X-GDR BarCamp sur les Défis d'Implémentation de l'IA – Sécurité, Fiabilité, Soutenabilité et Nouvelles Technologies

Reliability of Spiking Neural Network VLSI Implementations

Haralampos-G. STRATIGOPOULOS

Sorbonne Université, CNRS, LIP6

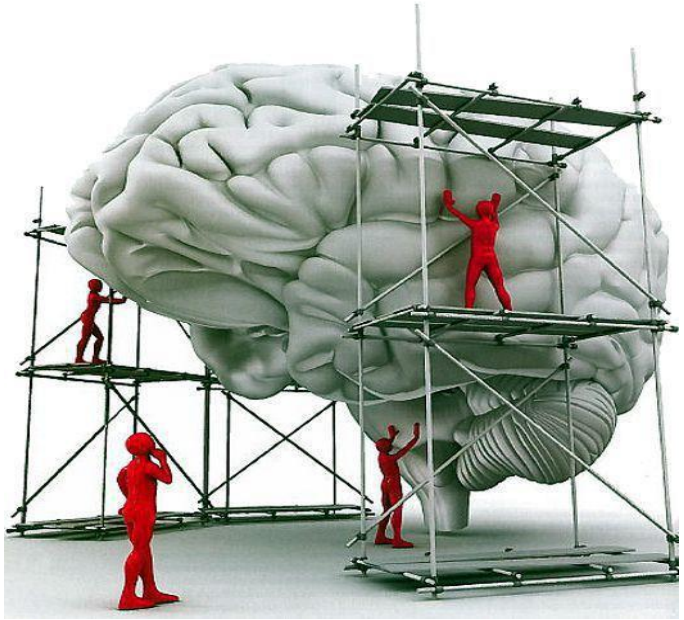
Paris, France

Outline

- Introduction to brain-inspired computing and SNNs
- Testability and reliability framework for SNNs:
  - Fault modeling
  - Fault injection frameworks
  - Reliability analysis
  - Testing strategies
  - Fault tolerance strategies
- Conclusions

Brain-inspired neuromorphic computing

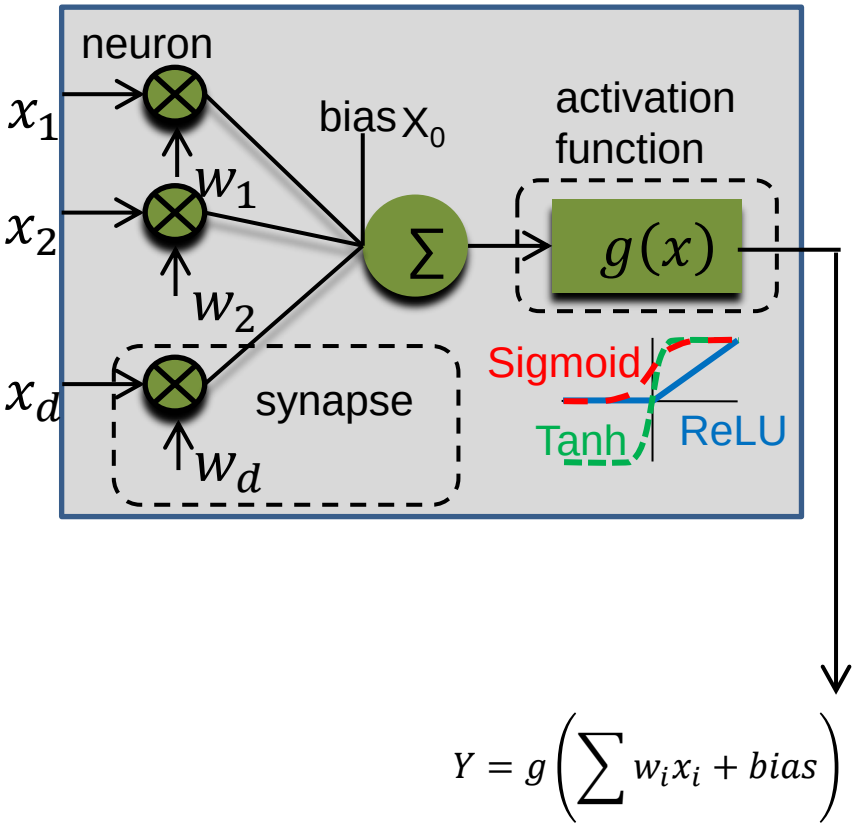
- Brain is the most brilliant computing machine
- Very “green”
- Computational power efficiency orders of magnitude higher than computers
- Brain has augmented capabilities (learning, produces ideas, error resilience…)



W. Gerstner, *Neuronal dynamics*

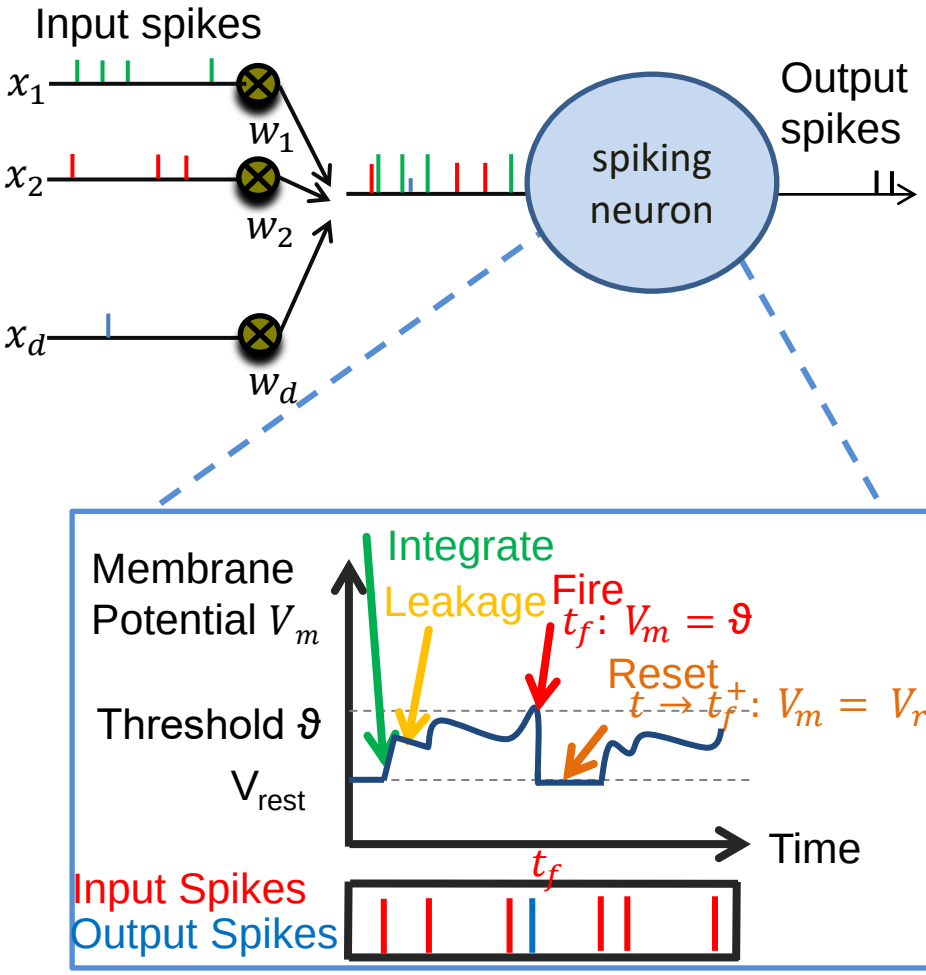
Spiking neural networks (SNNs)

ANNs

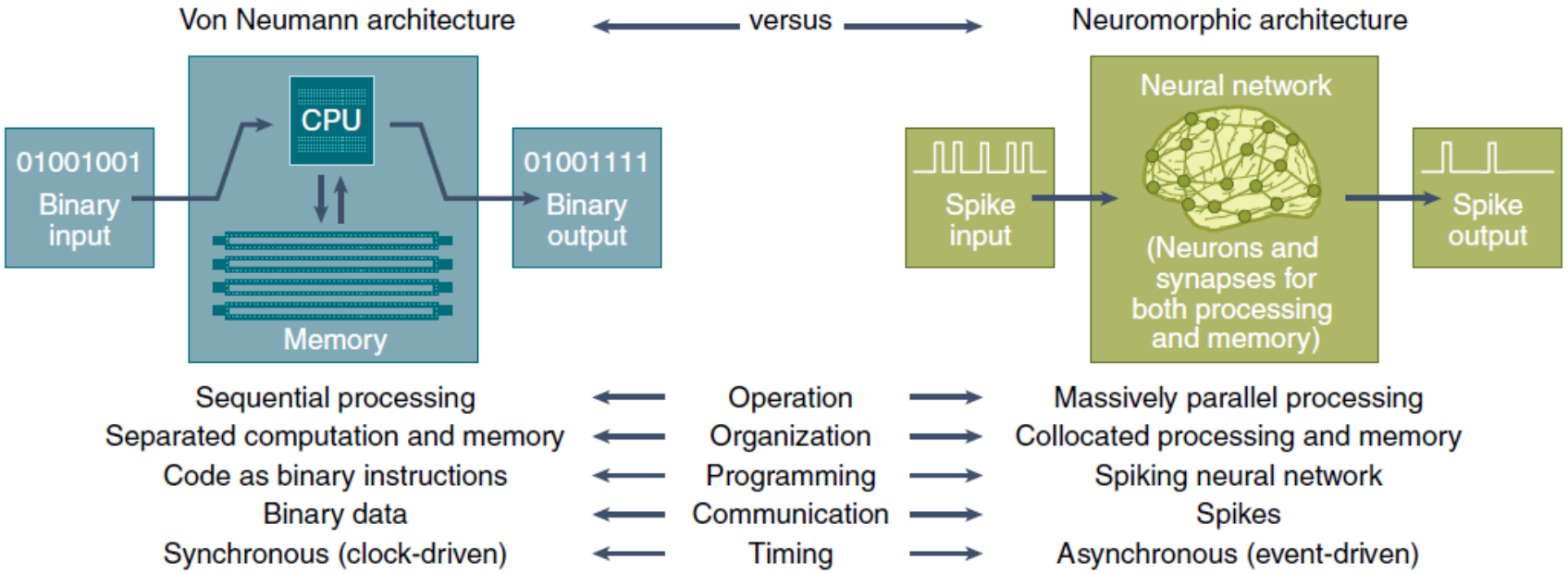


$$Y = g\left(\sum w_i x_i + bias\right)$$

SNNs



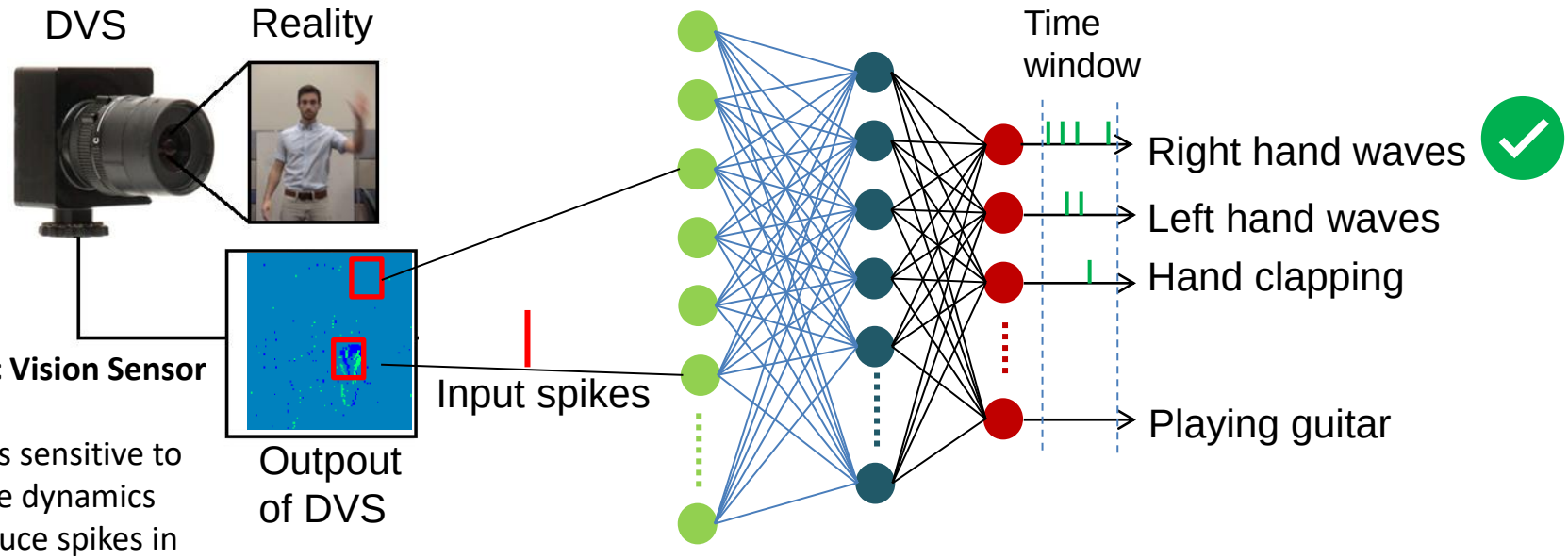
Von Neumann vs. neuromorphic architecture



- SNN is a dynamic system → maps well to speech and image recognition
- More energy efficient as it is asynchronous in operation
- Speed of computation improved due to event processing
- Challenges in learning remain
- Challenges in developing hardware

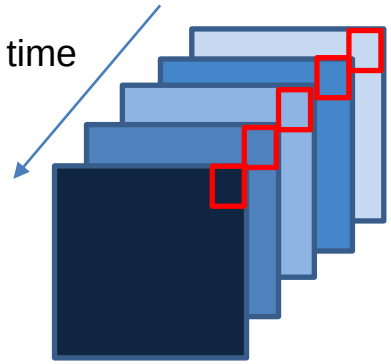
D. C. Schumman, Nature Computational Science, 2022

