http://www.jctjournals.com

ISSN(ONLINE):2278 - 3814

Volume-13, Issue-10, October 2024

JOURNAL OF COMPUTING TECHNOLOGIES (JCT)

International Journal

Page Number: 01-04

A Review on Digital Design and Optimization of VLSI Architecture

[1]Anubhav Chauhan, [2]Prof. Devendra Patle

[1]M.Tech Scholar VLSI , [2]Assistant Professor,

[1,2] Department of Electronics and communication

[1, 2] School of Engineering, SSSUTMS, Sehore M.P., India

Email id *anubhavchauhan591@gmail.com*

Abstract — This literature survey comprehensively explores the current state-of-the-art research and advancements in Very Large-Scale Integration (VLSI) design specifically tailored for Artificial Intelligence (AI) applications. The survey delves into existing methodologies, architectures, and technologies employed in designing efficient AI hardware. Key areas of focus include: Exploration of existing research: A thorough examination of published research papers, conference proceedings, and industry reports on VLSI design for AI accelerators. Identification of key trends and challenges: Identifying emerging trends, research gaps, and the major challenges faced in designing and implementing high-performance AI hardware .Analysis of different approaches: Analysis of various approaches to VLSI design for AI, including different architectural styles, circuit-level optimizations, and emerging technologies. Comparison of different AI accelerators: Comparison of the performance, efficiency, and suitability of different AI accelerator types, such as TPUs, GPUs, FPGAs, and ASICs, for various AI workloads. This literature survey aims to provide a comprehensive overview of the current landscape of VLSI design for AI and serve as a foundation for further research and development in this critical area. Note: This abstract is a concise summary of the literature survey section. You may need to adjust it further based on the specific focus and scope of your research paper. I hope this helps! Let me know if you have any other questions.

Keywords—AI, Artificial Intelligence, Machine Learning, Deep Learning, VLSI, Very Large-Scale Integration.

I. INTRODUCTION

Artificial Intelligence (AI) and Machine Learning (ML) have become essential technologies that are transforming various industries, from healthcare to finance, automotive, and entertainment. The exponential growth of data and the need for intelligent systems have driven the demand for high-performance hardware capable of running complex AI models efficiently. In the early days of AI development, traditional processors like Central Processing Units (CPUs) were sufficient for basic computations. However, as the complexity of AI algorithms increased, particularly in deep learning models, more specialized hardware became necessary.

The architecture and design of processors for AI workloads differ significantly from general-purpose processors. For AI tasks, the focus is on parallelism, low-latency data access, and high-speed computation. This has led to the development of AI accelerators, which are specialized hardware units designed to accelerate the execution of AI tasks such as matrix multiplications, convolutions, and activations that are common in neural networks.

II. Fundamentals of VLSI Design

Principles of VLSI Design in Digital Systems VLSI design is guided by several core principles aimed at maximizing the performance of digital systems while minimizing power consumption, area, and cost. The design process typically follows a hierarchical structure, starting with the most abstract functional requirements and gradually progressing toward physical implementation

A. Hierarchy

VLSI design employs a hierarchical approach to manage the complexity of digital systems. This approach divides the design process into different levels, from the system level down to the transistor level. The main levels of hierarchy are:

- **System level:** Specifies the overall function and architecture of the chip.
- **Register-Transfer Level (RTL):** Describes how data is transferred between registers in response to clock signals.
- **Gate level:** Details the logic gates used to implement the RTL description.
- **Transistor level:** Defines the behaviour of individual transistors and how they are connected.

B. Modularity

Modularity refers to breaking down a large design into smaller, reusable modules that can be independently developed and tested. This makes the design process more manageable and allows for the reuse of components, reducing time to market. In AI accelerators, modularity enables the integration of various functional blocks such as processing cores, memory controllers, and input/output interfaces.

C. Regularity

Regularity in VLSI design involves the use of repetitive structures to simplify the design process. For example, in AI accelerators, systolic arrays, which consist of regular, repetitive processing elements, are often used to perform matrix multiplications efficiently. Regular structures are easier to design and test, and they can improve the manufacturability of the chip.

