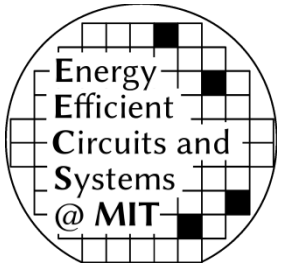
A Secure Digital In-Memory Compute (IMC) Macro with Protections for Side-Channel and Bus Probing Attacks

Maitreyi Ashok, Saurav Maji, Xin Zhang, John Cohn, Anantha Chandrakasan

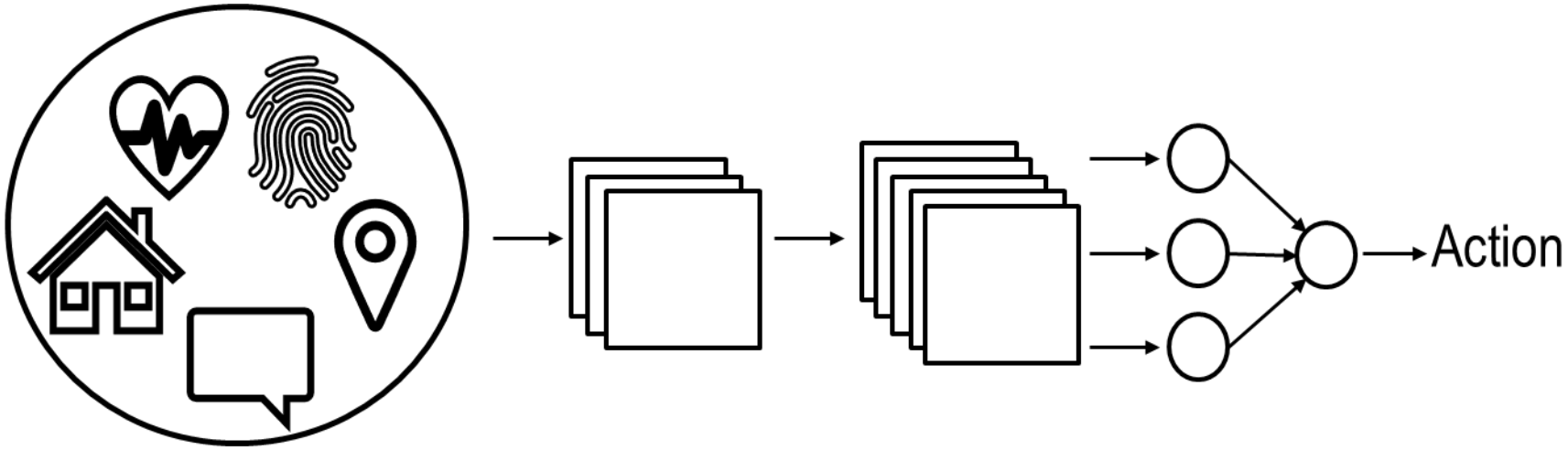
CICS Research Review, May 1st 2024
Presented at Custom Integrated Circuits Conference, April 2024



Outline

- **Motivation and Prior Work**
- Secure IMC Macro Features
  - Side-Channel Secure Boolean Shared Compute
  - Neural Network Model Security
  - Secret Key Generation On-Chip
- Measurement Results
- Conclusion

Security Threats for Machine Learning (ML)



Inputs to Model

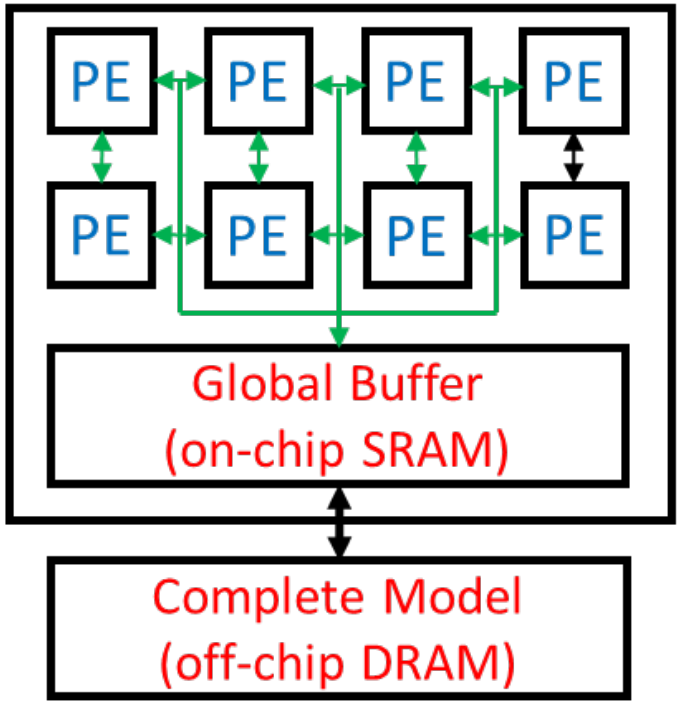
- Often collected from sensors on edge devices
- Private information should not be externally readable

Parameters of Model

- Contain information from private training datasets
- Knowledge can lead to adversarial attacks
- Competitive model IP

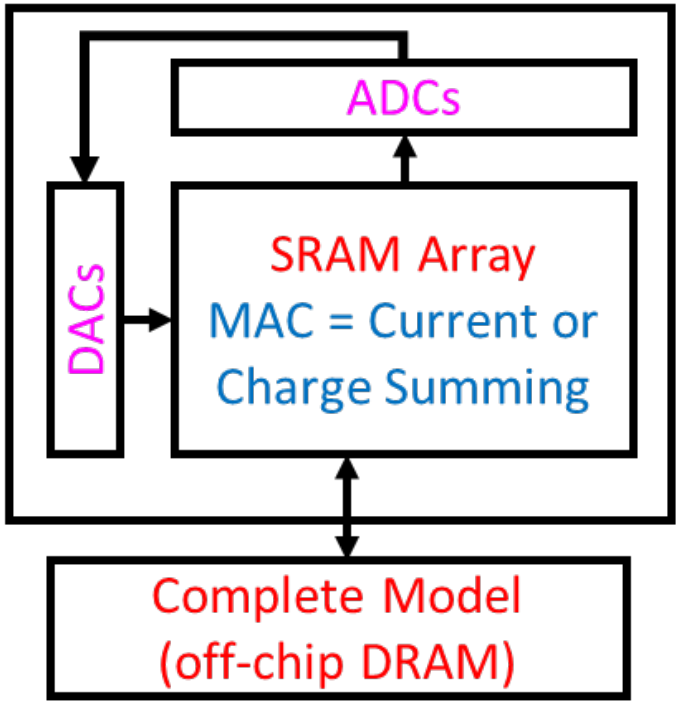
ML Accelerators + In-Memory Compute

Traditional ML Accelerators



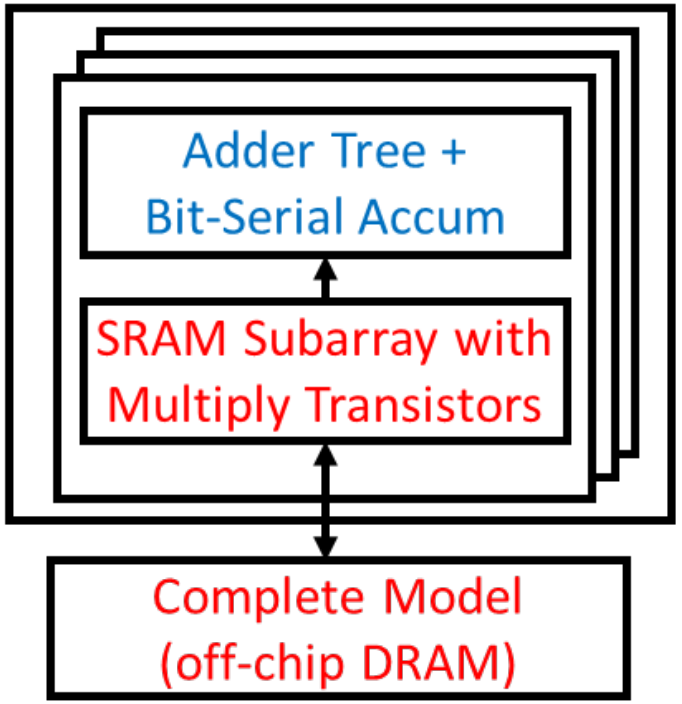
Compute and **Memory** separate
Large **NoC** → data movement energy limits the accelerator

