Design and Reliability Analysis of System-on-Chips for Artificial Intelligence Applications at the Edge

Submitted to
Graduate School of Informatics
Kyoto University

Quan Cheng

Publications

Journal Articles

[J1] Q. Cheng, M. Huang, C. Man, A. Shen, L. Dai, H. Yu, and M. Hashimoto, "Reliability Exploration of System-on-Chip With Multi-Bit-Width Accelerator for Multi-Precision Deep Neural Networks," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 70, no. 10, pp. 3978-3991, Oct. 2023, doi: 10.1109/TCSI.2023.3300899.

[J2] Q. Cheng, L. Dai, M. Huang, A. Shen, W. Mao, M. Hashimoto, and H. Yu, "A Low-Power Sparse Convolutional Neural Network Accelerator With Pre-Encoding Radix-4 Booth Multiplier," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 70, no. 6, pp. 2246-2250, June 2023, doi: 10.1109/TCSII.2022.3231361.

International Conference Papers

[I1] Q. Cheng, L. Lin, M. Huang, Q. Li, Z. Yang, L. Dai, H. Yu, Y. Chen, Y. Shi, M. Hashimoto, "A 13-34 TOPS/W Edge-AI Processor Featuring Booth-Value-Confined Accelerator, Near-Memory Computing, and Contiguity-Aware Mapping," 2024 IEEE Asian Solid-State Circuits Conference (A-SSCC), Hiroshima, Japan, 2024.

[I2] Q. Cheng, Q. Li, L. Lin, W. Liao, L. Dai, H. Yu, and M. Hashimoto, "How Accurately Can Soft Error Impact Be Estimated in Black-box/White-box Cases? – A Case Study with an Edge AI SoC –," 2024 61st ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 2024.

Summary

In recent years, the field of Artificial Intelligence (AI) has witnessed remarkable advancements, with a growing emphasis on large-volume computing. However, relying solely on cloud-based solutions for processing enormous data in Neural Networks (NNs) presents challenges such as high latency and the need for substantial bandwidth. These challenges underscore the importance of edge AI computing, where data processing occurs locally on the device. However, as NNs often involve tens of millions of weight and activation data, running AI applications on edge platforms often demands substantial computational resources, leading to significant hardware overhead and memory footprint. Besides computational resources, storing excessive weight and activation data often requires large-volume memories. In addition, on-chip memories, with their high data storage density and advanced manufacturing processes, are particularly more vulnerable to soft errors caused by particle radiation, such as neutrons and alpha particles. These soft errors in memories can drastically lead to data corruption (e.g., bit-flip), impacting the system's reliability, especially for those resource-constrained edge platforms without reliability-hardening mechanisms. Therefore, the challenges lie in how to realize energy-efficient edge platforms with high reliability for AI applications, especially for safety-critical applications. Ensuring that AI algorithms can efficiently process data while maintaining reliability is crucial for the effective deployment of safety-critical AI applications at the edge, such as those in autonomous vehicles, medical devices, and industrial control systems. Namely, integrating AI algorithms into safety-critical applications requires comprehensive exploration of the reliability and computational efficiency of AI-enabled systems at the edge. Therefore, the endeavors of this dissertation focus on addressing the challenges posed by safety-critical AI applications at the edge, providing valuable insights for developing robust, fault-tolerant, and efficient AI System-on-Chips (SoCs) in critical domains.

