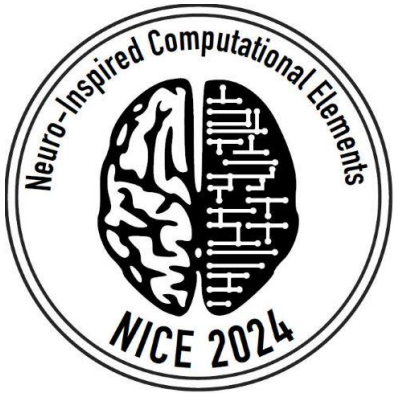


Quantized Context Based LIF Neurons for Recurrent Spiking Neural Networks in 45nm

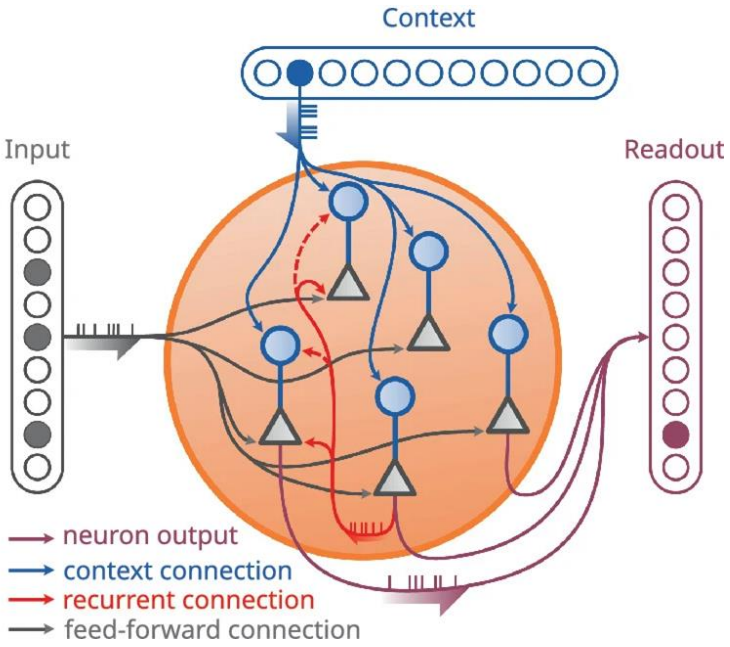
Sai Sukruth Bezugam*, Yihao Wu*, JaeBum Yoo*,
Dmitri Strukov, Bongjin Kim
Department of ECE
University of California Santa Barbara
saisukruthbezugam@ieee.org
(* Contributed Equally)

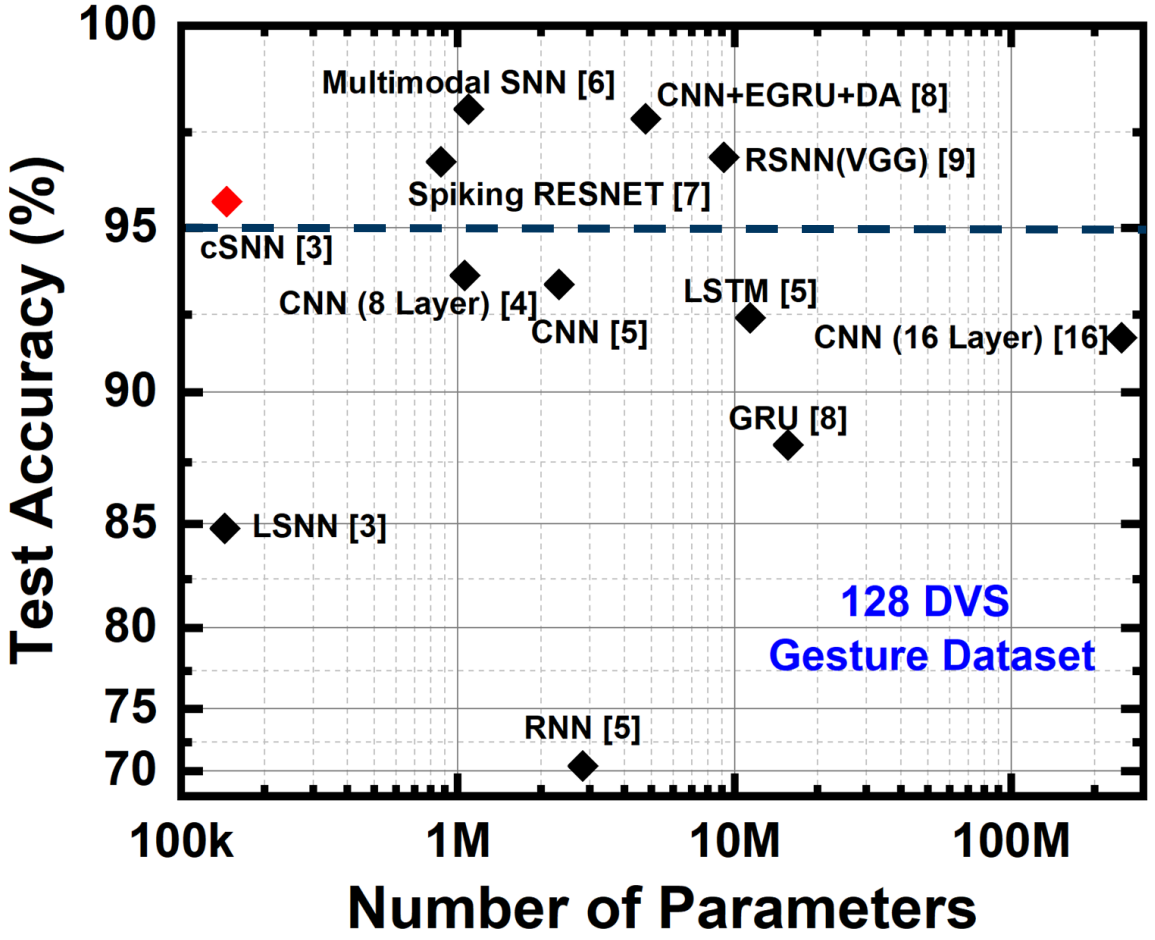
Without Context
Find something?

With Context
Find something **<u>to play music with.</u>**

Basic Idea : Context helps in finding things easier.

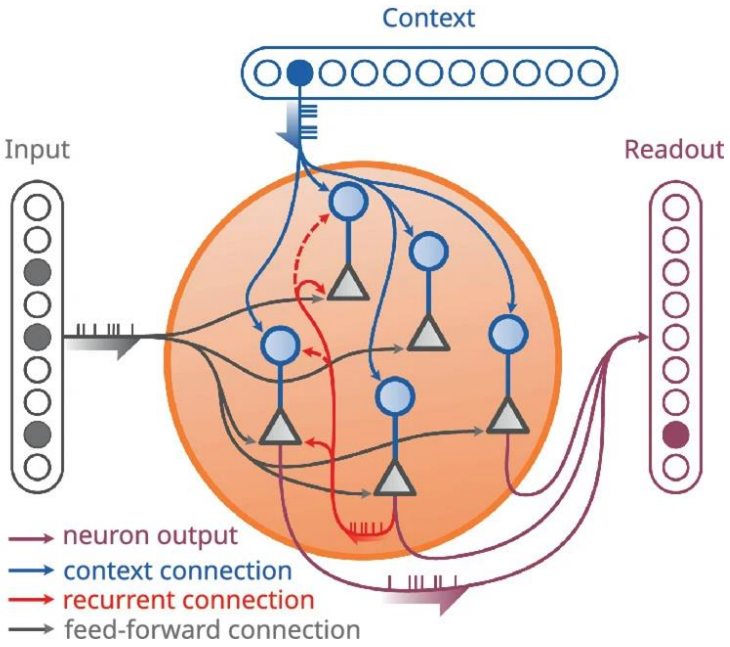
Is it with the SNNs too?



