AI Hardware: Test, Reliability and Security (AI-TREATS)

May 28th, 2021 (virtual event), 15:00-19:35 CET

Program

15:00	**Introduction** *Haralampos Stratigopoulos, Sorbonne Université, CNRS, LIP6* *Ioana Vatajelu, Université Grenoble Alpes, CNRS, TIMA*
15:05	**Invited presentations 1** *Session chair: Fei Su, Intel, USA* ***15:05 Improving reliability of DNN accelerators*** *Kanad Basu, UT Dallas, USA* ***15:30*** **Co-design of Quantized Neural Networks for Reliable Inference on FPGAs** *Giulio Gambardella, Xilinx, Ireland* ***15:55*** **Embracing the Unreliability of Memory Devices for Neuromorphic Computing** *Damien Querlioz, Université Paris Saclay, France* ***16:20*** **Software Based Dependability Management of Neuromorphic Computing** *Anup Das, Drexel University, USA*
16:45	**Break**
17:00	**Invited presentations 2** *Session chair: Martin Andraud, Aalto University, Finland* ***17:00*** **Robustness through Errors: Enhancing Machine Learning Security through Approximate Computing** *Ihsen Alouani, Polytechnic University Hauts-De-France, France* ***17:25*** **Reliability and Security for Machine Learning Systems** *Muhammad Shafique, NYU Abu Dhabi, UAE, NYU, USA* ***17:50*** **Functional Criticality Classification of Structural Faults in AI Accelerators ("ML for ML")** *Krish Chakrabarty, Duke University, USA*
18:15	**Break**
18:30	**Panel:** ***Test and dependability for AI chips: any different from traditional chips?*** *Moderator: Mehdi Tahoori, KIT, Germany* **Panelists:** - *Alberto Bosio, INL, Ecole Centrale de Lyon, France* - *Yiorgos Makris, UT Dallas, USA* - *Siddharth Garg, NYU, USA* - *Eric Zhang, Horizon Robotics, Canada* - *Fadi Maamari, Synopsys, USA*
19:30	**Closing remarks** *Haralampos Stratigopoulos, Sorbonne Université, CNRS, LIP6* *Ioana Vatajelu, Université Grenoble Alpes, CNRS, TIMA*

Detailed program

15:00 – 15:05 Introduction

Haralampos Stratigopoulos, Sorbonne Université, CNRS, LIP6
Ioana Vatajelu, Université Grenoble Alpes, CNRS, TIMA

15:05 – 16:45 Invited presentations 1

Session chair: Fei Su, Intel, USA

15:05 - Improving reliability of DNN accelerators

Kanad Basu, UT Dallas, USA

Abstract: High accuracy and ever-increasing computing power have made deep neural networks (DNNs) the algorithm of choice for various machine learning, computer vision, and image processing applications across the computing spectrum. To this end, Google developed the tensor processing unit (TPU) to accelerate the computationally intensive matrix multiplication operation of a DNN on its systolic array architecture. Faults manifested in the datapath of such a systolic array due to latent manufacturing defects or single-event effects may lead to subversion of reliability. Although DNNs are known to resist minor perturbations with their inherent fault-tolerant characteristics, we show that the classification accuracy of the model plummets from 97.4% to 7.75% with a minimal fault rate of 0.0003% in the accelerator, implying catastrophic circumstances when deployed across mission-critical systems. In this talk, first, we will explore the impact of faults in different components of a DNN accelerator. Next, we will provide two techniques to dynamically detect reliability degradation due to faults in mission mode.

15:30 - Co-design of Quantized Neural Networks for Reliable Inference on FPGAs

Giulio Gambardella, Xilinx, Ireland

Abstract: “Quantized Neural Networks (QNNs) are increasingly being adopted and deployed especially on embedded devices, thanks to their high accuracy, but also since they have significantly lower compute and memory requirements compared to their floating point equivalents showcasing the need of co-design of neural networks for efficient implementation of inference. For safety critical applications, reliability has to be considered additionally to performance requirements and hardware cost. In this talk we’ll explore different possibilities to tackle reliable inference of QNNs on FPGAs by mean of different co-design strategies. “

