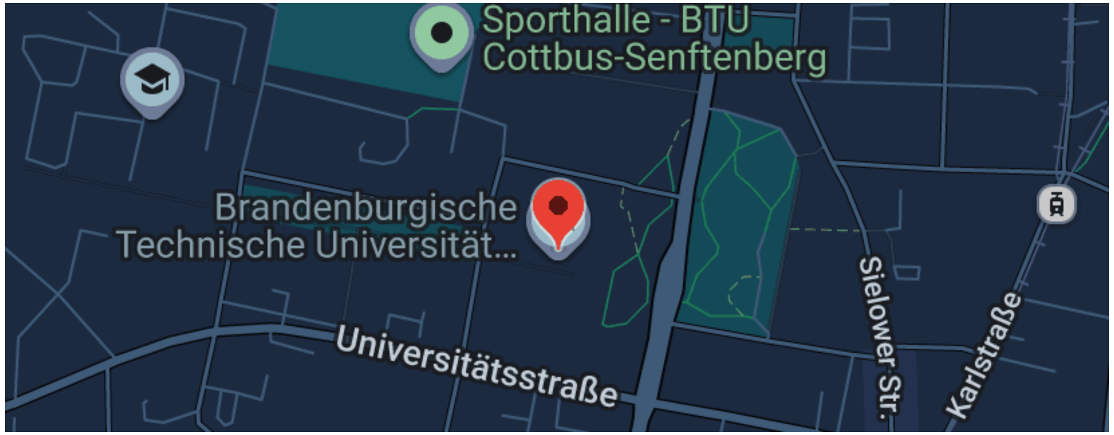
Reliability Vs. Security .. Friends or Foes?

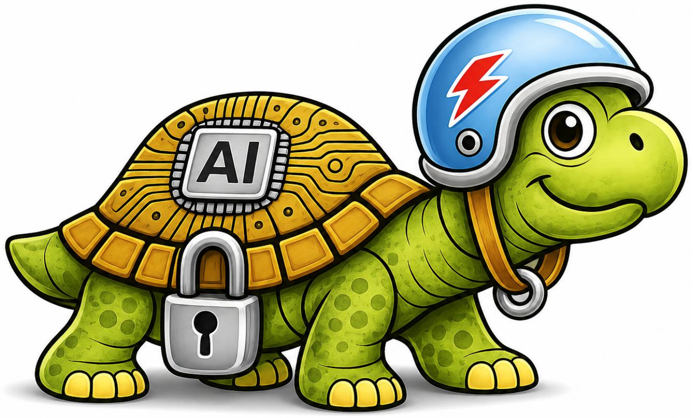
Mahdi Taheri
BTU Cottbus-Senftenberg
Tallinn University of Technology

13.07.2026

RSS Team and Location

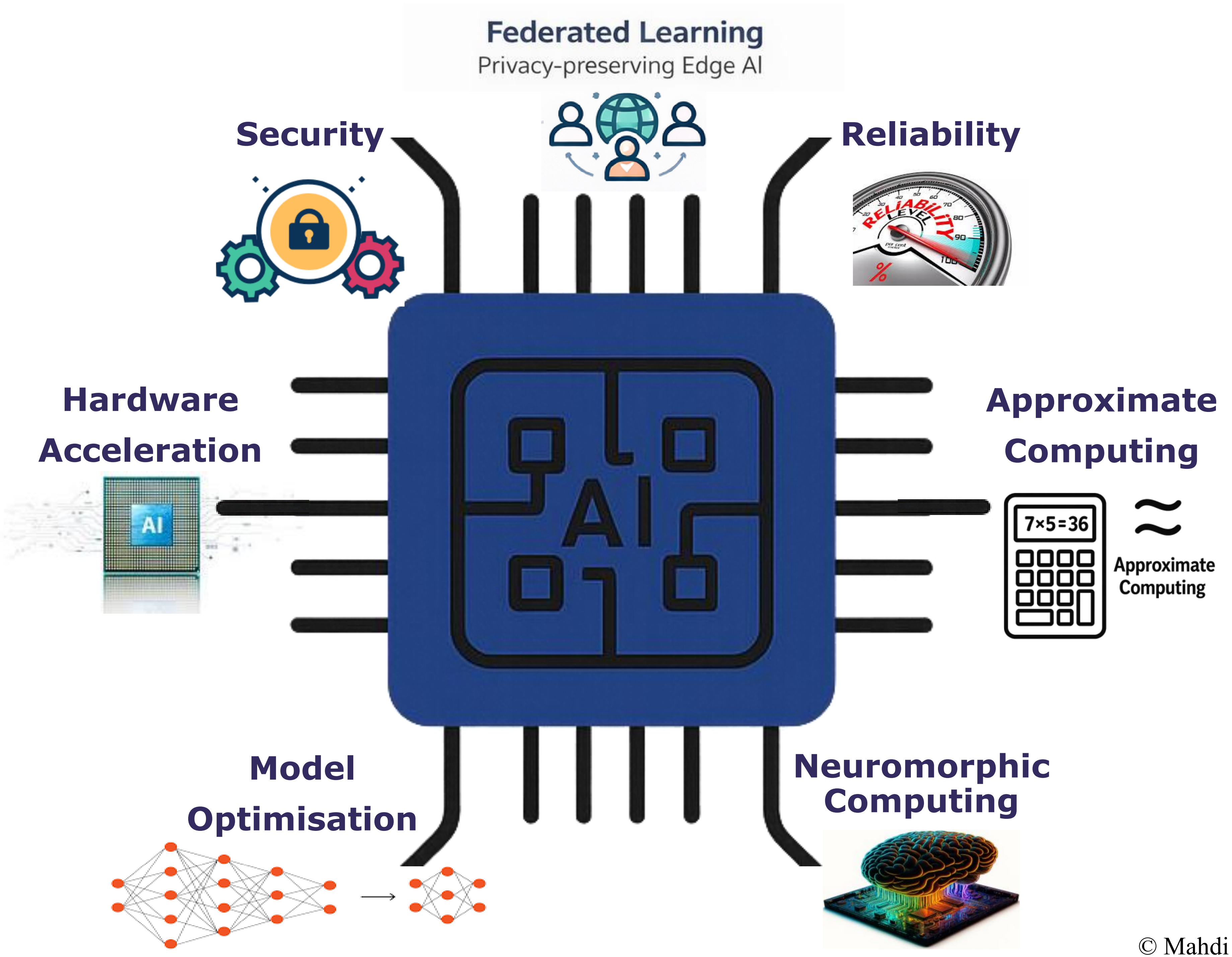
Location





Core Team

Topics

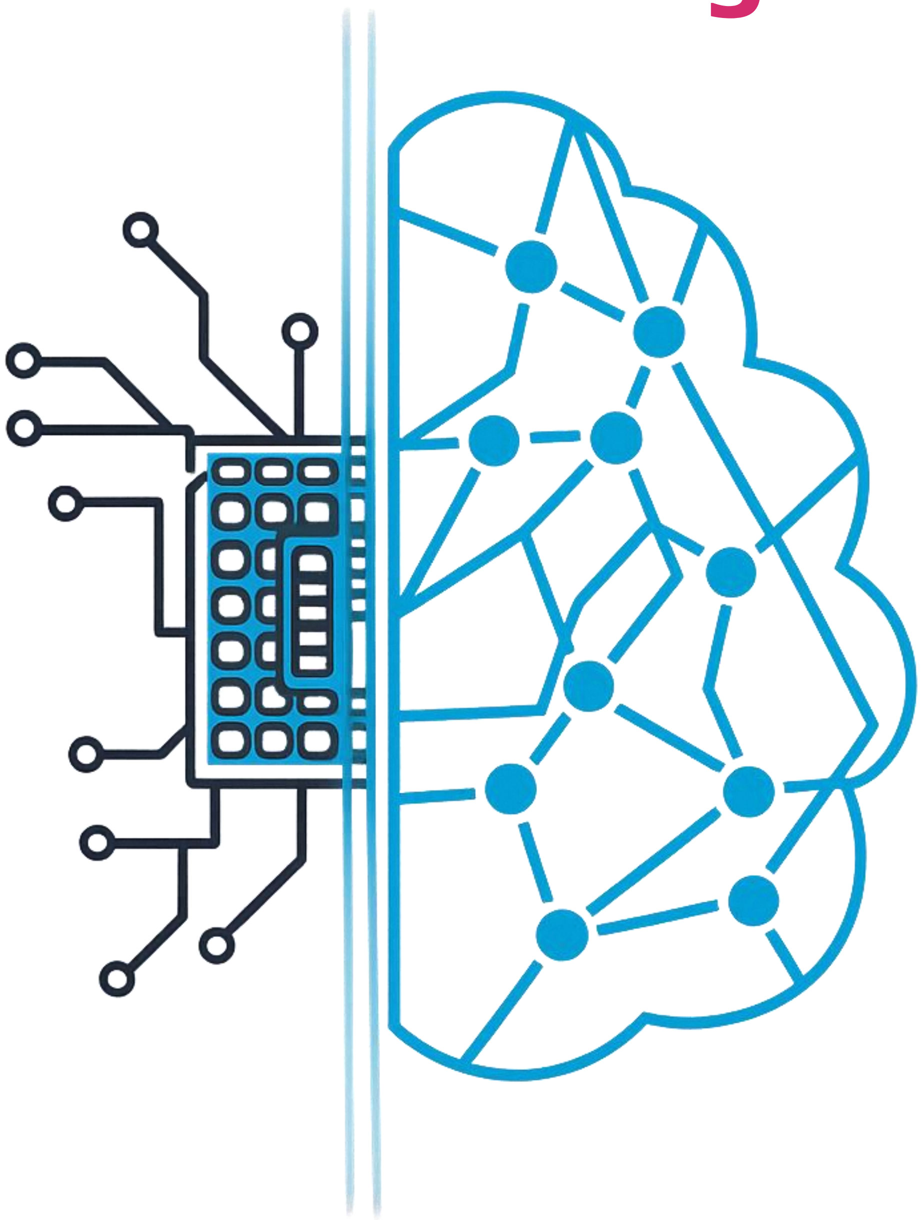

Federated Learning
Privacy-preserving Edge AI
Security
Reliability
Hardware Acceleration
AI
Approximate Computing
7×5=36
≈
Approximate Computing
Model Optimisation
Neuromorphic Computing

Projects

- **TIRAMISU** (Training and Innovation in Reliable and Efficient Chip Design for Edge AI) – MSCA-DN
- **TAICHIP** (Boosting TalTech Capacity in Reliable and Efficient AI-Chip Design) - EU Twinning Action
- **AI-DISCO** (Edge/Cloud AI for Distributed Sensing & Computing) - BMFTR

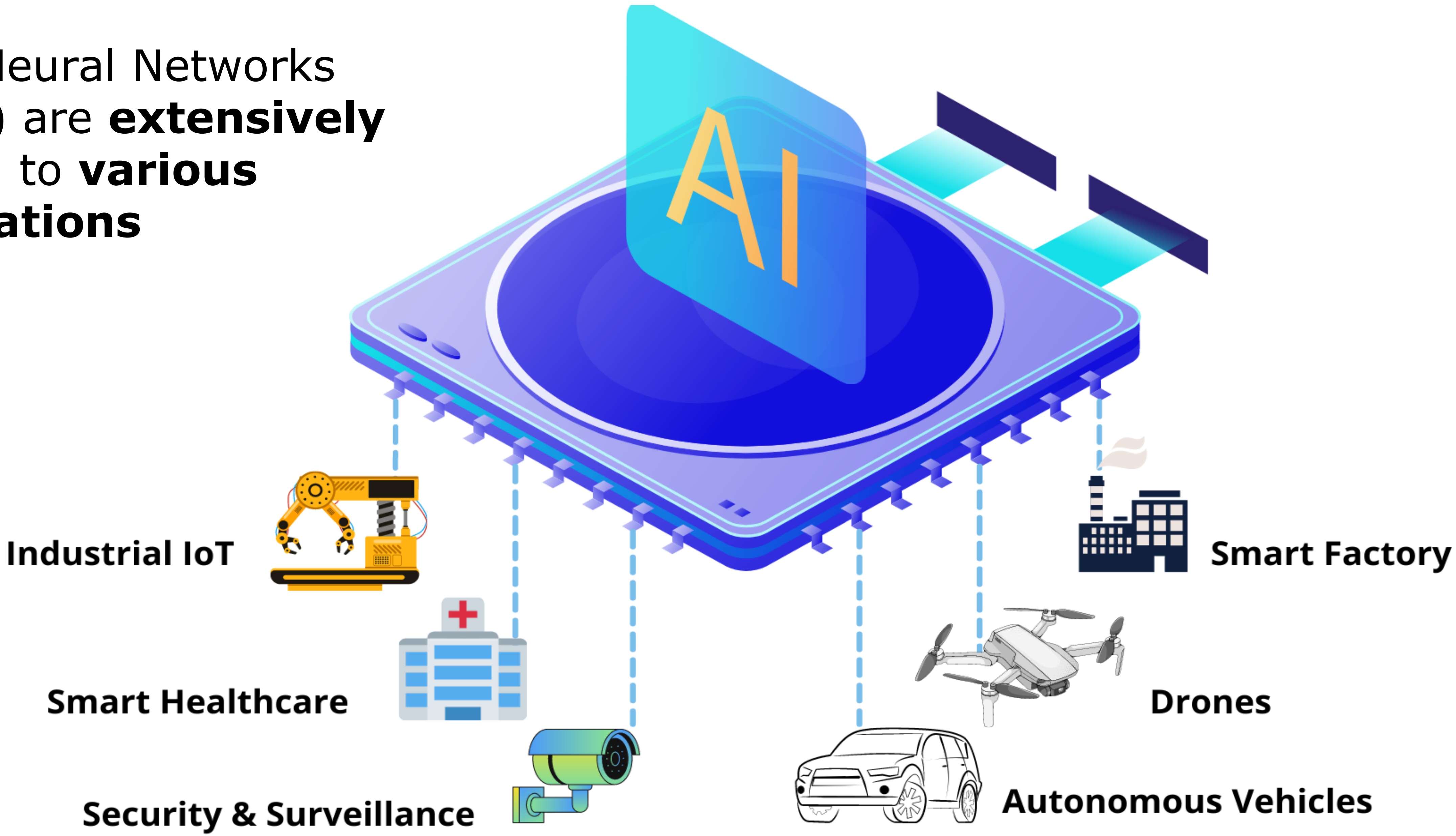
Agenda

- ❑Motivation
- ❑Proposed Methodologies
- ❑Conclusion
- ❑Q&A



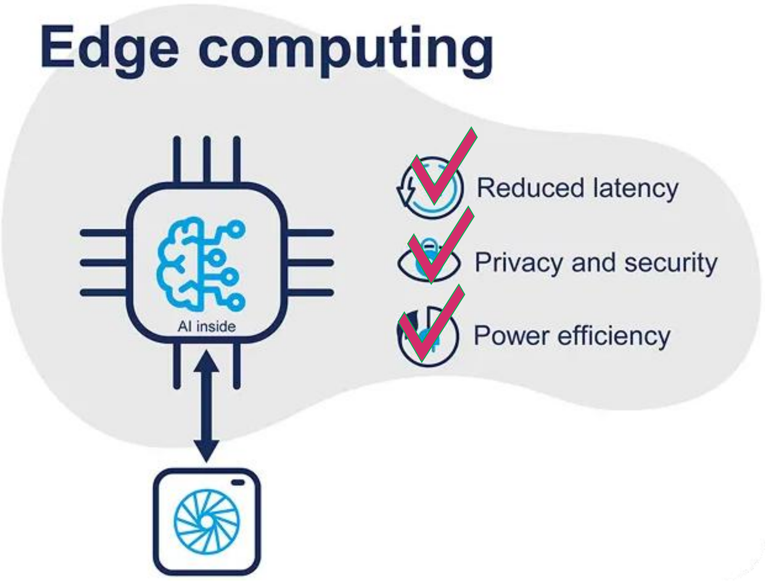
Motivation

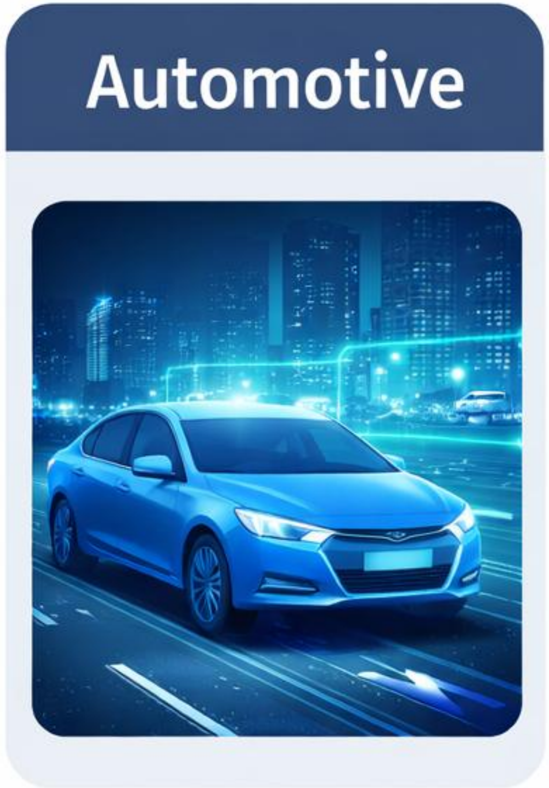
❑ Deep Neural Networks (DNNs) are **extensively** applied to **various applications**