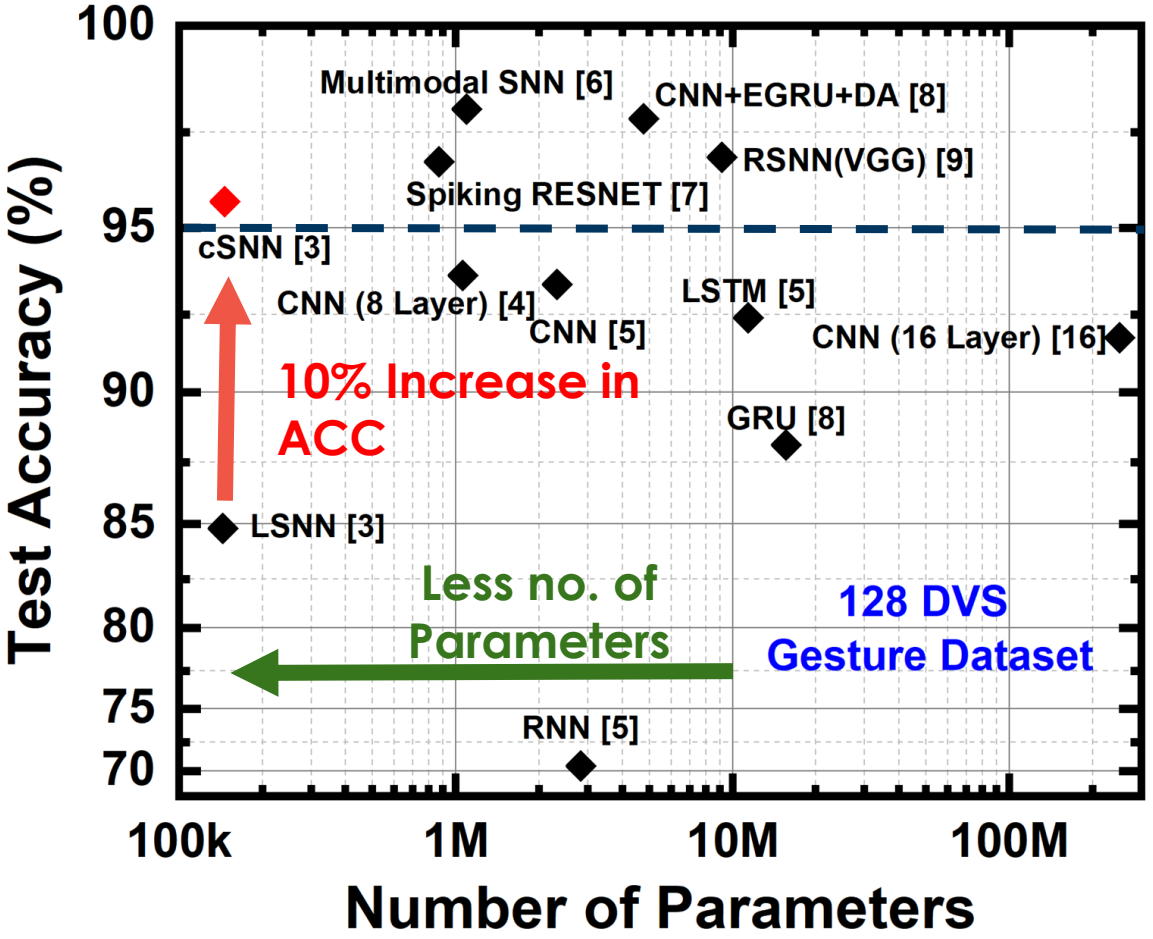


Ferrand, Romain, et al. "Context-dependent computations in spiking neural networks with apical modulation." International Conference on Artificial Neural Networks. Cham: Springer Nature Switzerland, 2023.

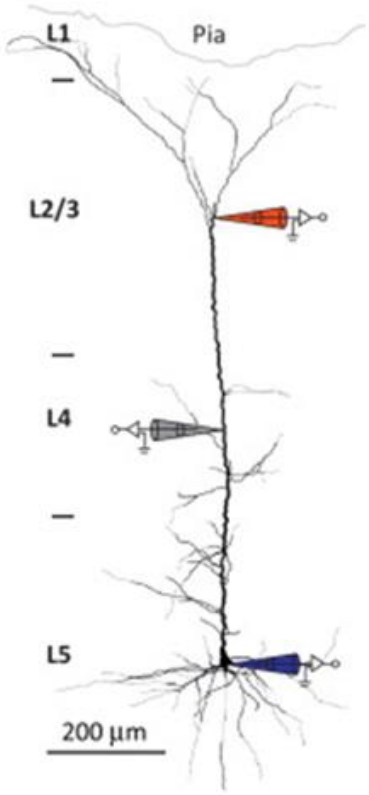
Is it with the SNNs too? - Yes



Ferrand, Romain, et al. "Context-dependent computations in spiking neural networks with apical modulation." International Conference on Artificial Neural Networks. Cham: Springer Nature Switzerland, 2023.

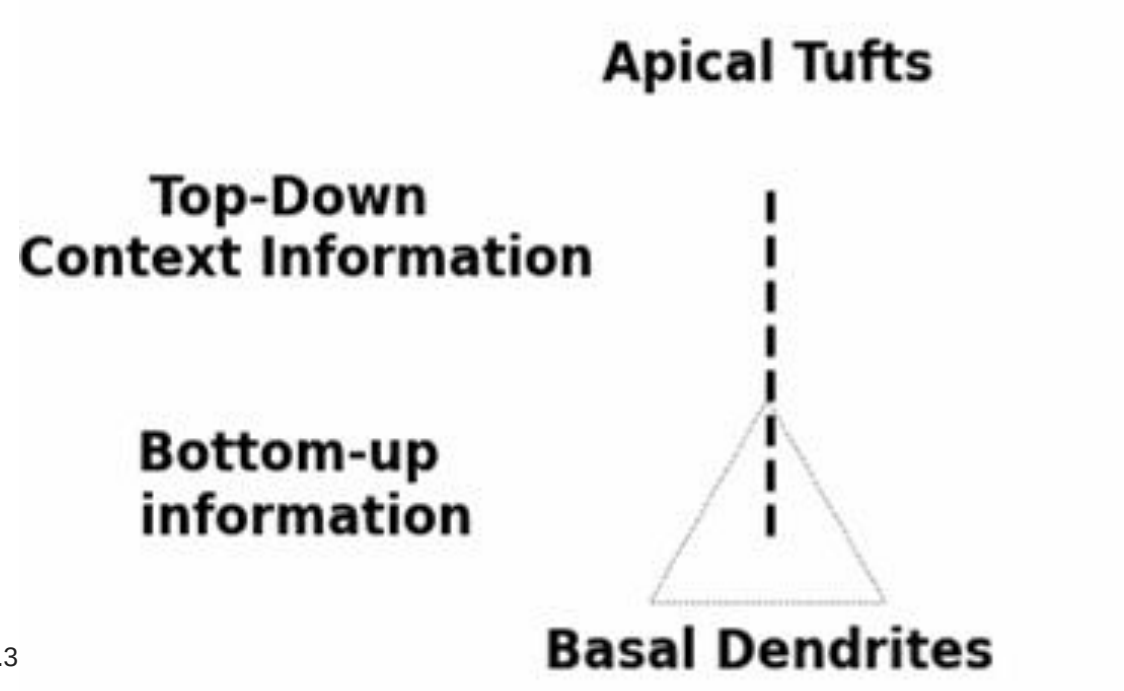


How does it work ?



Larkum, Matthew. "A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex." *Trends in neurosciences* 36.3 (2013): 141-151.

Inspired by dual integration of information using 2 compartment Neocortical Pyramidal Neuron



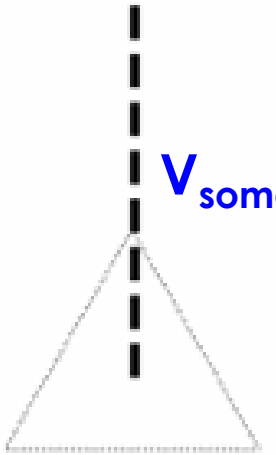
Context Based LIF

Mathematics of Context based Leaky Integrate and Fire Neuron Model

Apical Compartment

Context Input

$I_{context}(t+dt)$



$I_{som}(t+dt)$

Stimuli Input

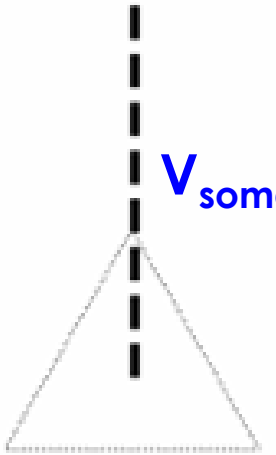
Somatic Compartment

$$V_{ap}(t+dt) = \alpha_{leak} * V_{ap}(t) + (1-\alpha) * I_{context}(t+dt)$$

$$V_{soma}(t+dt) = \beta_{leak} * V_{soma}(t) + (1-\beta_{leak}) * \text{Relu}\left[V_{ap}(t+dt) * I_{som}(t+dt)\right] - S_j(t) V_{th}$$

$$\text{if } V_{soma}(t+dt) > V_{th} \quad S_j(t) = 1$$

Mathematics of Context based Leaky Integrate and Fire Neuron Model



Apical Compartment

Context Input

$I_{context}(t+dt)$

$I_{som}(t+dt)$

Stimuli Input

Somatic Compartment

$$V_{ap}(t+dt) = \alpha_{leak} * V_{ap}(t) + (1-\alpha) * I_{context}(t+dt)$$

$$V_{soma}(t+dt) = \beta_{leak} * V_{soma}(t) + (1-\beta_{leak}) * Relu [V_{ap}(t+dt) * I_{som}(t+dt)] - S_j(t) V_{th}$$

