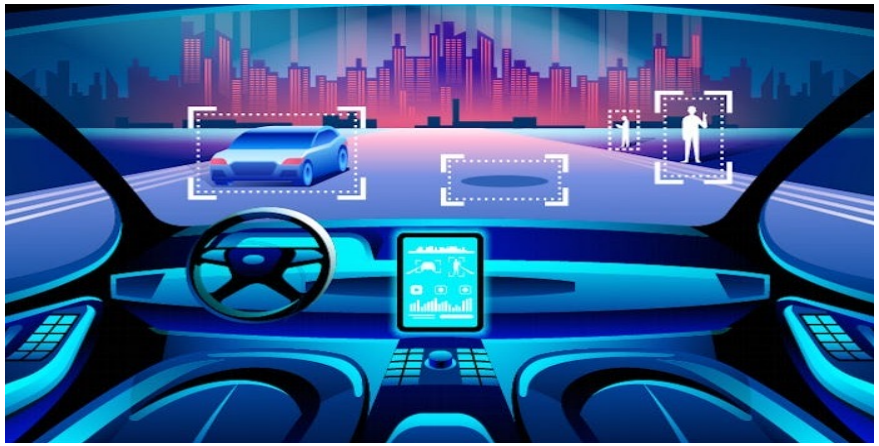
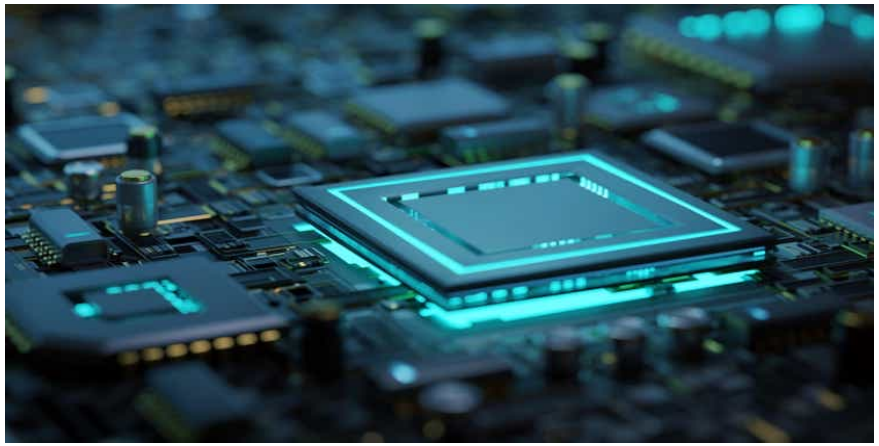
TensorFI+: A Scalable Fault Injection Framework for Modern Deep Learning Neural Networks

Sabuj Laskar, Md Hasanur Rahman, Guanpeng Li

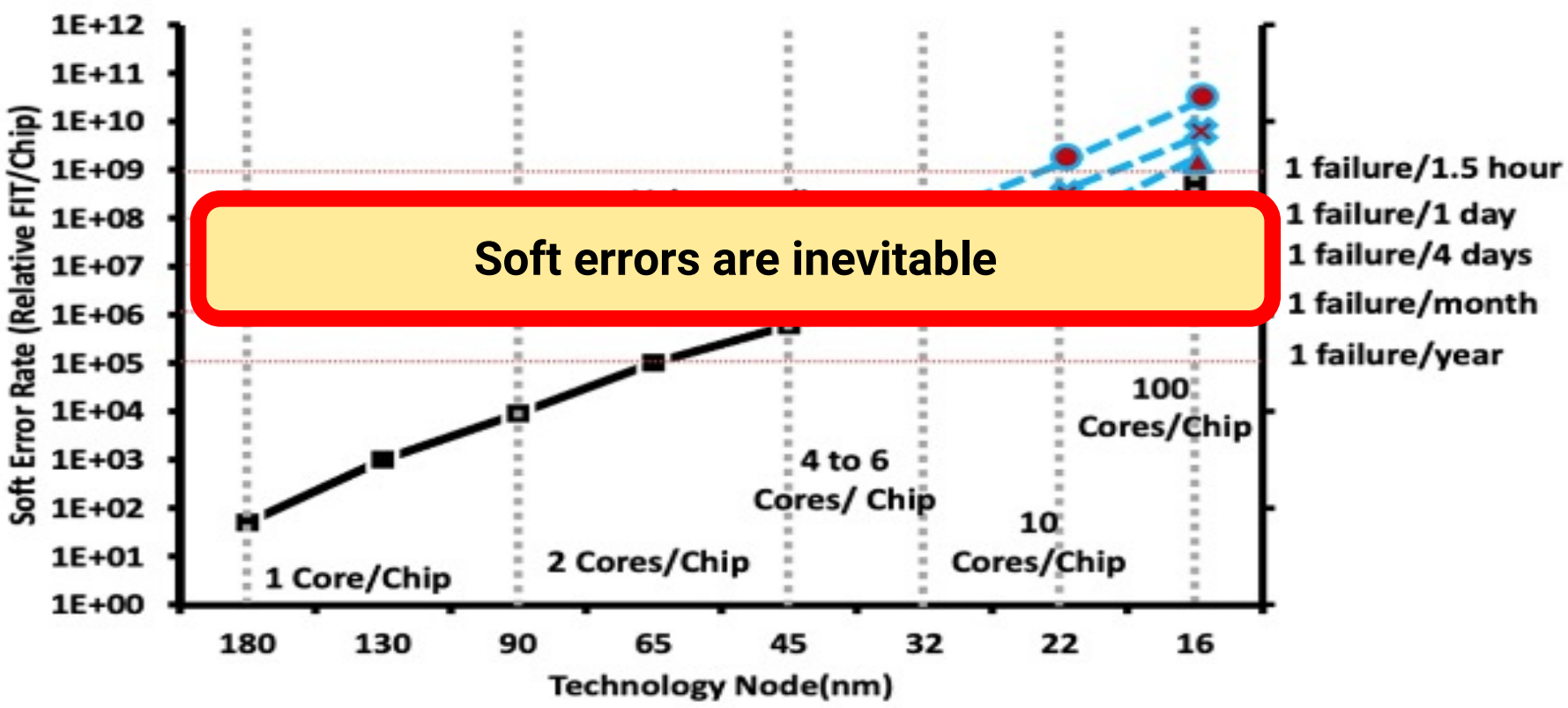
Motivation

- DNN has been increasingly deployed in many areas
  - Computer vision, NLP, autonomous vehicles (AVs)
- DNN reliability becomes important
  - ISO 26262 safety standard requires no more than 10 FIT (Failure in every 10^9 hours)





