. When it comes to symbol error-rate performance for Zero Forcing (ZF) Multiple Input Multiple Output (MIMO) uplink detection, these LC approximate multiplication circuits exhibit robust performance, particularly LC-BW circuits, which closely match the accuracy of ZF detection, followed by LC-Booth circuits.

1. Introduction

Modern applications involving multimedia and communication systems require extensive data processing and therefore need high computing resources. Also, energy dissipation has become a fundamental barrier to scale computing performance across multiple hardware computing platforms. To accommodate the requirement for next-generation computing applications, several techniques are put forth that leverage the existing hardware capability to provide demanding application performance, while simultaneously optimizing resource and energy efficiency. Approximate computing [1,2] is one such emerging technology that enables explicit control over the energy and hardware resource consumption for the intended application by incurring penalties in the system accuracy. The philosophy behind approximate computing is that application systems do not always have to provide the most accurate results for acceptable system performance [3].

In a logic circuit, the system functionality of the circuit is represented by a Boolean logic function. Logic circuits have been conventionally designed using deterministic logic gates with the input/output deterministic bit signals being logic '1' or '0'. The work [4] explores a pre-synthesis iterative approach to Application Specific Integrated Circuit (ASIC) approximation by pruning the logic gates based on the probability of active logic. Probabilistic bit signals are transformed to intended probabilities with deterministic combinational logic [5]. The characteristic of Boolean functions to output signal logic '1' and '0' with certain probability is exploited in [6] to form Probabilistic Boolean Logic (PBL) gate models. PBL employs implicit probabilistic logic gates. However, randomized circuits [7] use random input bits with deterministic gates for circuit output approximation, where the circuit operates with random inputs. Furthermore, gate-level methodologies are suited for ASIC implementation and cannot be ported for FPGA implementation because of architectural differences between both [8].

In reconfigurable computing, FPGA hardware implementation is characterized by adaptable hardware reconfiguration with uniform fundamental hardware structures, unlike ASIC hardware implementation. Resource and energy optimization of Register Transfer Level (RTL) circuits is a strategic task for both ASIC and FPGA implementations. At the RTL stage, ASIC circuits are composed of logic gates, while the FPGA circuits are composed of Configurable Logic Block (CLB), primarily comprising of LUTs, which are spread across the FPGA fabric. For a specific RTL architecture, ASIC hardware implementation is inherently about 35× efficient in terms of silicon area utilization, 14× efficient in terms of dynamic power consumption and about 4× faster than FPGA hardware implementation [8]. Hence, the ASIC implementations show asymmetric gains when they are ported to FPGA and vice versa. The

☆ This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), in part by Prompt, in part by the Canadian Foundation for Innovation (CFI), in part by CMC Microsystems, in part by OPAL-RT Technologies Inc. and in part by Hydro-Québec. Laboratoire des signaux et systèmes intégrés and Chaire de recherche sur les signaux et l'intelligence des systèmes haute performance (www.uqtr.ca/lssi).

* Corresponding author.
E-mail addresses: Abhinav.Kulkarni@uqtr.ca (A. Kulkarni), Messaoud.Ahmed.Ouameur@uqtr.ca (M.A. Ouameur), daniel.massicotte@uqtr.ca (D. Massicotte).

https://doi.org/10.1016/j.mejo.2024.106135
Received 30 October 2023; Received in revised form 14 January 2024; Accepted 18 February 2024
Available online 19 February 2024

proposed work is focused on the discussion of comparative techniques for FPGA hardware implementation. Therefore, *can the probabilistic nature of Boolean logic be utilized for inducing approximation in specific* FPGA *circuits?*

For the computationally intensive operation of arithmetic multiplication, optimal utilization of FPGA resources and power can be achieved either by efficiently implementing RTL of the circuit or by introducing approximation in the accurate RTL [9]. Dedicated heterogeneous ASIC DSP blocks are embedded in the FPGA fabric for efficient multiplication operation. However, these blocks improve the area and power consumption efficiency without providing any further scope for approximation based on the FPGA application [8]. Also, applications have to stringently adhere to the operand bit-width of the DSP block for significant efficiency gains. Hence, research on soft-core FPGA multiplication circuits is gaining ground.

1.1. Preliminary

The multiplication operation of multiplicand and multiplier generates Partial Product (PP)s, which are accumulated to produce the multiplication product. Carry Save Adder (CSA) structure is employed with a configuration like Wallace [10] or Dadda [11] for accumulation of PPs. Booth algorithm encodes the circuit operand bits which effectively reduces the total number of PPs required for the computation of the multiplication product. Unsigned multiplication involves unsigned operands with unsigned multiplication product, whereas signed multiplication is characterized by signed operands with signed multiplication product. Multiplier architectures have been adapted for FPGA by utilizing LUT and Carry Chain (CC) (comprising several Carry Chain Unit (CCU)s) primitives for the accurate multiplication operation. Guidelines for efficient implementation of accurate unsigned multiplication for FPGA implementation [12] are suggested in the form of efficient mapping and arithmetic transformation for improvement in the logic density of the implementation. However, this approach requires an explicit compressor tree for the accumulation of PPs and the implementation is suited for a 3-input LUT [13]. In the Multiply And Accumulate (MAC) architecture of FPGA multiplication circuits, the PP bits are simultaneously generated and accumulated using LUTs and CCUs. An accurate unsigned multiplication circuit [14] using Booth encoding eliminates the need for an explicit compressor tree for PPs accumulation by employing MAC operation and utilizes a contemporary 6-input LUT structure for FPGA implementation. However, its RTL implementation suffers from the under-utilization of LUT inputs for PP bit generation and is specifically intended for the unsigned multiplication operation.

For achieving signed multiplication using unsigned circuit implementation, the operands in two's complement format are segregated into sign and magnitude using sign-converters. The unsigned multiplication is performed with the magnitude of these operands and the multiplication product is converted back to two's complement by the sign-converter using signs of operands. Signed multiplication operation using sign-converters adds computational overhead for multiplication operation [15,16] in terms of increased LUT resource consumption. A smart approach for signed multiplication is by modifying the PPs generation by employing sign-extension for PPs using BW method, which avoids any kind of explicit sign conversion of operands. Also, the Booth encoding technique is modified for PPs generation [17], further reducing the number of PPs for signed multiplication.

1.2. Relevant work

Efficient implementation of an accurate signed multiplication circuit employing the Booth algorithm [16] reduced the CC by 3 CCUs for the generation of PPs, thereby optimizing the hardware resource consumption. However, an external hardware entity like an adder or a compressor was required for PPs accumulation and the accuracy of the circuit thus depends on the PP accumulation methodology utilized. Using the FPGA implementation of an accurate unsigned multiplication circuit [14], an architecture using LUTs and CCs for accurate signed multiplication circuit based on Booth encoding [18] known as Booth-Opt saved LUTs by embedding the logic of the carry generating rightmost LUT and the leftmost LUT into adjacent LUT in a PP row, thereby achieving a reduction in CC. An improvement in critical path delay of the work in [18] was presented in the further work [19] by optimizing the PPs generation and employing PP reduction tree, however, it was characterized by increased LUT resource consumption.

Performance gains were obtained by introducing approximations in accurate multiplication architectures. Approximations can be employed in PP accumulation by using FPGA implementation of approximate adder [20–22] or by approximate compressor [23], while the PPs are generated accurately. Configurable signed multiplication circuit architectures [24] were built using a combination of accurate and four approximate compressors for PPs accumulation. As these circuit architectures utilized explicit PP accumulation method, the LUT resource consumption was variable with circuit accuracy for a particular bit-width operation.

The functional approximation was introduced in the accurate circuit while generation of PPs. The Booth-Opt circuit was further modified for Booth-Approx [18] by approximating its PPs generation, specifically by approximating the carry signal generation at the rightmost LUT for all PPs, except the last PP. Booth-Approx was characterized by a reduction in CC chain and LUT resource consumption. However, Booth-Approx was non-configurable to achieve a trade-off with multiplication accuracy and energy efficiency. AxBM circuits [25] functionally modified the accurate Radix-8 Booth encoding for FPGA implementation to alleviate the generation of challenging 3×Multiplicand. The 3×Multiplicand was approximated by 4×Multiplicand in AxBM1 and AxBM2, while −3×Multiplicand was approximated by −4×Multiplicand in AxBM1 and −2×Multiplicand in AxBM2 respectively. For generating multiples of multiplicand for AxBM1 and AxBM2, 4× signal was generated by implicit XNOR operation of 1× and 2× signals in LUT of PP bit generation, thereby mapping the Booth encoder to a two output 5-input LUT configuration. However, AxBM1 and AxBM2 were characterized by non-MAC PPs accumulation and very poor multiplication product accuracy.

The presented work introduces a heuristic methodology designed to induce approximation in arithmetic multiplication circuits suitable for FPGA implementation at the pre-synthesis stage. The study distinguishes the logic structure from hardware entities and analyzes the circuit logic. The methodology relies on Truth Probability (TP) values, allowing for the creation of modular approximate multiplication circuits based on a single configurable parameter, in contrast to static approximate circuits in the works [18,25]. Additionally, the proposed approach employs a MAC architecture for PP accumulation, eliminating the need for explicit PP accumulation found in works [24,25]. In the work [24], the reduction stages for PP accumulation increase with the operand bit-width and have under-utilization of inputs for approximate compressors in certain instances, which is not the case for MAC architecture implementations.

1.3. Contributions

The following contributions are presented in the current work:

- Devise LC methodology for heuristically approximating the accurate multiplication circuits for FPGA implementation. In this pioneering work, LC methodology systematically utilizes the probability of logic '1' of the LUT logic structure output to induce approximation in the accurate circuit.
- Proposition of novel LC-BW and LC-Booth approximate circuits for signed multiplication. LC-BW circuits are characterized by high multiplication accuracy, their approximation being controlled with finer granularity, while LC-Booth circuits are characterized by low LUT resource consumption.

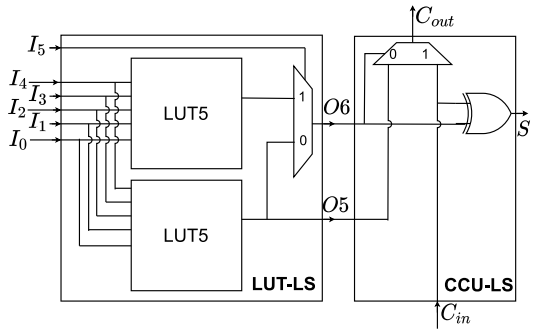

Fig. 1. LUT-LS and CCU-LS for Xilinx FPGA device [26].

- Analysis of proposed LC approximate circuits for LUT consumption and energy efficiency for 8-bit and 16-bit configurations with competitive signed approximate multiplication circuits. The approximation induced by LC methodology in accurate multiplication circuits is effective in reducing the LUT consumption and PDP.
- Evaluation of approximate signed circuit for symbol error-rate performance of MIMO uplink detection using ZF algorithm for 16-QAM and 64-QAM configuration. LC approximate signed circuits provides close to accurate ZF detection.

The paper is organized as follows; Section 1 discusses related work on approximate circuits techniques, motivating the need for softcore FPGA circuits. Section 2 explains the proposed methodology for the approximation of FPGA implementation of multiplication circuits in detail. Section 3 explains the application of the proposed methodology for signed circuits. Section 4 presents an error analysis of the proposed work with contemporary signed approximate multiplication circuits, while Section 5 discusses FPGA implementation analysis. Section 6 evaluates approximate signed multiplication circuits for MIMO uplink detection and discusses gains of the LC circuits for various MIMO configurations. Section 7 concludes the discussion by summarizing the LC methodology for FPGA implementation and its further potential.

2. Proposed methodology

2.1. Notations

$\mathcal{P}\{\mathcal{G}\}$ denotes the probability of occurrence of any event $\mathcal{G}$. The notation $|\mathbb{G}|$ denotes the cardinality of any finite set $\mathbb{G}$. Operators $\vee$ and $\wedge$ denote logical OR and AND operations. The symbol $\hat{.}$ denotes the approximate evaluation of the signal or value. Terms 'LUT', 'CC' and 'CCU' are used to denote the respective hardware entities henceforth, while the term with the suffix '-LS' denotes the logic structure of the corresponding hardware entity. The row of LUTs and CC which performs the MAC operation involving PP is denoted as MAC PP (MPP) for dexterity. $(.)^{(i)}$ denotes evaluation of any signal or value at the ith simulation step. $[.]_2$ denotes integer values represented in two's complement form.

2.2. FPGA hardware architecture

A FPGA device consists of CLBs spread across the FPGA fabric. For implementing arithmetic circuits, the carry logic is implemented using a dedicated CC chain comprising multiple CCUs. A CLB structure is comprised of slice units and each slice unit comprises four 6-input LUTs, a CC chain, and an associated circuit [26]. Every LUT is physically connected to a single CCU as shown in Fig. 1. For the Xilinx FPGA device, every LUT-LS can implement either two 5-input logic functions with shared inputs or a 6-input and 5-input logic function with shared inputs and shared logic values [27]. Every LUT is configured with a 64-bit hexadecimal value defining its characteristic LUT-LS.

In MAC architecture, MPPs are produced by rows of LUTs and CCs. The LUT-LS of the LUTs of MPP performs the MAC operation and the intermediate value produced is denoted as $\beta-$MPP value. The associated CC-LS of the CC chain of MPP transforms the $\beta-$MPP value into MPP value. The final MPP value is considered as the product of the multiplication operation. The $O6$ signal of the LUT-LS contributes a specific bit of the $\beta-$MPP value, while the S signal of the CCU-LS contributes a specific bit of the MPP value. A CCU-LS XORs the signal from LUT-LS with a prior carry signal and also computes the carry signal to be propagated to the next CCU-LS in the CC-LS chain. The CCU-LS as shown in Fig. 1 computes the following Boolean functions:

$$C_{out} = (\overline{O6} \wedge O5) \vee (C_{in} \wedge O6) \tag{1}$$

$$S = (C_{in} \wedge \overline{O6}) \vee (\overline{C_{in}} \wedge O6) \tag{2}$$

2.3. Probabilistic analysis

Signals of a FPGA multiplication circuit comprise of the inputs, interconnects and outputs. To investigate the potential for functional approximation in the circuit architecture, it is essential to develop a mathematical model to perform the probabilistic analysis of the circuit signals. The accurate multiplication circuit has inputs comprising of two N-bit signed operands and outputs forming one $2N$-bit multiplication product. Let the finite set $\mathbb{B}$ be defined such that $\mathbb{B} = \{0, 1\}$. Consider a FPGA multiplication circuit with $2N$ inputs, where each input is represented as $\psi_i \in \mathbb{B}$ for $i = 0, 1, \ldots, 2N$. The exhaustive simulation of the circuit comprises 2^{2N} unique operational states based on the input combinations. Let a particular simulation step in exhaustive simulation be denoted as ζ such that $0 \leq \zeta \leq 2^{2N} - 1$. A specific combination of $\psi_i^{(\zeta)}$ for a particular simulation step ζ is represented using a finite set $\mathbb{P}_\zeta = \{\psi_i^{(\zeta)} : 0 \leq i \leq 2N\}$. For the exhaustive simulation with $\mathbb{P}_\zeta$ for all ζ, the sample space is represented using a finite set $\mathbb{S} = \{\mathbb{P}_\zeta : 0 \leq \zeta \leq 2^{2N} - 1\}$.

Let $X \in \mathbb{B}$ represent a Boolean variable [28] for any signal of the circuit. The set of all such values of X due to exhaustive simulation is denoted by the finite set $\mathbb{X} = \{X^{(\zeta)} : 0 \leq \zeta \leq 2^{2N} - 1\}$. The number of logic '1' in the set $\mathbb{X}$ is $\mathbb{X}_1 = \{X^{(\zeta)} : 0 \leq \zeta \leq 2^{2N} - 1, X^{(\zeta)} = 1\}$. Let $[Y]_2$ be any arithmetic value computed in the circuit, which comprises of arithmetic combination of different signals. Then, $\mathbb{Y}$ denotes the finite set of values generated by $[Y]_2$ such that $\mathbb{Y} = \{[Y]_2^{(\zeta)} : 0 \leq \zeta \leq 2^{2N} - 1\}$.

The probability that X is logic '1' for the exhaustive simulation of the circuit is called the TP of X and evaluated as $\mathcal{P}\{X = 1\}$. The expression $\mathcal{T}(X)$ is used to imply $\mathcal{P}\{X = 1\}$ for dexterity. TP of X is computed as:

$$\mathcal{T}(X) = \frac{|\mathbb{X}_1|}{|\mathbb{X}|} = \frac{|\mathbb{X}_1|}{2^{2N}} \tag{3}$$

The arithmetic mean of X is computed as:

$$\mathcal{A}(X) = \sum_{X^{(\zeta)} \in \mathbb{X}} X^{(\zeta)} / |\mathbb{X}| \tag{4}$$

Similarly, the arithmetic mean of $[Y]_2$ is computed as:

$$\mathcal{A}([Y]_2) = \sum_{[Y]_2^{(\zeta)} \in \mathbb{Y}} [Y]_2^{(\zeta)} / |\mathbb{Y}| \tag{5}$$

But, since the sum of elements of $\mathbb{X}$ is equal to cardinality of $\mathbb{X}_1$:

$$\mathcal{T}(X) = \frac{|\mathbb{X}_1|}{|\mathbb{X}|} = \frac{\sum_{\zeta=0}^{2^{2N}-1} X^{(\zeta)}}{|\mathbb{X}|} = \mathcal{A}(X) \tag{6}$$

During circuit simulation, the calculation of the average value for any arithmetic value within the circuit is facilitated by Eq. (5). Eq. (6) allows for the computation of the TP value for any signal in the circuit.

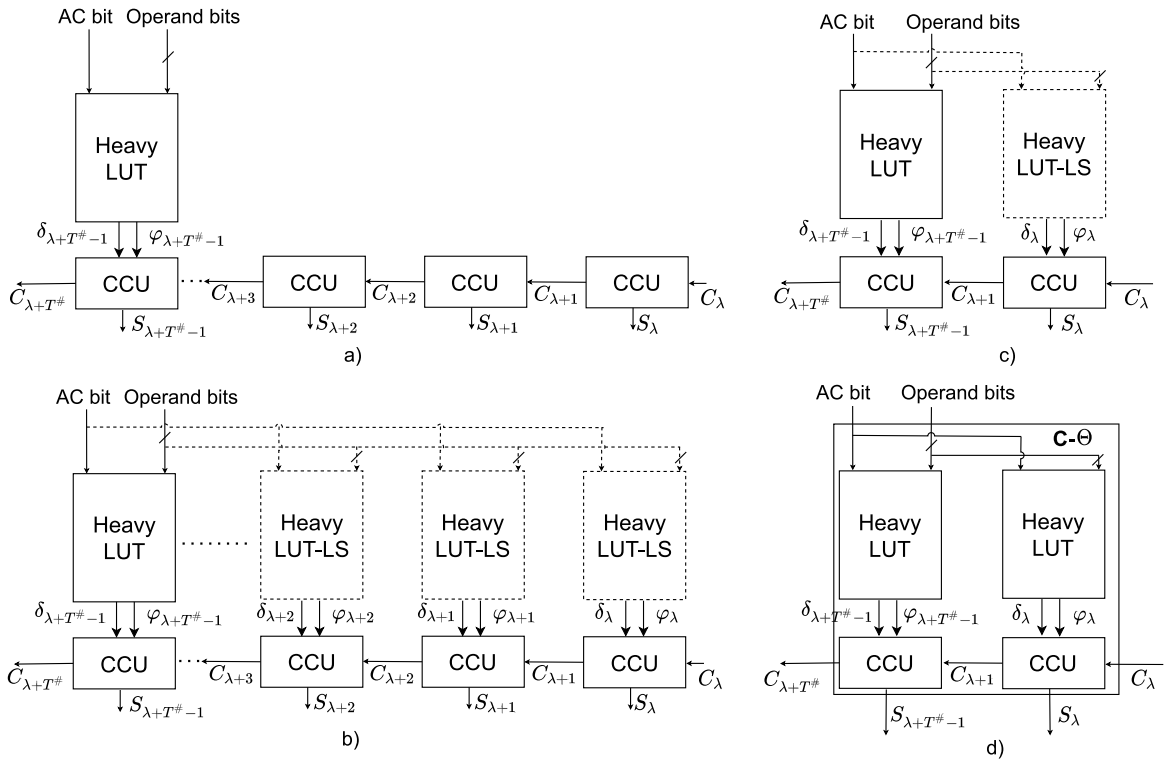

Fig. 2. (a) LUTs removed from MPP based on parameter $T^{\#}$ (b) Logic cloning of LUT-LS of Heavy LUT (c) CC chain reduction (d) C-Θ structure, where Θ represents LUT-LS configuration of Heavy LUT.

2.4. Error metrics

For evaluating approximate multiplication circuits, error metrics [29] capture the relative accuracy of the accurate and approximate multiplication product. For operands $[A]_2$ and $[B]_2$, $[\Phi]_2$ and $[\hat{\Phi}]_2$ represent the accurate and approximate multiplication product respectively. $\Delta_\zeta = [\Phi]_2^{(\zeta)} - [\hat{\Phi}]_2^{(\zeta)}$ computes the arithmetic difference between accurate and approximate multiplication product for the ζth simulation step. The sum of absolute errors in terms of the arithmetic mean of the multiplication product is computed using triangle inequality as:

$$\begin{aligned}\sum_{\zeta=0}^{2^{2N}-1} |\Delta_\zeta| &\geq |\sum_{\zeta=0}^{2^{2N}-1} \Delta_\zeta| \\ &= |\sum_{\zeta=0}^{2^{2N}-1} \Delta_\zeta| + Y \\ &= |\sum_{\zeta=0}^{2^{2N}-1} [\Phi]_2^{(\zeta)} - \sum_{\zeta=0}^{2^{2N}-1} [\hat{\Phi}]_2^{(\zeta)}| + Y \\ &= |\mathcal{A}([\Phi]_2) - \mathcal{A}([\hat{\Phi}]_2)| 2^{2N} + Y \end{aligned} \tag{7}$$

where $Y \geq 0$ is an arbitrary constant.

Mean Error Distance (MED) considers the averaging effect of the sum of absolute errors over the complete operand range, which is useful in measuring the implementation accuracy of the approximate circuit.

$$MED = \frac{1}{2N} \sum_{\zeta=0}^{2^{2N}-1} |\Delta_\zeta| \tag{8}$$

Relative Error Distance (RED) computes the relative error distance of an approximate multiplication product with respect to its accurate multiplication product (RED is valid only for non-zero accurate multiplication output). Mean Relative Error Distance (MRED) measures the arithmetic mean of RED for all valid accurate multiplication products.

$$MRED = \frac{1}{2N} \sum_{\zeta=0}^{2^{2N}-1} \frac{|\Delta_\zeta|}{|[\Phi]_2^{(\zeta)}|} \tag{9}$$

RED and MRED are useful in evaluating approximate circuit designs with varying operand ranges. Normalized Mean Error Distance (NMED) is a metric that normalizes MED with the maximum value of the accurate multiplication product over the operand range. Also, the maximum relative error (Max Rel) of an approximate multiplier circuit is a useful metric that pinpoints the worst-case operation of the approximate circuit design.

2.5. Logic cloning

Exhaustive simulation involves simulating the accurate multiplication circuit with a range of 2^{2N} distinct input combinations. Consider bit signals $a_i, b_i, \phi_i \in \mathbb{B}$ for $i = 0, 1, \ldots, N-1$. For signed multiplication, N bit operand multiplicand and circuit are represented in base-2 notation as $[A]_2 = -a_{N-1}2^{N-1} + \sum_{i=0}^{N-2} a_i 2^i$ and $[B]_2 = -b_{N-1}2^{N-1} + \sum_{i=0}^{N-2} b_i 2^i$ respectively, while the multiplication product is denoted as $[\Phi]_2$. The accurate multiplication circuit is prototyped and verified for hardware implementation by mapping the multiplication architecture to the FPGA architecture. The input bit signals of the circuit architecture are mapped such that $\psi_i \leftarrow a_i$ when $0 < i \leq N-1$ and $\psi_i \leftarrow b_i$ when $N < i \leq 2N-1$. The verified FPGA circuit is modeled in the C environment to compute the TP value of the $O6$ signal of the LUT-LSs of the accurate multiplication circuit. The LUT-LS function blocks are constructed to represent the Boolean logic function by using LUT-LS hexadecimal configuration value. The CCU-LS function blocks are constructed using the Eqs. (1) and (2). The FPGA interconnects are represented with an indexed 2-dimensional array structure. During exhaustive simulation, the occurrence of logic '1' at every $O6$ signal is recorded to compute its TP using Eq. (3).

