TWINRELECT

Soft Error Reliability Analysis & Hardening of NVDLA MAC Units

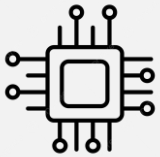
Nikolaos Chatzivangelis,
PhD Candidate, University of Thessaly

Co-funded by the European Union
UK Research and Innovation

Why Harden DNN Accelerators?

Radiation in Critical Systems

Avionics, space & automotive applications face increasing radiation exposure as semiconductor nodes shrink — smaller transistors mean lower charge thresholds and higher soft error susceptibility.



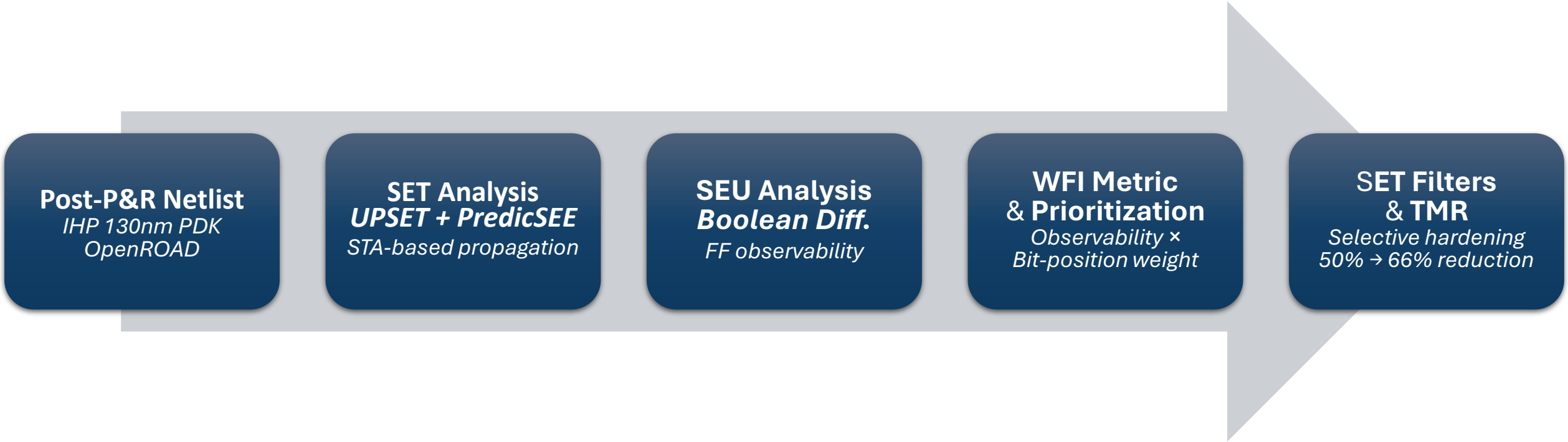
NVDLA & MAC Units Are the Core

MAC units are the primary computational blocks of DNN accelerators. A single fault can propagate through many neurons and significantly corrupt the final inference output.

Two Distinct Fault Classes Must Be Addressed

Single Event Transients (SETs) corrupt combinational logic transiently; Single Event Upsets (SEUs) permanently flip flip-flop state — each requires a dedicated analysis and mitigation approach.

Proposed Methodology at a Glance



Key differentiator:

*Unlike prior cross-layer DNN studies, our approach operates on the **post-place-and-route netlist**, capturing physical-aware SET generation (PredicSEE) and formal BDD-based SEU analysis in a unified gate-level flow.*

NVDLA MAC Unit Architecture

NVDLA

Open hardware architecture for DNN acceleration, offering highly parallel computation.

MAC Units

Perform multiply–accumulate operations — the core arithmetic of every convolution layer.

Parallel arrays

Organized with pipelined accumulation; memory elements store intermediate results between stages.