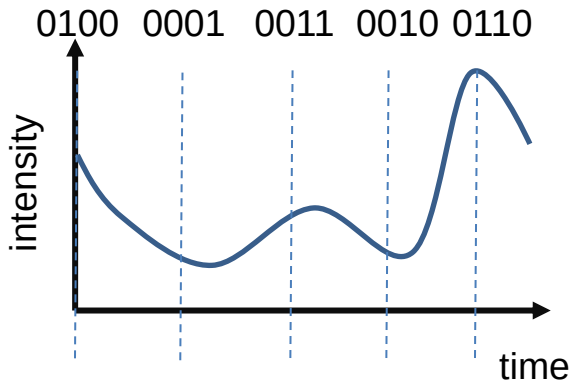
Data encoding into spikes

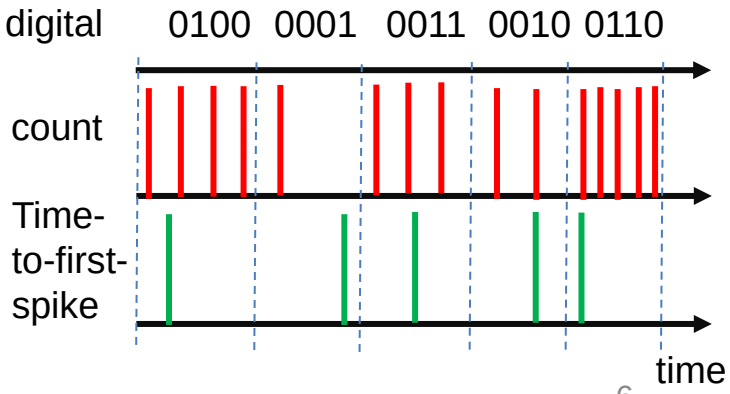


Dynamic Vision Sensor (DVS):

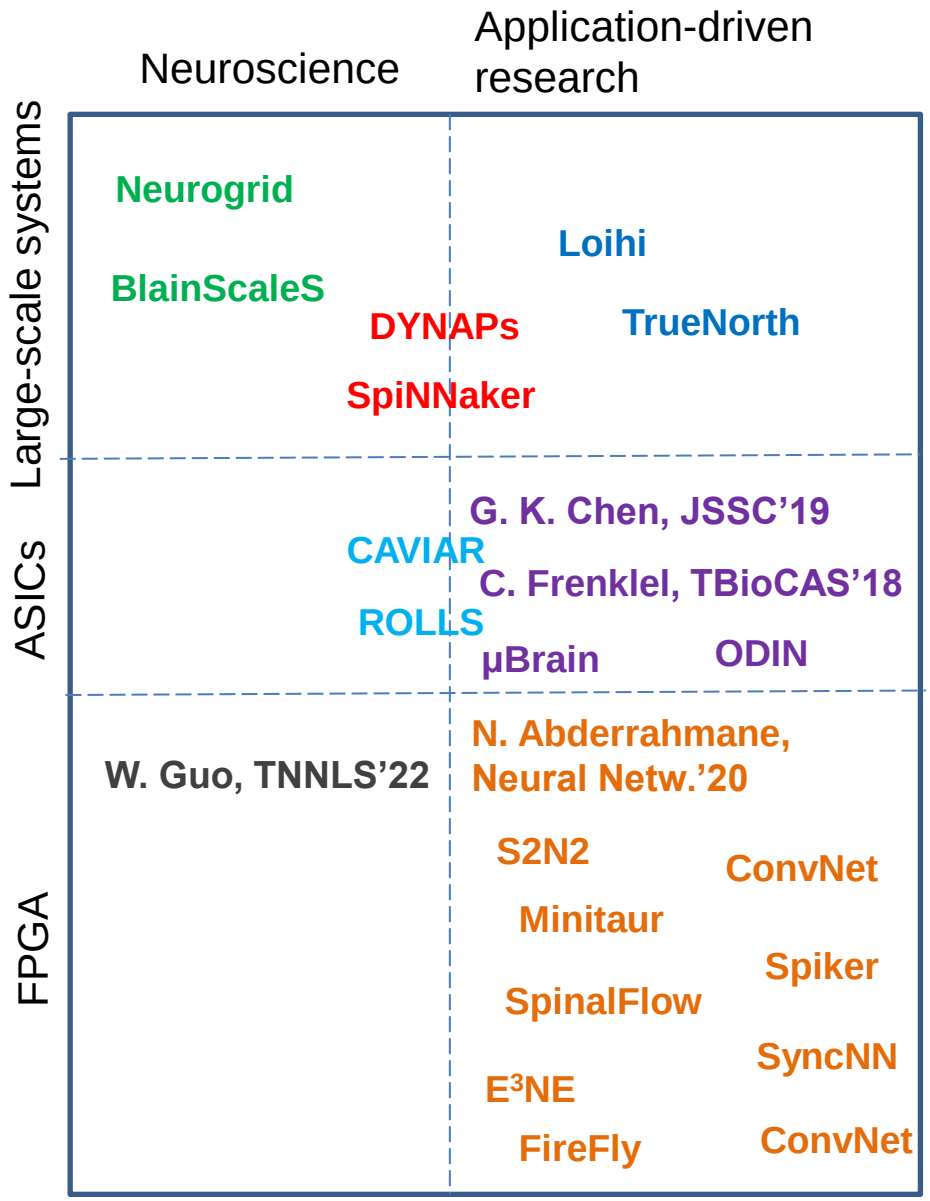
- Pixels sensitive to scene dynamics
- Produce spikes in response to the temporal intensity changes







Landscape of neuromorphic hardware



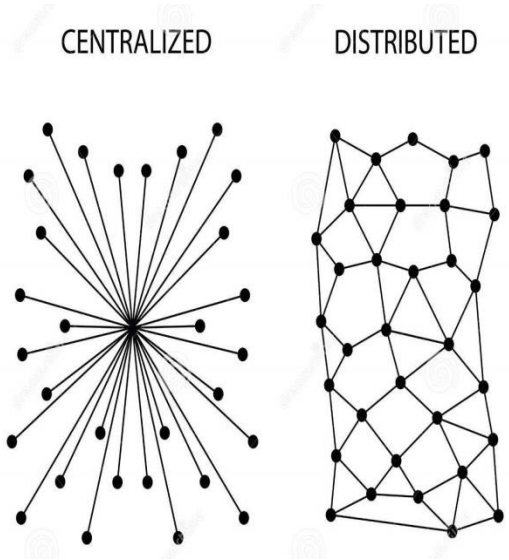
	SpiNNaker	Loihi	TrueNorth
Neurons/ core	36K	130K	1M
Synapses /core	2.8M	130M	256M
Cores/ chip	144	128	4096
Chips/ board	56	768	4096
Neurons	2.5B	100M	4B
Synapses	200B	100B	1T

Are AI hardware accelerators fault-tolerant?

They inherit the re… fault-tolerance capabilities of …gical brain

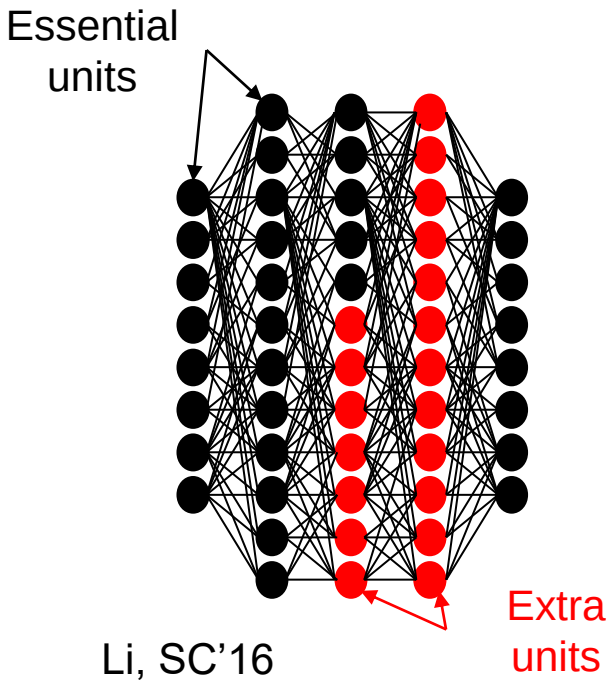
False

Distributed Processing



Biology-Inspired Architectures

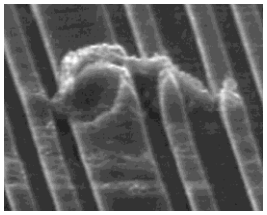
Overprovisioning

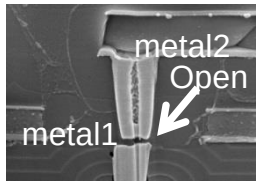


- Many fault simulation experiments have shown that this assumption is false

Li, SC'16
Reagan, DAC'18
Zhang, D&T'19
Vatajelu, VTS'19
Neggaz, D&T'20

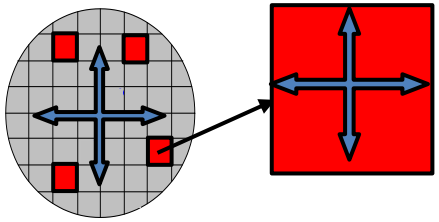
Hardware-level faults

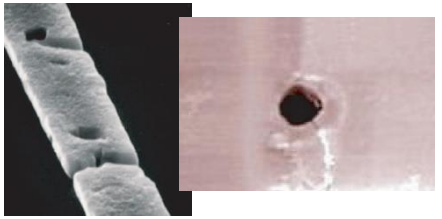


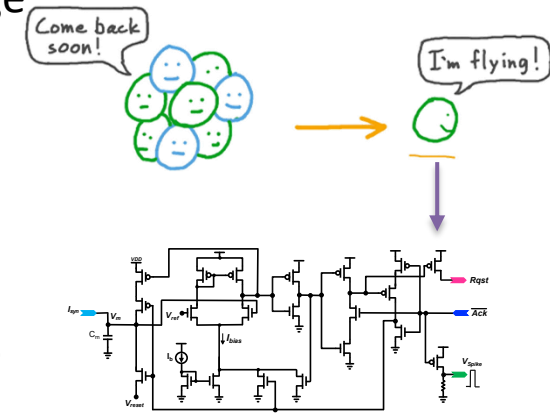


Hardware Faults

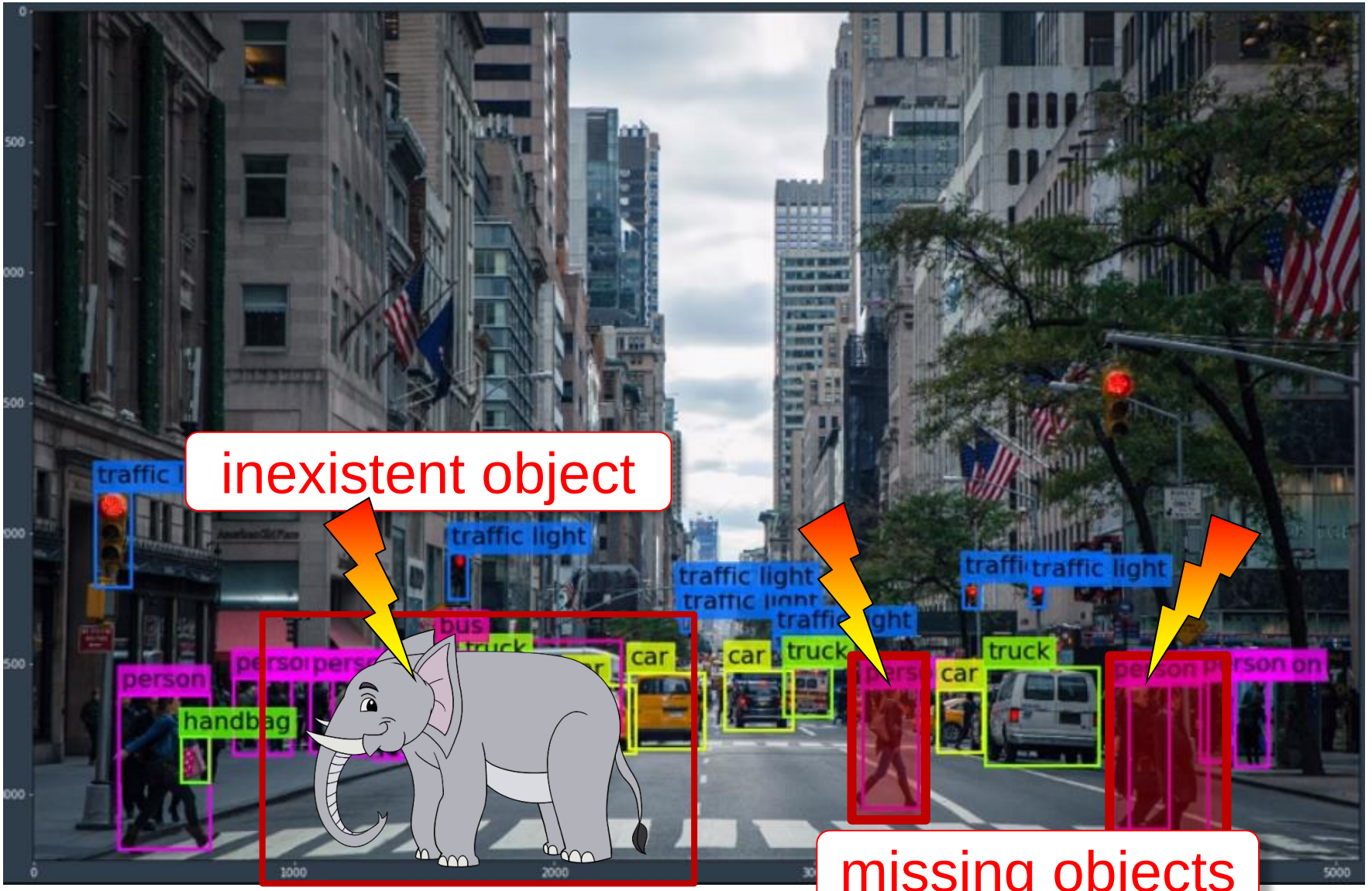
- During Manufacturing
  - Process variations
  - Defects
- In the Field
  - Aging
  - Radiation
  - Reduced-voltage memory operations
  - Low endurance (memristor synapses)
- Before Training
  - Can be overcome through training
- After Training
  - Can be catastrophic





