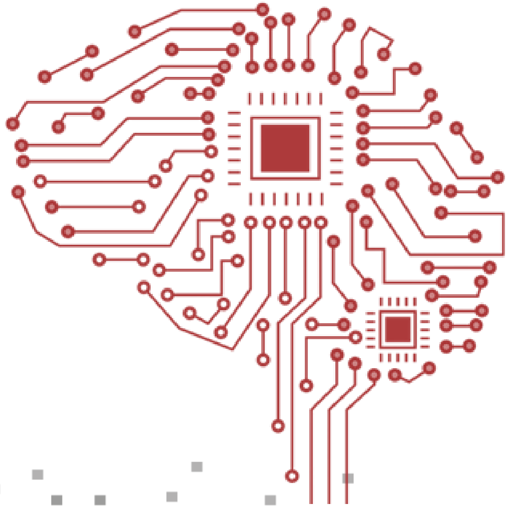
CryptArchi 2022



LASER FAULT INJECTION AGAINST EMBEDDED NEURAL NETWORK MODEL

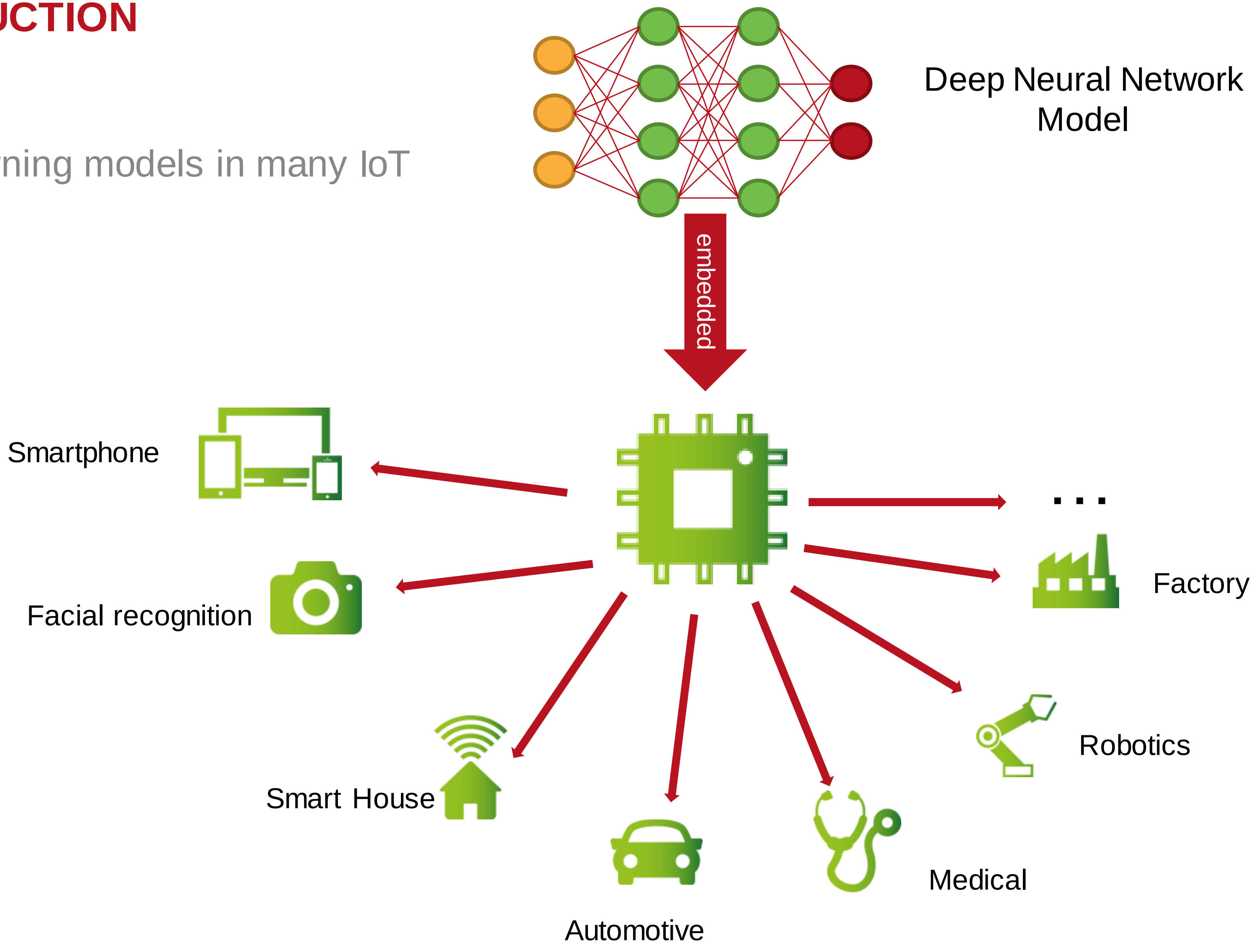
Mathieu Dumont, Pierre Alain Moëllic, Jean-Max Dutertre, Raphaël Viera

mathieu.dumont@cea.fr

30/05/2022

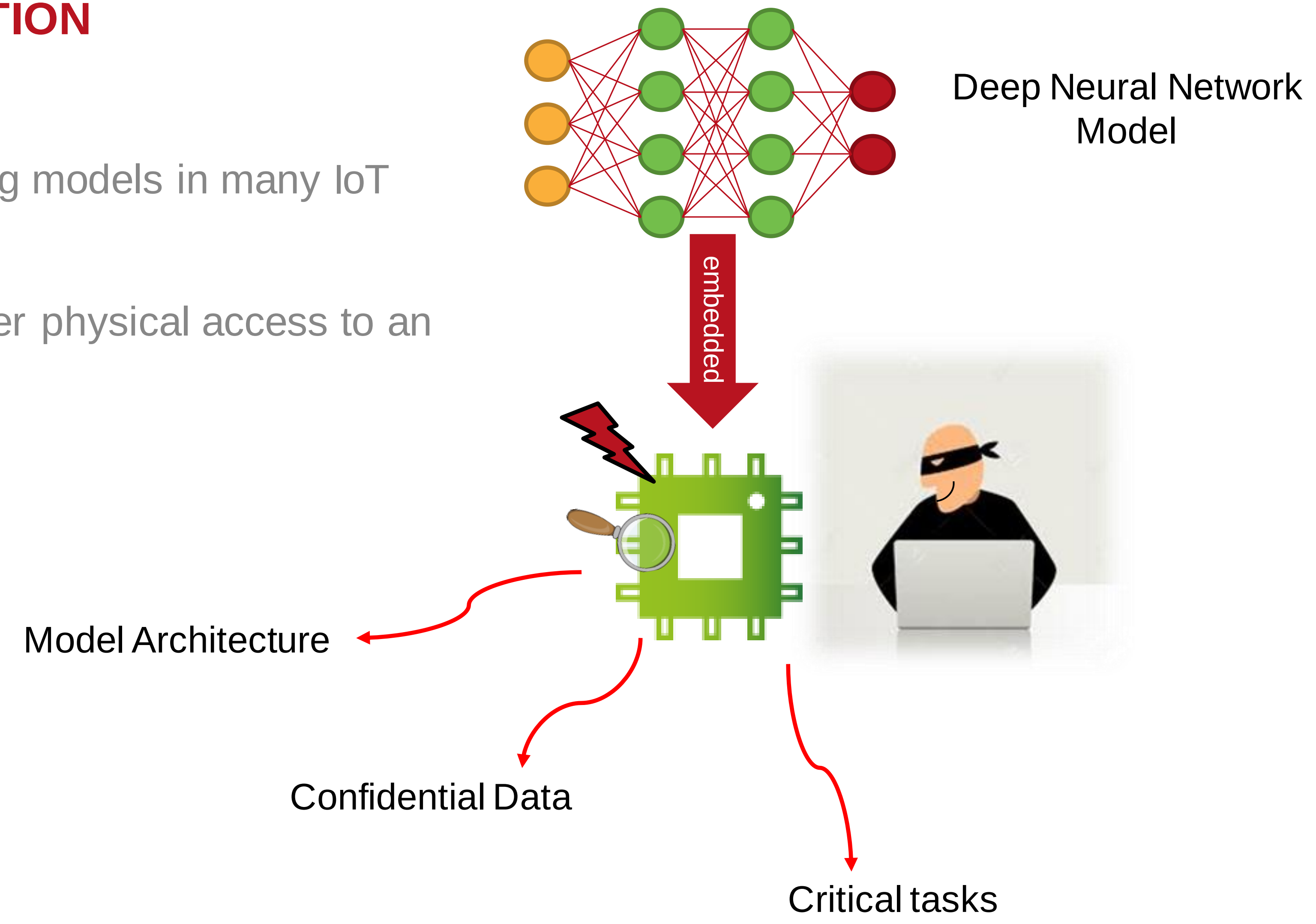
INTRODUCTION

- 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

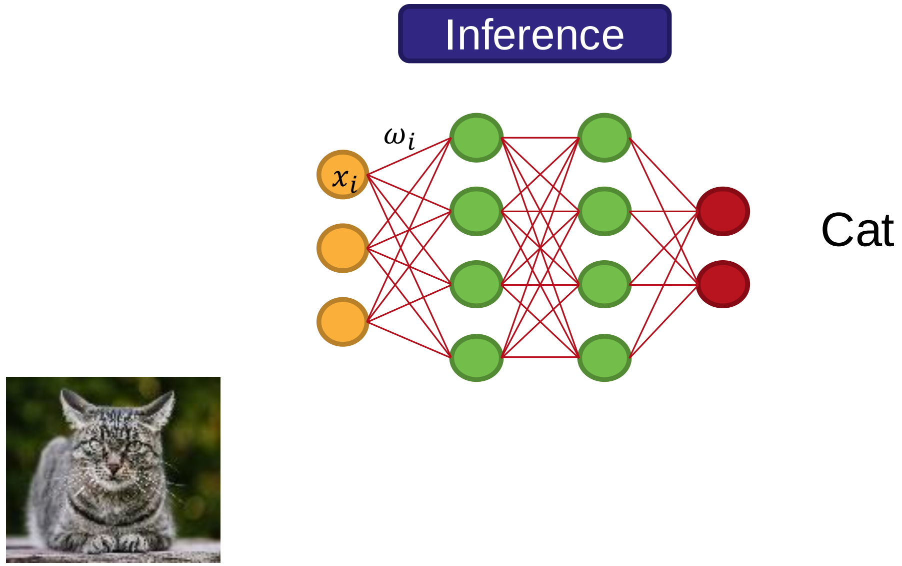


- Context
- Bit-set fault model
- Laser Fault Injection on embedded neural network
- Conclusion

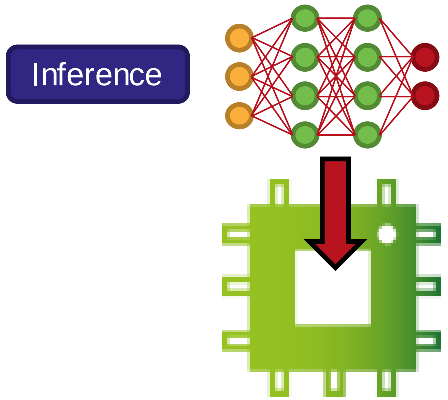
CONTEXT

Attack on machine learning models

- Adversarial Example (software attack) is a major threat against DNN. **Massive research efforts** on that field.