For any nth β-MPP of the accurate circuit, the $O6$ signal of the LUT-LS is mapped to φ signal such that $\varphi \leftarrow O6$. The $O5$ signal of the

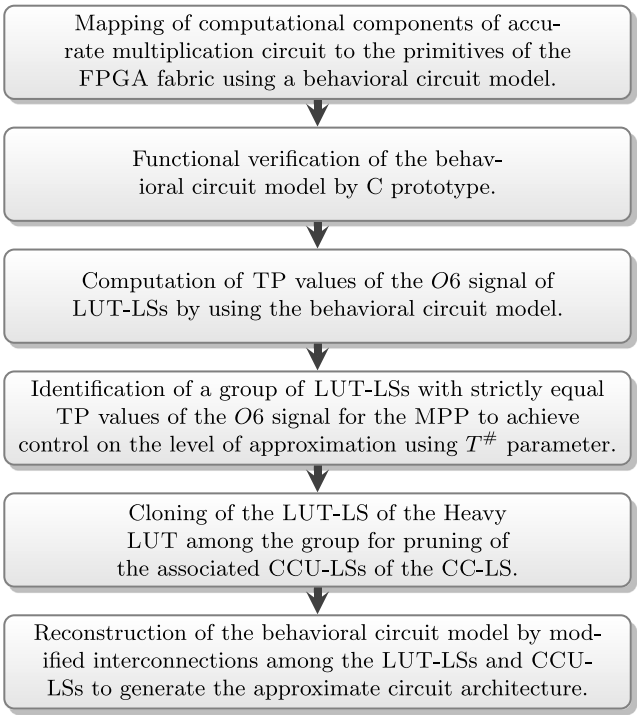

Fig. 3. LC methodology for approximation in FPGA implementation of multiplication circuit as explained in Section *2.6*.

LUT-LS is mapped to δ signal such that $\delta \leftarrow O5$, whose only function is to provide the Accumulator (AC) bit supplied to the LUT-LS for MAC operation to the CCU-LS. For any nth $\beta-$MPP of the circuit, let $\mathcal{T}(\varphi_\lambda) = \mathcal{T}(\varphi_{\lambda+1}) = \cdots \mathcal{T}(\varphi_{\lambda+T-1})$ for any $\lambda > 0$ for T LUT-LSs. Out of T LUT-LSs, with a group of $T^\#$ LUT-LSs such that $0 \leq T^\# \leq T$, the arithmetic mean value of $\beta-$MPP can be effectively approximated with only LUT-LS corresponding to $\varphi_{\lambda+T^\#-1}$ signal as given in Lemma. Thus, the $O6$ signal value of $(T^\# - 1)$ LUT-LSs are redundant to approximate the arithmetic mean of $\beta-$MPP. The LUT corresponding to $\varphi_{\lambda+T^\#-1}$ signal is termed as Heavy LUT and $(T^\# - 1)$ LUTs are removed in the MPP row. Parameter $T^\#$ controls the amount of approximation of the arithmetic mean of $\beta-$MPP, with $T^\# = 0$ resulting in no approximation, while $T^\# = T$ resulting in highest amount of approximation. Approximation in every nth $\beta-$MPP contributes to inducing approximation in the arithmetic mean value of the final multiplication product.

The approximation of the nth $\beta-$MPP makes several LUTs redundant thereby leaving CCU-LSs with dangling input signals as shown in Fig. 2a, which are interlinked to propagate the carry signal. To analyze the redundancy in the CC-LS of the nth $\beta-$MPP, the Heavy LUT-LS is cloned as shown in Fig. 2b. Consider $T^\#$ CCU-LSs of the corresponding $T^\#$ LUT-LSs with equal TP. δ_i and φ_i represent inputs to the ith CCU, while C_i and C_{i+1} represent the C_{in} and C_{out} respectively. Due to logic cloning procedure, $\varphi_\lambda = \varphi_{\lambda+1} = \cdots = \varphi_{\lambda+T^\#-1} \equiv \varphi_\#$. For nth MPP, if $\delta_\lambda = \delta_{\lambda+1} = \cdots = \delta_{\lambda+T^\#-1} \equiv \delta_\#$, then in such case, the Eqs. (1) and (2) for λth CCU-LS become:

$$C_{\lambda+1} = (\overline{\varphi_\#} \wedge \delta_\#) \vee (C_\lambda \wedge \varphi_\#) \tag{10}$$

$$S_\lambda = C_\lambda \oplus \varphi_\# \tag{11}$$

For evaluating the $(\lambda+2)$th carry signal, the Eq. (10) becomes:

$$\begin{aligned} C_{\lambda+2} &= (\overline{\varphi_\#} \wedge \delta_\#) \vee (C_{\lambda+1} \wedge \varphi_\#) \\ &= (\overline{\varphi_\#} \wedge \delta_\#) \vee (((\overline{\varphi_\#} \wedge \delta_\#) \vee (C_\lambda \wedge \varphi_\#)) \wedge \varphi_\#) \\ &= (\overline{\varphi_\#} \wedge \delta_\#) \vee (C_\lambda \wedge \varphi_\#) \end{aligned} \tag{12}$$

It is inferred from Eqs. (10) and (12) that $C_{\lambda+1} = C_{\lambda+2} = \cdots = C_{\lambda+T^\#}$ and $C_{\lambda+T^\#}$ can be equated with $C_{\lambda+1}$. Hence, from Eq. (11), it is also inferred that $S_{\lambda+1} = S_{\lambda+2} = \cdots = S_{\lambda+T^\#}$. Thus, CCU-LS for computing C_{i+1} and S_i for $\lambda+1 \leq i \leq \lambda+T^\#-2$ can be deemed redundant and the corresponding CCUs are pruned as shown in Fig. 2c. However, every LUT is physically connected to a single CCU, hence an additional LUT is required to supply the δ_λ and φ_λ signals as shown in Fig. 2d.

2.6. Overview of LC methodology

The LC methodology for approximating FPGA implementation of multiplication circuits is outlined in Fig. 3. In the initial phases of the multiplication algorithm to architecture mapping, an apt behavioral circuit model intended for implementation on an FPGA is developed, integrating LUT and CC primitives. The behavioral circuit model is prototyped in C by utilizing the logic structure of the primitives. The model undergoes functional verification to guarantee accurate behavior as intended by the multiplication algorithm. Through this careful verification, it is ensured that the chosen combination of primitives effectively captures the desired computational logic of the multiplication algorithm. Subsequently, the approximation process commences by calculating TP values for the output signal $O6$ within the LUT-LSs. Specifically, the TP values are computed by using circuit inputs through analysis as outlined in Section 2.3 using the behavioral circuit model. Upon analyzing the TP values of the MPPs of the behavioral circuit model, a set of LUT-LSs with identical TP values is identified for each MPP. Using a single parameter $T^\#$, a selective removal of a specified number of LUT-LS is employed. This results in corresponding CCU-LSs dangling without any input signals. Hence, a redundancy analysis is initiated as the Heavy LUT-LS is cloned, enabling a detailed examination of redundancy in the CC-LS. This process results in the subsequent removal of redundant CCU-LSs.

The behavioral circuit model obtained at this stage comprises LUT-LS and CCU-LS barring the redundant ones. The interconnects obliterated by redundant logic structures are modified, ensuring adherence to the physical constraints imposed by the FPGA architecture. The methodology generates various approximate circuit versions for multiplication based on the $T^\#$ parameter, providing a range of trade-offs between computational accuracy and resource utilization.

3. Logic cloning for signed multiplication circuits

3.1. Logic cloning for BW circuit

For the computation of the signed multiplication of two operands without the requirement of sign converters, the BW method is an efficient method for generating N PPs, which are accumulated to generate the final product $[\Phi]_2$. For BW architecture, LUT-LSs are configured for implementing two 5-input Boolean functions as shown in Fig. 4. LUT-LS of L configuration as shown in Fig. 5a performs the AND logic operation to produce a PP output using operand bits and performs its XOR logic operation with the AC bit to produce the $O6$ signal, comprising the $\beta-$MPP bit.

The functionality of controlled negation logic operation of the AC bit and its XOR logic operation with the output of NAND logic operation of the operand bits is configured for LUT-LS of M configuration as shown in Fig. 5b. Constant logic '1' signal is provided by LUT-LS of J configuration as shown in Fig. 5c. Using LUT-LS configurations of L, M and J, FPGA implementation of BW circuit for 4-bit and 8-bit multiplication operation as shown in Figs. 6 and 7 respectively are constructed. Considering $O6$ signal of LUT-LS of the BW multiplication circuit as X, TP value of this signal for every LUT-LS is evaluated using Eq. (6).

For the application of LC methodology for approximation in BW circuit, TP of LUT-LS of L, M and J configuration for 4-bit and 8-bit BW circuit are shown in Fig. 6 and Fig. 7 respectively. For a particular MPP row, all LUT-LSs with L configuration with an equal TP pose potential for the LC methodology. For every MPP_i where $0 \leq i < N-1$,

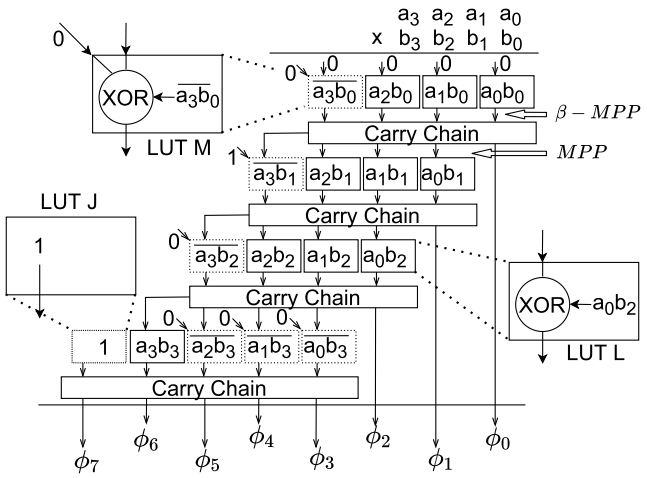

Fig. 4. Mapping of MPPs of signed multiplication circuit with the LUT-LS configurations using BW method for 4-bit operands. PP output is added using XOR operation in LUT-LS of L configuration. LUT-LS of M configuration employs MAC operation for negated PP output and controlled negation of AC bit. LUT J is utilized for supplying logic '1' signal for the most significant MPP bit of the last MPP.

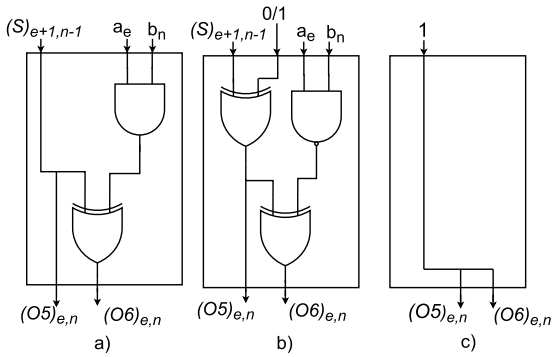

Fig. 5. LUT configuration for BW circuit with (a) L (b) M and (c) J.

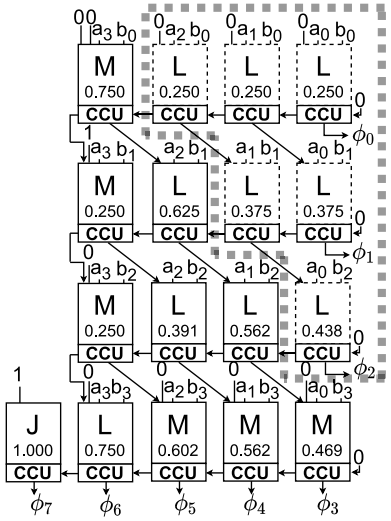

Fig. 6. Signed multiplication circuit of BW circuit for 4-bit operands. TP for $O6$ signal of LUT-LS of every LUT shown on its depiction.

$T = N - i - 1$ LUT-LSs of L configuration have equal TP with $\lambda = 0$. In version 1 of the approximate BW circuit, the LC methodology is applied by considering $T^{\#} = T$, resulting in saving of $(T-2)$ LUTs per MPP of the accurate BW circuit, while the $(T-1)$th LUT is considered as the Heavy LUT. Version 1 of the approximate BW circuit is represented as LC-BW-1, the 8-bit LC-BW-1 shown in Fig. 8a. In version 2 of the approximate BW circuit, LC methodology is applied by considering $T^{\#} = T - 1$, thereby saving $(T-3)$ LUTs per MPP of the accurate BW circuit, while the $(T-2)$th LUT is considered as the Heavy LUT. Version 2 of the approximate BW circuit is represented as LC-BW-2, the 8-bit LC-BW-2 shown in Fig. 8b.

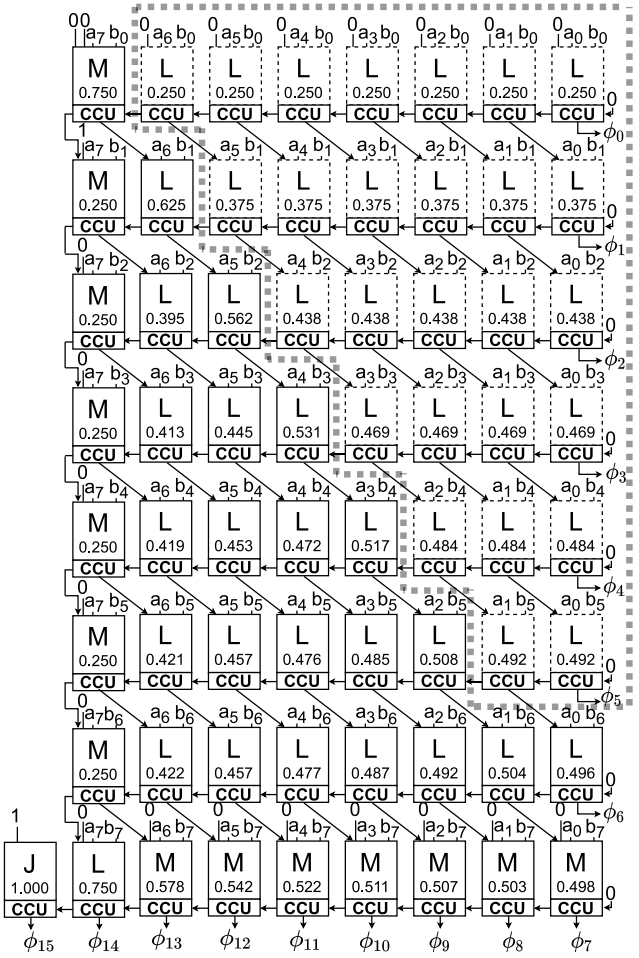

Fig. 7. BW circuit for signed multiplication with 8-bit operands.

3.2. Logic cloning for Booth circuit

In BW multiplication, N PPs of the multiplier are computed for operands with N bits. It is challenging to efficiently utilize LUT inputs for FPGA implementation of the Boolean functions involved in the BW circuit. A more efficient method of signed multiplication circuit, the Booth algorithm [17] encodes the operand multiplier bits, such that the number of PPs required for multiplication product is reduced. In Radix-4 Booth encoding [17], the operand circuit is encoded as follows:

$$[B]_2 = \sum_{i=0}^{N/2-1} [B_i']_2 2^{(2i)} \tag{13}$$

where $[B_i']_2$ for the ith bit position is represented as:

$$[B_i']_2 = -2b_{2i+1} + b_{2i} + b_{2i-1} \tag{14}$$

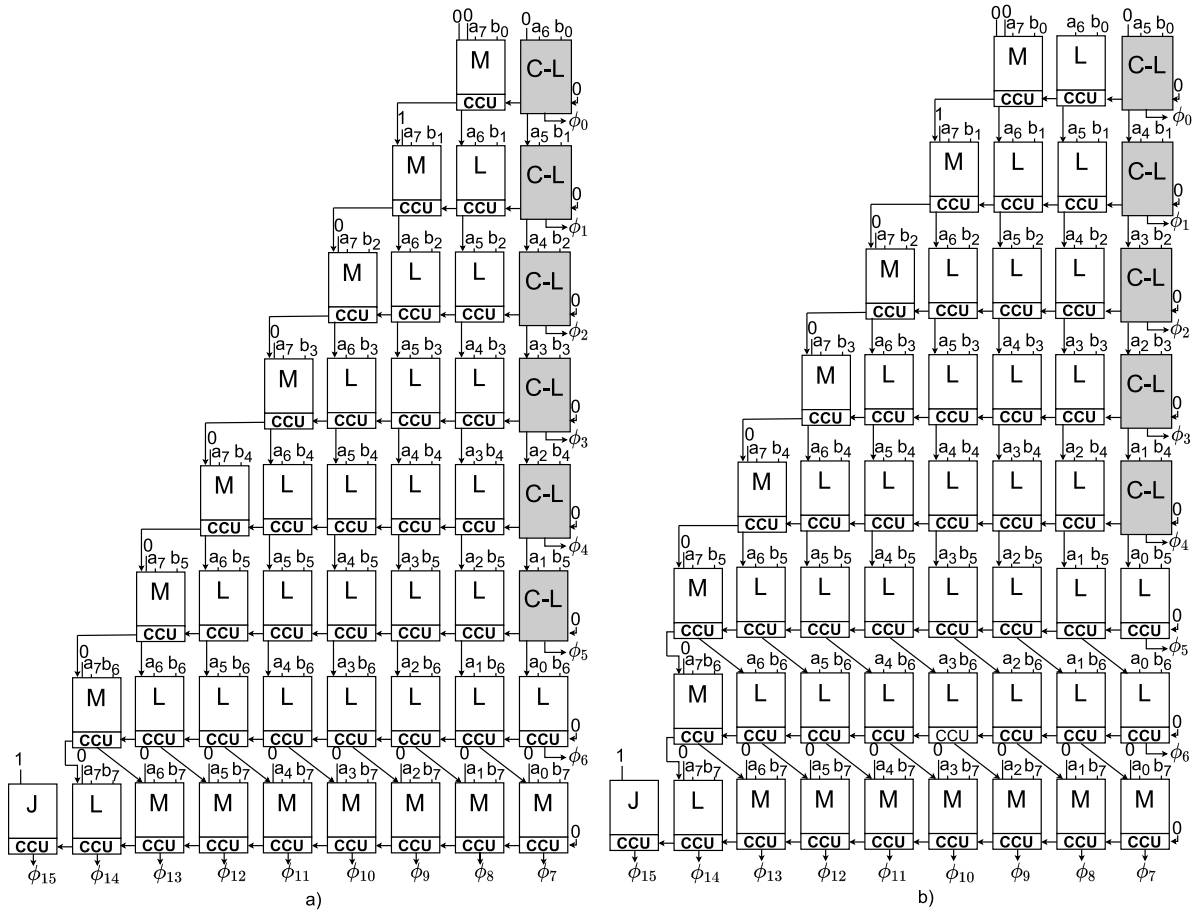

Fig. 8. LC-BW circuits for 8-bit operands (a) LC-BW-1 and (b) LC-BW-2.

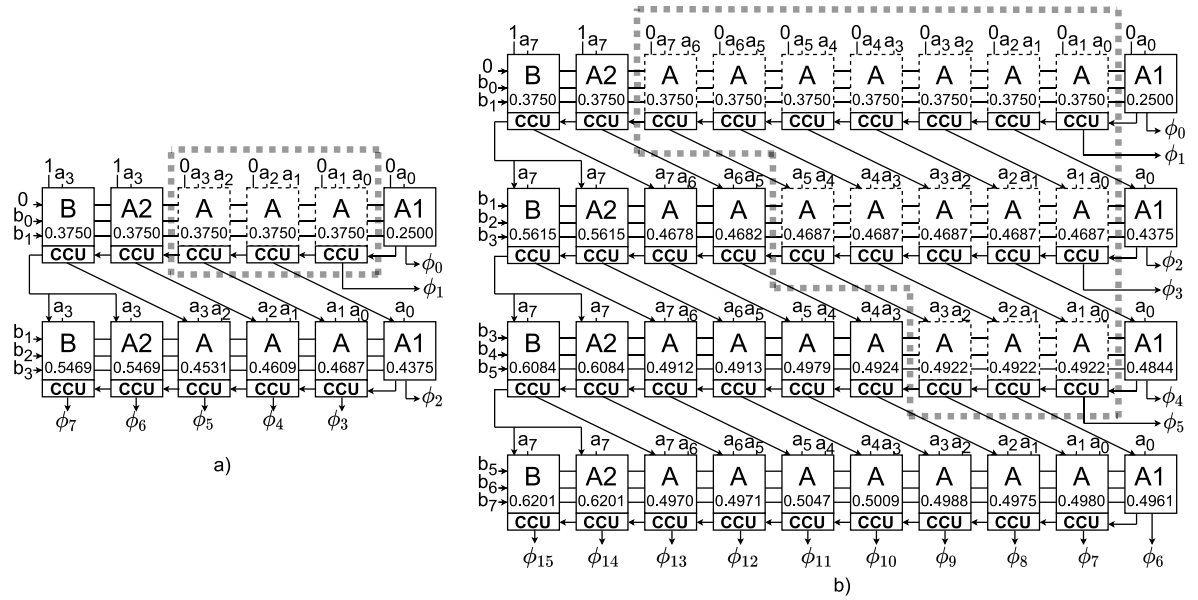

Fig. 9. Signed multiplication circuit of Booth-Opt [18] circuit with TP of $O6$ signal of LUT-LS*s* evaluated for (a) 4-bit (b) 8-bit operation.

The product is calculated as :

$$[\Phi]_2 = \sum_{i=0}^{N/2-1} [A]_2[B_i']_2 2^{(2i)} \tag{15}$$

Booth-Opt [18] is utilized as FPGA implementation of a signed multiplication circuit utilizing the Booth algorithm to demonstrate the applicability of LC methodology.

For every LUT-LS of the Booth-Opt circuit, TP value of the $O6$ signals are computed using Eq. (6) by considering $O6$ signal as X for analysis.

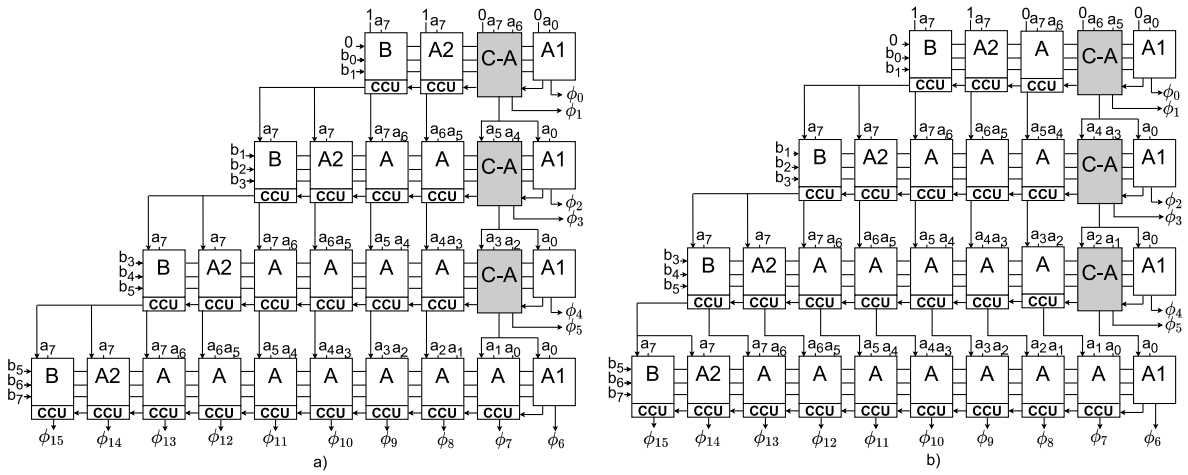

Fig. 10. LC-Booth circuits for 8-bit operands (a) LC-Booth-1 (b) LC-Booth-2.

Table 1
Evaluation of approximate circuits for multiplication product accuracy.

Multiplier	$N = 8$				$N = 16$			
	MED	MRED	NMED	Max Rel	MED	MRED	NMED	Max Rel
MUL_14_4[a] [24]	512	0.9393	0.0313	2118	7.007+e6	0.0301	652.569−e5	1.000
MUL_14_6[a] [24]	100	0.2317	0.0061	1094	2.765+e6	0.0169	257.543−e5	1.000
MUL_13_4[a] [24]	647	1.2014	0.0395	2118	7.518+e6	0.0333	700.189−e5	1.000
MUL_13_6[a] [24]	112	0.2486	0.0068	838	3.270+e6	0.0200	304.518−e5	1.000
Booth-Approx [18]	89	0.0949	0.0054	6	0.023+e6	0.0010	2.146−e5	0.097+e6
AxBM1 [25]	1229	2.2104	0.0750	72	294.880+e6	11.4910	27 470.000−e5	0.037+e6
AxBM2[b] [25]	1200	2.1893	0.0732	72	286.675+e6	11.4473	26 698.000−e5	0.037+e6
LC-BW-1	37	0.0584	0.0022	63	0.015+e6	0.0011	1.357−e5	0.016+e6
LC-BW-2	16	0.0279	0.0010	31	0.007+e6	0.0058	0.655−e5	0.008+e6
LC-Booth-1	82	0.0830	0.0050	2	0.034+e6	0.0016	3.184−e5	3.000
LC-Booth-2	37	0.0451	0.0023	2	0.017+e6	0.0009	1.543−e5	3.000

[a] 16-bit variant is built using a combination of M8s variants.

[b] For 16-bit, nine LSBs are truncated and '1' is added to the tenth bit for error compensation.

Proposed approximate multiplication circuits are shown in boldface.

TP values for LUT-LSs are evaluated for Booth-Opt circuit as shown in Fig. 9a and Fig. 9b for 4-bit and 8-bit multiplication operations respectively. The TP for $O6$ signal is constant across a chunk of LUT-LSs with A configuration for every MPP row. Precisely, for MPP_i with $i = 0, 1, \ldots N/2 - 1$ for $\lambda = 1$, about $T = N - 1 - 2i$ LUT-LSs have equal TP. For version 1 of the approximate Booth circuit, LC methodology is applied to the accurate Booth circuit by considering $T^{\#} = T$ LUT-LSs of A configuration for every MPP, thereby saving $(T - 2)$ LUTs and CCUs per MPP. The Tth LUT is considered as the Heavy LUT for each MPP. Version 1 of the approximate Booth circuit is represented as LC-Booth-1. For version 2 of the approximate Booth circuit, LC methodology is applied to the accurate Booth circuit by considering $T^{\#} = (T - 1)$ LUT-LSs of configuration A for every MPP, thereby saving $(T - 3)$ LUTs and CCUs per MPP. The $(T - 1)$th LUT of MPP is considered as the Heavy LUT. Version 2 of the approximate Booth circuit is represented as LC-Booth-2. Both versions of the 8-bit LC based circuits are shown in Fig. 10a and Fig. 10b.

