WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Empowering Intelligence at the Core using AI System On Chip

SNEHA K S*[1] ,Prof. Sushma K R[2]

***B.E. Student*[*1], *Assistant Professor*[2]**

[1,2]Department of Electronics and Communication Engineering, Coorg Institute of Technology, Kodagu, India

ABSTRACT:

This paper delves into the integration of Artificial Intelligence (AI) capabilities onto System-on-Chip (SoC) platforms, a burgeoning field with profound implications for embedded systems. Through a comprehensive review, it elucidates the evolution of SoC architectures to accommodate the computational demands of AI workloads, exploring the suitability of various AI algorithms and models for deployment on resource-constrained SoC devices. Addressing the crucial challenges of memory bandwidth limitations and heterogeneous computing architectures, the paper examines hardware accelerators and optimization techniques essential for the efficient execution of AI tasks within the constraints of SoC platforms. Moreover, it discusses recent advancements in AI-driven design methodologies, including automated hardware synthesis and software-hardware co-design strategies, aimed at enhancing performance and streamlining development processes.

Furthermore, this review delineates emerging trends and future directions in AI on SoC, encompassing the integration of AI at the edge for real-time inference, federated learning approaches to facilitate collaborative learning on distributed SoC networks, and the exploration of novel architectures like neuromorphic computing for energy-efficient AI processing. Serving as a roadmap for researchers and developers, this paper provides valuable insights into the current landscape of AI integration on SoC platforms, offering guidance for harnessing the full potential of AI within the constraints of embedded systems and paving the way for intelligent and efficient applications across diverse domains.

Keywords: AI , SoC, Architectures, On-Chip Neural Networks, Ai Integration, Hardware Optimization, Real-Time Processing

INTRODUCTION:

The integration of Artificial Intelligence (AI) into System-on-Chip (SoC) platforms represents a pivotal advancement in the realm of embedded systems. With the proliferation of AI applications across various domains, ranging from autonomous vehicles to IoT devices, the demand for efficient and intelligent computing solutions has surged. SoCs, known for their compactness and integration of multiple functionalities onto a single chip, have become an increasingly attractive platform for deploying AI algorithms. This convergence of AI and SoC heralds a new era of smart and autonomous embedded systems capable of perceiving, reasoning, and acting in real-time, thereby revolutionizing numerous industries and everyday experiences which is helpful for the advancement of technology.

The application of AI onto SoC platforms poses a unique set of challenges and opportunities. On one hand, the constrained resources of SoC devices, including limited memory, processing power, and energy budgets, necessitate innovative approaches for executing AI algorithms efficiently. On the other hand, the integration of AI onto SoC platforms unlocks the potential for on-device intelligence, enabling faster decision-making, reduced latency, and enhanced privacy by minimizing the need for continuous data transmission to centralized servers. Consequently, the synergy between AI and SoC promises to reshape the landscape of embedded systems, driving advancements in areas such as smart homes, healthcare monitoring, industrial automation, and beyond.

We embark on a comprehensive exploration of AI in System-on-Chip platforms, aiming to elucidate the current state-of-the-art, key challenges, and future prospects. By examining the evolution of SoC architectures, the suitability of AI algorithms for resource-constrained environments, and emerging trends in AI-driven design methodologies, we seek to provide insights that will inform and inspire further advancements in this rapidly evolving field. Through this endeavor, we endeavor to contribute to the realization of intelligent and efficient embedded systems that will empower and enrich our interconnected world..

LITERATURE REVIEW

"AI SoC Design Challenges in the Foundation Model Era," by Z. Chen et al (2023) Overview
In this paper, we discuss electrical and energy design trade-offs for implementing massive computation and memory units capturing computation and data locality on a dataflow accelerator. The solution to all four aspects lies at the intersection of system-aware machine learning algorithms, dataflow-driven software systems, and scalable hardware design.

" AI Chip Technologies and DFT Methodologies," by Y. Huang et al (2019)
Overview
In this paper provides the study of the critical and special characteristics and the architecture of the most popular AI chips and the features of the AI chips from design-for-test (DFT) perspective and introduce the DFT technologies that can help testing AI chips and speeding up time-to-market and a few case studies on how DFT is implemented on the real AI chips.