Analog In-Memory Compute



Compute and **Memory** co-located for low data movement energy
High **DAC/ADC** energy, noise, mismatch limit precision

Digital In-Memory Compute



Compute and **Memory** interleaved for lower data movement energy
Does not limit precision and benefits from technology scaling

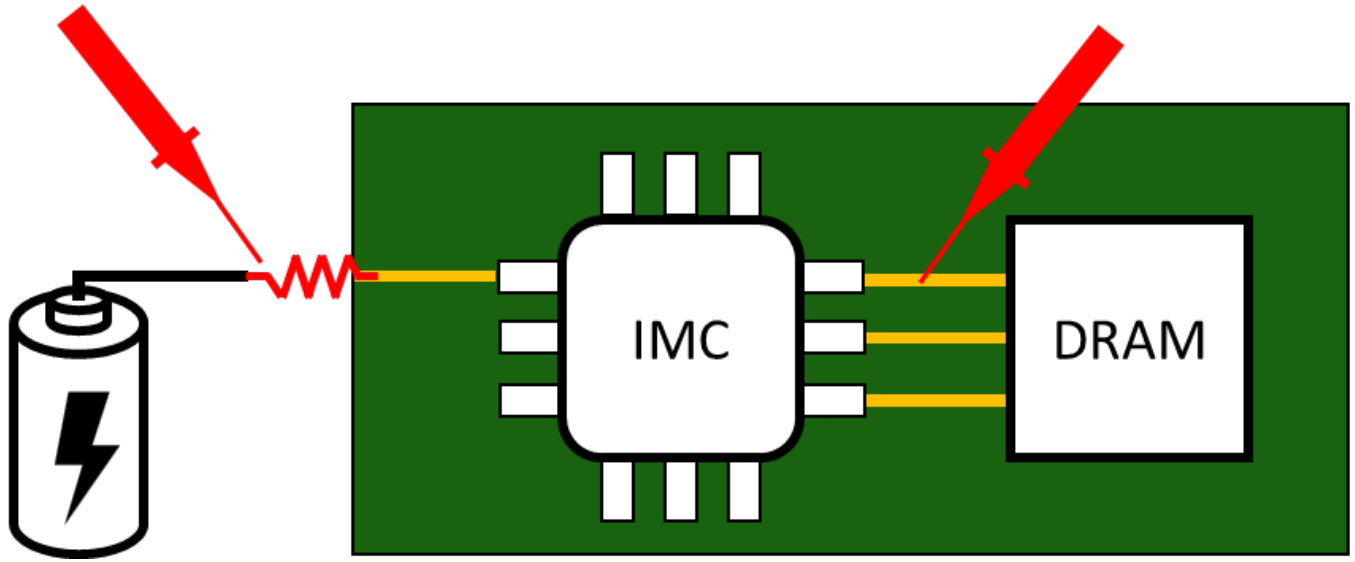
Threat Model: Potential Attack Points

Physical Side Channel Attacks (SCA)

- Correlate circuit currents to operation/data
- Can be invasive (power) or non-invasive (electromagnetic)

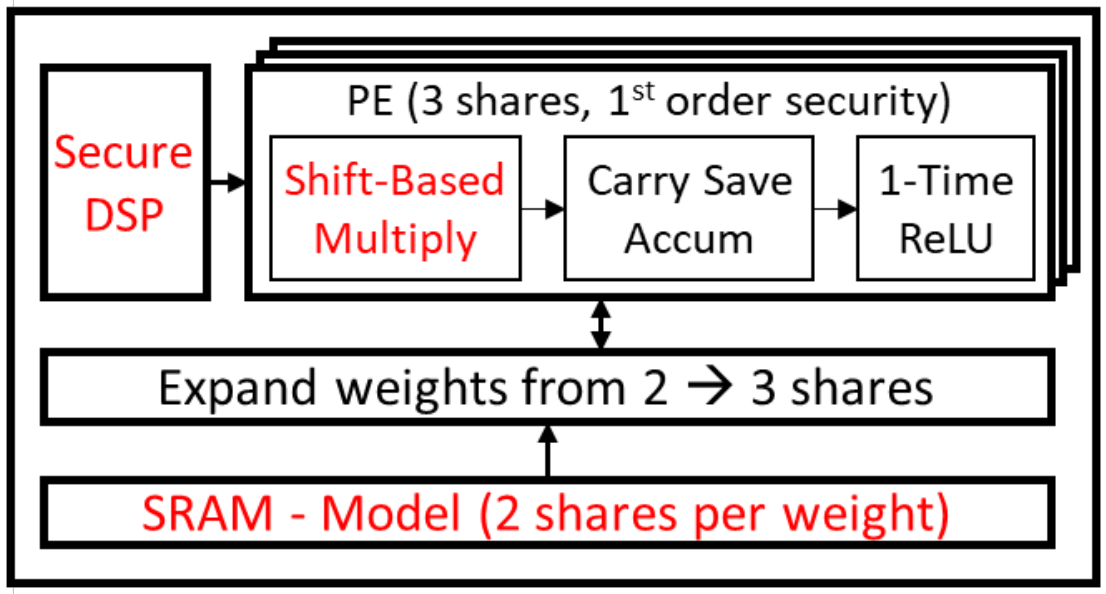
Memory Bus Probing Attacks (BPA)

- Directly measure data transfer between logic and off-chip memory IC
- Allow full reconstruction of large ML models stored externally



Challenge #1 – Side-Channel Attack Security

Threshold Implementation Based Accelerator
(S. Maji et al., ISSCC 2022)

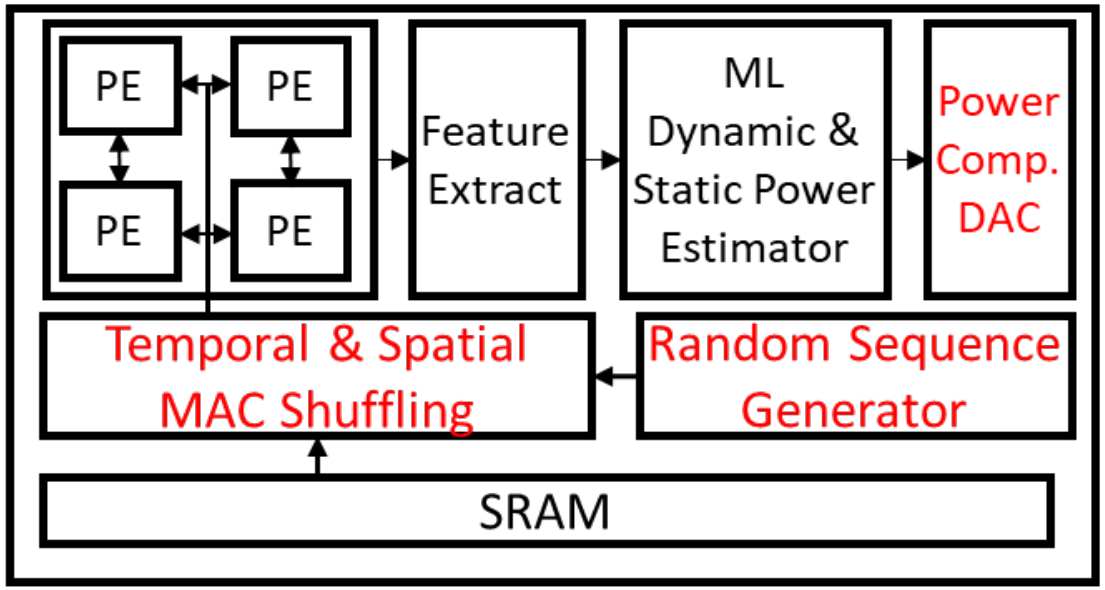


✓ Secure up to infinite attack samples

✗ Limit weight to powers of 2

✗ **Constant random bits for security**

Compensation and Shuffling Based Accelerator
(Q. Fang et al., VLSI 2023)



