16th IEEE International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC-2023)

Singapore University of Technology and Design, Singapore

A Novel Yield Improvement Approach for 3D Stacking Neuromorphic Architecture

Ngo-Doanh NGUYEN, Khanh N. Dang

The University of Aizu
Aizuwakamatsu, Fukushima, Japan

Email: {m5262108; khanh}@u-aizu.ac.jp

2023-12-20

Agenda

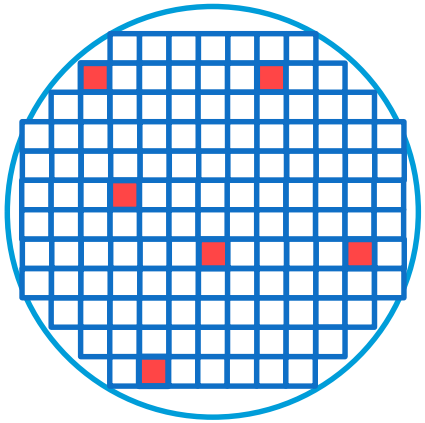
1. Motivations
2. Approach & Methodology
3. Evaluations
4. Conclusions

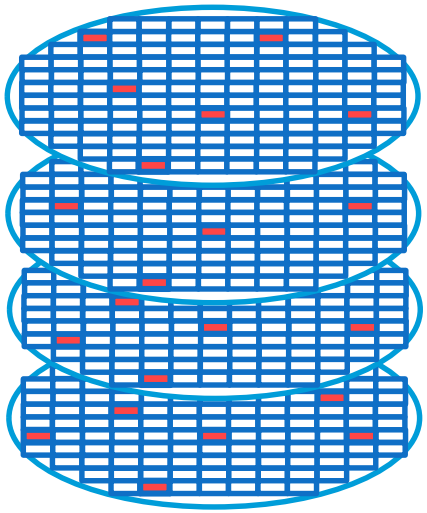
Agenda

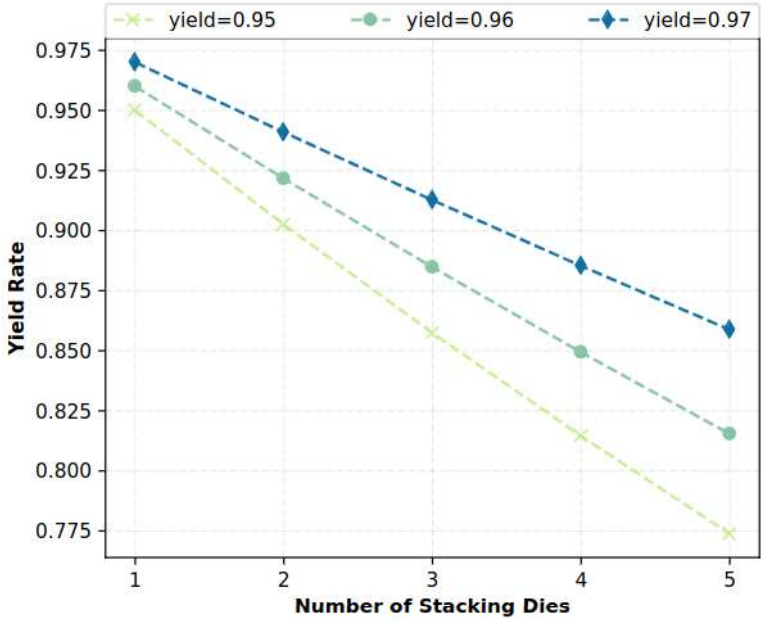
1. Motivations
2. Approach & Methodology
3. Evaluations
4. Conclusions

Motivation (1)

