International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 9 Issue 3, May-Jun 2025 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Emerging VLSI Technologies for High Performance AI and ML Applications: Survey Paper

D Sathya Preetham, Ananya R, Anshu Naikodi, Archana C K

Department of Electronics and Communication Engineering,
Dayananda Sagar Academy of Technology and Management, Bangalore, Karnataka, India

ABSTRACT

Modern market technology evolution requires CMOS-based semiconductor manufacturing to use more effective and smarter Electronic Design Automation (EDA) methods because of rising efficiency requirements. This paper examines how cutting-edge technologies consisting of machine learning (ML), artificial intelligence (AI), edge computing and neuromorphic systems function in Very Large Scale Integration (VLSI) and embedded system designs. The emphasis on sustainability happens through "design-based equivalent scaling" as well as AI implementations in chip production and power improvement with AVS alongside in-situ detection and task-memory scheduling. The paper details how FPGAs and MPSoCs bring performance benefits to hardware systems while examining new memory solutions consisting of resistive RAM and in-memory computing technology that help bypass traditional von Neumann system constraints. The joint optimization between hardware and software technology leads to meaningful applications which detect ASD while improving biomedical imaging. The paper demonstrates through diverse academic studies that ML models with energy-efficient circuit designs and edge AI represent future semiconductor and embedded systems standards.

How to cite this paper: D Sathya Preetham | Ananya R | Anshu Naikodi | Archana C K "Emerging VLSI Technologies for High Performance AI and ML Applications: Survey Paper" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-9 | Issue-3, June 2025, pp.669-674, URL: www.ijtsrd.com/papers/ijtsrd80043.pdf

IJTSRD80043

INTRODUCTION

Market technological changes lead to significant difficulties in product scheduling which puts heavy operational strains on production processes in CMOS-based semiconductor and electronics industries. Sustainable development of the Electronics Design Automation industry requires the use of "design-based equivalent scaling" methods. Machine learning has demonstrated potential to become a useful solution that operates within active design tools as they support their application domains. The analysis examines ML methods used in chip manufacturing with specific focus on benefits and manufacturing techniques for IC production. The power requirements and data usage of modern deep learning systems exceed their potential because cloud-based ML analytics outperform traditional computing in terms of scalability yet they generate extended delays due to server connection requirements. The market demands advanced edge devices for quick and efficient data processing as new technology grows because auto systems along with drones, robots, etc. need these devices to work effectively. This research evaluates enhanced AI methods while studying fast methods for narrow equations as well as binary and tensor processing for new information alongside hardware prototype assessments based on field-programmable gate arrays and CMOS-ASICs. The paper provides a brief overview of potential future outlooks concerning resistive random access memory devices.[1]. The neuromorphic systems named Dynamic Adaptive Neural Network Arrays (DANNAs) imitate spiking neural behavior while using evolutionary optimization techniques for their design. Array elements support complete reconfiguration for neuron and synapse operations as well as fan-out functions and their respective parameters can be customized for inter-element connectivity. DANNAs operate as neuromorphic systems that run on Field Programmable Gate Arrays (FPGAs) but display restrictions with regards to scalability and performance levels. The development of a semi-custom Very Large Scale Integration

(VLSI) design solves existing implementation problems to enable more advanced functionality. The VLSI hardware surpasses FPGA frameworks through its ability to yield three key advantages: increased element capacity by 50 times and double the speed along with matched power efficiency. The system enables continuous real-time checks of each array component[2].

Li-passivation in zigzag GaN nanoribbons significantly modifies their electronic properties, enhancing Fermi velocity and reducing effective mass to improve carrier mobility. DFT investigations further show strong gas adsorption and charge transfer, highlighting their potential as high-performance nanosensors[3-4].

The Multiprocessor System-on-Chip (MPSoC) integrates various instruction-set processors into a single integrated circuit which executes most functions in complex electronic systems. Each MPSoC architecture follows the design pattern of specific embedded functions. The central aspect of customization depends on memory system setup for the on-chip component. Embedded systems now employ software-managed scratchpad memory (SPM) instead of traditional caches because SPM demonstrates better efficiency and energy efficiency together with more predictable timings. To reach peak system performance it is vital to distribute the on-chip SPM budget among processors according to application requirements. Design processes traditionally separate task scheduling from SPM partitioning while these two elements form a closely related dynamic system. An Integer Linear Programming (ILP) formulation provides this research with an integrated method for mapping tasks and scheduling as well as partitioning SPM and managing data distribution. The proposed method reveals the best performance configuration which demonstrates performance enhancement up to 80% through integrating task scheduling with SPM optimization in embedded systems.[5]