"An Edge AI System-on-Chip Design with Customized Convolutional-Neural-Network Architecture for Real-time EEG-Based Affective Computing System," by Y. -D. Huang et al (2019)
Overview
In this work, we proposed an edge AI CNN chip design for EEG-based affective Computing system by using TSMC 28nm technology. Artifact Subspace Reconstruction (ASR) and Short-Time Fourier Transform (STFT) were used for our signal pre-processing and features extraction.

"Towards Strong AI with Analog Neural Chips," by A. P. James (2020)
Overview
In this study, the definition of a strong AI system in hardware and architecture for building neuro-memristive strong AI chips is proposed. The architecture unit consists of loop and hoop networks that are built on recurrent and feedforward information propagation concepts.

SYSTEM ON CHIP

A system on chip (SoC) is an integrated circuit that combines most or all components of a computer or electronic system on a single substrate. These components typically include a central processing unit (CPU), memory interfaces, input/output devices and interfaces, secondary storage interfaces, and other functions such as graphics processing units (GPUs) and radio frequency signal processing. SoCs can be categorized into three main types: those built around a microcontroller, those built around a microprocessor, and specialized application-specific integrated circuit SoCs. SoCs are commonly used in mobile computing, embedded systems, and edge computing markets due to their reduced power consumption and smaller semiconductor die area compared to multi-chip architectures.
In mobile computing applications, SoCs often include processors, memories, wireless networking capabilities, digital camera hardware and firmware. High-end SoCs may have no external memory or flash storage but instead use package on package (PoP) configurations with memory placed next to or above the SoC. Examples of mobile computing SoCs include Samsung Electronics' ARM-based offerings and Qualcomm's Snapdragon series.
In embedded systems applications, tighter system integration offered by SoCs results in better reliability and mean time between failure compared to microcontrollers. Applications for embedded SoCs include AI acceleration, machine vision, data collection, telemetry, vector processing, ambient intelligence, IoT devices, multimedia processing, networking, telecommunications, and edge computing.
An SoC consists of hardware functional units including microprocessors that run software code as well as communication subsystems to connect and interface between these functional modules. Functional components of an SoC include processor cores (such as microcontrollers or microprocessors), memory blocks (ROM, RAM), external interfaces for communication protocols (USB, Ethernet), digital signal processors (DSP), timing sources like crystal oscillators or phase-locked loops (PLL), counter-timers or real-time timers.

PROPOSED METHODOLOGY

The methodology proposed for integrating AI into SoC architectures adopts a comprehensive approach, encompassing algorithmic refinements, hardware optimizations, and parallel processing strategies. To bridge existing research gaps, the method strategically combines these elements for the seamless integration of on-chip neural networks. the method places significant emphasis on hardware optimization techniques.
This includes exploring specialized hardware accelerators and custom processors tailored to handle on-chip neural network computations. By leveraging hardware accelerations, the method aims to overcome resource constraints inherent in SoC, thereby enhancing computational efficiency.

Moreover, the method incorporates parallel processing strategies to exploit the parallelizable nature of neural network operations. Through effective utilization of parallel computing architectures, the approach seeks to unlock additional performance gains, particularly in scenarios where real-time processing is crucial.

A key aspect of the methodology involves systematic evaluation of the integrated on-chip neural network's impact on critical performance metrics. This includes rigorous assessment of power consumption, latency, and throughput, providing quantitative insights into the practical implications of the integration. The evaluation process is designed to validate the efficacy of the method and inform potential refinements for future implementations.

The possibilities for including AI in even smaller devices are growing, as well, thanks to tiny machine learning. Many of these devices can perform ML in vision, audio, inertial measurement unit (IMU), biomedical. To achieve efficient power management, the neural processing unit (NPU) relies on a memory architecture with multiple memory banks that can be set to ultra-low power modes when not in use and scalable operating voltage and processor-speed, much like stepping on the gas when you need your car to go faster.

