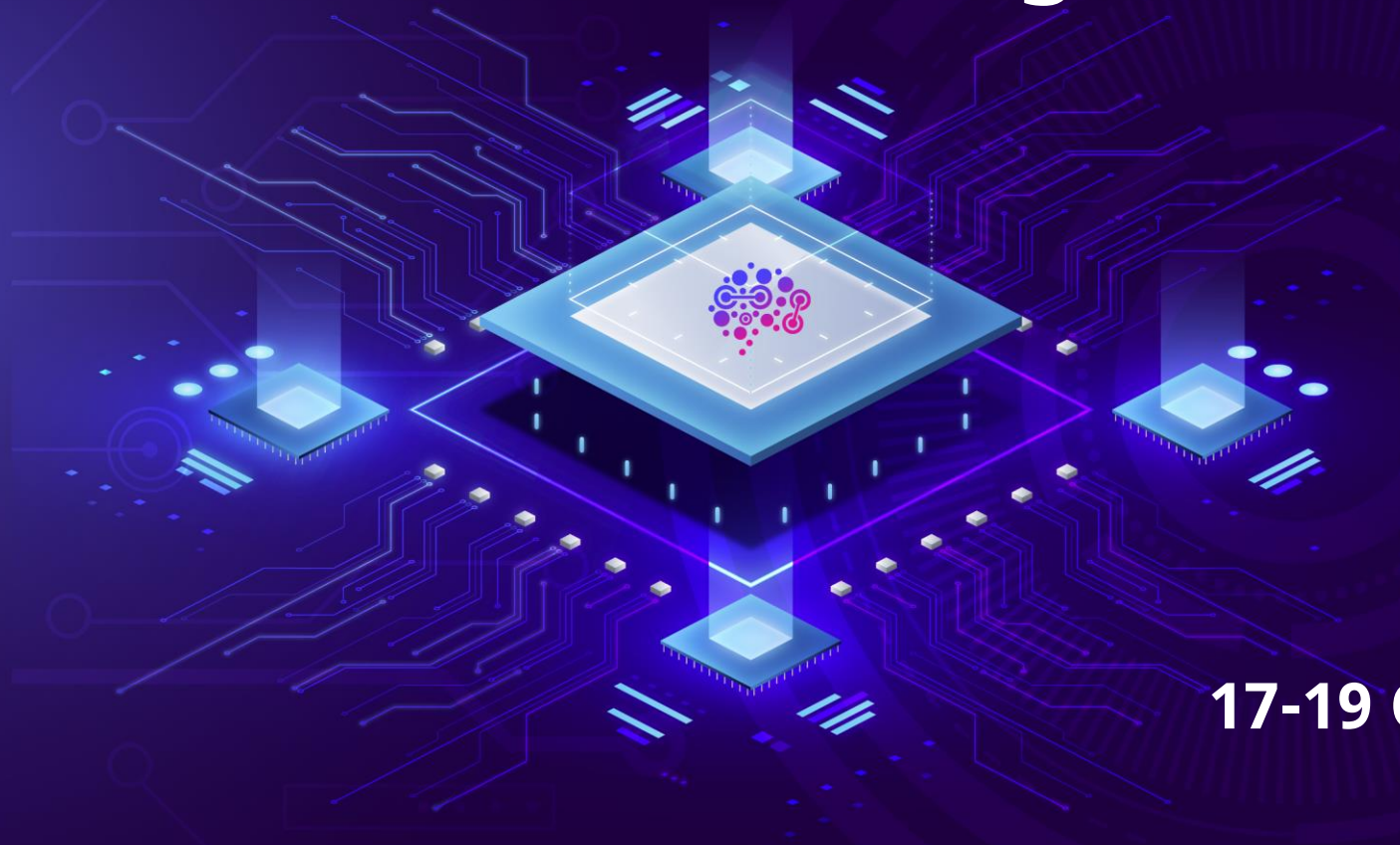
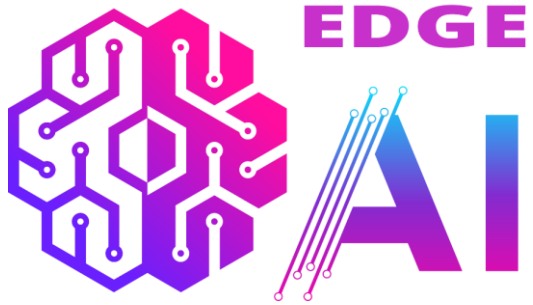
European Conference on EDGE AI Technologies and Applications - EEAI



Advancing emerging edge AI technologies and driving next generation intelligent applications.