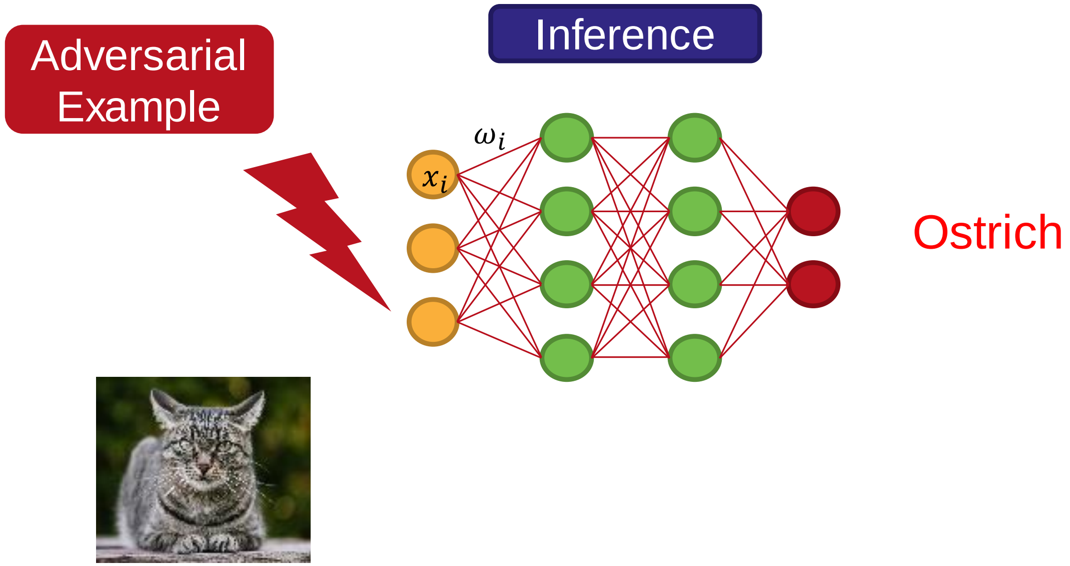
- Physical attacks (hardware attack) constitute new threats against DNN. **Upcoming works.**



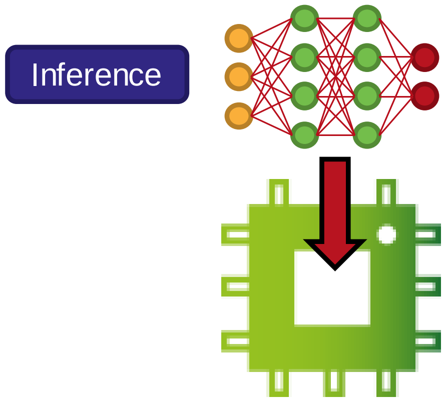
CONTEXT

Attack on machine learning models

- Adversarial Example (software attack) is a major threat against DNN. **Massive research efforts** on that field.



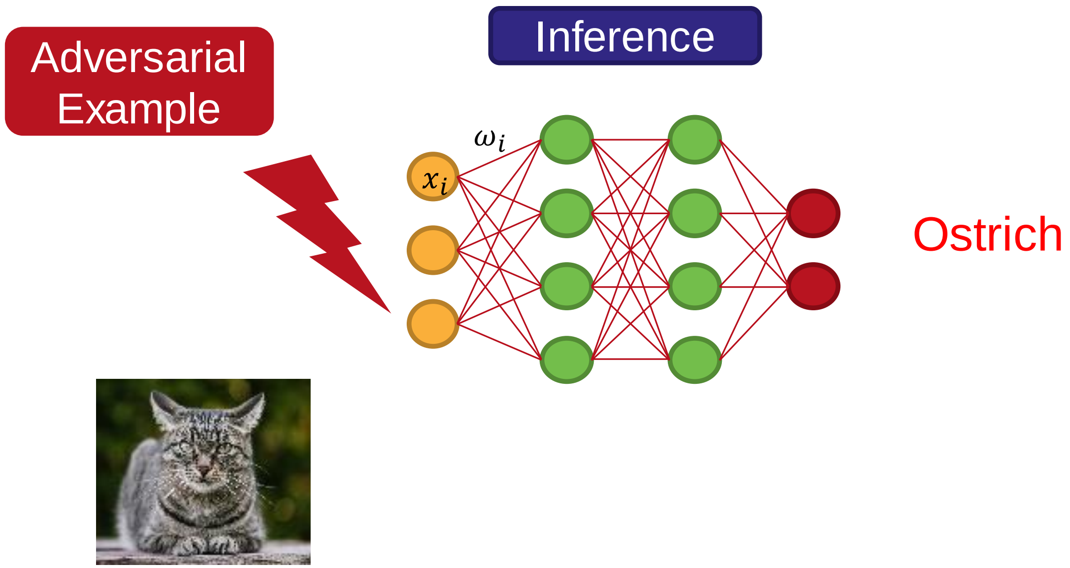
- Physical attacks (hardware attack) constitute new threats against DNN. **Upcoming works.**

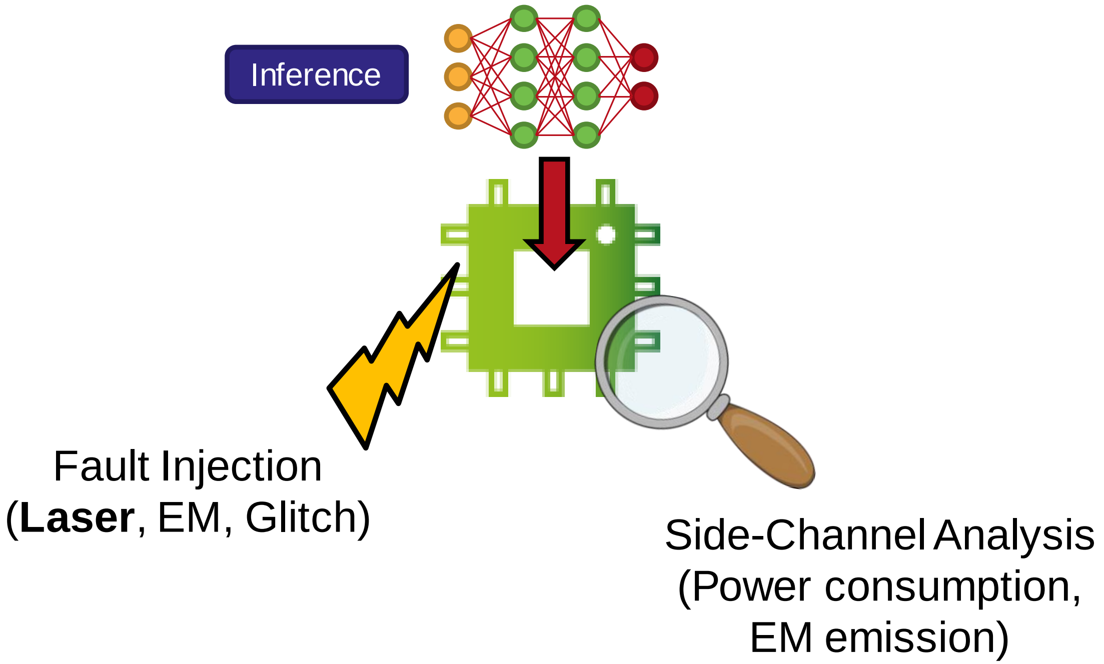


CONTEXT

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. **Upcoming works.**



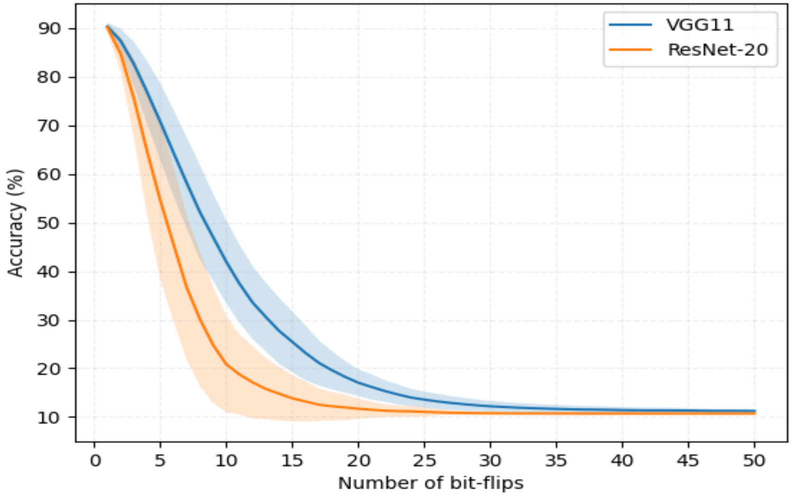


CONTEXT

➢ **State of the Art of Fault Injection on embedded Neural Network**

ATTACK AGAINST **INTEGRITY**

- Simulation **parameter-based** attack :
  - First in 2017 [1], Single Bias Attack & Gradient Descent Attack.
  - **Bit-Flip Attack (BFA)** by *Rakin et al.* [2] with Progressive Bit Search method.
- Physical Fault injection on network **activation function** :
  - **Laser Fault Injection** by *Jakub Breier et al* [3].
  - **Clock Glitching** [4].
- **RowHammer** attack by *Rakin et al.* [5]



Bit-Flip Attack simulation

CONTEXT

State of the Art of Fault Injection on embedded Neural Network

ATTACK AGAINST **CONFIDENTIALITY**

- Only one model reverse engineering method with fault injection: SNIFF [6], *Breier et al.*
  - Parameters recovery of the last layer only.
  - Need to know all previous parameters.

- As AES key recovery, **Machine Learning model data reverse** will be soon a critical topic

CONTEXT

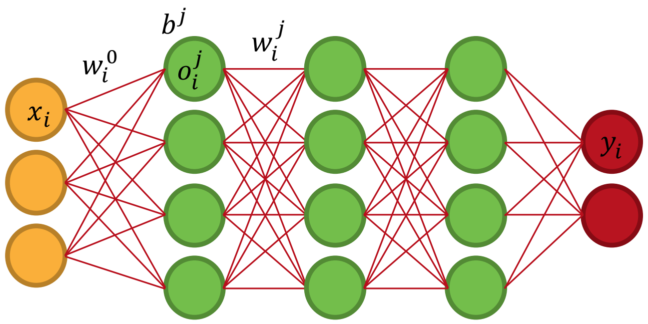
- **State of the Art of Fault Injection on embedded Neural Network**