II. LITERATURE REVIEW

Jyotishman Saikia et. Al., 2024, For state-of-the-art artificial intelligence (AI) accelerators, there have been large advances in both all-digital and analog/mixed-signal circuit-based designs. This article presents a practical overview and comparison of recent digital and analog AI accelerators. We first introduce hardware-efficient AI algorithms, which have been targeted for many AI hardware designs. Next, we present a survey of 1) all-digital AI accelerators, including designs with new dataflow, low precision, and sparsity, and 2) analog/mixed-signal AI accelerators featuring switch-capacitor circuits and in-memory computing (IMC) with ADCs. Recent advances of AI accelerators in both digital and analog design approaches are summarized, and emerging AI accelerator designs are discussed **[1].**

Manar Abu Talib et. Al., Artificial intelligence (AI) tools play a significant role in the recent evolution of smart systems. AI solutions are pushing towards a significant shift in many fields such as healthcare, autonomous airplanes and vehicles, security, marketing customer profiling and other diverse areas. One of the main challenges hindering the AI potential is the demand for high-performance computation resources. Recently, hardware accelerators are developed in order to provide the needed computational power for the AI and ML tools. In the literature, hardware accelerators are built using FPGAs, GPUs and ASICs to accelerate computationally intensive tasks. These accelerators provide high-performance hardware while preserving the required accuracy **[2].**

Raju Machupalli et. Al., Deep neural networks (DNNs) have become an essential tool in artificial intelligence, with a wide range of applications such as computer vision, medical diagnosis, security, robotics, and autonomous vehicle. The DNNs deliver the state-of-the-art performance in many applications. The complexity of the DNN models generally increases with application complexity and deployment of complex DNN models requires high computational power. General-purpose processors are unable to process complex DNNs within the required throughput, latency, and power budget. Therefore, domain-specific hardware accelerators are required to provide high computational resources with superior energy efficiency and throughput within a small chip area. In this paper, existing DNN hardware accelerators are reviewed and classified based on the optimization techniques used in their implementations. Each optimization technique generally improves one or more specific performance parameter(s) **[3].**

MS Akhoon et. Al., Artificial intelligence (AI) application is accelerated by high-performance VLSI architectures, which allow for real-time inference, analysis, and decision-making across a wide range of disciplines. The design, development, and implementation of VLSI architectures for AI and ML applications are examined in this paper, with an emphasis on scalability, efficiency, and practicality. The study's primary goals are to examine architectural paradigms, optimization strategies, energy-efficient design concepts, performance evaluation approaches, and practical uses of high-performance VLSI architectures for AI and ML. A thorough analysis of the body of research, case studies, and policy implications about VLSI design for AI and ML applications are all part of the methodology. Principal discoveries emphasize the variety of architectural paradigms, optimization strategies, and practical uses of high-performance VLSI architectures, along with their implementation difficulties and policy ramifications. The significance of ethical deliberations, adherence to regulations, and international cooperation in guaranteeing the conscientious and fair application of artificial intelligence and machine learning is highlighted by policy ramifications. By offering insights into the design, optimization, deployment, and policy implications of high-performance VLSI architectures for AI and ML applications, this study advances our collective

understanding of these technologies and the field of AI-driven technologies **[4]**.

Archika Malhotra et. Al., The Very Large Scale Integration (VLSI) industry has started adapting the Artificial Intelligence (AI) techniques in design automation as it provides the opportunity to transform the whole chip design methodology. It has been seen that in System-On-Chip (SoC), in order to add ML algorithms to increase its efficiency, there is a need to reduce the existing power consumption of the hardware as well. Hence, this makes AI an integral part of the VLSI industry. An extensive review has been conducted on various aspects of AI in the field of VLSI. This paper throws light on how AI has marked its way on various subfields of VLSI, namely, analog, digital and physical design. We have also taken into account the recent machine learning and deep learning techniques incorporated in VLSI **[5].**

Yiran Chen et. Al., Due to the availability of big data and the rapid growth of computing power, artificial intelligence (AI) has regained tremendous attention and investment. Machine learning (ML) approaches have been successfully applied to solve many problems in academia and in industry. Although the explosion of big data applications is driving the development of ML, it also imposes severe challenges of data processing speed and scalability on conventional computer systems. Computing platforms that are dedicatedly designed for AI applications have been considered, ranging from a complement to von Neumann platforms to a "must-have" and stand-alone technical solution. These platforms, which belong to a larger category named "domain-specific computing," focus on specific customization for AI. In this article, we focus on summarizing the recent advances in accelerator designs for deep neural networks (DNNs)—that is, DNN accelerators **[6].**

