Deep-Dup: An Adversarial Weight Duplication Attack Framework to Crush Deep Neural Network in Multi-Tenant FPGA

Adnan Siraj Rakin*[1], Yukui Luo*[2], Xiaolin Xu[2], Deliang Fan[1]

*Joint First Authors

[1]Arizona State University

[2]North Eastern University

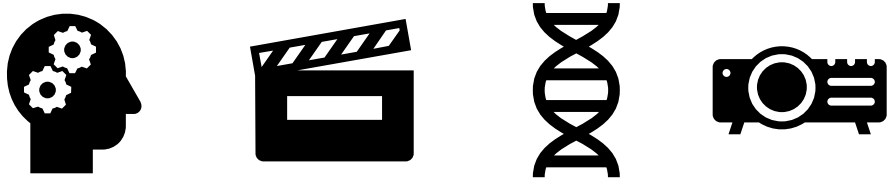
Outline

- **Introduction**
- **Threat Model**
- Hardware Fault Injection in Multi-Tenant FPGA
- Algorithm to search vulnerable weight index
- End-to-End Attack Framework
- Experimental Results
- Summary

Background : Machine Learning

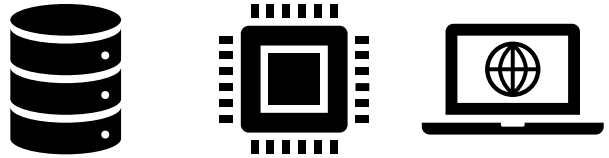
Machine Learning (ML) applications:

1. Computer Vision
2. Robotics
3. Medical Applications
4. Speech Recognition



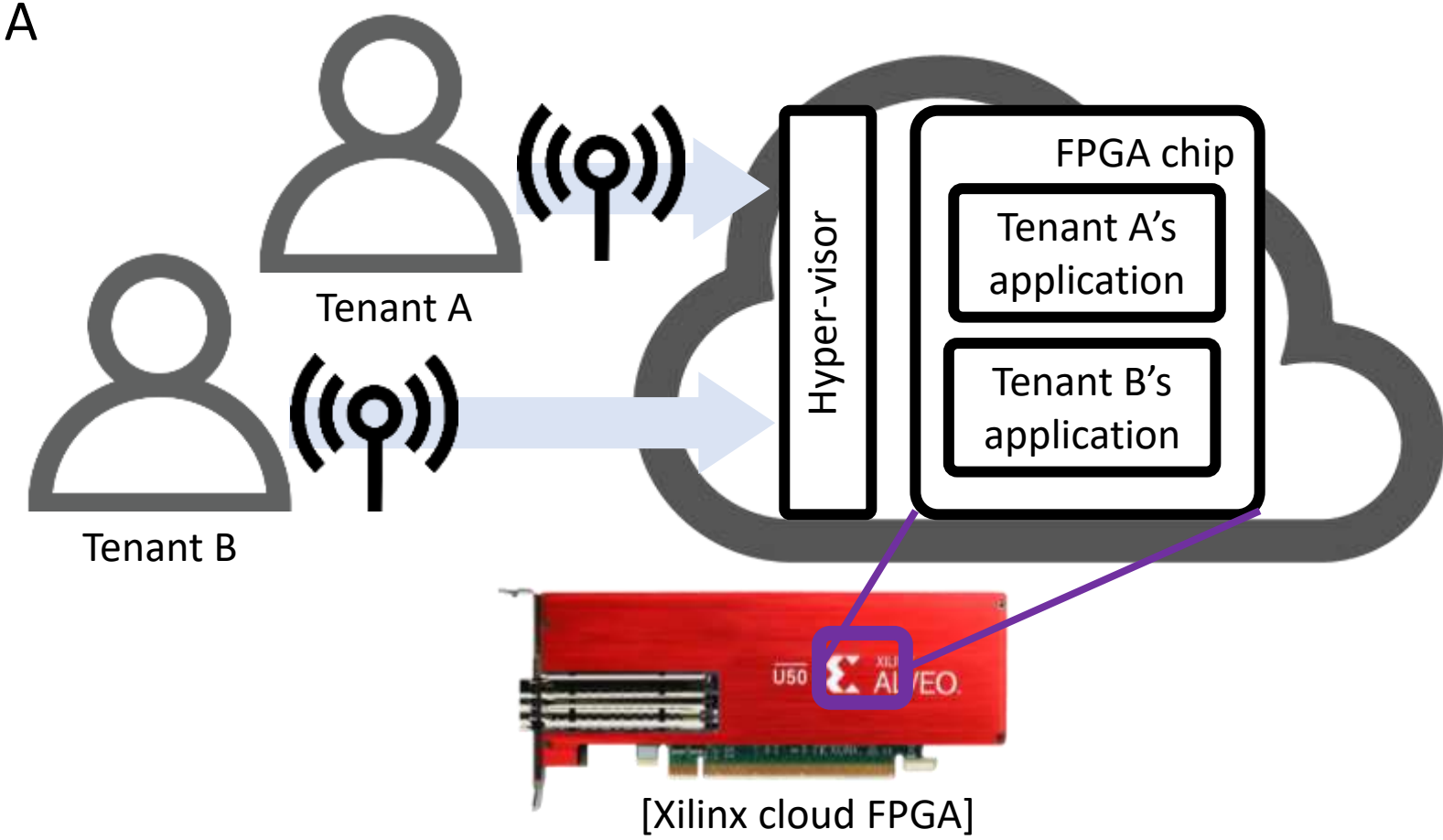
ML as a Service:

1. Amazon AWS AI
2. Google AI
3. Microsoft Azure ML



Background : Multi-Tenant FPGA

A promising cloud FPGA architecture: multi-tenant FPGA [Khawaja, A. OSDI'18] [Zha, Y. ASPLOS' 20]



[Xilinx cloud FPGA]

Potential security issues from hardware side:

