

# An FPGA-Based Weightless Neural Network for Edge Network Intrusion Detection

Zachary Susskind<sup>1</sup>, Aman Arora<sup>1</sup>, Alan T. L. Bacellar<sup>2</sup>, Diego L. C. Dutra<sup>2</sup>, Igor D. S. Miranda<sup>3</sup>, Mauricio Breternitz Jr.<sup>4</sup>, Priscila M. V. Lima<sup>2</sup>, Felipe M. G. França<sup>2,5</sup>, and Lizy K. John<sup>1</sup>

1- UT Austin, Austin, USA; 2- UFRJ, Rio de Janeiro, Brazil; 3- UFRB, Cruz das Almas, Brazil; 4- ISCTE, Lisbon, Portugal; 5- IT-Porto, Porto, Portugal

## Introduction

- Algorithms for mobile networking are increasingly moving towards the edge.
- 6G emphasizes adaptable algorithms for specific user scenarios, motivating broader use of FPGAs.
- We propose the FPGA-based Weightless Intrusion Warden (FWIW), based on *weightless* neural networks, for detecting anomalous network traffic on edge devices.

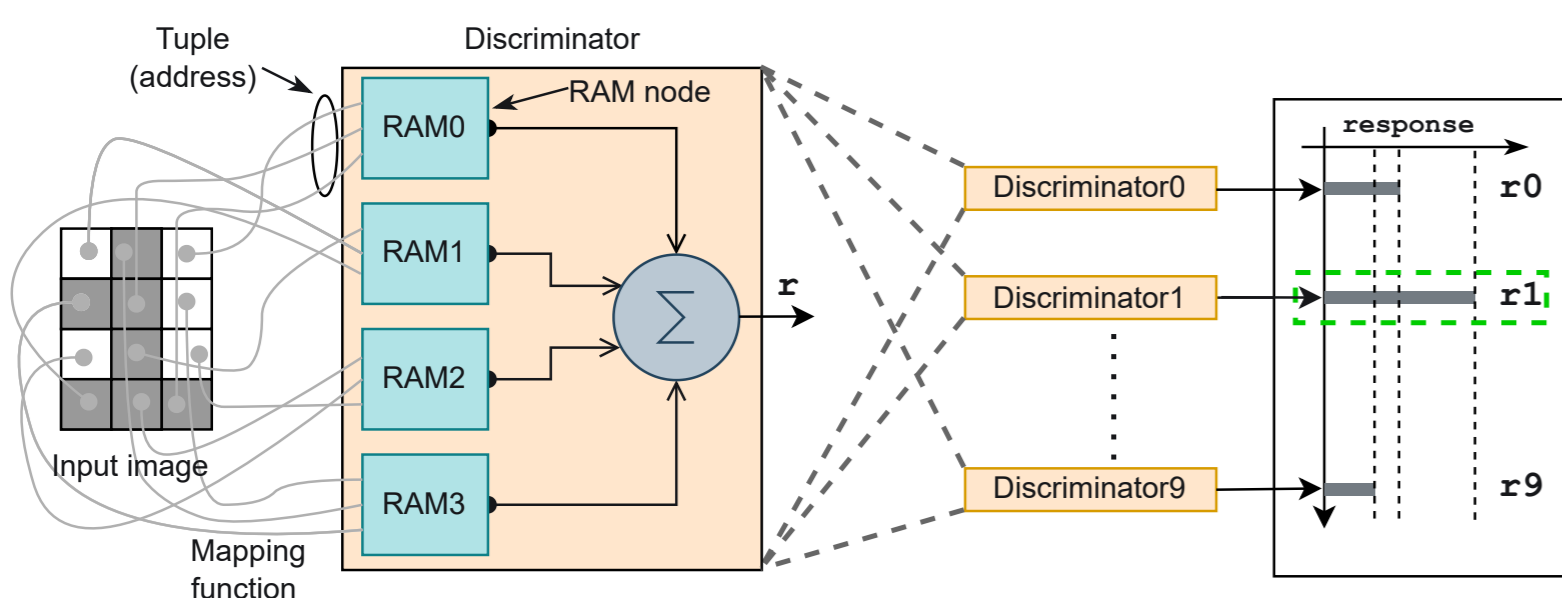


Figure 1: High-level view of a basic WNN architecture

- FWIW builds on the BTHOWeN [3] WNN architecture, which in turn builds on WiSARD (Fig. 1).
- WiSARD is a WNN for classification workloads which trains specialized submodels called *discriminators* for each output class.

