ReMeCo: Reliable memristor-based in-memory neuromorphic computation

Ali BanaGozar*
Eindhoven University of Technology, The Netherlands
a.banagozar@tue.nl

Seyed Hossein Hashemi Shadmehri*
University of Tehran, Iran
seyed.hashemi.ho@ut.ac.ir

Sander Stuijk
Eindhoven University of Technology, The Netherlands
s.stuijk@tue.nl

Mehdi Kamal
USC, CA, USA
mehdi.kamal@usc.edu

Ali Afzali-Kusha
University of Tehran, Iran
afzali@ut.ac.ir

Henk Corporaal
Eindhoven University of Technology
h.corporaal@tue.nl

ASIA SOUTH PACIFIC DAC DESIGN AUTOMATION CONFERENCE

Outline

- Introduction
- DSE (Contribution I)
- Related work
- ReMeCo (Contribution II)
- Results (Contribution III)
- Conclusion

Introduction

Motivation

Problem statement

Boom of ANN*

- Ubiquitous (e.g., NLP, CV, etc.)
- More complex

*ANN: Artificial Neural Network

Boom of ANN

- Ubiquitous (e.g., NLP, CV, etc.)
- More complex

Inefficient execution on conventional PEs*

- Memory Wall

*PE: Processing Element

Boom of ANN

- Ubiquitous (e.g., NLP, CV, etc.)
- More complex

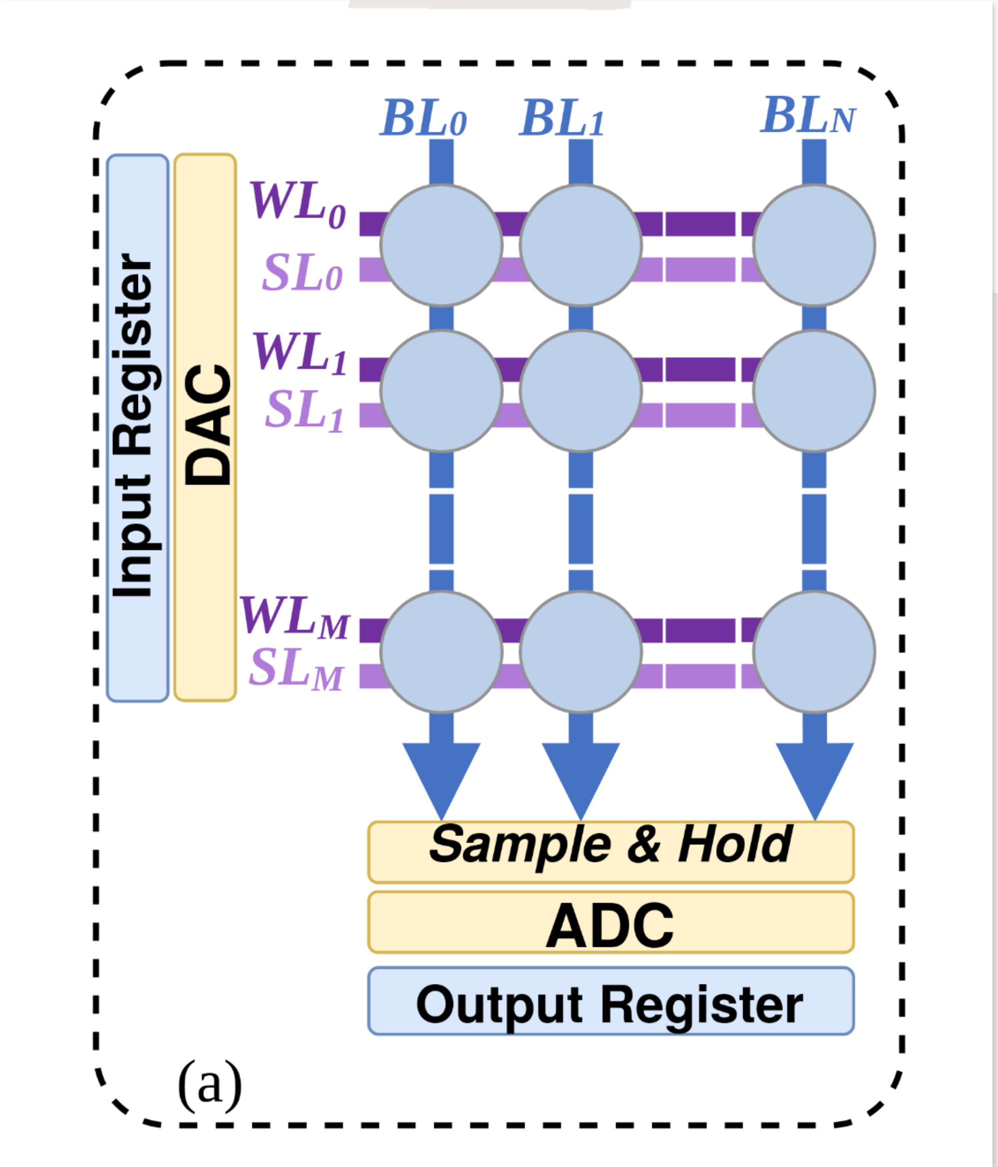
Inefficient execution on conventional PEs