Motivation

- AI inference is increasingly moving to the edge



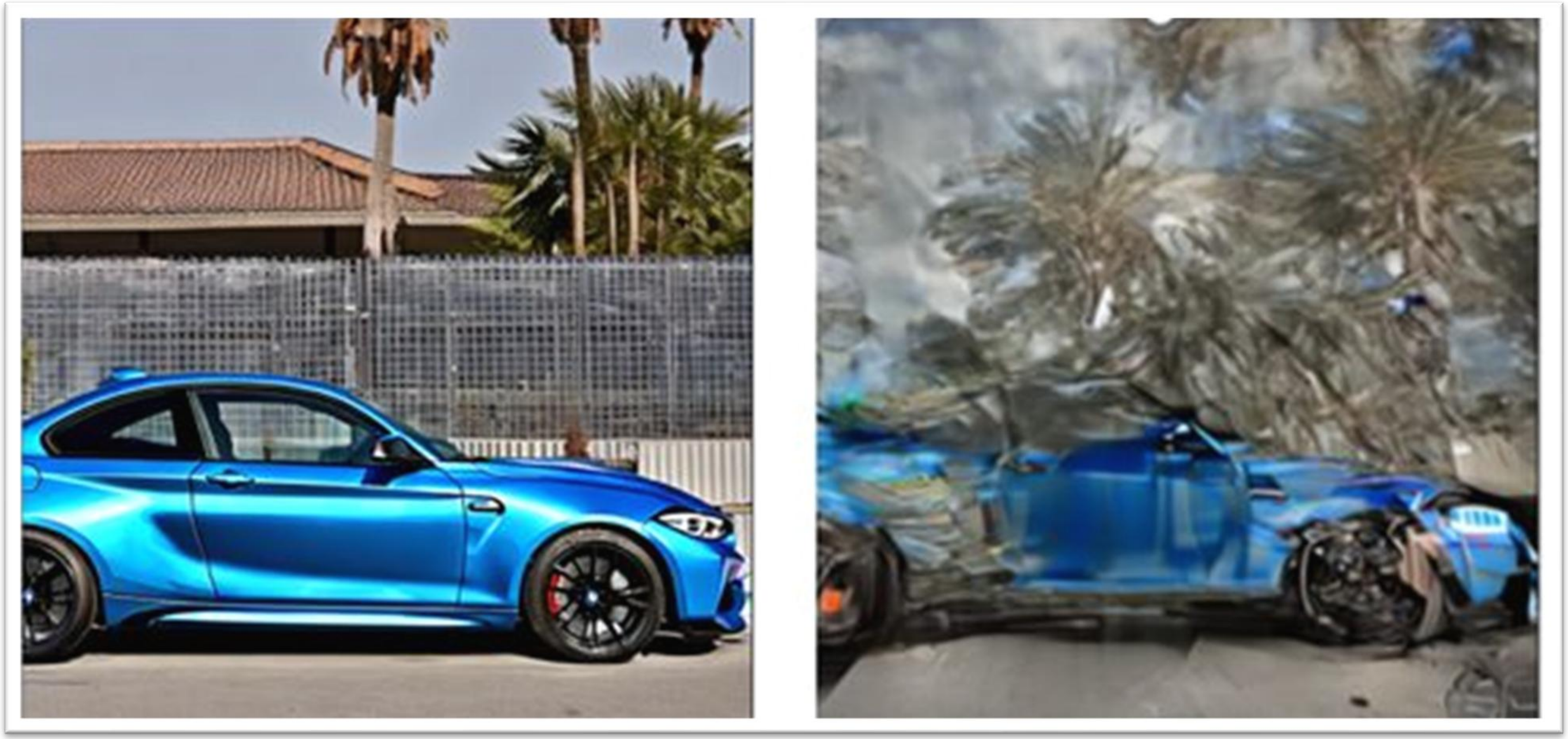


Safety- and mission-critical applications require extra attention

CEVA Inc., “2025 Edge AI Technology Report,” CEVA, 2024. [Online]

Motivation

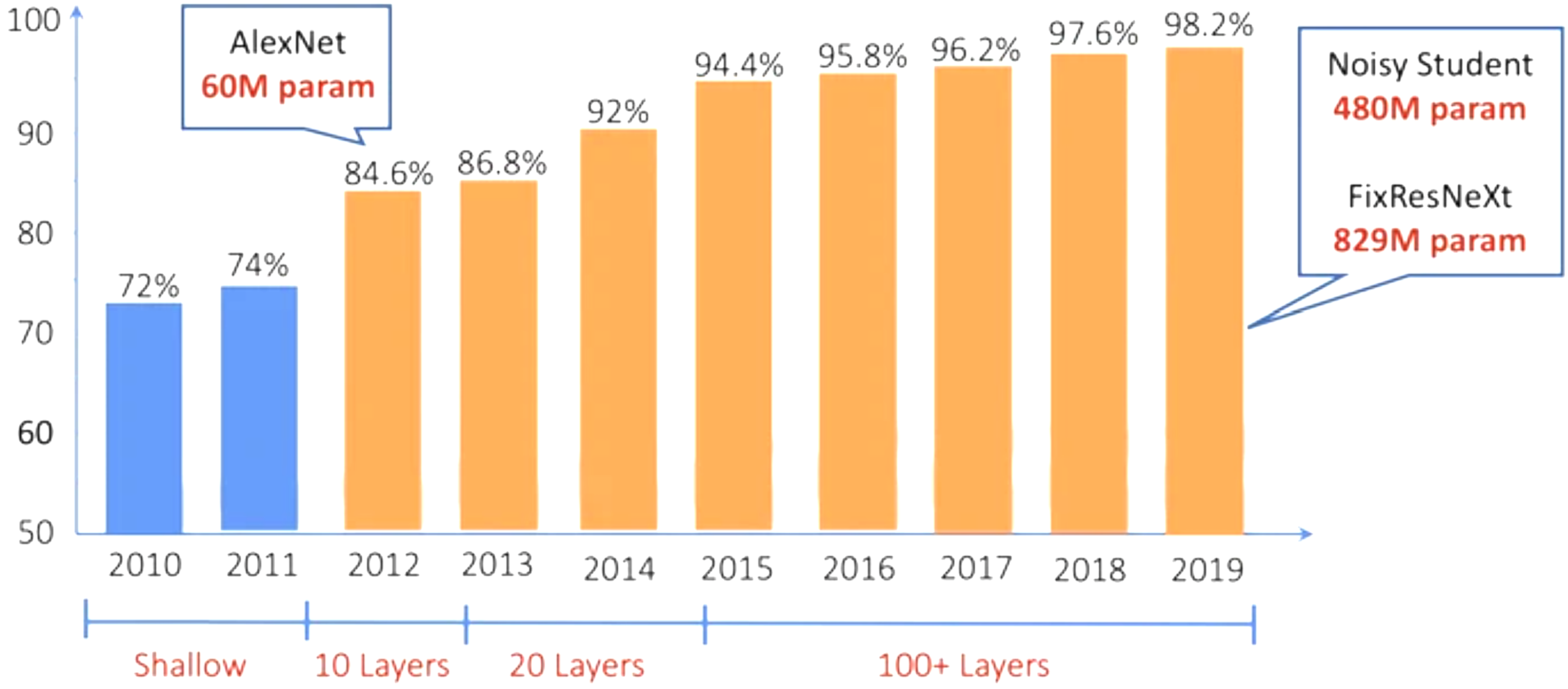
❑DNNs in safety, and mission critical applications raise **reliability** and **Security** concerns



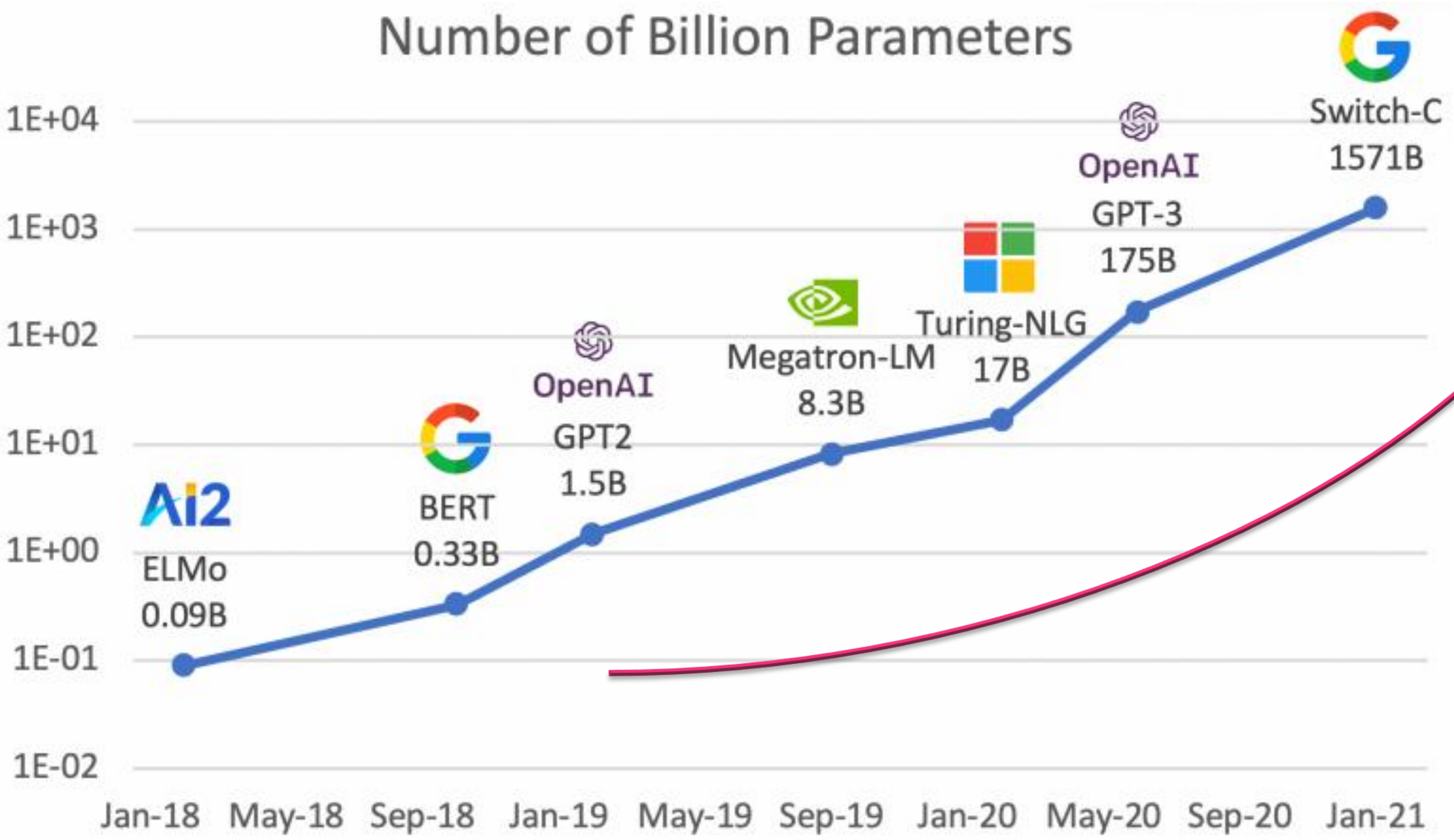
Z.Gao, L.Yuan, P.Reviriego, S.Liu, F.Lombardi, *"Dependability Evaluation of Stable Diffusion with Soft Errors on the Model Parameters"*, arXiv.org, March 2024

- generated with the same prompt, same random seed, but…
- **flipping a single bit** among billions totally changes the output image

Extra Percent …Extra Parameters!



The Parameter Inflation Era!

