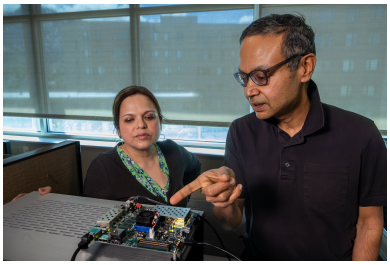
STRIVE: Empowering a Low Power Tensor Processing Unit with Fault Detection and Error Resilience

04/01/2026



STRIVE: Empowering a Low Power Tensor Processing Unit with Fault Detection and Error Resilience

Electrical & Computer Engineering

Abstract

The error-resilient systolic array seeks to solve the challenges of fault-prone AI inference architectures at the edge. It achieves this through a low overhead detection technique for faulty Multiply-Accumulate (MAC) units in a systolic array AI hardware. This opens opportunities for device level process variation (PV) and fault tolerance at the circuit-architectural layer. The technology enables the correct functioning of AI hardware during inference, even in the presence of faulty elements. Overall, this technology enhances resilience in low power AI operations.

Problem

Diminishing power budgets combined with manufacturing variability for edge AI can cause frequent faults. The fault-prone computations jeopardize the inference accuracy of DNN models, without viable solutions via conventional techniques.

Solution

USU Researchers have designed a unique fault-resilient pipeline for AI hardware acceleration. By transforming the pipeline design to relax compute-timing restrictions, there is greater throughput and higher levels of fault resiliency.

Value Proposition

This tech makes AI- architecture in edge environments more resilient to faults. It permits post-fabrication fault detection, calibration and functionally correct AI hardware operation in the presence of calibrated faulty elements.

Benefit

Global use of Internet of Things devices and artificial intelligence like deep neural networks (DNN) continue to expand rapidly. However, diminishing power budgets combined with manufacturing variability for edge AI can cause frequent faults. Fault-prone computations jeopardize the inference accuracy of DNN models, making architecture for edge AI less reliable. However, conventional solutions for preventing or catching faults, like guard banding, are not effective in edge environments.

USU researchers introduce a novel fault detection technique in a systolic array AI hardware accelerator. This technology also includes techniques for error reduction to enable accurate AI inference in the hardware accelerator. By hopping or correcting computations in faulty MAC elements, this technology enhances the resilience of edge AI architectures against faults.

This technology is meant to improve the reliability and efficiency of edge AI. Companies looking to create edge AI hardware can use this technology to give their product increased reliability and efficiency. Edge AI devices include industries like smartphone and tablets, wearables, medical devices, and smart home and other Internet of Things devices.

Market Application

This technology is meant to improve the reliability and efficiency of edge AI. Companies looking to create edge AI hardware can use this technology to give their product increased reliability and efficiency. Edge AI devices include industries like smartphone and tablets, wearables, medical devices, and smart home and other Internet of Things devices.

Inventors

Sanghamitra Roy

Noel Daniel Gundi

Koushik Chakraborty

USU Department: **Electrical and Computer Engineering**

Funding

This invention was made with government support under grant no. CNS2106237 awarded by the National Science Foundation. The government has certain rights in the invention.

USU Reference No. C25026

Publications

- Gundi et al., STRIVE: Empowering a Low Power Tensor Processing Unit with Fault Detection and Error Resilience. ACM Trans. Des. Autom. Electron. Syst. 30, 2, Article 16 (March 2025), 25 pages.Gundi et al., STRIVE: Enabling Choke Point Detection and Timing Error Resilience in a Low-Power Tensor Processing Unit, 2023 60th ACM/IEEE Design Automation Conference (DAC)

Intellectual Property

Status: Filed

Pending U.S. Non-Provisional Patent Application filed January 2025