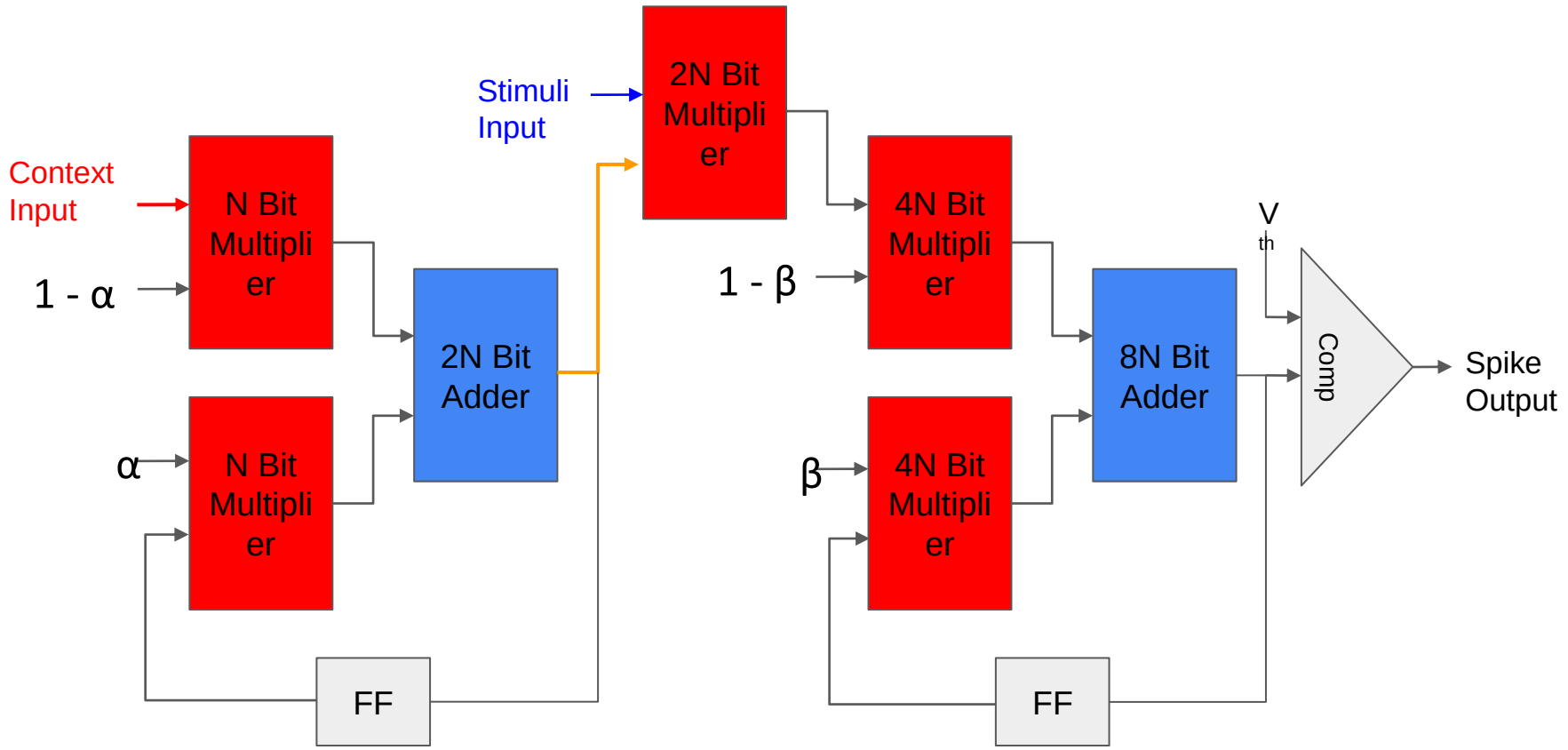
$$\text{if } V_{soma}(t+dt) > V_{th} \quad S_j(t) = 1$$

Amplification/ Attenuation of somatic input based on context/ Apical compartment Voltage

Lets build hardware for it
CLIF Digital design (No Brainer)

$$V_{ap}(t+dt) = \alpha_{leak} * V_{ap}(t) + (1-\alpha) * I_{context}(t+dt)$$

$$V_{soma}(t+dt) = \beta_{leak} * V_{soma}(t) + (1-\beta_{leak}) * [V_{ap}(t+dt) * I_{som}(t+dt)] - S_j(t)V_{th}$$