DFT-based studies demonstrate that Indium Nitride nanoribbons can effectively detect gases like CO, CO_2, NO, and NO_2 due to notable charge transfer and band structure modulation. Similarly, Scandium Nitride monolayers show strong adsorption sensitivity toward toxic gases such as NH_3, AsH_3, BF_3, and Bcl_3. Zigzag silicon carbide nanoribbons exhibit enhanced gas sensing performance through improved electronic response to hazardous gas molecules, making them promising for advanced sensor applications[6-8].

The research examines how Adaptive Voltage Scaling technology can be implemented onto standard FPGAs which originally lacked built-in voltage adjustment capabilities. The designed power management framework operates through an adjusted design process which integrates in-situ detectors alongside real-time clock management reprogramming capabilities. Adaptive Voltage Scaling (AVS) stands as a power-efficient methodology which lets devices change their operational voltage and frequency dynamically so they can adjust to their workload together with process fluctuations and environmental conditions which operate under closed-loop management. The power management strategy AVS exceeds Dynamic Voltage and Frequency Scaling (DVFS) since AVS allows better precision and energy-saving ability. The application of AVS with in-situ detectors on FPGAs enables power and energy savings exceeding 85% relative to running systems at nominal voltage with fixed frequency operation. The in-situ detector system provides a more dependable and straightforward approach than delay line channels for critical path replication since it removes the necessity of manual delay calibration.[9]

Processing and memory units in traditional von Neumann computing systems create substantial energy costs along with time delays because of data movement which has become even more severe as data-intensive applications expand including artificial intelligence workloads. A transformation in hardware design requires an alternative solution that uses in-memory computing to solve this issue. The physical characteristics of memory devices allow direct calculation of specific operations that happen directly in memory space. Researchers investigate both conceptual memory systems based on charge behavior and those based on resistance properties to use in this field. This review presents an extensive investigation of the basic computation capabilities provided by these memory devices while showing their uses across scientific computing, signal processing, optimization work and machine learning, deep learning and stochastic computing.[10]

Previous studies about power reduction focused their research on single components instead of analyzing complete system structures. The following paper establishes a framework which allows modeling power behavior at the system design level. Three essential components make up the model: a collection of system resources together with an environmental workload specification as well as a power management policy that serves as its core element. The model translates into simulation-based software for power consumption evaluation of the system. We have developed a method that aims to enhance power management policy efficiency. The optimization algorithm operates through multiple iterations with

the power estimation engine to build customized power management practices for systems from specific descriptions. The proposed method was tested on a low-power portable device through which we attained power estimation accuracy reaching 10% while power management policies led to a 23% power reduction.[11]

VLSI industry continues to adopt Artificial Intelligence (AI) techniques into its design automation system because this opens opportunities to transform chip design approaches. System-on-Chip (SoC) architectures validate the essential status of Artificial Intelligence in VLSI development since the integration requires power-saving measures for existing hardware along with machine learning algorithms for efficiency improvements. The analysis covers a complete assessment of how AI technology impacts VLSI through three fundamental subdomains namely analog design and digital design as well as physical design. This paper investigates modern deep learning and machine learning approaches which have been adopted in VLSI research.[12]

Field Programmable Gate Arrays (FPGAs) function as programmable logic devices which allow user post-manufacturing configuration to execute diverse operations from basic logic functions to complex systems-on-chip functions and AI applications. The appearance of FPGAs in scientific literature in 1992 has resulted in more than 70,000 related documents now available in major indexes such as Scopus and Clarivate Web of Science. The technology enables magnetic suspension systems that redefine the kilogram measurement and supports navigation systems used in Mars rovers. Our paper utilizes ScientoPy to perform scientometric research on FPGA-related literature which appeared between 1992 and 2018. Our study focuses on 150 leading application categories grouped into digital control, communication interfaces, networking, computer security, cryptography, machine learning, digital signal processing, image and video processing, big data and computer algorithms and several other subfields. The paper includes a study of application trends with historical data analysis of these applications since 1992.[13]

Density Functional Theory (DFT) investigations reveal that Cu and Fe doping in boron nitride nanoribbons (BNNRs) significantly enhances their electrical conductivity, making them suitable candidates for nanoscale interconnects in advanced integrated circuits. Ab-initio studies on aluminum nitride nanoribbons (AlNNRs) demonstrate their potential in implementing reconfigurable logic gates due to tunable electronic properties under external stimuli. Additionally, the design of a FinFET-based operational amplifier (Op-Amp) using 22 nm high-k dielectric technology shows promising results in reducing leakage currents and enhancing performance, offering a robust solution for low-power, high-efficiency analog circuit applications[14-16].

