INTEGRITY CHARACTERIZATION OF EMBEDDED NEURAL NETWORK AGAINST LASER FAULT INJECTION

Mathieu Dumont, Pierre Alain Moëllic, Kévin Hector, Jean-Max Dutertre

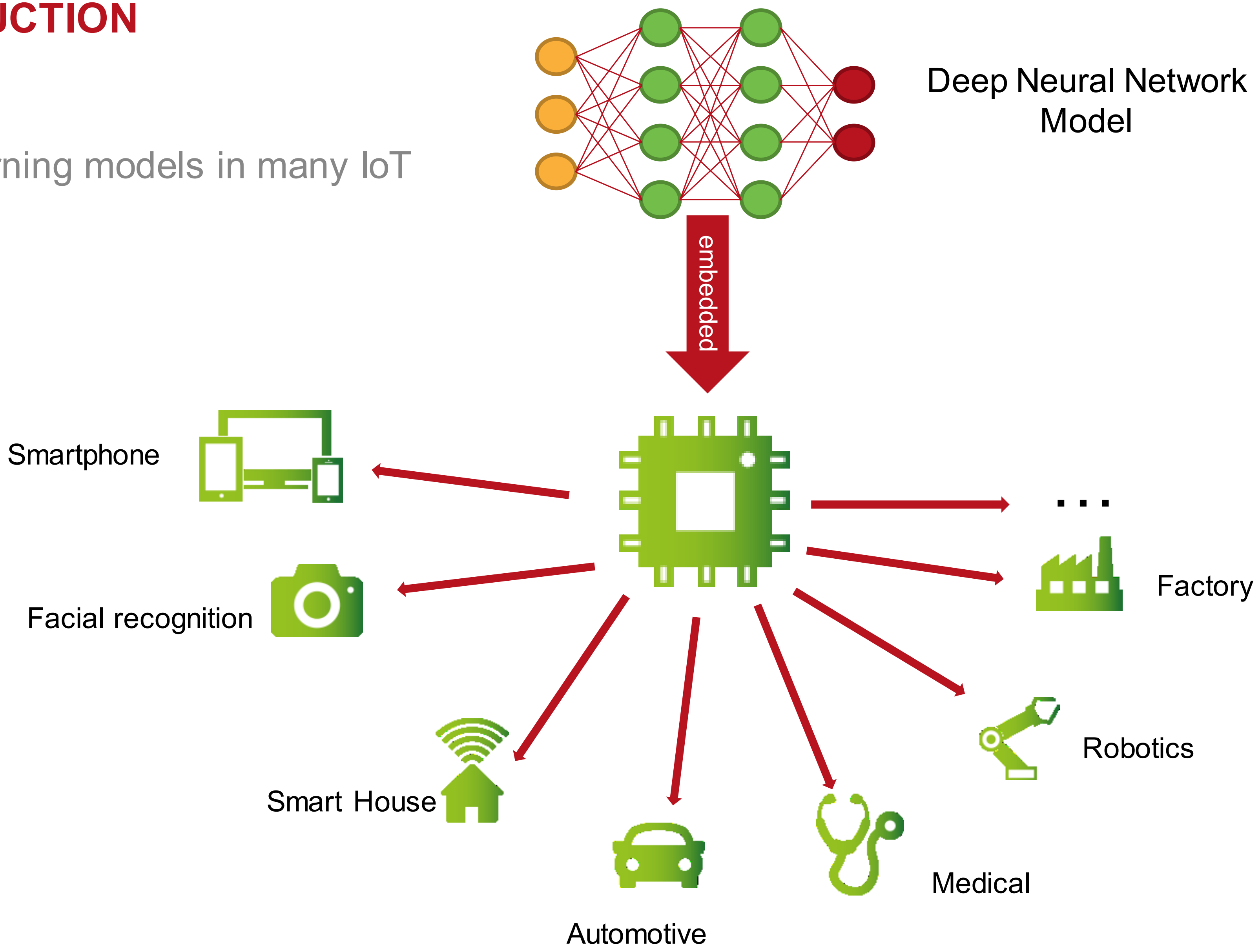
mathieu.dumont@cea.fr

JAIF 2022

09/11/2022

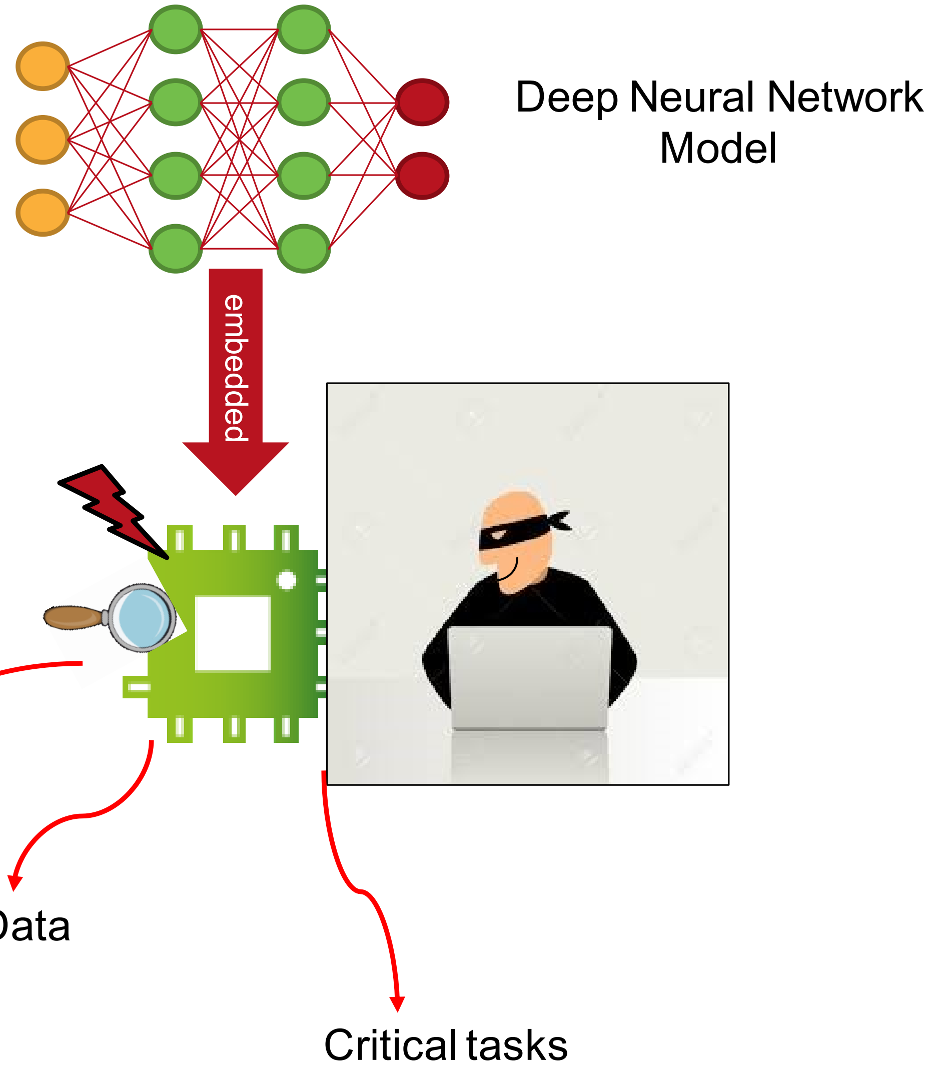
INTRODUCTION

- Deployment of Machine Learning models in many IoT devices.



INTRODUCTION

- Deployment of Machine Learning models in many IoT devices.
- Embedded Neural Networks offer physical access to an attacker.
- Ongoing standardization, regulation (European AI Act*), certification actions.