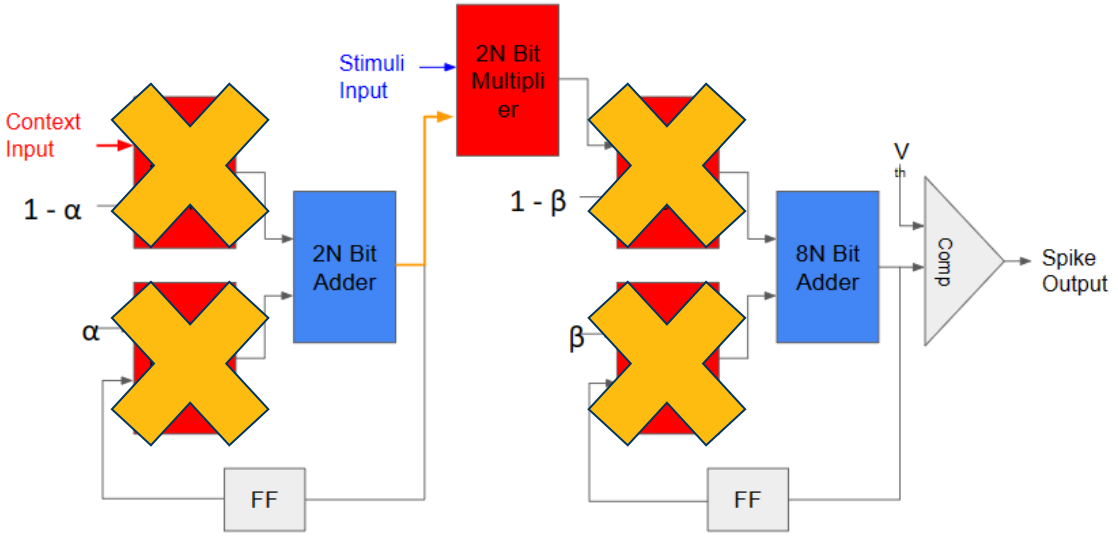


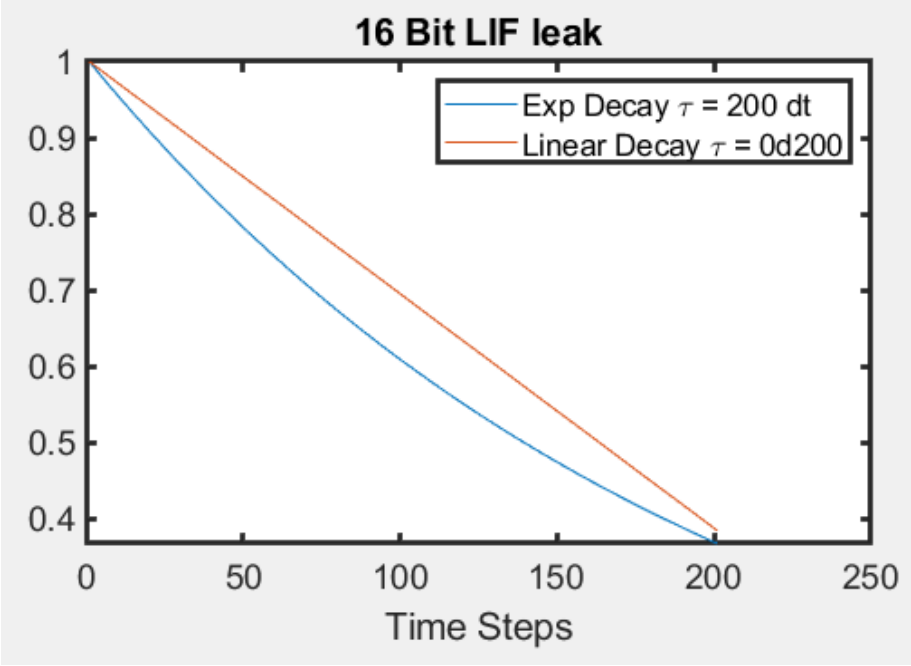
If N starts at 8 ends at 64 bits

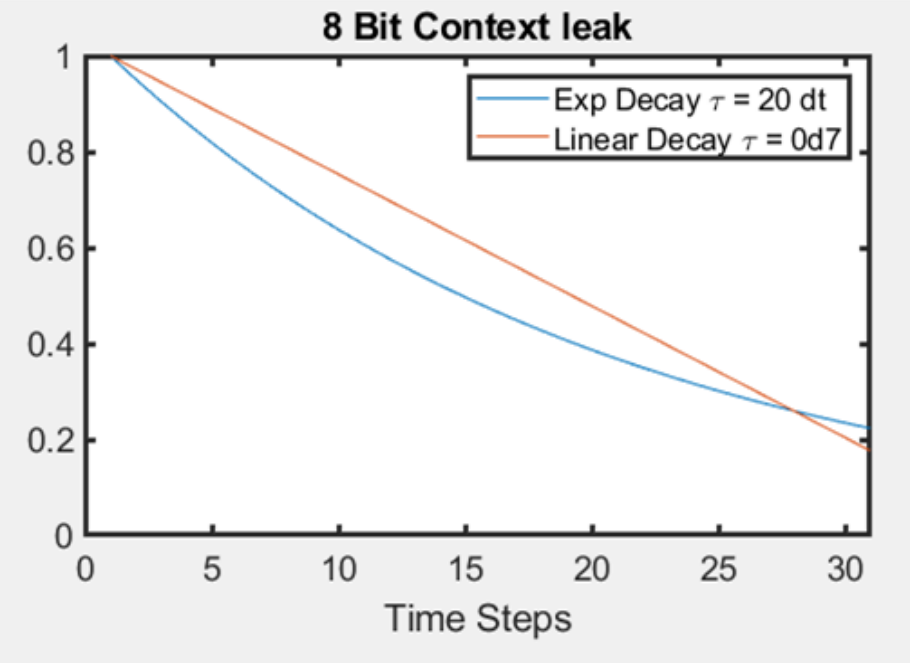
N -> 8N bits

Proposed qCLIF

(a)Linear Decay



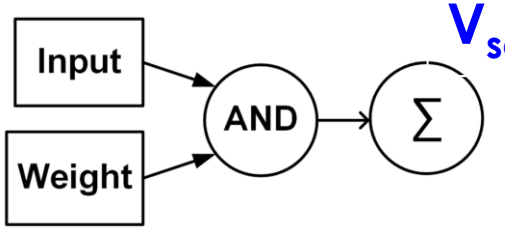




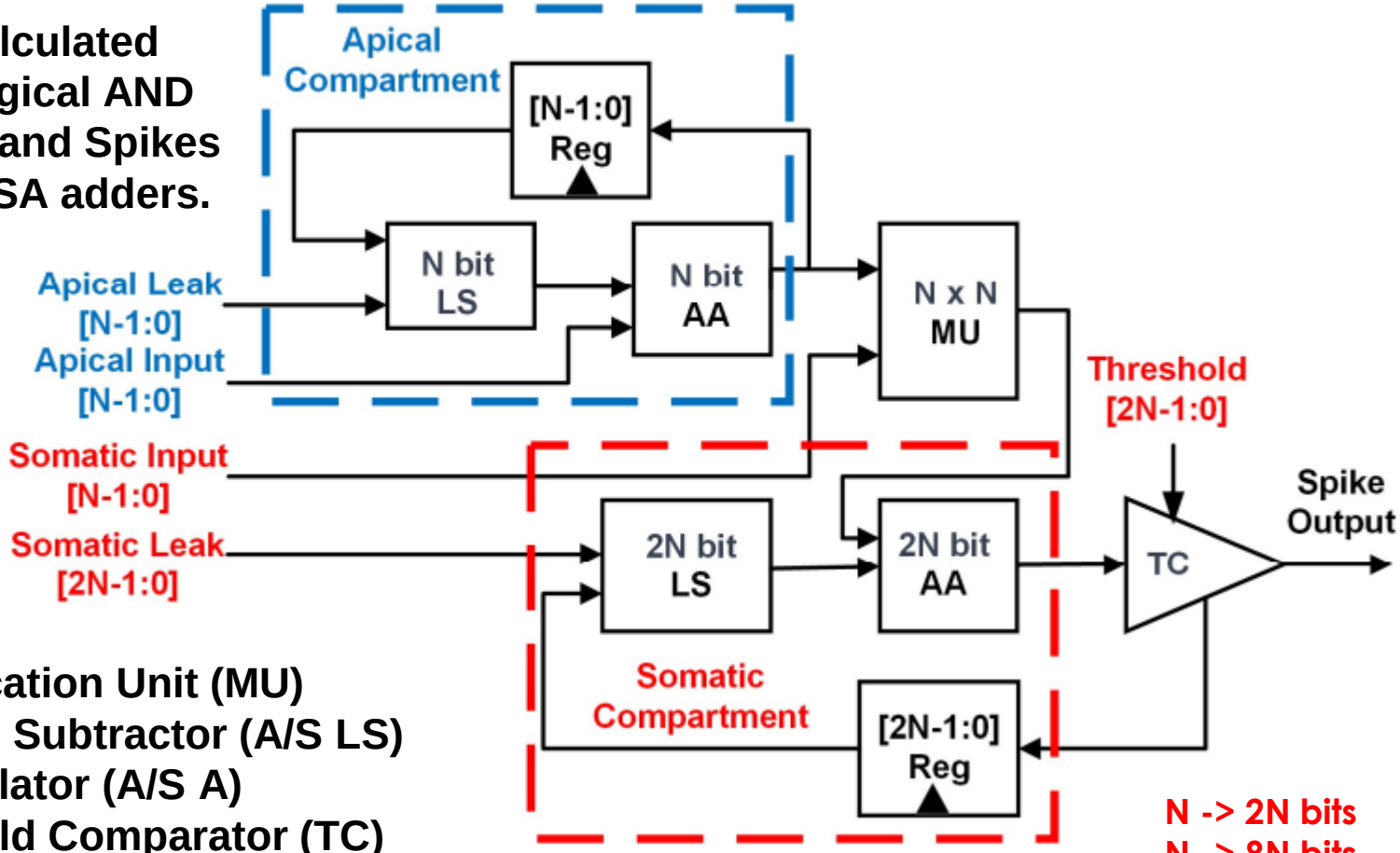
Proposed qCLIF

$$V_{ap}(t+dt) = V_{ap}(t) - a_{leak} + I_{context}(t+dt)$$

$$V_{soma}(t+dt) = V_{soma}(t) - \beta_{leak} + [V_{ap}(t+dt) * I_{som}(t+dt)]$$



SWM - Current Calculated through simple logical AND between Weights and Spikes and added with CSA adders.

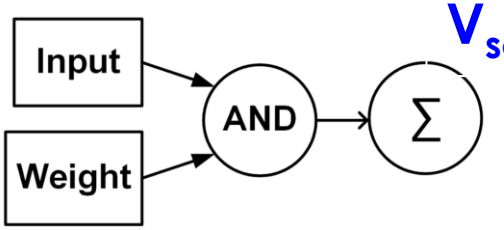


Multiplication Unit (MU)
Leakage Subtractor (A/S LS)
Accumulator (A/S A)
Threshold Comparator (TC)

Proposed qCLIF

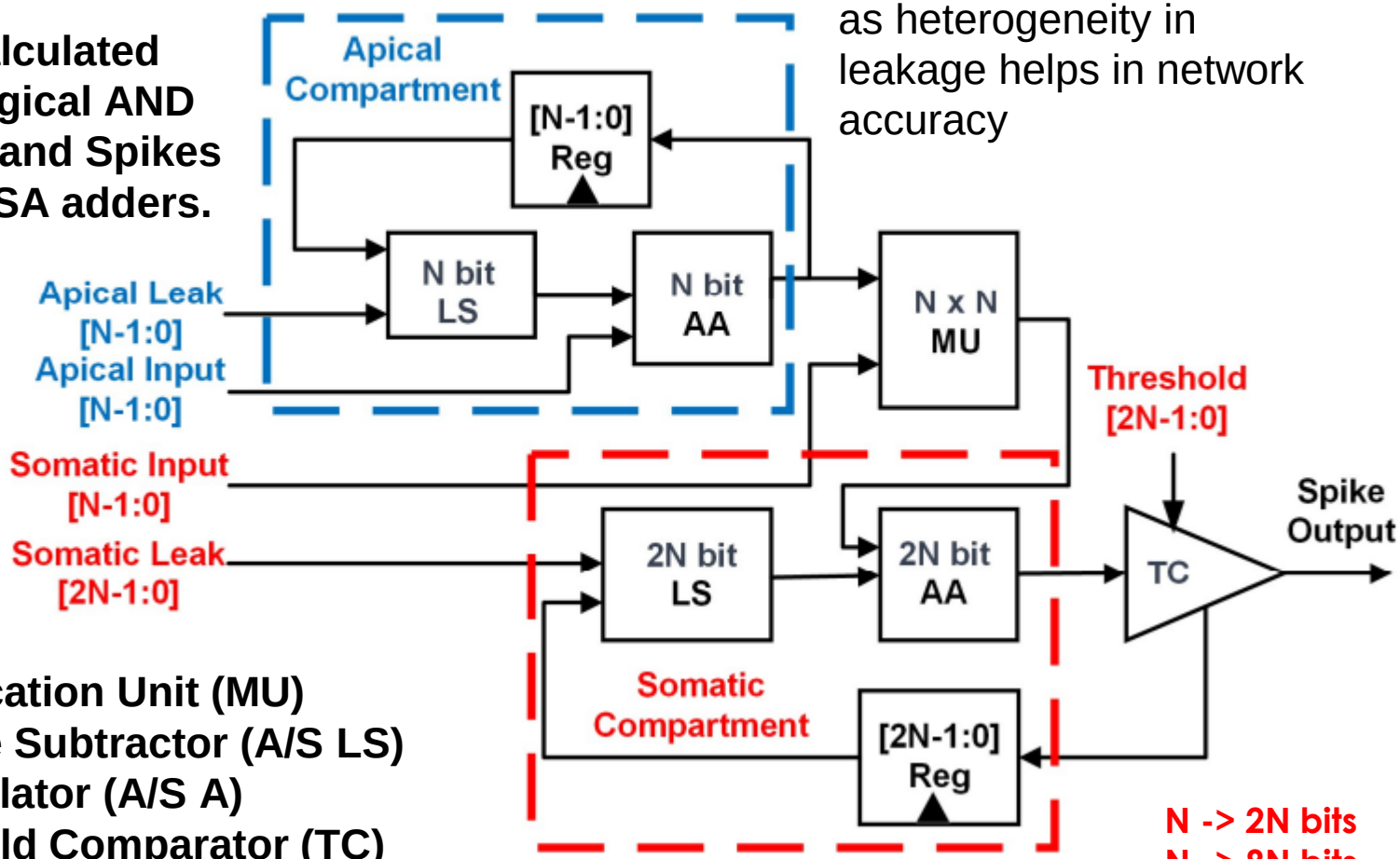
$$V_{ap}(t+dt) = V_{ap}(t) - a_{leak} + I_{context}(t+dt)$$

$$V_{soma}(t+dt) = V_{soma}(t) - \beta_{leak} + [V_{ap}(t+dt) * I_{som}(t+dt)]$$



SWM - Current Calculated through simple logical AND between Weights and Spikes and added with CSA adders.

