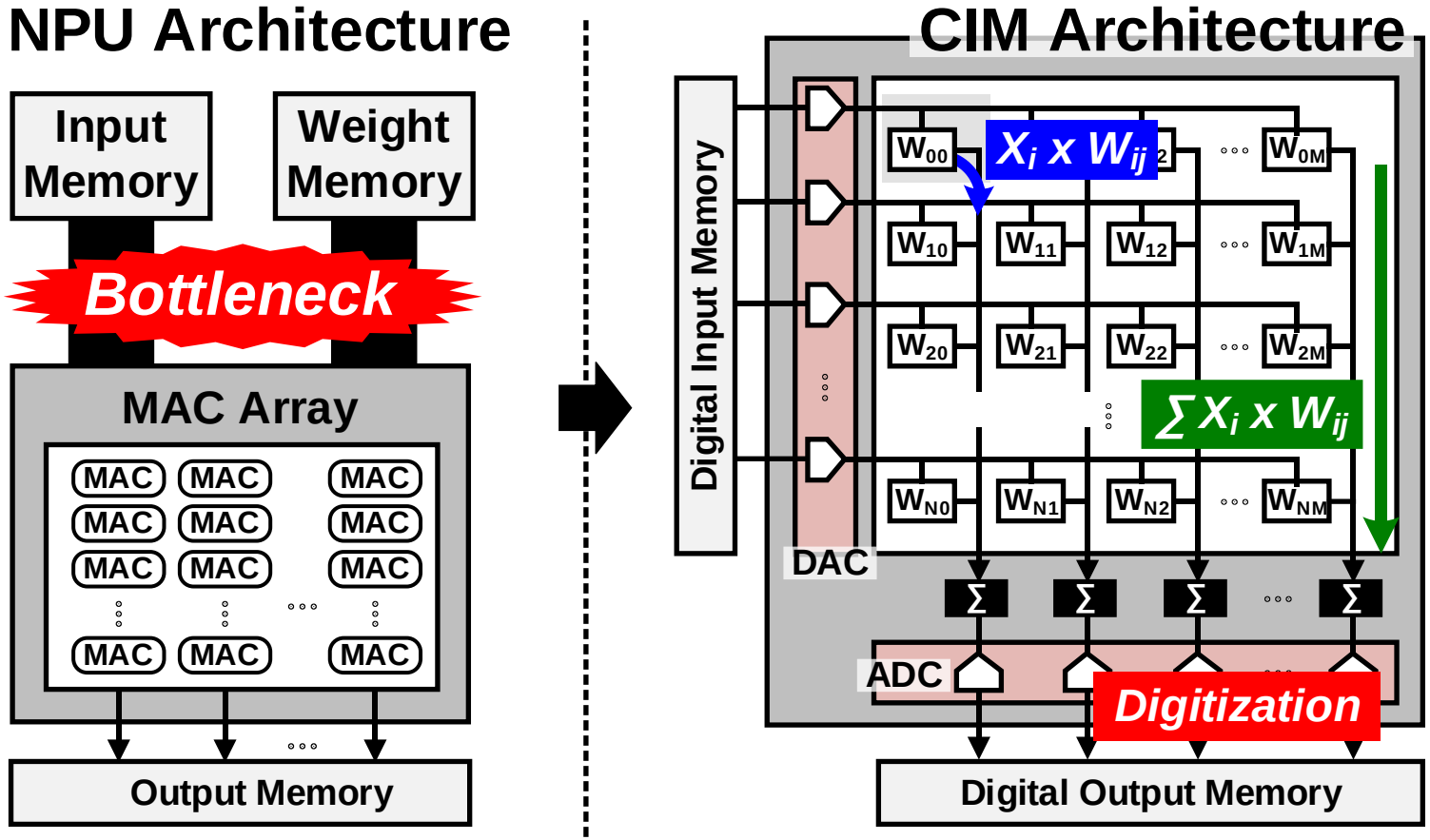
Neuro-CIM: A 310.4 TOPS/W Neuromorphic Computing-in-Memory Processor with Low WL/BL activity and Digital-Analog Mixed-mode Neuron Firing

Sangyeob Kim[1], Sangjin Kim[1], Soyeon Um[1], Soyeon Kim[1], Kwantae Kim[2], and Hoi-Jun Yoo[1]

[1] School of Electrical Engineering, KAIST
[2] Institute of Neuroinformatics, University of Zurich and ETH Zurich

Computing-in-Memory (CIM) Accelerator

- **Multi WLs Driving ➔ Low Energy Efficiency by ADC (<100 TOPS/W)**



Pros

- MEM Access Reduction
- Analog Accumulation

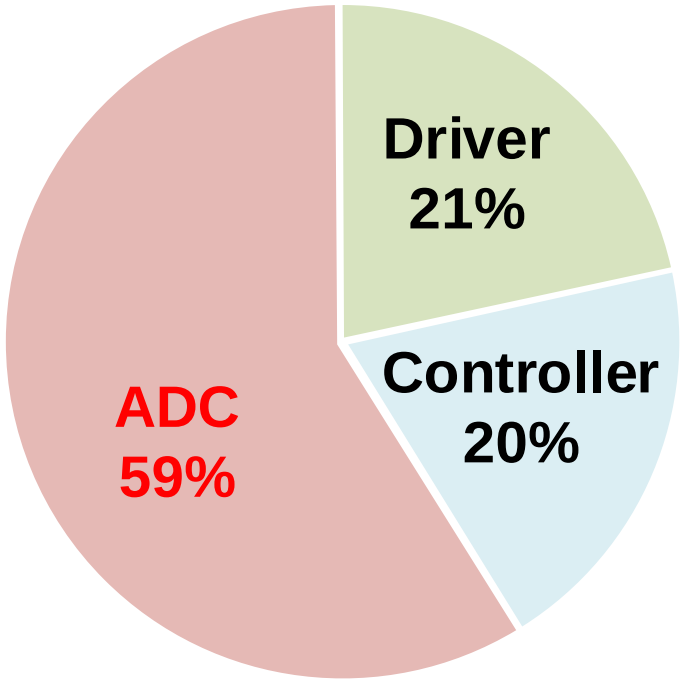
Cons

- 1 WL➔ Multi Cells Active
- 1 Col.➔ Multi WLs Active
- ADC/DAC
  - ➔ Large Power
  - ➔ Large Area

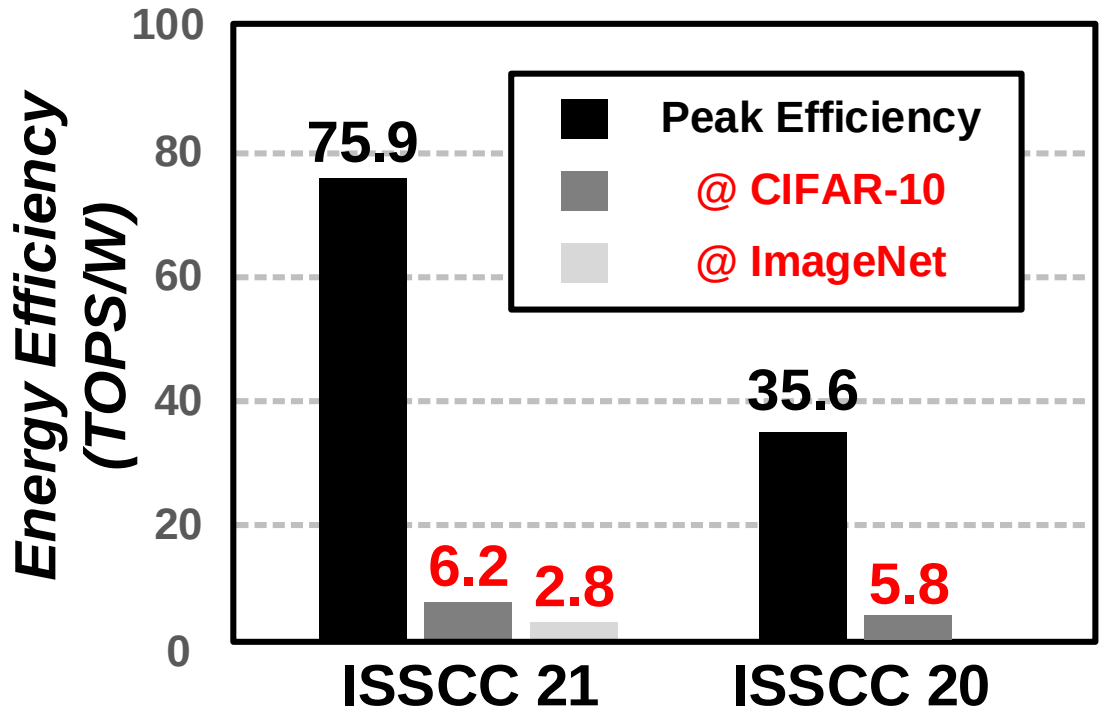
Limitation of Previous CIMs

1. **High Precision ADC is Required for Digital Output Activations**
2. **Low Energy Efficiency due to Low Sparsity in Real Conditions**

<Power Breakdown (VLSI 21)>

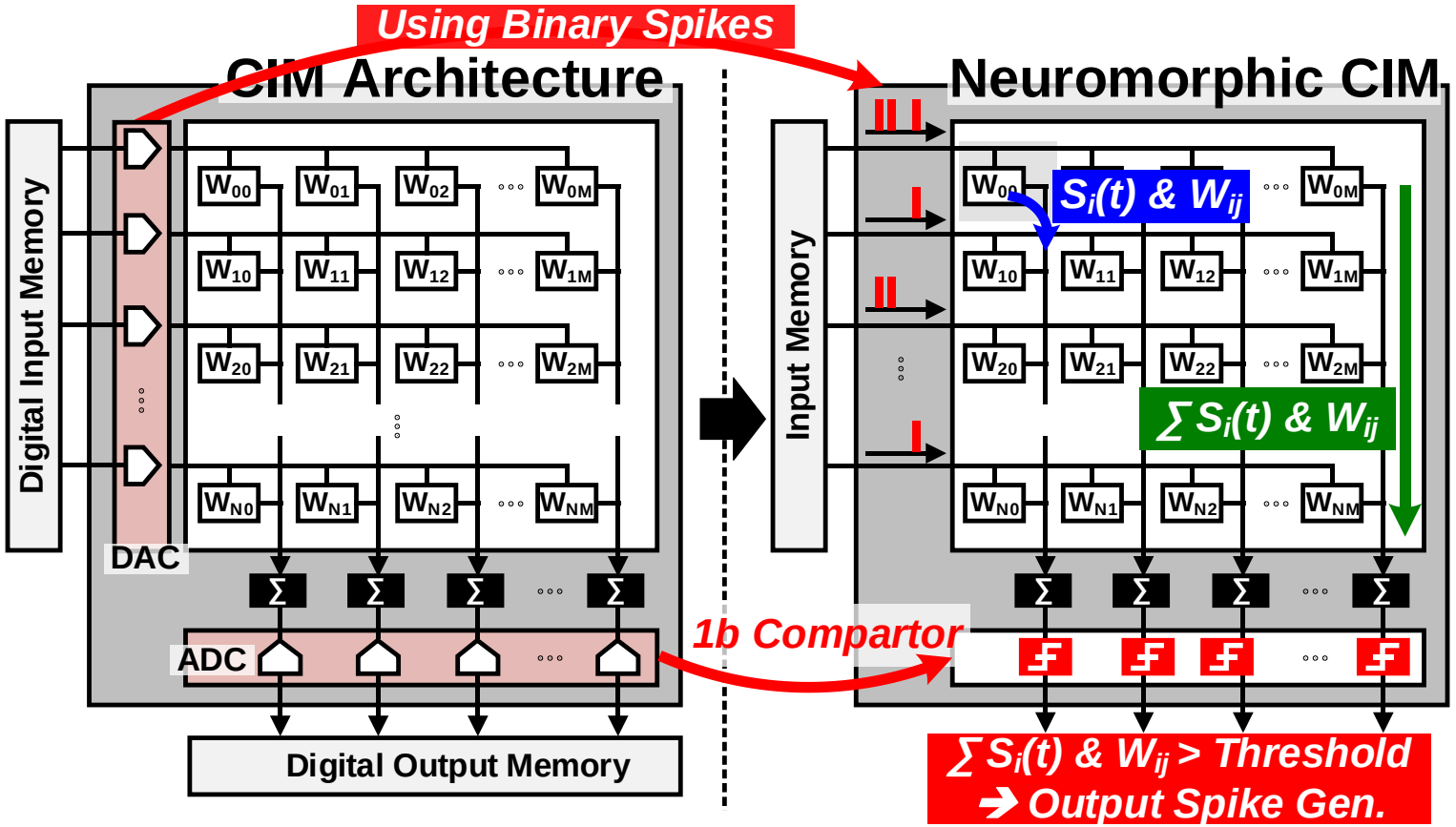


<Low Performance @ Real Application>



Neuromorphic CIM Processor

- **ADC and DAC are Not Necessary ➔ Power/Area Reduction**
- **Event-driven operation ➔ Input sparsity, but low weight sparsity**