▬ Focus on practical # of attack samples

✓ Voltage Scaling Agnostic

✗ **Constant random bits for security**

Challenge #1 – Side-Channel Attack Security

Threshold Implementation Based Accelerator
(*S. Maji et al., ISSCC 2022)*

Compensation and Shuffling Based Accelerator
(*Q. Fang et al., VLSI 2023)*

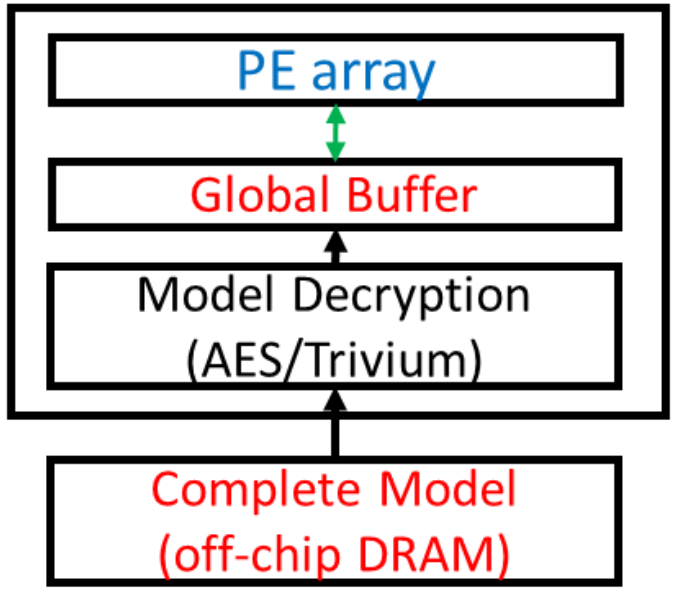
Prior work on SCA secure ML accelerators focus on Von Neumann Architectures

Protections do not necessarily scale to the high parallelism requirements of In-Memory Compute

Challenge #2 – Bus-Probing Attack (BPA) Security

Cryptographic Cipher for Model Decryption

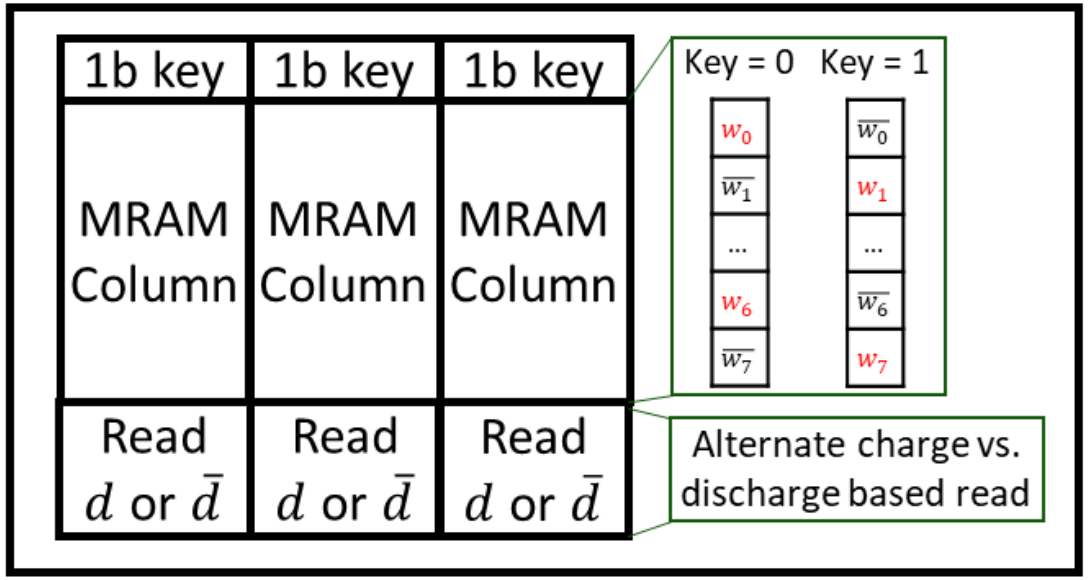
(S. Maji et al., ISSCC 2022; Q. Fang et al., VLSI 2023)



- ✓ Cryptographic security guarantees
- ✗ AES not lightweight, Trivium not standard
- ✗ Need secret key from off-chip

MRAM-Based XOR for Model Decryption

(Y-C Chiu et al., ISSCC 2022)



- ✓ Lightweight integrated security for IMC
- ✗ Uses insecure One-Time Pad
- ✗ Need secret key from off-chip

