
# TALLINN UNIVERSITY OF TECHNOLOGY
# TAICHIP TEAM, COMPUTER SYSTEMS

Jaan Raik
Department of Computer Systems
Tallinn University of Technology, ESTONIA

# Estonia: location


## Quick facts

- About half the size of Ireland
- Population: 1.3M
- Independence 1918 (occupied 1940-91)
- Member of EU, NATO, Euro
- Consistently highly ranked for quality of life, education, press freedom, digitalisation of public services and the prevalence of technology companies

# Tallinn: Medieval Heritage

- ✓ **Best preserved** Medieval Old Town in Northern Europe
- ✓ **UNESCO´s World Heritage** since 1997

# TalTech

- Established in 1918 as engineering college
- Second biggest Estonian university with ~12 000 students.
- A campus university with technology park
- 3 main cores:
  - Technology
  - Natural Sciences
  - Social Sciences
- Highest ranked university of technology in the Baltics by QS rankings

TAL TECH

# Department of Computer Systems

- 3 research groups for computer systems’ verification, reliability and security

*Prof. J. Raik*

*Prof. M. Jenihhin*

*Dr. Levent Aksoy*

- Professors

*Dr. T. Ghasempouri*

*Prof. G. Jervan (Dean)*

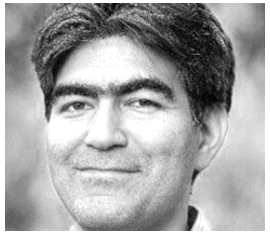
*Prof. P. Ellervee*

- Senior researchers:

*Dr. M. H. “Mojtaba” Ahmadilivani*

*Dr. A. Jutman*

- Adjunct professors:



*Prof. M. Daneshtalab*

*Prof. T. Hollstein*

- **TAICHIP** - Boosting TalTech Capacity in ***Reliable and Efficient AI-Chip Design***
  - Horizon Twinning action (Coordinator: **Maksim Jenihhin, TalTech**)
  - Duration: 36 months, start on September 1, 2024
  - Budget: 1,7 MEUR (EC contribution: 1,234,567.89 EUR)
  - *https://taichip.taltech.ee/*

# AI IN CLOUD AND EDGE


- **Cloud AI:** Data centers

- **Edge AI:** Embedded devices
- Strict constraints:
  - ❑ limited computational resources
  - ❑ strict energy budget
  - ❑ real-time latency requirements

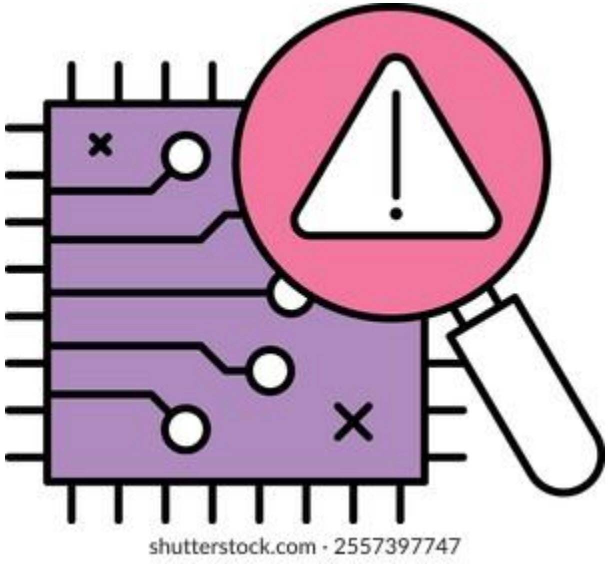
Edge:


TAL TECH

<u>**2023 Trend 2: Edge AI**</u> *"Gartner predicts that **more than 55%** of all data analysis by deep neural networks will occur at the point of capture in an edge system by 2025, up from less than 10% in 2021."*

# WHY RELIABILITY MATTERS?

- Cloud AI
  - Business calculation, highly critical
  - Silent data corruptions an increasing issue. Not captured by error detection.
  - Leads to ever increasing costs on screening
- Edge AI
  - Many applications in safety-critical domain
    - Automotive, drones, medical, …
  - Failures may lead to catastrophic consequences