- Low yield rate is one of the most critical issues in 3D-IC design*







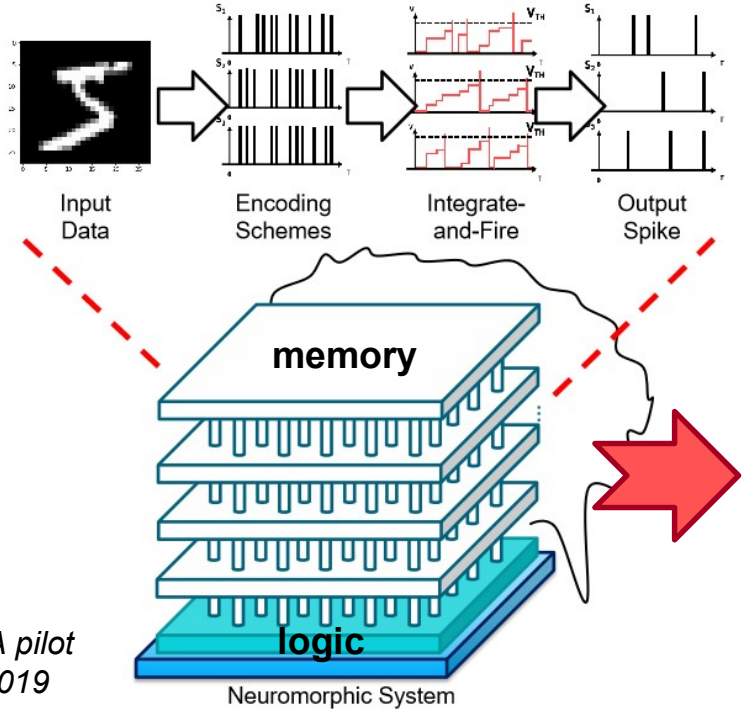
Defects in 1 die **Defects in multiple dies** **Overall yield rate**

=> Increase yield rate for 3D-IC (SNN) designs

*A.C. Hsieh & T. Hwang "TSV Redundancy: Architecture and Design Issues in 3-D IC"

Motivation (2)

- Spiking Neural Networks have the noise resilient characteristic*
  - Effectively against the manufacturing defects
  - With memory-on-logic architecture, defects of upper dies appear on memory blocks



Defects on memory blocks can be treated as noise in Spiking Neural Networks

*T. Wunderlich and et al., "Demonstrating advantages of neuromorphic computation: A pilot study," Frontiers in Neuroscience, vol. 13, 2019

=> **Increase yield rate for SNN systems with acceptable accuracy degradation**

Agenda

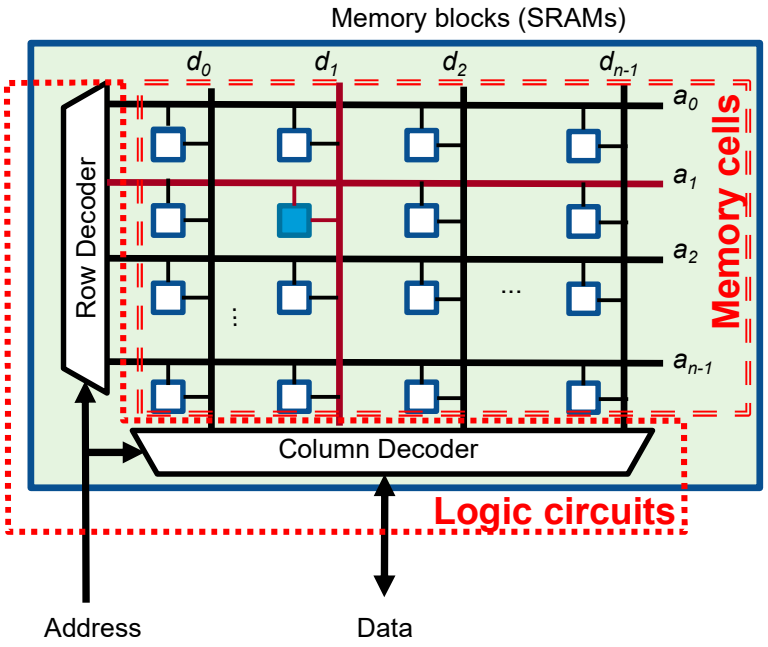
1. Motivations
2. Approach & Methodology
3. Evaluations
4. Conclusions

Approach

- Defects cause the stuck-bit (stay at '0' or '1') or bridging faults event*
- There are two main regions for defects in memory blocks
  - Memory cells
  - Logic circuits
- Solution:
  - Memory cells - Treat them as noises

    => Ultilize the noise-resilient characteristic of SNN
  - Logic circuits - Power-gate them to remove the errors

    => Turn off the memory in which the defects appear



*J.C. M.Li & E. McCluskey “Diagnoisis of resistive-open and stuck-open defects in digital CMOS ICs”

Evaluated Architectures (1)