15:55 - Embracing the Unreliability of Memory Devices for Neuromorphic Computing

Damien Querlioz, Université Paris Saclay, France

Abstract: The emergence of novel memories such as memristors or resistive memory (RRAM) opens the way to highly energy-efficient computation, near- or in-memory. However, this type of computation is not compatible with the use of conventional error correction codes (ECC) and has to deal with the severe unreliability of emerging memories. To address this issue, inspired by the architecture of animal brains, we present a manufactured differential hybrid CMOS/RRAM memory architecture suitable for neural network implementation that functions without formal ECC. We also highlight that using highly error-prone programming conditions only slightly reduces network accuracy, while bringing important benefits in terms of energy efficiency. Then, we present a second approach where the probabilistic nature of RRAM, instead of being mitigated, can be fully exploited to implement a type of probabilistic learning. We show that the inherent variability in hafnium oxide RRAM can naturally implement the sampling step in the Metropolis-Hastings Markov Chain Monte Carlo algorithm, and train experimentally an array of 16,384 RRAM cells to recognize images of cancerous tissues using this technique. These results highlight the interest in fully embracing the unreliable nature of emerging devices in neuromorphic designs.

References:

T. Hirtzlin, M. Bocquet, B. Penkovskyi, J.-O. Klein, E. Nowak, E. Vianello, J.-M. Portal and D. Querlioz, "Digital Biologically Plausible Implementation of Binarized Neural Networks with Differential Hafnium Oxide Resistive Memory Arrays", Frontiers in Neuroscience, Vol. 13, p. 1383, 2020.

T. Dalgaty, N. Castellani, C. Turck, K.-E. Harabi, D. Querlioz, E. Vianello, "In situ learning using intrinsic memristor variability via Markov chain Monte Carlo sampling", Nature Electronics, Vol. 4, p. 151, 2021.

16:20 – Software Based Dependability Management of Neuromorphic Computing

Anup Das, Drexel University, USA

Abstract: As process technology continues to scale aggressively, circuit aging in a neuromorphic hardware due to negative bias temperature instability (NBTI) and time-dependent dielectric breakdown (TDDB) is becoming a critical reliability issue and is expected to proliferate when using non-volatile memory (NVM) for synaptic storage. This is because NVM devices require high voltages and currents to access their synaptic weights, which further accelerate the circuit aging in neuromorphic hardware. Current methods for qualifying reliability are overly conservative, since they estimate circuit aging considering worst-case operating conditions and unnecessarily constrain performance. This talk will introduce RENEU, a reliability-oriented software approach to mapping machine learning applications to neuromorphic hardware, with the aim of improving system-wide reliability, without compromising key performance metrics such as execution time of these applications on the hardware.

16:45 – 17:00 Break

17:00 – 18:15 Invited presentations 2

Session chair: Martin Andraud, Aalto University, Finland

17:00 - Robustness through Errors: Enhancing Machine Learning Security through Approximate Computing

Ihsen Alouani, Polytechnic University Hauts-De-France, France

Abstract: Owing to their demonstrated efficiency, an increasing number of machine-learning and deep-learning architectures have been deployed to tackling a wide range of complex real-life problems. However, these systems suffer from a vulnerability to adversarial attacks: malicious additive noise that forces the output to a wrong label. State-of-the-art defenses such as adversarial training, gradient masking and pre-processing defenses are either vulnerable to white-box attack setting, or require substantial overhead in time and computation.

To address this problem, we propose a new type of defenses that is based on Approximate Computing paradigm. We leverage internal circuit-induced noise to either smooth the decision boundary or make it stochastic in time. Specifically, we present two defenses: the first is based on approximate convolution in image recognition deep neural networks, and the second uses Voltage Overscaling to harden Malware Detection-dedicated neural networks against evasive malware.

We show that such techniques make machine learning system more resilient to both white-box and black-box adversarial attacks while saving power as a by-product gain and without any retraining or fine-tuning effort.

17:25 - Reliability and Security for Machine Learning Systems

Muhammad Shafique, NYU Abu Dhabi, UAE, NYU, USA

Abstract: Deep Learning (DL) has emerged as the state-of-the-art for many Artificial Intelligence (AI) applications. Now-a-days models trained using DL are being used/research for safety-critical applications where even a single failure can lead to catastrophic results. Therefore, it is vital to ensure the robustness of DNNs against a wide range of dependability threats that include hardware-induced threats such as soft errors, aging, and manufacturing defects. Moreover, in the era of growing cyber-security threats, the intelligent

features of a smart CPS and IoT system face new type of attacks, requiring novel design principles for enabling Robust Machine Learning.