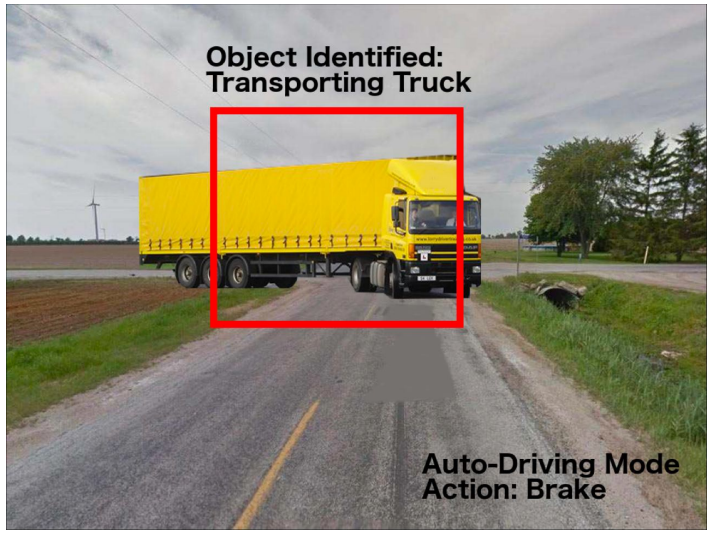
Soft Error



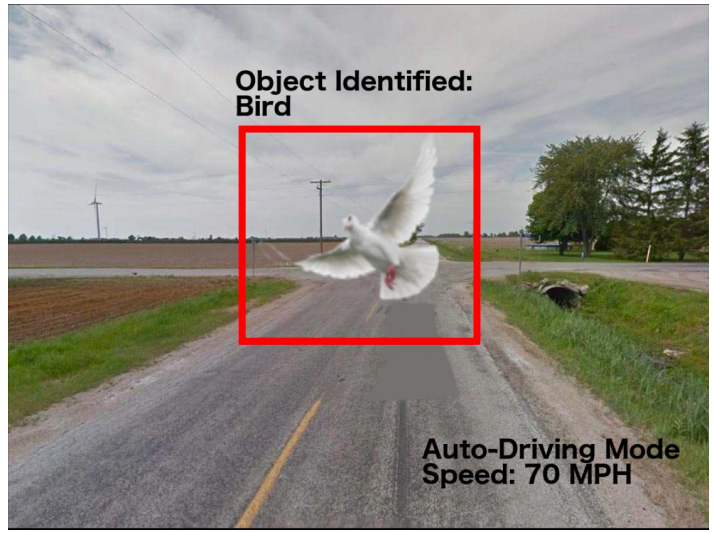
Taken from [1]

Consequences of Error Propagation in DNNs

- Single-bit fault[2] → Misclassification of image



Fault-free prediction label: Truck



Faulty predicted label: Bird

- Reliability assessment: hardware vs software level
  - Software implemented fault injection (FI) simulation has lower cost

[2] *Understanding Error Propagation in Deep Learning Neural Network (DNN) Accelerators and Applications, SC'17*

Existing DNN Reliability Measurement Tools

TensorFI[3]

- A fault injector for TensorFlow applications
- Specifically, for TensorFlow 1 applications

TensorFI 2[4]

- A fault injector for T
- This only supports sequential models

Need Support to inject faults in non-sequential DNN models with TensorFlow 2

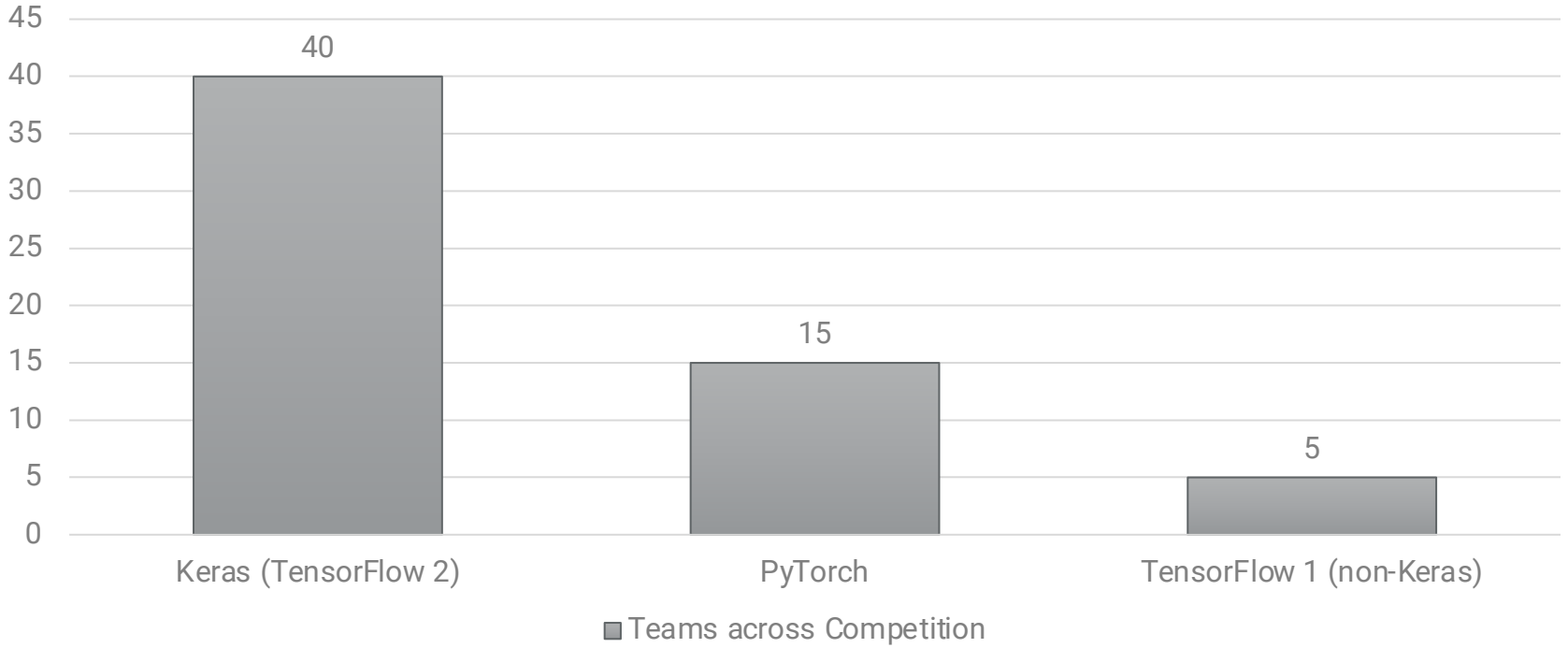
Most DNN models are non-sequential

- Sequential: VGG16, VGG19
- Non-Sequential: ResNet50, ResNet101, GoogleNet, Xception, DenseNet121, DeseNet169, MobileNet