Pros

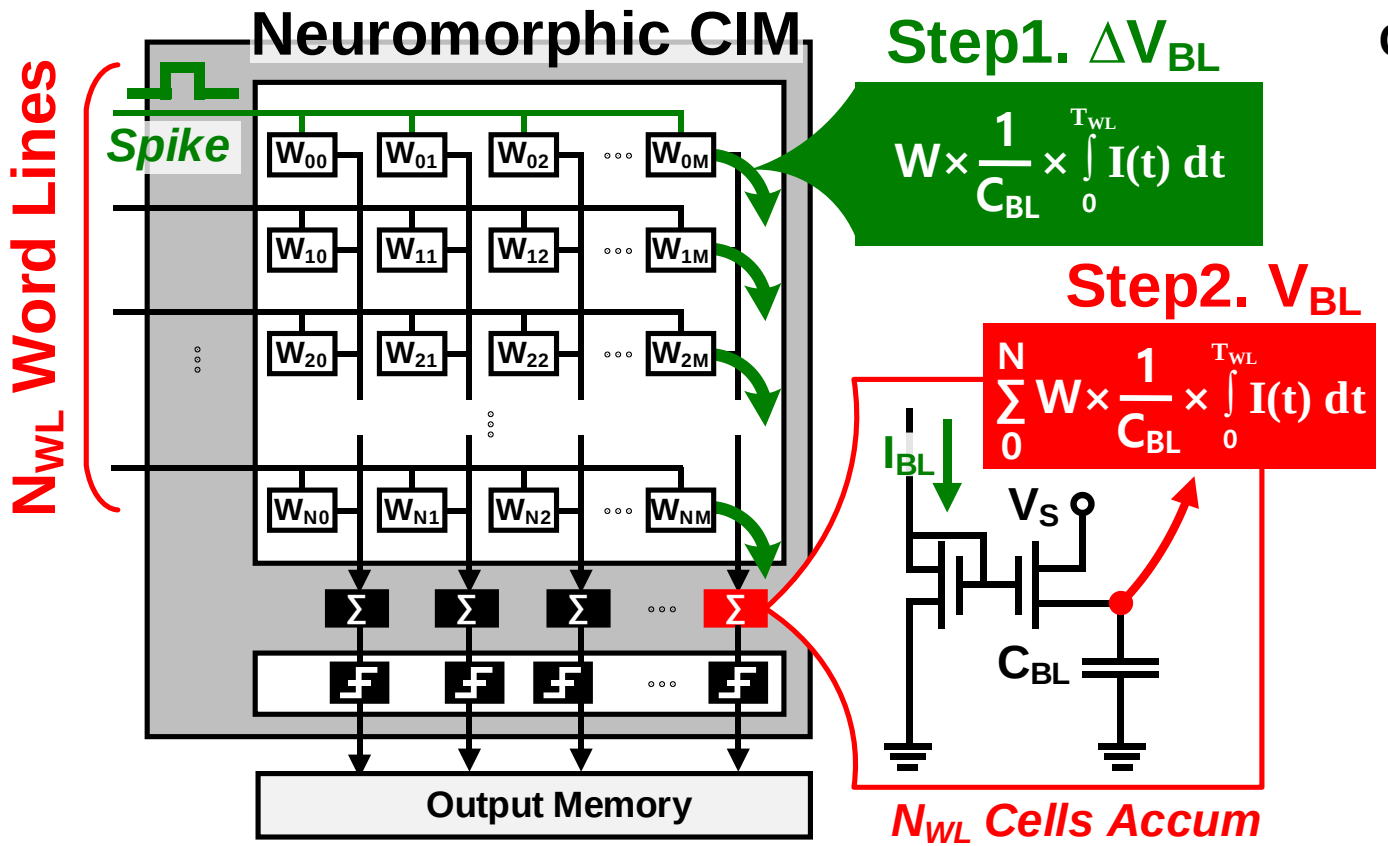
- **High Input Sparsity**
- **No ADC/DAC**

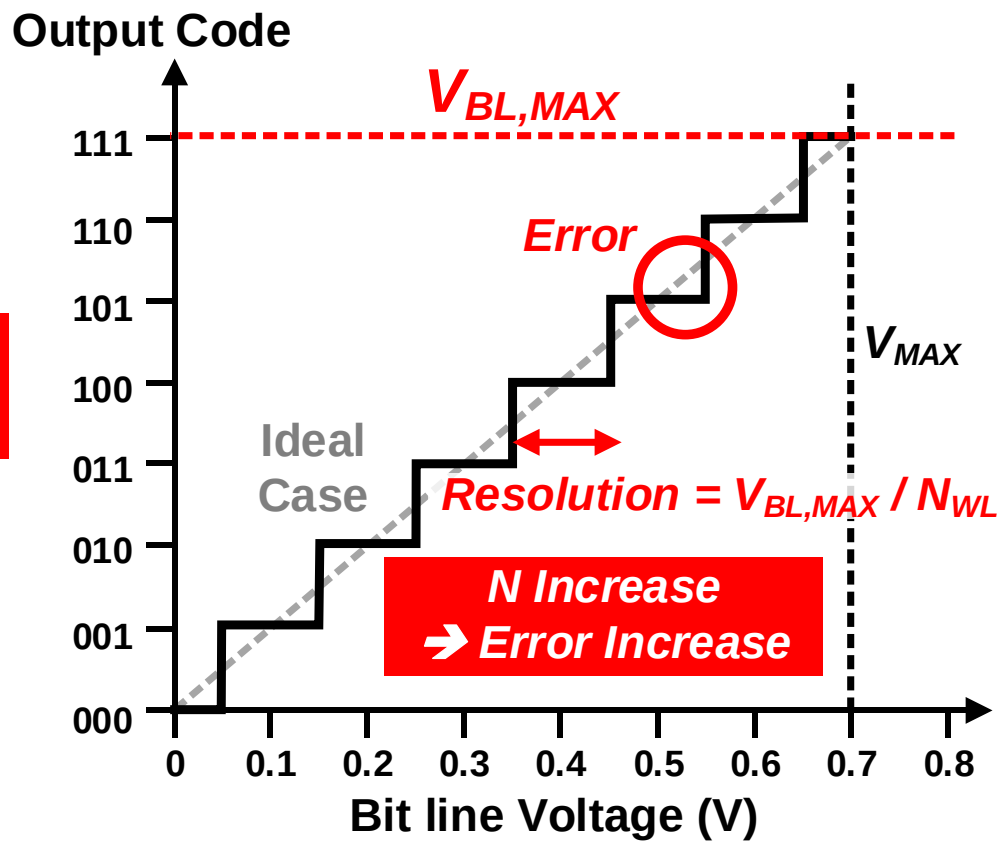
Cons

- **1 WL ➔ Multi Cells**
- **1 Col. ➔ Multi WLs**
- **Limited range of V_{BL}**

Limited V_{BL} Range of Neuromorphic CIM

- **V_{BL} Resolution and Range are Limited by the Number of WL (N_{WL})**
  - Range of V_{BL} : 0 ~ N_{WL} x ΔV_{BL}, Resolution of V_{BL} : $V_{BL,MAX}$ / N_{WL}

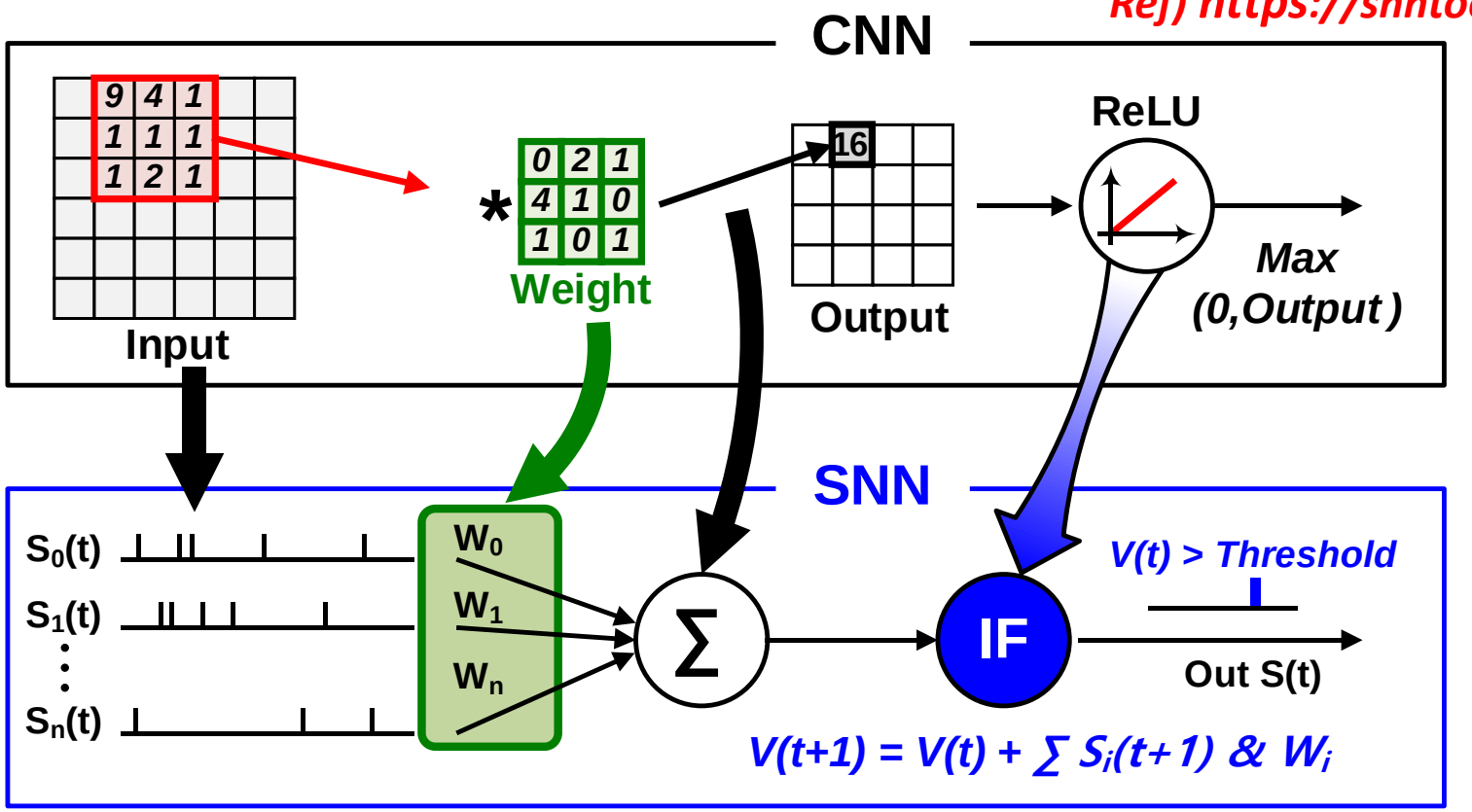




High Accuracy of Neuromorphic CIM

- **Spiking-Neural-Network (SNN) Conversion from trained CNN**
  - After CNN training, **transfer weight to SNN ➔ Highly Accurate SNN [1,2]**

Ref) https://snntoolbox.readthedocs.io/en/latest/



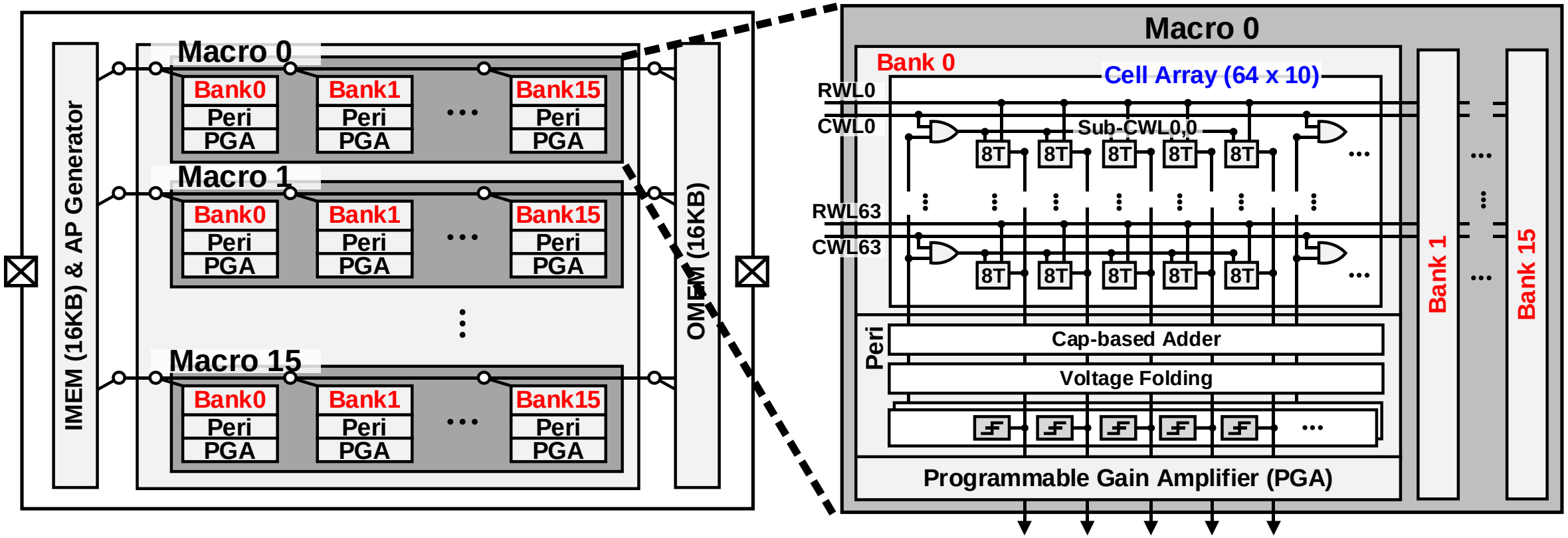
Dataset	Type	Acc.(%)
Cifar-10	CNN	90.80
	SNN	90.05[1]
Cifar-100	CNN	71.82
	SNN	69.67[2]
ImageNet	CNN	70.08
	SNN	69.00[2]