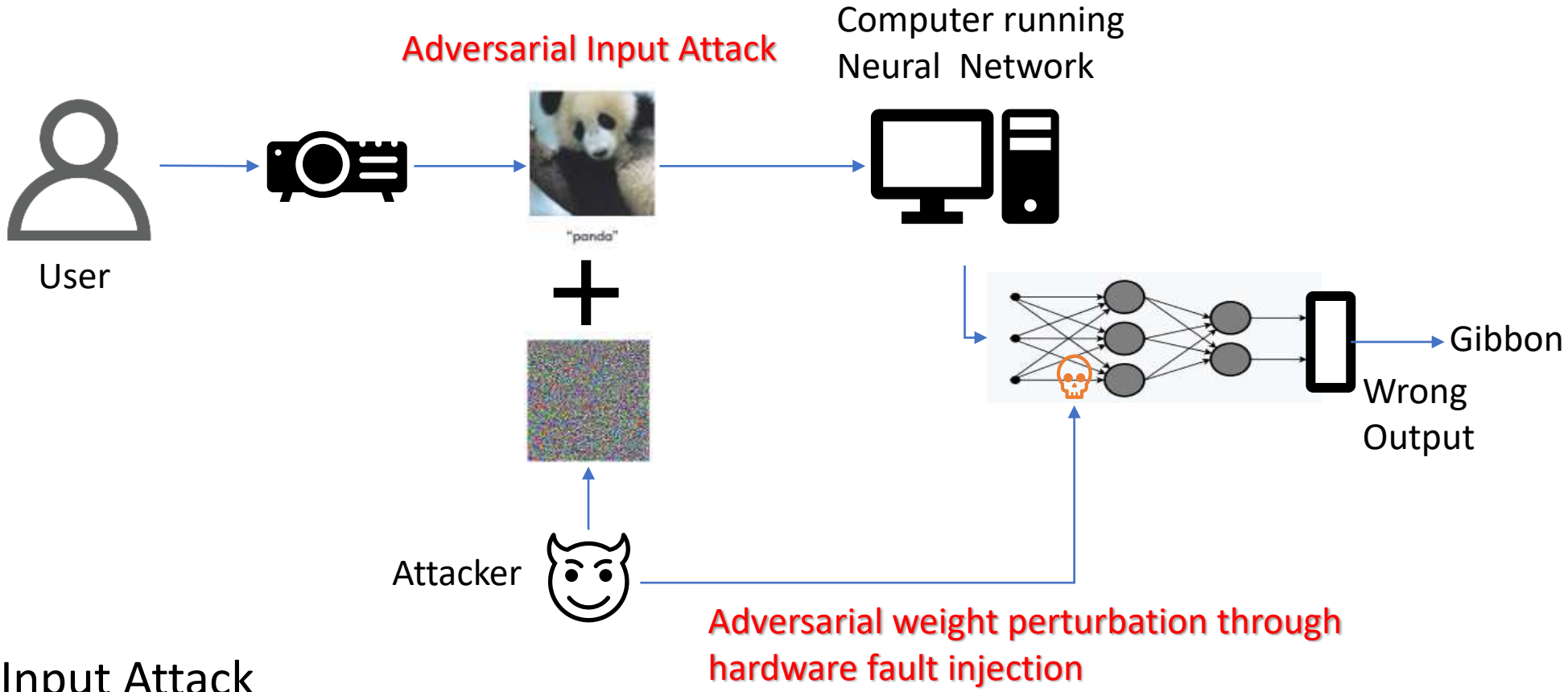
1. Side-channel attack

[Mark Z. SP' 18]

2. Fault injection attack

[Mahmoud, D. DATE' 19]

Potential Security Challenges



- Adversarial Input Attack
- Adversarial Weight Perturbations ***(Our Goal)***
- Backdoor/ Trojan Attack (Input + weights)

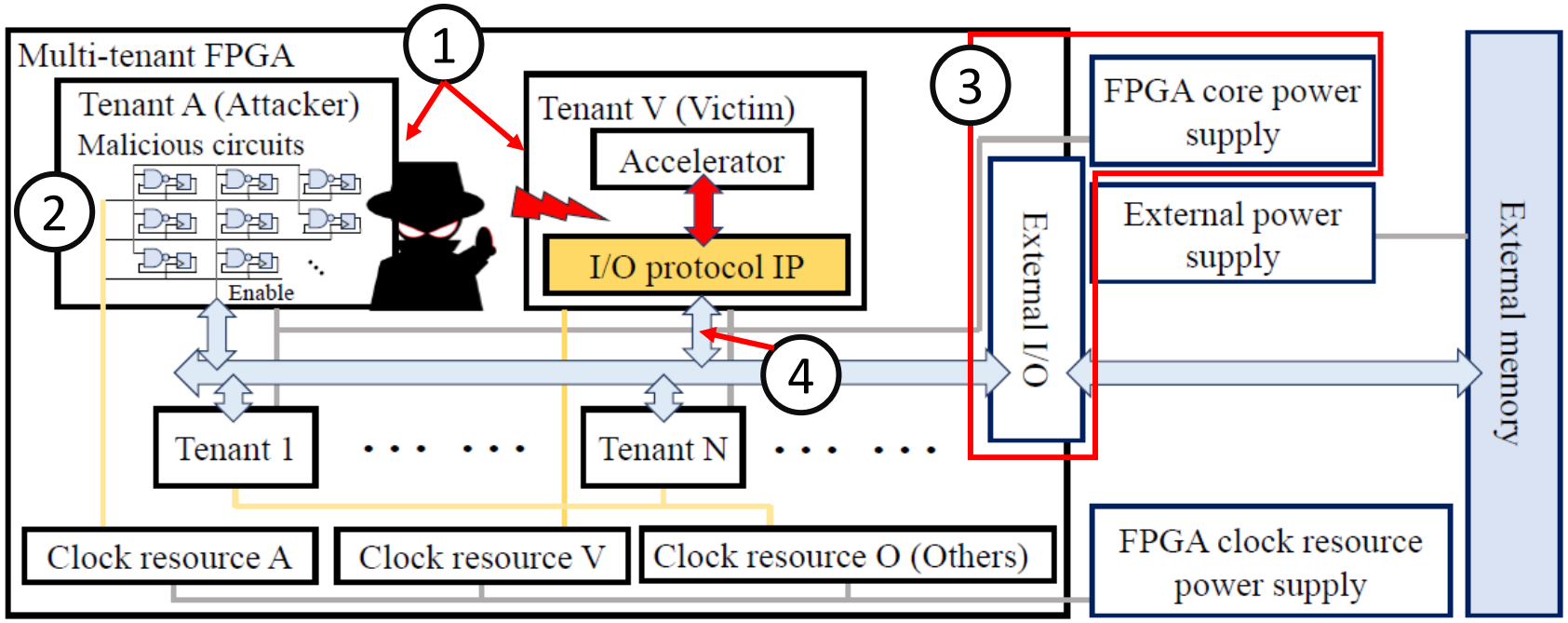
Our Motivation:

- **Domain of Attack:** Adversarial Weight perturbation

- **Objective:** Compromise the performance of machine learning model (e.g., degrading the test accuracy of deep neural network (DNN))

- **Application:** Multi-tenant Cloud FPGA Server.