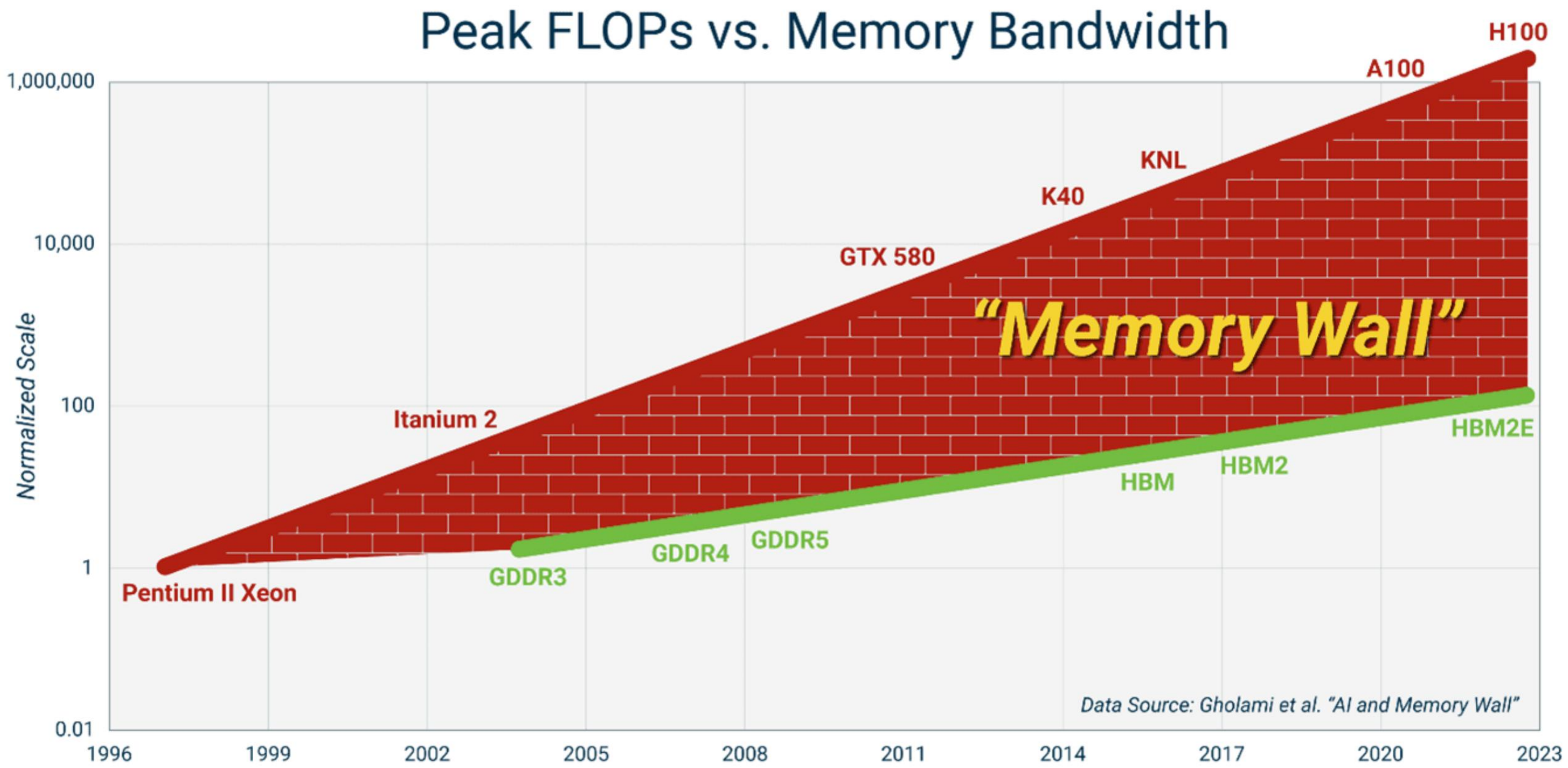


~ Trillion

NOWADAYS

AI Runs Fast… Memory Walks

3–5 million×

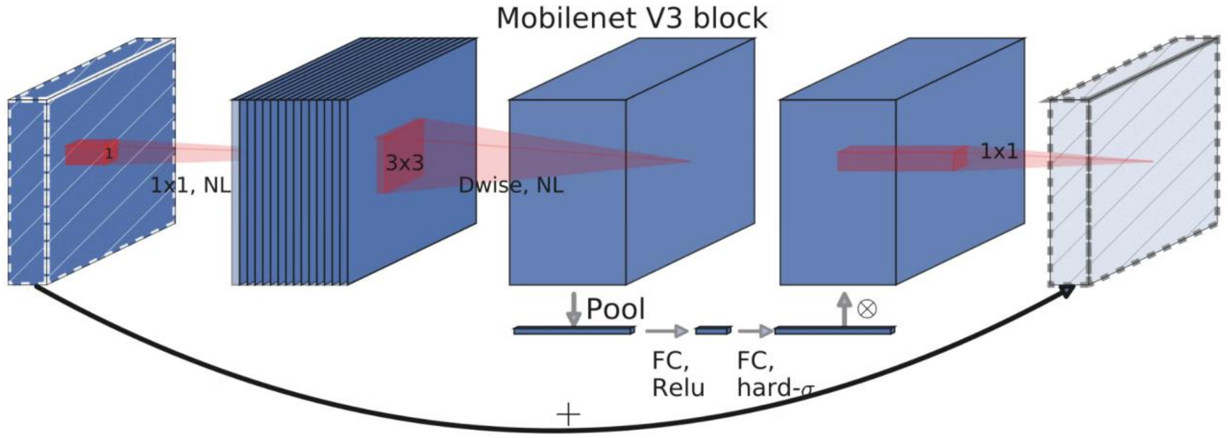


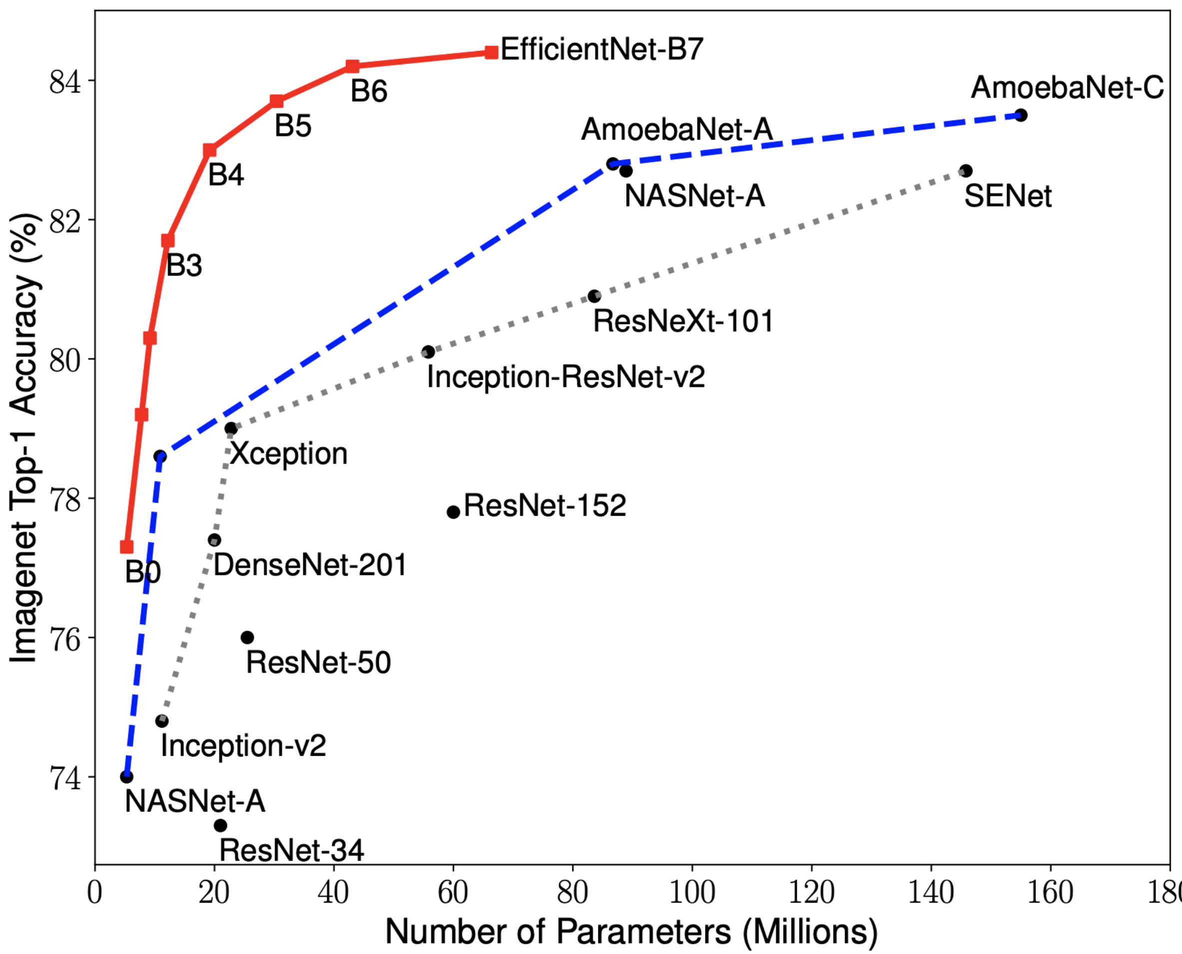
Rubin / HBM4-class

2026

Model compression and Acceleration

- Knowledge Distillation
- Neural Architectural Search
- Pruning
- Quantization





Optimisation and dual- impact on resilient

❑Quantization

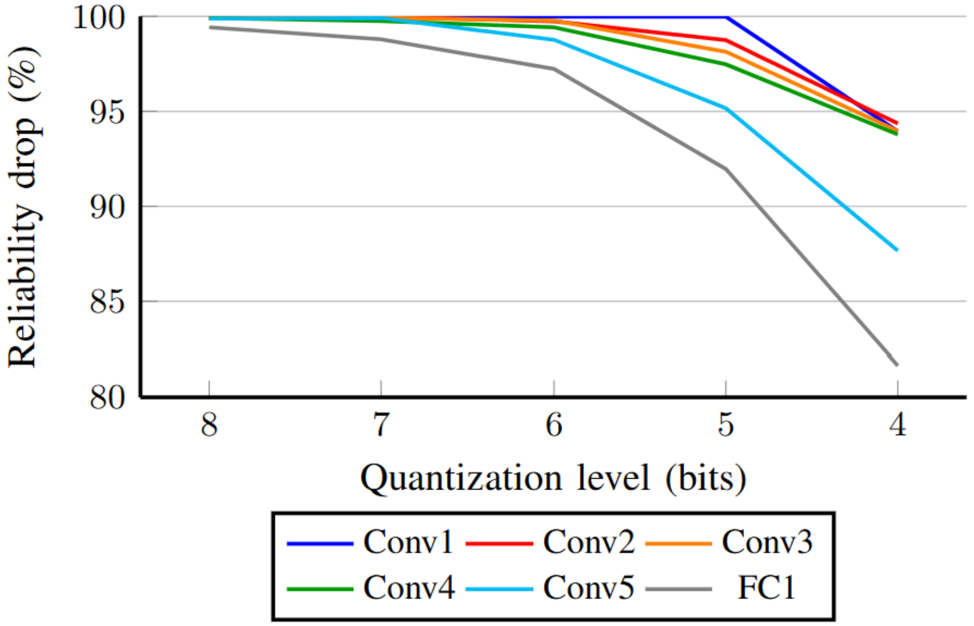
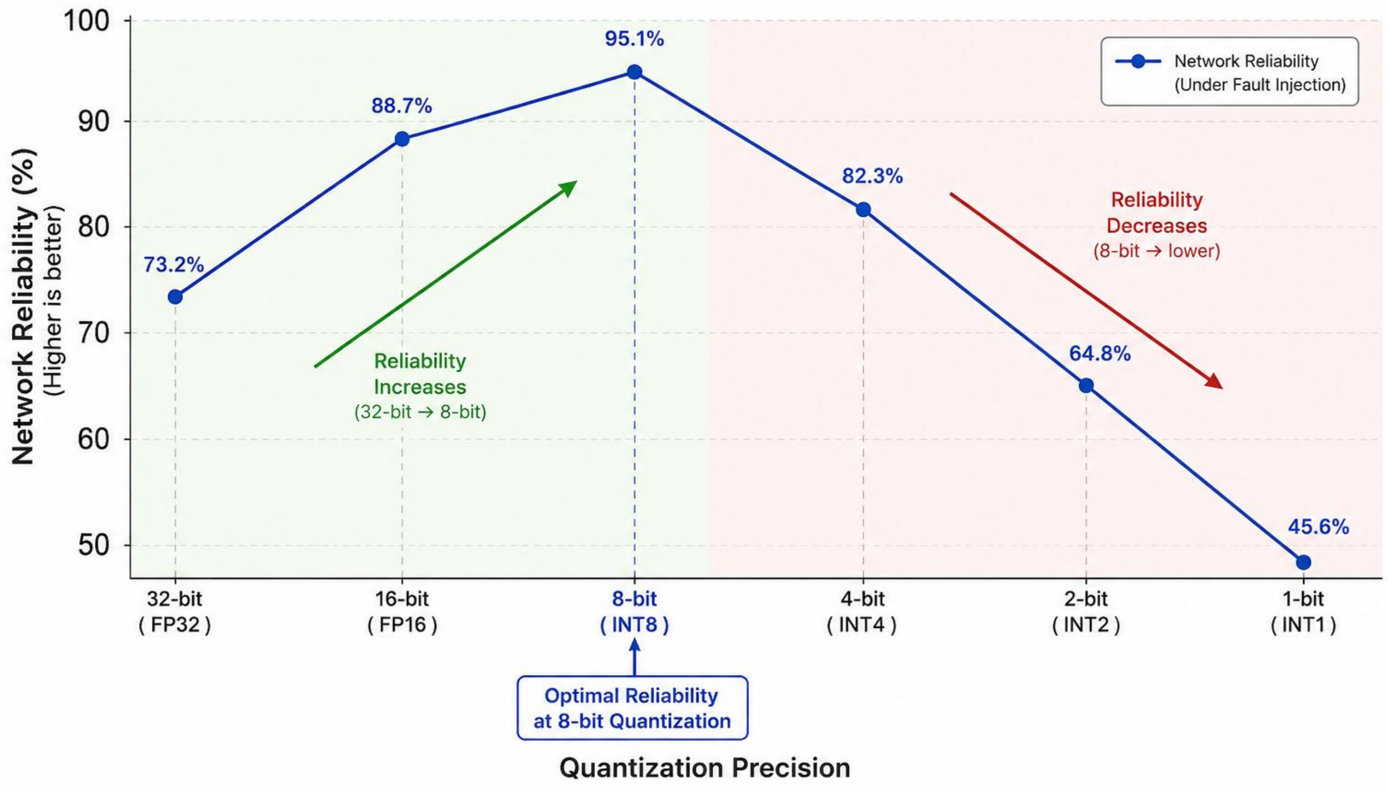


Fig. 5: AlexNet layer-level reports of reliability drop (%) based on different quantization levels (unprotected design)

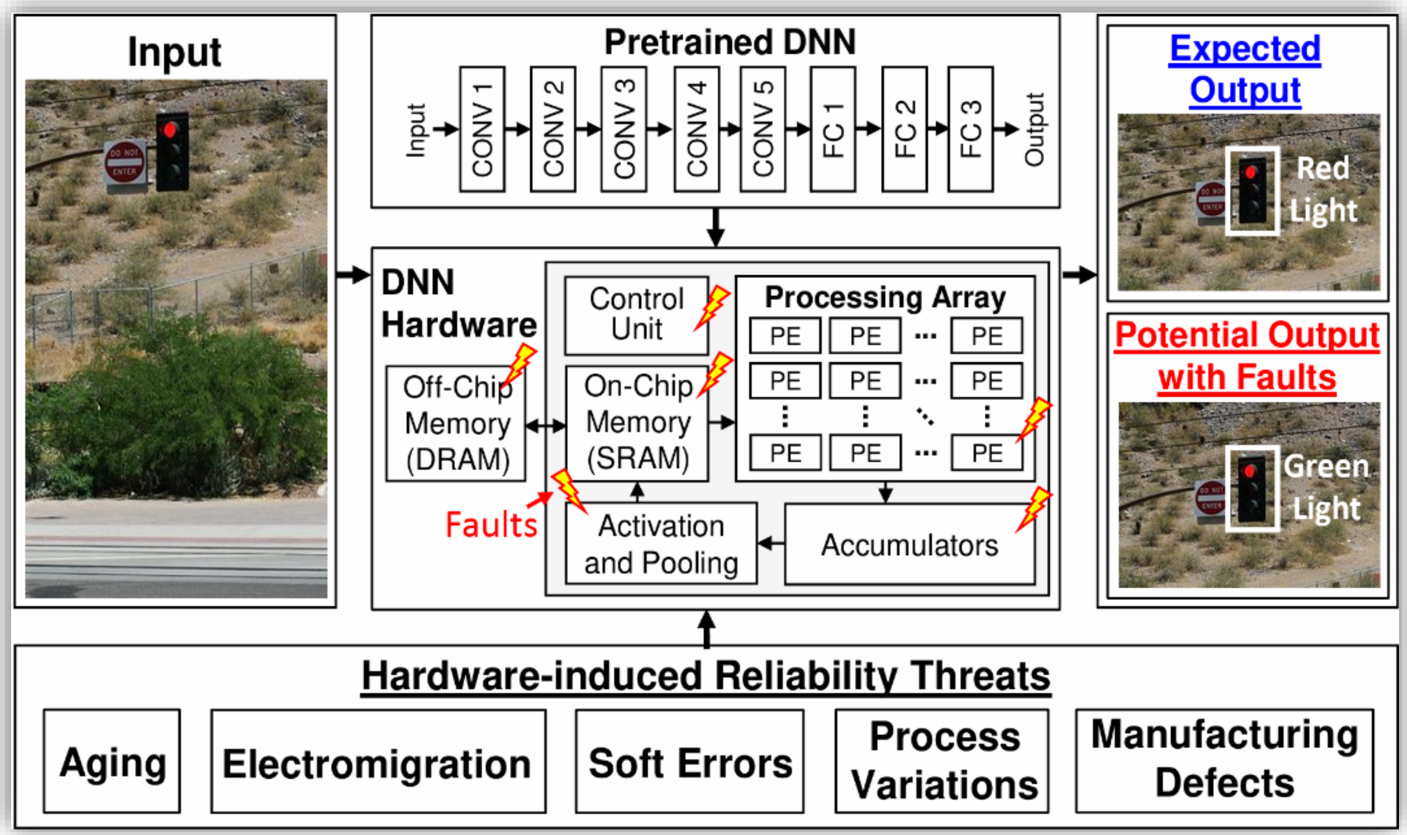
For Visualisation. Not real values!



Taheri, et .al., "Exploration of Activation Fault Reliability," ISQED, 2025.

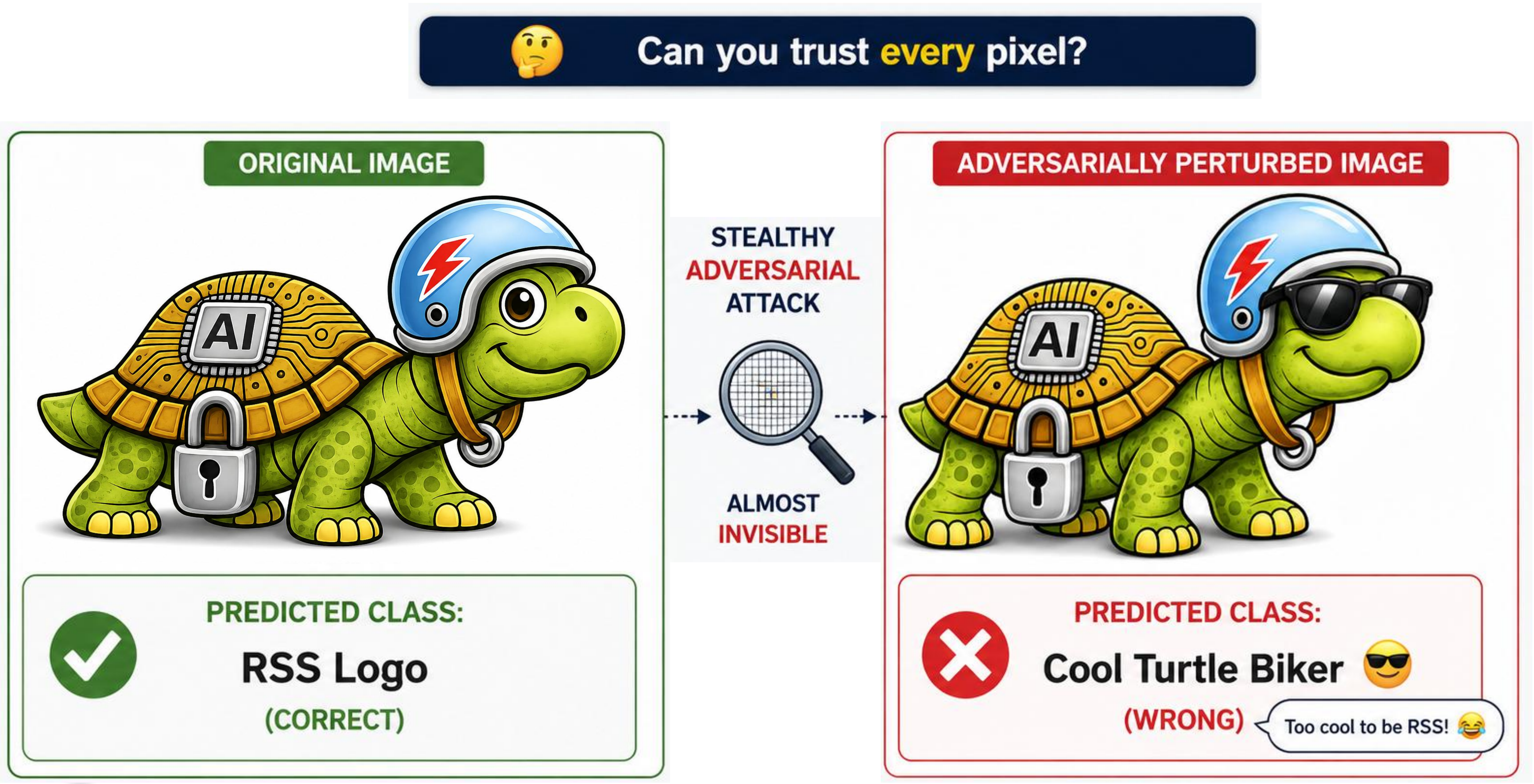
Faults in DNN-HAs

- ❑ Model parameters memory (weights, biases, etc.)
  - SEUs accumulate
- ❑ Buffers for inputs, activation and pooling; accumulators
  - SEUs are overwritten
- ❑ Logic of PEs and Control Units
  - FPGA: configuration memory