Focusing on computational efficiency in the realm of AI SoC at the edge, both the arithmetic unit and the data flow are key factors targeting performance optimization. At the arithmetic unit level, approximate Multiply-and-Accumulate (MAC) operations offer a valuable solution by leveraging the intrinsic approximations presented in NNs. These operations effectively optimize latency, power consumption, and area utilization without significantly affecting accuracy. Furthermore, adopting mixed-precision computing can also contribute to the deployment of NNs for edge devices. In many comput-

ing scenarios, high-precision data computations, not always necessary, also require high overhead in edge devices. Using low-precision computations while maintaining low accuracy loss yields great performance benefits. Besides, to leverage the approximate and quantization-compatible features of NNs, Neural Architecture Search (NAS) methods offer a promising approach by reducing data precision with limited accuracy loss, minimizing memory footprint, and mitigating hardware resource consumption. Thus, integrating NAS-based methods with Multi-precision (MP) MAC and/or approximate MAC units emerges as a viable solution for resource-constrained edge applications. At the data-flow level, a well-organized data flow plays a pivotal role in achieving computational efficiency. Systolic data flow, as exemplified by the Tensor Processing Unit (TPU), emerges as a powerful strategy, enhancing speed and power efficiency for matrix and convolutional operations while aligning seamlessly with the data-reusable feature of NNs. On the other hand, to address the limitations of traditional Von Neumann architectures, novel approaches like near-memory computing (NMC) technology have been introduced. NMC, featuring distributed memory blocks near Processing Elements (PEs), ensures larger bandwidth and balanced latency in the PE array. Therefore, our work embraces both approximate MAC operations, mixed-precision computing, and advanced data-flow designs, such as systolic data flow and near-memory computing, holistically enhance the computational efficiency of AI applications at the edge.

Ensuring the reliability of AI SoCs, particularly for safety-critical applications at the edge, is important as these applications become more prevalent. Striking a delicate balance between performance, accuracy, and reliability is essential. The approximate feature of NNs suggests that not all soft errors are catastrophic, and their impact on NNs' output may vary due to the differing fault tolerance capabilities of each weight and activation in the NNs. Consequently, some errors may be tolerant with minimal impact on the network's performance, while others could lead to significant deviations in the output. Simultaneously, highly precision-reduced NNs introduce complexities as each bit needs to carry more information, necessitating a thorough analysis of the reliability of these NNs. Besides, in edge devices, SoCs commonly consist of microcontrollers and peripherals, with hardware AI accelerators integrated as peripherals. Conducting reliability analysis, especially for AI applications, can be intricate due to potential black-box testing limitations. To address this, we designed various forms of AI SoCs and conducted fault injection experiments and radiation experiments from a white-box view, offering deeper insights into the reliability of our SoCs. Namely, comprehensive exploration is essential for identifying weak points and developing efficient error detection and mitigation techniques with minimal overhead. Thus, our work aims to explore the reliability and mitigation techniques of AI SoCs at the edge.

First, this dissertation aims to design energy-efficient AI accelerators catering to common operators such as convolution, pooling, and transformer operations. These operators are fundamental components in various NN architectures, playing crucial roles in tasks like feature extraction, down-sampling, and sequence modeling. By optimizing

for these widely used operators, the designed accelerators can effectively enhance the performance and efficiency of a broad range of AI applications. Thus, three distinct AI accelerators are introduced to evaluate performance across different aspects: multiplier design, data flow, and structural sparsity. Firstly, a multi-bit-width (MBW) Booth accelerator is proposed, supporting INT2/4/8 computation. Leveraging low-precision computation in NNs significantly reduces power consumption, and NAS effectively generates mixed-precision NNs with a high proportion of low-precision layers, enhancing power efficiency with minimal accuracy loss. The accelerator, implemented on Microchip MPF300T FPGA, achieves peak performances of 25.6 Giga operations per second (GOPS) for INT8, 51.2 GOPS for INT4, and 102.4 GOPS for INT2 in the convolutional layer, respectively. Secondly, recognizing that common NNs exhibit sparsity exceeding 60%, a low-power sparse accelerator is introduced to maximize MAC-unit utilization and power efficiency. This accelerator incorporates a radix-4 Booth multiplier for pre-encoding weights, reducing partial product count and encoder power consumption. Implemented on a 28nm process, the accelerator achieves 7.0325 Tera operations per second per Watt (TOPS/W) at 50% sparsity and scales up to 14.3720 TOPS/W at 87.5% sparsity. Finally, a Booth-value-confined MP accelerator is presented with a near-memory-computing-friendly data flow. The Booth-value-confined PE minimizes power and area by 82% and 70%, respectively, supporting MP computations. The NMC-friendly data flow, incorporating a partial sum scheduler, facilitates data scheduling in various NN models.