17-19 October 2023, Athens, Greece

European Conference on EDGE AI Technologies and Applications - EEAI



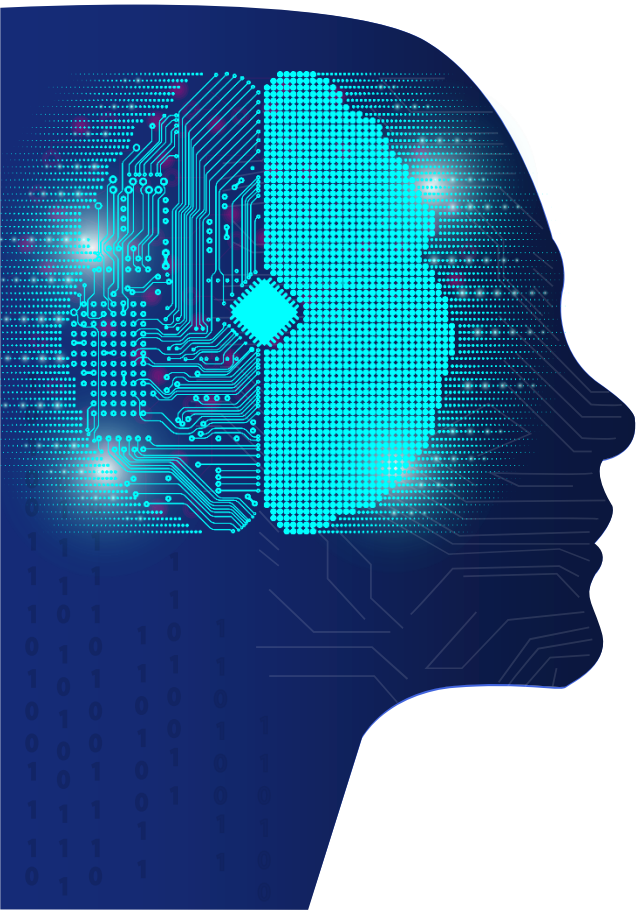
Cascaded Look Up Table Distillation of P4 Deep Neural Network Switches

Lorenzo De Marinis, Emilio Paolini, Rana Abu Bakar, Filippo Cugini, Francesco Paolucci, Scuola Superiore Sant'Anna, CNIT, ITALY

17-19 October 2023 Athens, Greece

Copyright © 2023

Presentation Outline



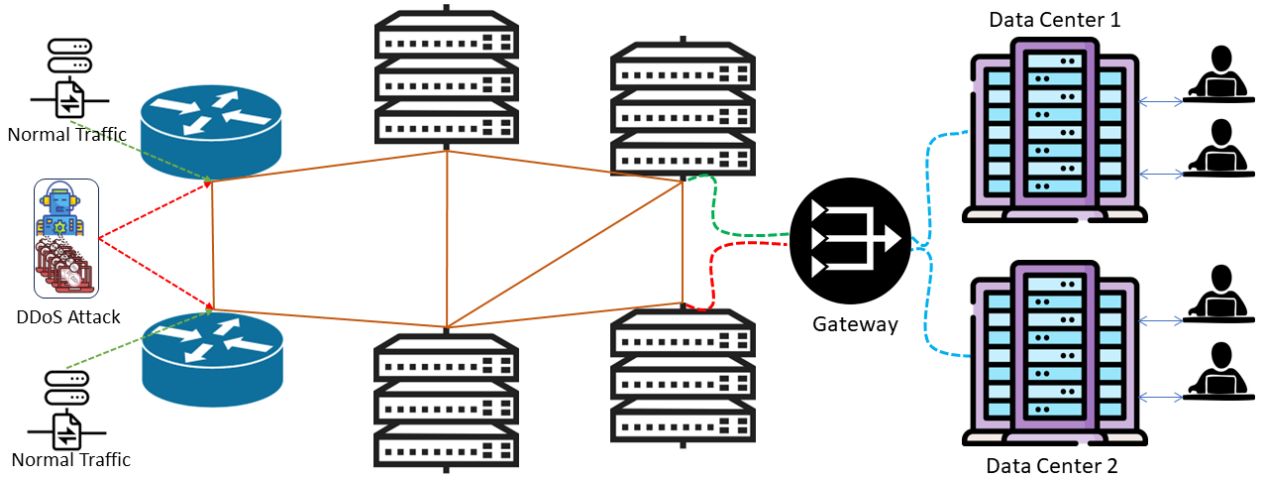
- Introduction
- State-of-the-art
- Research Goal
- Methodology
- Design and Implementation
- Experimental Results
- Future Research
- Conclusion