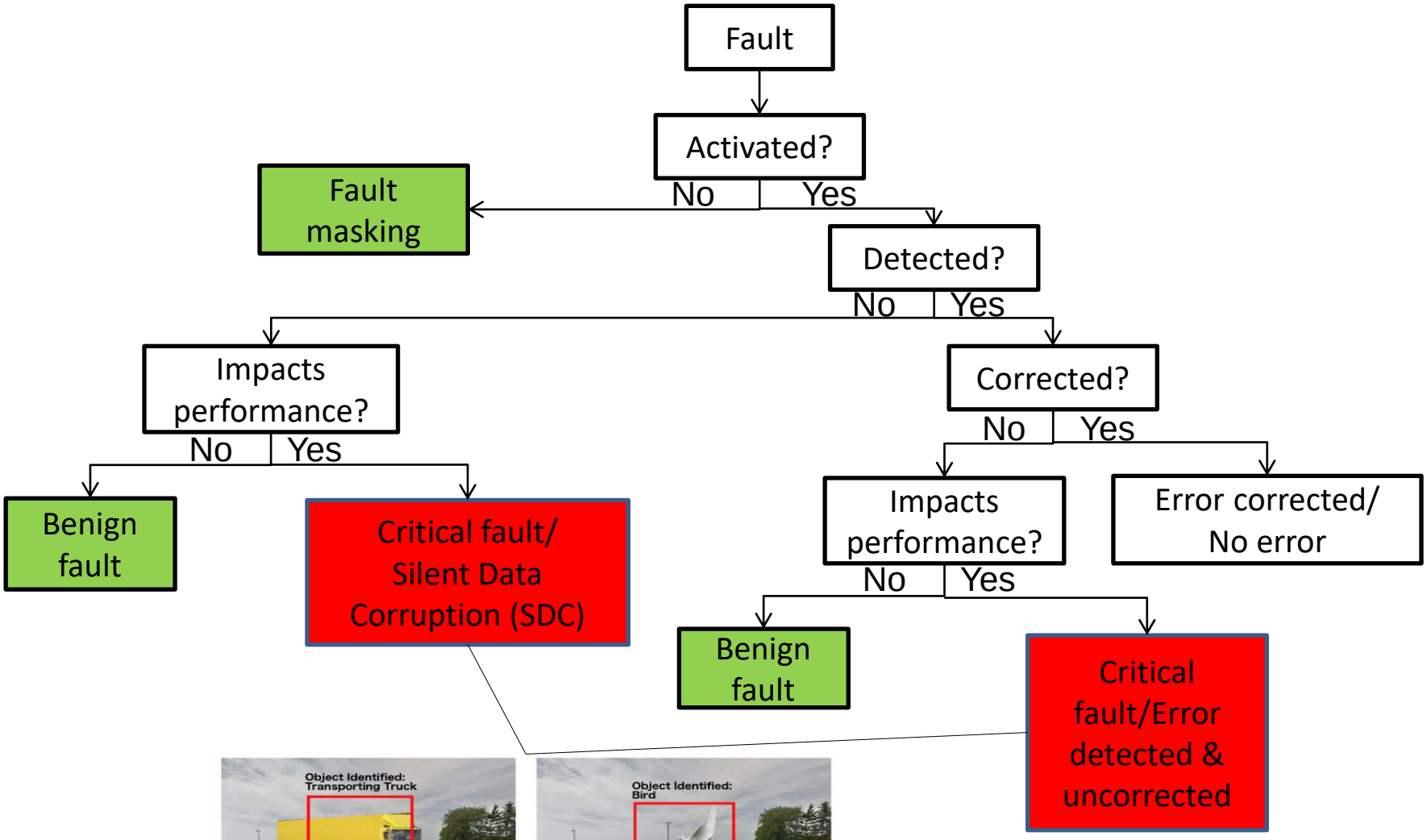


Neural networks reliability



Source: Paolo Rech, University of Trento

Not all faults are equal!



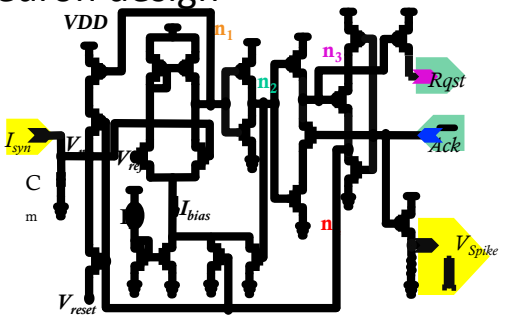
Fault-free

Critical fault

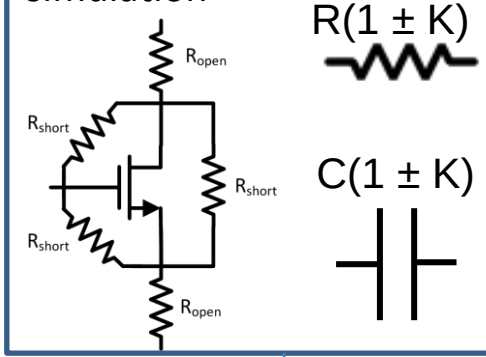
Testability and Reliability Framework

Fault modeling

Neuron design



Transistor-level fault simulation



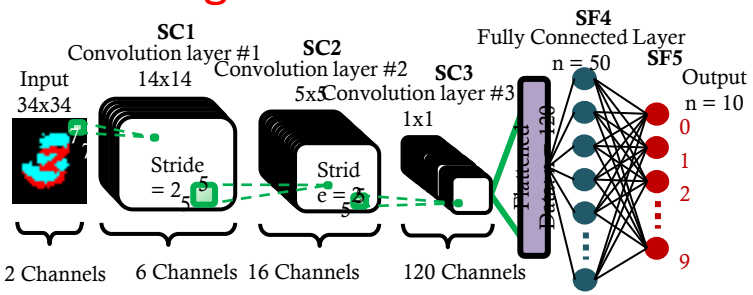
R(1 ± K)

C(1 ± K)

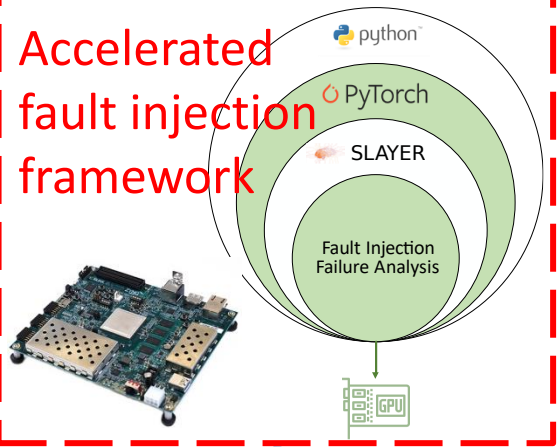
Faulty behaviors

Behavioral –level fault model

SNN designs



Accelerated fault injection framework

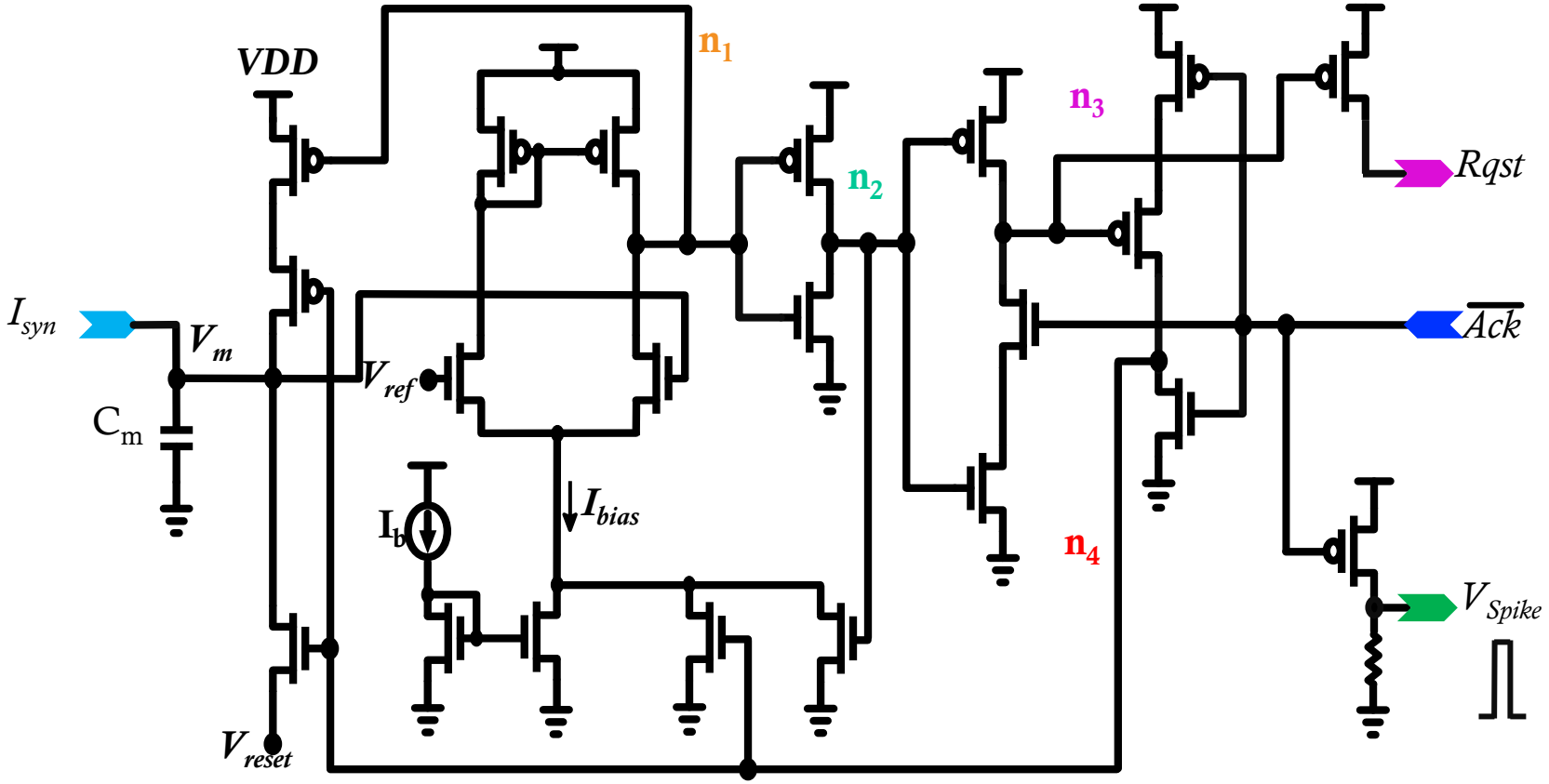


Reliability assessment

Testing & Fault tolerance

- Testing
- Model-based fault tolerance
- Hardware-based fault tolerance

I&F spiking neuron circuit

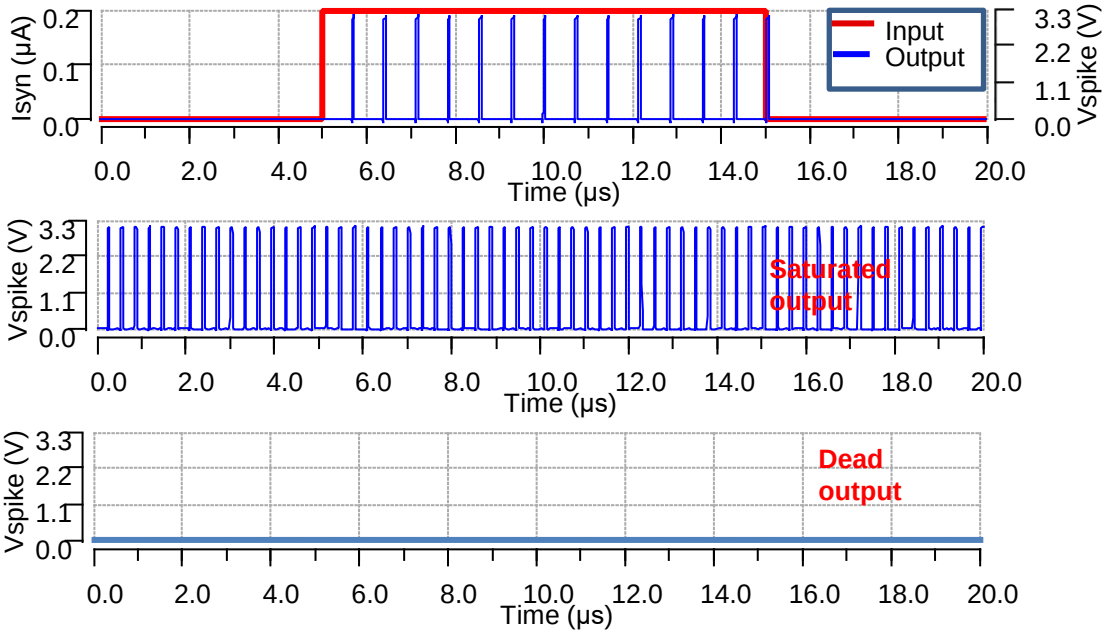


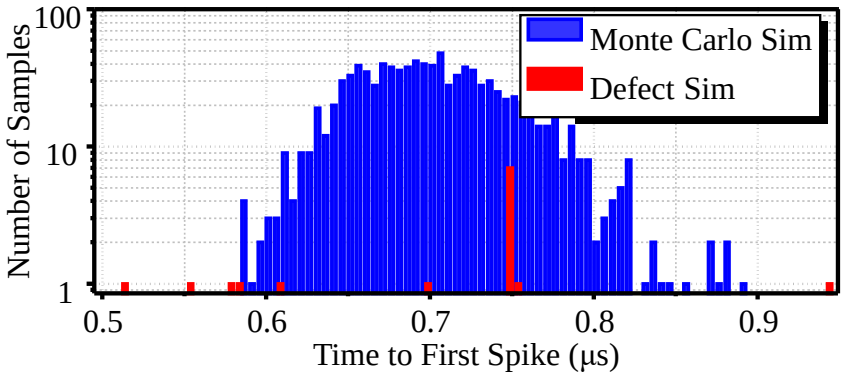
R. Serrano-Gotarredona, *TCAS'06*

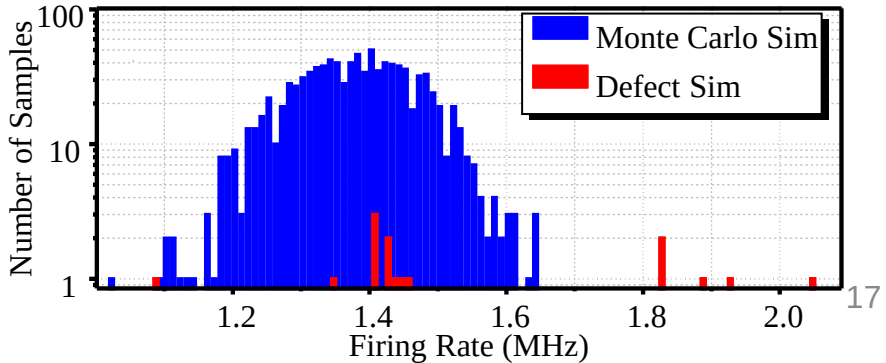
Transistor-level fault simulation

- 1000 MC runs using PDK
- Defect simulation using DefectSim by Siemens S. Sunter, *TCAS-I*'16
- Two types of faulty behaviors:
  - Catastrophic: neuron non-functional (observed for 31 defects)
  - Parametric: output spike train with timing variations (observed for MC and 15 defects)

