Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/39/20230597

Application of machine learning to SRAM circuit design

Haohan Wang

School of Medical Information Engineering, Shandong First Medical University, Tai'an City, 271000, China

1910841107@mail.sit.edu.cn

Abstract. This paper discusses the application of machine learning in Static Random Access Memory (SRAM) circuit design. It mentions several aspects of application analysis, including SRAM design for machine learning applications, SRAM PUF design for machine learning modelling attacks, lightweight fault-tolerant mechanism for NVDLA-based edge AI chips, interconnection resource allocation of SRAM FPGAs using reinforcement learning and Markov decision process, and SRAM power optimization and privacy security using machine learning. This article conclude that machine learning can be used to model and optimize SRAM performance, improve system reliability and stability through fault detection and fault tolerance mechanisms, provide a more comprehensive and efficient approach to testing and verification, and automatically correct and repair errors. The significance of machine learning in SRAM design is to provide an intelligent method and tool that can speed up the design process, improve system performance, increase reliability, and drive the advancement of chip manufacturing and processes.

Keywords: SRAM, Machine Learning, Reinforcement Learning.

1. Introduction

In VLSI design, the key problems in SRAM design include the stability, read, and write delay, power consumption and reliability of the memory cell. SRAM design is an important research direction in VLSI field, involving circuit design, process design, layout design, timing analysis and optimization [1]. Combined with authoritative academic articles in VLSI, SRAM and machine learning related fields, this article mainly discusses the application of machine learning in SRAM circuit design.

The article mentions several aspects of application analysis, including: Firstly, SRAM design for machine learning applications, aiming to provide high energy efficiency and high bandwidth memory solutions. Secondly, SRAM PUF design for machine learning modeling attacks, with good attack resistance. Thirdly, a lightweight fault-tolerant mechanism is proposed. Fourthly, apply reinforcement learning and Markov decision process (MDP) to realize the interconnection resource allocation of SRAM FPGAs [2]. Finally, the application of machine learning in SRAM power optimization and privacy security is discussed. The significance of machine learning in VLSI applications is to provide an intelligent method and tool that can speed up the design process, improve system performance, increase reliability, and drive the advancement of chip manufacturing and processes.

DOI: 10.54254/2755-2721/39/20230597

2. The theoretical foundation of machine learning and SRAM

2.1. SRAM

SRAM is a common computer memory technology with theoretical foundations in semiconductor physics, circuit design and digital logic. SRAM uses trigger circuits to store and read data, and is often used for applications that require fast access, such as caches and register files. The basic memory unit structure of the SRAM is a flip-flop circuit, usually composed of six transistors [3]. This circuit can store a bit (0 or 1) and read and write data through the control input and read and write lines. The memory unit of the SRAM performs write and read operations by controlling the input signal. The write operation writes data to the storage unit, while the read operation obtains data from the storage unit. These operations involve the control and drive of components such as bit line, word line and transfer gate. The graph of serial SARM is shown in figure 1.

Figure 1. Graph of serial SARM (Photo/Picture credit: Original).

It usually organizes storage units in the form of arrays, where each storage unit has a unique address [4]. The address line is used to select a specific storage unit to be read or written to, so that access to the memory can be completed in a constant time. The read operation of the SRAM involves sensing and amplifying the storage unit to ensure that the stored data is read correctly [5]. The write operation requires making changes to the charge of the storage cell and ensuring that the data is written correctly. It requires a proper power supply and has a stable voltage and signal amplitude to ensure correct storage and read operation.

2.2. Machine Learning

The theoretical foundations of machine learning include statistics, probability theory, linear algebra, and optimization theory. It involves algorithms and methods for learning patterns and patterns from data. In machine learning, SRAM can be used as a hardware accelerator to store and manipulate model parameters or intermediate results, thereby increasing computational efficiency and throughput [6]. Combining machine learning algorithms with SRAM enables a faster model inference and training process. Machine learning processes large data sets to make predictions or decisions based on patterns and patterns in the extracted data. Machine learning can be applied to many aspects of SRAM design, such as layout and routing, fault detection, fault tolerant design, and power optimization.

3. Application analysis of machine learning in SRAM circuit design

3.1. Line routing and timing analysis

In order to address the challenge of applying machine learning to SRAM circuit wiring, Sung Kyu Lim propose that the newest ICs technology could overcome the limitations of integrating 2D/3D circuits. The 3D IC technology could be used to combine various devices like CMOS, RRAM and SRAM within

DOI: 10.54254/2755-2721/39/20230597

a single chip [7]. Including the generation of technology files, the use of pseudo3D flows for standard cell designs, and the use of macro-on-logic (MoL) flows for designs with hard macros. Heterogeneous 3D ICs will help improve timing and power performance, signal and power integrity, and explore potential integration techniques in the future.