→ Focus on robustness characterization

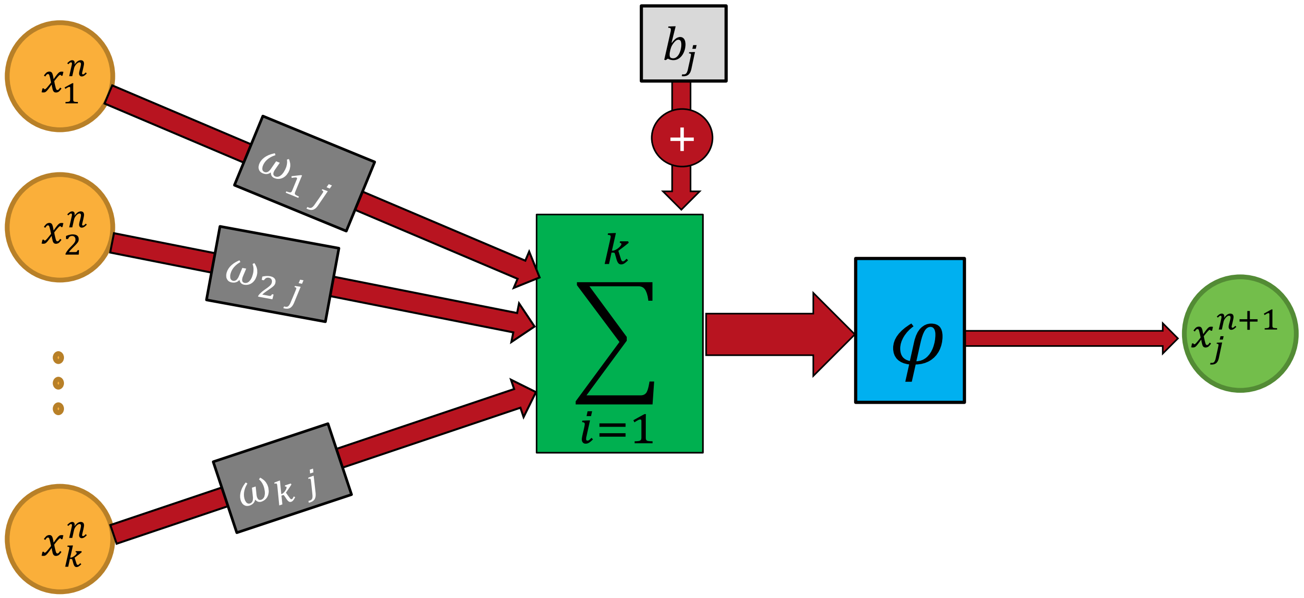
→ Fault on quantified networks

→ Stealthy and precise attack with minimum faults

CONTEXT



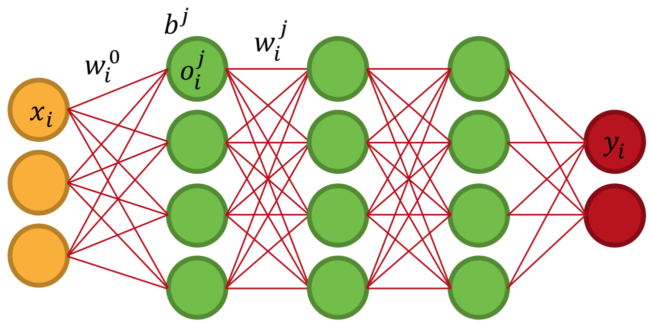
- **Which parameters to target on NN ?**
  - Typical neuron computation:

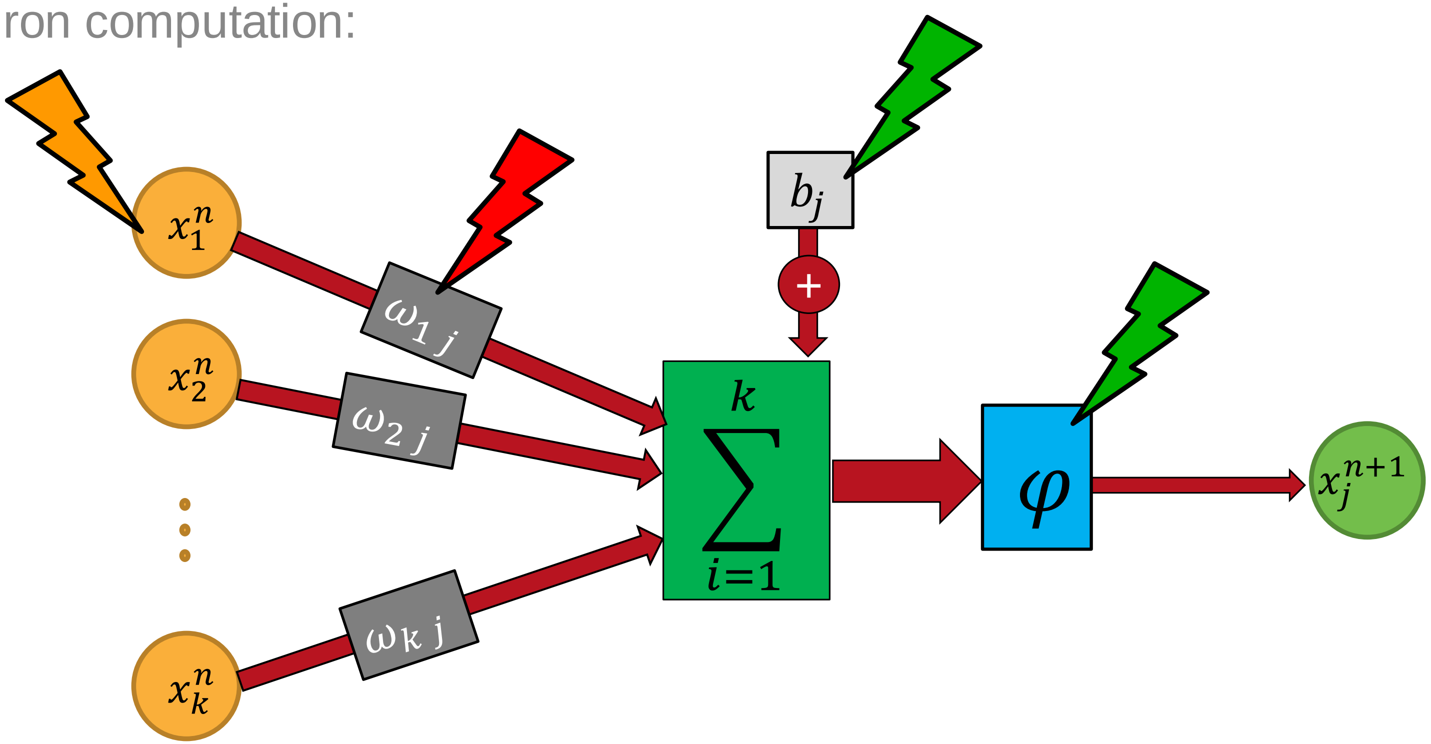


CONTEXT

➢ **Which parameters to target on NN ?**

- Typical neuron computation:





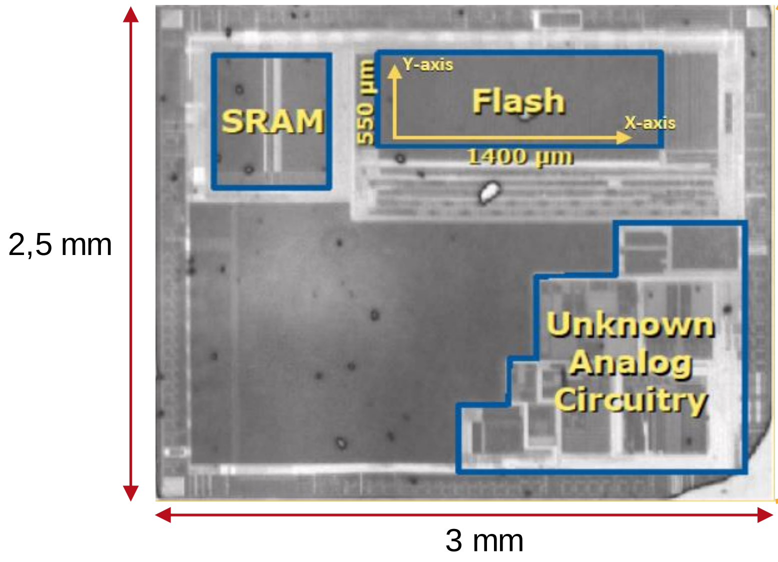
SUMMARY

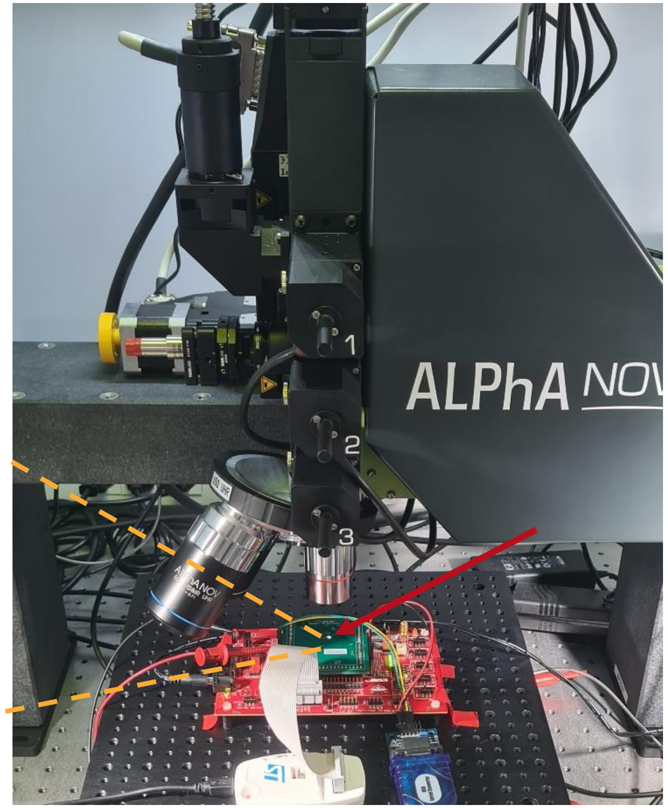
- Context
- Bit-set fault model
- Laser Fault Injection on embedded neural network
- Conclusion