Hardware Threat Model

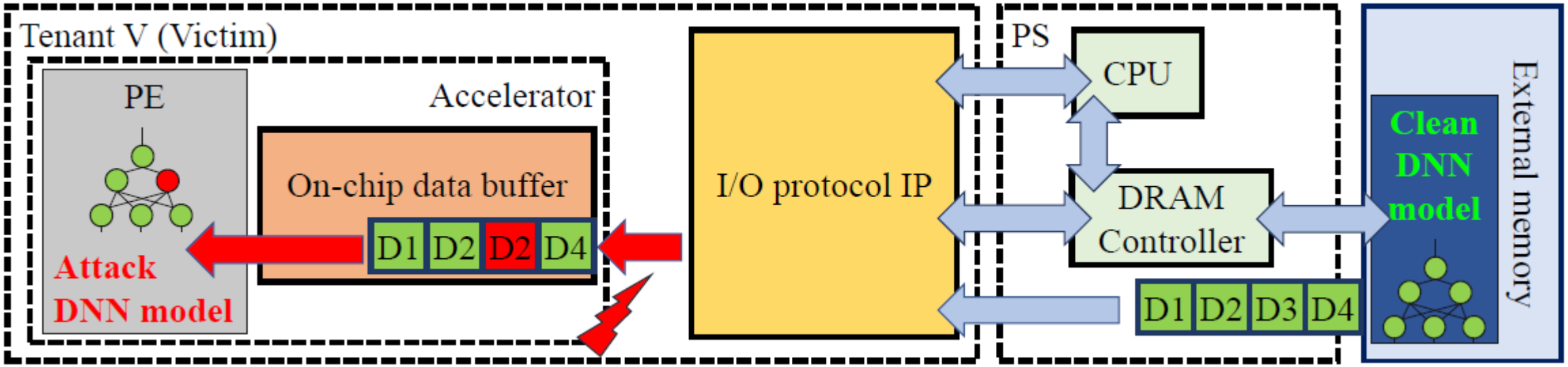


1. Multiple tenants co-reside on one FPGA chip.
2. Tenants have flexibility to program their applications.
3. All tenants share certain hardware resources
4. Logic cell and clock management cell have their own separate supplies.

Outline

- Introduction
- Threat Model
- **Hardware Fault Injection in Multi-Tenant FPGA**
- **Algorithm to search vulnerable weight index**
- **End-to-End Attack Framework**
- Experimental Results
- Summary

AWD Attack: attack overview



DNN application data flow

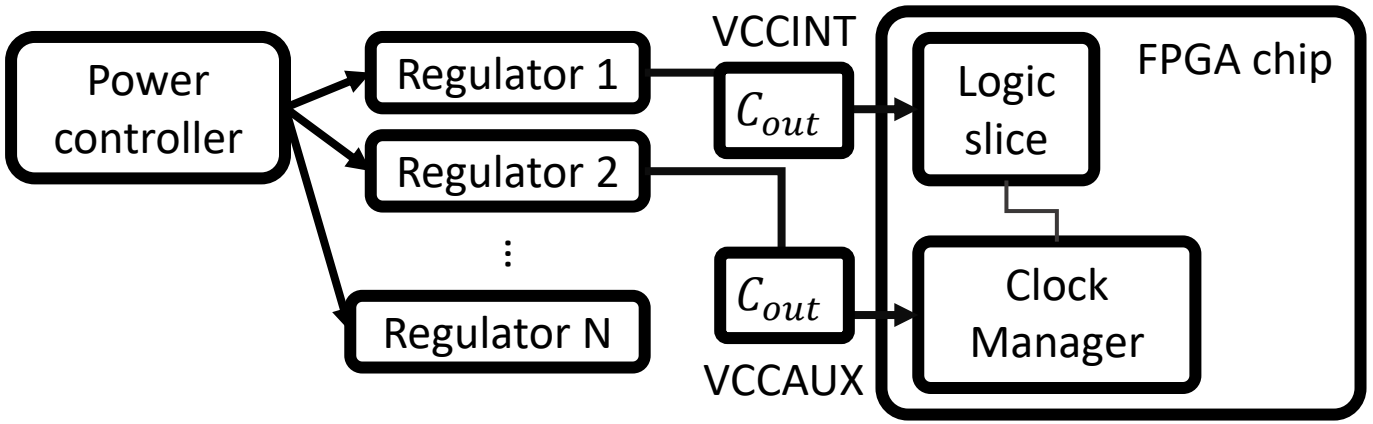
- Focusing on DNN accelerator.
- Targeting the data transmission process.

Power Distribution System (PDS)

PDS is the power source shared by the entire FPGAs chip.

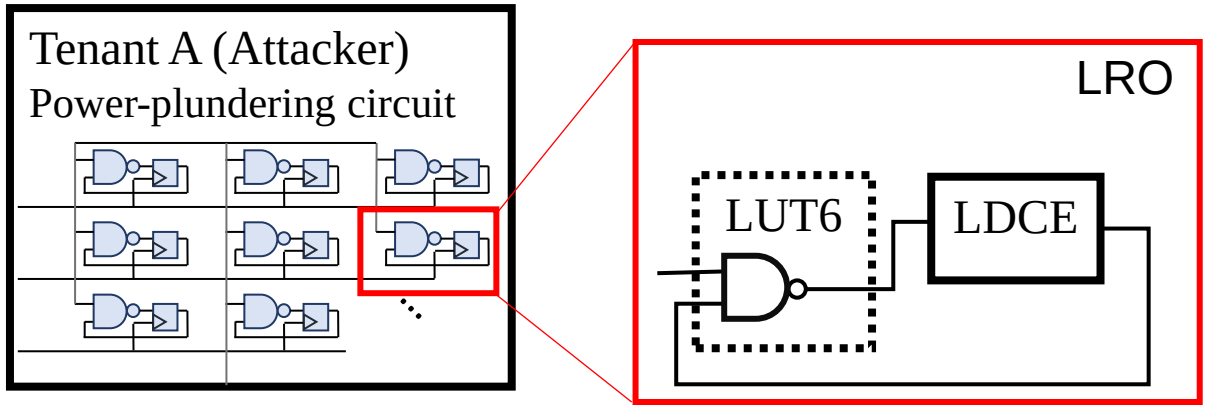
- Supply voltage fluctuations determined by:
  - Number of activated applications
  - Application's dynamic power consumption
- Aggressive FPGA utilization (e.g., overhead, power) leads to the supply voltage (VCCINT) drop