- *But how to assure a desired level of functional safety?*

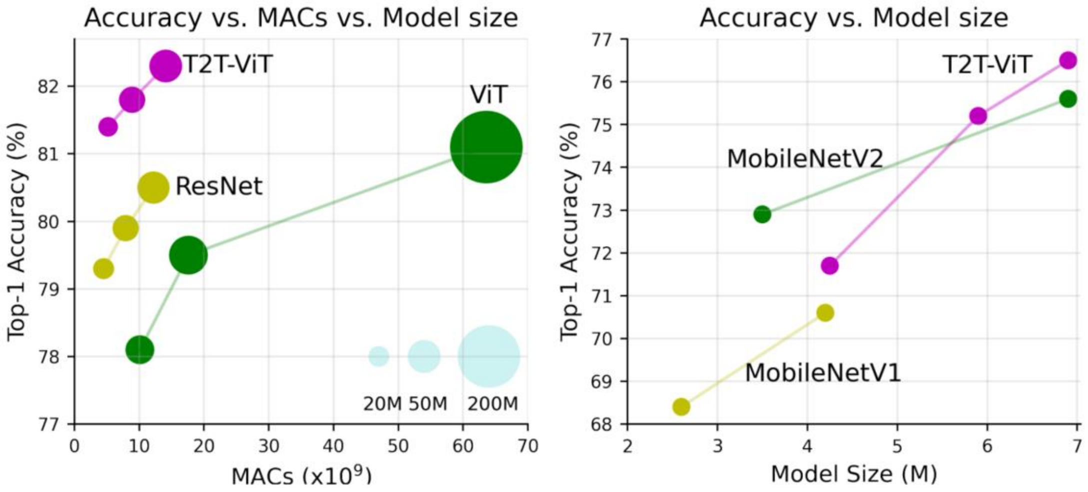
TAL TECH

# Reliability of DNNs: Recent Advancements in TalTech

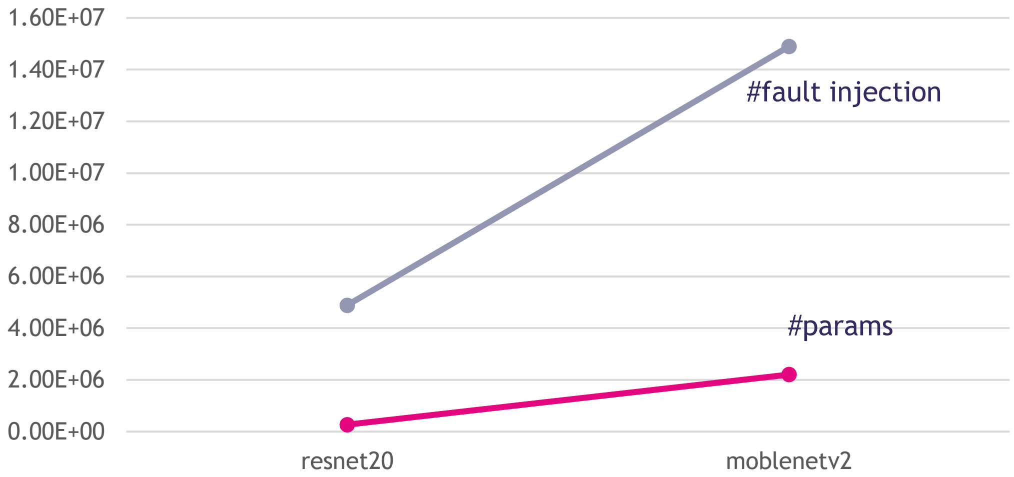
**Mohammad Hasan Ahmadilivani (Mojtaba)**
**Postdoctoral Researcher**
**Centre for Dependable Computing Systems**

TALLINN UNIVERSITY OF TECHNOLOGY

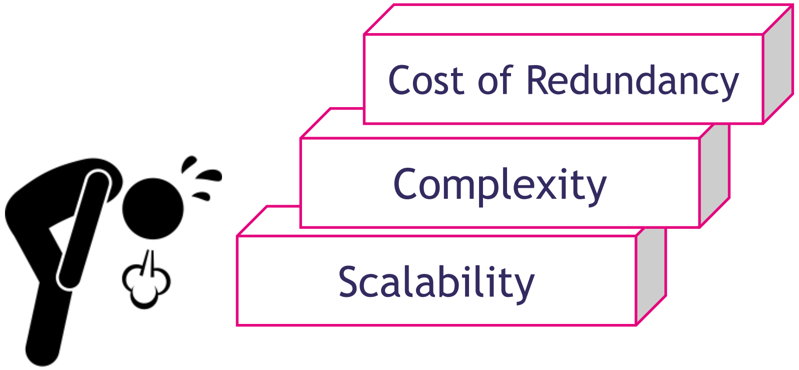
# Main Challenge




[Yuan, ICCV'21]




[Ruospo, Date'23]



# Open-Source Tools for Reliability Assessment and Enhancement of CNNs

Reliability Assessment

Reliability Enhancement

DeepVigor Framework

SentinelNN Framework

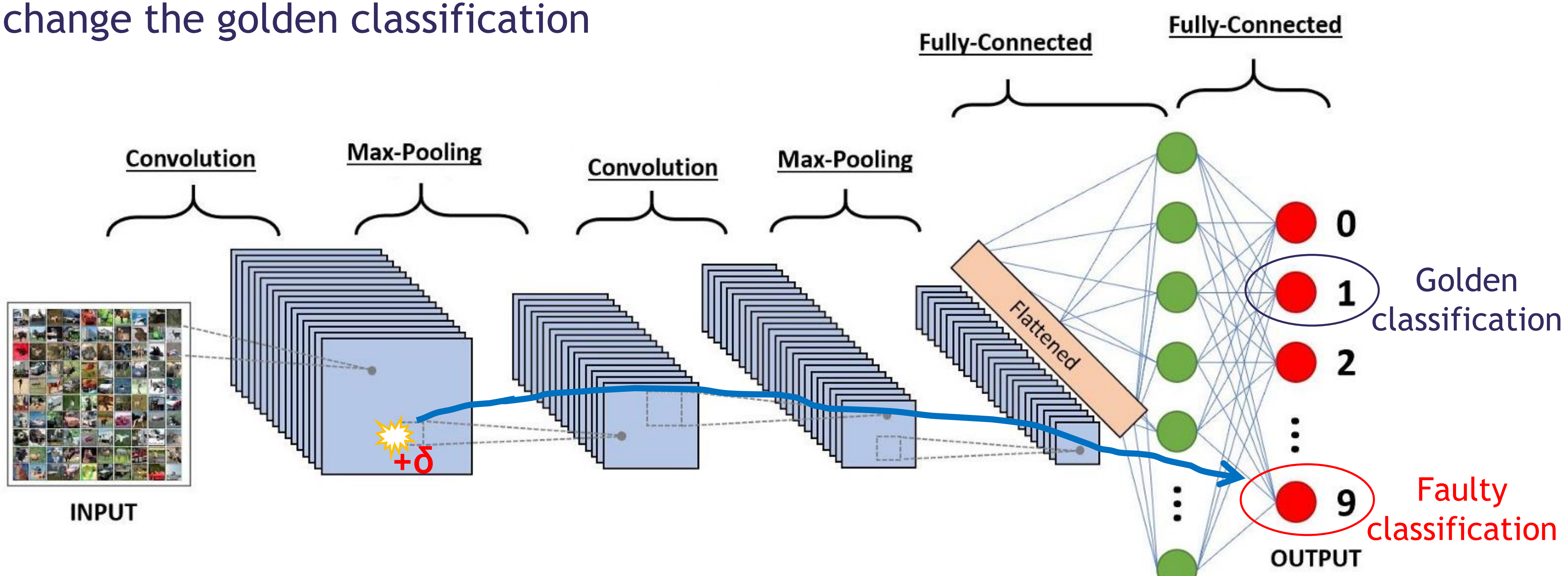
# DeepVigor+: Fault Resilience Analysis for CNNs