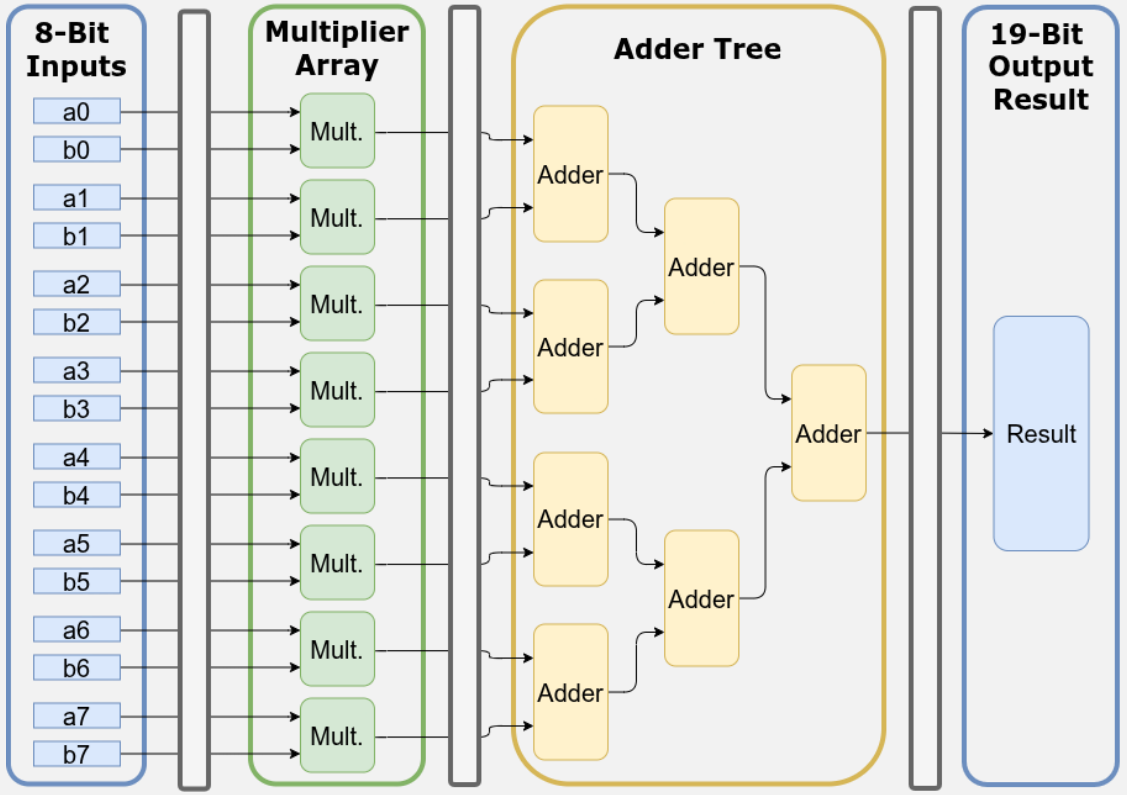
Fault impact

A single fault in one MAC can corrupt entire DNN inference outputs due to propagation.

Implementation

IHP 130 nm BiCMOS Open Source PDK via Synopsis Design Compiler synthesis flow.

MAC Pipeline Datapath



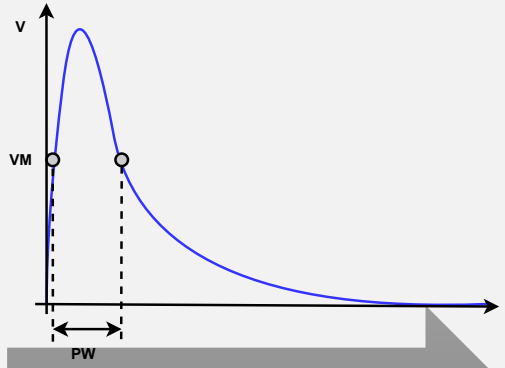
SET Analysis Overview

MAC Unit SET Analysis Flow

1

PredicSEE Tool

Modeling of SET generation for the technology cells to obtain realistic pulses caused by particle strikes



Voltage waveforms

2

UPSET Tool

SET pulse propagation and evaluation the error probability at the design's Flip-Flops

Latching Probability Formula

$$P_{latched} = PSE_T(Ei) \times \frac{PW_{SET} - WL_{atching}}{P_{clock}}$$

Quantifies the probability that an SET pulse is latched by a sequential element — combining logical masking probability with path electrical characteristics (from SPEF files) and timing window analysis.

Single Event Upset (SEU) Analysis

Approach: Boolean Difference on BDDs

- Models SEU as a bit-flip at flip-flop output
- Binary Decision Diagrams for exact Boolean difference computation
- Evaluates probability that a fault propagates to architectural observation points
- SAIF-derived signal probabilities reflect actual inference workload patterns
- Extension of UPSET: fully integrated with the SET analysis flow

Boolean Difference

$$\frac{\partial fj}{\partial xi} = fj(xi = 0, x') \oplus fj(xi = 1, x')$$

Evaluates to 1 when a change in FF_i output produces a change at endpoint e_j — i.e., the fault is observable.