Most embedded and portable image and video processing applications dedicate their energy to off-chip memory access instructions through load/store commands. The reduction of memory instructions and enhanced energy efficiency through performance-driven locality optimization techniques needs additional explicit energy-strategic approaches. A multi-bank memory system with multiple homogeneous processors is the focus of this study for handling large signal arrays. A compiler-based solution proposes power-saving methods which use memory bank operating modes. One primary obstacle within such systems emerges from implementing parallel processing alongside reduction of energy usage requirements. The simultaneous access for parallelism needs must compete against low-power bank entry requirements for energy reduction purposes. Our approach achieves memory trade-offs by implementing three comprehensive stages which locate parallel operations then distribute arrays across banks followed by data reordering procedures. The proposed solution demonstrates its ability to lower off-chip memory energy usage by a great extent while retaining full parallel processing capabilities.[17]

As the second most common internal cancer worldwide liver cancer stands as a major reason for mortality due to cancer-related illnesses. The diagnostic approach depends heavily on timely discovery and exact staging assessment in radiology-based practice. The research develops a dynamic segmentation approach that merges a customized Watershed method with Neutrosophic logic for performing liver segmentation on abdominal CT images. The method follows three essential operational stages starting with preprocessing then Neutrosophic transformation of CT images and concluding with post-processing. The preprocessing stage applies histogram equalization combined with median filtering to produce better image intensity together with noise reduction. The conversion process transforms the image into the Neutrosophic domain through creation of three membership sets. Mathematical morphology techniques join forces with the modified Watershed algorithm for refining both the truth image and segmenting the liver region properly during post-processing. The proposed method delivers performance results that measure approximately 95% accuracy based on different

evaluation metrics. This method provides better results than multiple existing segmentation procedures according to comparative analysis.[18]

Engineers across different fields continue to show growing interest in machine learning (ML) algorithms because such algorithms can create complex system models through historical data analysis. A new method for CMOS VLSI circuit power consumption estimation through passive ML models uses various circuit parameters. Supervised learning technologies provide the method with both efficient and precise power calculation abilities while maintaining system functionality. The document analyzes the random forest algorithm for power prediction in CMOS VLSI circuits through its relatively new application in this domain. The random forest model receives optimization through multi-objective NSGA-II algorithm implementation. The implemented model delivers testing errors between 1.4% and 6.8% along with a Mean Square Error value of 1.46e-06 which outperforms standard BPNN usage. Statistical evaluation of the approach through R coefficient and RMSE indicates its strong performance. The random forest model reveals excellent performance through its 0.99938 R-value and 0.000116 RMSE which establish it as an accurate choice to estimate CMOS VLSI power.[19]

Hardware systems initially under the label of neuromorphic computing were designed to remodel the neuro-biological structures of neural processes. The definition evolved so neuromorphic computing systems now include machines which execute models derived from biological inspiration such as neural networks with deep learning structures. The global scientific interest in neuromorphic computing grew rapidly because von Neumann technologies face limitations when handling cognitive applications. The document offers a historical review of neuromorphic computing advancements which includes discussions about models and hardware systems. The paper describes numerous implementation methods alongside practical implementation approaches. This paper examines future-shaping technologies that include new physical devices together with interdisciplinary computing architectures and novel devices which represent the potential course of neuromorphic computing evolution.[20]

The development of technology will result in leakage power becoming the major power consumption source in current processors. The majority of transistors used in on-chip designs are situated in caches making them the main targets to control leakage power. The fundamental problem remains unanswered regarding the full extent of which architectural together with circuit-level solutions can lower leakage power consumption. The paper examines the highest levels of leakage reduction that current technologies can accomplish. The strategic application of sleep and drowsy modes driven by perfect address trace information leads to a maximum reduction of instruction cache leakage to 3.6% alongside data cache leakage reaching 0.9% of their unmodified leakage levels. The model we developed provides a detailed parameterized analysis for determining optimal leakage savings in future technology nodes. We developed a prefetching-based approach to facilitate practical systems reaching their theoretical minimum leakages.[21]

