QUEEN'S UNIVERSITY BELFAST

CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

Trustworthy Hardware in the Age of AI

PROFESSOR MÁIRE O'NEILL
Regius Chair in Electronics and Computer Engineering,
Director, Centre for Secure Information Technologies (CSIT)
Director, Research Institute in Secure Hardware & Embedded Systems (RISE)

September 2025

CSIT

Delivered by Innovate UK and EPSRC

Invest Northern Ireland

Est 2009 as UK National **Innovation & Knowledge Centre** in Cyber Security Research

Mission: To couple major cyber security research breakthroughs with a unique model of innovation and commercialisation to drive economic and societal impact for the nation.

Recognition:

- Queen's Anniversary Prize in 2015 - for CSIT's work in strengthening global cyber security
- NCSC Academic Centre of Excellence in Cyber Security Research and Education
- Recognised in 2020 Royal Society report on successful 'Research & Innovation Clusters'
- Awarded Royal Irish Academy Gold Medal for Engineering Sciences in 2024

Strategic Partners:

THALES

ROLLS ROYCE

Qualcomm

QUEEN'S UNIVERSITY BELFAST | CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

Research & Innovation Model

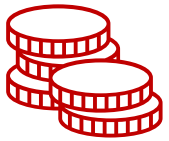
CSIT Director | **Operations Director**

Engineering → **Business Development**

Globally Leading Research Academics, RAs, PhDs	Industry Experienced Engineering Staff & Commercial Team	New Technology Products and Business Opportunities

~90 STAFF: ONE OF THE LARGEST CENTRES OF ITS KIND IN THE UK

CSIT CONTRIBUTED TO 3 IMPACT CASE STUDIES FOR REF 2021



SINCE NOV. 2022, CSIT (PHASE 3) HAS SECURED >£26.5M ADDITIONAL FUNDING

WE HAVE DELIVERED RAPID-RESPONSE PROJECTS TO >150 SMES & START-UPS ACROSS THE UK

QUEEN'S UNIVERSITY BELFAST | CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

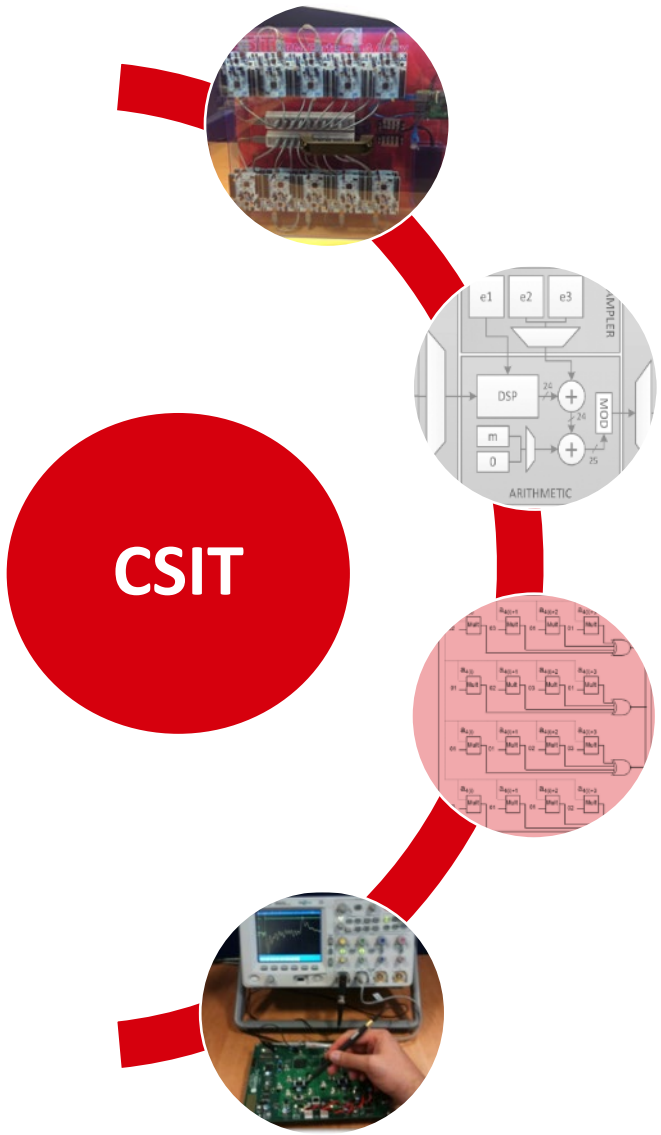
NI CYBER NI Cyber Security Ecosystem

- 2750+ cyber professionals working in
- 120+ companies
- £237m+ GVA to the local economy with an
- Average advertised salary £54,000

Target: 5000 Cyber security jobs in NI by 2030

QUEEN'S UNIVERSITY BELFAST | CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

Securing Complex Systems



Secure Connected Devices

- Trusted Hardware (PUF, HW Trojans, SCA, Approx Computing Security)
- Advanced cryptographic Architectures
  (Post-quantum, privacy preserving schemes, Homomorphic Encryption)

Networked Security Systems & ICS Security

- Cyber resilience, cloud security
- SDN-NFV security
- Malware detection, prevention and mitigation
- Resilience in ICS
- New forms of attacks/countermeasures for PLCs & legacy equipment in smart grids
- Digital Twins for IT-OT Security
- Security of safety-critical systems

Security Intelligence

- AI for Cybersecurity, Cybersecurity for AI
- Security & Assurance for Finance
- Video analytics for cyber-physical systems
- Autonomous Security

QUEEN'S UNIVERSITY BELFAST | CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

RESEARCH INSTITUTE FOR SECURE HARDWARE & EMBEDDED SYSTEMS

QUEEN'S UNIVERSITY BELFAST | CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

EPSRC Engineering and Physical Sciences Research Council

in association with National Cyber Security Centre

Aims

Hosted by CSIT since 2017

Conducts and supports **research in hardware & embedded systems security**

Projects

Affiliated projects at **Queen's University Belfast** and the **Universities of Manchester, Southampton & Birmingham, Cambridge, Surrey**

Community

Aims to grow UK Hardware Security **community** by bringing academia & industry together and facilitating **networking opportunities** nationally & internationally

Events

Summer Schools, Training roadshows, **International workshops**, UG competitions, **Impact competitions**, Innovation support

Policy

Supports **policy issues** relevant to RISE - working closely with the UK Department for Science Innovation and Technology (DSIT)

Trustworthy Hardware in the Age of AI

CSIT
CENTRE
FOR SECURE
INFORMATION
TECHNOLOGIES

China summons Nvidia over 'serious security issues' with chips

Meeting comes as US semiconductor giant seeks to revive sales in the country

The Cyberspace Administration of China requested Nvidia explain problems associated with the company's H20 chips, which were designed for China © Tyrone Siu/Reuters

Source: Financial Times, 31 July 2025

AUGUST 22, 2025 The GIST

Nvidia chief says H20 chip shipments to China not a security concern

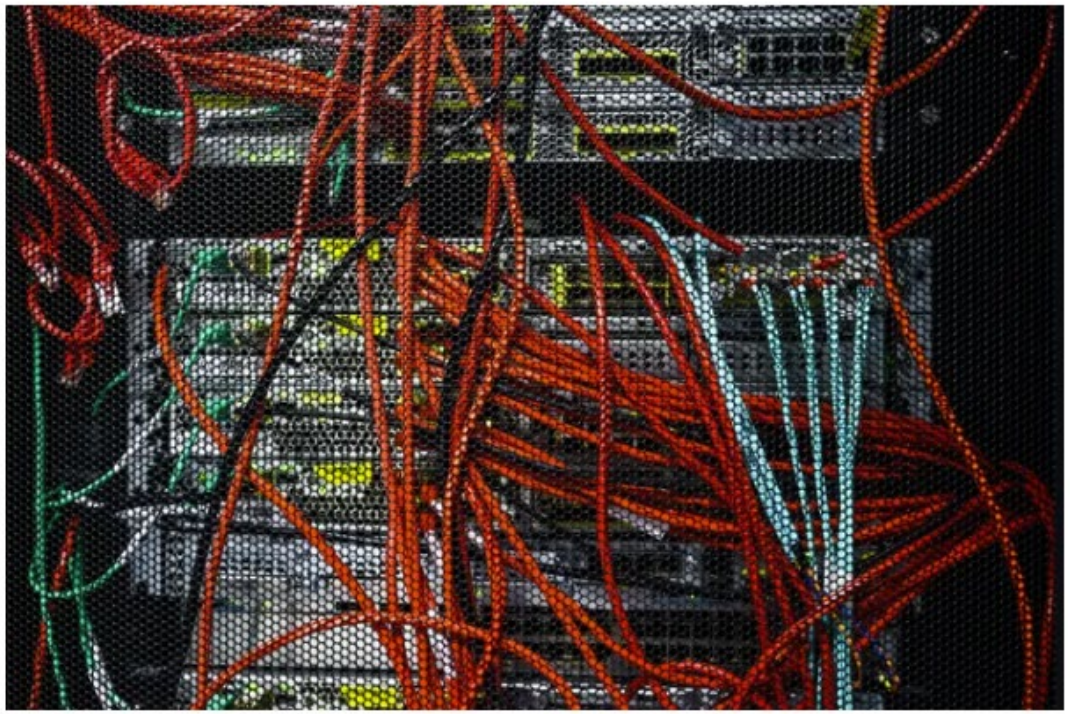
edited by Alexander Pol Editors' notes

NVIIDIA CEO – Jensen Huang - Source: https://techxplore.com/

CEO faces $15 million bill and prison time for counterfeit Cisco hardware scam

News By Ross Kelly published 7 June 2023

Onur Aksoy could face a lengthy prison sentence for his involvement in the counterfeit scheme



(Image credit: Getty Images)

Source: https://www.itpro.com/

CEO gets 6 years for selling counterfeit equipment that ended up on fighter jets

Updated: May. 03, 2024, 7:23 a.m. | Published: May. 02, 2024, 5:49 p.m.

By Matthew Enuco | NJ Advance Media for NJ.com

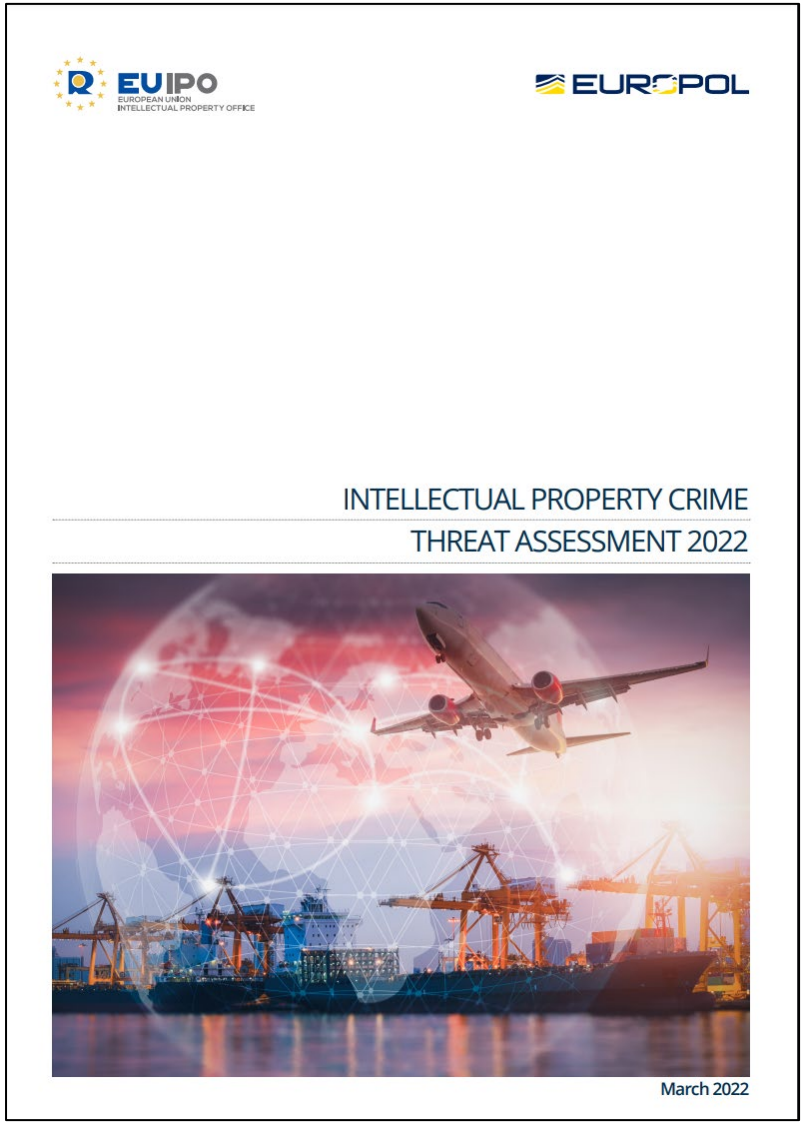
The CEO of dozens of companies, including some in New Jersey, was sentenced to over six years in prison after pleading guilty last summer to selling hundreds of millions of dollars worth of counterfeit networking equipment that made its way into sensitive places including U.S. fighter jets, federal authorities said Thursday.

Source: https://www.nj.com/

Counterfeit devices on the rise

Counterfeiters have been exploiting the global supply shortage in semiconductor chips and current geopolitical climate.