Observability Probability

$$P_{obs}(FFi \rightarrow ej) = P(\frac{\partial fj}{\partial xi} = 1) = \sum_{x \in V} \prod_{k=1}^{n} P_k(vk)$$

Sum over input vectors where Boolean difference = 1, weighted by workload-derived signal probabilities from SAIF simulation.

Flip-Flop & Design Vulnerability Metrics

Weighted Fault Impact (WFI) — Flip-Flop Sensitivity Metric

$$WFI(FFi) = \sum_{j \in Ei} W_j \times Po_{bs}(FFi \rightarrow ej)$$

$$Wj = \frac{Bit\ possition\ +\ 1}{Bus\ Width}$$

SET Criticality — SET_{crit}

$$SET_{crit}(FFi) = AVG(PlacthedSET_{}) \times WFI(FFi_{})$$

Combines SET capture probability with fault propagation impact — used to prioritize SET filter insertion.

Design Vulnerability — DWFI

$$DWFI = \sum_{i=1}^{n} WFI(FFi)$$

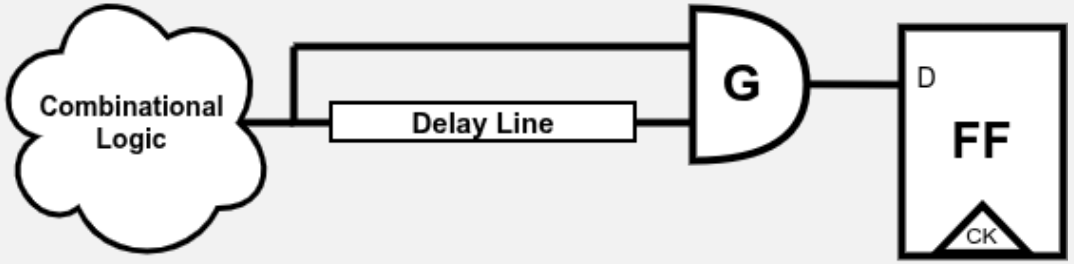
Design-wide cumulative vulnerability score. Higher DWFI = greater susceptibility. Primary metric tracked throughout hardening experiments.

Key Insight: MSB endpoints receive weight 1.0 while LSBs receive 1/B — a fault at bit 15 of an output accumulator causes exponentially more numerical damage than at bit 0. WFI directly reflects the expected numerical damage caused by an SEU at each flip-flop.

Selective Hardening Methodology

SET Filters

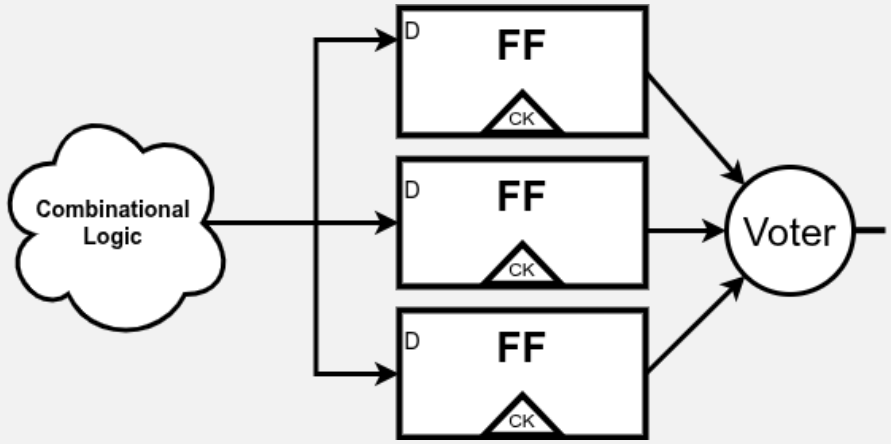
Prioritized by: SETcrit metric



- Guard gate + delay buffer chain before each FF
- Buffer count: $N_{buf} = \left\lceil \frac{PW_{avg}}{T_{buf}} \right\rceil$
- Achieves logical masking of transient pulses below average PW
- Custom IHP 130nm guard gate standard cell
- Average-based sizing -> Partial attenuation beyond average PW

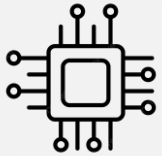
Triple Modular Redundancy (TMR)

Prioritized by: WFI metric



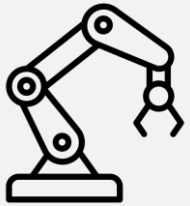
- Each critical FF replicated to 3 redundant storage elements
- Majority voter determines correct output value
- Replicas spatially separated — single strike cannot flip multiple
- Targets highest-WFI FFs first: output-stage & MSB registers
- Near-linear area/power scaling — predictable overhead

Experimental Setup



Technology Node

IHP 130nm BiCMOS Open Source PDK. Provides physical foundation and custom standard cells.