A. Bosio, et al. *"Emerging computing devices: Challenges and opportunities for test and reliability"*, ETS'21

Adversarial Attacks

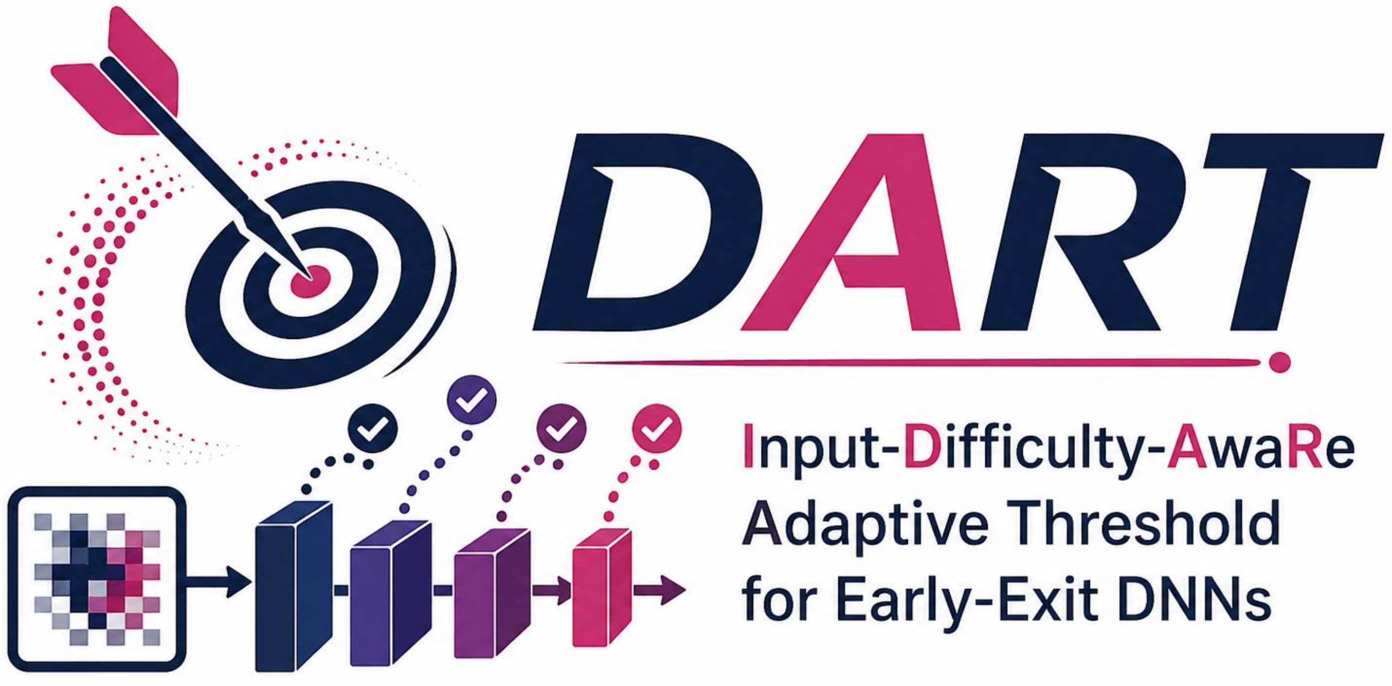


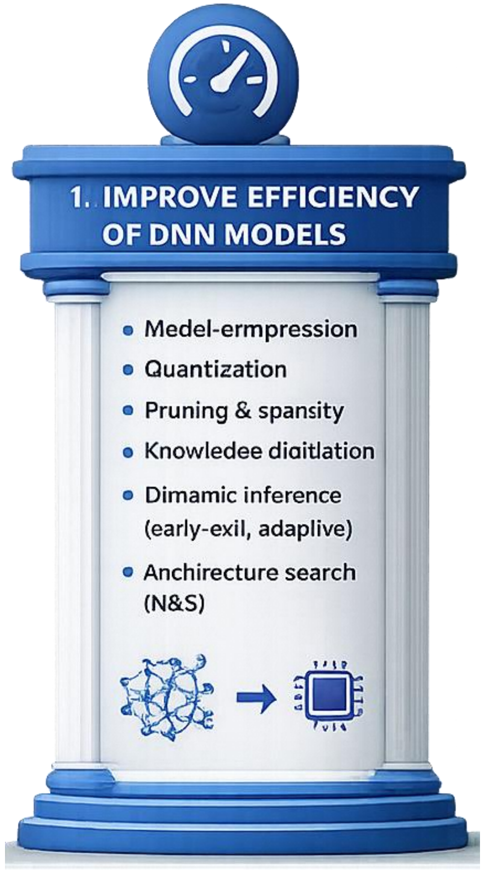
Pillars

1. IMPROVE EFFICIENCY OF DNN MODELS
Medel-ermpression
Quantization
Pruning & spansity
Knowledee diqitlation
Dimamic inference (early-exil, adaplive)
Anchirecture search (N&S)
2. ASSESS RELIABILITY AND SECURITY
Reliabtity evaluation (feurs- nalse, agiro)
Security assessment (adversanal, privacy, polsoning)
Mehlics & bendimarks
Analytical modeling & testing
3. ENHANCE RELIABILITY AND SECURITY
Bobust troining & regularization
Redundency & error mtileation
Adaphye syecision & ruirtime comro)
Secure interence (eneryption, loolalion)
Cordevign of rinvisty for trustworthiness.
COST EFFICIENCY
RESILIENT
Towards Efficien
re DNN Systems

Agenda

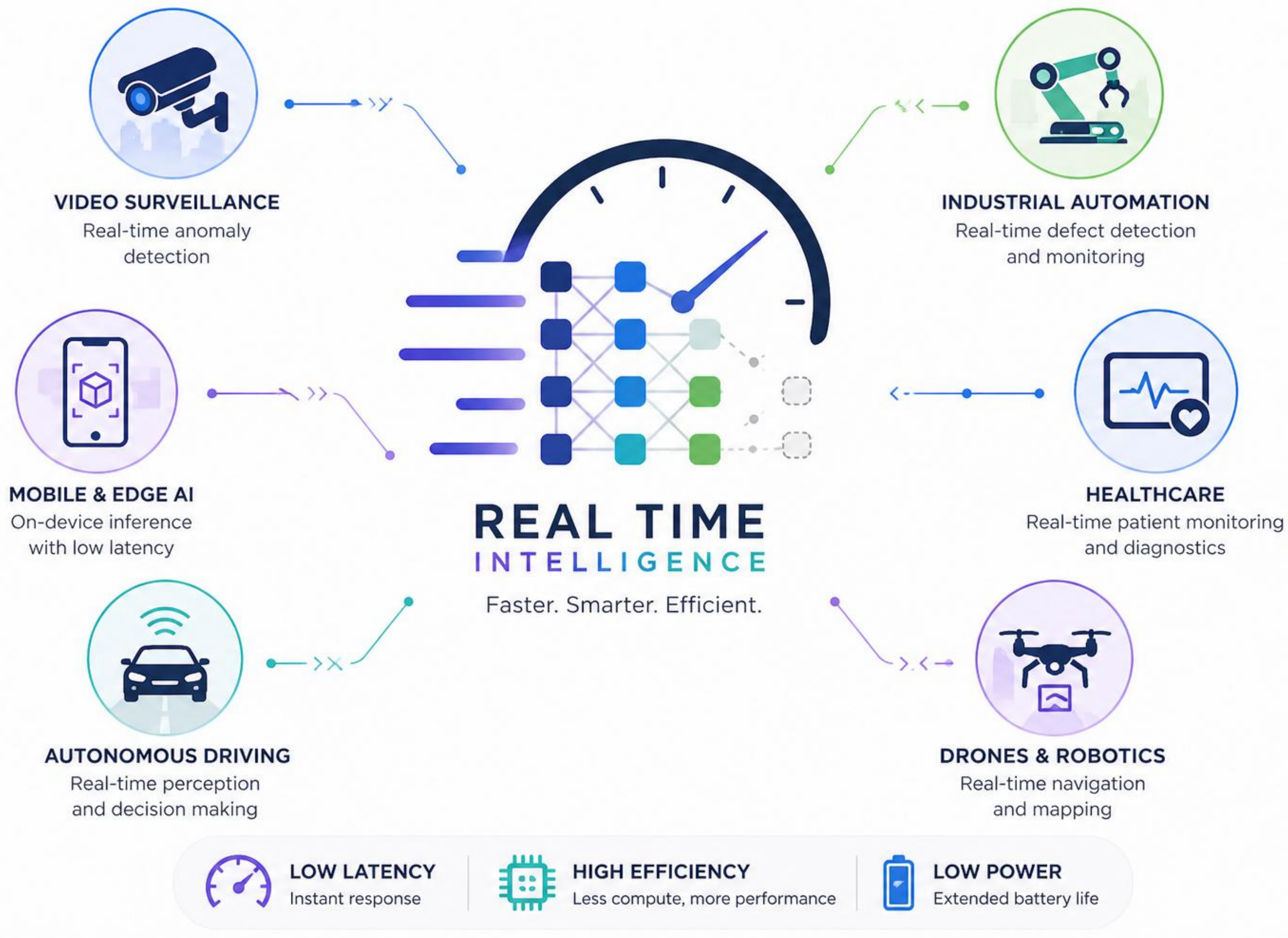
- **Dynamic Neural Networks**
- Approximate Computing





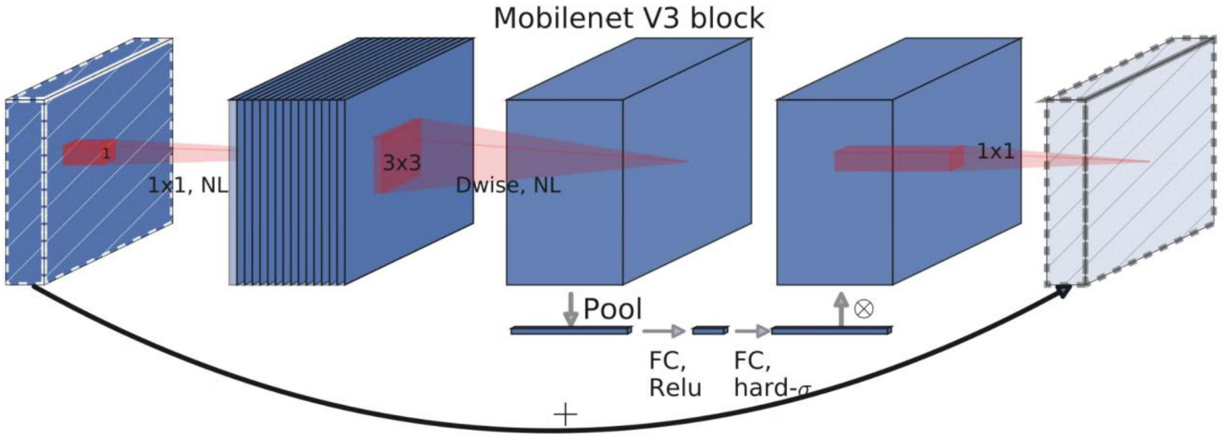
[1] **DART:** Input-Difficulty-Aware Adaptive Threshold for Early-Exit DNNs, **ICCAI** 2026
[2] **Adaptive Fault Resilience** for Early-Exit DNNs, **ATS/ITC-Asia** 2025

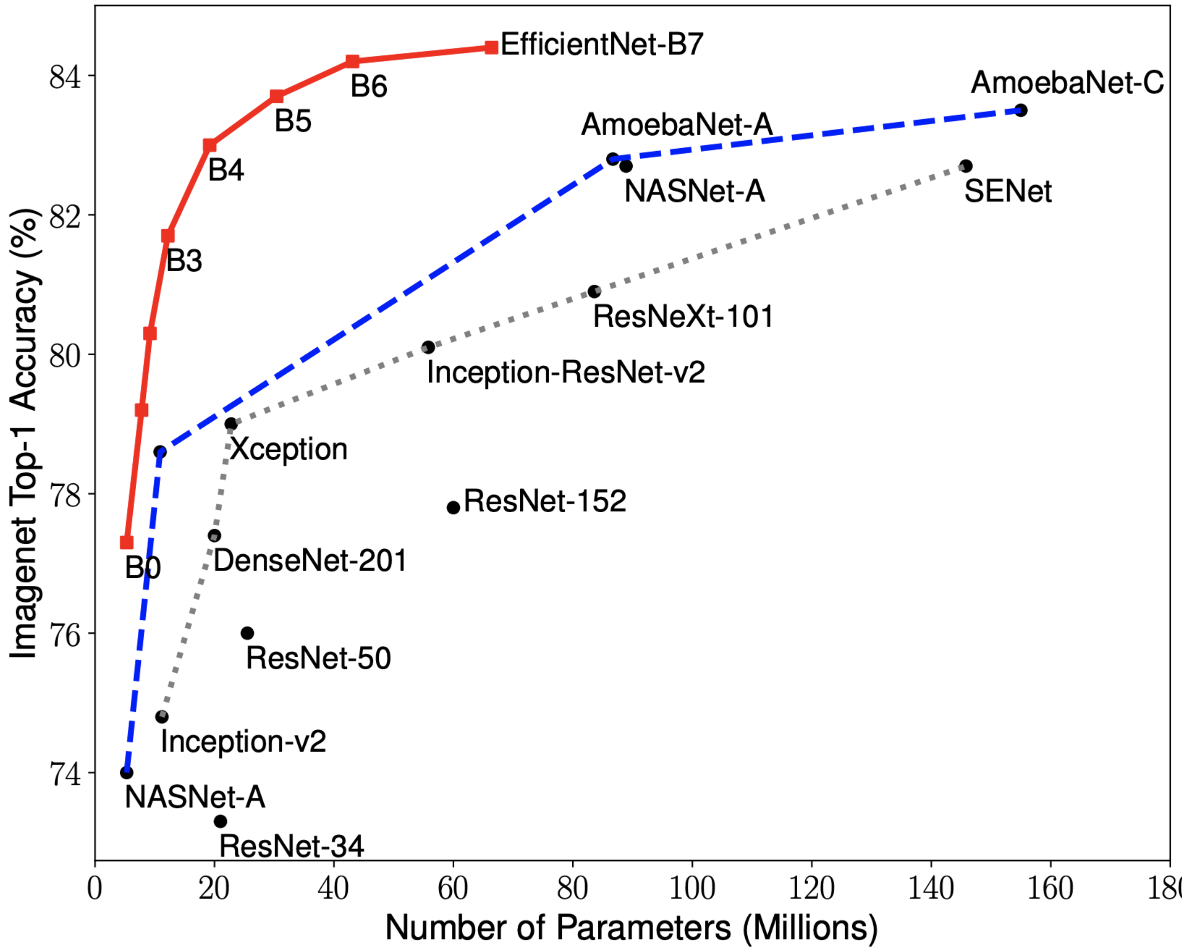
Real-time Applications



Model compression and Acceleration

- ❑Knowledge Distillation
- ❑Neural Architectural Search
- ❑Quantization
- ❑Pruning

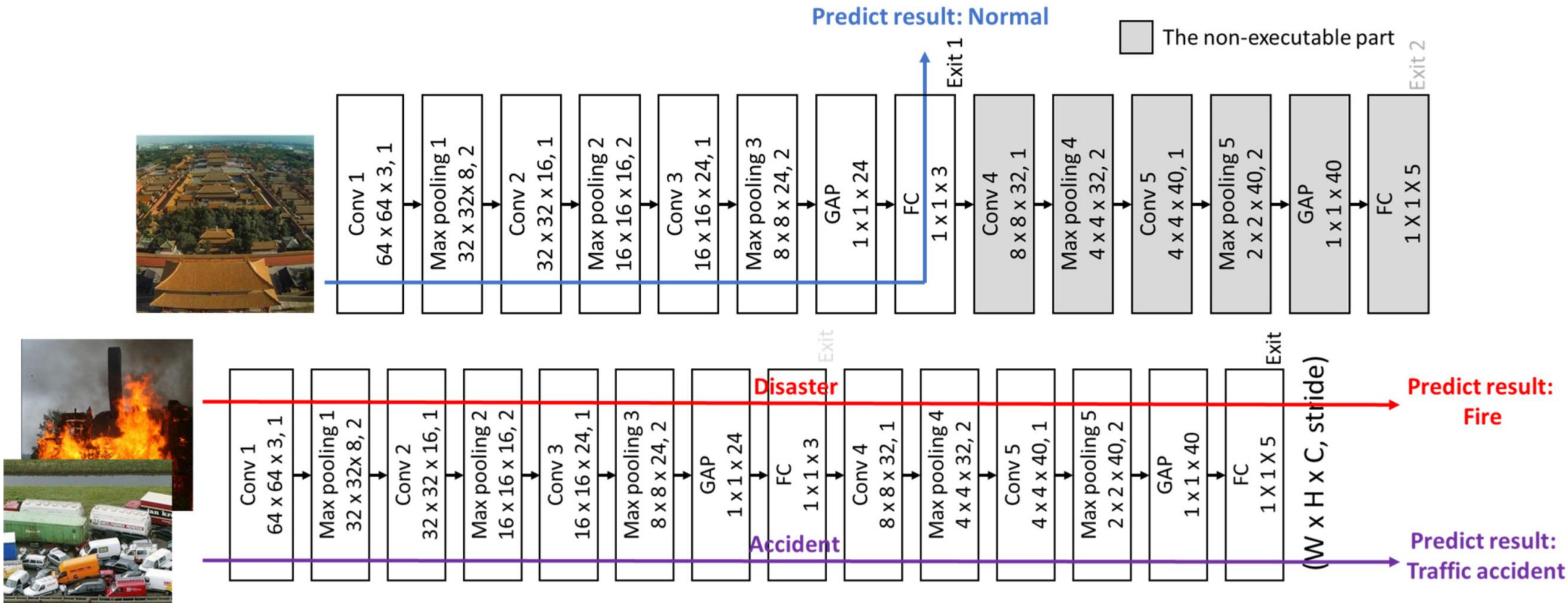




One-Size-Fits-All?