- A semi-analytical method for fault-resilience analysis of CNNs
  - Analysis accompanied by simulations
  - Provides metrics for parameters and activations
- Accelerator-agnostic, accurate, and metric-oriented:
  - Fine-grain Vulnerability Factor (VF) for neurons and layers
  - VF: probability of misclassification in the case of a fault occurrence
- An alternative to fault injection
- Validated extensively by fault injection

Ahmadilivani et al. *"DeepVigor: Vulnerability Value Ranges and Factors for DNNs' Reliability Assessment"*, *IEEE ETS, Italy, 2023.*

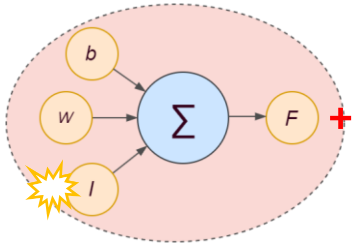
# DeepVigor+: Fault Resilience Analysis for CNNs

- The key idea:
  - Identifying the values (δ) for each neuron that change the golden classification




Ahmadilivani et al. *“DeepVigor: Vulnerability Value Ranges and Factors for DNNs' Reliability Assessment”*, *IEEE ETS, Italy, 2023.*

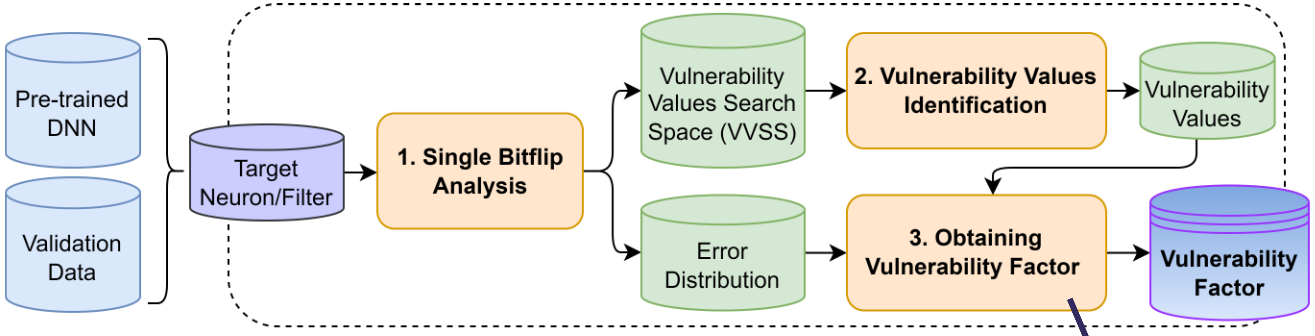
# DeepVigor+: Fast and Scalable Fault Resilience Analysis for CNNs



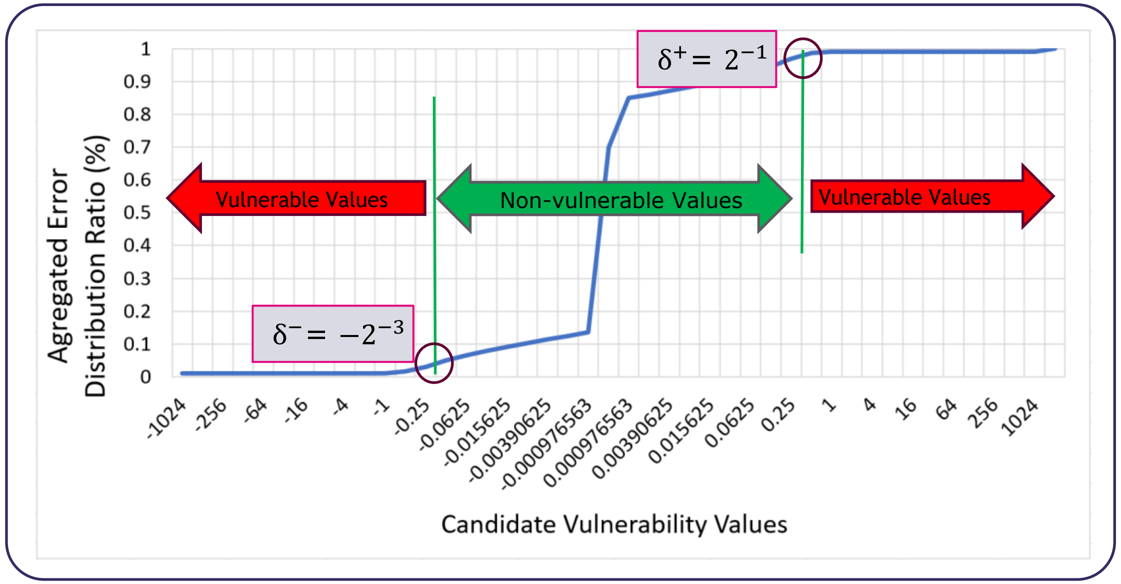

$+\delta \approx \pm 2^{\rho}; \rho \in \{\pm 1, \pm 2, \pm 3, \ldots\}$

Values with non-zero error distribution

Vulnerability values search space =
$\{-1024, -1, -0.5, \ldots, 0, +0.5, 1, 2, 1024\}$




$$VF_{target\_neuron} = 1 - (distribution\_ratio_{\delta+} - distribution\_ratio_{\delta-})$$




Data type

32-bit floating point

Ahmadilivani et al. *"DeepVigor+: Scalable and Accurate Semi-Analytical Fault Resilience Analysis for Deep Neural Network", arXiv preprint arXiv:2410.15742, 2024, under review.*