- Memory Wall

Computation In Memory

- SRAM/DRAM modification
- 3D stacked memories
- Memristors

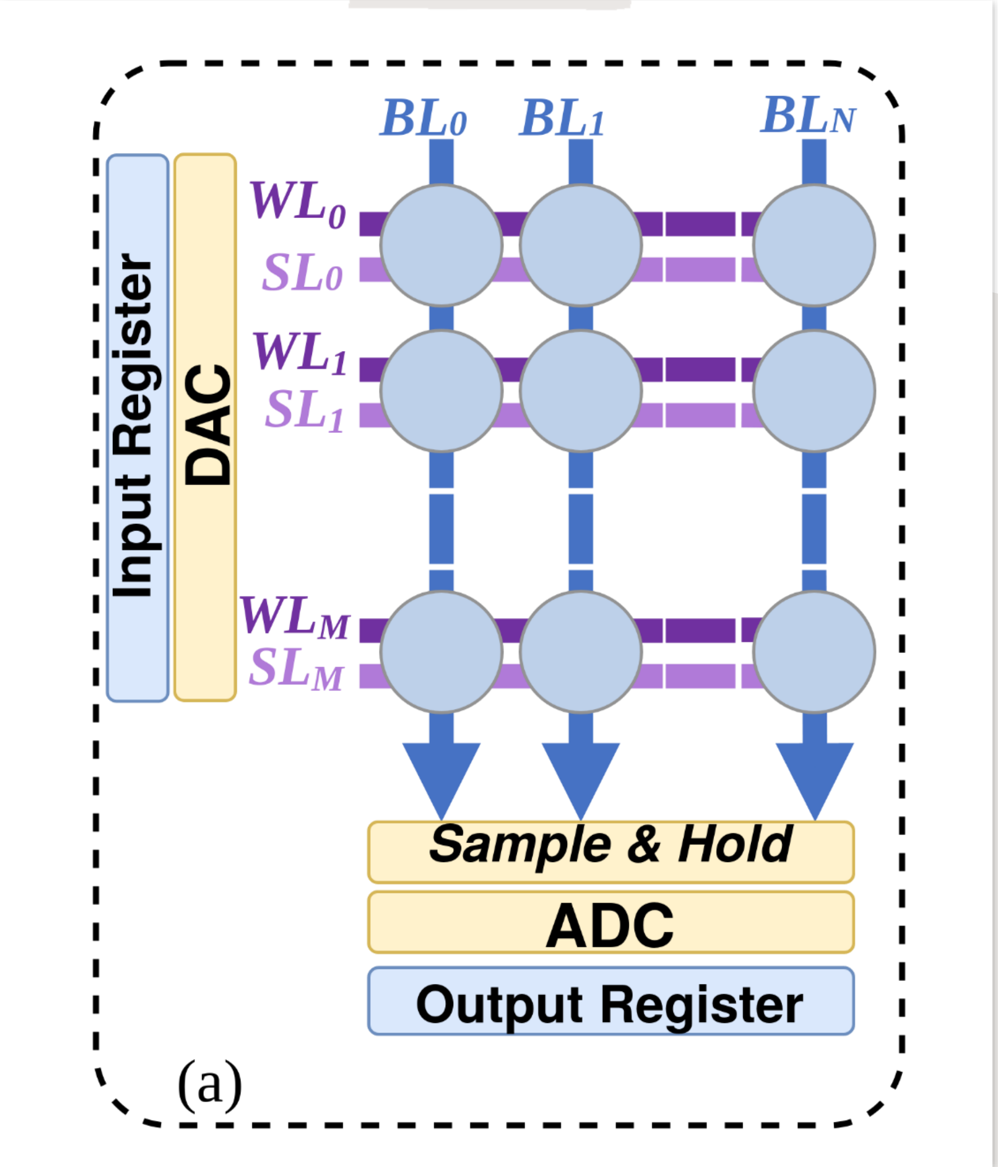
Memristor Crossbars



BL: Bit Line
WL: Word Line
SL: Source Line

Vector-Matrix Multiplication

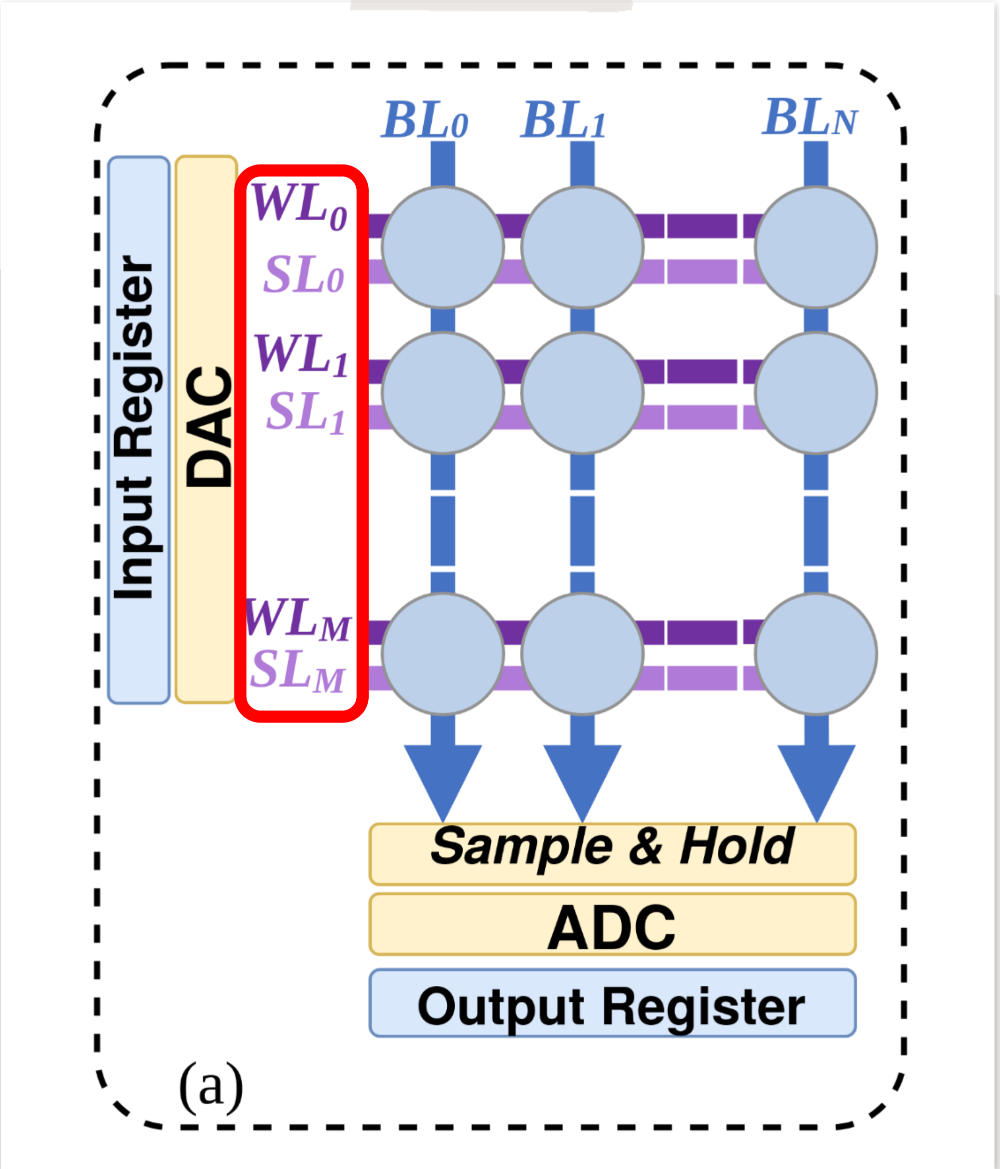
$$Out = In \times Weight$$



Vector-Matrix Multiplication

$$Out = In \times Weight$$

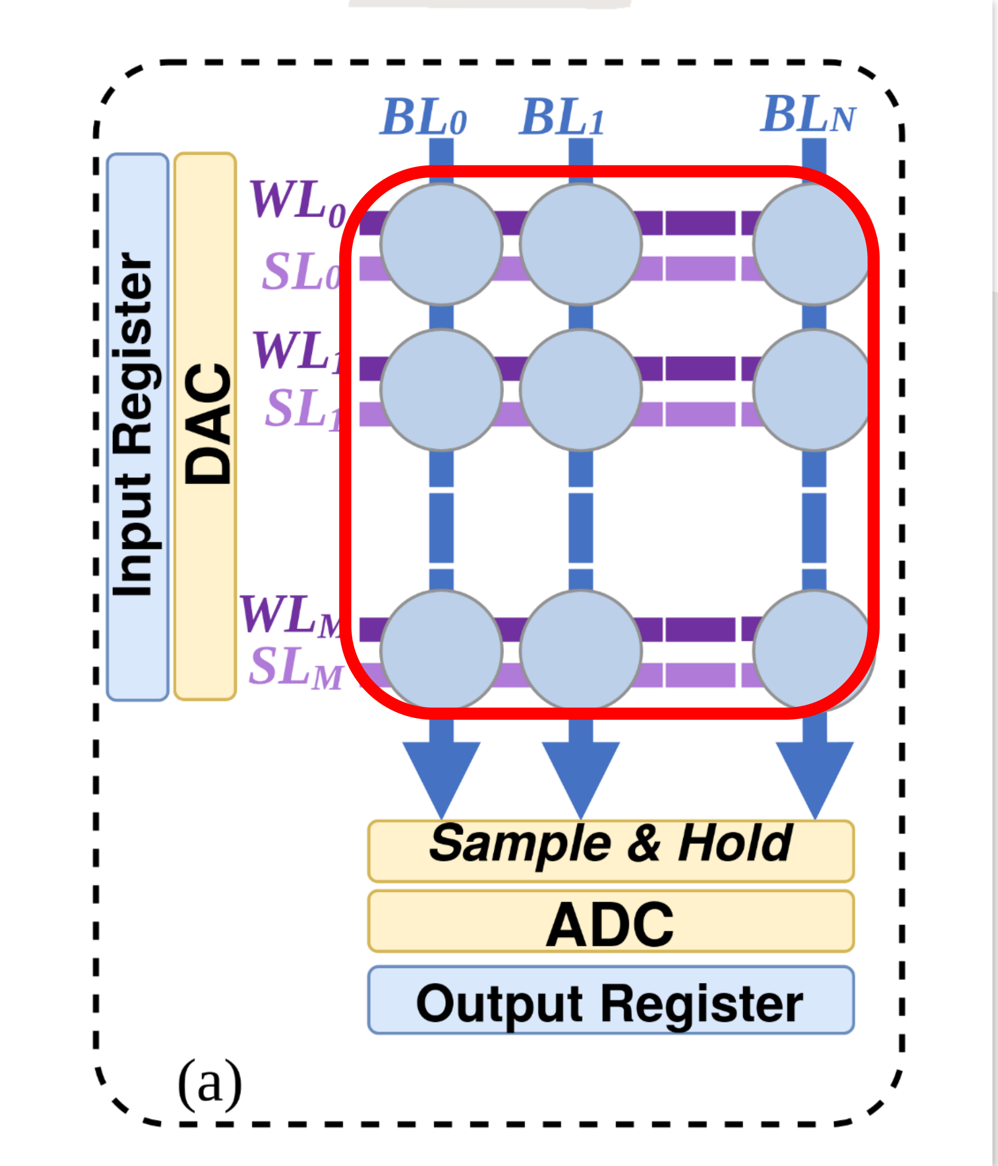
$In \to V$ (vector of voltages)



Vector-Matrix Multiplication

$$\boldsymbol{Out = In \times Weight}$$

$\boldsymbol{In \rightarrow V}$ (vector of voltages)
$\boldsymbol{Weight \rightarrow G}$ (matrix of conductances)