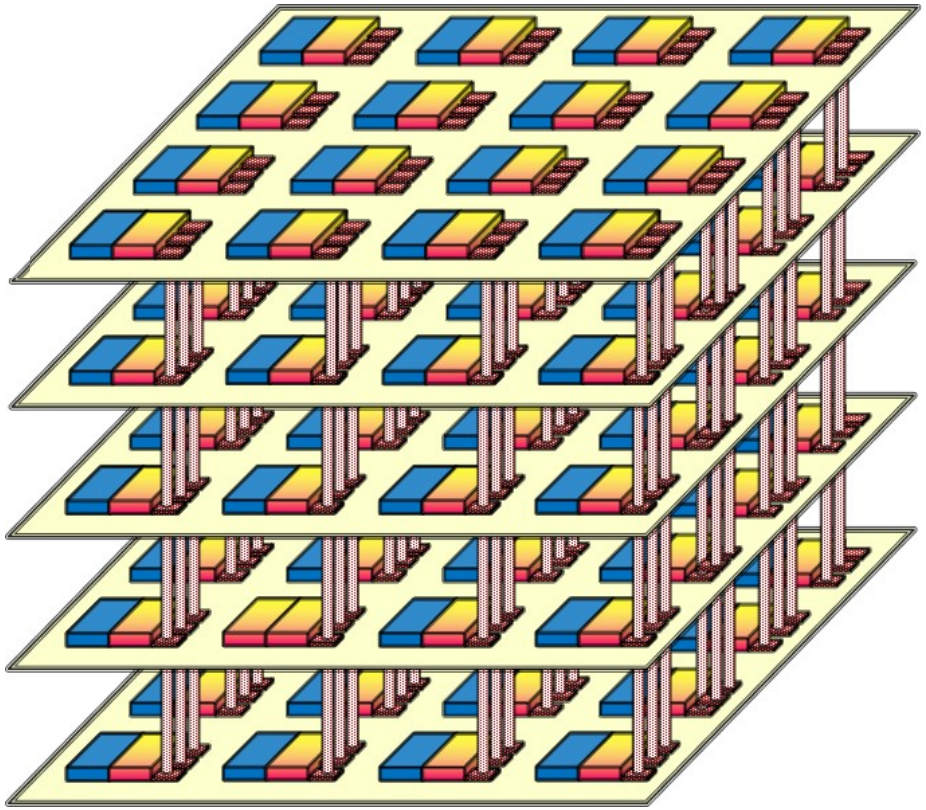
- We evaluate the accuracy-yield relationship in our three previous works

1. NASH: *3-D NoC-based neuromorphic system*

NASH architecture

Processing Elements

Memory Blocks



MSBs

LSBs

- Processing Elements & Memory blocks are in the same die/layer.

Evaluated Architectures (2)

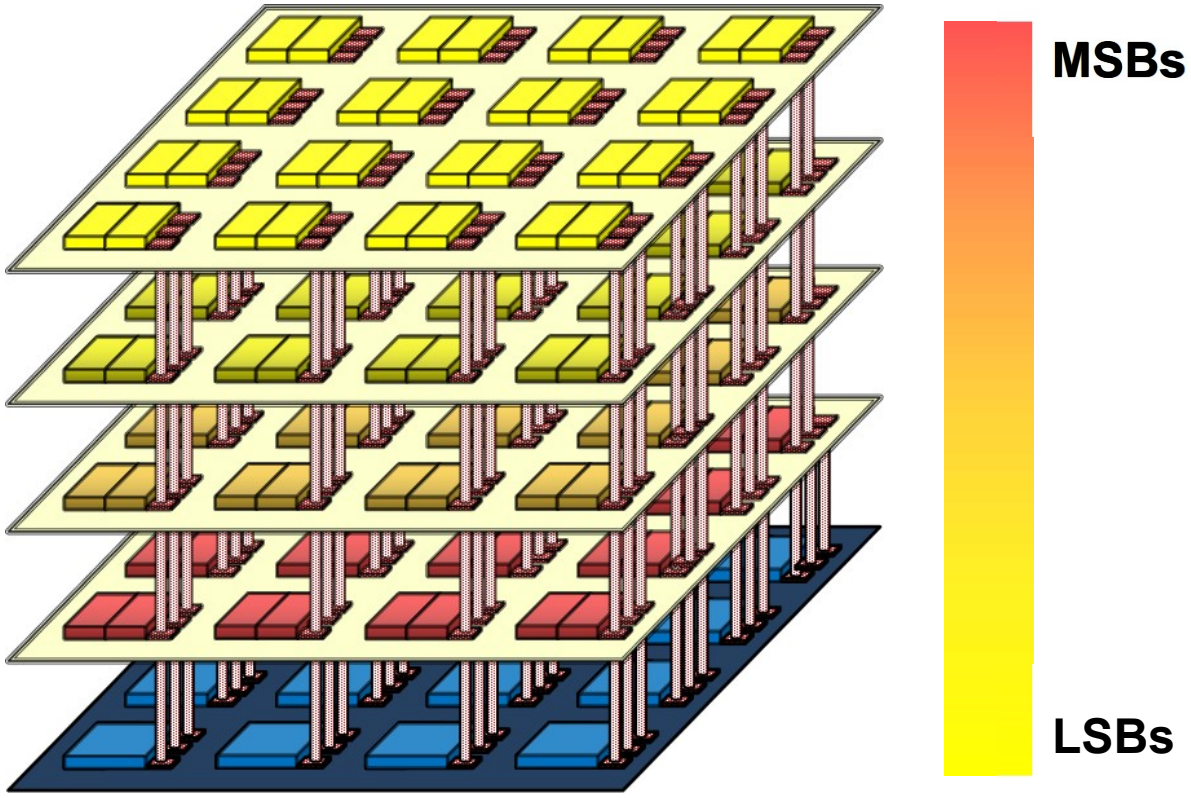
2. 3D-SCP: *3-D Spiking Computing Processor (with splitting synaptic weights in memory)*