Physical Design

Synthesis using Synopsis Design Compiler. Placement, routing, and parasitics extraction (SPEF) performed via OpenROAD.

Workload Simulation

Switching probabilities derived from SAIF files generated in Cadence Xcelium by simulating the MAC with uniform random inputs

Framework Integration

Post-placement netlist, SPEF, SAIF, and PDK data are integrated into the UPSET tool for automated hardening.

Experimental Analysis: *TMR Results*

Coverage	#TMR FFs	DWFI	Δ(DWFI)	Area (µm²)	Power (mW)	Timing (ns)
Baseline (0%)	0	246.15	—	30,598	3.02	4.96
25%	53	156.91	−36.3%	35,540	4.03	5.03
50%	106	83.14	−66.2%	40,336	5.03	5.03
75%	160	29.70	−87.9%	45,223	6.12	5.03
100%	213	0	−100%	50,109	6.81	5.20

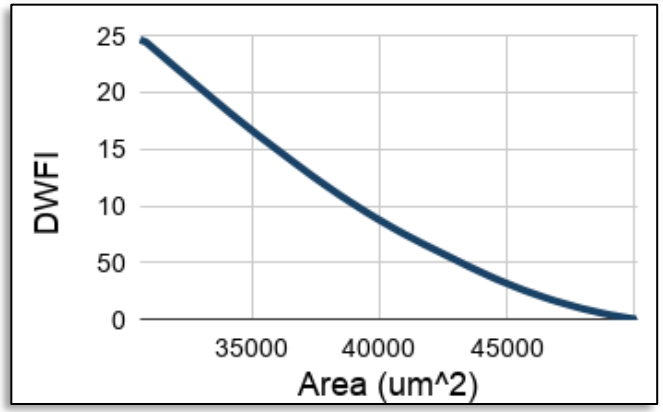
-66%
DWFI reduction at 50% TMR

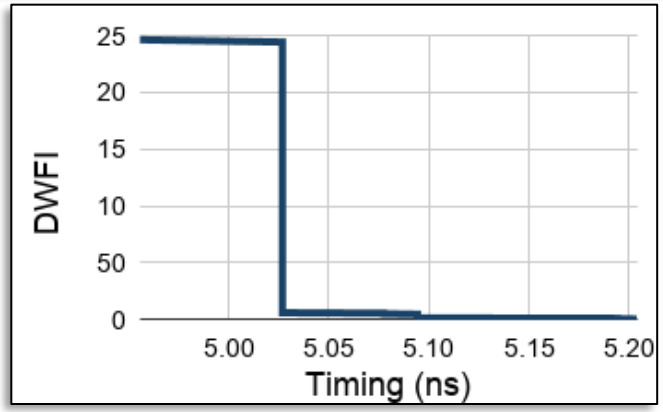
+32%
Area overhead at 50% TMR

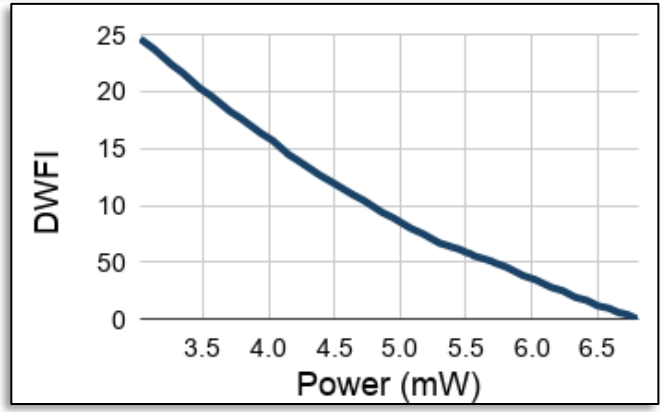
+5%
Max timing penalty (100%)

+125%
Power increase at full TMR

SEU Optimization Curves







Experimental Analysis: *SET Filter Results*

Coverage	# Filters	APW (×10³)	Δ(APW)	LASET (×10³)	Δ(LASET)	Area (um^2)	Power (mW)	Timing (ns)
Baseline (0%)	0	1087.9	—	107.9	—	30598	3.0	4.96
25%	36	728.5	−33.0%	72.0	−33.3%	36173	6.0	6.58
50%	72	366.3	−66.3%	35.8	−66.8%	40650	9.0	6.58
75%	108	294.9	−72.9%	28.6	−73.5%	43095	9.4	6.58
100%	144	284.1	−73.9%	28.0	−74.1%	44448	10.7	6.58