* https://artificialintelligenceact.eu/

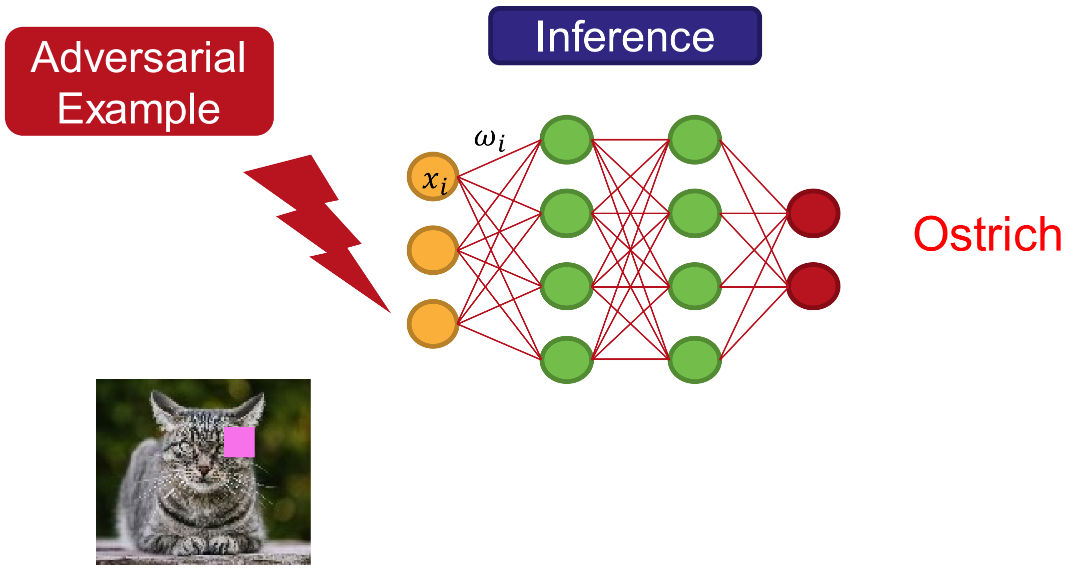
SUMMARY

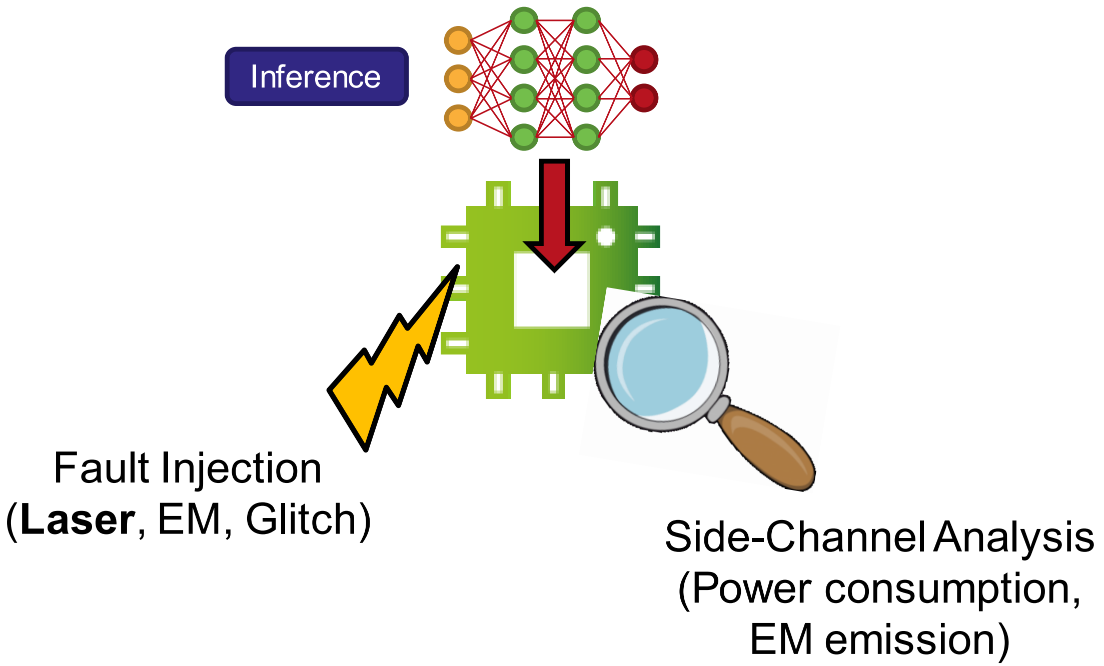
- **Context**
- Setup and Fault Model
- Targeting the Model Integrity with Laser Fault Injection
- Guided Laser Fault Injection
- Conclusion

CONTEXT

- **Attack on machine learning models**

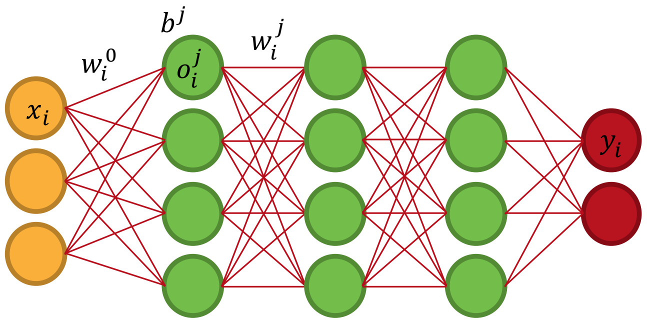
  - Adversarial Example (software attack) is a major threat against DNN. **Massive research efforts** on that field.
  - Physical attacks (hardware attack) constitute new threats against DNN. **Recent works.**

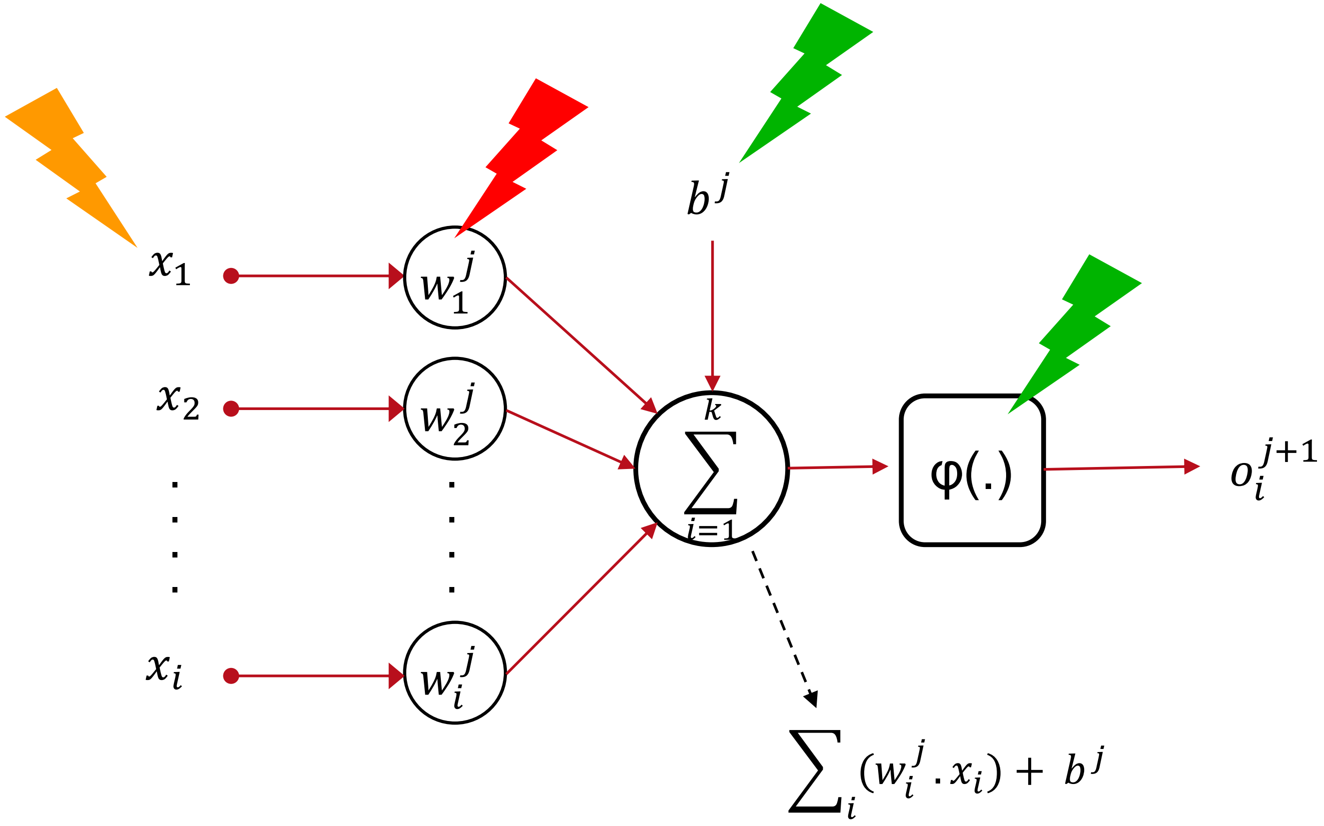




CONTEXT

- **Parameter-Based Attacks**
  - Typical neuron computation:

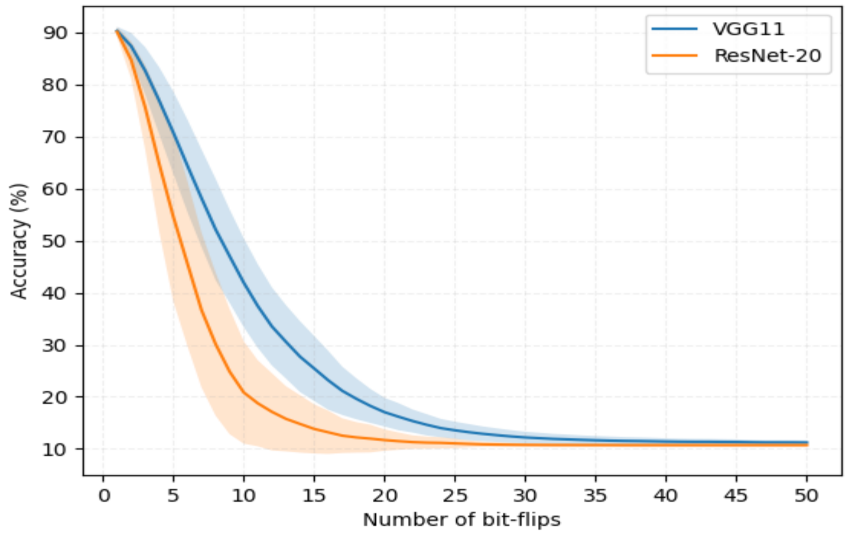




CONTEXT

Weight-based adversarial attacks

- Most of works are API-based attacks, first Liu *et al. (2017)* [1].
- Bit-Flip Attack (BFA) by Rakin *et al.* [2]:
  - Find the most sensitive bits to flip based on the loss gradient ranking of each bit $\nabla_b \mathcal{L}$
  - Decrease the model performance with few bit-flips



Bit-Flip Attack simulation (random-guess level = 10%)

CONTEXT

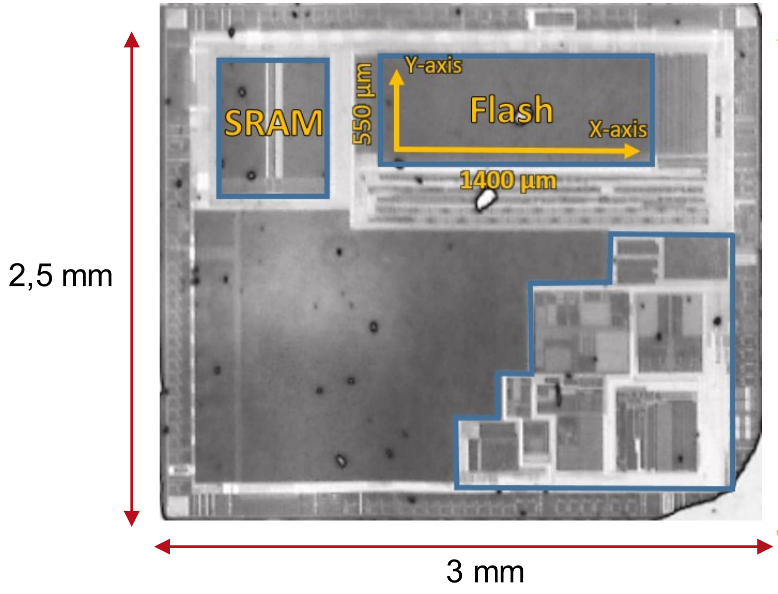
➢ Hardware Parameter-Based Attacks

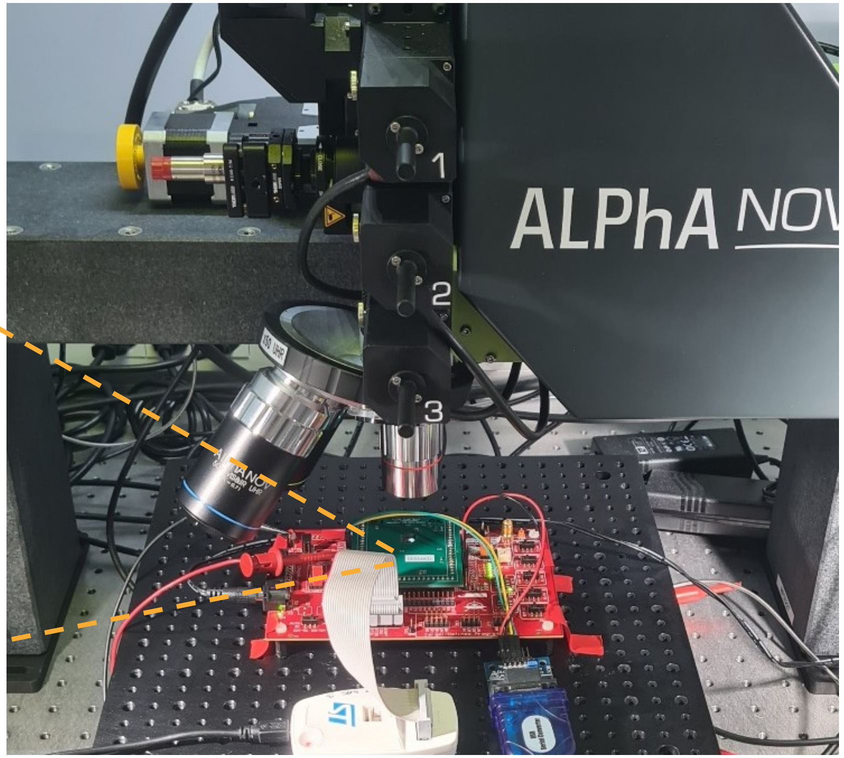
	Target	Model (Dataset)	Quantization	Simu/Exp	DUT	Comments
★ Breier *et al. [3]*	**Activation Function**	MLP (MNIST)	No	Simulation / **Laser**	ATMega 328P	Target last hidden layer. **Skip instruction**
Benevenuti *et al. [4]*	Whole model	MLP (IRIS)	No	Neutron irradiation / Laser	SRAM-Based FPGA	Safety-based.
★ Yao *et al. [5]*	**Weights**	CNN (MNIST, CIFAR10, ImageNet)	8-bit	BFA / **RowHammer**	Intel i7-3770 CPU (DRAM)	**Random-guess level** for 11 models with less than 20 bit-flips
Liu *et al. [6]*	Whole model (black-box)	CNNs (ImageNet)	8-bit	Clock Glitch	SoC (FPGA/ Cortex A53)	Black and gray box
Fukuda *et al. [7]*	Softmax function	CNN (MNIST)	No	Clock Glitch	ATMega128	Only last layer implemented in C
★ ***Ours works***	**Weights**	**MLP (IRIS, MNIST)**	**8-bit**	**BFA / Laser**	**32-bit MCU, Cortex-M**	**White-box. Precise attack with minimum faults**