IOWA

[3] Tensorfi: A configurable fault injector for tensorflow applications, ISSREW'18
[4] https://github.com/DependableSystemsLab/TensorFI2

Why Keras (TensorFlow 2)



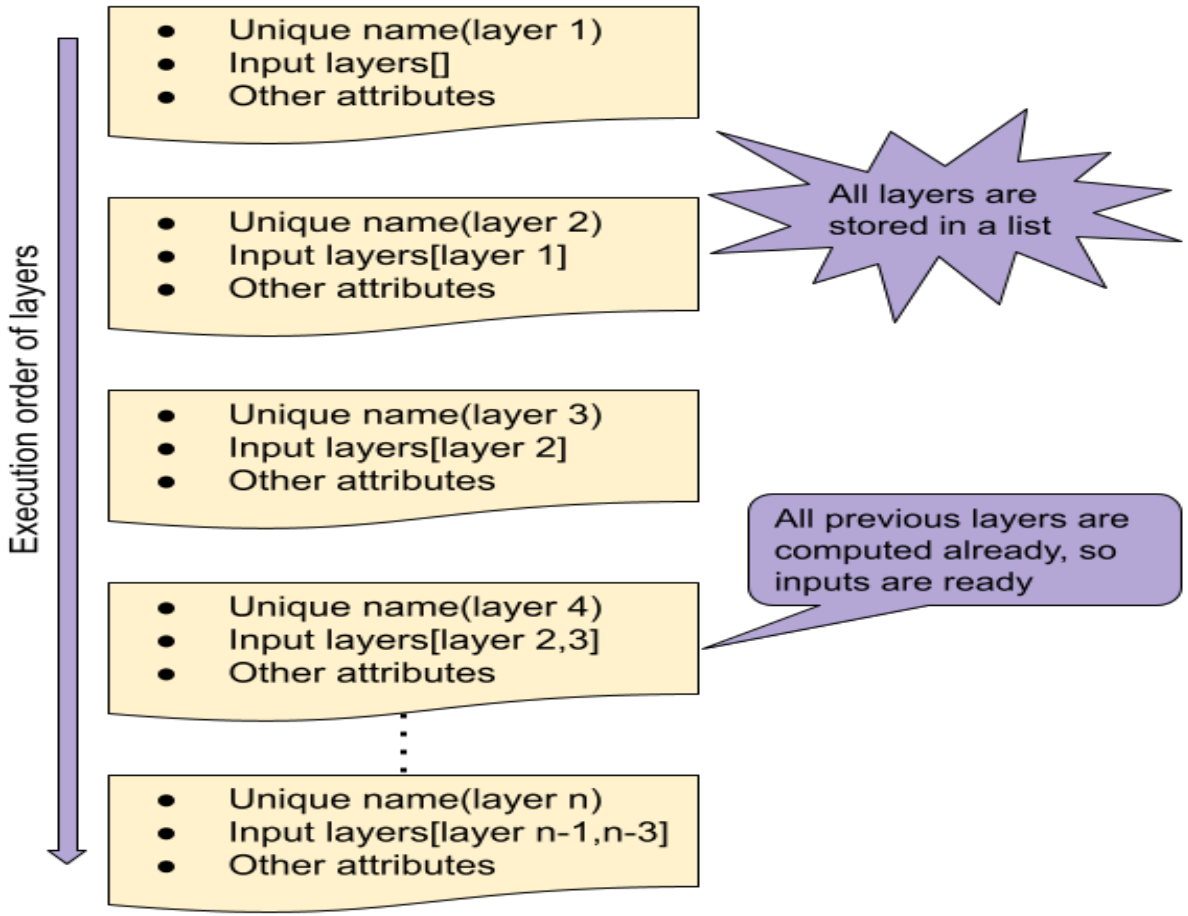
✓ Primary ML software tool used by top-5 teams on Kaggle in each competition in the last two years

Source: https://keras.io/why_keras/

Our Contributions

- Developed open-source tool, TensorFI+, to support FI in non-sequential DNN models
- Versatility: FI to any DNN models built with TensorFlow 2
- Performance optimization during FI