Copyright © 2023

Introduction

- DDoS Attack Scenario
  - Attackers use bots to generate large volumes of traffic
  - Traffic passthrough from different switch toward a data center
  - Attacker goal is to overwhelm the target and make it unavailable

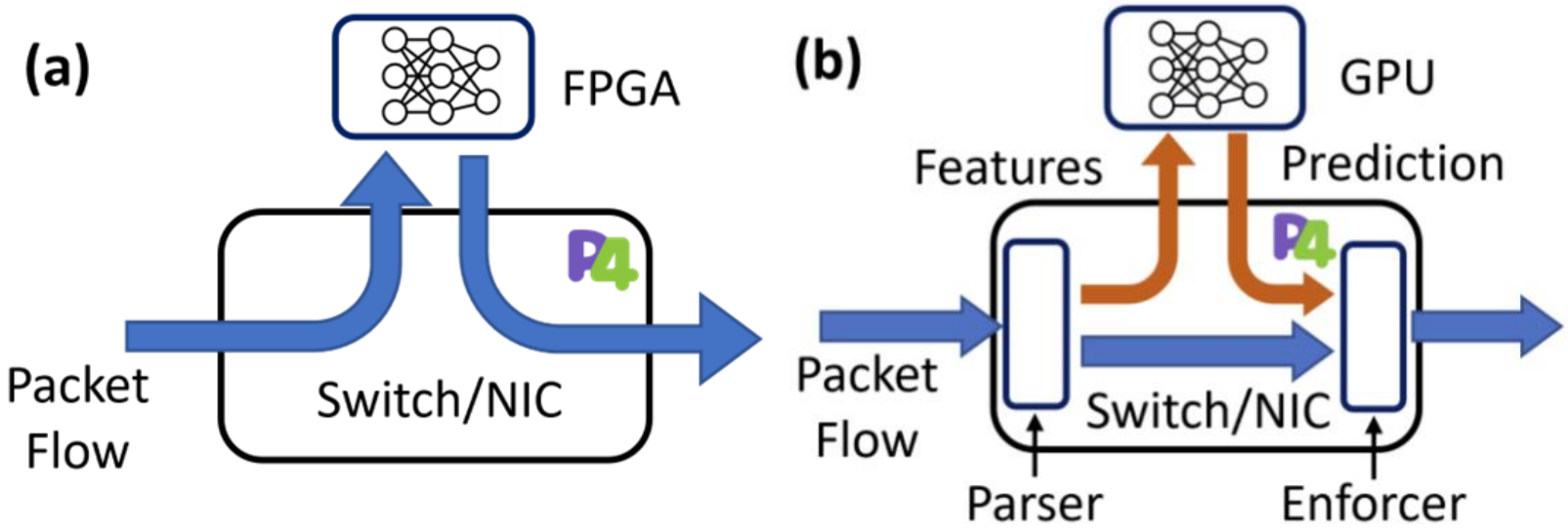
- Traditional DDoS Mitigation
  - Block traffic at the network gateway, but it disturbs legitimate traffic
  - There is no mechanism to detect DDoS attack at switch/router level



Copyright © 2023

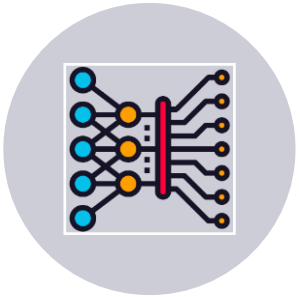
State-of-the-art

- DDoS has been addressed using Binary DNN or Machine Learning (e.g., SVMs, DTs, and RFs) running on an external FPGA.
- P4 Switches have been proposed to enable feature extraction, offloading DNN to a GPU
- However, offloading to external GPU still introduces drawbacks:
  a) Need for external servers
  b) Limited data rate
  c) Extra-latency
  d) Limited energy efficiency