Hardware-based attacks are major security threats to military, medical, government, transportation, and other critical and embedded systems applications



Need for Supply Chain Security

Globalisation of supply chains - use of overseas foundries, third party IP, third party test facilities

Supply chains susceptible to a range of hardware-based security threats:

- Hardware Trojans
- IP piracy
- IC overproduction or recycling
- reverse engineering
- Counterfeiting – devices could host malicious software, firmware or hardware
- Side-channel attacks

Chiplet Security

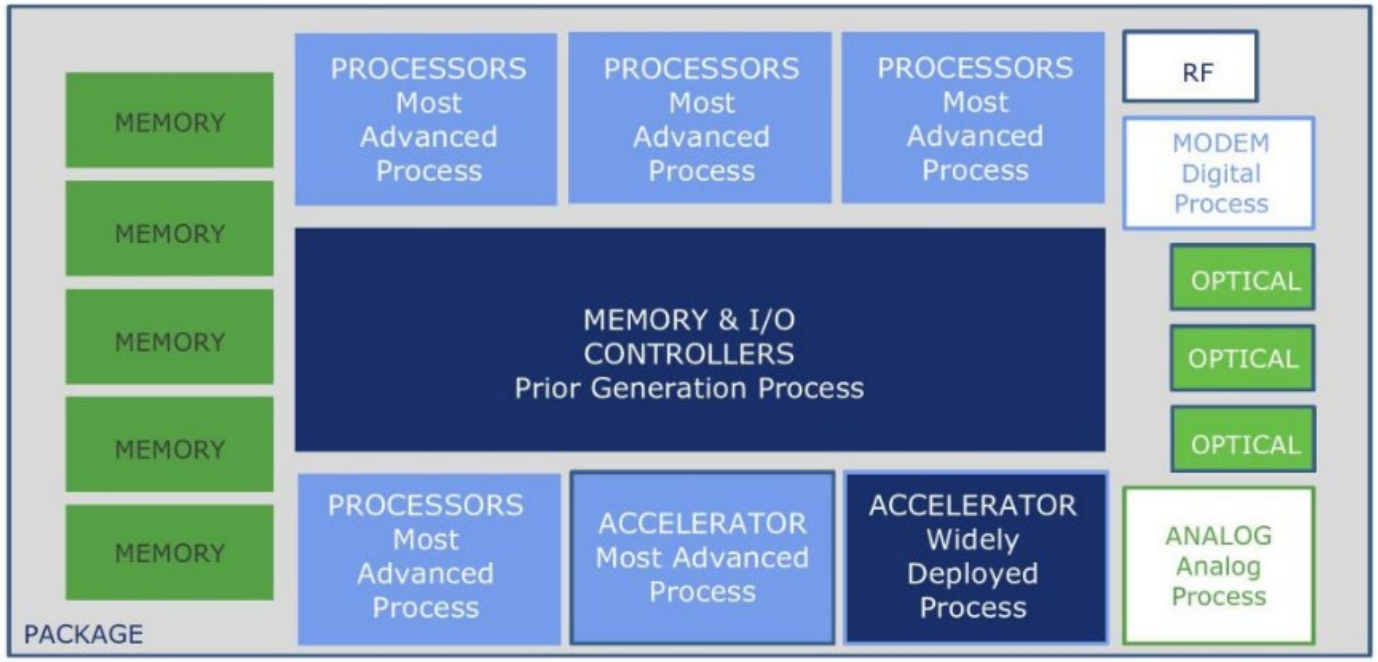
Popularity of chiplets growing

Susceptible to the same hardware-based security threats in addition to:

- chip-to-chip interface vulnerable to man-in-the-middle attacks

- Increased susceptibility to hardware trojans

Source: UCIe



Heterogeneous chiplet integration is the future of the semiconductor industry.

AI and Hardware Security?

Machine Learning (ML) has long been used in network and system security

- intrusion detection, anomaly detection, malware analysis (early 90s)

ML and DL in Hardware Security - significant advantages and implications

- ML and Side Channel Analysis (since ~2010)
- ML and Physical Unclonable Functions (PUFs) (since ~2010)
- ML and Hardware Trojan Detection (since ~2016)

CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

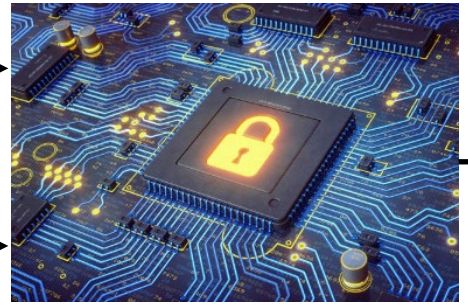
ML and Side Channel Analysis

CSIT
CENTRE
FOR SECURE
INFORMATION
TECHNOLOGIES

Side Channel Analysis (SCA)

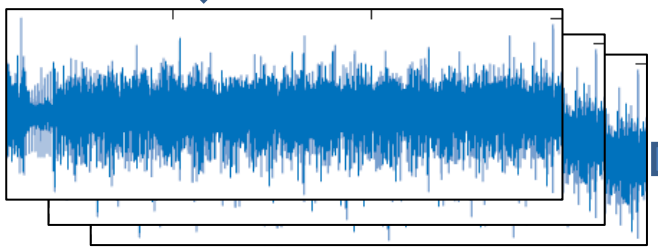
Key

Data

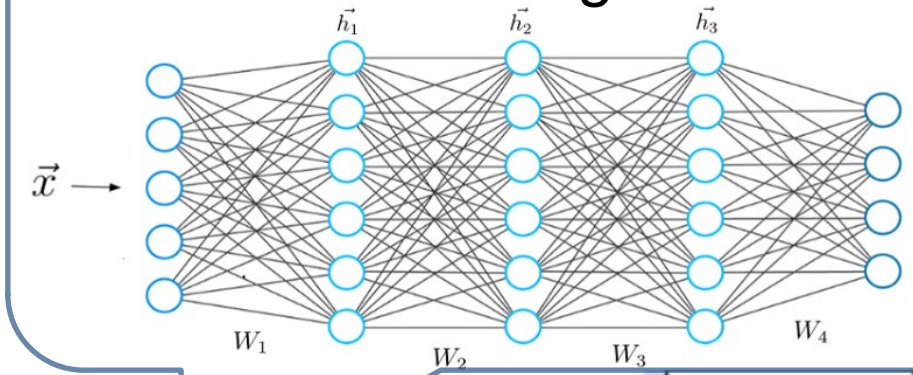


Ciphertext / plaintext

Power consumption / EM



Machine learning based SCA



No implementation knowledge required

Classify

Group "0"

Group "1"

In case incorrect key is assumed

Differential waveform

peak

In case correct key is assumed

Sensitive data

SCA-based attacks like DPA and CPA are well known since 1996

QUEEN'S UNIVERSITY BELFAST | CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

SCA Countermeasures against DPA/CPA

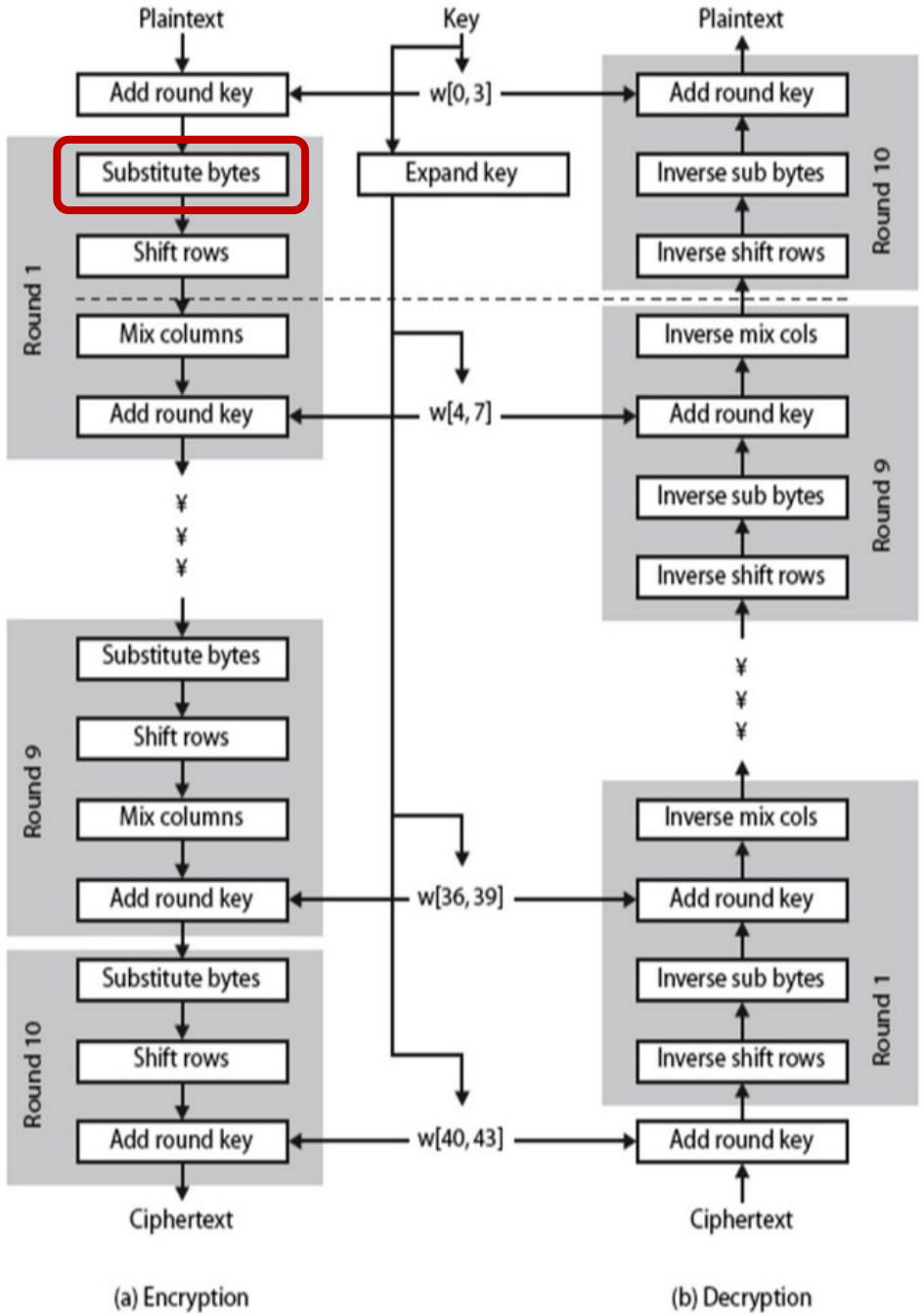
Masking

- a random mask is generated to conceal intermediate values, removing the correlation between the measurements and the secret data

Hiding

- aim is to make measurements look random or constant
- decreases the SNR only
- Timing (insert dummy operations, shuffling …)
- Amplitude (filters, pipelining …)

* ML approaches capable of **bypassing countermeasures** against DPA & CPA



Leakage of AES implementation

- Advanced Encryption Standard (AES) is safe in theory, but it is vulnerable under SCA.
- Non-linear and sensitive operation Sbox works with 8-bit sub-byte key.
- Sbox leaks hypothesis key via the relationship between the output value or its hamming weight, and side-channel information.
- The leakage can be trained using machine learning

Evaluated AES implementation with SCA countermeasure

- Used ANSSI SCA Database (ASCAD) – benchmarking reference for SCA community
- AES implementation on 8-bit AVR ATMega 8515 microprocessor
- Two **masks** are used for
  - Plaintext $\overline{p_i} = p_i \oplus m_i$
  - SBox $\overline{SBox(x)} = SBox(x \oplus m_{i.in}) \oplus m_{i.out}$

A-T. Hoang, N. Hanley, M.O'Neill, **Plaintext: A Missing Feature for Enhancing the Power of Deep Learning in Side-Channel Analysis?** IACR Transactions on Cryptographic Hardware and Embedded Systems (TCHES), 2020(4), 49-85

A-T. Hoang, N. Hanley, A. Khalid, D. Kundi, M.O'Neill, **Stacked Ensemble Model for Enhancing the DL based SCA.** 19th International Conference on Security and Cryptography, SECRYPT 2022, Lisbon, Portugal, July 11-13, 2022, pages 59–68, 2022

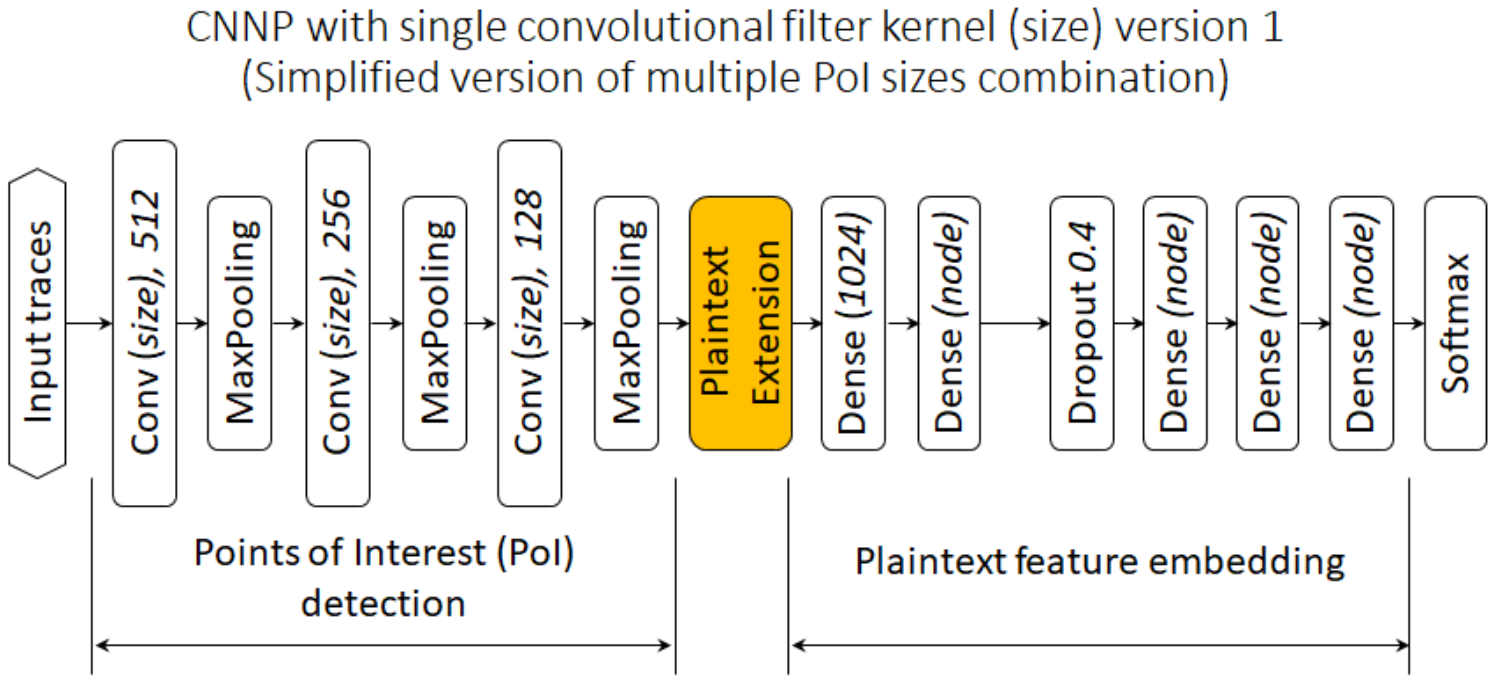
Attacker's knowledge & experimental conditions