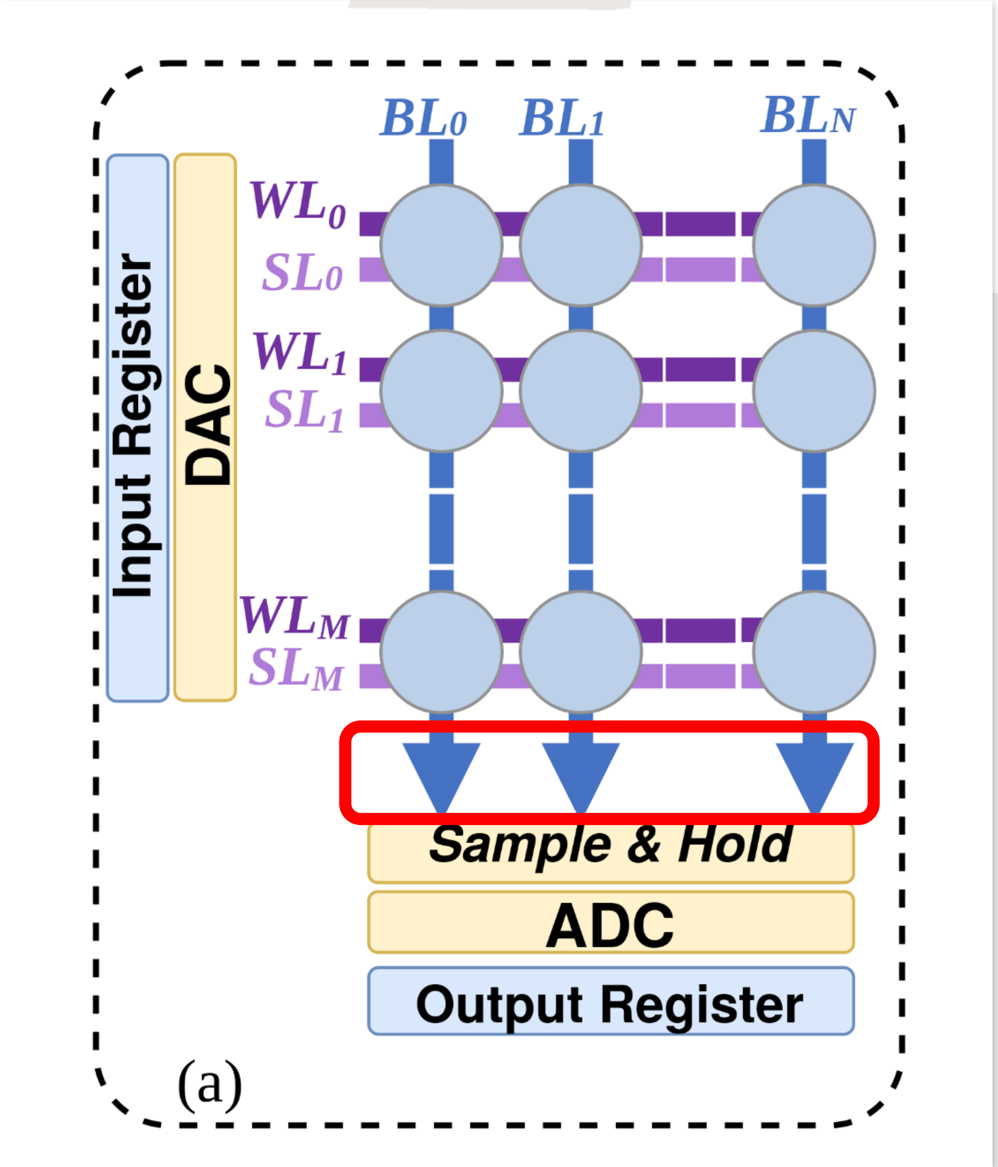


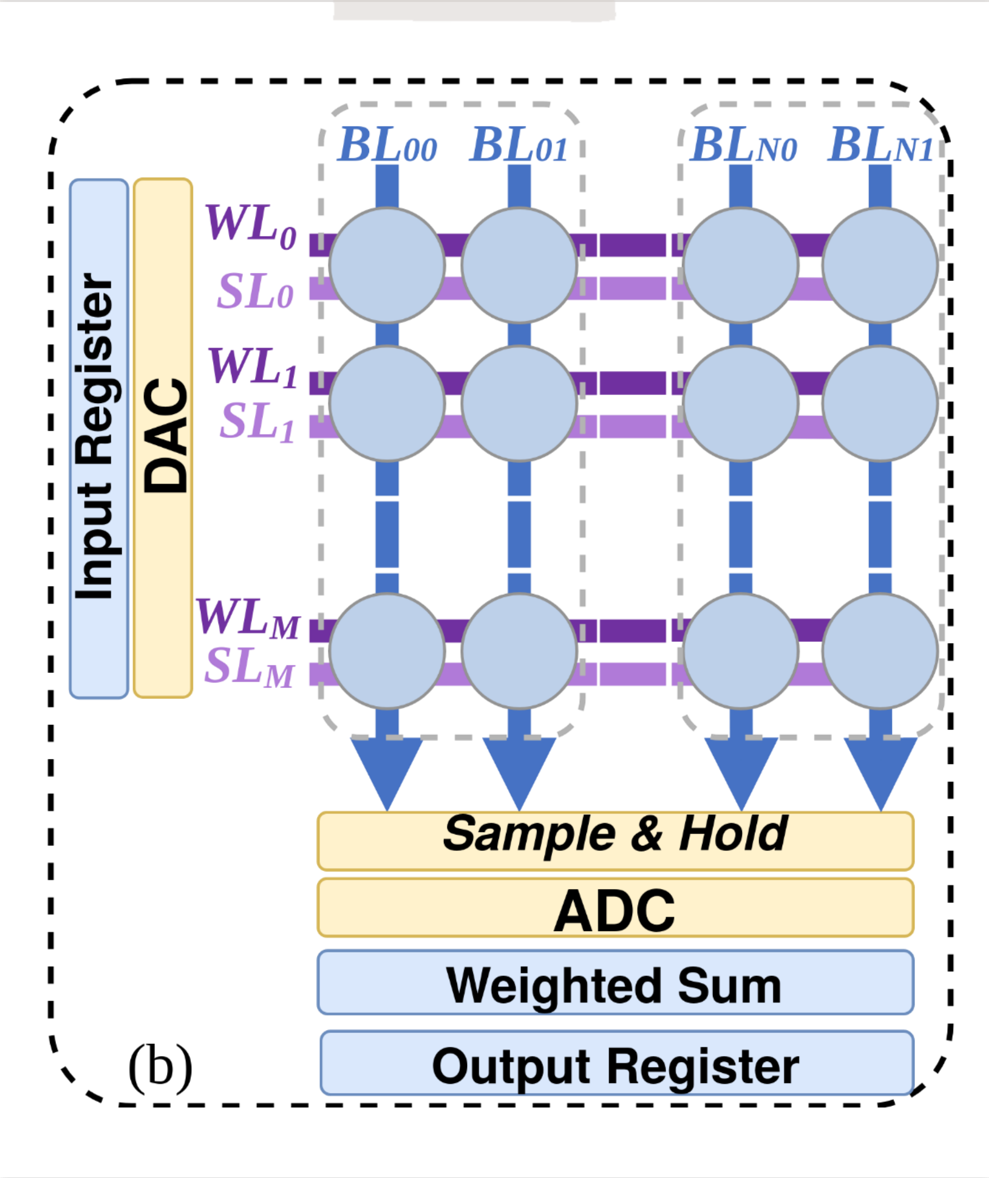
Vector-Matrix Multiplication

$$Out = In \times Weight$$

$In \rightarrow V$ (vector of voltages)
$Weight \rightarrow G$ (matrix of conductances)
$Out \rightarrow I$ (vector of currents)

$$I = V \times G$$





Bit-Slicing

ReMeCo: Reliable memristor-based in-memory neuromorphic computation, BanaGozar et. al.

Non-idealities

Stuck-at-fault

- **Freezes in a state**
- **Inaccurate weight mapping**

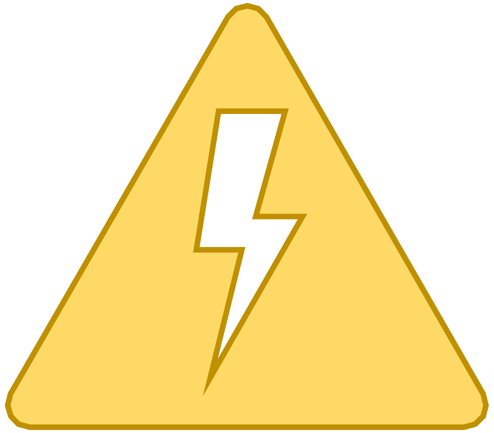
Non-idealities

Stuck-at-fault

- Freezes in a state
- Inaccurate weight mapping

IR-drop

- Wire resistance
- Output current deviation



Non-idealities

Stuck-at-fault

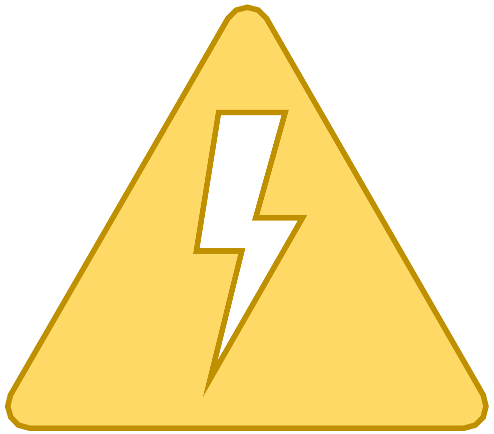
- **Freezes in a state**
- **Inaccurate weight mapping**

IR-drop

- **Wire resistance**
- **Output current deviation**

D2D variation

- **Uncertain weight mapping**



ReMeCo addresses these non-idealities using ***smart hardware redundancy***

ReMeCo

- ***Key idea:** apply redundancy to where it contributes most*

ReMeCo

- ***Key idea:* *apply redundancy to where it contributes most***
- ***Hence, limit hardware redundancy***

ReMeCo

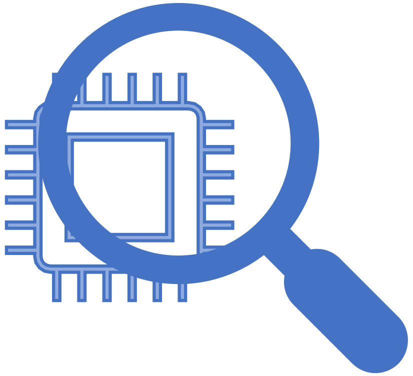
- ***Key idea:*** ***apply redundancy to where it contributes most***
- ***Hence, limit hardware redundancy***
- *By identifying* ***sensitive neurons*** *and* ***sensitive layers*** *in an ANN*

ReMeCo

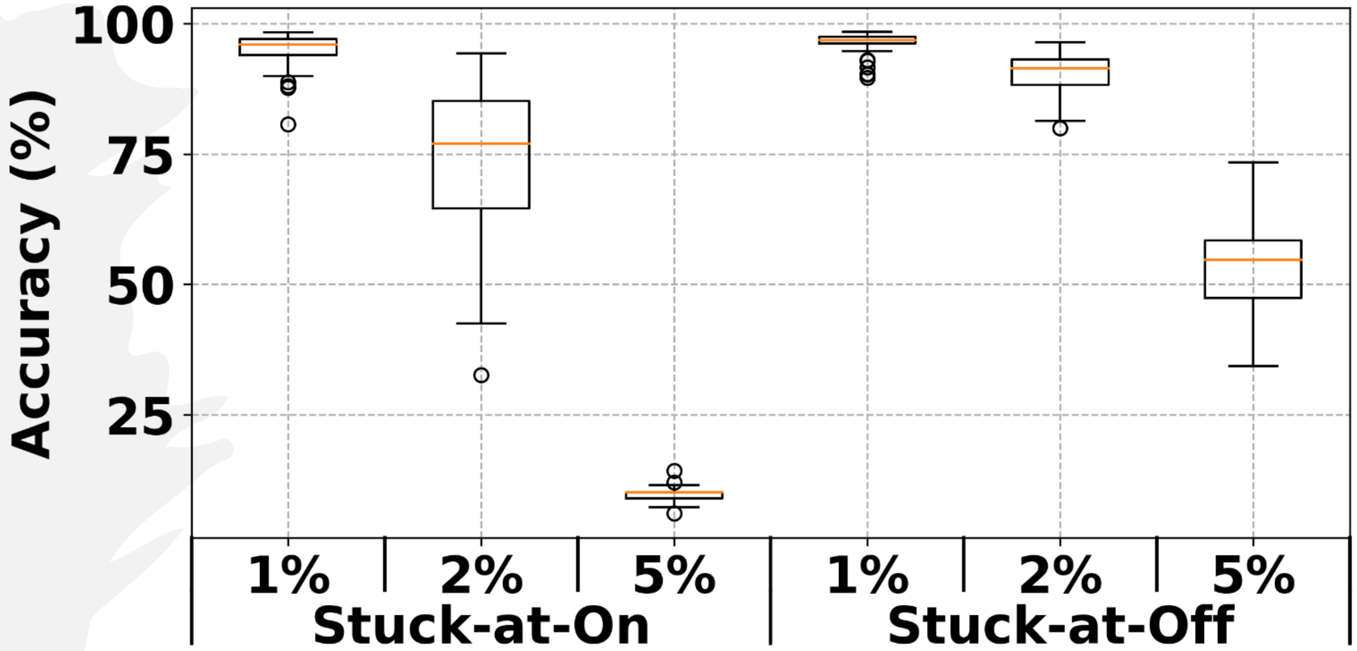
- ***Key idea:*** ***apply redundancy to where it contributes most***
- ***Hence, limit hardware redundancy***
- *By identifying* ***sensitive neurons*** *and* ***sensitive layers*** *in an ANN*
- ***NOTE:*** *Device independent*

Design space exploration of the effects and importance of different non-idealities