## The FWIW Model

FWIW replaces the LUT-based RAM nodes in WiSARD with Bloom filter-based, differentiable nodes, shown in Figure 2.

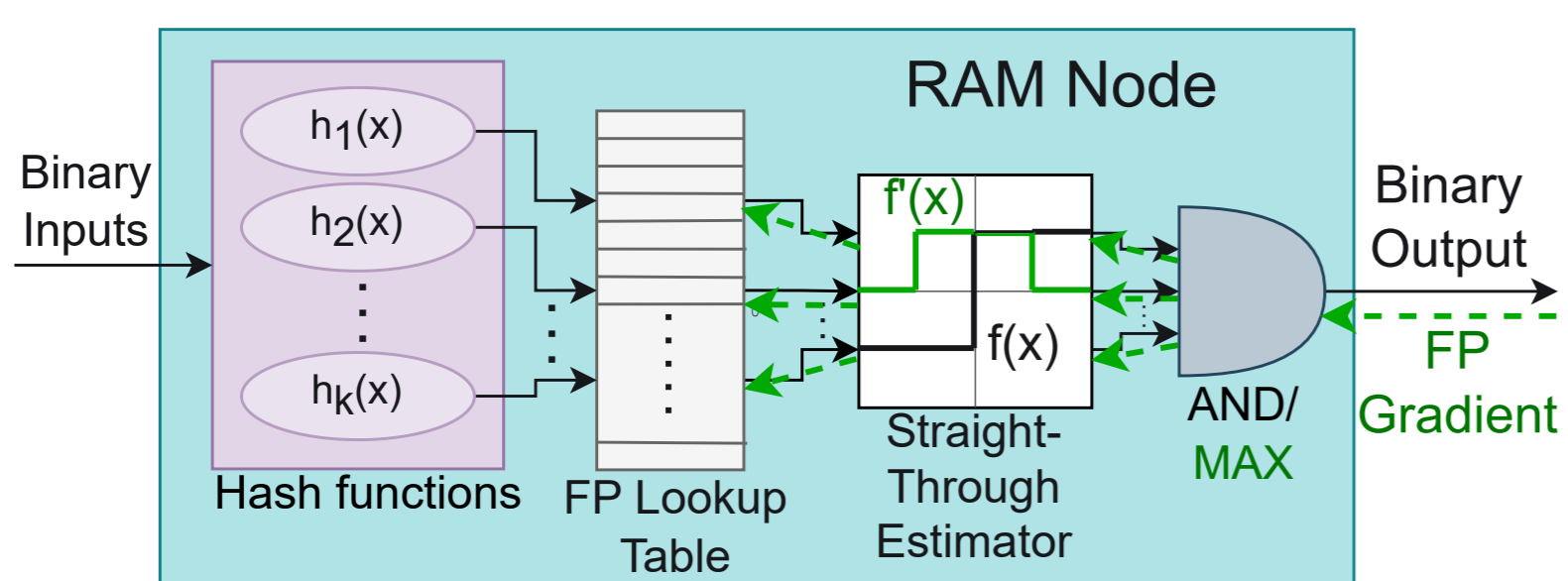


Figure 2: Multi-pass training for WNNs

FWIW incorporates two improvements over the prior work in BTHOWeN:

- FWIW uses a novel multi-pass WNN training rule based on the straight-through estimator.
- The WNN pruning techniques from [2] are used to reduce the final model size.