- ❑Ignoring complexity of Input
- ❑Fixed set of features
- ❑Fixed network size

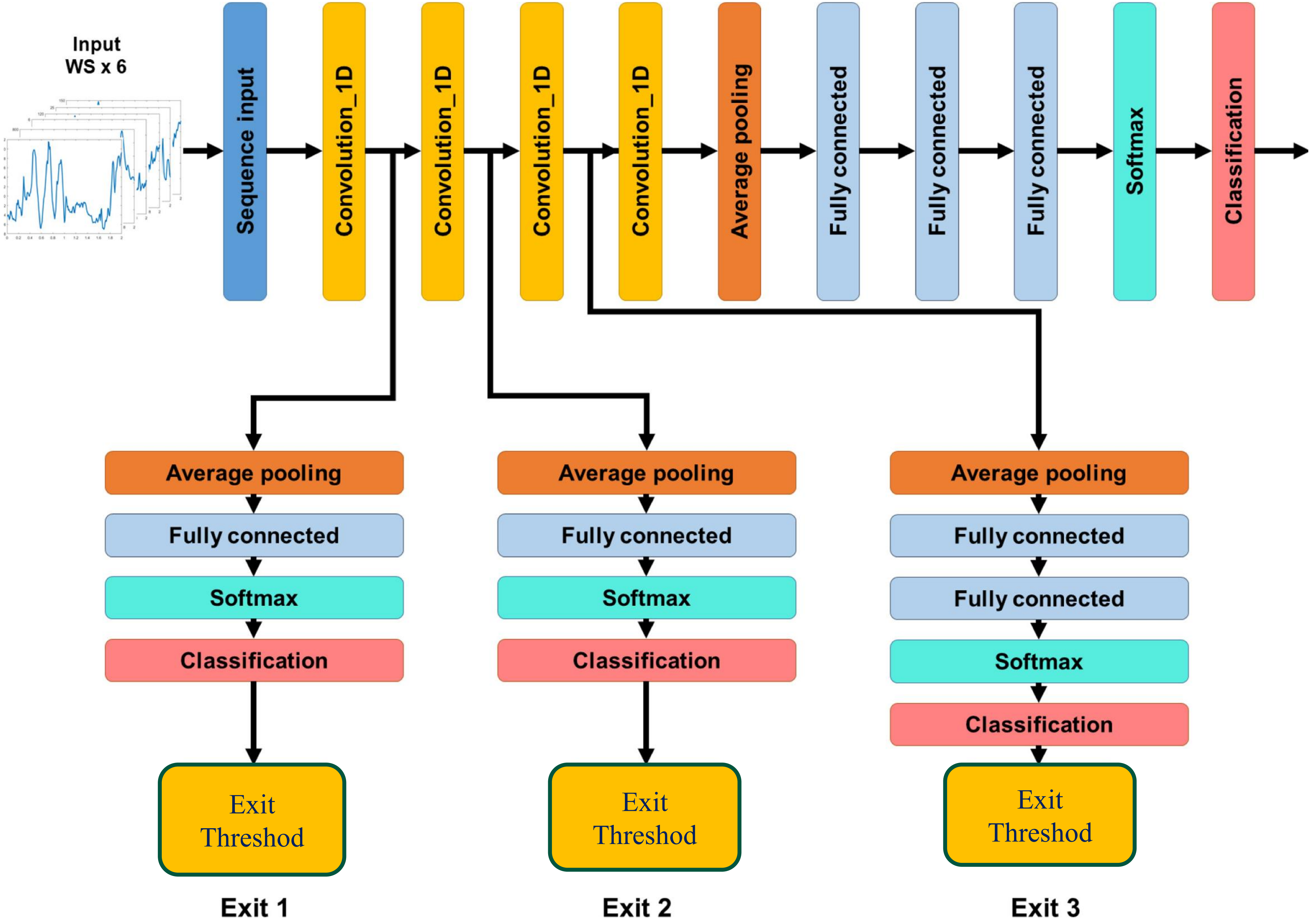
Early-Exit Dynamic Neural Networks



Manualy defined and hard-coded branch **locations** and **threshold** without **reliability** consideration

"BranchyNet: Fast inference via early exiting" *(S. Teerapittayanon et al., 2017)*

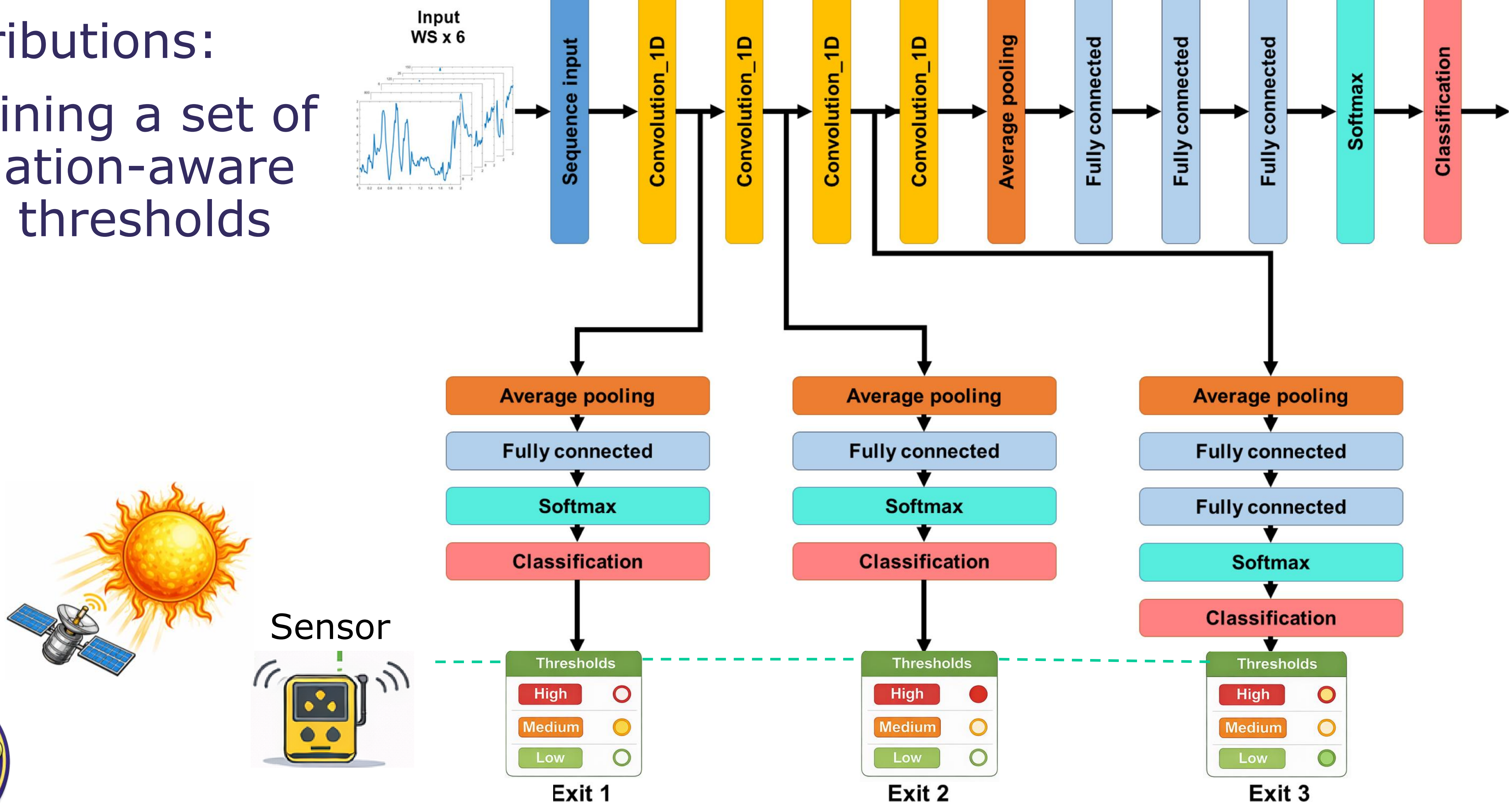
Early-Exit Dynamic Neural Networks



Early-Exit Dynamic Neural Networks

Contributions:

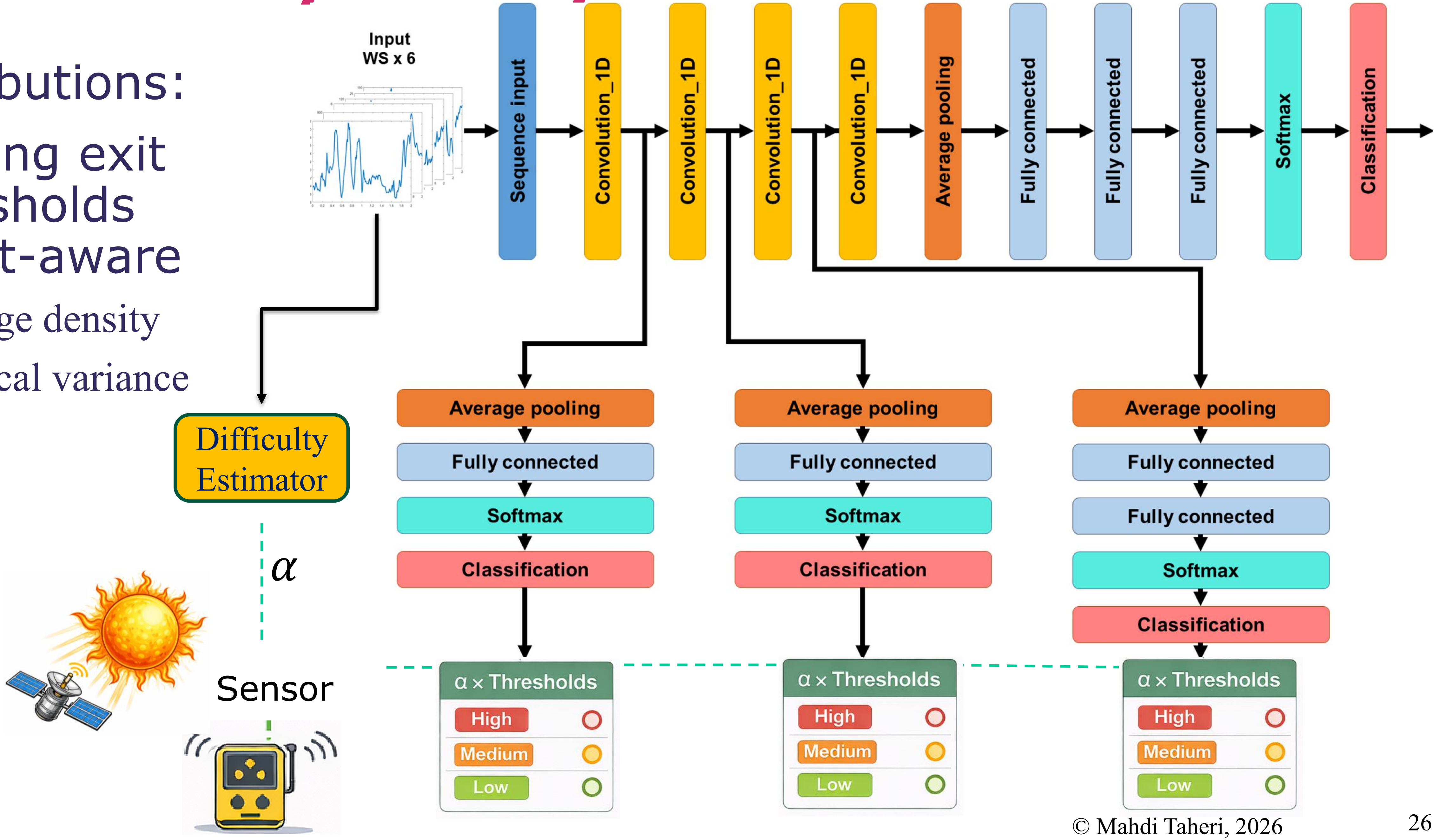
- ❑Defining a set of radiation-aware exit thresholds



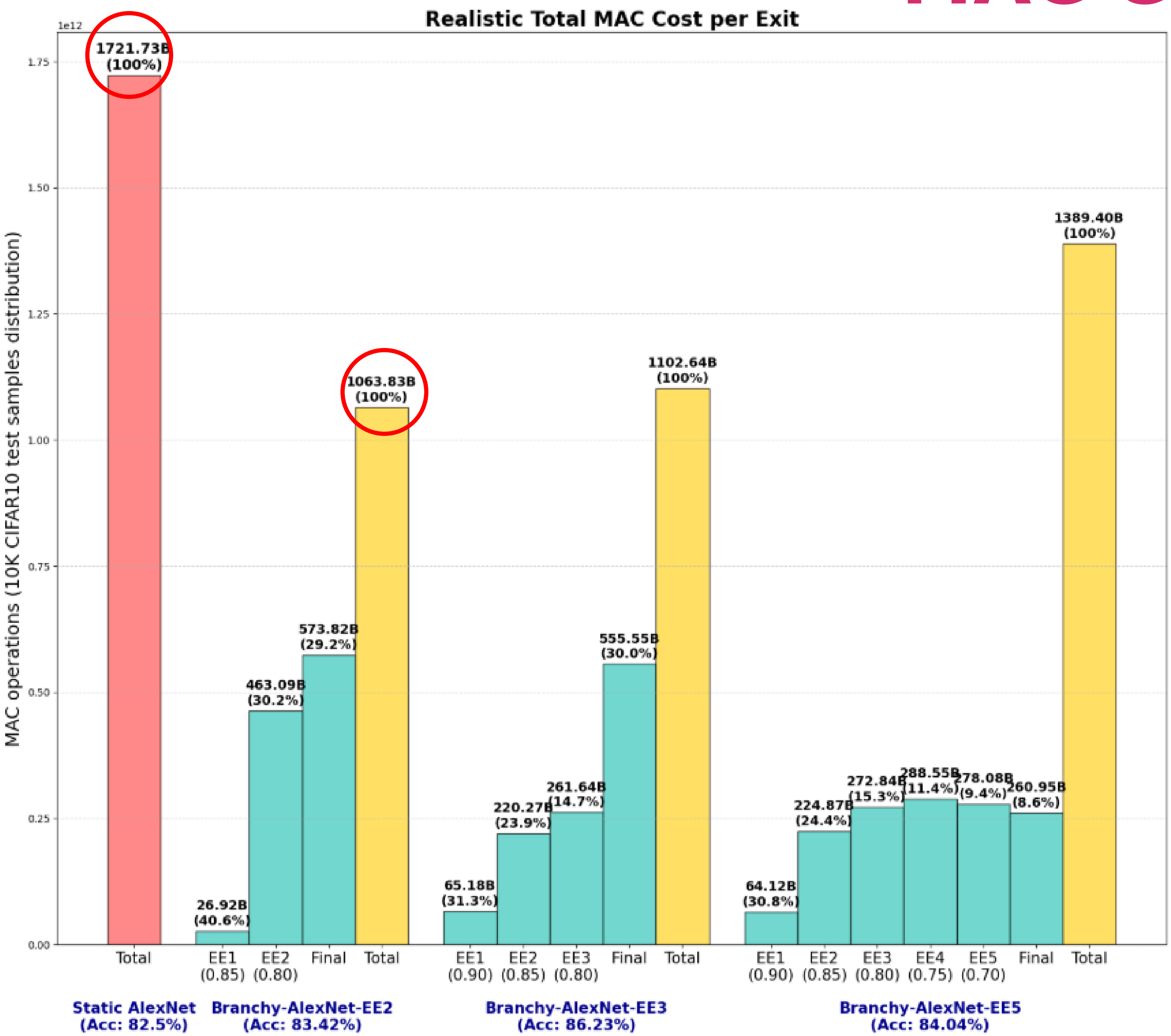
Early-Exit Dynamic Neural Networks

Contributions:

- ❑ Making exit thresholds Input-aware
  - Edge density
  - Local variance
  - …



MAC Calculations



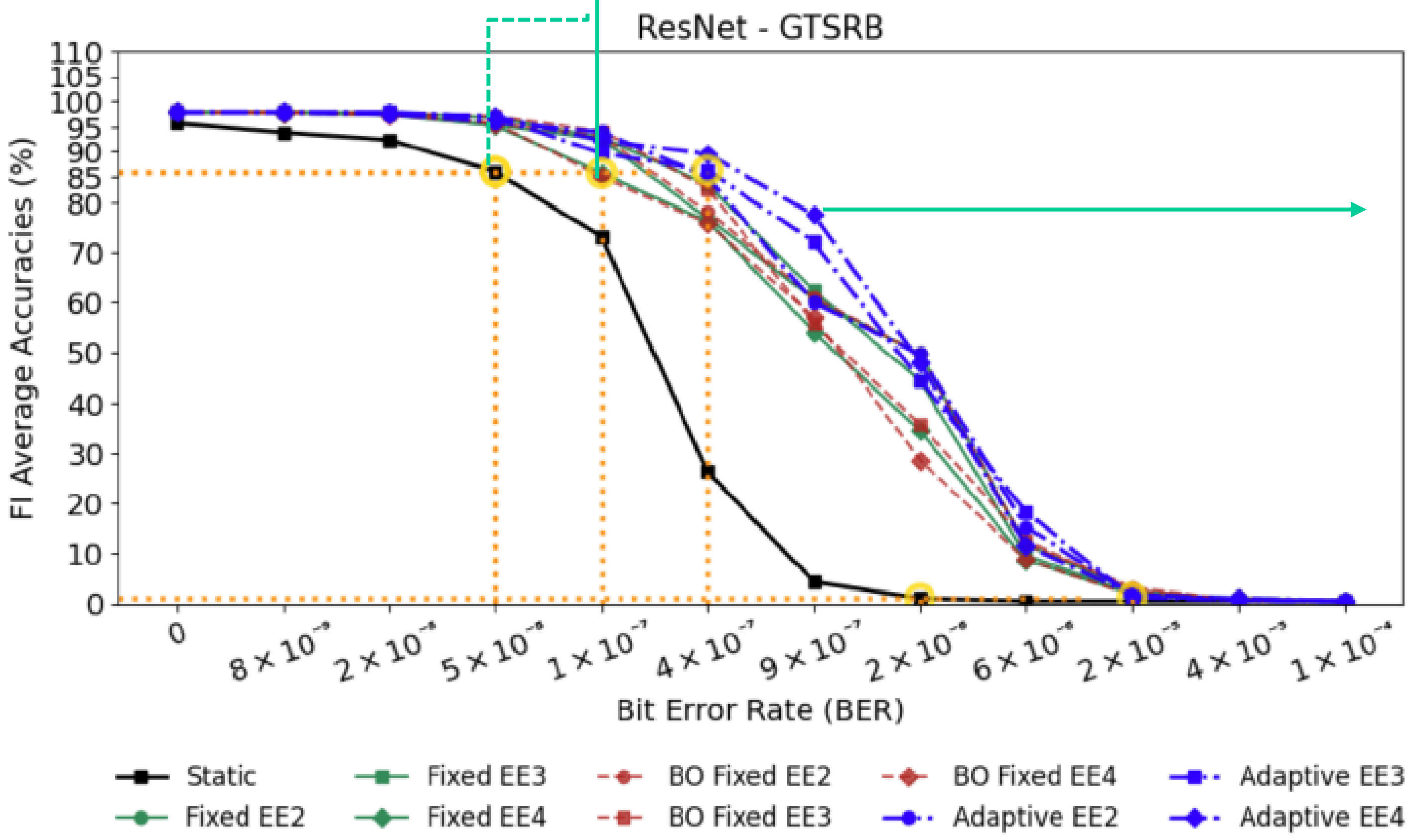
Experimental Results

- Performance Comparison Across Architectures and Datasets

Architecture	Method	Acc. (%)	Time (ms)	Energy (mJ)	Speedup (×)
CIFAR-10 Results					
AlexNet	Static	85.29	0.08	5.18	1.0×
	BranchyNet	83.15	0.07	4.57	1.14×
	DART	82.86	0.05	1.88	1.60×
ResNet-18	Static	88.32	0.27	17.66	1.0×
	BranchyNet	87.72	0.16	10.43	1.69×
	DART	85.35	0.12	7.53	2.25×
VGG-16	Static	79.16	0.20	9.41	1.0×
	BranchyNet	81.82	0.08	3.70	2.50×
	DART	80.20	0.06	2.54	3.33×