$$C_{out} = \frac{2 \times \Delta I_{out}}{f_{sw} \times \Delta V_{out}}$$



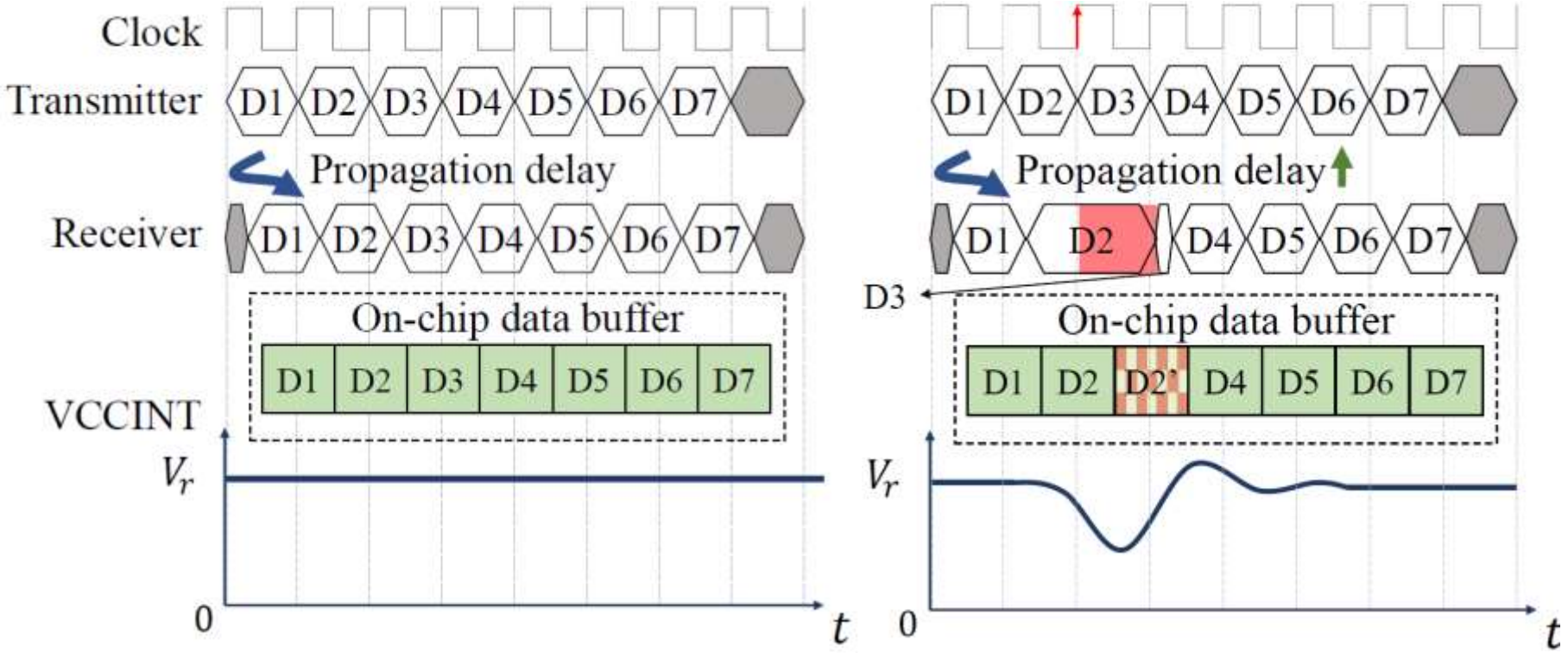
VCCINT: voltage supply of FPGAs internal components, such as LUT, charry-chain, etc.

AWD Attack: Power-plundering circuit



- A cloud-sanctioned power-plundering cell based on RO with a Latch (LRO)