Contribution I

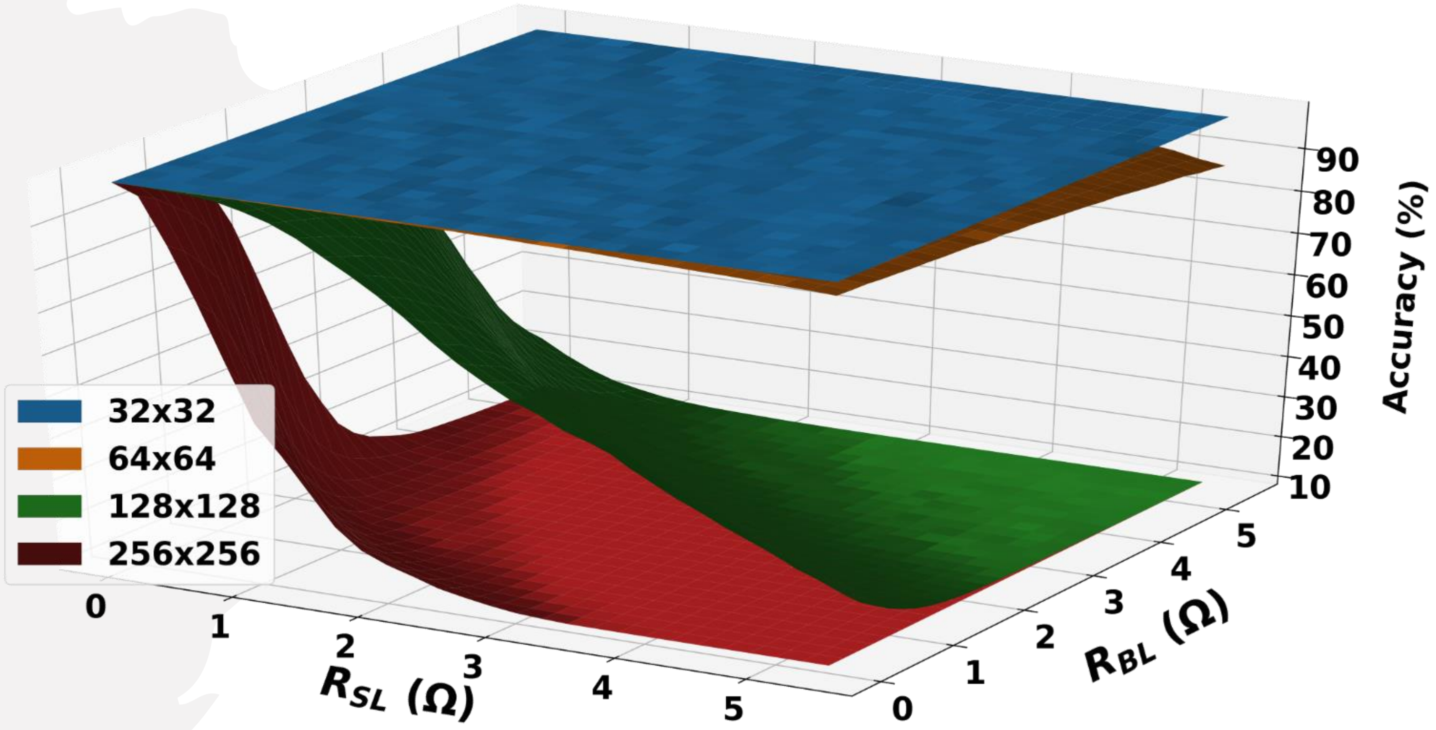


Stuck-at-fault

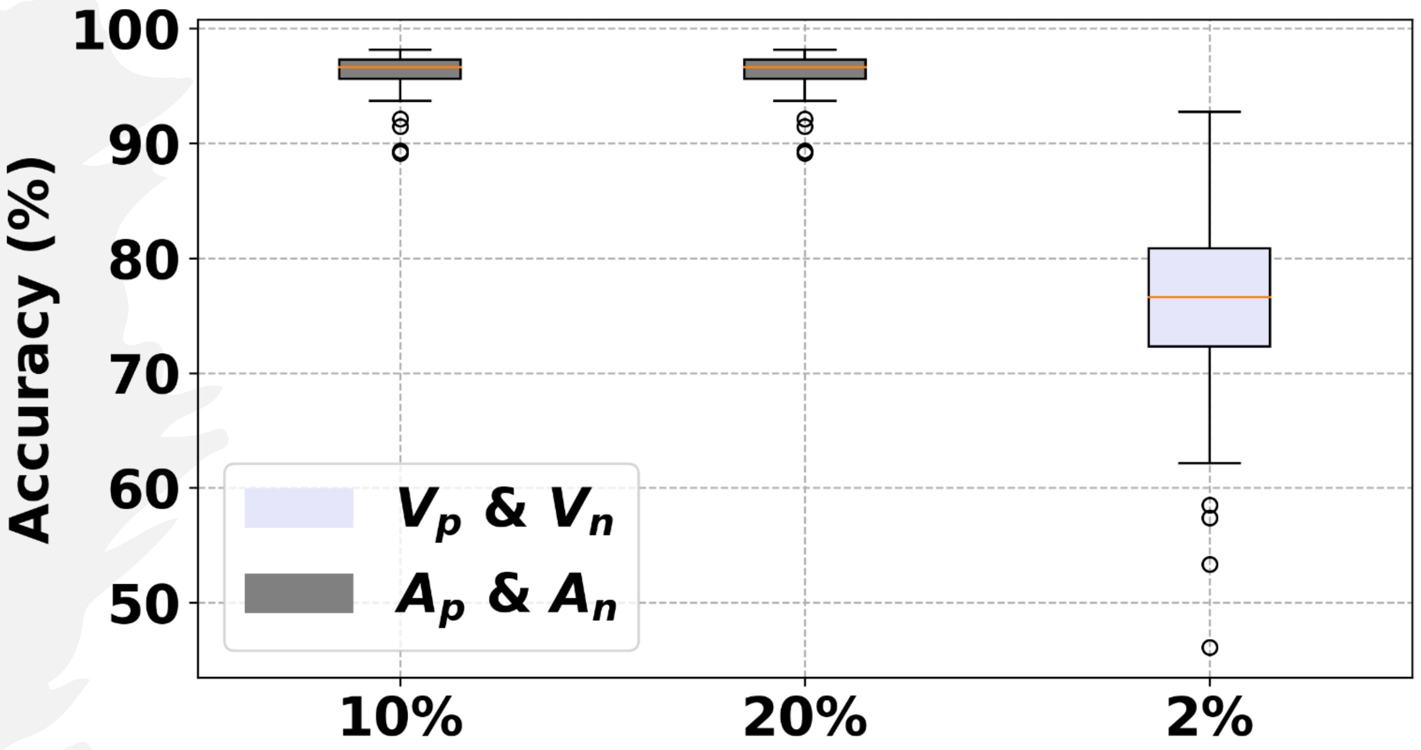


IR-drop

- *Wire resistance*



D2D variation



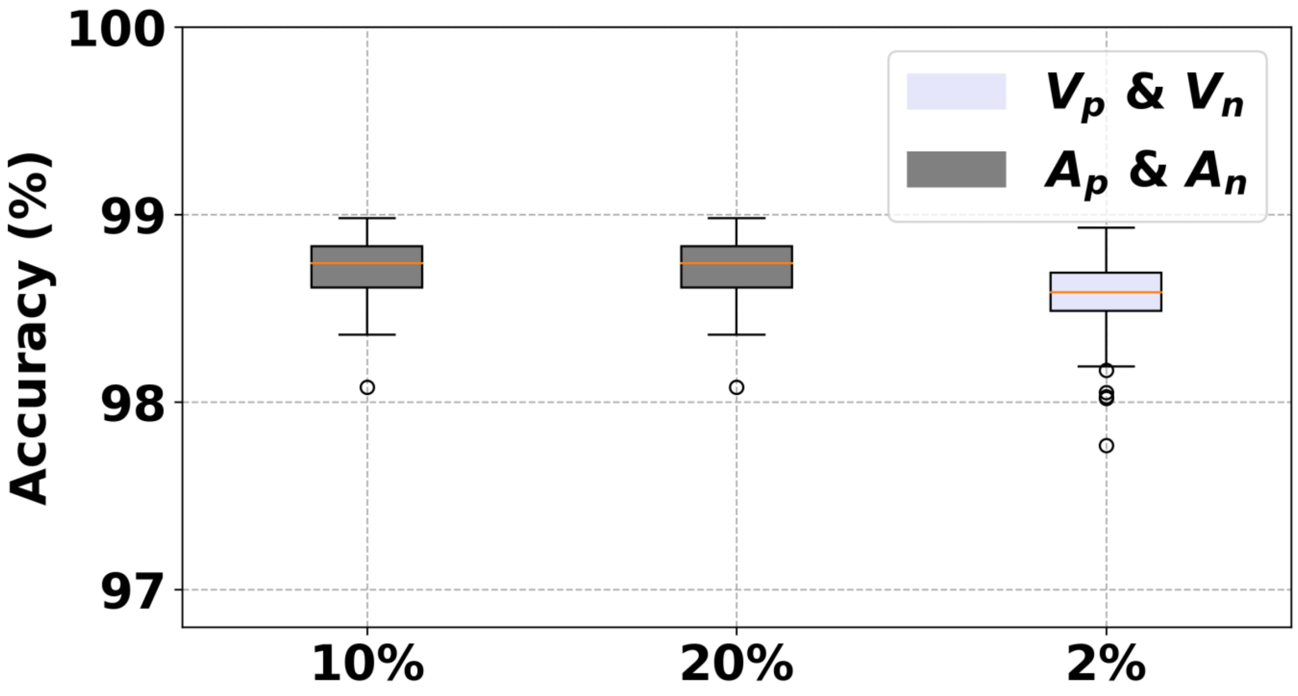
V_p: Set Threshold voltage

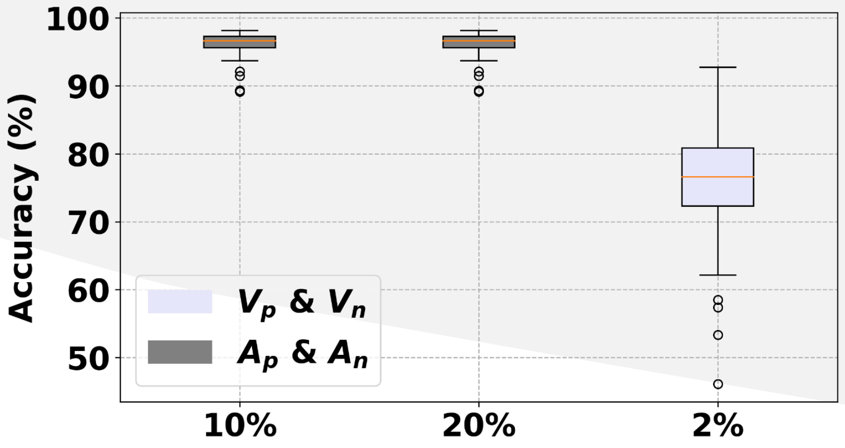
A_p: Set Modulation Factor

ReMeCo: Reliable memristor-based in-memory neuromorphic computation, BanaGozar et. al.

Write-and-verify

~~D2D variation~~





Related Work

Reliable Memristor-based Neuromorphic Design Using Variation- and Defect-Aware Training

Di Gao[1], Grace Li Zhang[2], Xunzhao Yin*[1], Bing Li[2], Ulf Schlichtmann[2], Cheng Zhuo*[1]
[1]College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China
[2]Department of Electronic Design Automation, Technical University of Munich, Munich, Germany
*Corresponding Email: xzyin1@zju.edu.cn, czhuo@zju.edu.cn

First author (Year)	Non-idealities	Model-aware training	Post-fabrication training	Remapping	Hardware redundancy	Sensitivity analysis	Time to market
Gaol (2021) [8]	SAF, PV*	X	X	X	-	X	High
Jin (2020) [10]	SAF	-	X	X	-	X	High
Xu (2021) [19]	SAF, PV	X	X	X	-	X	High
Liu (2017) [13]	SAF	X	X	X	X	X	High
Joksas (2020) [11]	SAF, PV, IR-drop	-	-	-	X	-	*Low*
ReMeCo	SAF, PV, IR-drop	-	-	-	X	X	*Low*

Reliable Memristor-based Neuromorphic Design Using Var

On Improving Fault Tolerance of Memristor Crossbar Based Neural Network Designs by Target Sparsifying

Song Jin
North China Electric Power University
Baoding, P. R. China
jinsong@ncepu.edu.cn

Songwei Pei
Beijing University of Posts and Telecommunications
Beijing, P. R. China
peisongwei@bupt.edu.cn

Yu Wang
North China Electric Power University
Baoding, P. R. China
wangyu@ncepu.edu.cn

First author (Year)	Non-idealities	Model-aware training	Post-fabrication training	Remapping	Hardware redundancy	Sensitivity analysis	Time to market
Gaol (2021) [8]	SAF, PV*	X	X	X	-	X	High
Jin (2020) [10]	SAF	-	X	X	-	X	High
Xu (2021) [19]	SAF, PV	X	X	X	-	X	High
Liu (2017) [13]	SAF	X	X	X	X	X	High
Joksas (2020) [11]	SAF, PV, IR-drop	-	-	-	X	-	*Low*
ReMeCo	SAF, PV, IR-drop	-	-	-	X	X	*Low*

ReMeCo: Reliable memristor-based in-memory neuromorphic computation, BanaGozar et. al.