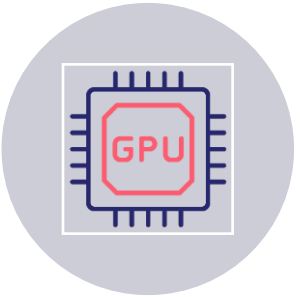


Copyright © 2023

Current issues

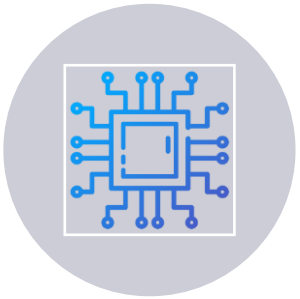


Network functions for QoS or cyber security attack detection based on AI need to operate at wire speed

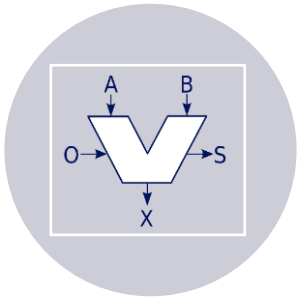


Current network implementations either:

- Rely on external CPU/GPU, deeply impacting latency
- Are limited to basic functionalities not supporting in-network AI



Programmable pipeline chipsets represent a potential technology to address these limitations



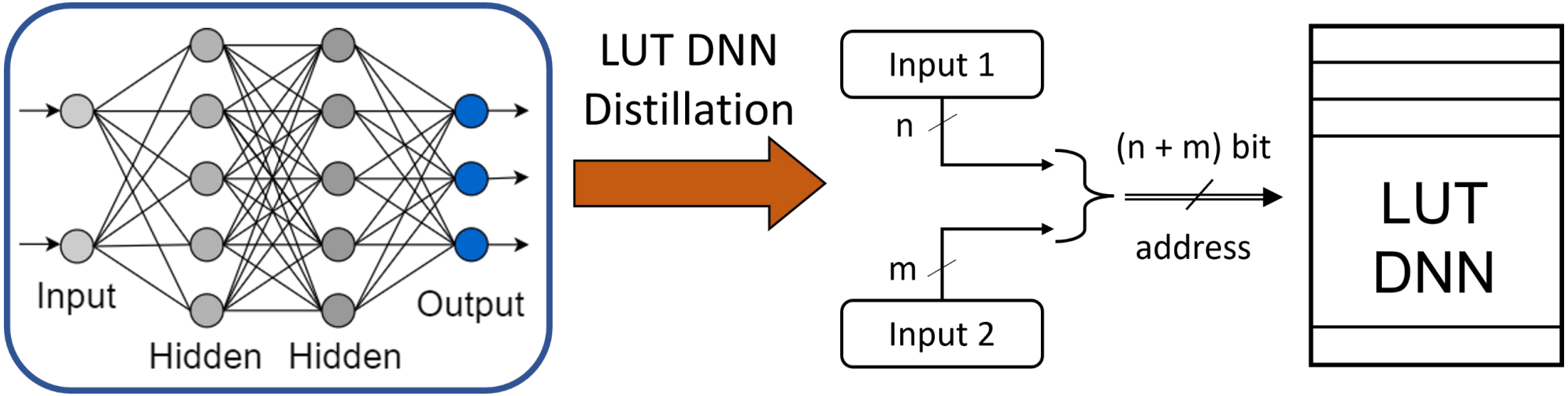
Current hardware chipsets do not have an Arithmetic Logic Unit (ALU)

- To operate at wire-speed, DNNs need to be computed using available in-network functions which are limited to match-actions in current hardware
- We propose a method to deploy DNNs within hardware pipelines with strictly limited arithmetic capabilities exploiting only Look-up Tables (LUT).