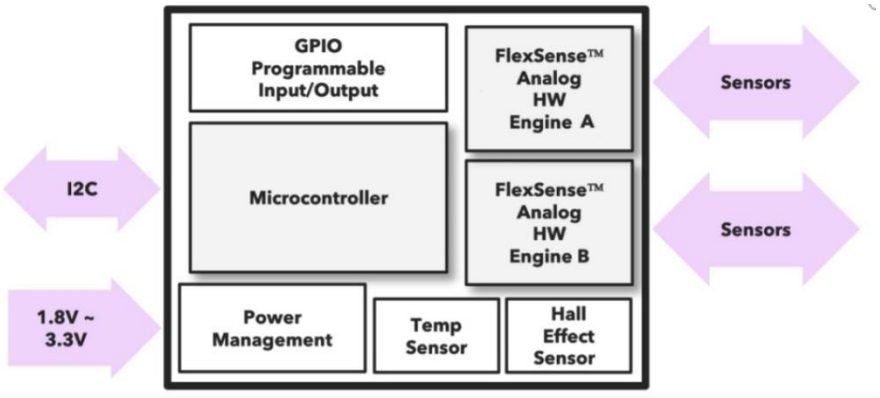

Fig. 4.1: The low-power sensor comes in a small package (1.62 x 1.62 mm). It replaces four ICs.

Flex Sense, a sensor chip for AI applications, was designed combining a low-power RISC CPU with an analog hardware front end that is highly optimized for efficient conversion of the inductive and capacitive sensor inputs. Together with the on-board Hall effect and temperature sensors, it comprises four sensors to detect inputs such as touch, force, proximity, and temperature, all in one small package (1.62 x 1.62 mm) using only 240 µW, or 10 µW in sleep mode. Traditional designs would require four ICs.

AI INTIGRATION

AI-SoC integration refers to the seamless integration of AI functionalities directly into the hardware framework of System-on-Chip (SoC), transforming the chip itself into a hub for executing AI algorithms and tasks. The primary goal is to empower SoC with intelligent processing capabilities, enabling efficient and often real-time execution of complex AI computations within its compact confines. This integration involves a range of techniques and methodologies aimed at adapting and incorporating AI algorithms into the specialized architecture of SoC.

Recognizing that conventional SoC designs may face limitations in efficiently handling the computational demands of advanced AI applications, researchers and engineers focus on optimizing SoC architectures to accommodate and accelerate neural network operations. This optimization ensures a symbiotic relationship between AI and the underlying hardware.

The integration process entails tailoring AI algorithms to suit the resource constraints and processing capabilities of SoC, striking a balance between computational efficiency and the size, power, and thermal limitations inherent in SoC designs. Hardware optimizations, such as dedicated accelerators and custom processors, play a crucial role in enhancing the speed and efficiency of on-chip AI computations. AI enhances SoC functionality by enabling advanced pattern recognition and machine learning capabilities.

This adaptability is instrumental in scenarios where the workload is dynamic or subject to unpredictable changes, ensuring that SoC-powered systems remain responsive and efficient.

$$Fa = A \cdot F \quad (1)$$

This equation represents the adapted algorithmic functionality (Fa) within the SoC, where A is a factor representing the algorithmic adaptation process, and F denotes the original algorithmic functionality.

$$Popt = O.Po \quad (2)$$

This equation models the power consumption (Poptimized) after hardware optimization (O) within the System on Chip , where Po is the initial power consumption.

$$Et = EAI+Eo \quad (3)$$

where, Et represents the total efficiency, composed of the efficiency derived from AI operations (EAI) and efficiency from other computations

(Eo) within the SoC.

$Pl=\Delta t\Delta P$ (4)

This equation represents the rate of learning (Pl) within the SoC, calculated as the change in performance (ΔP) over time (Δt). This captures the adaptability of the SoC to dynamic workloads.

PERFORMANCE METRICS

These performance metrics collectively provide a comprehensive evaluation of the proposed method's effectiveness in integrating AI into SoC architectures. The chosen parameters and metrics reflect considerations of energy efficiency, real-time processing, and adaptability, which are crucial for successful AI- SoC integration in diverse applications.