Mingn Kang and Naresh R. Shanbhag proposed to implement (RF) machine learning classifiers using in-memory computing methods. The main innovation of their proposed method lies in the use of DSS and balanced decision trees [8]. Through their prototype IC implementation, they achieve a remarkable energy savings of 3.1 times compared to conventional digital architecture along with a speed-up factor of 2.2 times. This results in a significantly lower energy-delay product (EDP) by approximately 6.8 times at an accuracy level exceeding 93%. These techniques effectively contribute towards achieving energy savings, reduced decision delay, and improved energy efficiency when compared to traditional digital architectures.

3.2. SRAM testing and fault diagnosis

Memory-based Physical Unclonable Functions (PUFs), such as SRAM PUFs, are highly attractive due to their extensive usage in embedded systems and minimal additional hardware requirements. SRAM PUFs utilize the initial values of SRAM memory cells as a unique digital fingerprint. Our objective is to enable authentication of devices even in the presence of data loss caused by transmission errors, without relying on complex error-correcting techniques. Nico Mexis think Convolutional Neural Networks (CNN) can be employed for classifying and identifying SRAM PUF responses [9]. To address challenges related to noisy datasets and data corruption, researchers manually manipulated the data by converting a portion of the response data into zero values, simulating data corruption. These data were then transformed into binary arrays and grayscale images before being classified and identified using a CNN model. The proposed approach achieves an accuracy rate of 96.85% for intact images and 95.46% for noisy images. Furthermore, it demonstrates accurate classification capabilities even with up to a 30% loss in data integrity. However, when the corruption rate exceeds 30%, the performance of the model starts deteriorating significantly. By utilizing a combined dataset during training, we can enhance the classification performance for both corrupted and intact images; however, accurately classifying images with 40% corruption remains challenging using our current model configuration. Overall, this method presents a novel approach towards authenticating SRAM PUFs while partially addressing concerns associated with data corruption.

There are certain challenges, such as the NP-complete problem, that make it quite challenging to find the best solution. The utilization of reinforcement learning in testing and diagnosing. In an article, Aokai Shi discusses how RL-based MDP is employed to implement interconnect resource configurations for SRAM FPGAs [9]. By utilizing dynamic programming, the authors successfully solve the MDP and achieve comprehensive coverage of interconnect resources. This innovative approach optimizes configuration numbers and addresses the challenge of achieving full coverage in NP-complete problems like FPGA testing. Experimental results demonstrate that this method effectively optimizes configuration numbers towards a theoretical value while ensuring complete coverage, making it applicable to NP-complete issues encountered during FPGA testing.

3.3. Fault recovery and fault tolerance

AnQi Zhang proposes the utilization of CNT SRAM memory [10]. Modifications are made to the NeuroSim tool to analyze how different fault patterns affect the memory units after recovering from stuck-at faults. The paper also presents mechanisms for fault recovery that aim to improve system accuracy. By leveraging emerging technology like CNT SRAM, it becomes possible to effectively address high fault rates and enhance overall system accuracy. The study examines various fault patterns and locations while offering techniques for optimizing system accuracy. Despite the availability of open-source NVDLA and ONNC, challenges still exist during the mass production stage of AI chip development, such as issues with SRAM MBIST fail and scan-chain fail leading to reduced productivity.

DOI: 10.54254/2755-2721/39/20230597

To tackle internal SRAM stuck faults, Luba Tang proposes a lightweight fault-tolerance mechanism [11]. By employing ONNC, the authors suggest employing lightweight recovery mechanisms like adjacent bit mechanism and barrel shift to address input faults. Additionally, they explore SRAM Analyzer for reconfiguring and integrating these fault recovery mechanisms into the SRAM compiler to ensure usage in CNN [12]. Experimental results demonstrate that this method can optimize configuration numbers towards achieving theoretical coverage with full effectiveness, making it suitable for NP-complete problems such as FPGA testing.

3.4. SRAM power optimization and privacy security

The major portion of power dissipation in deep learning networks is attributed to the read operations involved in fetching network weights. Saranyu Chattopadhyay and Kaustav Brahma think STT-MRAM can optimize power consumption. Voltage-mode sensing can achieve a high level of Tunnel Magnetoresistance (TMR) by utilizing a read operation with lower voltage and thicker oxide. The combination of reduced voltage during read operations and increased thickness of the oxide layer contributes to the achievement of higher TMR in voltage-mode sensing. Firstly, it allows for more aggressive scaling down and causes the energy consumption decreased. What's more, there is no static current path which could lead further reduction in power usage [13].

To address the vulnerability of PUF-based protocols to modeling attacks, Jimson Mathew proposes a SRAM PUF that exhibits strong resistance against such attacks. Their main research approach involves designing and implementing the 2SPUF architecture, which is a Two Round SRAM PUF [14]. A key innovation presented in this paper is the introduction of a two-round SRAM PUF architecture called 2SPUF that effectively mitigates machine learning modeling attacks. This unique feature enhances the resilience of the 2SPUF architecture against modeling attacks while experimental results validate its effectiveness in resisting such attacks by demonstrating improved accuracy.