S. Elsayed, *IOLTS*'20



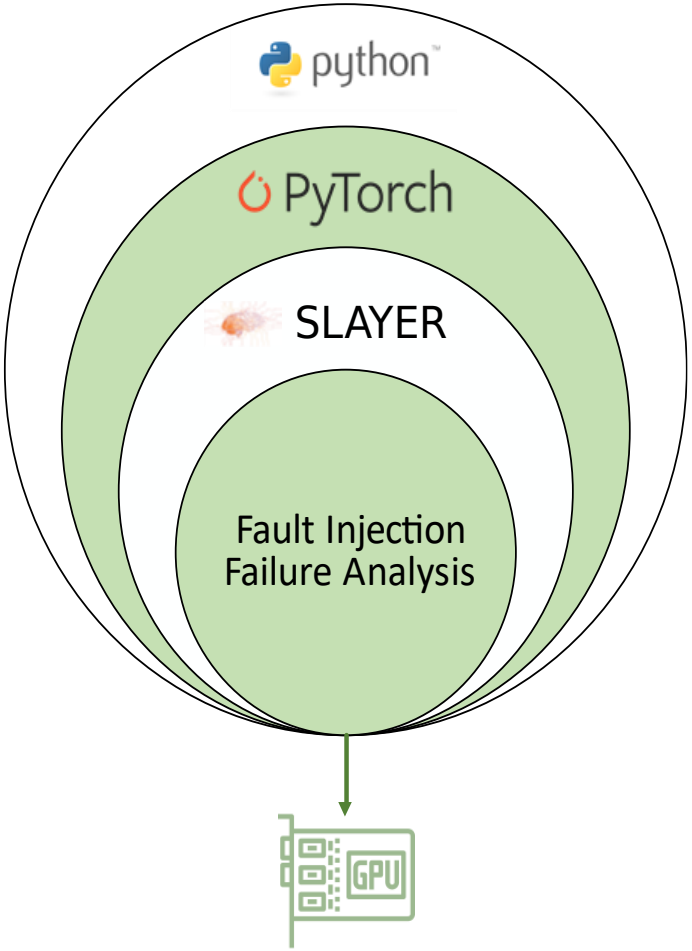




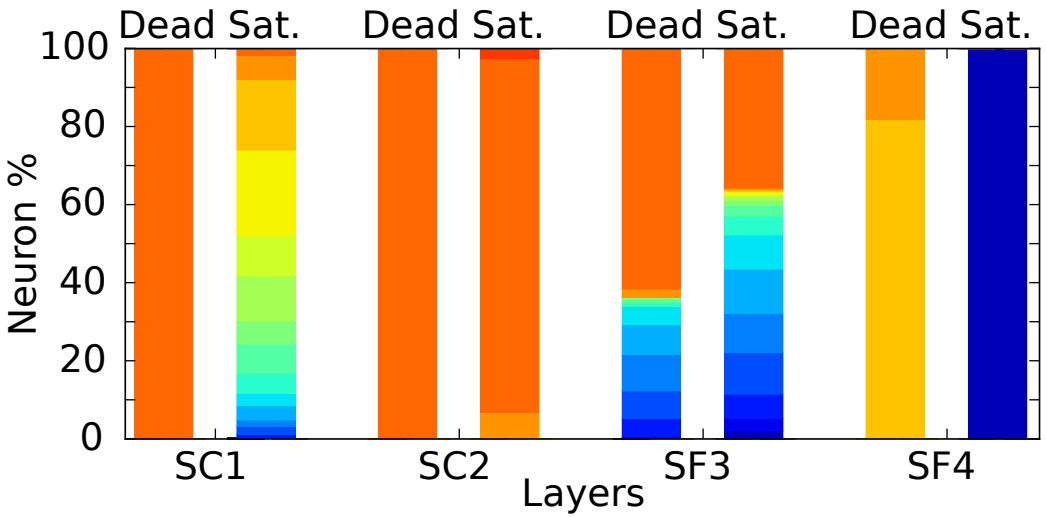
Software Fault Injection Framework

- SNNs modeled in Python using primitives from the Spike LAYer Error Reassignment (SLAYER) and PyTorch frameworks

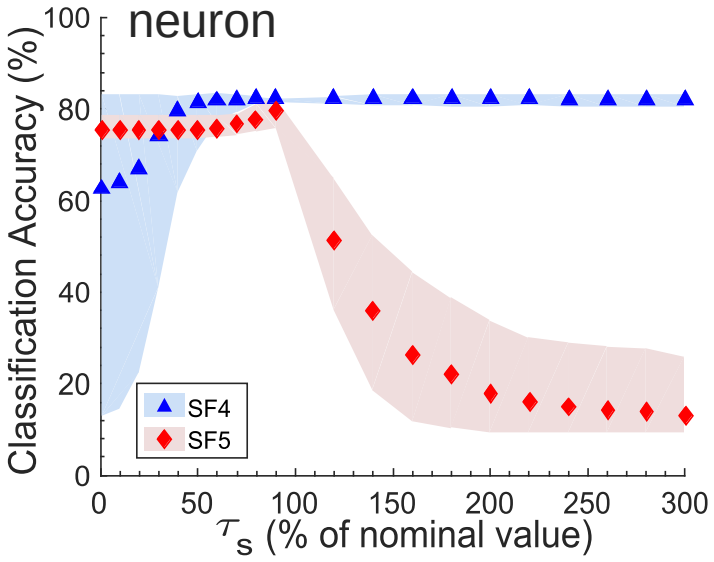
  S. B. Shrestha, *NeurIPS*'18
- Fault injection framework built on top of the SLAYER and PyTorch frameworks
- Fault injection and simulation are performed by customizing the flow of computations according to the faulty behavior
- Single and multiple faults
- Extendible fault model library
- Large-scale fault simulation acceleration: early stopping, late start, GPU
- Metric: classification accuracy drop for test set



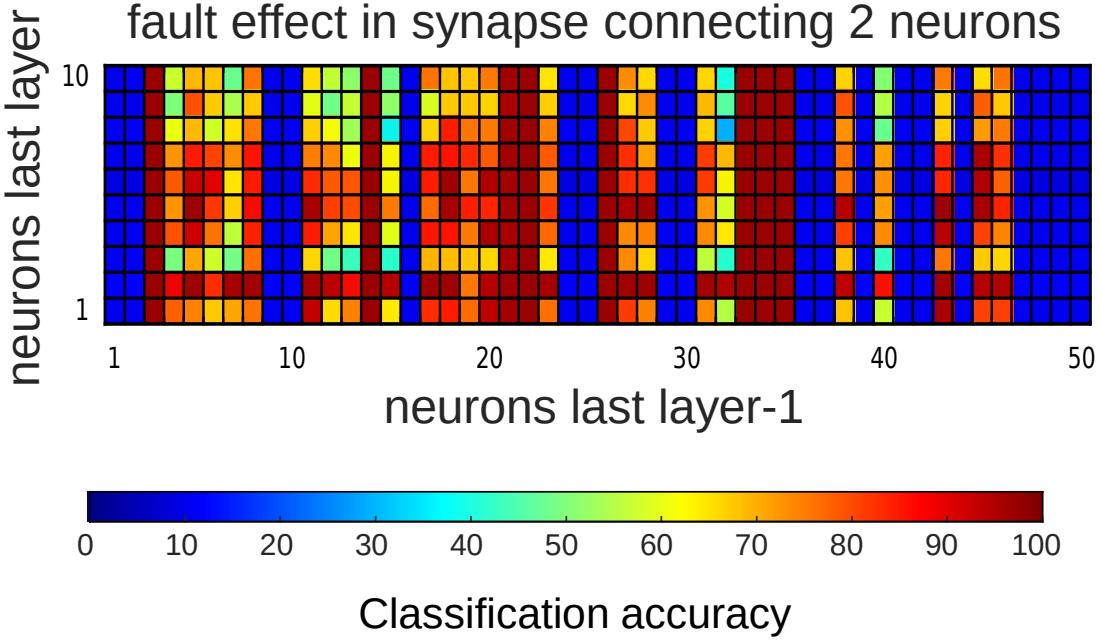
Reliability assessment



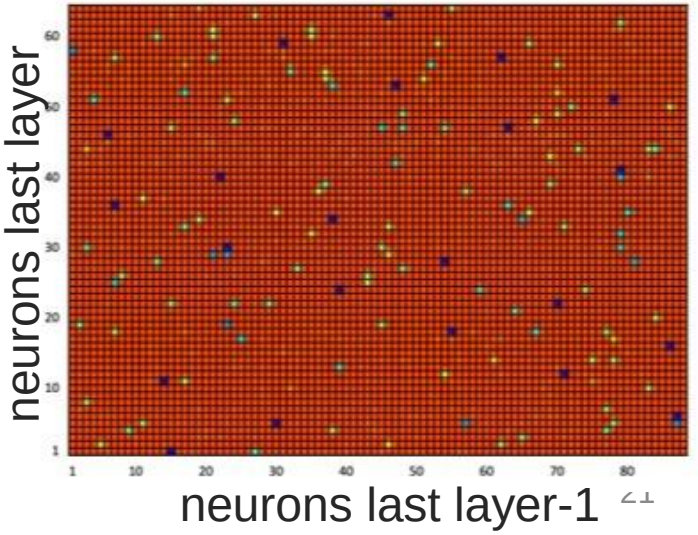
Variation in time constant of neuron



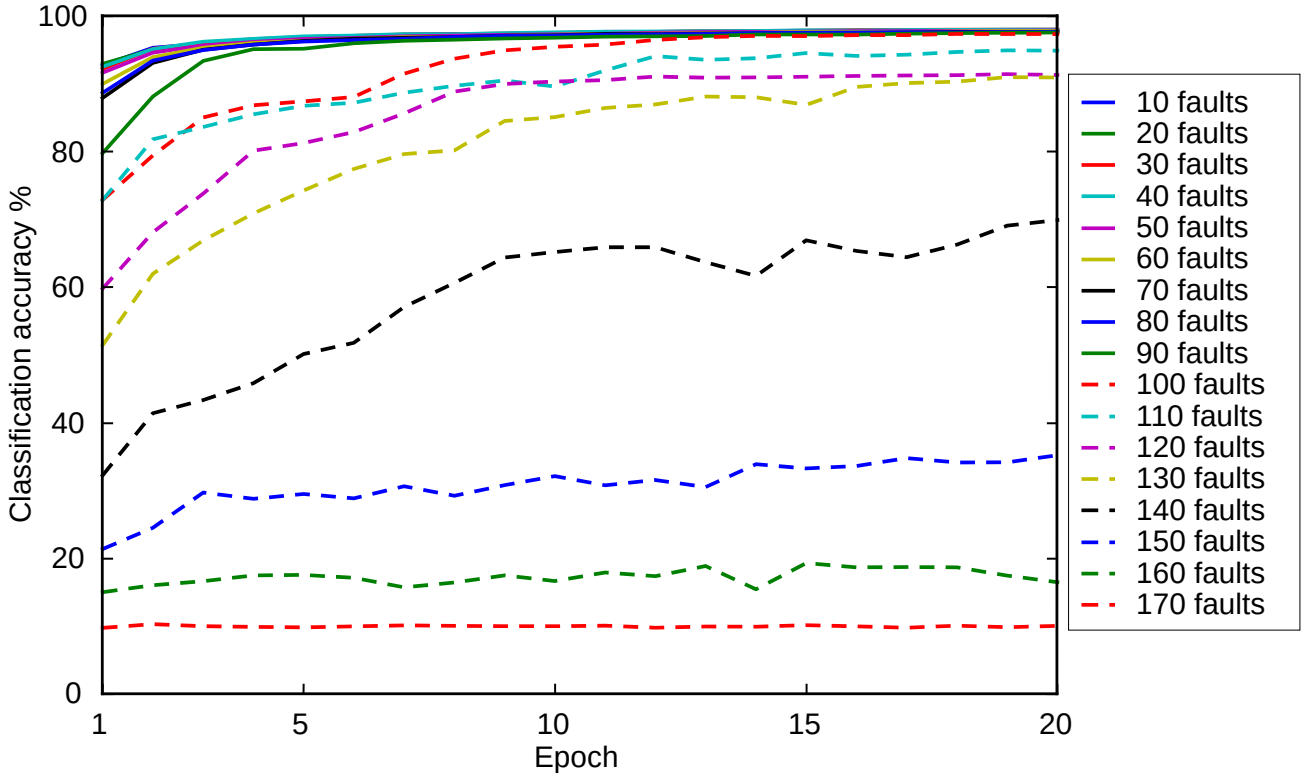
fault effect in synapse connecting 2 neurons