3D-SCP architecture

Processing Elements

Memory Blocks

- Processing Elements & Memory blocks are in the different dies/layers.
- Different levels of weight bit are placed in different dies/layers.



Evaluated Architectures (3)

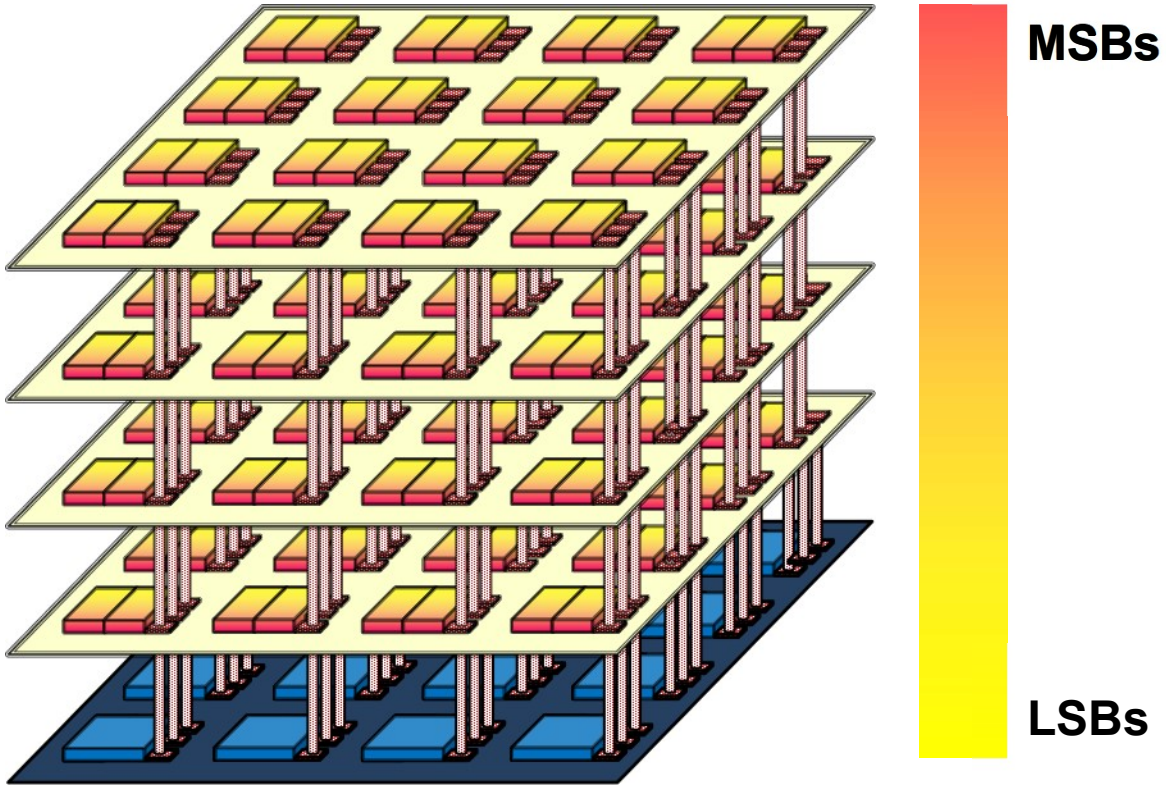
3. NSW 3D-SCP: *Non-Splitting-Weight 3-D Spiking Computing Processor*

NSW 3D-SCP architecture

Processing Elements

Memory Blocks

- Processing Elements & Memory blocks are in the different die/layer.
- Different levels of weight bit are placed in the same die/layer.