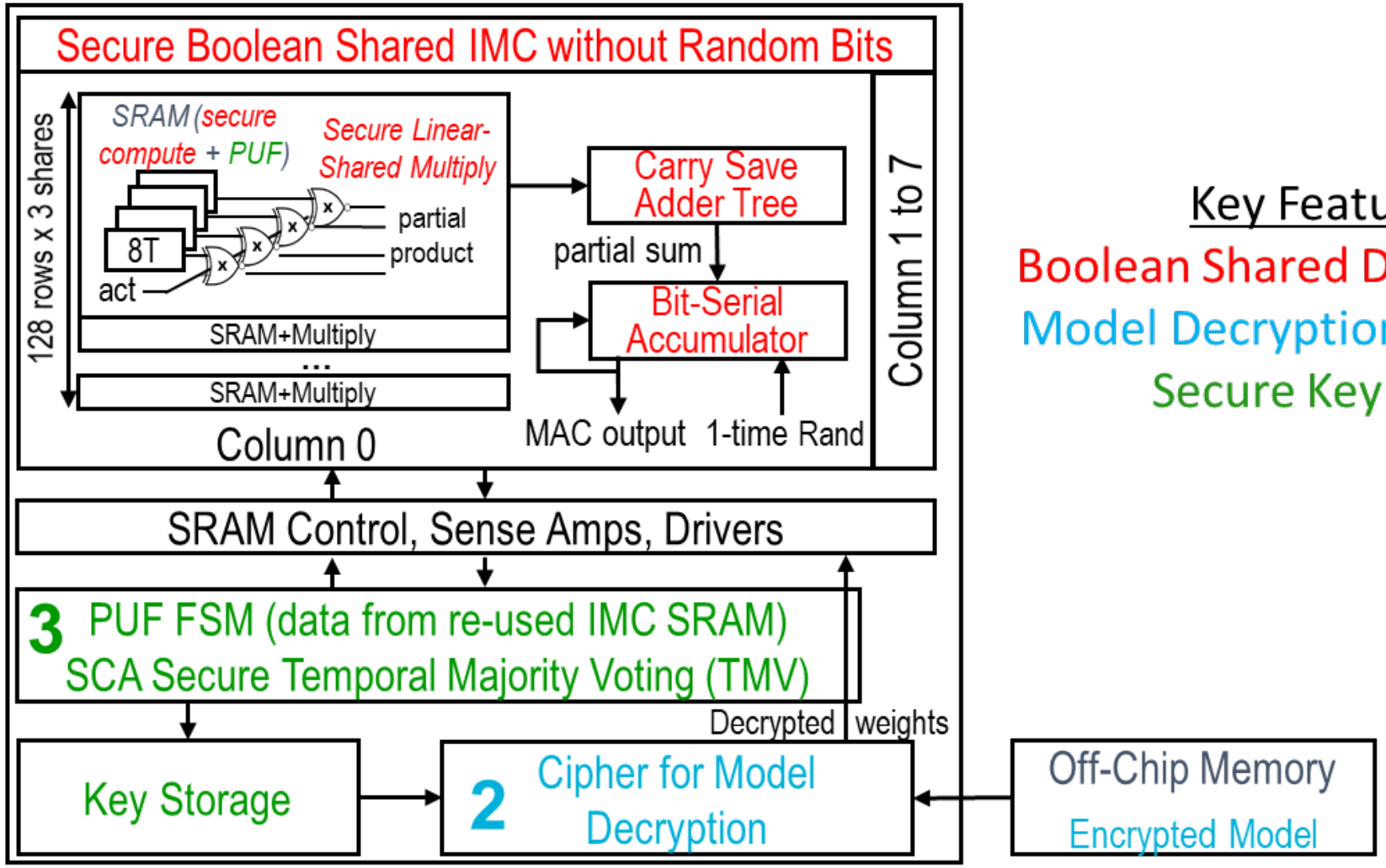
Solution – Key Features

- Boolean Shared Digital IMC for SCA Security
  - No constant random bits
  - No impact to neural network accuracy
- Model Decryption On-Chip for BPA Security
  - Lightweight cipher implemented in SCA secure fashion
- Secure Key Generated On-Chip
  - Used to maintain statistical security guarantees of cipher
  - Reuses IMC SRAM for minimal overhead Physically Unclonable Function

Outline

- Motivation and Prior Work
- **Secure IMC Macro Features**
  - **Side-Channel Secure Boolean Shared Compute**
  - **Neural Network Model Security**
  - **Secret Key Generation On-Chip**
- Measurement Results
- Conclusion

Macro Architecture



Key Features of Secure IMC

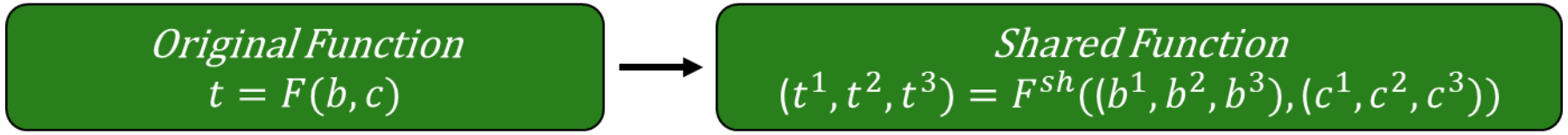
Boolean Shared Digital IMC for SCA Security

Model Decryption On-Chip for BPA Security

Secure Key Generated On-Chip

Feature 1: Side-Channel Secure Compute

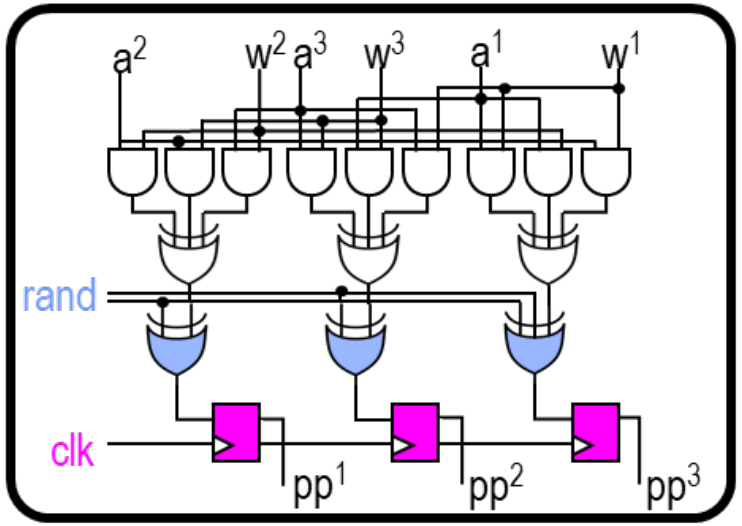
- Boolean Sharing
  - Split each data bit and computation into separately computed ***shares***
  - Total power consumption is unrelated to actual data



- Properties (for Practical Levels of Security)
  - Correctness:
    - If $b = b^1 \oplus b^2 \oplus b^3, c = c^1 \oplus c^2 \oplus c^3,$ Then $t = t^1 \oplus t^2 \oplus t^3$
  - Non-Completeness:
    - If $F^{sh} = \{F^1, F^2, F^3\}$, Then each of F^j does not include all shares of each input
  - Approximate Uniformity:
    - For each sub-circuit, each shared output has the same distribution bias as the unshared output
    - Outputs that are not jointly uniform are not combined directly

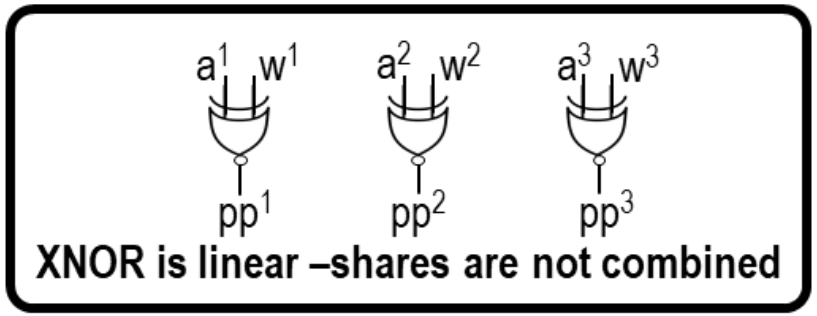
Feature 1a: Side-Channel Secure Multiply

Conventional: Shared AND gate



- ✗ Random bit refreshing to maintain uniformity
- ✗ Registers to maintain non-completeness
- ✗ 1 multiply = 48 gate-equivalents
- ✓ Standard bit-serial multiply for digital IMC

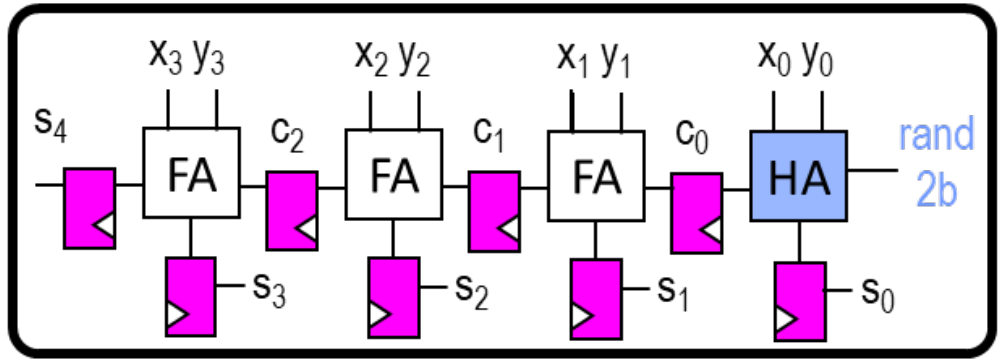
Proposed: Shared XNOR gate



- ✓ Shares maintain uniformity without random bits
- ✓ Can cascade to next gate without registers
- ✓ 1 multiply = 6 gate-equivalents
- ▬ Need data format conversion at macro interface
  Negligible effect on NN accuracy