[1] J. Wu et al. "Progressive tandem", TPAMI 2021

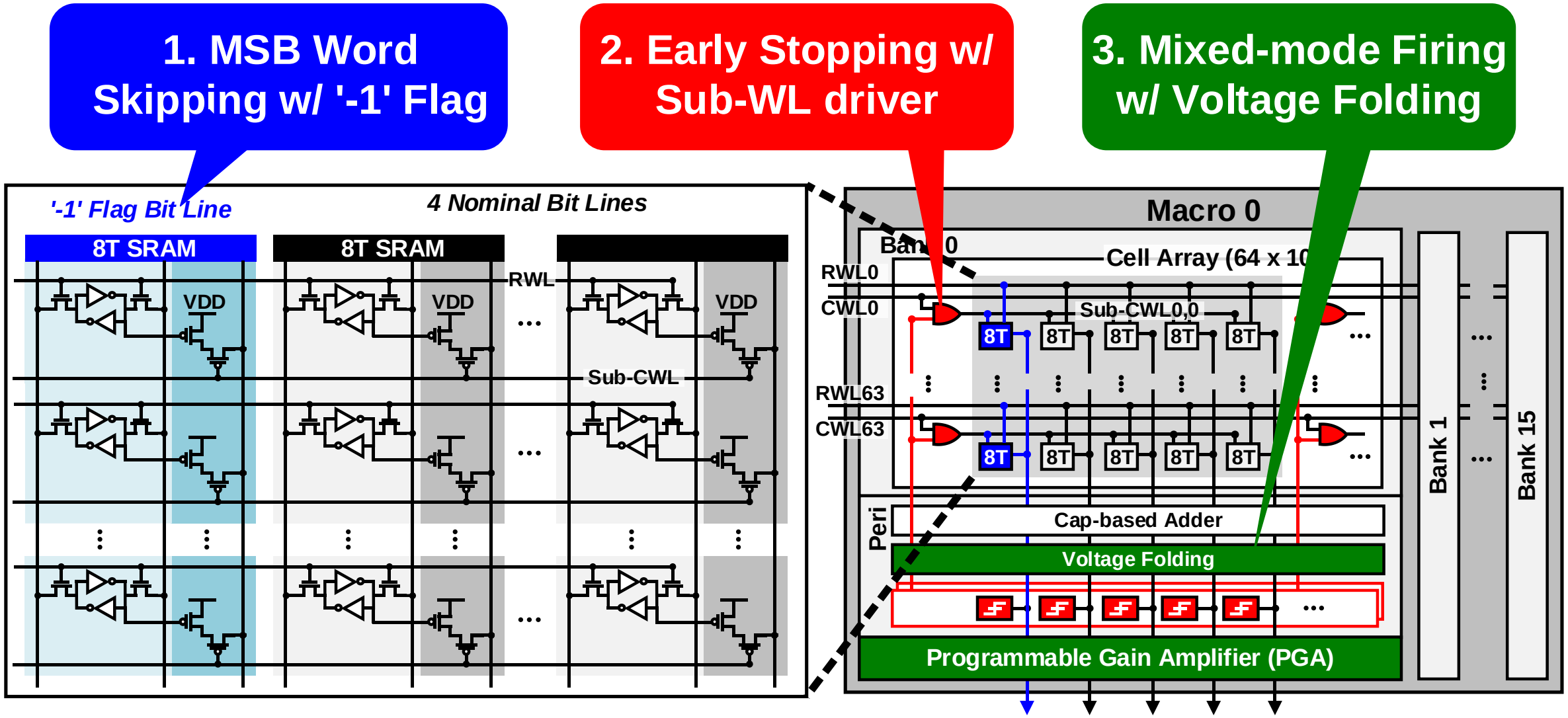
[2] N. Rathi et al. "DIET-SNN", TNNLS 2021

Overall Architecture

- 16 Macros with 16 Banks ➔ Each Bank has 64x10 8T cell array
- Cap-based Adder, Voltage Folding Logic, Comparator array, PGA

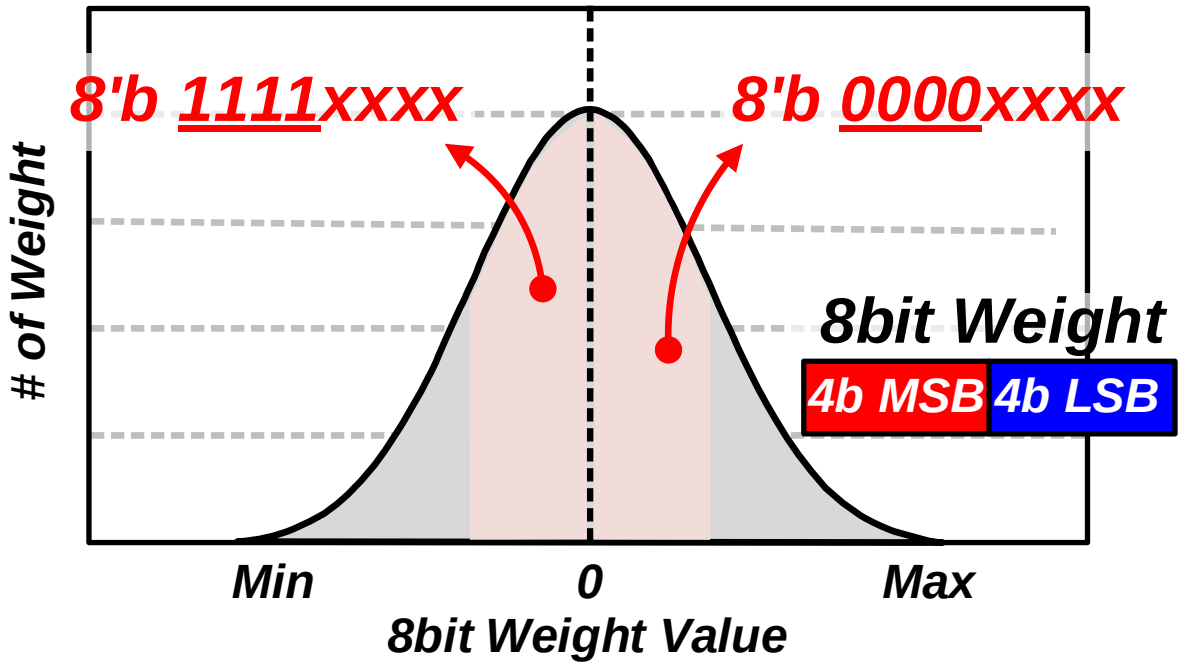


Overall Architecture

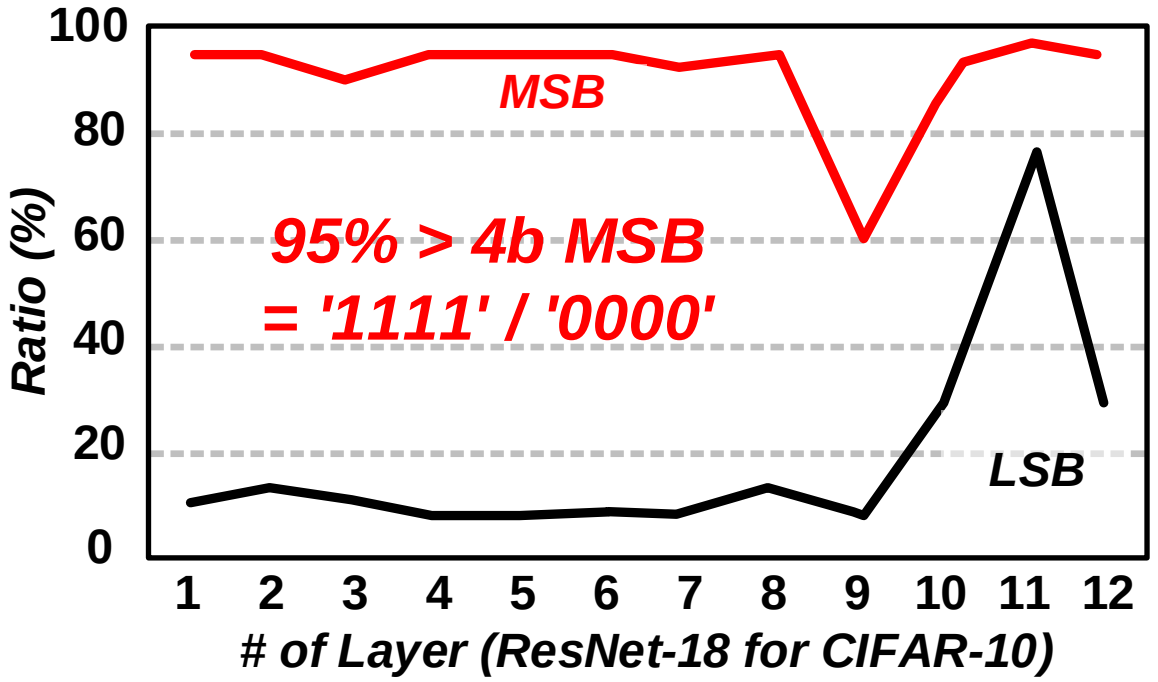


Characteristics of Weight Stored in CIM

- **High Negative Sign Extended Bits (‘1111xxxx’) Ratio of MSB Part**
  - Weight has gaussian distribution ➔ Most MSB has negative sign extended bits
  - **45% of total computation power is consumed by processing negative sign**