Yield/Defect probabilities

- Have:

$$Y_{1\_die} + D_{1\_die} = 1.0 \qquad (1)$$

- Yield of k-dies:

$$Y_{k\_dies} = \prod_{i=0}^{k-1} Y_{i^{th}\_die} \qquad (2)$$

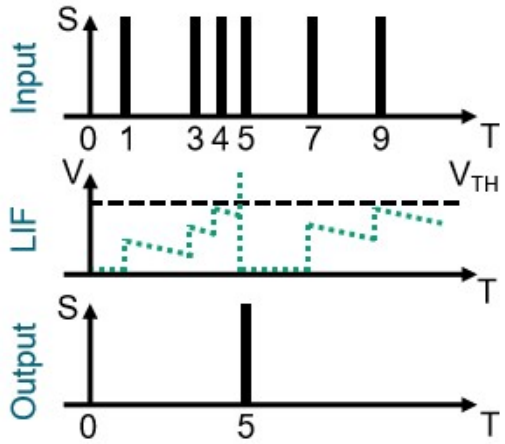
- Defect of k-dies

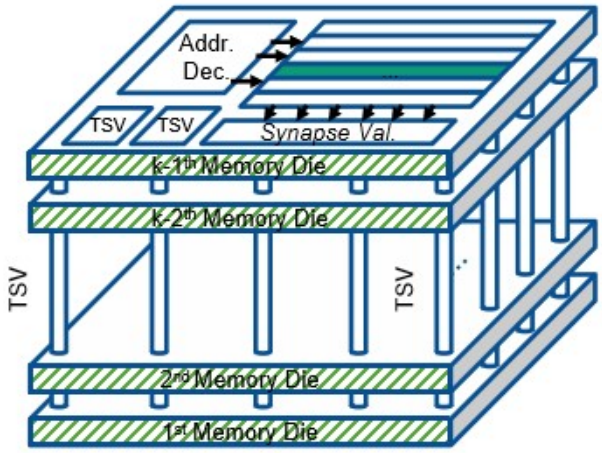
$$D_{k\_dies} = 1 - \prod_{i=0}^{k-1} (1 - D_i) \qquad (3)$$

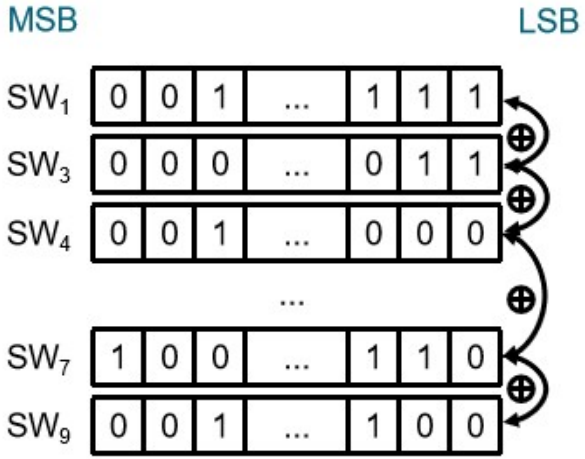
- Insert faults based on defect rate with the ratio of logic and memory component

Defective Operations (1)

- Normal operations of SNN without defects
  - *Accumulation till reaching threshold based on spike event*