Keras (TensorFlow 2) Execution Flow

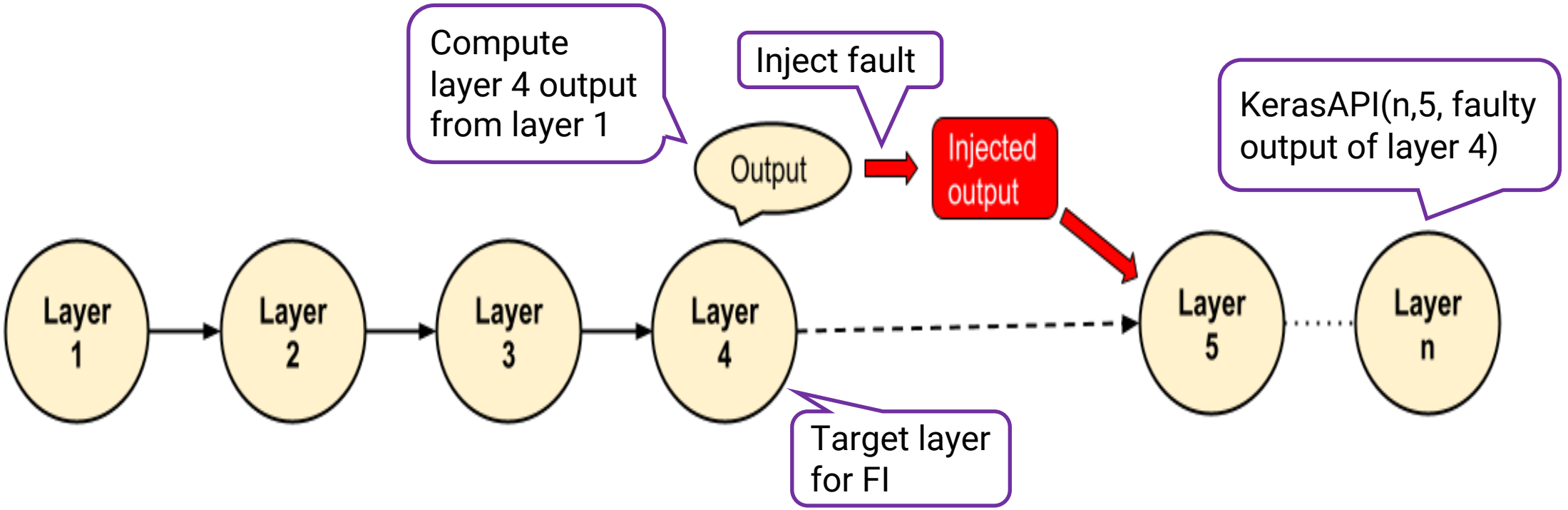


Keras execution flow without TensorFI+

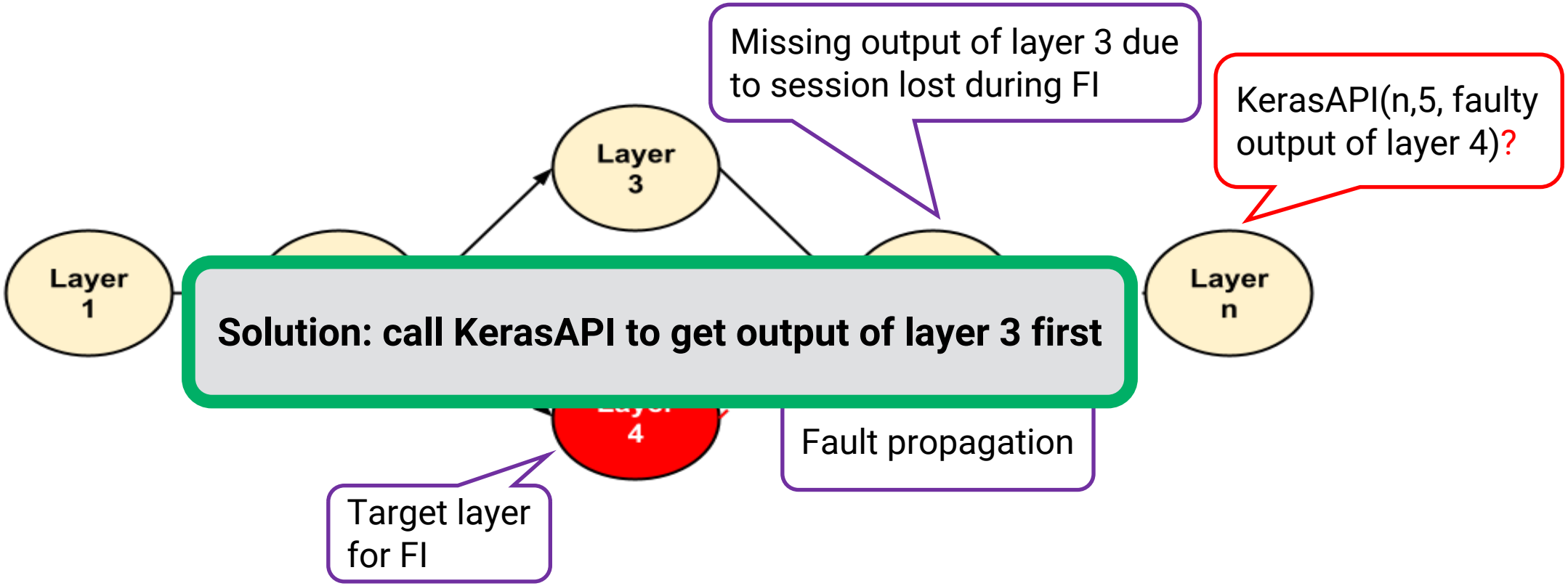
Keras Execution Flow Changes with TensorFI+

- Operators' structure changes in TensorFlow 2 are not allowed
- Need Keras API for fault injection and propagation
  - Output (layer D) = KerasAPI(Destination layer D, Source layer S, Input values of S)
  - KerasAPI call to get output of target layer t
  - Random bit fip of output of layer t
  - Previous session gone, need API calls to propagate faulty output to final layer