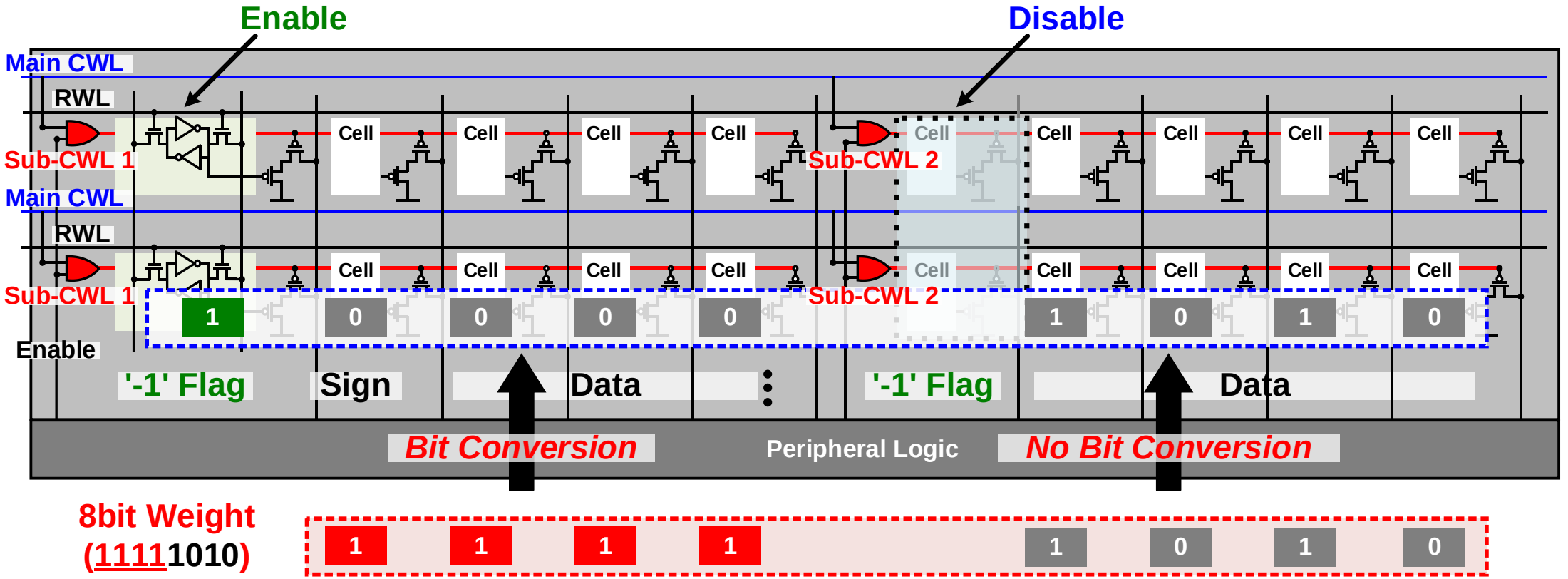
<Weight Distribution>



<Sign Extended Bits Ratio>

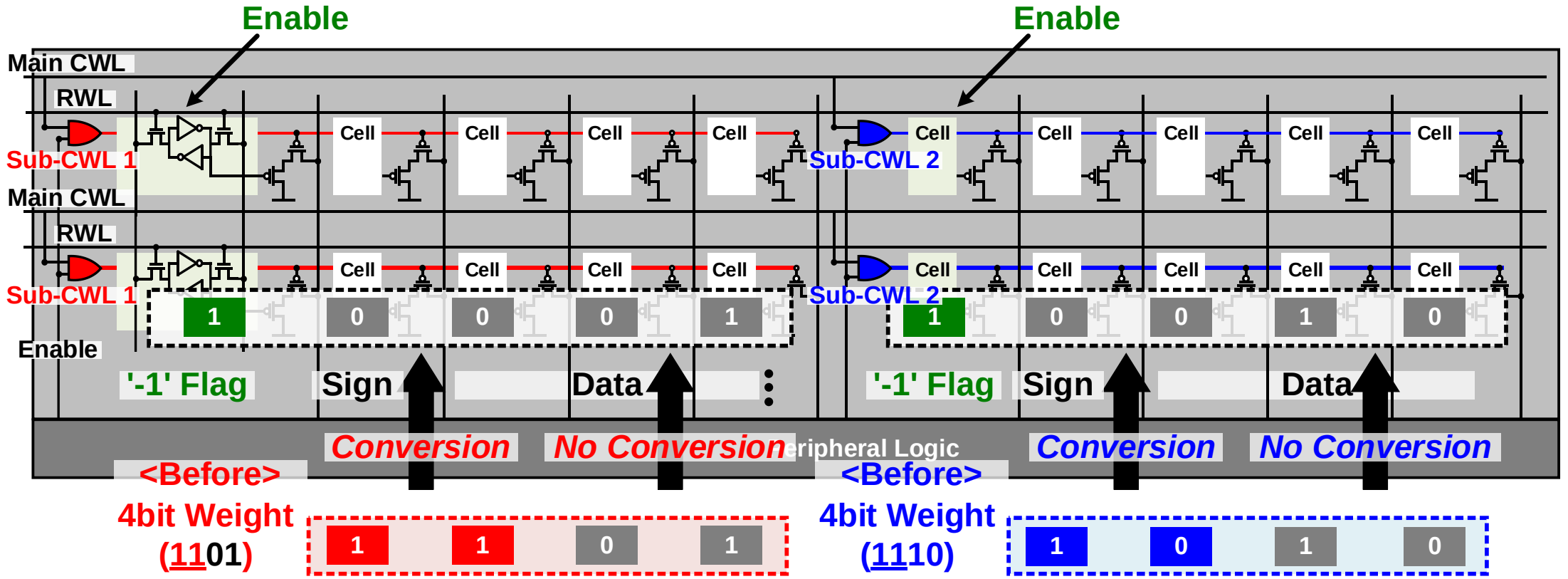
MSB Word Skipping (MWS) with '-1' Flag

- **MSB BL Activity Reduction**
  - MSB **'-1' Flag BL enables BL voltage not to switch.**
  - **4bit Negative sign extend bits ('1111') ➔ 5bit '-1' flag + zeros ('10000')**



MSB Word Skipping (MWS) with '-1' Flag

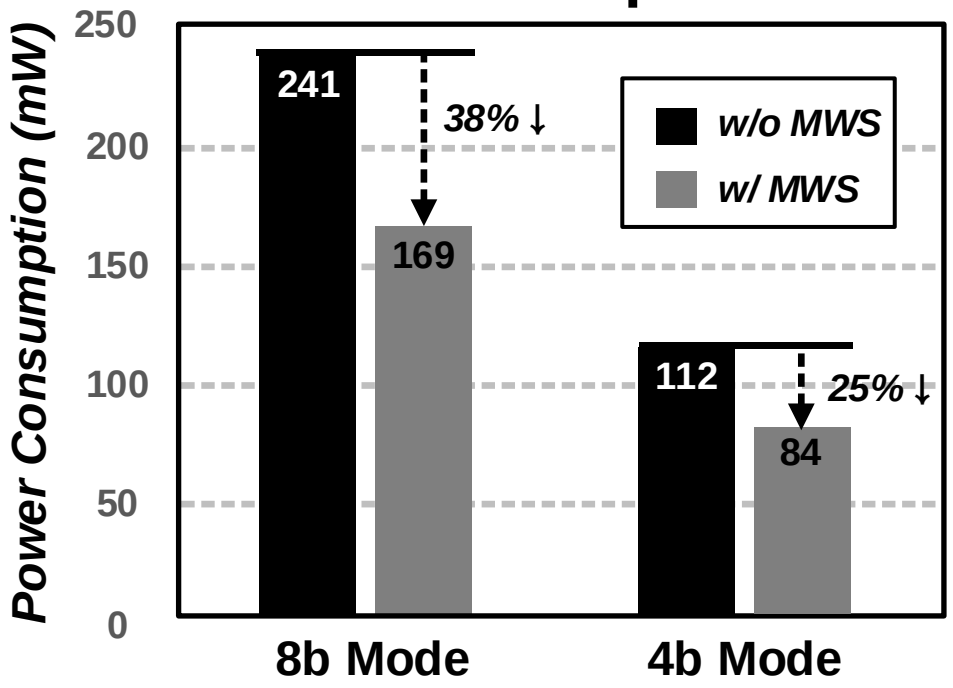
- **MSB BL Activity Reduction**
  - Two **'-1' Flag BLs enable BL voltage not to switch.**
  - **2bit Negative sign extend bits ('11') ➔ 3bit '-1' flag + zeros ('100')**



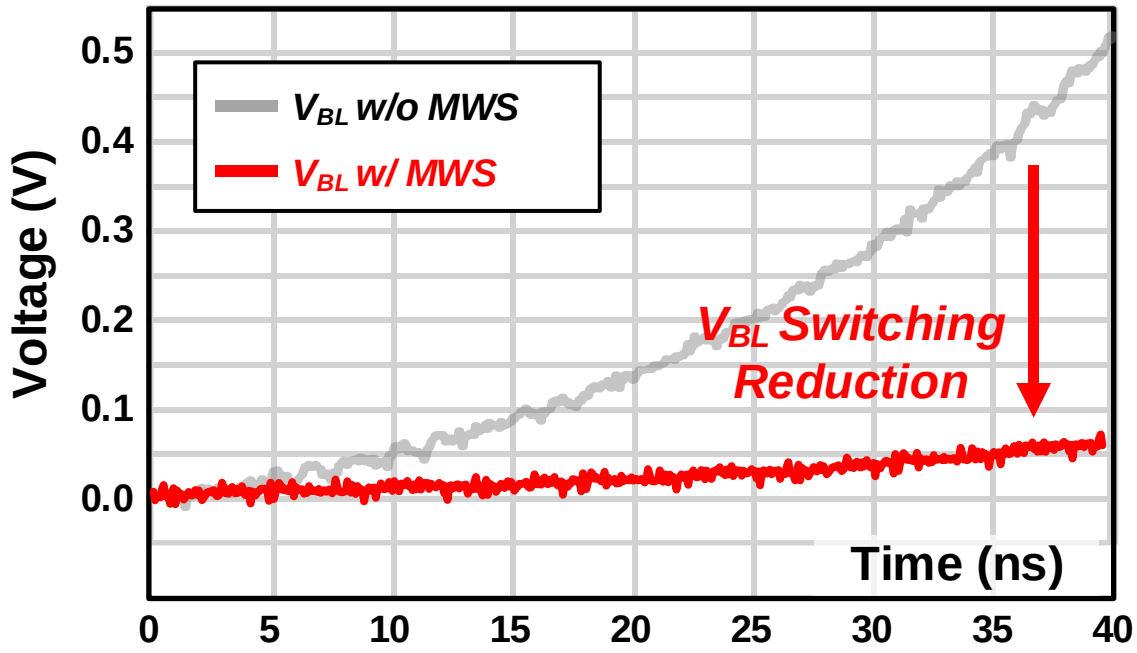
MSB Word Skipping (MWS) with '-1' Flag

- **Performance of MSB Word Skipping**
  - **38% power consumption reduction** @ 8b weight mode case
  - **25% power consumption reduction** @ 4b weight mode case

<Power Consumption>



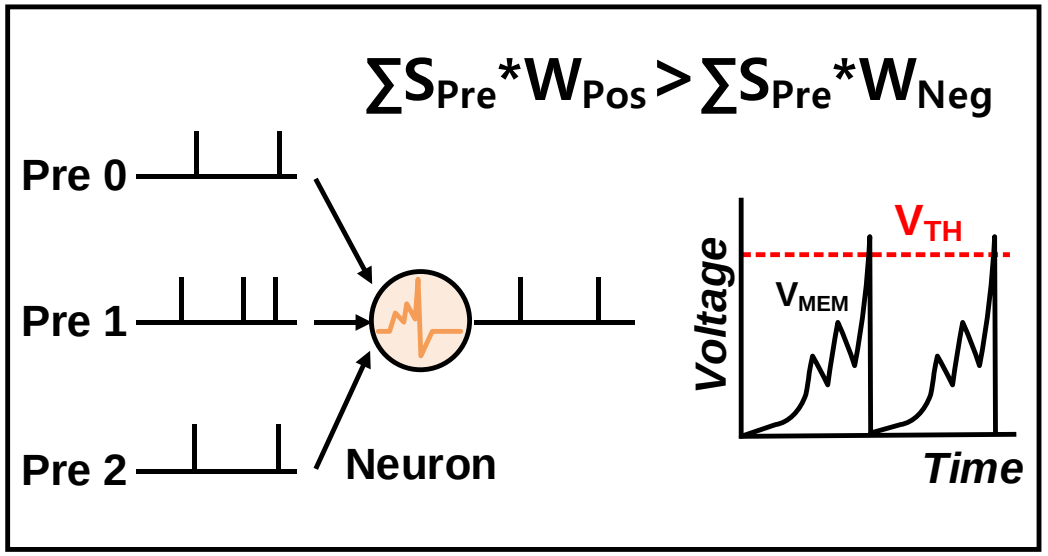
<Measured Waveforms for MWS>



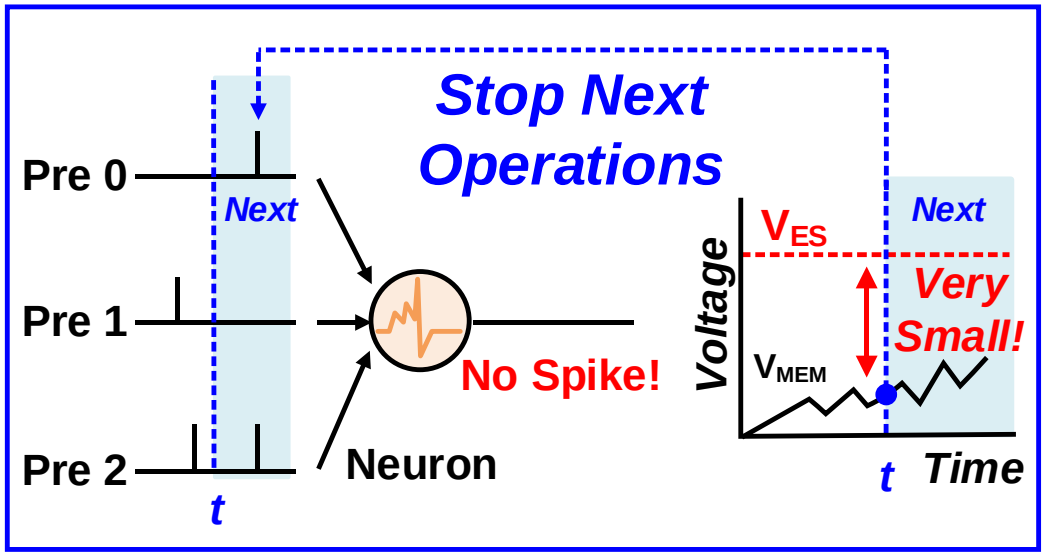
Motivation of Early Stopping (ES)

- **Power Reduction by Eliminating Redundant Operation**
  - Small membrane voltage (V_{MEM}) neuron ➔ No output spike
  - **If $V_{MEM} < V_{ES}$ (Hyper Param) @ T_{ES}** ➔ Early stopping