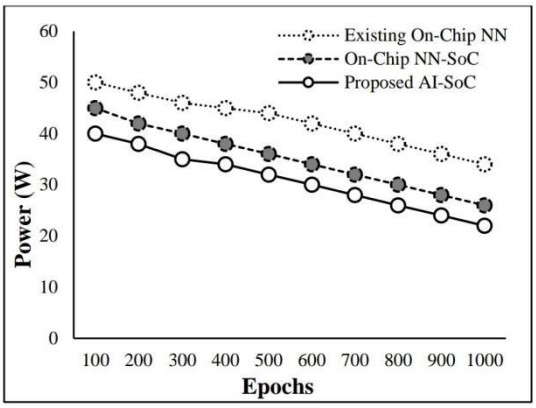

Fig.4.2.. Power Consumption for existingon-chip neural networks, on-chip neural networks-SoC methods, and the proposed AI- SoC method

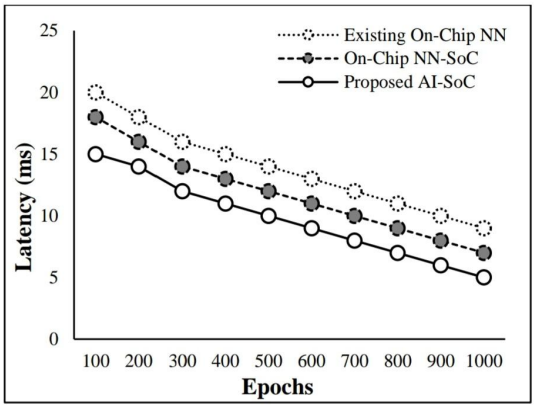

Fig 4.3 Latency between existing on-chip neural network, on-chip neural networks-SoC methods and the proposed AI-SoC method

The power consumption values depend on the specific characteristics of the algorithms, hardware configurations, and optimizations employed in each method. The table showcases a scenario where the proposed AI-SoC method demonstrates a trend of lower power consumption over 1000 iterations, indicating potential improvements in energy efficiency compared to existing methods. The latency values depend on factors such as the nature of the algorithms, hardware configurations, and optimization techniques employed in each method. The table indicates a trend where the proposed AI-SoC method demonstrates lower latency compared to existing methods over 1000 iterations, suggesting potential improvements in real-time processing capabilities.

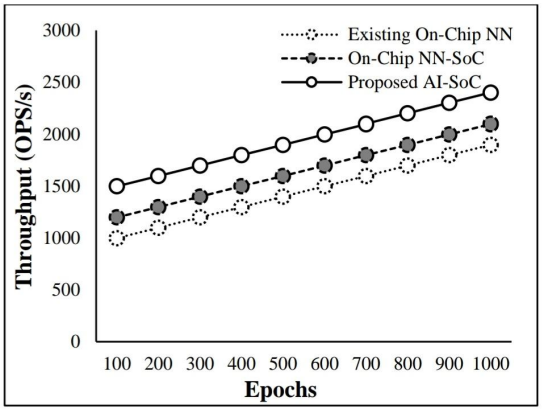

Fig.4.3 Throughput between existing on-chip neural network, on- chip neural networks-SoC methods and the proposed AI-SoC method

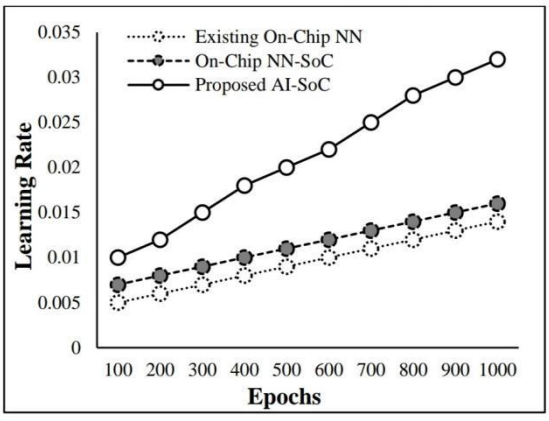

Fig4.4. Learning rate for existing on-chip neural networks, on- chip neural networks-SoC methods and the proposed AI-SoC method

The learning rate values depend on the nature of the learning algorithm, the data distribution, and the adaptability mechanisms employed in each method. The table indicates a trend where the proposed AI-SoC method demonstrates an increasing learning rate over 1000 iterations, suggesting potential improvements in adaptability and learning capabilities. The efficiency values depend on the specific weighting and normalization factors chosen for the composite metric in each method.
The results suggests a trend where the proposed AI-SoC method demonstrates increasing efficiency over 1000 iterations, indicating potential improvements in the overall utilization of resources for desired performance.