Copyright © 2023

Methodology

- This method distills input-quantized neural network into a LUT
- The method leverages the constraint of integer-encoded inputs as an advantage
- Inputs are paired and treated as a compound address of the final LUT
- The table is generated by simply collecting all the DNN outputs for all the input configurations, in this way no information is lost

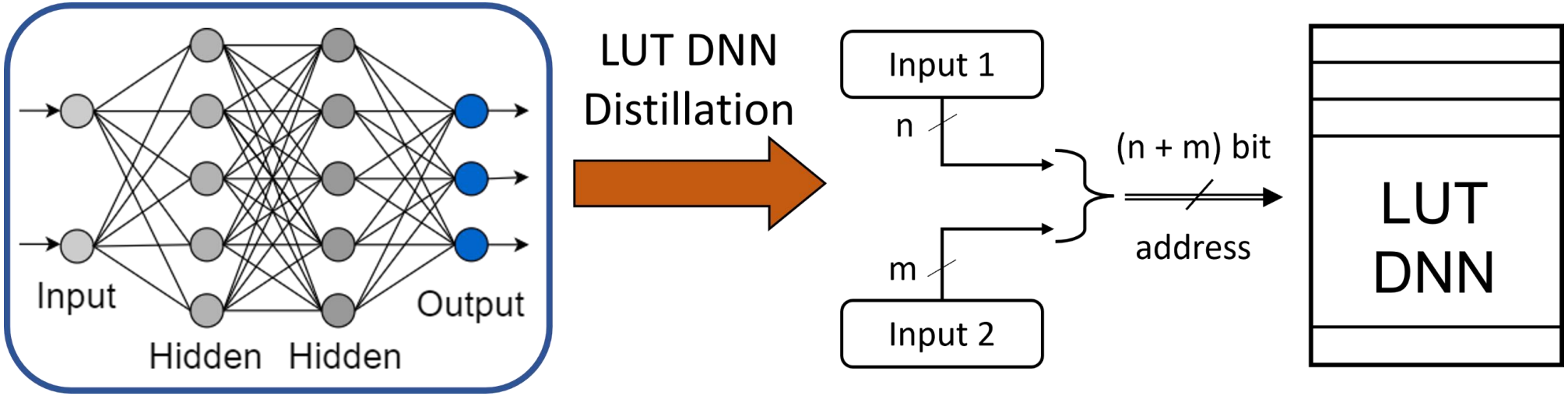


Copyright © 2023

Requirements and Drawbacks

1. Memory consumption (LUT size) grows exponentially with the number of bits composing the inputs
2. Input features should be integer-quantized
3. The number of outputs gives a linear increase of the final LUT

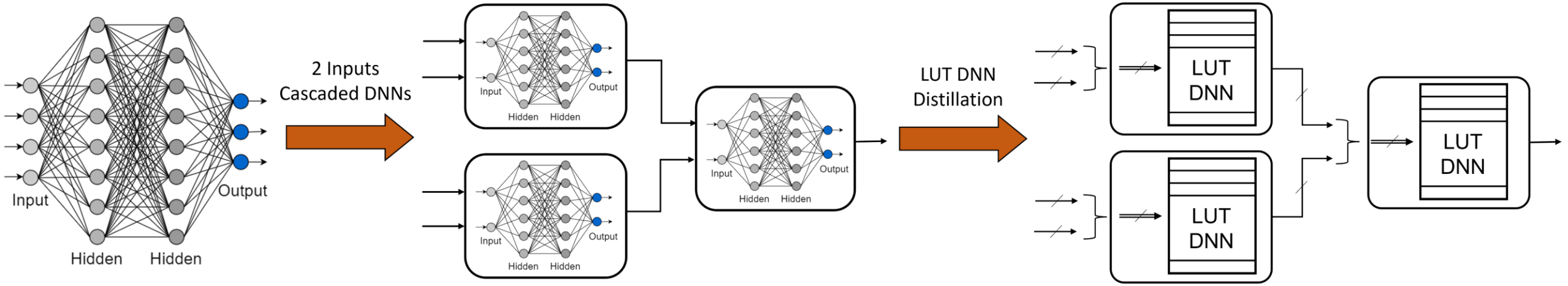
✓ No constraint on DNN complexity nor type