AWD Attack: Analysis

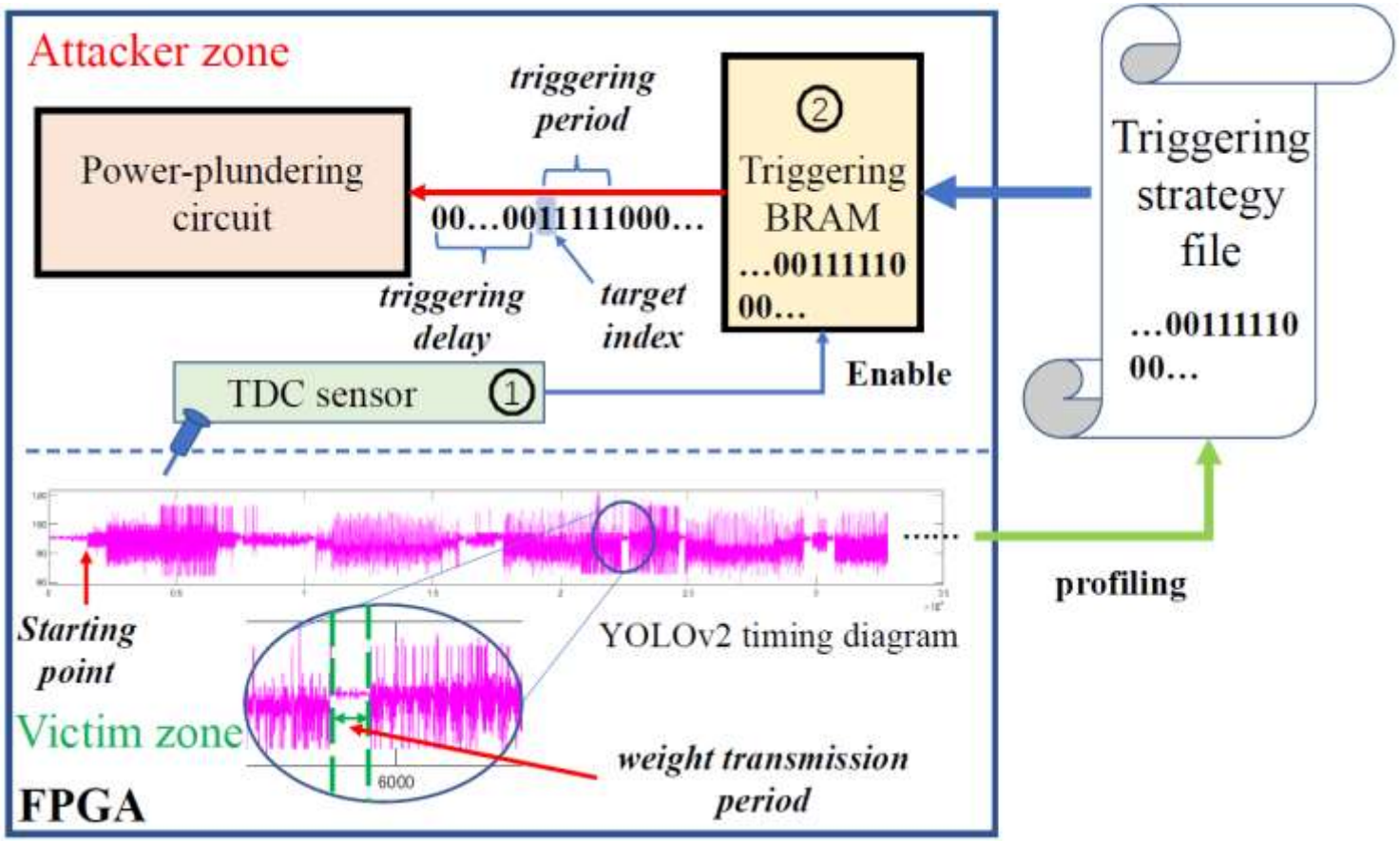


AMBA AXI4 data read channel:

External memory > on-chip data buffer (BRAM)

- Each DNN weight package (Di) is transmitted and received in a separate clock cycle.
- Voltage glitch incurs more propagation delay to the transmission of D2, which also shortens the next package D3.

AWD Attack: On-chip delay sensor



① Use a time-to-digital converter (TDC) [Gnad D. RE. FPL' 17] to trace the behavior of the target DNN application.

② Pre-configure the triggering strategy.

③ Start the attack by TDC detecting the DNN application starting point.

Attack success rate

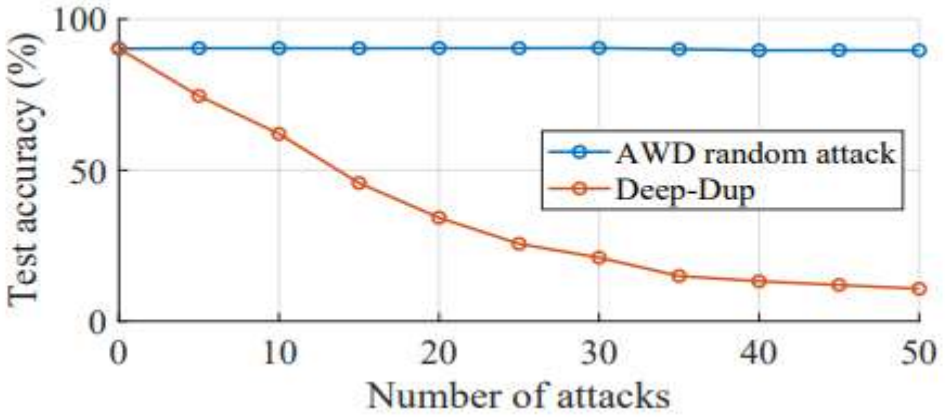
1.How successful our AWD Attack is in injecting the fault?

•LRO fault injection rate is 58.91%.

Even successful AWD fault attacks applied on random DNN weight index fails to achieve desired objective.

2.How successful a Random AWD is in depleting the DNN performance?

•Attacking VGG-11 on CIFAR-10 dataset



P-DES Algorithm

Questions:

- How to Use AWD Attack Efficiently?
- Can we improve Random AWD Attack ?

Challenges:

Can we conduct a black-box attack without any information of the target machine learning model?

Solution:

- Evolutionary Search Algorithm
- We propose:

  Progressive Differential Evolution Search (P-DES)