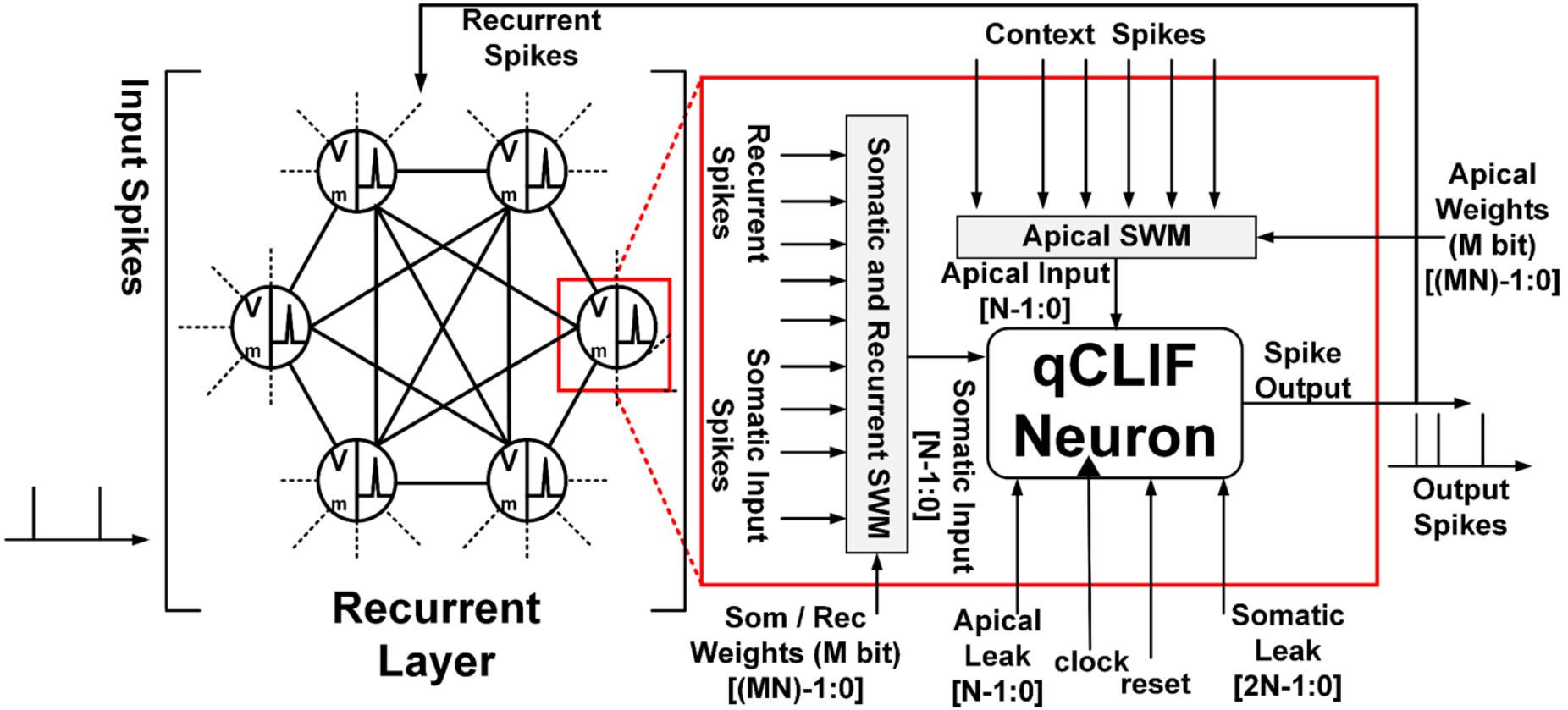
Not freezing the leakage as heterogeneity in leakage helps in network accuracy



Multiplication Unit (MU)
Leakage Subtractor (A/S LS)
Accumulator (A/S A)
Threshold Comparator (TC)

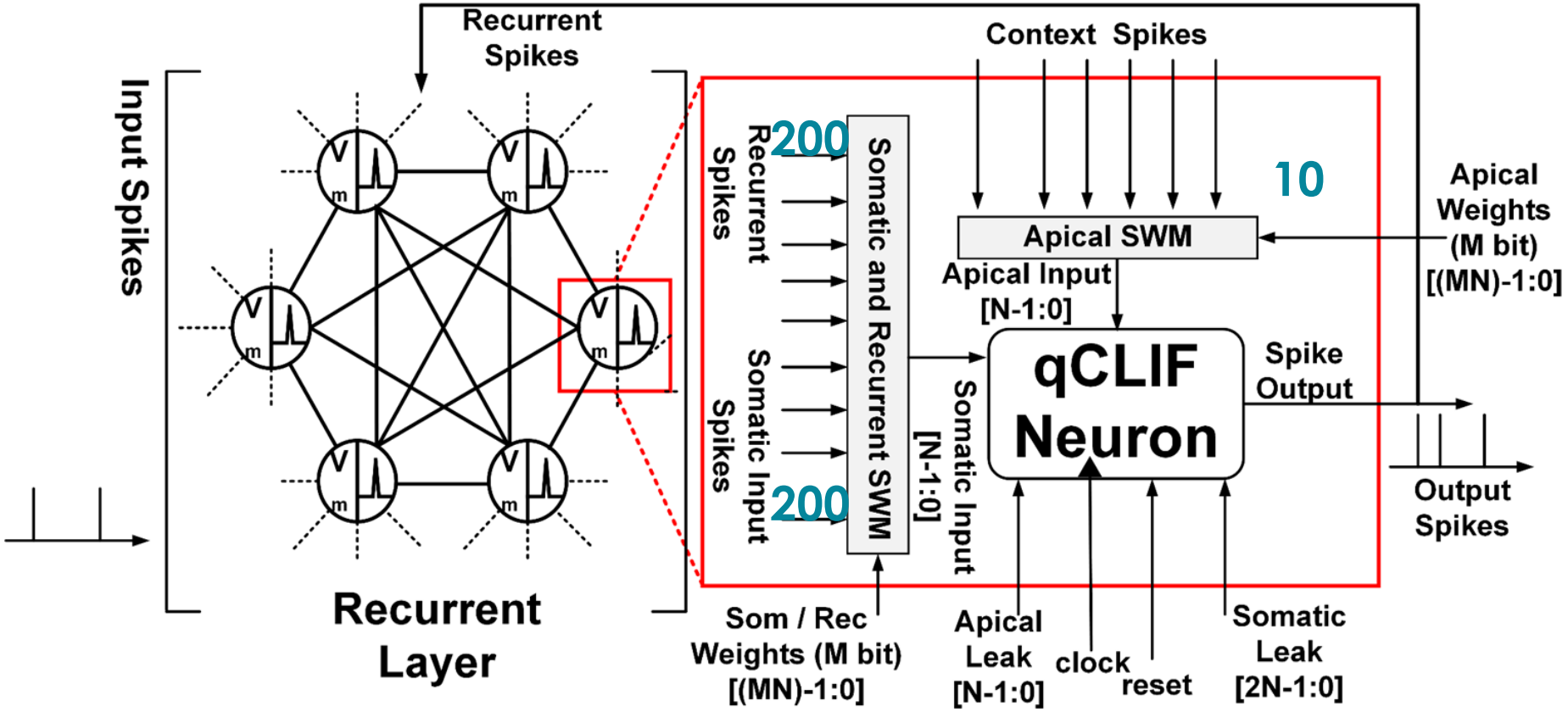
N -> 2N bits
N -> 8N bits
(Orignal)

Architecture of qCLIF Neuron layer



What should be the bit width of modules?