BIT-SET 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

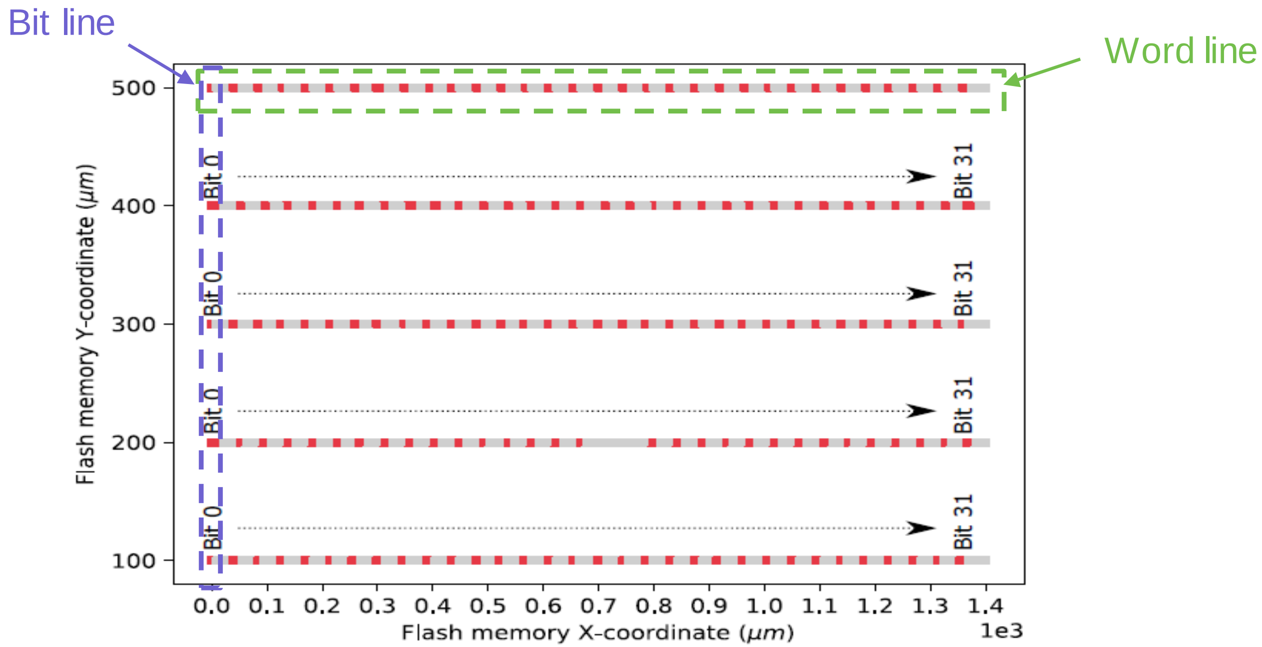




BIT-SET FAULT MODEL

➢ Bit-set fault model [7] at the read time

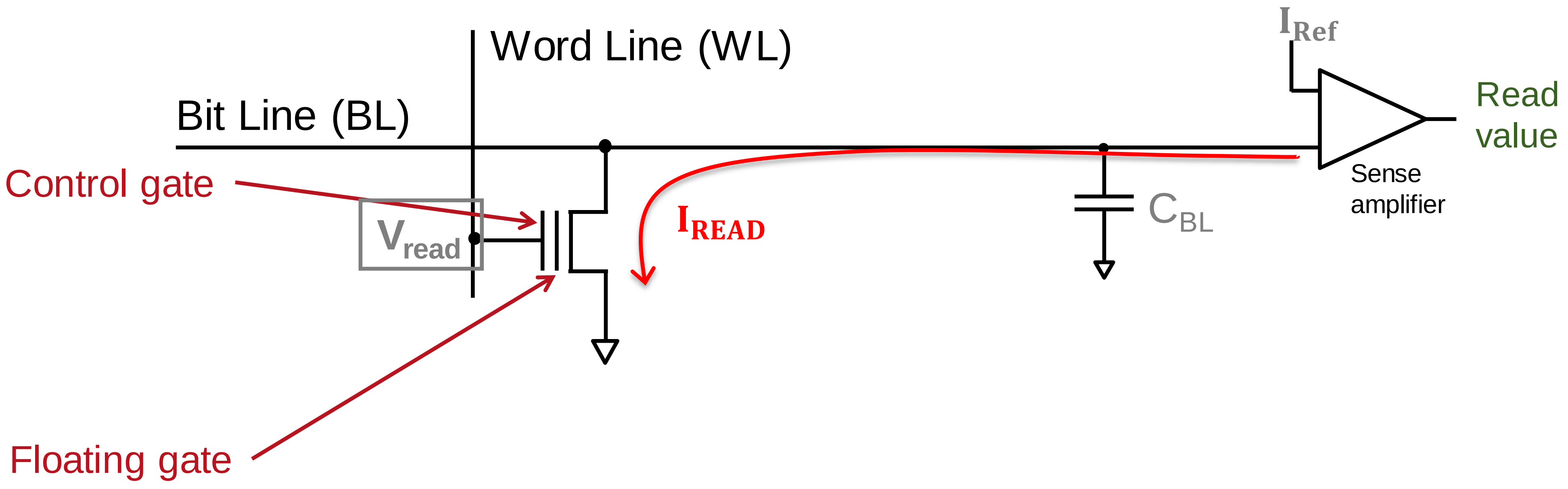
- Flash memory : constituted of Bit lines (common to all registers) and Word lines (32 bits register).
- A 32-bits word (with only 0) is loaded from the Flash memory and stored in r0 register. Shot at the “ldr” instruction.
- Every bit, from 0 to 31, is forced to 1 one after another, along the X-axis. No difference on Y-axis.



Optical Lens x5 (Spot of 15µm)
Pulse power : 200 mA (~120mW)
Pulse Width : 200 ns