Prashray Nagar et. Al., The capabilities of artificial intelligence (AI) and machine learning (ML) algorithms are constantly expanding, necessitating efficient and high-performance hardware systems. We have investigated the creation of hardware accelerators based on VLSI that are intended to effectively manage the heavy workloads of machine learning jobs, also explored low-power VLSI architectures that preserve computing capabilities while reducing power consumption to solve energy efficiency issues in AI and ML systems. To balance performance and energy utilization, power management strategies and circuit design improvements are analysed. The study emphasizes hardware-software co-design techniques, considering the integration of VLSI-based hardware accelerators with software frameworks to obtain optimal performance and flexibility, to address the complexity and scalability of AI and ML systems. We also examined the cutting-edge VLSI technologies that have the potential to support powerful AI and ML applications. The speed and effectiveness of AI and ML algorithms could be improved significantly by these technologies, which include neuromorphic computing, approximation computing, and in-memory computing **[7]**.

Arun SADANAND Tigadi et. Al., VLSI creates AIML applications, emphasising their impact on various stages of the design flow. The paper starts with an overview of VLSI design and its various stages, then moves on to an overview of AIML concepts and their relevance to VLSI design. The paper then delves into the use of AIML at various stages of the design process. In addition, the paper discusses the challenges and limitations of using AIML in VLSI design, such as data availability and scalability. Overall, the paper emphasises the power of AIML techniques in VLSI design and how they can significantly improve the performance and efficiency of modern integrated circuits. Many problems in various fields have been solved by artificial intelligence (AI). The AI principle is based on human intelligence, which is interpreted **[8].**

IV. VLSI Architectures for AI Accelerators

A. Parallel Processing in VLSI Architectures - Parallelism is a cornerstone of VLSI design for AI accelerators. AI workloads, especially those involving neural networks and deep learning models, require the processing of vast amounts of data concurrently. VLSI architectures leverage multiple levels of parallelism to maximize throughput and minimize latency.

B. Instruction-Level Parallelism (ILP)- Instruction-level parallelism (ILP) refers to the ability of a processor to execute multiple instructions simultaneously. In VLSI architectures, ILP is achieved by pipelining, where different stages of instruction execution (e.g., fetch, decode, execute) are overlapped. Modern AI accelerators, particularly those based on vector processing, use ILP to accelerate operations like matrix multiplications and convolutions.

C. Data-Level Parallelism (DLP)- Data-level parallelism (DLP) involves performing the same operation on multiple data points simultaneously. AI accelerators rely heavily on DLP, as many AI tasks involve processing large datasets. VLSI architectures implement DLP through techniques like SIMD (Single Instruction, Multiple Data) and systolic arrays, where large matrices are processed in parallel.

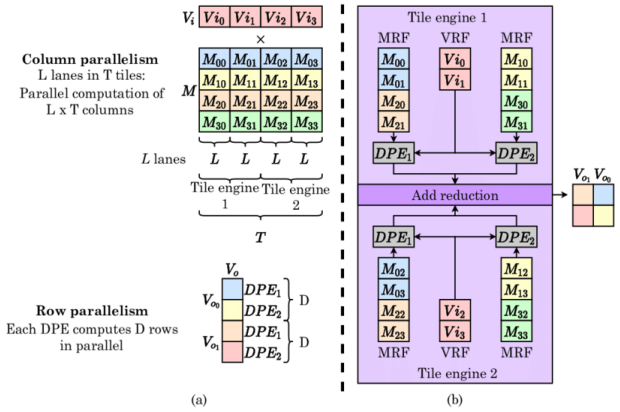

Figure1: Parallelism-in-vector-matrix-multiplication

V. CONCLUSION

VLSI technology has fundamentally transformed the landscape of AI accelerators, enabling the rapid growth and deployment of AI across various sectors. The principles and innovations discussed in this Research paper underscore the critical role of VLSI in advancing AI hardware and driving the development of more powerful, efficient, and scalable accelerators. As we look to the future, it is clear that continued innovation in VLSI design will be key to unlocking the full potential of AI, empowering new applications and driving breakthroughs in machine learning, robotics, autonomous systems, and beyond.

References

[1] Bishnoi L, Narayan Singh S (2018) Artificial intelligence techniques used in medical sciences: a review. In: 8th International Conference on Cloud Computing, Data Science and Engineering (Confluence), pp 106–113.

[2] Parker DS (1989) Integrating AI and DBMS through stream processing. In: Proceedings of Fifth International Conference on Data Engineering.