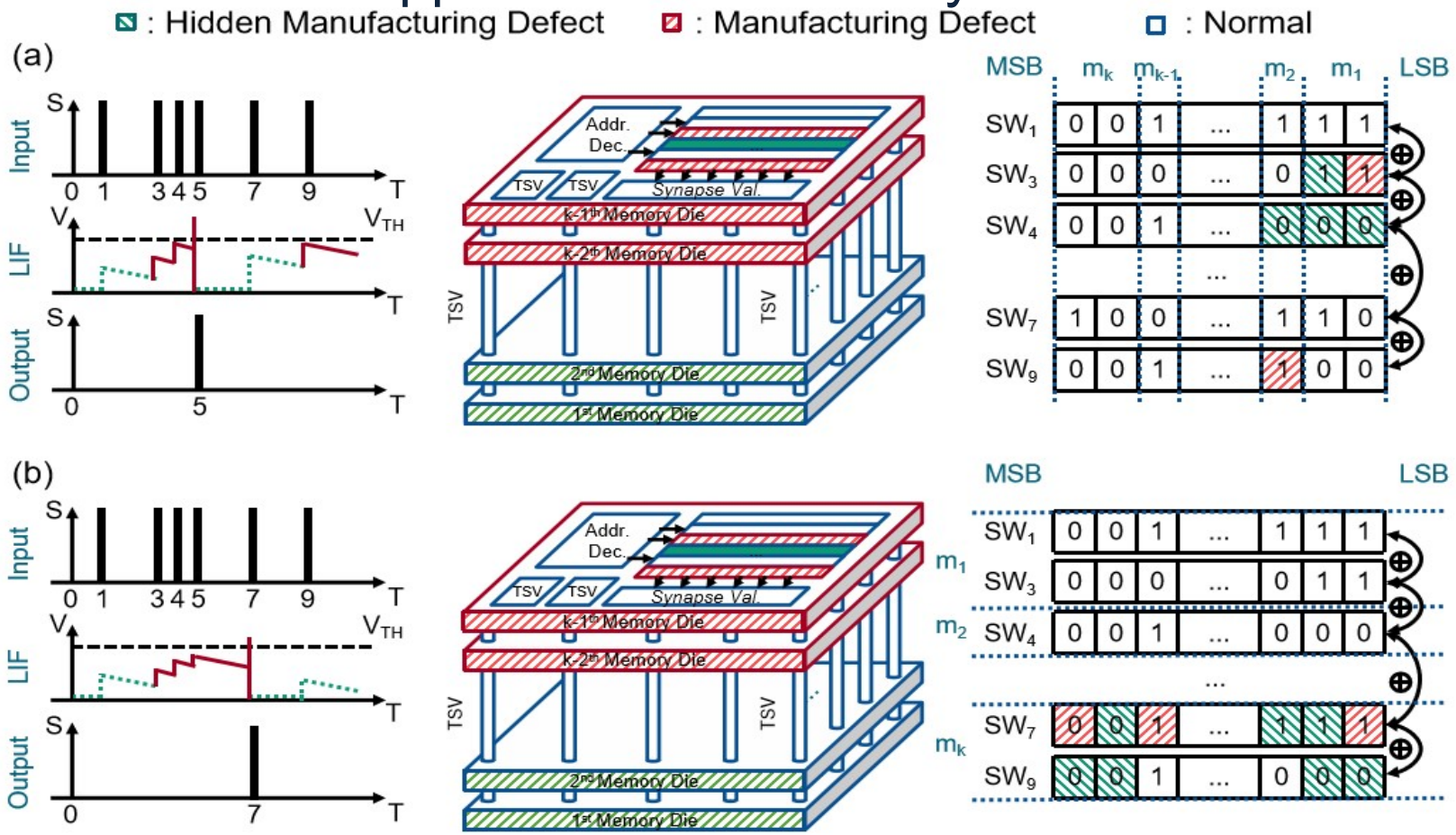


- Defective operations
  - 3D-SCP - Split weights => *Control/Protect the most significant bits (MSBs)*
  - NSW 3D-SCP & NASH => *Cannot protect MSBs because the defects appear randomly*

Value	Original	Flipped bit position			
		MSB	3^{rd} bit	5^{th} bit	LSB
Binary	10101100	00101100	10001100	10100100	10101101
Float	-0.34375	0.34375	-0.09375	-0.28125	-0.3515625
Diff. (%)	0 (0%)	+0.6875 (+200%)	+0.25 (+72.727%)	+0.0625 (+18.182%)	+0.0078125 (+2.273%)

Defective Operations (2)