# DeepVigor+: Fast and Scalable Fault Resilience Analysis for CNNs

- Sampling is conducted:
  - Channel sampling ratio
  - #neurons: log(#neurons in output channel)

Channel sampling ratio=10% ensures 1% total error

- Achieves Model VF (MVF):
  - Up to 26.9 times fewer simulations than best statistical FI
- Derives MVF in a few minutes (GPU NVIDIA A100)
  - ~46 minutes for ResNet34 on ImageNet

Comparing number of simulations

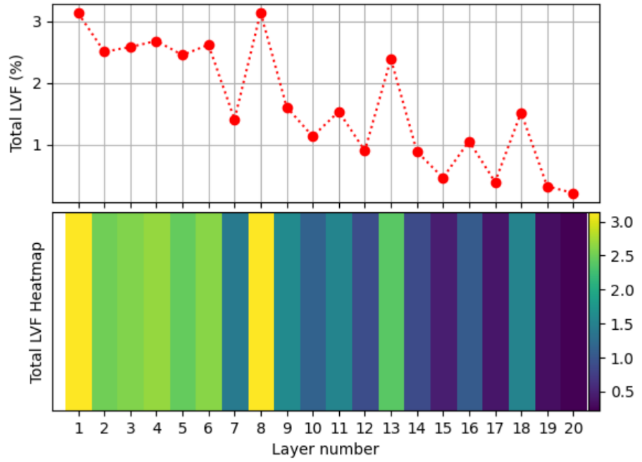
| CNN-dataset | Best statistical FI | DeepVigor+ | Improvement |
|---|---|---|---|
| VGG11-CIFAR10 | 138,107 | 5,634 | **24.5x** |
| VGG16-CIFAR100 | 218,664 | 11,759 | **18.6x** |
| ResNet18-CIFAR100 | 346,070 | 17,702 | **19.5x** |
| MobileNetV2-CIFAR100 | 712,264 | 26,441 | **26.9x** |
| ResNet18-ImageNet | 359,772 | 23,858 | **15.1x** |
| ResNet34-ImageNet | 653,867 | 43,005 | **15.2x** |

TAL TECH

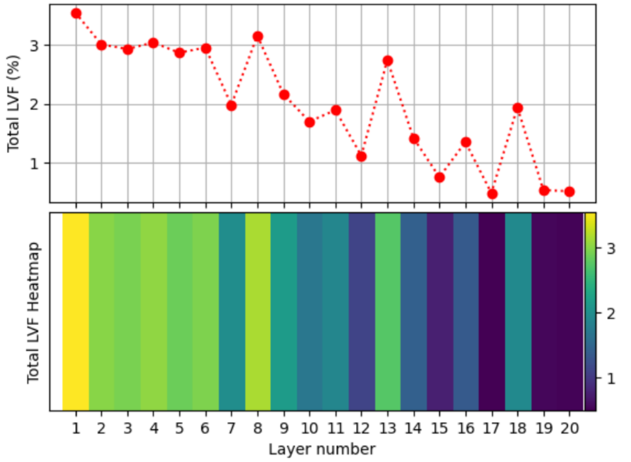
Ahmadilivani et al. *"DeepVigor+: Scalable and Accurate Semi-Analytical Fault Resilience Analysis for Deep Neural Network", arXiv preprint arXiv:2410.15742, 2024, under review.*

# DeepVigor+: Fast and Scalable Fault Resilience Analysis for CNNs

- Comparative analysis for CNNs:

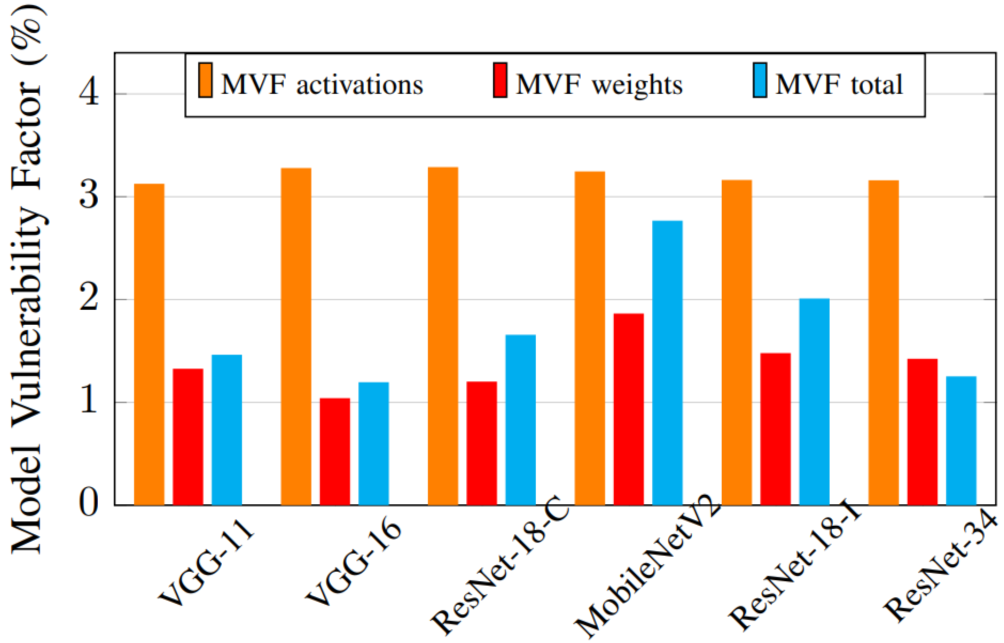



(a) ResNet-18-C




(b) ResNet-18-I

Layer-wise resilience analysis comparison




Model-wise resilience analysis comparison

Open source: https://github.com/mhahmadilivany/DeepVigor

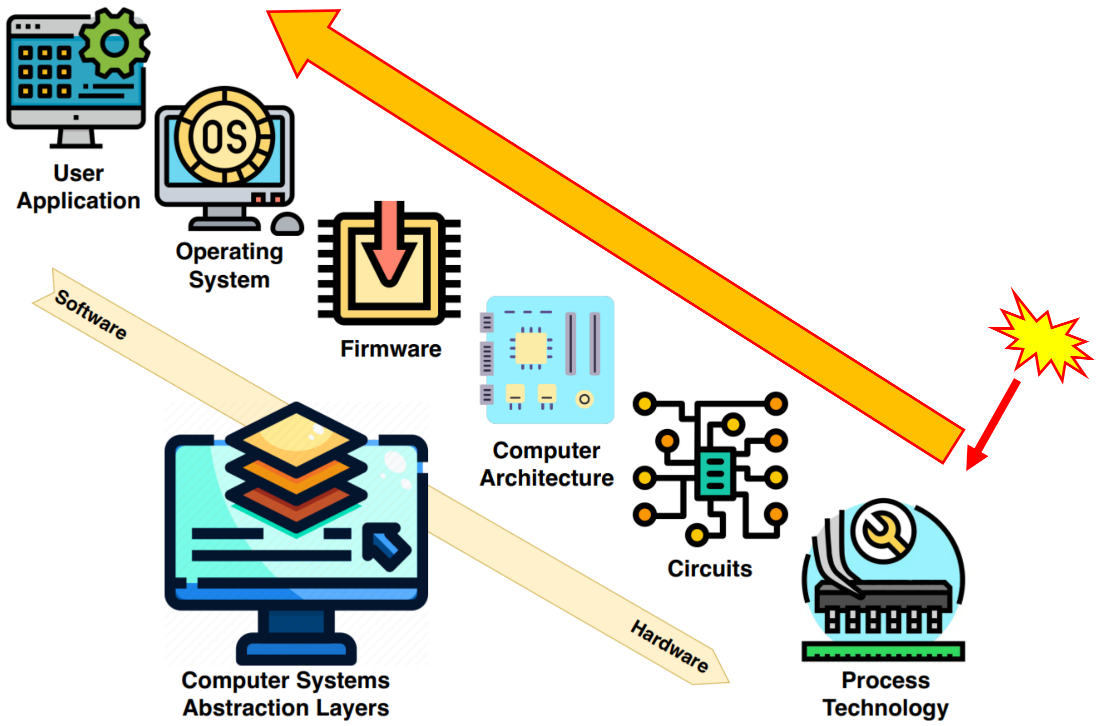
PyTorch

TAL TECH

Ahmadilivani et al. *"DeepVigor+: Scalable and Accurate Semi-Analytical Fault Resilience Analysis for Deep Neural Network"*, *arXiv preprint arXiv:2410.15742, 2024, under review.*

# Model-Level Fault Tolerance for CNNs

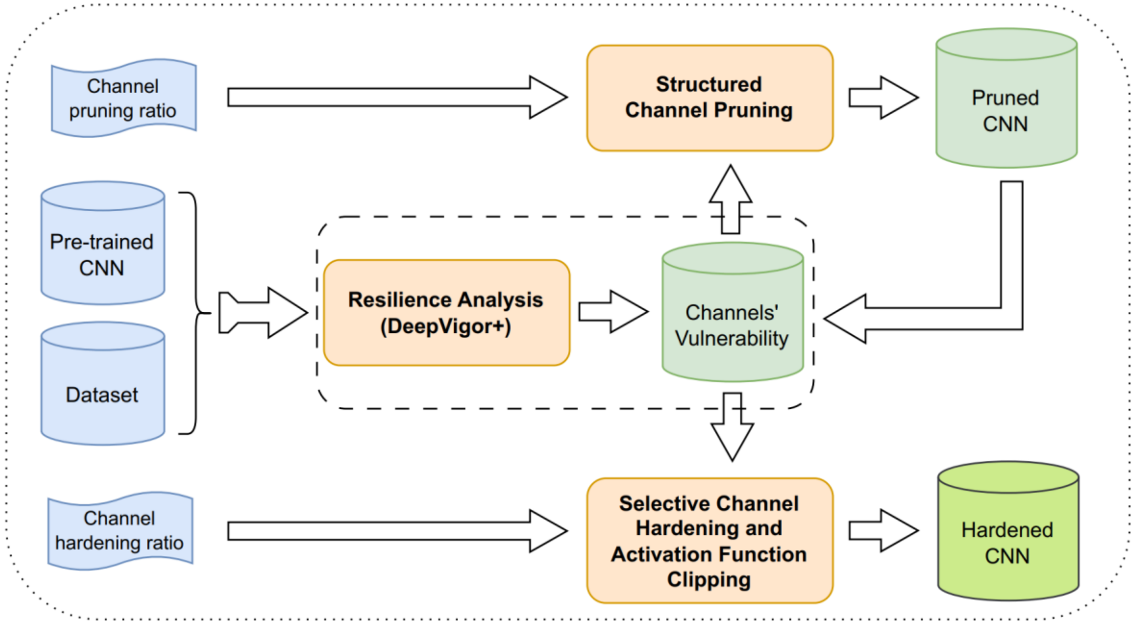
- Faults occur at hardware, propagating up to the application layer
- Faults propagation can be prevented at any system layer
- Software Implemented Hardware Fault Tolerance:
  - Conduct error detection and correction at software layer
  - Cost-effective alternative for pre-designed accelerators cores and commercial-off-the-shelf processors
  - The structure of CNNs are modified to achieve inherent fault tolerance