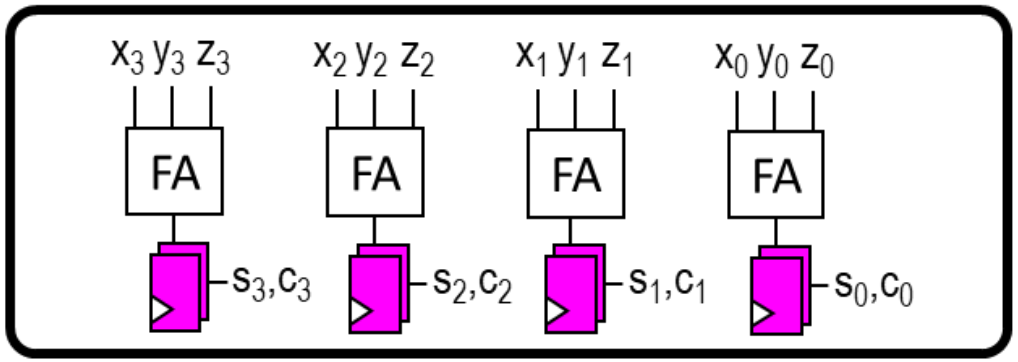
Feature 1b: Side-Channel Secure Adder Tree

Conventional: Ripple Carry Adder



- ✗ Random bit refreshing for half adder uniformity
- ✗ More pipeline stages: 16 clock latency

Proposed: Carry Save Adder



- ✓ Only use full adder, individual circuits uniform
- ✓ Fewer pipeline stages: 10 clock latency

Sum and Carry are not Jointly Uniform

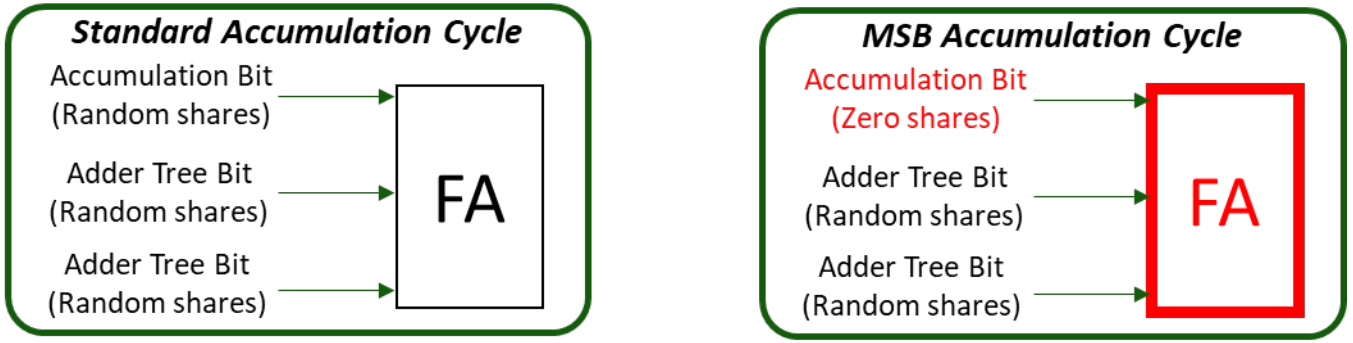
Combine outputs of same adder after they become approximately jointly uniform

Not “infinitely” secure…BUT

Enough security against practical attackers without random bits!

Feature 1c: Side-Channel Secure Bit-Serial Accumulator

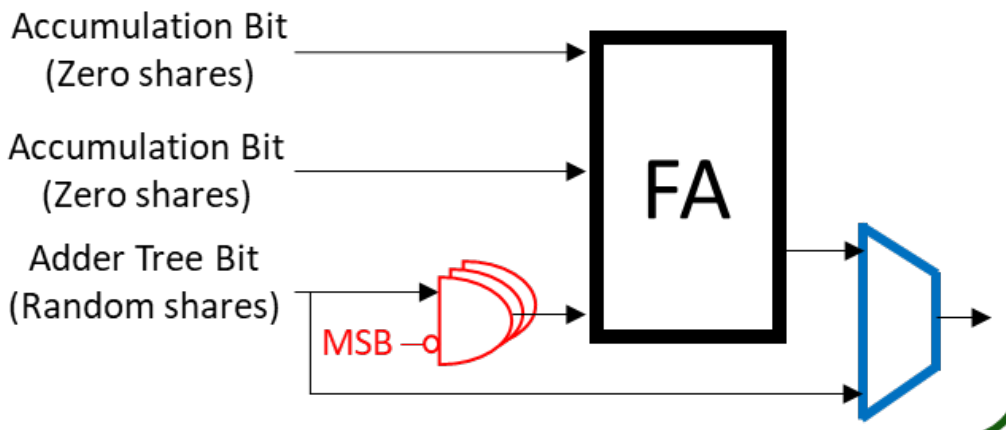
- Add binary weighted sums from adder tree for multi-bit activations
  - Partial sums in modified carry-save format
  - Performs last few layers of adder tree as part of accumulate to save on latency
- Eliminate any half adders
  - Full adders are natively secure, do not require random bits
- Adding with known {0,0,0} shares for MSB of activation - Insecure



Feature 1c: Side-Channel Secure Bit-Serial Accumulator

Solution 1

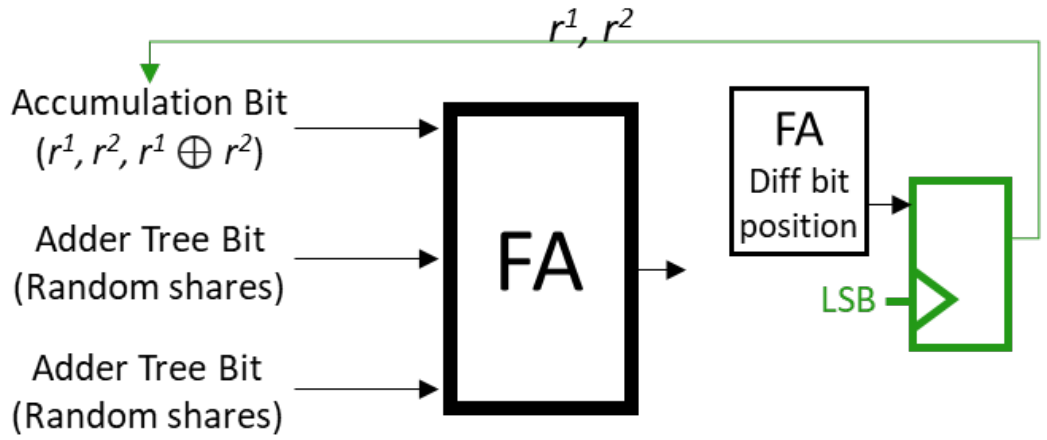
FA Gate + Direct assignment of partial sum