- Evaluate on 2 upper defective dies/layers



=> NASH & NSW 3D-SCP may get delayed in output spike

Agenda

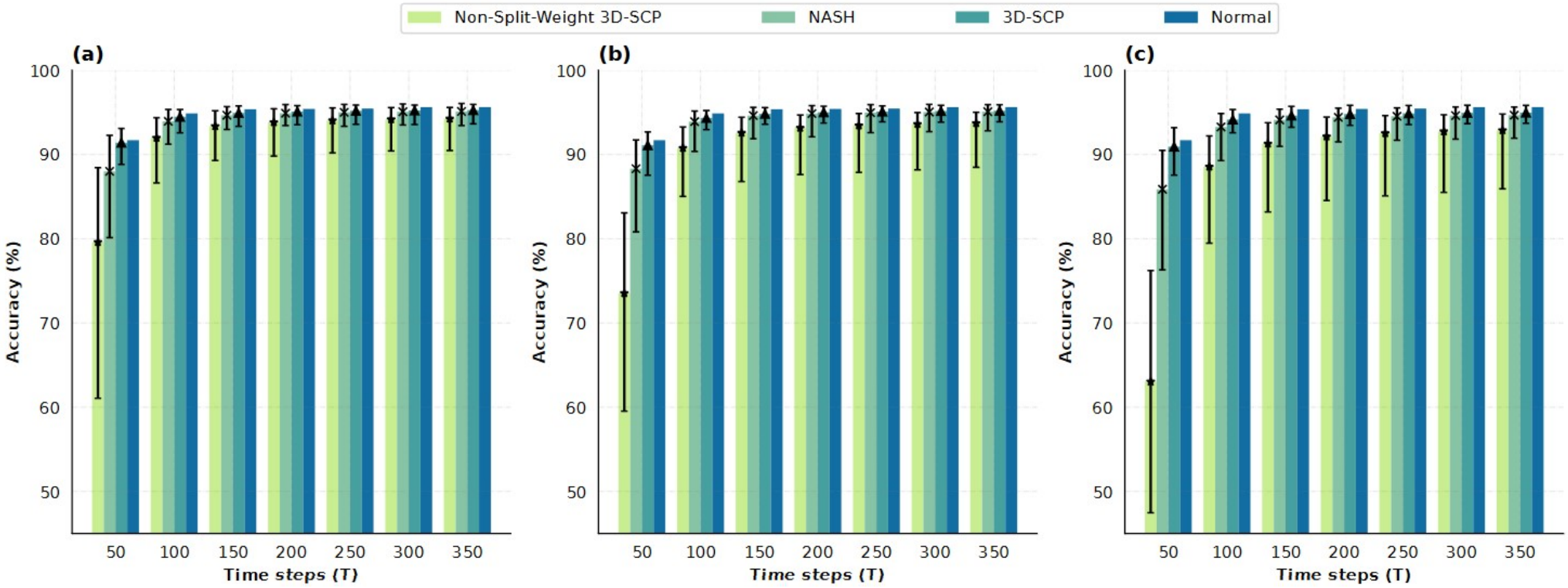
1. Motivations
2. Approach & Methodology
3. Evaluations
4. Conclusions

Evaluation methodology

- Dataset: MNIST
- SNN config. = 784:64:10 & 784:128:10
- Insert faults uniformly based on three assumption yields with MonteCarlo simulation (1000 times)
  - 0.97 (3% at faults);
  - 0.96 (4% at faults);
  - 0.95 (5% at faults).
- Assume:
  - Five stacking layers/dies in 3 architectures
  - Accept two defective upper layers/dies
- Evaluate:
  - Accuracy transformation with three yield rates

Result (1)

- The accuracy of our systems with config. = [784:64:10]