4. Simulation results

Competitive circuit designs are evaluated as per the error metrics in Table 1. For signed multiplication of N bit operands, the operand range is $[-2^{N-1}, 2^{N-1} - 1]$, while the multiplication product occupies $2N$ bits. As the error metrics are dependent on the operand range of the circuit according to Eqs. (8) and (9), the error performance of the approximate multiplication circuits is also dependent on the operand range. For the 8-bit signed configuration of approximate compressor based multiplication circuits [24], PPs are generated using the BW method. For 16-bit signed configuration, four 8-bit unsigned approximate multiplication circuits (built using the approximate compressors) are utilized while the sign computation is managed by the additional 17th bit of the PP. The signal ordering of the approximate compressors affects the error performance of the circuit. Booth-Approx [18] is approximated from the FPGA implementation of the Booth-Opt. For AxBM [25], the Booth encoding of accurate Radix-8 is modified for AxBM1 and AxBM2 encoders.

Table 2
LUT resource consumption of accurate and proposed LC circuits for N bit operands (N is even).

Multiplier	#LUT
BW	$(N^2 + 1)$
BOOTH-OPT	$(N^2 + 2N)/2$
LC-BW-1	$(N^2 + 5N - 4)/2$
LC-BW-2	$(N^2 + 7N - 10)/2$
LC-Booth-1	$(N^2 + 8N - 4)/4$
LC-Booth-2	$(N^2 + 10N - 8)/4$

For the 8-bit configuration, the LC-BW circuits have the lowest MED, MRED and NMED among all the circuits, with LC-BW-2 of LC-BW being the lowest than LC-BW-1. LC-Booth circuits show the best error performance after LC-BW circuits, where LC-BW-2 performs better than LC-BW-1. However, the maximum relative error is lower for the LC-Booth circuits than that for the LC-BW circuits. Booth-Approx performs best after LC-BW and LC-Booth circuits for the MED, MRED and NMED error metrics, its maximum relative error lies in between that of LC-BW and LC-Booth. Approximate compressor based multiplication circuits

(MUL_xy_k; where x and y denote the approximate compressors used and k denotes the approximation factor) perform better after Booth-Approx, LC-BW and LC-Booth circuits. AxBM1 and AxBM2 have the lowest error performance, AxBM2 performs better than AxBM1. For circuits with 16-bit operands; MED, MRED and NMED follow a similar pattern in error performance as the circuits with 8-bit operands. The maximum relative error varies unevenly among the circuits. It decreases with an increase in bit-width for multiplication circuits using approximate compressors. However, it increases for AxBM1, AxBM2, Booth-Approx, LC-BW and LC-Booth circuits.

5. Hardware implementation

For demonstrating the hardware evaluation of approximate circuits, Xilinx Vivado 2021.1 tool is used for synthesis and implementation, with XILINX VIRTEX-7 FPGA device xc7vx690tffg1930-3. FPGA circuits are prototyped in VHSIC Hardware Description Language (VHDL). For RTL synthesis, LUT6_2 and CARRY4 primitives are used as components. Critical Path Delay (CPD) is the longest path delay in the circuit, measured between the input and output signal of the circuit. The signal propagation delay of LUT is independent of its LUT-LS configuration.

Conventional multiplication circuits are combinational in nature, hence the logic delay of the circuit is clock invariant and the circuit delay must be lesser than the clock cycle period. However, for estimating the routing delay, the design tool optimizes the critical path based on the timing constraint of the inputs. Hence, the timing constraint is adjusted to get the best possible delay value of the critical path for a particular circuit configuration. To get the precise CPD value, initially, the timing slack value is nullified at the synthesis level by calibrating the timing constraint. In the next stage, implementation is initiated with the timing constraint obtained at the synthesis level for nullifying the timing slack. At the subsequent iterations of the synthesis/implementation, the timing constraint is updated with the data path delay from the previous iteration of synthesis/implementation. This process is repeated until the design tool provides a relatively minimum value for the data path delay, to be considered as CPD.

For power calculation, simulation configuration with a supply voltage of 1 V with Linear Feet per Minute (LFM) of 250 along with a heat sink is considered. Precise dynamic power values are computed by implementing multiple instances of RTL circuit design. The logic delay time of the Xilinx IP [30] in the area-optimized mode is chosen as the clock period to evaluate the power consumption of circuits.

The LUT consumption of non-MAC circuits of AxBM and approximate compressor based multiplication circuits not only depends on the PP generation but also on the PP accumulation method. However, the MAC circuits have a comprehensive architecture with implicit PP accumulation. LUT resource consumption of accurate BW and Booth circuits and their derived accurate LC approximate circuits for N bit operands are depicted in Table 2. The efficiency of LUT resource consumption of MAC approximate multiplication circuits is shown in Fig. 11. In terms of LUT savings, LC-BW circuits show a logistic growth, while the Booth-Approx shows an exponential decay with an increase in the circuit bit-width N. LC-BW circuits achieve more LUT savings as compared to LC-Booth circuits. Also, version 1 of LC-BW and LC-Booth shows more LUT saving than version 2.

To evaluate the performance of LC methodology, circuit configurations are shown in Table 3. As compared to the accurate BW circuit, the gains in LUT consumption for LC-BW-1 are 5.88% for 4-bit, 23.07% for 8-bit and 35.40% for 16-bit configuration. LUT consumption of LC-BW-2 for 4-bit configuration is similar to that of accurate BW circuit, while the gains are 15.38% for 8-bit and 30.40% for 16-bit configuration. LUT consumption for LC-BW-1 is lower than that for LC-BW-2 for similar configurations. Similarly, as compared to the accurate Booth circuit, gains in LUT consumption for LC-Booth-1 are 8.33% for 4-bit, 32.5% for 8-bit and 34.02% for 16-bit configuration. LUT consumption of LC-Booth-2 for 4-bit configuration is equal to that of the accurate Booth

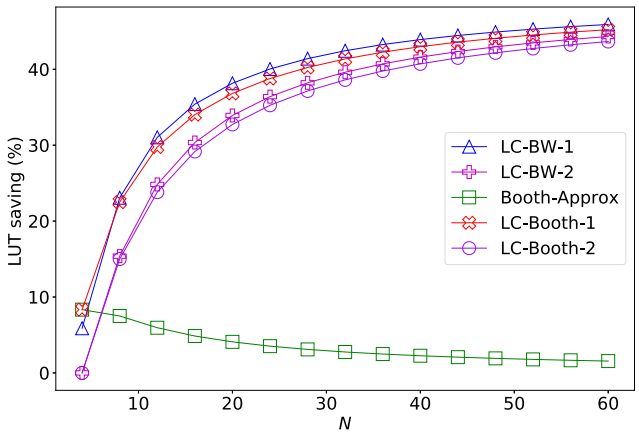

Fig. 11. LUT saving due to induced approximation of MAC signed approximate circuits.

circuit, while gains in LUT consumption are 15% for 8-bit and 29.17% for 16-bit configuration. Gains in LUT resource consumption increase with an increase in bit-width for all approximate circuits.

The Power Delay Product (PDP) of a circuit is the product of its CPD and power consumption. For 4-bit circuit configurations, in terms of PDP efficiency, LC approximate circuits have similar PDP efficiency compared to the accurate BW circuit, while LC-Booth-1 and LC-Booth-2 circuits are 20% and 3% more efficient than the accurate Booth circuit. For 8-bit circuit configuration, LC-BW-1 and LC-BW-2 show an efficiency of 23% and 4% in PDP respectively as compared to accurate BW circuit, while LC-Booth-1 and LC-Booth-2 show an average of 18% efficiency in PDP as compared to accurate Booth circuit. For 16-bit circuit configuration, LC-BW-1 and LC-BW-2 show an average of 36.5% PDP efficiency as compared to the accurate BW circuit, while LC-Booth-1 and LC-Booth-2 show an average of 29% PDP efficiency as compared to accurate Booth circuit. Efficiency in PDP increases with an increase in bit-width for all circuit configurations.

On comparing LUT consumption for competitive approximate multiplication circuits for 8-bit and 16-bit configurations, it is lowest for LC-Booth circuits, followed by Booth-Approx. AxBM1 and AxBM2 have comparable LUT consumption with LC-BW circuits, for 16-bit circuit configuration. For the 8-bit circuit configuration, the CPD of multiplication circuits with approximate compressors is higher than that of Booth-Approx and does not increase drastically with an increase in the bit-width of the operands. CPD performance of LC-circuits is comparable to multiplication circuits utilizing approximate compressors, however, the former's CPD increases more drastically as compared to the latter on increasing the bit-width. For the 16-bit circuit configuration, the low power consumption is more critical than low CPD in lowering the PDP of multiplication circuits with approximate compressors as compared to the PDP of AxBM1 and AxBM2 circuits. PDP of LC-Booth circuits lies in between that of multiplication circuits with approximate compressors and Booth-Approx. LC-BW circuits are characterized by the highest CPD, compensated by their better error performance.

Fig. 12 captures the relationship between PDP, NMED and LUT resource consumption of approximate multiplication circuits. LC-Booth circuits have the lowest LUT consumption with average PDP and low NMED. LC-BW circuits have the lowest NMED, average LUT consumption and highest PDP. LUT consumption and PDP increase while NMED decreases with the version of approximation for LC-Booth and LC-BW circuits. While having a comparable NMED with an LC-based approximate circuit, Booth-Approx has an average LUT consumption and PDP. Multiplication circuits with approximate circuits (MUL_xy_k) are characterized by the lowest PDP and high LUT consumption. AxBM1 and AxBM2 circuits have an overall average PDP and LUT consumption, however, have an extremely high NMED for 16-bit operands.

Table 3
Hardware implementation analysis of signed approximate multiplication circuits on XILINX VIRTEX-7 FPGA device. CPD in ns, Power in W and PDP in pJ.

	Multiplication circuits	$N = 4$				$N = 8$				$N = 16$			
		#LUT	CPD	Power	PDP	#LUT	CPD	Power	PDP	#LUT	CPD	Power	PDP
Accurate	Xilinx(Speed) [30]	19	1.85	0.34	0.63	73	2.76	1.35	3.73	281	3.73	3.70	13.78
	Xilinx(Area) [30]	25	2.82	0.60	1.69	86	3.41	1.95	6.66	314	5.03	5.50	27.69
	BW	17	2.71	0.23	0.63	65	5.67	1.20	6.80	257	11.64	3.55	41.31
	Booth-Sign-1 [16]	18	1.65	0.68	1.13	66	2.80	2.36	6.60	243	4.48	5.92	26.52
	Booth-Sign-2 [19]	14	2.59	0.51	1.31	54	4.37	1.66	7.26	208	5.28	5.90	31.14
	Booth-Opt [18]	12	2.31	0.26	0.60	40	4.13	1.00	4.13	144	7.96	2.75	21.90
Approximate	MUL_14_4[a] [24]	–	–	–	–	48	4.15	0.30	1.25	208	5.85	1.30	7.60
	MUL_14_6[a] [24]	–	–	–	–	59	4.65	0.43	2.01	212	5.85	1.32	7.72
	MUL_13_4[a] [24]	–	–	–	–	45	3.92	0.26	1.02	163	5.62	1.00	5.62
	MUL_13_6[a] [24]	–	–	–	–	57	4.57	0.35	1.60	176	5.63	1.04	5.86
	Booth-Approx [18]	–	–	–	–	37	3.41	1.24	4.23	137	6.88	2.78	19.13
	AxBM1 [25]	–	–	–	–	–	–	–	–	194	3.68	4.90	18.03
	AxBM2 [25]	–	–	–	–	–	–	–	–	161	3.45	4.12	14.21
	LC-BW-1	16	2.76	0.23	0.64	50	5.20	1.00	5.20	166	10.65	2.45	26.10
	LC-BW-2	17	2.72	0.23	0.63	55	5.39	1.00	5.39	179	10.84	2.45	26.55
	LC-Booth-1	11	2.08	0.23	0.48	27	4.04	0.85	3.43	95	7.95	1.90	15.10
	LC-Booth-2	12	2.23	0.26	0.58	34	3.93	0.85	3.34	102	8.22	1.90	15.61

[a] Clock frequency of 100 MHz.

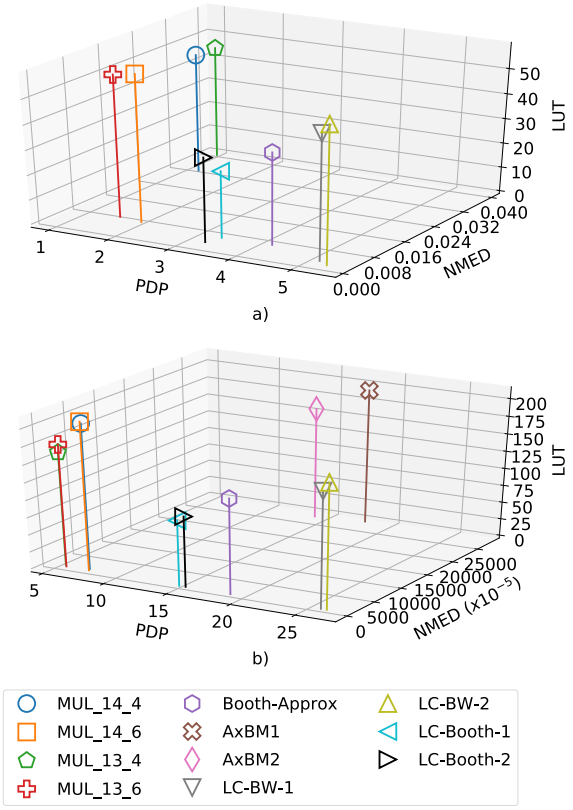

Fig. 12. Comparative analysis of PDP, NMED and LUT resource consumption for (a) 8-bit (b) 16-bit configuration. PDP is given in pJ.

6. Case study: Massive MIMO detection

Being a promising concept for future cellular networks, massive MIMO has been at the forefront in substantially improving both spectral and energy efficiencies of communication systems. A MIMO system is characterized by several Base Station (BS) antennas interacting with several User Equipment (UE)s over a wireless communication channel by spatial multiplexing [31]. The power dissipation at MIMO BS is significant when the system has to be scaled on a massive level for servicing an increasing number of UEs [32]. MIMO uplink detection with the ZF algorithm is explored to evaluate the impact of approximate signed multiplication circuits on symbol error-rate performance.

Approximate circuits are evaluated for the ZF MIMO uplink detection algorithm. For a generic BS model, a BS with B antennas servicing U number of UE is considered. Every UE has a single antenna. Accordingly, $\mathbf{y} = \mathbf{Hx} + \mathbf{n}$ represents MIMO uplink signal at BS, where $\mathbf{y} \in \mathbb{C}^{B\times 1}$ is the vector representing receive signal over B antennas of BS, $\mathbf{x} \in \mathbb{C}^{U\times 1}$ is transmit signal estimate from UEs to BS. $\mathbf{H} \in \mathbb{C}^{B\times U}$ is the channel model whose entries follow identical distribution (i.i.d). $\mathbf{n} \in \mathbb{C}^{B\times 1}$ is the channel noise. The ZF MIMO uplink signal is obtained as:

$$\hat{\mathbf{x}} = (\mathbf{H}^H\mathbf{H})^{-1}\mathbf{H}^H\mathbf{y} \tag{16}$$

Typically, the more critical case is when the number of transmit antennas is negligible to number of receive antennas i.e. $B \gg U$. In such scenarios, statistical methods of optimization are used to estimate the transmit vector $\hat{\mathbf{x}}$ from receive signal $\mathbf{y}$. Symbol error rate performance of MIMO uplink signal detector improves with increasing the number of BS antennas as the single-user channels become more decorrelated and approaches optimal uplink detection performance when the number of BS antennas is infinite [33]. However, this scenario is practically not possible to engineer due to resource constraints, signal processing complexity and energy requirements. Hence, optimization is introduced in the MIMO uplink signal detection to achieve optimal detection with bounded resource constraints.

Approximate multiplication circuits functionally implemented in the C environment are compiled into a dynamic library for accelerated simulation, which is linked to high-level Python implementation of the MIMO uplink detection, channel matrix being randomly generated for simulation. The circuit implementation in the C environment is interfaced with the Python environment using ctypes library. The 64-bit data type *unsigned long long* is used in C to handle multiplication operands and the multiplication product is interfaced with *c_int64* data type provided by ctypes library. To compute Eq. (16), matrix inverse operation is performed using the state-of-art Gauss-Jordan elimination method and accurate multiplication operations are replaced with approximate multiplication operations. To perform a multiplication operation, the operands are left bit shifted by $N-1$ and truncated to the nearest signed integer since all the operands are in the range $(-1,1)$. After the multiplication operation, the product is right bit-shifted by $2(N-1)$.

Simulation result with 5000 randomly generated symbols for ZF MIMO uplink detection using various approximate multipliers circuits is presented in Fig. 13. For the 8-bit simulation of 16-QAM and 64-QAM based MIMO detection, circuits with approximate compressors,

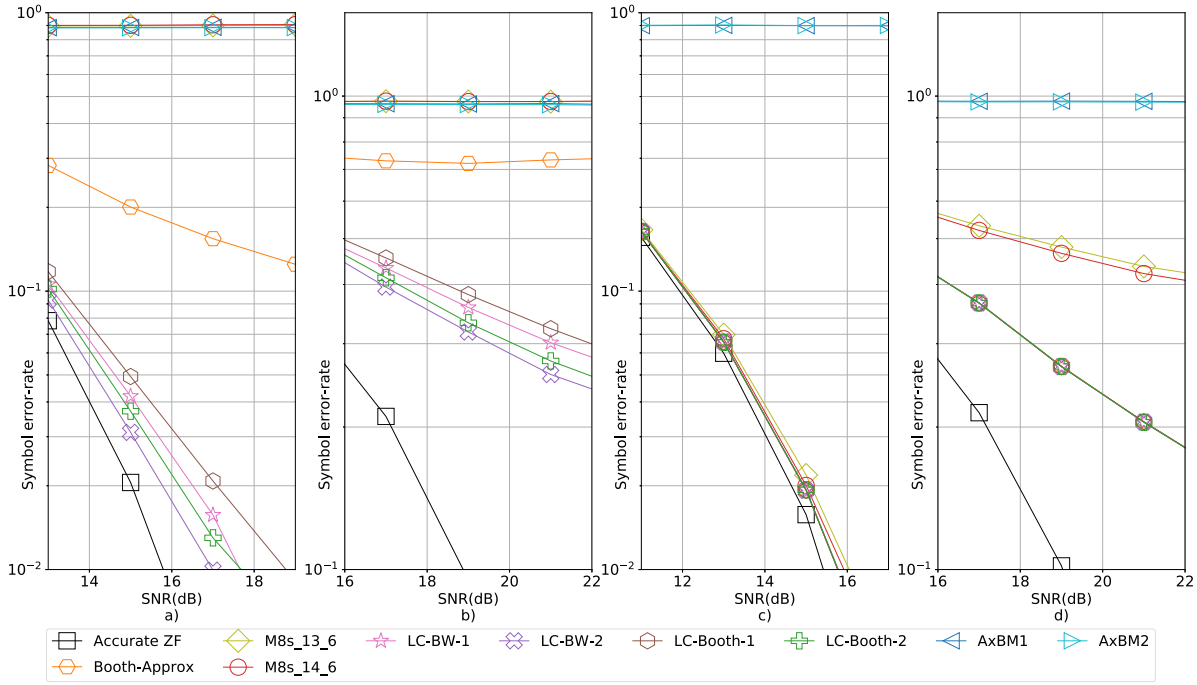

Fig. 13. Symbol error rate vs. SNR simulation for ZF MIMO uplink detection using 8-bit circuits for MIMO configuration $B = 128$, $U = 4$ given as (a) 16-QAM (b) 64-QAM and 16-bit circuits for MIMO configuration $B = 128$, $U = 4$ given as (c) 16-QAM (d) 64-QAM.

AxBM1 and AxBM1 show the worst error-rate performance. Booth-Approx diverges for higher Signal-to-Noise Ratio (SNR) for 16-QAM and 64-QAM configurations. LC based circuits converge towards a 1% symbol error-rate at a difference of less than 3 dB for 16-QAM as compared to the accurate ZF MIMO uplink detector. Symbol error-rate performance of version 2 of both LC-BW and LC-Booth is better than that of version 1 of both circuits respectively. ZF detector with LC-BW circuits provides more optimal detection than ZF detector with LC-Booth circuits. For 16-bit ZF MIMO uplink detection, detectors of all approximate circuits except AxBM1 and AxBM2, achieve a 1% symbol error-rate at an SNR difference of less than 1 dB for 16-QAM. Approximation in circuits increases inter-symbol interference which is prominently sensitive to higher QAM modulations.

ZF detection with LC-BW circuits shows close to accurate ZF detection performance, specifically at 16-QAM configuration for both 8-bit and 16-bit ZF detection, but the LC-BW circuits are characterized by relatively high PDP. ZF detection with LC-Booth circuits show close to that of ZF detection with LC-BW and the LC-Booth circuits have a relatively lower PDP than LC-BW circuits. Multiplication circuits with approximate compressors have the lowest PDP, however, their ZF detection stringently requires 16-bit operation for convergence. LC-Booth circuits are more advantageous than Booth-Approx for ZF detection symbol error rate performance as well as PDP efficiency.

7. Conclusion

LC methodology is an approximation technique employed on the accurate multiplication circuit for FPGA implementation to harness gains in energy and resource consumption by penalizing the accuracy of the multiplication product. The application of LC methodology for approximate signed multiplication circuits based on BW and Booth provides a systematic control over LUT resource consumption and PDP. LC-BW circuits show the best performance in terms of multiplication accuracy by cutting down on the PDP of the accurate BW, however, have more LUT consumption than other competitive approximate circuits. LC-Booth circuits require fewer LUTs and have lower PDP as compared to LC-BW circuits. However, since the number of MPPs is less in LC-Booth circuits, the granularity of approximation is lower for them and hence they incur more penalty in error performance as compared to LC-BW circuits. LC methodology provides explicit control over the approximation for LUT consumption, energy efficiency and accuracy for LC-BW and LC-Booth circuits, by selectively choosing the eligible LUTs of a MPP row for the particular circuit by using $T^{\#}$ parameter. Massive MIMO detection involves computationally intensive matrix-matrix and matrix–vector multiplication operations, which are fundamentally optimized by substituting accurate multiplication with approximate multiplication. For the application of approximate circuits for MIMO uplink detection, the ZF detector with LC-BW circuits achieves close to accurate ZF detection performance. It is also observed that the error performance of approximate multiplication circuits is critical for ZF detection for high QAM and low bit-width configurations. It can also be inferred that no single error metric can be used to optimally choose the approximate multiplication circuit for the ZF MIMO detection. CPD for circuits increase with bit-width and can prove a critical factor in configuring the circuit with high clock frequency. The proposed work presents a heuristic methodology for approximating arithmetic multiplication circuits for FPGA and such analysis for higher bit-widths is a topic of further research exploration with future computing capabilities.

CRediT authorship contribution statement

Abhinav Kulkarni: Conceptualization, Methodology, Investigation, Visualization, Writing – Original Draft. **Messaoud Ahmed Ouameur:** Writing – review & editing, Supervision, Resources, Project administration. **Daniel Massicotte:** Writing – review & editing, Supervision, Funding acquisition, Resources.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Daniel Massicotte reports financial support was provided by Natural Sciences and Engineering Research Council of Canada (NSERC), Prompt, Canadian Foundation for Innovation (CFI), CMC Microsystems,

Opal-RT Technologies Inc. and Hydro-Québec. The other authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The entire code employed for this work is accessible as open-source on GitHub at: https://github.com/abhinav333/Logic-Cloning.