Traditional fault-mitigation techniques that are based on redundancy (e.g., Dual Modular Redundancy (DMR) and Triple Modular Redundancy (TMR)) are not effective for DNN-based applications due to their huge overheads arising from redundant hardware/execution of compute-intensive DNNs and synchronization issues. Other techniques such as Instruction Duplication (ID) and use of Error-Correcting Codes (ECC) also pose similar issues that lead to noticeable degradation in system's performance and power-/energy-efficiency. To address this, alternate techniques need to be developed that exploit the intrinsic characteristics of DNNs to boost their fault -resilience at low cost. These techniques should be integrated in the current state-of-the-art systems without affecting their performance-/power-efficiency. Towards this, this talk will provides an overview of different techniques for enhancing the dependability of DNN-based systems against different vulnerabilities. It also covers how each technique can contribute to the overall resilience of a DNN-based system, and how can they be integrated to offer protection against a wide spectrum of hardware-induced dependability threats while incurring low design-time and run-time overheads. Towards the end, I will provide a quick overview of different security aspects of the machine learning systems deployed in Smart CPS and IoT, specifically at the Edge.

17:50 - Functional Criticality Classification of Structural Faults in AI Accelerators ("ML for ML")

Krish Chakrabarty, Duke University, USA

Abstract: The ubiquitous application of deep neural networks (DNN) has led to a rise in demand for artificial intelligence (AI) accelerators. For example, the Tensor Processing Unit from Google, based on a systolic array and its variants are of considerable interest for DNN inferencing using AI accelerators. DNN-specific functional criticality analysis identifies faults that cause measurable and significant deviations from acceptable requirements. The criticality of faults can thus be used to guide test grading and quality assessment of AI accelerators at various stages of the product lifecycle. Fault-criticality assessment method can be used to target domain-specific use-cases—different models, e.g., AlexNet, GoogleNet, ResNet, etc., mapped to the accelerator for applications such as digit classification and pattern recognition. As the functional criticality of faults depends on the type of dataset, DNN model, and model-to-hardware mapping, individual catalogs containing lists of critical and benign faults can be pre-generated for every domain-specific use-case. These catalogs can then be used by the customer as a reference for assessing functional fault-criticality.

This presentation will examine the problem of classifying structural faults in an systolic-array accelerator based on their functional criticality. The speaker will first analyze pin-level faults in the processing elements (PEs) of a systolic array. Simulation results for the LeNet network with 32-bit and 16-bit data paths applied to the MNIST dataset show that over 93% of the pin-level structural faults in a PE are functionally benign. The speaker will next analyze the impact of stuck-at faults in the netlist of the PE and present a two-tier machine-learning (ML) based method to assess the functional criticality of these faults. The problem of minimizing misclassification will be investigated by utilizing generative adversarial networks (GANs). A two-tier ML/GAN-based criticality assessment method leads to less than 1% test escapes during functional criticality evaluation of structural faults. In the last part of the talk, the speaker will present a transferable multi-tier machine-learning framework that leverages graph convolutional networks (GCNs) for quick assessment of the functional criticality of structural faults. This framework obviates the need for tedious feature engineering and achieves up to 90% classification accuracy with negligible misclassification of critical faults.

Joint work with Arjun Chaudhuri (Duke University), Jonti Talukdar (Duke University), Fei Su (Intel), Jinwook Jung (IBM) and Gi-Joon Nam (IBM)

18:15 – 18:30 Break

18:30 – 19:30 Panel: Test and dependability for AI chips: any different from traditional chips?

Moderator: Mehdi Tahoori, KIT, Germany

Panelists:

- Alberto Bosio, *INL, Ecole Centrale de Lyon, France*
- Yiorgos Makris, *UT Dallas, USA*
- Siddharth Garg, *NYU, USA*
- Eric Zhang, *Horizon Robotics, Canada*
- Fadi Maamari, *Synopsys, USA*

19:30 – 19:35 Closing remarks

Haralampos Stratigopoulos, Sorbonne Université, CNRS, LIP6
Ioana Vatajelu, Université Grenoble Alpes, CNRS, TIMA