Experimental Results

DyNN with higher fault resilience than Static NN

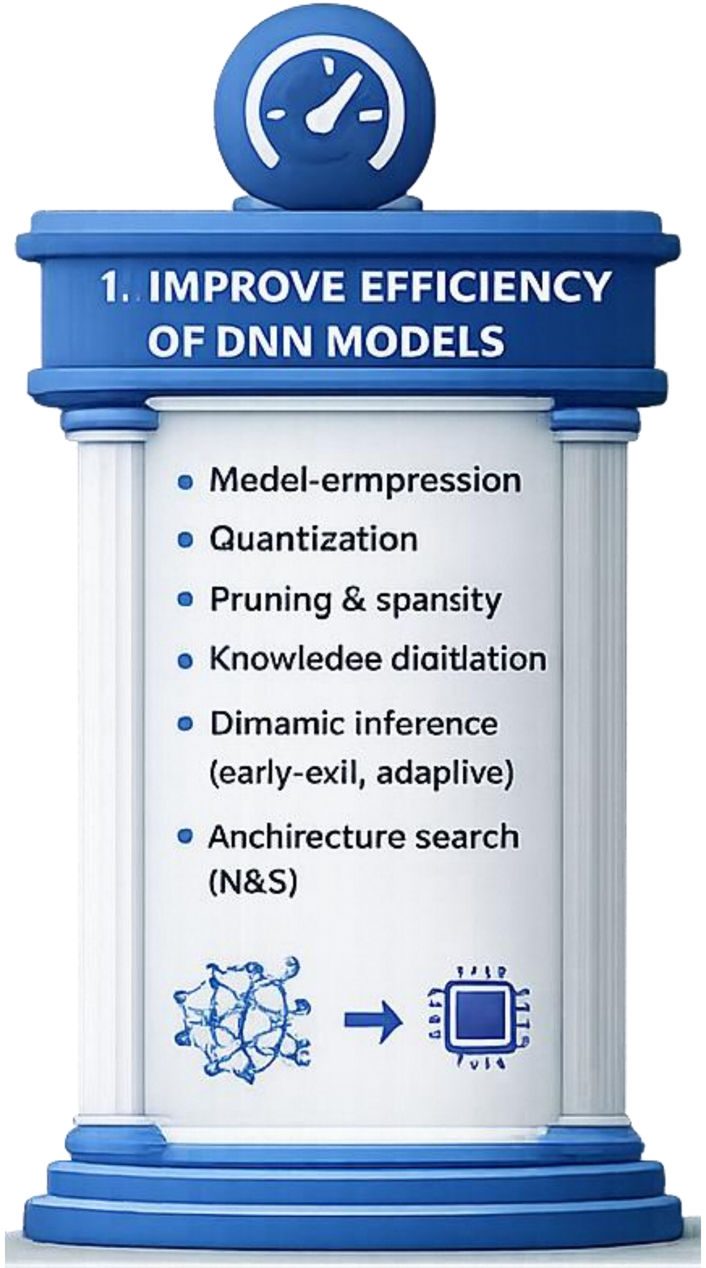


Fault Adaptive Exit Thresholds demonstrating highest fault resilience

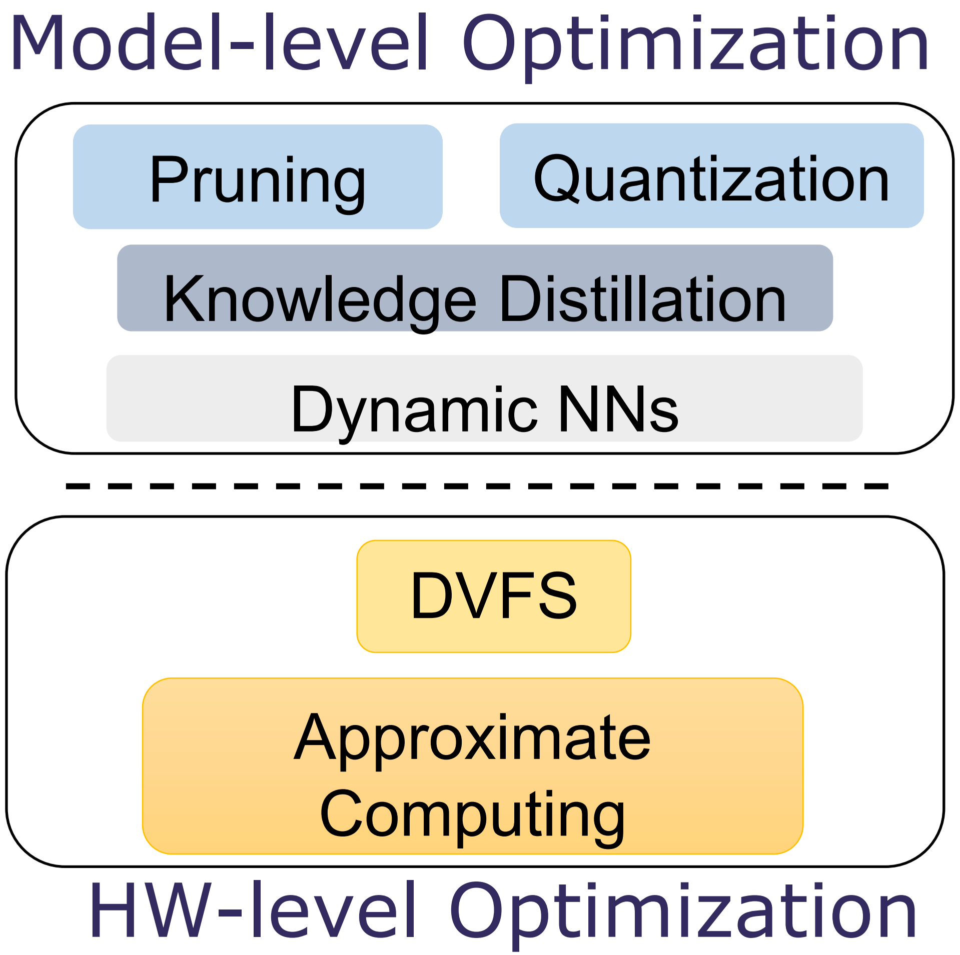
Agenda

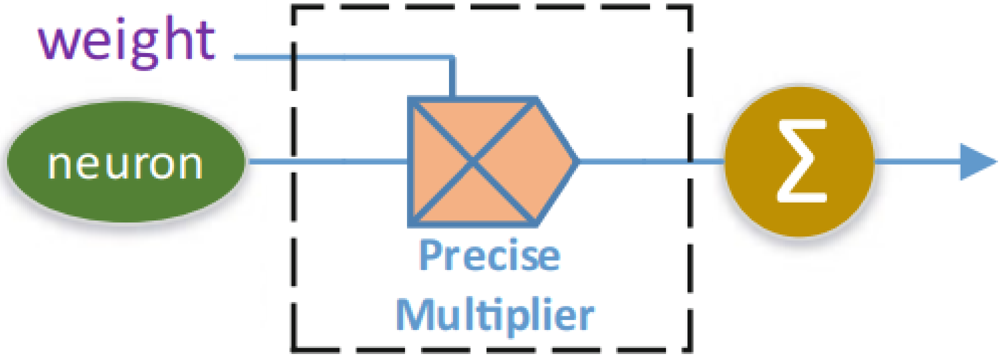
- ❑ Dynamic Neural Networks
- ❑ **Approximate Computing**

[3] **DeepAxe**: A Framework for Exploration of Approximation and Reliability Trade-offs in DNN Accelerators, **ISQED 2023**
[4] **HAWX**: A Hardware-Aware Framework for Fast and Scalable Approximation of DNNs, **DATE 2026** (Verona, Italy)

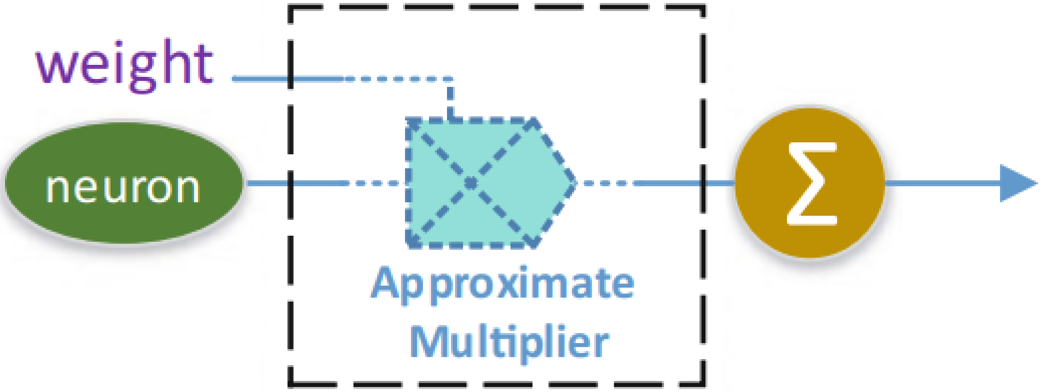


Approximate Configuration Search





Precise hardware model



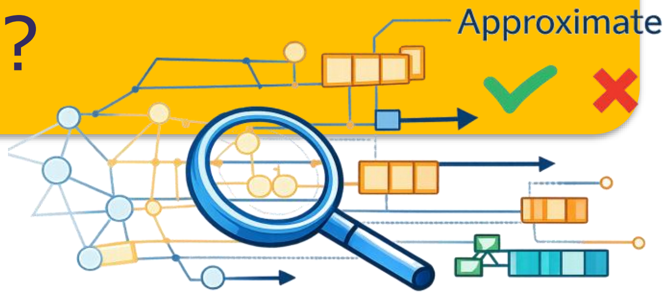
Approximate hardware model

Liu, W., et. Al., 2018. Approximate Computing and Its Application to Hardware Security.

Approximate Configuration Search

- Approximation Challenges in Safety-Critical Edge AI

Where and how much approximation is safe for the model?



How does approximation affect reliability?

Approximate Configuration Search

HAWX

- Fully automated, scalable hardware-aware DSE framework
- Rapid identification of high-quality AxC configurations
- Predictive models for accuracy, power, and area
- Quantitative scoring at:
  - Operation
  - Filter
  - Layer
  - Model level
- User-guided exploration with custom AxC blocks

DeepAxe

- Performs systematic reliability analysis under faults on HAWX output
- Identifies Pareto-optimal designs balancing:
  - Accuracy
  - Reliability
  - Hardware efficiency
- Automatic generation of synthesizable HLS code

Approximate Configuration Search

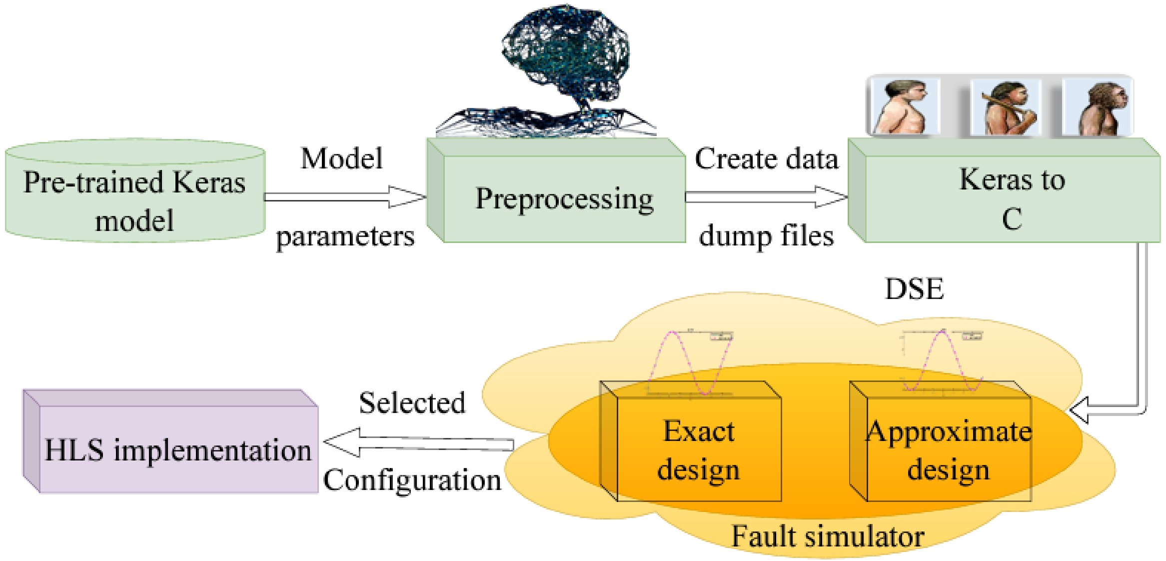
HAWX

DeepAxe

$$hawx(o,i) = \left|\frac{\partial L}{\partial o}\right| \times |\Delta out(o,i)|$$

$$|\Delta out(o,i)| = |out(o,exact)_i - out(o,axc)_i|$$

$$\nabla L_o = \frac{\partial \mathcal{L}}{\partial o}$$



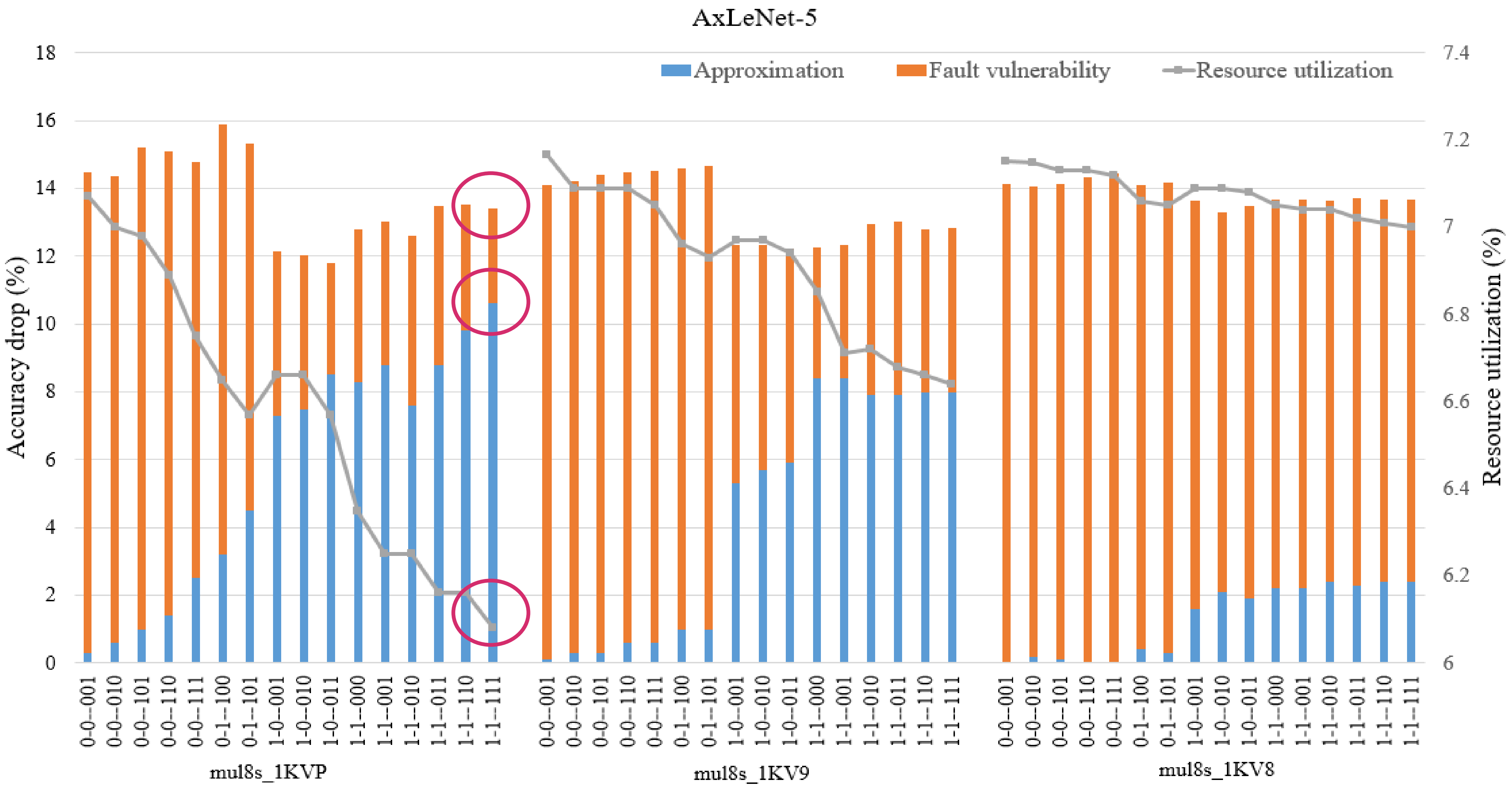
Experimental Results

	Model	Layers	No. Multiplications	Accuracy
MNIST	LeNet-5	2 Conv, 3 FC	416,520	**97.80**
CIFAR10	VGG-11	8 Conv, 3 FC	171,679,744	**92.23**
GTSRB	ResNet-18	20 Conv, 1 FC	37,526,476	**92.98**
	EfficientLiteNet-B0	46 Conv, 1 FC	346,525,024	**99.80**