## Inference Accelerator Architecture

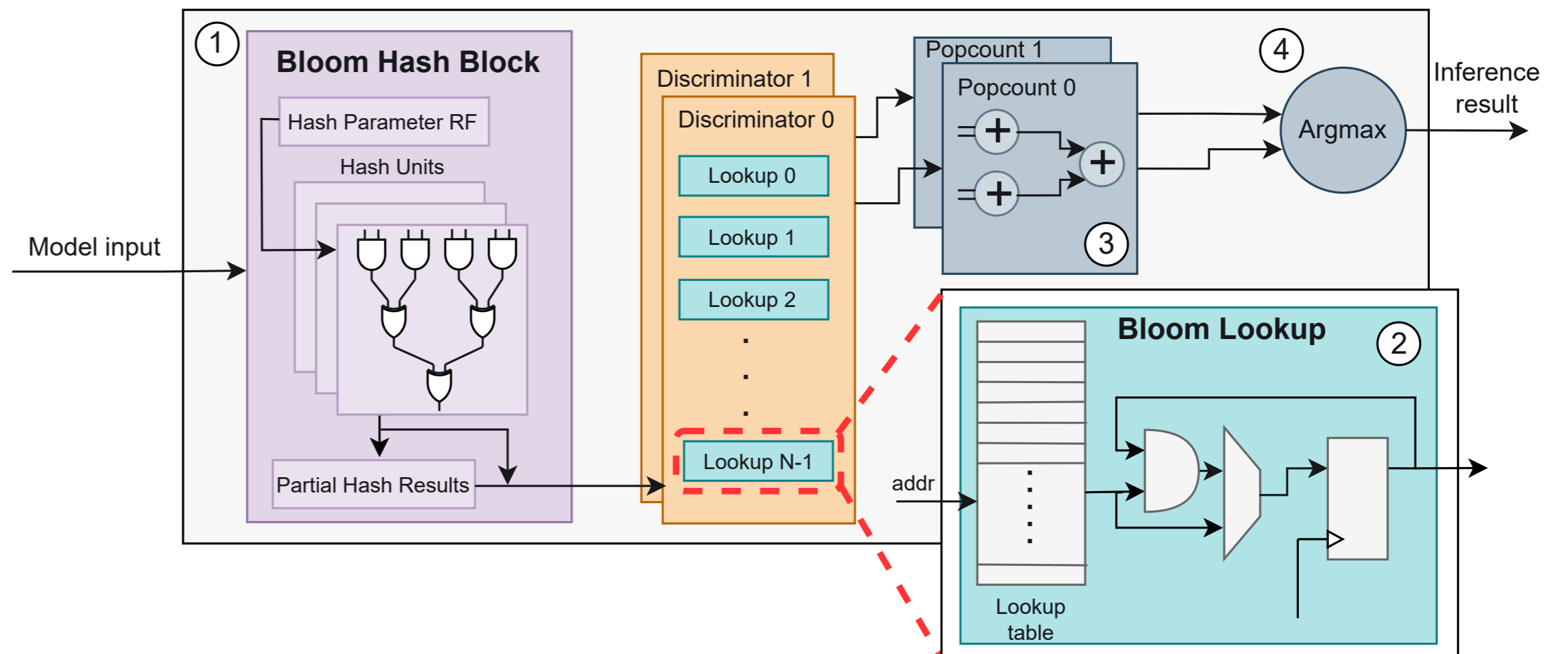


Figure 3: FWIW Accelerator Architecture

- The hardware architecture of the FWIW inference accelerator is shown in Figure 3:
1. Binary inputs are grouped and hashed to form Bloom filter addresses.
  2. Small LUTs are indexed using these addresses, and the AND across all hash functions is computed, implementing the lookup component of the Bloom filter.
  3. The popcounts of the LUT outputs in each discriminator are computed.
  4. The discriminator corresponding to the larger of the two responses (“normal” or “anomalous”) is chosen as the predicted class.