SETUP AND FAULT MODEL

➢ Laser bench setup

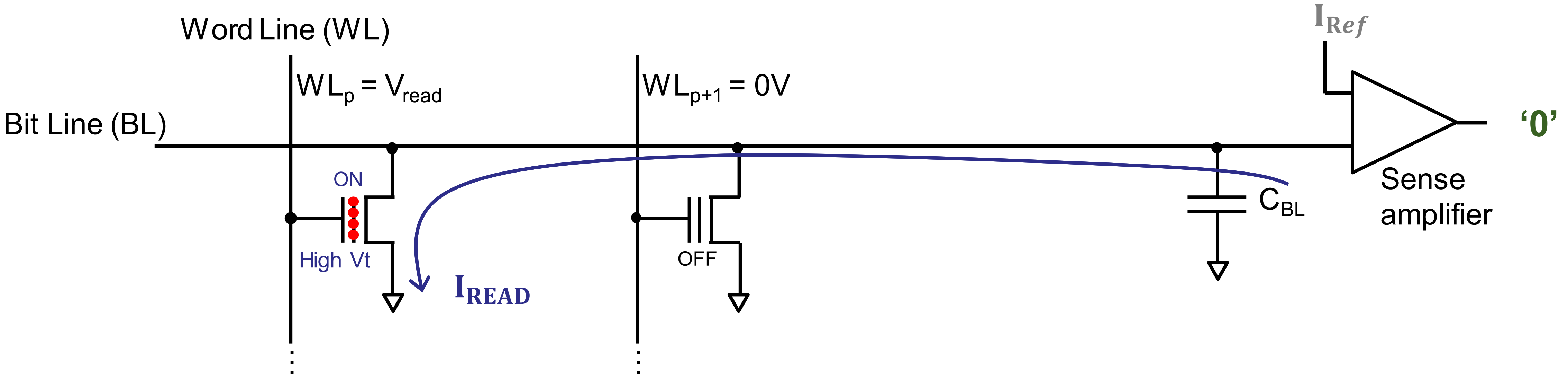
- Laser with two independent laser spots at 1064nm (near IR).
- Target : ARM Cortex M3 running at 8MHz. CMOS 90nm.
  - Flash : 128kb NOR Flash
  - Open backside





SETUP AND FAULT MODEL

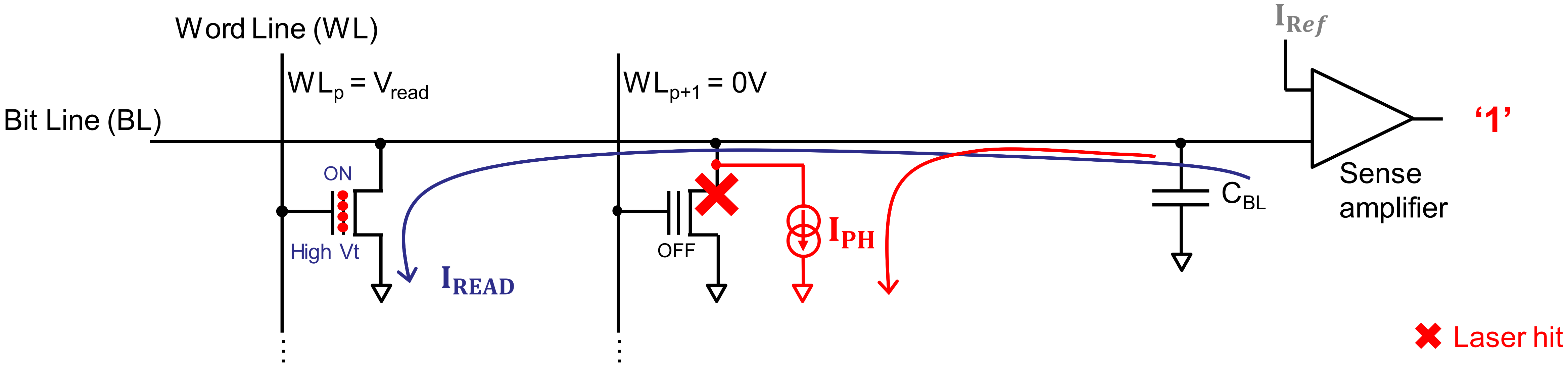
Bit-set fault model [8]



- Floating gate charged, low read current : $\mathbf{I_{READ}} < I_{Ref}$ → Read value : '0'

SETUP AND FAULT MODEL

Bit-set fault model [8]



- Floating gate charged, low read current : $I_{READ} < I_{Ref}$ → Read value : '0'
- Additionnal I_{PH} current : $I_{READ} + I_{PH} > I_{Ref}$ → Read value : '1'

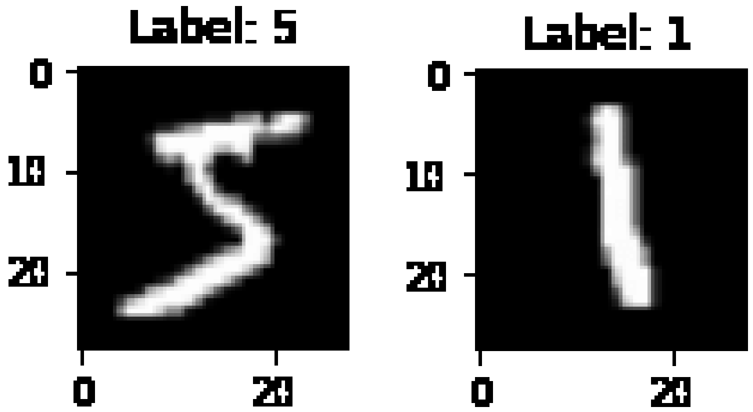
One-way (unidirectional) fault model

→ Bit-set fault model

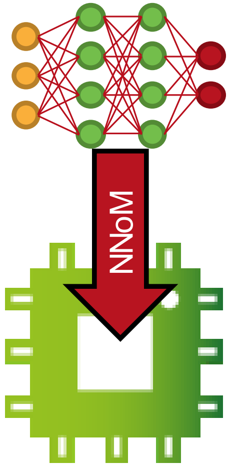
SETUP AND FAULT MODEL

Datasets and models

- **IRIS Dataset :** small network, 4 inputs and 3 outputs
  - Only few neurons and one hidden layer is sufficient
- **MNIST Dataset :** 28x28 digits images (‘0’,…’9’)
  - MLP network, one deep layer of 10 neurons, ReLu activation



SETUP AND FAULT MODEL



MCU implementation

- Need access to library → NNoM
  - 8-bit quantization
  - White-box access to inference code

- During the multiplication $\left(w_i^j . x_i\right)$ the load "`ldr`" instruction of the weight value is surrounded by a trigger

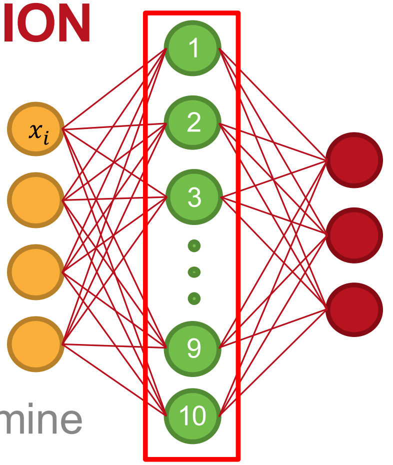
Part of C code of Weighted-sum computation during inference

```
while (rowCnt){    //loop on all neuron parameters
    for (int j = 0; j < dim_vec; j++){
        q7_t inA = *pA++;    //input value load to inA
        q7_t inB = *pB++;    //weight value load to inB
        ip_out += inA*inB;   //Mul input x weight
    }
    [...]
    rowCnt--;}
```

```
;q7_t inB = *pB++        ;Weight n+1 initialization
ldr      r3, [r7, #80]   ;Loading the weight address
adds     r2, r3, #1
str      r2, [r7, #80]
ldrsb.w  r3, [r3]        ;Weight value loading.
strb     r3, [r7,#23]
```

- Context
- Setup and Fault Model
- **Targeting the Model Integrity with Laser Fault Injection**
- Guided Laser Fault Injection
- Conclusion