Reliable Memristor-based Neuromorphic Design

On Improving Fault Tolerance of Memristor

Reliability-Driven Neuromorphic Computing Systems Design

Qi Xu[1], Junpeng Wang[1], Hao Geng[2], Song Chen[1], Xiaoqing Wen[3]
[1]School of Microelectronics, University of Science and Technology of China
[2]Department of Computer Science and Engineering, The Chinese University of Hong Kong
[3]Department of Computer Science and Networks, Kyushu Institute of Technology
{xuqi@,wjp97@mail,songch@}ustc.edu.cn, hgeng@cse.cuhk.edu.hk, wen@cse.kyutech.ac.jp

First author (Year)	Non-idealities	Model-aware training	Post-fabrication training	Remapping	Hardware redundancy	Sensitivity analysis	Time to market
Gaol (2021) [8]	SAF, PV*	X	X	X	-	X	High
Jin (2020) [10]	SAF	-	X	X	-	X	High
Xu (2021) [19]	SAF, PV	X	X	X	-	X	High
Liu (2017) [13]	SAF	X	X	X	X	X	High
Joksas (2020) [11]	SAF, PV, IR-drop	-	-	-	X	-	*Low*
ReMeCo	SAF, PV, IR-drop	-	-	-	X	X	*Low*

ReMeCo: Reliable memristor-based in-memory neuromorphic computation,
BanaGozar et. al.

Reliabili

Rescuing Memristor-based Neuromorphic Design with High Defects

Chenchen Liu, †Miao Hu, †John Paul Strachan and §Hai (Helen) Li
Department of Electrical and Computer Engineering, University of Pittsburgh
†Hewlett Packard Laboratories, §Department of Electrical and Computer Engineering, Duke University
CHL192@pitt.edu, †{miao.hu, john-paul.strachan}@hpe.com, §hai.li@duke.edu

School of Microelectronics, University of Science and Technology of China
[2]Department of Computer Science and Engineering, The Chinese University of Hong Kong
[3]Department of Computer Science and Networks, Kyushu Institute of Technology
{xuqi@,wjp97@mail,songch@}ustc.edu.cn, hgeng@cse.cuhk.edu.hk, wen@cse.kyutech.ac.jp

First author (Year)	Non-idealities	Model-aware training	Post-fabrication training	Remapping	Hardware redundancy	Sensitivity analysis	Time to market
Gaol (2021) [8]	SAF, PV*	X	X	X	-	X	High
Jin (2020) [10]	SAF	-	X	X	-	X	High
Xu (2021) [19]	SAF, PV	X	X	X	-	X	High
Liu (2017) [13]	SAF	X	X	X	X	X	High
Joksas (2020) [11]	SAF, PV, IR-drop	-	-	-	X	-	*Low*
ReMeCo	SAF, PV, IR-drop	-	-	-	X	X	*Low*

nature COMMUNICATIONS

Committee machines—a universal method to deal with non-idealities in memristor-based neural networks

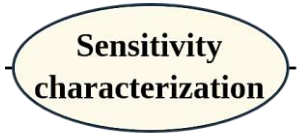
D. Joksas[1✉], P. Freitas[2], Z. Chai[2], W. H. Ng[1], M. Buckwell[1], C. Li[3], W. D. Zhang[2], Q. Xia[3], A. J. Kenyon[1] & A. Mehonic[1✉]

First author (Year)	Non-idealities	Model-aware training	Post-fabrication training	Remapping	Hardware redundancy	Sensitivity analysis	Time to market
GaoI (2021) [8]	SAF, PV*	X	X	X	-	X	High
Jin (2020) [10]	SAF	-	X	X	-	X	High
Xu (2021) [19]	SAF, PV	X	X	X	-	X	High
Liu (2017) [13]	SAF	X	X	X	X	X	High
Joksas (2020) [11]	SAF, PV, IR-drop	-	-	-	X	-	*Low*
ReMeCo	SAF, PV, IR-drop	-	-	-	X	X	*Low*

ReMeCo: a novel redundancy-based framework to improve the reliability of memristor-based ANN PEs by addressing the non-idealities

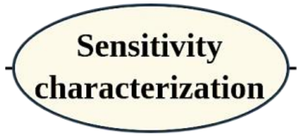
Contribution II

ReMeCo Flow



0. Sensitivity analysis

ReMeCo Flow



0. Sensitivity analysis

- Using the ***back-propagation*** ANN training algorithm
  - Feed forward path
  - Error calculation
  - Backward propagation of error to neurons

ReMeCo Flow

Sensitivity characterization

0. *Sensitivity analysis*
   - Using the ***back-propagation*** ANN training algorithm
     - Feed forward path
     - Error calculation
     - Backward propagation of error

Calculate the average error contribution of individual neurons

$$Sens._{neuron} = AVR_{j=0}^{N}\left(\frac{\partial E}{\partial \omega_{i,j}}\right)$$

ReMeCo Flow

Sensitivity characterization

0. Sensitivity analysis

- Using the ***back-propagation*** ANN training algorithm
  - Feed forward path
  - Error calculation
  - Backward propagation of error

Calculate the average error contribution of individual neurons

$$Sens._{neuron} = AVR_{j=0}^{N}\left(\frac{\partial E}{\partial \omega_{i,j}}\right)$$

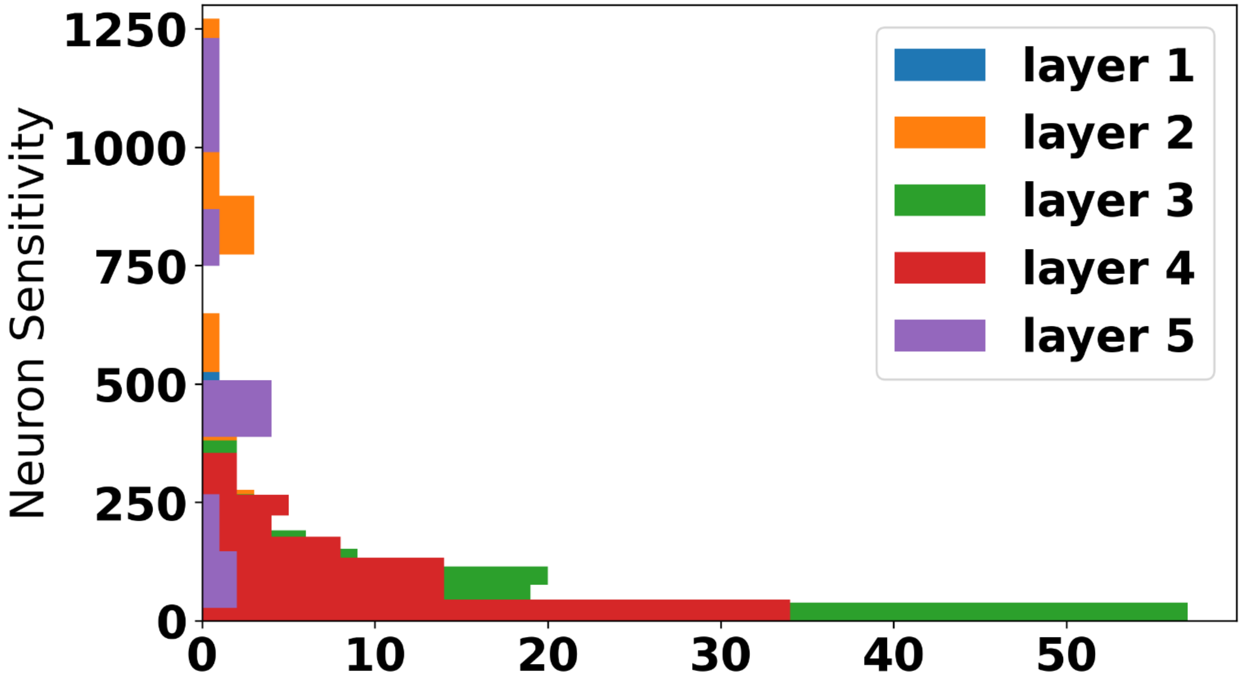
sensitivity of an ANN layer = average of the sensitivity of all its output neurons

ReMeCo Flow

Sensitivity characterization

0. Sensitivity analysis

- *Absolute neuron sensitivity histogram for LeNet.*

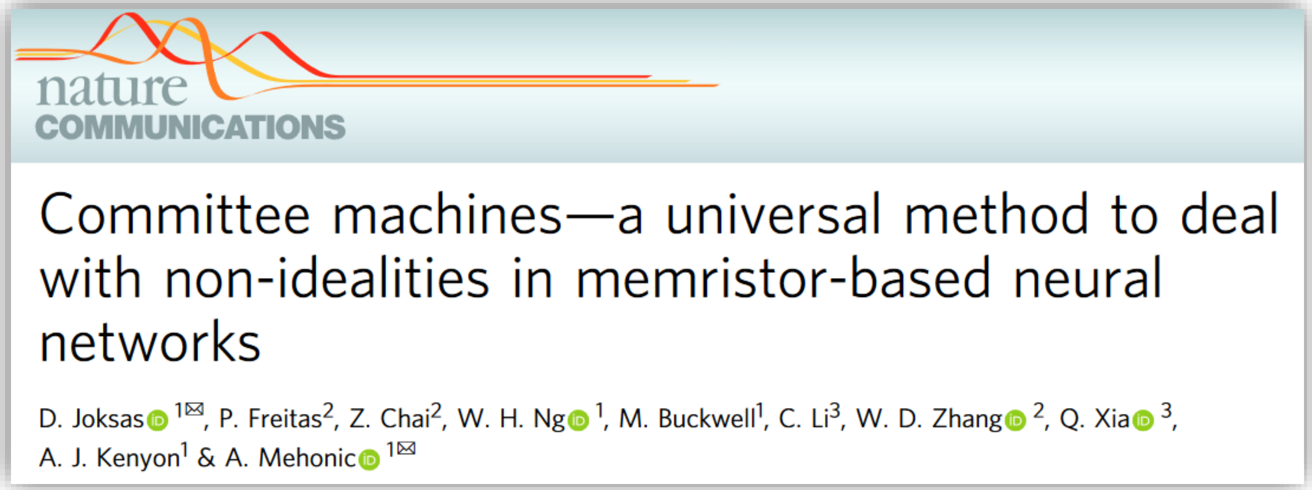


ReMeCo Flow

Committee Machine (100% Redundancy)

Sensitivity characterization

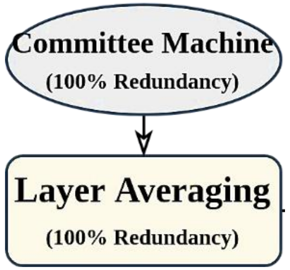
0. *Sensitivity analysis*
1. *Start from the Committee Machine*
   - Implements 1 ANN several times (2 times)
   - Reports the average final output

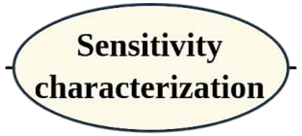

nature COMMUNICATIONS

Committee machines—a universal method to deal with non-idealities in memristor-based neural networks