Appendix A

A.1. Lemma

The following derivation explains that if a group of LUT-LSs of the MPP have equal TP value of $O6$ signal, then the arithmetic mean of β-MPP can be approximated without using all LUT-LSs of the MPP. Let the multiplication product $[\Phi]_2$ have N_R MPPs, each comprised of N_C bits. The nth MPP denoted as $[\Phi_n]_2$ is comprised of the nth β-MPP denoted as $\sum_{i=0}^{n-1}[\Phi_i]_2$ and the CC offset denoted as C_n. Therefore, $[\Phi_n]_2 = [C_n]_2 + \sum_{i=0}^{n-1}[\Phi_i]_2$ with $[\Phi_0]_2 = 0$ and the final product given as $[\Phi]_2 = [\Phi_{n=N_R}]_2$. ρ_i represents the power values of radix 2 such that $\rho_0 < \rho_1 < \cdots < \rho_{N_C-1}$. Hence, nth β-MPP is represented as $\sum_{i=0}^{n-1}[\Phi_i]_2 = -\varphi_{N_C-1}2^{\rho_{N_C-1}} + \sum_{i=0}^{N_C-2}\varphi_i 2^{\rho_i}$. The value of the nth $\beta-$MPP for the ζth simulation step is computed as:

$$\sum_{i=0}^{n-1}[\Phi_i^{(\zeta)}]_2 = -\varphi_{N_C-1}^{(\zeta)}2^{\rho_{N_C-1}} + \sum_{i=0}^{N_C-2}\varphi_i^{(\zeta)}2^{\rho_i} \tag{17}$$

The arithmetic mean of nth $\beta-$MPP evaluated over the exhaustive simulation range:

$$\mathcal{A}\left(\sum_{i=0}^{n-1}[\Phi_i]_2\right) = -\mathcal{A}(\varphi_{N_C-1})2^{\rho_{N_C-1}} + \sum_{i=0}^{N_C-2}\mathcal{A}(\varphi_i)2^{\rho_i} \tag{18}$$

From Eq. (6), as $\mathcal{T}(\varphi_\lambda) = \cdots = \mathcal{T}(\varphi_{\lambda+T^\#-1}) \implies \mathcal{A}(\varphi_\lambda) = \cdots = \mathcal{A}(\varphi_{\lambda+T^\#-1})$

Hence, Eq. (18) is transformed as:

$$\mathcal{A}\left(\sum_{i=0}^{n-1}[\Phi_i]_2\right) = -\mathcal{A}(\varphi_{N_C-1})2^{\rho_{N_C-1}} + \sum_{i=\lambda+T^\#}^{N_C-2}\mathcal{A}(\varphi_i)2^{\rho_i} \cdots + \underbrace{\mathcal{A}(\varphi_{\lambda+T^\#-1})\sum_{i=\lambda}^{\lambda+T^\#-1}2^{\rho_i}}_{\text{critical term}} + \sum_{i=0}^{\lambda-1}\mathcal{A}(\varphi_i)2^{\rho_i} \tag{19}$$

The critical term in Eq. (19) is approximated as:

$$\mathcal{A}(\varphi_{\lambda+T^\#-1})\sum_{i=\lambda}^{\lambda+T^\#-1}2^{\rho_i} \approx \mathcal{A}(\varphi_{\lambda+T^\#-1})2^{\rho_{\lambda+T^\#-1}} \tag{20}$$

Using Eq. (20) in Eq. (19):

$$\mathcal{A}\left(\sum_{i=0}^{n-1}[\widehat{\Phi}_i]_2\right) = -\mathcal{A}(\varphi_{N_C-1})2^{\rho_{N_C-1}} + \sum_{i=\lambda+T^\#}^{N_C-2}\mathcal{A}(\varphi_i)2^{\rho_i} \cdots + \underbrace{\mathcal{A}(\varphi_{\lambda+T^\#-1})2^{\rho_{\lambda+T^\#-1}}}_{\text{approximated term}} + \sum_{i=0}^{\lambda-1}\mathcal{A}(\varphi_i)2^{\rho_i} \tag{21}$$

Thus for every nth MPP, by using the approximation in Eq. (20), an error is generated in the $\beta-$MPP value with arithmetic mean given as:

$$\mathcal{A}\left(\sum_{i=0}^{n-1}[\Phi_i]_2\right) - \mathcal{A}\left(\sum_{i=0}^{n-1}[\widehat{\Phi}_i]_2\right) = \mathcal{A}(\varphi_{\lambda+T^\#-1})\sum_{i=\lambda}^{\lambda+T^\#-2}2^{\rho_i} \tag{22}$$

The arithmetic mean error in Eq. (22) is accumulated at every $\beta-$MPP till the final multiplication product value is computed, provided $\mathcal{T}(\varphi_\lambda) = \cdots = \mathcal{T}(\varphi_{\lambda+T^\#-1})$ for every MPP.

References

[1] J. Han, M. Orshansky, Approximate computing: an emerging paradigm for energy-efficient design, in: 2013 18th IEEE European Test Symposium, ETS, 2013, pp. 1–6, http://dx.doi.org/10.1109/ETS.2013.6569370.

[2] H. Jiang, F.J.H. Santiago, H. Mo, L. Liu, J. Han, Approximate Arithmetic Circuits: A Survey, Characterization, and Recent Applications, Proc. IEEE 108 (12) (2020) 2108–2135, http://dx.doi.org/10.1109/JPROC.2020.3006451.

[3] V.K. Chippa, S.T. Chakradhar, K. Roy, A. Raghunathan, Analysis and characterization of inherent application resilience for approximate computing, in: 2013 50th ACM/EDAC/IEEE Design Automation Conference, DAC, 2013, pp. 1–9, http://dx.doi.org/10.1145/2463209.2488873.

[4] A. Lingamneni, C. Enz, K. Palem, C. Piguet, Synthesizing Parsimonious Inexact Circuits through Probabilistic Design Techniques, ACM Trans. Embed. Comput. Syst. 12 (2s) (2013) http://dx.doi.org/10.1145/2465787.2465795.

[5] W. Qian, M.D. Riedel, H. Zhou, J. Bruck, Transforming Probabilities With Combinational Logic, IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. 30 (9) (2011) 1279–1292, http://dx.doi.org/10.1109/TCAD.2011.2144630.

[6] L.N. Chakrapani, K.V. Palem, A Probabilistic Boolean Logic and Its Meaning, Rice University, Department of Computer Science Technical Report, 2008.

[7] D. Gardy, Random boolean expressions, in: Discrete Mathematics and Theoretical Computer Science, Discrete Mathematics and Theoretical Computer Science, 2005, pp. 1–36.

[8] I. Kuon, J. Rose, Measuring the Gap Between FPGAs and ASICs, IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. 26 (2) (2007) 203–215, http://dx.doi.org/10.1109/TCAD.2006.884574.

[9] S. Mittal, A Survey of Techniques for Approximate Computing, ACM Comput. Surv. 48 (4) (2016) http://dx.doi.org/10.1145/2893356.

[10] C.S. Wallace, A Suggestion for a Fast Multiplier, IEEE Trans. Electron. Comput. EC-13 (1) (1964) 14–17, http://dx.doi.org/10.1109/PGEC.1964.263830.

[11] L. Dadda, Some schemes for parallel multipliers, Alta Freq. 34 (1965) 349–356.

[12] H. Parandeh-Afshar, P. Ienne, Measuring and Reducing the Performance Gap between Embedded and Soft Multipliers on FPGAs, in: 2011 21st International Conference on Field Programmable Logic and Applications, 2011, pp. 225–231, http://dx.doi.org/10.1109/FPL.2011.48.

[13] Intel-Altera, Stratix III Device Handbook, volume I, 2.2, 2011.

[14] M. Kumm, S. Abbas, P. Zipf, An Efficient Softcore Multiplier Architecture for Xilinx FPGAs, in: 2015 IEEE 22nd Symposium on Computer Arithmetic, 2015, pp. 18–25, http://dx.doi.org/10.1109/ARITH.2015.17.

[15] J. Hu, W. Qian, A new approximate adder with low relative error and correct sign calculation, in: 2015 Design, Automation Test in Europe Conference Exhibition, DATE, 2015, pp. 1449–1454.

[16] S. Ullah, T.D.A. Nguyen, A. Kumar, Energy-Efficient Low-Latency Signed Multiplier for FPGA-Based Hardware Accelerators, IEEE Embed. Syst. Lett. 13 (2) (2021) 41–44, http://dx.doi.org/10.1109/LES.2020.2995053.

[17] G.W. Bewick, Fast Multiplication: Algorithms and Implementation (Ph.D. thesis), Stanford University, 1994.

[18] S. Ullah, H. Schmidl, S.S. Sahoo, S. Rehman, A. Kumar, Area-Optimized Accurate and Approximate Softcore Signed Multiplier Architectures, IEEE Trans. Comput. 70 (3) (2021) 384–392, http://dx.doi.org/10.1109/TC.2020.2988404.

[19] S. Ullah, S. Rehman, M. Shafique, A. Kumar, High-Performance Accurate and Approximate Multipliers for FPGA-Based Hardware Accelerators, IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. 41 (2) (2022) 211–224, http://dx.doi.org/10.1109/TCAD.2021.3056337.

[20] B.S. Prabakaran, S. Rehman, M.A. Hanif, S. Ullah, G. Mazaheri, A. Kumar, M. Shafique, DeMAS: An efficient design methodology for building approximate adders for FPGA-based systems, in: 2018 Design, Automation Test in Europe Conference Exhibition, DATE, 2018, pp. 917–920, http://dx.doi.org/10.23919/DATE.2018.8342140.

[21] S. Boroumand, H.P. Afshar, P. Brisk, Approximate quaternary addition with the fast carry chains of FPGAs, in: 2018 Design, Automation Test in Europe Conference Exhibition, DATE, 2018, pp. 577–580, http://dx.doi.org/10.23919/DATE.2018.8342073.

[22] S. Balasubramani, U. Jagadeeshan, U. Krishnamoorthy, Performance optimized approximate multiplier architecture ST-AxM - based on statistical analysis and static compensation, Microelectron. Reliab. 151 (2023) 115277, http://dx.doi.org/10.1016/j.microrel.2023.115277.

[23] S. Venkatachalam, S.-B. Ko, Design of Power and Area Efficient Approximate Multipliers, IEEE Trans. Very Large Scale Integr. (VLSI) Syst. 25 (5) (2017) 1782–1786, http://dx.doi.org/10.1109/TVLSI.2016.2643639.

[24] N. Van Toan, J.-G. Lee, FPGA-based multi-level approximate multipliers for high-performance error-resilient applications, IEEE Access 8 (2020) 25481–25497.

[25] H. Waris, C. Wang, W. Liu, F. Lombardi, AxBMs: Approximate radix-8 Booth Multipliers for High-Performance FPGA-Based Accelerators, IEEE Trans. Circuits Syst. II 68 (5) (2021) 1566–1570, http://dx.doi.org/10.1109/TCSII.2021.3065333.

[26] Xilinx, 7 Series FPGAs Configurable Logic Block, User Guide, UG474 (v1.8) , 2016.

[27] Xilinx, Vivado Design Suite 7 Series FPGA and Zynq-7000 SoC, Libraries Guide, UG953 (v2021.2), 2021.

[28] D.V. Hall, Digital Circuits and Systems, Glencoe/McGraw-Hill School Publishing Company, 1989.

Chapter 5 - Systematic analysis of impact of approximate multiplication on wireless signal detection.

5.1 Résumé Long

5.1.1 Contexte de la Recherche

La QoS dans la réception de signaux sans fil englobe un ensemble de composants qui définissent et gèrent collectivement le niveau de QoS. Des métriques telles que le BER, le SER et le Frame Error Rate (FER) sont utilisées pour mesurer la QoS, influençant directement la fidélité des signaux reçus. Surveiller et contrôler ces facteurs est crucial pour maintenir la fiabilité et la stabilité de la détection des signaux, en particulier dans les applications sensibles à la précision. La prise de conscience de la consommation d'énergie au niveau algorithmique pour diverses applications sans fil est analysée comme un facteur significatif pour l'évaluation de la QoS. Étant donné le rôle critique des opérations de multiplication dans ces systèmes, l'application de techniques de calcul approximatif aux opérations de multiplication peut entraîner des améliorations au niveau du système en termes d'efficacité énergétique (EE) et de consommation de ressources.

Les unités fonctionnelles arithmétiques, telles que celles effectuant des multiplications, fonctionnent intrinsèquement comme des systèmes non linéaires, et l'introduction d'approximations peut créer des irrégularités dans ces systèmes. Le défi réside dans le fait que les métriques statistiques actuelles ne capturent souvent pas complètement ces irrégularités causées par l'approximation, ce qui freine l'adoption généralisée des circuits arithmétiques approximatifs en raison des préoccupations concernant la fiabilité au niveau du système. Pour améliorer la fiabilité du système et prendre des décisions éclairées sur les techniques d'approximation, il est crucial de mesurer avec précision les irrégularités

introduites par l'approximation dans les systèmes de calcul. De plus, l'utilisation du calcul approximatif pour la multiplication dans la détection de signaux peut réduire la fidélité du signal lorsque la résilience aux erreurs du système est dépassée, entraînant une diminution de la fiabilité et de la stabilité du système.

5.1.2 Méthodologie

Ce travail analyse systématiquement l'impact de l'utilisation de la multiplication approximative pour la détection de signaux sans fil. L'introduction de la multiplication approximative introduit des irrégularités dans la détection, qui sont modélisées à l'aide d'un modèle de bruit constant. Ce modèle est évalué pour le TM utilisé dans la détection QPSK MMSE. La fidélité du signal, évaluée en termes de SER, est analysée à l'aide de ce modèle. Une analyse d'approximation évalue les bénéfices obtenus en tronquant le circuit de multiplication précis.

Le travail introduit des métriques de résilience conçues pour évaluer l'efficacité du TM à réaliser des gains d'approximation significatifs avec un minimum de SER. Ces métriques sont formulées sur la base de la fidélité du signal et des gains d'approximation. L'analyse systématique englobe la fidélité du signal, l'approximation et la résilience. Toutes les simulations nécessaires au projet sont réalisées en Python, en utilisant la bibliothèque NumPy pour les calculs numériques et Matplotlib pour la visualisation des données.

5.1.3 Synthèse Complète

La relation entre la fidélité du signal et les conditions de canal indique que, pour un gain de canal fixe, une augmentation du SNR entraîne un élargissement de l'intervalle de bornes. À mesure que le gain du canal augmente, la borne inférieure du SER diminue ; cependant, l'intervalle de bornes global s'élargit considérablement. Ainsi, une

approximation accrue entraîne une plus grande dégradation du SER lorsque les gains du canal augmentent. Lorsque le niveau d'approximation augmente, le système démontre une meilleure résilience à des gains de canal élevés pour un SNR constant. Les niveaux d'approximation inférieurs maintiennent leur résilience à mesure que le SNR augmente pour un gain de canal fixe.

Le gain d'approximation et l'EE s'améliorent avec un niveau d'approximation plus élevé pour toutes les valeurs de largeur de bit. L'analyse de Normalized Resiliency Ratio (NRR) montre que les niveaux d'approximation inférieurs présentent une résilience plus faible à faible SNR mais s'améliorent à un SNR élevé. En revanche, les niveaux d'approximation plus élevés montrent plus de résilience dans des scénarios à faible SNR. À mesure que le SNR et le gain de canal augmentent, l'étendue des creux de résilience augmente également. Dans des conditions de faible SNR, NRR affiche des creux minimaux, qui deviennent plus prononcés avec l'augmentation du gain de canal. Les niveaux d'approximation plus élevés sont avantageux à faible SNR, tandis que les niveaux d'approximation inférieurs ont tendance à mieux performer à SNR élevé. À mesure que la largeur de bit des opérandes augmente, le taux d'augmentation de Average Normalized Resiliency Ratio (ANRR) ralentit. Les creux de résilience dans ANRR deviennent plus prononcés et les creux de résilience diminuent avec une largeur de bit plus élevée. L'analyse systématique a été résumée dans le Tableau 5-2.

Grâce à une analyse complète au niveau du système englobant la fidélité du signal, l'approximation et la résilience, la recherche vise à établir des lignes directrices concernant les implications de l'adoption des opérations de multiplication approximatives pour la détection de signaux sans fil. Globalement, ce travail représente un effort pionnier dans son domaine, établissant un cadre fondamental pour les recherches futures visant à relier

Table 5-1 Analyse Systématique.

Analyse	Métrique	Description
Fidélité du Signal	*SER*	L'intervalle de bornes augmente avec le SNR et le gain du canal, indiquant une dégradation accrue du SER.
Approximation	Gain d'Approximation	Le gain d'approximation augmente avec M pour tout N, tandis que le taux d'augmentation diminue avec N.
Approximation	Efficacité Énergétique	L'efficacité énergétique augmente avec M pour tout N, tandis que le taux d'augmentation diminue avec N.
Résilience	NRR	Un NRR élevé est atteint par un faible M dans un régime de SNR élevé et un M élevé dans un régime de SNR faible.
Résilience	ANRR	Le taux d'augmentation de ANRR diminue avec N et les creux de résilience augmentent avec N.

le niveau d'approximation à la QoS dans la détection des signaux.

5.1.4 Droits d'Auteur

5.2 Long Abstract

5.2.1 Research Context

QoS in wireless signal reception encompasses a set of components that collectively define and manage the level of QoS. Metrics such as BER, SER, and FER are employed to gauge QoS, directly influencing the fidelity of received signals. Monitoring and controlling these factors are critical for maintaining the reliability and stability of signal detection, particularly in applications that are accuracy sensitive. Energy consumption awareness at the algorithmic level for various wireless applications is being analyzed as a significant factor for QoS assessment. Given the critical role of multiplication operations in these systems, the application of approximation computing techniques to multiplication operations can yield system-level improvements in EE and resource consumption.

Arithmetic functional units, such as those performing multiplication, inherently operate as nonlinear systems, and introducing approximations can create irregularities within these systems. The challenge is that current statistical metrics often fail to fully capture these irregularities caused by approximation, which impedes the widespread adoption of approximate arithmetic circuits due to concerns about system-level reliability. To improve system reliability and make informed decisions about approximation techniques, it is crucial to accurately measure the irregularities introduced by approximation in computing systems. Also, using approximate computing for multiplication in signal detection can reduce signal fidelity when the system's error resilience is exceeded, resulting in decreased system reliability and stability.

5.2.2 Methodology

This work systematically analyzes the impact of employing approximate multiplication for wireless signal detection. The introduction of approximate multiplication introduces irregularities in detection, which are modeled using a constant noise model. The constant noise model is evaluated for TM used in QPSK MMSE detection. Signal fidelity, assessed in terms of SER, is analyzed using this model. An approximation analysis evaluates the benefits gained from truncating the accurate multiplication circuit.

The work introduces resilience metrics designed to evaluate the effectiveness of TM in achieving significant approximation gains with minimal SER. These metrics are formulated based on signal fidelity and approximation gains. The systematic analysis encompasses signal fidelity, approximation, and resiliency. All simulations required for the project are performed in Python, using the NumPy library for numerical computations and Matplotlib for data visualization.

5.2.3 Comprehensive Synthesis

The relationship between signal fidelity and channel conditions indicates that, for a fixed channel gain, an increase in SNR leads to a wider bound interval. As the channel gain increases, the lower bound of the SER decreases; however, the overall bound interval expands significantly. Hence, greater approximation results in greater SER degradation when channel gains are increased. When approximation level is increased, the system demonstrates improved resiliency at higher channel gains for a constant SNR. Lower approximation level maintain their resiliency as SNR increases for a fixed channel gain.

Approximation gain and EE improve with higher approximation level across all values of bit-width. The NRR analysis shows that lower approximation levels exhibit lower resilience in low SNR but improves at higher SNR level. Conversely, higher approximation level demonstrate more resilience in low SNR scenarios. As SNR and channel gain increase, the extent of resiliency dips also expands. In lower SNR conditions, NRR displays minimal dips, which become more pronounced with increased channel gain. Higher approximation levels are advantageous in low SNR, while lower approximation levels tend to perform better under high SNR. As bit-width of operands increase, the rate of increase in ANRR slows down. Resiliency dips in ANRR become more pronounced and resiliency trough decreases with higher bit-width. The systematic analysis has been summarized in Table 5-2.

Through a comprehensive system-level analysis encompassing signal fidelity, approximation, and resiliency, the research aims to establish guidelines regarding the implications of adopting approximate multiplication operations for wireless signal detection. Overall, this work represents a pioneering effort in its domain, establishing a foundational framework for future research aimed at linking the level of approximation

Table 5-2 Systematic Analysis.

Analysis	**Metric**	**Description**
Signal Fidelity	*SER*	The bound interval increases with both the SNR and channel gain, indicating greater degradation in SER.
Approximation	Approximation Gain	Approximation gain increases with M for all N, while the rate of increase decreases with N.
Approximation	Energy Efficiency	Energy efficiency increases with M for all N, while the rate of increase decreases with N.
Resiliency	NRR	High NRR is achieved by low M in high SNR regime and high M in low SNR regime.
Resiliency	ANRR	Rate of increase in ANRR decreases with N and resiliency dips increase with N.

with the QOS in signal detection.

5.2.4 Copyright

electronics

MDPI

Article

Energy Efficient Wireless Signal Detection: A Revisit through the Lens of Approximate Computing

Abhinav Kulkarni *, Messaoud Ahmed Ouameur and Daniel Massicotte *

Electrical and Computer Engineering Department, Université du Québec à Trois-Rivières, Trois-Rivières, QC G9A 5H7, Canada; messaoud.ahmed.ouameur@uqtr.ca
* Correspondence: abhinav.kulkarni@uqtr.ca (A.K.); daniel.massicotte@uqtr.ca (D.M.)

Abstract: In the pursuit of energy efficiency in next-generation communication systems, approximate computing is emerging as a promising technique. In the proposed work, efforts are made to address the challenge of bridging the gap between the level of approximation and the Quality-of-Service (QoS) of the system. The application of approximate multiplication to wireless signal detection is explored systematically, illustrated by employing Truncated Multiplication (TM) on Quadrature Phase Shift Keying (QPSK) Minimum Mean Square Error (MMSE) detection. The irregularities induced by approximation in the multiplication operation employed in wireless signal detection are captured by the Approximate Multiplication Noise (AMN) model, which aids in the analysis of signal fidelity and resiliency of the system. The energy efficiency gains through approximation are highlighted in the approximation analysis. Signal fidelity analysis provides the capability to predict system output for varying levels of approximation, which aids in improving the stability of the system. The higher approximation levels are advantageous in low Signal-to-Noise Ratio (SNR) regimes, whereas lower approximation levels prove beneficial in high SNR regimes.

Keywords: wireless signal detection; approximate computing; energy efficiency; arithmetic multiplication; noise; resiliency

Citation: Kulkarni, A.; Ouameur, M.A.; Massicotte, D. Energy Efficient Wireless Signal Detection: A Revisit through the Lens of Approximate Computing. *Electronics* **2024**, *13*, 1274. https://doi.org/10.3390/electronics13071274

Academic Editor: Eneko Iradier

Received: 28 February 2024
Revised: 21 March 2024
Accepted: 26 March 2024
Published: 29 March 2024

1. Introduction.

The explosive growth of mobile traffic, driven by the emergence of new services and applications [1] is propelling the development of next-generation communication systems. This surge in mobile traffic necessitates an increase in the communication system capability [2]. Crucial enabling technologies for next-generation communication systems operate at the spectrum-level, protocol-level, and infrastructure-level [3]. On the spectrum level, efforts are concentrated on increasing carrier frequencies, while at the protocol level, gains are projected through adjustments in data packet packaging. Infrastructure-level advancements in hardware technology play a pivotal role in enabling next-generation communication systems, where energy consumption is a key design factor impacting the scalability of the system [4].

The work [5] emphasizes the increasing importance of energy efficiency in system design, aligning it with considerations of spectral efficiency and spatial reuse. In wireless communication systems, energy consumption at Base Station (BS) and core networks became a noteworthy concern due to the imperative for extensive coverage, heightened BS density in populated areas, and management of data. Remarkably, a substantial portion, ranging from 60% to 80%, of the overall energy consumption is attributed solely to BS [6]. This energy consumption in active mode at the physical layer of BS is linked to baseband processing, RF processing, and signal power amplification. The average baseband energy consumption for a 4G LTE and 5G NR BS is approximately 150 and 220 Watts, respectively, constituting up to 5–15% of the total BS energy consumption [7]. Moreover, the energy efficiency of RF and power amplification is influenced by the efficiency of baseband processing [8].

The work [9] underscores the importance of integrating energy-efficient techniques into the design of baseband processing, addressing both operational costs and environmental concerns related to BS operation. Enhancing the energy efficiency of baseband processing, which plays a crucial role in BS operation, has a cascading effect on the overall energy consumption of the BS. Likewise, since wireless signal detection is fundamental to baseband signal processing, enhancing the energy efficiency of the former leads to an improvement in the energy efficiency of the latter. Going further, a system-wide perspective needs to include a real-time assessment and control of the energy consumption for wireless signal detection to cater to the incorporation of future technologies.

The following contributions are presented in this work:

- Modeling of a constant noise model for wireless signal detection to evaluate the impact of irregularities caused by approximate multiplication. AMN is a novel constant noise model that effectively captures irregularities of TM for QPSK MMSE signal detection.
- Gauging the effect of TM on Symbol Error Rate (SER) of QPSK MMSE signal detection. The derived analytical expression computes SER by using AMN.
- Proposition of resiliency metrics to provide insights into resilient TM configurations for QPSK MMSE signal detection. A TM configuration characterized by a low level of approximation proves advantageous in high SNR regimes, whereas one with a high level of approximation is preferable in low SNR regimes.

In Section 1, the motivation for the work is established, and the contributions of the proposed work are laid out. Section 2 outlines the related work. In Section 3, the methodology is detailed, utilizing components to derive entities and primary metrics for the proposed work. Section 4 presents the derivation of secondary metrics and analysis. Finally, in Section 5, the conclusion of the proposed work is presented.

2. Related Work

QoS in wireless communication encompasses a set of components that collectively define and manage the level of service quality. Bit Error Rate (BER), SER, and Frame Error Rate (FER) are key factors in ensuring the QoS, directly impacting received signal fidelity [10]. Monitoring and controlling these factors is crucial for maintaining the reliability and stability of communication systems, especially in accuracy-sensitive applications. Several efforts are made in integrating energy efficiency with QoS. Energy awareness has been induced at the algorithmic level for Internet-of-Things (IoT) application and analyzed as a QoS factor in the work [11]. The work [12] highlights the integrated approach of QoS parameters including energy efficiency and their influence on the dynamic network condition and mobility of wireless sensor nodes. The work [13] explores the development of an optimal radio resource allocation method in 5G LTE networks based on adaptive selection of channel bandwidth depending on the QoS requirements.