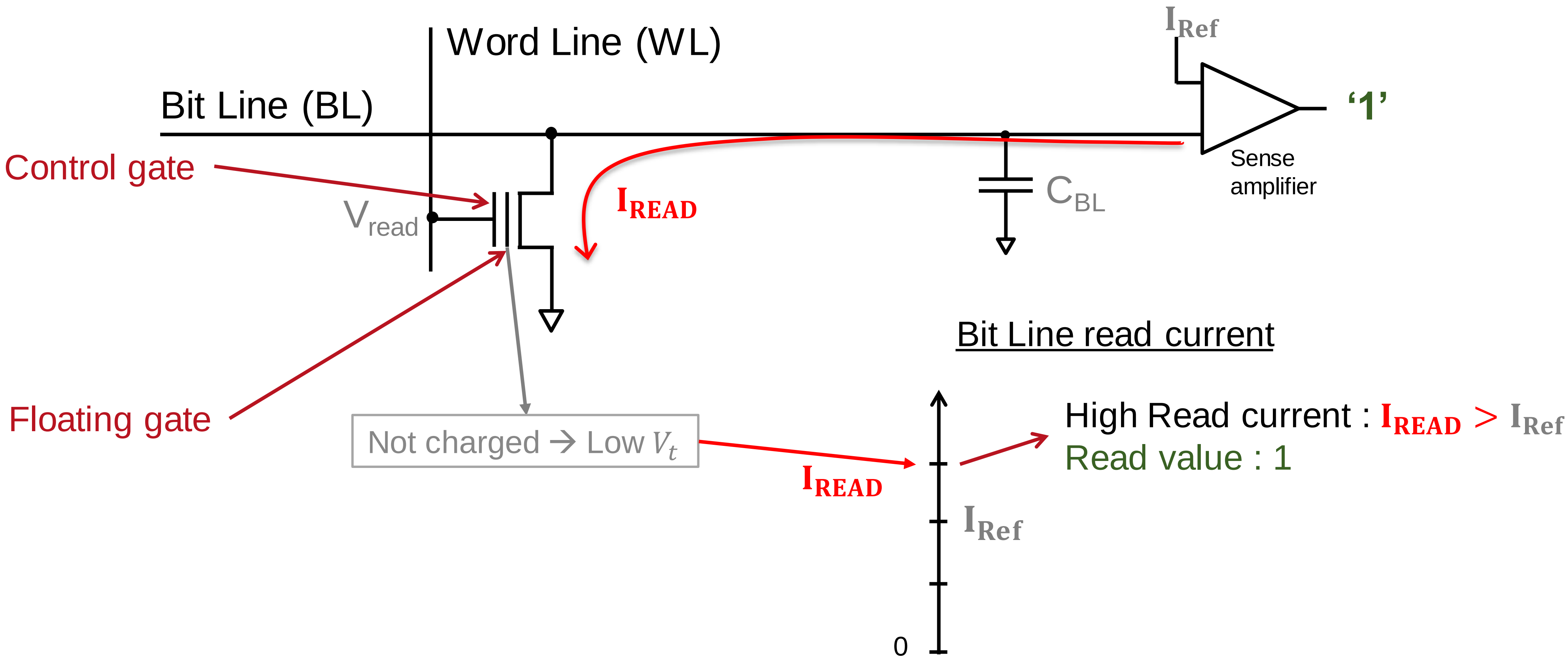
BIT-SET FAULT MODEL

- Read operation explanation



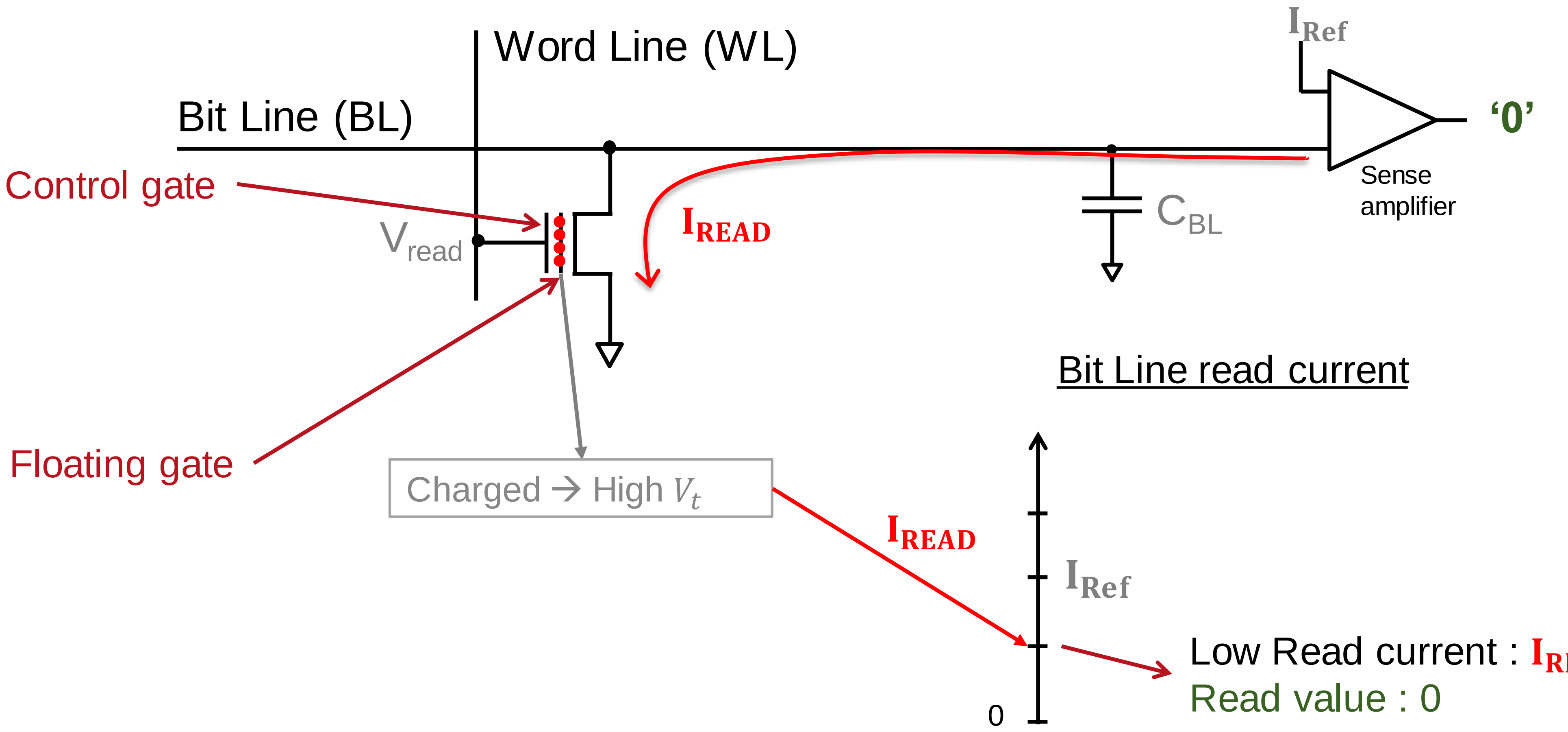
BIT-SET FAULT MODEL

➢ Read operation explanation



BIT-SET FAULT MODEL

➢ Read operation explanation

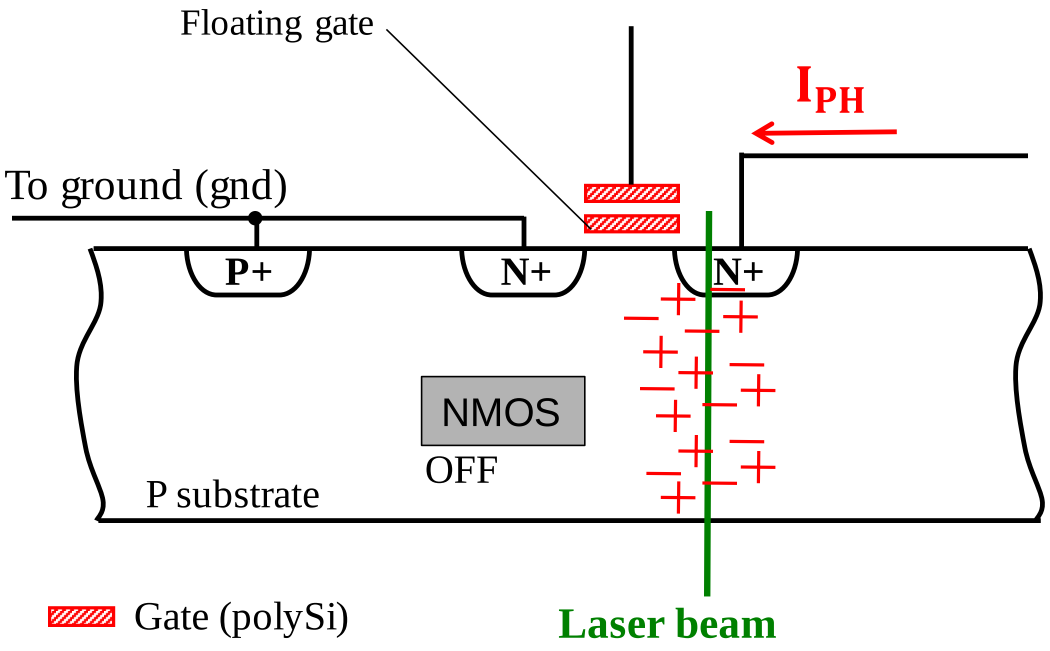


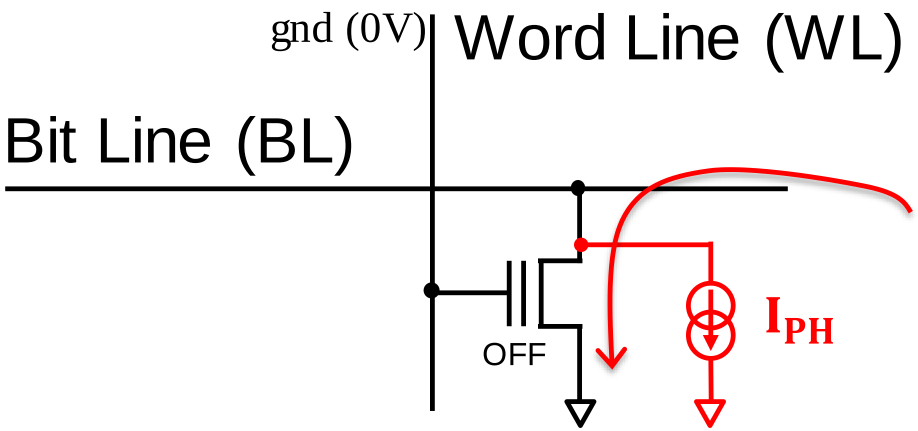
Low Read current : $I_{READ} < I_{Ref}$

Read value : 0

BIT-SET FAULT MODEL

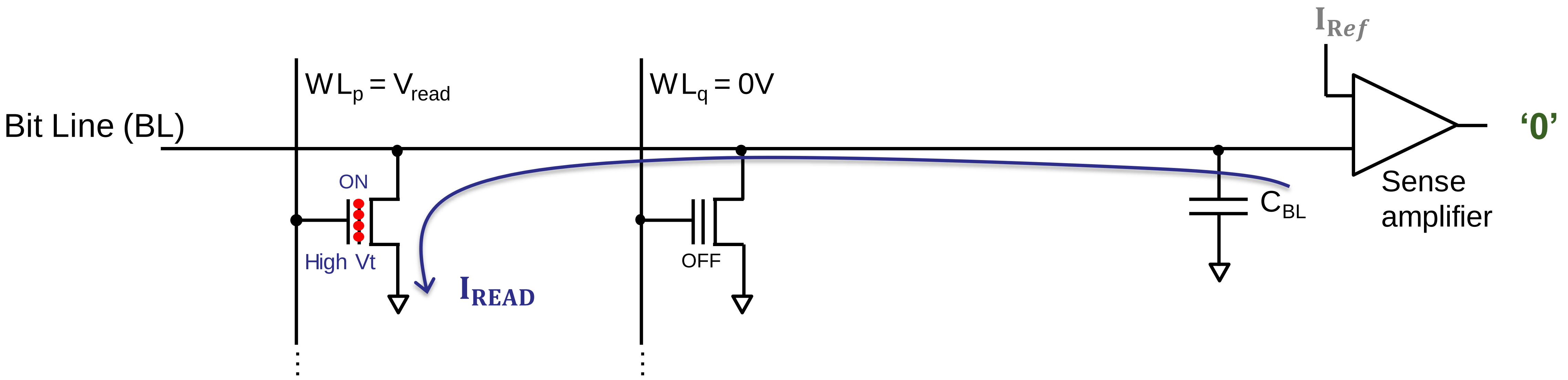
- Effect of Laser shot on Floating Gate NMOS explanation





BIT-SET FAULT MODEL

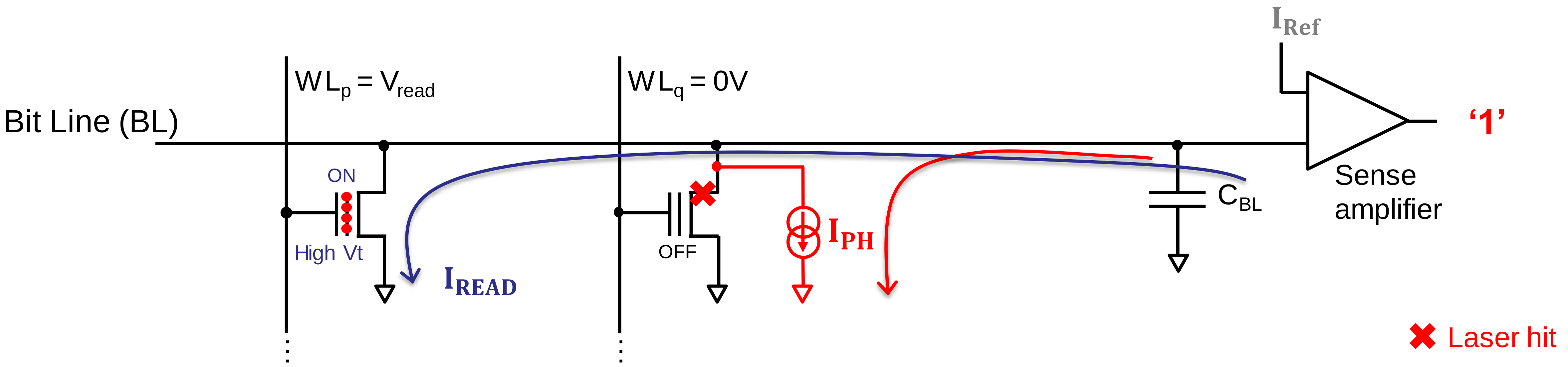
Bit-set fault model explanation



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

BIT-SET FAULT MODEL

➢ Bit-set fault model explanation



- 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

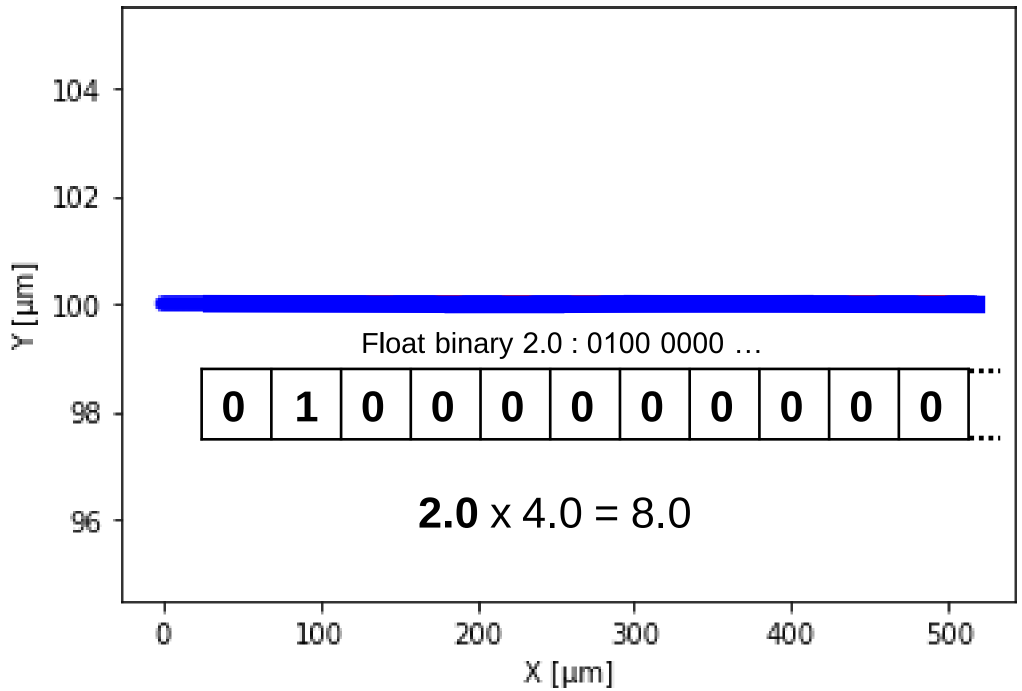
→ Bit-set fault model

BIT-SET FAULT MODEL

Application of bit-set fault model on a float multiplication

- Parallel with neural network multiplication $\left(w_i^j . x_i\right)$ with weight w = 2.0 and input x = 4.0.
- Laser shot during the load (`ldr`) instruction of the “weight” value, before the float multiplication.

```
ldr     r5, [r6, #0]
mov     r1, r5
mov     r0, r9
bl      8000210 <__aeabi_fmul>
str     r0, [sp, #20]
```