- Assumption about attacker:
  - Knows plaintext / ciphertext
  - Aware of SCA countermeasure but not aware of the detailed design and random mask value
  - Can profile keys on the implementation

- Hypothesis keys are ranked using Maximum likelihood score

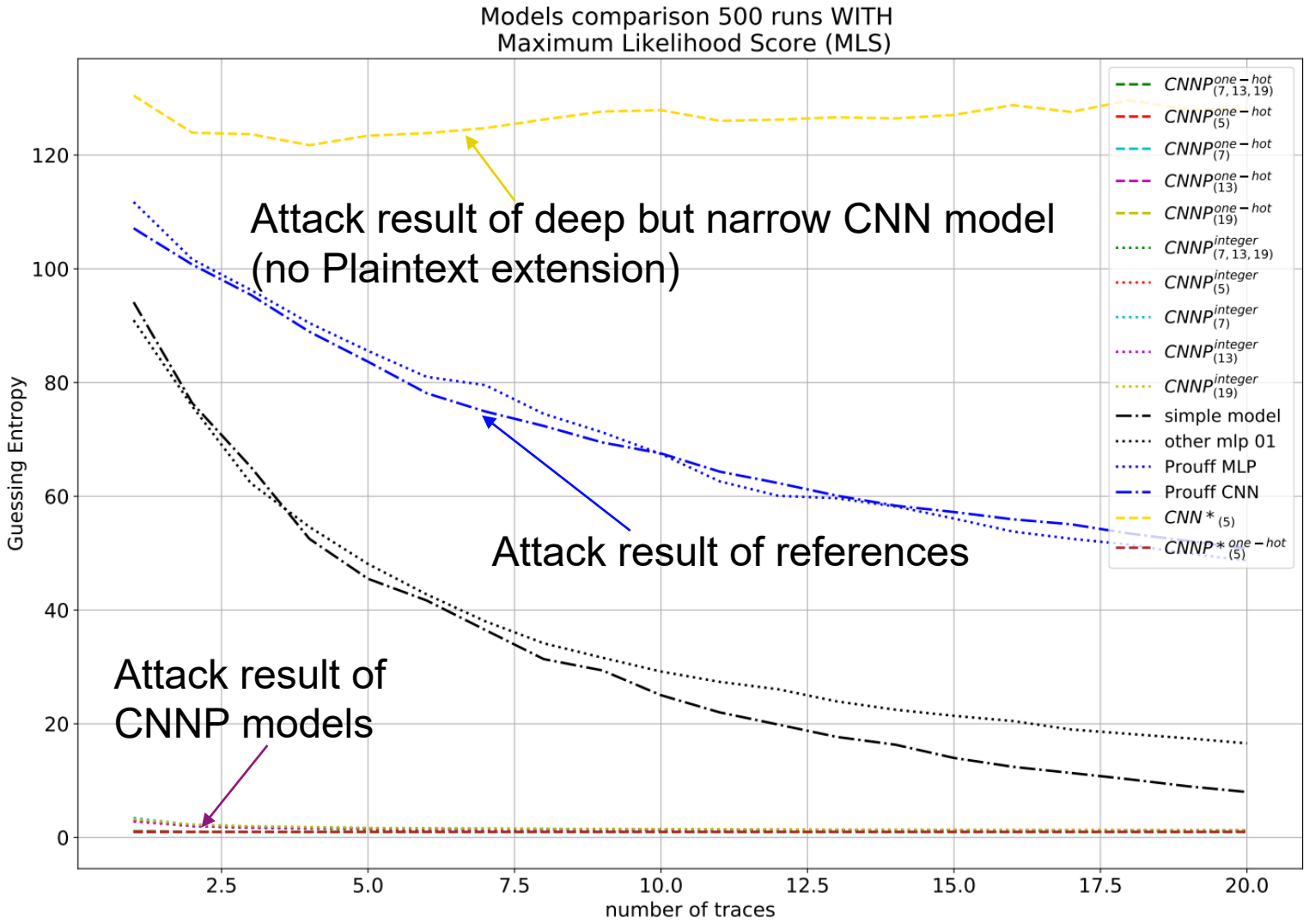
CNN with Plaintext extension (CNNP)

- Three convolutional layers
- The number of convolutional filters reduces from 512 to 128
- Maxpooling is used for feature finding
- Feature map extended with Plaintext features



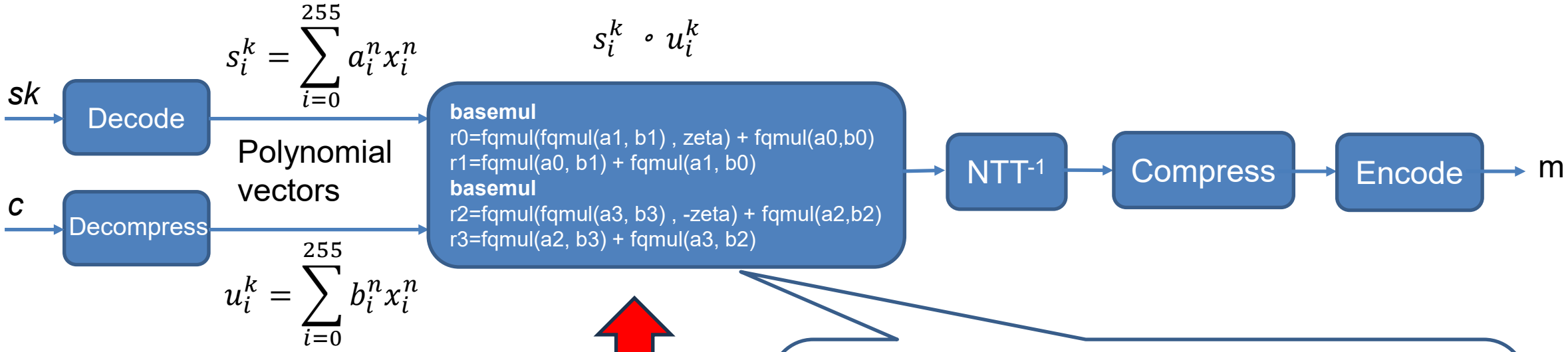
- Five fully-connected layers are used to compile the features extracted from the previous layers
- Over-fitting is prevented by using dropout

Evaluation of CNNP models on ASCAD fixed key dataset



- CNNP model can reveal the secret key within 2 traces
- CNNP models relies on the bijection $S[(.) \oplus K]$ to reveal K without using traces

Evaluated Kyber - Post Quantum Cryptography Decryption Implementation



- This ***polynomial multiplication*** operation receives 4 coefficients from **the key** and 4 others from ciphertext
- Each coefficient of **the key** is ***point-wise multiplied*** with two coefficients of ciphertext