As there has been a recent outburst to employ Artificial Intelligence (AI) techniques for intelligent automation, this approach has also been explored to achieve energy efficiency in communication systems [14,15]. However, the direct application of this approach at the physical layer of communication systems can incur additional computational overheads [16] related to ancillary data processing, which could potentially negate the energy efficiency benefits obtained at the foremost. Moreover, baseband processing of the physical layer is also being incorporated into resource-constrained edge computing devices, where the power budget is very stringent, but the performance constraints might be relaxed. The work [17] underscores the challenges of meeting the commercial budget requirements of power consumption of communication systems designed for THz frequency, which limits their operating frequency.

Approximate computing intentionally introduces errors into systems to enhance energy and resource consumption efficiency, albeit at the expense of tunable accuracy loss [18,19]. This methodology relies on the error-resilient nature of applications and employs disciplined methods for inserting errors into the system. The approach spans hardware, software, and cross-layer methodologies across diverse application domains to

achieve efficiency improvements [20]. The work [21] delves into approximate computing techniques, exploring security issues and analyzing the impact of application level on neural network processing, as well as image, speech, and baseband signal processing. The work [22] specifically investigates the analysis of the impact of approximate computing on application quality through a three-step process involving error characterization, error propagation, and linking errors with the quality metric of the application. The work [23] explores the integration of approximation techniques with conventional computing tasks to enhance the efficient utilization of computational infrastructure. By characterizing a library of approximated operators, the work [24] proposes a Bayesian model for predicting error propagation. Their work demonstrates improved accuracy evaluations and computational efficiency, positioning it as a valuable tool for design space exploration in approximate computing.

In the context of next-generation communication systems, which prioritize flexible performance targets for enhanced energy efficiency [5,25], the use of approximate computing techniques becomes relevant, which is another approach for improving energy efficiency. These techniques allow a controlled trade-off in system performance, contributing to improved efficiency in communication systems. A comprehensive survey [26] on the potential of approximate computing techniques for existing and future B5G communication highlights SER as a crucial Key Performance Indicator (KPI) for channel-related problems while energy efficiency as a prime KPI for resource allocation. By employing approximate computing techniques within a fixed power budget for communication systems, it becomes possible to reduce overall system power consumption. This reduction in power consumption consequently frees up additional capacity for system scaling within the same power budget.

Recent advancements in approximate computing techniques for communication systems are summarized in Table 1. The decoder based on the Successive Cancellation (SC) algorithm enhances the Forward Error Correction (FEC) performance of polar codes; however, it limits the throughput of its hardware implementations. To tackle this challenge, configurable approximation units are introduced in optimized computation function blocks used in the SC algorithm to improve the throughput of the decoder in the work [27]. The work in [28] harnesses the error-resilient nature of the inherent Fast Fourier Transform (FFT) operation in the industrial wireless communication system to demonstrate the potential of approximate computing. The exact add/subtract operators in the butterfly structure of the FFT are replaced with approximate adders, and the impact of the modified FFT operations is analyzed at the system level. Additionally, the work emphasizes the challenges related to system reliability and suitable error metrics, stressing the need to establish a connection between the characteristics of approximate adders and system performance. The work in [29] explores the application of approximate computing in the expectation propagation algorithm used for the Sparse Code Multiple Access (SCMA) receiver. By employing approximation techniques, the complexity of the algorithm is reduced, which is characterized by the number of arithmetic operations. Approximations are incorporated into the expectation propagation algorithm at the variable and function node updates, as well as the log-likelihood ratio calculation, to decrease algorithmic complexity. Moreover, parameter optimizations are proposed to strike a balance between detection performance and algorithm complexity. In the work [30], exact computing units are substituted with approximate ones in Root Raised Cosine (RRC) Finite Impulse Response (FIR) filters used for pulse shaping at the BS and decoders/equalizers at the User Equipment (UE) in Single Input Single Output (SISO) and Multiple Input Multiple Output (MIMO) 6G downlink operations. The BER performance of the proposed approximate computing-empowered 6G SISO downlink is superior to its MIMO counterpart, where the induced approximations achieve substantial power savings. The BER performance degradation is more pronounced in the high SNR regime compared to the low/medium SNR regime. The work in [31] exploits gradient bounds to propose a novel encoding scheme for Quadrature Amplitude Modulation (QAM) mapping in the communication system required for a

federated learning model. The results highlight the significance of quantifying the effects of approximation on the overall application. In a fixed SNR scenario, the test accuracy of the model deteriorates as the QAM modulation order increases.

Table 1. Approximate computing techniques for communication systems.

Work	Description	Approximation	Modulation	QoS
Zhou (2018) [27]	Throughput improvement by utilizing approximate computation blocks for decoding FEC polar codes.	SC decoder.	-	FER
Hao (2019) [28]	Reliability assessment on utilizing approximate adders for industrial wireless communication.	FFT.	QPSK	FER
Xiao (2019) [29]	Complexity reduction of expectation propagation algorithm utilized for SCMA.	Variable and functional node update, log likelihood ratio computation.	-	BER
Idrees (2021) [30]	Gains by utilizing approximate computing units in digital signal processing filters in 6G downlink operation.	FIR filter at pulse shaping/equalization/decoding.	BPSK, QPSK, 8-PSK	BER (link), Structural Similarity Index and Correlation Coefficient (Image transmission)
Ma (2023) [31]	Approximate communication scheme for federated learning application.	Gradients.	QPSK, 16-QAM, 256-QAM	Test accuracy

While approximation techniques can enhance performance at the expense of reduced accuracy, their application tends to diminish the reliability of the system. In the work [32], strategies for testing approximate circuits are delineated, emphasizing the critical role of reliability in the application of approximation techniques to any system. To address the reliability challenge, it is necessary to precisely estimate the accuracy of the approximate system in correlation to the level of induced approximation. The work [33] delves into approximation techniques grounded in the determinism of system accuracy and the granularity control provided by these techniques. The objective is to bolster the reliability of approximate systems by accurately estimating system accuracy reflecting the level of approximation.

Multiplication operations are pivotal in communication systems, influencing overall processing capability and the efficiency of handling complex mathematical transformations. Precision in these operations is vital for maintaining system accuracy, directly impacting signal fidelity during reception. Moreover, the resource-intensive nature of multiplication operations consumes substantial computational resources. In the context of energy efficient communication system design, particularly in environments with resource constraints and battery-powered devices, optimizing the efficiency of multiplication operations becomes paramount. Therefore, the introduction of approximations in multiplication operations presents an opportunity to improve the overall energy efficiency of communication systems.

Arithmetic functional units performing operations like multiplication inherently function as nonlinear static systems and approximating them introduces irregularities in the system. The challenge lies in the insufficient capacity of current statistical metrics to fully capture these irregularities arising from approximation, creating a barrier to the

widespread adoption of approximate arithmetic units as concerns about system reliability emerge. In one such attempt, the work [34] seeks to assess the impact of a nonlinear approximate adder on the application using statistical metrics. Accurately measuring the irregularities introduced in the computing system due to approximation is essential for enhancing the reliability of approximated systems and making more informed choices regarding the selection of approximation techniques.

Noise models are prevalent to model the cause of irregularities in the system, which may be deterministic or non-deterministic [35]. The peak value evaluation error method in wireless receivers was studied in the work [36] for urban noise impulses using past experimental data for different frequency bands, modulation schemes, and bit rates. Non-Gaussian noise was statistically modeled in the work [37] for signal processing applications. The work [38] provides an overview of impulse noise and its models, highlighting their similarities and differences in communication systems and by comparing the performance of single-carrier and multi-carrier communication systems under impulse noise. A computationally intensive Gaussian mixture model is employed to model the impulsive noise for computing analytical expression of SER [39].

In the systematic analysis depicted in Figure 1, the preliminary entities are derived from components. These entities are utilized to formulate primary metrics, while secondary metrics are derived from the primary ones. All analyses are conducted using primary and secondary metrics. The proposed work introduces an AMN model to characterize irregularities resulting from the approximate multiplication operation in wireless signal detection as shown in Figure 2. The TM structure is used as an approximate multiplication technique, while the MMSE technique is utilized for signal detection in the proposed work. Utilizing the AMN, a closed-loop expression for the SER is derived for QPSK MMSE signal detection with TM. This expression serves to assess the resilience of various configurations of TM for MMSE detection.

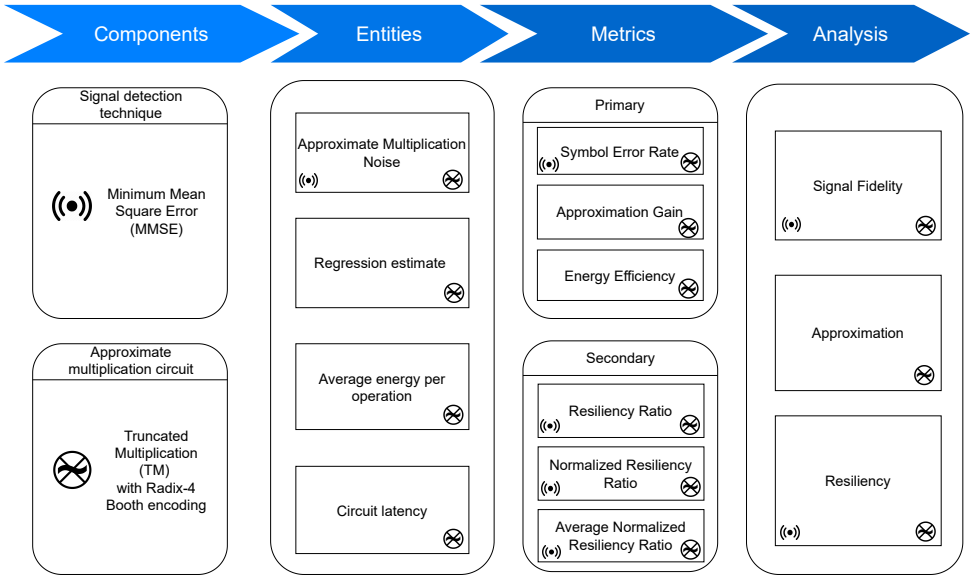

Figure 1. System analysis for energy-efficient wireless signal detection using approximate multiplication circuit. The relationship between entities, metrics, and analysis with components is explicitly demonstrated.

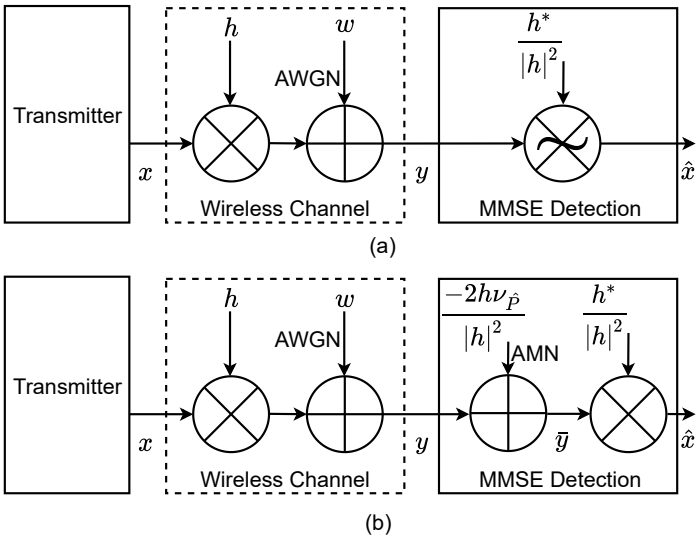

Figure 2. (**a**) MMSE detection using approximate multiplication. (**b**) Equivalent model for computing $\hat{x}$ using AMN and accurate multiplication.

3. Methodology

3.1. Preliminary

The notation $|.|$ signifies the absolute value, while $\mathbb{E}$ denotes the expectation operator. The symbols $\Re$ and $\Im$ represent the real and imaginary components, respectively. $\mathcal{P}\{\mathscr{G}\}$ denotes the probability of occurrence of any event $\mathscr{G}$. $[.]_2$ denotes integer values represented in two's complement form. The Probablity Distribution Function (PDF) of normal distribution with mean μ and variance σ^2 is represented as $\mathcal{N}(\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$, where exp(.) is the exponential function. Adding a constant k to $\mathcal{N}(\mu,\sigma)$ results in a new distribution, $\mathcal{N}(k+\mu,\sigma)$. Multiplying a normal distribution by k yields new normal distribution $\mathcal{N}(k\mu,k\sigma)$. For two normal distributions $\mathcal{N}(\mu_1,\sigma_1)$ and $\mathcal{N}(\mu_2,\sigma_2)$, the resultant normal distribution is given as $\mathcal{N}(\mu_1,\sigma_1)+\mathcal{N}(\mu_2,\sigma_2) = \mathcal{N}(\mu_1+\mu_2, \sqrt{\sigma_1^2+\sigma_2^2})$. The complementary error function, erfc(.), evaluates to 0 as ∞ is approached and to 2 as $-\infty$ is approached. Using the symmetry property of erfc(.), it can be inferred that $\text{erfc}(-x) = 2 - \text{erfc}(x)$. The integration of a normal distribution within the interval $[a,b]$ is expressed as the definite integral [40]:

$$\int_a^b \mathcal{N}(\mu,\sigma) = \frac{1}{2}\left(\text{erfc}\left(\frac{a-\mu}{\sigma\sqrt{2}}\right) - \text{erfc}\left(\frac{b-\mu}{\sigma\sqrt{2}}\right)\right) \tag{1}$$

Cov(.) represents the covariance operation. The linear operations performed on a complex random variable apply to its real and imaginary components. The symbol $\approx$ is used to denote the approximate multiplication of two operands.

3.2. Truncated Multiplication

Consider bit signals $a_i, b_i, \phi_i \in \{0,1\}$ for $i = 0,1,\ldots,N-1$. In the context of signed multiplication, the N-bit multiplicand and multiplier operands are represented in two's complement form as $[A]_2 = -a_{N-1}2^{N-1} + \sum_{i=0}^{N-2} a_i 2^i$ and $[B]_2 = -b_{N-1}2^{N-1} + \sum_{i=0}^{N-2} b_i 2^i$, respectively. The multiplication product is denoted as $[P]_2 = -\phi_{2N-1}2^{2N-1} + \sum_{i=0}^{2N-2} \phi_i 2^i$. In the Radix-4 Booth algorithm for encoding the multiplier [41], the multiplication product $[P]_2$ is computed using the Algorithm 1.

Algorithm 1 Signed multiplication using Radix-4 Booth algorithm.

1: $[A]_2$ and $[B]_2$ are N-bit multiplicand and multiplier.
2: **procedure** $([A]_2, [B]_2, N)$
3: Initialize product $[P]_2$ as 0 with $2N$ bits.
4: Initialize b_{-1} as 0.
5: **for** $i \leftarrow 0$ to $N/2 - 1$ **do** ▷ N^2 clock cycles
6: **if** $b_{2i+1}b_{2i}b_{2i-1} = 001$ **or** $b_{2i+1}b_{2i}b_{2i-1} = 010$ **then**
7: $[P]_2 \leftarrow [P]_2 + ([A]_2 \ll 2i)$ ▷ $2N$ clock cycles
8: **else if** $b_{2i+1}b_{2i}b_{2i-1} = 101$ **or** $b_{2i+1}b_{2i}b_{2i-1} = 110$ **then**
9: $[P]_2 \leftarrow [P]_2 - ([A]_2 \ll 2i)$ ▷ $2N$ clock cycles
10: **else if** $b_{2i+1}b_{2i}b_{2i-1} = 011$ **then**
11: $[P]_2 \leftarrow [P]_2 + ([A]_2 \ll 4i)$ ▷ $2N$ clock cycles
12: **else if** $b_{2i+1}b_{2i}b_{2i-1} = 100$ **then**
13: $[P]_2 \leftarrow [P]_2 - ([A]_2 \ll 4i)$ ▷ $2N$ clock cycles
14: **end if**
15: **end for**
16: **return** $[P]_2$
17: **end procedure**

The multiplication operation is approximated for TM by truncating the M least significant bits of every partial product generated by the multiplication operation, as illustrated in Figure 3. Thus, a TM configuration is decided by M. The TM product is obtained as $[\hat{P}]_2 = -\phi_{2N-1}2^{2N-1} + \sum_{i=M}^{2N-2} \phi_i 2^i$.

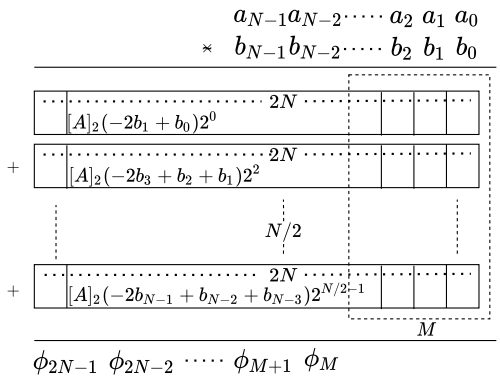

Figure 3. TM implemented using Radix-4 Booth algorithm for signed N bit operands $[A]_2$ and $[B]_2$.

However, the operands used in baseband processing are rational numbers with a fractional part. Therefore, N-bit signed multiplication with fixed-point representation is considered for baseband processing, with $N/2$ bits allocated for the integer part and $N/2$ bits for the fractional part. Consequently, the operands for baseband processing are scaled by a factor of $2^{N/2}$ to convert them into two's complement form. The multiplication product is then converted back to fixed-point representation from two's complement form by using a scaling factor of 2^{-N}. With operands $[A]_2 = A \cdot 2^{N/2}$ and $[B]_2 = B \cdot 2^{N/2}$, the accurate multiplication is represented as $P = ([A]_2[B]_2)2^{-N}$, while the TM product is represented as $\hat{P} = ([A]_2 \divideontimes [B]_2)2^{-N}$.

The mean of the error between accurate multiplication and TM is given by $\mu_{\hat{p}} = \mathbb{E}\{[A]_2[B]_2\} - \mathbb{E}\{[A]_2 \divideontimes [B]_2\}$. Considering the two's complement form of operands, the exact value of $\mu_{\hat{p}}$ is obtained through exhaustive simulation of all possible multiplication values for a particular N, as given in Table 2. For exhaustive simulation, all integer values from range $[-2^{N-1}, 2^{N-1} - 1)$ are considered for multiplicand and multiplier operand. However, when it is not feasible to perform exhaustive simulation for a particular N, $\mu_{\hat{p}}$ is point estimated with selective multiplication values. For point estimating $\mu_{\hat{p}}$, K_N samples

from range $[-2^{N-1}, 2^{N-1}-1)$ are considered for multiplicand and multiplier operand with an equal step size of $\frac{2^{N-1}-1}{K_N/2-1} = \frac{2^N-2}{K_N-2}$. The mean of the error for fixed-point representation is given by:

$$\nu_{\hat{P}} = \mathbb{E}\{P - \hat{P}\} = \mathbb{E}\{P\} - \mathbb{E}\{\hat{P}\} = \mu_{\hat{P}} 2^{-N} \tag{2}$$

Table 2. Error mean $\nu_{\hat{P}}$ values for $N = 8, 12, 16, 20$. $\nu_{\hat{P}}$ values for $N = 8, 12$ are calculated with exhaustive simulation of operands, while $\nu_{\hat{P}}$ values for $N = 16, 20$ are point estimated using $K_N = 4096$ and represented by †.

	$\nu_{\hat{P}}$			
M	$N = 8$	$N = 12$	N=16 †	N=20 †
1	9.77×10^{-4}	6.10×10^{-5}	3.81×10^{-6}	2.38×10^{-7}
2	3.91×10^{-3}	2.44×10^{-4}	1.57×10^{-5}	9.56×10^{-7}
3	1.37×10^{-2}	8.54×10^{-4}	5.64×10^{-5}	3.35×10^{-6}
4	3.71×10^{-2}	2.32×10^{-3}	1.55×10^{-4}	9.09×10^{-6}
5	9.96×10^{-2}	6.23×10^{-3}	4.13×10^{-4}	2.44×10^{-5}
6	2.40×10^{-1}	1.50×10^{-2}	9.88×10^{-4}	5.89×10^{-5}
7	5.84×10^{-1}	3.65×10^{-2}	2.38×10^{-3}	1.43×10^{-4}
8	1.33×10^{0}	8.34×10^{-2}	5.41×10^{-3}	3.27×10^{-4}
9	-	1.93×10^{-1}	1.24×10^{-2}	7.56×10^{-4}
10	-	4.27×10^{-1}	2.75×10^{-2}	1.67×10^{-3}
11	-	9.58×10^{-1}	6.15×10^{-2}	3.76×10^{-3}
12	-	2.08×10^{0}	1.33×10^{-1}	8.16×10^{-3}
13	-	-	2.93×10^{-1}	1.80×10^{-2}
14	-	-	6.27×10^{-1}	3.85×10^{-2}
15	-	-	1.36×10^{0}	8.35×10^{-2}
16	-	-	2.88×10^{0}	1.77×10^{-1}
17	-	-	-	3.81×10^{-1}
18	-	-	-	8.04×10^{-1}
19	-	-	-	1.71×10^{0}
20	-	-	-	3.59×10^{0}

3.2.1. Regression Estimate for $\hat{P}$

The $\hat{P}$ values are obtained through the operation of TM. To derive an analytical expression for the SER, there is a necessity for a mathematical model representing $\hat{P}$. The $\hat{P}$ values exhibit slight deviation from the P values for a specific operand pair. This relationship between approximate and accurate multiplication values is depicted in Figure 4 for $N = 8$, where P and $\hat{P}$ values display linear co-variation. This inference can be extended to other values of N without loss of generality. Utilizing this sufficient statistical information about covariance, it is possible to estimate $\hat{P}$ from P as the predictor variable in a linear regression model [42].

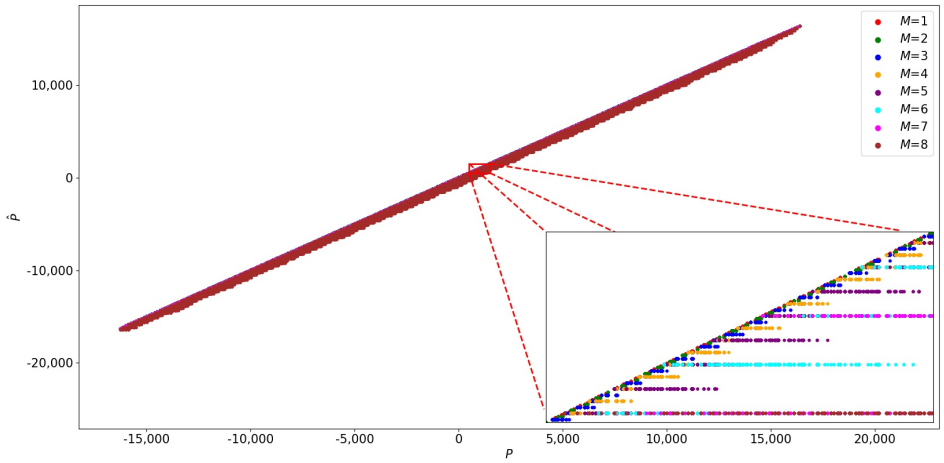

Figure 4. Variance of accurate multiplication values P and TM values $\hat{P}$ for $N = 8$ for varying M. P and $\hat{P}$ have linear covariance.

Consider a linear regression model $\hat{P} = \beta_0 + \beta_1 P + \epsilon$, where β_0 and β_1 are regression coefficients and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is the error term. The estimated regression model is given as $\hat{P} \approx \beta_0 + \beta_1 P$ and the values of estimated regression coefficients $\hat{\beta_0}$ and $\hat{\beta_1}$ are calculated as follows:

$$\hat{\beta_1} = \frac{\text{Cov}(\hat{P}, P)}{\text{Cov}(P, P)} \tag{3}$$

and

$$\hat{\beta_0} = \mathbb{E}\{\hat{P}\} - \hat{\beta_1}\mathbb{E}\{P\} \tag{4}$$

Since $\text{Cov}(\hat{P}, P) \approx \text{Cov}(P, P)$ as inferred from Figure 4, it implies that $\hat{\beta_1} \approx 1$; hence, $\hat{\beta_0}$ can be evaluated from Equation (4) and Equation (2) as follows:

$$\hat{\beta_0} = \mathbb{E}\{\hat{P}\} - \mathbb{E}\{P\} = -\nu_{\hat{p}} \tag{5}$$

The estimated regression model can be calculated as follows:

$$\hat{P} \approx P - \nu_{\hat{p}} \tag{6}$$

The linear regression model for $\hat{P}$ in Equation (6) provides a crude estimate of the multiplication product values of TM, which will be used for SER computation.

3.2.2. Energy Efficiency

The energy efficiency of TM for a particular N and M is computed with respect to the accurate multiplication. For the analysis of energy efficiency, an approximate expression is computed henceforth. For any multiplication operation performed using a digital logic processor, the total power consumed can be expressed as $E \cdot \#\,\text{multiplication/sec}$, where E is the average energy consumed per multiplication operation. The latency T of the multiplication operation is computed as the product of total clock cycles and the processor clock frequency. The number of multiplication operations per second is thus $\frac{1}{T}$. Low E signifies a more energy-efficient multiplication operation.