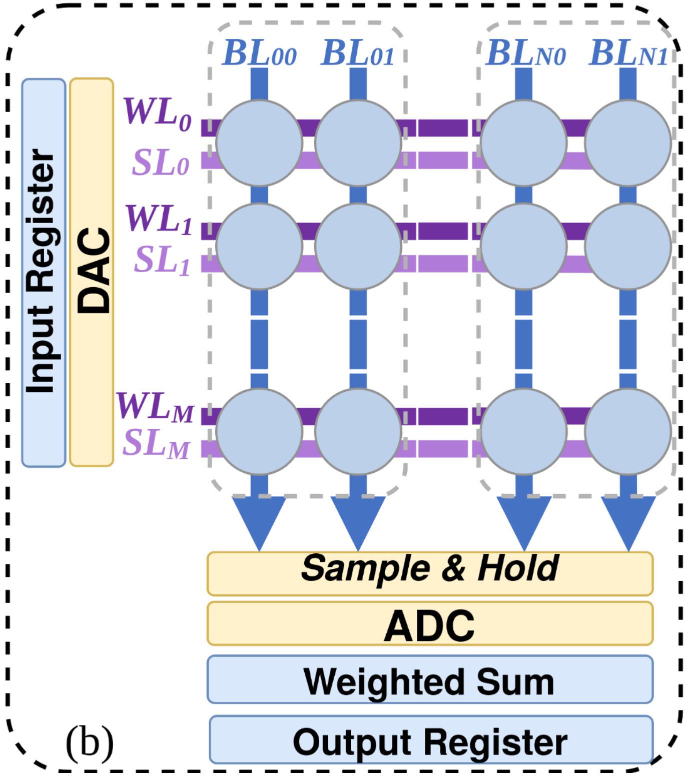
D. Joksas[1], P. Freitas[2], Z. Chai[2], W. H. Ng[1], M. Buckwell[1], C. Li[3], W. D. Zhang[2], Q. Xia[3], A. J. Kenyon[1] & A. Mehonic[1]

ReMeCo Flow

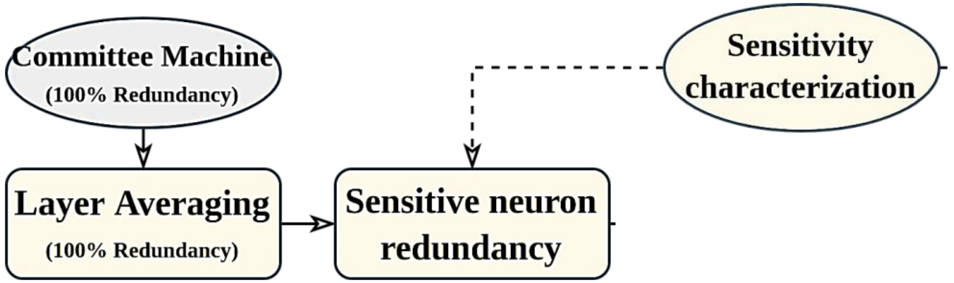




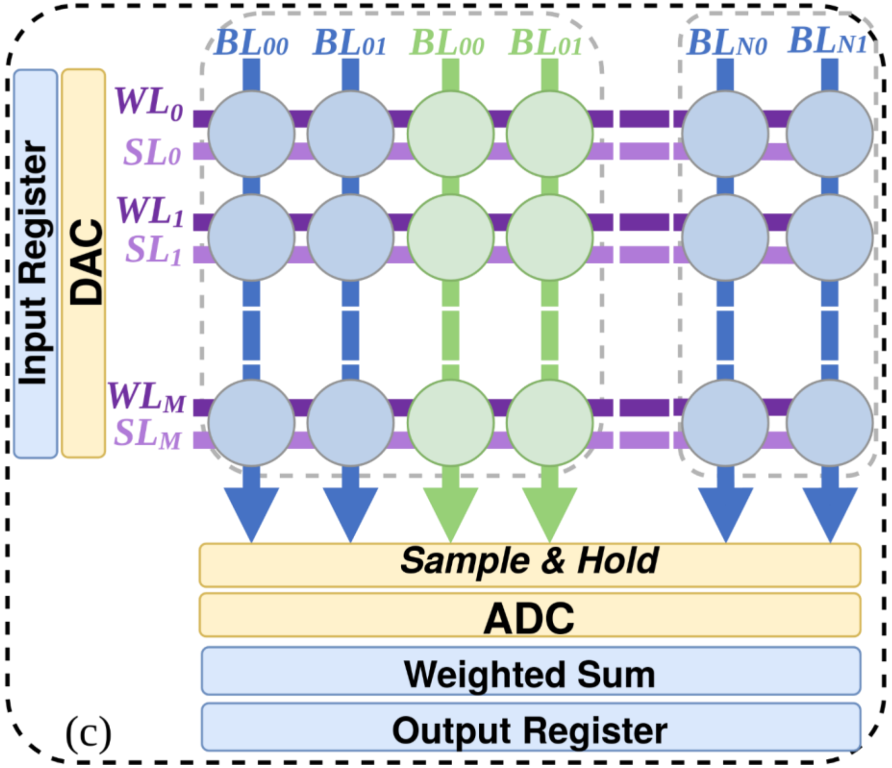
0. *Sensitivity analysis*
1. *Start from the Committee Machine*
2. *Layer averaging*



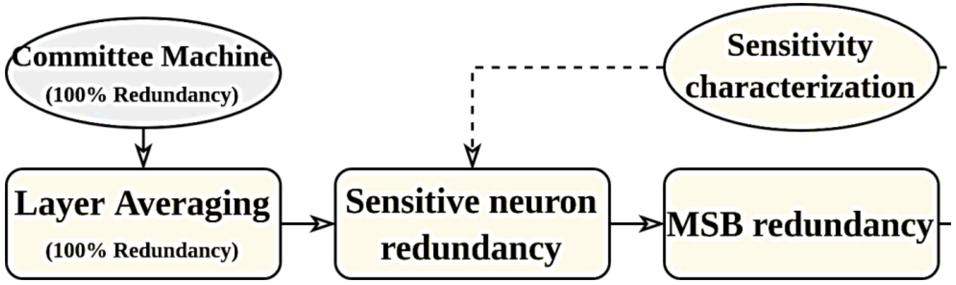
ReMeCo Flow



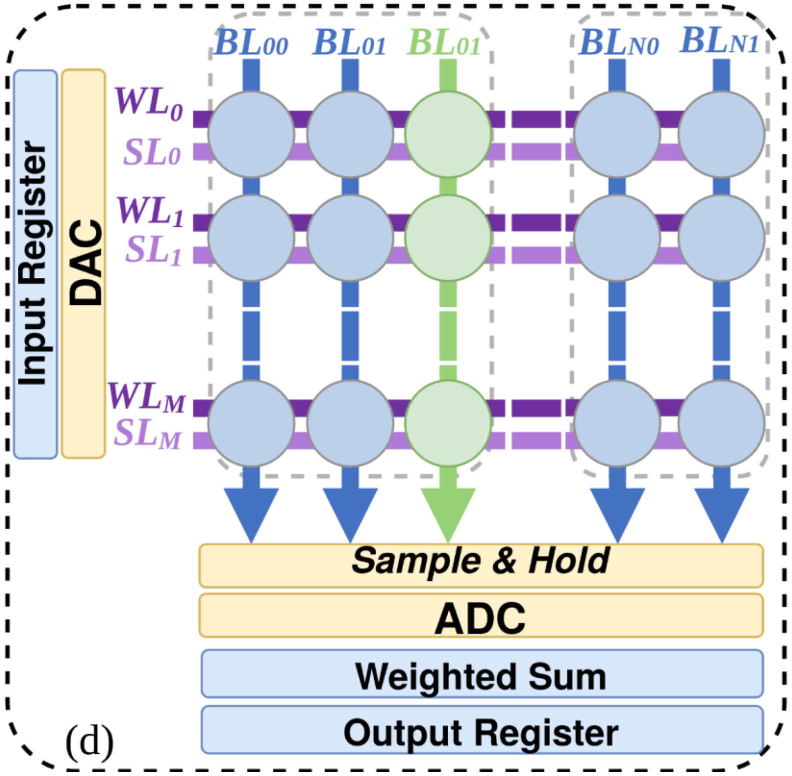
0. Sensitivity analysis
1. Start from the Committee Machine
2. Layer averaging
3. Sensitive neuron redundancy



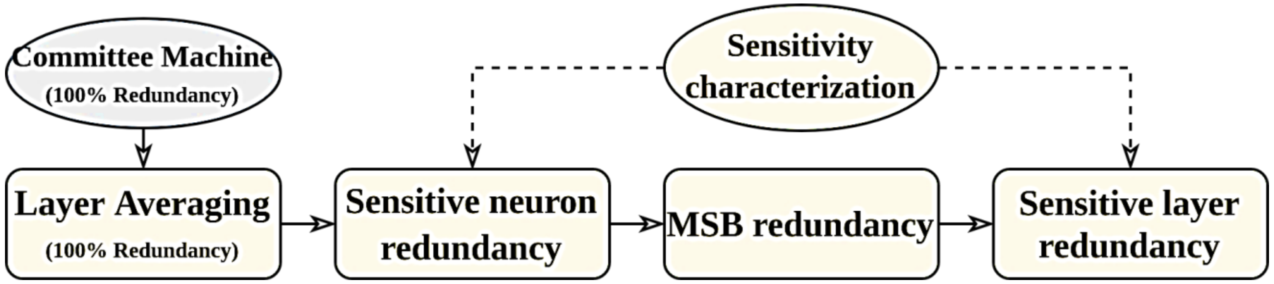
ReMeCo Flow



0. Sensitivity analysis
1. Start from the Committee Machine
2. Layer averaging
3. Sensitive neuron redundancy
4. Most Significant Bit Redundancy



ReMeCo Flow



0. *Sensitivity analysis*
1. *Start from the Committee Machine*
2. *Layer averaging*
3. *Sensitive neuron redundancy*
4. *Most Significant Bit Redundancy*
5. *Sensitive layer redundancy*

Evaluations using LeNet and AlexNet (trained/tested with the MNIST and CIFAR-10 data-sets) show up to 98.5% accuracy recovery!