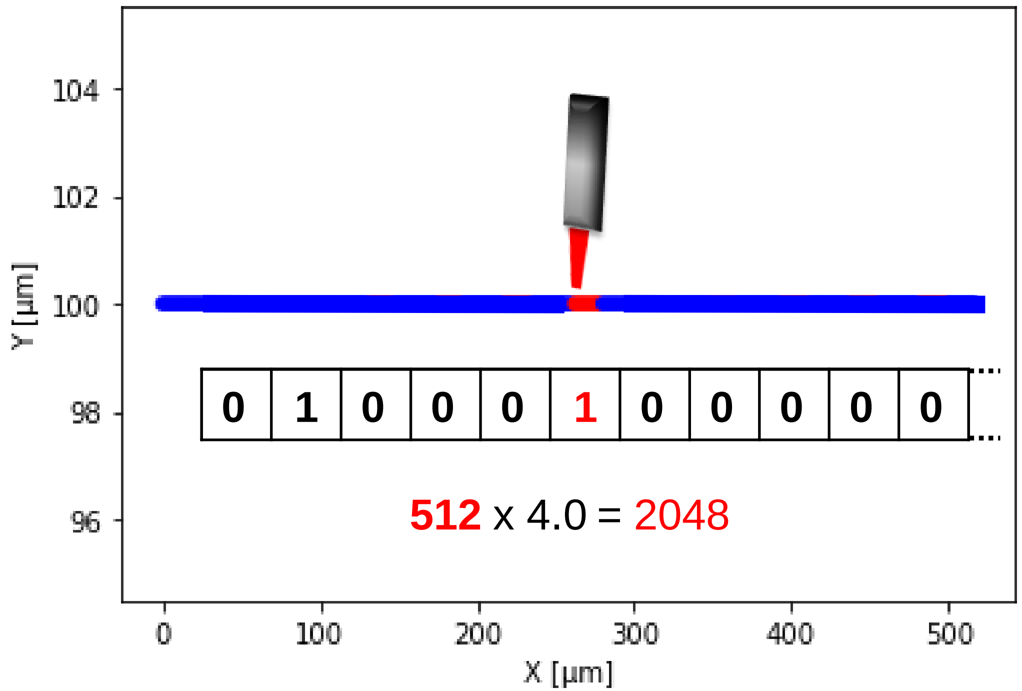


BIT-SET FAULT MODEL

Application of bit-set fault model on a float multiplication

- Parallel with neural network multiplication $\left(w_i^j . x_i\right)$ with weight w = 2.0 and input x = 4.0.
- Laser shot during the load (`ldr`) instruction of the "weight" value, before the float multiplication.

```
ldr          r5, [r6, #0]
mov          r1, r5
mov          r0, r9
bl           8000210 <__aeabi_fmul>
str          r0, [sp, #20]
```



Optical Lens x20 (Spot of 3µm)
Pulse power : 200 mA (~120mW)
Pulse Width : 200 ns
Delay : 1000 ns

- ✓ The bit-set fault model could induce huge value variation
- ✓ Bit-set is induced on every bit all along the X-axis.

SUMMARY

- Context
- Bit-set fault model
- Laser Fault Injection on embedded neural network
- Conclusion

➢ The targeted Neural Network

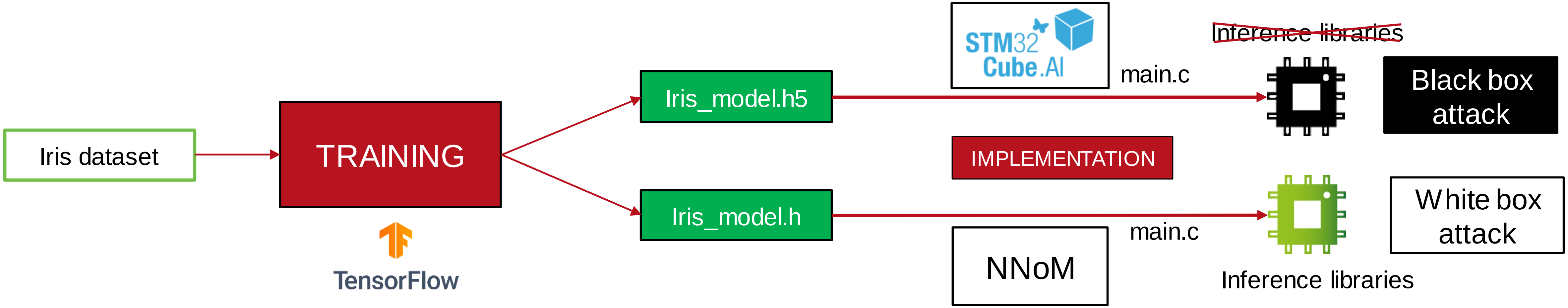
- Iris NN: small network, 4 inputs and 3 outputs
  - Multi-Layer Perceptron (Fully-Connected neural network)
  - Only few neurons and one hidden layer is sufficient.

LASER FAULT INJECTION ON EMBEDDED NEURAL NETWORK

The targeted Neural Network

- Iris NN: small network, 4 inputs and 3 outputs
  - Multi-Layer Perceptron (Fully-Connected neural network)
  - Only few neurons and one hidden layer is sufficient

- Need access to inference computation libraries



LASER FAULT INJECTION ON EMBEDDED NEURAL NETWORK

➢ The targeted Neural Network

- Iris NN: small network, 4 inputs and 3 outputs
  - Multi-Layer Perceptron (Fully-Connected neural network)
  - Only few neurons and one hidden layer is sufficient

- Need access to inference computation libraries

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

Multiplication loop $w_i^j . x_i$ during inference in a fully-connected layer.

C code

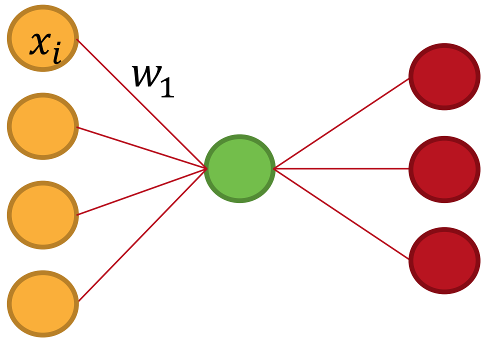
```
for (int j = 0; j < dim_vec; j++) {
        q7_t inA = *pA++;
        q7_t inB = *pB++;
        ip_out += inA * inB; }
```

Load of weight value

Assembly

```
ldr     r3, [r7, #72]
adds    r2, r3, #1
str     r2, [r7, #72]
ldrb    r3, [r3, #0]
strb    r3, [r7, #30]
```

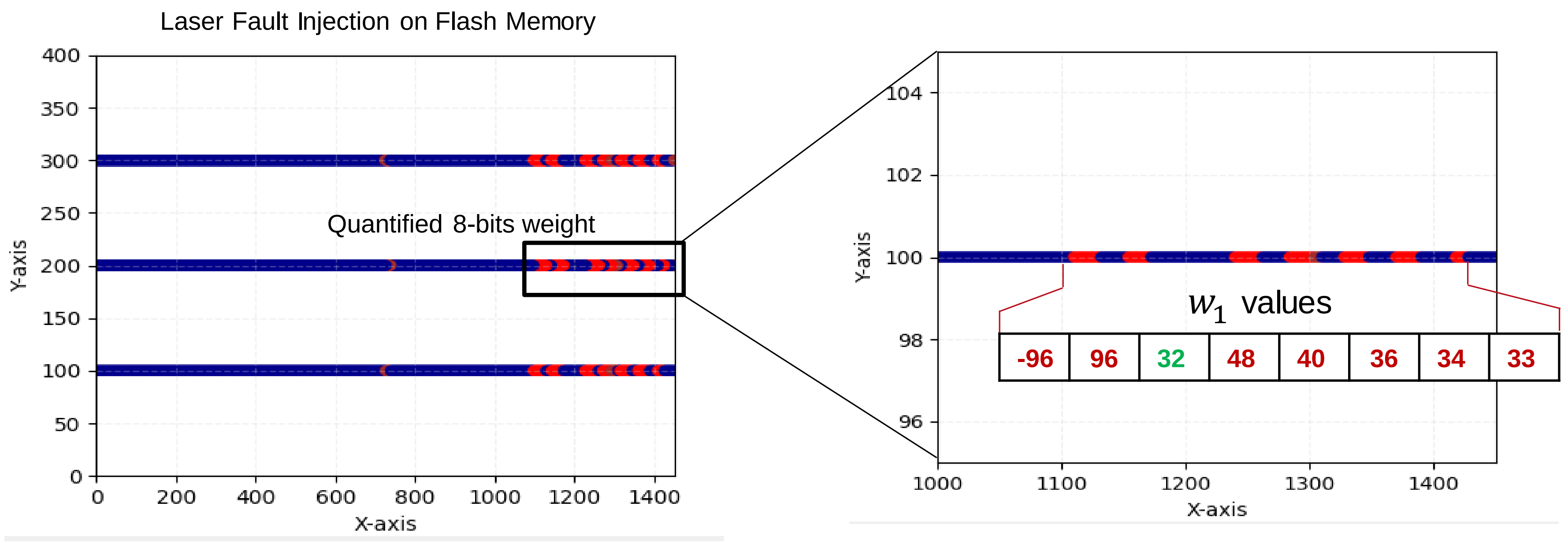
LASER FAULT INJECTION ON EMBEDDED NEURAL NETWORK



➢ Laser fault injection characterization on one weight

- A **laser shot** is induced during the load of only **one** of the four weights

w_1 = 32 $(00100000)_2$

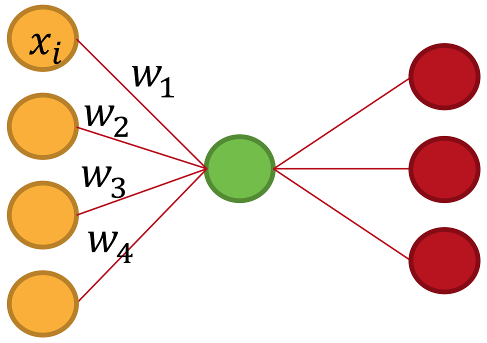


✓ *Bit-sets* induced on the weight of an embedded neural network.

✓ Variation of the weight value depending on the laser spot position on the X-axis.

Optical Lens x5 (Spot of 15µm)
Pulse power : 300 mA (~170mW)
Pulse Width : 200 ns
Delay : 500 ns
Step on X = 2µm

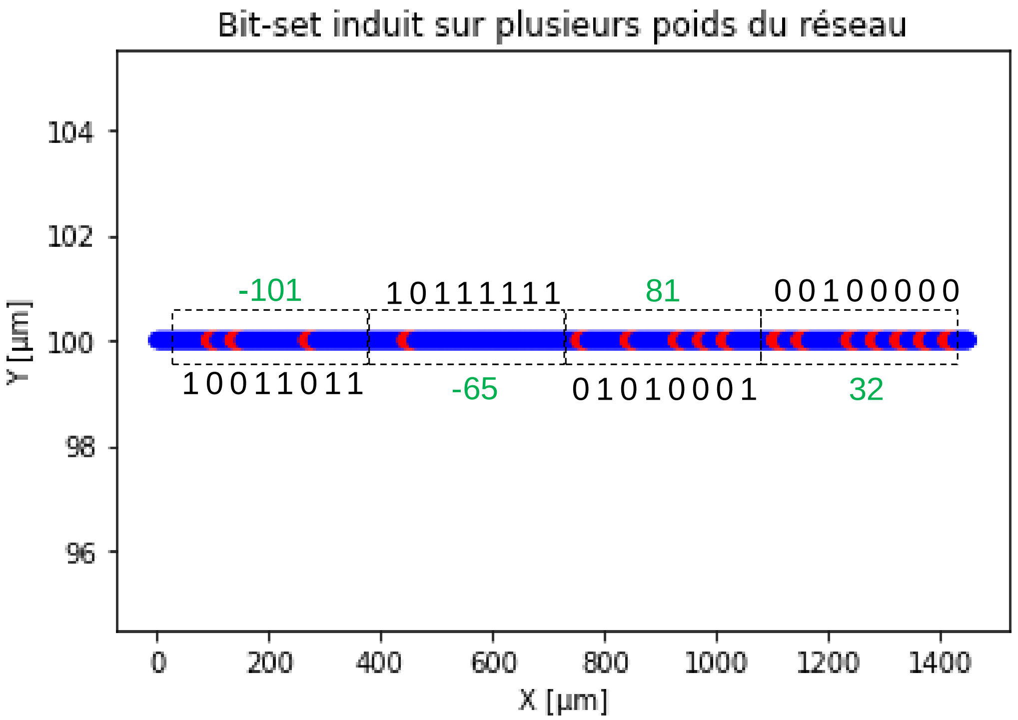
LASER FAULT INJECTION ON EMBEDDED NEURAL NETWORK



Laser fault injection characterization on several weights

- A **laser shot** is induced during the load of **every** weight from neurons

w_i = [32, 81, -65, -101]