For accurate multiplication operation using the Radix-4 Booth algorithm, the total clock cycles consumed by the multiplication operation stem from the steps outlined in Algorithm 1. Primarily, every partial product computation consumes about $2N$ clock cycles

for a typical ripple carry addition operation. As there are $N/2$ partial products to be computed, the total clock cycles consumed are about $N/2(2N) = N^2$. In the case of TM as shown in Figure 3, the operand bit-width for partial product addition is $2N - M$. Hence, the total clock cycles for TM are:

$$\frac{N}{2}(2N - M) = N^2 - \frac{NM}{2} \tag{7}$$

For a processor operating at a clock frequency of F Hz, the latency of accurate multiplication and TM is computed as:

$$T_P = N^2 F,\ T_{\hat{p}} = \left(N^2 - \frac{NM}{2}\right)F \tag{8}$$

For a power budget of Q Joule intended for multiplication operation on the processor, consider both accurate multiplication and TM for a specific task. Then, the relation between energy per operation for accurate multiplication denoted as E_P and $E_{\hat{p}}$ for TM can be derived as follows:

$$Q = \frac{E_P}{T_P}, \quad Q = \frac{E_{\hat{p}}}{T_{\hat{p}}} \implies \frac{E_P}{T_P} = \frac{E_{\hat{p}}}{T_{\hat{p}}} \tag{9}$$

To quantify the reduction in average energy consumed per multiplication operation, the fractional gain due to approximation caused by TM can be computed as approximation gain as follows:

$$\text{Approximation Gain} = \frac{E_P}{E_{\hat{p}}} = \frac{T_P}{T_{\hat{p}}} = \frac{N^2 F}{\left(N^2 - \frac{NM}{2}\right)F} = \frac{2N}{2N - M} \tag{10}$$

The relative change in $E_{\hat{p}}$ with respect to E_P is calculated as:

$$\frac{E_P - E_{\hat{p}}}{E_P} = 1 - \frac{E_{\hat{p}}}{E_P} \tag{11}$$

Expressing Equation (11) in terms of N and M as:

$$\text{Energy Efficiency} = \frac{E_P - E_{\hat{p}}}{E_P} = \frac{M}{2N} \tag{12}$$

The expressions for approximation gain and energy efficiency enable us to compare the gains achieved due to approximation for multiplication operation with accurate multiplication and TM.

3.3. SER Expression

For the QPSK constellation of symbols with each symbol represented by 2 bits, the alphabets for symbols are represented by set $\mathcal{X} = \{-1 - j1, -1 + j1, 1 - j1, 1 + j1\}$. Let x be the signal representing the transmit symbol such that $x \in \mathcal{X}$ considering Additive White Gaussian Noise (AWGN) channel and constant fading channel h. E_s represents the energy per symbol. AWGN for the receiver system is represented by w and the random variable W is used to represent the values of w such that $W \sim \mathcal{CN}(0, \sigma_n^2)$, where $\sigma_n^2 = N_0$ and N_0 being the Noise Spectral Density (NSD). E_s/N_0 is used to denote the SNR. The noise variance is evenly distributed between real and imaginary components of the symbol such that the variance per component is $\frac{\sigma_n^2}{2}$. The scaling factor for energy normalization for x is given as $S = \sqrt{\frac{E_s}{2}}$.

The received signal is represented as:

$$y = Shx + w \tag{13}$$

Consider AMN as $\delta = \delta_{\Re} + j\delta_{\Im}$. The received signal with AMN:

$$\overline{y} = Shx + w + \delta = y + \delta \tag{14}$$

MMSE detection [43] is a signal detection technique derived using optimization of the mean square error between transmit and receive symbols. MMSE detection achieves near-optimal performance while maintaining low computational complexity, rendering it highly suitable for implementation across a wide array of computing platforms, particularly those with limited computing resources. Additionally, MMSE detection exhibits high robustness across various SNR conditions, particularly in scenarios characterized by low SNR levels. This robustness enhances error resilience within the system, which enables expansion of the scope for approximation with high reliability.

The transmit symbol is estimated using MMSE at the receiver represented by $\hat{x}$ such that:

$$\hat{x} = \frac{1}{S}\frac{\mathbb{E}\{|x|^2\}}{\mathbb{E}\{|x|^2\}|h|^2 + N_0}h^* y = \frac{h^*}{S|h|^2}y \tag{15}$$

as $\mathbb{E}\{|x|^2\}|h|^2 \gg N_0$.

3.3.1. AMN Model

The MMSE computes $\hat{x}$ for transmit symbol x; however, by employing TM at the detection stage, an approximated $\hat{x}$ is obtained. To compute expression for δ, the signal $\hat{x}$ approximated by using TM at the receiver as shown in Figure 2a is equated to equivalent model of $\hat{x}$ approximated by employing AMN and accurate multiplication as shown in Figure 2b. The approximate value of $\hat{x}$ using TM at the receiver is computed as follows:

$$\begin{aligned}\hat{x} &\approx \frac{h^*}{S|h|^2} \approx y = \frac{1}{S|h|^2}(h_{\Re} \approx y_{\Re} + h_{\Im} \approx y_{\Im}) + \frac{j}{S|h|^2}(h_{\Re} \approx y_{\Im} - h_{\Im} \approx y_{\Re}) \\ &\approx \left(\frac{h_{\Re}}{S|h|^2}y_{\Re} + \frac{h_{\Im}}{S|h|^2}y_{\Im} - \frac{2\nu_{\hat{p}}}{S|h|^2}\right) + j\left(\frac{h_{\Re}}{S|h|^2}y_{\Im} - \frac{h_{\Im}}{S|h|^2}y_{\Re}\right)\end{aligned} \tag{16}$$

The approximate value of $\hat{x}$ using accurate multiplication by employing AMN is computed as follows:

$$\begin{aligned}\hat{x} &\approx \frac{h^*}{S|h|^2}\overline{y} = \left(\frac{h_{\Re}}{S|h|^2}\overline{y}_{\Re} + \frac{h_{\Im}}{S|h|^2}\overline{y}_{\Im}\right) + j\left(\frac{h_{\Re}}{S|h|^2}\overline{y}_{\Im} - \frac{h_{\Im}}{S|h|^2}\overline{y}_{\Re}\right) \\ &\approx \left(\frac{h_{\Re}}{S|h|^2}y_{\Re} + \frac{h_{\Im}}{S|h|^2}y_{\Im} + \frac{h_{\Re}\delta_{\Re} + h_{\Im}\delta_{\Im}}{S|h|^2}\right) + \\ &j\left(\frac{h_{\Re}}{S|h|^2}y_{\Im} - \frac{h_{\Im}}{S|h|^2}y_{\Re} + \frac{h_{\Re}\delta_{\Im} - h_{\Im}\delta_{\Re}}{S|h|^2}\right)\end{aligned} \tag{17}$$

Equating Equations (16) and (17):

$$h_{\Re}\delta_{\Re} + h_{\Im}\delta_{\Im} = -2\nu_{\hat{p}} \tag{18}$$

and

$$h_{\Re}\delta_{\Im} - h_{\Im}\delta_{\Re} = 0 \tag{19}$$

From Equations (18) and (19), the values of $\delta_{\Re}$, $\delta_{\Im}$ and eventually δ can be evaluated as:

$$\delta = \frac{-2\nu_{\hat{p}}h_{\Re}}{|h|^2} - j\frac{2\nu_{\hat{p}}h_{\Im}}{|h|^2} = \frac{-2h\nu_{\hat{p}}}{|h|^2} \tag{20}$$

AMN facilitates use of signal $\bar{y}$ in $\hat{x}$ detection.

3.3.2. SER Evaluation Using AMN

With signal x, consider the approximate detection of $\hat{x}$ using MMSE detection by using signal $\bar{y}$. For AWGN channel, the random variable for $\bar{y}$ is modeled as $\overline{Y} = Shx + W + \delta$.

Separating $\overline{Y}$ into real and imaginary components $\overline{Y}_{\Re} \sim \mathcal{N}(Sh_{\Re}x_{\Re} - Sh_{\Im}x_{\Im} + \delta_{\Re}, \sigma_n/\sqrt{2})$ and $\overline{Y}_{\Im} \sim \mathcal{N}(Sh_{\Re}x_{\Im} + Sh_{\Im}x_{\Re} + \delta_{\Im}, \sigma_n/\sqrt{2})$. Hence,

$$\hat{x} \approx \frac{h^*}{S|h|^2}\overline{y} \implies \hat{X} \approx \frac{h^*}{S|h|^2}\overline{Y} \tag{21}$$

$$\hat{X} \approx \left(\frac{h_{\Re}}{S|h|^2}\overline{Y}_{\Re} + \frac{h_{\Im}}{S|h|^2}\overline{Y}_{\Im}\right) + j\left(\frac{h_{\Re}}{S|h|^2}\overline{Y}_{\Im} - \frac{h_{\Im}}{S|h|^2}\overline{Y}_{\Re}\right) \tag{22}$$

The linear combination of two random variables also results in a random variable [44]. Using the PDF for $\overline{Y}_{\Re}$ and $\overline{Y}_{\Im}$ in Equation (22) and δ, the PDF expressions for real and imaginary parts for $\hat{X}$ are evaluated as follows.

$$\begin{aligned} f_{\hat{X}_{\Re}} &\approx \frac{1}{S|h|^2}\mathcal{N}\left(h_{\Re}(Sh_{\Re}x_{\Re} - Sh_{\Im}x_{\Im} + \delta_{\Re}) + h_{\Im}(Sh_{\Re}x_{\Im} + Sh_{\Im}x_{\Re} + \delta_{\Im}), \frac{\sigma_n}{\sqrt{2}}\sqrt{h_{\Re}^2 + h_{\Im}^2}\right) \\ &\approx \frac{1}{S|h|^2}\mathcal{N}\left(x_{\Re}S(h_{\Re}^2 + h_{\Im}^2) - 2\,\nu_{\hat{p}}, \frac{\sigma_n}{\sqrt{2}}|h|\right) \\ &\approx \mathcal{N}\left(x_{\Re} - \frac{2\nu_{\hat{p}}}{S|h|^2}, \frac{\sigma_n}{\sqrt{2}S|h|}\right) \end{aligned} \tag{23}$$

$$\begin{aligned} f_{\hat{X}_{\Im}} &\approx \frac{1}{S|h|^2}\mathcal{N}\left(h_{\Re}(Sh_{\Re}x_{\Im} + Sh_{\Im}x_{\Re} + \delta_{\Im}) - h_{\Im}(Sh_{\Re}x_{\Re} - Sh_{\Im}x_{\Im} + \delta_{\Re}), \frac{\sigma_n}{\sqrt{2}}\sqrt{h_{\Re}^2 + h_{\Im}^2}\right) \\ &\approx \frac{1}{S|h|^2}\mathcal{N}\left(x_{\Im}S(h_{\Re}^2 + h_{\Im}^2), \frac{\sigma_n}{\sqrt{2}}\sqrt{h_{\Re}^2 + h_{\Im}^2}\right) \\ &\approx \mathcal{N}\left(x_{\Im}, \frac{\sigma_n}{\sqrt{2}S|h|}\right) \end{aligned} \tag{24}$$

To compute the area under the inference regions for QPSK as shown in Figure 5, consider integrals $I1$, $I2$, $I3$, and $I4$. Using $f_{\hat{X}_{\Re}}$ and $f_{\hat{X}_{\Im}}$ when $x_{\Re}, x_{\Im} = \pm 1$, these integrals are computed as follows:

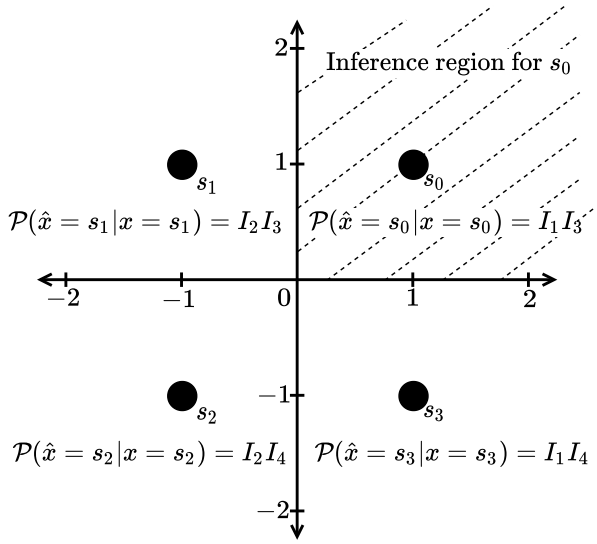

Figure 5. Inference regions for the QPSK constellation.

$$I1 = \int_0^{\infty} f_{\hat{X}_{\Re}}\Big|_{x_{\Re}=1} = \int_0^{\infty} \mathcal{N}\left(1 - \frac{2\nu_{\hat{p}}}{S|h|^2}, \frac{\sigma_n}{\sqrt{2}S|h|}\right)$$
$$= \frac{1}{2}\text{erfc}\left(\frac{-S|h|^2 + 2\nu_{\hat{p}}}{\sigma_n|h|}\right) \tag{25}$$

$$I2 = \int_{-\infty}^{0} f_{\hat{X}_{\Re}}\Big|_{x_{\Re}=-1} = \int_{-\infty}^{0} \mathcal{N}\left(-1 - \frac{2\nu_{\hat{p}}}{S|h|^2}, \frac{\sigma_n}{\sqrt{2}S|h|}\right)$$
$$= \frac{1}{2}\left(2 - \text{erfc}\left(\frac{S|h|^2 + 2\nu_{\hat{p}}}{\sigma_n|h|}\right)\right) = \frac{1}{2}\text{erfc}\left(\frac{-S|h|^2 - 2\nu_{\hat{p}}}{\sigma_n|h|}\right) \tag{26}$$

$$I3 = \int_0^{\infty} f_{\hat{X}_{\Im}}\Big|_{x_{\Im}=1} = \int_0^{\infty} \mathcal{N}\left(1, \frac{\sigma_n}{\sqrt{2}S|h|}\right) = \frac{1}{2}\text{erfc}\left(\frac{-S|h|}{\sigma_n}\right) \tag{27}$$

$$I4 = \int_{-\infty}^{0} f_{\hat{X}_{\Im}}\Big|_{x_{\Im}=-1} = \int_{-\infty}^{0} \mathcal{N}\left(-1, \frac{\sigma_n}{\sqrt{2}S|h|}\right)$$
$$= \frac{1}{2}\left(2 - \text{erfc}\left(\frac{S|h|}{\sigma_n}\right)\right) = \frac{1}{2}\text{erfc}\left(\frac{-S|h|}{\sigma_n}\right) \tag{28}$$

These integrals are used in evaluating the probability of correct symbol detection by the receiver for every symbol in $\mathcal{X}$. The SER for QPSK is computed by evaluating the union probability that each of the symbols in $\mathcal{X}$ is transmitted and received correctly at the receiver [45]. The SER expression is computed as follows:

$$SER = 1 - \frac{1}{4}\sum_{i=1}^{4} \mathcal{P}(\hat{x} = s_i | x = s_i)$$
$$= 1 - \frac{1}{4}(I_1 I_3 + I_2 I_3 + I_2 I_4 + I_1 I_4) = 1 - \frac{1}{4}(I3 + I4)(I1 + I2)$$
$$= 1 - \frac{1}{8}\text{erfc}\left(\frac{-S|h|}{\sigma_n}\right)\left(\text{erfc}\left(\frac{-S|h|^2 + 2\nu_{\hat{p}}}{\sigma_n|h|}\right) + \text{erfc}\left(\frac{-S|h|^2 - 2\nu_{\hat{p}}}{\sigma_n|h|}\right)\right) \tag{29}$$

Substituting the values of S and σ_n,

$$= 1 - \frac{1}{8}\text{erfc}\left(-|h|\sqrt{\frac{E_s}{2N_0}}\right)\left(\text{erfc}\left(\frac{-|h|^2\sqrt{E_s} + 2\sqrt{2}\nu_{\hat{p}}}{|h|\sqrt{2N_0}}\right) + \right.$$
$$\left.\text{erfc}\left(\frac{-|h|^2\sqrt{E_s} - 2\sqrt{2}\nu_{\hat{p}}}{|h|\sqrt{2N_0}}\right)\right) \tag{30}$$

SER is computed by simulation in Figure 2a using 5000 symbols in Python. The analytical expression of Equation (30) tracks SER obtained using simulation for $N = 8, 16$ as shown in Figure 6. It can be observed that for both $N = 8$ and $N = 16$, the SER starts increasing with an increase in SNR after $M = N/2$.

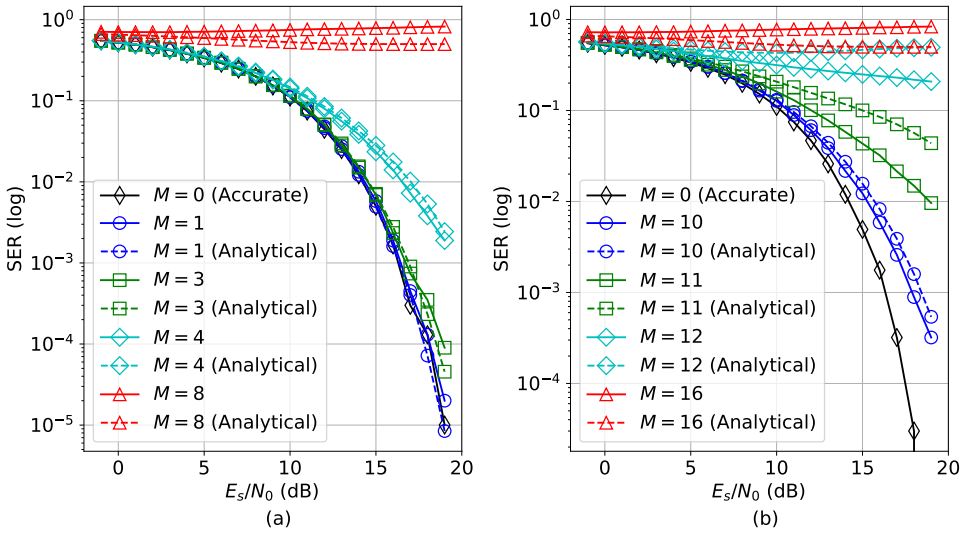

Figure 6. Comparison of SER for $N = 8, 16$ with channel gain $|h|^2 = 0.25$ for varying M calculated analytically using Equation (30) and by simulation for (**a**) $N = 8$; (**b**) $N = 16$.

4. Analysis

4.1. Signal Fidelity Analysis

The SER of the QPSK MMSE signal detection using TM is evaluated in Figure 7 under varying channel gain conditions. The SER for a specific N, a specific channel gain and a specific SNR is bounded within an interval. The SER exhibits resilience for a range of M, forming the lower bound and begins to degrade thereafter with an increase in M. Furthermore, after a certain value of M, it becomes constant with a further increase in M, establishing the upper bound of SER.

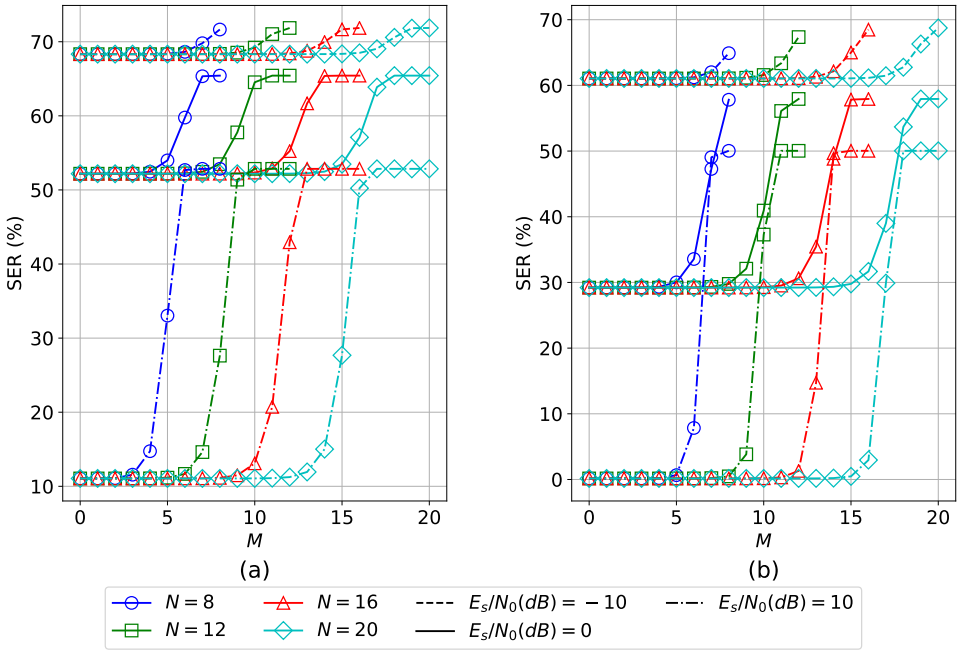

Figure 7. Signal fidelity analysis for QPSK MMSE signal detection using TM for varying M for $N = 8, 16, 12$ and 20 for channel gain (**a**) $|h|^2 = 0.25$; (**b**) $|h|^2 = 1$.

For a given channel gain, the bound interval increases with an increase in SNR. Although the lower bound of SER decreases with an increase in channel gain, the bound interval increases significantly. This implies that approximation causes more degradation of SER for high channel gain. However, the resiliency is sustained for higher values of M when channel gain is increased for the same SNR. Additionally, resiliency is maintained for lower values of M, when SNR is increased for the same channel gain.

4.2. Approximation Analysis

TM configurations for $N = 8, 12, 16$ and 20 are analyzed for approximation gain and energy efficiency in Figure 8. Approximation gain is assessed using Equation (10) and energy efficiency is evaluated using Equation (12). Both approximation gain and energy efficiency increase with an augmentation in M for all values of N. However, the rate of increase in approximation gain and energy efficiency decreases as N increases. This observation implies that the approximation gain and energy efficiency are lower for higher values of N for a specific value of M.

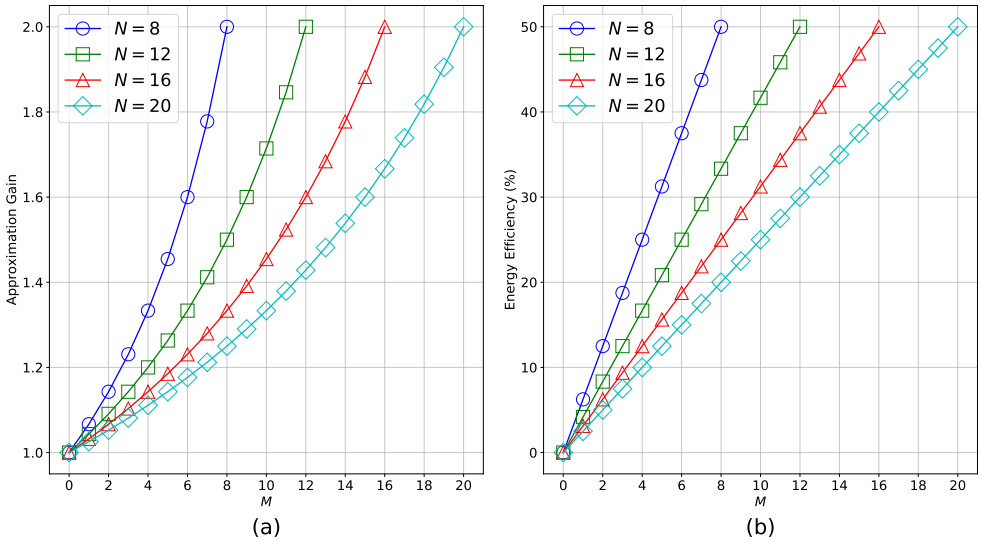

Figure 8. Approximation analysis with varying M truncation bits: (**a**) Approximation Gain; (**b**) Energy efficiency.

4.3. Resiliency Analysis

For achieving high energy efficiency of the TM configuration in QPSK MMSE signal detection, it is crucial to choose a configuration that exhibits a high approximation gain for high energy efficiency. However, high approximation gain necessitates a corresponding increase in the value of M, resulting in a rise in the SER. Balancing a high approximation gain and minimizing the SER is key to achieving optimal performance when utilizing a TM for QPSK MMSE signal detection. The resiliency of TM configuration can be considered high if it provides high approximation gain without degrading much of the SER.

The Resiliency Ratio (RR) serves as a metric to quantify the resiliency of TM operation for QPSK MMSE detection for a specific combination of M, SNR and channel gain computed as follows:

$$\text{RR}\,|_{\text{SNR},|h|^2} = \frac{\text{Approximation Gain}}{SER\,|_{\text{SNR},|h|^2}} \tag{31}$$

The Normalized Resiliency Ratio (NRR), defined as the ratio of RR to the maximum RR for a given h and SNR across all values of M scales the RR in range of $[0, 1]$ and is computed as follows:

$$\mathrm{NRR}\,|_{\mathrm{SNR},|h|^2} = \frac{\mathrm{RR}\,|_{\mathrm{SNR},|h|^2}}{\max\limits_{M \le N}\left(\mathrm{RR}\,|_{\mathrm{SNR},|h|^2}\right)} \tag{32}$$

While the NRR is a useful metric to analyze resiliency, it depends on SNR and channel gain. Hence, Average Normalized Resiliency Ratio (ANRR) provides insights into overall resiliency of the TM and is computed by averaging NRR values for K_{SNR} samples of SNR in range $[-10, 10)$ and $K_{|h|}$ samples of channel gain in range of $[0, 1)$ as follows:

$$\mathrm{ANRR} = \frac{1}{K_{\mathrm{SNR}} K_{|h|}} \sum_{|h|=0}^{1} \sum_{\mathrm{SNR}=-10}^{10} \left(\mathrm{NRR}\,|_{\mathrm{SNR},|h|^2}\right) \tag{33}$$

The graph of NRR and ANRR exhibits a distinct resiliency dip, commencing at the resiliency crest and concluding at the resiliency trough. An examination of Figure 9 reveals that for lower values of M, NRR tends to be lower for lower SNR regime, as depicted in Figure 9a,d, in contrast to its behavior in high SNR regime seen in Figure 9c,f. Conversely, at higher values of M, NRR significantly rises at low SNR compared to its performance at high SNR. The resilience of approximations with low M is more pronounced in the high SNR regime, while those with high M are more resilient in the low SNR regime. The extent of the resiliency dips expands with an increase in both SNR and channel gain. Also, the value of M required to reach the resiliency crest rises with SNR.