Solution 2

Generate approximately random share of 0

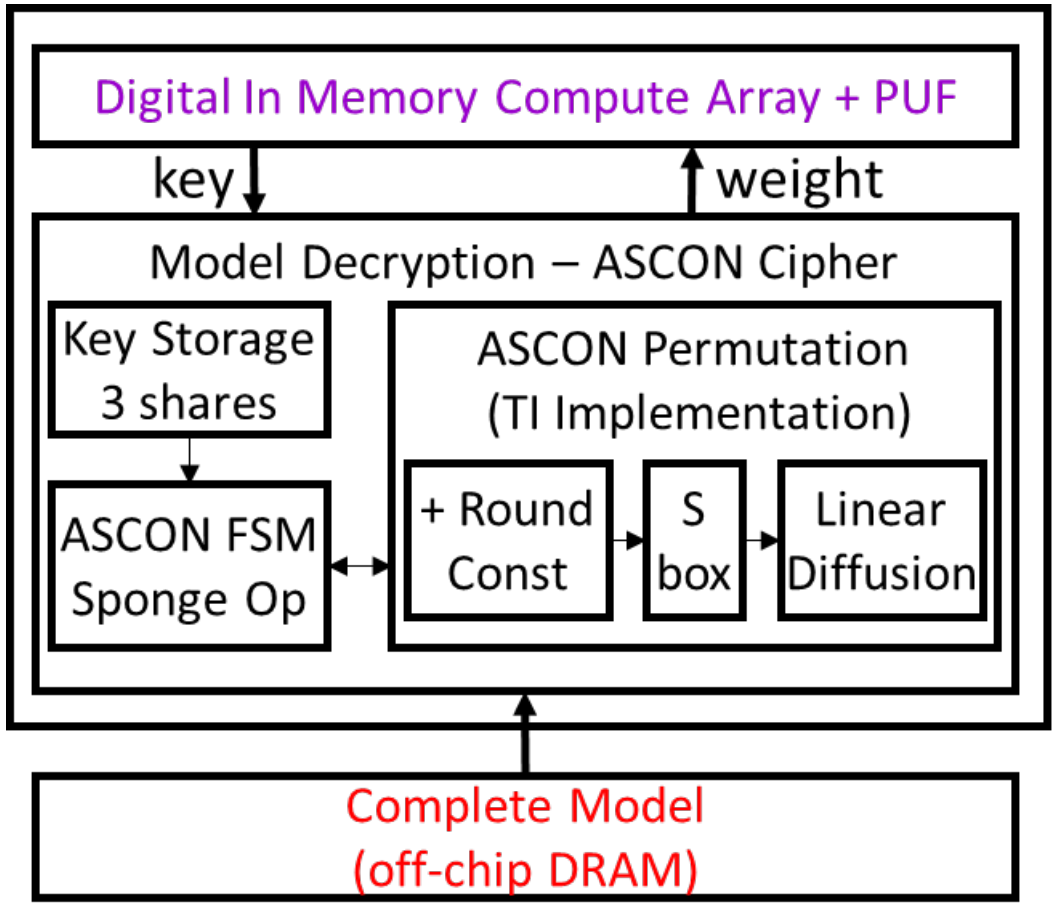
(r^1, r^2, $r^1 \oplus r^2$) from prior compute



True random bits only required one-time at start-up

Inspired by "Changing of the Guards" J. Daemen et al., CHES 2017.

Feature 2: Model Security

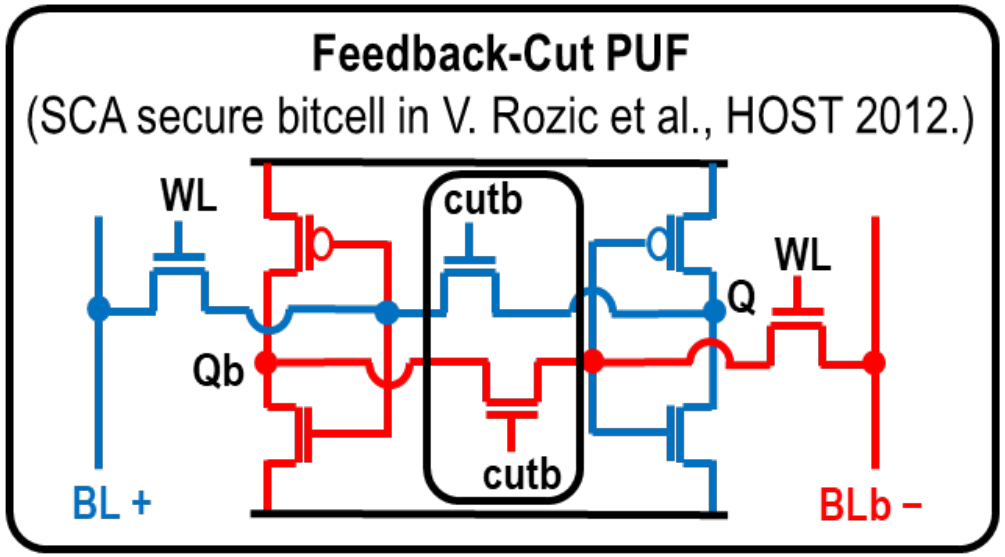


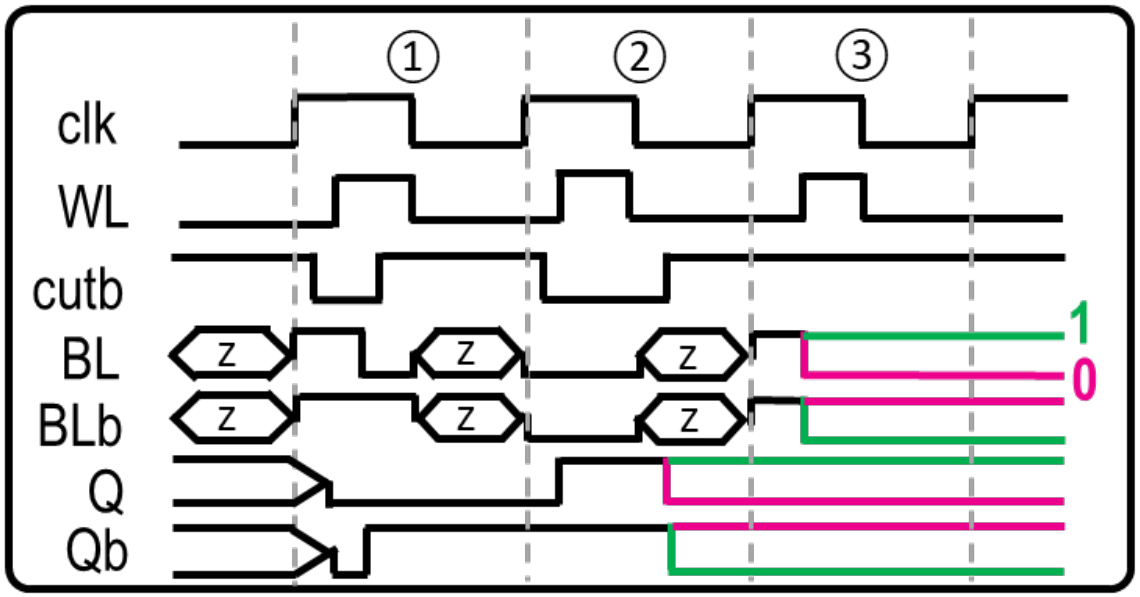
ASCON Cipher

- Selected by NIST for standardized lightweight cryptography
- Low overhead side-channel security
  - Low logic algebraic order
  - Permutation: AND, XOR, Shift
  - No constant random bits required with "Changing of the Guards", J. Daemen et al., CHES 2017.