TALTECH

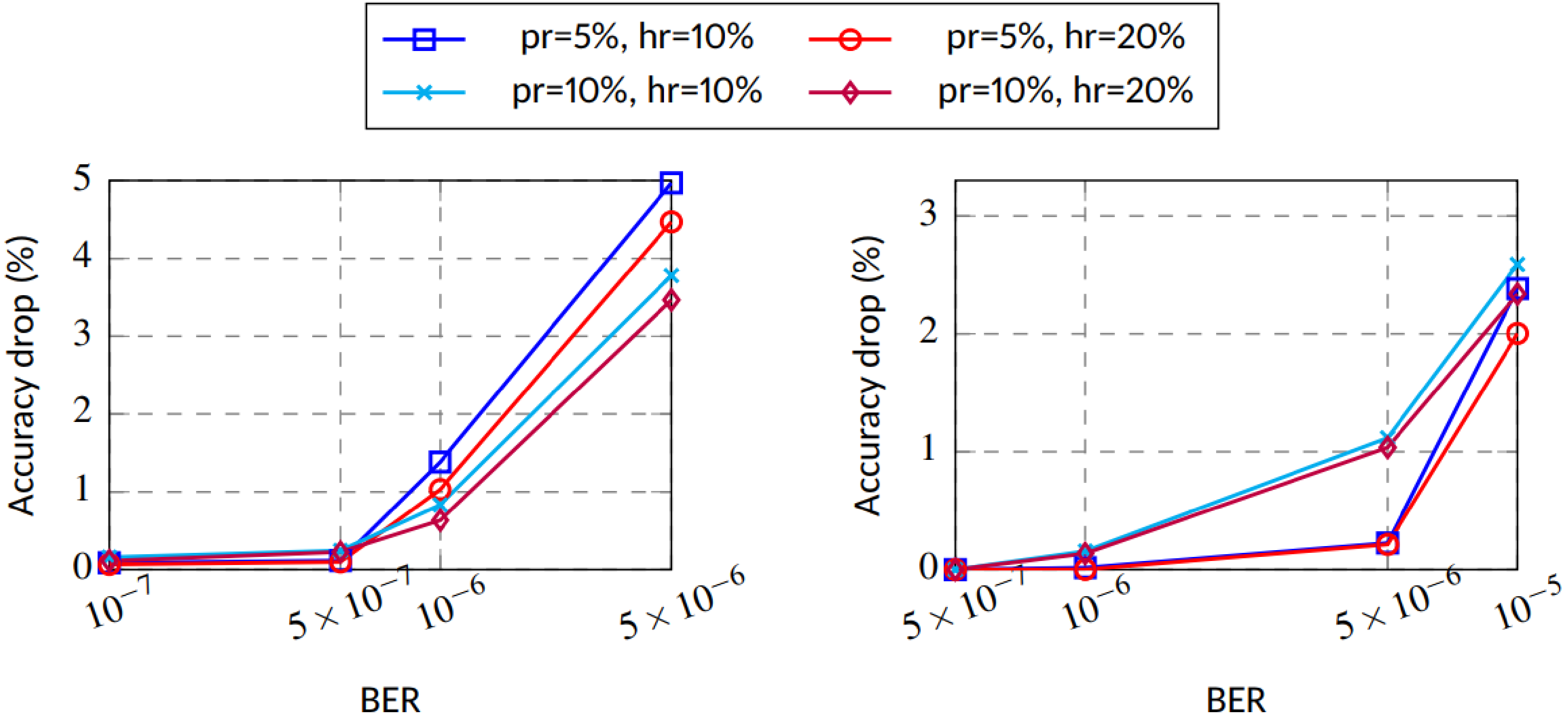
# SentinelNN: CNN Model Analysis and Hardening

- Cost-Effective Fault Tolerance for CNNs
- An open-source framework to produce hardened CNNs, integrating:
  - DeepVigor+ for resilience analysis
  - Structured channel pruning
  - Selective channel duplication and correction
  - Range restriction




Open source: https://github.com/mhahmadilivany/SentinelNN PyTorch

# SentinelNN: CNN Model Analysis and Hardening

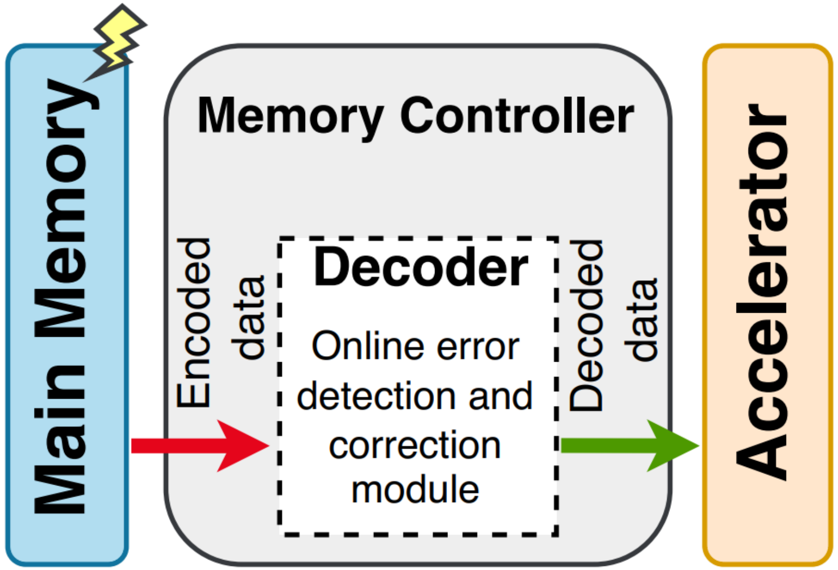



(a) VGG-19 (b) ResNet-20

Ahmadilivani et al. “Cost-Effective Fault Tolerance for CNNs Using Paameter Vulnerability Based Hardening and Pruning”, IOLTS, France, 2024.