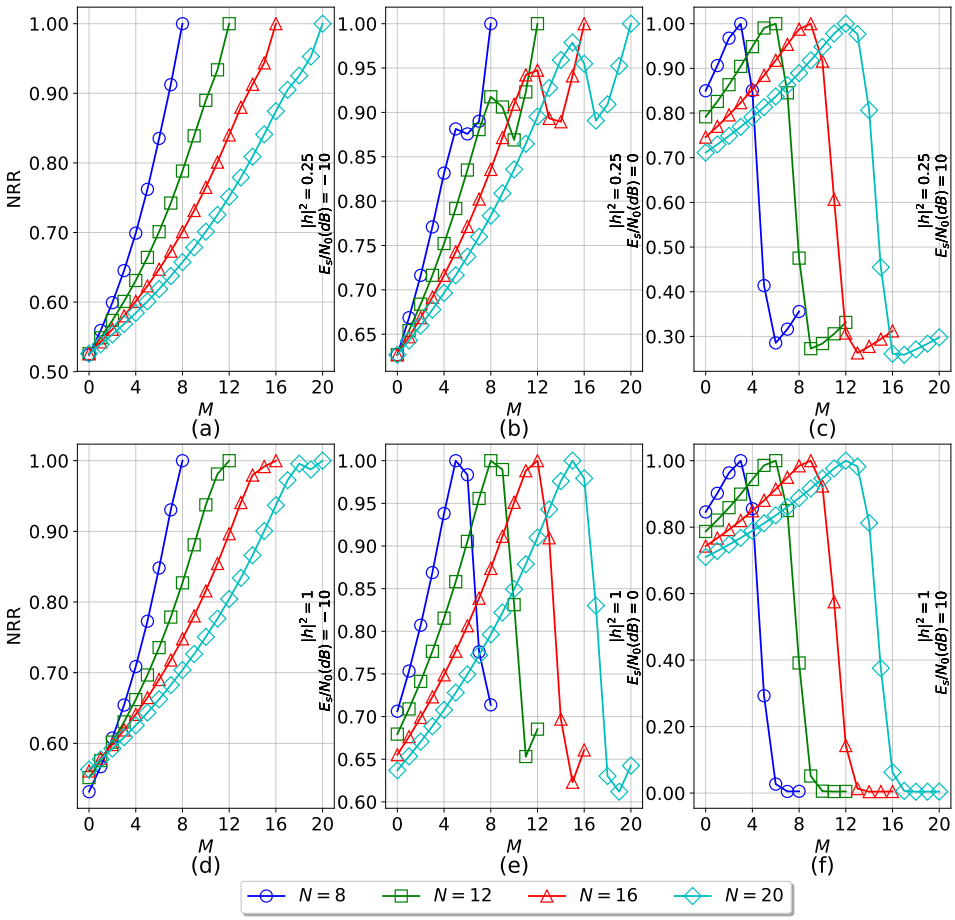

Figure 9. NRR analysis for QPSK MMSE signal detection using TM for varying M. NRR evaluated for **(a)** $|h|^2 = 0.25$, $E_s/N_0(dB) = -10$; **(b)** $|h|^2 = 0.25$, $E_s/N_0(dB) = 0$; **(c)** $|h|^2 = 0.25$, $E_s/N_0(dB) = 10$; **(d)** $|h|^2 = 1$, $E_s/N_0(dB) = -10$; **(e)** $|h|^2 = 1$, $E_s/N_0(dB) = 0$; **(f)** $|h|^2 = 1$, $E_s/N_0(dB) = 10$.

In the lower SNR regime, NRR experiences low resiliency dips, while dips begin to form with an increase in channel gain. NRR approaches 0 after the resiliency trough for high M in high SNR regime and high channel gain. The resiliency trough decreases with an increase in SNR for all values of N. After a resiliency dip, surpassing the NRR beyond the resiliency crest towards 1 becomes increasingly difficult with an increase in SNR and channel gain. Until the resiliency crest is achieved, the rate of increase of NRR amplifies with an increase in both SNR and channel gain. Approximation with higher M is advantageous in low SNR regime, while that with lower M is advantageous in high SNR regime.

From the ANRR analysis presented in Figure 10 using $K_{\text{SNR}} = 20$ and $K_{|h|} = 100$, it is evident that the rate of increase in ANRR diminishes with an increase in N. Resiliency dips for ANRR intensify with an increase in N and the resiliency trough decreases with an increase in N.

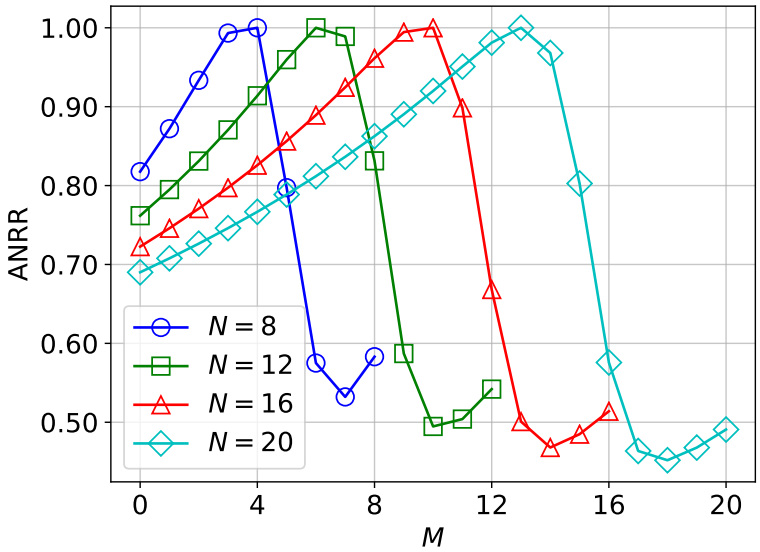

Figure 10. ANRR analysis for QPSK MMSE signal detection using TM for varying M with $K_{\text{SNR}} = 20$ and $K_{|h|} = 100$.

5. Conclusions

The proposed work systematically investigates the impact of employing approximate multiplication, represented by TM, on wireless signal detection, represented by QPSK MMSE detection. It aims to address the challenge of linking approximation level with QoS of the system through analyses of signal fidelity, resilience, and approximation. The study evaluates the AMN model to understand the effects of irregularities induced in the system by approximation and derives an analytical expression for the SER using the AMN model for signal fidelity analysis. A summary of all the simulations conducted for the proposed work is depicted in Table 3.

Signal fidelity analysis forecasts system output at different approximation levels, thereby enhancing the stability of the system. Energy efficiency increases with higher levels of approximation in TM. However, an increase in TM approximation leads to degradation of SER, and system reliability begins to decline after error resilience has been exhausted beyond a certain level of approximation. Resilience metrics capture this phenomenon and provide insights into reliable approximate configurations. Higher levels of approximation exhibit more resilience in low SNR scenarios, while lower levels of approximation are more resilient in high SNR scenarios. In the context of the proposed work, future avenues for exploration include extending the AMN model to accommodate high modulation schemes, diverse channel models, a variety of approximate multiplication schemes, and alternative receiver techniques.

Table 3. Simulation Summary.

Analysis	Metric	Parameter	Description
		Entities	
Computation	$\nu_{\hat{p}}$	$N = 8, 12$; $M = 1 \ldots N$; $K_N = 2^N$; Operand range $[-2^{N-1}, 2^{N-1} - 1)$	As given in Table 2, $\nu_{\hat{p}}$ values are computed by exhaustive simulation.
Computation	$\nu_{\hat{p}}$	$N = 16, 20$; $M = 1 \ldots N$; $K_N = 4096$; Operand range:$[-2^{N-1}, 2^{N-1} - 1)$; step size = $\frac{2^N-2}{K_N-2}$	As given in Table 2, $\nu_{\hat{p}}$ values are computed by point estimation.
Covariance	P, $\hat{P}$	$N = 8$; $M = 1 \ldots N$; $K_N = 2^N$; Operand range $[-2^{N-1}, 2^{N-1} - 1)$	As depicted in Figure 4, P and $\hat{P}$ have linear covariance.
Verification	SER	$N = 8, 16$; $M = 1 \ldots N$; $\lvert h \rvert^2 = 0.25$, 5000 symbols per $E_s/N_0(dB)$	As depicted in Figure 6, the analytical expression for SER tracks the SER computed by simulation.
		System	
Signal Fidelity	SER	$N = 8, 12, 16, 20$; $M = 1 \ldots N$; $\lvert h \rvert^2 = 0.25, 1$; $E_s/N_0(dB) = -10, 0, 10$	As depicted in Figure 7, the bound interval increases with both the SNR and channel gain, which indicates a greater degradation in SER.
Approximation	Approximation Gain	$N = 8, 12, 16, 20$; $M = 1 \ldots N$	As shown in Figure 8a, approximation gain increases with M for all N, while the rate of increase decreases with N.
Approximation	Energy Efficiency	$N = 8, 12, 16, 20$; $M = 1 \ldots N$	As shown in Figure 8b, energy efficiency increases with M for all N, while the rate of increase decreases with N.
Resiliency	NRR	$N = 8, 12, 16, 20$; $M = 1 \ldots N$; $\lvert h \rvert^2 = 0.25, 1$; $E_s/N_0(dB) = -10, 0, 10$	As shown in Figure 9, high NRR is achieved by low M in high SNR regime and high M in low SNR regime.
Resiliency	ANRR	$N = 8, 12, 16, 20$; $M = 1 \ldots N$; $K_{\text{SNR}} = 20$; $K_{\lvert h \rvert} = 100$	As shown in Figure 10, rate of increase in ANRR decreases with N and resiliency dips increase with N.

Author Contributions: Conceptualization, A.K.; methodology, A.K.; software, A.K.; validation, A.K.; formal analysis, A.K.; investigation, A.K.; resources, M.A.O. and D.M.; data curation, A.K.; writing—original draft preparation, A.K.; writing—review and editing, M.A.O. and D.M.; visualization, A.K.; supervision, M.A.O. and D.M.; project administration, M.A.O.; funding acquisition, D.M. All authors have read and agreed to the published version of the manuscript.

Funding: The research was funded in part by Natural Sciences and Engineering Research Council of Canada (NSERC): CRSNG-RDCPJ 514758-17, in part by Prompt, in part by Canadian Foundation for Innovation (CFI), in part by CMC Microsystems, in part by Opal-RT Technologies Inc. and in part by Hydro-Québec.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Jiang, W.; Han, B.; Habibi, M.A.; Schotten, H.D. The Road Towards 6G: A Comprehensive Survey. *IEEE Open J. Commun. Soc.* **2021**, *2*, 334–366. [CrossRef]
2. Chowdhury, M.Z.; Shahjalal, M.; Ahmed, S.; Jang, Y.M. 6G Wireless Communication Systems: Applications, Requirements, Technologies, Challenges, and Research Directions. *IEEE Open J. Commun. Soc.* **2020**, *1*, 957–975. [CrossRef]
3. Rajatheva, N.; Atzeni, I.; Bjornson, E.; Bourdoux, A.; Buzzi, S.; Dore, J.B.; Erkucuk, S.; Fuentes, M.; Guan, K.; Hu, Y.; et al. White paper on broadband connectivity in 6G. *arXiv* **2020**, arXiv:2004.14247.

4. Andersson, C.; Bengtsson, J.; Byström, G.; Frenger, P.; Jading, Y.; Nordenström, M. Improving energy performance in 5G networks and beyond. *Ericsson Technol. Rev.* **2022**, *2022*, 2–11. [CrossRef]
5. Viswanathan, H.; Mogensen, P.E. Communications in the 6G era. *IEEE Access* **2020**, *8*, 57063–57074. [CrossRef]
6. López-Pérez, D.; De Domenico, A.; Piovesan, N.; Xinli, G.; Bao, H.; Qitao, S.; Debbah, M. A survey on 5G radio access network energy efficiency: Massive MIMO, lean carrier design, sleep modes, and machine learning. *IEEE Commun. Surv. Tutor.* **2022**, *24*, 653–697. [CrossRef]
7. Chih-Lin, I.; Han, S.; Bian, S. Energy-efficient 5G for a greener future. *Nat. Electron.* **2020**, *3*, 182–184.
8. Seskar, I.; Patwary, M.; Dutta, A.; Chaparadza, R.; Elkotab, M. INGR Roadmap. In Proceedings of the 2022 IEEE Future Networks World Forum (FNWF), Montreal, QC, Canada, 10–14 October 2022 ; pp. 1–38. [CrossRef]
9. Cheng, X.; Hu, Y.; Varga, L. 5G network deployment and the associated energy consumption in the UK: A complex systems' exploration. *Technol. Forecast. Soc. Chang.* **2022**, *180*, 121672. [CrossRef]
10. Khalili, R.; Salamatian, K. A new analytic approach to evaluation of packet error rate in wireless networks. In Proceedings of the 3rd Annual Communication Networks and Services Research Conference (CNSR'05), Halifax, NS, Canada, 16–18 May 2005; pp. 333–338.
11. Sodhro, A.H.; Obaidat, M.S.; Abbasi, Q.H.; Pace, P.; Pirbhulal, S.; Fortino, G.; Imran, M.A.; Qaraqe, M. Quality of service optimization in an IoT-driven intelligent transportation system. *IEEE Wirel. Commun.* **2019**, *26*, 10–17. [CrossRef]
12. Pundir, M.; Sandhu, J.K. A systematic review of quality of service in wireless sensor networks using machine learning: Recent trend and future vision. *J. Netw. Comput. Appl.* **2021**, *188*, 103084. [CrossRef]
13. Beshley, H.; Beshley, M.; Medvetskyi, M.; Pyrih, J. QoS-aware optimal radio resource allocation method for machine-type communications in 5G LTE and beyond cellular networks. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 9966366. [CrossRef]
14. Bertin, E.; Crespi, N.; Magedanz, T. *Shaping Future 6G Networks: Needs, Impacts, and Technologies*; John Wiley & Sons: Hoboken, NJ, USA, 2021.
15. Chochliouros, I.P.; Kourtis, M.A.; Spiliopoulou, A.S.; Lazaridis, P.; Zaharis, Z.; Zarakovitis, C.; Kourtis, A. Energy efficiency concerns and trends in future 5G network infrastructures. *Energies* **2021**, *14*, 5392. [CrossRef]
16. He, S.; Zhang, Y.; Wang, J.; Zhang, J.; Ren, J.; Zhang, Y.; Zhuang, W.; Shen, X. A survey of millimeter-wave communication: Physical-layer technology specifications and enabling transmission technologies. *Proc. IEEE* **2021**, *109*, 1666–1705. [CrossRef]
17. Wang, Z.; Du, Y.; Wei, K.; Han, K.; Xu, X.; Wei, G.; Tong, W.; Zhu, P.; Ma, J.; Wang, J.; et al. Vision, application scenarios, and key technology trends for 6G mobile communications. *Sci. China Inf. Sci.* **2022**, *65*, 151301. [CrossRef]
18. Leon, V.; Hanif, M.A.; Armeniakos, G.; Jiao, X.; Shafique, M.; Pekmestzi, K.; Soudris, D. Approximate Computing Survey, Part II: Application-Specific & Architectural Approximation Techniques and Applications. *arXiv* **2023**, arXiv:2307.11128.
19. Mittal, S. A Survey of Techniques for Approximate Computing. *ACM Comput. Surv.* **2016**, *48*, 62. [CrossRef]
20. Que, H.H.; Jin, Y.; Wang, T.; Liu, M.K.; Yang, X.H.; Qiao, F. A Survey of Approximate Computing: From Arithmetic Units Design to High-Level Applications. *J. Comput. Sci. Technol.* **2023**, *38*, 251–272. [CrossRef]
21. Liu, W.; Lombardi, F. *Approximate Computing*; Springer: Berlin/Heidelberg, Germany, 2022.
22. Bosio, A.; Ménard, D.; Sentieys, O. *Approximate Computing Techniques: From Component-to Application-Level*; Springer International Publishing: Cham, Switzerland, 2022.
23. Zamani, A.R.; Petri, I.; Diaz-Montes, J.; Rana, O.; Parashar, M. Edge-supported approximate analysis for long running computations. In Proceedings of the 2017 IEEE 5th International Conference on Future Internet of Things and Cloud (FiCloud), Prague, Czech Republic, 21–23 August 2017; pp. 321–328.
24. Traiola, M.; Savino, A.; Di Carlo, S. Probabilistic estimation of the application-level impact of precision scaling in approximate computing applications. *Microelectron. Reliab.* **2019**, *102*, 113309. [CrossRef]
25. Damsgaard, H.J.; Ometov, A.; Nurmi, J. Approximation Opportunities in Edge Computing Hardware: A Systematic Literature Review. *ACM Comput. Surv.* **2023**, *55*, 1–49. [CrossRef]
26. Damsgaard, H.J.; Ometov, A.; Mowla, M.M.; Flizikowski, A.; Nurmi, J. Approximate computing in B5G and 6G wireless systems: A survey and future outlook. *Comput. Netw.* **2023**, *233*, 109872. [CrossRef]
27. Zhou, Y.; Chen, Z.; Lin, J.; Wang, Z. A high-speed successive-cancellation decoder for polar codes using approximate computing. *IEEE Trans. Circuits Syst. II Express Briefs* **2018**, *66*, 227–231. [CrossRef]
28. Hao, M.; Najafi, A.; García-Ortiz, A.; Karsthof, L.; Paul, S.; Rust, J. Reliability of an industrial wireless communication system using approximate units. In Proceedings of the 2019 29th International Symposium on Power and Timing Modeling, Optimization and Simulation (PATMOS), Rhodes, Greece, 1–3 July 2019; pp. 87–90.
29. Xiao, J.; Hu, J.; Han, K. Low complexity expectation propagation detection for SCMA using approximate computing. In Proceedings of the 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, USA, 9–13 December 2019; pp. 1–6.
30. Idrees, M.; Maqbool, M.M.; Bhatti, M.K.; Rahman, M.M.U.; Hafiz, R.; Shafique, M. An approximate-computing empowered green 6G downlink. *Phys. Commun.* **2021**, *49*, 101444. [CrossRef]
31. Ma, X.; Sun, H.; Hu, R.Q.; Qian, Y. Approximate Wireless Communication for Federated Learning. *arXiv* **2023**, arXiv:2304.03359.
32. Anghel, L.; Benabdenbi, M.; Bosio, A.; Traiola, M.; Vatajelu, E.I. Test and reliability in approximate computing. *J. Electron. Test.* **2018**, *34*, 375–387. [CrossRef]
33. Wyse, M .; Baixo, A.; Moreau, T.; Zorn, B.; Sampson, A.; Bornholt, J.; Ceze, L.; Oskin, M. Mapping and Modeling Approximate Computing Techniques. Available online : https://homes.cs.washington.edu/~luisceze/approx-darpa-report.pdf (accessed on 4 December 2023).

34. Bruestel, M.; Kumar, A. Accounting for systematic errors in approximate computing. In Proceedings of the Design, Automation & Test in Europe Conference & Exhibition (DATE), Lausanne, Switzerland, 27–31 March 2017; pp. 298–301.
35. Scales, J.A.; Snieder, R. What is noise? *Geophysics* **1998**, *63*, 1122–1124. [CrossRef]
36. Nakamura, S. A study of errors caused by impulsive noise and a simple estimation method for digital mobile communications. *IEEE Trans. Veh. Technol.* **1996**, *45*, 310–317. [CrossRef]
37. Middleton, D. Non-Gaussian noise models in signal processing for telecommunications: New methods an results for class A and class B noise models. *IEEE Trans. Inf. Theory* **1999**, *45*, 1129–1149. [CrossRef]
38. Shongwe, T.; Vinck, A.H.; Ferreira, H.C. A study on impulse noise and its models. *SAIEE Afr. Res. J.* **2015**, *106*, 119–131. [CrossRef]
39. Rozic, N.; Banelli, P.; Begusic, D.; Radic, J. GMM-Based Symbol Error Rate Analysis for Multicarrier Systems with Impulsive Noise Suppression. *IEEE Trans. Veh. Technol.* **2022**, *71*, 13060–13076. [CrossRef]
40. Abramowitz, M. *Abramowitz and Stegun: Handbook of Mathematical Functions*; United States Department of Commerce: Washington, DC, USA , 1972; Volume 10.
41. Bewick, G.W. Fast Multiplication: Algorithms and Implementation. Ph.D. Thesis, Stanford University, Stanford, CA, USA, 1994.
42. Seber, G.A.; Lee, A.J. *Linear Regression Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2003; Volume 330.
43. Tse, D.; Viswanath, P. *Fundamentals of Wireless Communication*; Cambridge University Press: Cambridge, UK, 2005.
44. Springer, M.D. *The Algebra of Random Variables*; Wiley: Hoboken, NJ, USA, 1979.
45. Haykin, S. *Digital Communications*; Wiley: New York, NY, USA, 1988.

Chapter 6 - Concluding Remarks

6.1 General Discussion

The scope for further improvement in DN includes improving its latency and throughput. With antenna scaling to extremely large number, the advantages of using soft detection methods, such as MAP approaches begin diminishing in terms of SER. Linear computational complexity methods like DN can provide significant gains in SER in such scenario. However, exploration of this aspect can be done for DN with increased computing capabilities and tools to conduct such simulation and development. A focus on optimizing interconnect bandwidth between BS antennas is essential along with improving detection accuracy and throughput. Investigating hybrid topologies could yield more flexible performance gains in MIMO systems.

When designing signed multiplication circuits for FPGA using the LC methodology, a key challenge lies in computing the TP values either analytically or through selective simulation of operands. It would also be valuable to explore the potential gains of the LC methodology in the context of ASIC implementation by adapting the methodology for the ASIC design flow. The performance of approximate multiplication circuits is generally evaluated using Pareto-optimal analysis. Further research into novel metrics for assessing error rate performance and hardware implementation efficiency could lead to a unified metric for evaluating the suitability of an approximate multiplication circuits for specific applications.

For a systematic analysis of the effects of approximate multiplication circuits on wireless signal detection, the use of the AMN model should be further explored to accommodate high modulation schemes, diverse channel models, various approximate multiplication schemes, and novel receiver techniques. Analytically computing the mean

error value $v_{\hat{P}}$ could strengthen the mathematical estimation of $\hat{P}$. Further analysis could also focus on effectively capturing the nonlinear and spatial correlations that arise in MIMO systems to improve the overall analysis, in which the current work provides a guideline.

6.2 Conclusion

The research investigates strategies to improve EE in B5G for MIMO wireless signal detection. Drawing on a thorough analysis of current literature, this research posits two level approaches each offering individual gains in KPIs: algorithm decentralization and approximate computing. These approaches are the focus of three research articles aimed at advancing the overarching goal of enhancing the EE. The work paves way for exploring further application of approximate computing techniques for D-MIMO detection for heightened EE, making it a topic of further interest.

The algorithm decentralization approach is applied to MIMO detection to score on KPIs for B5G. This research introduces a novel D-MIMO algorithm and its hardware implementation topologies based on centralized Newton optimization algorithm. In the second approach, the work explores the application of approximate computing technique for arithmetic signed multiplication operations, which are phenomenal in deciding the QoS of wireless signal detection. A heuristic methodology is proposed to induce approximation into accurate signed multiplication circuits for FPGA, and the system-level impact of these approximate multiplication circuits on MIMO signal detection is analyzed. To determine the appropriate level of approximation relative to QoS in signal detection, a systematic framework is proposed for SISO. Drawing from the conceptual understanding of this framework for SISO, guideline is derived in Appendix. A for selecting an optimal level of approximation in MIMO by considering the tractability

of the problem.

The first article introduces the DN algorithm for D-MIMO uplink detection, achieving performance close to ZF signal detection with relatively low computational complexity. Additionally, it proposes and analyzes two hardware topologies—ring and star—for FPGA implementation across various configuration scenarios. The second article presents LC methodology for creating approximate FPGA softcore circuits designed for signed multiplication operations. LC-BW circuits prioritize high accuracy, while LC-Booth circuits primarily focus on minimizing hardware consumption. Furthermore, it analyzes the application of approximate multiplication circuits for ZF MIMO uplink signal detection, studying the impact on signal fidelity. The third article systematically investigates the effects of approximate multiplication on wireless signal detection, particularly in QPSK MMSE detection utilizing TM operations. The AMN model is employed to characterize system irregularities induced by approximation in wireless signal detection. Signal fidelity, approximation, and resilience analyses contribute to maintaining system stability and reliability.

6.3 Future Work

The research work can lead to future possibilities which are being discussed in contemporary literature:

- **Robustness to realistic channel environments** [13]: Investigate methods to enhance robustness to channel estimation errors and varying channel conditions. Develop algorithms that dynamically adapt to changes in CSI without significant performance degradation. Consider colored noise when designing algorithms for signal detection.