Feature 3: Secret Key Generation Reusing IMC

Physically Unclonable Function: Unique and Repeatable Response per Challenge





① *Secure Write Reset*: Write fixed value to remove data dependence (use SCA secure write)

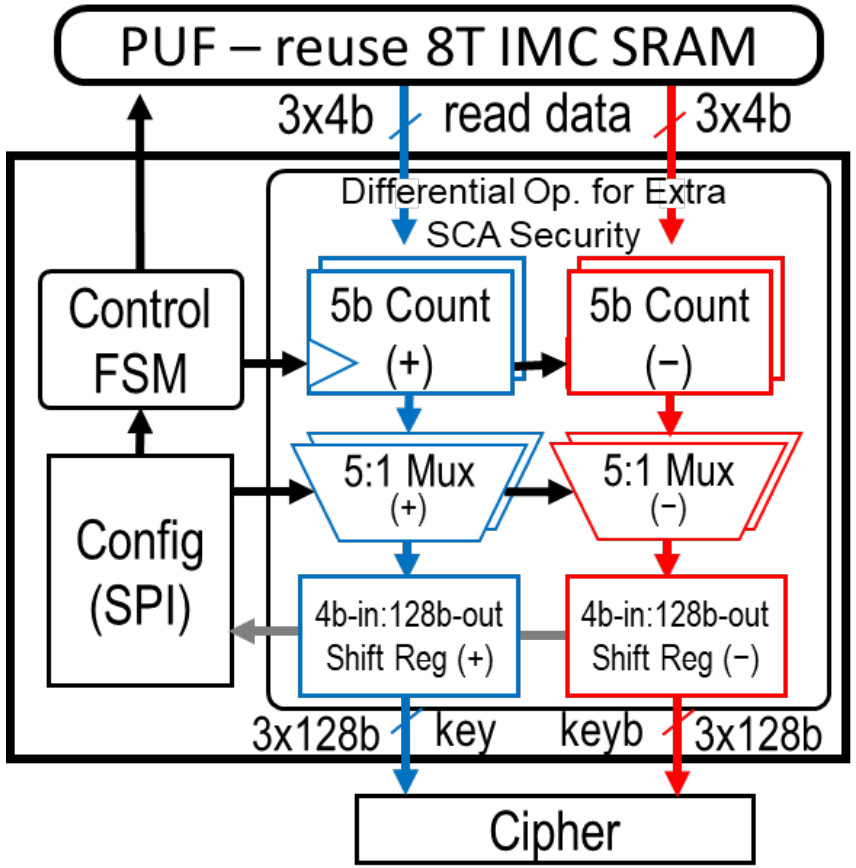
② *PUF Evaluation*: Cut and reconnect feedback transistor

Settles to 0 or 1 since + or – side stronger due to local mismatch

③ *Standard Read*: PUF data from SRAM through differential sense amplifiers

Feature 3: Secret Key Generation Reusing IMC

- Cells are susceptible to noise
  - Especially those with lower mismatch
- Temporal Majority Voting
  - Evaluate PUF multiple times
  - Choose more common data value
- Key value shared with manufacturer in secure environment only during initial configuration
- Generate several keys per macro with different addresses (multiple CRP per chip)
- Configuration option for desired noise tolerance and error correction capability



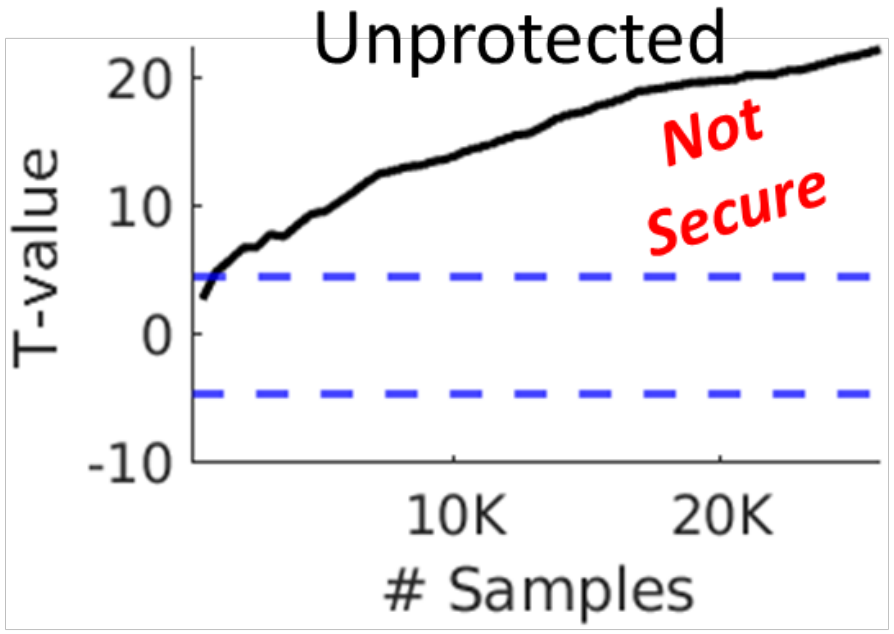
**ECC not included on-chip

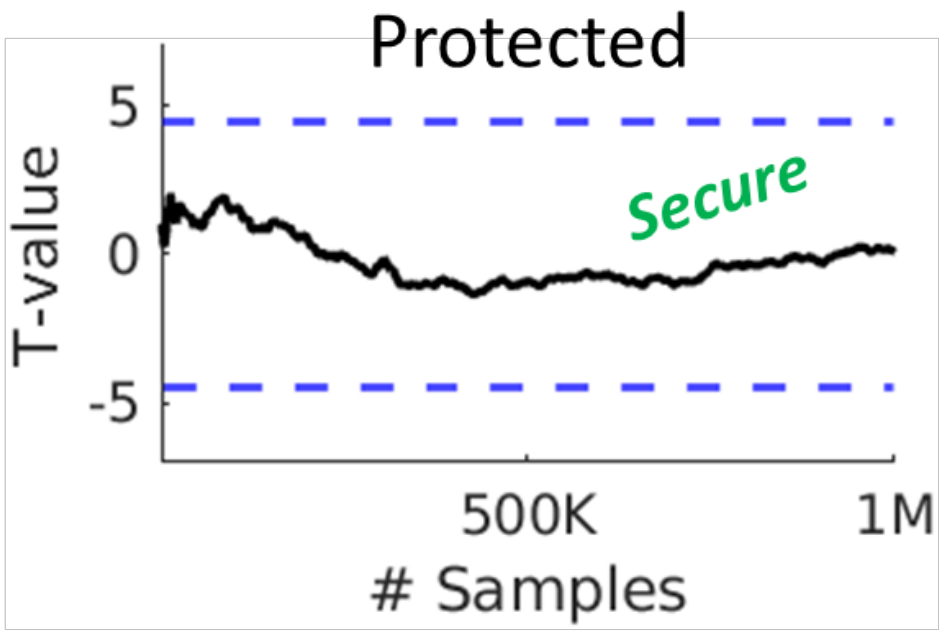
Outline

- Motivation and Prior Work
- Secure IMC Macro Features
  - Side-Channel Secure Boolean Shared Compute
  - Neural Network Model Security
  - Secret Key Generation On-Chip
- **Measurement Results**
- Conclusion