MSB flip in synapse connecting 2 neurons

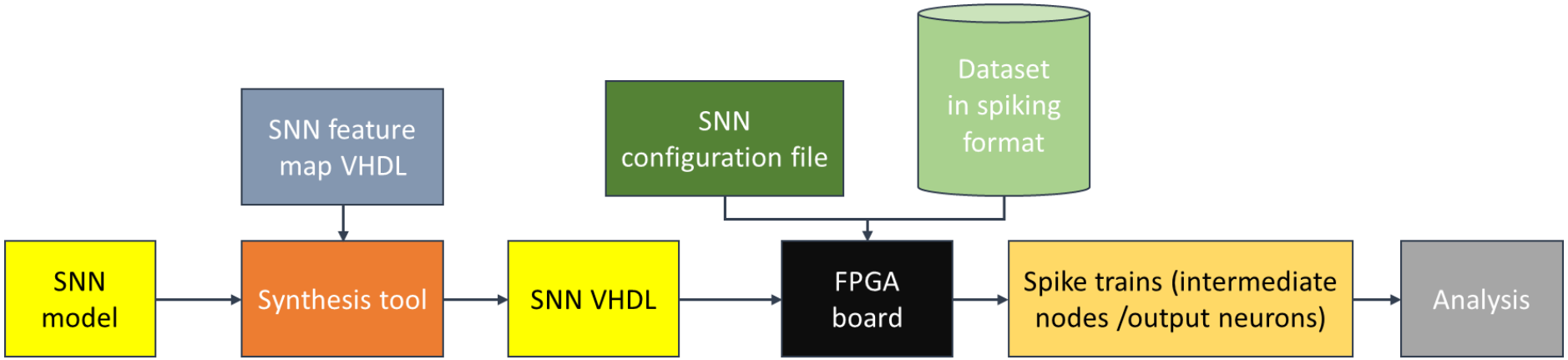


Faults occurring before training



- Networks can compensate for a high fault rate if faults occur before the training

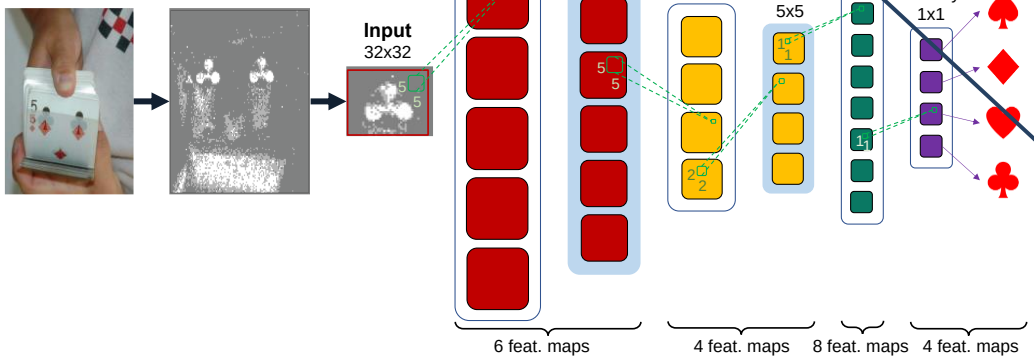
SNN hardware accelerator design framework

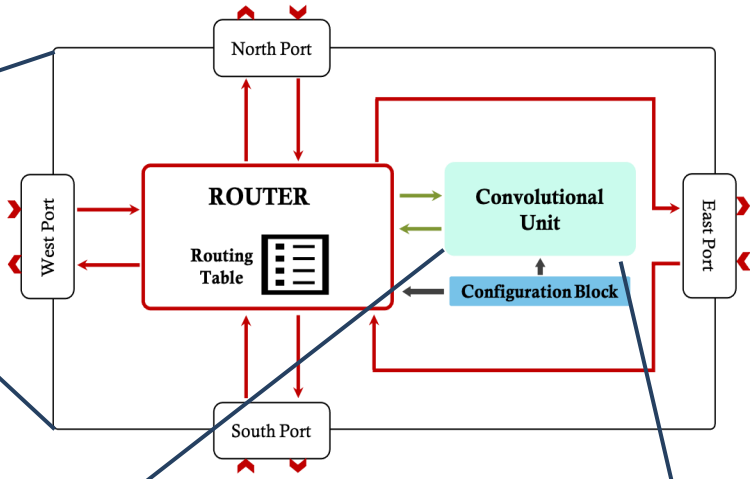


- End-to-end model-to-VHDL automated synthesis of arbitrary SNN
- FPGA implementation
- Fully synthesizable for an ASIC implementation
- Will be released as open-source

SNN hardware accelerator architecture

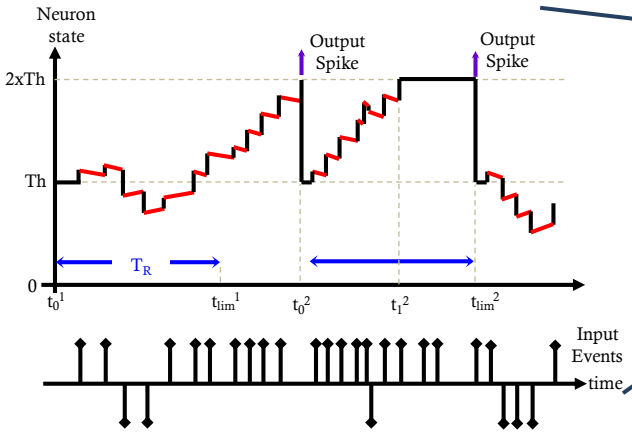
Dynamic Vision Sensor (DVS)

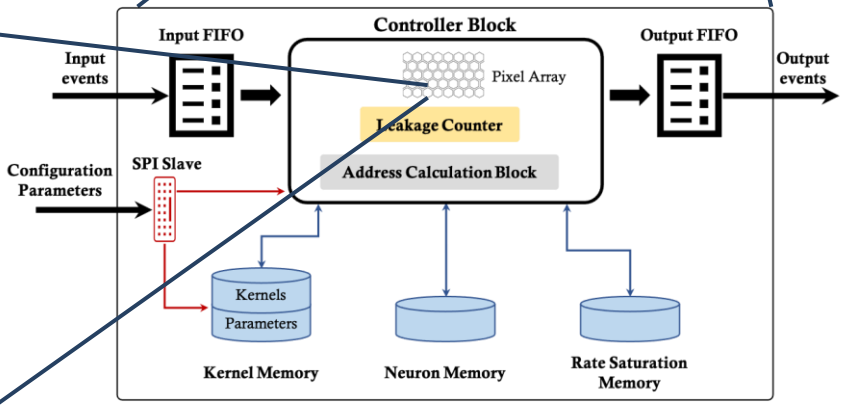




Supported features:

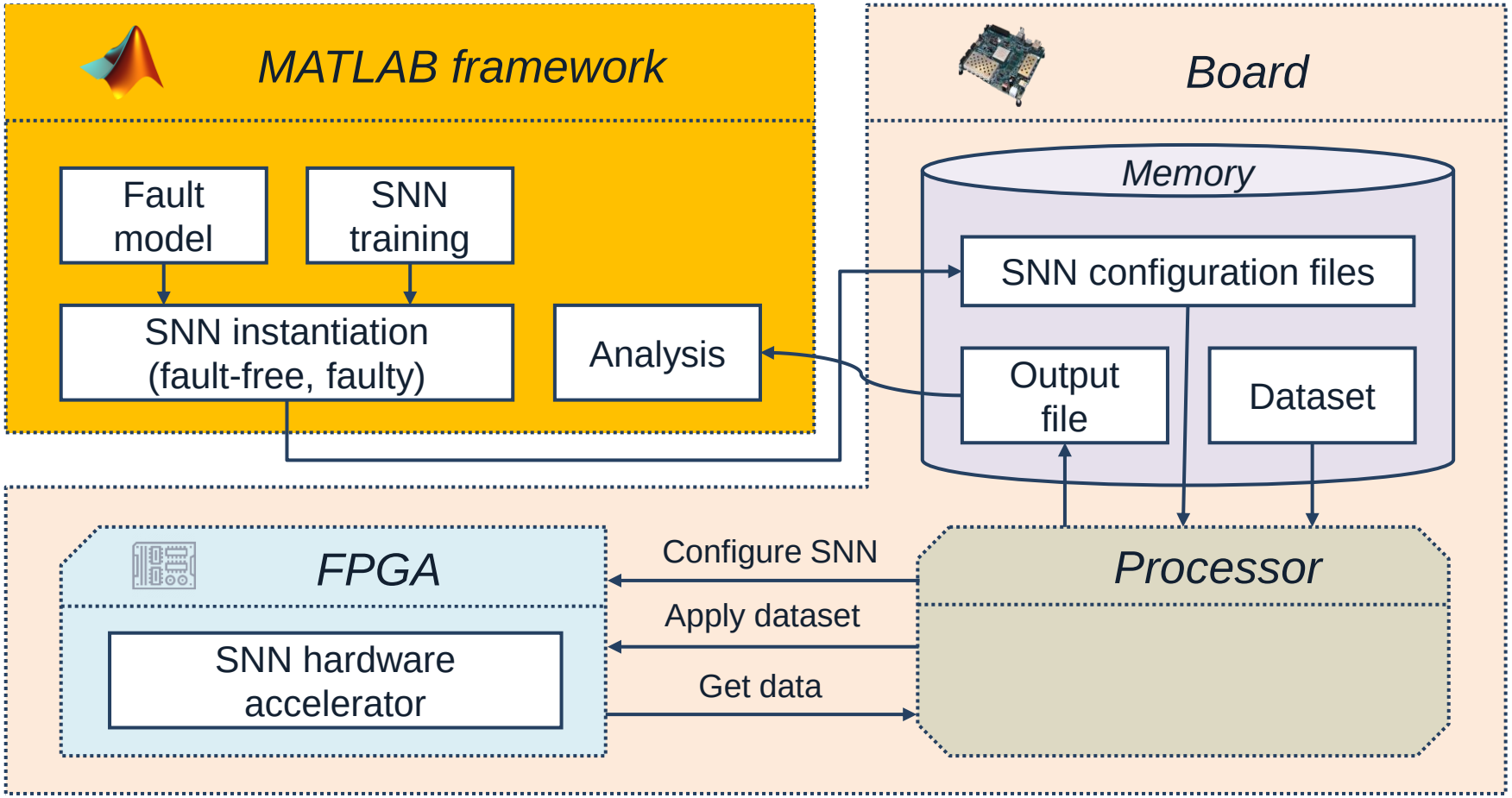
- Convolution
- Spike generation
- Leakage
- Rate saturation





L. Camuñas-Mesa, Front. Neurosci. ’18

SNN hardware experimentation platform

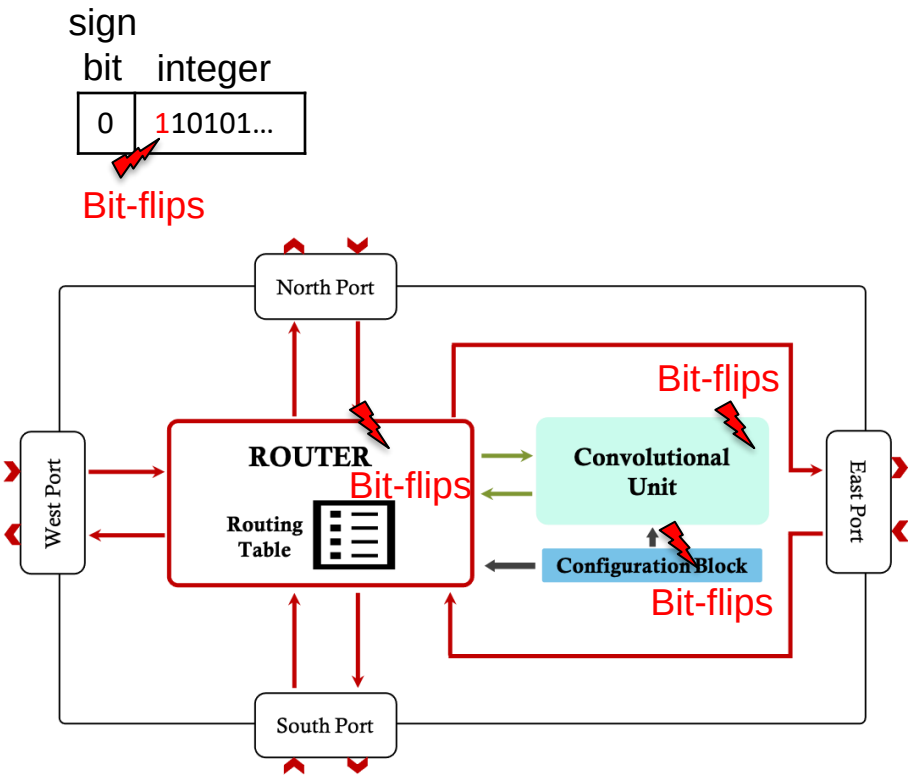


Reliability analysis of SNN hardware accelerator

T. Spyrou, *DATE*'22

- Each node is configurable through a set of 8-bit parameters
- Parameters are stored in memory blocks inside the node:

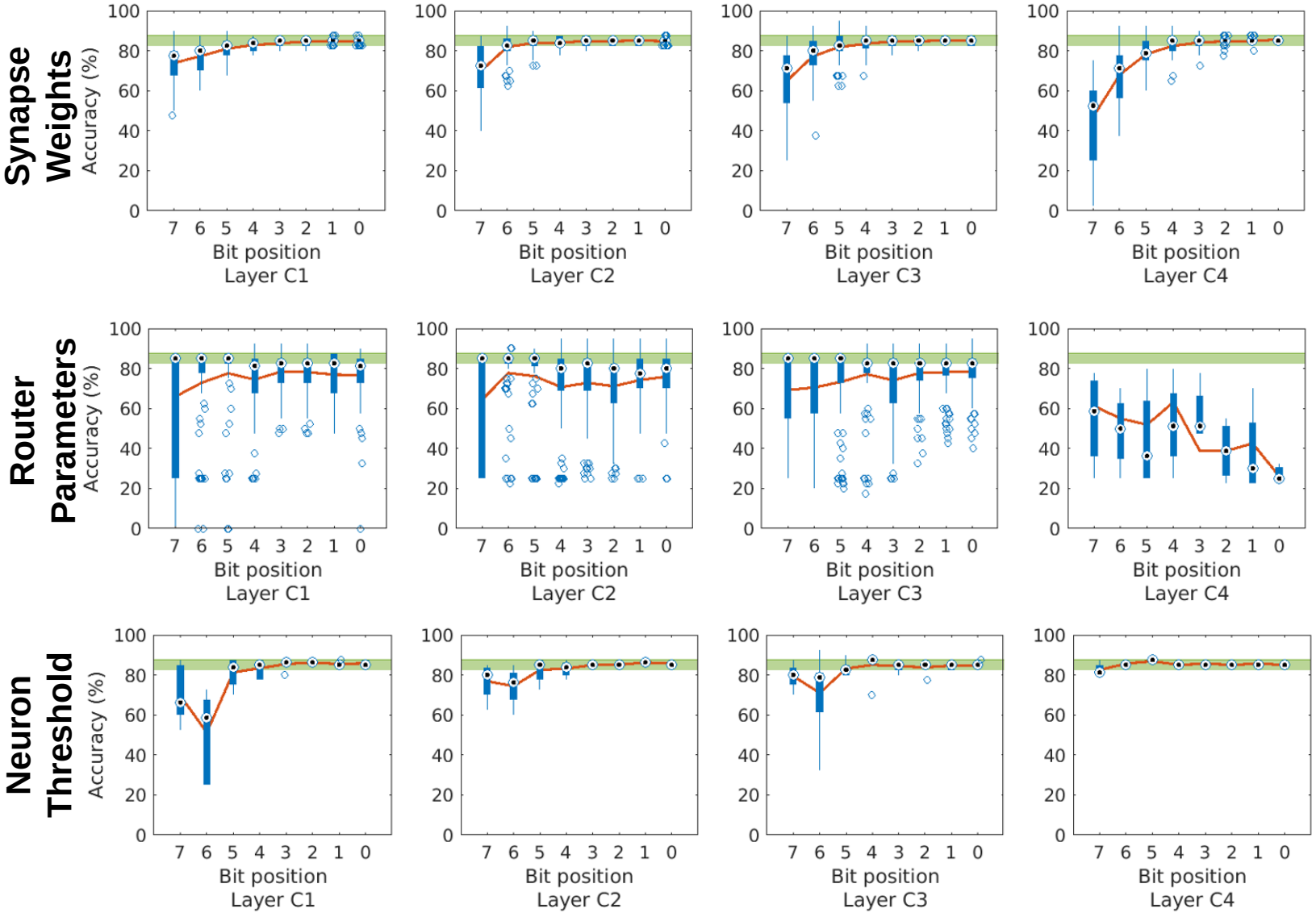
Memory	Purpose
Splitter Parameters	input split information to first layer
Router Parameters	routing information in the nodes' mesh
Neuron Parameters	key features of the neurons within the node
Kernel Parameters	kernels structural characteristics
Synapse Weights	values of the synaptic weights



- Fault model: bit-flips in memories
  - Single bit-flips across different bit positions
  - Multiple bit-flips with a BER probability

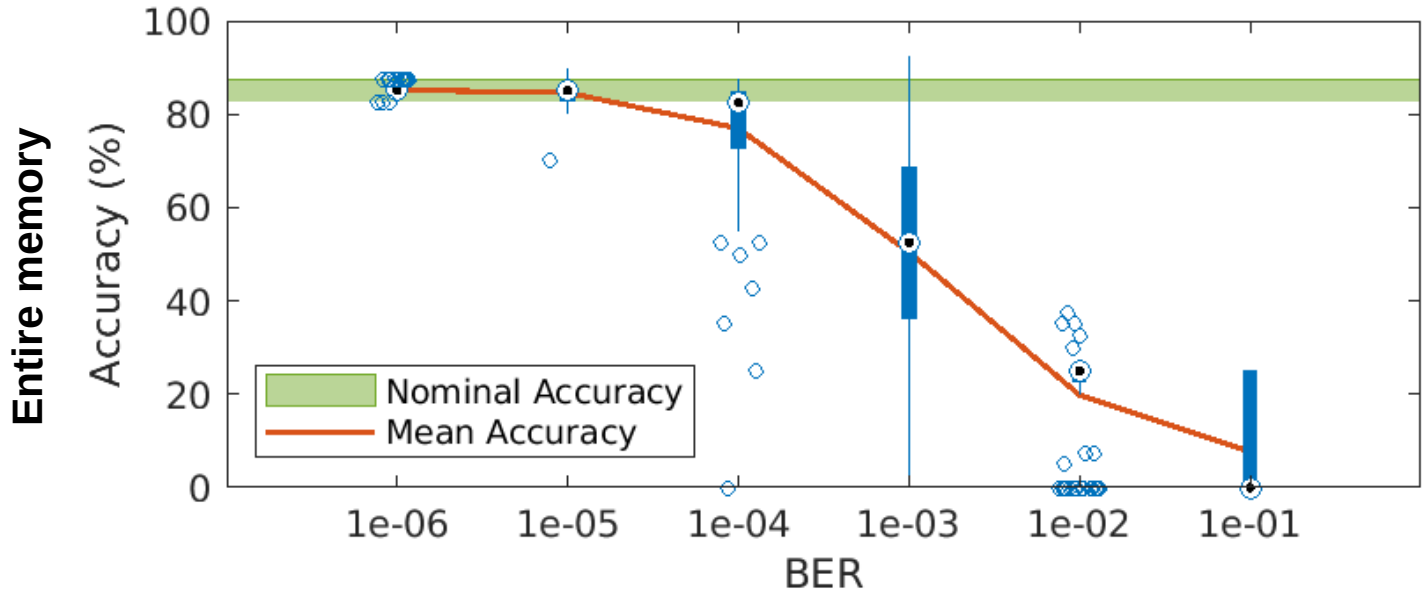
Reliability analysis results

T. Spyrou, *DATE*'22



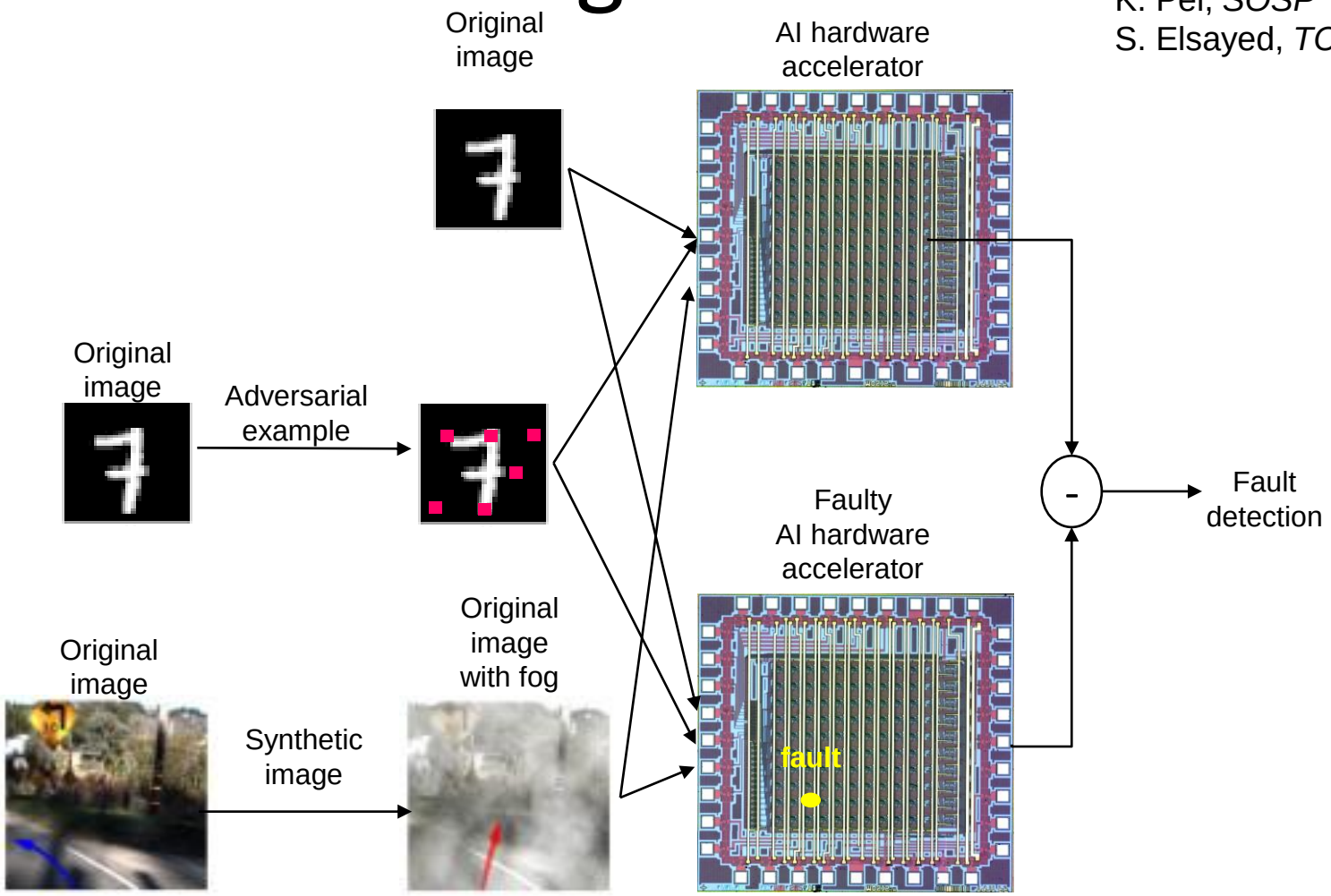
Reliability analysis results (cont'd)

T. Spyrou, *DATE*'22



Functional testing

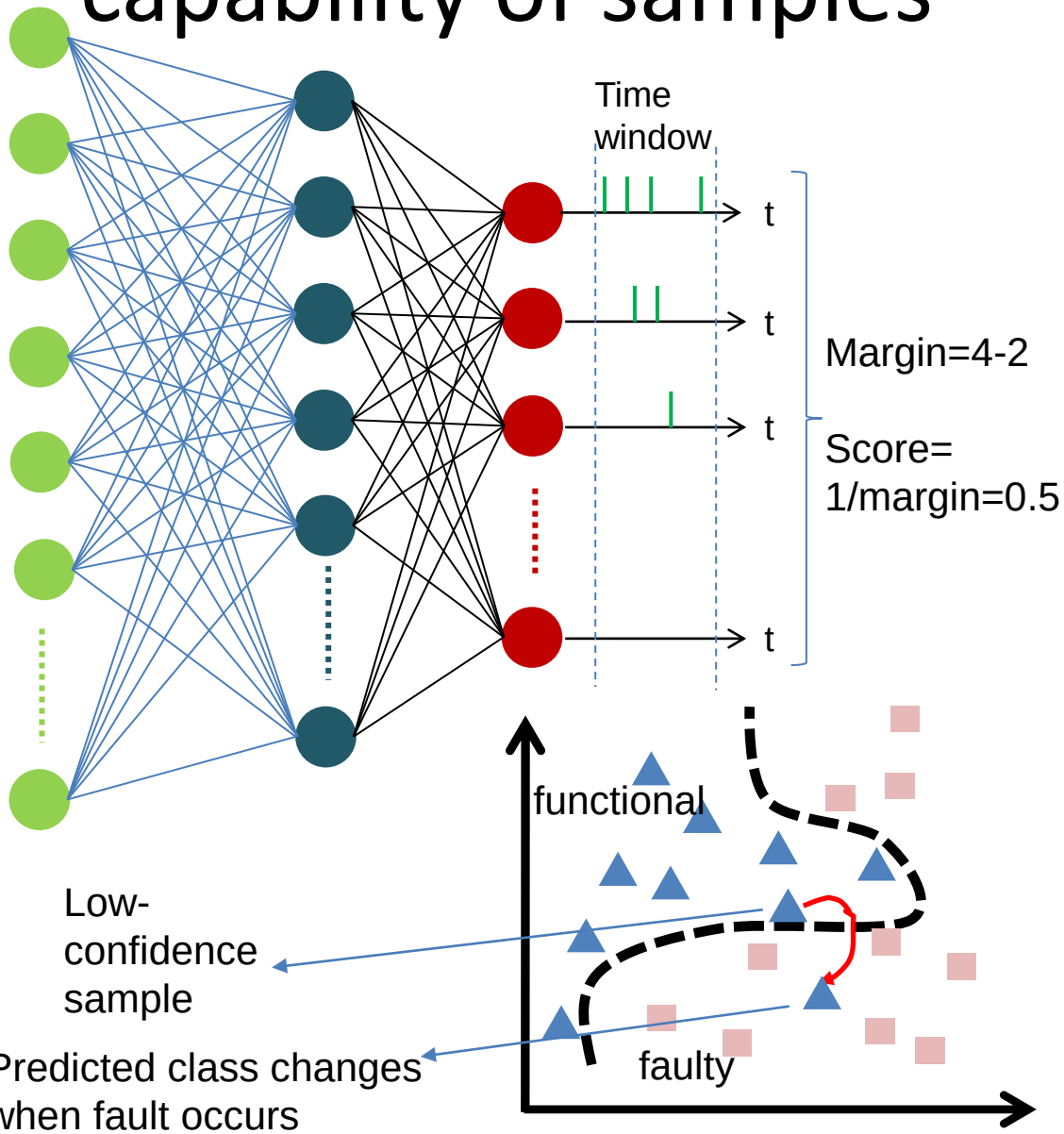
S. Kundu, *TVLSI*'21
H.-Y. Tseng, *ICCAD*'21
Y. Tian, *ICSE*'18
K. Pei, *SOSP*'17
S. Elsayed, *TCAD*'23