TARGETING THE MODEL INTEGRITY WITH LASER FAULT INJECTION

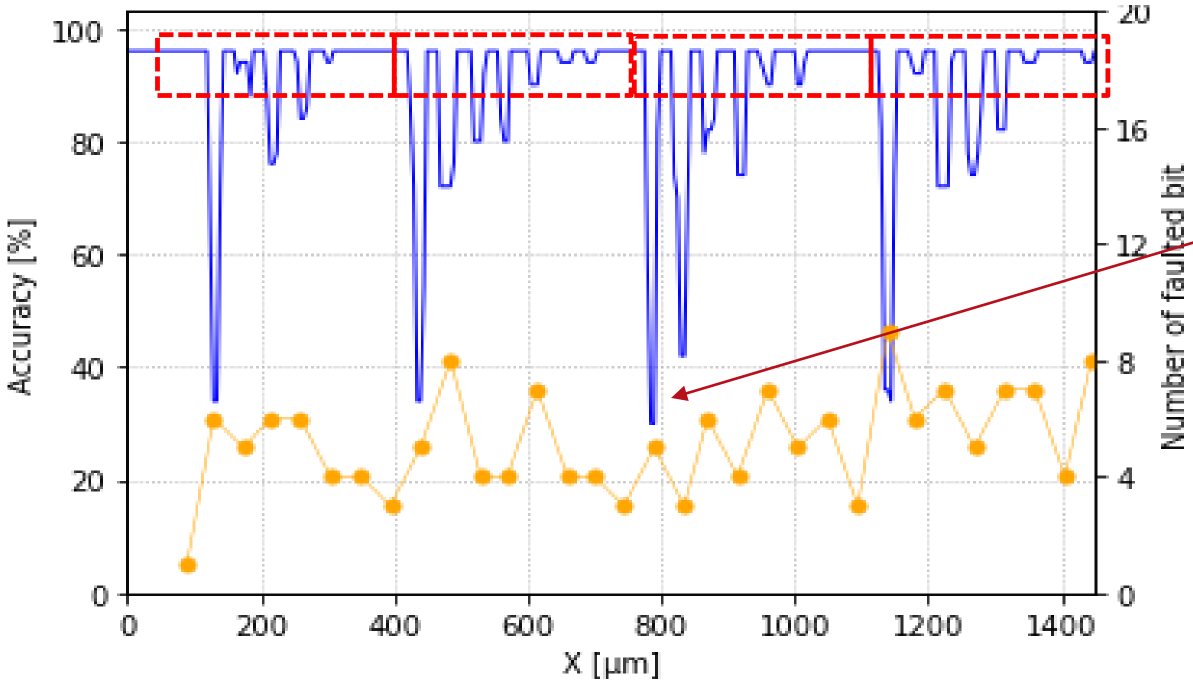


➢ Laser fault injection characterization on Multi Layer Perceptron

- Iris model with one deep layer of 10 neurons (**40 weights** on the first layer).
- The laser spot move along the X-Axis of the flash memory (with a step of 2µm).
  - At each X-step, **50 inferences** are performed and outputs compared with software results to determine the **embedded model accuracy**.
  - During one inference, **all weight loading** (‘`ldr`’) trigger a laser shot.

- Accuracy of embedded model without attack = 95%
- Total number of bits = 320bits

Optical Lens x5 (Spot of 15µm)
Pulse power : 170mW
Pulse Width : 200 ns
Delay : 500 ns
Step on X = 2µm



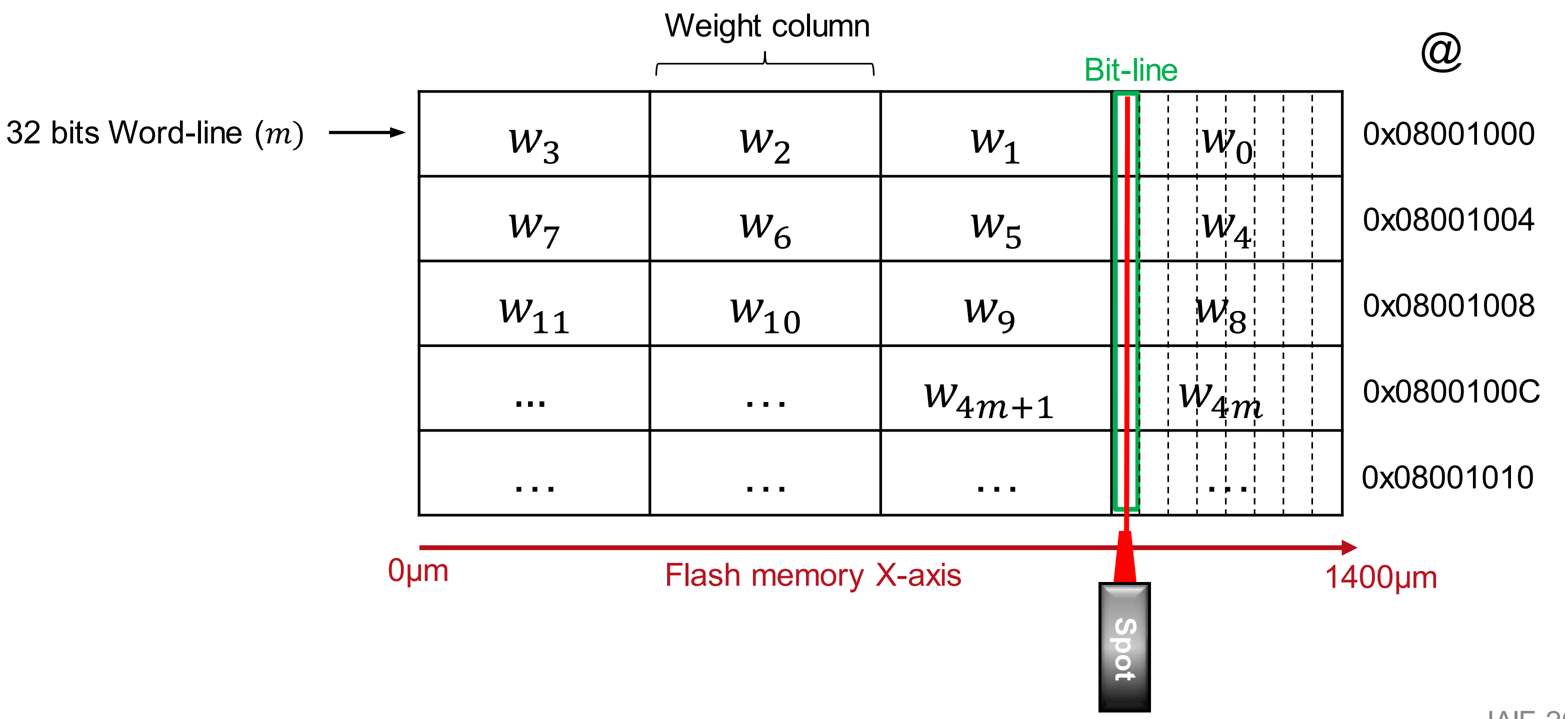
Faults number average = 5 faults

✓ Drop accuracy to 30%, with only **5 faulted bits** (1,6% of faulted bits)

TARGETING THE MODEL INTEGRITY WITH LASER FAULT INJECTION

Laser fault injection characterization on Multi Layer Perceptron

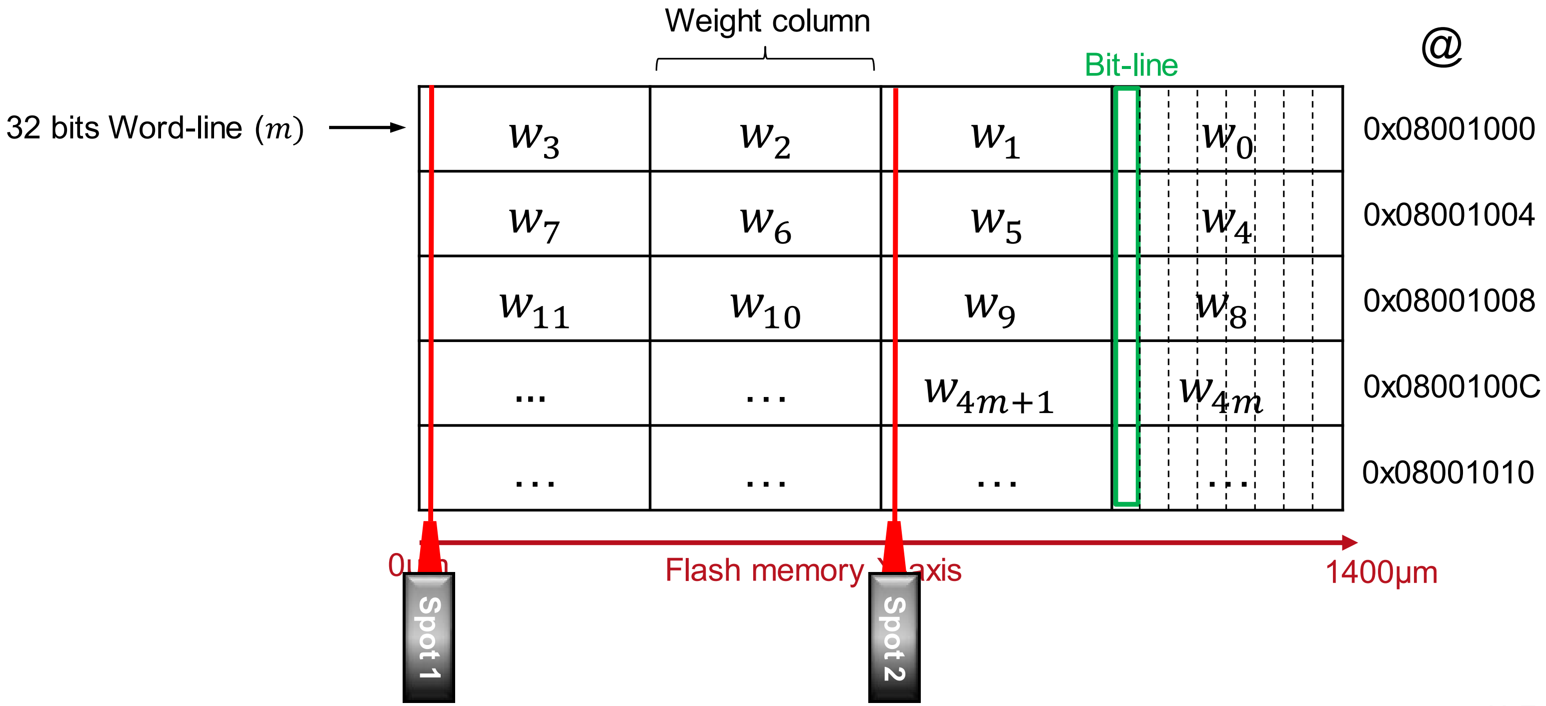
- LFI characterization limitation : Due to memory flash storage architecture, only **1/4** of all weights could be faulted during one inference.