<Firing Neuron>



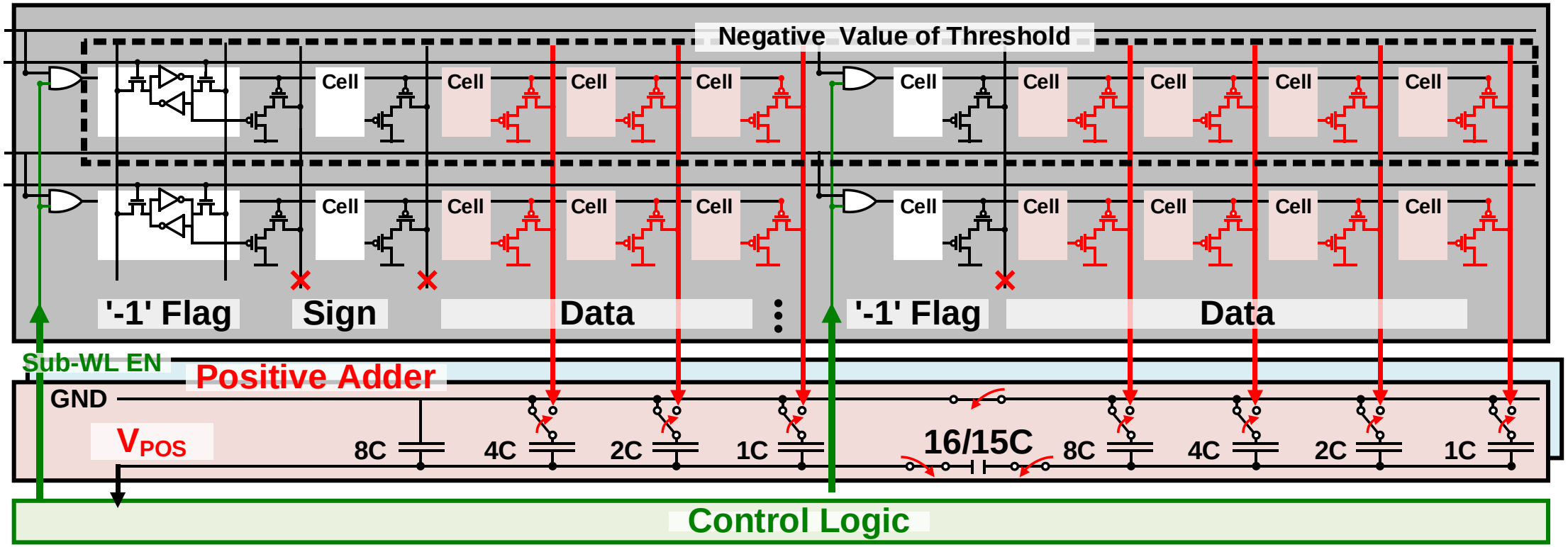
<Non-Firing Neuron w/ ES>



Early Stopping ➔ Power Reduction

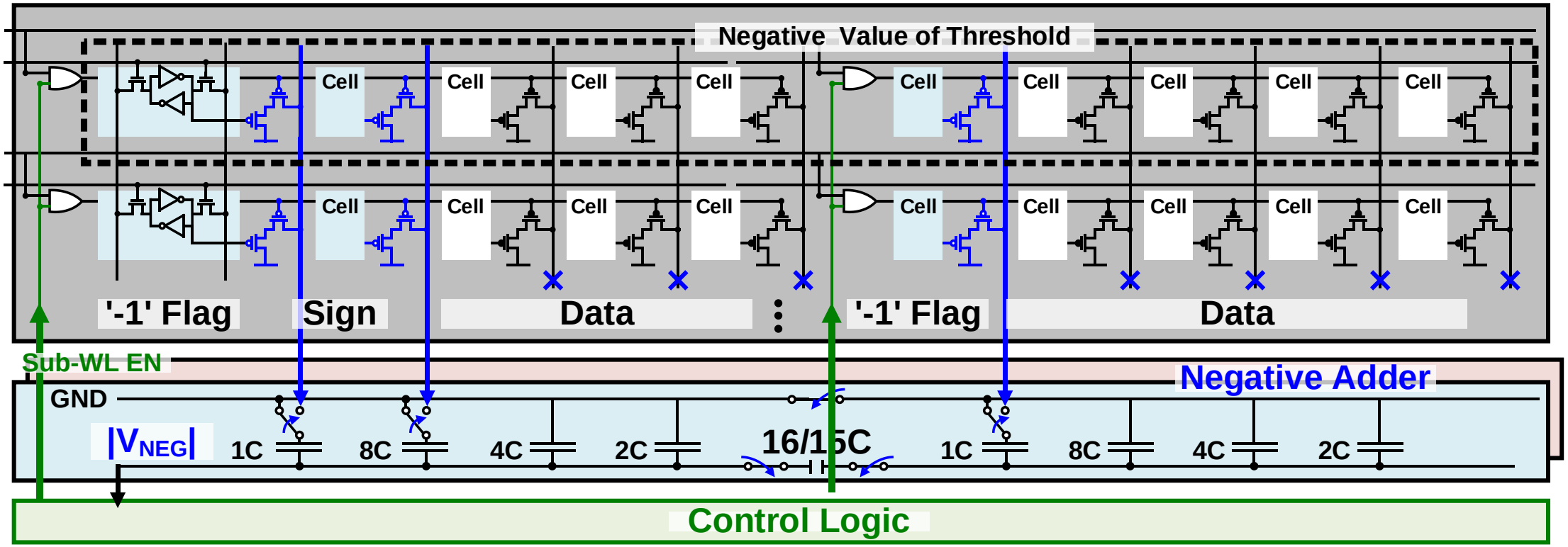
(1) V_{POS} Generation

- **Connecting Bridge Cap. and Accumulating 2's Complement Weight**
  - Only BL voltage of data bit lines ➔ pos. adder (Flag & Sign bit not transferred)
  - Positive adder generates **positive part of (∑ W&S - Threshold)**



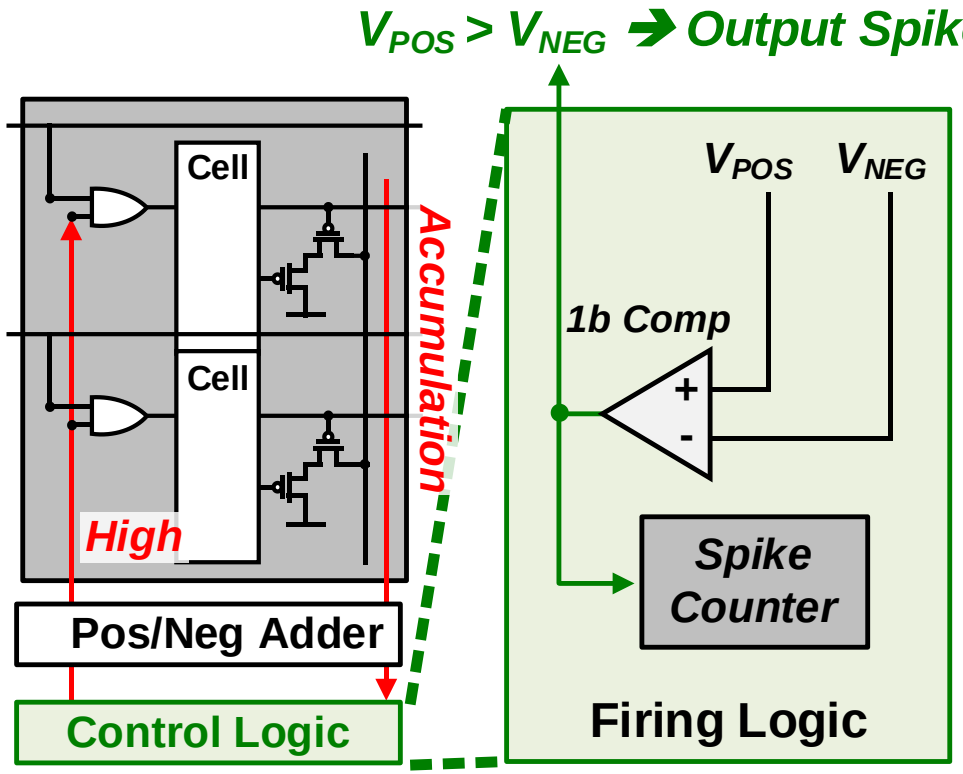
(2) V_{NEG} Generation

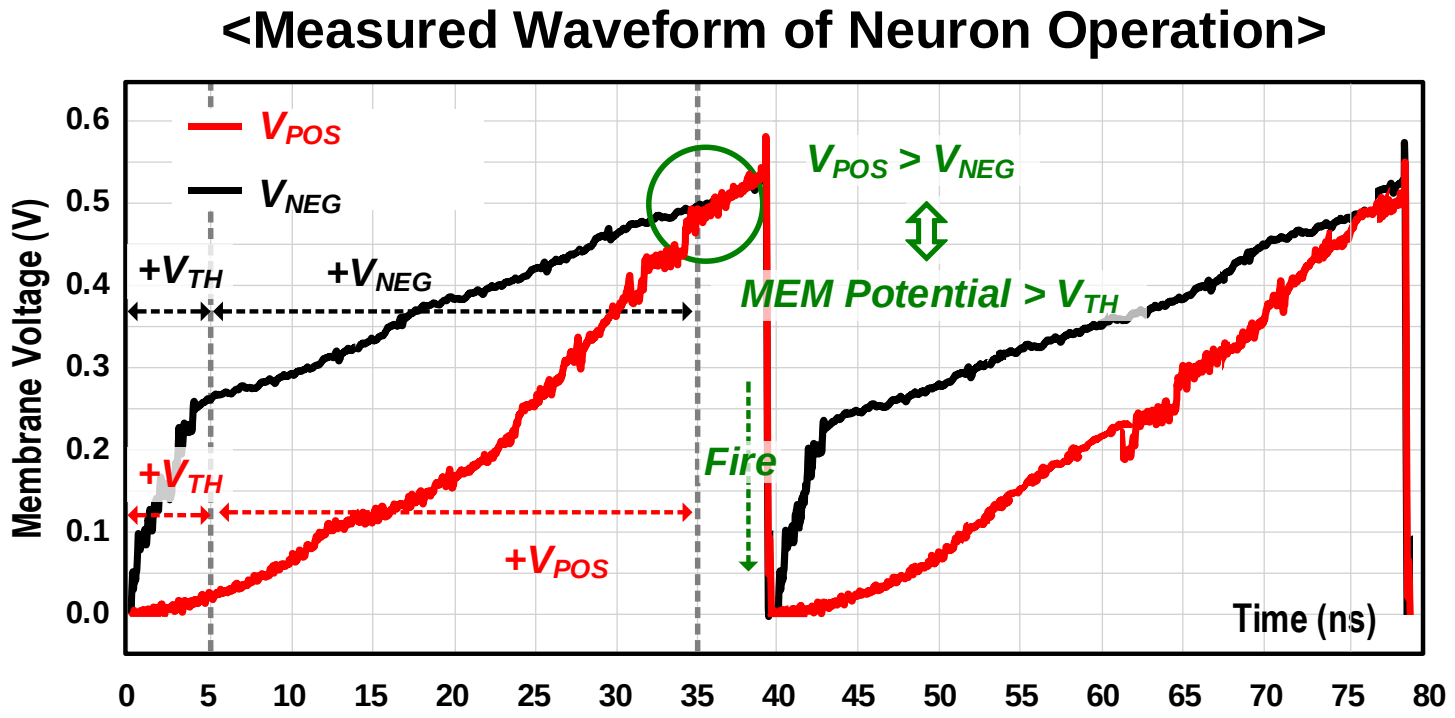
- **Connecting Bridge Cap. and Accumulating 2's Complement Weight**
  - BL voltage of '-1' flag bits and sign bit are transferred to neg. adder
  - Negative adder generates **negative part of (∑ W&S - Threshold)**