For Kyber512, $k = 2$

sk – 768 bytes

c – 768 bytes

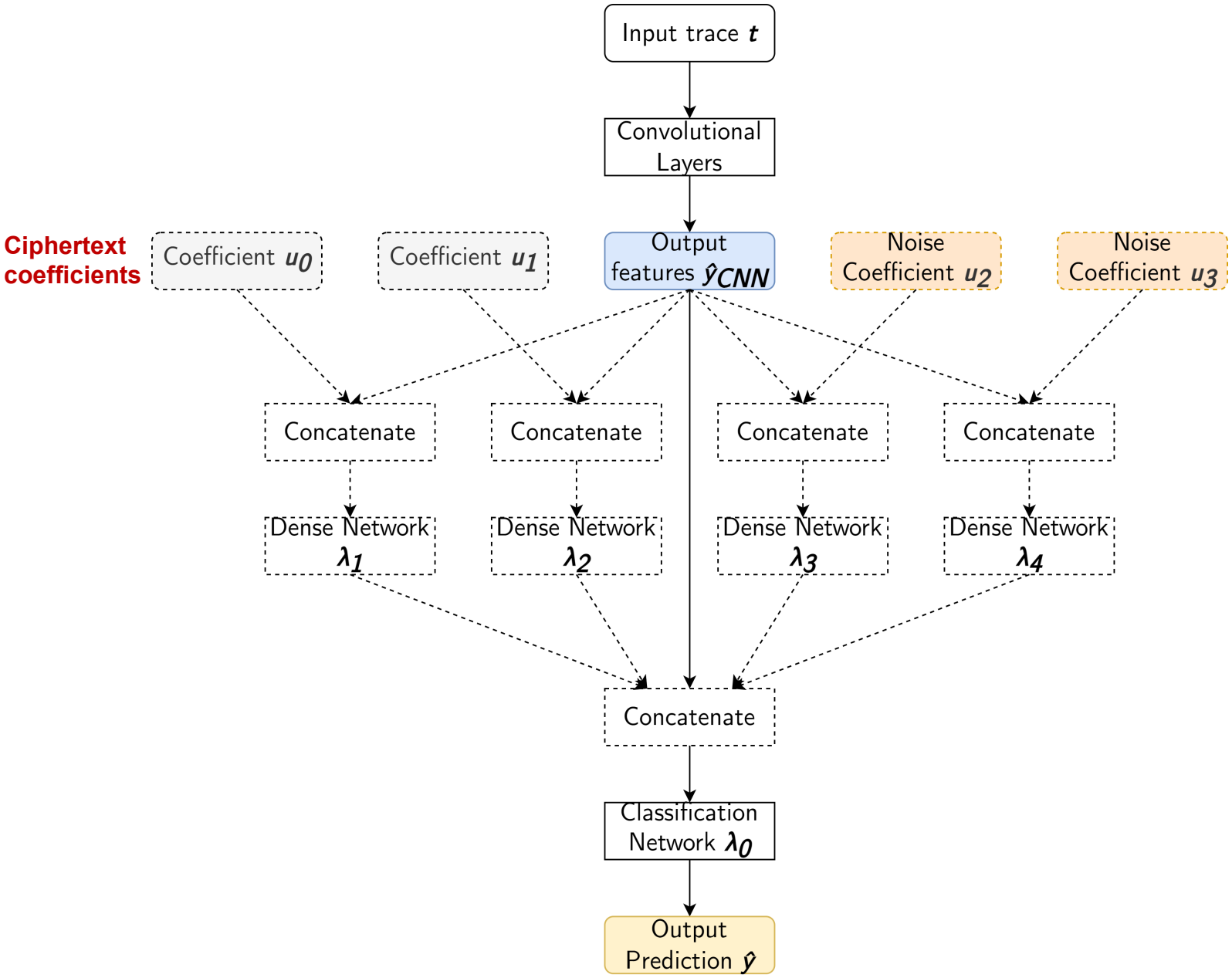
s_i^n - 512 12-bit coefficients

u_i^n - 512 12-bit coefficients

m – 64 bytes

PQM4 library: https://github.com/mupq/pqm4

CNN with Known Ciphertext Model

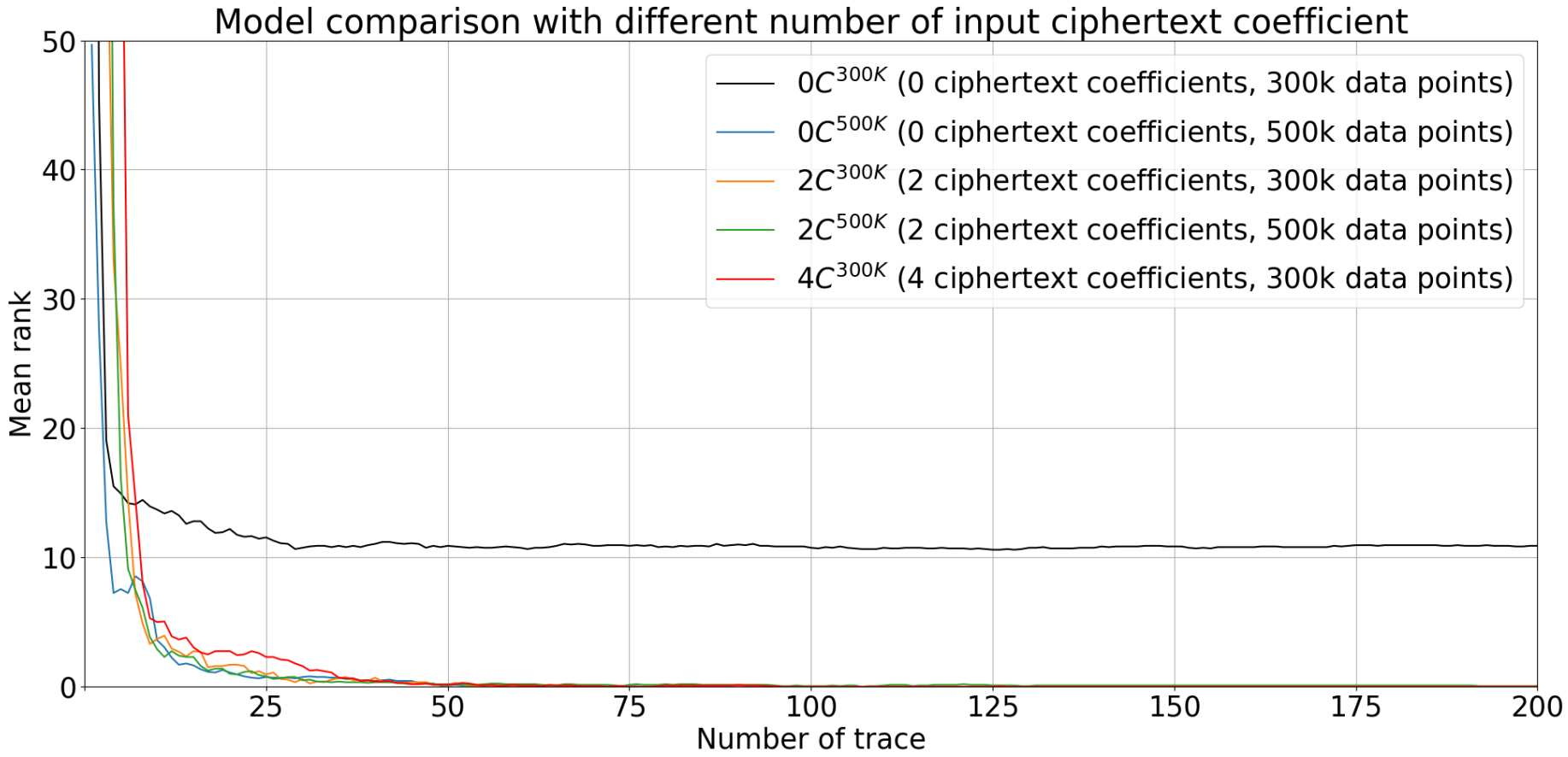


- Convolutional filter kernel size 3
- MaxPooling is used for local point of interest selection
- Convolutional layers have 64, 128, 256 and 512 filters
- Five fully connected layers of 1024 and 512 neurons each
- Activation function: ReLu

- First CNN-SCA attack against Kyber (ML-KEM)

QUEEN'S UNIVERSITY BELFAST | CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

CNN-based SCA with Ciphertext Knowledge



Hoang, A-T, Kennaway, M, Pham, T.D, Khalid, A, Rafferty, C, O'Neill, M., **Deep Learning Enhanced Side Channel Analysis on CRYSTALS-Kyber,** 25th International Symposium on Quality Electronic Design (ISQED, 2024)

QUEEN'S UNIVERSITY BELFAST | CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

Countermeasures against ML and DL-based SCA attacks?

ML-based countermeasures can be used to thwart ML-based attacks!

- Countermeasures based on adversarial attacks
  - add adversarial perturbations to the crypto implementation

- Reinforcement learning approach to construct low-cost hiding countermeasure combinations
  - finds the best combination of countermeasures within a specific budget

ML and Physical Unclonable Functions (PUFs)

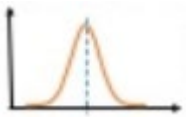
CSIT
CENTRE FOR SECURE INFORMATION TECHNOLOGIES

What is a PUF?

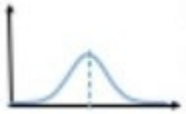
A PUF (Physical Unclonable Function) is a digital circuit that uses manufacturing process variations to generate a unique digital fingerprint.

Process Variations

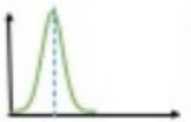
Oxide



Doping



Feature size



PUF

000000000000

70db46a1208fe753

29d3a6720fc43805

39e01f6290ac89d0

No two chips should give the same response when supplied with the same challenge.

CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

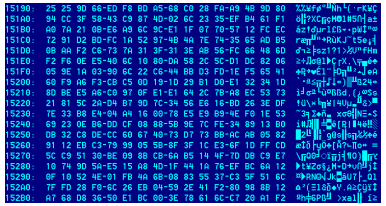
PUF Applications

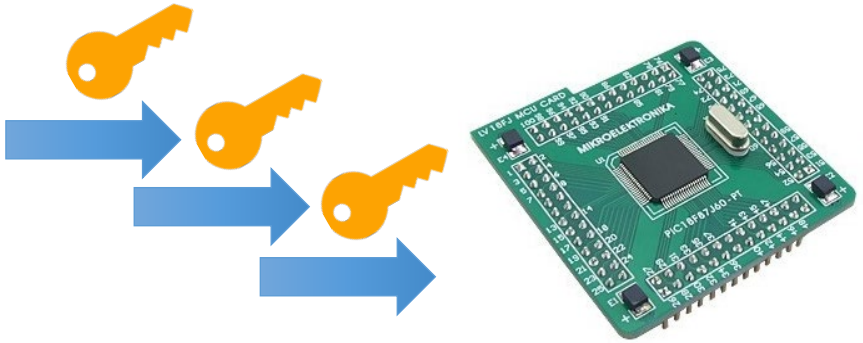
Anti-cloning/Anti-tamper

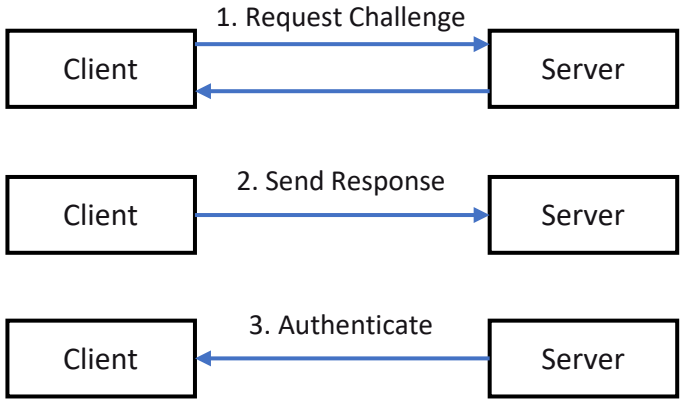
Key generation/Memoryless key storage

Lightweight device authentication and unique identification

Data Provenance/Incident Tracing

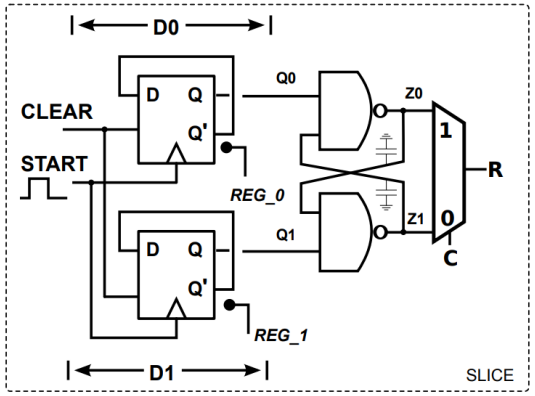






PUF in Practice

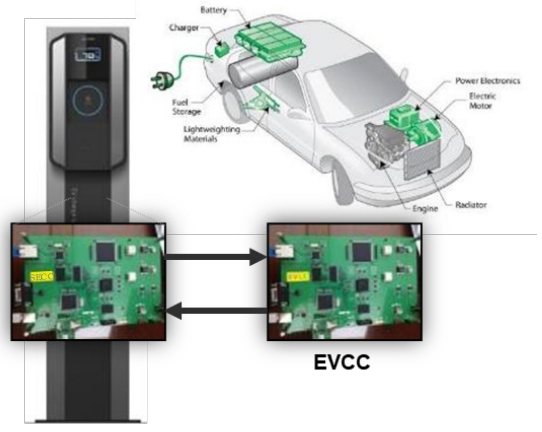
PUF designs



PoC Demonstrators

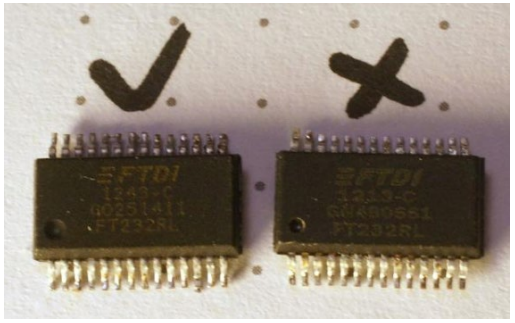
Applications

EV Charging system security



Smart meter security

Anti-Counterfeiting



Collaboration/Co-creation

ROLLS ROYCE
THALES
LG CNS
Fraunhofer AISEC
IT R&D Global Leader ETRI

PUF in Practice

AMD

Versal Adaptive SoC Technical Reference Manual (AM011)

Physically Unclonable Function

The AMD Versal™ device contains a physically unclonable function (PUF). The PUF creates a signature (or fingerprint) of each device that is unique to that device. Its value is unknown to AMD and device user by design and can be used as a key encryption key (KEK). This KEK is 256 bits in length with 256 bits of entropy and is used to encrypt the users red key allowing its storage in black (encrypted) form. The black key can be stored in either eFUSEs, BBRAM, or external storage.
PUF outputs a user accessible unique ID that is cryptographically isolated from the PUF KEK itself despite using the same entropy source. While unique to each device, it is not considered a *secret* and does not have the same access protections as the KEK itself.
For additional details, see the *Versal Adaptive SoC Security Manual* (UG1508).
This manual requires an active NDA to download from the Design Security Lounge.

Technology Brief

Security

altera
An Intel Company

Security: Protecting Your IP with Agilex™ 5 FPGAs

Author
Mark Frost
Product Marketing Manager
Altera Corporation

Table of Contents

The Need for Security 1
Overview of Security Features in Agilex 5 FPGAs 1
Secure Device Manager 2
Protecting Your IP 3
Bitstream Authentication 3
Bitstream Encryption 3
Platform Attestation 3
Cryptographic Primitive Services 4
Conclusion 4

The Need for Security

Security is a foundational requirement in many applications and vertical markets, such as industrial, broadcast, communications, and in the data center. To address these needs, the Agilex™ 5 FPGA family has been designed from the ground up with security in mind, supporting many of the latest security standards.

Why should companies care about security, and how does it apply to FPGAs? FPGA designers have long relied on the concept of "security through obscurity," depending on the reconfigurable fabric and unpublished bit stream formats to keep their intellectual property (IP) secure. But with the proliferation of FPGAs in all markets, and with the ever-increasing number of cyberattacks, particularly in the areas of IP theft and espionage, implementing security within an FPGA design has never been more relevant.

These trends make the need for adopting dedicated security within the FPGA more pressing:

- An increased reliance on third-party contract manufacturers, often a long way (in both distance and in the supply chain) from where the FPGA design has been created, which can lead to concerns about IP theft, cloning, and overbuilding
- Increased remote access in an interconnected world is a double-edged sword, often opening products up to cyberattack
- As the use cases for FPGAs continue to proliferate, security researchers and cyber criminals are exposing device vulnerabilities
- It's increasingly important to have assurance that all connected devices are running current designs and software
- Corporate mandates are increasingly focused on device and application security

The security features built into Agilex 5 FPGAs can help to address these concerns.

Overview of Security Features in Agilex 5 FPGAs

Agilex 5 FPGAs support the following device-level security features:

Security Features	Security Benefit
Authentication	Helps ensure the origin, integrity, and validity of bitstreams
Bitstream Encryption	Helps protect design IP from cloning or reverse engineering
Physically Unclonable Function (PUF)	Device-unique, undiscoverable value used for device identity and additional key storage protection
Cryptographic Services	Secure Device Manager provides access to hardened cryptographic IP to the FPGA logic and HPS software
Attestation	Industry standard DICE and SPDM protocols provide a traceable device identity and measurements of device state and configuration
Secure Provisioning	Use SPDM Secure Channels to send bitstream encryption and cryptographic service keys in untrusted environments
Physical Anti-tamper	Detection of environmental changes and responses to help resist physical intrusion

Table 1. Overview of Agilex 5 device security features

Product Overview

SYNOPSYS
Silicon to Software

Physically Unclonable Function (PUF) Solution for ARC EM Processors

Highlights

- Secure and reliable PUF-based crypto key generation
- Physical fingerprint and entropy extraction from embedded SRAM
- Pure firmware implementation leveraging Synopsys SecureShield technology
- Optional high-performance implementation with Synopsys ARC CryptoPack acceleration
- Chip identification based on Fuzzy Identifier

Target Applications

- IoT
- Wearables
- Mobile
- Microcontrollers
- Sensors

Technology

- TSMC, UMC, Intel, Samsung
- 180nm, 150nm, 130nm, 90nm, 65nm, 45nm, 40nm, 28nm, 16nm, 14nm

PUF for Integrated Circuits

Tiny variations in a semiconductor manufacturing process make each transistor and each piece of silicon unique. These variations are random and uncontrollable, so it is impossible to make an exact clone of an integrated circuit (IC), hence we refer to this as a Physically Unclonable Function or PUF. These variations can be amplified and measured with standard embedded Static Random-Access Memory (SRAM) cells and the startup behavior of on chip SRAM results in a unique pattern that is analogous to a fingerprint for the IC.