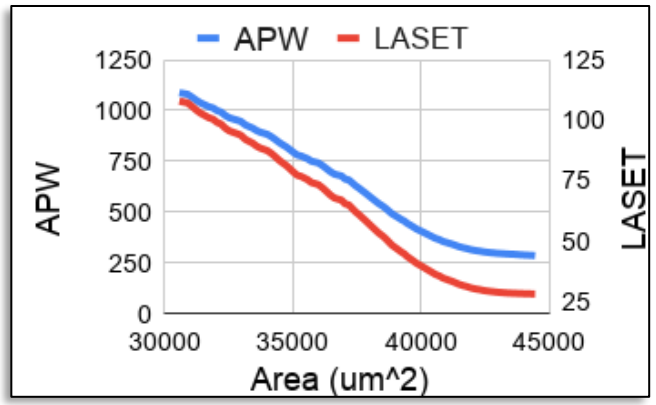
-66%
APW reduction at 50% Filtering

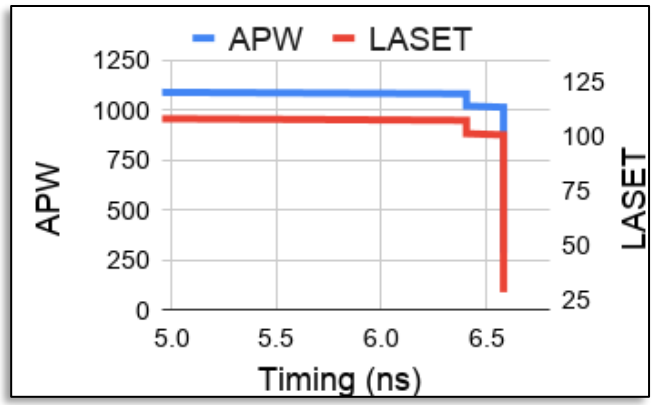
+33%
Area overhead at 50% Filtering

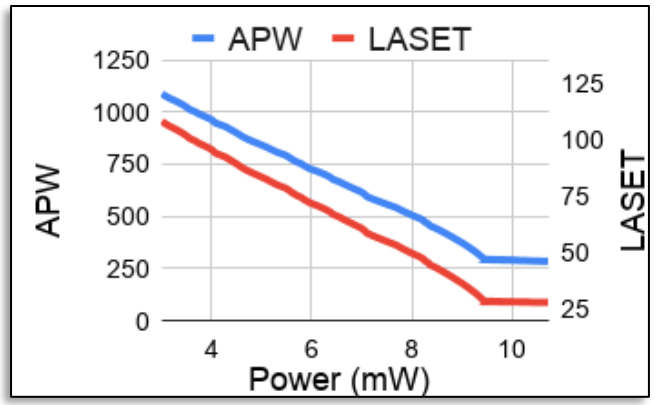
+33%
Max timing penalty (100%)

+257%
Power increase at full Filtering

SET Optimization Curves







Key Takeaways

1. Gate-level post-P&R methodology combining STA-based SET propagation (UPSET) with formal Boolean difference SEU observability analysis.

2. WFI metric unifies fault observability and arithmetic bit-position significance: A single score directly reflecting expected numerical damage per flip-flop.

3. 50% TMR → 66.2% DWFI reduction, 31.8% area overhead. 50% SET filters → 66.3% APW reduction. Both achieved at the same coverage level.

4. TMR: minimal timing impact (≤4.8%), predictable scaling, ideal for tight-timing designs. SET filters: fixed +32.7% timing penalty, saturation above 75%.

5. Combined 50% TMR + 50% SET filters is the optimal operating point: >66% reduction in both SEU and SET vulnerability simultaneously.

Thank you!

Contact Details:
chnikolaos@uth.gr
twinrelect@gmail.com

Websites:
twin-relect.uth.gr

SCAN ME

TWINRELECT
TWINNING FOR EXCELLENCE IN RELIABLE ELECTRONICS

UNIVERSITY OF THESSALY FOUNDED 1984

CAS lab
University of Thessaly
Greece

Co-funded by the European Union

UK Research and Innovation

Grant Agreement Nº 101160314