Copyright © 2023

Scaling Strategy

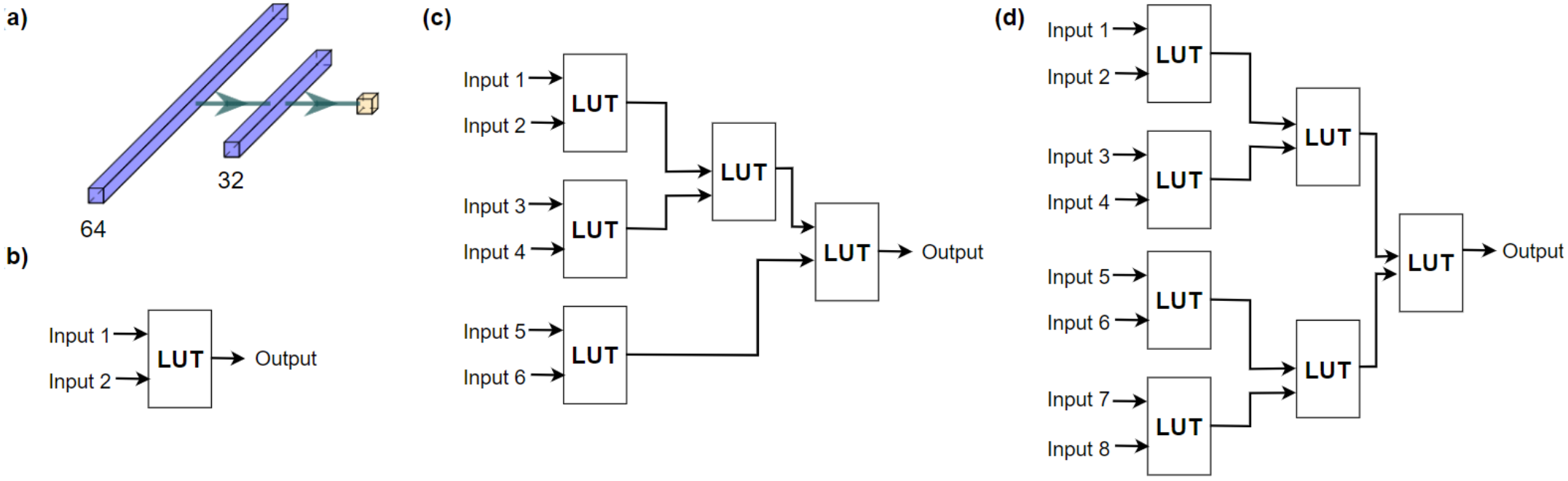
- In practical cases, several input features are required for DNN tasks
- To avoid memory consumption problems with higher number of inputs we propose the use of a hierarchical method
- The DNN model is substituted by a hierarchical structure made of 2-inputs DNNs
- Finally, each DNN is distilled in a LUT



Copyright © 2023

Demonstration with DDoS Mitigation

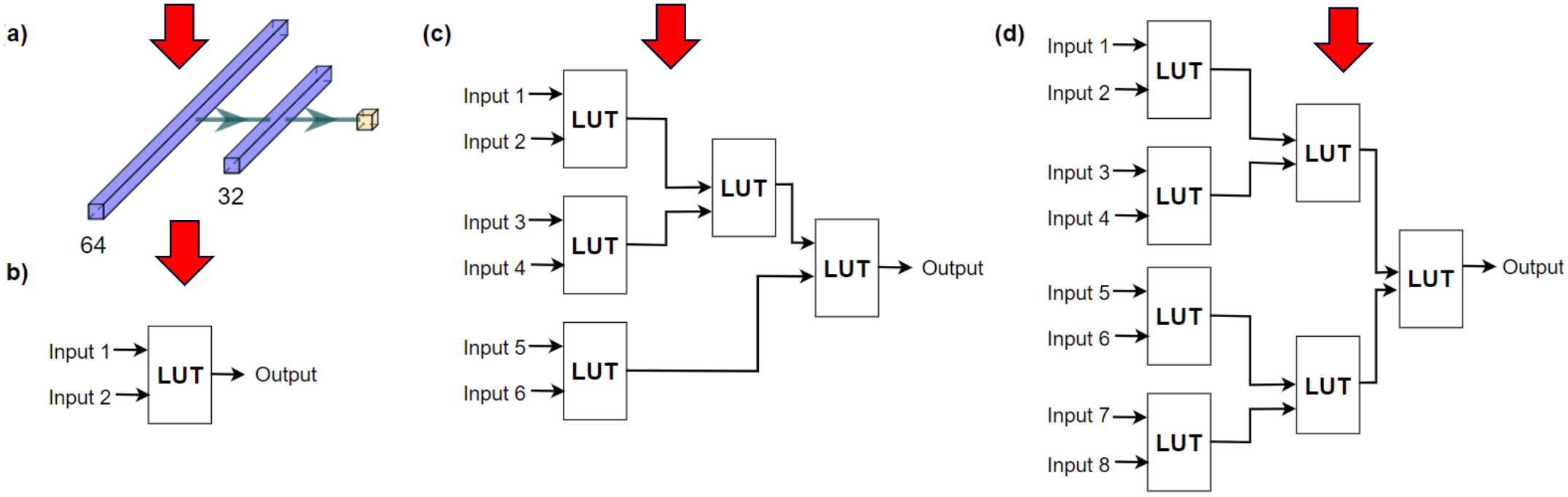
- We use the UNSW-NB15 Network Intrusion Detection dataset
- The DNN is trained as a Firewall, categorizing benign and malicious packets
- We chose 8 features, 6 stateless and 2 stateful
- Depending on type of feature used we developed 3 architectures