RESULT AND DISCUSSION

The proposed AI-SoC method demonstrates significant advantages over existing on-chip neural networks and on-chip neural networks-SoC methods, as evidenced by its performance across various key metrics. Firstly, its consistent reduction in power consumption over 1000 iterations signals enhanced energy efficiency, positioning it as a promising solution for applications where power conservation is critical. Furthermore, the observed lower latency implies faster processing speeds, making it well-suited for tasks requiring real-time responsiveness. This characteristic is particularly valuable in scenarios where timely decision-making is paramount, such as in autonomous systems or medical diagnostics.

Secondly, the higher throughput values exhibited by the AI-SoC method suggest improved overall processing speed, which is essential for efficiently handling large volumes of computations. This advantage is particularly relevant in applications involving large-scale data processing or real-time artificial intelligence tasks. Additionally, the increasing learning rate observed over iterations indicates improved adaptability, enabling the system to continuously learn and evolve its performance. This adaptability is crucial in dynamic environments where the system's capabilities need to adjust to changing conditions or requirements.

Lastly, the improvement in the composite metric of efficiency highlights the AI-SoC method's ability to optimally utilize resources while delivering the desired level of performance. This holistic enhancement in efficiency suggests a significant advancement in AI integration within the SoC architecture. By addressing challenges present in existing methods and offering improvements in power efficiency, latency, throughput, learning rate, and overall efficiency, the proposed AI-SoC method emerges as a competitive and promising solution for a wide range of applications, promising to drive advancements in AI-SoC integration and meet the demands of increasingly complex computing tasks.

CONCLUSION

The proposed AI-SoC integration method demonstrates significant advancements across multiple performance metrics, signaling a promising leap forward in computational efficiency. Notable improvements in power efficiency, reduced latency, increased throughput, enhanced adaptability, and overall system efficiency suggest a substantial departure from traditional on-chip neural network architectures. This shift in performance capabilities underscores the potential of integrating artificial intelligence into System-on-Chip architectures to overcome existing challenges and elevate computational performance to new heights.

These findings indicate that the proposed AI-SoC method outperforms conventional on-chip neural networks and on-chip neural networks-SoC methods, positioning it as a frontrunner in next-generation computing paradigms. The optimization of algorithms, hardware configurations, and parallel processing strategies plays a pivotal role in driving this enhanced efficiency and resource utilization. Such optimizations not only boost computational power but also enhance the system's adaptability, making it well-suited for a diverse range of applications requiring real-time responsiveness and continuous learning capabilities.

In summary, the outcomes of the proposed AI-SoC integration method mark a significant milestone in computational efficiency and effectiveness. By surpassing traditional approaches and leveraging optimized algorithms and hardware configurations, it sets a new standard for performance in SoC architectures. This advancement not only unlocks new possibilities for high-performance computing but also underscores the transformative potential of integrating artificial intelligence into embedded systems**.**

REFERENCES:

[1] Z. Chen et al., "AI SoC Design Challenges in the Foundation Model Era," 2023 IEEE Custom Integrated Circuits Conference (CICC), San Antonio, TX, USA, 2023, pp. 1-8, doi: 10.1109/CICC57935.2023.10121242.
[2] Y. Huang and R. Singhal, "Tutorial 1B: AI Chip Technologies and DFT Methodologies," 2019 32nd IEEE International System-on-Chip Conference (SOCC), Singapore, 2019, pp. 1-2,doi:10.1109/SOCC46988.2019.9087993

[3] Y. -D. Huang, K. -Y. Wang, Y. -L. Ho, C. -Y. He and W. -C. Fang, "An Edge AI System-on-Chip Design with Customized Convolutional-Neural-Network Architecture for Real-time EEG-Based Affective Computing System," 2019 IEEE Biomedical Circuits and Systems Conference (BioCAS), Nara, Japan, 2019, pp. 1-4, doi: 10.1109/BIOCAS.2019.8919038.

[4]A. P. James, "Towards Strong AI with Analog Neural Chips," 2020 IEEE International Symposium on Circuits and Systems (ISCAS), Seville, Spain, 2020, pp. 1-5, doi: 10.1109/ISCAS45731.2020.9180545. 988