- Use existing samples in training/testing sets or craft new samples that can detect faults
- Fault is detected if responses of nominal/faulty chips differ

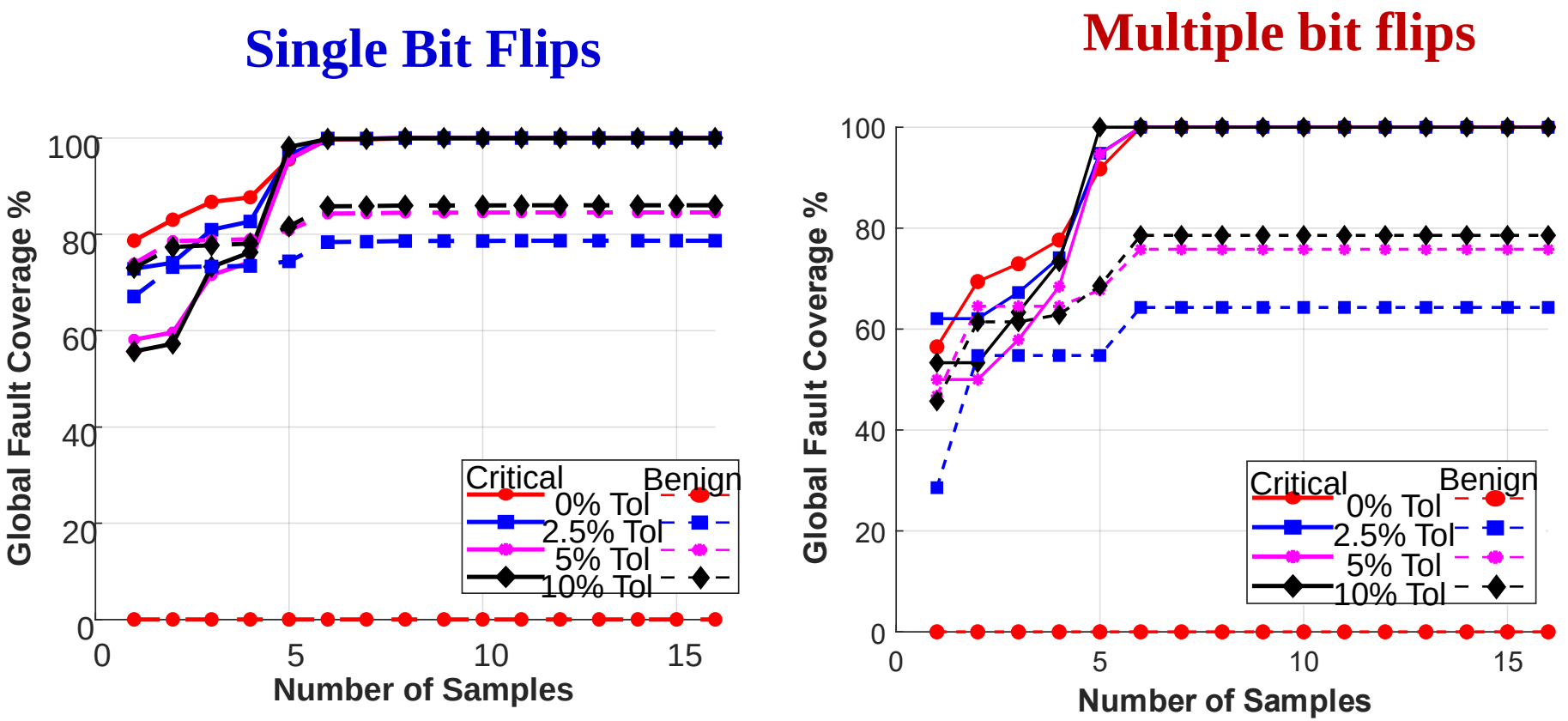
ATPG based on ranking fault detection capability of samples

S. Elsayed, *TCAD*'23



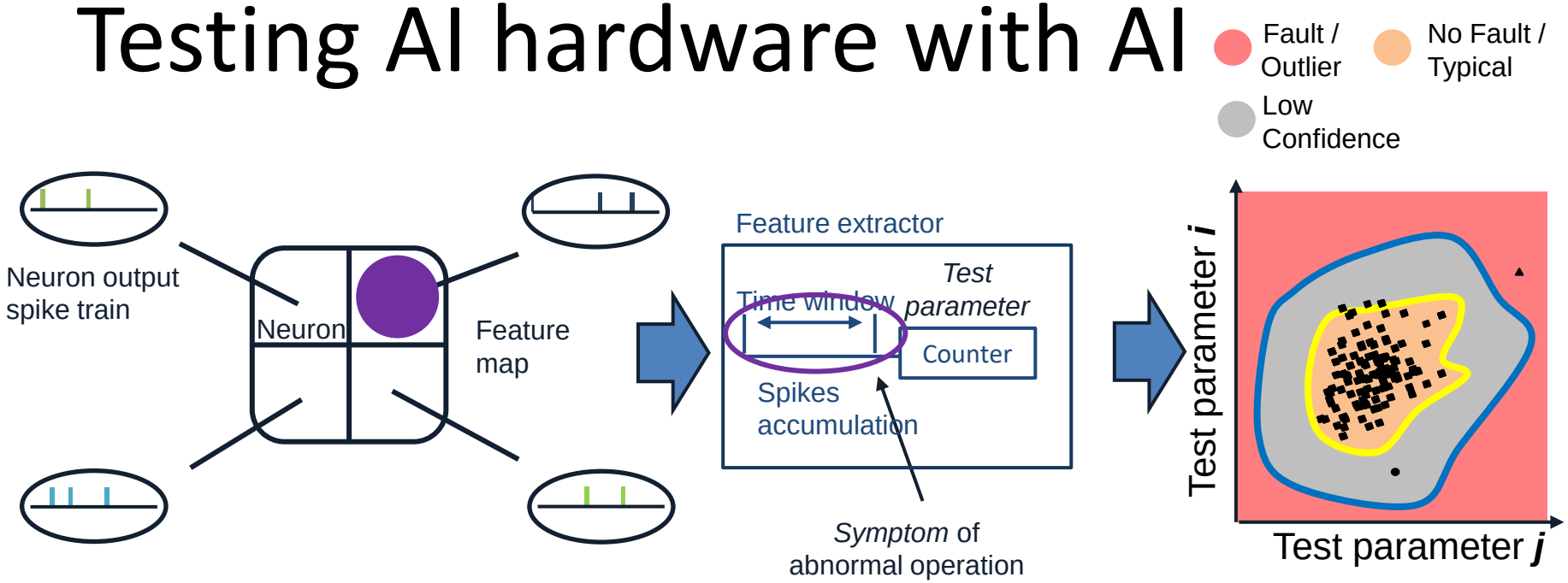
- Assess the fault coverage of an input sample <u>with no fault simulation</u>
- Fault coverage ∝ prediction confidence
- Proposed criterion: difference in output spikes between top-1 and top-2 classes
- Rank samples based on confidence in ascending order
- Add samples in the test-set according to ranking until fault coverage maximizes

Results on SNN hardware accelerator



- The global cumulative fault coverage curves quickly reach 100%
- 6 samples suffice to detect all critical faults and a high percentage of benign faults

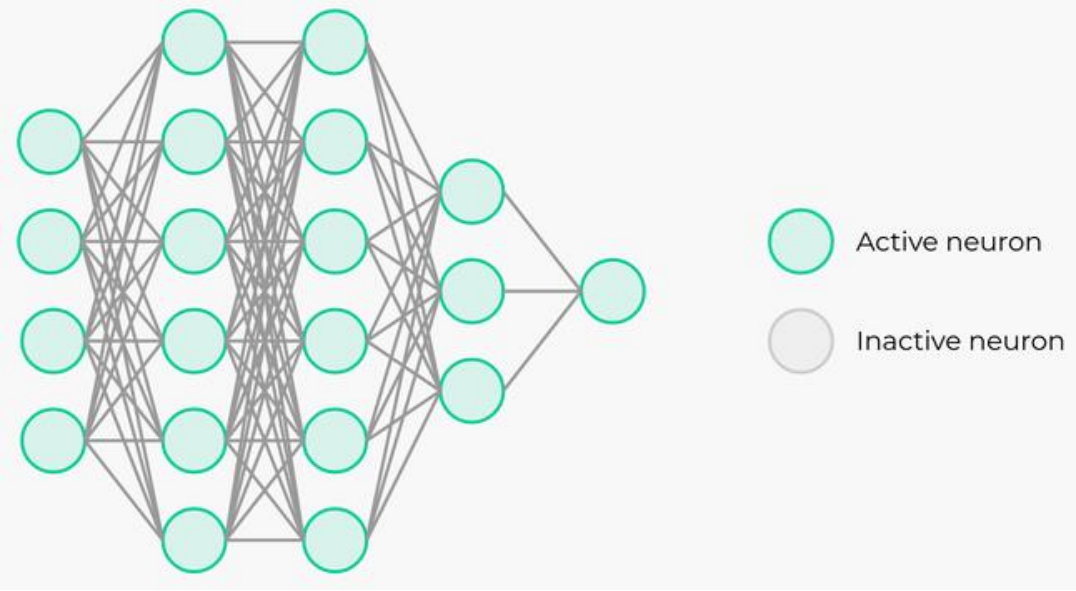
Testing AI hardware with AI



- Symptom detection
- Test parameter is the cumulative spike count at feature map output
- Use a system of two one-class classifiers for mapping test parameters to a decision
- One-shot decision (fault or no fault) with high confidence
- If low-confidence execute a reply operation to resolve ambiguity

Training with faults: dropout

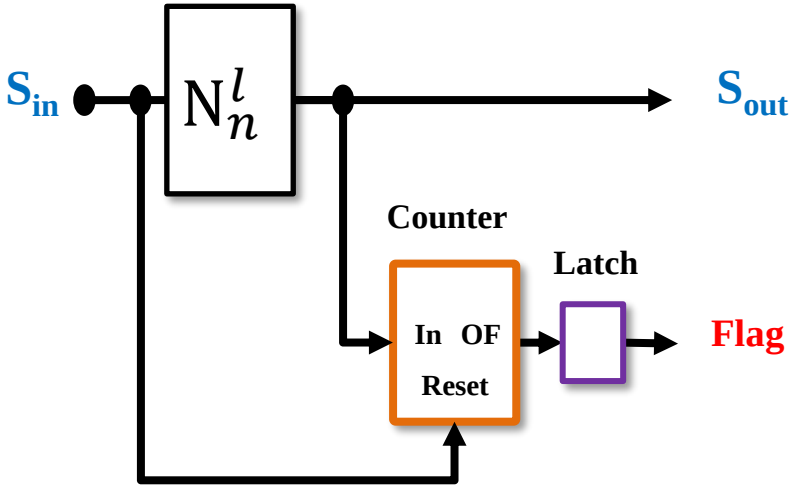
N. Srivastava, *JMLR*'14
T. Spyrou, *DATE'*21

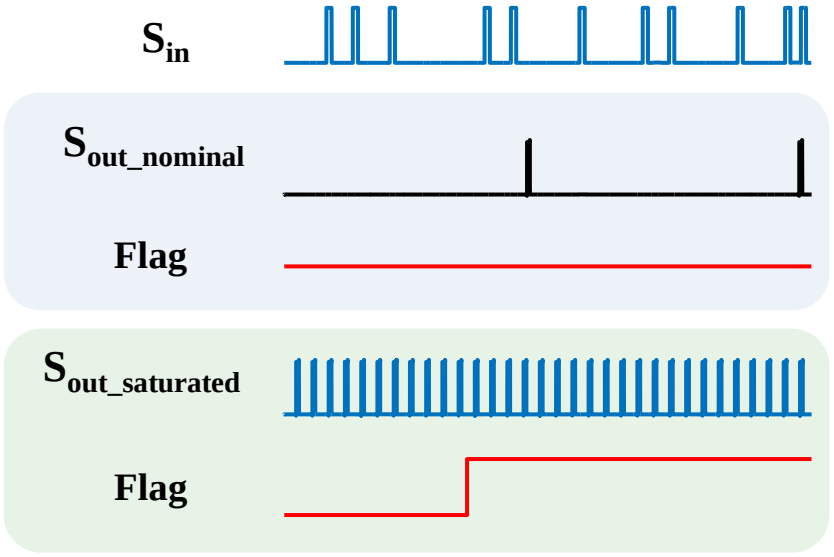


- Training with dropout: temporarily removing neurons during training along with their connections
- Nullifies the effect of dead neuron faults in all hidden layers:
  - Distribution of computational load among the neurons of the network
  - More uniform and sparse spiking activity across the network

