IDEC Chip Design Contest

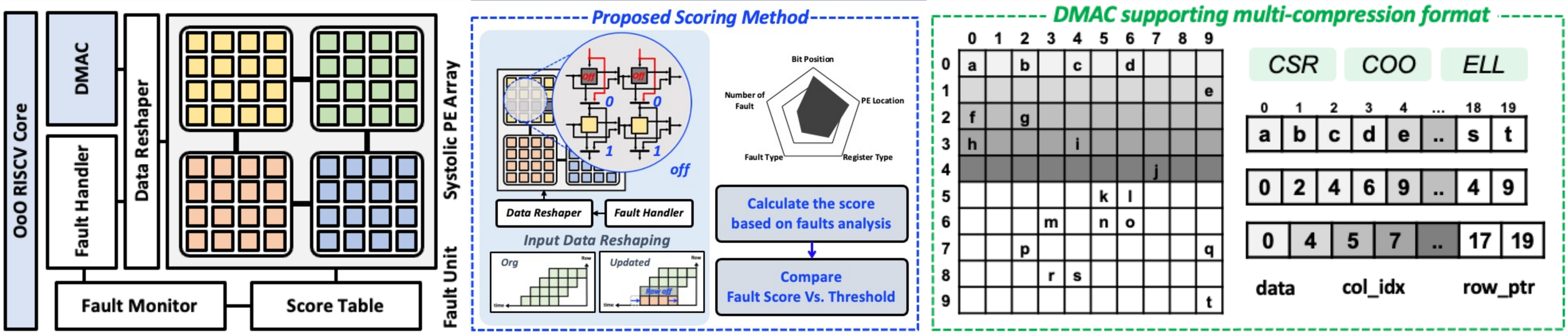
Sparsity-Aware Multi-Compression Format Support Fault-tolerant Neural Processing Unit with Custom Out-of-Order RISC-V CPU Core

Sunyoung Park, Hyunji Kim, Sujin Kim and Ji-Hoon Kim
Department of Electronic and Electrical Engineering, Ewha Womans University, Seoul 03760, South Korea
Email: {sunyoung_p, hyunjikim0223, iamsujinkim}@ewhain.net, jihoonkim@ewha.ac.kr

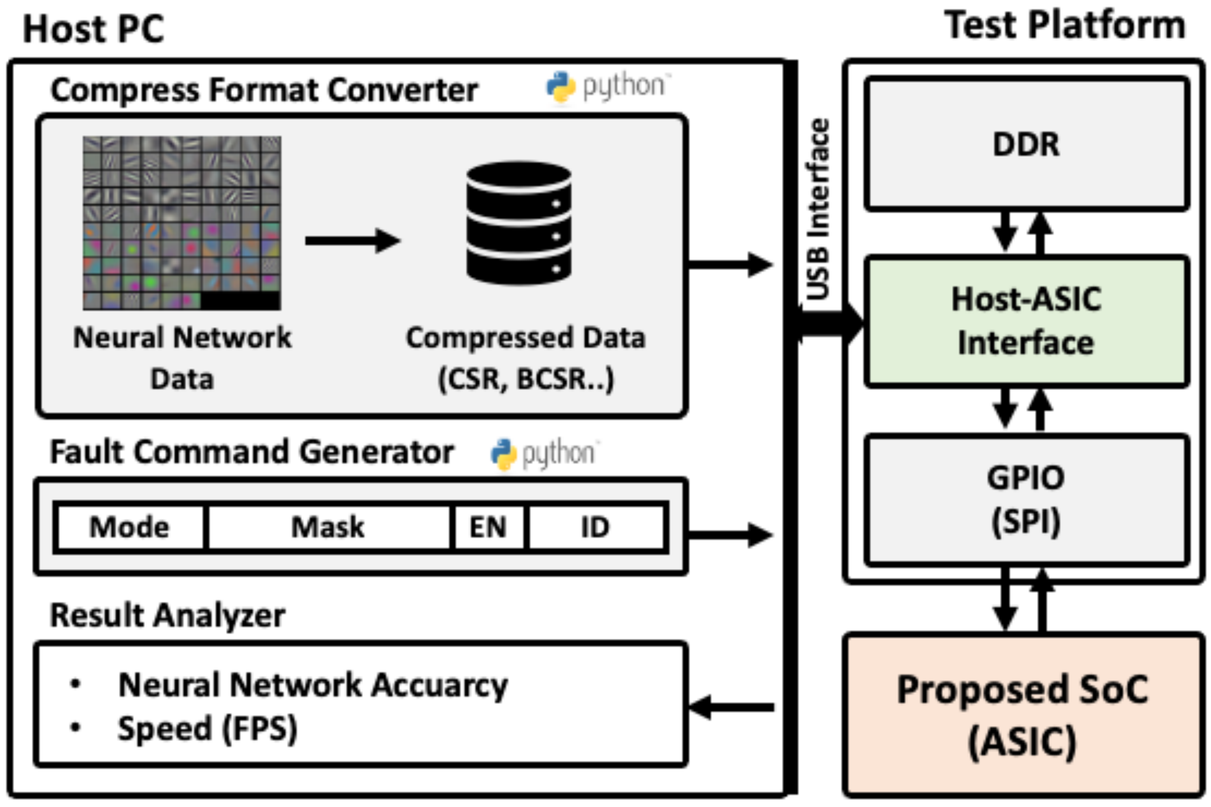
Introduction

Recently, neural networks have become widely used in safety-critical and real-time applications, such as automotive driving cars and drones. As these applications can cause significant problems due to incorrect predictions, the need for a fault-tolerant neural processing unit (NPU) has arisen. Meanwhile, data compression is essential to reduce the latency of the NPU and the number of DRAM accesses. The sparsity of the matrix used in neural network operations exhibits various patterns depending on the algorithm and learning method used, leading to an increase in the efficiency of DRAM storage and operation. To accomplish this, we propose a fault-tolerant NPU that includes a DMAC supporting sparsity-aware compression format conversion. Additionally, a custom RISC-V CPU core that effectively supports out-of-order (OoO) handles preprocessing and postprocessing for on-chip neural network workloads.

Architecture



Methods



Results

Technology	Samsung 28nm CMOS
Core Voltage	1.0V
Max. Operating Frequency	400MHz
Circuit Type	Digital (Cell-based) Design
On-Chip Memory	100KB (NPU) + 150KB (Global Buffer) + 262KB (Cache)
Core Size	3.5mm x 3.5 mm

Reference

[1] Zhang, Jeff Jun, et al. "Analyzing and mitigating the impact of permanent faults on a systolic array based neural network accelerator." *2018 IEEE 36th VLSI Test Symposium (VTS)*. IEEE, 2018.

Conclusion

1.This work contributes to fault-tolerant NPU in two ways: a compression-support DMAC design that reduces the number of DRAM accesses, resulting in reduced latency and efficient power consumption, and a custom RISC-V OoO CPU core that speeds up the pre/post processing for neural networks.
2.The test was conducted in a fault-occurrence environment by disabling specific rows/columns, and the impact of faults on the neural network was measured as a score by counting the results that exceed a certain value. The design supports both normal and split modes to reduce the range of fault spreading to neighboring processing elements (PEs) based on fault characteristics.

This work was supported by the Industrial Fundamental Technology Development Program (No. 20019367, Development of Low Power AI Architecture for AIoT) funded by the Ministry of Trade, Industry & Energy (MOTIE) of Korea.

The chip fabrication and EDA tool were supported by the IC Design Education Center(IDEC), Korea.

반도체설계교육센터 IC DESIGN EDUCATION CENTER