TARGETING THE MODEL INTEGRITY WITH LASER FAULT INJECTION

➢ Laser fault injection characterization on Multi Layer Perceptron

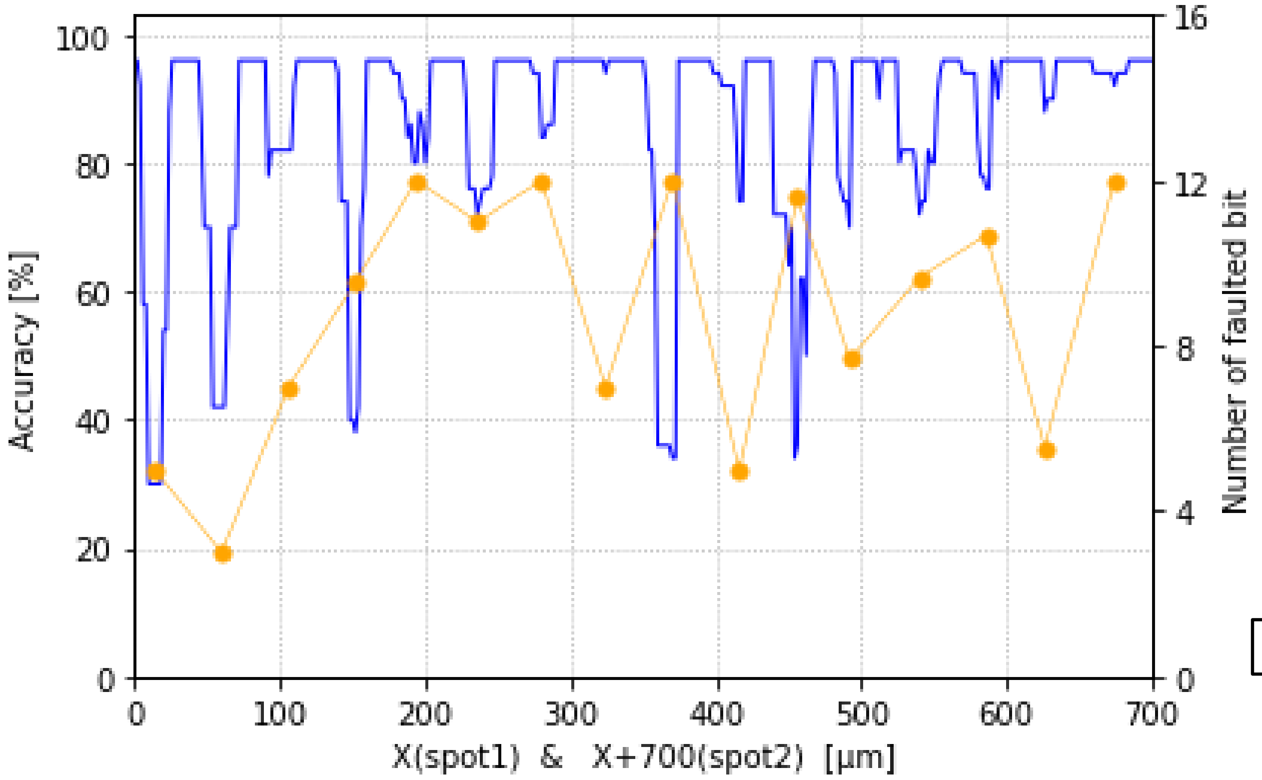
- LFI characterization limitation : Due to memory flash storage architecture, only **1/4** of all weights could be faulted during one inference.
- With the two spots, 2 weights columns could be targeted, leading to **1/2** of the weights that be can faulted.



TARGETING THE MODEL INTEGRITY WITH LASER FAULT INJECTION

➢ Bi-spot Laser fault injection characterization on Multi Layer Perceptron

- Both spots are moved together from 0 to 700µm for Spot1 (from 700 to 1400µm for Spot2) and shot at the same time.



✓ More faults are induced with bi-spot.

✓ Huge accuracy drop happened, not only on high order bit.

✓ No drop accuracy below 30%.

Faults number average = 9 faults

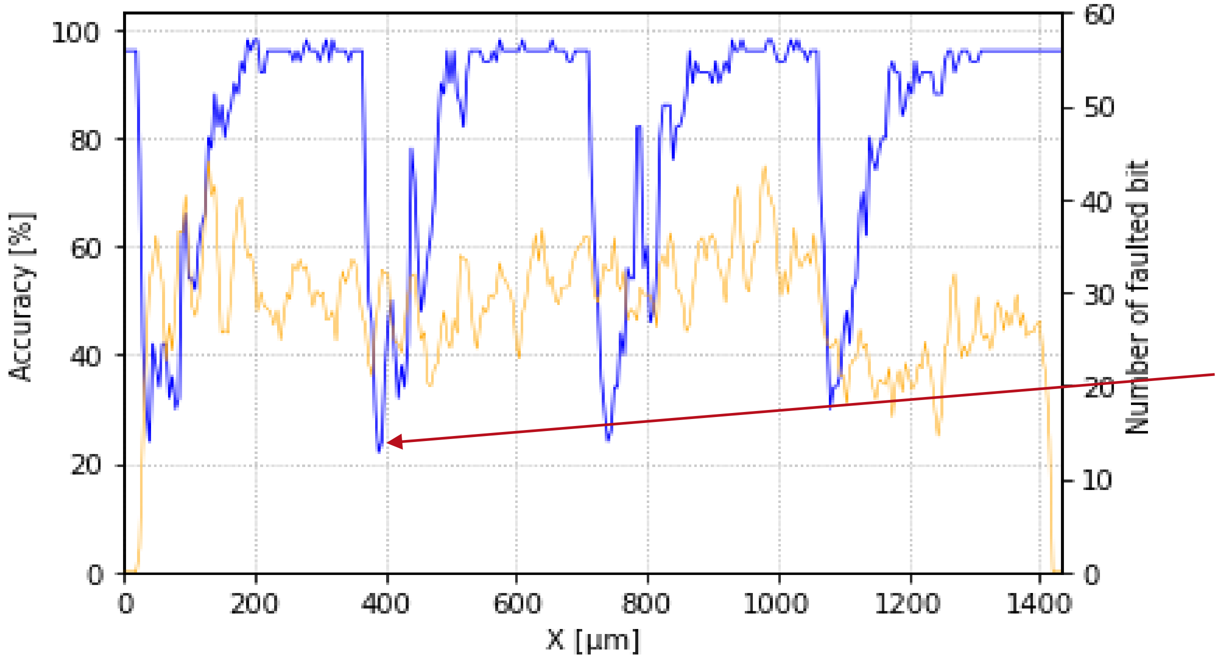
<u>For both lens :</u>
Optical Lens x5 (Spot of 15µm)
Pulse power : 170mW
Pulse Width : 200 ns
Delay : 500 ns
Step on X = 2µm

TARGETING THE MODEL INTEGRITY WITH LASER FAULT INJECTION

➢ Laser fault injection characterization on MNIST Model

- Robustness evaluation of **MNIST** MLP 8-bit model. 50 neurons on the targeted layer.
- Embedded accuracy : 96%
- **500 weights** are targeted. **100 inferences** are performed at each X-position.