On-line testing using in-situ monitors

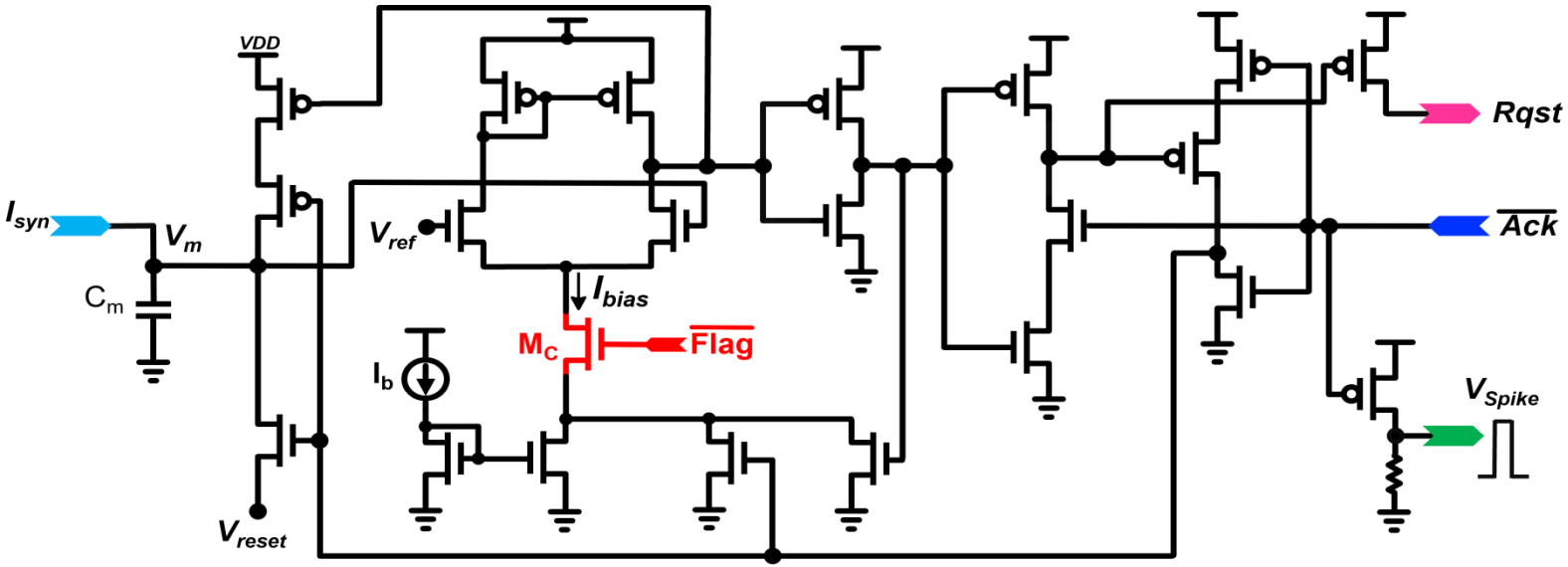
- Count the number of spikes a neuron produces between two successive inputs
- A saturated neuron will produce spikes with higher frequency than usual: counter overflows before an incoming spike resets it again
- Exploits temporal dependency between the input and output of a spiking neuron





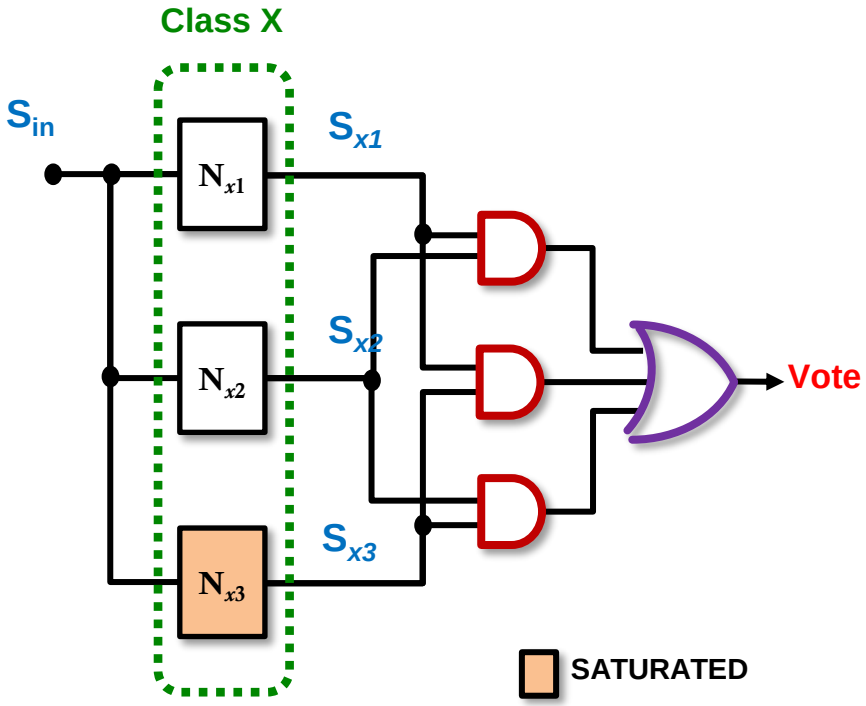
Error recovery using fault masking

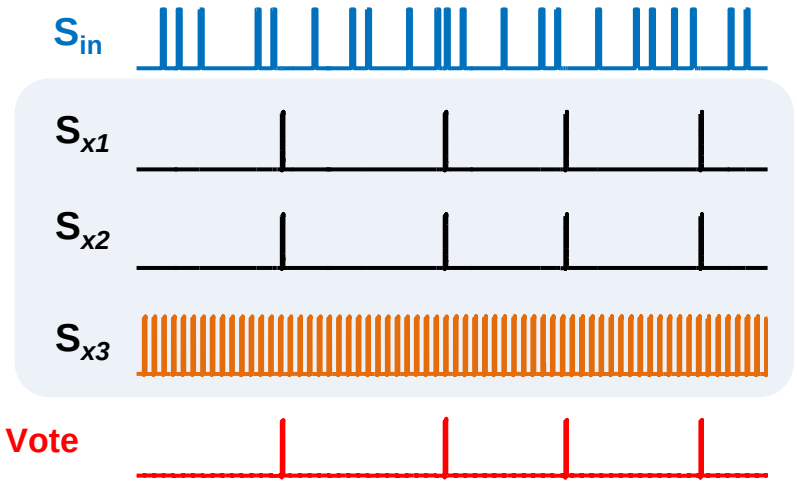
- Saturated neurons are more critical than dead neurons & dead neurons can be nullified using dropout
- "Fault Hopping" concept: saturated neuron fault is translated to a dead neuron fault
- One single transistor is added to the neuron to switch-it off when a saturation "Flag" signal is raised
- Dead neurons do not consume energy



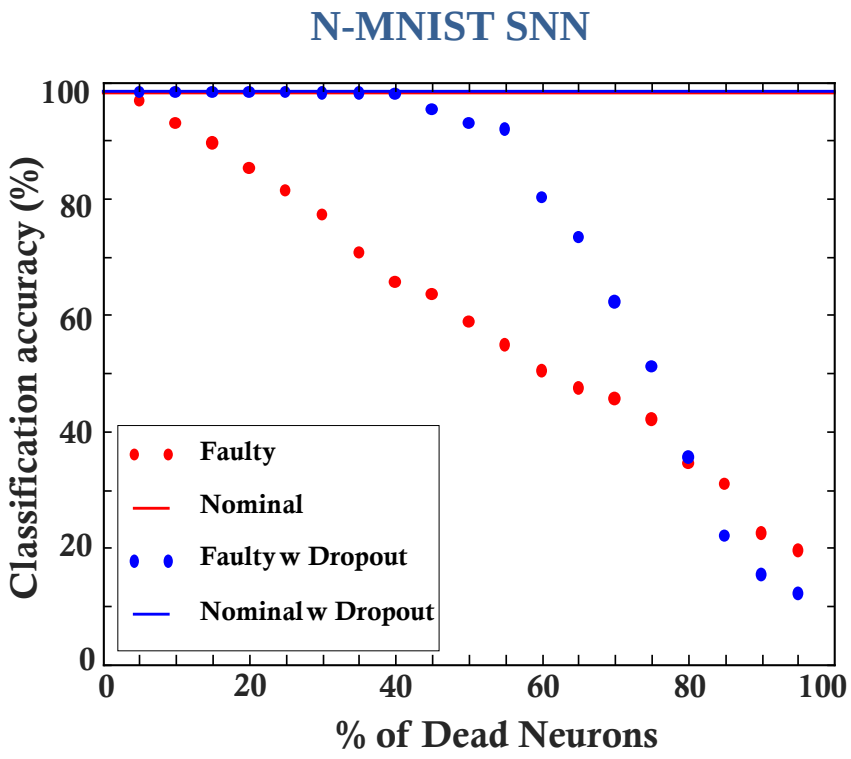
Redundancy-based fault tolerance

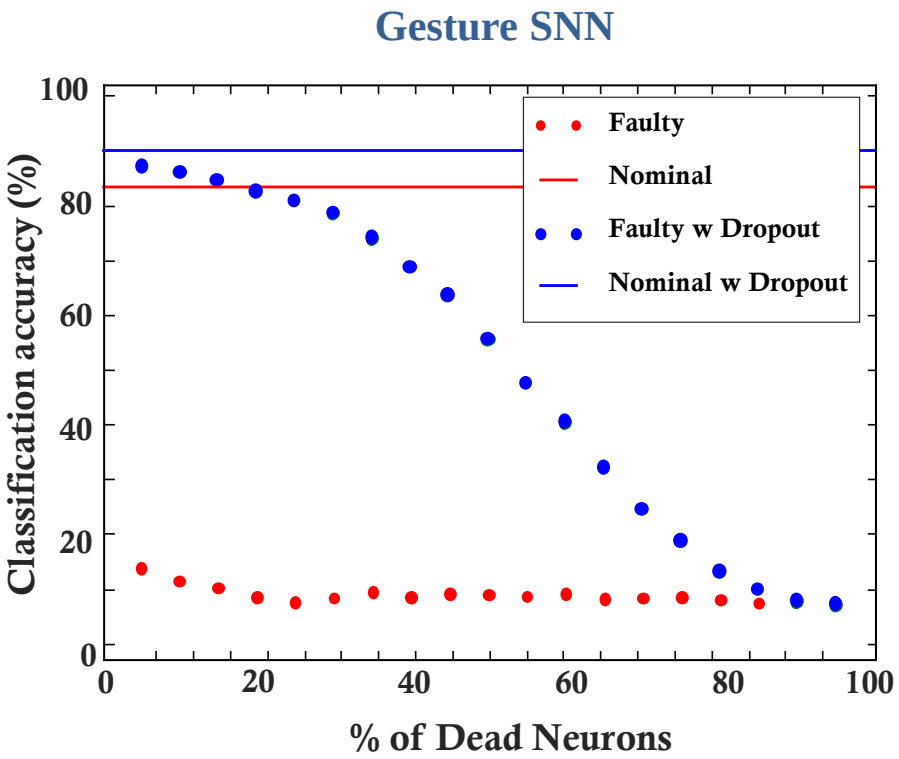
- Triple Modular Redundancy:
  - 3 identical neurons vote for the decision of each class
  - majority decides
- Output layer is usually smaller in size than whole network (0.57% for the N-MNIST SNN and 0.04% for the IBM's Gesture SNN)
- Area overhead is negligible



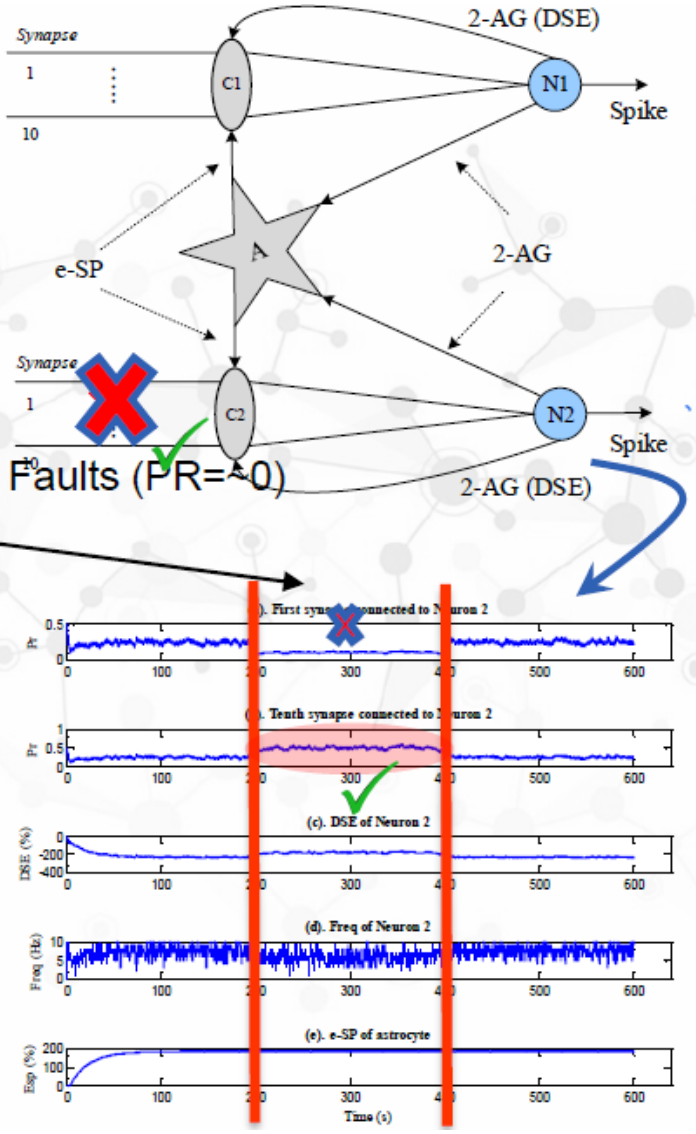


Multiple fault scenario





Astrocyte neural networks



Temporary Faults injected

PR of Synapse #1

PR of Synapse #10

Neuron #2 under **80% fault rate** with temporary faults.
(*80% - severely damaged*)

J. Liu, *TNNLS*'18
S. Karim, *DATE*'20
J. Harkin & M. Trefzer, Tutorial *DATE*'23

Conclusions

- SNNs for neuromorphic edge computing
- SNN hardware accelerators are emerging
- Frameworks for accelerator design and fault injection
- Testability and fault tolerance concepts still at an early stage
- Acknowledgments:
  - Collaboration with the University of Sevillia
  - PhD Students: Sarah Elsayed, Theofilos Spyrou, Spyridon Raptis, Paul Kling
  - Sorbonne Center for Artificial Intelligence (SCAI)
  - ANR RE-TRUSTING
  - Horizon Europe dAIEDGE

Further reading

- H.-G. Stratigopoulos, T. Spyrou, and S. Raptis, “Testing and reliability of spiking neural networks: A review of the state-of-the-art,” Proc. *IEEE Int. Symp. Defect Fault Toler. {VLSI} Nanotechnol. Syst. (DFT)*, Jaun Les Pins, France, Oct. 2023.
- F. Su, C. Liu, and H.-G. Stratigopoulos, “Testability and dependability of AI hardware: Survey, trends, challenges, and perspectives,” *IEEE Des. Test*, vol. 40, no. 2, pp. 8–58, Apr. 2023.