Side Channel Security

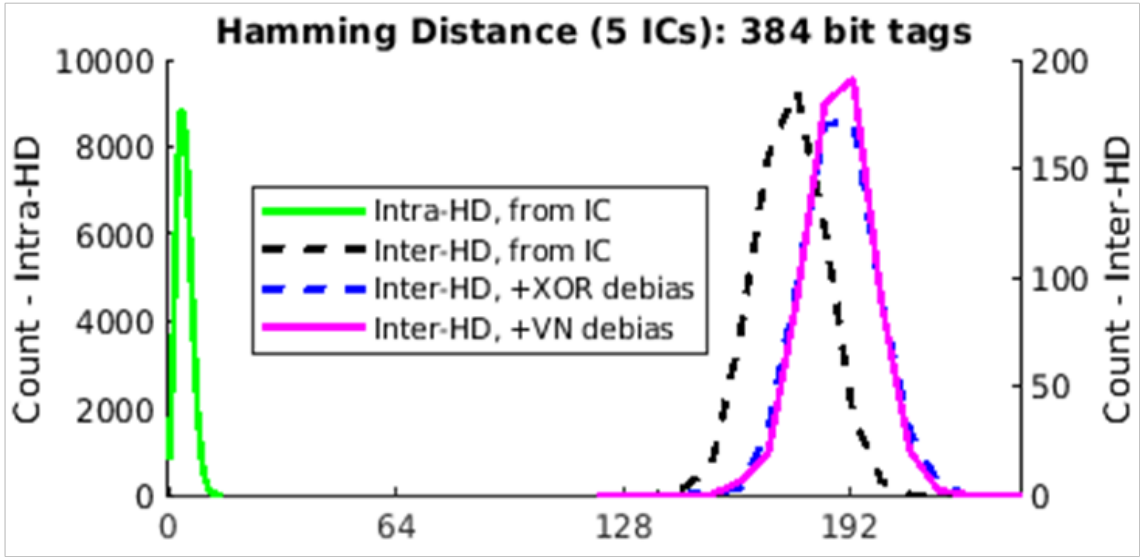
- Test Vector Leakage Assessment
  - Measure power side channel leakage with fixed and random compute inputs
  - Calculate significance of difference between both types of inputs, $t = \frac{\mu_{fixed} - \mu_{random}}{\sigma_{fixed} - \sigma_{random} / \sqrt{\#\ samples}}$
  - $t > 4.5$ indicates statistically significant difference





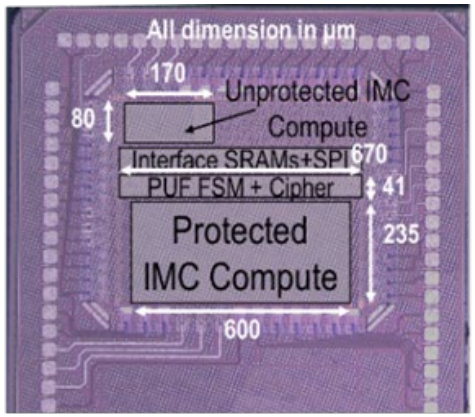
Physical Unclonable Function Security

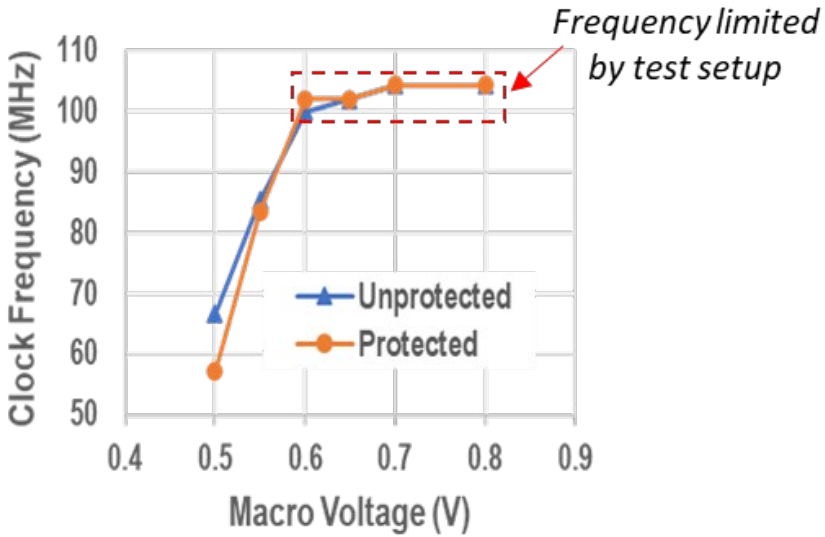
- NIST 800-22 Tests pass for 5/5 chips
- BER = 0% with off-chip BCH code
- SCA Security (test with CNN attack)
  - PUF Read/Evaluation always differential → inherently secure
  - Write uses SCA secure feedback-cut
  - Boolean sharing of key/SRAM keeps exact value private
  - Differential operation of temporal majority voting FSM keeps data statistics private



Key	Not Shared	Shared	**Shared & Diff.**
RMSE from FSM	250K train	1M train	
Unshared HW	1.38	4.90	**7.20**
Shared HW	N/A	2.35	**32.13**

Die Micrograph + Performance





Technology		14nm CMOS
Area (excluding IO pads)		1.06 mm^2
Throughput (GOPS) 0.55 V, 80 MHz	Unprotected	41.0 (4b weight, 1b act) 9.10 (4b weight, 8b act)
	Protected	81.9 (4b weight, 1b act) 10.2 (4b weight, 8b act)
Energy Efficiency (TOPS/W) 0.55 V, 80 MHz	Unprotected	90.2 (4b weight, 1b act) 14.4 (4b weight, 8b act)
	Protected	6.94 (4b weight, 1b act) 0.89 (4b weight, 8b act)
Area Efficiency (TOPS/mm2) 0.55 V, 80 MHz	Unprotected	3.01 (4b weight, 1b act) 0.67 (4b weight, 8b act)
	Protected	0.49 (4b weight, 1b act) 0.061 (4b weight, 8b act)

Comparison to Prior Work

		ISSCC'22	ISSCC'22	VLSI'23	This Work
Process		ASIC, 28nm	ASIC, 22nm	ASIC, 40nm	ASIC, 14nm
Operation		0.60-0.95 V, 10-125 MHz	0.65-0.8 V	0.70-0.90V, 0.05-50 MHz	0.50-0.80 V, 57-100 MHz
Architecture		Von Neumann	Analog IMC	Von Neumann	Digital IMC
Precisions		8b act; 8b wgt power-of-2	1/8b act; 2/4/8b wgt	mult-bit	1-8b act; 4/8/12/16b wgt
Threat Model		SCA, BPA	BPA	SCA, BPA	SCA, BPA
Security Protection		Boolean Mask + Trivium	XOR/XNOR Weight Encryption	Shuffling; Dual power compensation	Boolean Shared + ASCON
Random Bits		Each clk (Trivium)	N/A	Each clk (RSG)	1-time Only
Security Overhead	**TOPS/mm^2**	2.30x (6.1x mult)	*Not reported*	2.27x	6.2x (1b act)
	TOPS/W	5.48x	*Not reported*	1.76x	13.0x (1b act)
Security Level		CPA (>>2M),TVLA (>>2M)	One-Time Pad	CPA (>200M),TVLA (55M)	CPA (>>1M)*,TVLA (>>1M)*

*Tested in ideally high attack SNR condition, security will last significantly longer for realistic operation

Outline

- Motivation and Prior Work
- Secure IMC Macro Features
  - Side-Channel Secure Boolean Shared Compute
  - Neural Network Model Security
  - Secret Key Generation On-Chip
- Measurement Results
- **Conclusion**

Conclusion + Future Work

- Generalized IMC solution for ML with *Privacy & Integrity*
  - Side Channel and Bus Probing Attack Security for In Memory Compute
  - No random bits from PRNGs required
  - No limitations on neural network accuracy
- Future Improvements
  - More exploration of tradeoffs between security and area/energy overheads
  - Usage of approximate compute for further exploitation of natively secure compute gates

Acknowledgements

- This work is supported by the MIT-IBM Watson AI Lab, NSF GRFP (Grant No. 1745302), and MathWorks Engineering Fellowship
- The authors would like to express our gratitude to the IBM teams:
  - Thank Kevin Tien, Cliff Osborn, Seiji Munetoh, Kohji Hosokawa for tape-out support and valuable discussions;
  - Thank Cyril Cabral, Kai Schleupen, John Timmerwilke for packaging solutions;
  - Thank Dirk Pfeiffer, Daniel Friedman, Dan Dechene, David Cox, Mukesh Khare for management support
- The authors thank members of the Energy Efficient Circuits and Systems Group at MIT for their valuable discussion and feedback.

- Please reach out to maitreyi@mit.edu with any questions or feedback!