# ECC Alternatives for Vision Transformers (ViTs)

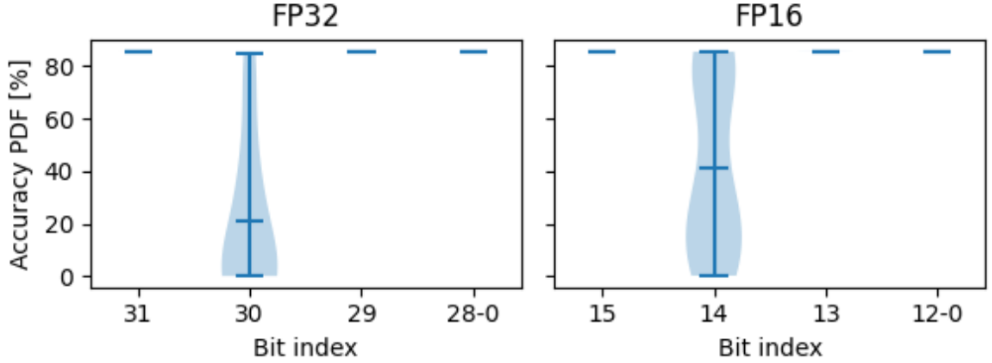
- ViTs are recent models that possess millions to billions of parameters
- Error correction Codes (ECCs) are too costly
  - They require extra memory to store the parity bits
  - For ViT-base model, ~350MB extra memory is needed
- We propose an effective and efficient alternative to ECC that significantly improve the reliability of ViTs



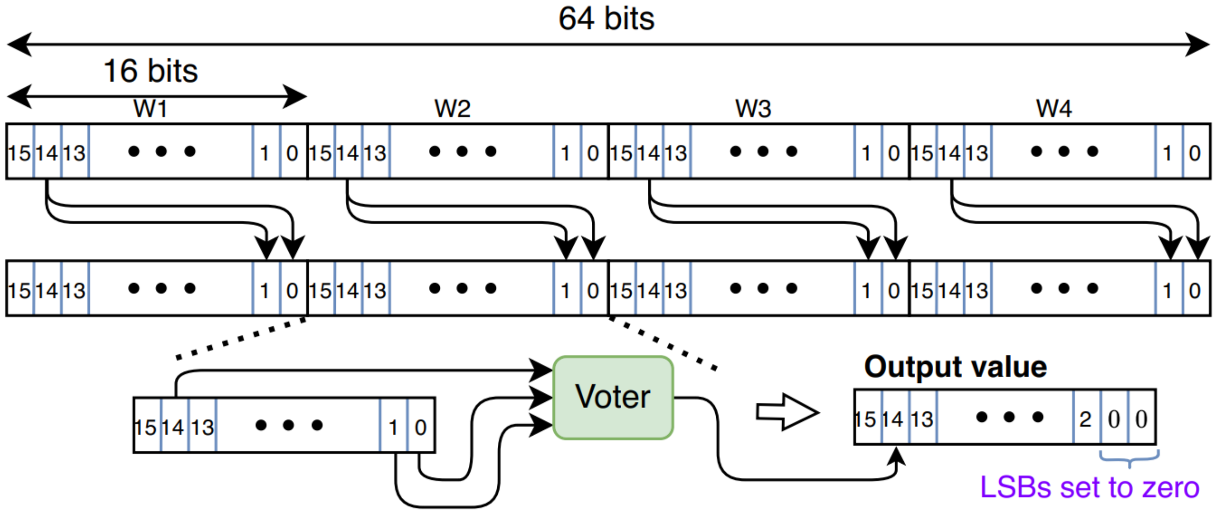

Ahmadilivani et al. “Uncovering the Limits of ECCs in Vision Transformers and a Zero-Cost Reliability Enhancement”, DATE, 2026.

# ECC Alternatives for Vision Transformers

- Highest exponent in floating-point data type is the most vulnerable one




- MSET method
  - Most significant exponent bit is triplicated and stored in LSBs
  - The triplicated bits are corrected by a voter