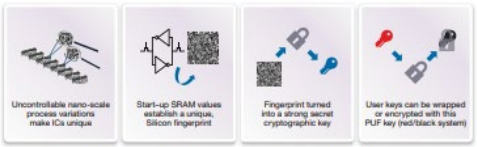


Figure 1. Flow of PUF technology used for secure key management

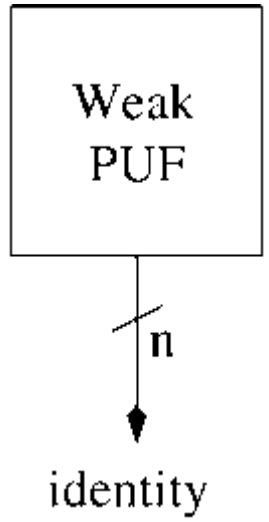
Physically Unclonable Function Solution for ARC EM Processors

The Physically Unclonable Function (PUF) solution from Intrinsic-ID is available for DesignWare® ARC® EM Processors and enables designers to extract a unique device fingerprint from standard embedded SRAM. This fingerprint can be used as a device identifier or as a cryptographic key. In the latter case, it effectively creates a secure key vault without the need to add non-volatile memory (NVM) or a dedicated security core. In combination with ARC EM Processor security options such as the Enhanced Security Package and CryptoPack, the PUF solution provides a high-performance, low-power security engine for protecting low-power IoT edge nodes such as wearables or smart home devices.

Identification with Fuzzy-ID

The startup pattern from an SRAM PUF can be used to uniquely identify a chip. Some of the bits in the pattern are unstable, so the matching has to be done using software known as the Fuzzy Identifier algorithm. This algorithm converts the unique but variable fuzzy identifier into a unique, collision-free fixed identifier comparable to a chip Identifier like the Electronic Chip ID (ECID).

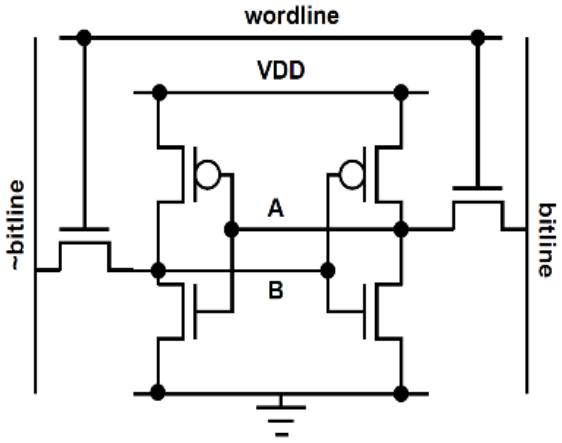
PUF Classification: Identity Vs Challenge-Response



Weak PUF / Identity PUF

- Typically have no (or one fixed) challenge
  - e.g. SRAM PUF, Butterfly PUF.
- Assumed an attacker *cannot* access the responses of "Weak" PUFs as one or few CRPs could be used to build a model of the security system
- Applications include:
  - Identity generation
  - RNG seed
  - Non volatile key storage

SRAM PUF[1]

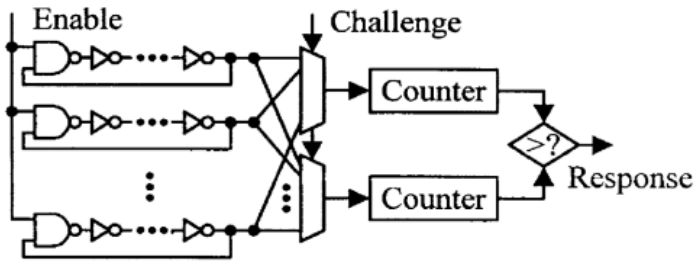


[1]G. Jorge, K. Sandeep S, S. Geert-Jan, and T. Pim. *FPGA intrinsic PUFs and their use for IP protection*. Springer, 2007.

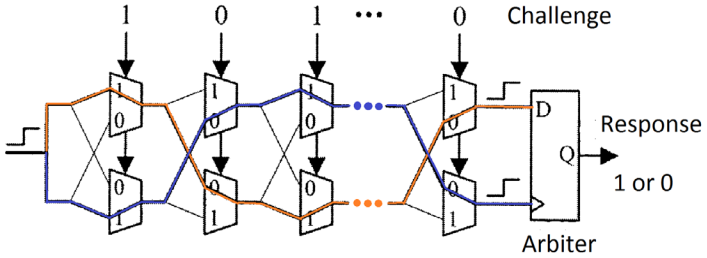
PUF Classification: Identity Vs Challenge-Response

Strong PUF/Challenge Response PUFs

- May have many possible challenge response pairs (CRPs)
  - e.g. Arbiter PUF, Ring Oscillator PUF
- With access to the CRPs, it *should be infeasible* to model the system and determine the CRPs of a strong PUF
- Applications include *challenge-response authentication*



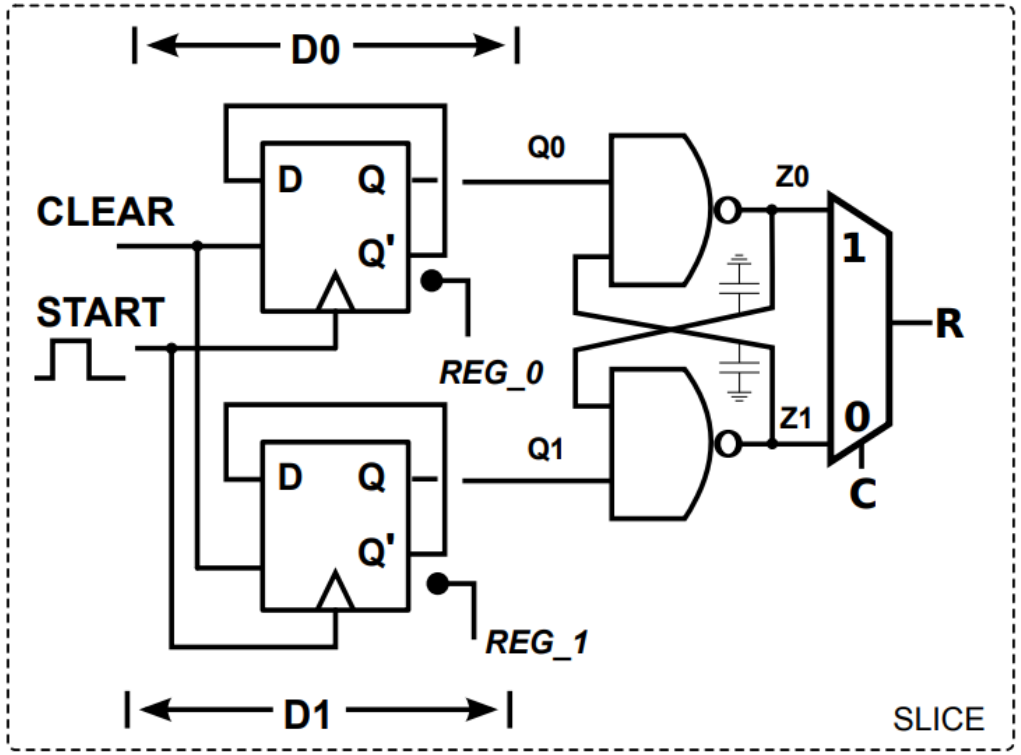
Ring Oscillator (RO) PUF[2]



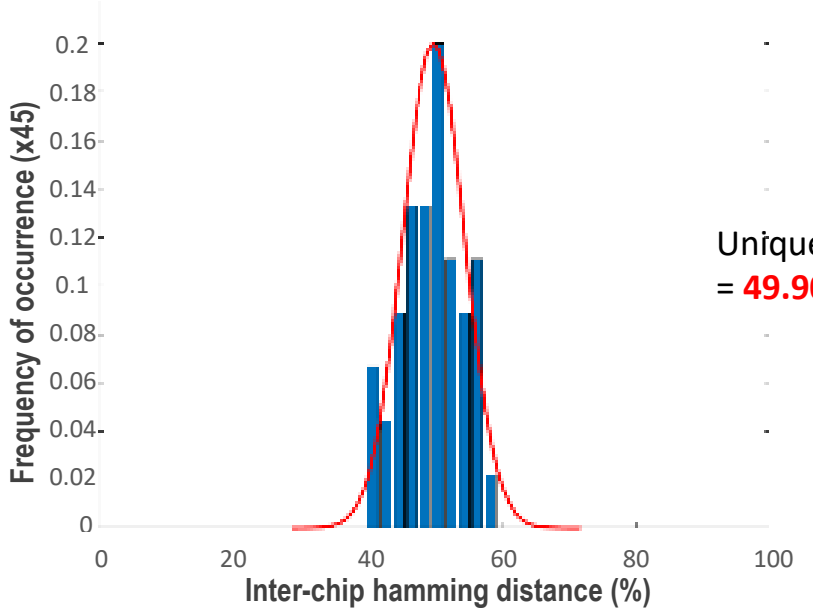
Arbiter PUF[3]

[2]G.E. Suh, S Devadas. Physical unclonable functions for device authentication and secret key generation. In Proc. 44th ACM/IEEE DAC, pp 9–14, San Diego, 2007.
[3]G. Blaise, C. Dwaine, Marten V. D., D. Srinivas. Silicon physical random functions. In Proc.9th ACM Conference on Computer and Communications Security, CCS 2002, Washington DC, US, 2002.

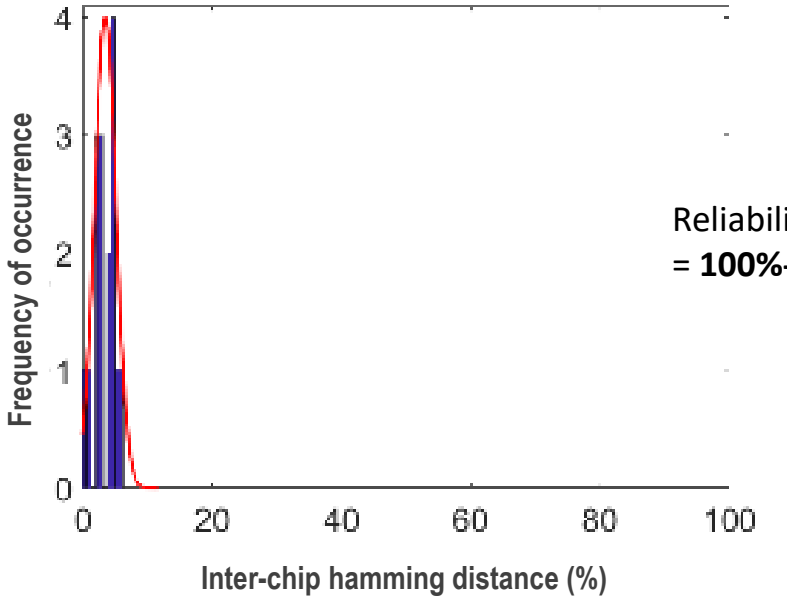
Identity PUF (PicoPUF)



1-bit PicoPUF Design



Uniqueness
= **49.90%**



Reliability
= **100%-3.47%** = **96.53%**

To generate a 128-bit response, it costs 128 slices, **8.95%** of hardware resources on a Spartan-6 (0.01% on Artix-7)

C. Gu, C.H. Chang, W. Liu, N. Hanley, J.Miskelly, M. O'Neill, *A Large Scale Comprehensive Evaluation of Single Slice Ring Oscillator and PicoPUF Bit Cells on 28nm Xilinx FPGAs*, Journal of Cryptographic Engineering, 2020.

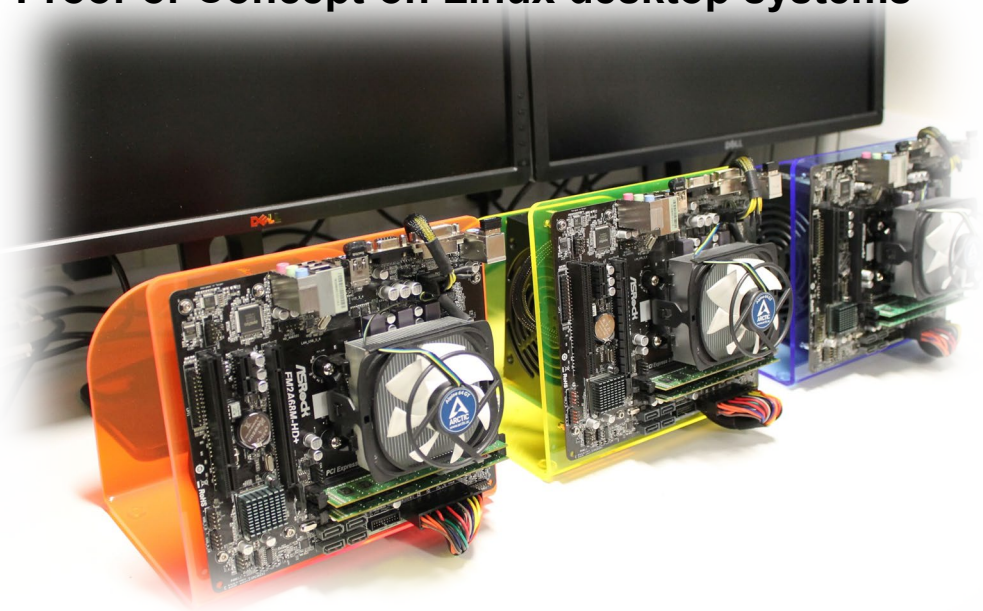
CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

DRAM Latency PUF

- DRAM cells are gated capacitors - manuf. process variation in rates of charge and discharge is entropy source
- Generate hardware-rooted ID through controlled read errors
- Challenge = Set of timings + memory locations to read
- Response = Error pattern

- Fast generation of large reliable identifiers (1.2ms per Kb)
- Near ideal uniqueness
- Highly reliable - >99% for 8Kb ID

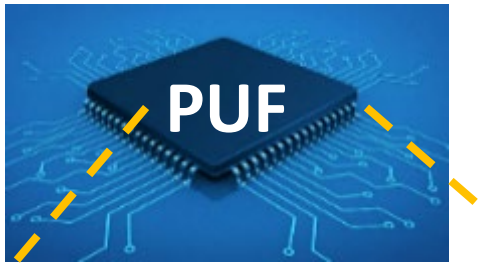
Proof-of-Concept on Linux desktop systems



Miskelly, J., & O'Neill, M, *Fast DRAM PUFs on Commodity Devices*. IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems, 3566-3576, 2020