P-DES: Initialization

- **Initialize** Random weight index

$w_1, w_2, \ldots\ldots w_z$

Fitness Function

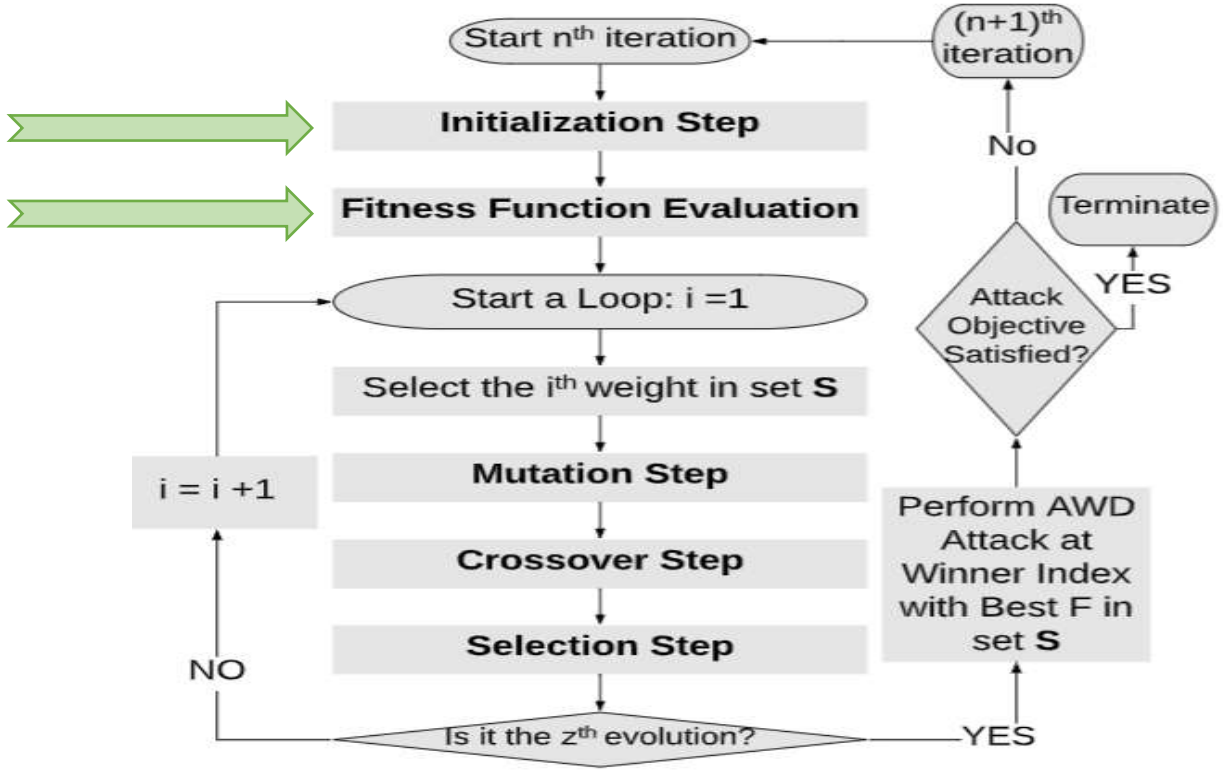
- Evaluate the Fitness Function

At each weight index after AWD.

- We adopt the neural network loss

Function as the Fitness Function *L*:

$$\max_{\theta} E_x L(f_\theta(x), t)$$



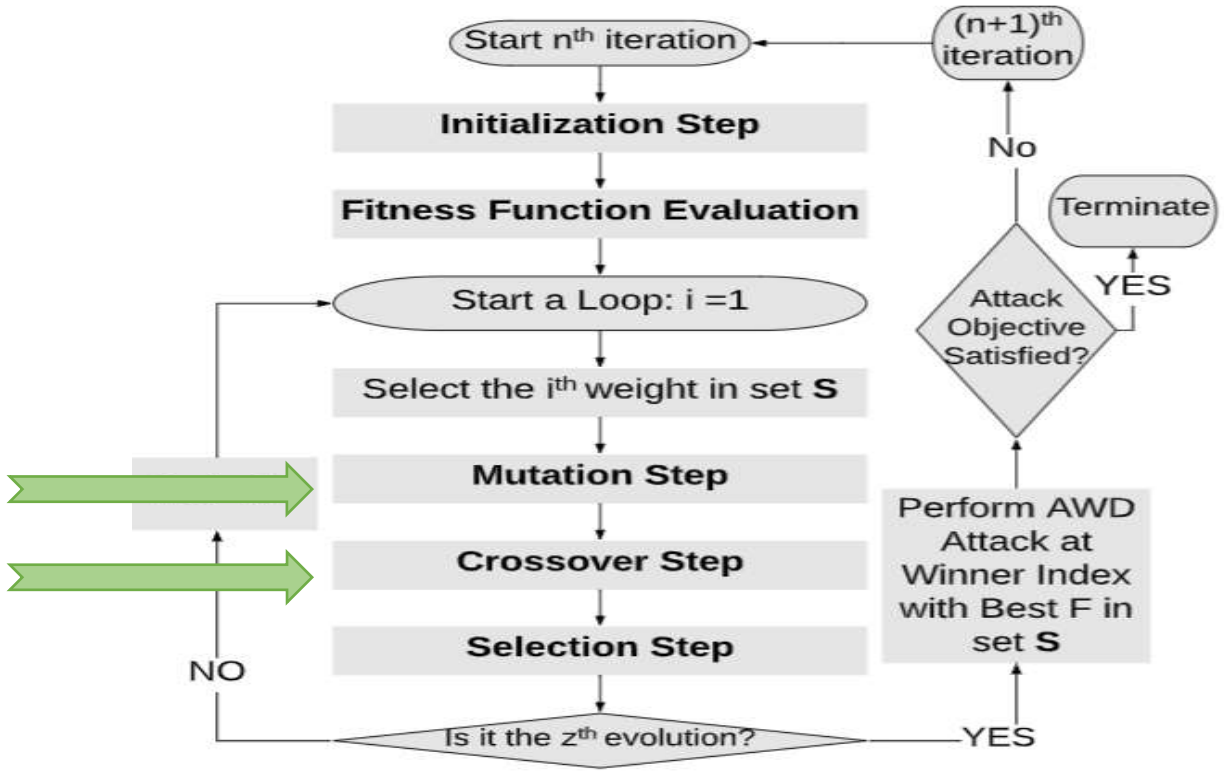
P-DES: Evolution

- **Mutation:** The goal is to create a new Weight candidate with better fitness function. A sample mutation Strategy:

$w_{new} = w_a + \alpha (w_b - w_c)$

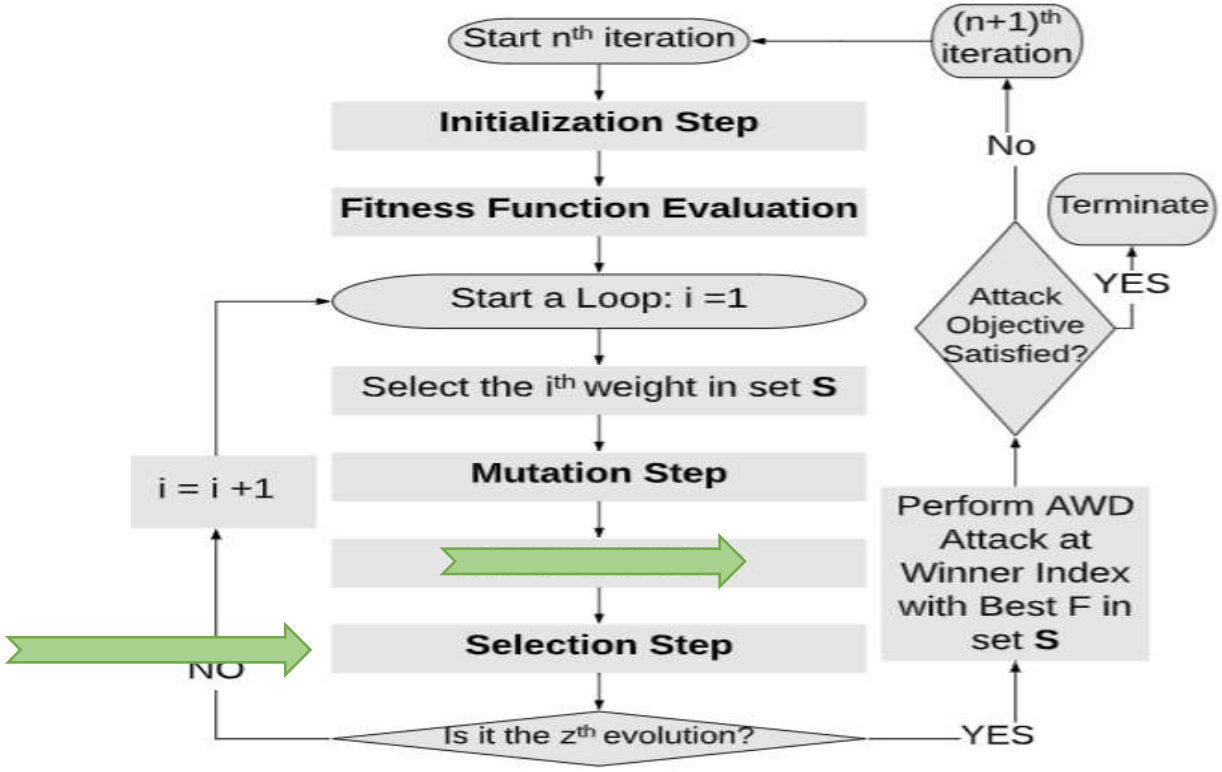
a,b,c are random numbers between zero to z.