Targeted network size
(S) → 200, (Rec) → 200, (Context) → 10

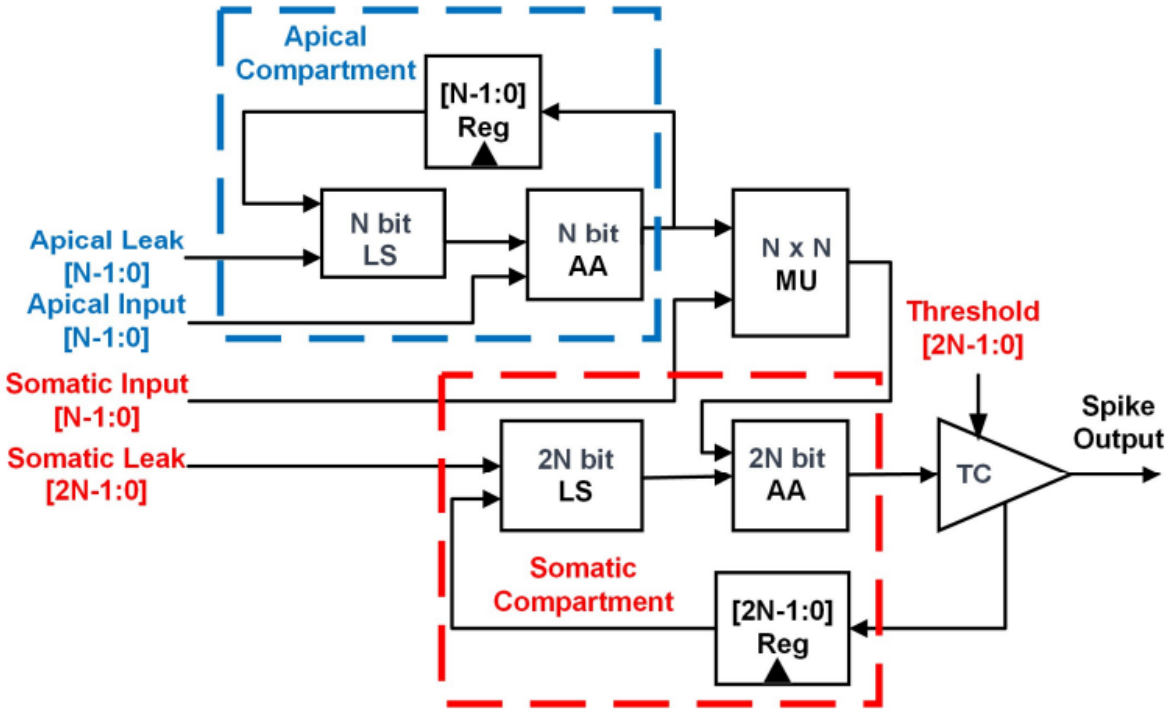


What should be the bit width of modules?

Assume K bit weights

Somatic Inputs worst case → 400*K

→ N → log2(400)*K → ~64bits

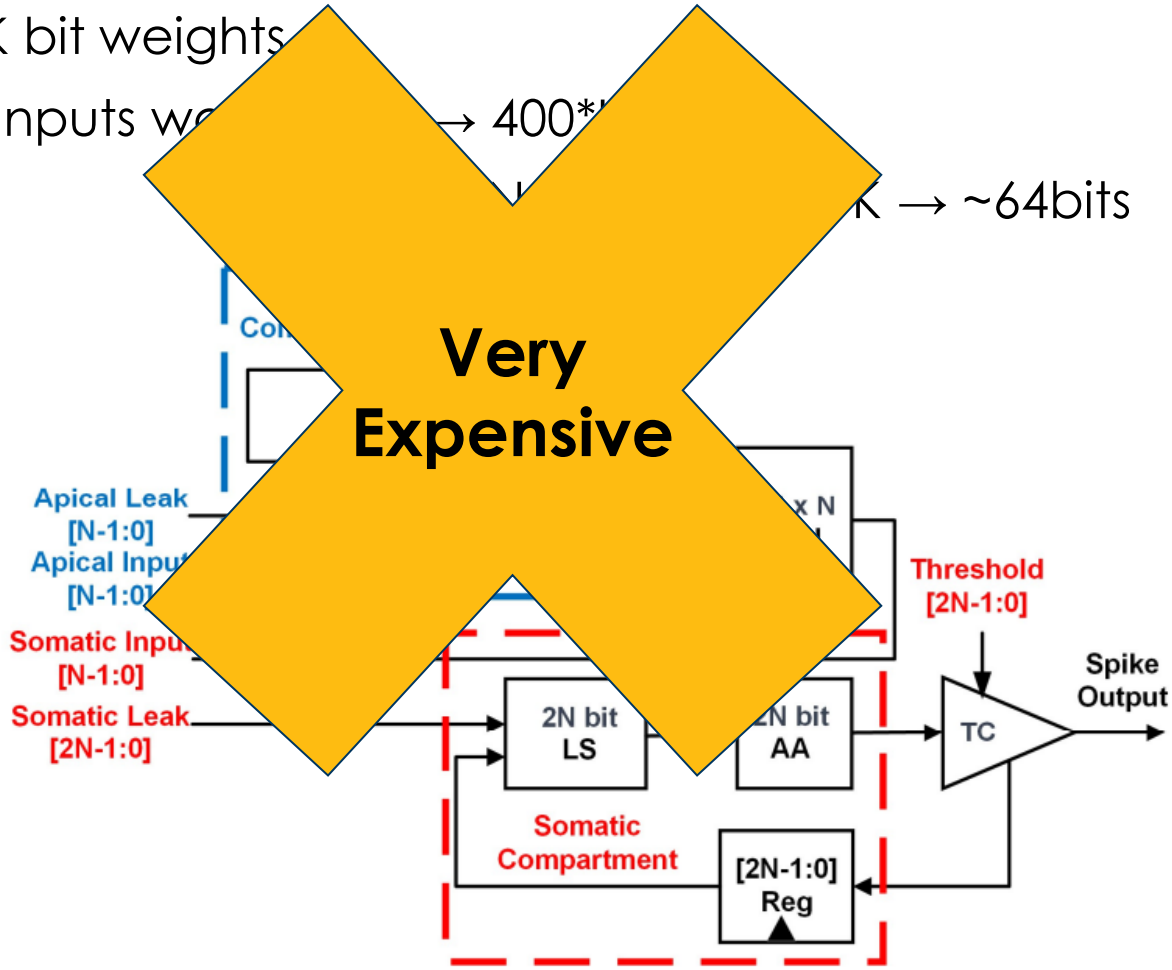


What should be the bit width of modules?

Assume K bit weights

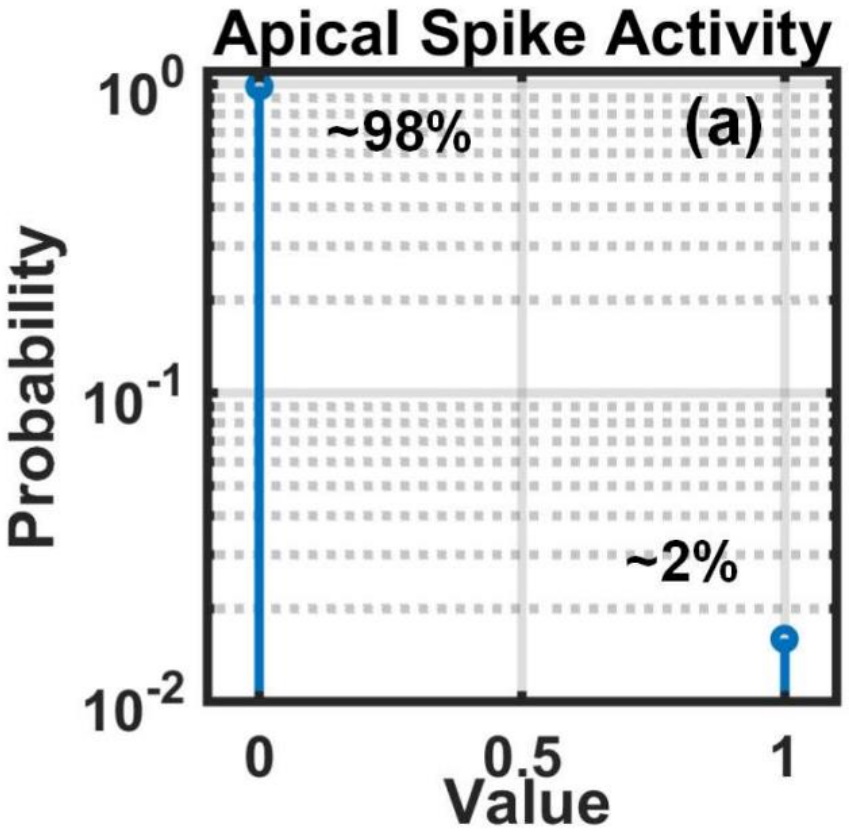
Somatic Inputs wo → 400*

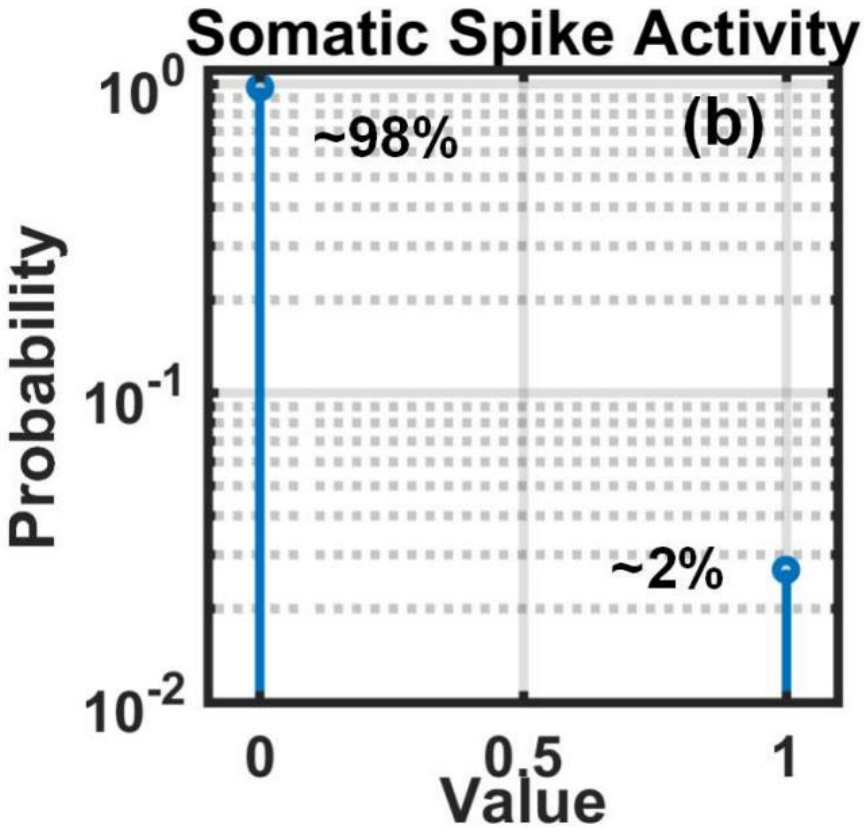
K → ~64bits



What should be the bit width of modules?