Contribution III

Benchmarks

- **LeNet** and **AlexNet** used for assessing ReMeCo
- Trained/tested with **MNIST** and **CIFAR-10** datasets
- **8-bit quantization**
- Achieved **software accuracy** 98.8% and 80.1%

ANN	Data-set	Layers		Parameters	FLOPs
		Conv.	Dense		
LeNet	MNIST	2	3	866 K	28 M
AlexNet	CIFAR-10	5	3	62 M	1.5 B

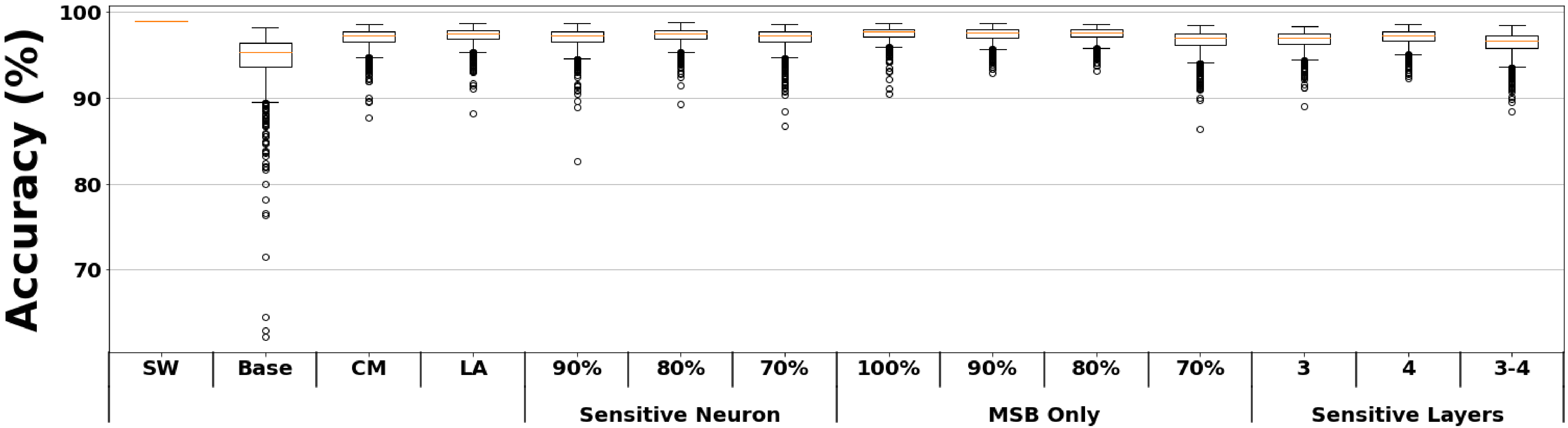
Comparison

- We compare ReMeCo with Committee Machine (CM)
- Compare the ***recovered accuracy*** vs ***induced overhead***
- The lowest overhead for CM is considered (100% redundancy)

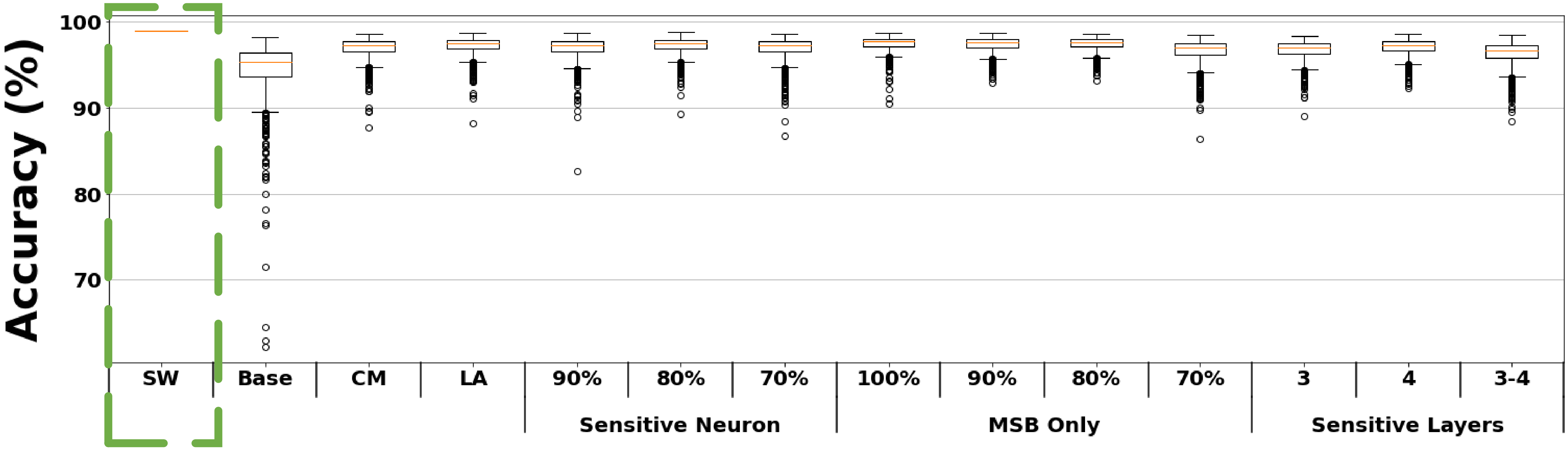
Implementation Notes

- 4-bit memrisotrs are use (2 columns bit-slicing)
- Implement benchmarks on several **different corssbar models**
  - The source-line and bit-line resistance $\in N(\mu = 2, \sigma^2 = 0.5)$
  - In each circuit model, 1% of devices are randomly considered to be stuck
  - 16.2% of all stuck devices being stuck-at-off and 83.8% being stuck-at-on