- **Crossover Step:** Ensures the new weight index generated is in a valid range.

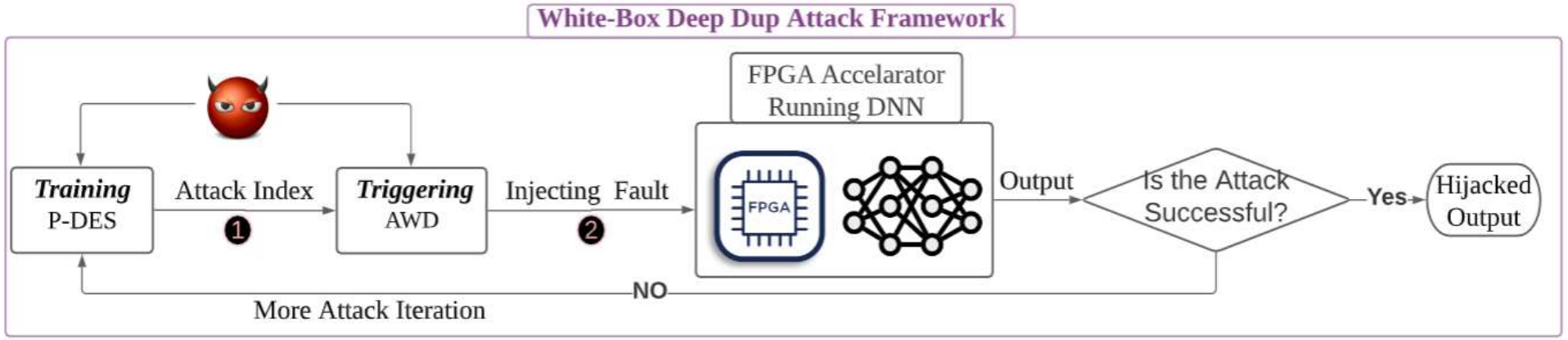


P-DES: Termination

- Select or skip the new mutant index
- The value of Fitness Function works as a decision rule
- Finally, the AWD attack is Conducted at the winner index after z evolution.



End-to-End Attack: White-Box

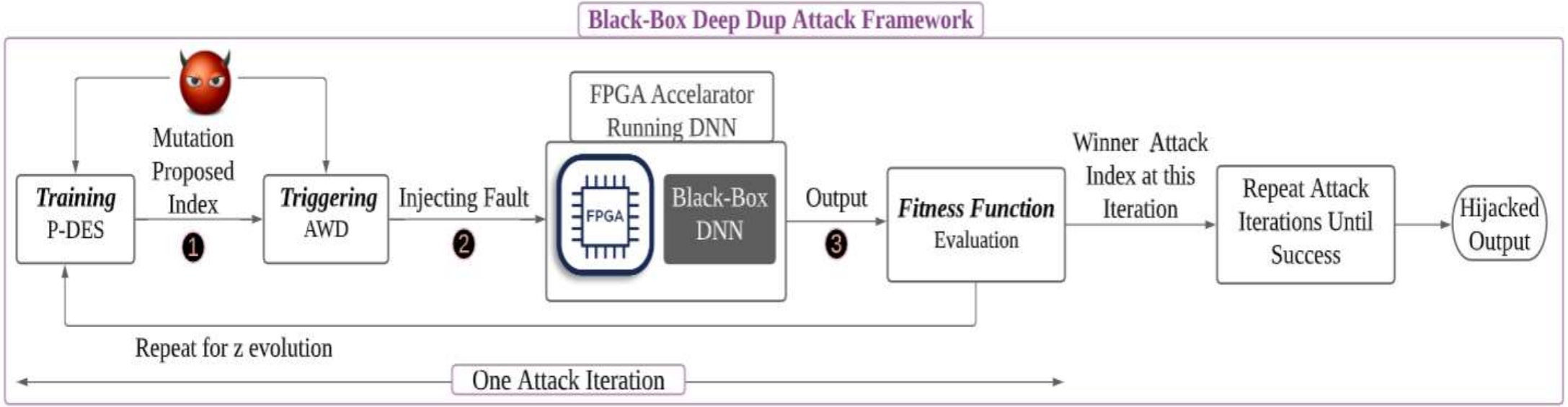


Attacker Knowledge:

Information	White-Box	Black-Box
Model Architecture	Yes	No
Model weights and biases	Yes	No
Gradient Information	Yes	No
Batch-Norm and other parameters	Yes	NO

- In a white-box, the P-DES algorithm operates on an off-line model (replica of the target model)

End-to-End Attack: Black-Box



- Fitness function is evaluated directly on the deployed model in FPGA.
- Mutation proposed candidate ⟹ AWD attack on that candidate ⟹ Evaluate Fitness Function

Outline

- Introduction
- Threat Model
- Hardware Fault Injection in Multi-Tenant FPGA
- Algorithm to search vulnerable weight index
- End-to-End Attack Framework
- **Experimental Results**
- **Summary**