## Results

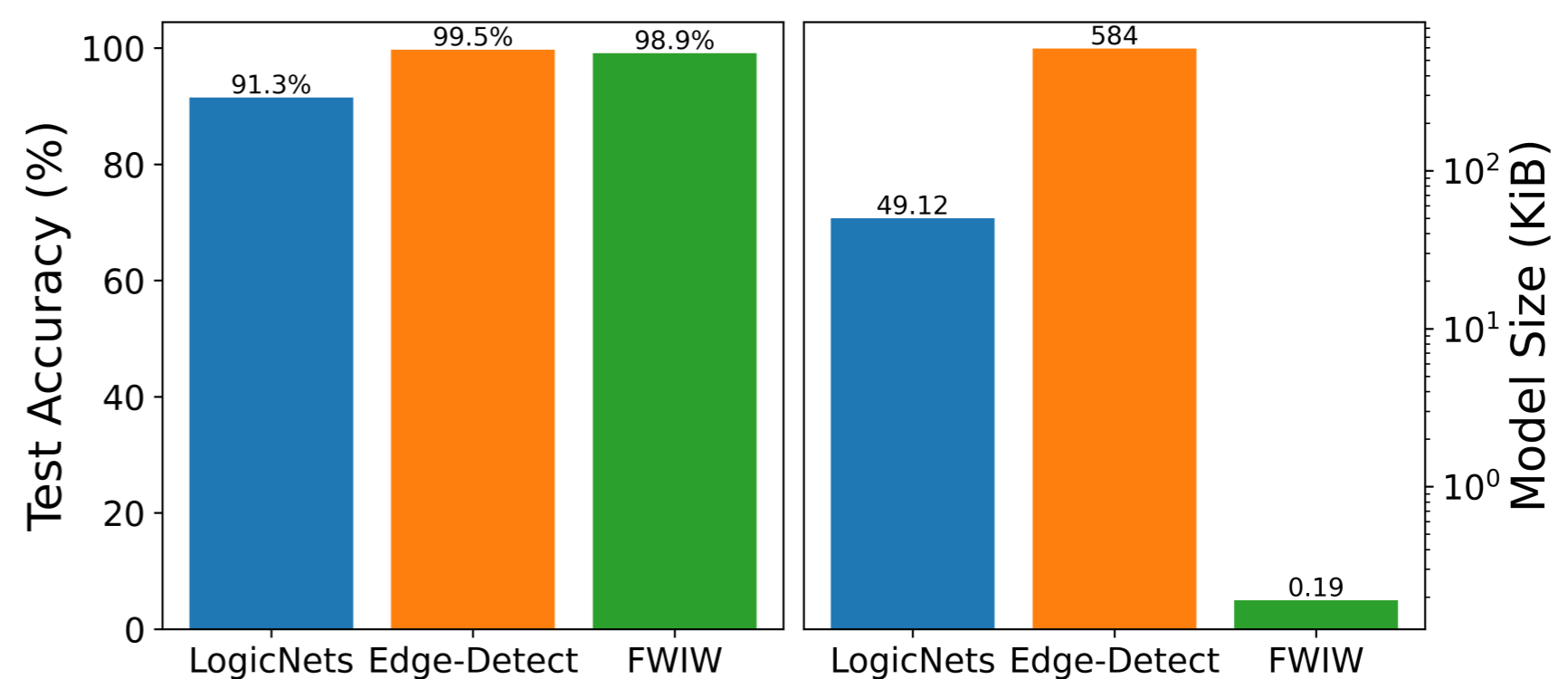


Figure 4: Comparison to Prior Work: LogicNets [4] and EdgeDetect [1]

- FWIW achieves 98.5% classification accuracy on the UNSW-NB15 dataset with a parameter footprint of just 192 bytes and a FNR of 0.10%.
- Compared to LogicNets, the best FPGA-based prior work, FWIW reduces error by 7.9x and model parameter size by 262x.
- Compared to Edge-Detect, the overall most accurate prior work, FWIW is 0.6% less accurate but has a parameter size >3000x smaller

Model Name	Clock (MHz)	Bus Width	Initiation Interval	Dynamic Energy (pJ/Sample)	LUTs	FFs	Latency (ns)
LogicNets	471	593b	1	654	15,949	1,274	10.5
FWIW	740	160b	1	73	269	538	10.8

Table 1: FPGA Implementation Results

- FWIW reduces energy per inference by 9.0x and LUT usage by 59x compared to LogicNets on a Xilinx Virtex UltraScale+ FPGA.
- Overall, FWIW demonstrates the viability of WNNs for edge intrusion detection.

## References

- [1] Praneet Singh et al. “Edge-Detect: Edge-Centric Network Intrusion Detection using Deep Neural Network”. In: *2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC)*. 2021, pp. 1–6. doi: 10.1109/CCNC49032.2021.9369469.
- [2] Zachary Susskind et al. “Pruning Weightless Neural Networks”. In: *ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. 2022, pp. 37–42. doi: http://dx.doi.org/10.14428/esann/2022.ES2022-55.
- [3] Zachary Susskind et al. “Weightless Neural Networks for Efficient Edge Inference”. In: *31st International Conference on Parallel Architectures and Compilation Techniques (PACT)*. 2022. doi: https://doi.org/10.1145/3559009.3569680.
- [4] Yaman Umuroglu et al. “LogicNets: Co-Designed Neural Networks and Circuits for Extreme-Throughput Applications”. In: *2020 30th International Conference on Field-Programmable Logic and Applications (FPL)* (2020), pp. 291–297.

Acknowledgment: This research was supported in part by Semiconductor Research Corporation (SRC) Task 3015.001/3016.001.