(3) Stopping before T_{ES}

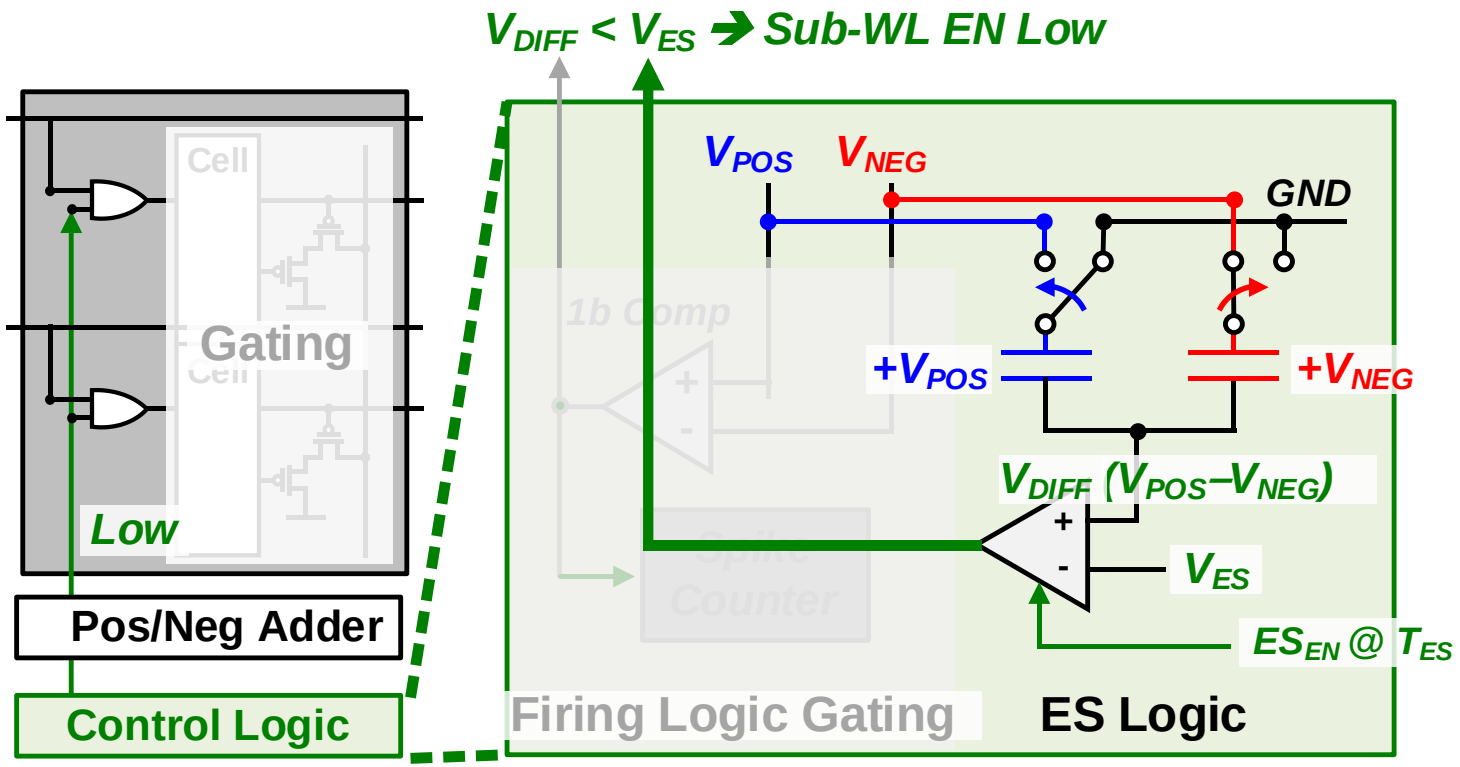
- **1-bit Analog Comparator Generates Output Spike**
  - V_{POS} > V_{NEG} in comparator ➔ output spike firing
  - Spike counter stores the number of output spikes for output memory

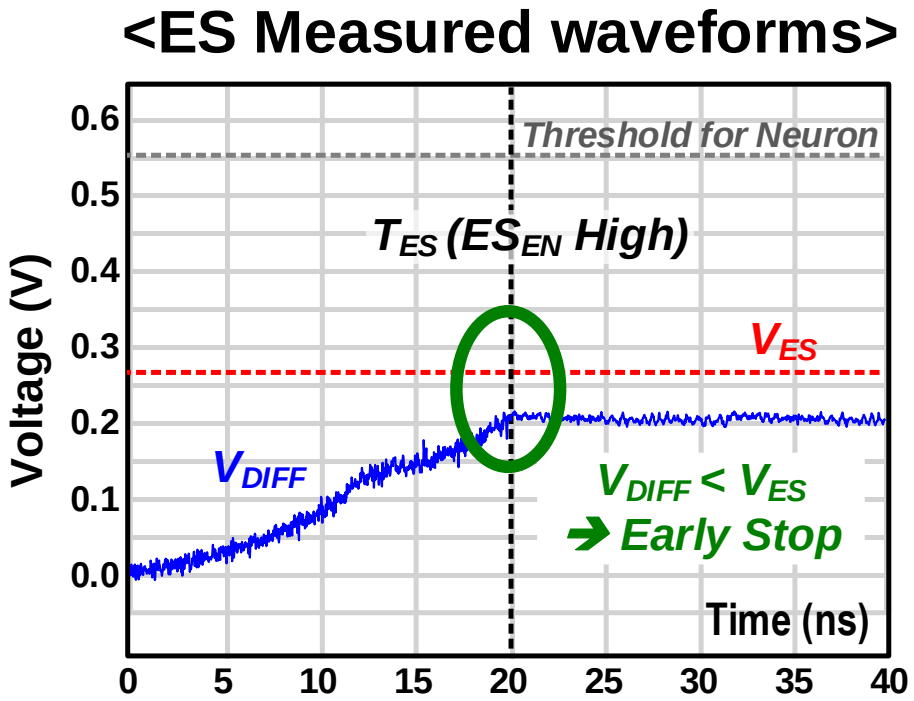




(4) Stopping @ T_{ES}

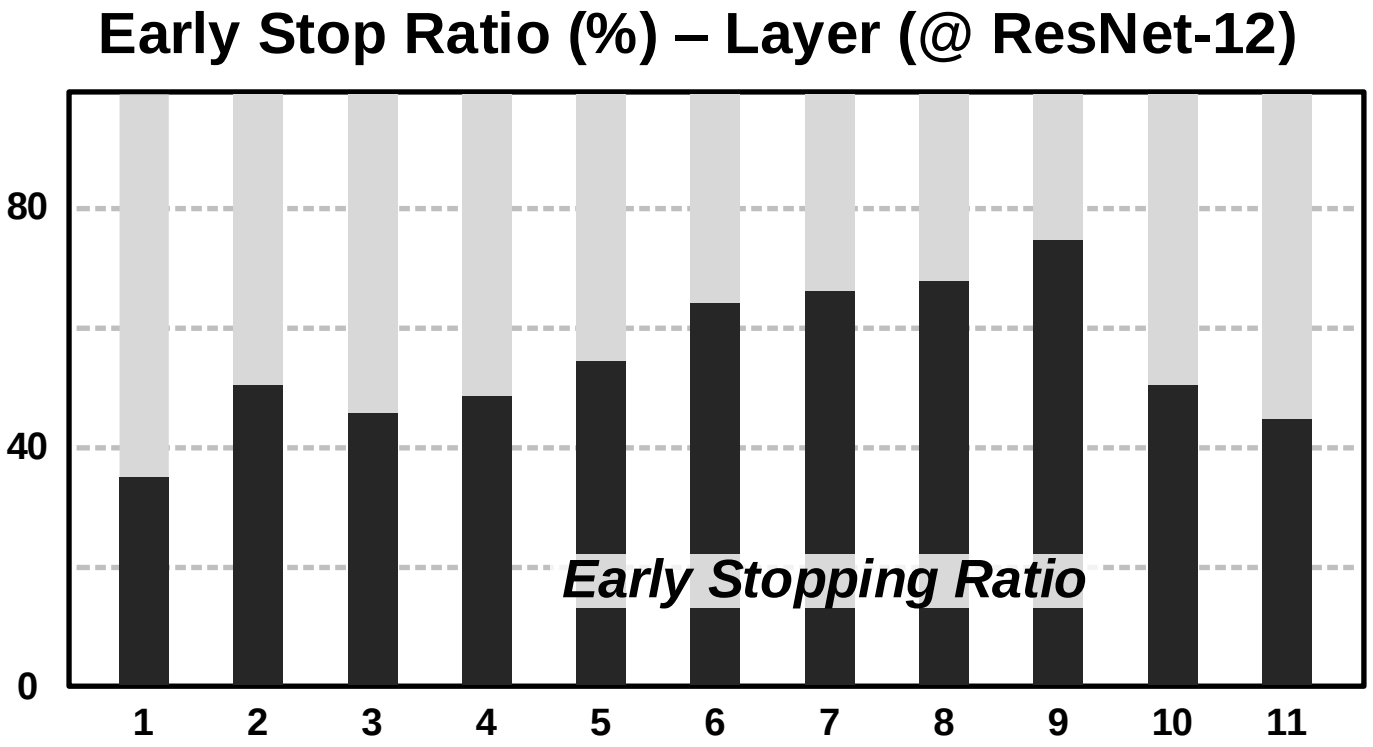
- **Predicting Non-Firing Neuron and Stopping Neuron Operation**
  - **At T_{ES}**, compare V_{DIFF} and early stop voltage (V_{ES})
  - **If $V_{DIFF} < V_{ES}$, processing is stopped to reduce power**

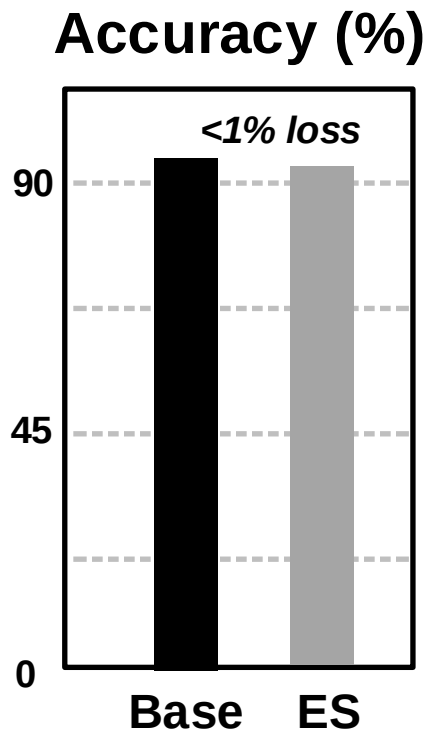


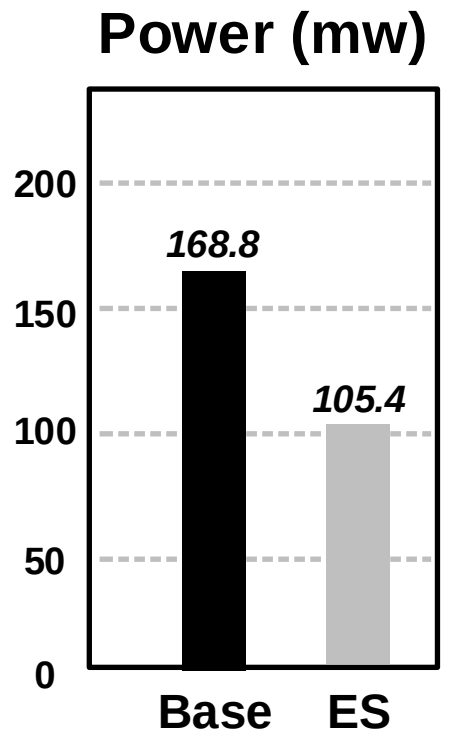


Performance of Early Stopping

- **Reducing Power Consumption by Early Stopping (@ CIFAR-10)**
  - **50~70 % of neurons** in each layer are **early terminated** by prediction
  - **37.6% power consumption is reduced** by early Stopping