Experimental Setup

- Dataset and DNN models
  - CIFAR-10, ImageNet, COCO
  - ResNet-20 (18,50), VGG-11, MobileNetV2, YOLOv2
- Multi-tenant FPGA prototype
  - Xilinx MPSoC device
  - The acceleration core DPU (150MHz/300MHz),

  and manual made YOLOv2 (180MHz)

Results

Black-Box Attack on YOLOv2 using LRO cell		
Initial mAP	Post-Attack mAP	# of Attacks
0.428	0.14	63

- Attacking a YOLOv2 object recognition model
- Performing weight duplication at P-DES searched index
- The person class no-longer being recognized correctly

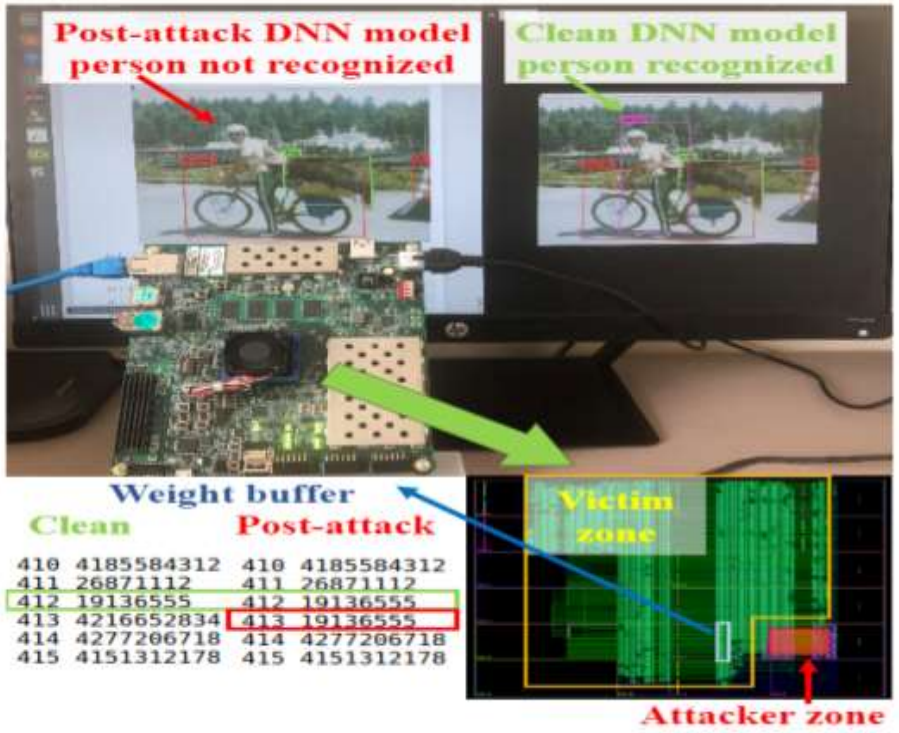


Image Classification Results

- Successful across a wide range of Deep Learning Architectures.
- Compact architectures like MobileNet-V2 are more vulnerable.

White-Box Attack on Image Recognition				Un-Targeted Attack		Targeted Attack			
Dataset	Network	# of Parameters	TA (%)	Post-Attack TA (%)	# of Attacks	Post-Attack TA (%)	Target Class(t_s)	ASR (%)	# of Attacks
CIFAR-10	ResNet-20	0.27 M	90.77	10.92	28	21.63	Bird	99.2	14
	VGG-11	132 M	90.38	10.94	77	23.68	Horse	98.6	63
ImageNet	MobileNetV2	2.1 M	70.79	0.19	1	8.93	Lesser Panda	100.0	1
	ReNet-18	11 M	69.35	0.18	106	34.45	Ostrich	100.0	13
	ReNet-50	23 M	72.97	0.19	175	30.57	Ostrich	100.0	20

Summary

- A ***novel attack surface*** on multi-tenant FPGA.
- A ***novel algorithm*** to locate vulnerable weight index in deep neural networks.
- Our Deep-Dup attack framework can conduct ***black-box attack***.
- Deep-Dup can completely ***deplete DNN inference*** performance to random guess.

Reference:

- [Xilinx cloud FPGA] https://forums.xilinx.com/t5/Xilinx-Xclusive-Blog/Accelerating-Cloud-Applications-with-Nimbix-and-Samsung/ba-p/1089617
- [Khawaja, A. OSDI'18] Khawaja, A., Landgraf, J., Prakash, R., Wei, M., Schkufza, E., & Rossbach, C. J. (2018). Sharing, protection, and compatibility for reconfigurable fabric with amorphos. In *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI' 18)*
- [Zha, Y. ASPLOS' 20] Zha, Yue, and Jing Li. "Virtualizing FPGAs in the cloud." Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems. 2020. (ASPLOS' 20)
- [Mark Z. SP' 18] Mark Zhao and G Edward Suh. Fpga-based remote power side-channel attacks. In2018 IEEE Symposiumon Security and Privacy (SP' 18)
- [Gnad D. RE. FPL' 17] Gnad, Dennis RE, Fabian Oboril, and Mehdi B. Tahoori. "Voltage drop-based fault attacks on FPGAs using valid bitstreams." 2017 27th International Conference on Field Programmable Logic and Applications (FPL' 17)
- [Provelengios, G. FPGA' 19] Provelengios, G., Ramesh, C., Patil, S. B., Eguro, K., Tessier, R., & Holcomb, D. (2019, February). Characterization of long wire data leakage in deep submicron FPGAs. In Proceedings of the 2019 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA' 19)
- [Mahmoud, D. DATE' 19] Mahmoud, Dina, and Mirjana Stojilović. "Timing violation induced faults in multi-tenant FPGAs." *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE' 19)*
- Open AI (https://openai.com/blog/adversarial-example-research/)

More Information

Questions?

Adnan Siraj Rakin (asrakin@asu.edu)

Dr. Deliang Fan (dfan@asu.edu)

Website:

https://dfan.engineering.asu.edu/ai-security/

Yukui Luo (luo.yuk@northeastern.edu)

Dr. Xiaolin Xu (x.xu@northeastern.edu)

Website:

https://www.xiaolinxu.com/

This work is supported in part by the *National Science Foundation* under Grant No.2019548 and No.2043183.