Optical Lens x5 (Spot of 15µm)
Pulse power : 300 mA (~170mW)
Pulse Width : 200 ns
Delay : 500 ns

✓ Every weights of the network could be **precisely** faulted

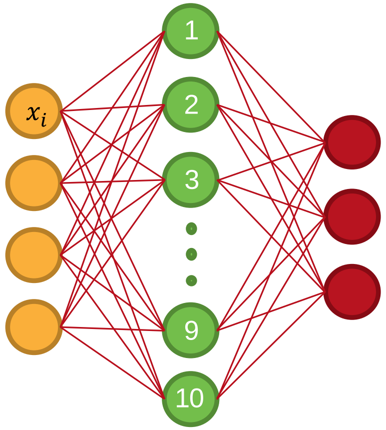
✓ With **bi-spot** we can induce 2 bit-sets at the same time:

→ on 1 weight:

Example: $w_2 = 81\ (b'01010001)$
After bi-spot attack;
$w_2 = 241\ (b'11110001)$

→ On 2 different weights

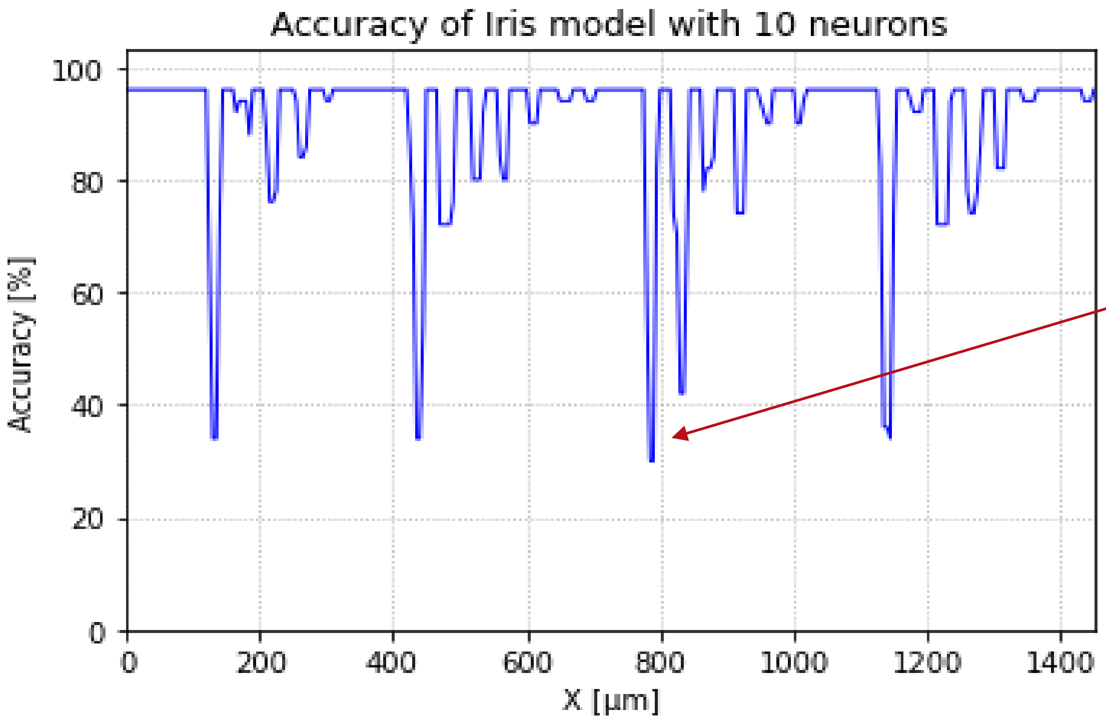
LASER FAULT INJECTION ON EMBEDDED NEURAL NETWORK



➢ Neural network robustness characterization against Laser Fault Injection

- 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.

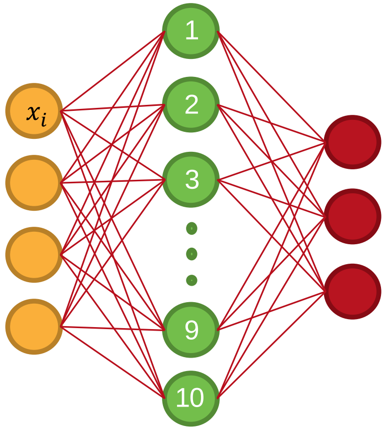
- Accuracy of embedded model without attack = 93%
- Total number of bits = 320bits



✓ Drop accuracy to 30%.

Optical Lens x5 (Spot of 15µm)
Pulse power : 300 mA (~170mW)
Pulse Width : 200 ns
Delay : 500 ns
Step on X = 2µm

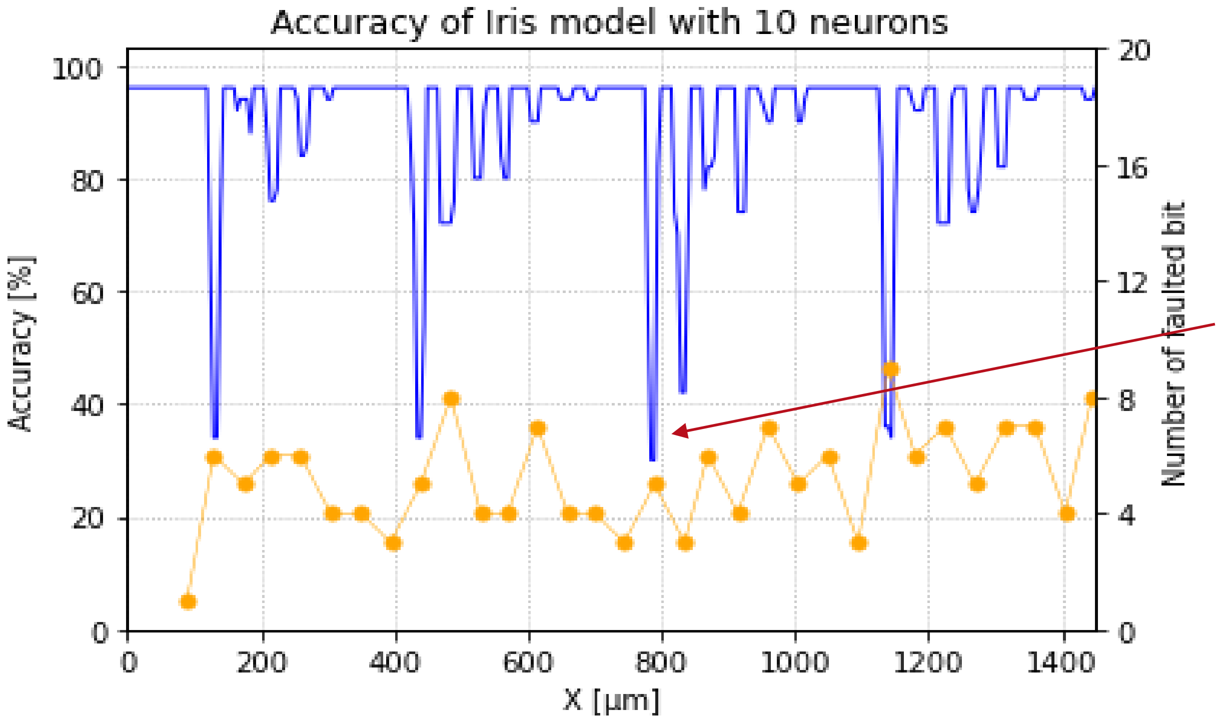
LASER FAULT INJECTION ON EMBEDDED NEURAL NETWORK



➢ Neural network robustness characterization against Laser Fault Injection

- 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.

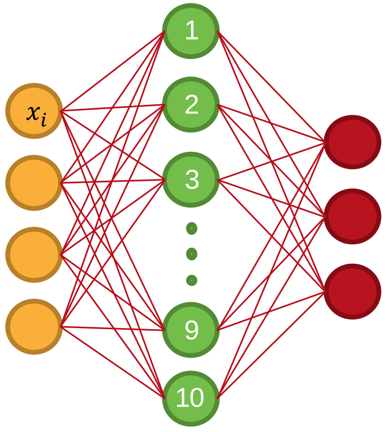
- Accuracy of embedded model without attack = 93%
- Total number of bits = 320bits



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

Optical Lens x5 (Spot of 15µm)
Pulse power : 300 mA (~170mW)
Pulse Width : 200 ns
Delay : 500 ns
Step on X = 2µm

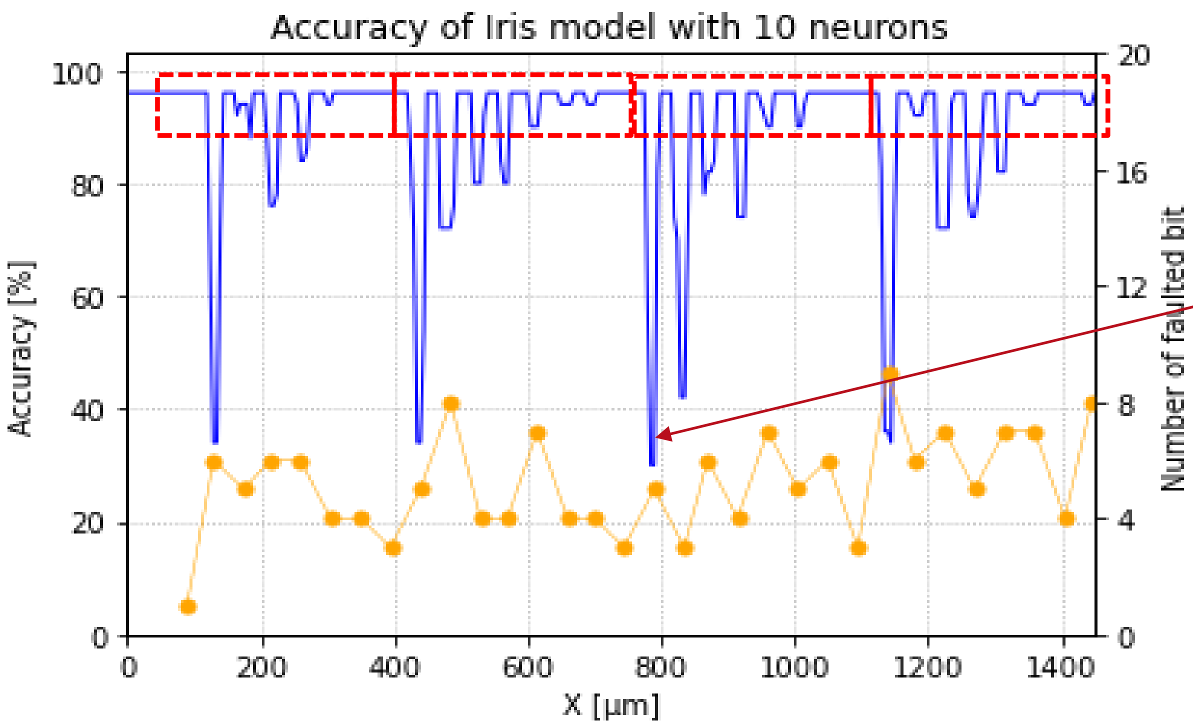
LASER FAULT INJECTION ON EMBEDDED NEURAL NETWORK



➢ Neural network robustness characterization against Laser Fault Injection

- 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.