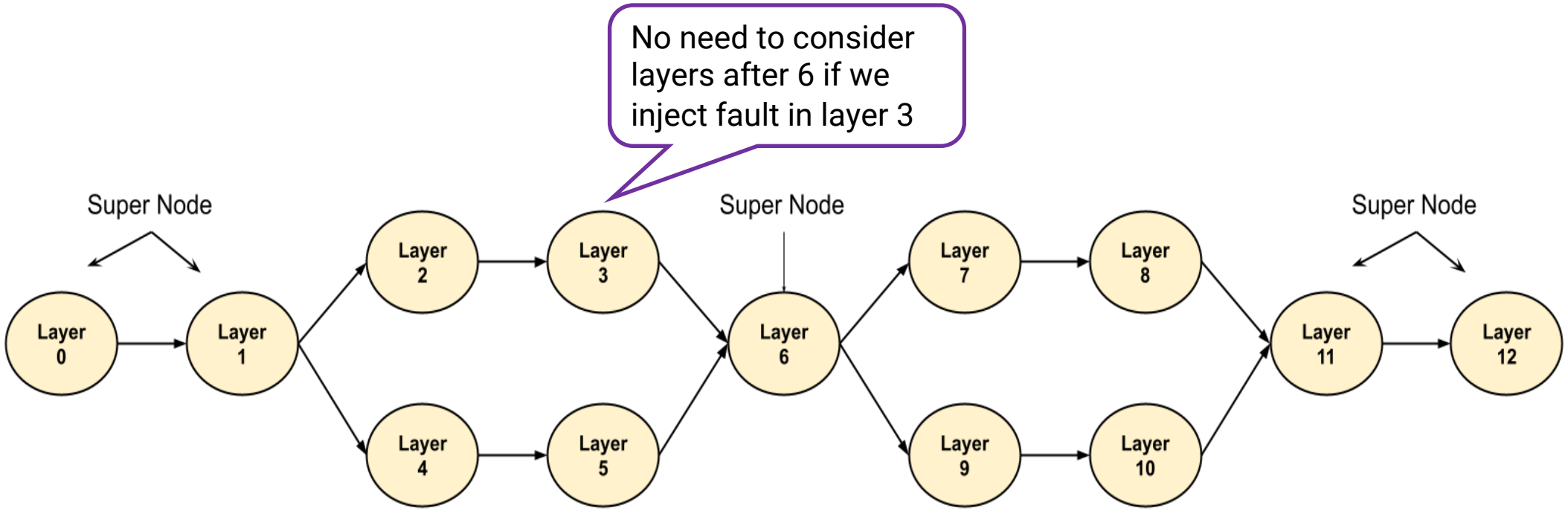
FI in a Sequential Model



Issues in FI in Non-sequential Model

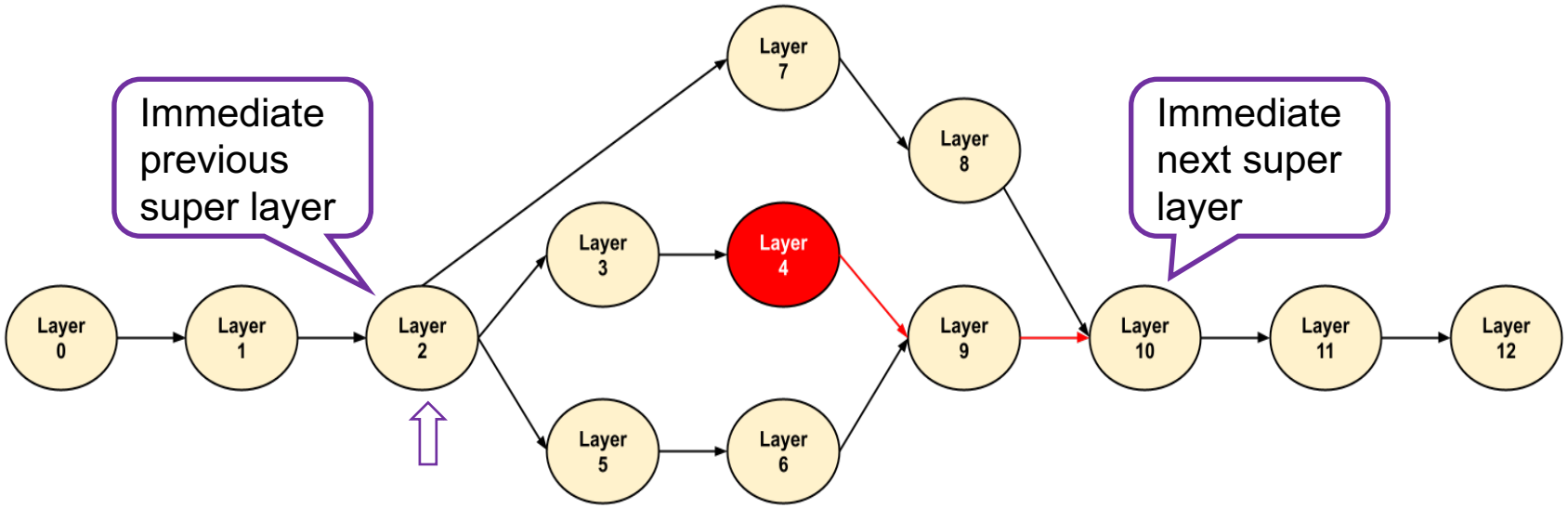


Solution: Super Layer

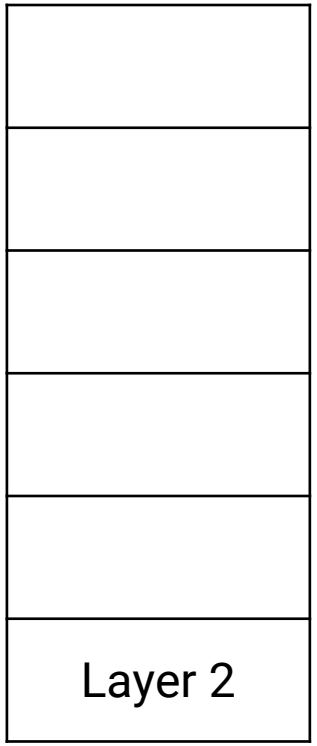


- Super layers are not part of any branch
- Any layer after a super layer is not dependent on any layer prior to super layer