Next, this dissertation presents three SoCs incorporating the developed AI accelerators for AI applications. In flash-based FPGA design, a high-performance SoC with a MBW AI accelerator is proposed. A 2-stage open-source RISC-V core is selected as a centric controller, and a 16×8 PE array is realized for parallel computation. In ASIC design, an edge AI SoC is developed to accelerate MP NNs. Measured on a 22nm testchip with 16 PEs, the peak power efficiency is 25.76/28.61/29.11 TOPS/W when running DarkNet19/ResNet34/VGG16 networks. Furthermore, an optimized work about a multi-core multi-mode SoC is proposed for versatile applications. Accelerator mode is realized for the acceleration of convolutional operation and matrix multiplication with a peak power efficiency of 6.463 TOPS/W. The multi-core mode features a fully programmable structure supporting all general-purposed computations with a peak power efficiency of 5.44 TOPS/W. Furthermore, the dual-core-lock-step (DCLS) mode is adapted for safety-critical applications. Besides, the JTAG debugging port and scan chain are integrated into the chip to facilitate reliability analysis and debugging.

Finally, this dissertation presents the reliability assessment of our SoCs through fault injection (FI) experiments and irradiation experiments (neutron and alpha). The assessment aims to offer valuable insights and serve as an essential reference for future reliability-aware designs. First, the work evaluates the reliability of an FPGA-based AI SoC featuring the MBW accelerator. FI results highlight control and status registers (CSRs) as the most vulnerable component in the accelerator, with layers employing

INT2/4 data being more susceptible than INT8 data due to the increased information carried by each bit in low-precision compared to high-precision. Additionally, the study reveals that MBW networks are more fragile than single-precision networks, and high bits in low-precision are more prone to misclassification errors. Neutron radiation experiments pinpoint the RISC-V core, particularly the Instruction/Data Tightly Coupled Memory (I/DTCM), as the most sensitive component in the SoC. The bit-flips of CSRs in the accelerator contribute to critical Detected Unrecoverable Errors (DUEs), such as dead loops and unresponsiveness. Interestingly, weights and activations exhibit the lowest cross section to system failures in the entire SoC, implying their influence can be disregarded for MBW CNNs in safety-critical AI applications for edge devices. For the ASIC design of a single-core with Booth-value-confined MP accelerator, the dissertation offers a comprehensive evaluation of the reliability of a 22nm AI SoC from both pre-silicon and post-silicon perspectives, with post-silicon alpha experimental results serving as a golden reference. Notably, the I/DTCM emerges as the most vulnerable component, storing static program data, NN model information, and cache data. Flash execute-in-place (FlashXIP) execution mode leads to critical errors in the SoC due to the non-responsiveness of AI tasks caused by upsets in DTCM. The FI experiment on NNs indicates that the MP NN is considerably more reliable than other system components, suggesting minimal impact on the system from bit-flips in accelerator memories related to NN data. White-box and black-box simulation results align, emphasizing their applicability to Commercial off-the-shelf (COTS) platforms without error detection or correction mechanisms, as long as FI can operate on any memory at any given time.

Overall, this dissertation aims to achieve energy-efficient AI SoC designs with necessary and sufficient reliability consideration. The accelerator design leverages sparsity, MP computing, and approximate computing techniques. Additionally, an exploration of high-efficiency data flow enhances data reuse and facilitates efficient NN mapping. In SoC design, a proposed multi-core multi-mode is investigated to support efficient general-purpose computations and fault-tolerant AI applications. Furthermore, fault injection and radiation experiments are conducted to analyze the reliability of the proposed SoCs, providing valuable insights for designing reliable and energy-efficient hardware at the edge.