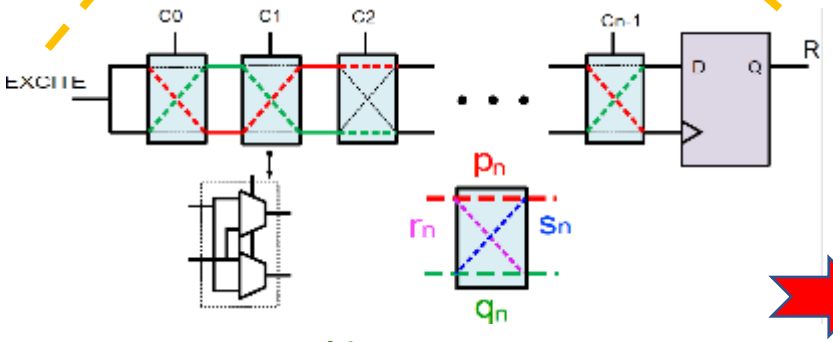
Challenge-Response PUFS are vulnerable to ML attacks

Device



Challenge

Response



Arbiter PUF

PUF CRPs

Challenge	Response
10011…110	10110…101
00110…101	00101…010
10110…001	10101…000
…	
01101…101	10001…111

Linear Function

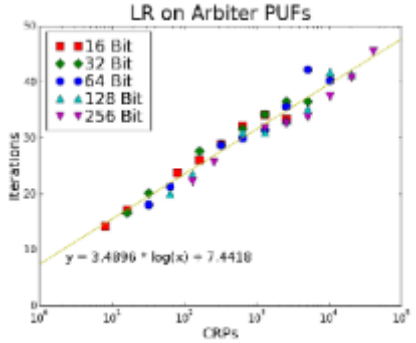
$$\Delta(n) = P \cdot \boldsymbol{\omega}^{T}$$

$$\omega_i = \alpha_i + \beta_{i-1}, 2 \le i \le n$$

$$\alpha_i = \frac{p_i - q_i - r_i + s_i}{2}$$

$$\beta_i = \frac{p_i - q_i + r_i - s_i}{2}$$

Machine Learning Algorithm



Logistic Regression (LR) on Arbiter PUFs

ML Method	Bit Length	Prediction Rate	CRPs	Training Time
LR	64	95%	640	0.01 sec
		99%	2,555	0.13 sec
		99.9%	18,050	0.60 sec
LR	128	95%	1,350	0.06 sec
		99%	5,570	0.51 sec
		99.9%	39,200	2.10 sec

LR Prediction Results

U. Rührmair, J. Sölter, F. Sehnke, X. Xu, A. Mahmoud, V. Stoyanova, G. Dror, J. Schmidhuber, W. Burleson, S. Devadas, “PUF modelling attacks on simulated and silicon data,” IEEE Trans Information Forensics and Security, vol. 8(11), pp.1876–1891, 2013.

CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

ML- Attack Resistant PUF Design Approaches

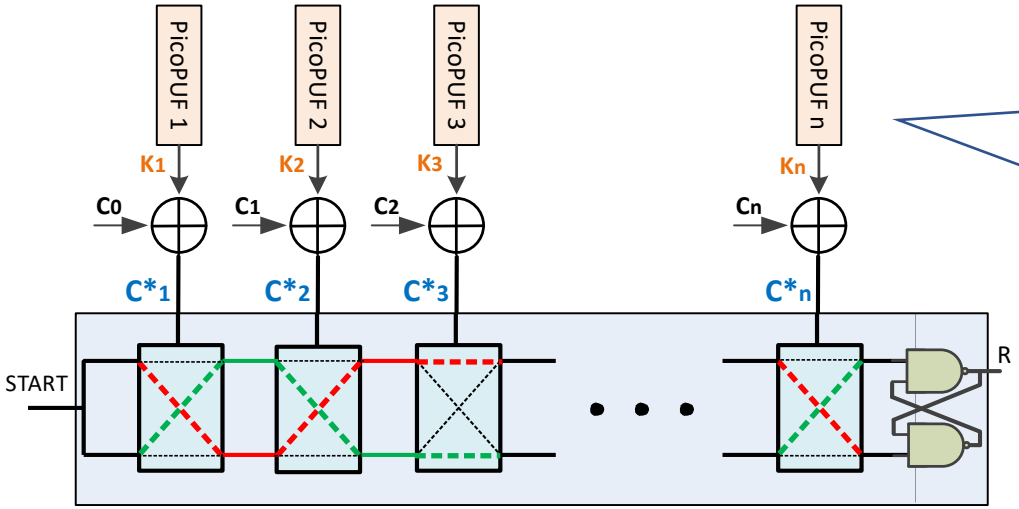
- Obfuscate the challenge/response (e.g. XOR Arbiter PUF)
  - Use a weak PUF
  - All XOR APUF shown to be susceptible to reliability based CMA-ES attacks (based on challenge-reliability pairs)

- Increase complexity of the PUF design
  - if too complex, PUF design is no longer a lightweight primitive

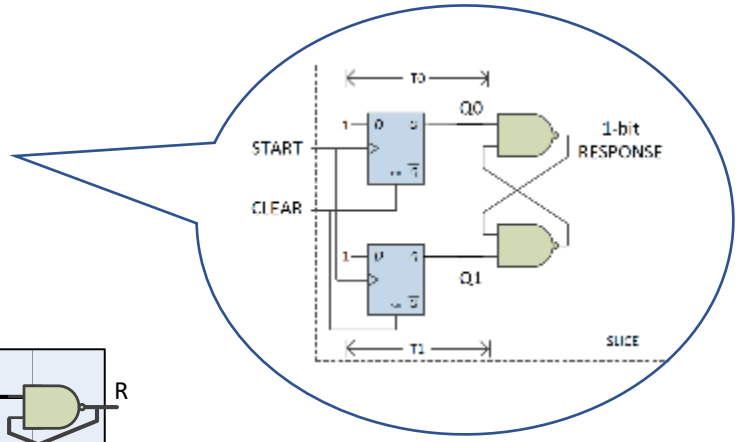
- Deception techniques

ML- Attack Resistant PUF - Challenge obfuscation

Arbiter-based multi-PUF (MPUF) design - utilises an Identity PUF to obfuscate the challenges to the Challenge/Response PUF
=> harder to model than the conventional Arbiter PUF using ML attacks.



Proposed 1-bit MPUF Design



1-bit PicoPUF Circuit Design

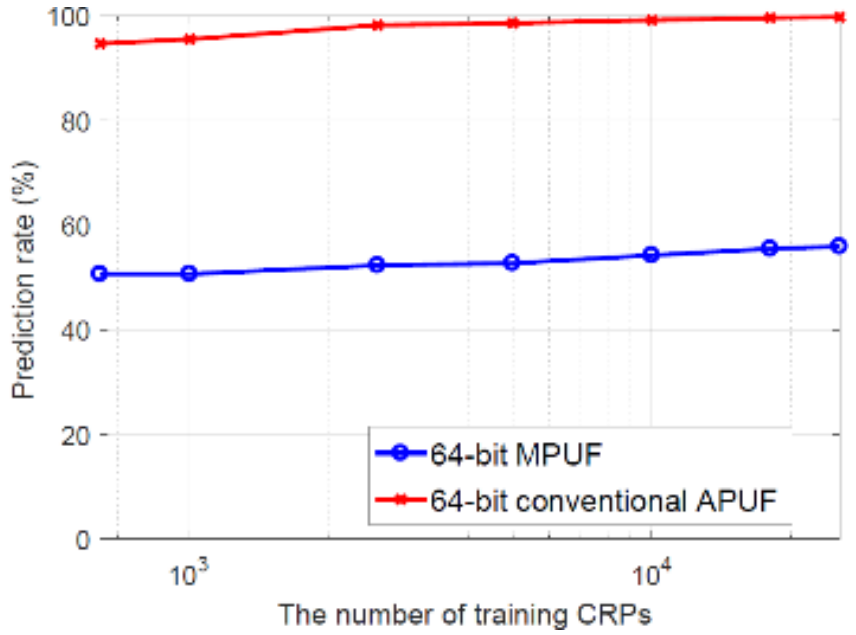
The responses of the PicoPUFs are used to mask the original challenges C_i

Q. Ma, C. Gu, N. Hanley, C. Wang, W. Liu and M. O'Neill, "*A machine learning attack resistant multi-PUF design on FPGA*," 2018 23rd Asia and South Pacific Design Automation Conference (ASP-DAC), Jeju, 2018, pp. 97-104.

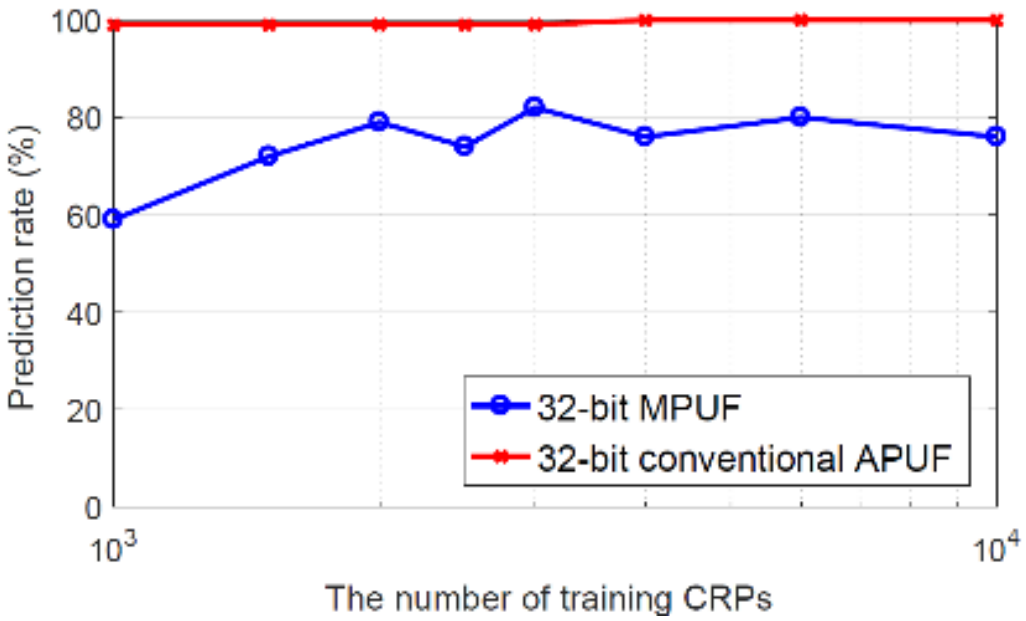
CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

ML- Attack Resistant PUF - Challenge obfuscation

Most common ML-based attacks applied to PUF:
- Logistic regression (LR)
- Covariance matrix adaptation evolution strategies (CMA-ES)



Prediction rates for conventional Arbiter- PUF and proposed MPUF designs using LR

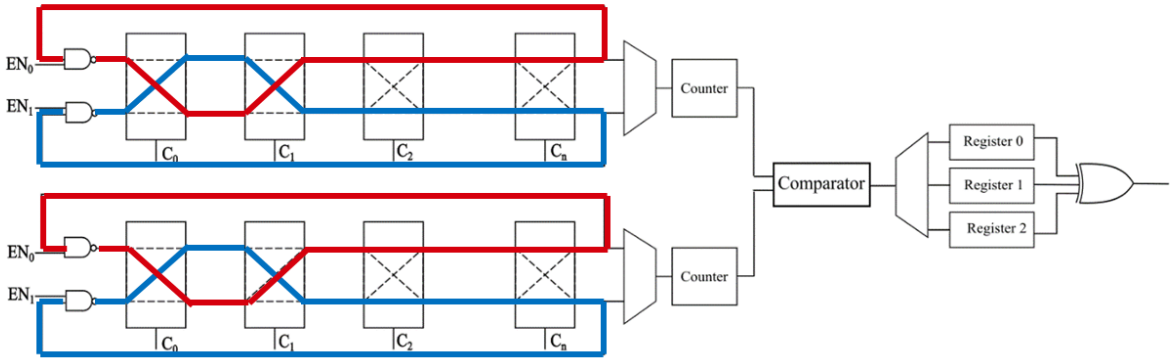


Prediction rates for conventional Arbiter- PUF and proposed MPUF designs using CMA-ES

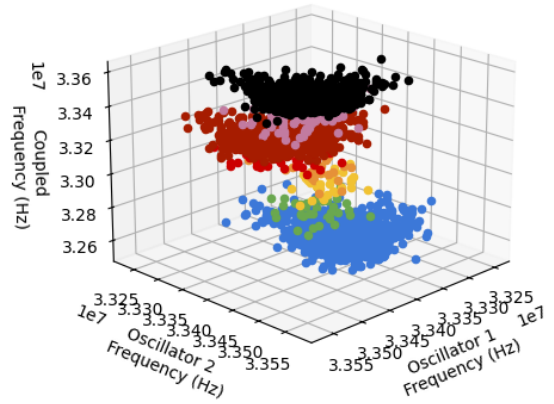
ML- Attack Resistant PUF – Increasing complexity

Mutually Coupled Configurable Ring Oscillator (CRO) PUF - uses 2 CRO PUFs placed sufficiently close to become coupled

Implemented & tested on 100 Xilinx ZYBO Z7 boards (on XC7Z010 FPGAs)



Knowing natural frequency of the independent oscillators does not necessarily allow coupled frequency to be predicted - overlap due to coupled frequencies adds to complexity