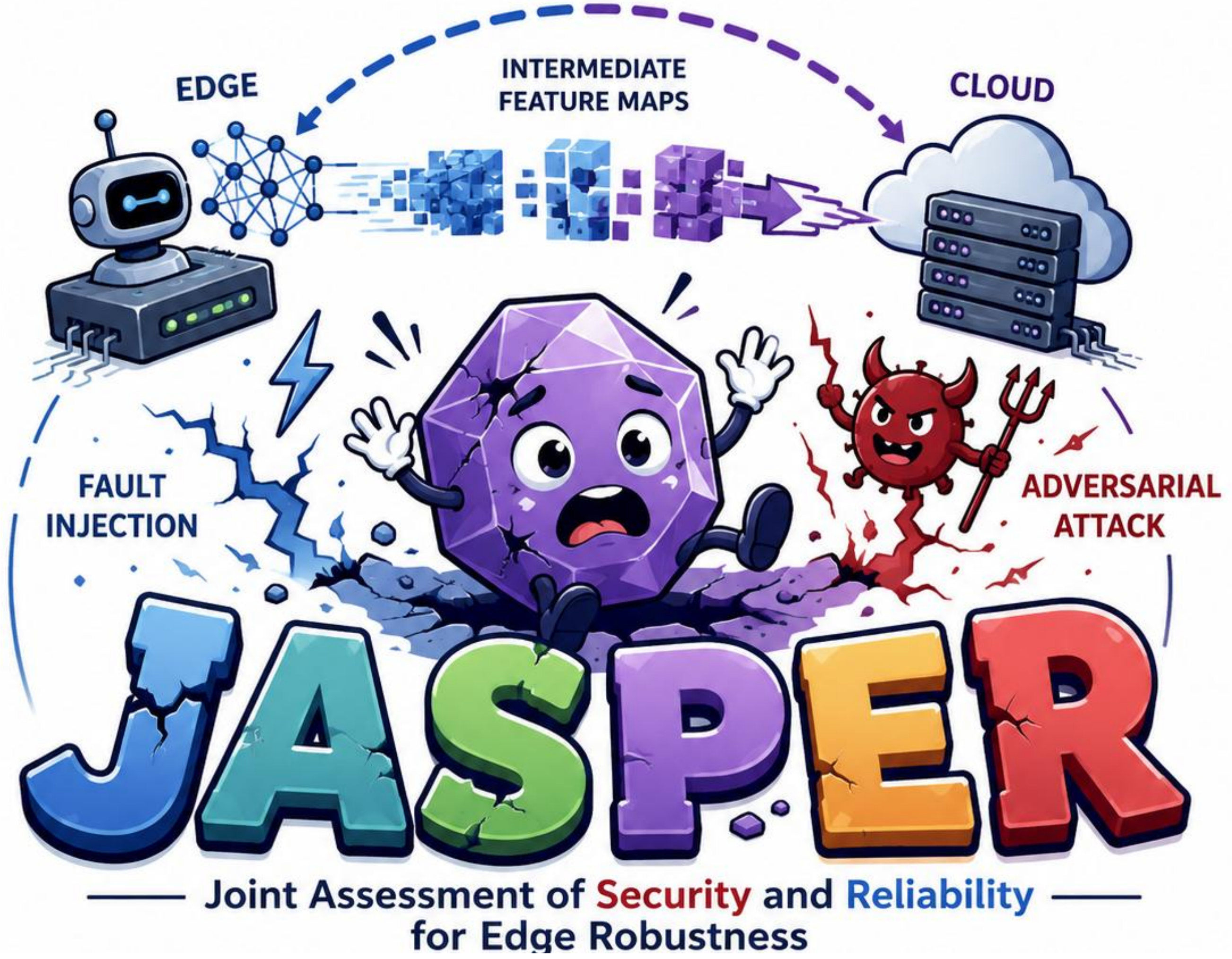
❑ HAWX vs. SOTA

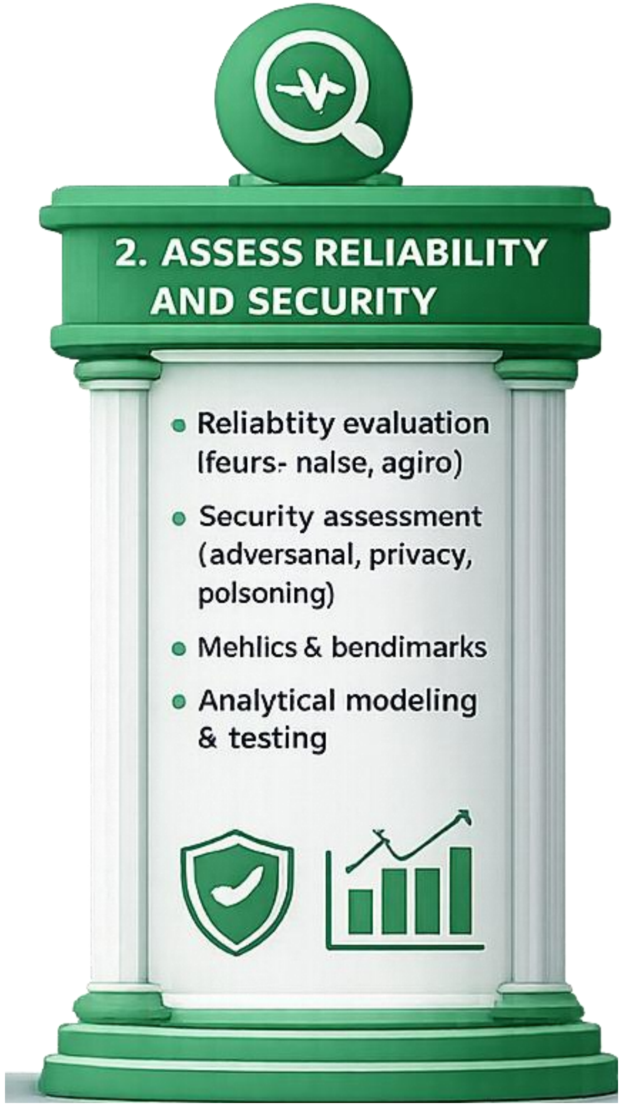
Feature	Prior Works	HAWX
Search Granularity	Model/Layer only	Operation, Filter, Layer, Model
Hardware Awareness	Limited or none	Spatial and Temporal Support
Speedup vs. Exhaustive	Up to 10^6x (Model Level)	10^{5000}x (Filter Level)
Heterogeneous Support	Limited	Full Support

Experimental Results



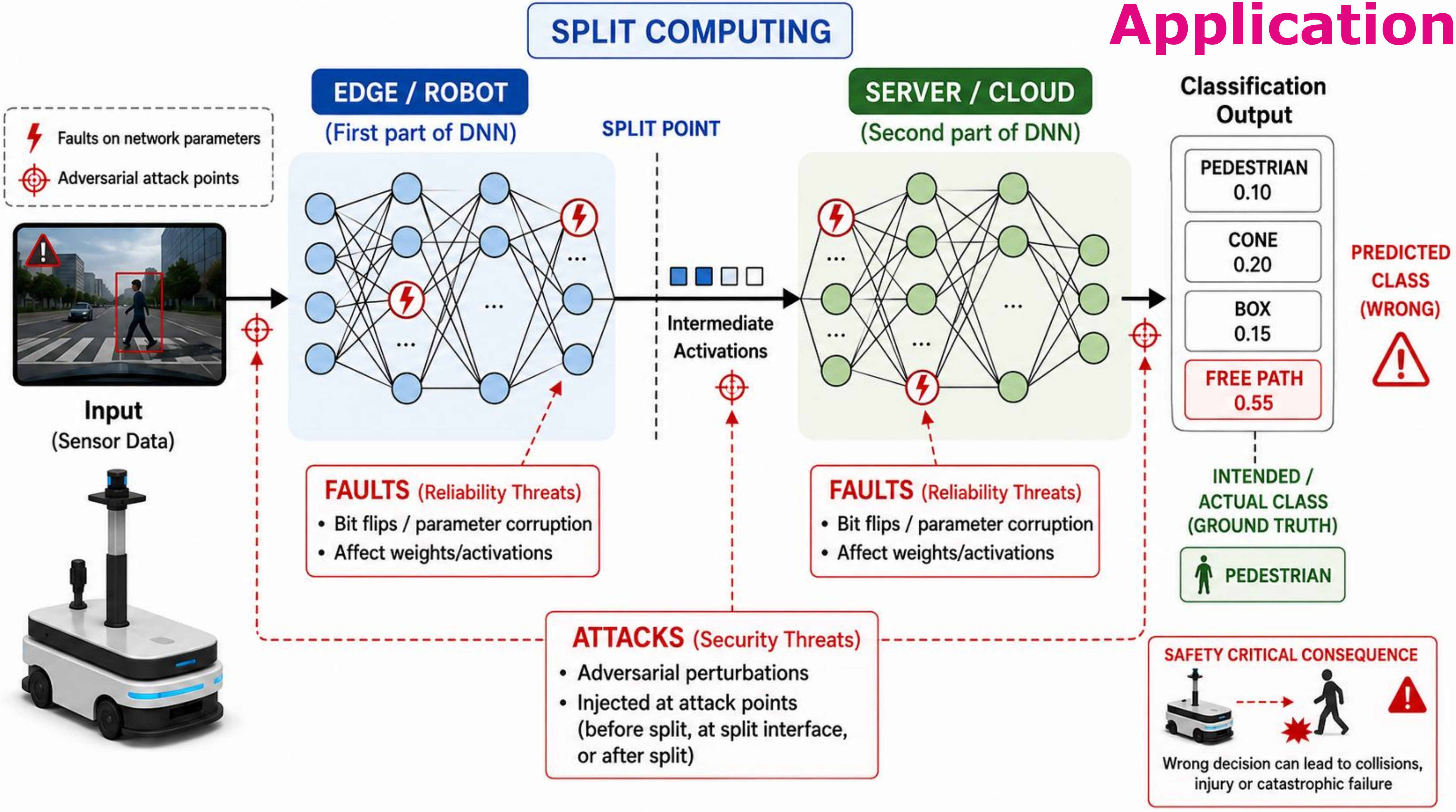
❑Assessment and Joint Resilience Metric





[3] **JASPER**: A unified metric for DNN resilience, **DFTS 2026,** Accepted

Application

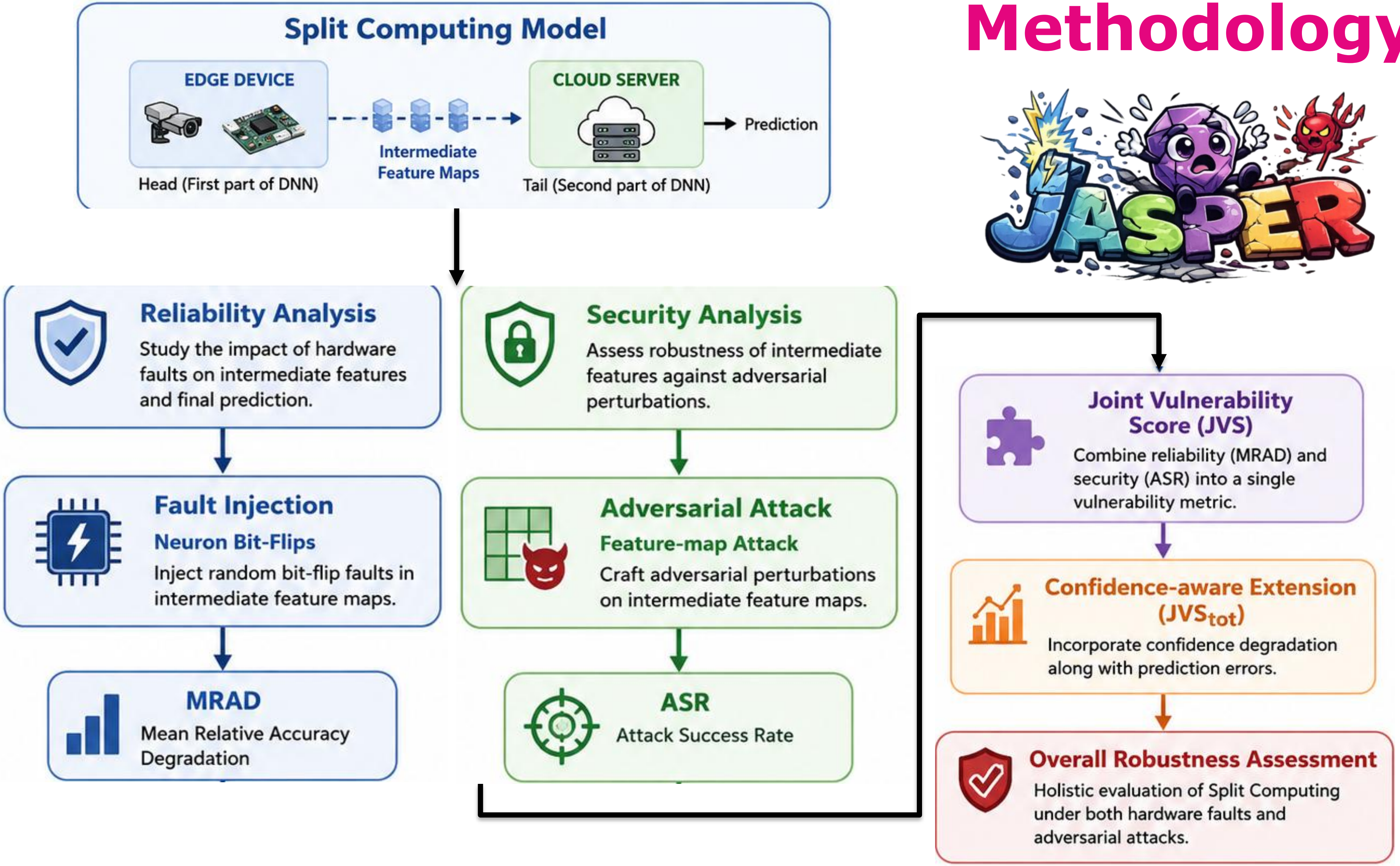


Related Works

- Current Literature is considering limited aspects

Area	Existing Work	Limitation
Split Computing	Compression, communication efficiency, latency	Ignore robustness
Reliability	Fault injection, SDC analysis, MRAD	Ignore adversarial attacks
Security	Intermediate-feature adversarial attacks, ASR	Ignore hardware faults
Joint Reliability & Security	Very limited	No unified quantitative metric

Methodology



Joint Resilience Metric

- MRAD normalized reliability vulnerability score of the model under fault
- ASR quantifies the system's vulnerability under adversarial attack

$$JVS = \sqrt{MRAD \cdot ASR}$$

$$MRAD \in [0, 100]$$

$$ASR \in [0, 100]$$

Network Confidence is also important

- If prediction remains correct under perturbations, and how confident the prediction is