Maximal Accuracy drop = 22%
Faults number average = 29 faults **(175mW)**



- ✓ Model precision can be significantly decrease on a deeper typical model.
- ✓ Drop of accuracy of 22% with 28 faults (0,6% of faulted bits)
- ✓ **Brute-force** attack strategy is limited.

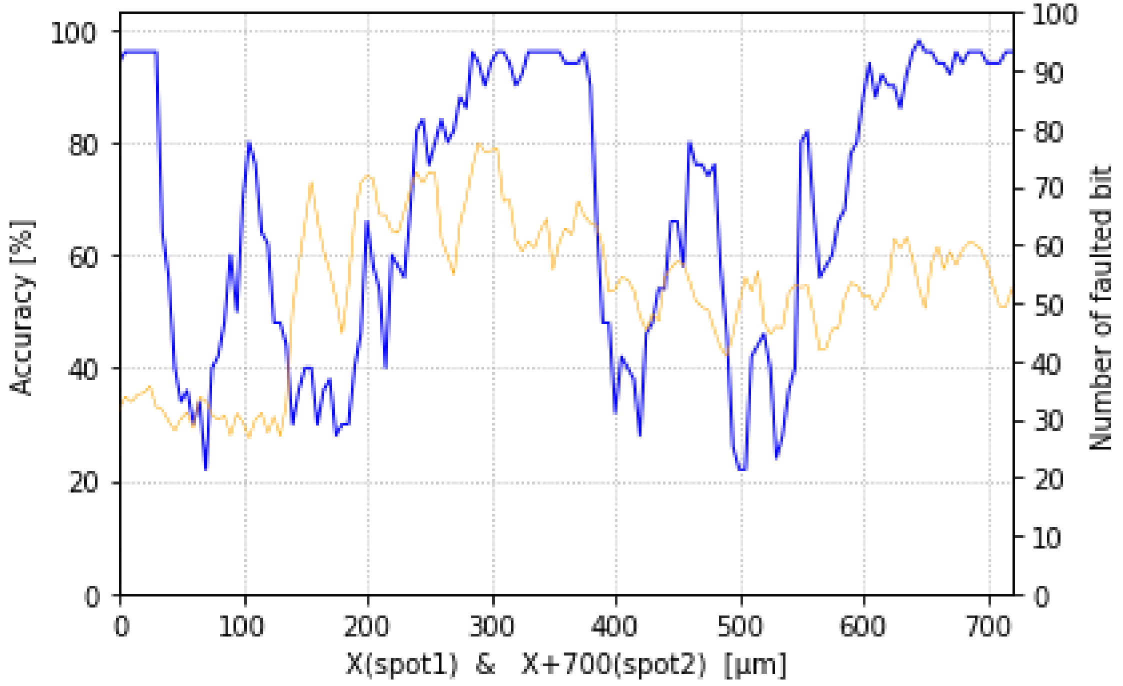
Optical Lens x5 (Spot of 15µm)
Pulse power : **140mW – 175mW**
Pulse Width : 200 ns
Delay : 930 ns
Step on X = 2µm

TARGETING THE MODEL INTEGRITY WITH LASER FAULT INJECTION

Bi-spot Laser fault injection characterization on MNIST Model

- Same experiment with both spots on the MNIST Model.

Maximal Accuracy drop = 22%
Faults number average = 53 faults (Bi-Spot)



✓ **Brute-force** attack strategy is limited.

For both lens :
Optical Lens x5 (Spot of 15µm)
Pulse power : ~170mW
Pulse Width : 200 ns
Delay : 930 ns
Step on X = 2µm

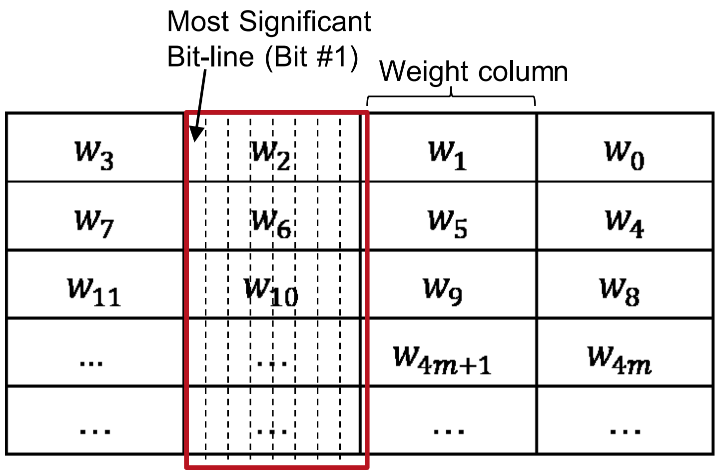
SUMMARY

- Context
- Setup and Fault Model
- Targeting the Model Integrity with Laser Fault Injection
- **Guided Laser Fault Injection**
- Conclusion

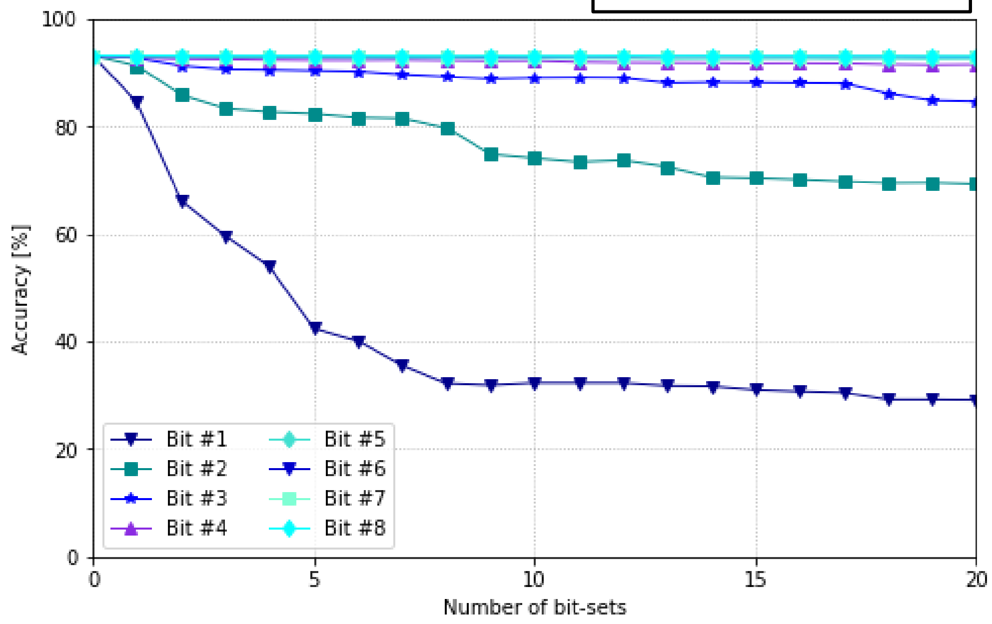
MODEL CHARACTERIZATION WITH GUIDED LASER FAULT INJECTION

➢ Simulation BSCA : Bit-Set Constrained Attack

- Based on BFA, the **most sensitive bits** of the model are identified.
- To be experimentally evaluate, bits (previously identified by BFA) are sorted by weights columns and bit lines.
- **Adversarial budget** is fixed to 20 bit-sets.
- All bit-lines from one weight column are targeted in **simulation**.