Thanks to sparse activity of SNN

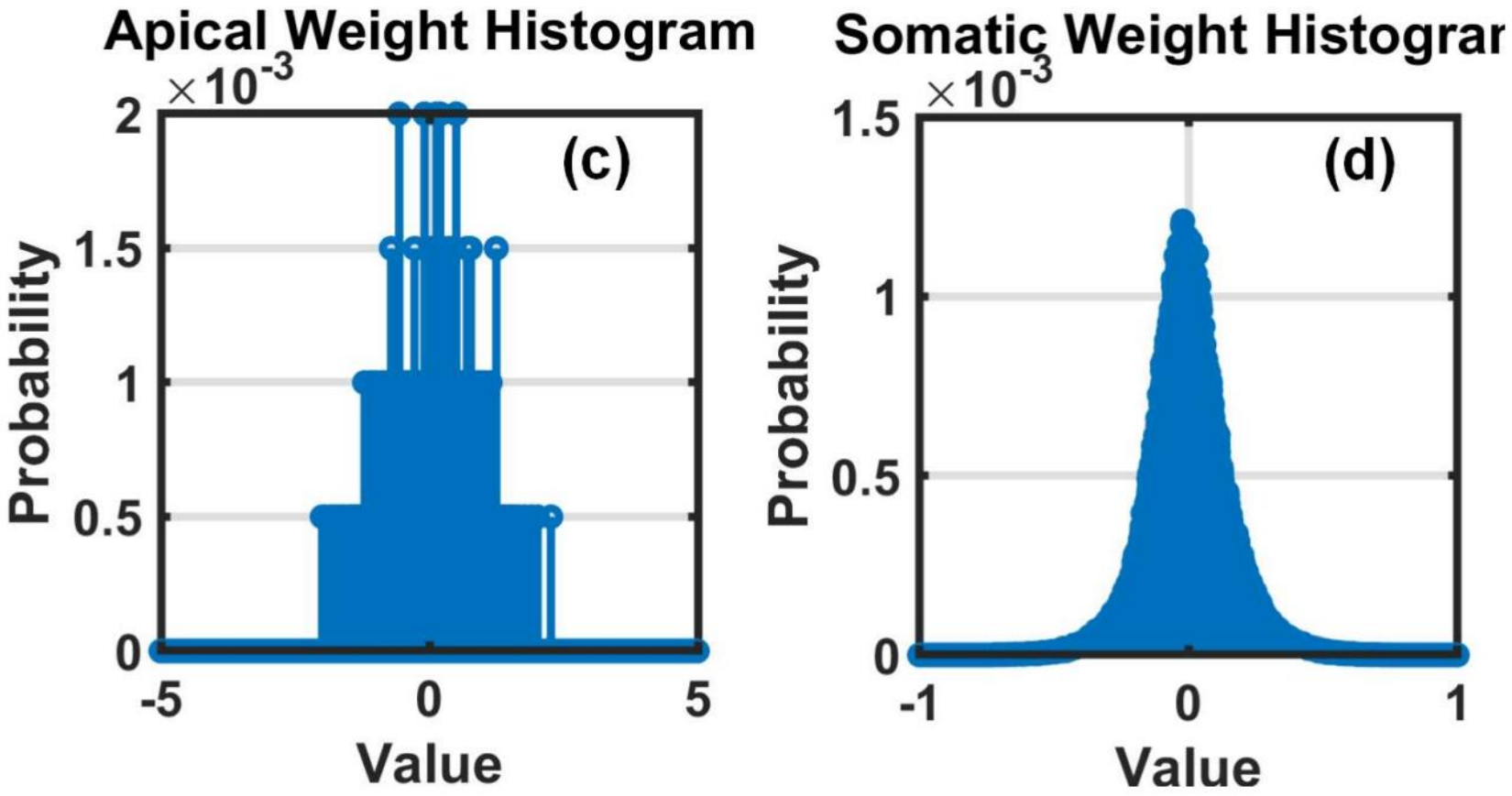




Hence
4 bits weights → N = 4
8 bits weights → N = 8

Quantization of weights

Normal distribution (Region bounded quantization)



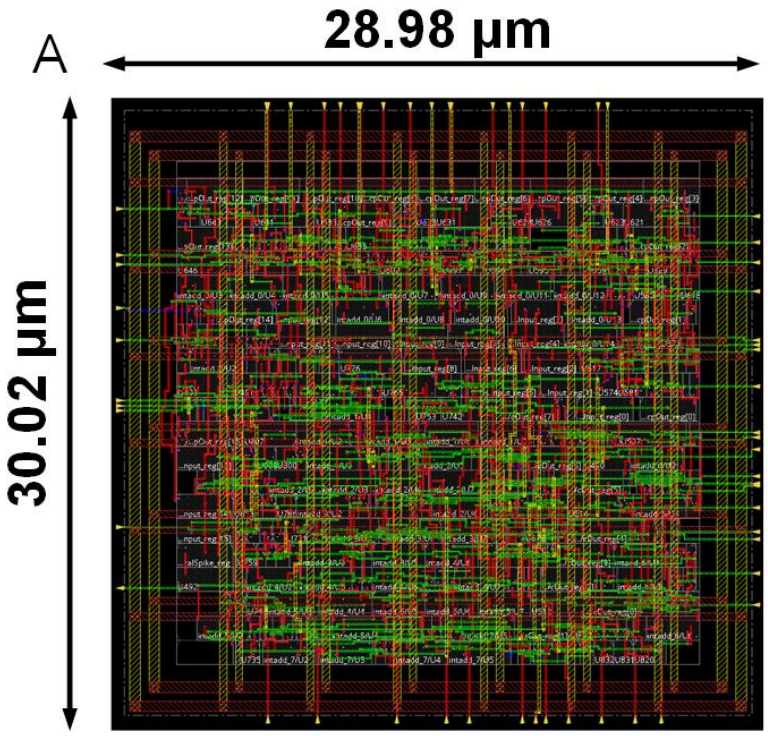
EFFECT OF QUANTIZATION ON NETWORK PERFORMANCE

Quantization Aware Training done for both neuron states and weights

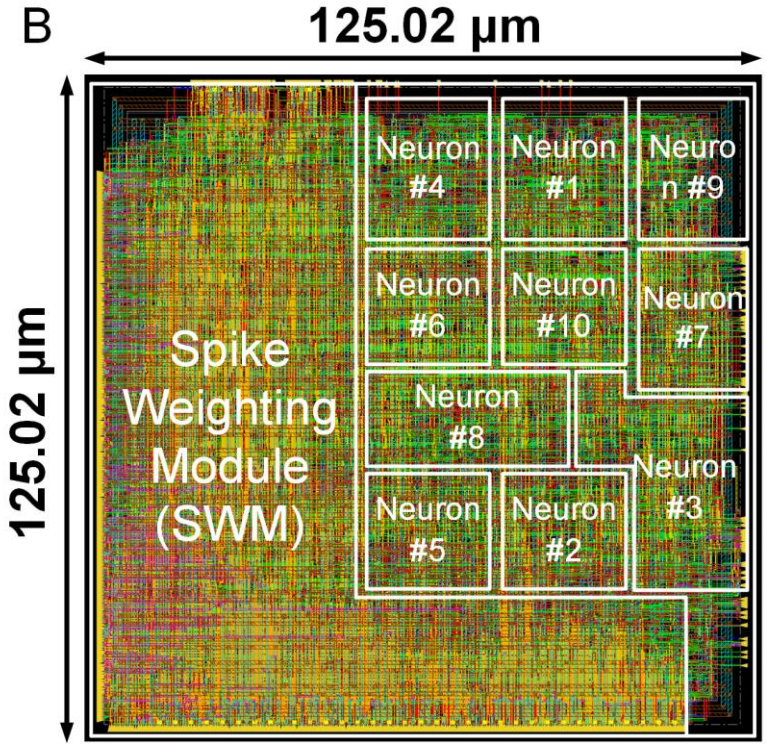
Precision Level	Neuron Quantization Accurancy (%)	Weight and Neuron Quantization Accuracy (%)
Full Precision	94.5	94.5
16-bit	93.4	93
8-bit	92	90
4-bit	77.5	73
2-bit	55	N/A