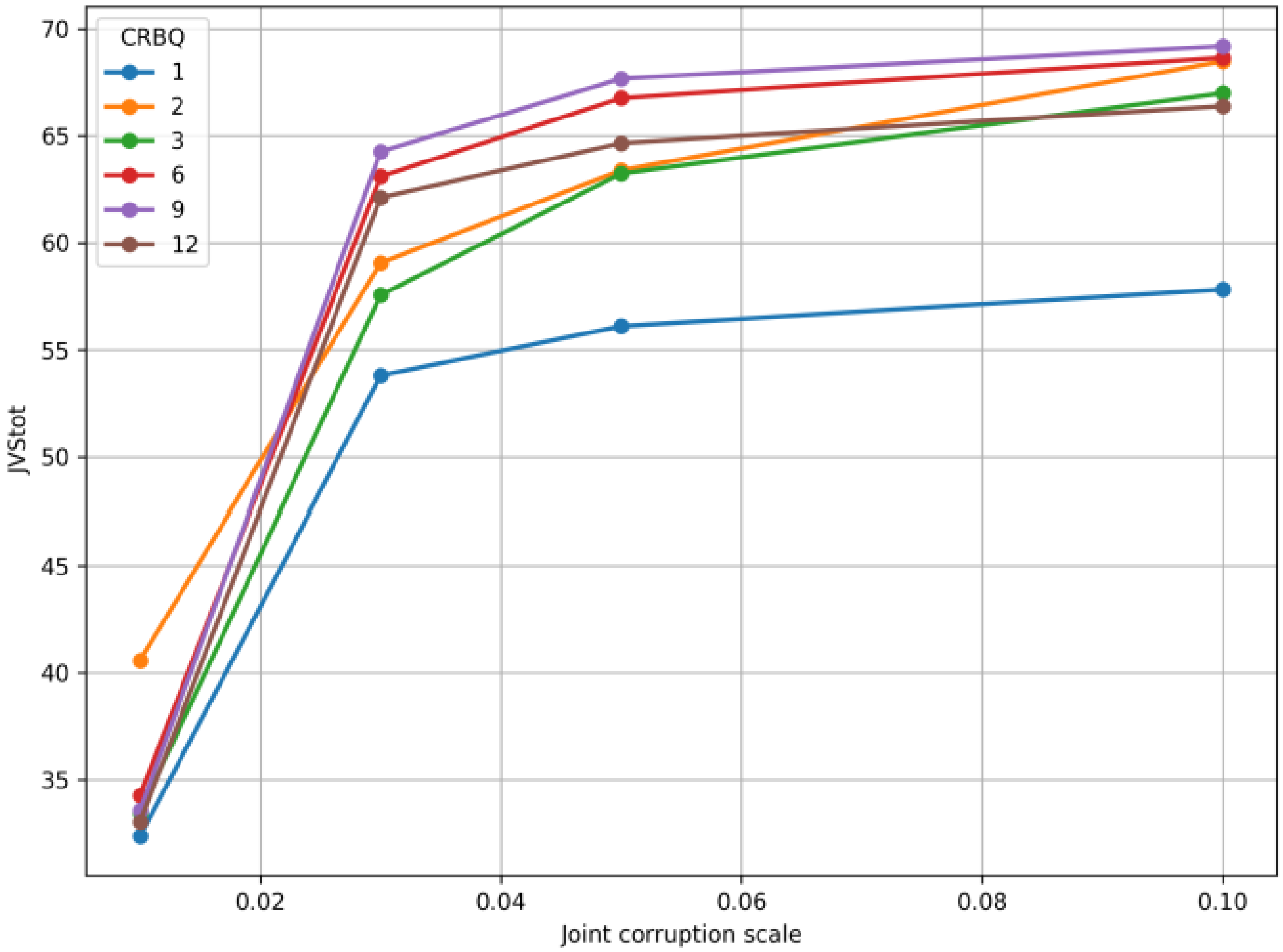
$$JVS_{tot} = \sqrt{R_{tot} \cdot S_{tot}}.$$

$$R_{tot} = \alpha * MRAD + (1 - \alpha) * (100 - R_{\Delta}),$$

$$S_{tot} = \beta * ASR + (1 - \beta) * (100 - S_{\Delta}),$$

Results

- increasing the compression of the transmitted feature maps amplifies the combined impact of reliability and security threats



(a) CRBQ configuration.

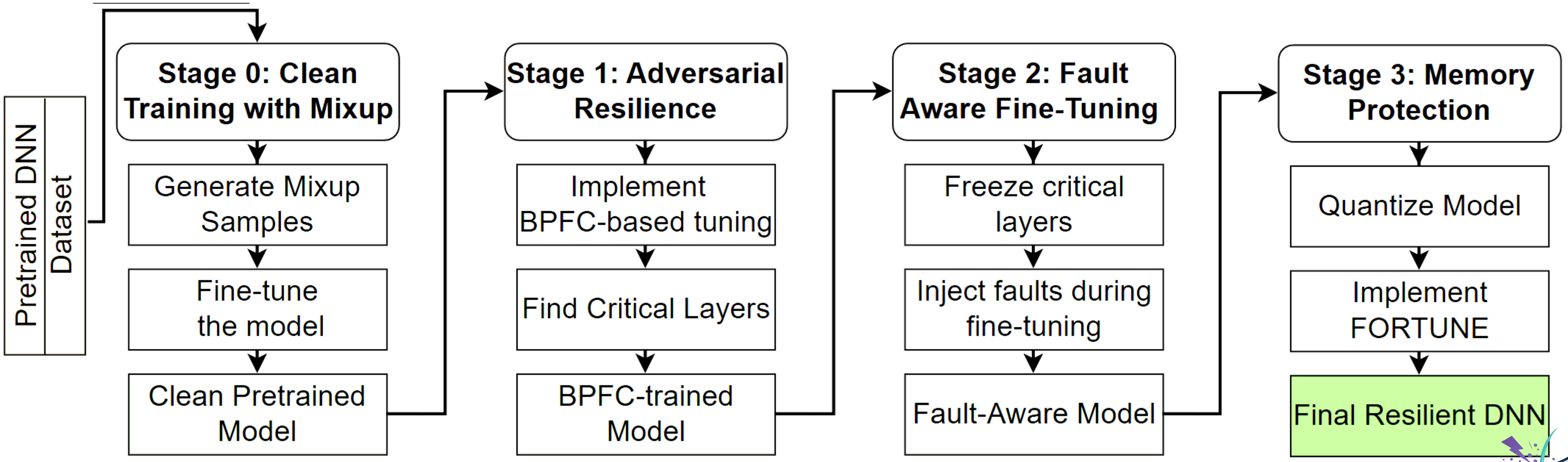
Agenda

❑Resilience Enhancement

[4] **RESQ**: A Unified Framework for REliability- and Security DNNs, **LATS 2025**

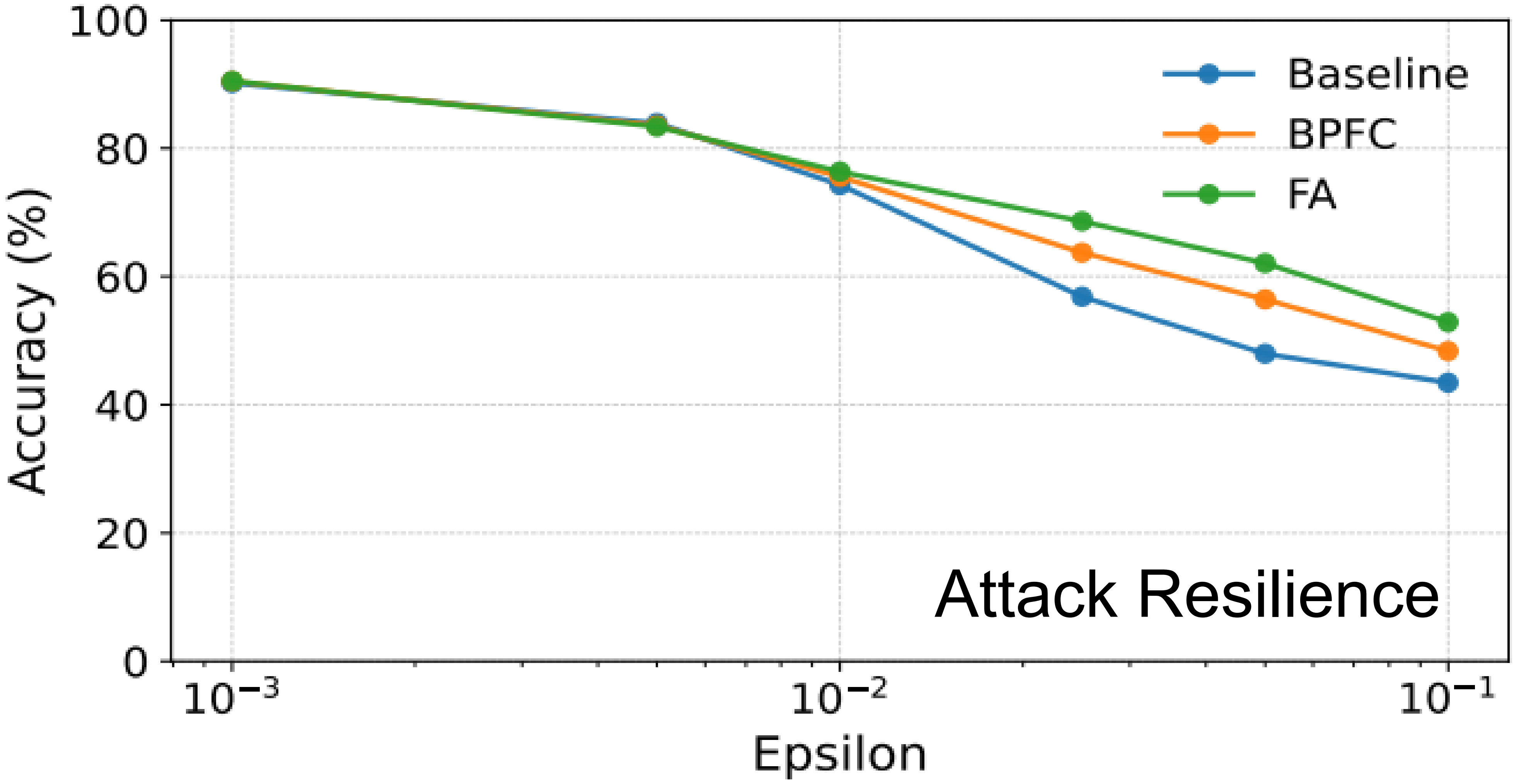
Let´s RESQ!

- A unified three-stage framework that produces a quantized DNN with balanced **fault** and **attack robustness**



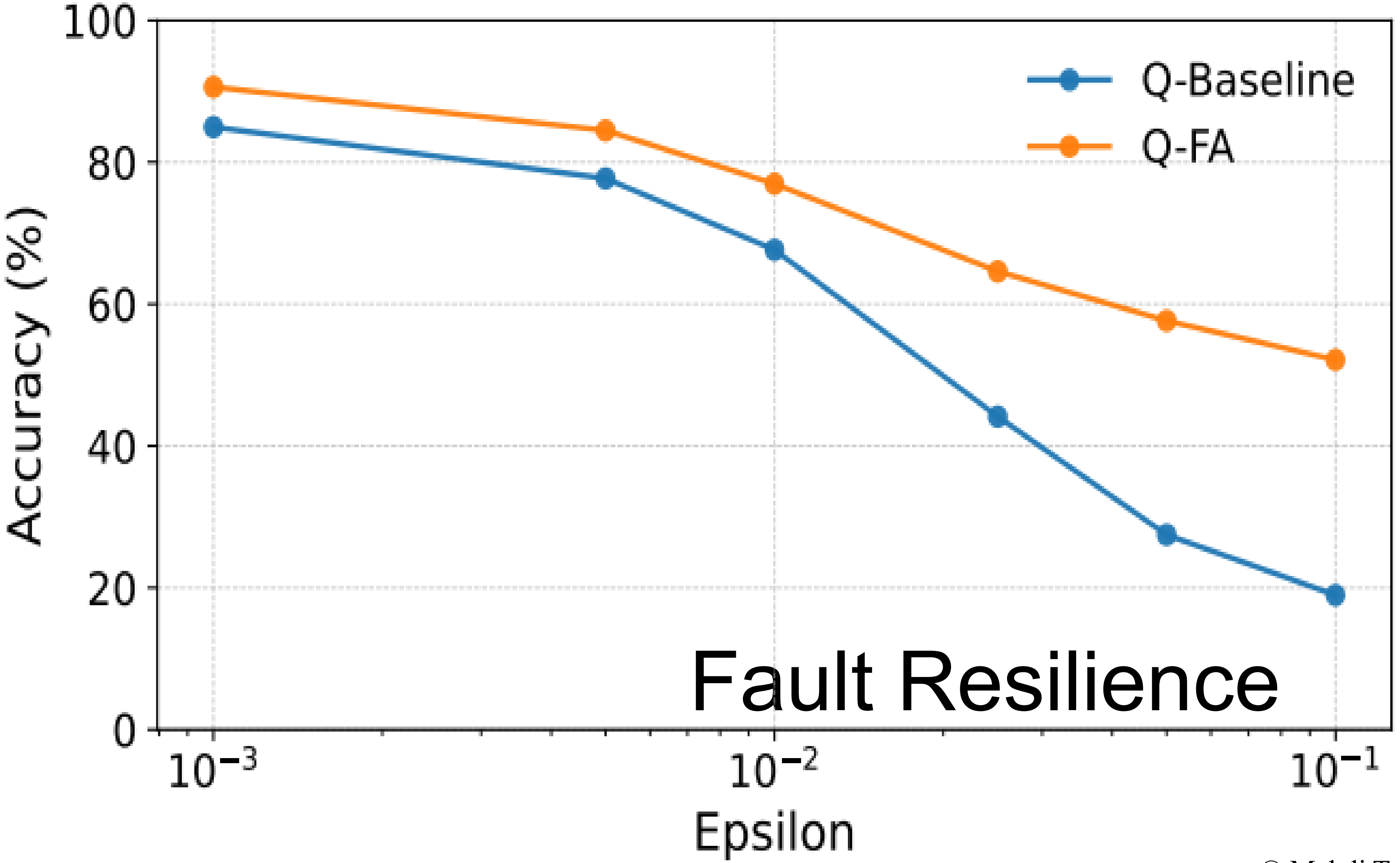
Results

- Different benchmarks like SwinTiny and ResNet-18



Results

- Different benchmarks like SwinTiny and ResNet-18



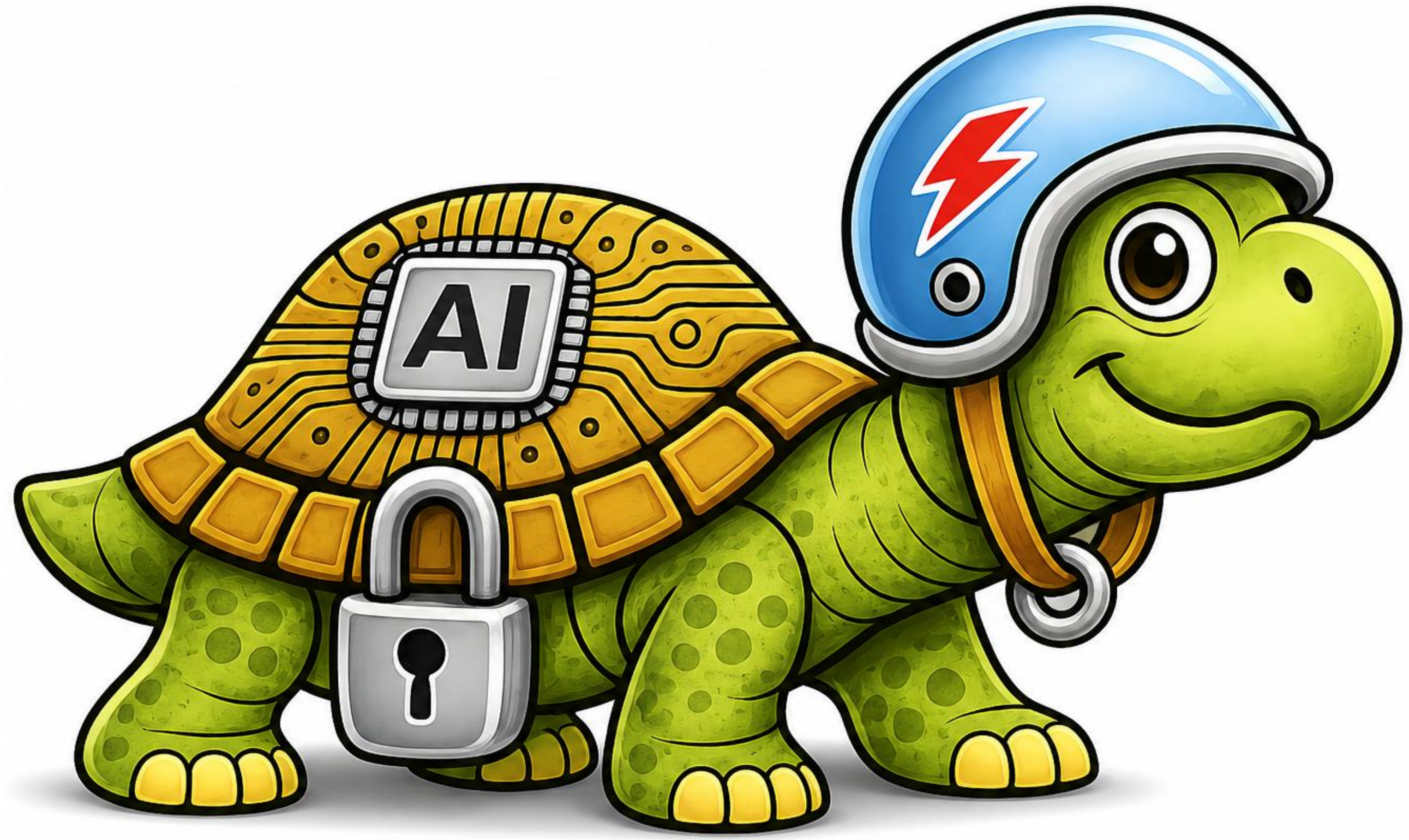
Conclusion

- **Emerging AI architectures:** improved efficiency **and** resilience
- **Approximate Computing (AxC):** optimization opportunity with new reliability and security challenges
- **Joint resilience assessment:** new metrics beyond isolated reliability or security
  - **JVS / JVStot** for unified vulnerability evaluation
- **RESQ**: unified framework for reliability and security enhancement
  - ↑ **10.35%** adversarial resilience
  - ↑ **12.47%** fault resilience
  - **0%** clean accuracy loss

THANK YOU

DR. M. TAHERI
BTU COTTBUS-SENFTENBERG
TALTECH
TAHERI@B-TU.DE

Find me here!!



"True speed is predictable, secure, and reliable execution"