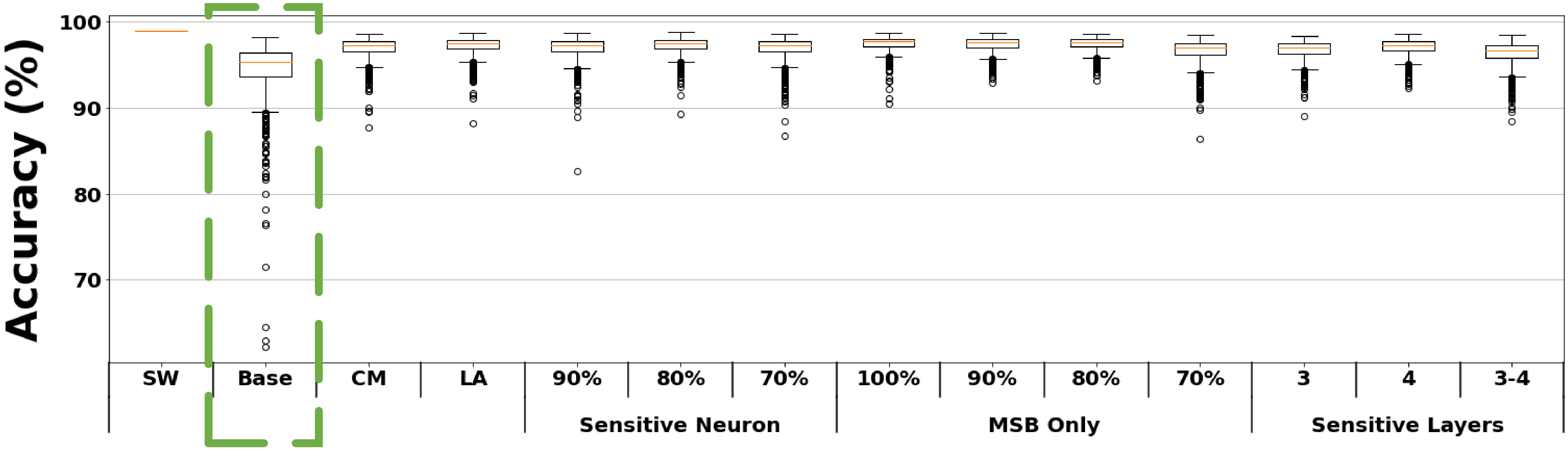
Results
LeNet



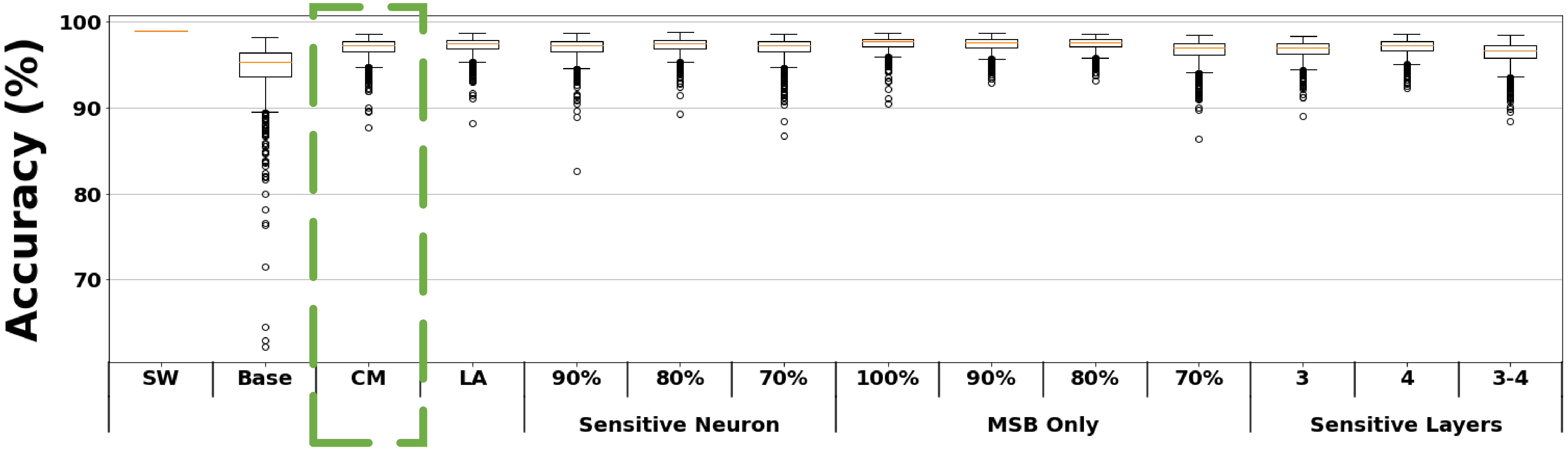
Results
LeNet



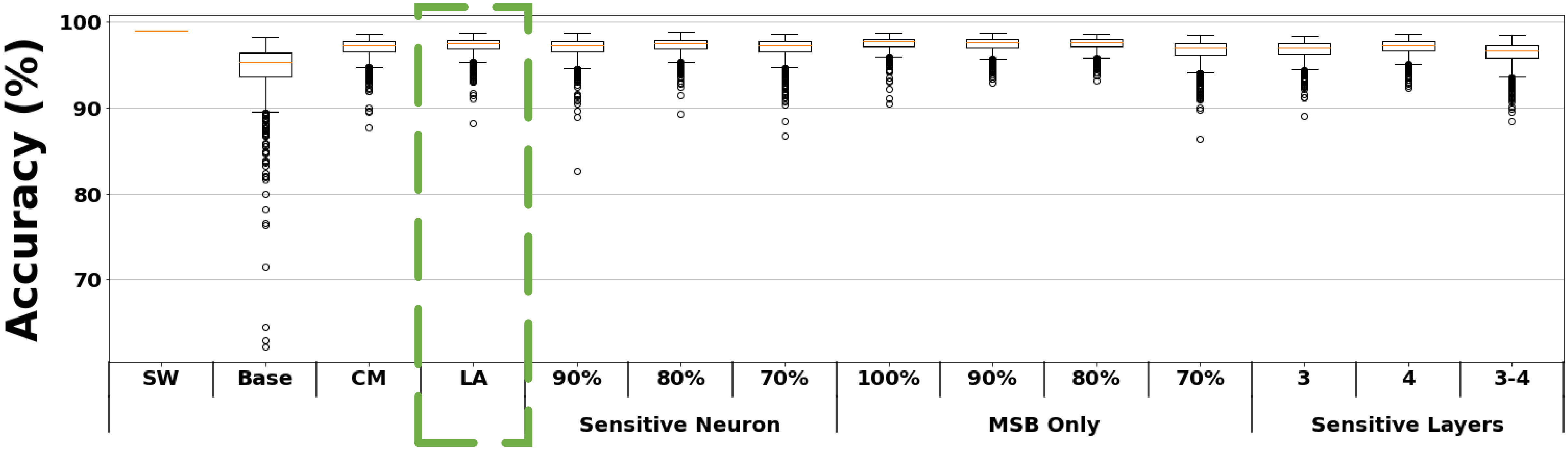
Results
LeNet



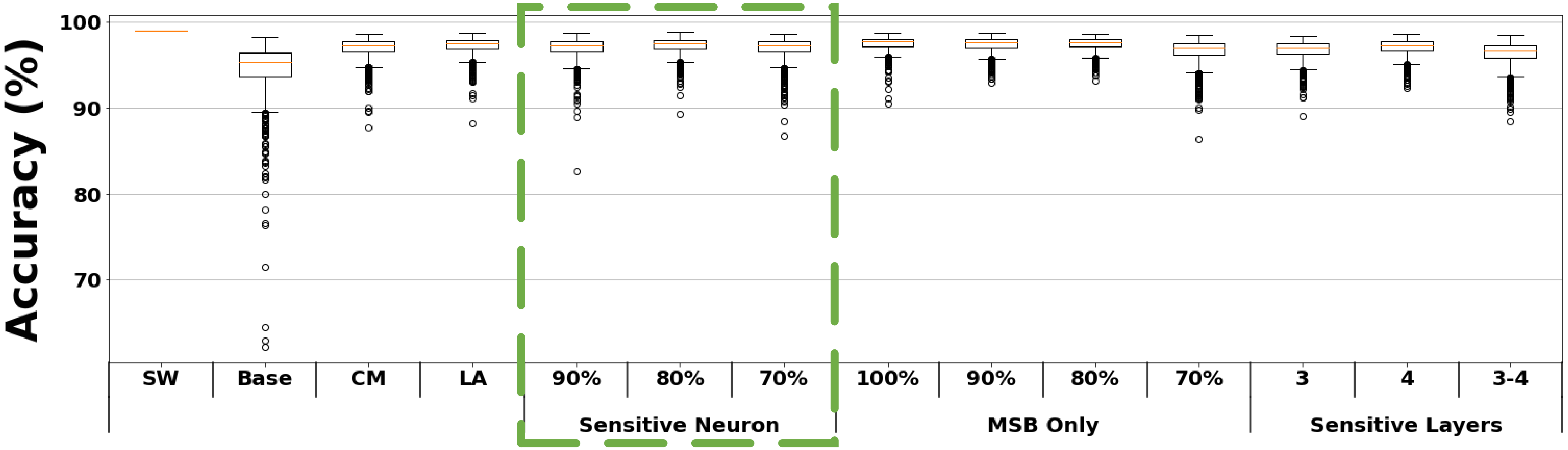
Results
LeNet



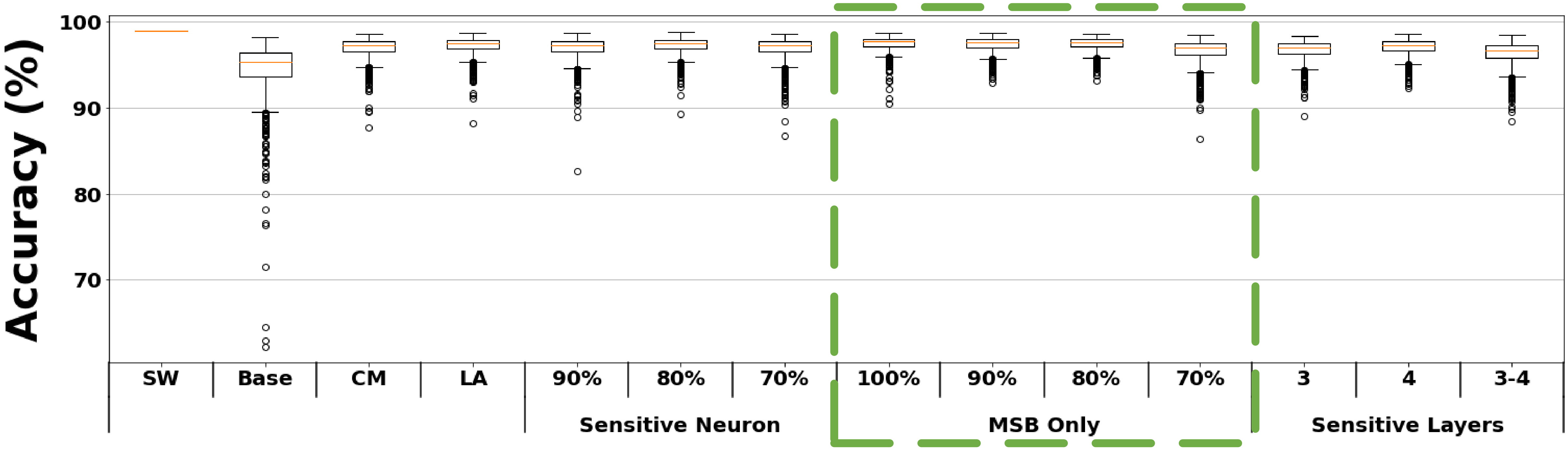
Results
LeNet



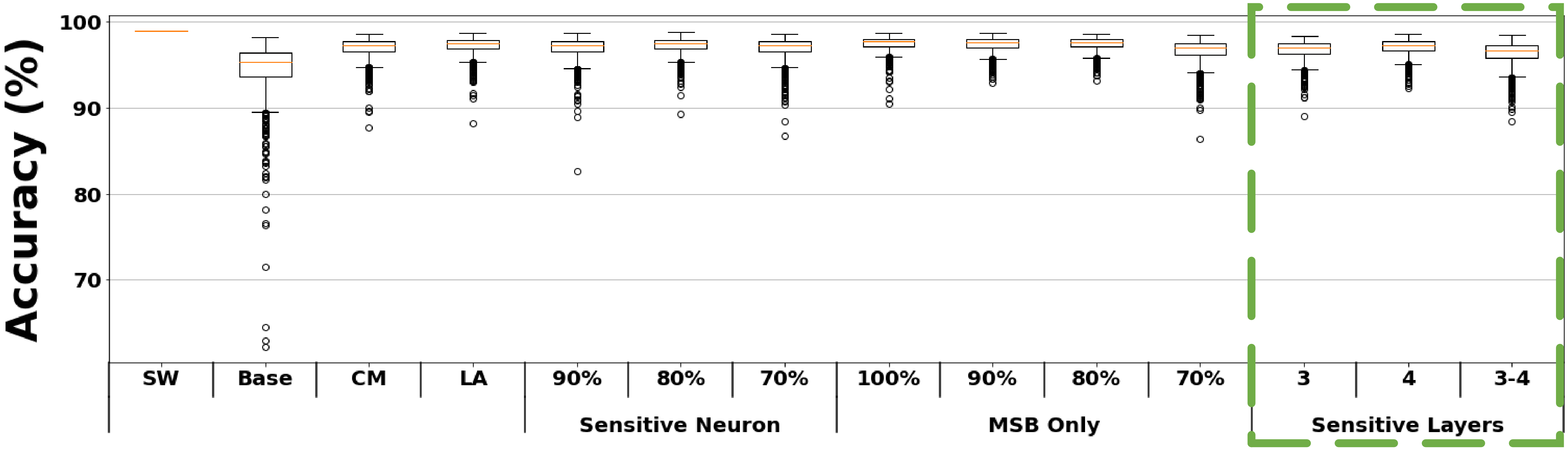
Results
LeNet



Results
LeNet

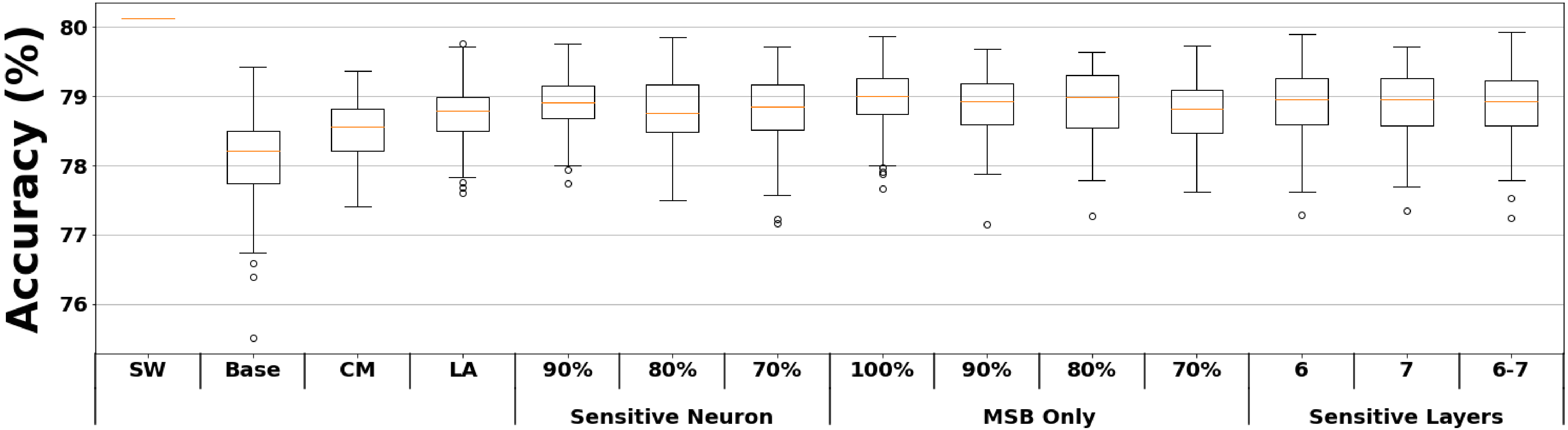


Results
LeNet



ReMeCo: Reliable memristor-based in-memory neuromorphic computation,
BanaGozar et. al.

Results
AlexNet



Results

ANN	Metric	Base (non-ideal)	Committee Machine	Layer Averaging	Sensitive neuron	MSB (80%)	Sensitive layer
LeNet	**Accur.**	94.4	96.9	97.2	97.2	**97.4**	96.4
	Area	7.23e-3	13.63e-3	13.64e-3	12.35e-3	9.79e-3	**7.53e-3**
	Energy	897.1	1791.2	1791.6	1612.7	1254.9	**1211.1**
AlexNet	**Accur.**	78.11	78.5	78.7	78.8	**78.9**	**78.9**
	Area	1.41	2.78	2.78	2.51	1.96	**1.45**
	Energy	1.84	3.68	3.68	3.31	2.58	**2.54**

Conclusion and Future Work

- We performed DSE to quantify the effects of different non-idealities
- We presented ReMeCo framework that addresses non-idealities.
- We achieved up to 98.5% accuracy recovery!
- We reduce HW overhead by +20x compared to CM!
- Future work, we perform tests on even deeper networks.