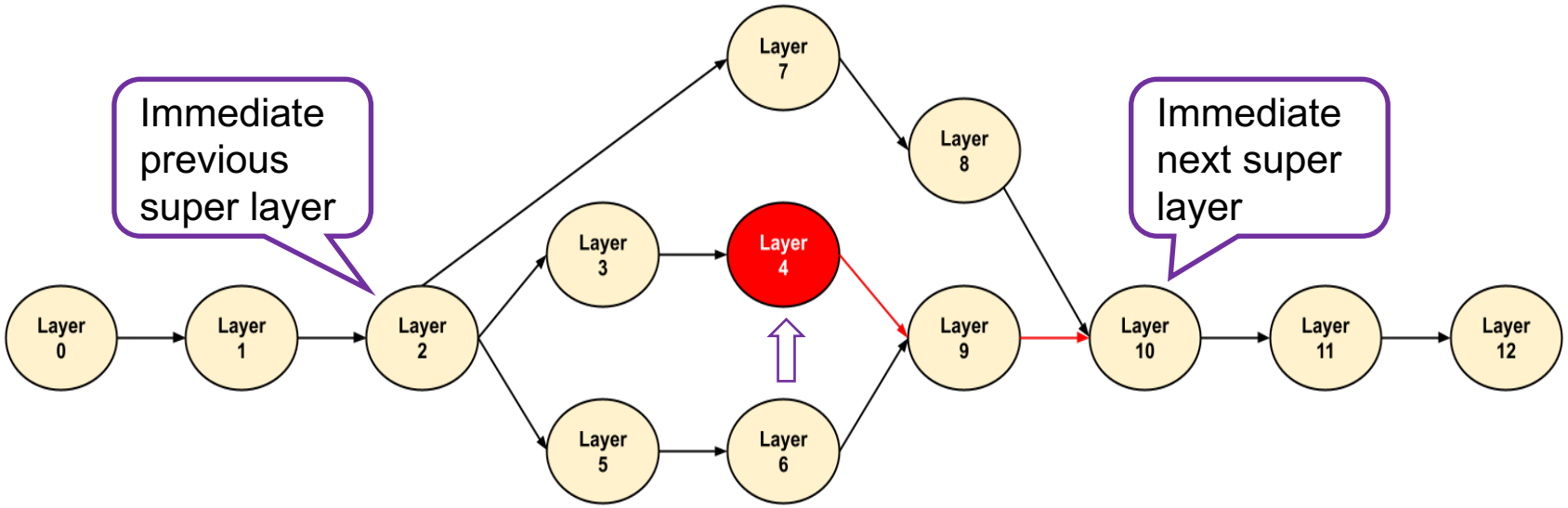
Simulation of FI with TensorFI+



MDict



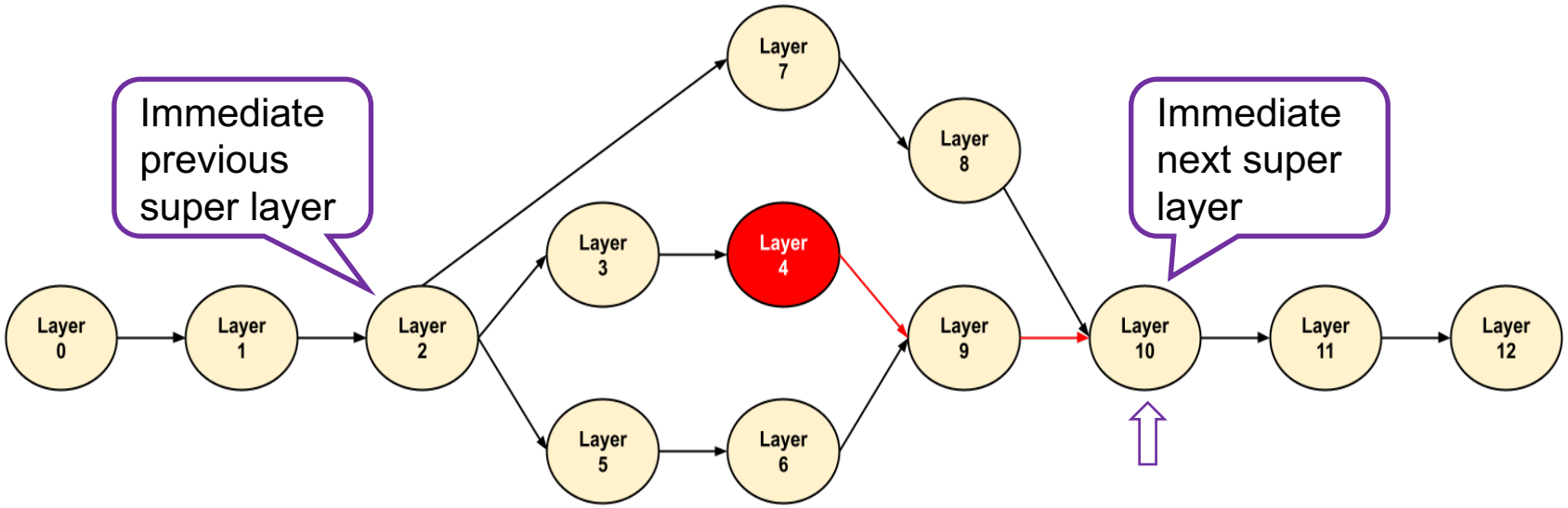
Simulation of FI with TensorFI+



MDict

Layer 4
Layer 2

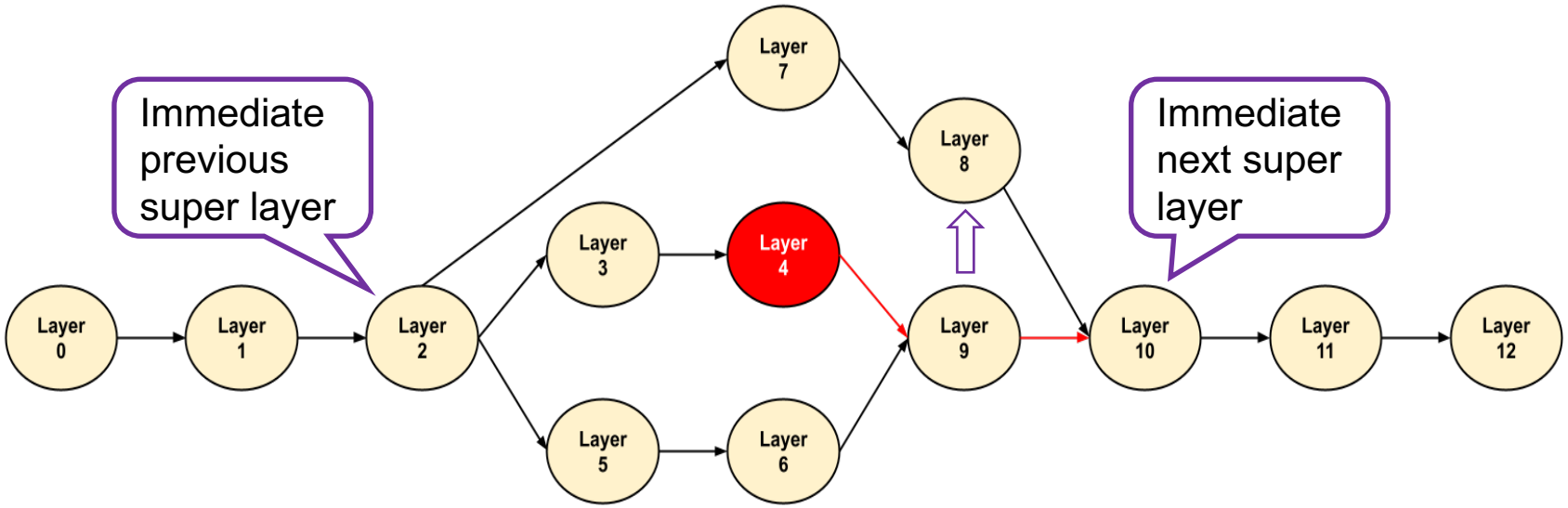
Simulation of FI technique of TensorFI+



MDict

Layer 4
Layer 2

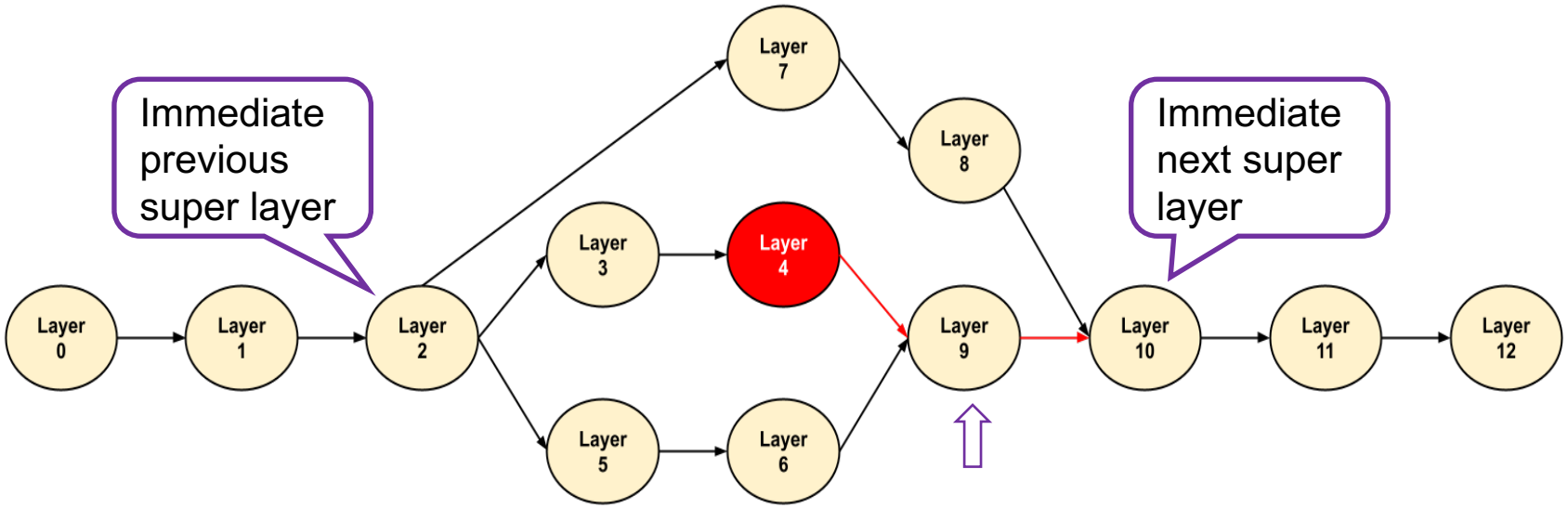
Simulation of FI technique of TensorFI+



MDict

Layer 8
Layer 4
Layer 2

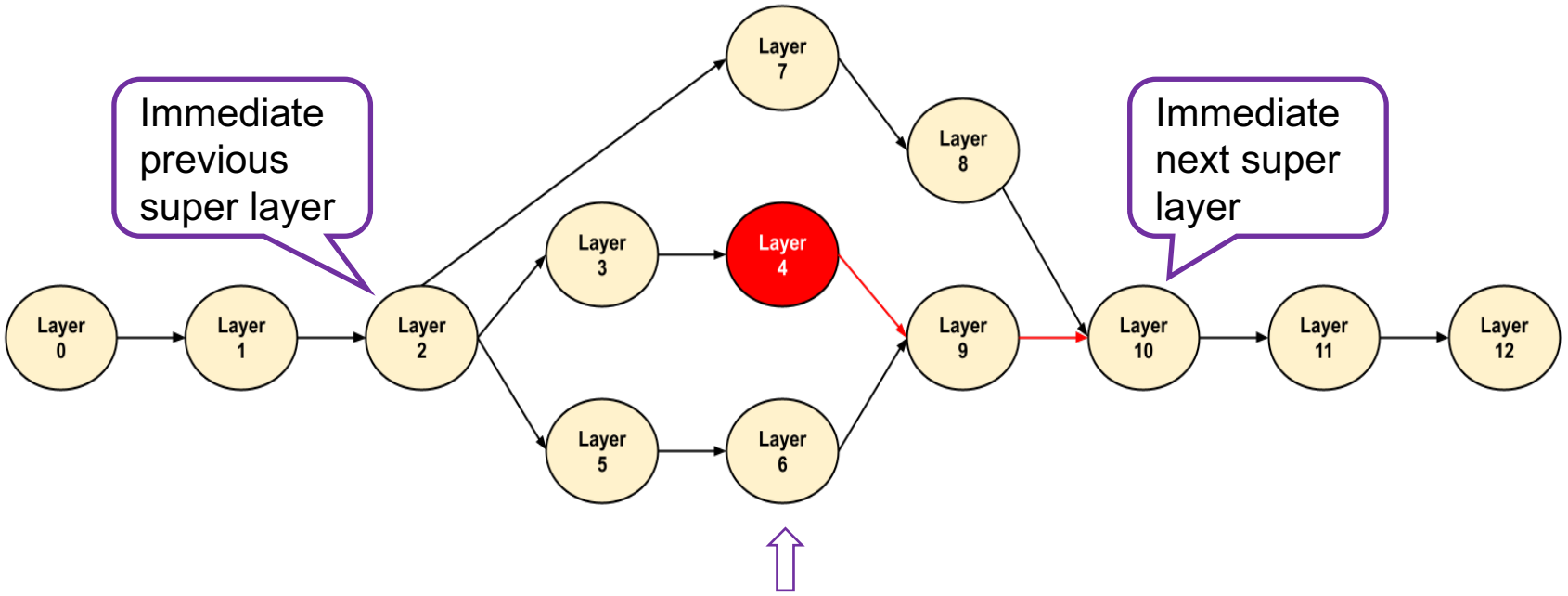
Simulation of FI technique of TensorFI+



MDict

Layer 8
Layer 4
Layer 2

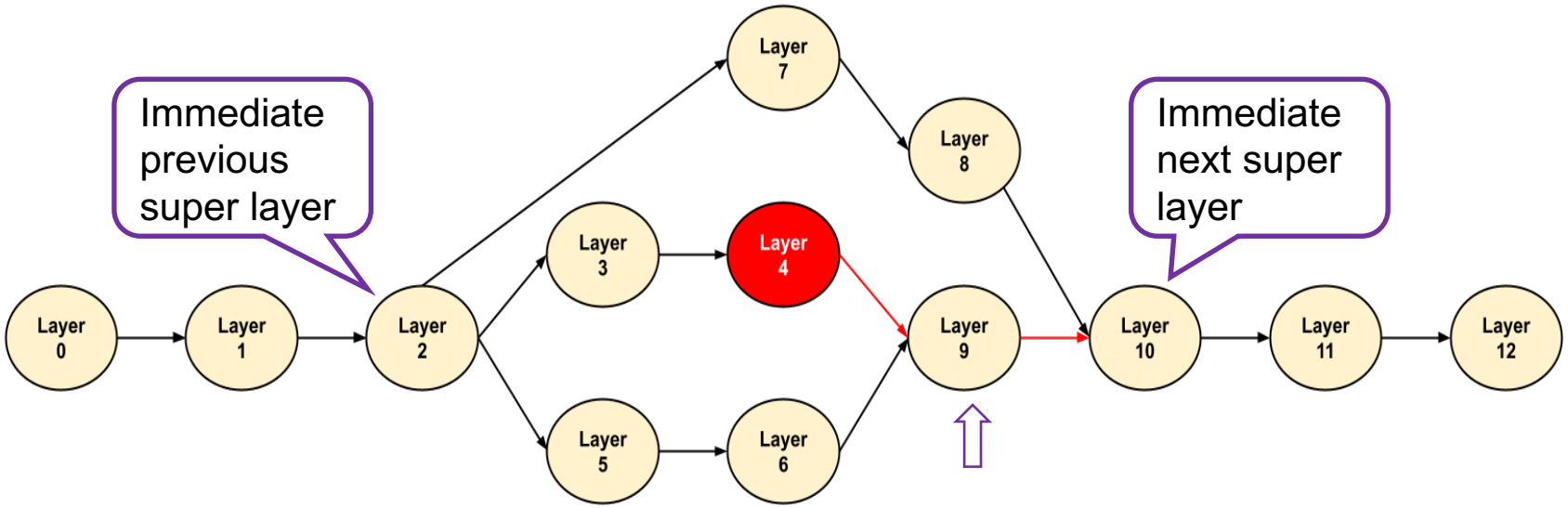
Simulation of FI technique of TensorFI+



MDict

Layer 6
Layer 8
Layer 4
Layer 2

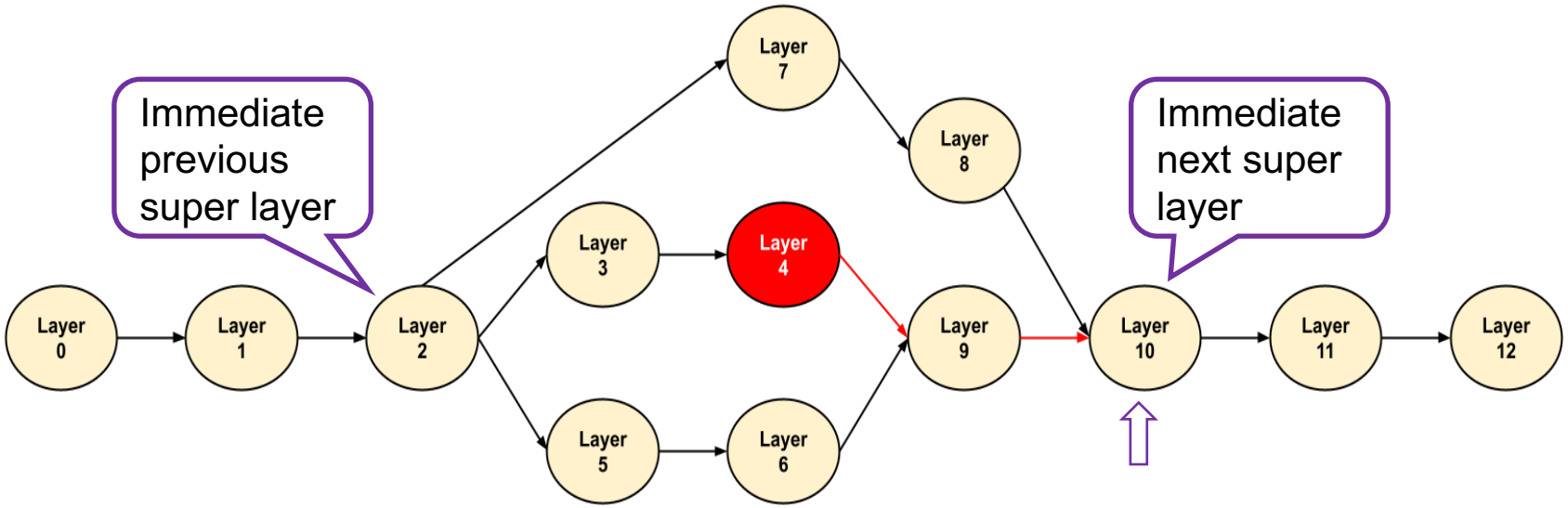
Simulation of FI technique of TensorFI+



MDict

Layer 9
Layer 6
Layer 8
Layer 4
Layer 2

Simulation of TensorFI+



MDict

Layer 10
Layer 9
Layer 6
Layer 8
Layer 4
Layer 2