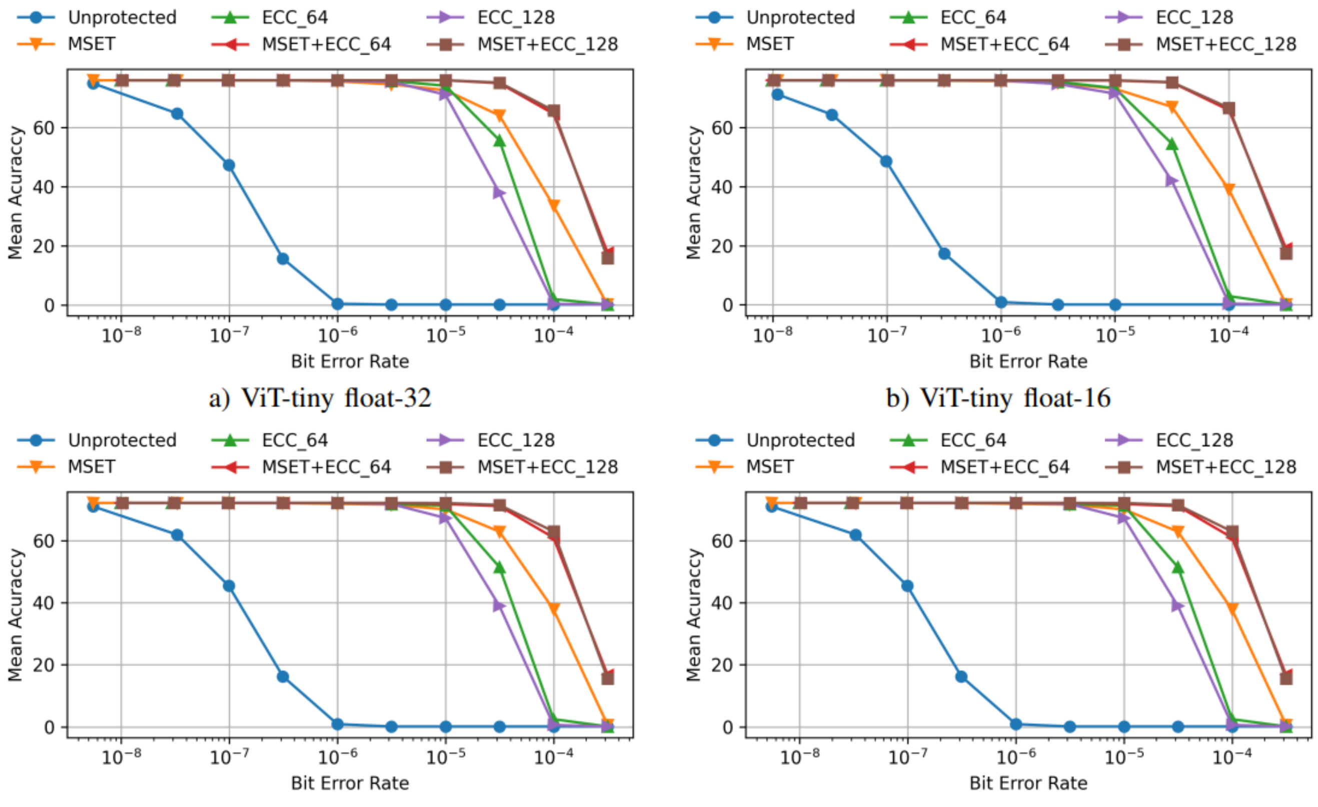



TAL TECH

Ahmadilivani et al. "Uncovering the Limits of ECCs in Vision Transformers and a Zero-Cost Reliability Enhancement", DATE, 2026.

# ECC Alternatives for Vision Transformers

- Results under FI into parameters




a) ViT-tiny float-32

b) ViT-tiny float-16

c) DeiT-tiny float-32

d) DeiT-tiny float-16

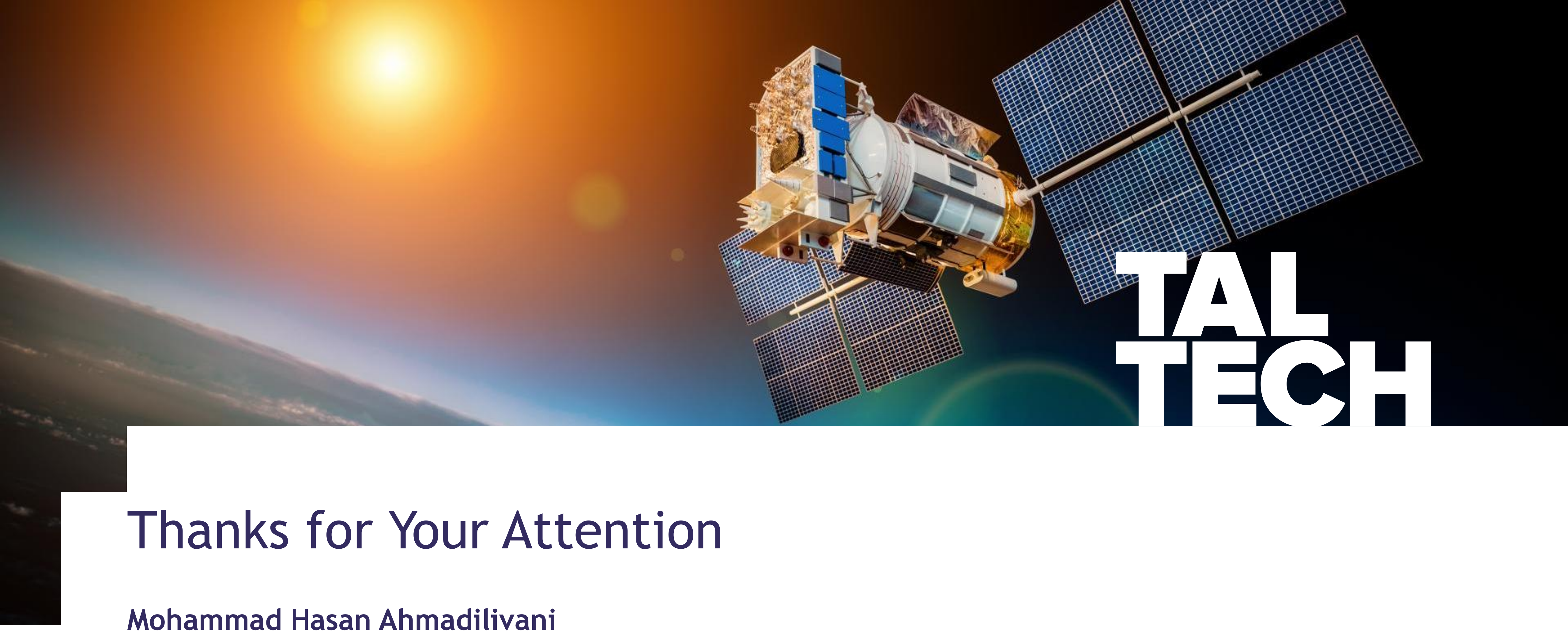
TAL TECH

Ahmadilivani et al. “Uncovering the Limits of ECCs in Vision Transformers and a Zero-Cost Reliability Enhancement”, DATE, 2026.

# Conclusions

- Methods to analyze and improve reliability for various DNNs
  - DeepVigor+
  - SentinelNN
  - MSET
- Our open-source tools:
  - DeepVigor: https://github.com/mhahmadilivany/DeepVigor
  - SentinelNN: https://github.com/mhahmadilivany/SentinelNN
  - FaultForge: https://github.com/rezzubs/faultforge
  - RReLU: https://github.com/hamidmousavi0/reliable-relu-toolbox
  - ML-FRAMES: https://github.com/mhahmadilivany/ML-FRAMES
  - FT-Sparse: https://github.com/mhahmadilivany/FT-Sparse
  - Reliability and security: https://github.com/mhahmadilivany/bitflip-attack-on-fault-tolerant-DNNs

TAL TECH



# Thanks for Your Attention

**Mohammad** Hasan **Ahmadilivani**
Tallinn University of Technology

Email: mohammad.ahmadilivani@taltech.ee
GitHub: github.com/mhahmadilivany