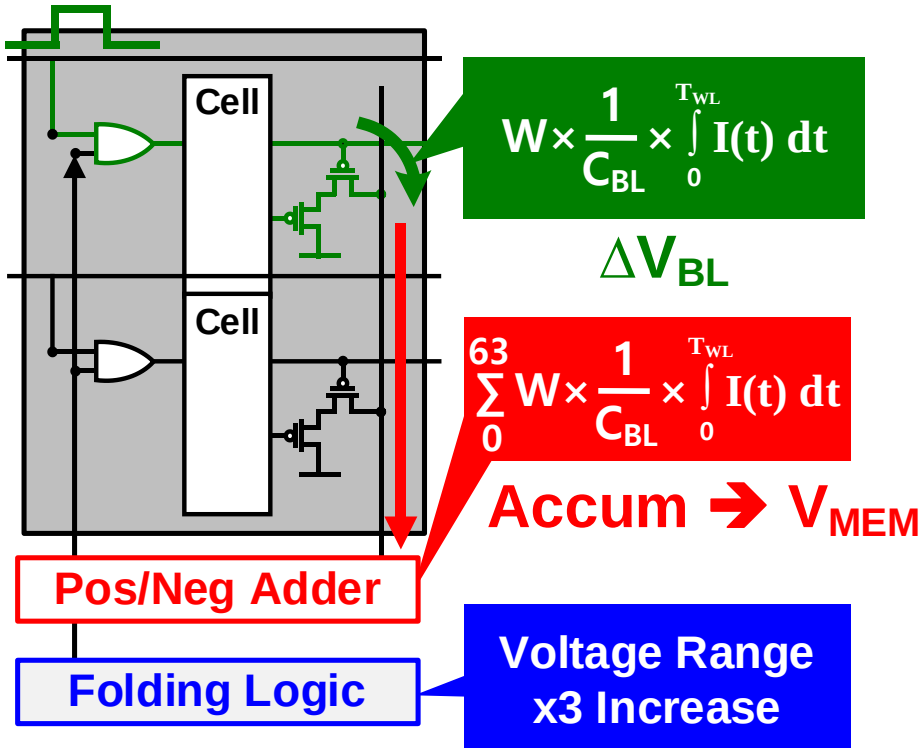


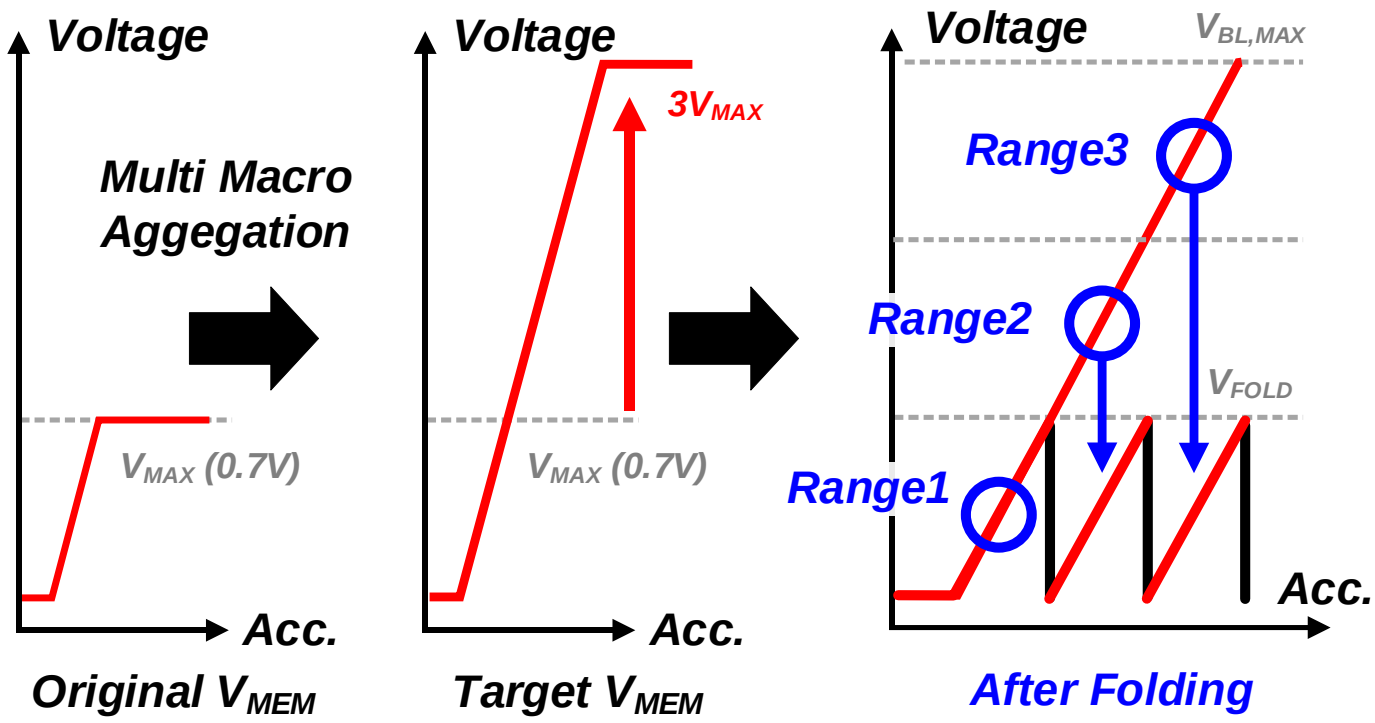




Motivation of Voltage Folding

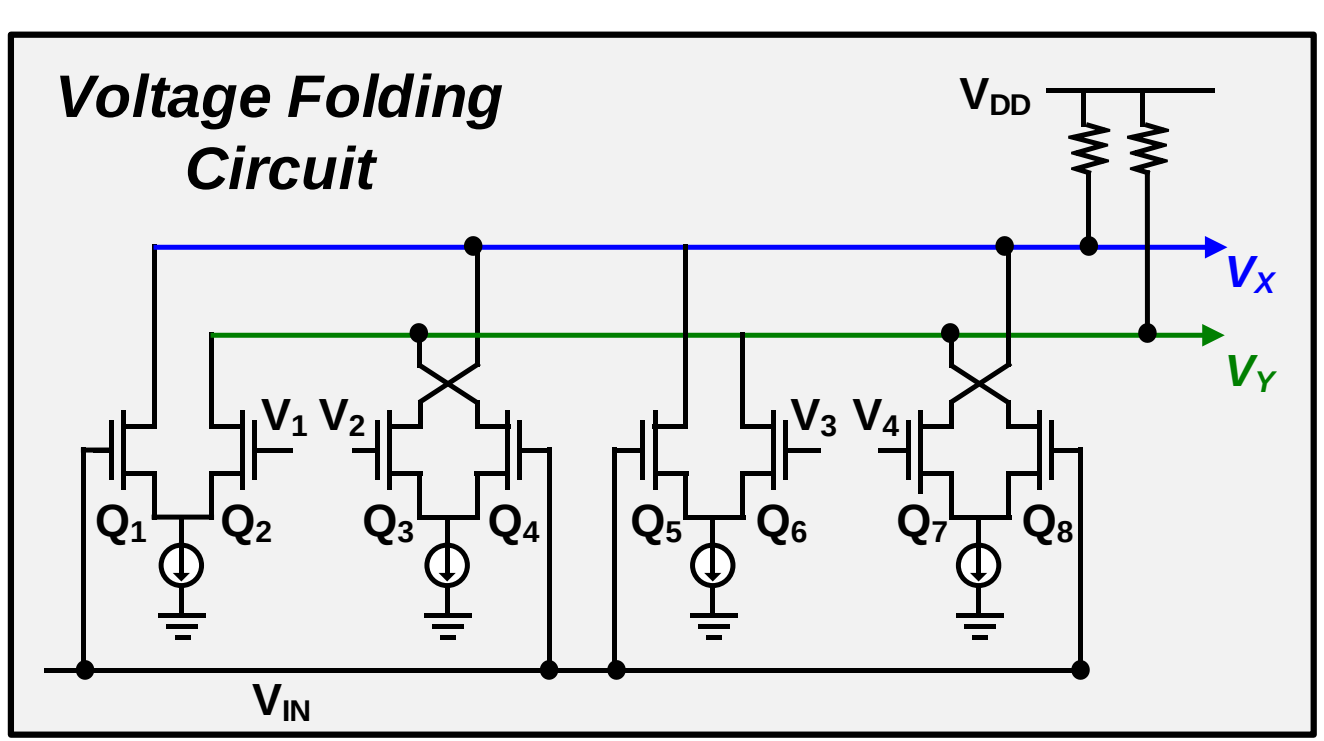
- **Virtually Increasing V_{LSB} of Membrane Voltage (V_{MEM})**
  - **By Voltage Folding V_{MEM} ➔ Folding Count + Residue Voltage**
  - **Amplifying the residue voltage ➔ Increasing the range and V_{LSB} of V_{MEM}**

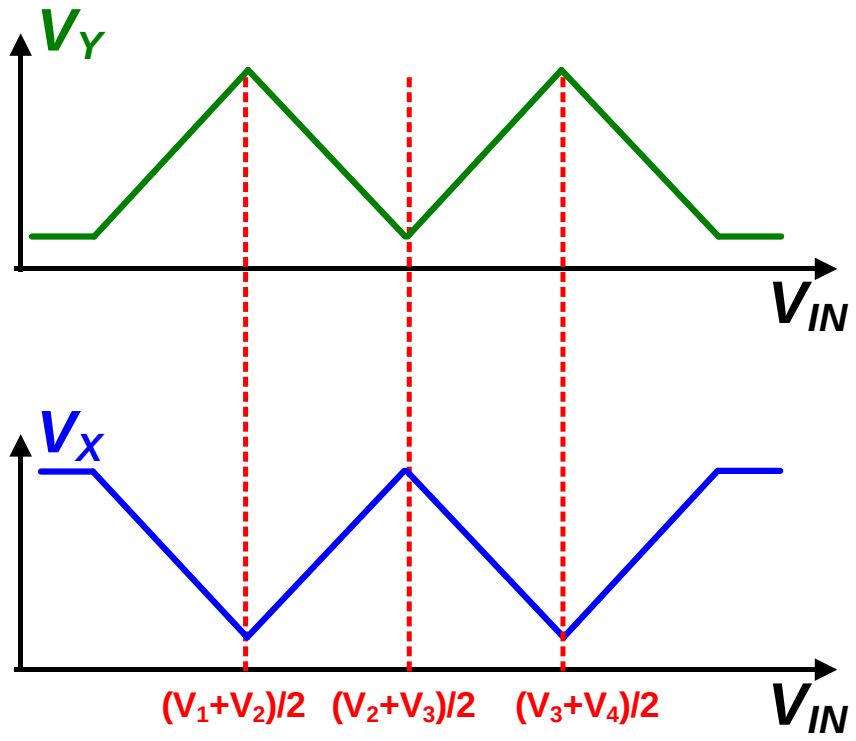




Voltage Folding

- **Virtually Increasing Range of Membrane Voltage (V_{MEM})**
  - **V_{IN} is folded @ $(V_1+V_2)/2$, $(V_2+V_3)/2$, $(V_3+V_4)/2$ …**
  - Generating two voltages (V_X and V_Y) with a phase difference of 180 degrees

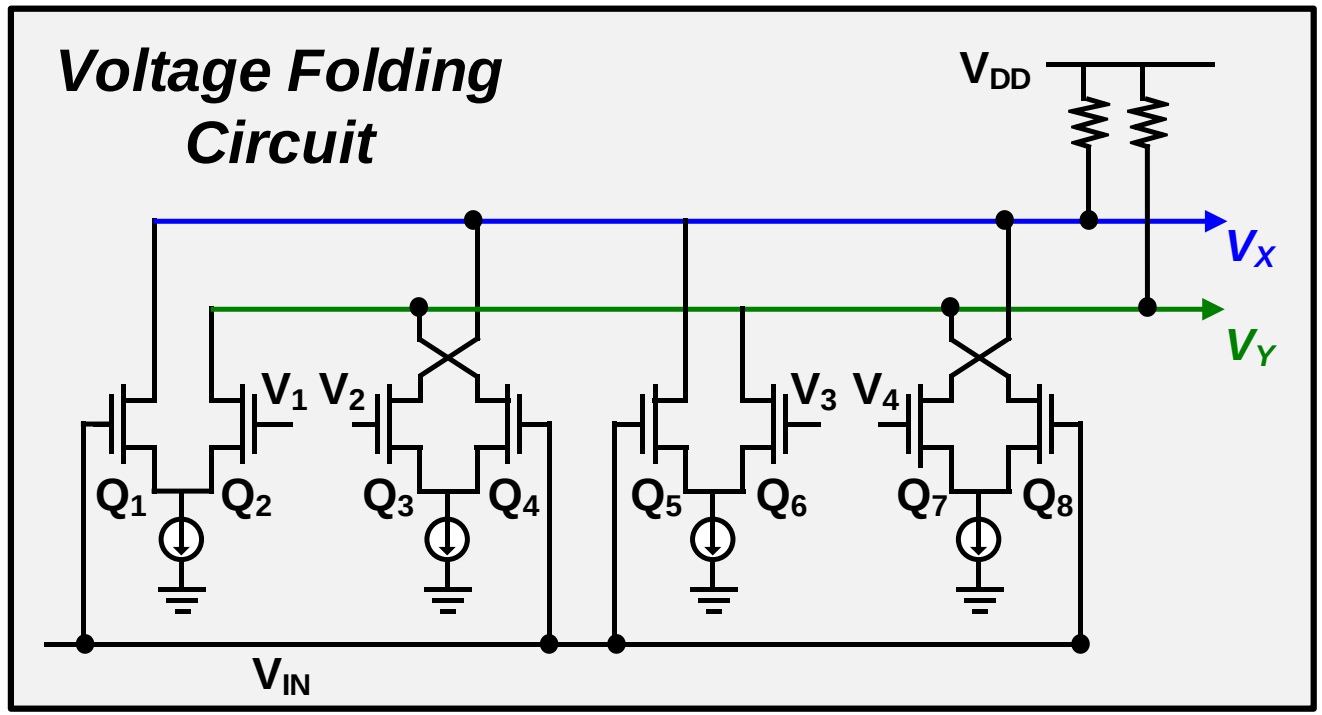


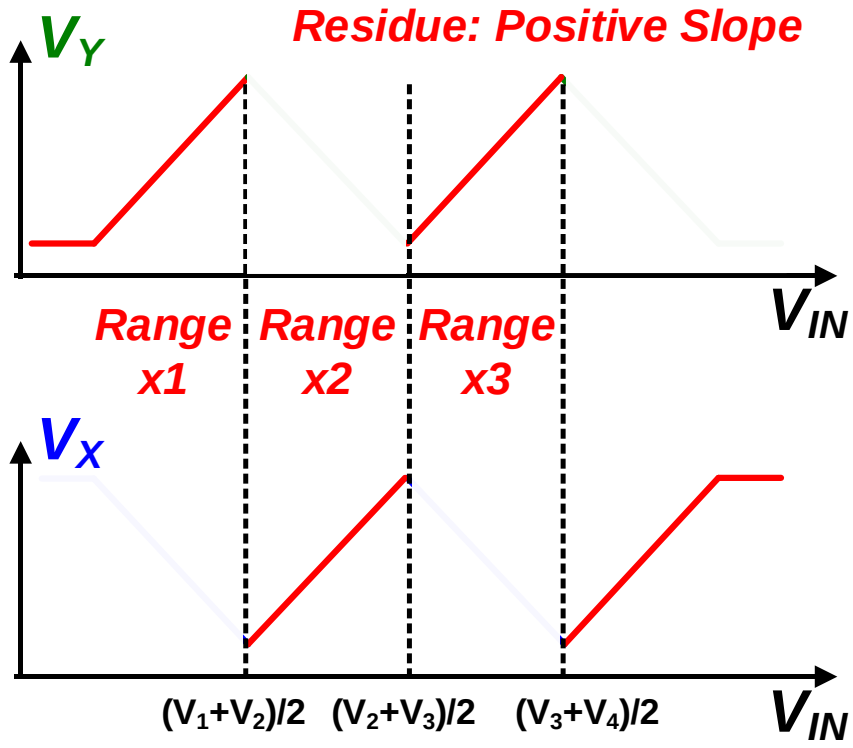


Voltage Folding

- **Virtually Increasing Range of Membrane Voltage (V_{MEM})**
  - Positive slope voltage selection between V_X and V_Y in Folding Circuit