👉 Finally Compute the output of layer 12 using only one KerasAPI(12, 10, inputs(10)) call

Benchmark & Experimental Setup

- Demonstrated on 30 popular DNN models
  - VGGNets, ResNets, DenseNets, Inception, Xception
- 3 open-source widely used datasets
  - CIFAR-100, ImageNet, GTSRB(traffic sign)
- 3000 random fault injections per DNN model
- Measured Silent Data Corruption(SDC) rate in the evaluation
  - Prediction mismatch from the fault free DNN inference

Results: SDC rates

Dataset	Model	Top-1 Accuracy	SDC Rate
ImageNet	VG16(Sequential)	71.18%	3.53%
	ResNet50(Non-sequential)	74.76%	1.43%
	DenseNet121(Non-sequential)	75.04%	1.20%
CIFAR-100	VGG19(Sequential)	71.53%	1.23%
	GoogleNet(Non-sequential)	76.70%	1.57%
	Xception(Non-sequential)	77.96%	2.00%
GTSRB	VGG16(Sequential)	97.57%	0.80%
	ResNet101(Non-sequential)	98.55%	0.67%

SDC rates range from 0.53% to 2.07% (error bars range from 0.10% to 2.95%) across different DNN models

Results: Performance Overhead

Dataset	Model	Overhead
ImageNet	VG16(Sequential)	1.81x
	ResNet50(Non-Sequential)	5.37x
	DenseNet121(Non-Sequential)	17.04x
CIFAR-100	VGG19(Sequential)	2.78x
	GoogleNet(Non-Sequential)	18.58x
	Xception(Non-Sequential)	10.29x
GTSRB	VGG16(Sequential)	2.45x
	ResNet101(Non-Sequential)	4.10x

On average, fault injected inference time is 7.62x higher than fault free inference

Conclusion

- Built a FI tool, TensorFI+, for both sequential and non-sequential DNN resilience evaluation
  - Demonstrated on 30 popular both sequential and non-sequential DNN models with 3 widely used datasets
- TensorFI+ is ready for research in DNN resilience
  - Open source at https://github.com/sabuj7177/TensorFIPlus

Md Hasanur Rahman
2nd Year CS PhD Student
University of Iowa
https://hasanur-rahman.github.io

IOWA