IoT has expanded beyond measure since its rapid growth because Internet of Everything (IoE) technology now connects billions of smart devices to the internet network. Traditional cloud computing models have reached their capacity limit because of which they now experience slow speeds and increased bandwidth consumption together with privacy and security issues. Modern intelligent systems require more than cloud computing because it cannot efficiently handle their diverse and complex data processing needs. A contemporary solution to these problems exists in edge computing since it moves data processing near the data collection and user locations. Edge computing system allows distributed data storage and processing within network boundaries that decrease response delays and distributes workload from central cloud facilities. The study examines all recent research alongside advancements that exist in edge computing. The research starts with explaining the main idea of edge computing while comparing it to cloud computing before analyzing edge computing architecture and key technologies for implementation and security and privacy topics and real-world application examples.[22]

Early brain imaging diagnosis of Autism Spectrum Disorder (ASD) serves as a vital tool for reducing its social communication impact. The paper develops a deep learning method which employs Magnetic Resonance Imaging (MRI) brain scans to identify Autism Spectrum Disorder. A Deep Convolutional Neural Network (CNN) functions together with a Dwarf Mongoose optimized Residual Network (DM-ResNet) for the analysis process. An initial step includes preprocessing MRI input images so they contain only brain tissues. The brain image segmentation process uses Fuzzy C-Means (FCM) in combination with Gaussian Mixture Model (GMM) to divide the image into subgroups for improving accuracy during classification. A subsequent process

divides the segmented volumes between specific cortical along with subcortical defined regions. VGG-16 network serves as the feature extraction method through its utilization of small convolutional filters that enhance the detection of complex discriminative features in the data. The extracted Functional connectivity features use Regions of Interest (ROI) to provide input to DM-ResNet for classification. With the help of Dwarf Mongoose optimization the algorithm achieves remarkable results by optimizing network hyperparameters thus enhancing classification accuracy rates. The proposed method reaches an autism detection accuracy level of 99.83%.[23]

Conclusion

The review explores in detail the industrial transformation happening to the semiconductor and embedded systems industry that results from artificial intelligence combined with intricate hardware frameworks. ML and neuromorphic computing technologies engage in a strong partnership that proves mutually beneficial for DANNAs and CMOS systems across various areas of application. Modern electronic systems need power efficiency as a fundamental requirement which receives multidimensional optimization solutions that cover scheduling operations and memory adaptability and AVS implementations for minimal power leakage mechanisms. The combination of in-memory computing and edge computing becomes an effective approach to resolve data movement problems with the added benefit of decentralized cloud system operation and reduced latency. FPGAs continue expanding into different applications which demonstrates the vital position of reconfigurable logic in current research and operational systems. AI models serve as an illustration of cutting-edge technology because they enable ASD diagnosis through MRI testing within healthcare environments. Future intelligent systems require important interdisciplinary innovations between these technologies to address their performance as well as power and scalability limitations.

References:

[1] Miller, A. (1989, April). From expert assistant to design verification: applications of AI to VLSI design. In Proceedings. IEEE Energy and Information Technologies in the Southeast'(pp. 406-410). IEEE.

[2] Dean, M. E., & Daffron, C. (2016, July). A VLSI design for neuromorphic computing. In 2016 IEEE Computer Society Annual Symposium on VLSI (ISVLSI) (pp. 87-92). IEEE.

[3] M. Jatkar, K. K. Jha and S. K. Patra, "Fermi Velocity and Effective Mass Variations in ZGaN Ribbons: Influence of Li-Passivation," in IEEE Access, vol. 9, pp. 154857-154863, 2021, doi:10.1109/ACCESS.2021.3128294.

[4] M. Jatkar, K. K. Jha and S. K. Patra, "DFT Investigation on Targeted Gas Molecules Based on Zigzag GaN Nanoribbons for Nano Sensors," in IEEE Journal of the Electron Devices Society, vol. 10, pp. 139-145, 2022, doi:10.1109/JEDS.2022.3144014.

[5] Suhendra, V., Raghavan, C., & Mitra, T. (2006, October). Integrated scratchpad memory optimization and task scheduling for MPSoC architectures. In Proceedings of the 2006 international conference on Compilers, architecture and synthesis for embedded systems (pp. 401-410).