4. Conclusion

This paper combined with authoritative academic articles in VLSI, SRAM and machine learning related fields to analyze the specific application examples of machine learning in SRAM circuit design. For Optimized design, machine learning can be applied to circuit and chip design processes to automatically discover and optimize circuit structure, layout, and characteristics by analyzing large amounts of data and patterns. This can improve circuit performance, power consumption and reliability. For Fault detection and fault tolerance, machine learning can be used to develop fault detection algorithms and fault tolerance mechanisms to improve system reliability and stability. For Testing and verification, machine learning can identify potential errors, vulnerabilities, and anomalies to provide a more comprehensive and efficient approach to testing and verification by learning from large test samples and historical data, machine learning algorithms. For Error correction and repair, machine learning can automatically correct and repair errors by learning known error patterns and solutions. This saves time and labor costs and improves the maintainability of the system.

Machine learning has a broad application prospect in SRAM design. Firstly, by using machine learning, SRAM performance can be modeled and optimized. Machine learning algorithms can analyze large amounts of SRAM performance data and discover the effects of different design parameters on performance to guide the design process. Secondly, Machine learning can identify patterns and related factors that cause failures by analyzing large-scale reliability data. This can help designers take appropriate measures to improve SRAM reliability. Thirdly, Machine learning can help designers find the best power optimization strategy by learning and predicting the relationship between different design parameters and power consumption. This approach can reduce the contribution of SRAM to the overall chip power consumption, thereby extending the chip's battery life or reducing heat dissipation requirements. It is important to note that the application of machine learning in SRAM design is still in the stage of continuous development and exploration. As the technology advances and more research is carried out, it can expect to see more innovative applications of machine learning in SRAM design.

DOI: 10.54254/2755-2721/39/20230597

References

[1] Wu Zhao, Ling Chen, Xin Guo, et al. Knowledge push technology based on quality function knowledge deployment. Proceedings of the Institution of Mechanical Engineers, Part C. Journal of mechanical engineering science, 2019, 233(4): 1119-1138

[2] K.G. Devi, "Artificial Intelligence Trends for Data Analytics using Machine Learning and Deep Learning Approaches", CRC Press, 2020

[3] V. Govindaraj and B. Arunadevi, "Machine Learning Based Power Estimation for CMOS VLSI Circuits", Applied Artificial Intelligence, Vol. 35, No. 13, pp. 1043-1055, 2021.

[4] W. Lu, D. S. Jeong, M. Kozicki. A CAM full custom design with new unit and matching line. IEEE Electron Device Letters, 2008, 21(5): 735-746R.

[5] Mina, C. Jabbour and G. Sakr, "A Review of Machine Learning Techniques in Analog Integrated Circuit Design Automation", Electronics, vol. 11, no. 3, p. 435, 2022.

[6] Zhengping Li, Mingming Xie, Yongliang Zhou, Robust activating timing for SRAM SA with replica cell voltage boosted circuit. IEICE Electronics Express, Vol. 13, No.10: 1-7, 2021.

[7] Gauthaman Murali; Sung Kyu Lim, "Heterogeneous 3D ICs: Current Status and Future Directions for Physical Design Technologies", IEEE DATE, 2021.

[8] Mingu Kang, Sujan K. Gonugondla, Naresh R. Shanbhag. "A 19.4 nJ/decision 364K decisions/s in-memory random forest classifier in 6T SRAM array", ESSCIRC 2017 - 43rd IEEE European Solid State Circuits Conference, 2017

[9] Aiwu Ruan; Aokai Shi; Liang Qin; Shiyang Xu; Yifan Zhao, "A Reinforcement Learning-Based Markov-Decision Process (MDP) Implementation for SRAM FPGAs", IEEE Transactions on Circuits and Systems II: Express Briefs,2020

[10] An Qi Zhang; Amr M.S. Tosson; Lan Wei, "Error Resilience and Recovery of Process Induced Stuck-at Faults in MLP Neural Networks using Emerging Technology", IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH), 2021

[11] V. Govindaraj, B. Arunadevi. "Machine Learning Based Power Estimation for CMOS VLSI Circuits", Applied Artificial Intelligence, 2021

[12] B. Van Essen, C. Macaraeg, M. Gokhale, and R. Prenger, "Accelerating a random forest classifier: Multi-core, GP-GPU, or FPGA?", IEEE FCCM, 2012.

[13] Saranyu Chattopadhyay; Kaustav Brahma; Arkaprova Ray; Mrigank Sharad, "STT-MRAM for Low Power Access for Read-Intensive Parallel Deep-Learning Architectures", IEEE International Symposium on Nanoelectronics and Information Systems, 2017

[14] Vikash Kumar Rai; Somanath Tripathy; Jimson Mathew, "2SPUF: Machine Learning Attack Resistant SRAM PUF", IEEE ISEA-ISAP, 2020.