Same MNIST model



MODEL CHARACTERIZATION WITH GUIDED LASER FAULT INJECTION

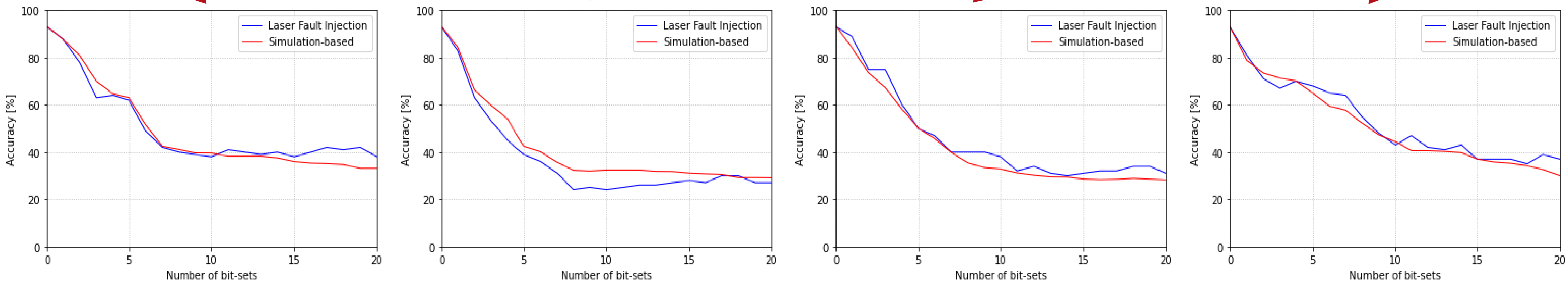
Experimental BSCA : Bit-Set Constrained Attack

- Laser shot is triggered only for the **selected** 20 weights, depending on the chosen weight column/bit-line.
- We target the MSB of each of the 4 weight columns, by changing the laser X-position.

Optical Lens x5 (Spot of 15µm)
Pulse power : **360 mW**
Pulse Width : 200 ns
Delay : 930 ns

w_3	w_2	w_1	w_0
w_7	w_6	w_5	w_4
w_{11}	w_{10}	w_9	w_8
...	...	w_{4m+1}	w_{4m}
...	...	...	...

0µm — Flash memory X-axis — 1400µm

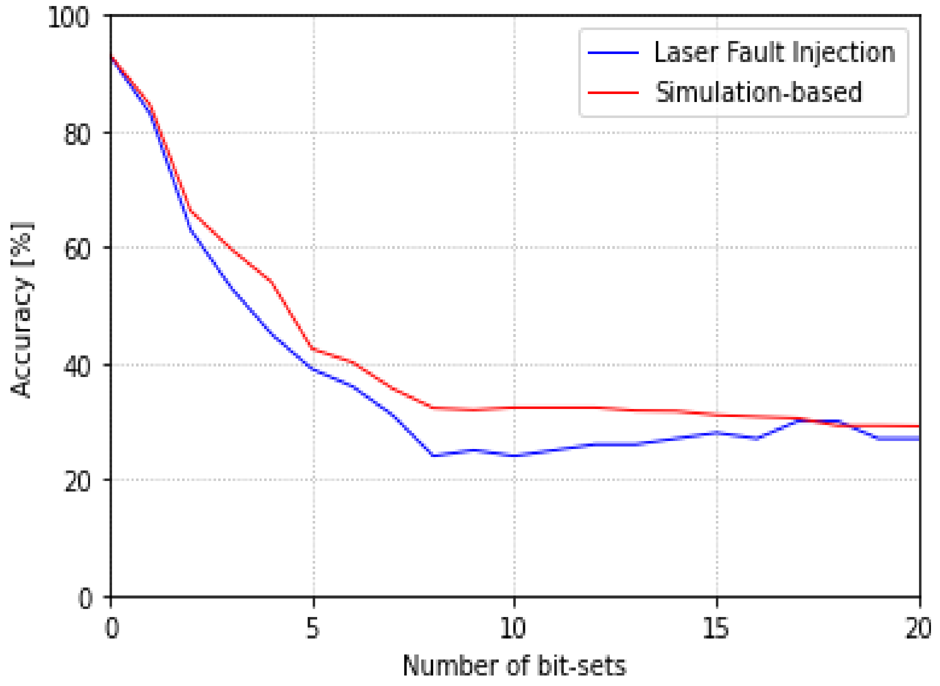


X = 30 µm — *X = 400µm* — *X = 760µm* — *X = 1060 µm*

MODEL CHARACTERIZATION WITH GUIDED LASER FAULT INJECTION

Experimental BSCA : Bit-Set Constrained Attack

- Laser shot is triggered only for the **selected** 20 weights, depending on the chosen weight column/bit-line.
- We target the MSB of each of the 4 weight columns, by changing the laser X-position.
- Focus on the MSB of the 2^{nd} weight column.



X = 400µm

- ✓ Experimental and simulation results are quite similar.
- ✓ 5 bit-sets (0,1% faulted bits) accuracy drops to 39%. 10 bits-sets : 24%.
- ✓ After 10 bit-sets accuracy not decrease → model level of robustness

CONCLUSION

- **Integrity evaluation** of embedded neural network is still in its **infancy.**
  - Laser injection and bit-set fault model are powerful means to assess the **robustness** of an embedded model.
- First **experimental** characterization of **weight-based adversarial** attack with a laser fault injection.
- With **bi-spot laser** characterization, more weights can be faulted in the same inference.
- With the Bit-Set Constraint Attack we can **guide** the laser fault injection.
  - **High accordance** between simulation and practical results.
  - Only **few bits** are necessary to significantly decrease the model's accuracy.
- Basis for developing reliable evaluation methodology for future standardization and certification schemes of embedded AI-system.

ONGOING WORKS

- Robustness characterization on Convolutional Neural Network.
- Other attack vectors (Instructions, activation functions…).
- Evaluate state-of-the-art defense strategies against fault injection in a ML model context.
- Model reverse engineering with fault injection.

THANK YOU

JAIF 2022

INTEGRITY CHARACTERIZATION OF EMBEDDED NEURAL NETWORK AGAINST LASER FAULT INJECTION

Mathieu DUMONT // CEA LETI // mathieu.dumont@cea.fr

CEA-Leti, technology research institute
Commissariat à l'énergie atomique et aux énergies alternatives
Minatec Campus | 17 avenue des Martyrs | 38054 Grenoble Cedex | France
www.leti-cea.com

REFERENCES

- **[1]** Y. Liu, L. Wei, B. Luo, and Q. Xu, *Fault injection attack on deep neural network*, IEEE/ACM International Conference on Computer- Aided Design, Digest of Technical Papers, ICCAD, 2017.
- **[2]** A. S. Rakin, Z. He, and D. Fan, *Bit-flip attack: Crushing neural network with progressive bit search,* in IEEE International Conference on Computer Vision, 2019.
- **[3]** J. Breier, X. Hou, D. Jap, L. Ma, S. Bhasin, and Y. Liu, *Practical fault attack on deep neural networks,* in Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, 2018.
- **[4]** F. Benevenuti, F. Libano, V. Pouget, F. L. Kastensmidt, and P. Rech, *Comparative analysis of inference errors in a neural network implemented in sram-based fpga induced by neutron irradiation and fault injection methods*, in 31st Symposium on Integrated Circuits and Systems Design SBCCI, 2018.
- **[5]** F. Yao, A. S. Rakin, and D. Fan*, DeepHammer: Depleting the Intelligence of Deep Neural Networks through Targeted Chain of Bit Flips*. 29th USENIX Security Symposium, 2020.
- **[6]** W. Liu, C.-H. Chang, F. Zhang, and X. Lou, *Imperceptible misclassification attack on deep learning accelerator by glitch injection*, Design Automation Conference DAC, 2020.
- **[7]** Y. Fukuda, K. Yoshida, and T. Fujino, *Fault injection attacks utilizing waveform pattern matching against neural networks processing on microcontroller*, Transactions on Fundamentals of Electronics, Communications and Compute Sciences, 2022.
- **[8]** B. Colombier, A. Menu, J. M. Dutertre, P. A. Moellic, J. B. Rigaud, and J. L. Danger, *Laser-induced Single-bit Faults in Flash Memory: Instructions Corruption on a 32-bit Microcontroller*, IEEE International Symposium on Hardware Oriented Security and Trust, HOST, 2019.