[3] Rao Q, Frtunikj J (2018) Deep learning for self-driving cars. In: Proceedings of the 1st International Workshop on Software Engineering for AI in Autonomous Systems—SEFAIS ’18.

[4] Misra J, Saha I (2010) Artificial neural networks in hardware: a survey of two decades of progress. Neurocomputing 74(1–3):239–255.

[5] Baji T (2018) Evolution of the GPU device widely used in AI and massive parallel processing. In: IEEE 2nd Electron Devices Technology and Manufacturing Conference (EDTM).

[6] Shawahna A, Sait SM, El-Maleh A (2019) FPGA-based accelerators of deep learning networks for learning and classification: a review. IEEE Access 7:7823–7859.

[7] Mittal S (2018) A survey of FPGA-based accelerators for convolutional neural networks. Neural Comput Appl 32(4):1109–1139.

[8] Nurvitadhi E, Venkatesh G, Sim J, Marr D, Huang R, Ong Gee Hock J, Liew YT, Srivatsan K, Moss D, Subhaschandra S, Boudoukh G (2017) Can FPGAs beat GPUs in accelerating next-generation deep neural networks? In: Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays—FPGA ’17.

[9] Lacey G, Taylor G, Areibi S (2016) Deep learning on FPGAs: past, present, and future, pp 1–8.

[10] Kryjak T, Komorkiewicz M, Gorgon M (2012) FPGA implementation of real-time headshoulder detection using local binary patterns, SVM and foreground object detection. In: Conference on Design and Architectures for Signal and Image Processing (DASIP), pp 1–8.

[11] Jain V, Patel D (2016) A GPU based implementation of robust face detection system. Procedia Comput Sci 87:156–163.

[12] Lescano G, Santana P, Costaguta R (2017) Analysis of a GPU implementation of Viola-Jones’ algorithm for features selection. J Comput Sci Technol 17(1):68–73.

[13] Svab J, Krajnik T, Faigl J, Preucil L (2009) FPGA based speeded up robust features. Presented at the 2009 IEEE International Conference on Technologies for Practical Robot Applications (TePRA).

[14] Gu Q, Takaki T, Ishii I (2013) Fast FPGA-based multiobject feature extraction. IEEE Trans Circuits Syst Video Technol 23(1):30–45.

[15] Kryjak T, Gorgon M (2013) Real-time implementation of the ViBe foreground object segmentation algorithm. In: Federated Conference on Computer Science and Information Systems (FedCSIS), pp 591–596.

[16] Saqib F, Dutta A, Plusquellic J, Ortiz P, Pattichis MS (2015) Pipelined decision tree classification accelerator implementation in FPGA (DT-CAIF). IEEE Trans Comput 64(1):280–285.

[17] Pan J, Lauterbach C, Manocha D (2010) g-Planner: real-time motion planning and global navigation using GPUs. In: Proceedings of AAAI Conference on Artificial Intelligence 1245–1251.

[18] Vasumathi B, Moorthi S (2012) Implementation of hybrid ANN-PSO algorithm on FPGA for harmonic estimation. Eng Appl ArtifIntell 25(3):476–483.

[19] Wang Y, Xu J, Han Y, Li H, Li X (2016) DeepBurning: automatic generation of FPGA-based learning accelerators for the neural network family, pp 1–6.

[20] Abdelouahab K, Pelcat M, Serot J, Bourrasset C, Berry F (2017) Tactics to directly map CNN graphs on embedded FPGAs. IEEE Embed Syst Lett 9(4):113–116.

[21] Sharma H et al (2016) From High-level deep neural models to FPGAs. In: 49th Annual IEEE/ACM International Symposium on Microarchitecture, pp 1–12.

[22] Zeng H, Zhang C, Prasanna V (2018) Fast generation of high throughput customized deep learning accelerators on FPGAs. In: International Conference on Reconfigurable Computing FPGAs, ReConFig 2017, vol 2018-Janua, pp 1–8.

[23] Art P (2011) Artificial neural network acceleration on FPGA using custom instruction, pp 450–455.

[24] Cadambi S, Graf HP (2010) A programmable parallel accelerator for learning and classification, pp 273–283.

[25] Motamedi M, Gysel P, Akella V, Ghiasi S (2016) Design space exploration of FPGA-based deep convolutional neural networks. In: Proceeding of