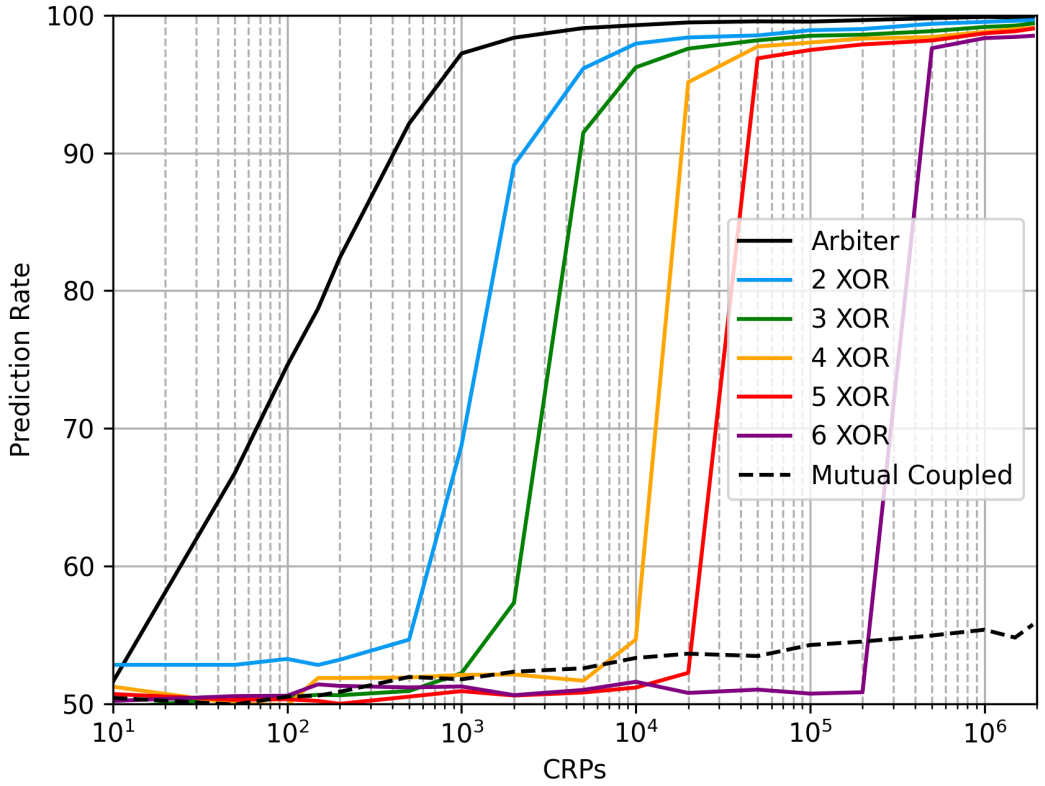
Mutual Coupled Oscillator frequency Vs Sub Oscillators frequencies

PUF testing in Temperature Chamber

CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

ML- Attack Resistant PUF – Increasing complexity

- Machine learning resistance is the measure of how many CRPs are needed to be able to predict a PUF accurately.
- This was tested using a MLP model using 3 layers each a size of (500,1000,500) with each layer using a tanh activation function.

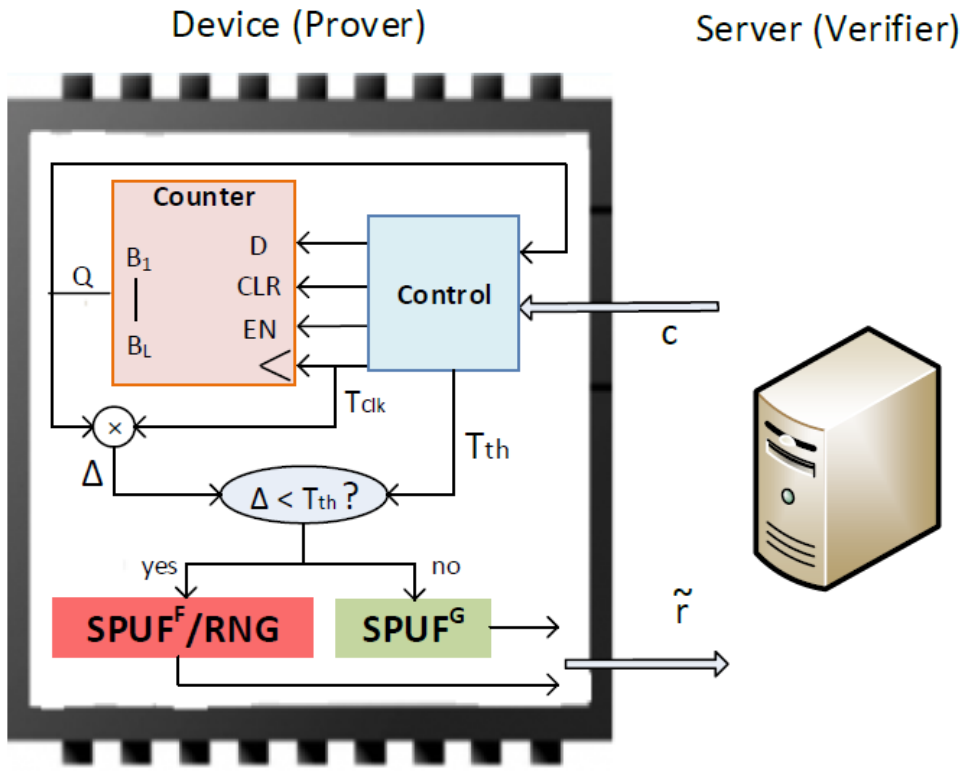


Moore, J, Miskelly, J, O'Neill, M, Gu, C, A novel FPGA mutually coupled configurable ring oscillator PUF, Asian Hardware Oriented Security and Trust Symposium (AsianHOST), 2024

CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

ML- Attack Resistant PUF – Deception Protocols

- Device detects an adversary sending continuous authentication requests
- Generate some responses from a deceptive PUF design and others generated from real PUF.
- Adversary will be deceived into deriving a fake PUF model from the collected data.
- Lightweight and do not require error-correction or sophisticated cryptographic algorithms

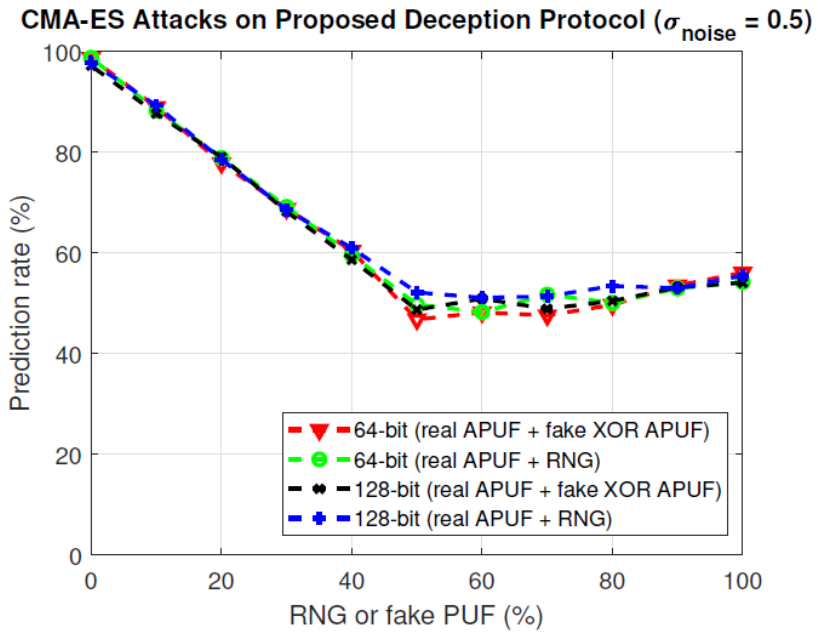


C Gu, C.H. Chang, W. Liu, S. Yu, Q. Ma, M. O'Neill, A Modeling Attack Resistant Deception Technique for Securing PUF based Authentication , 2019 Asian Hardware Oriented Security and Trust Symposium (AsianHOST)

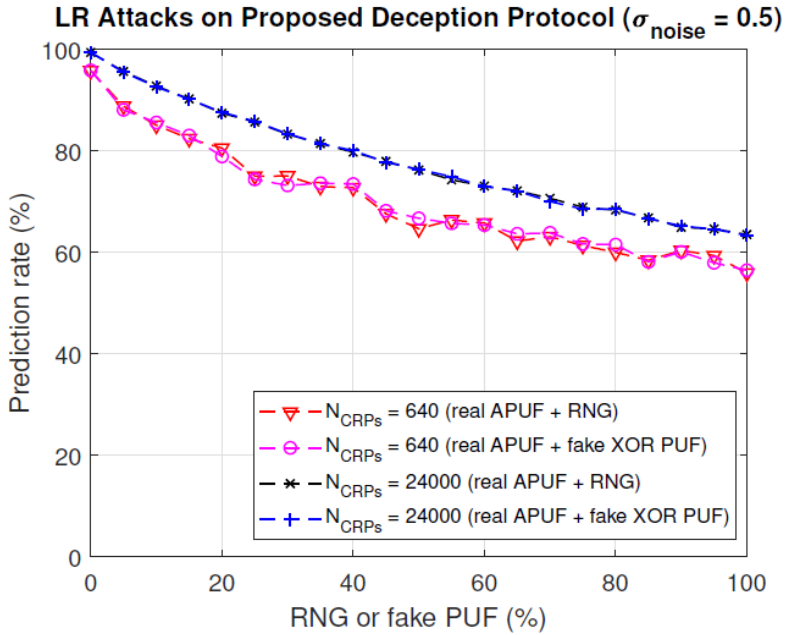
C Gu, C.H. Chang, W. Liu, S. Yu, Q. Ma, M. O'Neill, A Modeling Attack Resistant Deception Technique for Securing Lightweight-PUF based Authentication, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), 2020

CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES

ML- Attack Resistant PUF – Deception Protocol



The CMA-ES attack results for the proposed deception protocol by applying different challenge bit lengths, 64-bit and 128-bit, as well as utilizing different strategies (RNG and fake PUF).



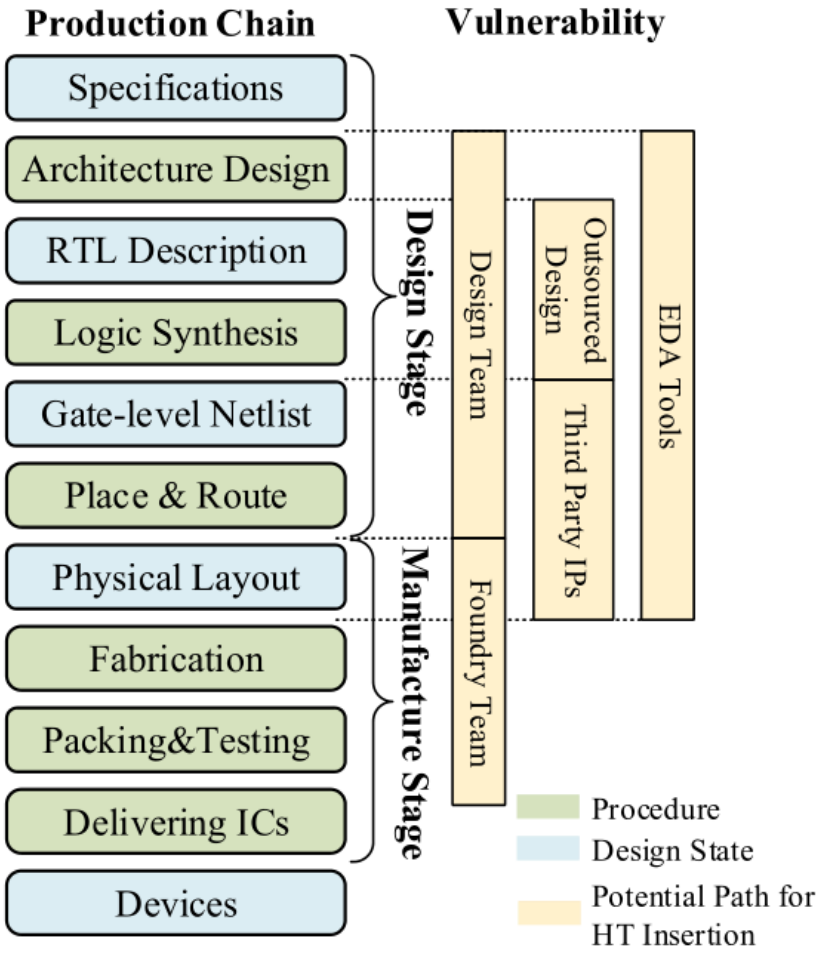
The LR attack results for the proposed deception protocol utilizing different strategies, RNG and fake PUF. The y-axis shows the achieved prediction rate of the LR attacks based on different percentages of fake information mixed with the training responses.

ML and Hardware Trojan Detection

CSIT
CENTRE
FOR SECURE
INFORMATION
TECHNOLOGIES

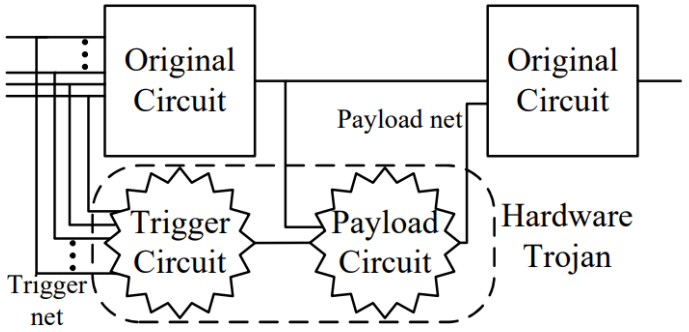
Hardware Trojan Detection

CSIT CENTRE FOR SECURE INFORMATION TECHNOLOGIES



Hardware Trojan

- Additional circuit inserted into an IC design at RTL or gate level for malicious purposes;
- Malicious modification of a circuit.
- Usually stealthy to escape verification and manufacturing test processes
- Detection is very difficult – there may be no Trojan-free reference for comparison



Previous Work –HT Feature Extraction

- Most of the previous research extract HT features by statistical analysis of netlist information.
- Most of them need knowledge driven approaches for the features selection and weight adjustment.