Copyright © 2023

Design and Implementation

- The basic 2-input DNN consists of 3 layers, with 64, 32, and 1 neurons
- Final model for the stateless-only case
- Final model for the stateful-only case
- Final model for the combined features case



Copyright © 2023

Experimental Results

- To reduce further the memory consumption, we considered also the case of a reduced bitwidth of 6 bits per input on average
- The model with Sateful-only features provides an F1-score lower than 50%
- The models with Sateless-only and combined features performs much better in both cases with an accuracy below 92 %

Considered Features	F1-score 8-bit inputs	F1-score with reduced bitwidh
Stateful	42.96 %	
Stateless	93.75 %	92.19 %
Combined	93.26 %	93.91 %

Copyright © 2023

Future Research

- Investigate the use of other LUT distillation techniques to improve the accuracy and/or efficiency of the proposed method.
- Explore the LUT distillation method for other network security tasks, such as intrusion detection and prevention.
- Deploy the proposed solution in different P4-enabled hardware platforms (e.g., Tofino switches, smart NICs, Data Processing Units) evaluating the performance in terms of required hardware resources, throughput, latency and power savings

Copyright © 2023

Conclusion

- Developed a knowledge distillation technique to convert DNNs into a cascaded architecture based on LUTs, with no information loss and with a resource usage independent from the DNN complexity.
- Demonstrated the working principle of the proposed technique by developing a ML model to act as a DDoS attack mitigator task with the UNSW-NB15 dataset.
- Achieved a F1-score above 92% for the stateless-only and aggregated features, even when using reduced bitwidths.

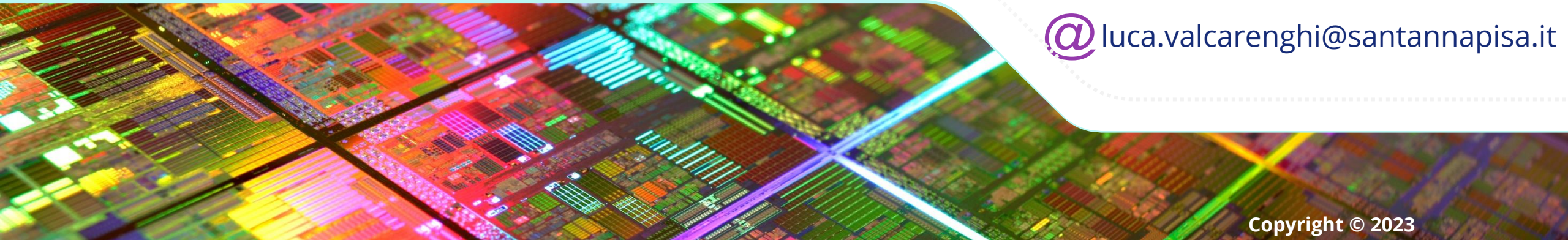
- “Cascaded Look Up Table Disitillation fo P4 Deep Neural Network Switches”, to be presented at Globecom 2023, December 2023, Kuala Lumpur, Malaysia

Copyright © 2023

EDGE AI

Thank You

For your attention



@ luca.valcarenghi@santannapisa.it

Copyright © 2023