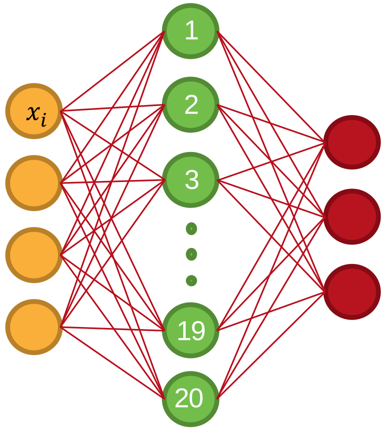
- Accuracy of embedded model without attack = 93%
- Total number of bits = 320bits



✓ Drop accuracy to **30%**, with only **5 faulted bit.**

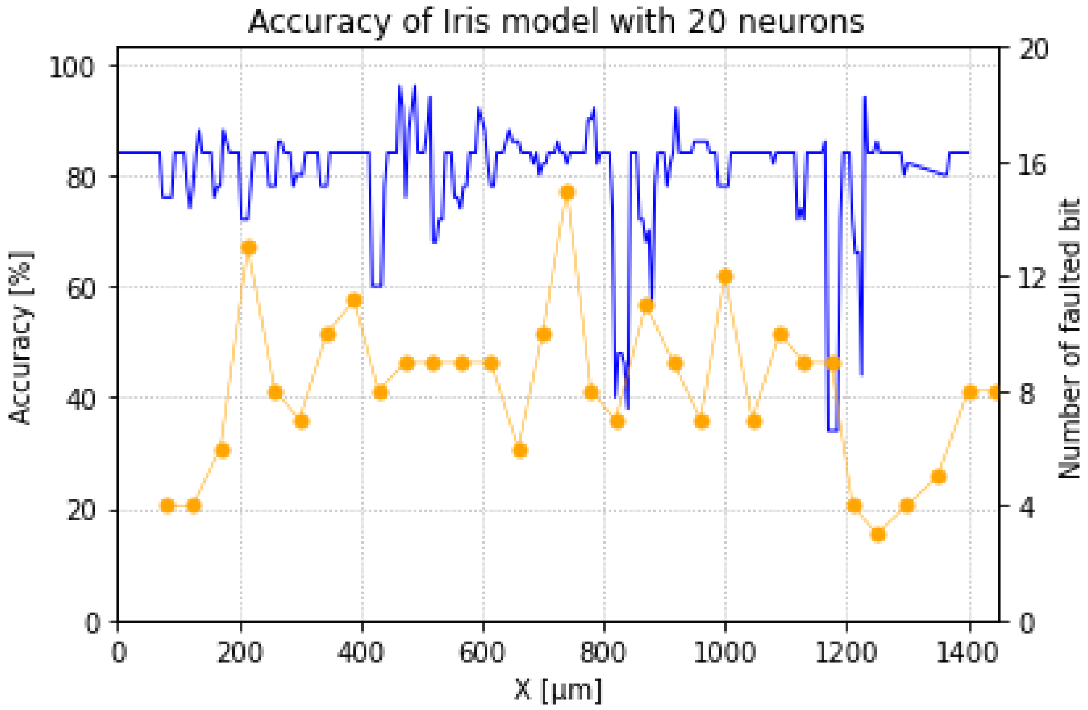
Optical Lens x5 (Spot of 15µm)
Pulse power : 300 mA (~170mW)
Pulse Width : 200 ns
Delay : 500 ns
Step on X = 2µm

LASER FAULT INJECTION ON EMBEDDED NEURAL NETWORK



➢ Neural network robustness characterization against Laser Fault Injection

- Iris model with one deep layer of **20** neurons (**80 weights** on the first layer).



Optical Lens x5 (Spot of 15µm)
Pulse power : 300 mA (170mW)
Pulse Width : 200 ns
Delay : 500 ns
Step on X = 2µm

LASER FAULT INJECTION ON EMBEDDED NEURAL NETWORK

➢ Neural network robustness characterization against Laser Fault Injection

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

Flash Memory with 8-bit weights stored

w_1	w_2	w_3	w_4
w_5	w_6	w_7	w_i
w_{i+1}	$\dots$	$\dots$	$\dots$

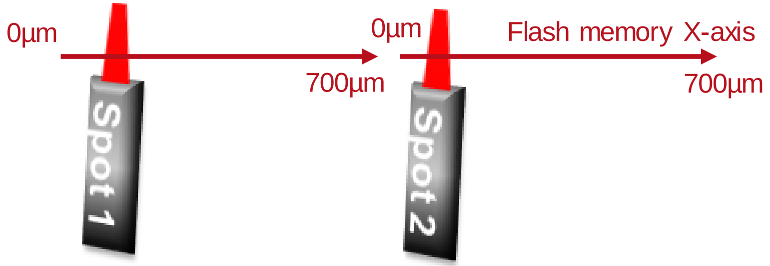
LASER FAULT INJECTION ON EMBEDDED NEURAL NETWORK

➢ Neural network robustness characterization against Laser Fault Injection

- 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.

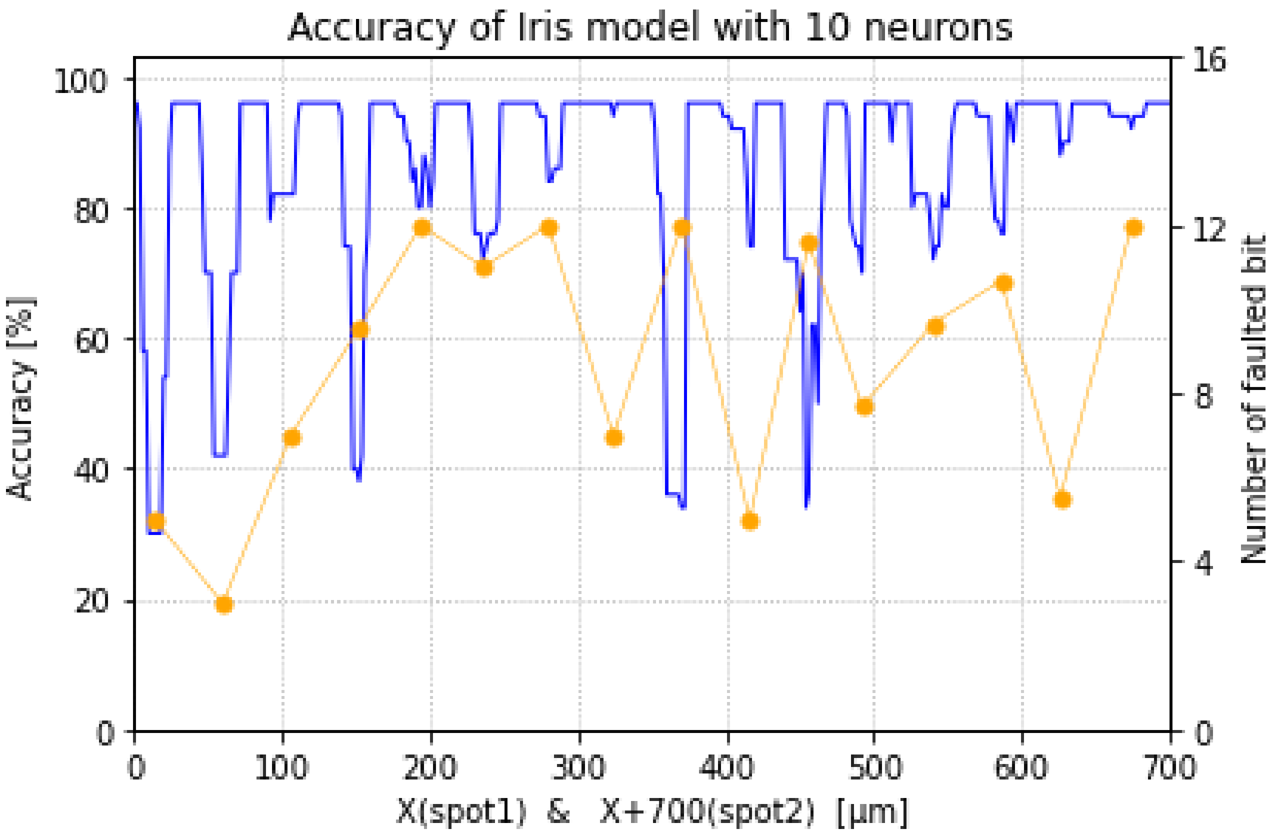
Flash Memory with 8-bit weights stored

w_1	w_2	w_3	w_4
w_5	w_6	w_7	w_i
w_{i+1}	$\dots$	$\dots$	$\dots$



Neural network robustness characterization against Laser Fault Injection bi-spot

- Study of the model accuracy under **bi-spot laser** characterization. Iris model with one deep layer of 10 neurons.



For both lens :
Optical Lens x5 (Spot of 15µm)
Pulse power : 300 mA (~170mW)
Pulse Width : 200 ns
Delay : 500 ns
Step on X = 2µm

- ✓ More faults are induced with bi-spot.
- ✓ Huge accuracy drop happened not only on high order bit.
- ✓ No drop accuracy below 30%.

- Context
- Bit-set fault model
- Laser Fault Injection on embedded neural network
- Conclusion

CONCLUSION

- **Fault injection analysis** on embedded neural network is still in its **infancy.**
  - Laser fault injection is a powerful mean to assess the **robustness** of an embedded model.
- Bit-set fault model allows to induce **precise** and **repeatable** faults on the **weights** of a neural network.
- We achieve an **accuracy drop** of a neural network with a laser fault injection targeting the weights.
- With **bi-spot laser** characterization, more weights can be faulted in the same inference.

ONGOING WORKS

- Use simulations to predict the most sensitive bits to fault with laser fault injection.
- Robustness characterization on deeper neural networks
- Other attack vectors (Instructions, activation functions…)
- Model reverse engineering with fault injection
- Evaluate state-of-the-art defense strategies against fault injection in a ML model context

THANK YOU

CryptArchi 2022

LASER FAULT INJECTION AGAINST EMBEDDED NEURAL NETWORK MODEL

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]** X. Hou, J. Breier, D. Jap, L. Ma, S. Bhasin, and Y. Liu, *Security Evaluation of Deep Neural Network Resistance against Laser Fault Injection*, Proceedings of the International Symposium on the Physical and Failure Analysis of Integrated Circuits, IPFA, 2020.
- **[4]** Yao, Fan, A.Rakin, et al. *,DeepHammer: Depleting the Intelligence of Deep Neural Networks through Targeted* Chain of Bit Flips." 29th USENIX Security Symposium, 2020.
- **[5]** Y. Fukuda, K. Yoshida, and T. Fujino, *Fault Injection Attacks Utilizing Waveform Pattern Matching against Neural Networks Processing on Microcontroller*, IEICE Transactions on Fundamentals of Electronics Communications and Computer Science, 2022.
- **[6]** J. Breier, D. Jap, X. Hou, S. Bhasin and Y. Liu, *SNIFF: Reverse Engineering of Neural Networks With Fault Attacks,* in IEEE Transactions on Reliability, 2021.
- **[7]** 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.