- **Lower interconnect bandwidth** [23]: Explore decentralized algorithms with reduced interconnect bandwidth to improve scalability of D-MIMO.

- **Integration with emerging technologies** [30], [26]: Integrate MIMO detection with emerging technologies such as OTFS, ML frameworks, sensing, and Intelligent Reflecting Surfaces (IRS) to achieve synergistic gains.

- **Utilization of novel electromagnetic modelling techniques** [13]: Leverage distributed processing benefits while mitigating issues such as high inter-symbol interference or spatial non-stationarities.

- **Configurable architectures** [70]: Develop approximate computing enabled wireless signal detection architectures that can be configured to varying network conditions. Investigate novel technique to configure these architectures intelligently.

- **Utilization of ML Methods for Design Space Exploration for EE** [27], [71], [72]: Explore ML techniques for decision making in approximate computing enabled MIMO detection systems.

- **Increased Interoperability** [30]: Develop techniques for seamless cross-platform solution deployment and enhance interoperability between resource-constrained IoT devices and large-scale systems to improve application integration.

References

[1] Mohammed Banafaa, Ibraheem Shayea, Jafri Din, Marwan Hadri Azmi, Abdulaziz Alashbi, Yousef Ibrahim Daradkeh, and Abdulraqeb Alhammadi. 6g mobile communication technology: Requirements, targets, applications, challenges, advantages, and opportunities. Alexandria Engineering Journal, 64:245–274, 2023.

[2] Zakria Qadir, Khoa N Le, Nasir Saeed, and Hafiz Suliman Munawar. Towards 6g internet of things: Recent advances, use cases, and open challenges. ICT express, 9(3):296–312, 2023.

[3] Nandana Rajatheva, Italo Atzeni, Emil Bjornson, Andre Bourdoux, Stefano Buzzi, Jean-Baptiste Dore, Serhat Erkucuk, Manuel Fuentes, Ke Guan, Yuzhou Hu, et al. White paper on broadband connectivity in 6g. arXiv preprint arXiv:2004.14247, 2020.

[4] Cecilia Andersson, Jonas Bengtsson, Greger Byström, Pål Frenger, Ylva Jading, and My Nordenström. Improving energy performance in 5g networks and beyond. Ericsson Technology Review, 2022(8):2–11, 2022.

[5] Josip Lorincz, Tonko Garma, and Goran Petrovic. Measurements and modelling of base station power consumption under real traffic loads. Sensors, 12(4):4281–4310, 2012.

[6] Arsalan Ahmed and Marceau Coupechoux. The long road to sobriety: Estimating the operational power consumption of cellular base stations in france. In 2023 International Conference on ICT for Sustainability (ICT4S), pages 188–196. IEEE, 2023.

[7] Shuangfeng Han, Sen Bian, et al. Energy-efficient 5g for a greener future. Nature Electronics, 3(4):182–184, 2020.

[8] Ivan Seskar, Mohammad Patwary, Ashutosh Dutta, Ranganai Chaparadza, and Muslim Elkotab. Ingr roadmap. In 2022 IEEE Future Networks World Forum (FNWF), pages 1–38, 2022.

[9] J. Han and M. Orshansky. Approximate computing: An emerging paradigm for energy-efficient design. In 2013 18th IEEE European Test Symposium (ETS), pages 1–6, 2013.

[10] Sparsh Mittal. A Survey of Techniques for Approximate Computing. ACM Comput. Surv., 48(4), March 2016.

[11] V. K. Chippa, S. T. Chakradhar, K. Roy, and A. Raghunathan. Analysis and

characterization of inherent application resilience for approximate computing. In 2013 50th ACM/EDAC/IEEE Design Automation Conference (DAC), pages 1–9, 2013.

[12] Xiaoyuan Cheng, Yukun Hu, and Liz Varga. 5g network deployment and the associated energy consumption in the uk: A complex systems' exploration. Technological Forecasting and Social Change, 180:121672, 2022.

[13] Yu Han, Shi Jin, Michail Matthaiou, Tony QS Quek, and Chao-Kai Wen. Toward extra large-scale mimo: New channel properties and low-cost designs. IEEE Internet of Things Journal, 10(16):14569–14594, 2023.

[14] Kaipeng Li, Oscar Castaneda, Charles Jeon, Joseph R Cavallaro, and Christoph Studer. Decentralized coordinate-descent data detection and precoding for massive mu-mimo. In 2019 IEEE International Symposium on Circuits and Systems (ISCAS), pages 1–5. IEEE, 2019.

[15] Kaipeng Li, Rishi R Sharan, Yujun Chen, Tom Goldstein, Joseph R Cavallaro, and Christoph Studer. Decentralized baseband processing for massive mu-mimo systems. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 7(4):491–507, 2017.

[16] Messaoud Ahmed Ouameur and Daniel Massicotte. Efficient distributed processing for large scale mimo detection. In 2019 27th European Signal Processing Conference (EUSIPCO), pages 1–5. IEEE, 2019.

[17] J. Rodríguez Sanchez, F. Rusek, O. Edfors, M. Sarajlić, and L. Liu. Decentralized massive mimo processing exploring daisy-chain architecture and recursive algorithms. IEEE Transactions on Signal Processing, 68:687–700, 2020.

[18] Charles Jeon, Kaipeng Li, Joseph R Cavallaro, and Christoph Studer. Decentralized equalization with feedforward architectures for massive mu-mimo. IEEE Transactions on Signal Processing, 67(17):4418–4432, 2019.

[19] Charles Jeon, Ramina Ghods, Arian Maleki, and Christoph Studer. Optimal data detection in large mimo. arXiv preprint arXiv:1811.01917, 2018.

[20] Hanqing Wang, Alva Kosasih, Chao-Kai Wen, Shi Jin, and Wibowo Hardjawana. Expectation propagation detector for extra-large scale massive mimo. IEEE Transactions on Wireless Communications, 19(3):2036–2051, 2020.

[21] Zhenyu Zhang, Hua Li, Yuanyuan Dong, Xiyuan Wang, and Xiaoming Dai. Decentralized signal detection via expectation propagation algorithm for uplink massive mimo systems. IEEE Transactions on Vehicular Technology,

69(10):11233–11240, 2020.

[22] Pascal Seidel, Steffen Paul, and Jochen Rust. Decentralized massive mimo uplink signal estimation by binary multistep synthesis. In 2019 53rd Asilomar Conference on Signals, Systems, and Computers, pages 1967–1971. IEEE, 2019.

[23] Erik Bertilsson, Oscar Gustafsson, and Erik G Larsson. A scalable architecture for massive mimo base stations using distributed processing. In 2016 50th Asilomar Conference on Signals, Systems and Computers, pages 864–868. IEEE, 2016.

[24] Qiyu Yang, Jiayi Yan, Xia Zhang, and Hekun Zhang. Decentralized detection for b5g massive mimo: When local computation meets iterative algorithm. Physical Communication, 51:101554, 2022.

[25] Victor Croisfelt, Taufik Abrão, Abolfazl Amiri, Elisabeth de Carvalho, and Petar Popovski. Decentralized design of fast iterative receivers for massive and extreme-large mimo systems. arXiv preprint arXiv:2107.11349, 2021.

[26] Shuo Li, Lixia Xiao, Chunlin He, Liuke Li, and Tao Jiang. Approximated expectation propagation assisted decentralized signal detection for uplink massive mimo-otfs systems. IEEE Transactions on Vehicular Technology, 2024.

[27] Zhilong Liu, Jiayi Zhang, Ziheng Liu, Hongyang Du, Zhe Wang, Dusit Niyato, Mohsen Guizani, and Bo Ai. Cell-free xl-mimo meets multi-agent reinforcement learning: Architectures, challenges, and future directions. IEEE Wireless Communications, 2024.

[28] Zakir Hussain Shaik and Erik G Larsson. Decentralized algorithms for out-of-system interference suppression in distributed mimo. IEEE Wireless Communications Letters, 2024.

[29] Zhenyu Zhang, Yuanyuan Dong, Keping Long, Xiyuan Wang, and Xiaoming Dai. Decentralized baseband processing with gaussian message passing detection for uplink massive mu-mimo systems. IEEE Transactions on Vehicular Technology, 71(2):2152–2157, 2021.

[30] Zhiyuan Zhai, Xiaojun Yuan, and Xin Wang. Decentralized federated learning via mimo over-the-air computation: Consensus analysis and performance optimization. IEEE Transactions on Wireless Communications, 2024.

[31] Hua Li, Yuanyuan Dong, Caihong Gong, Xiyuan Wang, and Xiaoming Dai. Decentralized groupwise expectation propagation detector for uplink massive mu-mimo systems. IEEE Internet of Things Journal, 10(6):5393–5405, 2022.

[32] Xiaotong Zhao, Xin Guan, Mian Li, and Qingjiang Shi. Decentralized linear mmse equalizer under colored noise for massive mimo systems. In 2021 IEEE Global Communications Conference (GLOBECOM), pages 01–06. IEEE, 2021.

[33] Kang Zheng, Hao Gao, Shuai Cui, Jiaheng Wang, and Yongming Huang. Decentralized recursive mmse equalizer for massive mimo systems. In 2022 IEEE 22nd International Conference on Communication Technology (ICCT), pages 152–156. IEEE, 2022.

[34] Soumyadeep Datta, Dheeraj Naidu Amudala, Ekant Sharma, Rohit Budhiraja, and Shivendra S Panwar. Full-duplex cell-free massive mimo systems: Analysis and decentralized optimization. IEEE Open Journal of the Communications Society, 3:31–50, 2021.

[35] JC De Luna Ducoing, Chathura Jayawardena, Marcin Filo, and Konstantinos Nikitopoulos. Towards 6g mimo systems, with massively-parallel non-linear processing. In 2021 IEEE International Mediterranean Conference on Communications and Networking (MeditCom), pages 300–305. IEEE, 2021.

[36] Abolfazl Amiri, Carles Navarro Manchón, and Elisabeth De Carvalho. Uncoordinated and decentralized processing in extra-large mimo arrays. IEEE Wireless Communications Letters, 11(1):81–85, 2021.

[37] Ke Wang Helmersson, Pål Frenger, and Anders Helmersson. Uplink d-mimo processing using kalman filter combining. In GLOBECOM 2022-2022 IEEE Global Communications Conference, pages 1703–1708. IEEE, 2022.

[38] Salim Ullah, Tuan Duy Anh Nguyen, and Akash Kumar. Energy-efficient low-latency signed multiplier for fpga-based hardware accelerators. IEEE Embedded Systems Letters, 13(2):41–44, 2020.

[39] Martin Kumm, Shahid Abbas, and Peter Zipf. An efficient softcore multiplier architecture for xilinx fpgas. In 2015 IEEE 22nd Symposium on Computer Arithmetic, pages 18–25. IEEE, 2015.

[40] Salim Ullah, Hendrik Schmidl, Siva Satyendra Sahoo, Semeen Rehman, and Akash Kumar. Area-optimized accurate and approximate softcore signed multiplier architectures. IEEE Transactions on Computers, 70(3):384–392, 2020.

[41] Salim Ullah, Semeen Rehman, Muhammad Shafique, and Akash Kumar. High-performance accurate and approximate multipliers for fpga-based hardware accelerators. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 41(2):211–224, 2021.

[42] Bharath Srinivas Prabakaran, Semeen Rehman, Muhammad Abdullah Hanif, Salim Ullah, Ghazal Mazaheri, Akash Kumar, and Muhammad Shafique. Demas: An efficient design methodology for building approximate adders for fpga-based systems. In 2018 Design, Automation & Test in Europe Conference & Exhibition (DATE), pages 917–920. IEEE, 2018.

[43] Sina Boroumand, Hadi P Afshar, and Philip Brisk. Approximate quaternary addition with the fast carry chains of fpgas. In 2018 Design, Automation & Test in Europe Conference & Exhibition (DATE), pages 577–580. IEEE, 2018.

[44] Sukanya Balasubramani, Uma Jagadeeshan, and Umapathi Krishnamoorthy. Performance optimized approximate multiplier architecture st-axm-based on statistical analysis and static compensation. Microelectronics Reliability, 151:115277, 2023.

[45] Suganthi Venkatachalam and Seok-Bum Ko. Design of power and area efficient approximate multipliers. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 25(5):1782–1786, 2017.

[46] Nguyen Van Toan and Jeong-Gun Lee. Fpga-based multi-level approximate multipliers for high-performance error-resilient applications. IEEE Access, 8:25481–25497, 2020.

[47] Haroon Waris, Chenghua Wang, Weiqiang Liu, and Fabrizio Lombardi. Axbms: Approximate radix-8 booth multipliers for high-performance fpga-based accelerators. IEEE Transactions on Circuits and Systems II: Express Briefs, 68(5):1566–1570, 2021.

[48] Zainab Aizaz and Kavita Khare. Asmpec: Approximate-sum based mapping of partial products with error correction for softcore multipliers on fpgas. IEEE Transactions on Circuits and Systems II: Express Briefs, 2023.

[49] Siva Satyendra Sahoo, Salim Ullah, Soumyo Bhattacharjee, and Akash Kumar. Axocs: Scaling fpga-based approximate operators using configuration supersampling. IEEE Transactions on Circuits and Systems I: Regular Papers, 2024.

[50] Harish Viswanathan and Preben E Mogensen. Communications in the 6g era. IEEE Access, 8:57063–57074, 2020.

[51] Hans Jakob Damsgaard, Aleksandr Ometov, and Jari Nurmi. Approximation opportunities in edge computing hardware: A systematic literature review. ACM Computing Surveys, 55(12):1–49, 2023.

[52] Hans Jakob Damsgaard, Aleksandr Ometov, Md Munjure Mowla, Adam

Flizikowski, and Jari Nurmi. Approximate computing in b5g and 6g wireless systems: A survey and future outlook. Computer Networks, page 109872, 2023.

[53] Yangcan Zhou, Zhiyu Chen, Jun Lin, and Zhongfeng Wang. A high-speed successive-cancellation decoder for polar codes using approximate computing. IEEE Transactions on Circuits and Systems II: Express Briefs, 66(2):227–231, 2018.

[54] Mingjie Hao, Ardalan Najafi, Alberto García-Ortiz, Ludwig Karsthof, Steffen Paul, and Jochen Rust. Reliability of an industrial wireless communication system using approximate units. In 2019 29th International Symposium on Power and Timing Modeling, Optimization and Simulation (PATMOS), pages 87–90. IEEE, 2019.

[55] Jie Xiao, Jianhao Hu, and Kaining Han. Low complexity expectation propagation detection for scma using approximate computing. In 2019 IEEE Global Communications Conference (GLOBECOM), pages 1–6. IEEE, 2019.

[56] Maryam Idrees, Mohammed Manzar Maqbool, Muhammad Khurram Bhatti, M Mahboob Ur Rahman, Rehan Hafiz, and Muhammad Shafique. An approximate-computing empowered green 6g downlink. Physical Communication, 49:101444, 2021.

[57] Xiang Ma, Haijian Sun, Rose Qingyang Hu, and Yi Qian. Approximate wireless communication for federated learning. arXiv preprint arXiv:2304.03359, 2023.

[58] Lorena Anghel, Mounir Benabdenbi, Alberto Bosio, Marcello Traiola, and Elena Ioana Vatajelu. Test and reliability in approximate computing. Journal of Electronic Testing, 34(4):375–387, 2018.

[59] Mark Wyse André Baixo Thierry Moreau, Bill Zorn Adrian Sampson James Bornholt, and Luis Ceze Mark Oskin. Mapping and modeling approximate computing techniques.

[60] M. A. Albreem, M. Juntti, and S. Shahabuddin. Massive MIMO Detection Techniques: A Survey. IEEE Communications Surveys Tutorials, 21(4):3109–3132, 2019.

[61] S. Yang and L. Hanzo. Fifty years of mimo detection: The road to large-scale mimos. IEEE Communications Surveys Tutorials, 17(4):1941–1988, 2015.

[62] E. Bjornson, L. Sanguinetti, H. Wymeersch, J. Hoydis, and T. Marzetta. Massive mimo is a reality—what is next?: Five promising research directions for antenna arrays. Digital Signal Processing, 94:3–20, 2019.

[63] K. Li, J. McNaney, C. Tarver, O. Castaneda, C. Jeon, J. R. Cavallaro, and

C. Studer. Design trade-offs for decentralized baseband processing in massive mu-mimo systems. pages 906–912, 2019.

[64] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson. Scaling up mimo: Opportunities and challenges with very large arrays. IEEE Signal Processing Magazine, 30(1):40–60, Jan 2013.

[65] K. Li, R. R. Sharan, Y. Chen, T. Goldstein, J. R. Cavallaro, and C. Studer. Decentralized baseband processing for massive mu-mimo systems. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 7(4):491–507, Dec 2017.

[66] Abhinav Kulkarni, Messaoud Ahmed Ouameur, and Daniel Massicotte. Hardware topologies for decentralized large-scale mimo detection using newton method. IEEE Transactions on Circuits and Systems I: Regular Papers, 68(9):3732–3745, 2021.

[67] Ian Kuon and Jonathan Rose. Measuring the Gap Between FPGAs and ASICs. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 26(2):203–215, 2007.

[68] Abhinav Kulkarni, Messaoud Ahmed Ouameur, and Daniel Massicotte. Logic cloning based approximate signed multiplication circuits for fpga. Microelectronics Journal, 145:106135, 2024.

[69] Abhinav Kulkarni, Messaoud Ahmed Ouameur, and Daniel Massicotte. Energy efficient wireless signal detection: A revisit through the lens of approximate computing. Electronics, 13(7), 2024.

[70] Bharath Srinivas Prabakaran, Vojtech Mrazek, Zdenek Vasicek, Lukas Sekanina, and Muhammad Shafique. Xel-fpgas: An end-to-end automated exploration framework for approximate accelerators in fpga-based systems. In 2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD), pages 1–9. IEEE, 2023.

[71] Ioannis P Chochliouros, Michail-Alexandros Kourtis, Anastasia S Spiliopoulou, Pavlos Lazaridis, Zaharias Zaharis, Charilaos Zarakovitis, and Anastasios Kourtis. Energy efficiency concerns and trends in future 5g network infrastructures. Energies, 14(17):5392, 2021.

[72] Emmanuel Bertin, Noël Crespi, and Thomas Magedanz. Shaping Future 6G Networks: Needs, Impacts, and Technologies. John Wiley & Sons, 2021.

Appendix A - Titre de l'annexe A

Approximate multiplication for general matrix-vector operation.

This section demonstrates the analysis of the effect of approximate multiplication on matrix-vector operations in the context of MIMO uplink signal detection. Unlike SISO systems, where signals are transmitted and received through a single antenna, MIMO systems involve simultaneous transmission and reception of multiple signals through multiple antennas. The interaction between these signals results in nonlinear effects such as inter-channel and spatial interference. Having multiple antennas at both the transmitter and receiver introduces a vast number of possible channel states. Each channel state corresponds to a different combination of fading coefficients, interference levels, and noise characteristics, posing challenges in deriving a closed-form expression for SER that accurately.

Quantitatively, the QoS of MIMO systems can be assessed in terms of metrics like SER, BER, FER or Mean Square Error (MSE). However, as closed form expression is challenging to obtain, an upper bound on the MSE is evaluated henceforth, which provides valuable insights for application of approximate multiplication for MIMO uplink signal detection. Accordingly, the concept of AMN is extended to MIMO uplink system and an upperbound on MSE is derived using the AMN.

For a MIMO uplink detection system with U transmit and B receiver antennas, channel matrix $\mathbf{H} \in \mathbb{C}^{B \times U}$ is represented as:

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1U} \\ h_{21} & h_{22} & \cdots & h_{2U} \\ \vdots & \vdots & \ddots & \vdots \\ h_{B1} & h_{B2} & \cdots & h_{BU} \end{bmatrix}$$

Hermitian matrix $\mathbf{H}^H \in \mathbb{C}^{U \times B}$ is represented as:

$$\mathbf{H}^H = \begin{bmatrix} h_{11}^* & h_{21}^* & \cdots & h_{B1}^* \\ h_{12}^* & h_{22}^* & \cdots & h_{B2}^* \\ \vdots & \vdots & \ddots & \vdots \\ h_{1U}^* & h_{2U}^* & \cdots & h_{BU}^* \end{bmatrix}$$

With transmit signal $\mathbf{x}$, the receive signal $\mathbf{y}$ is given by:

$$\mathbf{y} = S\mathbf{H}\mathbf{x} + \mathbf{w} \tag{A.1}$$

Detected signal after application of Matched Filter (MF) is obtained as:

$$\hat{\mathbf{x}} = \frac{1}{S}\mathbf{H}^H\mathbf{y} \tag{A.2}$$

However, after application of approximate multiplication at the receiver instead of accurate multiplication, the estimate of $\hat{\mathbf{x}}$ denoted as $\hat{\hat{\mathbf{x}}}$ is computed as:

$$\hat{\hat{\mathbf{x}}} = \frac{1}{S}\mathbf{H}^H \approx \mathbf{y} \tag{A.3}$$

$$\hat{\mathbf{x}} = \frac{1}{S}\begin{bmatrix} h_{11}^* & h_{21}^* & \cdots & h_{M1}^* \\ h_{12}^* & h_{22}^* & \cdots & h_{M2}^* \\ \vdots & \vdots & \ddots & \vdots \\ h_{1K}^* & h_{2K}^* & \cdots & h_{MK}^* \end{bmatrix} \propto \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = \frac{1}{S}\begin{bmatrix} h_{11}^* \propto y_1 + h_{21}^* \propto y_2 + \cdots + h_{M1}^* \propto y_M \\ h_{12}^* \propto y_1 + h_{22}^* \propto y_2 + \cdots + h_{M2}^* \propto y_M \\ \vdots \\ h_{1K}^* \propto y_1 + h_{2K}^* \propto y_2 + \cdots + h_{MK}^* \propto y_M \end{bmatrix}$$

$$= \frac{1}{S}\begin{bmatrix} h_{11}^* y_1 + h_{21}^* y_2 + \cdots + h_{M1}^* y_M \\ h_{12}^* y_1 + h_{22}^* y_2 + \cdots + h_{M2}^* y_M \\ \vdots \\ h_{1K}^* y_1 + h_{2K}^* y_2 + \cdots + h_{MK}^* y_M \end{bmatrix} + \frac{1}{S}\begin{bmatrix} -2M\nu_{\hat{P}} \\ -2M\nu_{\hat{P}} \\ \vdots \\ -2M\nu_{\hat{P}} \end{bmatrix} = \frac{1}{S}\mathbf{H}^H\mathbf{y} + \frac{1}{S}\begin{bmatrix} -2M\nu_{\hat{P}} \\ -2M\nu_{\hat{P}} \\ \vdots \\ -2M\nu_{\hat{P}} \end{bmatrix} \quad \text{(A.4)}$$

When AMN is used, the received signal is $\bar{\mathbf{y}}$ given as:

$$\bar{\mathbf{y}} = S\mathbf{H}\mathbf{x} + \mathbf{w} + \delta = \mathbf{y} + \delta \quad \text{(A.5)}$$

$$\hat{\bar{\mathbf{x}}} = \frac{1}{S}\mathbf{H}^H\bar{\mathbf{y}} = \frac{1}{S}\mathbf{H}^H(\mathbf{y} + \delta) = \frac{1}{S}\mathbf{H}^H\mathbf{y} + \frac{1}{S}\mathbf{H}^H\delta = \hat{\mathbf{x}} + \frac{1}{S}\mathbf{H}^H\delta \quad \text{(A.6)}$$

Comparing eq. (A.4) and (A.6),

$$\mathbf{H}^H \delta = \begin{bmatrix} -2M\nu_{\hat{P}} \\ -2M\nu_{\hat{P}} \\ \vdots \\ -2M\nu_{\hat{P}} \end{bmatrix} \implies \delta = (\mathbf{H}^H)^{-1} \begin{bmatrix} -2M\nu_{\hat{P}} \\ -2M\nu_{\hat{P}} \\ \vdots \\ -2M\nu_{\hat{P}} \end{bmatrix} \tag{A.7}$$

The Degradation MSE is defined as the D-MSE between the output by detection scheme and that by employing the approximate multiplication.

$$\begin{aligned} \text{D-MSE} &\leq \frac{1}{K}\mathbb{E}\left\{(\hat{\mathbf{x}}-\hat{\hat{\mathbf{x}}})^T(\hat{\mathbf{x}}-\hat{\hat{\mathbf{x}}})\right\} \\ &\leq \frac{1}{K}\mathbb{E}\left\{\sum_{i=1}^{K}(\hat{x}_i-\hat{\hat{x}}_i)^2\right\} \\ &\leq \frac{1}{K}\mathbb{E}\left\{\sum_{i=1}^{K}\left(\frac{2M\nu_{\hat{P}}}{S}\right)^2\right\} \\ &\leq \frac{4M^2\nu_{\hat{P}}^2}{S^2} \end{aligned}$$

$$\frac{S}{2M}\sqrt{\text{D-MSE}} \leq \nu_{\hat{P}} \tag{A.8}$$

The above equation provides a lower bound on the value of $\nu_{\hat{P}}$, given a specific D-MSE to be tolerated in the system. The availability of $\nu_{\hat{P}}$ aids in the choice of appropriate TM configuration based on the required D-MSE.

Appendix B - Titre de l'annexe B

Hardware Topologies for Decentralized Large-Scale MIMO Detection Using Newton Method

Author: Abhinav Kulkarni

Publication: IEEE Transactions on Circuits and Systems I: Regular Papers

Publisher: IEEE

Date: September 2021

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BACK CLOSE WINDOW