[6] K. K. Jha, M. Jatkar, P. Athreya, T. M. P. and S. K. Jain, "Detection of Gas Molecules (CO, CO2, NO, and NO2) Using Indium Nitride Nanoribbons for Sensing Device Applications," in IEEE Sensors Journal, vol. 23, no. 19, pp. 22660-22667, 1 Oct.1, 2023, doi:10.1109/JSEN.2023.3307761.

[7] Pratham Gowtham, Mandar Jatkar, DFT based study to sense harmful gases (NH3, AsH3, BF3, BCl3) using Scandium Nitride monolayer for sensing device applications, Micro and Nanostructures, Volume 201, 2025, 208100, ISSN 2773-0123, https://doi.org/10.1016/j.micrna.2025.208100.

[8] Jatkar, M. Improving the sensor capability of zigzag silicon carbide nanoribbon for the detection of harmful gases. Discover Electronics 2, 7 (2025). https://doi.org/10.1007/s44291-025-00047-0.

[9] Nunez-Yanez, J. L. (2014). Adaptive voltage scaling with in-situ detectors in commercial FPGAs. IEEE Transactions on Computers, 64(1), 45-53.

[10] Sebastian, A., Le Gallo, M., Khaddam-Aljameh, R., & Eleftheriou, E. (2020). Memory devices and applications for in-memory computing. Nature nanotechnology, 15(7), 529-544.

[11] Benini, L., Hodgson, R., & Siegel, P. (1998, August). System-level power estimation and optimization. In Proceedings of the 1998 international symposium on Low power electronics and design (pp. 173-178).

[12] Malhotra, A., & Singh, A. (2022, March). Implementation of AI in the field of VLSI: A Review. In 2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T) (pp. 1-5). IEEE.

[13] Ruiz-Rosero, J., Ramirez-Gonzalez, G., & Khanna, R. (2019). Field programmable gate array applications—A scientometric review. Computation, 7(4), 63.

[14] Mandar Jatkar, P.Y. Mallikarjun, ”Optimized Cu/Fe doped Boron Nitride Nanoribbons as nanoscale interconnect: DFT Investigation, Materials Science in Semiconductor Processing, Volume 186, 2025, 109050, ISSN 1369-8001,https://doi.org/10.1016/j.mssp.2024.109050.

[15] Sudhir Rai, Kamal K. Jha, Mandar Jatkar, Ab-initio investigation on aluminum nitride nanoribbons for reconfigurable logic gates, Diamond and Related Materials, Volume 152, 2025, 111966, ISSN 0925-9635, https://doi.org/10.1016/j.diamond.2025.111966.

[16] R. Rambola and M. Jatkar, "An Effective Synchronization of ERP in Textile Industries," *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2018, pp. 969-973, doi:10.1109/ICECA.2018.8474686.

[17] De La Luz, V., Kandemir, M., & Sezer, U. (2001, August). Improving off-chip memory energy behavior in a multi-processor, multi-bank environment. In International Workshop on Languages and Compilers for Parallel Computing (pp. 100-114). Berlin, Heidelberg: Springer Berlin Heidelberg.

[18] Sayed, G. I., Ali, M. A., Gaber, T., Hassanien, A. E., & Snasel, V. (2015, December). A hybrid segmentation approach based on Neutrosophic sets and modified watershed: A case of abdominal CT Liver parenchyma. In 2015 11th international computer engineering conference (ICENCO) (pp. 144-149). IEEE.

[19] Govindaraj, V., & Arunadevi, B. (2021). Machine learning based power estimation for CMOS VLSI circuits. Applied Artificial Intelligence, 35(13), 1043-1055.

[20] Chen, Y., Li, H. H., Wu, C., Song, C., Li, S., Min, C., ... & Liu, X. (2018). Neuromorphic computing's yesterday, today, and tomorrow–an evolutional view. Integration, 61, 49-61.

[21] Meng, Y., Sherwood, T., & Kastner, R. (2005, February). On the limits of leakage power reduction in caches. In 11th International Symposium on High-Performance Computer Architecture (pp. 154-165). IEEE.

[22] Cao, K., Liu, Y., Meng, G., & Sun, Q. (2020). An overview on edge computing research. IEEE access, 8, 85714-85728.

[23] Jain, S., Tripathy, H. K., Mallik, S., Qin, H., Shaalan, Y., & Shaalan, K. (2023). Autism detection of mri brain images using hybrid deep cnn with dm-resnet classifier. IEEE Access, 11, 117741-117751.