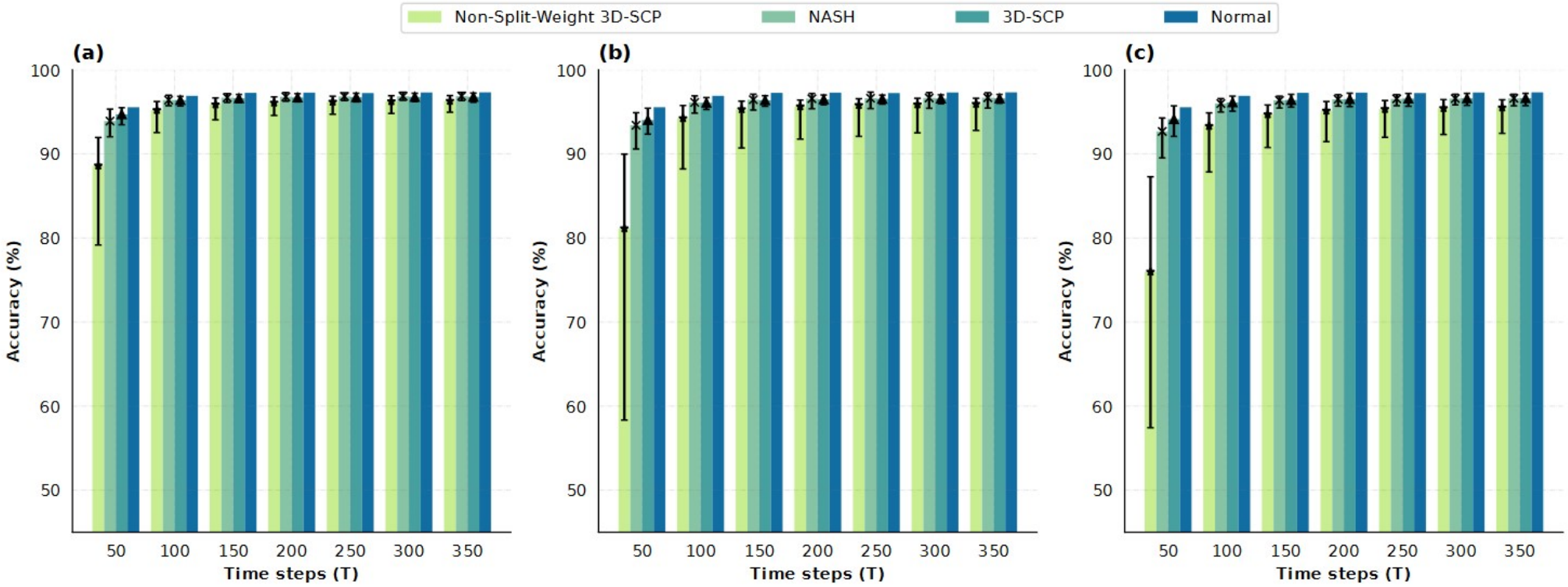


SNN = [784:64:10]. (a) Yield = 0.97; (b) Yield = 0.96; (c) Yield = 0.95

=> Great impact on early but less in long run

Result (2)

- The accuracy of our systems with config. = [784:128:10]



SNN = [784:128:10]. (a) Yield = 0.97; (b) Yield = 0.96; (c) Yield = 0.95

=> Same fault rate but less impact on a bigger FFW model

Result (3)

- Yield improvement with 2 defective layers
  - Configuration of 784:64:10

Defective Rate per Die	Average Accuracy (Loss)			Yield Per Die	Est. Yield	Actual Yield Improv.
	NSW NASH-3DM	NASH	NASH-3DM			
$D_1 = 0.03$	94.24% (-1.38%)	95.11% (-0.51%)	95.27% (-0.35%)	0.97	0.8587	0.9126 (+5.39%)
$D_2 = 0.04$	93.70% (-1.92%)	95.08% (-0.54%)	95.19% (-0.43%)	0.96	0.8153	0.8847 (+6.94%)
$D_3 = 0.05$	92.86% (-2.76%)	94.67% (-0.95%)	94.97% (-0.65%)	0.95	0.7737	0.8573 (+8.36%)

=> Reduce 0.35-2.76% in accuracy for increasing 5.39-8.36% in yield

Comparison

- **SNN configuration = [784:128:10]**

Parameters	Seo *et al.* [11]	Kim *et al.* [13]	TrueNorth [17]	Loihi [12]	ODIN [14]	Karimi *et al.* [16]	Our works		
							NASH [6]	NSW NASH-3DM	NASH-3DM [7]
Benchmark	MNIST	MNIST	MNIST	MNIST	MNIST	MNIST	MNIST		
Accuracy (%)	77.2	84.5	91.94	96	84	99.2	96.61 - 96.88	95.5 - 96.29	96.6 - 96.79
Neuron Model	LIF	IF	IF	DenMem	LIF & Izhikevicz	LIF	LIF		
Synaptic Weight Storage	1-bit SRAM	4, 5, 14-bit SRAM	1-bit SRAM	1-to-9-bit SRAM	4-bit SRAM	CTT twin-cell	8-bit SRAM		
Interconnect	2D	2D	2D	2D	2D	2D	3D		
Implementation	Digital	Digital	Digital	Digital	Digital	Analog Mix-signal	Digital		
Learning Rule	On-chip STDP	Stochastic Gradient Descent	Unsupervised	On-chip STDP	On-chip Stochastic SDSP	Off-chip	On-chip & Off-chip		
Technology	45nm SOI	65nm	28nm	14nm FinFET	28nm FD-SOI	22nm FD-SOI	45nm		

=> Maintain high accuray compared to other works

Agenda

1. Motivations
2. Approach & Methodology
3. Evaluations
4. Conclusions

Conclusion

- 3D-IC design has a critical problem in yield rate
- Accept defects in die(s)/layer(s) to increase the yield rate
  - Evaluate on our three previous works *(NASH; 3D-SCP; NSW 3D-SCP)*
  - Reduce **0.35-2.76%** in accuracy for increase **5.39-8.36%** in yield (784:64:10)
  - Average accuracy in 784:128:10
    - *NASH = 96.61 - 96.88%*
    - *NSW 3D-SCP = 95.5 - 96.29%*
    - *3D-SCP = 96.6 - 96.79%*
- Future works:
  - Propose a methodology and framework for faulty network-of-memory to increase the reliability of neuromorphic systems
    - A methodology maintains an acceptable accuracy without increasing hardware area significantly
    - A framework supports designers at the beginning to choose memory architectures/technologies at ease

Thank you

for your attention.

Q&A