* All results on proposed qCLIF

Hardware results - Layout 45nm FreePDK



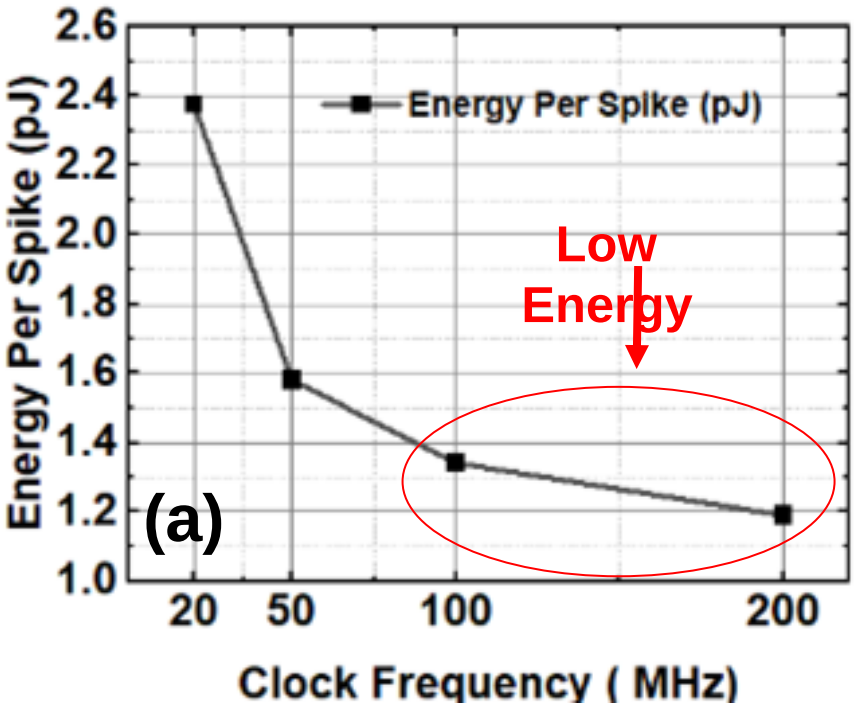
Single qCLIF



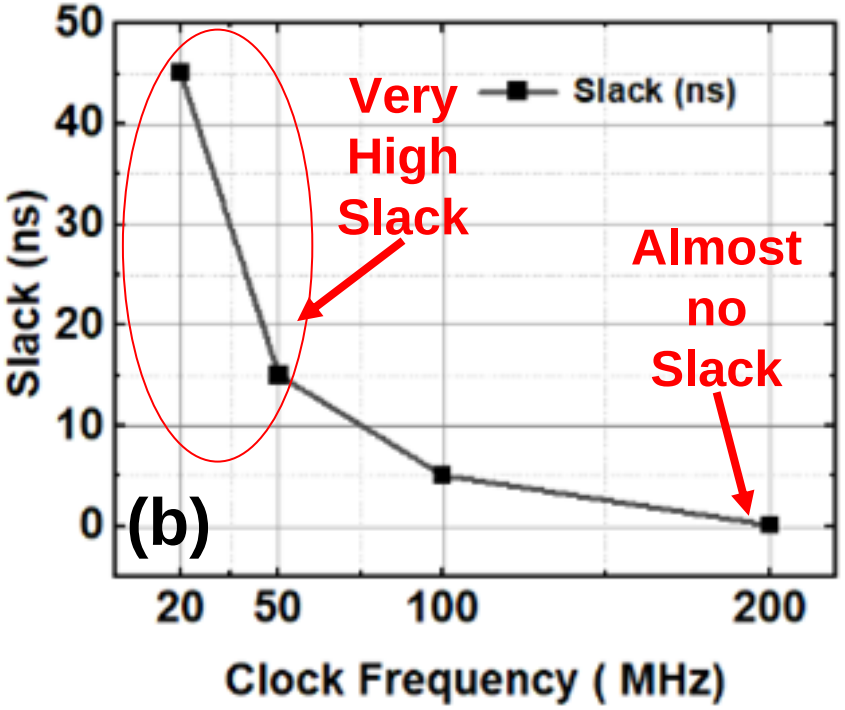
10 qCLIF neurons network

Hardware results -Performance of Layer of 10 qCLIF neurons

ESOP vs Freq.



Timing vs Freq.



Synapses	250, 8bit
Area (mm^2)	0.125*0.125

Hardware results - Scalability of Design

Clock Frequency	No. of qCLIF	Synapses	Area (mm^2)	Slack (ns)	Total Power (mW)	Energy Per Spike (pJ)
100 MHz	**10**	250, 8bit	0.125*0.125	5.10	1.315	1.342
100 MHz	**200**	82K, 8bit	1.925*1.925	4.07	358.0	17.9

20X Neurons and 328x Synapses but 15.4X increase in area

Sublinear increase in Total Power Consumption

Hardware results - Scalability of Design (Precision)

Clock Frequency	No. of qCLIF	Synapses	Area (mm^2)	Slack (ns)	Total Power (mW)	Energy Per Spike (pJ)
100 MHz	200	82K, 8bit	1.925*1.925	4.07	358.0	17.9
100 MHz	200	82K, 4bit	1.365*1.365	6.45	174.0	8.7

Energy reduction > 50%

Slack increased by 2.4 ns -> Lower precision may even operate @ 200 MHz

Comparison with literature

	[18]	[19]	[20]	[21]	[22]	This work	This work
	Fabricated	Fabricated	Fabricated	Fabricated	Fabricated	Simulated	Simulated
Technology (nm)	65	90	65	10	28	45	45
Neuron count	650	400	410	4096	1M	200	200
Network Type	FF SNN	FF SNN	SNN	FF SNN	FF SNN	cRSNN	cRSNN
Neuron Type	IF	Stochastic	IF	LIF	LIF	qCLIF	qCLIF
Synapse count	67k	313k	N//A	1M	256M	82k	82 k
Precision	6 bit	1bit	4 bit	7 bit	4 bit	4 bit	8 bit
Area (mm2)	1.99	0.15	10.08	1.72	430	1.86	3.71
Clock frequency	70KHz@ 0.52V	37.5MHz	20MHz	105MHz @ 0.5V	1KHz@ 1.05V	100MHz@ 1.1V	100MHz@ 1.1V
Energy per SOP (pJ)	1.5	8.4	N//A	3.8	26	8.7	17.9
Dataset	GSCD (4 Keywords)	GSCD (2 Keywords)	GSCD (10 Keywords)	TIMIT (4 Keywords)	TDIGIT (4 classes)	DVS Gesture (10 Classes)	DVS Gesture (10 Classes)
Accuracy (%)	91.8	94.6	90.2	94	90.8	73	90

- Proposed quantized cLIF digital implementation and First cRSNN implementation (Simulated).
- Although the neuron model is complex relatively low energy per spike consumption compared to literature.
- Careful hardware software codesign helped in network optimization.

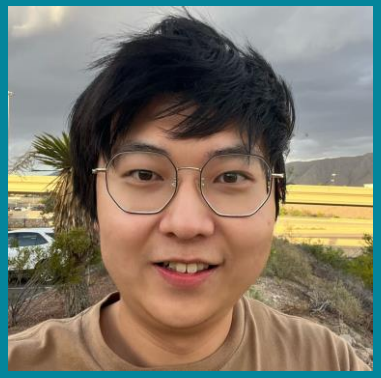
Limitations of the work / Future work

- The accumulator occupies a significant portion of the area.
  - Space-efficient alternatives, such as sparse accumulators or in-memory computing (e.g., memristor crossbar architectures), could be explored.

- All results were simulated on an open source 45 nm technology node.
  - Fabrication using a smaller technology node may further optimize performance.

- Deviation from true asynchronous nature of neuromorphic system, Synchronous behavior between apical and somatic compartment is expected in current design.

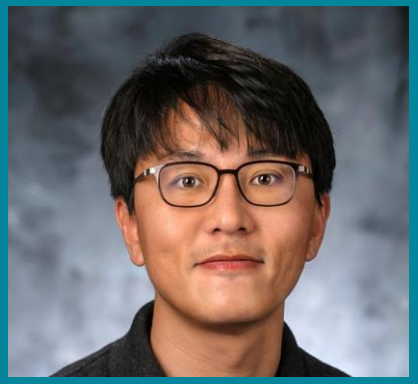
Thanks to my co authors



Yihao Wu

JaeBum Yoo

Dmitri Strukov



Bongjin Kim

This work is outcome of course taught by Prof. Bongjin Kim at UC Santa Barbara during Fall 2023.

All authors would like to thank discussions with Prof. Robert Legenstein, George Hutchinson and Tinish Bhattacharya.

I was funded by National Science Foundation BRAID award #2318152

Thank you
Any questions please feel free to reach out to me at

saisukruthbezugam@ieee.org
sbezugam@ucsb.edu

UC SANTA BARBARA