A SUMMARIZATION OF GATE-LEVEL HT FEATURES EXTRACTED IN PREVIOUS RESEARCH

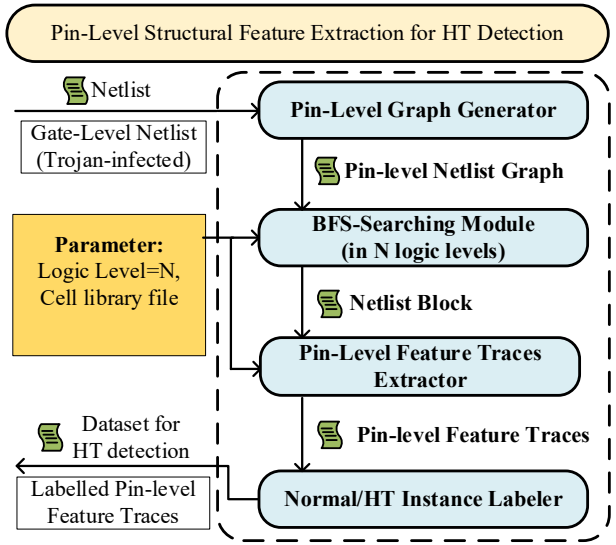
Ref.	Feature Type	Features	Detection method
[1]	statistical	controllability,observability	K-means clustering
[2]	statistical	controllability,observability	Bagged Trees
[3]	statistical	controllability, switching probability	K-means clustering
[4]	statistical	controllability,observability, number of specific cells	SVM
[5]	statistical	LGFi, FFi, FFo, PI, PO	SVM
[6]	statistical	LGFi, FFi, FFo, PI, PO	Ensemble-learning
[7]	statistical	11 numerical features	Random forest
[8]	statistical	11 numerical features	Neural Networks
[9]	structural, statistical	two-level AONN gates, number of specific paths	Score-based

[1] H. Salmani, "Cotd: Reference-free hardware Trojan detection and recovery based on controllability and observability in gate-level netlist "
[2] C. H. Kok, "Classification of Trojan nets based on scoap values using supervised learning"
[3] Y. He "Trigger identification using difference-amplified controllability and dynamic transition probability for hardware Trojan detection"
[4] X. Xie, "Hardware Trojans classification based on controllability and observability in gate-level netlist"
[5] K. Hasegawa, "Hardware Trojans classification for gate-level netlists based on machine learning"
[6] Y. Wang, "Ensemble-learning-based hardware Trojans detection method by detecting the trigger nets"
[7] K. Hasegawa "Trojan-feature extraction at gate-level netlists and its application to hardware-Trojan detection using random forest classifier"
[8] K. Hasegawa, "Hardware Trojans classification for gate-level netlists using multilayer neural networks"
[9] Q. Liu, "A hardware Trojan detection method based on structural features of Trojan and host circuits"

DL-based HT Detection Methods (Data-driven)

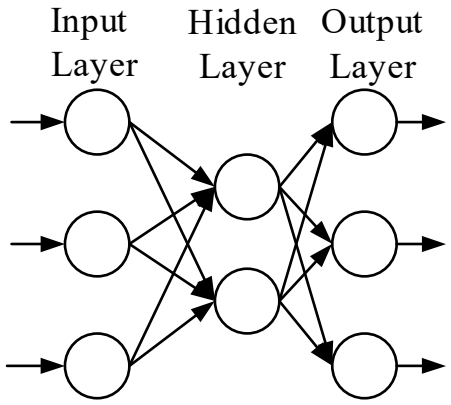
- ❑ Data-driven HT detection that can effectively detect HTs
- ❑ Automatically extracts features and learns model
- ❑ Natural language processing (NLP) technique for information encoding;
- ❑ DL-based classification models for HT detection (tested using LSTM and CNN models).

Netlists



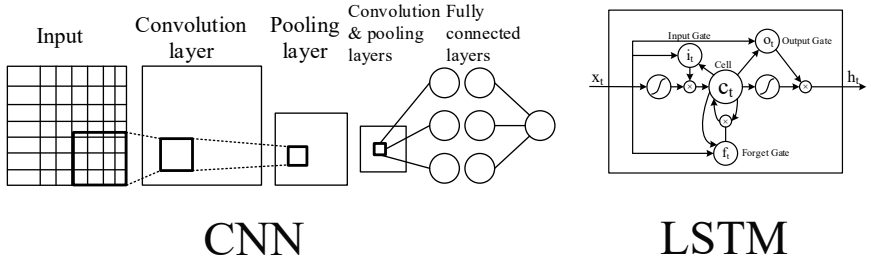
Netlist Information Extraction Strategy

PCP Feature Traces



NLP-based Feature traces Encoding

Encoded Feature Traces



Deep Learning Module for HT Detection

Shichao Yu, Chongyan Gu, Weiqiang Liu and Maire O'Neill, **A Novel Feature Extraction Strategy for Hardware Trojan Detection**," In Proc. IEEE Int. Symp. Circuits and Systems (ISCAS), pages 1-5, Seville, Spain, Oct. 2020

Shichao Yu, Chongyan Gu, Weiqiang Liu and Maire O'Neill, **Deep Learning-based Hardware Trojan Detection with Block-based Netlist Information Extraction**," In IEEE Trans. Emerg. Topics Comput., Oct. 2021

DL-Based HT Detection System Evaluation

- Trust-Hub LEDA library containing 914 HT-infected netlist samples are utilized for evaluation.

HT Types	Netlist	Num. of Components				Netlist	Num. of Components			
		TN	FN	TP	FP		TN	FN	TP	FP
Combinational Trojan-infected Dataset	c2670_T093	776	4	5	0	s15850_T003	2984	4	3	1
	s15850_T012	2985	3	5	0	c6288_T041	2416	0	9	0
	c2670_T016	775	1	6	1	c6288_T066	2416	0	5	0
	c2670_T073	769	1	7	7	s1423_T008	480	3	4	0
	c2670_T054	776	0	6	0	s1423_T003	480	1	6	0
	c2670_T095	775	0	6	1	s15850_T009	2984	4	4	1
	c3540_T087	1134	4	6	0	s1423_T011	480	1	5	0
	c3540_T005	1133	0	9	1	s1423_T005	480	1	4	0
	c3540_T015	1133	1	7	1	s1423_T014	480	0	5	0
	c3540_T012	1129	0	5	5	s13207_T002	2309	1	4	1
	c3540_T017	1133	3	6	1	s35932_T015	6838	4	4	1
	c5315_T004	2307	1	7	0	s13207_T013	2310	5	6	0
	c5315_T047	2306	0	8	1	s35932_T006	6838	2	5	1
	c5315_T064	2306	0	6	1	s13207_T014	2310	0	6	0
	c5315_T057	2306	0	6	1	s35932_T005	6838	3	4	1
	s15850_T014	2984	3	1	1	s13207_T005	2310	0	7	0
	c5315_T063	2306	0	8	1	s35932_T018	6838	5	4	1
	c6288_T049	2415	0	6	1	s13207_T011	2310	2	4	0
	c6288_T048	2416	0	6	0	s35932_T016	6838	1	5	1
	c6288_T082	2416	0	5	0	s15850_T002	2985	0	7	0
	Total	TNR=0.9997, TPR=0.7929, NPV=0.9994, PPV=0.8775								

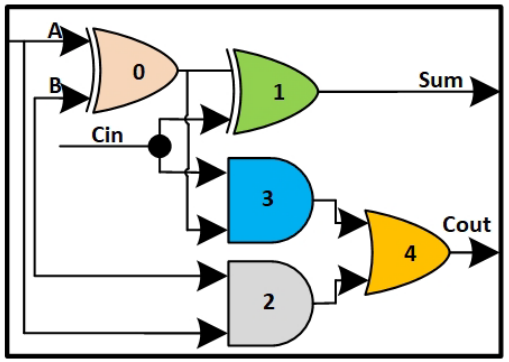
HT Types	Netlist	Num. of Components				Netlist	Num. of Components			
		TN	FN	TP	FP		TN	FN	TP	FP
Sequential (non-scan) Trojan-infected Dataset	s1423_T408	480	4	53	0	s15850_T417	2985	2	22	0
	s15850_T439	2985	0	35	0	s13207_T462	2309	4	57	1
	s15850_T450	2985	3	30	0	s35932_T414	6839	7	76	0
	s1423_T405	480	6	101	0	s13207_T440	2310	1	20	0
	s1423_T429	479	5	84	1	s35932_T402	6836	5	68	3
	s1423_T418	478	5	61	2	s13207_T449	2310	0	18	0
	s1423_T412	480	1	41	0	s35932_T421	6836	0	32	3
	s15850_T468	2984	2	18	1	s13207_T484	2310	4	8	0
	s1423_T407	480	1	16	0	s35932_T413	6839	2	60	0
	s1423_T411	480	1	19	0	s13207_T444	2310	1	16	0
	s1423_T421	480	5	19	0	s35932_T408	6839	8	75	0
	s1423_T422	480	1	19	0	s13207_T473	2310	2	10	0
	s1423_T413	480	0	18	0	s15850_T406	2985	9	40	0
	s15850_T434	2985	2	8	0	s35932_T430	6839	0	21	0
	s13207_T425	2310	0	41	0	s35932_T435	6839	1	22	0
	s13207_T468	2310	1	22	0	s15850_T429	2984	9	92	1
	s15850_T475	2984	2	21	1	s35932_T427	6839	0	22	0
	s13207_T461	2310	0	21	0	s15850_T443	2983	0	33	2
	s15850_T433	2985	1	20	0	s35932_T411	6839	1	21	0
	s13207_T450	2309	7	93	1	s35932_T434	6839	0	18	0
	Total	TNR=0.9999, TPR=0.9346, NPV=0.9992, PPV=0.9892								

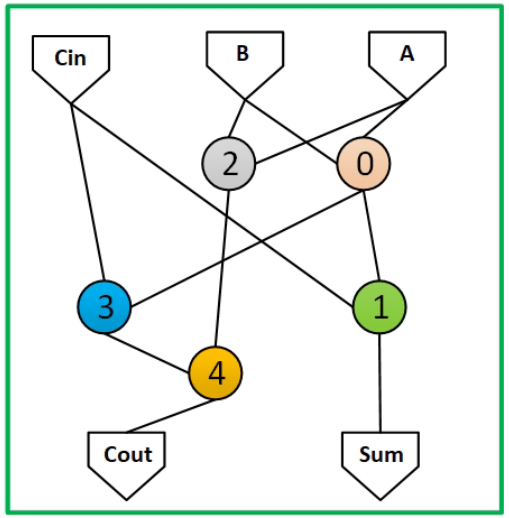
- 79% TPR, 99% TNR, 87% PPV and 99% NPV for **combinational** Trojan detection (40 training samples/40 validating samples, 5 epochs, LSTM);
- 93% TPR, 99% TNR, 98% PPV and 99% NPV for **sequential** Trojan detection (40 training samples/40 validating samples, 5 epochs, LSTM)

GNN Framework for Hardware Trojan Detection

```
module FA(
   input A,
   input B,
   input Cin,
   output Sum,
   output Cout
);

wire xor1_out;
wire and1_out;
wire and2_out;

// XOR gates for Sum
xor xor1 (xor1_out, A, B);
xor xor2(Sum, xor1_out, Cin);

// AND gates for Cout
and and1(and1_out, A, B);
and and2(and2_out, xor1_out, Cin);

// OR gate for Cout
or or1(Cout, and1_out, and2_out);

endmodule
```

```
xor xor1 (xor1_out, A, B);
xor xor2(Sum, xor1_out, Cin);
and and1(and1_out, A, B);
and and2(and2_out, xor1_out, Cin);
or or1(Cout, and1_out, and2_out);
```





- GNN model trained on a mixed dataset comprising Combinational and Sequential HTs
- Enables multi-type trojan detection within a single model.
- To compute subgraph embeddings a Graph Attention Network (GAT) is employed
- Enables nodes to selectively aggregate information from their neighbours through a learned attention mechanism.

Conclusions and Future Research Directions

Conclusion

ML/DL has a major role to play in Hardware Security

- AI techniques can be used to attack hardware security
  - DL-based side channel attacks – can bypass traditional SCA countermeasures
  - ML-based modelling attacks of PUFs
- AI can be used to aid hardware security
  - Hardware Trojan Detection
  - New PUF designs
  - Thwarting ML-based side channel attacks
- However, AI circuitry on hardware platforms is itself vulnerable to SCA attacks
  - Hyperparameters and weights of well-trained ML/DL models are valuable

Future Research Directions

ML and side channel analysis

- further investigate ML-based SCA attacks of post-quantum and advanced crypto implementations
- how to cost-effectively thwart ML-based SCA attacks?

ML and hardware Trojan detection

- generic HW Trojan detection approaches for the design-stage
- consider AI detection approaches resilient to adversarial HTs
- can AI-based approaches be used to detect Trojans at other stages of the IC manufacturing process?
- can AI-based HW Trojan detection approaches be embedded into EDA design tools?

Future Research Directions

ML and Physical Unclonable Functions

- further research on using ML-based approaches to cost-effectively thwart ML-based modelling attacks of strong PUF designs
- research on novel PUF designs using ML/DL approaches

Vulnerability of ML/DL Models on Hardware Platforms

- research on hardware-based attacks of edge ML/DL implementations
- use of approximate ML architectures to improve their security?
- is it possible to target ML/DL models on multi-tenant FPGA cloud?

Need further research to understand full capability of AI-based approaches in *attacking* and *defending* hardware security to allow us to deliver truly trustworthy hardware.

Acknowledgement

Chongyan Gu

Anh-Tuan Hoang

Ayesha Khalid

Ciara Rafferty

Shichao Yu

Jack Miskelly

Phil Hodgers

Neil Hanley

Zain Shabbir

James Moore

Thank you