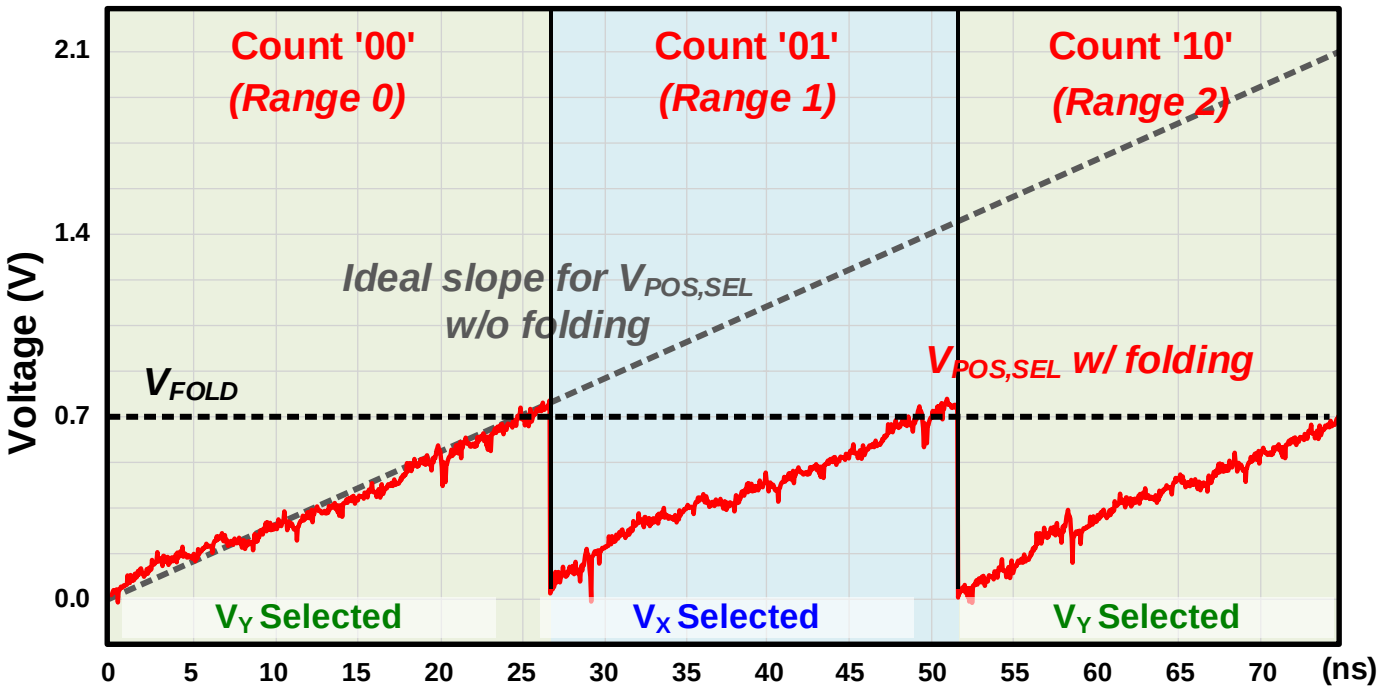
➔ **increasing range of V_{IN} continuously**



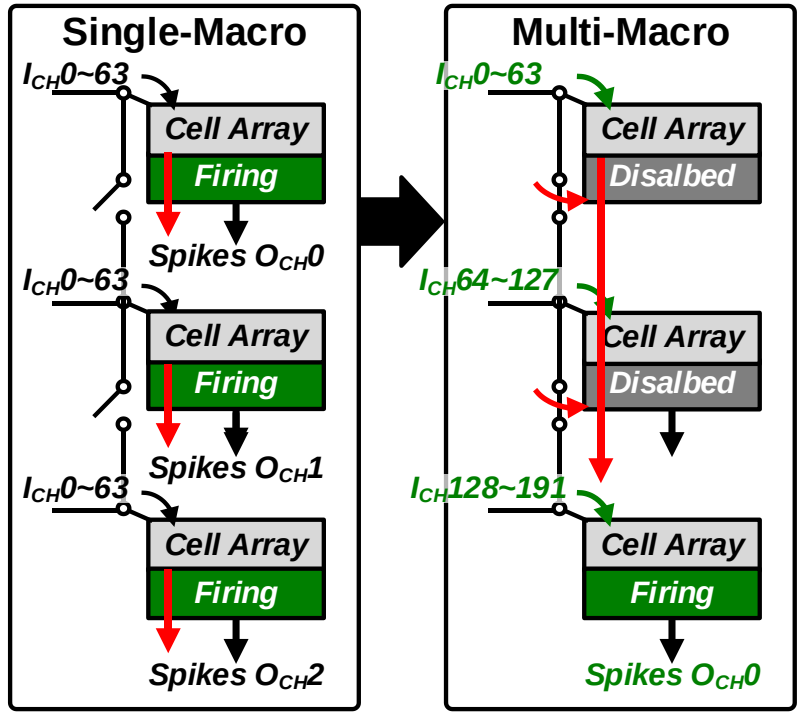


Measured Waveforms of Voltage Folding

- **Increasing the Range of Voltage and Generating Folding Count**
  - V_{MEM} ➔ **Analog** Folded output voltage (V_{SEL}) + **Digital** folding count
  - **High Virtual Range ➔ Multi-Macro Aggregation w/o High Precision ADC**



<Measured Waveform>



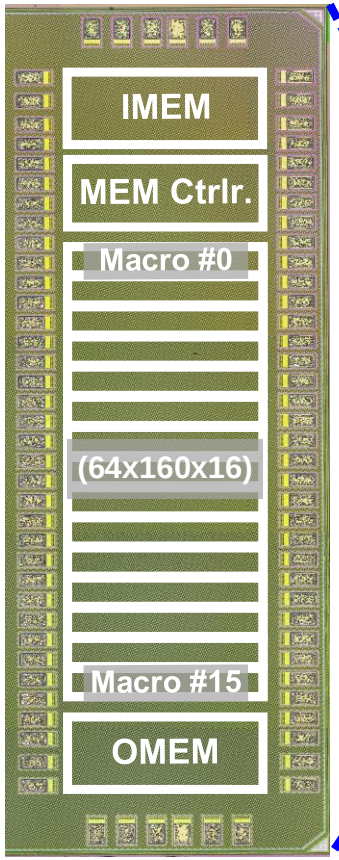
High Reconfigurability

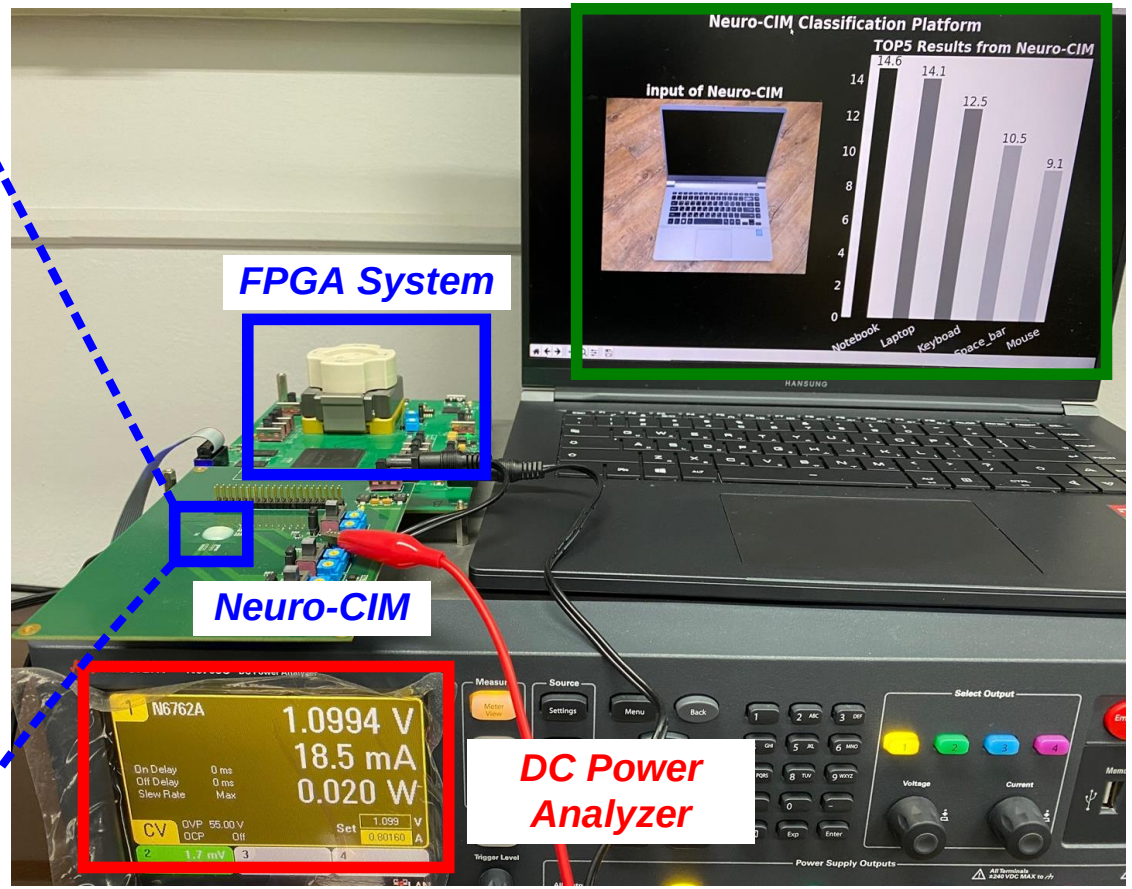
Chip Summary

- **Chip Photograph and Performance**

Technology	28nm
Die Area	3228μm x 900μm
Single Macro Area	0.048
Total CIM Storage	32 KB
Digital SRAM	32 KB
Supply	1.1V
Frequency	200MHz
Macro Power (mW)	15.8[1)] – 36.2[2)]
System Power (mW)	105.4[1)] – 241.4[2)]
System Energy Efficiency (TOPS/W)	I=4b, W=1b: 310.4[3)] I=4b, W=4b: 124.2[3)] I=4b, W=8b: 62.1[3)]

1) w/ MWS & ES 2) w/o MSB & ES
3) CIFAR-10 with ResNet-18





Conclusion

- **Neuro-CIM: An Energy-Efficient Neuromorphic CIM+SNN Processor**
  - **CIM: Reducing memory access and multiple WL driving**
  - **SNN: Generating input sparsity and eliminating high-precision ADC**

- **For Energy-Efficient Neuromorphic CIM Processing**
  - **MSB Word Skipping ➔ Reducing 25~38% power consumption**
  - **Early Stopping ➔ Reducing 37% power consumption**
  - **Mixed-mode Neuron Firing ➔ Increasing the voltage range x3**

A 310.4 TOPS/W 1034.6 TOPS/W Bank Neuromorphic CIM Processor for Energy Efficient Neural Network Processing

Thank You!

- **Questions? Feel Free to Contact Me!**
  - E-mail: sangyeob.kim@kaist.ac.kr
  - LinkedIn: https://www.linkedin.com/in/sangyeob-kim-871818179/
  - Zoom Meeting: https://us05web.zoom.us/j/3753663353?pwd=dzNlMFh3M0pIeWJiM0dVZmxLNVVoUT09

- **Acknowledgement**
  - This work was supported by the Samsung Electronics. (IO201207-07799-01)